Diagonally implicit Runge–Kutta methods for 3D shallow water applications *

P.J. van der Houwen and B.P. Sommeijer

CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Received March 1999; revised June 1999

Communicated by K. Burrage

We construct A-stable and L-stable diagonally implicit Runge–Kutta methods of which the diagonal vector in the Butcher matrix has a minimal maximum norm. If the implicit Runge–Kutta relations are iteratively solved by means of the approximately factorized Newton process, then such iterated Runge–Kutta methods are suitable methods for integrating shallow water problems in the sense that the stability boundary is relatively large and that the usually quite fine vertical resolution of the discretized spatial domain is not involved in the stability condition.

Keywords: numerical analysis, shallow water problems, DIRK methods, stability

AMS subject classification: 65L06, 65L20, 65M12, 65M20

1. Introduction

If the hydrodynamical equations modelling shallow water flow or the transport equations modelling transport of pollutants in shallow water are discretized in space, then the resulting system of ordinary differential equations (ODEs) is highly stiff due to the relatively fine vertical resolution usually needed in the spatial discretization of the physical domain. This requires an implicit time integrator in order to cope with the stiff vertical terms. We shall focus on the family of A-stable and L-stable diagonally implicit Runge–Kutta (DIRK) methods. The implicit relations will be solved by modified Newton where the system matrix in the linear Newton system is replaced by an approximate factorization tuned to the shallow water application. The resulting approximately factorized Newton iteration method (AFN iteration) was analysed in [4,6]. In these papers it was shown that for three-dimensional problems a convergence condition has to be satisfied. Hence, the stability region of AFN iterated DIRK methods is the intersection of the AFN convergence region and the DIRK stability region. Thus, even if the underlying DIRK method is A-stable or L-stable, we have

* The investigations in this paper were supported by the NCF (Foundation Nationale Computer Faciliteiten). Work carried out under project MAS 1.2 – "Numerical Algorithms for Surface Water Quality Modelling".

to satisfy a stability condition on the time step Δt. This condition is of the form $\Delta t \leqslant \beta c \min\{\Delta x_1, \Delta x_2\}$, where Δx_1 and Δx_2 denote the horizontal mesh sizes in the spatial grid, c is a constant determined by the problem and the spatial discretization formulas, and β is the stability boundary determined by the AFN-DIRK method. Note that this stability condition does not contain the usually quite small vertical mesh size Δx_3. Furthermore, AFN iteration is highly parallel and vectorizable, so that it provides an attractive approach for integrating the large systems of shallow water ODEs on parallel supercomputers.

It turns out that the stability boundary is given by $\beta \approx 0.65\rho^{-1}(T)$, where $\rho(T)$ is the spectral radius of the (lower triangular) Butcher matrix T of the DIRK method. Hence, the maximal stable time step increases as $\rho(T)$ is smaller. This motivated us to look for A- and L-stable DIRK methods with minimal $\rho(T)$.

2. Shallow water applications

We briefly describe two shallow water applications, viz. the hydrodynamical equations modelling shallow water flow and the transport equations modelling transport of pollutants in shallow water.

2.1. Shallow water equations

The mathematical model describing the hydrodynamics in shallow water is defined by an initial-boundary value problem for the system of three-dimensional partial differential equations (PDEs)

$$\begin{aligned}
\frac{\partial u}{\partial t} &= L(u,v,w)u + \omega v - g\frac{\partial}{\partial x_1}\zeta + \tau_1,\\
\frac{\partial v}{\partial t} &= L(u,v,w)v - \omega u - g\frac{\partial}{\partial x_2}\zeta + \tau_2,\\
\frac{\partial}{\partial t}\zeta &= -\int_{-d}^{\zeta}\frac{\partial}{\partial x_1}u(t,x_1,x_2,x_3)\,\mathrm{d}x_3 - \int_{-d}^{\zeta}\frac{\partial}{\partial x_2}v(t,x_1,x_2,x_3)\,\mathrm{d}x_3,\\
L(u,v,w) &:= -\left(u\frac{\partial}{\partial x_1} + v\frac{\partial}{\partial x_2} + w\frac{\partial}{\partial x_3}\right) + \frac{\partial\varepsilon_1\partial}{\partial x_1^2} + \frac{\partial\varepsilon_2\partial}{\partial x_2^2} + \frac{\partial\varepsilon_3\partial}{\partial x_3^2}.
\end{aligned} \tag{2.1}$$

Here (u,v,w) represent the fluid velocities in the x_1, x_2, x_3 directions (x_3 denotes the vertical direction), w is defined by requiring that the velocity field is divergence free, ζ represents the water elevation, (τ_1, τ_2) external forcing terms, $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ diffusion coefficients in the x_1, x_2, x_3 directions, g denotes the acceleration due to gravity, ω the Coriolis parameter, and d the depth function.

The boundary is assumed to consist of coastal and ocean parts which are both assumed to be vertical. The boundary conditions prescribe the water elevation at ocean boundaries and require the velocity field to be normal to the coastal boundaries. Furthermore, at the sea surface and at sea bed, we impose the usual free surface and bottom friction conditions (see [13]).

2.2. *Shallow water transport equations*

The mathematical model describing transport processes of salinity, pollutants, etc., combined with their bio-chemical interactions, is defined by an initial-boundary value problem for the system of three-dimensional advection–diffusion–reaction equations

$$\frac{\partial c_\mu}{\partial t} = L(u,v,w)c_\mu + g_\mu(x_1,x_2,x_3,t,c_1,\ldots,c_m), \quad \mu = 1,\ldots,m, \tag{2.2}$$

where $L(u,v,w)$ is the same operator as defined in (2.1) with (u,v,w) denoting the velocities (assumed to be divergence free) and the c_μ represent the concentrations of the contaminants. The terms g_μ describe chemical reactions, emissions from sources, etc., and, therefore, depend on the concentrations. In shallow water applications, the g_μ are *nonstiff*. The mutual coupling of the equations in the system (2.2) is due to these functions g_μ. The boundary conditions are either of Dirichlet type or of Neumann type. Both the velocities (u,v,w) and diffusion coefficients $(\varepsilon_1,\varepsilon_2,\varepsilon_3)$ are assumed to be known in advance.

3. Time integration

First, the physical domain on which the PDEs are defined is replaced by a set of $N_1N_2N_3$ Cartesian grid points with mesh sizes Δx_1, Δx_2, and Δx_3. On this grid, the spatial derivative terms can be discretized by finite differences or finite element approximations. For the shallow water equations (2.1) this results in a semidiscrete, $N_1N_2(2N_3+1)$-dimensional initial value problem (briefly IVP) for the system of ODEs

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{U}}{\mathrm{d}t} &= A(\mathbf{U},\mathbf{V},\mathbf{W})\mathbf{U} + \omega\mathbf{V} - gB_1\mathbf{Z} + T_1(t), &\quad \mathbf{U}(t_0) &= \mathbf{U}_0,\\
\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t} &= A(\mathbf{U},\mathbf{V},\mathbf{W})\mathbf{V} - \omega\mathbf{U} - gB_2\mathbf{Z} + T_2(t), &\quad \mathbf{V}(t_0) &= \mathbf{V}_0,\\
\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}t} &= -C_1(\mathbf{Z})\mathbf{U} - C_2(\mathbf{Z})\mathbf{V}, &\quad \mathbf{Z}(t_0) &= \mathbf{Z}_0,
\end{aligned} \tag{3.1}$$

where $\mathbf{W}$ is defined by $\mathbf{W} = -E_1\mathbf{U} - E_2\mathbf{V}$. Here $\mathbf{U}$, $\mathbf{V}$ contain the horizontal velocity components at all $N_1N_2N_3$ grid points, $\mathbf{Z}$ contains the elevation at the N_1N_2 horizontal grid points, T_1 and T_2 represent the external forces at the grid points including the inhomogeneous parts of the boundary conditions, A, C_1, and C_2 are matrices depending on the velocity or elevation values, and B_1, B_2, E_1 and E_2 are constant matrices. The matrices A, C_1, and C_2 also take the coastal, free surface and bottom friction conditions into account.

