Stability Analysis of a Difference Scheme for Three-Dimensional Advection-Diffusion Problems [1]

Dedicated to Cor Baayen at the Occasion of his Retirement

as our Scientific Director

J.G. Verwer
B.P. Sommeijer
CWI

1 Introduction

1.1 General

The authors of this contribution belong to the research group *Discretization of Evolution Problems* of CWI's *Numerical Mathematics Department.* This research group focuses on fundamental and applied research into numerical methods for evolutionary differential equations. Both ordinary and partial differential equations are treated. In recent years much attention is devoted to large-scale applications and high performance computing. In this connection, an important research subject concerns *Transport Problems in Environmental Applications* which are constituted by systems of time-dependent partial differential equations of the advection-diffusion-reaction type. Numerical research for this type of problems is important for the simulation and prediction of the chemistry and transport of hazardous pollutants in the atmosphere, groundwater and shallow water. Because the systems are usually three-dimensional in space and usually contain many components, one for each chemical or biological constituent in the model, they are extremely CPU and memory intensive and in fact belong to the computationally most expensive models in environmental research and fluid dynamics. Consequently, high performance computing on powerful vector and parallel computers is an important field of research for these applications.

Moreover, when new methods and techniques designed for high-performance use on such computers are developed, also their fundamental numerical properties need to be investigated, notably their stability, consistency and convergence properties. The present contribution provides an example of such a theoretical investigation. This paper deals with a linear stability analysis of a method recently designed in our group for the numerical integration of transport problems in shallow water on vector and parallel computers. To appreciate the

[1] This research was supported by Cray Research Inc. under grant CRG 94.04 via the Stichting Nationale Computerfaciliteiten (National Computing Facilities Foundation, NCF).

complete paper the reader should have a numerical background. Fortunately, the linear stability analysis for difference schemes of the type considered here is based on the well-known Fourier method as proposed by J. Von Neumann (see [8], which is one of the earliest papers where the Fourier method is applied to finite-difference equations). This means that an important part of the paper, viz. Section 3, should be accessible, and hopefully is of some interest, for many readers without any numerical background.

Section 3 is almost self-contained. Here we study the problem of determining the location of the zeros of a polynomial relative to the unit circle in the complex plane. This problem is of long standing (see Schur [11]) and of great practical relevance in applied mathematics (see Miller [6]). In our case we have to deal with a quadratic polynomial whose coefficients are complex-valued functions of a real variable, a phase angle. These functions are determined by the difference scheme and contain so-called advection and diffusion parameters. The question is what conditions should be imposed on these coefficient functions, and hence on their defining parameters, such that the two zeros lie on the unit disc for all phase angles. The resulting conditions determine the critical stepsize for the linear time step stability of the difference scheme. The analysis to solve this stability question shows interesting aspects and surprising results.

1.2 Research contents

In [12] and [13] an odd-even-line hopscotch (OELH) method is developed and implemented for the efficient numerical solution of three-space dimensional advection-diffusion problems modeling the transport of pollutants and suspended material in shallow water. A special feature of this OELH method is that it is explicit for the horizontal transport and implicit for the vertical transport. The implicitness in the vertical direction is necessary to avoid a too stringent stability restriction on the time step. This implicitness gives rise to the solution of a large set of tridiagonal systems, one for every grid point in the horizontal plane. The solution of this large set of tridiagonal systems can be vectorized and parallelized over the horizontal grid, which results in a very good performance [13]. In the comparison with other techniques discussed in [12, 13], the method has been shown superior.

In neither of the aforementioned two papers a comprehensive stability analysis is given. The purpose of the present paper is to fill up this gap. For the general, constant coefficient, linear advection-diffusion model problem we will derive sufficient and necessary conditions for von Neumann stability in the strict sense. Strict means that the stability property we investigate requires the absolute value of amplification factors less than or equal to one. The stability analysis is based on an equivalence with an associated scheme which is composed of the leap-frog, the Du Fort-Frankel, and the Crank-Nicholson scheme. The actual Fourier analysis is carried out for this associated scheme and appears to be rather intricate. For example, the resulting expressions for critical stepsizes reveal that the presence of horizontal diffusion generally leads to a

smaller value, in spite of the fact that we have unconditional stability for pure diffusion problems.

2 THE OELH METHOD FORMULATED FOR THE MODEL PROBLEM

We consider the 3D, constant coefficient, scalar advection-diffusion model problem

$$u_t + q_1 u_x + q_2 u_y + q_3 u_z = \epsilon_1 u_{xx} + \epsilon_2 u_{yy} + \epsilon_3 u_{zz}. \tag{1}$$

Let

$$\frac{d}{dt} U_{ijk} = L_h U_{ijk} \tag{2}$$

be the semi-discrete approximation, resulting from the use of 2nd-order central differences at the uniformly spaced gridpoints

$$(x_i, y_j, z_k) = (ih_1, jh_2, kh_3).$$

The basic formula [1, 2, 3, 4] defining the OELH method studied in [12, 13] then reads

$$U_{\underline{i}}^{n+1} = U_{\underline{i}}^{n} + \tau\theta_{\underline{i}}^{n} L_h U_{\underline{i}}^{n} + \tau\theta_{\underline{i}}^{n+1} L_h U_{\underline{i}}^{n+1}, \tag{3}$$

where $\underline{i} = (i, j, k)$, $\tau = t_{n+1} - t_n$, and the hopscotch parameter $\theta_{\underline{i}}^{n}$ is defined by

$$\theta_{\underline{i}}^{n} = \begin{cases} 1 & \text{for odd values of } \; n + i + j, \\ 0 & \text{for even values of } \; n + i + j. \end{cases} \tag{4}$$

Notice that the subscript k is not involved in this definition, i.e., all gridpoints on a vertical gridline have the same θ-value. If we consider only the odd points (in the space-time grid), then the forward Euler rule results,

$$U_{\underline{i}}^{n+1} = U_{\underline{i}}^{n} + \tau L_h U_{\underline{i}}^{n}, \tag{5}$$

and at the even points, for the same n, we have the backward Euler rule

$$U_{\underline{i}}^{n+1} = U_{\underline{i}}^{n} + \tau L_h U_{\underline{i}}^{n+1}. \tag{6}$$

