Explicit and Implicit High-Resolution Finite Element Schemes Based on the Flux-Corrected-Transport Algorithm

Dmitri Kuzmin and Stefan Turek

Institute of Applied Mathematics (LS III), University of Dortmund
Vogelpothsweg 87, D-44227, Dortmund, Germany

Abstract. A new approach to flux correction for finite elements is presented. The low-order transport operator is constructed from the discrete high-order operator by elimination of negative off-diagonal entries, so as to enforce the M-matrix property. The corresponding antidiffusive terms can be decomposed into a sum of internodal fluxes (rather than element contributions). Thereby essentially one-dimensional flux correction tools can be applied on unstructured meshes. The proposed algorithm guarantees mass conservation and makes it possible to design both explicit and implicit FEM-FCT schemes based on a unified limiting procedure.

1 Introduction

A variety of CFD applications involve transport of quantities which must be conserved and remain positive. According to the Godunov theorem, linear monotonicity preserving methods are at most first-order accurate, so that the results are corrupted by excessive numerical diffusion. At the same time, high-order methods tend to produce spurious overshoots and undershoots in regions with steep gradients. The Flux-Corrected-Transport (FCT) algorithm overcomes these difficulties by adaptively switching between high- and low-order discretizations. Building on the FEM-FCT methodology introduced by Löhner et al. [5], we will derive an alternative formulation applicable to both explicit and implicit time-stepping [3],[4].

2 State of the Art

The concept of flux correction can be traced back to the celebrated SHASTA scheme of Boris and Book [1]. The crux of this algorithm consists in adding compensating antidiffusion to a monotone low-order solution, so as to obtain high-order accuracy on smooth solutions and preclude the arising of nonphysical oscillations in proximity to shocks and discontinuities. Unlike other high-resolution schemes, the FCT approach carries over to multidimensional problems [6] and finite element discretizations on unstructured meshes [5]. However, the classical FEM-FCT procedure is inherently explicit, which makes it extremely inefficient for computing steady-state solutions or solving problems with strongly varying velocities and/or mesh sizes.

3 Positivity Criterion

In order to derive a family of FEM-FCT schemes for arbitrary time-stepping, we need to introduce a rigorous positivity criterion. A particularly useful mathematical tool is provided by the concept of an M-matrix which is closely related to the discrete maximum principle for conservation laws.

Definition. A nonsingular discrete operator $A \in \mathbb{R}^{n\times n}$ is called an M-matrix if $a_{ij} \leq 0$ for $i \neq j$ and all the entries of A^{-1} are nonnegative.

If A is strictly diagonally dominant and $a_{ii} > 0$, while $a_{ij} \leq 0$ for $i \neq j$, then A is an M-matrix. Note that for M-matrices $Ax \geq 0$ implies that $x \geq 0$.

Assume that the numerical scheme can be represented in abstract matrix operator form as

$$Lu^{n+1} = Ru^n. \tag{1}$$

Then it is obvious that the positivity of u^n is inherited by u^{n+1} as long as L is an M-matrix and all entries of R are nonnegative.

4 Low-Order Scheme

An important ingredient of the FCT algorithm is the positivity-preserving low-order scheme. It can be constructed by adding artificial diffusion to a high-order discretization. Consider the convection-diffusion equation

$$\frac{\partial u}{\partial t} + \nabla\cdot(\mathbf{v}u) = \nabla\cdot(\epsilon\,\nabla u) \tag{2}$$

discretized in space by the Galerkin method. The semi-discrete problem reads

$$M_C\frac{du}{dt} = K^H u, \tag{3}$$

where M_C is the consistent mass matrix, and K^H denotes the high-order transport operator. Let the low-order counterpart of this scheme be given by

$$M_L\frac{du}{dt} = (K^H + D)u = K^L u, \tag{4}$$

where M_L stands for the row-sum lumped mass matrix, while D represents a tensor of artificial dissipation defined as

$$d_{ii} = -\sum_{k\neq i} d_{ik}, \qquad d_{ij} = d_{ji} = \max\{0, -k^H_{ij}, -k^H_{ji}\}, \qquad \forall\ i < j. \tag{5}$$

It is designed so as to eliminate all negative off-diagonal coefficients of the high-order operator while maintaining mass conservation.

After the discretization in time by the standard θ-scheme, we obtain a discrete scheme of form (1), which preserves positivity if the time step is

small enough. In particular, solutions to incompressible flow problems are local extremum diminishing and positivity preserving provided that

$$\Delta t \leq \frac{1}{1-\theta} \min_i \{-m_i/k_{ii}^L \mid k_{ii}^L < 0\}. \tag{6}$$

The fully implicit backward Euler method is unconditionally positive.

