The Iterative Solution of Fully Implicit Discretizations of Three-Dimensional Transport Models

P.J. van der Houwen, B.P. Sommeijer, J. Kok
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

We investigate the use of approximate factorization and diagonalizing techniques for solving iteratively fully implicit numerical models of three-dimensional transport-chemistry problems. In particular, we investigate various possibilities that can take advantage of the parallelization and vectorization facilities offered by parallel vector computers.

AMS Subject Classification (1991): 65M06, 65M12, 65M20
CR Subject Classification (1991): G.1.7, G.1.8
Keywords and Phrases: Transport models, shallow water, 3D, implicit, approximate factorization, HPCN
Note: The investigations reported in this paper were partly supported by the Dutch HPCN Program.

1. Introduction

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 3D advection-diffusion-reaction equations

$$\text{(1.1a)} \quad \frac{\partial c_\mu}{\partial t} = L(u,v,w)c_\mu + g_\mu(x,y,z,t,c_1,...,c_m), \quad \mu = 1, \ldots, m,$$

$$\text{(1.1b)} \quad L(u,v,w)c_\mu := -u\frac{\partial}{\partial x}c_\mu - v\frac{\partial}{\partial y}c_\mu - w\frac{\partial}{\partial z}c_\mu + \frac{\partial}{\partial x}\left(\varepsilon_x\frac{\partial c_\mu}{\partial x}\right) + \frac{\partial}{\partial y}\left(\varepsilon_y\frac{\partial c_\mu}{\partial y}\right) + \frac{\partial}{\partial z}\left(\varepsilon_z\frac{\partial c_\mu}{\partial z}\right).$$

Here, the various quantities are defined as follows:

c_μ	concentrations of the contaminants,
u, v, w	local fluid velocities in x, y, z directions (assumed to be divergence free),
$\varepsilon_x, \varepsilon_y, \varepsilon_z$	diffusion coefficients in x, y, z directions,
g_μ	reaction terms (e.g. chemical interactions) and emissions from sources.

The boundary conditions are assumed to be of Dirichlet type and the velocities u, v, w, and diffusion coefficients $\varepsilon_x, \varepsilon_y, \varepsilon_z$ are assumed to be known in advance. The terms g_μ describe chemical reactions, emissions from sources, etc., and therefore depend on the concentrations. The mutual coupling of the equations in the system (1.1) is due to the functions g_μ.

Along the lines described in [4] and [6], we replace in (1.1) the physical domain by a set of N Cartesian grid points with mesh sizes Δx, Δy, and Δz, the advection terms by upwind-biased $\kappa = 1/3$ discretizations (see [5]), and the diffusion terms by symmetric three-point discretizations. This results in a semidiscrete, mN-dimensional initial value problem (IVP)

(1.2a) $$\frac{d\mathbf{C}(t)}{dt} = \mathbf{F}(t,\mathbf{C}(t)) := \mathbf{H}(t,\mathbf{C}(t)) + \mathbf{G}(t,\mathbf{C}(t)), \quad \mathbf{C}(t_0) = \mathbf{C}_0.$$

Here, $\mathbf{C}$ contains the m concentrations c_μ at all N grid points, $\mathbf{C}_0$ defines the initial values, $\mathbf{H}(t,\mathbf{C}(t))$ represents the advection-diffusion terms, and $\mathbf{G}(t,\mathbf{C}(t))$ contains the reaction terms and emissions from sources. $\mathbf{H}(t,\mathbf{C})$ is linear in $\mathbf{C}$ with a block-diagonal matrix of coefficients. The m diagonal blocks in this matrix are all the same (assuming the same boundary conditions and diffusion coefficients for the various species); furthermore, these blocks have dimension N and contain (at most) only 13 nonzero diagonals (see also Section 2.1). In fact, from the definition of the operator L it follows that $\mathbf{H}(t,\mathbf{C})$ can be split into three terms corresponding with the derivatives with respect to x, y and z, respectively (dimensional splitting). Hence, $\mathbf{H}(t,\mathbf{C})$ can be written as

(1.2b) $$\mathbf{H}(t,\mathbf{C}) = (X(t) + Y(t) + Z(t))\mathbf{C},$$

where the matrices X, Y and Z are again block-diagonal, each with m identical blocks containing 5 nonzero diagonals.

In general, $\mathbf{G}(t,\mathbf{C})$ is nonlinear in $\mathbf{C}$, but at each grid point, it only depends on the m concentrations c_μ at that particular grid point. The reaction term $\mathbf{G}$ may be considered as nonstiff, because the chemistry in shallow water transport problems usually has large time constants. However, the advection-diffusion term $\mathbf{H}$ introduces stiffness, due to the relatively small vertical mesh size Δz.

In order to cope with the stiffness of the IVP (1.2), we shall use for the time discretization an *implicit* discretization formula. Since transport problems usually are advection dominated, this implicit formula should at least be A-stable and preferably L-stable. The choice of such a highly stable time discretization formula now depends on the required order of time accuracy. If second-order accuracy suffices, one may use in the first integration step the (selfstarting) first-order, L-stable backward Euler method, and in all subsequent steps, the second-order, L-stable backward differentiation formula (BDF). Thus,

(1.3a) $$\mathbf{R}(t_{n+1},\mathbf{C}_{n+1}) = \mathbf{0},$$

where

(1.3b)
$$\mathbf{R}(t_1,\mathbf{C}) := \mathbf{C} - \Delta t\, \mathbf{F}(t_1,\mathbf{C}) - \mathbf{C}_0, \quad n = 0,$$
$$\mathbf{R}(t_{n+1},\mathbf{C}) := \mathbf{C} - \tfrac{2}{3}\Delta t\, \mathbf{F}(t_{n+1},\mathbf{C}) - \tfrac{1}{3}[4\mathbf{C}_n - \mathbf{C}_{n-1}], \quad n \geq 1.$$

If third-order time accuracy is desired, we may use the (selfstarting) L-stable, two-stage Radau IIA discretization

(1.4a) $$\mathbf{R}_1(t_{n+1/3},t_{n+1},\mathbf{A}_{n+1},\mathbf{C}_{n+1}) = \mathbf{0}, \quad \mathbf{R}_2(t_{n+1/3},t_{n+1},\mathbf{A}_{n+1},\mathbf{C}_{n+1}) = \mathbf{0},$$

where $\mathbf{A}_{n+1}$ may be considered as an auxiliary vector and where the residual functions $\mathbf{R}_1$ and $\mathbf{R}_2$ are defined by

$$\mathbf{R}_1(t_{n+1/3},t_{n+1},\mathbf{A},\mathbf{C}) := \mathbf{A} - \tfrac{5}{12}\Delta t\,\mathbf{F}(t_{n+1/3},\mathbf{A}) + \tfrac{1}{12}\Delta t\,\mathbf{F}(t_{n+1},\mathbf{C}) - \mathbf{C}_n,$$

(1.4b)

$$\mathbf{R}_2(t_{n+1/3},t_{n+1},\mathbf{A},\mathbf{C}) := \mathbf{C} - \tfrac{3}{4}\Delta t\,\mathbf{F}(t_{n+1/3},\mathbf{A}) - \tfrac{1}{4}\Delta t\,\mathbf{F}(t_{n+1},\mathbf{C}) - \mathbf{C}_n.$$

The aim of this paper is to develop efficient iterative methods for solving the systems (1.3) and (1.4).

