DELFT UNIVERSITY OF TECHNOLOGY

REPORT 08-17

A local theta scheme for advection problems with strongly varying meshes and velocity profiles

P. van Slingerland M. Borsboom C. Vuik

ISSN 1389-6520

Reports of the Department of Applied Mathematical Analysis

Delft 2008

A local theta scheme for advection problems with strongly varying meshes and velocity profiles

P. van Slingerland[a], M. Borsboom[a], C. Vuik[b]

[a] *Deltares, Rotterdamseweg 185 2629 HD Delft, The Netherlands*

[b] *Delft University of Technology, J.M. Burgerscentrum, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft Institute of Applied Mathematics, Mekelweg 4, 2628 CD Delft, The Netherlands*

Abstract

A generalisation of the theta time discretisation scheme is proposed for advection problems with strongly varying meshes and velocity profiles.

Starting point is the theta upwind scheme, which is conservative, stable, positivity-preserving, and free of spurious oscillations for any given time step, provided that theta is sufficiently large. The main disadvantage of this scheme is the accompanying numerical diffusion, which increases with theta. Because theta is a constant, the 'worst' *local* properties bring down the effectiveness of the scheme in the *entire* computational domain. The presented generalisation of the theta scheme to the local theta scheme uses a space- and time-dependent theta coefficient, instead of a constant value. Based on local mesh and velocity properties, the local theta coefficients are chosen as small as possible, to minimize the numerical diffusion, but sufficiently large, to ensure that the scheme is stable, positivity-preserving, and non-oscillatory without restricting the time step. To improve the accuracy even more, the flux corrected transport algorithm is incorporated into the scheme. For a sufficiently small time step, the resulting local theta upwind FCT scheme coincides with an original Euler forward upwind FCT scheme. For a larger time step, for which the latter is no longer applicable, the local theta upwind FCT scheme achieves a remarkably high accuracy, as is illustrated by numerical examples.

Altogether, the generalisation of the theta scheme to the local theta scheme provides new flexibility in the construction of numerical schemes for advection problems.

Key words: advection equation, theta scheme, upwind discretisation, flux corrected transport, numerical diffusion, stability, positivity, wiggles, conservation, discrete maximum principle

1. Introduction

Unphysical numerical diffusion errors seem inescapable in the design of a numerical discretisation of advection terms, if the scheme needs to be Conservative, Stable, Positivity-preserving, and free of spurious Wiggles (CSPW) for any time step. Reversely, without restricting the time step, it seems impossible to obtain a method that suffers hardly from numerical diffusion and that is CSPW at the same time.

This is problematic for many real-life advection dominated problems, for which the velocity profile is usually strongly varying, and for which the complex geometry of the spatial domain often requires a variety of sizes and shapes in the numerical mesh. Unfavourable *local* properties of the mesh or the velocity profile can lead to to an unacceptable inaccuracy in the *entire* spatial domain, or to an infeasible computational time.

Before moving on to the suggested solution, the problem is substantiated by discussing certain existing finite volume schemes for the advection equation. For an extensive overview, see e.g. [6,7,12,17].

The Euler forward upwind scheme (see e.g. [12, p. 73]) can be considered to be the most basic scheme that approximates the solution to the advection equation. An important disadvantage of this scheme is that the time step needs to be restricted in order for the scheme to be CSPW (see also Section 2 later on). Roughly speaking, the upperbound for the time step is determined by the single cell for which the velocity is largest in proportion to the size of the cell volume. This condition is also known as the CFL condition, which is named after Courant, Friedrichs, and Lewy [3], and which "simply states that the method must be used in such a way that information has a chance to propagate at the correct physical speeds" [12, p. 68]. A second drawback of the Euler forward upwind scheme is the accompanying numerical diffusion, which lowers the accuracy (for an illustration, see e.g. [11, p. 26] or Section 6).

To improve the accuracy of the first-order upwind scheme above, it seems straightforward to use a higher-order spatial discretisation. However, higher order schemes have a tendency to create spurious wiggles and unphysical negative results. Godunov's order barrier theorem (see e.g. [17, p. 341]) states that linear (explicit or implicit) one-step second-order accurate schemes for the advection equation cannot be monotonicity-preserving (unless the so-called CFL number is a natural number). Indeed, numerical examples [11, p. 20-32] show that higher-order methods such as Lax-Wendroff, Fromm, higher-order upwind(-biased), and higher-order central schemes show spurious oscillations.

To tackle this problem, Boris and Book [2] have devoped the Flux Corrected Transport (FCT) algorithm, which has been generalised to multi-dimensional problems by Zalesak [18]. The main idea is to update the Euler forward upwind solution with a non-linear limited flux correction that is based on a higher order scheme. The pupose of the limiter is to prevent the introduction of new local extrema. Ever since, many different limiters have been developed. A large class of limiters is based on Harten's [5] concept of Total Variation Non-Increasing (TVNI) schemes, which later on became known as Total Variation Diminishing (TVD) schemes. The idea is that a TVD scheme is also monotonicity-preserving. An overview of a large number of TVD limiters has been given by Sweby [14]. Nonetheless, although the FCT algorithm improves the accuracy of the Euler forward upwind scheme, the restriction on the time step remains unchanged.

A scheme that does not have this drawback is the theta upwind scheme (see e.g. [6,

p. 35, 215]), which reduces to the Euler forward upwind scheme if the parameter theta is equal to zero. This scheme is CSPW for any given time step, provided that theta is sufficiently large (see also Section 2 later on). The single cell for which the velocity is largest in proportion to the size of the cell volume now determines a lowerbound for theta, instead of an upperbound for the time step. The main disadvantage of this scheme is, again, the accompanying numerical diffusion, which grows with theta.

Similar to the Euler forward upwind scheme, the numerical diffusion of the theta upwind scheme can be reduced with the help of an FCT approach. Kuzmin et al. [9,8,10] have suggested such a strategy, which includes an iterative method to deal with the implicit flux correction. However, they use a fixed value of theta in combination with adaptive time step control, which means that the time step is restricted again.

This article proposes a generalisation of the theta scheme to the local theta scheme, which, in combination with an FCT approach, for advection problems with strong variations in the mesh or the velocity profile, is efficient in the sense that it is Conservative, Stable, Positivity-preserving, and free of spurious Wiggles (CSPW) without a restriction on the time step, and that is also accurate in the sense that it introduces a locally minimised amount of numerical diffusion.

The outline of this paper is as follows. First of all, the cause of the problem mentioned at the beginnning of this section is illustrated by the relation between numerical diffusion and the requirements to be CSPW for the one-dimensional theta upwind scheme (Section 2). This relation reveals why a constant value of theta is unsuitable for problems with strong variations in the mesh or the velocity. The proposed solution is to make theta space- and time-dependent, which leads to the local theta scheme (Section 3). Moreover, theoretical conditions for the local theta coefficients are derived for which the scheme is CSPW. After that, a practical strategy to choose the coefficients is provided in terms of explicit expressions, which leads to the local theta upwind scheme (Section 4). To increase the accuracy even more, the FCT algorithm is incorporated into the scheme (Section 5). The performance of the resulting local theta upwind FCT scheme is illustrated by means of two test cases (Section 6). Finally, a number of conclusions is given (Section 7).

2. Theta upwind scheme: revealing the disadvantages of a constant theta

The previous section described the problem under consideration as a trade-off between numerical diffusion and a restriction on the time step.

This section illustrates the cause of this problem by considering the relation between numerical diffusion and the requirements to be CSPW for the one-dimensional theta upwind scheme.

To this end, consider the following one-dimensional advection problem with a constant velocity $u > 0$:

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} = 0.$$

Application of the theta upwind scheme (for more details, see Definition 4.1 later on for the one-dimensional case, a constant theta and $x_1 < x_2 < ... < x_I$) yields:

$$\frac{c_i^{n+1} - c_i^n}{\Delta t} + \theta u \frac{c_i^{n+1} - c_{i-1}^{n+1}}{\Delta x} + (1-\theta) u \frac{c_i^n - c_{i-1}^n}{\Delta x} = 0, \tag{1}$$

where Δt denotes the constant time step, Δx indicates the constant cell width, and $\theta \in [0, 1]$ is a constant parameter.

The theta upwind scheme is CSPW, provided that the following relation is satisfied (for the proof, see Theorem 3.14 later on):

$$\theta \geq 1 - \frac{\Delta x}{u \Delta t}. \tag{2}$$

In other words, either θ should be sufficiently large, or the time step needs to be sufficiently small.

The lowest order error of the theta upwind scheme has the character of diffusion. The corresponding numerical diffusion coefficient can be computed by means of the modified equation (for the proof, see Proposition A.1 later on):

$$\text{‘numerical diffusion coefficient’} = \frac{u \Delta x}{2} \left(1 - (1 - 2\theta) \frac{u \Delta t}{\Delta x} \right). \tag{3}$$

Note that the numerical diffusion grows with θ.

Figure 1 shows a contour plot of the scaled numerical diffusion coefficient $1-(1-2\theta)\frac{u\Delta t}{\Delta x}$ as a function of the parameter θ and the so-called CFL-number $\frac{u \Delta t}{\Delta x}$. The white line indicates the circumstances for which the numerical diffusion is zero. Underneath (and on the right of) this line, the numerical diffusion is positive. Above the line, the numerical diffusion is negative, which can usually be associated with instability. Below (and on the right of) the black line, the scheme is CSPW according to (2). Apparently, this latter condition forces the numerical diffusion to be large.

Altogether, a certain amount of numerical diffusion is inescapable in order for the theta upwind scheme to be CSPW. By satisfying (2) for a minimal value of θ, the scheme is unconditionally CSPW at the cost of the least possible amount of numerical diffusion. However, because θ is a constant, in case of a non-equidistant mesh, the smallest cell determines the value of θ in the entire computational domain. For the larger cells, this value is unnecessarily large. As a consequence, the numerical diffusion is unnecessarily large for these cells. This effect becomes stronger as the variations in the mesh are larger. Similarly, for multi-dimensional problems, high variations in the velocity profile can lead to needless inaccuracy.