In a similar way the transport equations (2.2) can be approximated by the system of ODEs (cf. [6])

$$\frac{\mathrm{d}\mathbf{C}(t)}{\mathrm{d}t} = \mathbf{F}\bigl(t,\mathbf{C}(t)\bigr) + \mathbf{G}\bigl(t,\mathbf{C}(t)\bigr), \qquad \mathbf{C}(t_0) = \mathbf{C}_0, \tag{3.2}$$

where $\mathbf{C}$ contains the m concentrations c_μ at all $N_1N_2N_3$ grid points, $\mathbf{F}(t, \mathbf{C}(t))$ contains the discretization of the operator L, and $\mathbf{G}(t, \mathbf{C}(t))$ contains the reaction terms and emissions from sources.

In the description of methods for integrating the IVPs (3.1) and (3.2) it is convenient to write them in the compact form

$$\begin{aligned} &\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y}), && \mathbf{f}(t, \mathbf{y}) := \mathbf{f}_1(t, \mathbf{y}) + \mathbf{f}_2(t, \mathbf{y}) + \mathbf{f}_3(t, \mathbf{y}) + \mathbf{f}_4(t, \mathbf{y}), \\ &\mathbf{y}(t_0) = \mathbf{y}_0, && \mathbf{y}, \mathbf{f}_k \in \mathbb{R}^N, \end{aligned} \tag{3.3}$$

where $\mathbf{y}$ contains the $N = N_1N_2(2N_3 + 1)$ components of $(\mathbf{U}, \mathbf{V}, \mathbf{Z})$ in the case (3.1) or the $N = mN_1N_2N_3$ concentrations $\mathbf{C}$ in the case (3.2). In both cases, $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{f}_3$ contain the spatial derivative terms with respect to the x_1, x_2 and x_3 directions, respectively, $\mathbf{f}_4$ represents the forcing terms (T_1, T_2) or the reaction/emission term $\mathbf{G}$. Hence, the function $\mathbf{f}_4$ is nonstiff. Furthermore, the function $\mathbf{f}_3$ corresponding with the vertical spatial direction is highly stiff, whereas the functions $\mathbf{f}_1$ and $\mathbf{f}_2$ corresponding with the horizontal spatial directions are less stiff. As a consequence, the spectral radius of the Jacobian matrix $\partial\mathbf{f}_3/\partial\mathbf{y}$ is much larger than the spectral radius of $\partial\mathbf{f}_1/\partial\mathbf{y}$ and $\partial\mathbf{f}_2/\partial\mathbf{y}$. The reason is that in shallow seas the grid size in the vertical direction is several orders of magnitude smaller than in the horizontal directions.

3.1. Implicit integration methods

In order to cope with the stiffness of the IVP (3.3), we shall use for the time discretization an *implicit* integration formula. Since the PDEs (2.1) and (2.2) are convection dominated, this implicit formula should at least be A-stable and preferably L-stable.

Our approach is based on the iterative solution of the implicit relations by the approximately factorized Newton (AFN) method analysed in [4,6]. This approach applies to a large class of implicit integration methods. In this paper, we shall consider methods that can be written in the form

$$\mathbf{R}_n(\Delta t, \mathbf{Y}_{n+1}) = \mathbf{0}, \qquad \mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t \boldsymbol{\Phi}_n(\Delta t, \mathbf{Y}_{n+1}), \tag{3.4}$$

where $\mathbf{R}_n$ and $\boldsymbol{\Phi}_n$ are given functions depending on Δt and $\mathbf{Y}_{n+1}$, and where $\mathbf{R}_n$ is such that its Jacobian satisfies the relation

$$\frac{\partial \mathbf{R}_n}{\partial \mathbf{Y}_{n+1}} = I - \Delta t\left(T \otimes \frac{\partial \mathbf{f}(t_n, \mathbf{y}_n)}{\partial \mathbf{y}_n}\right) + O\big((\Delta t)^2\big). \tag{3.5}$$

Here, Δt denotes the time step $t_{n+1} - t_n$ and

$$\mathbf{Y}_{n+1} = \big(\mathbf{y}_{n+1,1}^{\mathrm{T}}, \ldots, \mathbf{y}_{n+1,s}^{\mathrm{T}}\big)^{\mathrm{T}}$$

is the so-called stage vector with s components $\mathbf{y}_{n+1,i}$ representing approximations to the solution $\mathbf{y}(t)$ of (3.3) at the points $t_n + c_i\Delta t$, where the c_i, $i = 1, \ldots, s$, are given numbers. Furthermore, $\otimes$ denotes the Kronecker product (direct matrix product), T is an arbitrary *diagonal* or *lower triangular* s-by-s matrix with nonnegative diagonal entries, and I is the sN-by-sN identity matrix (in the sequel, identity matrices of arbitrary order will be denoted by I, but its order will always be clear from the context). The class of methods {(3.4), (3.5)} contains all linear multistep and all diagonally implicit Runge–Kutta methods, and many other useful integration methods.

In applying (3.4), the main effort goes into the iterative solution of the equation $\mathbf{R}_n(\Delta t, \mathbf{Y}_{n+1}) = \mathbf{0}$ (note that the formula for the step point value $\mathbf{y}_{n+1}$ is explicit). In the AFN method used in this paper, the matrix T, and in particular its spectral radius, plays a crucial role. For future reference, we present this matrix for a few methods from the literature which are suitable for the time integration of shallow water PDEs.

3.1.1. LM method

We start with a one-parameter family of zero-stable and L-stable linear multistep (LM) methods,

$$\mathbf{y}_{n+1} - b_0 \Delta t\, \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) = (2 - b_0)\mathbf{y}_n + (b_0 - 1)\mathbf{y}_{n-1}, \quad \frac{2}{3} \leqslant b_0 < 2. \tag{3.6}$$

Evidently, this method is of the form {(3.4), (3.5)} with $\mathbf{Y}_{n+1} = \mathbf{y}_{n+1}$ and $T = b_0$. For $\frac{2}{3} < b_0 < 2$, these methods have order of accuracy $p = 1$. For $b_0 = \frac{2}{3}$ the second-order accurate backward differentiation formula (BDF) is obtained. In practice, the BDF is the recommended method, but in section 3.3 we shall use (3.6) with $b_0 \neq \frac{2}{3}$ in order to illustrate the effect of $b_0 = \rho(T)$ on the stability of the iterated LM method.

3.1.2. Nørsett method

Another example of a second-order, L-stable method is provided by the DIRK method of Nørsett [9],

$$\begin{aligned}
&\mathbf{y}_{n+c} - c\Delta t\, \mathbf{f}(t_{n+c}, \mathbf{y}_{n+c}) = \mathbf{y}_n, \\
&\mathbf{y}_{n+1-c} - (1 - 2c)\Delta t\, \mathbf{f}(t_{n+c}, \mathbf{y}_{n+c}) - c\Delta t\, \mathbf{f}(t_{n+1-c}, \mathbf{y}_{n+1-c}) = \mathbf{y}_n, \\
&\quad c = 1 \pm \frac{1}{2}\sqrt{2}, \\
&\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{1}{2}\Delta t\, \mathbf{f}(t_{n+c}, \mathbf{y}_{n+c}) + \frac{1}{2}\Delta t\, \mathbf{f}(t_{n+1-c}, \mathbf{y}_{n+1-c}).
\end{aligned} \tag{3.7}$$

This method is of the form {(3.4), (3.5)} with

$$\mathbf{Y}_{n+1} = \begin{pmatrix} \mathbf{y}_{n+c} \\ \mathbf{y}_{n+1-c} \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} c \\ 1 - c \end{pmatrix}, \qquad T = \begin{pmatrix} c & 0 \\ 1 - 2c & c \end{pmatrix}.$$

3.1.3. DIM method

A method designed for integrating stiff IVPs on parallel computers [12] is the third-order, strongly A-stable, diagonally implicit (DIM) method,

$$\begin{aligned}
&\mathbf{y}_{n+c} - \frac{462}{660}\Delta t\,\mathbf{f}(t_{n+c}, \mathbf{y}_{n+c}) \\
&\quad = \mathbf{y}_n + \Delta t\left(\frac{441}{660}\mathbf{f}(t_{n+c-1}, \mathbf{y}_{n+c-1}) + \frac{483}{660}\mathbf{f}(t_n, \mathbf{y}_n)\right), \quad c = \frac{21}{10}, \\
&\mathbf{y}_{n+1} - \frac{143}{66}\Delta t\,\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) \\
&\quad = \mathbf{y}_n + \Delta t\left(-\frac{100}{66}\mathbf{f}(t_{n+c-1}, \mathbf{y}_{n+c-1}) + \frac{23}{66}\mathbf{f}(t_n, \mathbf{y}_n)\right),
\end{aligned} \qquad (3.8)$$

which takes the form {(3.4), (3.5)} by defining

$$\mathbf{Y}_{n+1} = \begin{pmatrix} \mathbf{y}_{n+c} \\ \mathbf{y}_{n+1} \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} c \\ 1 \end{pmatrix}, \qquad T = \frac{1}{660}\begin{pmatrix} 462 & 0 \\ 0 & 1430 \end{pmatrix}.$$