Consequently, by first applying the explicit forward Euler method at all odd points, and subsequently the implicit backward Euler method at all even points, we have carried out one step with (2.3). The merit of the method lies in the fact that the implicit step is only implicit for the vertical direction. This follows from the 3-point coupling in the horizontal directions and from the definition of the $\theta_{\underline{i}}^{n}$. If we remove the third dimension, then we recover the odd-even-hopscotch scheme (OEH) which is scalarly implicit. Note that the OEH scheme for the 3D problem results if we replace $(n + i + j)$ in $\theta_{\underline{i}}^{n}$ by $(n + i + j + k)$. The stability of the OEH scheme applied to (2.1) has been studied in [14].

The von Neumann stability approach cannot be carried out for (2.3) as it stands. Following [3, 14], we therefore derive an equivalent formula which

does admit Fourier analysis. First introduce, for $m = 1, 2, 3$, the advection parameter c_m and the diffusion parameter σ_m,

$$c_m = \frac{\tau q_m}{h_m},\ \sigma_m = \frac{\tau \epsilon_m}{h_m^2}, \tag{7}$$

and the difference operators H_m and δ_m^2,

$$H_1 U_{\underline{i}} = U_{i+1jk} - U_{i-1jk},\ \text{etc.} \tag{8}$$

$$\delta_1^2 U_{\underline{i}} = U_{i+1jk} - 2U_{ijk} + U_{i-1jk},\ \text{etc.} \tag{9}$$

We then may express $\tau L_h U_{\underline{i}}$ as

$$\tau L_h U_{\underline{i}} = \sum_{m=1}^{3} \left(-\tfrac{1}{2} c_m H_m + \sigma_m \delta_m^2\right) U_{\underline{i}}. \tag{10}$$

Next introduce, in addition to (2.3), the OELH formula for the next time step

$$U_{\underline{i}}^{n+2} = U_{\underline{i}}^{n+1} + \tau \theta_{\underline{i}}^{n+1} L_h U_{\underline{i}}^{n+1} + \tau \theta_{\underline{i}}^{n+2} L_h U_{\underline{i}}^{n+2}. \tag{11}$$

Using (2.3), (2.4) and (2.11), for the odd points we then can write, considering time levels n and $n+2$,

$$U_{\underline{i}}^{n+2} = U_{\underline{i}}^{n} + \tau L_h \left(U_{\underline{i}}^{n} + U_{\underline{i}}^{n+2}\right). \tag{12}$$

Likewise, for the even points we find

$$U_{\underline{i}}^{n+2} = 2U_{\underline{i}}^{n+1} - U_{\underline{i}}^{n}. \tag{13}$$

Next we elaborate the odd-point formula (2.12). Using (2.13) to eliminate variables at even points, an elementary calculation with (2.10) shows that (2.12) can be written as

$$\begin{aligned}(1+\sigma)\, U_{\underline{i}}^{n+2} = (1-\sigma)\, U_{\underline{i}}^{n} + (4\sigma_1\mu_1 + 4\sigma_2\mu_2)\, U_{\underline{i}}^{n+1} - \\ (c_1 H_1 + c_2 H_2)\, U_{\underline{i}}^{n+1} + \left(-\tfrac{1}{2} c_3 H_3 + \sigma_3 \delta_3^2\right) \left(U_{\underline{i}}^{n} + U_{\underline{i}}^{n+2}\right),\end{aligned} \tag{14}$$

where μ_m is the averaging operator

$$\mu_1 U_{\underline{i}} = \tfrac{1}{2}\left(U_{i+1jk} + U_{i-1jk}\right),\ \text{etc.} \tag{15}$$

and

$$\sigma = 2\left(\sigma_1 + \sigma_2\right). \tag{16}$$

It is important to note that in (2.14) only variables at odd numbered points appear. This means that the solution defined by (2.3), can first be computed

by means of (2.14) at the complete set of odd points, and thereafter at the complete set of even points by means of (cf. (2.13))

$$U_{\underline{i}}^{n+1} = \tfrac{1}{2}\left(U_{\underline{i}}^{n} + U_{\underline{i}}^{n+2}\right). \tag{17}$$

Hence for the stability analysis we may proceed with the odd-point scheme (2.14), because the sets of even and odd points are decoupled.

We see that this odd-point scheme is composed of the leap-frog scheme for the horizontal advection part,

$$U_{\underline{i}}^{n+2} = U_{\underline{i}}^{n} - (c_1 H_1 + c_2 H_2)\, U_{\underline{i}}^{n+1}, \tag{18}$$

of the Du Fort-Frankel scheme for the horizontal diffusion part,

$$(1+\sigma)\, U_{\underline{i}}^{n+2} = (1-\sigma)\, U_{\underline{i}}^{n} + (4\sigma_1\mu_1 + 4\sigma_2\mu_2)\, U_{\underline{i}}^{n+1}, \tag{19}$$

and of the Crank-Nicholson scheme, with stepsize 2τ, for the vertical advection and diffusion part,

$$U_{\underline{i}}^{n+2} = U_{\underline{i}}^{n} + \left(-\tfrac{1}{2}c_3 H_3 + \sigma_3\delta_3^2\right)\left(U_{\underline{i}}^{n} + U_{\underline{i}}^{n+2}\right). \tag{20}$$

Consequently, in view of the unconditional stability of the Crank-Nicholson and Du Fort-Frankel scheme, at first sight one might expect that the critical stepsize for stability equals that of the leap-frog scheme (2.18). In the next section we will prove that this is indeed true if there is no horizontal diffusion. However, if horizontal diffusion terms are present, then the situation turns out to be more complicated. We will show that in this case the critical stepsize is generally smaller.