It can be readily verified that the classical upwind scheme is recovered for the pure convection equation in 1D. At the same time, the proposed technique is applicable to arbitrary meshes and multidimensional problems. In addition, the resulting scheme is less diffusive than upwind for problems with diffusion. In fact, the operators K^H and K^L are identical in diffusion dominated cases, since the coefficients are non-negative from the outset.

5 Flux-based FEM-FCT formulation

The backward Euler and the Crank-Nicolson schemes are unconditionally stable and can be used as a high-order method in conjunction with the Galerkin spatial discretization. At the same time, the forward Euler scheme needs to be stabilized by a proper amount of streamline diffusion. In this paper, we use the Lax–Wendroff method, whereby this stabilization corresponds to the second-order time derivative in the Taylor series expansion.

The high-order transport operator can be transformed into a low-order one as explained above. The resulting methods are related by the formula

$$(M_L - \theta \Delta t K^L) u^H = (M_L + (1-\theta)\Delta t K^L) u^n + F(u^H, u^n), \tag{7}$$

where the antidiffusion responsible for high spatial accuracy is given by

$$F(u^H, u^n) = -(M_C - M_L)\Delta u^H - \Delta t (K^L - K^H)[\theta u^H + (1-\theta)u^n] + \Delta t\, S u^n.$$

Here S stands for the streamline diffusion operator which is present only in the fully explicit scheme. Similar to the matrices $M_C - M_L$ and $K^L - K^H$, it is symmetric and features zero row/column sums. This enables us to decompose the antidiffusive terms into a sum of fluxes [4]

$$\begin{aligned} f_{ij} = &-m_{ij}(\Delta u_j^H - \Delta u_i^H) - \Delta t\, d_{ij}[\theta(u_j^H - u_i^H) + (1-\theta)(u_j^n - u_i^n)] \\ &+ \Delta t\, s_{ij}(u_j^n - u_i^n), \qquad f_{ji} = -f_{ij}, \quad i < j. \end{aligned} \tag{8}$$

These raw antidiffusive fluxes offset the errors induced by mass lumping, 'upwinding', and first-order time discretization (for the explicit scheme). Coefficients m_{ij}, d_{ij}, and s_{ij} denote the entries of the consistent mass matrix, artificial diffusion and streamline diffusion operators, respectively.

Hence, the flux-corrected version of (7) can be written in the form

$$m_i u_i^{n+1} - \theta \Delta t \sum_j k_{ij}^L u_j^{n+1} = m_i \tilde{u}_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \qquad \alpha_{ji} = \alpha_{ij}, \tag{9}$$

where α_{ij} denote the correction factors (to be defined below), while $\tilde{u}$ represents the positivity-preserving solution to the explicit subproblem

$$m_i \tilde{u}_i = m_i u_i^n + (1-\theta)\Delta t \sum_j k_{ij}^L u_j^n. \tag{10}$$

In essence, $\tilde{u}$ corresponds to an intermediate solution computed at the time instant $t^{n+1-\theta}$ by the explicit low-order scheme.

The newly introduced family of FCT schemes distinguishes itself in that it is applicable to explicit and implicit time discretizations alike. The fully explicit scheme is consistent with the standard FCT methodology. Note that implicit schemes require solving *two* non-symmetric linear systems per time step: one for the high-order solution u^H and one for the final solution u^{n+1}. Nevertheless, implicit methods are typically more efficient than explicit ones because larger time steps are admissible.

6 Limiting Strategy

It is obvious that the success of the FCT algorithm depends on the positivity of the provisional solution $\tilde{u}$ and on the choice of correction factors α_{ij}. For $\tilde{u}$ to be positive, the time step must satisfy condition (6) unless the scheme is fully implicit. The left-hand side of our schemes contains a monotone low-order operator and poses no hazard to positivity. Therefore, it is sufficient to require that the operator of the right-hand side be non-negative. This can be accomplished by tuning the correction factors as proposed by Zalesak [6].

It is worthwhile to start the limiting process with cancelling all antidiffusive fluxes directed down the gradient of $\tilde{u}$:

$$f_{ij} := 0, \quad \text{if} \quad f_{ij}(\tilde{u}_i - \tilde{u}_j) < 0. \tag{11}$$

This optional test should be applied *before* the flux correction step. It prevents the smoothing of the low-order solution which can lead to the arising of spurious ripples even though the positivity is preserved [2].