2. The iteration process

In order to solve the nonlinear systems (1.3) or (1.4) there are two often used approaches: fixed-point iteration and modified Newton iteration. Let us first consider the fixed-point iteration process for the BDF discretization (1.3)

$$\mathbf{C}^{(\nu)} = \mathbf{C}^{(\nu-1)} - \mathbf{R}(t_{n+1},\mathbf{C}^{(\nu-1)}), \quad \nu = 1, 2, \ldots . \tag{2.1}$$

This iteration process is relatively cheap, highly vectorizable, and highly parallelizable. However, it will only converge if Δt is extremely small. To make this more precise, we ignore the nonstiff chemistry and the usually small diffusion terms. Then, fixed-point iteration will converge if the CFL number Q, defined as

$$Q := \Delta t \left(\frac{|u|}{\Delta x} + \frac{|v|}{\Delta y} + \frac{|w|}{\Delta z} \right), \tag{2.2}$$

is about 1, say. Due to the usually small mesh size Δz, such a convergence condition imposes a severe restriction on the time step Δt, unless $w = 0$. For example, in practical situations $|w|/\Delta z$ can be as large as 0.1 sec^{-1}, forcing this process to take timesteps less than 10 seconds to achieve $Q \le 1$ (the horizontal advection terms are less dangerous, because the horizontal mesh sizes usually are a factor 100 or 1000 larger). Evidently, we have to discard the fixed-point iteration process.

At the other end of the scale, we can generate successive approximations $\mathbf{C}^{(\nu)}$ to $\mathbf{C}_{n+1}$ by means of the modified Newton process. Let us again consider the BDF discretization (1.3) for which modified Newton reads

$$\mathbf{C}^{(\nu)} = \mathbf{C}^{(\nu-1)} - J^{-1}\mathbf{R}(t_{n+1},\mathbf{C}^{(\nu-1)}), \quad J := I - \alpha\Delta t \frac{\partial \mathbf{F}}{\partial \mathbf{C}}, \quad \nu = 1, 2, \ldots , \tag{2.3}$$

where $\partial\mathbf{F}/\partial\mathbf{C} = \partial(\mathbf{H}+\mathbf{G})/\partial\mathbf{C}$ is the Jacobian and where $\alpha = 1$ for $n = 0$ and $\alpha = 2/3$ for $n \ge 1$. Each Newton iteration requires the solution of a linear system with J as its matrix of coefficients. In the case of *linear* interaction terms, only *one* Newton iteration suffices, because the advection-diffusion terms are also linear. But also in the case of nonlinear interaction, the Newton process is expected to converge under rather mild conditions on the time step Δt. However, due to the different coupling of the unknowns in the functions **H** and **G**, and due to the fact that we are dealing with *three* spatial dimensions, the linear algebra involved in solving the linear Newton systems is extremely expensive, so that we also have to drop the modified Newton algorithm.

In this paper, we shall describe an iteration scheme that is in some sense a compromise between the two extreme cases of fixed-point iteration and modified Newton iteration. This iteration scheme will be separately discussed for the BDF discretization (1.3) and the Radau discretization (1.4).

2.1. The BDF discretization

By observing that the modified Newton process (2.3) can be interpreted as fixed-point iteration in which the residual term $\mathbf{R}(t_{n+1},\mathbf{C}^{(\nu-1)})$ is preconditioned by the matrix J, we are led to define for (1.3) a preconditioned fixed-point iteration process of the form

$$(2.4)\qquad \mathbf{C}^{(\nu)} = \mathbf{C}^{(\nu-1)} - \tilde{J}^{-1}\mathbf{R}(t_{n+1},\mathbf{C}^{(\nu-1)}),\quad \nu = 1, 2, \ldots ,$$

where the matrix $\tilde{J}^{-1}$ should remove the stiffness from the residual term. Evidently, any nonsingular matrix $\tilde{J}$ defines a *consistent* iteration scheme, that is, if the iteration scheme converges, then it converges to the solution of (1.3). The matrix $\tilde{J}$ may also be interpreted as a smoothing matrix which removes the high frequencies from the residual term.

Our first concern is to choose $\tilde{J}$ such that we have convergence as $\Delta z \to 0$. Suppose that we define $\tilde{J}$ by the 'vertical-preconditioning' matrix $\tilde{J} = I - \alpha\Delta tZ$, where Z is the matrix occurring in (1.2b). Then, (2.4) takes the form

$$(2.4')\qquad (I - \alpha\Delta tZ)\,(\mathbf{C}^{(\nu)} - \mathbf{C}^{(\nu-1)}) = -\,\mathbf{R}(t_{n+1},\mathbf{C}^{(\nu-1)}),\quad \nu = 1, 2, \ldots .$$

Since the matrix Z only represents the vertical coupling in the advection diffusion terms, the linear system (2.4') can be decoupled into N_{xy} 'vertical' linear pentadiagonal subsystems, where N_{xy} is the number of grid points in the horizontal plane. Therefore, the solution of the system (2.4') has a considerable amount of intrinsic parallelism and vectorization, and can be done extremely fast on parallel vector computers like Cray architectures.

Although the vertical advection (and diffusion) terms are the most dangerous ones, it is not recommendable to forget about the horizontal advection (and diffusion) terms. Therefore, we shall apply a similar 'horizontal' preconditioning of the residual term. This results into the iteration process

$$(2.5)\qquad (I - \alpha\Delta tX)(I - \alpha\Delta tY)(I - \alpha\Delta tZ)(\mathbf{C}^{(\nu)} - \mathbf{C}^{(\nu-1)}) = -\,\mathbf{R}(t_{n+1},\mathbf{C}^{(\nu-1)}),\quad \nu = 1, 2, \ldots .$$

Each iteration requires the sequential solution of three linear systems, but by virtue of the structure of the matrices X, Y and Z (cf. (1.2b)), these systems can be decoupled respectively into N_{xy}, N_{xz} and N_{yz} linear pentadiagonal subsystems (compare the decoupling in (2.4')). Here, N_{yz} and N_{xz} are defined in a similar way as N_{xy} and denote the numbers of gridpoints in the vertical 'north-south' and 'east-west' planes.

The process (2.5) may be considered as the method of Approximate Factorizations applied to the modified Newton process (2.3) (cf. [2, p. 439]). This approach replaces the expensive linear system in (2.3) by the set of pentadiagonal subsystems involved in (2.5).