3 Strict von Neumann Stability

Substitution of the Fourier mode

$$U_{\underline{i}}^{n} = \xi^n e^{I(\omega_1 x_i + \omega_2 y_j + \omega_3 z_k)}, \quad I^2 = -1, \tag{21}$$

into scheme (2.14) leads to the characteristic polynomial

$$f(\xi) = a_0 + a_1\xi + a_2\xi^2 \tag{22}$$

with coefficients

$$\begin{aligned} a_0 &= -1 + \sigma - 2\sigma_3(\cos\theta_3 - 1) + Ic_3\sin\theta_3, \\ a_1 &= \sum_{m=1}^{2} -4\sigma_m\cos\theta_m + 2Ic_m\sin\theta_m, \\ a_2 &= 1 + \sigma - 2\sigma_3(\cos\theta_3 - 1) + Ic_3\sin\theta_3, \end{aligned} \tag{23}$$

where $\theta_m = \omega_m h_m$ denotes the phase angle. The specific stability property we will investigate is von Neumann stability in the strict sense:

DEFINITION 1 *Method (2.14) is called von Neumann stable if the zeroes* ξ_1, ξ_2 *of the characteristic polynomial (3.2) satisfy*

$$|\xi_1|,\ |\xi_2| \leq 1 \ \textit{for all}\ \ |\theta_m| \leq \pi,\ \ m = 1, 2, 3. \tag{24}$$

Hence strict means that the stability property we investigate requires the absolute value of amplification factors less than or equal to one. In literature, this is also called 'practical' or 'modified' von Neumann stability [9, 7, 5]. Note that the original von Neumann condition is weaker as it requires $|\xi| \leq 1 + O(\tau)$ [9]. As is well known, for advection-diffusion problems this weaker condition can lead to unacceptably large errors [7]. Strict stability is also more natural here, since Fourier modes of the true solution cannot grow in time either.

For the von Neumann analysis we will use results from [6]. We therefore introduce the polynomial

$$f^{\star}(\xi) = \overline{a}_2 + \overline{a}_1 \xi + \overline{a}_0 \xi^2, \tag{25}$$

and the so-called first reduced polynomial

$$f_1(\xi) = \overline{a}_2 a_1 - \overline{a}_1 a_0 + (\overline{a}_2 a_2 - \overline{a}_0 a_0)\, \xi, \tag{26}$$

where

$$\overline{a}_2 a_1 - \overline{a}_1 a_0 = -8 \sum_{m=1}^{2} \sigma_m \cos\theta_m \ + I \left(8 c_3 \sin\theta_3 \sum_{m=1}^{2} \sigma_m \cos\theta_m \right) + I \left(4\,(\sigma + 2\sigma_3 - 2\sigma_3 \cos\theta_3) \sum_{m=1}^{2} c_m \sin\theta_m \right) \tag{27}$$

and

$$\overline{a}_2 a_2 - \overline{a}_0 a_0 = 4\,(\sigma + 2\sigma_3 - 2\sigma_3 \cos\theta_3)\,. \tag{28}$$

Note that in the pure advection case the first reduced polynomial vanishes, because then $\sigma_m = 0$ for $m = 1, 2, 3$.

In the remainder of this section we will prove and discuss two stability theorems. Theorem 1 deals with the case where horizontal diffusion is absent ($\epsilon_1 = 0, \epsilon_2 = 0$ and $\epsilon_3 \geq 0$). In Theorem 2 we consider the remaining cases where diffusion exists in at least one of the two horizontal directions ($\epsilon_1 \geq 0, \epsilon_2 \geq 0, \epsilon_3 \geq 0$ and $\epsilon_1 + \epsilon_2 > 0$). In both theorems all velocities c_m may take on arbitrary values, including zero.

THEOREM 1 *Suppose* $\epsilon_1 = 0, \epsilon_2 = 0$ *and* $\epsilon_3 \geq 0$. *Then we have von Neumann stability if and only if*

$$|c_1| + |c_2| \leq 1. \tag{29}$$

PROOF. We distinguish the two cases $\epsilon_3 = 0$ and $\epsilon_3 > 0$. First suppose $\epsilon_3 = 0$. Then the first reduced polynomial $f_1 \equiv 0$, so that according to case (ii) of Th.

6.1 from [6], there holds $|\xi_1|, |\xi_2| \leq 1$, if and only if the root ξ_0 of the derivative polynomial f' satisfies $|\xi_0| \leq 1$. Since $\xi_0 = -a_1/2a_2$ we find

$$|\xi_0|^2 = \frac{\left(\sum_{m=1}^{2} c_m \sin\theta_m\right)^2}{1 + c_3^2 \sin^2\theta_3}, \tag{30}$$

which immediately proves the theorem for the case $\epsilon_3 = 0$. Next suppose $\epsilon_3 > 0$. Two subcases then must be distinguished, viz. phase angle $\theta_3 = 0$ and $\theta_3 \neq 0$. If $\theta_3 = 0$, then again $f_1 \equiv 0$ and the proof goes the same as above. If $\theta_3 \neq 0$, then f_1 does not vanish so that now case (i) of Th. 6.1 from [6] applies. That is, $|\xi_1|, |\xi_2| \leq 1$, if and only if

$$\begin{array}{ll} (a) & |f^*(0)| > |f(0)| \text{ and} \\ (b) & \text{The root } \xi_0 \text{ of } f_1 \text{ satisfies } |\xi_0| \leq 1. \end{array}$$

Condition (a) means $|\bar{a}_2| > |a_0|$ or, according to (3.8),

$$|a_2|^2 - |a_0|^2 = \bar{a}_2 a_2 - \bar{a}_0 a_0 = 4\left(\sigma + 2\sigma_3 - 2\sigma_3 \cos\theta_3\right) > 0. \tag{31}$$

