

On the negative weighting factors in the Muskingum-Cunge scheme

Sur les facteurs de pondération négatifs dans le schéma Muskingum-Cunge

SÁNDOR SZÉL, *Water Resources Research Centre, Kvassay Jenő út 1, H-1095 Budapest, Hungary*

CSABA GÁSPÁR, *Széchenyi István College, Department of Mathematics, P.O.Box 701, H-9007 Győr, Hungary*

ABSTRACT

The Muskingum-Cunge scheme applied to the one-dimensional unsteady advection-diffusion equation is investigated. To eliminate the numerical diffusion, the coefficients of the scheme are defined in such a way that the scheme does not contain the weighting parameters explicitly, but the Courant and Péclet numbers only. If one of the weighting factors is prescribed, the other should be necessarily negative in a lot of cases, which does not affect the applicability of the scheme. It is shown that the accuracy can be increased further, the numerical oscillations can also be eliminated by prescribing a simple relationship between the Courant and Péclet numbers. Sufficient conditions for strong stability are also presented.

RÉSUMÉ

Dans l'article on analyse le schéma Muskingum-Cunge appliqué à l'équation unidimensionnelle de convection-diffusion. Afin éliminer la diffusion numérique les coefficients du schéma sont exprimés exclusivement en termes des Nombres de Courant et de Péclet et non pas en termes habituels contenant explicitement les paramètres de pondération. La définition arbitraire de l'un des facteurs de pondération entraîne, dans bien des cas, des valeurs négatives d'autres facteurs sans que l'applicabilité du schéma soit pour autant invalidée. On démontre que la précision peut être améliorée au-delà et que des oscillations numériques peuvent être éliminées en imposant une relation simple entre les Nombres de Courant et de Péclet. On présente aussi les conditions fortes de la stabilité numériques.

Introduction

The Muskingum method is a traditional method to solve the kinematic wave equation discretised by finite differences. It was modified by Cunge [1] who used weighting parameters in such a way that the numerical diffusion of the scheme is equal to the physical diffusion, so that the scheme is consistent with the diffusion wave equation. See also [2]. Usually, non-negative weighting parameters are defined: however, there are cases when certain negative parameters give the optimal approximation, see [3]. The use of negative parameters seems to suffer from physical meaningless. Nevertheless, as we will show it, no contradiction arises from this definition and, in a lot of cases, this choice is the only way to avoid the false diffusion and preserve stability and/or accuracy. We will also show that, in the final form of the scheme, the weighting parameters play no role: instead, the properties of the scheme can be characterised by the Courant and Péclet numbers only. To illustrate the approach, simple numerical examples are also presented.

The Muskingum-Cunge scheme

We derive the Muskingum-Cunge scheme through the example of the unsteady advection-diffusion equation with constant coefficients. The equation can be written in the following form:

$$\frac{\partial u}{\partial t} + C \frac{\partial u}{\partial x} - D \frac{\partial^2 u}{\partial x^2} = 0 \quad (1)$$

where $u = u(t, x)$ is the physical quantity, the transport of which is under consideration (e.g. concentration; the discharge in the diffusion wave model etc.). C is the advective velocity, D stands for the diffusion coefficient. Introducing the differential operator

$$Lu := \frac{\partial u}{\partial t} + C \frac{\partial u}{\partial x} - D \frac{\partial^2 u}{\partial x^2} = 0 \quad (2)$$

Equation (1) can be written in the more compact form

$$Lu = 0. \quad (3)$$

For simplicity, Equation (3) is defined if the quarter-plane defined by the inequalities $t > 0, x > 0$. Along the boundary, initial as well as boundary conditions have to be imposed.

In the Muskingum-Cunge scheme, the differential operator L is approximated by the following finite difference operator:

$$L_{\Delta t, \Delta x} := C \cdot \frac{\varepsilon(u_{k+1}^j - u_k^j) + (1 - \varepsilon)(u_{k+1}^{j+1} - u_k^{j+1})}{\Delta x} + \frac{\Theta(u_k^{j+1} - u_k^j) + (1 - \Theta)(u_{k+1}^{j+1} - u_{k+1}^j)}{\Delta t} \quad (4)$$

where the scheme is defined on the grid $(t_j, x_k) := (j \cdot \Delta t, k \cdot \Delta x)$ ($j, k = 0, 1, 2, \dots$), and $\Delta t, \Delta x$ are the applied discrete steps in time and space, respectively. The dimensionless numbers ε and θ are weighting factors, which will be defined later. In the expression (4), u means the sequence of the discrete values u_k^j , which should approximate the values $u(t_j, x_k)$ ($j, k = 0, 1, 2, \dots$). Thus, the discretised form of Equation (3) is the following system of algebraic linear equations:

$$L_{\Delta t, \Delta x} u = 0.$$

Simple calculations show that this leads to the explicit recursion:

$$u_{k+1}^{j+1} := C_1 u_k^j + C_2 u_{k+1}^j + C_3 u_k^{j+1}, \quad (5)$$

Revision received November, 1999. Open for discussion till February 28, 2001.