Let us consider the maximum and minimum solution values at the stencil S_i which consists of the node i and its nearest neighbors:

$$u_i^{\substack{\max \\ \min}} = \begin{matrix}\max \\ \min\end{matrix} \tilde{u}_j, \qquad j \in S_i. \tag{12}$$

In contrast with the standard FCT theory, these extrema no longer represent bounds for the final solution. Nevertheless, all antidiffusive fluxes which try to accentuate a local maximum or minimum must be completely canceled:

$$\alpha_{ij} = 0, \quad \text{if} \quad \tilde{u}_i = u_i^{\max}, \quad f_{ij} > 0 \quad \text{or} \quad \tilde{u}_i = u_i^{\min}, \quad f_{ij} < 0. \tag{13}$$

If this applies to all fluxes into the node i, we are done. Otherwise, the remaining fluxes have to be limited so as to comply with the positivity constraint.

The right-hand side of our scheme (9) admits the following representation:

$$RHS = m_i\tilde{u}_i + \sum_{j\neq i}\alpha_{ij}f_{ij} = m_i\tilde{u}_i + c_iQ_i, \qquad c_i = \frac{\sum_{j\neq i}\alpha_{ij}f_{ij}}{Q_i}, \tag{14}$$

where the multiplier Q_i is chosen to be

$$Q_i = \begin{cases} Q_i^+ = u_i^{\max} - \tilde{u}_i, & \text{if } \sum_{j\neq i}\alpha_{ij}f_{ij} > 0, \\ Q_i^- = u_i^{\min} - \tilde{u}_i, & \text{if } \sum_{j\neq i}\alpha_{ij}f_{ij} < 0, \\ 1, & \text{if } \sum_{j\neq i}\alpha_{ij}f_{ij} = 0. \end{cases} \tag{15}$$

By virtue of (13), we have $Q_i \neq 0$, so that no division by zero takes place. Furthermore, the coefficient c_i is always non-negative. Let the local extremum $u_i^{\substack{\max\\ \min}}$ be attained at a node k adjacent to the node i. This yields

$$RHS = m_i\tilde{u}_i + c_i(\tilde{u}_k - \tilde{u}_i) = (m_i - c_i)\tilde{u}_i + c_i\tilde{u}_k, \qquad c_i \geq 0. \tag{16}$$

It follows that the entries of the right-hand side operator are non-negative provided that $m_i \geq c_i$. It remains to show that Zalesak's limiter does enforce this constraint. Auxiliary quantities $P_i^{\pm}$ and $R_i^{\pm}$ are defined as

$$P_i^{\pm} = \frac{1}{m_i}\sum_{j\neq i}\begin{matrix}\max\\ \min\end{matrix}\{0, f_{ij}\}, \qquad R_i^{\pm} = \begin{cases} \min\{1, Q_i^{\pm}/P_i^{\pm}\}, & \text{if } P_i^{\pm} \neq 0, \\ 0, & \text{if } P_i^{\pm} = 0. \end{cases} \tag{17}$$

A sufficiently safe choice of correction factors is given by

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\}, & \text{if } f_{ij} \geq 0, \\ \min\{R_j^+, R_i^-\}, & \text{if } f_{ij} < 0. \end{cases} \tag{18}$$

This limiter is independent of the number of spatial dimensions and can be easily implemented as a 'black-box' routine which computes the correction factors given an array of antidiffusive fluxes for each pair of nodes.

The condition (13) is automatically satisfied, since $Q_i^{\pm} = 0$ spells $R_i^{\pm} = 0$ and $\alpha_{ij} = 0$. Hence, any enhancement of local extrema is neutralized by the limiter. Furthermore, the following estimate holds:

$$\sum_{j\neq i}\alpha_{ij}f_{ij} \leq \sum_{j\neq i}\alpha_{ij}\max\{0, f_{ij}\} \leq m_iR_i^+P_i^+ \leq m_iQ_i^+. \tag{19}$$

In much the same way, it can be verified that

$$\sum_{j\neq i}\alpha_{ij}f_{ij} \geq \sum_{j\neq i}\alpha_{ij}\min\{0, f_{ij}\} \geq m_iR_i^-P_i^- \geq m_iQ_i^-. \tag{20}$$

This proves that $m_i \geq c_i$ is satisfied. The above interpretation of Zalesak's limiter demonstrates that it can be utilized in an implicit framework.