We immediately conclude that condition (a) is unconditionally true because the diffusion parameter σ_3 is positive and $\sigma = 0$. Generally, condition (b) is true if and only if

$$\begin{aligned} &\Big| -2\sum_{m=1}^{2} \sigma_m \cos\theta_m \; + I\left(2c_3 \sin\theta_3 \sum_{m=1}^{2} \sigma_m \cos\theta_m\right) + \\ &I\left(\left(\sigma + 2\sigma_3 - 2\sigma_3 \cos\theta_3\right) \sum_{m=1}^{2} c_m \sin\theta_m\right) \Big| \leq \sigma + 2\sigma_3 - 2\sigma_3 \cos\theta_3. \end{aligned} \tag{32}$$

Because $\sigma_1 = \sigma_2 = 0$ and $\sigma_3 > 0$, this inequality simply means that

$$\Big|\sum_{m=1}^{2} c_m \sin\theta_m\Big| \leq 1,$$

which immediately proves the theorem also for the case $\epsilon_3 > 0$. □

In the situation of Theorem 1 the Du Fort-Frankel scheme is absent in (2.14), so that only the leap-frog scheme and the Crank-Nicholson scheme as combined in (2.14) play a role. Theorem 1 nicely shows this. We see that the critical stepsize for von Neumann stability is determined by the familiar CFL condition of the leap-frog scheme (2.18),

$$\tau\left(\frac{|q_1|}{h_1} + \frac{|q_2|}{h_2}\right) \leq 1. \tag{33}$$

This is an optimal result in the sense that the vertical velocity q_3 and the vertical mesh width h_3 are absent in the stability condition, which is due to the unconditional stability of the Crank-Nicholson scheme. Especially h_3 should be absent, since in shallow water transport problems h_3 is significantly smaller than h_1 and h_2. This, in fact, was the motivation for developing the odd-even-line hopschotch method [12, 13]. Also note that in the case of pure advection ($\epsilon_m = 0, m = 1,2,3$) the characteristic polynomial f is conservative ($|\xi_1| = |\xi_2| = 1$) as long as (3.13) holds (Th. 6.4, [6]). If we impose strict inequality, then f is simple conservative (conservative and $\xi_1 \neq \xi_2$, see [6], Cor. 6.5). This means that in the case of pure advection the OELH scheme does not damp Fourier modes, which is a natural property because the true Fourier modes are not damped either. If $\epsilon_3 > 0$, then one of the amplification factors must lie in the open unit disc as long as (3.13) holds, since f_1 does not vanish. If we impose strict inequality in (3.13), then both factors lie in the open unit disc which means damping of Fourier modes similar as for the true solution.

Before we present Theorem 2, we first give a result due to [5] and repeat its proof here for reasons of self-containedness.

LEMMA 1 *Consider the finite, real-valued series*

$$S = 1 - \sum_{m=1}^{M} \alpha_m \theta_m^2 + \left(\sum_{m=1}^{M} c_m \theta_m \right)^2 .$$

Suppose $\alpha_m \geq 0$ *for all* $m = 1, ..., M$. *Then we have* $S \leq 1$ *for all* θ_m, *if and only if*

$$\sum_{m=1}^{M} \frac{c_m^2}{\alpha_m} \leq 1.$$

PROOF. Denote

$$\alpha = diag(\alpha_1, \ldots, \alpha_M), \vec{c} = (c_1, \ldots, c_M)^T, \vec{\theta} = (\theta_1, \ \ldots, \ \theta_M)^T.$$

Then S can be expressed as

$$S = 1 - \vec{\theta}^T (\alpha - \vec{c}\vec{c}^T)\vec{\theta}.$$

Thus, we have $S \leq 1$ for all $\vec{\theta}$, if and only if the matrix $\beta = \alpha - \vec{c}\vec{c}^T$ is non-negative definite. In particular, its diagonal elements $\alpha_m - c_m^2$ must be non-negative, so that $\alpha_m = 0$ implies $c_m = 0$ and the m-th dimension can be dropped. Hence in the remainder of the proof we may assume all $\alpha_m > 0$. If we then define

$$\gamma = \alpha^{-1/2} = \mathrm{diag}(\alpha_1^{-1/2}, \ldots, \alpha_M^{-1/2}),$$

we have $\beta = \alpha^{1/2}(I_M - \gamma\vec{c}\vec{c}^T\gamma)\alpha^{1/2}$ and the matrix

$$\beta' = I_M - \gamma\vec{c}\vec{c}^T\gamma = I_M - (\gamma\vec{c})(\gamma\vec{c})^T = I_M - \vec{d}\vec{d}^T,$$

where $\vec{d} = \gamma\vec{c}$, must also be non-negative. This, in turn, means non-negativity of

$$\vec{z}^T\beta'\vec{z} = \vec{z}^T\vec{z} - (\vec{d}^T\vec{z})^2$$

for all $\vec{z}$. We can deduce that this is true if and only if

$$\vec{d}^T\vec{d} \leq 1.$$

Sufficiency follows immediately from the Cauchy-Schwarz inequality

$$(\vec{d}^T\vec{z})^2 \leq (\vec{d}^T\vec{d})(\vec{z}^T\vec{z})$$

and necessity by selecting $z_m = cd_m$ for $m = 1, \ldots, M$, where c is an arbitrary constant. Since $\vec{d}^T\vec{d} = \sum c_m^2/\alpha_m$, the proof is complete. □

This lemma is used to prove necessity of inequality (3.14) in Theorem 2. Note that in certain cases the sum in (3.14) is infinite (division by $\sigma_m = 0$), implying that the interval for von Neumann stability is empty. This situation is discussed in more detail later on. We wish to emphasize that the proof of this theorem is inspired by the proof of the stability theorem in [5], which also uses the result of Lemma 1.