where the coefficients C_1, C_2, C_3 have the following forms:

$$C_1: = \frac{\frac{C\varepsilon}{\Delta x} + \frac{\Theta}{\Delta t}}{\frac{C(1-\varepsilon)}{\Delta x} + \frac{1-\Theta}{\Delta t}}, C_2: = \frac{-\frac{C\varepsilon}{\Delta x} + \frac{1-\Theta}{\Delta t}}{\frac{C(1-\varepsilon)}{\Delta x} + \frac{1-\Theta}{\Delta t}}, \quad (6)$$

$$C_3: = \frac{\frac{C(1-\varepsilon)}{\Delta x} - \frac{\Theta}{\Delta t}}{\frac{C(1-\varepsilon)}{\Delta x} + \frac{1-\Theta}{\Delta t}}$$

It should be emphasized that the formula in the right-hand side of Equation (4) is a linear combination of *first-order* finite differences with respect to both space and time variable. It does *not* contain any explicit approximation of second-order derivatives, consequently, the physical diffusion coefficient D does not appear in the scheme. Instead, the parameters ε and θ will be chosen in such a way that the false (numerical) diffusion of the scheme compensate the lack of the physical diffusion D in (4). As will be shown, this is not always possible. Roughly speaking, the condition of applicability is that Equation (1) is not diffusion-dominated. In addition to it, there are some restrictions also to the choice of the space and time steps Δx and Δt .

Now we are going to analyse the scheme (4). The aim is to approximate the second-order advection-diffusion operator L by the discrete operator $L_{\Delta t, \Delta x}$ which, however, contains difference schemes of first-order derivatives only.

Approximation

To investigate the approximation properties of the scheme (4), we invoke the Taylor formula in two variables. If $u = u(t, x)$ is smooth enough, the following formula is valid:

$$u(t_j + \tau, x_k + \xi) = u + u_t \tau + u_x \xi + \frac{1}{2} u_{tt} \tau^2 + u_{tx} \xi \tau + \frac{1}{2} u_{xx} \xi^2 + \frac{1}{6} u_{ttt} \tau^3 + \frac{1}{2} u_{ttx} \tau^2 \xi + \frac{1}{2} u_{txx} \tau \xi^2 + \frac{1}{6} u_{xxx} \xi^3 + \dots \quad (7)$$

where, for simplicity, we have applied the following notations:

$$u = u(t_j, x_k), u_t = \frac{\partial u}{\partial t}(t_j, x_k), u_{tt} = \frac{\partial^2 u}{\partial t^2}(t_j, x_k),$$

and so forth.

Applying the expansion formula (7) in the two-dimensional points $(t_j, x_{k+1}), (t_j, x_{k-1}), (t_{j+1}, x_{k+1})$, (i.e. by setting $\tau = 0, \xi = \Delta x; t = \Delta t, \xi = 0; \tau = \Delta t, \xi = \Delta x$) after some algebraic manipulations we obtain that the discrete operator $L_{\Delta t, \Delta x}$ (applied to the function u) can be rewritten as

$$L_{\Delta t, \Delta x} u = C u_x + u_t + \frac{1}{2} [u_{xx} \cdot C \Delta x + 2 u_{tx} \cdot ((1-\varepsilon) \cdot C \Delta t + (1-\Theta) \cdot \Delta x) + u_{tt} \Delta t] + \frac{1}{6} [u_{xxx} \cdot C \Delta x^2 + 3 u_{xxt} \Delta x \cdot ((1-\varepsilon) \cdot C \Delta t + (1-\Theta) \cdot \Delta x) + 3 u_{xtt} \Delta t \cdot ((1-\varepsilon) \cdot C \Delta t + (1-\Theta) \cdot \Delta x) + u_{ttt} \cdot \Delta t^2] + O((\Delta x + C \Delta t)^3) \quad (8)$$

Observe that ε and θ always appear in the same expression. Introducing the notation $\alpha = (1-\varepsilon) \cdot C \Delta t + (1-\Theta) \cdot \Delta x$, and replacing the first two terms $C u_x + u_t$ with $Lu + Du_{xx}$, (cf. Equation (2)), Equation (8) has the following simpler form:

$$L_{\Delta t, \Delta x} u = Lu + \frac{1}{2} [u_{xx} \cdot (2D + C \Delta x) + 2 \alpha u_{tx} + u_{tt} \cdot \Delta t] + \frac{1}{6} [u_{xxx} \cdot C \Delta x^2 + 3 \alpha u_{xxt} \cdot \Delta x + 3 \alpha u_{xtt} \cdot \Delta t + u_{ttt} \cdot \Delta t^2] + O((\Delta x + C \Delta t)^3) \quad (9)$$

This equation allows for an *error estimation* and shows how exactly the advection-diffusion operator L is approximated by the discrete operator $L_{\Delta t, \Delta x}$ on the function u , which is assumed to be smooth enough but no additional condition is required. The second and third term in the right-hand side vanish in first and second order, respectively, as both Δt and Δx tend to zero.