3.1.4. EBDF method

The third-order, L-stable method from the family of so-called extended backward differentiation formulas (EBDFs) proposed by Cash [3] is given by

$$\begin{aligned}
&\mathbf{u}_{n+1} - \frac{2}{3}\Delta t\,\mathbf{f}(t_{n+1}, \mathbf{u}_{n+1}) = \frac{4}{3}\mathbf{y}_n - \frac{1}{3}\mathbf{y}_{n-1}, \\
&\mathbf{u}_{n+2} - \frac{2}{3}\Delta t\,\mathbf{f}(t_{n+2}, \mathbf{u}_{n+2}) - \frac{4}{3}\mathbf{u}_{n+1} = -\frac{1}{3}\mathbf{y}_n, \\
&\mathbf{y}_{n+1} + \frac{4}{23}\Delta t\,\mathbf{f}(t_{n+2}, \mathbf{u}_{n+2}) - \frac{22}{23}\Delta t\,\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) = \frac{28}{23}\mathbf{y}_n - \frac{5}{23}\mathbf{y}_{n-1}.
\end{aligned} \qquad (3.9)$$

This method can be cast into the form {(3.4), (3.5)} with

$$\mathbf{Y}_{n+1} = \begin{pmatrix} \mathbf{u}_{n+1} \\ \mathbf{u}_{n+2} \\ \mathbf{y}_{n+1} \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \qquad T = \begin{pmatrix} \frac{2}{3} & 0 & 0 \\ \frac{8}{9} & \frac{2}{3} & 0 \\ 0 & -\frac{4}{23} & \frac{22}{23} \end{pmatrix}.$$

3.2. Iterative solution of the implicit relations

Our starting point for solving $\mathbf{Y}_{n+1}$ from (3.4) is the modified Newton process

$$\left(I - \Delta t T \otimes \frac{\partial \mathbf{f}(t_n, \mathbf{y}_n)}{\partial \mathbf{y}_n}\right)\left(\mathbf{Y}^j - \mathbf{Y}^{j-1}\right) = -\mathbf{R}_n\left(\Delta t, \mathbf{Y}^{j-1}\right), \quad j \geqslant 1, \qquad (3.10)$$

where $\mathbf{Y}^0$ is an initial approximation to be provided by some predictor formula. Due to the multidimensional structure of the underlying PDEs, a direct solution of the linear Newton systems in (3.10) is extremely costly. Therefore, we replace (3.10) by

$$\begin{aligned}
&\Pi\left(\mathbf{Y}^j - \mathbf{Y}^{j-1}\right) = -\mathbf{R}_n\left(\Delta t, \mathbf{Y}^{j-1}\right), \quad j \geqslant 1, \\
&\Pi := (I - \Delta t D \otimes J_1)(I - \Delta t D \otimes J_2)(I - \Delta t D \otimes J_3), \quad J_k \approx \frac{\partial \mathbf{f}_k(t_n, \mathbf{y}_n)}{\partial \mathbf{y}_n},
\end{aligned} \qquad (3.11)$$

where $D := \operatorname{diag}(d_1, \dots, d_s)$, d_i denoting the diagonal entries of T. The matrix Π is an approximate factorization of the matrix $I - \Delta t T \otimes \partial \mathbf{f}(t_n, \mathbf{y}_n)/\partial \mathbf{y}_n$. Since the function $\mathbf{f}_4$ in (3.3) is either independent of $\mathbf{y}$ or nonstiff, we have ignored its contribution to the approximate factorization. The iteration method (3.11) defines the *approximately factorized Newton (AFN) iteration* method.

Although AFN iteration requires the solution of three linear systems instead of the single linear Newton system in (3.10), the block structure of the system matrices $I - \Delta t D \otimes J_k$, $k = 1, 2, 3$, is such that solving these linear systems is not costly. Moreover, there is much scope for parallelism and vectorization. For example, for the transport problem (3.2), we have

$$I - \Delta t D \otimes J_k = \begin{pmatrix} I - \Delta t d_1 J_k & & \\ & \ddots & \\ & & I - \Delta t d_s J_k \end{pmatrix},$$

$$J_k = \begin{pmatrix} K_1 & & \\ & \ddots & \\ & & K_{M(k)} \end{pmatrix},$$

where the d_i are the diagonal entries of D, the matrices K_i are of dimension N_k, and $M(1) := mN_2N_3$, $M(2) := mN_1N_3$, $M(3) := mN_1N_2$. The matrices K_i corresponding with J_k only depend on the coordinate direction x_k, so that they can all be reduced to band matrices with small band width. Hence there is a considerable amount of intrinsic parallelism in the AFN algorithm (for details, we refer to [11]). Finally, note that intrinsic parallelism is due to the fact that we used the matrix D rather than the full matrix T in the definition of the matrix Π.

The AFN iteration process (3.11) was used in [6] and analysed in [4,7], where a number of convergence results have been derived for the test model

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = J_1\mathbf{y} + J_2\mathbf{y} + J_3\mathbf{y}, \quad J_k \text{ commuting.} \tag{3.12}$$

This test model is commonly used in the normal mode analysis of the stability of PDE methods. For our purpose, the following two results are important (eigenvalues and the spectral radius of matrices will be denoted by $\lambda(\cdot)$ and $\rho(\cdot)$, respectively).

Theorem 3.1. The iteration error of the AFN iteration process (3.11) applied to the model problem (3.12) satisfies the relations

$$\mathbf{Y}^j - \mathbf{Y}_{n+1} = \mathrm{O}((\Delta t)^{2j}), \quad j \geqslant 1, \qquad \text{if } T \text{ is diagonal,}$$

$$\mathbf{Y}^j - \mathbf{Y}_{n+1} = \begin{cases} \mathrm{O}((\Delta t)^j) & \text{for } 1 \leqslant j \leqslant s-1, \\ \mathrm{O}((\Delta t)^{2j+1-s}) & \text{for } j \geqslant s, \end{cases} \qquad \text{if } T \text{ is lower triangular.}$$

Theorem 3.2. The AFN iteration process (3.11) applied to (3.12) converges in the region

$$\mathbb{C} := \Big\{ \big(\lambda(J_1), \lambda(J_2), \lambda(J_3)\big) \colon \operatorname{Re}\big(\lambda(J_k)\big) \leqslant 0, \quad k = 1,2,3;$$

$$\big|\lambda(J_k)\big| \leqslant \frac{\gamma}{\Delta t\, \rho(T)}, \ k = 1,2 \Big\},$$

$$\gamma := \frac{1}{6}\Big(2 + \big(26 + 6\sqrt{33}\big)^{1/3} - 8\big(26 + 6\sqrt{33}\big)^{-1/3}\Big) \approx 0.65.$$

From this second result we immediately have the convergence condition

$$\Delta t \leqslant \frac{\gamma}{\rho(T) \max\{\rho(J_1), \rho(J_2)\}}, \quad \gamma \approx 0.65. \tag{3.13}$$

If the underlying integration method is A-stable (as we shall always assume), then this convergence condition also ensures the linear stability of the method. In such cases, the quantity $\beta_{\text{imag}} := \gamma \rho^{-1}(T)$ may be considered as the *imaginary stability boundary* of the AFN iterated method.

Let us apply (3.13) to convection dominated shallow water problems where $\rho(J_k) = O((\Delta x_k)^{-1})$.

Then, we obtain the convergence/stability condition

$$\Delta t \leqslant \beta_{\text{imag}} c \min\{\Delta x_1, \Delta x_2\}, \quad \beta_{\text{imag}} := \frac{\gamma}{\rho(T)}, \ \gamma \approx 0.65, \tag{3.14}$$

where c is a constant determined by the problem and the spatial discretization formulas. Note that this time step condition does not depend on the vertical resolution Δx_3.

The numerical experiments in section 3.3 and in [6,11] show that condition (3.13) is determining the maximal stepsize, rather than the accuracy of the numerical solution. This aspect will be analysed in [8], where we consider the effect of AFN iteration on the global error after a *finite* number of iterations, and in particular, the phenomenon of order reduction caused by the splitting errors.