THEOREM 2 *Suppose* $\epsilon_1, \epsilon_2, \epsilon_3 \geq 0$ *and* $\epsilon_1+\epsilon_2 > 0$. *Then we have von Neumann stability if and only if*

$$\sum_{m=1}^{3} \frac{c_m^2}{2\sigma_m/\sigma} \leq 1. \tag{34}$$

PROOF. Because $\sigma > 0$, the first reduced polynomial f_1 does not vanish so that case (i) of Th. 6.1 from [6] applies, similar as in the second part of the proof of Theorem 1 above. Hence, $|\xi_1|$, $|\xi_2| \leq 1$, if and only if inequalities (3.11) and (3.12) are true. We immediately conclude that inequality (3.11) is unconditionally true, because $\sigma > 0$ and $\sigma_3 \geq 0$. So our task is to check inequality (3.12). Denote

$$\begin{aligned} \sigma^\star &= \sigma + 2\sigma_3 - 2\sigma_3\cos\theta_3, \\ \sigma_m^\star &= 2\sigma_m/\sigma^\star, \quad m = 1, 2, \\ c_1^\star &= c_1, \quad c_2^\star = c_2, \quad c_3^\star = c_3 \sum_{m=1,2} \sigma_m^\star \cos\theta_m. \end{aligned}$$

Inequality (3.12) is equivalent to $|\mu| \leq 1$, where

$$\mu = \frac{\sigma}{\sigma^\star} - \sum_{m=1}^{2} \sigma_m^\star(1 - \cos\theta_m) - \sum_{m=1}^{3} Ic_m^\star \sin\theta_m. \tag{35}$$

Introduce the new diffusion parameter $\sigma_3^\star$ by writing

$$\frac{\sigma}{\sigma^\star} = 1 - \sigma_3^\star(1 - \cos\theta_3), \tag{36}$$

which implies the same expression as for $\sigma_1^\star$ and $\sigma_2^\star$,

$$\sigma_3^\star = \frac{2\sigma_3}{\sigma^\star}. \tag{37}$$

Note that for zero phase angle θ_3 the definition of $\sigma_3^\star$ through (3.16) is meaningless. However, from the limiting case

$$\sigma^\star = \sigma + \sigma_3\theta_3^2 + O(\theta_3^4), \quad \theta_3 \to 0$$

it follows, by substitution of (3.17) into (3.16), that expression (3.17) is also valid for $\theta_3 = 0$. Hence, for all phase angles we can write

$$\mu = 1 - \sum_{m=1}^{3} \sigma_m^\star(1 - \cos\theta_m) - \sum_{m=1}^{3} Ic_m^\star \sin\theta_m, \tag{38}$$

so that inequality (3.12) is true if and only if

$$|\mu|^2 = \left(1 - \sum_{m=1}^{3} \sigma_m^\star(1 - \cos\theta_m)\right)^2 + \left(\sum_{m=1}^{3} c_m^\star \sin\theta_m\right)^2 \leq 1. \tag{39}$$

Our task is now to prove that (3.14) is necessary and sufficient for (3.19). We will first establish necessity of (3.14). Consider the limiting case: $\theta_m \to 0$ with $|\theta_m| \leq \theta$ for $m = 1, 2, 3$. For $\theta_3 \to 0$ we have

$$\sigma_m^\star = \frac{2\sigma_m}{\sigma} + O(\theta_3^2) \text{ for } m = 1, 2, 3 \text{ and } c_3^\star = c_3 + O(\theta^2),$$

so that in the limiting case $|\mu|^2$ satisfies

$$|\mu|^2 = 1 - \sum_{m=1}^{3} \frac{2\sigma_m}{\sigma}\theta_m^2 + \left(\sum_{m=1}^{3} c_m\theta_m\right)^2 + O(\theta^4). \tag{40}$$

Set $\alpha_m = 2\sigma_m/\sigma$. Because $\sigma > 0$, we have $\alpha_m \geq 0$ for $m = 1, 2, 3$ and application of Lemma 1 immediately reveals the necessity of (3.14). In particular, if a $\alpha_m = 0$, then the corresponding c_m must be zero too, which means that the dimension is dropped. Hence, in the sufficiency part of the proof we will assume that all α_m are positive and observe that for a lower dimension the proof of sufficiency goes entirely similar.

To prove sufficiency of (3.14) we proceed as follows. Write

$$\sum_{m=1}^{3} c_m^\star \sin\theta_m = \sum_{m=1}^{2} \frac{c_m}{\sqrt{\alpha_m}}\sqrt{\alpha_m}\sin\theta_m + \frac{c_3}{\sqrt{\alpha_3}}\sqrt{\alpha_3}\left(\sum_{m=1}^{2} \sigma_m^\star \cos\theta_m\right)\sin\theta_3. \tag{41}$$

The Cauchy-Schwarz inequality then yields

$$\left(\sum_{m=1}^{3} c_m^\star \sin\theta_m\right)^2 \le \left(\sum_{m=1}^{3} \frac{c_m^2}{\alpha_m}\right)$$
$$\left(\sum_{m=1}^{2} \alpha_m \sin^2\theta_m + \alpha_3 \left(\sum_{m=1}^{2} \sigma_m^\star \cos\theta_m\right)^2 \sin^2\theta_3\right). \tag{42}$$

Set $y_m = \cos\theta_m$ and invoke (3.14). Using $\alpha_1 + \alpha_2 = 1$, we then can write

$$\left(\sum_{m=1}^{3} c_m^\star \sin\theta_m\right)^2 \le 1 - \alpha_1 y_1^2 - \alpha_2 y_2^2 + \alpha_3 \left(\sigma_1^\star y_1 + \sigma_2^\star y_2\right)^2 (1 - y_3^2). \tag{43}$$