For future reference, it is convenient to give an explicit expression for the matrices X, Y and Z occurring in (2.5). For simplicity, we present this expression for constant diffusion coefficients ε_x, ε_y

and ε_z and positive velocities u, v and w. Then, at a particular grid point $P_{ijk} := (i\Delta x, j\Delta y, k\Delta z)$, we have

$$\mathbf{H}_{ijk} = (X\mathbf{C})_{ijk} + (Y\mathbf{C})_{ijk} + (Z\mathbf{C})_{ijk}, \tag{2.6}$$

where

$$\begin{aligned}
(X\mathbf{C})_{ijk} &:= -(3B_1 + 2D_1)\,\mathbf{c}_{i,j,k} - (2B_1 - D_1)\,\mathbf{c}_{i+1,j,k} + (6B_1 + D_1)\,\mathbf{c}_{i-1,j,k} - B_1\mathbf{c}_{i-2,j,k},\\
(Y\mathbf{C})_{ijk} &:= -(3B_2 + 2D_2)\,\mathbf{c}_{i,j,k} - (2B_2 - D_2)\,\mathbf{c}_{i,j+1,k} + (6B_2 + D_2)\,\mathbf{c}_{i,j-1,k} - B_2\mathbf{c}_{i,j-2,k},\\
(Z\mathbf{C})_{ijk} &:= -(3B_3 + 2D_3)\,\mathbf{c}_{i,j,k} - (2B_3 - D_3)\,\mathbf{c}_{i,j,k+1} + (6B_3 + D_3)\,\mathbf{c}_{i,j,k-1} - B_3\mathbf{c}_{i,j,k-2},
\end{aligned}$$

and

$$B_1 := \frac{u}{6\Delta x}\ ,\ B_2 := \frac{v}{6\Delta y}\ ,\ B_3 := \frac{w}{6\Delta z}\ ,\ D_1 := \frac{\varepsilon_x}{(\Delta x)^2}\ ,\ D_2 := \frac{\varepsilon_y}{(\Delta y)^2}\ ,\ D_3 := \frac{\varepsilon_z}{(\Delta z)^2}\ .$$

Here, $\mathbf{c}_{i,j,k}$ refers to the m-dimensional vector of concentrations c_μ at the grid point P_{ijk}.

2.2. The Radau discretization

Next we consider the approach in which the underlying implicit time discretization formula is of Runge-Kutta type. As an example, we will discuss the third-order, L-stable, 2-stage Radau IIA method. Then, in each step, a system has to be solved, the dimension of which is twice as large as the IVP dimension (cf. (1.4a) and (1.4b)). To avoid the increase in the linear algebra work, the Runge-Kutta matrix appearing in the modified Newton method will be approximated by a diagonal matrix $\mathrm{diag}(\alpha_1,\alpha_2)$. For the 2-stage Radau method this results in the iteration process

$$\begin{aligned}
&\mathbf{A}^{(\nu)} = \mathbf{A}^{(\nu-1)} - J_1^{-1}\mathbf{R}_1(t_{n+1/3}, t_{n+1}, \mathbf{A}^{(\nu-1)}, \mathbf{C}^{(\nu-1)}),\ J_1 := I - \alpha_1\Delta t\,\frac{\partial \mathbf{F}}{\partial \mathbf{C}}\ ,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \nu = 1, 2, \ldots\ ,\\
&\mathbf{C}^{(\nu)} = \mathbf{C}^{(\nu-1)} - J_2^{-1}\mathbf{R}_2(t_{n+1/3}, t_{n+1}, \mathbf{A}^{(\nu-1)}, \mathbf{C}^{(\nu-1)}),\ J_2 := I - \alpha_2\Delta t\,\frac{\partial \mathbf{F}}{\partial \mathbf{C}}\ ,
\end{aligned} \tag{2.7}$$

where $\partial\mathbf{F}/\partial\mathbf{C} = \partial(\mathbf{H}+\mathbf{G})/\partial\mathbf{C}$ is again the Jacobian. Notice that both equations are of IVP dimension and, moreover, the new iterates $\mathbf{A}^{(\nu)}$ and $\mathbf{C}^{(\nu)}$ can be solved in parallel. In fact, these equations are of the same type as the modified Newton iteration (2.3) for the BDF method. In [3], we followed the same approach to construct parallel iteration methods to solve stiff ODEs with Runge-Kutta methods. In that paper it was shown that good results are obtained by requiring that (extremely) stiff components are optimally damped. For the Radau IIA method this leads to

$$\alpha_1 = (4 - \sqrt{6})/6, \quad \alpha_2 = (4 + \sqrt{6})/10. \tag{2.8}$$

In Section 3.2 we will argue that (2.8) is also a plausible choice in the present context.
Proceeding as in the previous section, the scheme (2.7) will be further simplified by the Approximate Factorization approach to obtain

(2.9a) $(I - \alpha_1\Delta tX)(I - \alpha_1\Delta tY)(I - \alpha_1\Delta tZ)(\mathbf{A}^{(\nu)} - \mathbf{A}^{(\nu-1)}) = -\mathbf{R}_1(t_{n+1/3},t_{n+1},\mathbf{A}^{(\nu-1)},\mathbf{C}^{(\nu-1)})$,

(2.9b) $(I - \alpha_2\Delta tX)(I - \alpha_2\Delta tY)(I - \alpha_2\Delta tZ)(\mathbf{C}^{(\nu)} - \mathbf{C}^{(\nu-1)}) = -\mathbf{R}_2(t_{n+1/3},t_{n+1},\mathbf{A}^{(\nu-1)},\mathbf{C}^{(\nu-1)})$,

where $\nu = 1, 2, \ldots$, and the matrices X, Y and Z are defined in (1.2b). Similar to the BDF-based method (cf. (2.5)), each iteration requires the successive solution of three linear systems; however, this can be done in parallel for (2.9a) and (2.9b).

3. Convergence analysis

Again, we discuss the BDF and Radau discretizations separately.