A remedy is to make theta space and time dependent, which is dicussed in the next section.

3. Local theta scheme: benefitting from a space- and time-dependent theta

The previous section discussed how strong variations in the mesh or the velocity profile can lead to unnecessary diffusion errors in the theta upwind approximation, if the scheme is required to be CSPW.

This section proposes a generalisation of the theta scheme to the local theta scheme (Definition 3.3), which is designed to prevent the introduction of unnecesary numerical diffusion. Moreover, theoritical conditions for the local theta coeffients are derived for which the scheme is CSPW (Theorem 3.14).

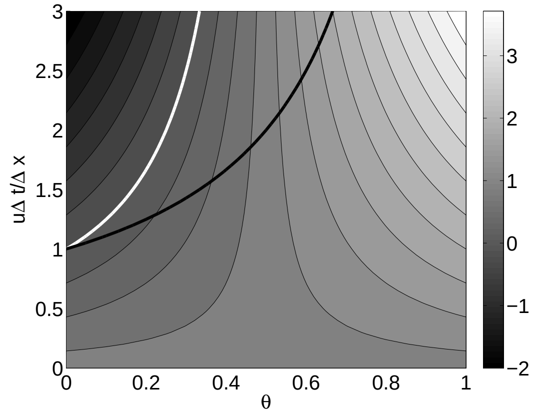


Fig. 1. Contour plot of the scaled numerical diffusion coefficient (3). The white line corresponds to zero numerical diffusion. The black line indicates the border of the region where the scheme is CSPW (2).

First of all, a mathematical discription of the advection problem under consideration is formulated (Definition 3.1). Next, notational aspects of a general one-step numerical scheme for an unstructured mesh are introduced (Case 3.2). After that, this framework is used to formulate the local theta scheme (Definition 3.3), which simply follows from the theta scheme by replacing theta by a space- and time-dependent coefficient.

Subsequently, theoritical conditions for the local theta coeffients are derived for which the scheme is CSPW. To be precise, a scheme is called CSPW if it is conservative (Definition 3.4), absolutely stable (Definition 3.5), positivity-preserving (Definition 3.6), and if it satisfies the local discrete maximum principle (Definition 3.7). The local discrete maximum principle states that each concentration c_i^n lies between the minimum and the maximum of the concentrations that it depends on. As a consequence, it excludes local extrema in the interior of the discrete space-time domain.

To show that a scheme is CSPW, the concept of the global discrete maximum principle (Definition 3.8) is usefull. The global discrete maximum principle states that, for each concentration c_i^n, there exists a path in the space-time stencil to either an initial or a boundary value such that the concentrations along the path are non-decreasing. A similar principle applies for non-increasing values. Hence, each concentration c_i^n lies between the minimum and the maximum of the discrete initial and boundary conditions. As a consequence, a scheme that satisfies the global discrete maximum principle is automatically absolutely stable and positivity-preserving (Theorem 3.9). Conveniently, for a scheme of positive type (Definition 3.11) that has a unique solution, the local discrete maximum principle implies the global discrete maximum principle (Lemma 3.10, Theorem 3.12), in which case the scheme is also absolutely stable and positivity-preserving by virtue of Theorem 3.9.

With the help of this conclusion, the local theta scheme is CSPW if it is conservative and of positive type, and if the scheme has a unique solition. The corresponding conditions for the local theta coefficients (Lemma 3.10 and Theorem 3.14) include a CFL-like condition.

Altogether, unlike the traditional theta scheme, the local theta scheme uses a different value of theta for each time step and for each face in the mesh. This provides more flexibility in the conditions under which the scheme is CSPW, which can lead to a more efficient scheme.

The next section provides a practical strategy to choose optimal coefficients in terms of explicit expressions.

Definition 3.1 (Advection model)
Consider a substance with a concentration $c(\underline{x}, t)$ that is dissolved in a fluid with a divergence-free velocity profile $\underline{u}(\underline{x}, t)$. Suppose that the transport of the substance is governed by the advection equation:

$$\frac{\partial c(\underline{x}, t)}{\partial t} + \nabla \cdot \big(\underline{u}(\underline{x}, t) c(\underline{x}, t)\big) = 0, \qquad t \in [0, T],\ \underline{x} \in \mathcal{D}.$$

The spactial domain $\mathcal{D} \subset \mathbb{R}^m$ $(m = 1, 2, 3)$ is assumed to be compact and connected. Additionally, an initial condition at $t = 0$, a Dirichlet boundary condition at the inflow boundary, and a Neumann boundary condition at the closed part of the boundary are assumed to be specified. ⌟

Case 3.2 (Linear one-step scheme)
Consider the advection equation (Definition 3.1). Subdivide the spatial domain $\mathcal{D}$ into m-dimensional simple polygons $\mathcal{V}_1, ..., \mathcal{V}_I \subset \mathcal{D}$. To be precise, these volumes are chosen such that the union of all volumes is equal to $\mathcal{D}$, and that the intersection of two different volumes is equal to the intersection of their boundaries. Next, suppose that the intersection of precisely K volumes $\mathcal{V}_{i_1}$, ..., $\mathcal{V}_{i_K}$ with the inflow boundary is non-empty and connected, and introduce the corresponding adjacent boundary elements:

$$\mathcal{V}_{I+k} = \mathcal{V}_{i_k} \cap \partial\mathcal{D} \neq \varnothing, \qquad k = 1, ..., K.$$

For a volume $\mathcal{V}$, the notation $|\mathcal{V}|$ will be used to indicate the volume of $\mathcal{V}$, the surface area of $\mathcal{V}$, the length of $\mathcal{V}$, or 1, depending on the nature of $\mathcal{V}$. Moreover, in favour of notational brevity, if $\mathcal{V}$ contains only a single element of $\mathcal{D}$, the integral of a function over $\mathcal{V}$ is defined as the evaluation of that function in that element of $\mathcal{D}$.

Moreover, let $x_i \in \mathcal{D}$ be the center of mass of $\mathcal{V}_i$ (for all $i = 1, ..., I + K$) and let $\mathcal{J}_i$ contain the indices of the neighbors of $\mathcal{V}_i$:

$$\mathcal{J}_i = \{j \in \{1, ..., I + K\} \setminus \{i\} | \mathcal{S}_{ij} := \partial\mathcal{V}_i \cap \partial\mathcal{V}_j \neq \varnothing\}, \qquad i = 1, ..., I.$$

Next, discretise the time domain according to

$$0 < t^0 < t^1 < ... < t^N < T,$$

and define the following averages of the concentration and the velocity (for all $i = 1, ..., I+K$ and $n = 0, 1, ..., N$):

$$\begin{aligned} c_i^n &= \frac{1}{|\mathcal{V}_i|} \int_{\mathcal{V}_i} c(\underline{x}, t^n)\, d\underline{x}, \\ u_{ij}^n &= \frac{1}{t^n - t^{n-1}} \int_{t^{n-1}}^{t^n} \left(\frac{1}{|\mathcal{S}_{ij}|} \int_{\mathcal{S}_{ij}} \underline{u}(\underline{x}, t) \cdot \underline{n}_{ij}\, d\underline{x} \right) dt, \qquad n \neq 0,\ j \in \mathcal{J}_i, \end{aligned}$$

where $\underline{n}_{ij}$ is the unit normal vector on $\mathcal{S}_{ij}$ that points in the direction of $\mathcal{V}_j$. Note that u_{ij}^n is averaged over the time imterval $[t^{n-1}, t^n]$. Next, consider a linear one-step scheme of the form (for $i = 1, .., I$):

$$a_{ii}^n c_i^n + \sum_{j\in\mathcal{A}_i^n} a_{ij}^n c_j^n = \sum_{j\in\mathcal{B}_i^n} b_{ij}^n c_j^{n-1}, \qquad a_{ij}^n \neq 0\,(j \in \mathcal{A}_i^n \cup \{i\}),\, b_{ij}^n \neq 0\,(j \in \mathcal{B}_i^n), \tag{4}$$

where $\mathcal{A}_i \subset \mathcal{J}_i$ and $\mathcal{B}_i \subset \mathcal{J}_i \cup \{i\}$. The scheme can also be written as a linear system of the following form:

$$A \underbrace{\begin{bmatrix} c_1^n \\ \vdots \\ c_I^n \end{bmatrix}}_{=:\underline{c}^n} = B \begin{bmatrix} c_1^{n-1} \\ \vdots \\ c_{I+K}^{n-1} \\ c_{I+1}^n \\ \vdots \\ c_{I+K}^n \end{bmatrix}, \tag{5}$$

where $A \in \mathbb{R}^{I\times I}$ and $B \in \mathbb{R}^{I\times(I+2K)}$ are matrices. ⌟

Definition 3.3 (Local theta scheme)
Consider Case 3.2. The *local theta scheme* reads (for $n = 1, ..., N$):

$$\frac{|\mathcal{V}_i|c_i^n - |\mathcal{V}_i|c_i^{n-1}}{t^n - t^{n-1}} = \sum_{j\in\mathcal{J}_i} -|\mathcal{S}_{ij}| \underbrace{\left(\left(1-\theta_{ij}^n\right)\phi_{ij}^{n,n-1} + \theta_{ij}^n \phi_{ij}^{n,n}\right)}_{=:\Phi_{ij}^n}. \tag{6}$$

Here, $\phi_{ij}^{n,m}$ is a numerical flux function that is assumed to be of the form:

$$-|\mathcal{S}_{ij}|\phi_{ij}^{n,m} = -\,\gamma_{ji}^n c_i^m + \gamma_{ij}^n c_j^m. \tag{7}$$

The coefficients $\theta_{ij}^n \in [0,1]$ are free to be chosen such that the scheme has nice properties.