Remark: Note again that here a *second-order* (advection-diffusion) operator is approximated by a discrete operator which contains *first-order* discretised derivatives only, i.e. it is apparently inconsistent with the differential operator (2). However, it can be made consistent by requiring some additional assumptions to the parameters of the scheme.

From Equation (9) it is clear that the terms in the right-hand side are not identically zero for arbitrary function u . In other words, it is impossible to choose ε and θ in such a way that even at least the first-order error term vanishes for *arbitrary* smooth function u . If, however, the function u is a solution of the differential equation (3), the estimation can be significantly improved. In this case, the derivative of u with respect to t can be expressed by the derivatives with respect to x only:

$$u_t = -C u_x + D u_{xx} \quad (10)$$

Using Equation (10) repeatedly, each term in the right-hand side of Equation (9) can be expressed in a similar way:

$$u_{tx} = -C u_{xx} + D u_{xxx}$$

$$u_{tt} = C^2 u_{xx} - 2CD u_{xxx} + D^2 u_{4x}$$

$$u_{ttx} = -C u_{xxx} + D u_{4x}$$

$$u_{ttx} = C^2 u_{xxx} - 2CD u_{4x} + D^2 u_{5x}$$

$$u_{ttt} = -C^3 u_{xxx} + 3C^2 D u_{4x} - 3CD^2 u_{5x} + D^3 u_{6x}$$

where $u_{4x} := u_{xxxx}$, and so on. Thus, Equation (9) can be rewritten as:

$$L_{\Delta t, \Delta x} u = Lu + \left[D + \frac{C\Delta x}{2} - C\alpha + \frac{C^2\Delta t}{2} \right] \cdot u_{xx} + \left[D\alpha - CD\Delta t + \frac{C\Delta x^2}{6} - \frac{C\alpha\Delta x}{2} + \frac{C^2\alpha\Delta t}{2} - \frac{C^3\Delta t^2}{6} \right] \cdot u_{xxx} + \dots \quad (11)$$

Equation (11) is again an error expression, which contains the derivatives of u with respect to x only. If the original Equation (1) is advection-dominated, then the constant D is small: more precisely, it should be assumed that D and $C\Delta x$ are of the same order. In this sense, the neglected terms in the right-hand side of (11) are all at least of order 3. The behaviour of the approximate solution is characterised by the first two terms. Namely, the first term generates *numerical diffusion*, while the second one is responsible for the presence of *numerical oscillations*.

Elimination of the numerical dispersion. Negative weighting factors

From Equation (11), it is clear that the parameters ε , θ can be chosen in such a way that the first term in the right-hand side vanishes, that is, no extra numerical diffusion arises (i.e. the numerical diffusion generated by the scheme exactly coincides with the physical diffusion). To do this, it is necessary and sufficient that the equality

$$\alpha = (1 - \varepsilon) \cdot C\Delta t + (1 - \Theta) \cdot \Delta x = \frac{D}{C} + \frac{\Delta x}{2} + \frac{C\Delta t}{2}$$

that is,

$$(1 - \varepsilon) \cdot \frac{C}{\Delta x} + (1 - \Theta) = \frac{D}{C\Delta x} + \frac{1}{2} + \frac{C\Delta t}{2\Delta x} \quad (12)$$

is satisfied. Introducing the well-known Courant and Péclet numbers by

$$Cr := \frac{C\Delta t}{\Delta x}, Pe := \frac{C\Delta x}{2D},$$

it can be easily seen that:

Proposition 1: The scheme (5) is free from numerical diffusion (i.e. the coefficient of u_{xx} in Equation (11) vanishes) if and only if

$$\left(\frac{1}{2} - \varepsilon\right) \cdot Cr + \left(\frac{1}{2} - \Theta\right) = \frac{1}{2Pe} \quad (13)$$

Since the coefficients of the scheme can obviously be written in the form

$$C_1 = \frac{\varepsilon \cdot Cr + \Theta}{(1 - \varepsilon) \cdot Cr + (1 - \Theta)},$$

$$C_2 = \frac{-\varepsilon \cdot Cr + (1 - \Theta)}{(1 - \varepsilon) \cdot Cr + (1 - \Theta)},$$

$$C_3 = \frac{(1 - \varepsilon) \cdot Cr - \Theta}{(1 - \varepsilon) \cdot Cr + (1 - \Theta)}$$

therefore, using Equation (13), we obtain the following result:

Proposition 2: If the condition (13) is satisfied, then the coefficients in the scheme (5) depend only on the Courant and Péclet numbers rather than the weighting factors ε and θ , and can be expressed as:

$$C_1 = \frac{1 + Cr - \frac{1}{Pe}}{1 + Cr + \frac{1}{Pe}}, C_2 = \frac{1 - Cr + \frac{1}{Pe}}{1 + Cr + \frac{1}{Pe}},$$

$$C_3 = \frac{-1 + Cr + \frac{1}{Pe}}{1 + Cr + \frac{1}{Pe}} \quad (14)$$

Remarks:

1. The condition (13) assures that the simple explicit scheme (5) approximates the second-order advection-diffusion equation (1) in spite of the fact that (5) contains neither the physical diffusion coefficient nor any discretisation for second-order derivatives. However, the accuracy of the estimation (11) is inherently limited since the expansion is based on the assumption that D and $C\Delta x$ are of the same order, which is often not the case.

2. If a pure advection equation is investigated i.e. $D = 0$, then the Péclet number becomes infinity. Setting the Courant number to 1, Scheme (5) reduces simply to

$$u_{k+1}^{j+1} := u_j^k.$$

3. It is usual to investigate the numerical diffusion for the pure advection equation, which is much simpler: the conclusion is that since the scheme (5) produces more or less false diffusion, it is unpractical to apply it to pure advection problems; instead, it should be applied to advective diffusion problems, the physical diffusion of which is equal to the numerical diffusion. However, this argument is not completely correct: this means only that the discretised operator $L_{\Delta t, \Delta x}$ approximates the original differential operator L on the solutions of *another differential equation* (i.e. the pure advection equation). In spite of this fact, the condition (13) can be derived from the above considerations, but the condition of the more accurate approximation (see later) cannot.

4. Since the values of ε and θ are not uniquely determined by the condition (13) (which requires only a linear relationship between them), the weighting factors can be defined in unusual ways: they may be negative or greater than 1. For

instance, it is usual to define $\epsilon := 1/2$; in this case, the condition (13) always results in a negative value for θ whenever the Péclet number is less than 1. This is not a contradiction: the scheme may exhibit very good accuracy and stability properties, as will be shown in the next section.

Elimination of the numerical oscillation

According to the above considerations, if the condition (13) is satisfied, then the scheme produces no numerical diffusion. The accuracy can be increased further if the second term in the right-hand side of Equation (11) could be also eliminated. Since the value of α has been set by (13) (when eliminating the first error term in (11)), it is clear that this goal can be achieved only by adjusting the other parameters of the scheme (i.e. Δt , Δx or the Courant and Péclet numbers, respectively). From (11), we obtain that the numerical oscillation vanishes if and only if

$$\left[D - \frac{C\Delta x}{2} + \frac{C^2\Delta t}{2} \right] \cdot \alpha = CD\Delta t - \frac{C\Delta x^2}{6} + \frac{C^3\Delta t^2}{6} \quad (15)$$

whence

$$\left[D - \frac{C\Delta x}{2} \cdot (Cr - 1) \right] \cdot [(1 - \epsilon) \cdot Cr + (1 - \Theta)] = D \cdot Cr - \frac{C\Delta x}{6} \cdot (Cr^2 - 1)$$

Using (13), this condition can be simplified by straightforward calculations, and we obtain:

Proposition 3: Assume that the condition (13) is satisfied. Then the scheme (5) is free from numerical oscillation (i.e. the coefficient of u_{xxx} in Equation (11) vanishes) if and only if

$$Cr^2 = 1 - \frac{3}{Pe^2} \quad (16)$$

In practice this condition means that either Δt or Δx is fixed, the other is determined uniquely by (16). In particular, we have:

$$\Delta t = \frac{1}{C} \sqrt{\Delta x^2 - \frac{12D^2}{C^2}}, \Delta x = \sqrt{C^2\Delta t^2 - \frac{12D^2}{C^2}}$$

Remarks:

1. In case of pure advection, the condition (16) is equivalent to the condition $Cr = 1$. In this case neither numerical diffusion nor oscillations are generated.
2. The condition (16) cannot be satisfied in all cases: obviously it is necessary that the inequality $Pe^2 > 3$ is valid (the equality would result in zero time step). This restriction for the Péclet number is not surprising. Recall that, in the derivation of (11) it was essential that D does *not* dominate over $C\Delta x$, which means a similar restriction. It should also be pointed out that Equation (16) is a rather strong condition, which always makes the Courant number less than 1 (allowing only

very small time steps in general). In other words, if we redefine the Courant number to be greater than 1, then the condition (16) can never be satisfied: if we set $Cr := 1$, then the condition (16) is satisfied only in the case of pure advection.