3.1. The BDF discretization

Let us define the iteration error $\mathbf{e}^{(\nu)} := \mathbf{C}^{(\nu)} - \mathbf{C}_{n+1}$. Then,

$$(3.1)\qquad (I - \alpha\Delta tX)(I - \alpha\Delta tY)(I - \alpha\Delta tZ)(\mathbf{e}^{(\nu)} - \mathbf{e}^{(\nu-1)}) = -\left(\mathbf{R}(t_{n+1},\mathbf{C}^{(\nu-1)}) - \mathbf{R}(t_{n+1},\mathbf{C}_{n+1})\right)$$
$$\approx -\left(I - \alpha\Delta t\,\frac{\partial \mathbf{H}}{\partial \mathbf{C}} - \alpha\Delta t\,\frac{\partial \mathbf{G}}{\partial \mathbf{C}}\right)\mathbf{e}^{(\nu-1)}.$$

By observing that $\partial\mathbf{H}/\partial\mathbf{C} = X + Y + Z$, where the matrices X, Y and Z are defined according to (1.2b), we obtain the error recursion

$$(3.2)\qquad \mathbf{e}^{(\nu)} = M\mathbf{e}^{(\nu-1)},$$
$$M := I - (I - \alpha\Delta tZ)^{-1}(I - \alpha\Delta tY)^{-1}(I - \alpha\Delta tX)^{-1}\left(I - \alpha\Delta t\,(X + Y + Z) - \alpha\Delta t\,\frac{\partial \mathbf{G}}{\partial \mathbf{C}}\right).$$

We have convergence if the eigenvalues μ of M are within the unit circle. Since the reaction terms are nonstiff, we ignore the term $\alpha\Delta t\,\partial\mathbf{G}/\partial\mathbf{C}$, so that a normal mode analysis can be applied. Let

$$M\exp(i(\omega_1 x+\omega_2 y+\omega_3 z)) = \mu\exp(i(\omega_1 x+\omega_2 y+\omega_3 z)),$$

$$\theta_1 := \omega_1\Delta x,\quad \theta_2 := \omega_2\Delta y,\quad \theta_3 := \omega_3\Delta z.$$

Then it follows from (3.2) and (2.6) that

$$(3.3)\qquad \mu = 1 - \frac{1 - \alpha\Delta t\,(\lambda_1(\theta_1) + \lambda_2(\theta_2) + \lambda_3(\theta_3))}{(1 - \alpha\Delta t\,\lambda_1(\theta_1))(1 - \alpha\Delta t\,\lambda_2(\theta_2))(1 - \alpha\Delta t\,\lambda_3(\theta_3))},\quad |\theta_j| \le \pi,\ j = 1, 2, 3,$$

where the $\lambda_j(\theta_j)$ represent the eigenvalues of the matrices X, Y and Z, i.e.

$$\lambda_j(\theta_j) = -B_j\gamma(\theta_j) + D_j\delta(\theta_j),$$

$$(3.4)\qquad \gamma(\theta) := 2e^{i\theta} + 3 - 6e^{-i\theta} + e^{-2i\theta} = 2(\cos(\theta) - 1)^2 + 2i\,(4 - \cos(\theta))\sin(\theta),$$

$$\delta(\theta) := e^{-i\theta} - 2 + e^{i\theta} = 2(\cos(\theta) - 1).$$

Recalling that the magnitude of B_3 is usually much larger than that of B_1 and B_2, we consider the limiting case where $B_3 = \infty$ (i.e., $\Delta z \to 0$ and $w \neq 0$), to obtain

$$\text{(3.5a)} \qquad \mu(\theta_1,\theta_2,\theta_3) = 1 - \frac{1}{(1 - \alpha\Delta t\, \lambda_1(\theta_1))(1 - \alpha\Delta t\, \lambda_2(\theta_2))}, \quad \theta_3 \neq 0,\ |\theta_j| \le \pi,\ j = 1, 2,$$

$$\text{(3.5b)} \qquad \mu(\theta_1,\theta_2,0) = 1 - \frac{1 - \alpha\Delta t\, (\lambda_1(\theta_1) + \lambda_2(\theta_2))}{(1 - \alpha\Delta t\, \lambda_1(\theta_1))(1 - \alpha\Delta t\, \lambda_2(\theta_2))}, \quad |\theta_j| \le \pi,\ j = 1, 2.$$

It can straightforwardly be verified that these expressions are on the open unit disk if

$$\text{(3.6a)} \qquad (\alpha\Delta t)^2\, \mathrm{Re}(\lambda_1(\theta_1)\lambda_2(\theta_2)) - \alpha\Delta t\, \mathrm{Re}(\lambda_1(\theta_1) + \lambda_2(\theta_2)) + \frac{1}{2} > 0, \quad |\theta_j| \le \pi,\ j = 1, 2.$$

$$\text{(3.6b)} \qquad \alpha\Delta t\, \mathrm{Re}(\lambda_j(\theta_j)) < \frac{1}{2},$$

The second inequality is always satisfied (see (3.4)) and the first inequality is satisfied if $\mathrm{Re}(\lambda_1(\theta_1)\lambda_2(\theta_2)) \ge 0$. If $\mathrm{Re}(\lambda_1(\theta_1)\lambda_2(\theta_2)) < 0$, then we obtain an upper bound for Δt. However, this upper bound is *positive* and certainly greater than $[-2\alpha^2 \min\{\mathrm{Re}(\lambda_1(\theta_1)\lambda_2(\theta_2))\}]^{-1/2}$. Since $\left|\mathrm{Re}(\lambda_1(\theta_1)\lambda_2(\theta_2))\right| \le \left|\lambda_1(\theta_1)\lambda_2(\theta_2)\right| \le (\gamma_0 B_1 + 4D_1)(\gamma_0 B_2 + 4D_2)$, where $\gamma_0 := \max\{|\gamma(\theta)|\}$, we conclude that a sufficient convergence condition is

$$\text{(3.7)} \qquad \Delta t \le \frac{1}{\alpha\sqrt{2(\gamma_0 B_1 + 4D_1)(\gamma_0 B_2 + 4D_2)}}, \quad \gamma_0 := \max\{|\gamma(\theta)|\} = 9.$$

Thus, we have the following convergence theorem:

Theorem 3.1. If $w \neq 0$ and $\Delta z \to 0$, then a sufficient condition for convergence is given by (3.7).

Remark 3.1. In the case where $w = 0$ (i.e., $B_3 = 0$) and a sufficiently small vertical diffusion coefficient, the eigenvalues μ can be approximated by (3.5b). We already showed that (3.5b) assumes values within the unit circle if (3.6b) is satisfied, so that we have unconditional convergence if $B_3 = 0$. ♦

In order to obtain some quantitative information on the size of the CFL numbers and the maximal convergent stepsize, consider a advection-dominated model problem $\{B_1 = B_2,\ B_3 = qB_1,\ D_j = 0\}$. Defining the quantity $b = \alpha B_1 \Delta t$, we obtain

$$\text{(3.3')} \qquad \mu(\theta_1,\theta_2,\theta_3) = 1 - \frac{1 + b\gamma(\theta_1) + b\gamma(\theta_2) + bq\gamma(\theta_3)}{(1 + b\gamma(\theta_1))(1 + b\gamma(\theta_2))(1 + bq\gamma(\theta_3))}, \quad |\theta_j| \le \pi,\ j = 1, 2, 3.$$

For this model problem, the CFL number is given by $Q = 6(2+q)B_1\Delta t = 6(2+q)b\alpha^{-1}$, so that for a given q, the maximal convergent CFL number can be computed by finding the largest value of b for

which $\mu(\theta_1,\theta_2,\theta_3)$ remains on the open unit disk when $(\theta_1,\theta_2,\theta_3)$ runs through the frequency space. In Table 3.1 we have listed the CFL number for a sequence of q-values. In addition, we listed the maximal values of Δt, i.e. $\Delta t = Q\big(6(2+q)B_1\big)^{-1}$.