Definition 3.4 (Conservative)
Consider Case 3.2. The scheme is called *conservative* if the scheme can be written in the form

$$\frac{|\mathcal{V}_i|c_i^n - |\mathcal{V}_i|c_i^{n-1}}{t^n - t^{n-1}} = \sum_{j\in\mathcal{J}_i} |\mathcal{S}_{ij}|\Psi_{ij}^n, \qquad i = 1, ..., I,$$

where Ψ_{ij}^n is anti-symmetric:

$$\Psi_{ij}^n = -\Psi_{ji}^n, \qquad i = 1, ..., I;\ j \in \mathcal{J}_i.$$

As a consequence, in the absence of an open boundary, the total change in mass is equal to zero. If a certain part of the boundary is open, the total change in mass is equal to the transport over the open boundary. A similar definition for one-dimensional schemes can be found in [4, p. 250] or [15, p. 171]. ⌟

Definition 3.5 (Absolutely stable)
Consider Case 3.2. The scheme is called *absolutely stable* with respect to a norm $\|.\|$, if there exist constants $k, \tau > 0$ (τ may depend on the spatial mesh size) such that, if

$$t^n - t^{n-1} \leq \tau, \qquad \forall n = 1, ..., N,$$

then,

$$\|\underline{w}^n\| \leq k\|\underline{w}^0\|, \qquad \forall n = 1, ..., N,$$

for any perturbation $\underline{w}^0$ of the initial condition $\underline{c}^0$ that yields a perturbation $\underline{w}^n$ of $\underline{c}^n$. See also [17, p. 168]. ⌟

Definition 3.6 (Positivity-preserving)
Consider Case 3.2. The scheme is called *positivity-preserving*, if the solution is nonnegative

$$c_i^n \geq 0, \qquad i = 1, .., I;\ n = 1, ..., N,$$

whenever the initial and boundary condtions are nonnegative

$$\begin{aligned} c_i^0 &\geq 0, && i = 1, ..., I, \\ c_i^n &\geq 0, && i = I+1, ..., I+K;\ n = 0, ..., N. \end{aligned}$$

See also [6, p. 118-122] or [4, p. 292]. ⌟

Definition 3.7 (Local discrete maximum principle)
Consider Case 3.2. The scheme satisfies the *local discrete maximum principle* if, for any solution to the scheme, either

$$c_i^n = c_j^n = c_k^{n-1}, \qquad j \in \mathcal{A}_i^n,\ k \in \mathcal{B}_i^n, \tag{8}$$

or

$$\min\left\{\min_{j\in\mathcal{A}_i^n} c_j^n, \min_{j\in\mathcal{B}_i^n} c_j^{n-1}\right\} < c_i^n < \max\left\{\max_{j\in\mathcal{A}_i^n} c_j^n, \max_{j\in\mathcal{B}_i^n} c_j^{n-1}\right\}, \tag{9}$$

for all $i = 1, ..., I$ and for all $n = 1, .., N$. A similar concept for explicit Euler forward schemes can be found in [1, p. 12-13]. ⌟

Definition 3.8 (Global discrete maximum principle)
Consider Case 3.2. The scheme satisfies the *global discrete maximum principle* if, for any solution to the scheme, for all $i = 1, ..., I$ and for all $n = 0, ..., N$, a non-decreasing path to an initial or a boundary value exists in the sense that there are $j_1, j_2, ..., j_E \in \{1, ..., I+K\}$ and $m_1, m_2, ..., m_E \in \{0, 1, ..., N\}$ such that

- $j_{e+1} \in \mathcal{A}_{j_e}^{m_e} \cup \mathcal{B}_{j_e}^{m_e}$,
- $m_{e+1} \leq m_e$,
- either $m_E = 0$ or $j_E \in \{I+1, ..., I+K\}$,
- $c_i^n \leq c_{j_1}^{m_1} \leq ... \leq c_{j_E}^{m_E}$,

and if, similarly, a non-increasing path exists. As a consequence, the scheme satisfies:

$$c_{\min} \leq c_i^n \leq c_{\max}, \qquad i = 1, ..., I;\ n = 0, ..., N, \tag{10}$$

where

$$c_{\min} := \min \left\{ \min_{j=1,\dots,I} \{c_j^0\}, \min_{\substack{j=I+1,\dots,I+K \\ n=0,\dots,N}} \{c_j^n\}, \right\},$$

$$c_{\max} := \max \left\{ \max_{j=1,\dots,I} \{c_j^0\}, \max_{\substack{j=I+1,\dots,I+K \\ n=0,\dots,N}} \{c_i^n\} \right\},$$

A similar concept for explicit Euler forward schemes can be found in [1, p. 12-13]. $\lrcorner$

Theorem 3.9
Consider Case 3.2. If the global discrete maximum principle (Definition 3.8) is satisfied, then, the scheme is positivity-preserving (Definition 3.6) and absolutely stable (Definition 3.5) with respect to $\|.\|_\infty$.

Proof:
A similar proof can be found in [17, p. 170-171]. That the scheme is positivity-preserving is an immediate consequence of Definition 3.8. That the scheme is absolutely stable with respect to $\|.\|_\infty$ can be shown as follows. Let $\underline{w}^0$ be a perturbation of the initial condition $\underline{c}^0$ that yields a perturbation $\underline{w}^n$ of $\underline{c}^n$. As both $\underline{c}^n$ and $\underline{c}^n + \underline{w}^n$ satisfy the (linear) scheme, subtraction yields that $\underline{w}^n$ also satisfies the scheme. Apply the global discrete maximum principle (10) to obtain:

$$\min_j \{w_j^0\} \leq w_i^n \leq \max_j \{w_j^0\}.$$

Hence,

$$\|\underline{w}^n\|_\infty \leq \|\underline{w}^0\|_\infty.$$

As a result, the scheme is absolutely stable with respect to $\|.\|_\infty$ (use $k = 1$ in 3.5). $\blacksquare$

Lemma 3.10
Let γ_j and c_j ($j = 1, ..., J$; $J \geq 2$) be reals satisfying

$$\sum_{j=1}^{J} \gamma_j c_j = 0, \tag{11}$$

$$\sum_{j=1}^{J} \gamma_j = 0, \tag{12}$$

$$\gamma_2, ... \gamma_J < 0.$$

Then, either

$$c_1 = c_2 = ... = c_J$$

or

$$\min_{j=2,\dots,J} c_j < c_1 < \max_{j=2,\dots,J} c_j.$$

Proof:
See also [17, Lemma 4.4.1]. First of all, note that

$$\gamma_1 = -\sum_{j=2}^{J} \gamma_j > 0.$$

Hence, (11) can be rewritten to obtain:

$$c_1 = \sum_{j=2}^{J} \underbrace{\frac{-\gamma_j}{\gamma_1}}_{>0} c_j.$$

Using (12), it follows that

$$\sum_{j=2}^{J} \frac{-\gamma_j}{\gamma_1} = \frac{\gamma_1}{\gamma_1} = 1.$$

Apparently, c_1 is a convex combination of $c_2, .., c_J$. As a result,

$$c_{\min} := \min_{j=2,...,J} c_j \leq c_1 \leq \max_{j=2,...,J} c_j =: c_{\max}.$$

Now, there are two options.

(i) If $c_{\min} = c_{\max}$, then, $c_1 = c_2 = ... = c_J$.

(ii) If $c_{\min} < c_{\max}$, then there exist $j_{\min}, j_{\max} \geq 2$ such that

$$c_{j_{\min}} = c_{\min} < c_{\max} = c_{j_{\max}}.$$

As a result,

$$\begin{aligned} c_1 &= \sum_{j=2}^{J} \frac{-\gamma_j}{\gamma_1} c_j \\ &\leq \sum_{j=2, j \neq j_{min}}^{J} \frac{-\gamma_j}{\gamma_1} c_{\max} + \frac{-\gamma_{j_{\min}}}{\gamma_1} c_{\min} \\ &< \sum_{j=2}^{J} \frac{-\gamma_j}{\gamma_1} c_{\max} = c_{\max} \end{aligned}$$

Similarly, it can be shown that $c_1 > c_{\min}$. ■

Definition 3.11 (Positive type)
Consider Case 3.2. The scheme is said to be of *positive type*, if

$$\begin{aligned} a_{ii}^n &> 0 && \\ a_{ij}^n &< 0, && j \in \mathcal{A}_i^n, \\ b_{ij}^n &> 0, && j \in \mathcal{B}_i^n, \\ a_{ii}^n + \sum_{j \in \mathcal{A}_i^n} a_{ij}^n + \sum_{j \in \mathcal{B}_i^n} -b_{ij}^n &= 0, && \end{aligned} \tag{13}$$

for all $i = 1, ..., I$. See also [17, p. 170]. ⌟

Theorem 3.12
Consider Case 3.2. If the scheme is of positive type (Definition 3.11), then it satisfies the local discrete maximum principle (Definition 3.7). If, furthermore, the matrix A in (5) is invertible, then the scheme also satisfies the global maximum principle (Definition 3.8), and, by virtue of Theorem 3.9, the scheme is also positivity-preserving and absolutely stable with respect to $\|.\|_\infty$.

<u>PROOF</u>:
The local discrete maximum principle follows with the help of Lemma 3.10. For a given concentration c_i^n, choose $\gamma_1 = a_{ii} > 0$ and $c_1 = c_i^n$. Furthermore, for the coefficients $c_2, ..., c_J$, choose the concentrations c_j^n with $j \in \mathcal{A}_i^n$ and c_j^{n-1} with $j \in \mathcal{B}_i^n$. The values of $\gamma_2, ..., \gamma_J$ are the corresponding coefficients a_{ij}^n and b_{ij}^n. By applying Lemma 3.10 in this manner for each concentration c_i^n, the local discrete maximum principle follows.