Further, using $\sigma^\star = \sigma + 2\sigma_3(1 - y_3)$, we have

$$\left(1 - \sum_{m=1}^{3} \sigma_m^\star (1 - \cos\theta_m)\right)^2 = \frac{1}{\sigma^{\star 2}} \left(2\sigma_1 y_1 + 2\sigma_2 y_2\right)^2, \tag{44}$$

so that there remains to prove

$$|\mu|^2 \le 1 + \frac{1}{\sigma^{\star 2}} \left(2\sigma_1 y_1 + 2\sigma_2 y_2\right)^2 + \alpha_3$$
$$\left(\sigma_1^\star y_1 + \sigma_2^\star y_2\right)^2 (1 - y_3)^2 - \alpha_1 y_1^2 - \alpha_2 y_2^2 \le 1 \tag{45}$$

for all $y_m \in [-1, 1]$, $m = 1, 2, 3$. Define $\vec{y} = (y_1, y_2)^T$ and $Y = \alpha_3(1 - y_3^2)$. Then the second inequality can be rewritten as

$$\vec{y}^T A \vec{y} \le 0, \tag{46}$$

where A is a symmetric two-by-two matrix with the entries

$$A_{11} = \frac{4(Y+1)}{\sigma^{\star 2}} \sigma_1^2 - \frac{2\sigma_1}{\sigma}, \quad A_{12} = \frac{4(Y+1)}{\sigma^{\star 2}} \sigma_1 \sigma_2,$$
$$A_{22} = \frac{4(Y+1)}{\sigma^{\star 2}} \sigma_2^2 - \frac{2\sigma_2}{\sigma}. \tag{47}$$

Note that the entries do depend on y_3, but not on $\vec{y}$. Hence, it is sufficient that A is non-positive definite for all $y_3 \in [-1, 1]$. Because $A_{12} > 0$, A is non-positive definite if

$$A_{11} + A_{12} \le 0 \quad \text{and} \quad A_{22} + A_{12} \le 0.$$

A trivial calculation shows that this is indeed the case for all $y_3 \in [-1, 1]$, which completes the proof of the theorem. □

Any case covered by Theorem 2 involves the Du Fort-Frankel scheme in (2.14) since $\sigma > 0$. We emphasize that this gives rise to curious and unexpected

stability results. Substitution of σ_m, c_m in (3.14) shows that the critical stepsize for von Neumann stability in all cases covered by Theorem 2 is determined by

$$\tau^2 \left(\sum_{m=1}^{3} \frac{q_m^2}{\epsilon_m} \quad \sum_{l=1}^{2} \frac{\epsilon_l}{h_l^2} \right) \leq 1. \tag{48}$$

First, we see that the vertical meshwidth h_3 is absent, which is advantageous as we explained in the discussion of Theorem 1. Second, for zero velocities (the pure diffusion case) we have unconditional stability, which is in complete agreement with the unconditional stability of the Du Fort-Frankel scheme (2.19) and the Crank-Nicholson scheme (2.20). However, if a velocity is not zero, then the corresponding diffusion parameter plays a role. Surprisingly, the critical stepsize determined by (3.28) is generally smaller than the one determined by the CFL condition (3.13) and in fact can be zero.

To see this, let us first suppose that $\epsilon_1, \epsilon_2, \epsilon_3$ are positive. Application of the Cauchy-Schwarz inequality to the CFL condition (3.13) then leads to (3.28) as follows,

$$\left(\sum_{l=1}^{2} \frac{\tau |q_l|}{h_l} \right)^2 = \left(\sum_{l=1}^{2} \frac{\tau |q_l| \sqrt{\epsilon_l}}{h_l \sqrt{\epsilon_l}} \right)^2 \leq$$

$$\sum_{l=1}^{2} \frac{\tau^2 q_l^2}{\epsilon_l} \quad \sum_{l=1}^{2} \frac{\epsilon_l}{h_l^2} \quad \sum_{m=1}^{3} \frac{\tau^2 q_m^2}{\epsilon_m} \quad \sum_{l=1}^{2} \frac{\epsilon_l}{h_l^2} \leq 1. \tag{49}$$

Generally (3.28) appears to be more restrictive, implying a smaller critical stepsize. We consider this curious because it means, for example, that adding artificial diffusion to the advection problem can have a destabilizing effect for the time integration, rather than working out stabilizing. A similar curious situation has been observed earlier in [10, 14]. Also note that if the three diffusion parameters are equal, then they cancel out in (3.28) so that the critical stepsize then even is independent of the diffusion, but yet smaller than in the case of the CFL condition. Of course, the difference between the two conditions is minor if

$$\frac{|q_1| h_1}{\epsilon_1} \approx \frac{|q_2| h_2}{\epsilon_2} \quad \text{and} \quad \frac{q_3^2}{\epsilon_3} << \min \left(\frac{q_1^2}{\epsilon_1}, \frac{q_2^2}{\epsilon_2} \right). \tag{50}$$

The observation that for cases covered by Theorem 2 the critical stepsize can even be zero, follows directly from inspection of (3.28). For example, if we take $q_1, q_2, q_3 \neq 0$, ϵ_1, ϵ_2 fixed and $\epsilon_3 \to 0$, then $\tau \to 0$ when satisfying the stability inequality. By also taking into account Theorem 1, we thus can formulate:

THEOREM 3 *For von Neumann stability it is necessary that either both ϵ_1 and ϵ_2 are zero or positive and if they are both positive, then it is required to have $\epsilon_3 > 0$ too.*

4 THE DU FORT-FRANKEL DEFICIENCY

We will further explain this curious stability result by relating it with the well-known Du Fort-Frankel deficiency, which describes the situation that for parabolic problems this method is only conditionally convergent, in spite of its unconditional stability (see [9], Sect.7.5).