3.3. Numerical illustration

In order to illustrate the effect of $\rho(T)$ on the stability of AFN iterated integration methods, we have applied the LM method (3.6) for various values of $\rho(T) = b_0$ to a three-dimensional transport test problem of the form (2.2) with two pollutants. A detailed description of this problem and its spatial discretization can be found in [11]. Here, we only mention that the vertical grid size $\Delta x_3 \approx 3.2$ m and that the horizontal grid sizes Δx_1 and Δx_2 varied from about 220 m to 110 m. The resulting ODE system (3.2) consists of 922000 equations. We integrated this system by the LM method (3.6) with $b_0 = \frac{3}{2}$, $b_0 = \frac{3}{4}$, and $b_0 = \frac{2}{3}$ from $t = 0$ to $t = 10$ h.

We first tried a fixed number of AFN iterations per step. For three iterations per step, table 1 presents the endpoint accuracies of the two pollutants expressed

Table 1
Values of scd for 3 AFN iterations per step at $t = 10$ h.

Δt in min.	$b_0 = \frac{3}{2}$	$b_0 = \frac{3}{4}$	$b_0 = \frac{2}{3}$
120	0.9/0.6	1.5/0.9	1.5/0.9
60	2.5/1.7	2.7/1.9	2.8/2.0
30	2.9/2.1	3.5/2.3	3.6/2.4

Table 2
Values of scd at $t = 10$ h and $t = 100$ h when iterated until convergence.

Δt in min.	$b_0 = \frac{3}{2}$	$b_0 = \frac{3}{4}$	$b_0 = \frac{2}{3}$	$b_0 = \frac{3}{2}$	$b_0 = \frac{3}{4}$	$b_0 = \frac{2}{3}$
15	*	*	*	*	*	*
12	*	1.9/2.0	4.6/4.1	*	*	2.8/1.4
10	*	4.1/3.6	4.7/4.4	*	4.7/3.8	5.4/4.5
6	*	4.3/3.9	5.0/4.7	*		
5	3.7/3.2	4.3/3.9	5.0/4.7	4.4/3.5		

in terms of significant correct digits (that is, the endpoint accuracy is written as scd $:= -\log(\text{absolute error})$). These results show that all three methods produce quite reasonable results for stepsizes that are extremely large with respect to the spatial discretization (the higher accuracies obtained as b_0 is smaller is due to a smaller error constant). However, these results are misleading. In fact, for larger stepsizes, a fixed-number-of-iterations strategy gradually becomes unstable as more steps are performed. The reason is that the resulting integration method is a (complicated) splitting method and, like many non-iterative time integration methods based on splitting, instabilities develop quite slowly, because the error amplification factors are often only slightly greater than 1. To illustrate this we continued the integration until negative scd-values were produced. For $\Delta t = 2$ h this happened at $t = 26$ h for $b_0 = \frac{3}{2}$ and at $t = 36$ h for $b_0 = \frac{3}{4}$ and $b_0 = \frac{2}{3}$.

A remedy for this unsatisfactory situation is a dynamic iteration strategy guaranteeing that the implicit relations in the underlying integration method are solved within a given tolerance. With such a strategy, we may rely on the stability condition (3.14). In the case of the LM method (3.6) with $b_0 = \frac{3}{2}$, $b_0 = \frac{3}{4}$ and $b_0 = \frac{2}{3}$, the imaginary stability boundary β_{imag} in (3.14) is approximately given by 0.43, 0.86 and 0.97, respectively. Thus, our stability theory predicts that the methods $\{(3.6), b_0 = \frac{3}{4}\}$ and $\{(3.6), b_0 = \frac{2}{3}\}$ will become stable for time steps of comparable size, whereas $\{(3.6), b_0 = \frac{3}{2}\}$ is predicted to require time steps that are twice as small. Table 2 (left hand part) presents the scd-values obtained at $t = 10$ h in the case of iteration until convergence (negative scd-values are indicated by $*$). The results in this table show that the methods with $b_0 = \frac{3}{2}$, $b_0 = \frac{3}{4}$ and $b_0 = \frac{2}{3}$ produce about 90% of the attainable accuracy for steps of 5, 10 and 12 min., respectively. This is in full agreement with the theory; that is, the maximal stable stepsize is proportional with $\rho^{-1}(T) = b_0^{-1}$. In order to show that these stepsizes are more or less "safe", we again continued the integration over a ten times longer integration interval. The right hand part of table 2 presents the results.

4. Stability functions with minimal $\rho(T)$ values

The preceding considerations strongly suggest the use of A-stable or L-stable integration methods of the form {(3.4), (3.5)} with minimal $\rho(T)$, so that β_{imag} is maximized. We shall focus on one-step methods. The linear stability of these methods is determined by their stability function. Here, we consider stability functions of the form

$$R(z) = \frac{P(z)}{Q(z)}, \quad Q(z) := \prod_{i=1}^{s}(1 - d_i z), \tag{4.1}$$

where P is a polynomial of degree $\leqslant s$ and where the d_i are the diagonal entries of the matrix T in (3.5). Stability functions of this type have been extensively analysed by Nørsett [9] and by Nørsett and Wolfbrandt [10]. Before discussing the special stability functions with *minimal* $\rho(T)$, we first summarize the relevant definitions and results from the literature.

4.1. Results from the literature

The most important class of methods with a stability function of the form (4.1) are the DIRK methods. These methods are of the form (3.4) and are generated by the Butcher tableau

$$\begin{array}{c|c} \mathbf{c} & T \\ \hline & \mathbf{b}^{\mathrm{T}} \end{array},$$

where $\mathbf{c} := T\mathbf{e}$ with $\mathbf{e}$ denoting the s-dimensional vector with unit entries, $\mathbf{b}$ is a given s-dimensional vector and where the Butcher matrix T is a lower triangular s-by-s matrix with nonnegative diagonal entries. If in (4.1) the d_i are all equal, then the corresponding DIRK method is often called *semi-explicit* or *singly-diagonal implicit* and *semi-implicit*, otherwise.

Nørsett and Wolfbrandt [10] showed that the maximum order of an s-stage semi-explicit DIRK is $s + 1$ and that for any semi-implicit DIRK of order s with stability function $R(z)$ of the form (4.1) where P is of degree s, the principal error is minimized if the d_i all equal d. This function can be expressed in terms of the Laguerre polynomials $L_s(x)$ and its derivatives $L_s^{(i)}(x)$, viz.

$$R(z) = \frac{(-1)^s \sum_{j=0}^{s} L_s^{(s-j)}(d^{-1})(dz)^j}{(1 - dz)^s}, \quad L_s(x) := \sum_{j=0}^{s} \frac{1}{j!}(-1)^j \binom{s}{j} x^j.$$

For our purposes, the stability functions with minimal d which are still A-stable or L-stable are of interest (recall that the shallow water applications (2.1) and (2.2) are convection dominated). Such stability functions will be called *optimal*. One can find in [2] and in [5, p. 103 ff.] the range of d-values for which A-stability or L-stability is preserved.

For future reference, we give a few of these optimal stability functions in explicit form. Since in shallow water problems we do not need more than second-order or third-order accuracy, we restrict our considerations to second-order and third-order consistent stability functions (R is called *consistent* of order p if $\mathrm{d}^i R(0)/\mathrm{d}z^i = 1$ for $i = 0, \ldots, p$). A comparison of the values of $\rho(T) = d$ and β_{imag} is given in table 3 in section 6.