The global discrete maximum principle is shown with the help of induction. To this end, note that, by definition, the non-decreasing paths exist for $n = 0$. Next, suppose that the claim is true to up to time t^{n-1} and consider c_i^n for some $i \in \{1, ..., I\}$. Now, the following iterative argument applies. First of all, note that the local discrete maximum principle (Definition 3.7) leads to one of the following two options:

1. (9) holds. There are two options:
   i. $c_i^n < c_j^{n-1}$ for some $j \in \mathcal{B}_i^n$. As a path with non-decreasing values exists for c_j^{n-1}, the same path can be used for c_i^n and the claim is true.
   ii. $c_i^n < c_j^n$ for some $j \in \mathcal{A}_i^n$. If j indicates a boundary element, a non-decreasing path to the boundary has been found, in only one step. If j indicates an interior element, this iterative argument can be repeated without visiting this index again, because $c_i^n < c_j^n$ is a strict inequality.
2. (8) holds. Again, there are two options:
   (a) $\mathcal{B}_i^n$ is not empty. Hence, $c_i^n = c_j^{n-1}$ for some $j \in \mathcal{B}_i^n$. As a path with non-decreasing values exists for c_j^{n-1}, the same path can be used for c_i^n and the claim is true.
   (b) $\mathcal{B}_i^n$ is empty. Then, there must be an index $j \in \mathcal{A}_i^n$ for which a non-decreasing path exists. To show this claim, in search of a contradiction, assume that there is no $j \in \mathcal{A}_i^n$ for which a non-decreasing path exists. Then, there must be an isolated set of cells in the sense that (5) can be reordered as

$$\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} \underline{c}_1^n \\ \underline{c}_2^n \end{bmatrix} = \begin{bmatrix} 0 \\ B_1 \end{bmatrix} \begin{bmatrix} c_1^{n-1} \\ \vdots \\ c_{I+K}^{n-1} \\ c_{I+1}^n \\ \vdots \\ c_{I+K}^n \end{bmatrix},$$

where the vector $\underline{c}_1^n$ contains the concentrations of the isolated cells. Because A is invertible, also A_1 is invertible, which leads to the following unique trivial solution:

$$\underline{c}_1^n = \underline{0}.$$

This contradicts the fact that each constant should be a solution:

$$\underline{c}_1^n = \begin{bmatrix} C \\ \vdots \\ C \end{bmatrix}, \qquad \text{for all } C \in \mathbb{R}.$$

The latter can be seen by considering (13) for each of the isolated cells and multiplying the equation with the arbitrary constant C. Here, it is crucial that the isolated cells do not depend on the boundary elements. Indeed, there must be an index $j \in \mathcal{A}_i^n$ for which a non-decreasing path exists.

In conclusion, a path with non-decreasing values can be found for all $i = 1, ..., I$ and all $n = 0, ..., N$. Similarly, a path with non-increasing values can be found. Thus, the global discrete maximum principle is satisfied. ■

Remark 3.13 (Unique solution)
If each solution depends on the previous time in the sense that $\mathcal{B}_i^n$ is non-empty for all $i = 1, ..., I$ and $n = 1, ..., N$, then, it follows from (13) that the matrix A is (strictly) diagonally dominant, thus invertible. However, it is not a necesary condition. ⌟

Theorem 3.14
Consider the local theta scheme (Definition 3.3). Suppose that, for all $i = 1, ..., I$ and $j \in \mathcal{J}_i$:

(i) the flux coefficients are nonnegative:

$$\gamma_{ij}^n \geq 0, \tag{14}$$

(ii) the flux coefficients satisfy:

$$\sum_{j \in \mathcal{J}_i} \Big(-\gamma_{ji}^n + \gamma_{ij}^n \Big) = 0, \tag{15}$$

(iii) the local theta coefficients are symmetric:

$$\theta_{ij}^n = \theta_{ji}^n, \tag{16}$$

(iv) the coefficients satisfy the following CFL-like condition:

$$|\mathcal{V}_i| - \sum_{j \in \mathcal{J}_i} (t^n - t^{n-1})(1 - \theta_{ij}^n)\gamma_{ji}^n \geq 0. \tag{17}$$

(v) the matrix A in the form (5) is invertible.

Then, the scheme is conservative (Definition 3.4), absolutely stable (Definition 3.5) with respect to $\|.\|_\infty$, and positivity-preserving (Definition 3.6), and it satisfies the local discrete maximum principle (Definition 3.7). So, the scheme is CSPW.

PROOF:
First, it will be shown that the scheme is conservative (Definition 3.4). To this end,

consider the form (6) and observe that the flux function $\phi_{ij}^{n,m}$ is anti-symmetric:

$$-|\mathcal{S}_{ij}|\phi_{ij}^{n,m} = -\gamma_{ji}^{n}c_{i}^{m} + \gamma_{ij}^{n}c_{j}^{m} = |\mathcal{S}_{ji}|\phi_{ji}^{n,m} = |\mathcal{S}_{ij}|\phi_{ji}^{n,m}. \tag{18}$$

Since, the local theta coefficients are symmetric, the flux Φ_{ij}^{n} is also anti symmetric:

$$\begin{aligned}\Phi_{ji}^{n} &= (1-\theta_{ji}^{n})|\mathcal{S}_{ji}|\phi_{ji}^{n,n-1} + \theta_{ji}^{n}|\mathcal{S}_{ji}|\phi_{ji}^{n,n} \\ &\overset{(16)}{=} (1-\theta_{ij}^{n})|\mathcal{S}_{ij}|\phi_{ji}^{n,n-1} + \theta_{ij}^{n}|\mathcal{S}_{ij}|\phi_{ji}^{n,n} \\ &\overset{(18)}{=} -(1-\theta_{ij}^{n})|\mathcal{S}_{ij}|\phi_{ij}^{n,n-1} - \theta_{ij}^{n}|\mathcal{S}_{ij}|\phi_{ij}^{n,n} = -\Phi_{ij}^{n}.\end{aligned} \tag{19}$$

As a result, the scheme is conservative.

To show the rest of the claim, the scheme is written in the form (4) where

$$\begin{aligned} a_{ii}^{n} =& |\mathcal{V}_i| + (t^n - t^{n-1}) \sum_{j\in\mathcal{J}_i} \theta_{ij}^{n}\gamma_{ji}^{n} > 0, \\ \mathcal{A}_i^n =& \Big\{ j \in \mathcal{J}_i \,|\, a_{ij}^{n} := -(t^n - t^{n-1})\theta_{ij}^{n}\gamma_{ij}^{n} \neq 0 \Big\}, \\ \mathcal{B}_i^n =& \Big\{ j \in \mathcal{J}_i \cup \{i\} \,|\, b_{ij}^{n} := (t^n - t^{n-1})(1-\theta_{ij}^{n})\gamma_{ij}^{n} \neq 0 \quad (j \neq i), \\ & b_{ii}^{n} := |\mathcal{V}_i| - (t^n - t^{n-1}) \sum_{j\in\mathcal{J}_i} (1-\theta_{ij}^{n})\gamma_{ji}^{n} \neq 0 \Big\}. \end{aligned}$$

Observe that the scheme is of positive type (Definition 3.11) under the given circumstances. Hence, Theorem 3.12 implies that the local discrete maximum principle and the global discrete maximum principle are satisfied and that the scheme is positivity-preserving and absolutely stable with respect to $\|.\|_\infty$. ■

4. Local theta upwind scheme: choosing the coefficients in practice

The previous section introduced the local theta scheme and derived theoretical conditions for the local theta coefficients that ensure that the scheme is CSPW.

This section provides a practical strategy to choose the coefficients in terms of explicit expressions, which results in the local theta upwind scheme (Definition 4.1).

The local theta upwind scheme (Definition 4.1) combines the local theta time discretisation with first order upwind flux functions, so that (14) is satisfied. Moreover, the local theta coefficients are chosen (nearly) as small as possible, to minimise the amount of numerical diffusion (cf. Section 2), but large enough to ensure that the CFL-like condition (17) and the symmetry condition (16) are satisfied. If, additionally, the scheme has a unique solution and if the velocity profile is conservative so that (15) is satisfied, it follows immediately from Theorem 3.14 that the local theta upwind scheme is CSPW (Theorem 4.4).

Altogether, the local theta upwind scheme is a practical version of the local theta scheme that is CSPW without a restriction on the time step. Additionally, the scheme suffers minimally from numerical diffusion in the sense that the local theta coeffcients are chosen as small as possible.

The next section incorporates the flux corrected transport algorithm into the model, to improve the accuracy even more.

Definition 4.1 (Local theta upwind scheme)
Consider Case 3.2. The local theta upwind scheme results from the local theta scheme (Definition 3.3) by using first order upwind flux functions:

$$\gamma_{ij}^n = |\mathcal{S}_{ij}| \max\{u_{ji}^n, 0\}. \tag{20}$$

Note that (14) is satisfied in this case. Moreover, the local theta coefficicients are chosen as follows. First, an auxiliary coefficient θ_i^n is chosen such that the the CFL-like condition (17) is satisfied for cell $\mathcal{V}_i$, if the local theta coefficients are equal to this vale θ_i^n at each face of the cell, i.e. if $\theta_{ij}^n = \theta_i^n$ for all $j \in \mathcal{J}_i$:

$$\theta_i^n = 1 - \frac{|V_i|}{(t^n - t^{n-1}) \sum_{j \in \mathcal{J}_i} \gamma_{ji}^n}.$$

The local theta coefficients are now chosen as the following maximum:

$$\theta_{ij}^n = \max\left\{0, \theta_i^n, \theta_j^n\right\},$$

in order to satisfy the symmetry condition (16).