Realistic values for B_1 and B_2 in a shallow water transport problem are $B_1 = B_2 = 1/6000$ (corresponding to, for example, $|u| = |v| = 1$ m.sec^{-1} and $\Delta x = \Delta y = 1000$ m). Then, the most critical q-value still allows for a timestep of approximately 16 minutes.

It is of interest to apply the sufficient condition (3.7) of Theorem 3.1 to the above model problem $\{B_1 = B_2,\ D_j = 0\}$. This yields $\Delta t B_1 \le 0.12$. Since $w \neq 0$ and $\Delta z \to 0$ implies that $q = \infty$, this sufficient condition should be compared with the true condition $\Delta t B_1 \le 0.18$ of Table 3.1. Thus, condition (3.7) prescribes stepsizes that are a factor 2/3 smaller than really necessary.

Table 3.1. CFL numbers and maximal convergent stepsizes for the BDF discretization.

q =	∞	100	50	20	10	5	2	1	0.5	0.1	0
Q =	∞	99	50	21	11.5	6.9	4.5	4.1	4.4	7.9	∞
$\Delta t B_1 \le$	0.18	0.16	0.16	0.16	0.16	0.16	0.18	0.22	0.29	0.62	∞

Furthermore, it is of interest to know which frequencies are responsible for the limitation of the time step. To get an impression, we again use the convection-dominated model problem $\{B_1=B_2,\ B_3=qB_1,\ D_j=0\}$, apply the maximal value for Δt and we consider the $|\mu|$-values as a function of θ_1,θ_2,and θ_3. In the Figures 3.1a,b,c $|\mu|$-values are plotted for q=100, 1, and 0.1, respectively. In these plots, the horizontal axes contain θ_1- and θ_3-values; θ_2 has been omitted since we observed that the maximal Δt was always obtained for $\theta_1=\theta_2$.

We see that for all q-values, $|\mu|$ does not critically depend on θ_3, except for $\theta_3=0$. In that case, the expression for μ reduces to (3.5b) for which it was shown that there is no restriction on the time step. Furthermore, we observe that the low frequencies in the θ_1-direction are perfectly damped; the highest 'horizontal' frequencies ($\theta_1 \approx \pi$) are treated satisfactorily for large q, but for small q, the damping of these frequencies tends to one. Hence, the price to be paid for unconditional convergence (which corresponds to q=0, see Table 3.1) is a poor convergence of the high-frequency error components. Finally, these plots clearly indicate that the mid-range in the frequency space causes the condition on the time step.

3.2. The Radau discretization

Proceeding as for the BDF discretization, we obtain for the iteration errors $\mathbf{e}_1^{(\nu)} := \mathbf{A}^{(\nu)} - \mathbf{A}_{n+1}$ and $\mathbf{e}_2^{(\nu)} := \mathbf{C}^{(\nu)} - \mathbf{C}_{n+1}$ the error recursion

$$(3.8)\qquad \begin{pmatrix} \mathbf{e}_1^{(\nu)} \\ \mathbf{e}_2^{(\nu)} \end{pmatrix} \approx M \begin{pmatrix} \mathbf{e}_1^{(\nu-1)} \\ \mathbf{e}_2^{(\nu-1)} \end{pmatrix}, \quad M := \begin{pmatrix} I - P_1^{-1}\left(I - \frac{5}{12}\Delta t\, S\right) & -\frac{1}{12}\Delta t\, P_1^{-1} S \\ \frac{3}{4}\Delta t\, P_2^{-1} S & I - P_2^{-1}\left(I - \frac{1}{4}\Delta t\, S\right) \end{pmatrix},$$

where

$$P_j := (I - \alpha_j \Delta t X)\,(I - \alpha_j \Delta t Y)\,(I - \alpha_j \Delta t Z), \quad S := X + Y + Z + \frac{\partial \mathbf{G}}{\partial \mathbf{C}}, \quad j = 1, 2.$$

Ignoring the reaction terms and applying a normal mode analysis, we see that the eigenvalues of M are given by the eigenvalues of the matrix

$$M^* := \begin{pmatrix} 1 - \dfrac{1 - \frac{5}{12}\Delta t\, \lambda(S)}{\lambda(P_1)} & -\frac{1}{12}\Delta t\, \dfrac{\lambda(S)}{\lambda(P_1)} \\ \frac{3}{4}\Delta t\, \dfrac{\lambda(S)}{\lambda(P_2)} & 1 - \dfrac{1 - \frac{1}{4}\Delta t\, \lambda(S)}{\lambda(P_2)} \end{pmatrix}, \tag{3.9}$$

where

$$\lambda(S) = \lambda_1(\theta_1) + \lambda_2(\theta_2) + \lambda_3(\theta_3),$$

$$\lambda(P_j) = \big(1 - \alpha_j \Delta t\, \lambda_1(\theta_1)\big)\big(1 - \alpha_j \Delta t\, \lambda_2(\theta_2)\big)\big(1 - \alpha_j \Delta t\, \lambda_3(\theta_3)\big), \quad j = 1, 2,$$

and where $\lambda_j(\theta_j)$ is defined in (3.4). Convergence requires that the eigenvalues μ of M^* are within the unit circle. Table 3.2 represents the analogue of Table 3.1 for the advection-dominated model problem $\{B_1 = B_2,\ B_3 = qB_1,\ D_j = 0\}$. A comparison of these tables reveals that the Radau discretization imposes slightly less restrictive convergence conditions than the BDF discretization.

Table 3.2. CFL numbers and maximal convergent stepsizes for the Radau discretization.

q =	∞	100	50	20	10	5	2	1	0.5	0.1	0
Q =	∞	107	55	23	12.6	7.5	4.8	4.4	4.8	8.7	∞
$\Delta t B_1 \leq$	0.19	0.18	0.17	0.17	0.17	0.18	0.20	0.24	0.31	0.69	∞

Similar to the preceding section, we want to know which frequencies give rise to the critical time step with respect to convergence. To that end, we present in Figure 3.2a a plot of the spectral radius of the matrix M* as a function of the spatial frequencies θ_1 and θ_3. Here we again used the above model problem with q=100 and the maximally allowed time step. We observe a similar behaviour as for the BDF case.