Summarising the above results, the scheme (5) is free from numerical diffusion, if the condition (13) is satisfied. In this case, the coefficients of the scheme are defined by the formulas (14). Moreover, if the condition (16) is also satisfied, then the scheme is free also from numerical oscillations. This can be achieved, however, only for not too small Péclet numbers, which are greater than $\sqrt{3}$.

Weak and strong stability of the Muskingum-Cunge scheme

Roughly speaking, the stability of a scheme means that, starting from an initial and boundary condition which are bounded (in time as well as in space), the scheme produces also bounded discrete solutions. Otherwise, the numerical errors grow quickly and destroy the discrete solution. This boundedness can be mathematically defined in different ways resulting in different stability concepts.

In the following we restrict the space domain of Equation (1) to the *finite* interval $(0, a)$. Assume that this interval is divided into N equidistant subintervals, that is, $\Delta x := a/N$. Denote by $\mathbf{u}^{(j)}$ the vector of the discrete u -values in the j th time step:

$$\mathbf{u}^{(j)} := (u_1^j, u_2^j, \dots, u_N^j).$$

Suppose for the moment that the (upper) boundary condition at $x = 0$ is zero at each time step. Then, from (5), it can be easily seen that the vector of the discrete solution at the $(j + 1)$ th time step can be expressed by the vector at the j th step in the following form:

$$\mathbf{u}^{(j+1)} = C_2\mathbf{u}^{(j)} + (C_1 + C_2C_3)B\mathbf{u}^{(j)} = [C_2I + (C_1 + C_2C_3)B]\mathbf{u}^{(j)} =: A\mathbf{u}^{(j)} \quad (17)$$

where I denotes the unit matrix and B stands for the following upper triangular matrix:

$$B = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ C_3 & 1 & 0 & \dots & 0 & 0 \\ C_3^2 & C_3 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ C_3^{N-2} & C_3^{N-3} & C_3^{N-4} & \dots & 1 & 0 \end{bmatrix} \quad (18)$$

It is well known that for the stability of the scheme, it is necessary that the absolute value of each eigenvalue is less than 1. Since the matrix B is idempotent (that is, its N th power is the zero matrix), it possesses only zero eigenvalues, consequently, all eigenvalues of A are equal to C_2 . Thus, from (14) it is obvi-

ous that if the advective velocity is positive i.e. $C > 0$ and the condition (13) is satisfied then

$$|C_2| < 1, \tag{19}$$

consequently, the discrete solution is bounded with respect to time.

Remark: If $C < 0$, then the scheme (5) is unstable in general, as can be shown through very simple examples. Practically this means that the recursion (5) can be performed in the direction of the advective velocity but not in the opposite direction. This fact points out the disadvantage of the scheme: no lower boundary conditions can be imposed, though the original differential equation does require it. However, if the transport in advection-dominated (with higher Péclet number), this phenomenon affects the solution only in a narrow vicinity of the lower boundary.

The other case (when the initial condition is identically zero) can be analysed in quite a similar way. Now suppose that, in the time domain, Equation (1) is investigated only on the finite interval $(0, T)$ which is subdivided into N timesteps. In this case, denote by $\mathbf{u}^{(k)}$ the vector of the discrete u -values taken at the k^{th} gridpoint:

$$\mathbf{u}^{(k)} := (u_{k,1}^1, u_{k,2}^2, \dots, u_{k,N}^N).$$

then it can be easily seen that

$$\mathbf{u}^{(k+1)} = C_3 \mathbf{u}^{(k)} + (C_1 + C_2 C_3) B \mathbf{u}^{(k)} \tag{20}$$

where

$$B = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ C_2 & 1 & 0 & \dots & 0 & 0 \\ C_2^2 & C_2 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ C_2^{N-2} & C_2^{N-3} & C_2^{N-4} & \dots & 1 & 0 \end{bmatrix} \tag{21}$$

The formulas (20)–(21) are analogous to the formulas (17)–(18): the only difference that the roles of C_2 and C_3 are replaced. Therefore for the stability it is necessary that the inequality

$$|C_3| < 1, \tag{22}$$

is valid. From (14) it follows that the condition (22) is always satisfied, whenever $C > 0$.

The general case can be derived easily from the above special cases. It is clear that every solution u of the original differential equation (1) can be expressed in the form $u = u_1 + u_2$, where

both u_1 and u_2 are solutions of (1), and u_1 satisfies the initial condition and vanishes at the upper boundary $x = 0$; in contrast to this, u_2 satisfies the upper boundary condition and vanishes in the time $t = 0$.