Example 4.2
Case 3.2 for $m = 1$ and a constant velocity $u > 0$, so that the advection equation reduces to:

$$\frac{\partial c}{\partial t} + u \frac{\partial c}{\partial x} = 0.$$

In this case, the local theta upwind scheme (Definition 4.1) reads:

$$\begin{aligned}
\frac{\Delta x_i c_i^n - \Delta x_i c_i^{n-1}}{\Delta t} &= -(1 - \theta_{i,i+1}^n) u (c_i^{n-1} + c_{i-1}^{n-1}) - \theta_{i,i+1}^n u (c_i^n + c_{i-1}^n), \\
\theta_{i+1,i}^n &= \theta_{i,i+1}^n = \max\left\{0, 1 - \frac{\Delta x_i}{u \Delta t}\right\},
\end{aligned}$$

where Δt is the constant time step and Δx_i is the width of cell $\mathcal{V}_i$. ⌟

Remark 4.3
In the local theta upwind scheme (4.1), θ_i^n 'blames' each face of cell i equally if (17) is not satisfied. Alternatively, only the face with the largest incoming flux could be blamed, by using a relatively large nonzero local theta coefficient for that face only. After that, the coefficients need to be updated to ensure symmetry. Another strategy is to use weighted coefficients in accordance with the size of the incoming flux. In other words, there remains a certain amount of freedom in the way that the coefficients are chosen. ⌟

Theorem 4.4
Consider the local theta upwind scheme (Definition 4.1). Suppose that the scheme has a unique solution (cf. Remark 3.13) and that the given discrete velocity profile u_{ij}^n is mass conservative in the sense that

$$\sum_{j \in \mathcal{J}_i} |\mathcal{S}_{ij}| u_{ij}^n = 0. \tag{21}$$

Then, the scheme is conservative (Definition 3.4), absolutely stable (Definition 3.5) with respect to $\|.\|_\infty$, and positivity-preserving (Definition 3.6), and it satisfies the local discrete maximum principle (Definition 3.7). In other words, the scheme is CSPW.
Proof:
This follows immediately from Theorem 3.14. ■

5. Local theta upwind FCT scheme: increasing the accuracy even more

The previous section proposed the local theta upwind scheme, which is CSPW for any time step, and prevents unnecesary numerical diffusion as much as possible due to its local approach. However, the order of accuracy of the scheme is not higher than one. This originates from the wish for a CSPW scheme, which is in conflict with the tendency of higher order schemes to create spurious wiggles and negative results.

Fortunately, similar problems have been tackeled before with the help of the Flux Correct Transport (FCT) algorithm. This algorithm combines a non-oscillatory first-order flux function with an accurate higher-order flux function by means of a nonlinear limiter, whose purpose it is to keep wiggles under control. Roughly speaking, it uses a convex combination of the two flux functions, instead of just one of them. This strategy can also be described as updating the first-order flux with a limited correction flux. This section improves the accuracy of the the local theta upwind scheme by incorporating an FCT approach into the scheme (Definition 5.1).

To upgrade the local theta upwind scheme with a an FCT approach, two elements are needed: a flux correction and a limiter.

The local theta upwind FCT scheme (Definition 5.1) uses the limiter that was proposed by Zalesak [18]. This limiter (practically) preserves the local discrete maximum principle, which implies that the scheme remains CSPW (see also Section 3).

The flux correction is obtained by computing the difference of the first-order upwind flux and a flux that corresponds to a stable higher-order scheme. The local theta upwind FCT scheme uses a combination of the Lax-Wendroff scheme and the central scheme in the following manner.

Lax-Wendroff is a popular flux corrector, as its first-order error (numerical diffusion) is equal to zero (cf. Proposition A.2 later on). However, the scheme is unstable if the 'traditional' CFL-condition is not satisfied. Therefore, in the local theta upwind FCT scheme, Lax-Wendroff flux correction should only be applied at the explicit faces ($\theta_{ij}^n = 0$).

An alternative flux correction is obtained with the help of the theta central scheme with $\theta = \frac{1}{2}$. Similar to the Lax-Wendroff scheme, it can be shown that this scheme does not introduce numerical diffusion (cf. Proposition A.2 for $\theta = \frac{1}{2}$ later on). Additionally, the scheme is unconditionally stable, so it could be applied in the entire domain. Nonetheless, a new difficulty arises. The flux correction now depends on the unknown solution at the new time. Here, it is chosen to avoid the necessity of solving an implicit nonlinear system by approximating the flux correction at the new time with the help of the first-order local theta upwind approximation. This way, the flux correction can be approximated in an explicit manner. Because this strategy introduces an unknown error, central flux correction is only applied at implicit faces. At explicit faces, Lax-Wendroff flux-correction is applied.

Altogether, for a sufficiently small time step, the local theta upwind FCT scheme coincides with the original explicit FCT scheme of Boris and Book [2]. For a larger time step, for which the latter is no longer applicable, the local theta upwind FCT scheme remains CSPW and should achieve a higher accuracy than the (local) theta upwind scheme.

The next section illustrates the performance of the local theta upwind (FCT) scheme by means of two numerical examples.

Definition 5.1 (Local theta FCT upwind scheme)
Consider Case 3.2. A local theta upwind FCT approximation of c_i^n can be obtained as follows.

(i) Compute a (first-order) local theta upwind approximation $\hat{c}_i^n$ by means of the local theta upwind scheme (Definition 4.1).

(ii) Compute the flux correction $\Delta\Phi_{ij}^n = \Psi_{ij}^n - \Phi_{ij}^n$ with the help of a (higher-order) scheme of the form:

$$\frac{|\mathcal{V}_i|c_i^n - |\mathcal{V}_i|c_i^{n-1}}{t^n - t^{n-1}} = \sum_{j\in\mathcal{J}_i} -|\mathcal{S}_{ij}|\Psi_{ij}^n.$$

At explicit faces ($\theta_{ij}^n = 0$), Lax-Wendroff-based flux correction is applied:

$$\Psi_{ij}^n =\Phi_{ij}^n + \left(\frac{1}{2} - \frac{u_{ij}(t^n - t^{n-1})}{2\|\underline{x}_j - \underline{x}_i\|_2}\right) u_{ij}^n \left(c_j^{n-1} - c_i^{n-1}\right). \tag{22}$$

At the implicit faces ($\theta_{ij}^n > 0$), the flux correction is based on central discretisation in combination with a Crank-Nicolson approach:

$$\Psi_{ij}^n = \frac{1}{2}u_{ij}^n \frac{c_i^{n-1} + c_j^{n-1}}{2} + \frac{1}{2}u_{ij}^n \frac{c_i^n + c_j^n}{2}. \tag{23}$$

The flux correction $\Delta\Phi_{ij}^n$ may depend on the unkown c_i^n. An explicit expression for the flux correction is obtained by using the approximation $c_i^n \approx \hat{c}_i^n$.

(iii) Compute the flux limiter l_{ij}^n (see also [18]):

i. Since $\Delta\Phi_{ij}^n$ often corrects the numerical diffusion of the first-order flux, it is sometimes referred to as anti-diffusion. Because an anti-diffusion term should not behave as a diffusion term, put $\Delta\Phi_{ij}^n = 0$ if

$$\Delta\Phi_{ij}^n(\hat{c}_i^n - \hat{c}_j^n) > 0.$$

If this prelimiting step is not performed, the flux correction may smooth the low-order solution, or it may cause small-scale numerical ripples [9, p.541]. Since it is unphysical for an anti-diffusion term to be directed from a higher concentration to a lower concentration, the effect of the adjustment above is minimal in practice [18, p. 342].

ii. Define an upper and a lower bound for c_i^n by means of the first-order approximation $\hat{c}_i^n$:

$$c_i^{\max} = \max_{j \in \mathcal{J}_i \cup \{i\}} \{\hat{c}_j^n\},$$
$$c_i^{\min} = \min_{j \in \mathcal{J}_i \cup \{i\}} \{\hat{c}_j^n\}.$$

The flux limiter will be such that c_i^n is bounded by these limits.

iii. The amount of mass that flows into cell $\mathcal{V}_i$ as a result of the flux correction (without the limiter) reads:

$$\lambda_i^+ = \sum_{j \in \mathcal{J}_i} (t^n - t^{n-1}) |\mathcal{S}_{ij}| \max\{0, -\Delta\Phi_{ij}^n\}.$$

The allowed mass increase is, however:

$$\mu_i^+ = |\mathcal{V}_i| \left(c_i^{\max} - \hat{c}_i^n\right).$$

Thus, the fraction of mass that is allowed to flow into the cell is given by:

$$\nu_i^+ = \begin{cases} \min\left\{1, \dfrac{\mu_i^+}{\lambda_i^+}\right\}, & \lambda_i^+ > 0, \\ 0, & \lambda_i^+ = 0, \end{cases}$$

Introduce analogue quantities for mass decrease:

$$\begin{aligned} \lambda_i^- &= \sum_{j \in \mathcal{J}_i} (t^n - t^{n-1}) |\mathcal{S}_{ij}| \max\{0, \Delta\Phi_{ij}^n\}, \\ \mu_i^- &= |\mathcal{V}_i| \left(\hat{c}_i^n - c_i^{\min}\right), \\ \nu_i^- &= \begin{cases} \min\left\{1, \dfrac{\mu_i^-}{\lambda_i^-}\right\}, & \lambda_i^- > 0, \\ 0, & \lambda_i^- = 0. \end{cases} \end{aligned}$$

iv. The limiter is the mass fraction that is allowed by both adjacent cells:

$$l_{ij}^n := \begin{cases} \min\{\nu_j^+, \nu_i^-\}, & \Delta\Phi_{ij}^n \geq 0, \\ \min\{\nu_i^+, \nu_j^-\}, & \Delta\Phi_{ij}^n < 0. \end{cases}$$

(iv) Compute the local theta FCT solution by means of:

$$\frac{|\mathcal{V}_i| c_i^n - |\mathcal{V}_i| \hat{c}_i^n}{t^n - t^{n-1}} = \sum_{j \in \mathcal{J}_i} -|\mathcal{S}_{ij}| l_{ij}^n \Delta\Phi_{ij}^n. \tag{24}$$

6. Numerical examples: putting the local theta scheme to the test

The previous section incorprated the FCT algorithm into the local theta upwind scheme to improve the accuracy.