The necessity of (3.14) or (3.28) has been established from the asymptotic relation (3.20) where all three phase angles $\theta_m \to 0$. This suggests to compute for this limiting case the maximum of the absolute value of the two amplification factors ξ_1, ξ_2 directly from the polynomial (3.2). Denote $\xi_{max} = \max(|\xi_1|, |\xi_2|)$. An elementary calculation then yields

$$\xi_{max} = 1 - \sum_{m=1}^{3} \sigma_m \theta_m^2 + \tfrac{1}{2}\sigma \left(\sum_{m=1}^{3} c_m \theta_m \right)^2 + O(\theta^3). \tag{51}$$

Indeed, use of Lemma 1 shows again the necessity of (3.14). However, expression (4.1) also reveals a link with the aforementioned convergence deficiency. To see this, consider the modified equation for scheme (2.14) (cf. [9], Sect. 7.5),

$$u_t + q_1 u_x + q_2 u_y + q_3 u_z = \epsilon_1 u_{xx} + \epsilon_2 u_{yy} + \epsilon_3 u_{zz} - \tfrac{1}{2}\sigma\tau u_{tt}. \tag{52}$$

This modified equation shows the convergence deficiency through the additional term $-\frac{1}{2}\sigma\tau u_{tt}$. To establish the link between our stability deficiency and the convergence deficiency, it now suffices to substitute a Fourier mode into (4.2) and to compute the associated continuous amplification factor for vanishing phase angles, similar as we did in the derivation of (4.1). We then find that the continuous amplification factor just equals (4.1), up to $O(\theta^3)$. Further, it then follows that the term which causes the instability, that is,

$$\tfrac{1}{2}\sigma \left(\sum_{m=1}^{3} c_m \theta_m \right)^2, \tag{53}$$

originates from the deficiency term $-\frac{1}{2}\sigma\tau u_{tt}$, although this term itself is independent of the velocities c_m. This means that also the modified equation is unstable if (3.14) is violated, in the sense that it admits growing Fourier modes in the low frequence range. This obviously implies that this then also must happen for scheme (2.14) when subjected to the von Neumann stability test.

Noteworthy is that if we bound the phase angles from below, say $\theta_m \geq \theta_0 > 0$, that then an interval $0 < \tau \leq \tau_0$ exists for which the amplification factors ξ_1, ξ_2 are strictly less than one. This follows from expression (3.18), since its real part is independent of τ and can be made < 1 by taking θ_0 sufficiently small, while the imaginary part can be made sufficiently small by taking τ_0 small enough. Hence, if we consider a fixed grid, then we can always achieve stability, but of course τ_0 becomes smaller if the grid is refined.

5 PRACTICAL CONSIDERATIONS

Strict von Neumann stability is known to have great practical relevance. There is no doubt that the von Neumann method is the best single technique (cf. [5]) for finding necessary conditions for stability if we are in a non-model situation, which in practice of course always happens. In this connection a natural question is, how bad actually is the stability deficiency for the OELH scheme. In other words, should we in practice consider the CFL condition (3.13) as a 'practical restriction', or should we take the more stringent condition (3.28) really serious.

Let τ_{cfl} and $\tau_{(3.28)}$ denote the critical stepsizes. Because the necessity of condition (3.14) shows up in the limiting case $\theta_m \to 0$, the maximum ξ_{max} as derived in (4.1) will be only marginally larger than one if $\tau_{(3.28)} < \tau \leq \tau_{cfl}$. However, there is a possibility that other critical combinations of phase angles exist, away from zero, which also lead to (3.14). Therefore we have computed approximate values of ξ_{max} (the maximum taken over all discrete θ-values) as a function of τ for several choices of ϵ_m, q_m, h_m. We indeed observed other critical θ-combinations away from zero. Yet, in all tests ξ_{max} appeared to become only marginally larger than one in the stepsize range $\tau_{(3.28)} < \tau \leq \tau_{cfl}$, similar as in the limiting case which led to (3.14).

Figure 1 shows a plot of $\xi_{max}(\tau)$ which is characteristic for the tests considered. We see that the overshoot due to violating (3.28) is practically insignificant. In the interval $\tau_{(3.28)} < \tau \leq \tau_{cfl}$ the overshoot of $\xi_{max}(\tau)$ is ≤ 0.001. However, as expected, we also see that $\tau > \tau_{cfl}$ will quickly result in severe instability. The fact that the CFL condition should be satisfied in general, thus also in all cases covered by Theorem 2, can be understood by computing (3.18) for special choices of the θ_m. For example, for $\theta_m = \frac{\pi}{2}, m = 1, 2, 3$, we get

$$\mu = 1 - \sum_{m=1}^{3} \sigma_m^\star - \sum_{m=1}^{3} I c_m^\star = -I(c_1 + c_2), \tag{54}$$

which trivially yields the CFL condition (3.13) for positive c_1, c_2 (cf. (3.19)).

We conclude that the more stringent condition (3.28) is only a theoretical curiosity. For the actual practice it will be of little importance since the instability that will occur by violation is so small that it will not be observed in actual computation, of course as long as the CFL condition (3.13) is satisfied. This condition is highly relevant for the actual practice and should always be obeyed. On the other hand, violation of (3.28) will only be noticable after an unrealistically large number of time steps. To illustrate this in actual integration, we applied the OELH integrator to the model equation (2.1), discretized on a uniform 40x40x10 grid, using periodic boundary conditions. The parameters in this experiment were set to the same values as in Figure 1 and the grid sizes to $(h_1, h_2, h_3) = (500, 500, 10)$. These values yield