7 Defect Correction

Many practical applications are described by *nonlinear* conservation laws. The simplest iterative treatment of nonlinearities is offered by the fixed point defect correction method. If we consider an abstract nonlinear system of the form $A(u)u = b$, then the basic nonlinear iteration can be formulated as

$$u^{(l+1)} = u^{(l)} - [C(u^{(l)})]^{-1}(A(u^{(l)})u^{(l)} - b), \tag{21}$$

where l is the iteration counter, and C is a suitably chosen 'preconditioner' which is typically 'inverted' by some iterative procedure. Taking

$$C(u^{(l)}) = M_L - \theta\Delta t K^L(u^{(l)})$$

yields the following iterative FEM-FCT algorithm for nonlinear problems

$$(M_L - \theta\Delta t K^L(u^{(l)}))u^{(l+1)} = (M_L + (1-\theta)\Delta t K^L(u^n))u^n + F(u^{(l)}, u^n). \tag{22}$$

The last term is composed from the (limited) antidiffusive fluxes. Flux correction can be performed after each outer iteration or just once after the high-order solution has converged. Even if the problem at hand is linear, it might be expedient to equip implicit schemes with an outer defect correction loop. Approximating the original matrix by a well-behaved preconditioner aids the convergence of iterative methods and results in a very robust solver.

8 Summary of the Algorithm

The proposed FEM-FCT procedure can be implemented as follows:

1. Discretize the governing equation by a high-order method.
2. Perform mass lumping and eliminate negative off-diagonal entries of the transport operator to construct the associated low-order scheme.
3. For $\theta < 1$, examine the diagonal entries of the low-order operator and adapt the time step so as to comply with the positivity condition.
4. Advance the solution in time by the high-order scheme to obtain u^H.
5. Assemble the raw antidiffusive fluxes f_{ij} for each pair of nodes.
6. Compute the positivity-preserving auxiliary solution $\tilde{u} = u^L(t^{n+1-\theta})$.
7. Cancel all antidiffusive fluxes directed down the gradient of $\tilde{u}$.
8. Apply Zalesak's limiter to calculate the correction factors α_{ij}.
9. Add the contribution of limited antidiffusive fluxes $\alpha_{ij}f_{ij}$ to the right-hand side of the low-order scheme.
10. For $\theta = 0$, scale the right-hand side by the diagonal matrix M_L^{-1}. Otherwise, solve the linear system for the end-of-step solution u^{n+1}.

In the nonlinear version of the algorithm, u^H is replaced by $u^{(l)}$, so that just one linear system per outer iteration (22) has to be solved. In addition, only the low-order matrix $C(u^{(l)})$ needs to be assembled and stored.

9 Iterative Solution

Explicit schemes do not require any advanced linear algebra tools, since the consistent mass matrix can be efficiently 'inverted' e.g. by just a few Jacobi-like iterations using the lumped mass matrix as a preconditioner [5]. Similarly, for relatively small time steps the nonsymmetric linear systems generated by implicit schemes can be solved by multigrid or BiCGSTAB methods with basic components like Jacobi, Gauß-Seidel or SOR. However, the large time steps afforded by the unconditionally positive backward Euler/FCT method may cause a severe deterioration of the matrix for the high-order system, so that iterative techniques may fail to converge. This can be rectified by resorting to defect correction and using ILU decomposition with appropriate renumbering as a smoother/preconditioner.

10 Steady-State Problems

The fully implicit FEM/FCT scheme is unconditionally positive but only first-order accurate in time, which makes it quite diffusive at large time steps. At the same time, it appears to be very attractive for computing solutions to steady-state convection-diffusion equations by 'time-marching'. In this case, the time step is merely an artificial parameter, whereas the accuracy of the converged solution depends entirely on the spatial discretization. To reduce the computational cost, the time steps should be chosen as large as possible. This makes explicit schemes non-competitive, since they are subject to a restrictive CFL-like condition. Moreover, the numerical solutions produced e.g. by the explicit Lax–Wendroff method are affected by streamline diffusion depending on the artificial (typically local) time step.

In practice, it is expedient to use the converged low-order solution as a 'predictor' for the time-dependent FEM-FCT algorithm. In this case, the cost of flux correction is minimized, since the initial guess should be close enough to the steady-state limit. Furthermore, the use of the consistent mass matrix is superfluous, since the temporal accuracy is not relevant anymore. Therefore, mass lumping is appropriate also for the high-order scheme.