This section illustrates the performance of the resulting local theta upwind FCT scheme by means of two periodic advection test problems. A three-dimensional real-life application in the form of salinity transport in an estuary near Hong Kong can be found in [13].

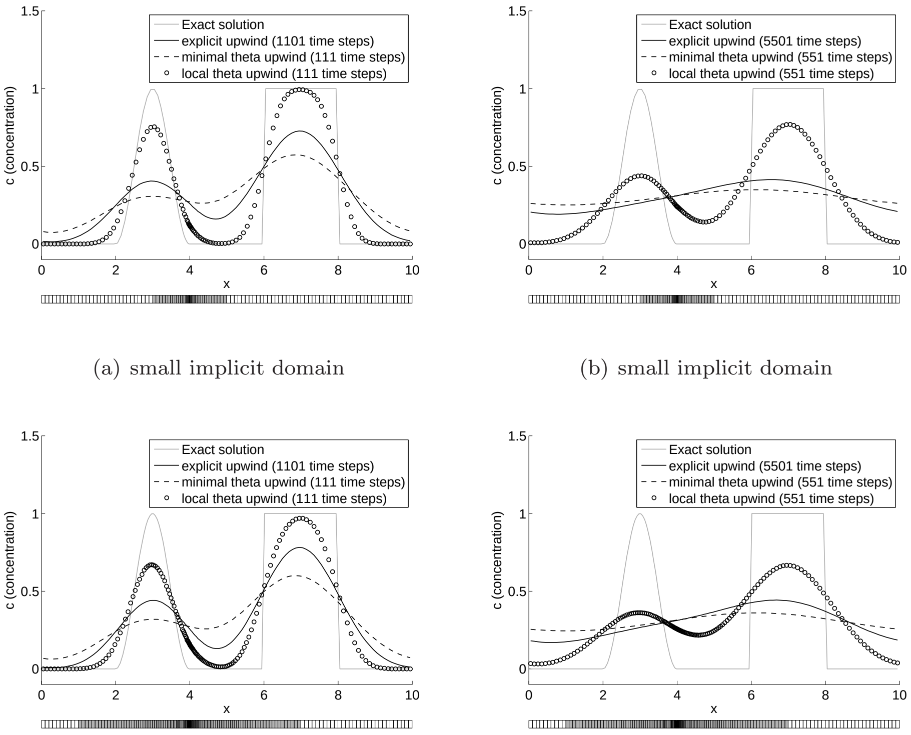
First, consider a one-dimensional advection problem with constant velocity $u = 1\,(\mathrm{ms}^{-1})$. The boundary conditions are chosen to be periodic, so the exact solution is the translation of the initial condition with a period of $10\,(\mathrm{s})$. Figure 2 shows the results after one and five periods, so the initial condition coincides with the displayed exact solution. Furthermore, the grid is illustrated below the chart. The gray cells are implicit in the sense that at least one of their faces uses a strictly positive local theta coefficient.

Figure 2 compares the results of the local theta upwind scheme (see Example 4.2) with the Euler forward upwind scheme (see Example 4.2 for local theta coefficients that are all equal to zero), and the mimimal theta upwind scheme (see Example 4.2 for the smallest possible constant value of theta for which the scheme is CSPW). The explicit upwind scheme requires a much smaller time step in order to be CSPW, and, as a consequence, many more time steps. The local theta upwind scheme suffers least from numerical diffusion without the restriction on the time step. Nonetheless, improvement of the accuracy by means of the flux corrected transport algorithm remains desirable.

Figure 3 displays the performance of the local theta upwind FCT scheme (see Definition 5.1) in comparison to the explicit FCT scheme of of Boris and Book [2] (see Definition 5.1 for local theta coefficients that are all equal to zero). Even though the latter uses a much larger time step, the accuracy of the schemes is quite comparable. As a matter of fact, the local theta upwind FCT scheme may achieve even higher accuracy than the explicit scheme, especially if the implicit part of the domain is not too large. This can be explained by the fact that the local theta upwind scheme forms a better starting point for the flux correction than the explicit upwind scheme, as is illustrated by Figure 2.

Next, the same schemes are applied for a two-dimensional advection problem with a constant angular velocity. The velocity is chosen such that the exact solution is the clockwise rotation of the initial condition with a period of six hours. An illustration of the initial condition, the grid, and the local theta coefficients can be found in Figure 4. The results after one and four periods are displayed in Figure 5 and Figure 6. At first glance, the performance of the schemes is comparable to that for the one-dimensional test case. However, this test case reveals a difference in behaviour of the FCT schemes in the explicit and implicit part of the spatial domain.

The explicit part of the spatial domain is at the center, because the velocity is relatively low in that area. This is also illustrated by Figure 4. As a consequence, aside from the time step, the two FCT schemes coincide in this area. At the same time, the local theta FCT scheme appears to perform better than the explicit FCT scheme. This can only be explained by the only difference between the two schemes: the fact that the local theta FCT scheme uses *larger* time step. Indeed, by considering Figure 1 for $\theta = 0$, it becomes



(a) small implicit domain

(b) small implicit domain

(c) large implicit domain

(d) large implicit domain

Fig. 2. Performance of the local theta upwind scheme for a one-dimensional advection problem with a constant velocity and periodic boundary conditions after one (left) and five (right) periods

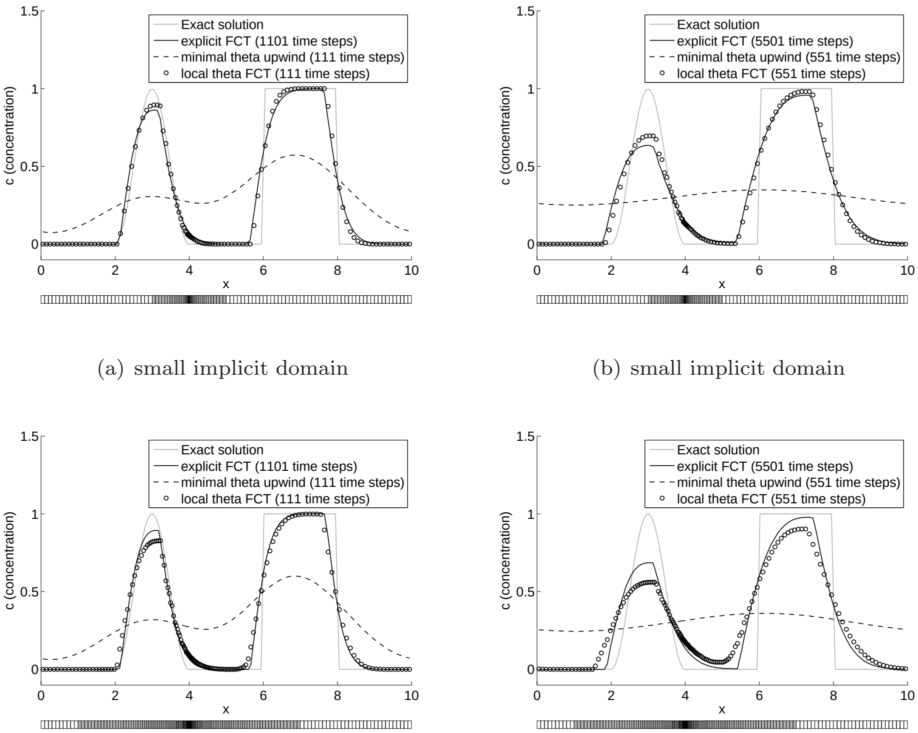
clear that a larger time step corresponds to less numerical diffusion. As a result, the local theta FCT scheme has a better starting point before the flux correction is applied.

In the implicit part of the spatial domain, the FCT schemes behave differently as far as diffusion and disperion errors are concerned. The explicit FCT scheme exhibits a larger diffusion error, whereas the local theta FCT scheme shows a larger dispersion error. A different limiter might lead to better dispersion characterics, but this is not investigated in this paper.

Altogether, the two numerical examples illustrate that the accuracy of the local theta upwind FCT scheme can measure up to that of the expensive explicit FCT scheme.

7. Conclusion

The local theta upwind FCT scheme (Definition 5.1) has been proposed for advection problems with strongly varying meshes and velocity profiles. The scheme is stable, positivity-preserving and free of spurious wiggles for any time step. Because the numerical diffusion is locally minimised, the accuracy can measure up to that of an expensive



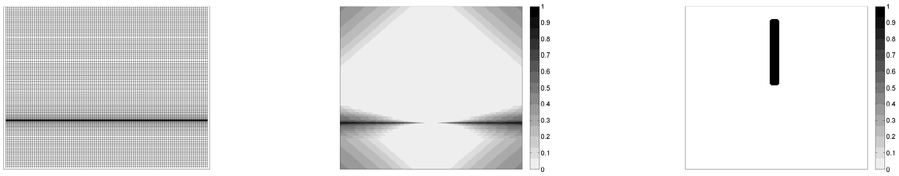
(a) small implicit domain

(b) small implicit domain

(c) large implicit domain

(d) large implicit domain

Fig. 3. Performance of the local theta upwind FCT scheme for a one-dimensional advection problem with a constant velocity and periodic boundary conditions after one (left) and five (right) periods