Next, we discuss the choice of the parameters α_1 and α_2, as introduced in Section 2.2. As explained there, these parameters were selected on the basis of previous research of related iteration methods to solve stiff ODEs. It is, however, not a priori evident that (2.8) is also the best choice in the present application. To see whether there exist values that allow for a larger time step, we performed a numerical search in the (α_1,α_2)-parameter space (again for the model problem with q=100). It turned out that $\alpha_1 = 0.19$, $\alpha_2 = 0.40$ are the optimal values for the Radau IIA method. These values give rise

to a bound of 0.22 for $\Delta t B_1$, which is an increase of ≈ 20% compared with the value 0.18 as listed in Table 3.2. There is, however, a price to be paid in taking these optimal α-values. As we see from Figure 3.2b (where we plotted the spectral radius of M* for α_1=0.19, α_2=0.40, q=100, and the largest possible time step), the low frequency components in the θ_1 direction are damped with a factor close to 0.6. This is an undesirable situation and indeed, using these α-values, we observed a reduced convergence behaviour in the numerical test problem, described in Section 4. The poor convergence of the low-frequency components is easily understood by analysing the matrix M* in (3.9) for $\theta_1=\theta_2=0$.
Then this matrix reduces to

$$M^* = z\,(I - zD)^{-1}(A - D)\,, \tag{3.9'}$$

where $z := \Delta t\, \lambda_3(\theta_3)$, $D := \mathrm{diag}(\alpha_1, \alpha_2)$ and A is the Radau matrix. Since $|z| = |B_3 \Delta t\, \gamma_3(\theta_3)|$ will take large values in many practical situations, the matrix M* behaves like $I - D^{-1}A$ (indeed, the "optimal" choice $(\alpha_1,\alpha_2)=(0.19,0.40)$ leads to eigenvalues ≈ 0.61 for this matrix). In [3], we studied exactly the same iteration matrix M* as given in (3.9'); the choice (2.8) resulted from requiring that $I-D^{-1}A$ possesses a zero spectrum. Therefore, we also adopt this choice in the present application, in spite of the fact that the maximally allowed time step in order to obtain convergence for the mid-range frequencies is slightly reduced.

4. Numerical experiments

Consider the test problem

$$\begin{aligned} &\frac{\partial c_1}{\partial t} + \mathbf{U}\cdot\nabla c_1 = \varepsilon\,\Delta c_1 + g_1(t,x,y,z) - k_1 c_1 c_2, \\ &\qquad\qquad 0 \le x, y \le L_h,\ -L_v \le z \le 0,\ 0 \le t \le T, \\ &\frac{\partial c_2}{\partial t} + \mathbf{U}\cdot\nabla c_2 = \varepsilon\,\Delta c_2 + g_2(t,x,y,z) - k_1 c_1 + k_2(1 - c_2), \end{aligned} \tag{4.1a}$$

where **U** = (u, v, w) denotes the divergence free velocity field, given in analytical form (see [1])

$$\begin{aligned} u(t,x,y,z) &= \{\ \ \tilde{y} + 3\,(\tilde{z} + 1/2)\,[\,(\tilde{x} - 1/2)^2 + (\tilde{y} - 1/2)^2 - p^2]\,\}\ d(t), \\ v(t,x,y,z) &= \{\ -\tilde{x} + 3\,(\tilde{z} + 1/2)\,[\,(\tilde{x} - 1/2)^2 + (\tilde{y} - 1/2)^2 - p^2]\,\}\ d(t), \\ w(t,x,y,z) &= -\,3\,L_v\,\tilde{z}\,(\tilde{z} + 1)\,\{(\tilde{x} - 1/2)/L_h + (\tilde{y} - 1/2)/L_h\}\ d(t), \end{aligned} \tag{4.1b}$$

where we used the scaled co-ordinates $\tilde{x} := x/L_h$, $\tilde{y} := y/L_h$, $\tilde{z} := z/L_v$, and p = 1/3 and $d(t) = \cos(2\pi t/T_p)$. The Dirichlet boundary conditions, the initial condition and the functions g_1 and g_2 are chosen in accordance with the prescribed analytical solution, which is of the form

$$c_\mu(t,x,y,z) = \exp\{\,\tilde{z}\,/\mu - f_\mu(t) - \gamma_\mu\,[(\,\tilde{x} - r(t)\,)^2 + (\,\tilde{y} - s(t)\,)^2\,]\},\quad \mu = 1, 2, \tag{4.1c}$$

with $f_2(t) = t/(T_b+t)$, $f_1(t) = 4\, f_2(t)$, $r(t) = [2 + \cos(2\pi t/T_p)]/4$, and $s(t) = [2 + \sin(2\pi t/T_p)]/4$.
In our experiments, we take the following values for the parameters: $L_h = 20\,000$, $L_v = 100$, $\varepsilon = 0.5$, $\gamma_1 = 80$, $\gamma_2 = 20$, $T_b = 32400$, and $T_p = 43200$. We use two grids, respectively with $N_x = N_y = 41$, $N_z = 6$ (coarse grid) and $N_x = N_y = 81$, $N_z = 11$ (fine grid). The length of the integration interval $T = 36000$. Realistic values for the reaction rate constants are: $k_1 = k_2 = 10^{-4}$.
The accuracy is measured by

$$cd_\mu := \text{minimum over all grid points } ({}^{10}\log |\text{absolute error for } c_\mu|), \qquad \mu = 1, 2.$$

The iteration processes in the BDF and Radau methods {(1.3),(2.5)} and {(1.4),(2.8),(2.9)} are applied with ν iterations in each step, where ν is specified in the tables of results. Since the Jacobian matrices $\partial\mathbf{H}/\partial\mathbf{C}$ only depend on t (and not on $\mathbf{C}$), they are evaluated at the new time levels. To start the iteration, we use the trivial prediction, i.e., $\mathbf{C}^{(0)} := \mathbf{C}_n$ in BDF and $\mathbf{A}^{(0)} := \mathbf{C}^{(0)} := \mathbf{C}_n$ in the Radau method. It turns out that this choice for $\mathbf{C}^{(0)}$ is more robust than using a (more accurate) extrapolation formula. In addition to these iteration methods, we apply the RBWLH method developed in [4] (see also [6]). For various values of Δt, the cd-values and the CPU time per step are given in the Tables 4.1. These CPU times are obtained when running the codes in vector mode on one processor of a CRAY C98/4256. Evidently, the new methods allow considerably larger stepsizes than the RBWLH method; even for timesteps as large as 1 hour, the new methods are able to produce results of reasonable accuracy. Another nice property of the new methods is that they do not show the Du Fort-Frankel inconsistency as observed in the RBWLH method. By this we mean that, if both Δt and the grid meshes are halved, then the accuracy of the BDF and Radau method increases proportionally. Furthermore, for $N \to \infty$ (that is, $\Delta t \to 0$) all methods show a third-order behaviour in space, as is to be expected from the upwind discretization of the advection terms, which form the most important part in the model. In case the temporal error dominates the spatial error (small N-values), we see that the Radau method is the most accurate one; this is obviously owing to its third-order accuracy in time. Finally, if we take into account that the two sets of linear systems for $\mathbf{A}^{(\nu)}$ and $\mathbf{C}^{(\nu)}$ in the Radau-based method can be solved in parallel, we see that per step both the BDF and the Radau method are cheaper than the RBWLH method.