11 Numerical Examples

As a standard one-dimensional example, consider convection of a discontinuous step function with unit velocity. The results produced by the explicit Lax–Wendroff/FCT and the implicit backward-Euler/FCT schemes are depicted in Fig. 1. Here and below, the dash-dotted line stands for the initial data, while the dotted line refers to the analytical solution. The computed profiles are completely free of oscillations. However, the BE/FCT method is seen to be diffusive because of the first-order time discretization. As the time step is refined, the accuracy approaches that of the LW/FCT scheme.

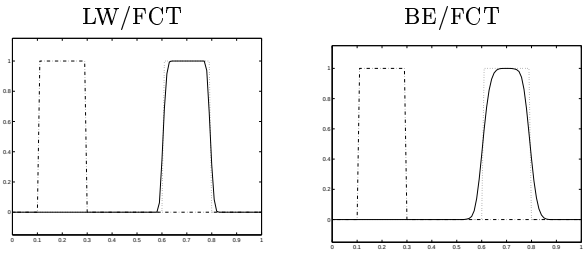

Fig. 1. Convection of a step function, $t = 0.5,\ \Delta x = 10^{-2},\ \Delta t = 10^{-3}$

For the inviscid Burgers equation, both FEM-FCT methods produce solutions of comparable quality (see Fig. 2). The nonlinearity was handled by the fixed point defect correction scheme as described above. Note that the shock propagates with correct speed, which implies that the mass is conserved.

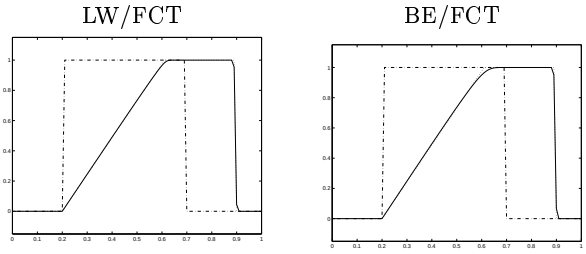

Fig. 2. Inviscid Burgers' equation, $\Delta x = 10^{-2},\ \Delta t = 10^{-3}$

Figure 3 illustrates the ability of the BE/FCT method to deal with singularly perturbed convection-diffusion problems. The exact steady-state solution exhibits a boundary layer due to the homogeneous Dirichlet boundary condition at the outflow boundary. This leads to nonphysical oscillations if the standard Galerkin method is applied. Remarkably, the flux-corrected solution obtained by the BE/FCT scheme is nodally exact.

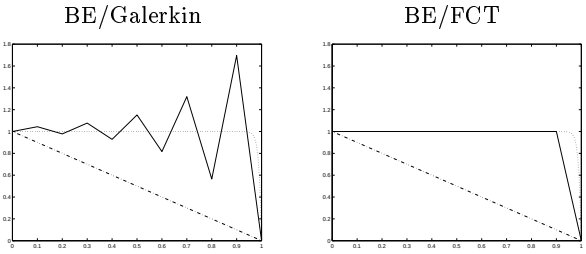

Fig. 3. Steady-state convection-diffusion in 1D, $\epsilon = 10^{-2},\ \Delta x = 0.1$

The first two-dimensional example is a standard benchmark problem [5],[6] which deals with advection of a solid body by the vortex $\mathbf{v} = (-y, x)$ in a square cavity $(-1, 1) \times (-1, 1)$. The counterclockwise rotation takes place about the origin. The initial condition is a cylinder with a slot defined by

$$u^0(x, y) = \begin{cases} 1, & R < 1/3 \text{ and } (|x| > 0.05 \text{ or } y > 0.5), \\ 0, & \text{otherwise}, \end{cases}$$

where $R = \sqrt{x^2 + (y - 1/3)^2}$. After one complete revolution the exact solution matches the initial data. The numerical results produced by the LW/FCT and BE/FCT schemes on a uniform mesh of 128×128 bilinear elements are shown in Fig. 4. They exhibit a sharp resolution of discontinuities and confirm that the methods are immune to spurious ripples.