(a) grid

(b) Local theta coefficients

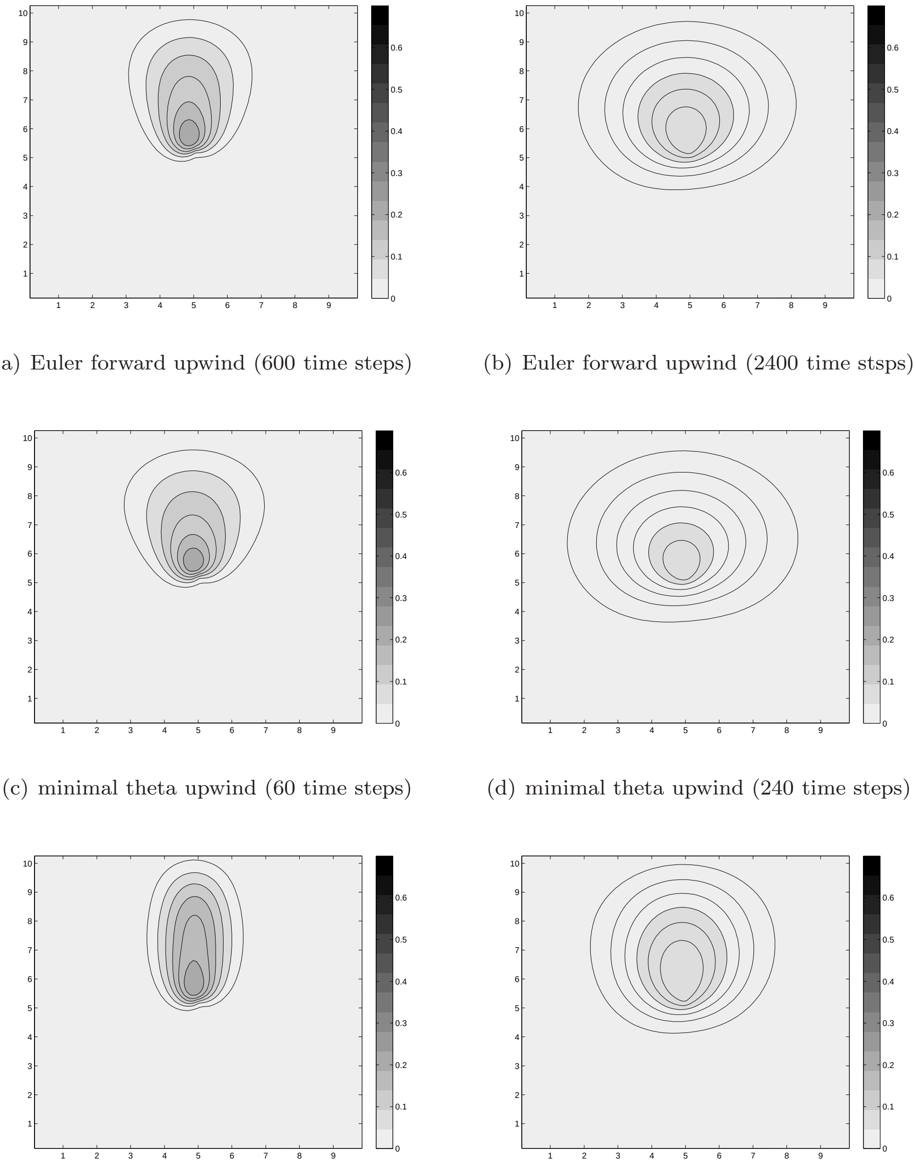
(c) initial condition

Fig. 4. Two-dimensional advection problem with a constant angular velocity

explicit FCT scheme. Nevertheless, there are several possibilities to improve the performance of the scheme even more.

First of all, a different FCT approach could be used. There is a large variety of limiters and higher-order schemes available that may lead to better results.

Moreover, the flux correction, which depends on the unknown new solution, could be computed in more advanced manner to obtain higher accuracy. This paper simply substitutes the known first-order local theta upwind solution. Alternatively, an iterative



(a) Euler forward upwind (600 time steps)

(b) Euler forward upwind (2400 time stsps)

(c) minimal theta upwind (60 time steps)

(d) minimal theta upwind (240 time steps)

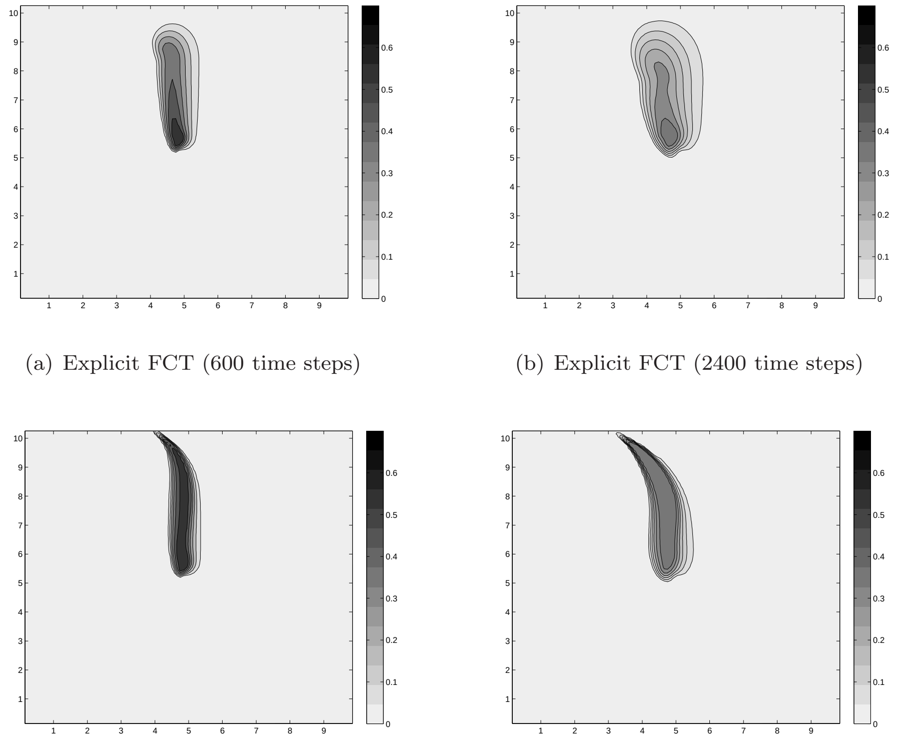
(e) Local theta upwind (60 time steps)

(f) Local theta upwind (240 times steps)

Fig. 5. Performance of the local theta upwind scheme for a two-dimensional advection problem with a constant angular velocity after one (left) and four (right) periods

approach similar to that of Kuzmin et al. [9,8,10] could be applied.

Additionally, there is a certain amount of freedom in the way that the local theta coefficients can be chosen. A different choice may lead to a better accuracy. For instance, the local theta coefficients could be chosen in accordance with the flux (cf. Remark 4.3).



(a) Explicit FCT (600 time steps)

(b) Explicit FCT (2400 time steps)

(c) Local theta upwind FCT (60 time steps)

(d) Local theta upwind FCT (240 time steps)

Fig. 6. Performance of the local theta upwind FCT scheme for a two-dimensional advection problem with a constant angular velocity after one (left) and four (right) periods

Finally, the nature of the local theta coefficients can be exploited when solving the linear system that results from the local theta upwind scheme. The linear system could be reordered to obtain a smaller implicit system and an explicit system. Such a reordering strategy may reduce the overall computational costs, especially when the explicit system is large. The latter is the case, if there are only a few cells for which the velocity is large in proportion to the size of the cell volume, which is typical for real-life applications.

Altogether, the local theta scheme is a time integration scheme that introduces new flexibility that can be explored and exploited in many ways.

Acknowledgements

This article is the result of an M.Sc. project that was carried out at WL|Delft Hydraulics. In particular, Menno Genseberger and Leo Postma have contributed greatly to this project.

Appendix A. Modified equation: revealing the nature of the lowest order error

Here, the modified equation equation [16] is derived for the theta upwind scheme, the Lax-Wendroff scheme, and the theta central scheme. This provides more inside into the nature of the errors of these schemes.

Proposition A.1
Consider Case 3.2 for one-dimensional case ($m = 1$), a constant velocity $u > 0$, a constant cell width $\Delta x > 0$ and a constant time step $\Delta t > 0$. Furthermore, consider the theta upwind scheme:

$$\frac{c_i^{n+1} - c_i^n}{\Delta t} + \theta u \frac{c_i^{n+1} - c_{i-1}^{n+1}}{\Delta x} + (1-\theta) u \frac{c_i^n - c_{i-1}^n}{\Delta x} = 0.$$

Let $\eta(x,t)$ be a smooth function such that $\eta(x_i, t^n) = c_i^n$. Then, the corresponding *modified equation* reads

$$\begin{aligned}\frac{\partial \eta}{\partial t}(x_i, t^n) + u \frac{\partial \eta}{\partial x}(x_i, t^n) = &\underbrace{\frac{u \Delta x}{2}\left(1 - (1-2\theta)\frac{u \Delta t}{\Delta x}\right)}_{\text{Numerical diffusion coefficient}} \frac{\partial^2 \eta}{\partial x^2}(x_i, t^n) \\ &+ O\left(\Delta t^2\right) + O\left(\Delta x^2\right) + O(\Delta x \Delta t).\end{aligned}$$

<u>Proof</u>:
Because $\eta(x_i, t^n) = c_i^n$,

$$\begin{aligned}&\frac{\eta(x_i, t^{n+1}) - \eta(x_i, t^n)}{\Delta t} \\ +\theta u &\frac{\eta(x_i, t^{n+1}) - \eta(x_{i-1}, t^{n+1})}{\Delta x} \\ +(1-\theta) u &\frac{\eta(x_i, t^n) - \eta(x_{i-1}, t^n)}{\Delta x} = 0.\end{aligned}$$