The conditions (19) and (22) are also *sufficient* conditions for stability. It is well known that if all eigenvalues of a matrix A are placed within the unit circle, then the powers of A tend elementwise to zero, therefore the iteration $\mathbf{u}^{(n+1)} = A\mathbf{u}^{(n)}$ results in a bounded vector sequence. However, the bounds of these sequences may be extremely large, therefore the scheme can be often considered practically unstable, though it produces bounded discrete solutions, if the conditions (19), (22) are fulfilled. This shows that the above stability concept is rather weak, so it should be strengthened.

Now let us turn to the first special case and investigate the following much stronger stability concept: the discrete solution is said to be stable, if some norm of the discrete solution does not increase in time, that is:

$$\|\mathbf{u}^{(j+1)}\| \leq \|\mathbf{u}^{(j)}\| \tag{23}$$

We will use the maximum norm of vectors defined by

$$\|\mathbf{u}\| := \max(|u_1|, |u_2|, \dots, |u_N|) \tag{24}$$

It is known that for arbitrary vector \mathbf{u} (having N elements) and matrix A (having N rows and N columns), the estimation

$$\|A\mathbf{u}\| \leq \|A\| \cdot \|\mathbf{u}\| \tag{25}$$

is valid, where $\|A\|$ stands for the associated matrix norm:

$$\|A\| := \max_{k=1, \dots, N} \sum_{j=1}^N |A_{kj}| \tag{26}$$

Consequently, in order that the (strong) stability condition (23) is satisfied, it is sufficient that the norm of the matrix A appearing in (17) does not exceed 1.

Proposition 4: If the condition (13) is satisfied, moreover,

$$Cr + \frac{1}{Pe} \geq 1 \quad \text{and} \quad Cr - \frac{1}{Pe} \leq 1 \tag{27}$$

then $\|A\| \leq 1$, that is, the strong stability condition is satisfied.

Proof: First we note that the first inequality of (27) implies that $C \geq 0$, which is necessary for stability. From (14), it can be easily seen that (27) is equivalent to the conditions $C_2 \geq 0$, $C_3 \geq 0$. By definition, the norm of A is as follows:

$$\|A\| = \max\{|C_2| + |C_1 + C_2 C_3| \cdot (1 + C_3 + C_3^2 + \dots + C_3^{k-2}): k = 1, 2, \dots, N\}$$

which can be obviously estimated in the following way:

$$\|A\| \leq |C_2| + |C_1 + C_2 C_3| \cdot \sum_{k=0}^{\infty} C_3^k = |C_2| + \frac{|C_1 + C_2 C_3|}{1 - C_3} \quad (28)$$

Again from (14), one can easily derive that

$$C_1 + C_2 C_3 = \frac{(Cr + 1)^2 - (Cr - 1)^2}{\left(1 + Cr + \frac{1}{Pe}\right)^2} = \frac{4Cr}{\left(1 + Cr + \frac{1}{Pe}\right)^2}$$

Substituting this expression into (28), we obtain:

$$\begin{aligned} \|A\| &\leq C_2 + \frac{1}{1 - C_3} \cdot \frac{4Cr}{\left(1 + Cr + \frac{1}{Pe}\right)^2} \\ &= \frac{1 - Cr + \frac{1}{Pe}}{1 + Cr + \frac{1}{Pe}} + \frac{1 - Cr + \frac{1}{Pe}}{2} \cdot \frac{4Cr}{\left(1 + Cr + \frac{1}{Pe}\right)^2} \\ &= \frac{1 - Cr + \frac{1}{Pe}}{1 + Cr + \frac{1}{Pe}} + \frac{2Cr}{1 + Cr + \frac{1}{Pe}} = 1 \end{aligned}$$

which proves the proposition.

A quite similar analysis shows that also in the second special case, the conditions (13) and (27) assure the strong stability.

Finally, let us investigate whether or not the stability condition (27) can be satisfied when the condition of the third-order approximation (Equation (16)) should also be fulfilled. Let us fix a Péclet number, and define the Courant number according to (16). Since the Courant number is now a less than 1, the second inequality in (27) is obviously fulfilled. It is thus sufficient to solve the inequality

$$1 - \sqrt{1 - \frac{3}{Pe^2}} \leq \frac{1}{Pe}$$

for the Péclet number. The two sides are equal for the value $Pe = 2$. Elementary calculations show that the above inequality is valid if and only if

$$Pe \geq 2 \quad (29)$$

In this case, from (29) and (16), we obtain an estimation for the Courant number:

$$\frac{1}{2} \leq Cr < 1.$$

Remark: The conditions (27), (29) are *sufficient* conditions for the strong stability. This means that (27) (respectively (29)) implies the inequality (23). However, the above analysis states nothing about the *necessity* of these conditions. Practical experiences show that these conditions can be often violated without arising instability.