Finally, let us illustrate the utility of the BE/FCT method by applying it to a steady-state convection-diffusion equation with $\epsilon = 10^{-3}$ in a unit square discretized by 32×32 bilinear elements. The velocity field is given by $\mathbf{v} = (\cos 10^o, \sin 10^o)$, and the boundary conditions read:

$$\frac{\partial u}{\partial y}(x, 1) = 0, \qquad u(x, 0) = u(1, y) = 0, \qquad u(0, y) = \begin{cases} 1, & y \geq 0.5, \\ 0, & y < 0.5. \end{cases}$$

A reasonable initial approximation is given by

$$u^0(x, y) = \begin{cases} 1 - x, & y \geq 0.5, \\ 0, & y < 0.5. \end{cases}$$

The numerical solutions obtained by the BE/Galerkin and BE/FCT schemes are displayed in Fig. 5. It is observed that the Galerkin method without flux correction gives rise to spurious oscillations in the boundary layer. This is obviously not the case for the flux-corrected solution, which is remarkably accurate and satisfies the discrete maximum principle.

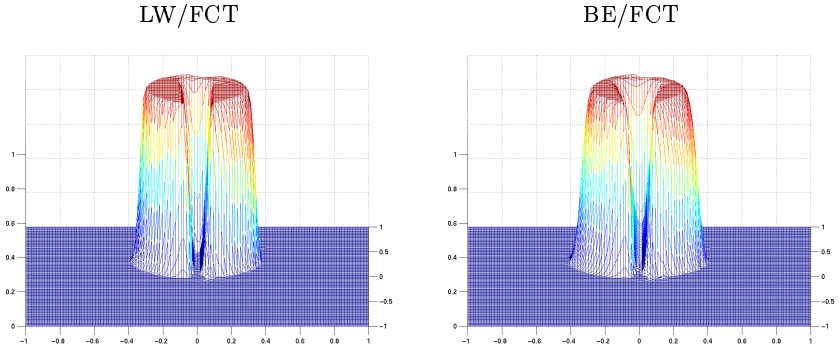

Fig. 4. Rotation of a cylinder with a slot, $t = 2\pi$, $\Delta t = 10^{-3}$

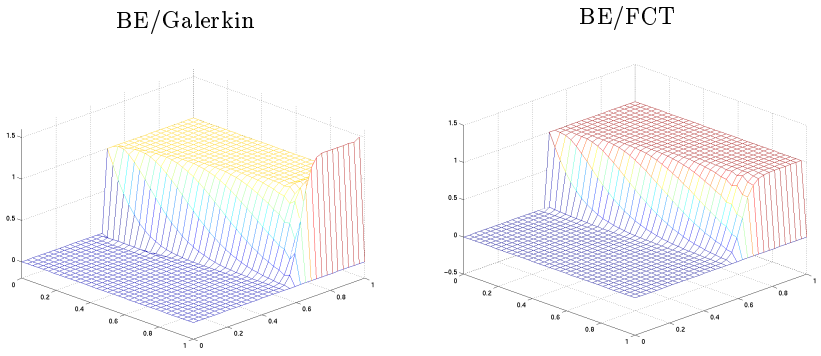

Fig. 5. Steady-state convection-diffusion in 2D

12 Conclusions

The flux-corrected-transport methodology was generalized to implicit finite element schemes using a rigorous positivity criterion. In contrast with the conventional element-based formulation, the antidiffusive terms were represented in terms of internodal fluxes. The discrete nature of the new approach makes it completely independent of the mesh and of the number of spatial dimensions. The fully implicit BE/FCT method is unconditionally positive, while the maximum admissible time step for other FEM-FCT schemes is readily computable. The presented numerical examples indicate that the best transient solutions are produced by explicit schemes, while steady-state problems call for a fully implicit treatment.

References

1. Boris J. P., Book D. L. (1973) Flux-Corrected Transport. I. SHASTA, A Fluid Transport Algorithm that Works. J. Comput. Phys. **11**, 38–69
2. DeVore C. R.(1998) An Improved Limiter for Multidimensional Flux-Corrected Transport. NASA Tech. Rep. AD-A360122
3. Kuzmin D. (2001) Positive Finite Element Schemes Based on the Flux-Corrected Transport Procedure. In: Bathe K. J. (Ed.) Computational Fluid and Solid Mechanics, First MIT Conference on Computational Fluid and Solid Mechanics, USA, June 12-15, Elsevier, 887-888
4. Kuzmin D., Turek S. (2002) Flux Correction Tools for Finite Elements. J. Comput. Phys. **175**, 1–34.
5. Löhner R., Morgan K., Peraire J., Vahdati M. (1987) Finite Element Flux-Corrected Transport (FEM-FCT) for the Euler and Navier-Stokes Equations. Int. J. Numer. Meth. Fluids **7**, 1093–1109
6. Zalesak S. T. (1979) Fully Multidimensional Flux-Corrected Transport Algorithms for Fluids. J. Comput. Phys. **31**, 335–362