The optimal, A-acceptable stability functions with $s = p - 1$, p and $p = 2, 3$ are given by

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}, \qquad \rho(T) = \frac{1}{2}, \qquad p = 2, \tag{4.2}$$

$$R(z) = \frac{1 + (1 - 2d)z + (\frac{1}{2} - 2d + d^2)z^2}{(1 - dz)^2}, \tag{4.3}$$

$$\rho(T) = d = \frac{1}{4} \Rightarrow p = 2, \qquad \rho(T) = d = \frac{1}{2} + \frac{1}{6}\sqrt{3} \Rightarrow p = 3,$$

$$R(z) = \frac{1 - \frac{1}{6}z^2 - \frac{1}{27}z^3}{(1 - \frac{1}{3}z)^3}, \qquad \rho(T) = \frac{1}{3}, \qquad p = 3, \tag{4.4}$$

and the optimal, L-acceptable stability functions with $s = p, p + 1$ and $p = 2, 3$ are given by

$$R(z) = \frac{1 + (1 - 2d)z}{(1 - dz)^2}, \qquad \rho(T) = d = 1 - \frac{1}{2}\sqrt{2} \Rightarrow p = 2. \tag{4.5}$$

$$R(z) = \frac{1 + (1 - 3d)z + (\frac{1}{2} - 3d + 3d^2)z^2}{(1 - dz)^3}, \tag{4.6}$$

$$\rho(T) = d = \frac{1}{12}\left(9 + 3\sqrt{3} - \sqrt{72 + 42\sqrt{3}}\right) \Rightarrow p = 2,$$

$$\rho(T) = d = 1 - \frac{1}{2}\sqrt{2}\big(\cos(\phi) - \sqrt{3}\sin(\phi)\big), \qquad \phi := \frac{1}{3}\arctan\left(\frac{1}{4}\sqrt{2}\right) \Rightarrow p = 3,$$

$$R(z) = \frac{1 + (1 - 4d)z + (\frac{1}{2} - 4d + 6d^2)z^2 + (\frac{1}{6} - 2d + 6d^2 - 4d^3)z^3}{(1 - dz)^4}, \tag{4.7}$$

$$\rho(T) = d \approx 0.223648 \Rightarrow p = 3.$$

4.2. Distinct diagonal entries d_i

The question arises whether we can decrease the value of $\rho(T)$ by using distinct diagonal entries in the matrix T. We consider the cases $s = 2$ and $s = 3$.

If $s = 2$ and if R has order of consistency $p \geqslant 2$, then

$$R(z) = \frac{1 + (1 - d_1 - d_2)z + \frac{1}{2}(1 - 2d_1 - 2d_2 + 2d_1 d_2)z^2}{(1 - d_1 z)(1 - d_2 z)},$$

so that the A-acceptability condition $|R(z)| \leqslant 1$ requires that the so-called E-polynomial $E(y) := |Q(\mathrm{i}y)|^2 - |P(\mathrm{i}y)|^2 \geqslant 0$ for all real values of y (cf., e.g., [5, p. 43]). Hence,

$$E(y) = \frac{1}{4}\left(4d_1^2 d_2^2 - (1 - 2d_1 - 2d_2 + 2d_1 d_2)^2\right)y^4 \geqslant 0$$

for all y. This leads to the condition $d_1 + d_2 \geqslant \frac{1}{2} \wedge 2(d_1 + d_2) - 4d_1 d_2 - 1 \leqslant 0$. By plotting in the (d_1, d_2)-plane the region where this condition is satisfied, it can be verified numerically that the value of $\rho(T) = \max\{d_1, d_2\}$ is minimized for $d_1 = d_2$. This yields the stability functions (4.3) and (4.5).

Likewise, we find for $s = 3$ and $p \geqslant 2$,

$$R(z) = \frac{1 + (1 - \sigma_1)z + \frac{1}{2}(1 - 2\sigma_1 + 2\sigma_2)z^2 + p_3 z^3}{(1 - d_1 z)(1 - d_2 z)(1 - d_3 z)}, \tag{4.8}$$

where $\sigma_1 := d_1 + d_2 + d_3$, $\sigma_2 := d_1 d_2 + d_1 d_3 + d_2 d_3$ and p_3 is a free parameter. This leads to the E-polynomial

$$E(y) = y^4\left(\sigma_4 y^2 - \frac{1}{4}N(d_1, d_2, d_3) + 2p_3(1 - \sigma_1)\right),$$
$$N(d_1, d_2, d_3) := (1 - 2\sigma_1 + 2\sigma_2)^2 - 4\sigma_5,$$

where $\sigma_4 := (d_1 d_2 d_3)^2 - p_3^2$ and $\sigma_5 := d_1^2 d_2^2 + d_1^2 d_3^2 + d_2^2 d_3^2$. Thus, the parameter p_3 should be such that $\sigma_4 \geqslant 0$ and $N(d_1, d_2, d_3) \leqslant 8p_3(1 - \sigma_1)$. Assuming that $\sigma_1 \leqslant 1$, we find for p_3 the inequalities

$$|p_3| \leqslant d_1 d_2 d_3 \quad \text{and} \quad p_3 \geqslant \frac{N(d_1, d_2, d_3)}{8(1 - d_1 - d_2 - d_3)}. \tag{4.9}$$

This leads to the condition

$$N(d_1, d_2, d_3) - 8d_1 d_2 d_3(1 - d_1 - d_2 - d_3) \leqslant 0. \tag{4.10}$$

Again, it can be verified that $\rho(T) = \max\{d_1, d_2, d_3\}$ with d_1, d_2 and d_3 satisfying (4.10) is minimized for $d_1 = d_2 = d_3$.

In the following, we restrict our considerations to stability functions with $d_1 = \cdots = d_s = d$.

4.3. Stability functions with still smaller $\rho(T)$

Among the stability functions listed in section 4.1, we miss the A-acceptable stability functions with $(s, p) = (3, 2), (4, 2), (4, 3)$ and the L-acceptable stability function with $(s,p)=(4,2)$. These functions possess still smaller $\rho(T)$-values and will be derived in the present section.

4.3.1. The A-acceptable case $(s,p)=(3,2)$

If $s=3$, then the stability function is of the form (4.8) for which the A-acceptability condition was shown to be given by (4.9) and (4.10). Setting $d_1=d_2=d_3=d$, condition (4.10) becomes

$$N(d,d,d)-8d^3(1-3d)=(6d-1)(2d-1)^3\leqslant 0,\quad d\leqslant\frac{1}{3}. \tag{4.11}$$

Hence, we have A-acceptability if $\frac{1}{6}\leqslant d\leqslant\frac{1}{3}$ and if p_3 satisfies (4.9) with $d_i=d$. This yields $p_3=\frac{1}{216}$ and $d=\frac{1}{6}$, leading to the A-acceptable, optimal stability function

$$R(z)=\frac{1+\frac{1}{2}z+\frac{1}{12}z^2+\frac{1}{216}z^3}{\left(1-\frac{1}{6}z\right)^3},\qquad \rho(T)=\frac{1}{6},\qquad p=2. \tag{4.12}$$

4.3.2. The A-acceptable cases $(s,p)=(4,2)$ *and* $(s,p)=(4,3)$

For $s=4$, $p\geqslant 2$ we have

$$R(z)=\frac{1+(1-4d)z+\left(\frac{1}{2}-4d+6d^2\right)z^2+p_3z^3+p_4z^4}{(1-dz)^4}, \tag{4.13}$$

where p_3 and p_4 are free parameters. The E-polynomial takes the form

$$\begin{aligned}E(y)&=y^4\left(\left(d^8-p_4^2\right)y^4+e_1y^2+e_2\right),\\ e_1&:=4d^6-p_3^2+12p_4d^2-8p_4d+p_4,\\ e_2&:=-30d^4+48d^3-22d^2+4(1-2p_3)d+2p_3-2p_4-\frac{1}{4}.\end{aligned} \tag{4.14}$$

Hence, we have A-acceptability if one of the following two conditions is satisfied:

$$|p_4|<d^4\quad\text{and}\quad g(d,p_3,p_4):=e_1^2-4e_2\left(d^8-p_4^2\right)\leqslant 0, \tag{4.15a}$$

$$|p_4|=d^4\quad\text{and}\quad e_1\geqslant 0\wedge e_2\geqslant 0. \tag{4.15b}$$

For a given value of d, (4.15a) determines a region in the (p_3,p_4)-plane. We verified numerically that this region converges to a single point in the (p_3,p_4)-plane as $d\downarrow\frac{1}{8}$. For $d=\frac{1}{8}$, this point is determined by the equation $g(\frac{1}{8},p_3,p_4)=0$, which is satisfied if $|p_4|-d^4=e_1=e_2=0$ (note that these equations imply that (4.15b) is fulfilled). From (4.14) it follows that $d=\frac{1}{8}$ is obtained if $p_3=\frac{1}{128}$ and $p_4=\frac{1}{4096}$, leading to the A-acceptable, optimal stability function

$$R(z)=\frac{1+\frac{1}{2}z+\frac{3}{32}z^2+\frac{1}{128}z^3+\frac{1}{4096}z^4}{\left(1-\frac{1}{8}z\right)^4},\qquad \rho(T)=\frac{1}{8},\qquad p=2. \tag{4.16}$$