$$\tau_{cfl} = 100.0, \quad \tau_{(3.28)} = 37.7.$$

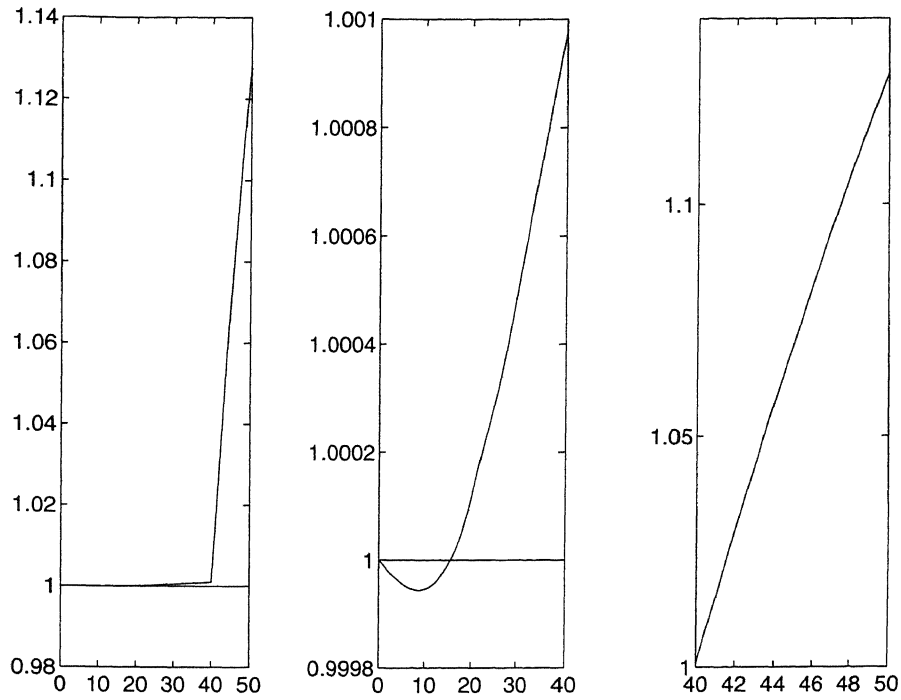

FIGURE 1. Plots of $\xi_{max}(\tau)$ for the parameters $(\epsilon_1, \epsilon_2, \epsilon_3) = (1.0, 0.5, 0.01)$, $(q_1, q_2, q_3) = (3, 2, 1)$. The grid sizes are $(h_1, h_2, h_3) = (200, 200, 1)$. This yields $\tau_{(3.28)} \approx 15.1$ and $\tau_{cfl} = 40.0$. The left plot covers the τ-interval $0 \leq \tau \leq 50$, the middle plot $0 \leq \tau \leq \tau_{cfl}$ and the right plot $\tau_{cfl} \leq \tau \leq 50$. The middle and the right plot show a finer scale in the vertical.

Obviously, $u \equiv 1$ is an exact solution for the test model. To study the long-term stability behaviour of the OELH method, we slightly perturbed the initial condition to $u(x, y, z) = 1.0 + \delta g(x, y, z)$, with g a smooth function with maximum modulus equal to 1.0 and $\delta = 10^{-5}$. Table 1 contains the values of the experimental amplification factors

$$\delta^{-1} \max_{\underline{i}} |U_{\underline{i}}^N - 1| \tag{55}$$

for various values of τ and N. Here $U_{\underline{i}}^N$ denotes the numerical solution at grid point $\underline{i}$ after N steps of length τ. The results are self evident. Violation of the CFL condition is disastrous, whereas violation of (3.28) leads to error growth, but only destroys the solution after an unrealistically large number of time steps.

	$\tau = 37$	$\tau = 100$	$\tau = 100.1$
$N = 10^4$	0.724	3.68	10^{185}
$N = 10^5$	0.497	870	
$N = 5.10^5$	0.362	10^{20}	

Table 1: Experimental amplification factors (5.2).

Finally, it is also of interest to recall the convergence deficiency, from which the OELH scheme also suffers. Presumably, this convergence deficiency is also of little relevance for the shallow water transport application. In this application the regular temporal and spatial truncation errors are expected to be larger than the error induced by the parasitic, non-physical term $\frac{1}{2}\sigma\tau u_{tt}$. For example, in the experiments reported in [12, 13] this error plays no role. Experiments where this error is shown, though, can be found in [14].

REFERENCES

1. A.R. Gourlay (1970). *Hopscotch: A Fast Second Order Partial Differential Equation Solver.* J. Inst. Math. Appl., 6, 375 - 390.
2. A.R. Gourlay (1971). *Some Recent Methods for the Numerical Solution of Time-Dependent Partial Differential Equations.* Proc. Roy. Soc. London A 323, 219 - 235.
3. A.R. Gourlay, J. LI. Morris (1972). *Hopscotch Difference Methods for Nonlinear Hyperbolic Systems.* IBM J. Res. Develop., 16, 349 - 353.
4. A.R. Gourlay (1977). *Splitting Methods for Time-Dependent Partial Differential Equations.* In: D. Jacobs, ed., The State of the Art in Numerical Analysis, Academic Press, 757 - 791.
5. A.C. Hindmarsh, P.M. Gresho, D.F. Griffiths (1984). *The Stability of Explicit Euler Time-Integration for Certain Finite-Difference Approximations of the Multi-Dimensional Advection-Diffusion Equation.* Int. J. Numer. Meth. in Fluids, 4, 853 - 897.
6. J.J.H. Miller (1971). *On the Location of Zeros of Certain Classes of Polynomials with Applications to Numerical Analysis.* J. Inst. Maths. Applics., 8, 397 - 406.

7. K.W. Morton (1980). *Stability of Finite Difference Approximations to a Diffusion-Convection Equation.* Int. J. Numer. Meth. in Engn. 15, 677 - 683.
8. J. Von Neumann, R.D. Richtmyer (1950). *A Method for the Numerical Calculations of Hydrodynamical Shocks.* J. Appl. Phys. 21, 232 - 237.
9. R.D. Richtmyer, K.W. Morton (1967). *Difference Methods for Initial Value Problems.* Interscience Publishers, New York.
10. U. Schumann (1975). *Linear Stability of Finite-Difference Equations for Three-Dimensional Flow Problems.* J. Comput. Phys., 18, 465 - 470.
11. J. Schur (1918). *Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind.* J. Reine u. angew. Math. 147, 205 - 232.
12. B.P. Sommeijer, P.J. van der Houwen, J. Kok (1993). *Time Integration of Three-Dimensional Numerical Transport Models.* Report NM-R9316, Centre for Mathematics and Computer Science, Amsterdam (to appear in Appl. Numer. Math.).
13. B.P. Sommeijer, J. Kok (1994). *Implementation and Performance of a Three-Dimensional Numerical Transport Model.* Report NM-R9402, Centre for Mathematics and Computer Science, Amsterdam (to appear in Int. J. Numer. Meth. in Fluids).
14. J.H.M. ten Thije Boonkkamp, J.G. Verwer (1987). *On the Odd-Even Hopscotch Scheme for the Numerical Integration of Time-Dependent Partial Differential Equations.* Appl. Numer. Math., 3, 183 - 193.