Numerical examples

We illustrate the above results via three examples. In all examples, Equation (1) has analytical solutions. Consider (1) supplied with zero initial condition and assume that the (upper) boundary condition is a Dirac distribution concentrated to the time $t = 0$. It is well known that the corresponding solution of (1) has the following form:

$$v_0(t, x) = \frac{x}{\sqrt{4\pi Dt^3}} \exp\left(-\frac{(x - Ct)^2}{4Dt}\right) \quad (30)$$

Applying Equation (30), it can be shown (see [4]) that, if the boundary condition is the unit step function at $t = 0$, and the initial condition is again identically zero, then the corresponding solution is as follows:

$$v_0(t, x) = \Phi\left(-\frac{x - Ct}{\sqrt{2Dt}}\right) + \exp\left(\frac{Cx}{D}\right) \Phi\left(-\frac{x + Ct}{\sqrt{2Dt}}\right) \quad (31)$$

where Φ denotes the usual error function:

$$\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^x \exp\left(-\frac{\xi^2}{2}\right) d\xi$$

The examples are based on the functions (30) and (31):

Example 1: Consider Equation (1) with the following parameters: $C = 2$ m/sec; $D = 10000$ m²/sec; $\Delta x = 1000$ m; $\Delta t = 500$ sec, and the total length $a = 200$ km. Then Equation (1) can be regarded as a diffusion wave model for a river reach and then D stands for the turbulent diffusion coefficient. Assume that the initial condition is zero and the (upper) boundary condition is the unit step starting at $t = 0$. In this example the Courant and Péclet number are equal to 1 and 0.1, respectively. To eliminate the numerical diffusion, the coefficients of the scheme (5) are defined by the formulas (14). It can be easily seen that the condition (28) of strong stability is satisfied. Figure 1. shows the analytical and numerical results at $t = 40000$ sec: the computational solution fits the analytical solution fairly well. We note that, using higher Courant numbers e.g. $Cr = 2$, no instability occurred, despite the condition (28) is not satisfied in this case. If, as usual, the weighting parameter ε is defined to be 0.5, then Equation (13) implies that the other weighting parameter $\theta = -4.5$. (Moreover, θ is always negative, whenever ε is between 0 and 1.) If θ is set to 0, in order that the values of parameters are "correct" (i.e. more traditional), this significantly undermines the exactness of the approximation, as it can

be seen in Figure 1. The picture becomes even worse if θ is set to be positive.

Since the Péclet number is small, the condition (16) cannot be satisfied. However, the solution is smooth enough, consequently, the second term in the right-hand side of (11) does not cause serious numerical problems.

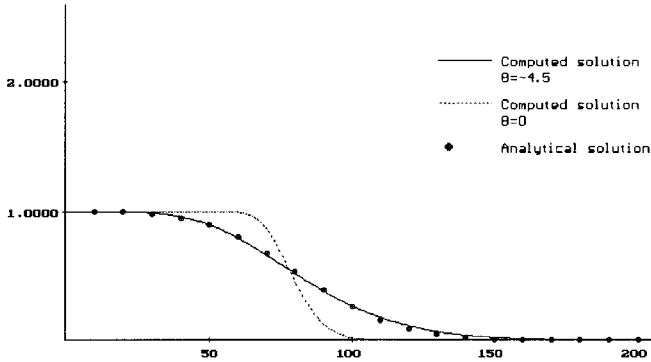


Fig. 1. Analytical and numerical solutions of Example 1, using different weighting parameters θ .

Example 2: Let the initial and the boundary conditions be the same as in Example 1, and let the parameters be as follows: $C := 2$ m/sec; $D := 250$ m²/sec; $\Delta x := 1000$ m; $a := 200$ km. The only difference is that the diffusion coefficient is much smaller, which makes the transport is more advection dominated. Now $Pe = 4$, therefore the condition (16) can be satisfied, which results in the optimal Courant number $Cr = 0.9013$. It should be pointed out that θ must be negative again, if $\epsilon = 1$. Figure 2. shows the computed and analytical solutions at $t = 40000$ sec: the computed solution is again exact enough. In this example, the effect of the numerical oscillation can be clearly seen, if the Courant number differs from the above optimal value. If, for instance, $Cr = 5$, a significant numerical oscillation appears behind the front (see Figure 2). On the other hand, if the actual Courant number is less then the optimal, e.g. $Cr = 0.5$, this causes a (much smaller) numerical oscillation before the front (not displayed in Figure 2).