Third-order consistent stability functions are obtained by setting $p_3 = \frac{1}{6} - 2d + 6d^2 - 4d^3$. Substitution into (4.15a) and plotting the region of admissible (d, p_4)-values reveals that d becomes smaller as $|p_4|$ approaches d^4. This leads us to use condition (4.15b). This straightforwardly yields that $d = \frac{1}{2} - \frac{1}{6}\sqrt{3}$, resulting in the A-acceptable, optimal stability function

$$\begin{aligned} R(z) &= \frac{1 + (1 - 4d)z + \left(\frac{1}{2} - 4d + 6d^2\right)z^2 + \left(\frac{1}{6} - 2d + 6d^2 - 4d^3\right)z^3 - d^4 z^4}{(1 - dz)^4}, \\ \rho(T) &= d = \frac{1}{2} - \frac{1}{6}\sqrt{3}, \qquad p = 3. \end{aligned} \tag{4.17}$$

4.3.3. The L-acceptable case $(s, p) = (4, 2)$

The corresponding stability function is given by (4.13) with $p_4 = 0$, so that the conditions (4.15) should be imposed with $p_4 = 0$. Since now only (4.15a) is relevant, we obtain the condition $g(d, p_3, 0) \leqslant 0$. The smallest value of d for which this inequality is fulfilled should satisfy the equations $g(d, p_3, 0) = 0$ and $\partial g(d, p_3, 0)/\partial p_3 = 0$, leading to $1 - 16d + 80d^2 - 128d^3 + 32d^4 = 0$, $p_3 = \frac{1}{8} - \frac{3}{2}d + 5d^2 - 4d^3$. The solution (d, p_3) with minimal d yields the L-acceptable, optimal stability function:

$$\begin{aligned} R(z) &= \frac{1 + (1 - 4d)z + \left(\frac{1}{2} - 4d + 6d^2\right)z^2 + \left(\frac{1}{8} - \frac{3}{2}d + 5d^2 - 4d^3\right)z^3}{(1 - dz)^4}, \\ \rho(T) &= d = 1 + \frac{1}{2}\sqrt{2} - \frac{1}{4}\sqrt{20 + 14\sqrt{2}}, \qquad p = 2. \end{aligned} \tag{4.18}$$

This d-value (≈ 0.13) is only slightly larger than the lower bound $d = \frac{1}{8}$ obtained in (4.16), but (4.18) is L-acceptable and (4.16) is only A-acceptable.

4.4. Conjecture

The stability functions (4.2), (4.3), (4.12) and (4.16) are second-order consistent and A-acceptable with the property that $|R(\mathrm{i}y)| = 1$ for all y. More generally, we have:

Theorem 4.1. Let

$$R(z) = \frac{P(z)}{Q(z)} = \frac{1 + p_1 z + p_2 z^2 + p_3 z^3 + \cdots + p_s z^s}{(1 - dz)^s}, \quad P \not\equiv Q, \ s \geqslant 1. \tag{4.19}$$

Then,

(a) $|R(\mathrm{i}y)| = 1$ for all real y and $d \geqslant 0$, if and only if $p_j = \binom{s}{j} d^j$.

(b) The A-acceptable stability function $R^*(z)$ defined by $\{(4.19),\ p_j := \binom{s}{j} d^j,\ d \geqslant 0\}$ is second-order consistent for $d = 1/(2s)$.

Proof. Evidently,

$$P(\mathrm{i}y) = \left(1 - p_2y^2 + p_4y^4 - \cdots\right) + \mathrm{i}y\left(p_1 - p_3y^2 + p_5y^4 - \cdots\right),$$
$$Q(\mathrm{i}y) = (1 - \mathrm{i}dy)^s = \left(1 - q_2y^2 + q_4y^4 - \cdots\right) + \mathrm{i}y\left(q_1 - q_3y^2 + q_5y^4 - \cdots\right),$$
$$q_j := \binom{s}{j}(-d)^j.$$

Hence, $p_j := \binom{s}{j}d^j$ implies that $|P(\mathrm{i}y)| = |Q(\mathrm{i}y)|$, or equivalently $|R(\mathrm{i}y)| = 1$, for all real y and $d \geqslant 0$, and all positive integers s. Conversely, if $|P(\mathrm{i}y)| = |Q(\mathrm{i}y)|$ for all y, then because $P \not\equiv Q$ it follows that $p_j = q_j$ for j even and $p_j = -q_j$ for j odd, so that $p_j := \binom{s}{j}d^j$. This proves part (a).

From part (a) it follows that the functions $\{(4.19),\ p_j := \binom{s}{j}d^j\}$ are A-acceptable for all $d \geqslant 0$. Furthermore, from the definition of consistency (cf. section 4.1) it follows that these functions are second-order consistent if $P'(0) = p_1 = 1 - sd$, $P''(0) = 2p_2 = 1 - 2sd + s(s-1)d^2$, $s \geqslant 1$. On substitution of

$$p_1 = \binom{s}{1}d = sd \quad \text{and} \quad p_2 = \binom{s}{2}d^2 = \frac{1}{2}s(s-1)d^2,$$

we find $d = 1/(2s)$, proving part (b). □

The question now arises whether the functions $R^*(z)$ defined in this theorem are optimal, that is, do there exist A-acceptable and second-order consistent stability functions of the form (4.19) with still *smaller* values for d? In fact, the functions (4.2), (4.3), (4.12) and (4.16), which are special cases of $R^*(z)$ for $s \leqslant 4$, show that $R^*(z)$ is optimal for $s \leqslant 4$. Since the functions $R^*(z)$ possess the weakest possible form of A-acceptability for all s, we conjecture that $R^*(z)$ is also optimal for all $s \geqslant 5$.

5. Construction of DIRK methods with prescribed stability function

Having available a number of optimal stability functions, we can construct corresponding families of one-step methods like the methods of Runge–Kutta, Rosenbrock, Obreschkov, etc. In this paper, we consider DIRK methods. Any DIRK method is second-order accurate if its stability function $R(z)$ is second-order consistent and third-order accurate if $R(z)$ is third-order consistent and if T satisfies

$$3\mathbf{b}^{\mathrm{T}}(T\mathbf{e})^2 = 1. \qquad (5.1)$$

For a given optimal stability function $R(z)$, we shall construct DIRK methods with a minimal number of stages s. Furthermore, in view of the shallow water applications we have in mind, we shall try to construct DIRK methods with a storage saving Butcher tableau. DIRK methods with an optimal stability functions will also be called *optimal*.

5.1. L-stable methods with $s = 2$

For two-stage L-stable methods, T, $\mathbf{b}$ and the stability function are given by

$$T = \begin{pmatrix} d & 0 \\ a & d \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} a \\ d \end{pmatrix}, \qquad R(z) = \frac{1 + (a-d)z}{(1-dz)^2}. \tag{5.2a}$$

Hence, we can only identify this function with the function (4.5). This yields

$$a = \frac{1}{2}\sqrt{2}, \qquad d = 1 - \frac{1}{2}\sqrt{2}. \tag{5.2b}$$

The method {(5.2a), (5.2b)} defines a second-order accurate, L-stable optimal method.

5.2. A-stable methods with $s = 2$

The arrays T, $\mathbf{b}$ and the stability function are given by

$$T = \begin{pmatrix} d & 0 \\ a & d \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} b \\ c \end{pmatrix},$$
$$R(z) = \frac{1 + (b + c - 2d)z + (ac - bd - cd + d^2)z^2}{(1-dz)^2}. \tag{5.3a}$$

Identification with $\{(4.3), d = \frac{1}{4}\}$ yields a family of second-order accurate, A-stable optimal methods

$$a = \frac{1}{4c}, \qquad b = 1 - c, \qquad d = \frac{1}{4} \tag{5.3b}$$

with free parameter c (note that in this case there is no storage saving value for free parameter). Likewise, we can identify $R(z)$ with $\{(4.3), d = \frac{1}{2} + \frac{1}{6}\sqrt{3}\}$, to obtain $a = -(1/6c)\sqrt{3}$ and $b = 1 - c$, again with free parameter c. In order to make the method third-order accurate, we have to impose condition (5.1). This yields $c = \frac{1}{2}$. Hence, (5.3a) with

$$a = -\frac{1}{3}\sqrt{3}, \qquad b = c = \frac{1}{2}, \qquad d = \frac{1}{2} + \frac{1}{6}\sqrt{3} \tag{5.3c}$$

defines a third-order accurate, strongly A-stable, optimal method. In fact, this method is identical with one of one third-order DIRK methods of Nørsett [9].