Using a Taylor expansion of η around t^n results in:

$$\begin{aligned}&\frac{\partial \eta}{\partial t}(x_i, t^n) + \frac{\Delta t}{2}\frac{\partial^2 \eta}{\partial t^2}(x_i, t^n) + \frac{\Delta t^2}{6}\frac{\partial^3 \eta}{\partial t^3}(x_i, \tau_1) \\ +\theta u \Bigg(&\frac{\eta(x_i, t^n) + \Delta t \frac{\partial \eta}{\partial t}(x_i, t^n) + \frac{\Delta t^2}{2}\frac{\partial^2 \eta}{\partial t^2}(x_i, t^n) + \frac{\Delta t^3}{6}\frac{\partial^3 \eta}{\partial t^3}(x_i, \tau_2)}{\Delta x} \\ - &\frac{\eta(x_{i-1}, t^n) + \Delta t \frac{\partial \eta}{\partial t}(x_{i-1}, t^n) + \frac{\Delta t^2}{2}\frac{\partial^2 \eta}{\partial t^2}(x_{i-1}, t^n) + \frac{\Delta t^3}{6}\frac{\partial^3 \eta}{\partial t^3}(x_{i-1}, \tau_3)}{\Delta x}\Bigg) \\ &+(1-\theta) u \frac{\eta(x_i, t^n) - \eta(x_{i-1}, t^n)}{\Delta x} = 0,\end{aligned}$$

for certain $\tau_1, \tau_2, \tau_3 \in [t^n, t^{n+1}]$. Using a Taylor expansion of η around x_i yields:

$$
\begin{aligned}
&\frac{\partial \eta}{\partial t}(x_i,t^n)+\frac{\Delta t}{2}\frac{\partial^2\eta}{\partial t^2}(x_i,t^n)+\frac{\Delta t^2}{6}\frac{\partial^3\eta}{\partial t^3}(x_i,\tau_1)\\
&+\theta u\Bigg(\frac{\partial \eta}{\partial x}(x_i,t^n)-\frac{\Delta x}{2}\frac{\partial^2\eta}{\partial x^2}(x_i,t^n)+\frac{\Delta x^2}{6}\frac{\partial^3\eta}{\partial x^3}(\xi_1,t^n)\\
&\qquad+\Delta t\frac{\partial}{\partial x}\frac{\partial \eta}{\partial t}(x_i,t^n)-\frac{\Delta t\Delta x}{2}\frac{\partial^2}{\partial x^2}\frac{\partial \eta}{\partial t}(\xi_2,t^n)\\
&\qquad+\frac{\Delta t^3}{6\Delta x}\frac{\partial^3\eta}{\partial t^3}(x_i,\tau_2)-\frac{\Delta t^3}{6\Delta x}\frac{\partial^3\eta}{\partial t^3}(\xi_3,\tau_3)\Bigg)\\
&+(1-\theta)u\left(\frac{\partial \eta}{\partial x}(x_i,t^n)-\frac{\Delta x}{2}\frac{\partial^2\eta}{\partial x^2}(x_i,t^n)+\frac{\Delta x^2}{6}\frac{\partial^3\eta}{\partial x^3}(\xi_4,t^n)\right)=0,
\end{aligned}
$$

for certain $\xi_1,\xi_2,\xi_3,\xi_4\in[x_{i-1},x_i]$. Rewriting gives:

$$
\begin{aligned}
\frac{\partial \eta}{\partial t}(x_i,t^n)+u\frac{\partial \eta}{\partial x}(x_i,t^n)=&\frac{u\Delta x}{2}\frac{\partial^2\eta}{\partial x^2}(x_i,t^n)\\
&-\frac{\Delta t}{2}\frac{\partial^2\eta}{\partial t^2}(x_i,t^n)-\theta u\Delta t\frac{\partial}{\partial x}\frac{\partial \eta}{\partial t}(x_i,t^n)\\
&-\frac{\Delta t^2}{6}\frac{\partial^3\eta}{\partial t^3}(x_i,\tau_1)\\
&-\theta u\Bigg(\frac{\Delta x^2}{6}\frac{\partial^3\eta}{\partial x^3}(\xi_1,t^n)-\frac{\Delta t\Delta x}{2}\frac{\partial^2}{\partial x^2}\frac{\partial \eta}{\partial t}(\xi_2,t^n)\\
&+\frac{\Delta t^3}{6\Delta x}\frac{\partial^3\eta}{\partial t^3}(x_i,\tau_2)-\frac{\Delta t^3}{6\Delta x}\frac{\partial^3\eta}{\partial t^3}(\xi_3,\tau_3)\Bigg)\\
&-(1-\theta)u\frac{\Delta x^2}{6}\frac{\partial^3\eta}{\partial x^3}(\xi_4,t^n).
\end{aligned}
$$

Note that

$$
\begin{aligned}
\frac{\partial \eta}{\partial t}(x_i,t^n)&=-u\frac{\partial \eta}{\partial x}(x_i,t^n)+O(\Delta t)+O(\Delta x)\\
\frac{\partial^2 \eta}{\partial t^2}(x_i,t^n)&=u^2\frac{\partial^2 \eta}{\partial x^2}(x_i,t^n)+O(\Delta t)+O(\Delta x).
\end{aligned}
$$

Substitution yields:

$$
\begin{aligned}
\frac{\partial \eta}{\partial t}(x_i,t^n)+u\frac{\partial \eta}{\partial x}(x_i,t^n)=&\frac{u\Delta x}{2}\frac{\partial^2\eta}{\partial x^2}(x_i,t^n)\\
&-\frac{u^2\Delta t}{2}\frac{\partial^2\eta}{\partial x^2}(x_i,t^n)+\theta u^2\Delta t\frac{\partial^2\eta}{\partial x^2}(x_i,t^n)\\
&+O\left(\Delta t^2\right)+O\left(\Delta x^2\right)+O(\Delta x\Delta t).
\end{aligned}
$$

Rewriting ends the proof. ■

Proposition A.2
Consider Case 3.2 for one-dimensional case ($m=1$), a constant velocity $u>0$, a constant cell width $\Delta x>0$ and a constant time step $\Delta t>0$. Furthermore, consider the Lax-Wendroff scheme:

$$
\Delta x\frac{c_i^n-c_i^{n-1}}{\Delta t}+u\frac{c_{i+1}^{n-1}-c_{i-1}^{n-1}}{2}-\frac{u^2\Delta t}{2}\frac{c_{i-1}^{n-1}-2c_i^{n-1}+c_{i+1}^{n-1}}{\Delta x}=0.
$$

Let $\eta(x,t)$ be a smooth function such that $\eta(x_i,t^n)=c_i^n$. Then, for all $i=1,...,I$ and for all $n=1,...,N$:

$$\frac{\partial\eta}{\partial t}(x_i,t^n)+u\frac{\partial\eta}{\partial x}(x_i,t^n)=O\left(\Delta t^2\right)+O\left(\Delta x^2\right)+O(\Delta x\Delta t).$$

As a consequence, the numerical diffusion is zero.

<u>Proof</u>:
The proof is similar to the proof of Proposition A.1. ■

Proposition A.3
Consider Case 3.2 for one-dimensional case ($m=1$), a constant velocity $u>0$, a constant cell width $\Delta x>0$ and a constant time step $\Delta t>0$. Furthermore, consider the theta central scheme:

$$\Delta x\frac{c_i^n-c_i^{n-1}}{\Delta t}+\theta u\frac{c_{i+1}^n-c_{i-1}^n}{2}+(1-\theta)u\frac{c_{i+1}^{n-1}-c_{i-1}^{n-1}}{2}=0.$$

Let $\eta(x,t)$ be a smooth function such that $\eta(x_i,t^n)=c_i^n$. Then, for all $i=1,...,I$ and for all $n=1,...,N$:

$$\frac{\partial\eta}{\partial t}(x_i,t^n)+u\frac{\partial\eta}{\partial x}(x_i,t^n)=\underbrace{\left(\theta-\frac{1}{2}\right)u^2\Delta t}_{\text{Numerical diffusion coefficient}}\frac{\partial^2\eta}{\partial x^2}(x_i,t^n)$$
$$+O\left(\Delta t^2\right)+O\left(\Delta x^2\right)+O(\Delta x\Delta t).$$

As a consequence, the numerical diffusion is zero for $\theta=\frac{1}{2}$.

<u>Proof</u>:
The proof is similar to the proof of Proposition A.1. ■

References

[1] T. Barth, M. Ohlberger, Finite volume methods: Foundation and analysis, in: Encyclopedia of Computational Mechanics, John Wiley & Sons, 2004.

[2] J. P. Boris, D. L. Book, Flux corrected transport. 1. shasta, a fluid transport algorithm that works, Journal of Computational Physics 11 (1973) 38–69.

[3] R. Courant, K. Friedrichs, H. Lewy, On the partial difference equations of mathematical physics, IBM Journal 11 (1969) 215–243.

[4] D. Durran, Numerical methods for wave equations in geophysical fluid dynamics, Springer, New York, 1999.

[5] A. Harten, High resolution schemes for hyperbolic conservation laws, Journal of computational physics 49 (1983) 357–393.

[6] W. Hundsdorfer, J. Verwer, Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations, Springer, Berlin, 2003.

[7] D. Kroner, Numerical schemes for conservation laws, Wiley-Teubner, Chichester, 1997.

[8] D. Kuzmin, M. Möller, S. Turek, High-resolution femfct schemes for multidimensional conservation laws, Journal of Computational Physics 175 (2004) 525–558.

[9] D. Kuzmin, S. Turek, Flux correction tools for finite elements, Journal of Computational Physics 175 (2002) 525–558.

[10] D. Kuzmin, S. Turek, High-resolution fem-tvd schemes based on a fully multidimensional flux limiter, Journal of Computational Physics 198 (2004) 131–158.

[11] B. Leonard, The ultimate conservative difference scheme applied to unsteady one-dimensional advection, Computer methods in applied mechanics and engineering 88 (1991) 17–74.

[12] R. LeVeque, Finitie Volume Methods for Hyperbolic Problems, Cambridge University Press, New York, 2002.

[13] P. Slingerland, An accurate an robust finite volume method for the advection diffusion equation, Master's thesis, Delft University of Technology (2007).

[14] P. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws, SIAM journal on numerical analysis 21 (1984) 995–1011.

[15] E. Toro, Riemann solvers and numerical methods for fluid dynamics, Springer, Berlin, 1997.

[16] R. Warming, B. Hyett, The modified equation approach to the stability and accuracy analysis of finite-difference methods, Journal of computational physics 14 (1974) 159–179.

[17] P. Wesseling, Principles of computational fluid dynamics, Springer, Berlin, 2001.

[18] S. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, Journal of Computational Physics 31 (1979) 335–362.