The numerical oscillation can be observed more clearly, if the boundary condition is a Dirac distribution at $t = 0$. This is illustrated by the last example:

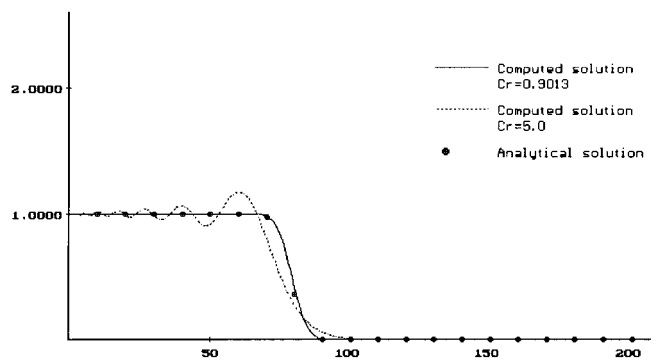


Fig. 2. Analytical and numerical solutions of Example 2, using different Courant numbers.

Example 3. Let the initial condition be zero again, but assume the boundary condition to be a Dirac distribution concentrated to the time $t = 0$. The parameters are as follows: $C := 2$ m/sec; $D := 150$ m²/sec; $\Delta x := 1000$ m; $a := 100$ km. The Péclet number is $Pe = 6.67$, and the optimal Courant number (satisfying the condition (16)) is $Cr = 0.9656$. Using this optimal Courant number, the scheme produces again good approximation, as it can be seen in Figure 3: however, other values of the Courant number cause numerical oscillations. Namely, if the Courant number is greater ($Cr := 2$ in the figure), then the oscillation appears behind the front, while it is smaller ($Cr := 0.5$), the oscillation appears before the front. Note that, in Examples 2 and 3, the presence of the numerical oscillations also shows, that the condition of the strong stability is violated: the approximate solution, however, remains bounded.

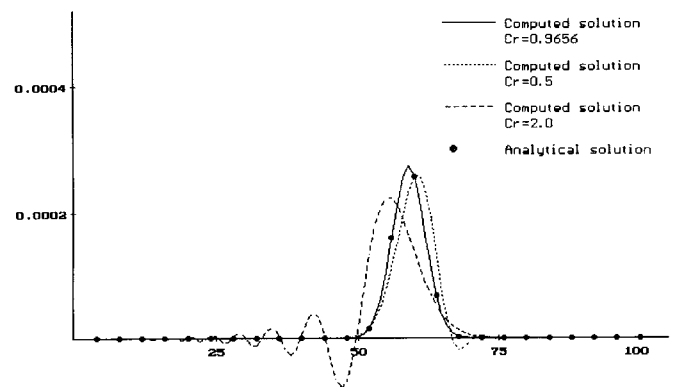


Fig. 3. Analytical and numerical solutions of Example 3, using different Courant numbers.

Conclusions

Though the scheme (5) contains neither the diffusion coefficient D nor expressions for second-order derivatives, it is possible to define the parameters of the scheme in such a way that it approximates the advection-diffusion equation (1). Instead of defining the weighting parameters ϵ and θ separately and in such a way that they are between 0 and 1, it is sufficient to require the condition (13). This makes the computed solution free from false diffusion. In this case, the coefficients of the scheme (5) can be determined by the expressions (14), which do not contain the weighting parameters any more, only the Courant and Péclet numbers. If one insists on computing the corresponding weighting parameters, their values may be even negative: however, the approximation properties of the scheme remain good in this apparently pathological situation as well. The accuracy can be increased further, if the numerical oscillation is also eliminated: this can be done by describing the condition (16). This condition can be satisfied only in advection dominated cases, more precisely, if $Pe^2 > 3$. Violating the condition (16) causes numerical oscillations. If the advective velocity is positive, the scheme is always stable in a weak sense. Sufficient conditions for a much stronger stability are given by (27) and (29).

Notations

C	advective velocity
D	diffusion coefficient
L	advection-diffusion operator
u	unknown function of the advection-diffusion equation
t, x	time and space variables
$\Delta t, \Delta x$	time step and grid size
ε, θ	weighting factors in the Muskingum-Cunge scheme
C_1, C_2, C_3	coefficients of the Muskingum-Cunge scheme
Cr, Pe	Courant and Péclet numbers
$\ \cdot\ $	maximum norm of vectors and matrices

References

- 1 CUNGE, J.A. (1969): On the subject of a flood propagation computation method (Muskingum Method). *Journal of Hydraulic Research*, Vol.7, No.2, pp. 205–230.
- 2 MILLER, W.A., CUNGE, J.A. (1975): Simplified equations of unsteady flow. In: *Unsteady Flow in Open Channels, Volume I*, Chapter 5, pp. 183–255. (Edited by K. MAHMOUD and V. YEVJEVICH) Water Resources Publ., Colorado, USA.
- 3 SZILÁGYI, J. (1992): Why can the weighting parameter of the Muskingum channel routing method be negative? *Journal of Hydrology*, Vol.138, pp. 145–151.
- 4 TODINI, E., BOSSI, A. (1986): PAB (Parabolic and Backwater) an unconditionally stable flood routing scheme particularly suited for real time forecasting and control. *Journal of Hydraulic Research*, Vol.24, No.5, pp. 405–424.