5.3. L-stable methods with $s = 3$

If we allow an $s = 3$ method to have three implicit stages, then

$$T = \begin{pmatrix} d & 0 & 0 \\ a & d & 0 \\ b & c & d \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} b \\ c \\ d \end{pmatrix},$$
$$R(z) = \frac{1 + (b + c - 2d)z + (ac - bd - cd + d^2)z^2}{(1-dz)^3} \tag{5.4a}$$

so that we can identify $R(z)$ with (4.6). This leads to a one-parameter family of second-order accurate, L-stable, optimal DIRK methods define by

$$a=\frac{1-4d+2d^2}{2c},\qquad b=1-c-d,\qquad d=\frac{1}{12}\left(9+3\sqrt{3}-\sqrt{72+42\sqrt{3}}\right).\tag{5.4b}$$

The choice $c=1-d$ (i.e., $b=0$) saves storage in an actual implementation.

For third-order accuracy, we impose condition (5.1). This determines the free parameter c and leads to a third-order accurate, L-stable, optimal method defined by

$$\begin{aligned} &a=\frac{1-4d+2d^2}{2c},\qquad b=1-c-d,\qquad c=\frac{3(1-4d+2d^2)^2}{4(1-6d+9d^2-3d^3)},\\ &d=1-\frac{1}{2}\sqrt{2}\left(\cos(\phi)-\sqrt{3}\sin(\phi)\right),\qquad \phi:=\frac{1}{3}\arctan\left(\frac{1}{4}\sqrt{2}\right). \end{aligned}\tag{5.4c}$$

5.4. A-stable methods with $s=3$

To construct methods with the stability functions (4.4) and (4.12), we consider

$$T=\begin{pmatrix} d & 0 & 0\\ a & d & 0\\ b & c & d\end{pmatrix},\qquad \mathbf{b}=\begin{pmatrix} e\\ f\\ g\end{pmatrix},\qquad R(z)=\frac{1+r_1z+r_2z^2+r_3z^3}{(1-dz)^3},\tag{5.5a}$$

$$r_1=\alpha_1-3d,\quad r_2=\alpha_2-2\alpha_1d+3d^2,\quad r_3=\alpha_3-\alpha_2d+\alpha_1d^2-d^3,$$

$$\alpha_1:=e+f+g,\qquad \alpha_2:=af+g(b+c),\qquad \alpha_3:=acg.\tag{5.6a}$$

We write the numerator of the prescribed stability function in the form $P(z)=1+p_1z+p_2z^2+p_3z^3$. Identification with the numerator of $R(z)$ in (5.5a) leads to

$$\alpha_1=p_1+3d,\qquad \alpha_2=p_2+2p_1d+3d^2,\qquad \alpha_3=p_3+p_2d+p_1d^2+d^3,\tag{5.6b}$$

so that the quantities α_i are completely determined by the prescribed stability function. Thus, if this stability function is second-order consistent, then {(5.5a), (5.6)} defines a three-parameter family of second-order methods whose stability function is given in (5.5a).

If the given stability function is third-order consistent, then we can make the method (5.5a) third-order accurate by imposing the condition (5.1). Using (5.6a), this additional condition becomes

$$fa^2+g(b+c)^2=\alpha_4,\qquad \alpha_4:=\frac{1}{3}-\alpha_1d^2-2\alpha_2d.\tag{5.7}$$

Solving (5.6a) and (5.7) for $e=0$, we obtain

$$\begin{aligned} &b=\frac{\alpha_2-af}{g}-c,\qquad c=\frac{\alpha_3}{ag},\qquad e=0,\\ &f=\frac{\alpha_1\alpha_4-\alpha_2^2}{\alpha_1a^2-2\alpha_2a+\alpha_4},\qquad g=\alpha_1-f. \end{aligned}\tag{5.8}$$

The formulas {(5.5a), (5.8)} define a one-parameter family of third-order methods with stability function $R(z)$ as is given in (5.5a). We are now ready to identify $R(z)$ with (4.12) and (4.4).

In the case (4.12) we have $\alpha_1 = 1$, $\alpha_2 = \frac{1}{3}$ and $\alpha_3 = \frac{1}{27}$. We propose to choose $b = e = f = 0$ yielding a storage saving method. Then, it follows from (5.6a) that

$$a = \frac{1}{9}, \qquad c = \frac{1}{3}, \qquad d = \frac{1}{6}, \qquad g = 1, \qquad b = e = f = 0. \tag{5.5b}$$

The formulas {(5.5a), (5.5b)} define a second-order accurate, A-stable, optimal method.

For the third-order consistent stability function (4.4), we have $\alpha_1 = 1$, $\alpha_2 = \frac{1}{6}$ and $\alpha_3 = -\frac{1}{18}$. On substitution into (5.8) we obtain a family of third-order accurate, A-stable, optimal methods with free parameter a. For example, $a = -\frac{1}{3}$ yields the method

$$\begin{aligned} &a = -\frac{1}{3}, \quad b = \frac{1}{9}, \quad c = \frac{2}{9}, \quad d = \frac{1}{3}, \\ &e = 0, \quad f = \frac{1}{4}, \quad g = \frac{3}{4}. \end{aligned} \tag{5.5c}$$

5.5. *L-stable methods with $s = 4$*

To construct methods with the stability functions (4.7) and (4.18) we consider

$$T = \begin{pmatrix} d & 0 & 0 & 0 \\ a & d & 0 & 0 \\ b & c & d & 0 \\ e & f & g & d \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} e \\ f \\ g \\ d \end{pmatrix},$$

$$R(z) = \frac{1 + r_1 z + r_2 z^2 + r_3 z^3}{(1 - dz)^4}, \tag{5.9a}$$

$$r_1 = \alpha_1 - 3d, \quad r_2 = \alpha_2 - 2\alpha_1 d + 3d^2, \quad r_3 = \alpha_3 - \alpha_2 d + \alpha_1 d^2 - d^3,$$

where α_1, α_2 and α_3 are defined in (5.6a). Writing the numerator of the prescribed stability functions in the form $P(z) = 1 + p_1 z + p_2 z^2 + p_3 z^3$ and identifying P with the numerator of $R(z)$ in (5.9a) again leads to the equations (5.6b). Thus, if the prescribed stability function is second-order consistent, then {(5.9a), (5.6)} defines a three-parameter family of second-order methods whose stability function is given in (5.9a). If the prescribed stability function is third-order consistent, then we achieve third-order accuracy by imposing the condition (5.1), i.e.,

$$fa^2 + g(b + c)^2 = \alpha_4, \quad \alpha_4 := \frac{1}{3} - \alpha_1 d^2 - 2\alpha_2 d - d(\alpha_1 + d)^2. \tag{5.10}$$

Solving (5.6a) and (5.10) for $e = 0$ yields the relations (5.8), so that the formulas $\{(5.9\text{a}), (5.8)\}$ define a one-parameter family of third-order methods with the stability function $R(z)$ of (5.9a).

In the case (4.18) we have $\alpha_1 = 1-d$, $\alpha_2 = \frac{1}{2}-2d+d^2$ and $\alpha_3 = \frac{1}{8}-d+2d^2+d^3$, so that solving (5.6a) with $b = e = f = 0$ leads to

$$a = \frac{\frac{1}{8} - d + 2d^2 + d^3}{\frac{1}{2} - 2d + d^2}, \qquad c = \frac{\frac{1}{2} - 2d + d^2}{1-d}, \qquad d = 1 + \frac{1}{2}\sqrt{2} - \frac{1}{4}\sqrt{20 + 14\sqrt{2}},$$
$$g = 1 - d, \qquad b = e = f = 0. \tag{5.9b}$$

The method $\{(5.9\text{a}), (5.9\text{b})\}$ defines a second-order accurate, L-stable, optimal method.

In the case (4.7), where $d \approx 0.223648$, $\alpha_1 = 1 - d$, $\alpha_2 = \frac{1}{2} - 2d + d^2$ and $\alpha_3 = \frac{1}{6} - \frac{3}{2}d + 3d^2 - d^3$, we find on substitution into (5.8) a family of third-order accurate, L-stable, optimal methods with free parameter a. For example, $a = \frac{1}{2}$ and $d = \frac{17}{76}$ (which is only slightly greater than $d = 0.223648$) yields

$$a = \frac{1}{2}, \quad b = \frac{12589505881}{70677472392}, \quad c = -\frac{6039885655}{70677472392}, \quad d = \frac{17}{76},$$
$$e = 0, \quad f = \frac{11552}{153145}, \quad g = \frac{8157603}{11639020}. \tag{5.9c}$$

The method $\{(5.9\text{a}), (5.9\text{c})\}$ defines a third-order accurate, L-stable, optimal method.