Table 4.1a. cd_1 / cd_2 - values obtained by the RBWLH method [6].
N = the number of timesteps ($\Delta t = T/N$); an unstable behaviour is indicated by an $*$.

Spatial grid	N=10	N=20	N=35	N=70	N=140	N=280	N=560	...	$N \to \infty$	CPU/step
$\text{Grid}_{\text{coarse}}$	*	*	2.8/1.8	3.5/2.4	3.8/3.0	3.8/3.6	3.8/4.2	...	3.8/4.3	0.062
$\text{Grid}_{\text{fine}}$	*	*	*	2.9/1.8	3.6/2.5	4.3/3.0	4.8/3.7	...	4.8/5.2	0.28

Table 4.1b. cd_1 / cd_2 - values obtained by the BDF method {(1.3),(2.5)}.
N = the number of timesteps ($\Delta t = T/N$).

Spatial grid	ν	N=10	N=20	N=35	N=70	N=140	N=280	N=560	...	N → ∞	CPU/step
$Grid_{coarse}$	2	1.7/1.4	2.0/1.8	2.4/2.2	2.9/2.8	3.4/3.4	3.7/3.9	3.8/4.1	...	3.8/4.3	0.024
$Grid_{fine}$	2	1.6/1.4	2.0/1.8	2.4/2.2	2.9/2.8	3.5/3.4	4.1/4.0	4.5/4.5	...	4.8/5.2	0.13

Table 4.1c. cd_1 / cd_2 - values obtained by the Radau method {(1.4),(2.8),(2.9)}.
N = the number of timesteps ($\Delta t = T/N$).

Spatial grid	ν	N=10	N=20	N=35	N=70	N=140	N=280	N=560	...	N → ∞	CPU/step
$Grid_{coarse}$	3	2.1/2.1	3.1/3.0	3.7/3.6	3.8/4.2	3.8/4.3		…		3.8/4.3	0.064
$Grid_{fine}$	3	2.0/2.0	2.8/2.6	3.5/3.2	4.2/4.0	4.7/4.6	4.8/5.1	…		4.8/5.2	0.33

References

[1] D. Dunsbergen, *Particle models for transport in three-dimensional shallow water flow*, Ph. D. Thesis, Delft Technical University, 1994.

[2] C. Hirsch, *Numerical Computation of Internal and External Flows*, Vol. 1: Fundamentals of Numerical Discretization, 1988, Wiley.

[3] P.J. van der Houwen and B.P. Sommeijer, *Iterated Runge-Kutta methods on parallel computers*, SIAM J. Sci. Stat. Comput. 12 (1991), 1000-1028.

[4] P.J. van der Houwen and B.P. Sommeijer, *Splitting methods for three-dimensional transport models with interaction terms*, CWI Report NM-R9516, submitted for publication.

[5] B. van Leer, *Upwind-difference methods for aerodynamic problems governed by the Euler equations*, Proceedings of the 15th AMS-SIAM Summer Seminar on Applied Mathematics, Scripps Institution of Oceanography, 1983 (B.E. Engquist, S. Osher and R.C.J. Sommerville, Eds.), Lectures in Applied Mathematics 22 - Part 2 (1985), 327-336, AMS, Providence, RI.

[6] B.P. Sommeijer and J. Kok, *Splitting methods for three-dimensional bio-chemical transport*, CWI Report NM-R9607, to appear in APNUM.

Centrum voor Wiskunde en Informatica

REPORTRAPPORT

The iterative solution of fully implicit discretizations of three-dimensional transport models