5.6. *A-stable methods with $s = 4$*

Finally, we construct methods with the stability functions (4.16) and (4.17). Consider the method

$$T = \begin{pmatrix} d & 0 & 0 & 0 \\ a & d & 0 & 0 \\ 0 & b & d & 0 \\ 0 & 0 & c & d \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ e \\ 1-e \end{pmatrix},$$
$$R(z) = \frac{1 + r_1 z + r_2 z^2 + r_3 z^3 + r_4 z^4}{(1 - dz)^4}, \tag{5.11a}$$
$$r_1 := 1 - 4d, \qquad r_2 := be + c(1-e) - 3d + 6d^2,$$
$$r_3 := bc(1-e) + abe - 2(be + c(1-e))d + 3d^2 - 4d^3,$$
$$r_4 := abc(1-e) - (abe + bc(1-e))d + (be + c(1-e))d^2 - d^3 + d^4.$$

Identification of $R(z)$ with (4.16) and setting $e = 0$ yields a second-order accurate, A-stable, optimal method of the form (5.11a) with

$$a = \frac{1}{16}, \qquad b = \frac{1}{6}, \qquad c = \frac{3}{8}, \qquad d = \frac{1}{8}, \qquad e = 0. \tag{5.11b}$$

Table 3
Characteristics of implicit integration methods.

Order	Method	s	Stability function	Stability	$\rho(T) \approx$	$\beta_{\text{imag}} \approx$
2	BDF {(3.6), $b_0 = \frac{2}{3}$}	1	–	L-stable	0.67	0.97
	Nørsett (3.7)	2	(4.5)	L-stable	0.29	2.21
	{(5.2a), (5.2b)}	2	(4.5)	L-stable	0.29	2.21
	{(5.3a), (5.3b)}	2	(4.3)	A-stable	0.25	2.59
	{(5.4a), (5.4b)}	3	(4.6)	L-stable	0.18	3.59
	{(5.5a), (5.5b)}	3	(4.12)	A-stable	0.17	3.88
	{(5.9a), (5.9b)}	4	(4.18)	L-stable	0.13	4.98
	{(5.11a), (5.11b)}	4	(4.16)	A-stable	0.13	5.18
3	DIM (3.8)	2	–	strongly A-stable	2.17	0.29
	Cash (3.9)	3	–	L-stable	0.96	0.67
	Nørsett {(5.3a), (5.3c)}	2	(4.3)	strongly A-stable	0.79	0.82
	{(5.4a), (5.4c)}	3	(4.6)	L-stable	0.44	1.48
	{(5.5a), (5.5c)}	3	(4.4)	A-stable	0.33	1.94
	{(5.9a), (5.9c)}	4	(4.7)	L-stable	0.22	2.89
	{(5.11a), (5.11c), (5.11d)}	4	(4.17)	A-stable	0.21	3.06

Similarly, identification of $R(z)$ with (4.17) yields a third-order accurate, A-stable, optimal method of the form (5.11a) with

$$\begin{aligned} a &= \frac{-c}{1 - 4c\sqrt{3} + 12c^2}, & b &= \frac{\sqrt{3} - 9c + 6c^2\sqrt{3}}{3(1 - 4c\sqrt{3} + 12c^2)}, \\ d &= \frac{1}{2} - \frac{1}{6}\sqrt{3}, & e &= \frac{1 - 4c\sqrt{3} + 12c^2}{2 - 4c\sqrt{3} + 12c^2}, \end{aligned} \tag{5.11c}$$

where c is a real zero of the equation

$$432c^5 - 360c^4\sqrt{3} + 18(25 + 2\sqrt{3})c^3 - 12(3 + 8\sqrt{3})c^2 + 24c - \sqrt{3} = 0. \tag{5.11d}$$

This equation possesses one real root, which is approximately given by $c \approx 0.545717$.

6. Summary of results

In this paper, we considered A-stable and L-stable DIRK methods of which the diagonal vector in the Butcher matrix has a minimal maximum norm. If the implicit relations are iteratively solved by means of the approximately factorized Newton process (3.11), then such DIRK methods possess stability properties which enable us to solve shallow water problems with relatively large time steps. Table 3 lists the main characteristics of all methods discussed in this paper. In particular, this table presents the value of $\rho(T)$ and the resulting imaginary stability boundary β_{imag} occurring in the stability condition (3.14). For the DIRK methods in this table, we see that for fixed order, the value of β_{imag} increases strongly with the number of stages s, the order 2 values being substantially larger than the order 3 values. Also note that for given order p and number of stages s, β_{imag} decreases only slightly when changing from A-stability to L-stability, so that the maximal stable stepsize is mainly determined by p and s.

Thus, we may conclude that A- and L-stable DIRK methods with minimal $\rho(T)$ are attractive candidates for integrating shallow water problems. In particular, the higher-stage methods with their relatively small $\rho(T)$, and therefore large imaginary stability boundaries, are suitable on parallel computer systems (we recall that theorem 4.1 implies the existence of second-order, A-stable methods with $\rho(T) = (2s)^{-1}$). Since the AFN process (3.11) is fully parallel over the stages (see the discussion of (3.11) in section 3.2), the number of stages is mainly determined by the number of processors available. The actual application of the methods proposed in this paper is subject of future research.

Finally, as observed by one of the referees of this paper, in the case of *parabolic* problems, where we need only A_0-stability or L_0-stability, some of the results of Bales et al. [1] are relevant. In fact, since for parabolic problems one is often satisfied with order 2 or 3 accuracy, it would be of interest to investigate whether the $\rho(T)$-values of table 3 can be improved.

Acknowledgement

The authors wish to thank the referees for their constructive remarks.

References

[1] L.A. Bales, O.A. Karakashian and S.M. Serbin, On the A_0-acceptability of rational approximations to the exponentional function with only real poles, BIT 28 (1988) 70–79.

[2] K. Burrage, A special family of Runge–Kutta methods for solving stiff differential equations, BIT 18 (1978) 22–41.

[3] J.R. Cash, On the integration of stiff systems of O.D.E.s using extended backward differentiation formulae, Numer. Math. 34 (1980) 235–246.

[4] C. Eichler-Liebenow, P.J. van der Houwen and B.P. Sommeijer, Analysis of approximate factorization in iteration methods, Appl. Numer. Math. 28 (1998) 245–258.

[5] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential–Algebraic Problems* (Springer, Berlin, 1991).

[6] P.J. van der Houwen, B.P. Sommeijer and J. Kok, The iterative solution of fully implicit discretizations of three-dimensional transport models, Appl. Numer. Math. 25 (1997) 243–256.

[7] P.J. van der Houwen and B.P. Sommeijer, Factorization in block-triangularly implicit methods for shallow water applications, Report MAS R9906, CWI, Amsterdam (1999, submitted for publication).

[8] P.J. van der Houwen and B.P. Sommeijer, Order reduction effects in approximate factorization (1999, in preparation).

[9] S.P. Nørsett, Semi-explicit Runge–Kutta methods, Report Mathematics and Computation No. 6/74, Department of Mathematics, University of Trondheim (1974).

[10] S.P. Nørsett and A. Wolfbrandt, Attainable order of rational approximations to the exponential function with only real poles, BIT 17 (1977) 200–208.

[11] B.P. Sommeijer, The iterative solution of fully implicit discretizations of three-dimensional transport models, in: *Parallel Computational Fluid Dynamics – Development and Applications of Parallel*

Technology, eds. C.A. Lin, A. Ecer, P. Fox, J. Periaux and N. Satofuka, *Proceedings of the 10th Int. Conf. on Parallel CFD*, May 1998, Hsinchu, Taiwan (Elsevier, Amsterdam, 1999) pp. 67–74.

[12] B.P. Sommeijer, W. Couzy and P.J. van der Houwen, A-stable parallel block methods for ordinary and integro-differential equations, Appl. Numer. Math. 9 (1992) 267–281.

[13] C.B. Vreugdenhil, *Numerical Methods for Shallow-Water Flow* (Kluwer, Dordrecht, 1994).