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

Department of Numerical Mathematics

NM-R9521 1995

Report NM-R9621
ISSN 0169-0388

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

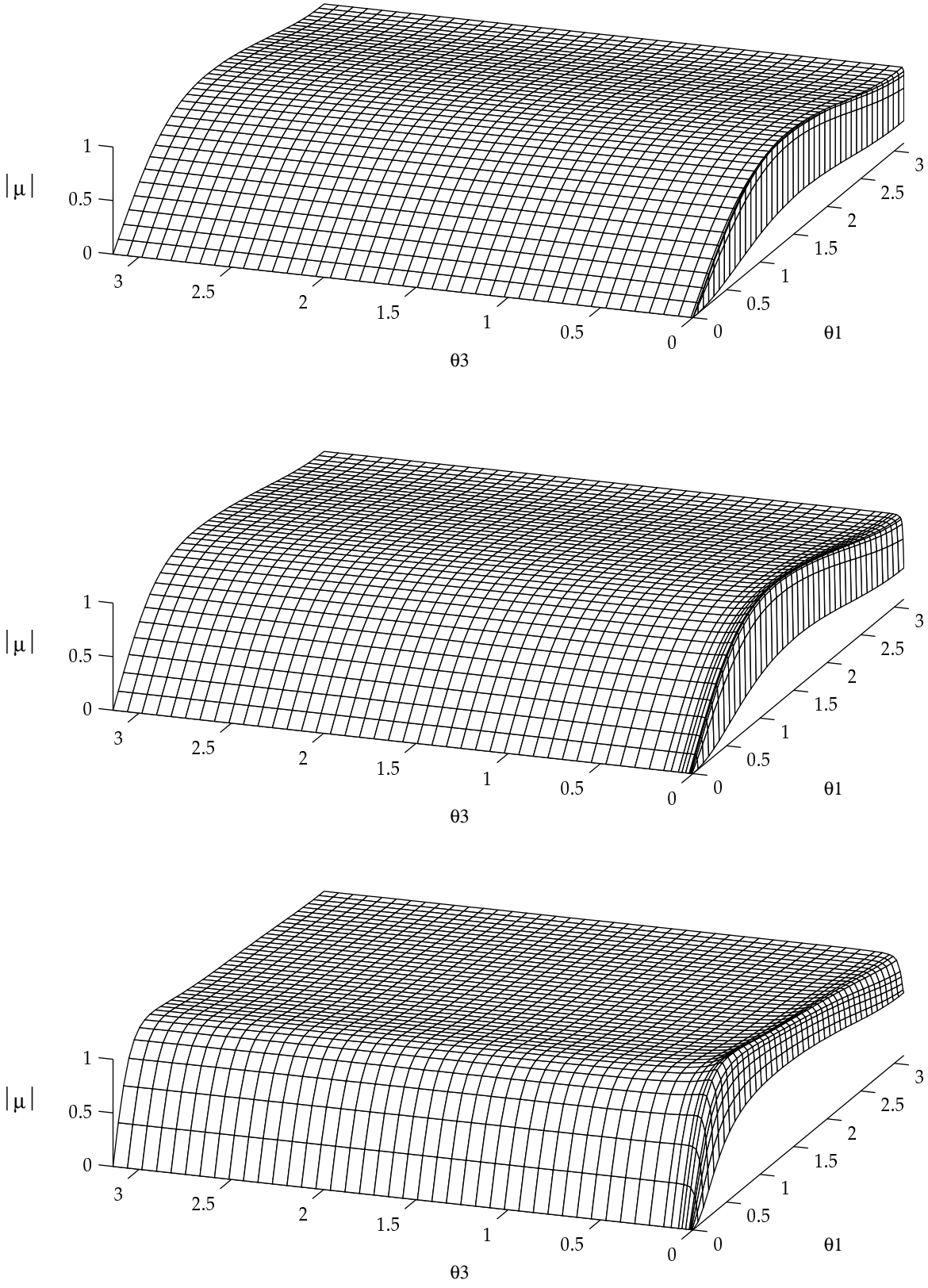

Fig 3.1 |μ|−values (3.3) for q = 100, q = 1, q = 0.1

Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

The iterative solution of fully implicit discretizations of three-dimensional transport models

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

Department of Numerical Mathematics

NM-R9621 1996

Report NM-R9621
ISSN 0169-0388

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

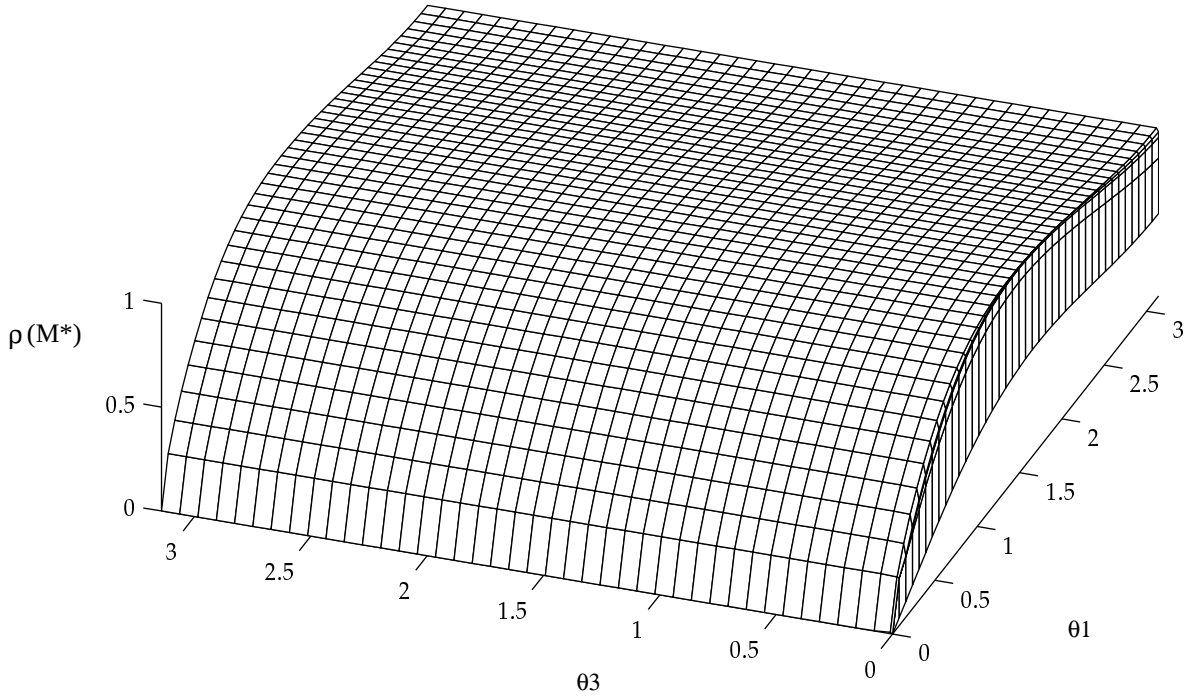

Fig 3.2a Spectral radius of M* (cf. (3.9)) with q = 100 and
$\alpha 1, \alpha 2$ according to (2.8)

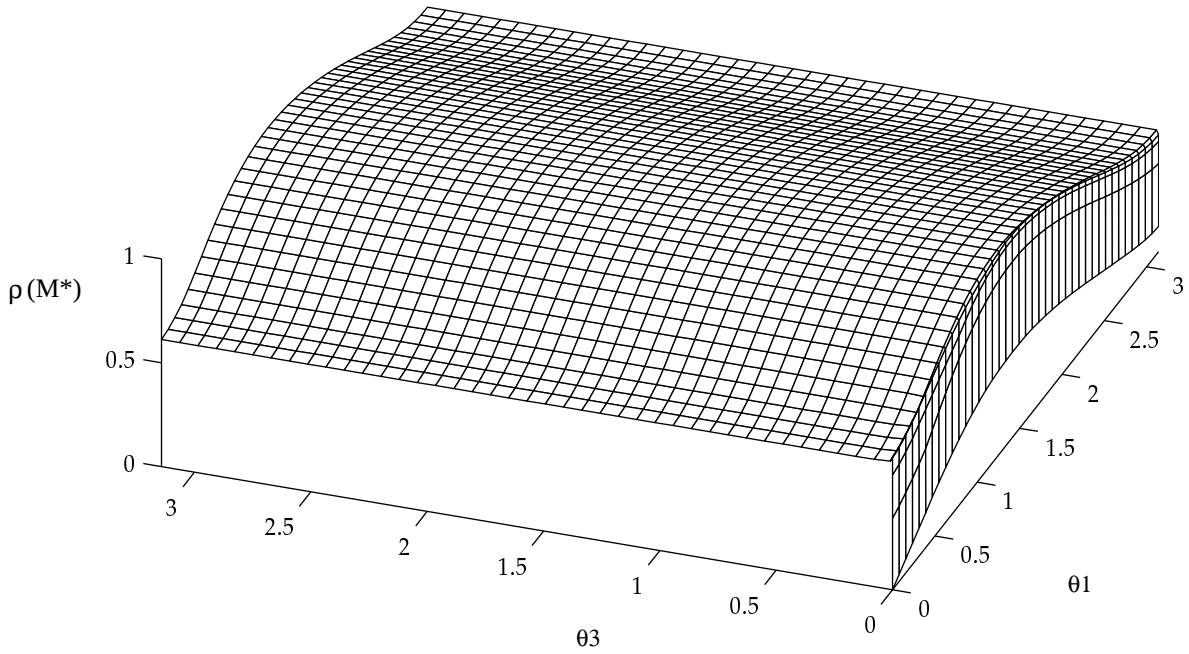

Fig 3.2b Spectral radius of M* (cf. (3.9)) with q = 100 and
$\alpha 1 = 0.19, \alpha 2 = 0.40$