

# Grid Orientation Effect in coupled Finite Volume Schemes

R. Eymard, C. Guichard and R. Masson.

September 14, 2011

## Abstract

The numerical simulation of two-phase flow in a porous medium may lead, when using structured grids, to the apparition of the so-called Grid Orientation Effect (GOE). We propose in this paper a procedure to eliminate this phenomenon, based on the use of new fluxes with a new stencil in the discrete version of the convection equation, without changing the discrete scheme for computing the pressure field. A mathematical study, based on a weak BV inequality using the new fluxes, shows the convergence of the modified scheme in a particular case. Finally, numerical results show the efficiency and the accuracy of the method.

## 1 Introduction

In the 1980's, numerous papers have been concerned with the so-called grid orientation effect, in the framework of oil reservoir simulation. This effect is due to the anisotropy of the numerical diffusion induced by the upstream weighting scheme, and the computation of a pressure field, solution to an elliptic equation in which the diffusion coefficient depends on the value of the convected unknown. This problem has been partly solved in the framework of industrial codes, in which the meshes are structured and regular (mainly based on squares and cubes). The literature on this problem is huge, and is impossible to exhaustively quote; let us only cite [4, 5, 8, 12, 13] and references therein. In the 2000's, a series of new schemes have been introduced in order to compute these coupled problems on general grids [1, 3, 6, 10]. But, in most of the cases, the non regular meshes conserve structured directions, although the shape of the control volumes is no longer that of a regular cube. This is the case for the Corner Point Geometries [11] widely used in industrial reservoir simulations. The control volumes which are commonly used in 3D reservoir simulations are generalised "hexahedra", in the sense that each of them is neighboured by 6 other control volumes. In this case, the stencil for the pressure resolution may have a 27-point stencil, using for instance a Multi-Point Flux Approximation (MPFA) scheme, see [1]. Nevertheless, selecting a 27-point stencil instead of a 7-point stencil for the pressure resolution has no influence on the Grid Orientation Effect, which results from the stencil used in upstream weighted mass exchanges coupled with the pressure resolution.

In order to overcome this problem, we study here a new method consisting in changing the stencil of the convection equation, without modifying the pressure equation. This method is presented on a simplified problem, modelling immiscible two-phase flow within a porous medium. Let  $\Omega \subset \mathbb{R}^d$  (with  $d = 2$  or  $3$ ) be the considered bounded open connected space domain, with a regular boundary denoted by  $\partial\Omega$ . We consider the following two-phase flow problem in  $\Omega$ ,

$$\begin{cases} u_t - \operatorname{div}(k_1(u)\Lambda\nabla p) & = \max(s, 0)f(c) + \min(s, 0)f(u) \\ (1 - u)_t - \operatorname{div}(k_2(u)\Lambda\nabla p) & = \max(s, 0)(1 - f(c)) + \min(s, 0)(1 - f(u)), \\ f(u) = \frac{k_1(u)}{k_1(u) + k_2(u)}, \end{cases} \quad (1)$$

where, for  $\mathbf{x} \in \Omega$  and  $t \geq 0$ ,  $u(\mathbf{x}, t) \in [0, 1]$  is the saturation of phase 1 (for example water), and therefore  $1 - u(\mathbf{x}, t)$  is the saturation of phase 2,  $k_1$  is the mobility of phase 1 (increasing function such that  $k_1(0) = 0$ ),  $k_2$  is the mobility of phase 2 (decreasing function such that  $k_2(1) = 0$ ),

the function  $s$  represents a volumic source term, corresponding to injection/pumping fluids into the domain,  $p$  is the common pressure of both phases (the capillary pressure is assumed to be negligible in front of the pressure gradients due to injection and production wells) and  $\Lambda(\mathbf{x})$  denotes the permeability tensor (that is defined by a symmetric positive definite matrix which may depend on the point  $\mathbf{x} \in \Omega$ ). The volumic composition of the injected fluid is tuned by the function  $c$ , assumed to vary between 0 and 1. We assume that there is no flow across the boundary, which corresponds to homogeneous Neumann boundary conditions.

We may see System (1) as the coupling of an elliptic problem with unknown  $p$  and a nonlinear scalar hyperbolic problem with unknown  $u$ ,

$$\begin{cases} \operatorname{div} \mathbf{v} = s & \text{with } \mathbf{v} = -(k_1(u) + k_2(u))\Lambda \nabla p, \\ u_t + \operatorname{div}(f(u)\mathbf{v}) = \max(s, 0)f(c) + \min(s, 0)f(u). \end{cases} \quad (2)$$

Let us now consider a coupled finite volume scheme for the approximation of Problem (1), written under the form (2):

$$\begin{aligned} \sum_{L,(K,L) \in S} F_{K,L}^{n+1} &= s_K^{n+1} \\ F_{K,L}^{n+1} + F_{L,K}^{n+1} &= 0 \\ |K| (u_K^{n+1} - u_K^n) + \tau^n \sum_{L,(K,L) \in S} &\left( f(u_K^m)(F_{K,L}^{n+1})^{(+)} - f(u_L^m)(F_{L,K}^{n+1})^{(+)} \right) = \\ &\tau^n ((s_K^{n+1})^{(+)} f(c_K^{n+1}) - (s_K^{n+1})^{(-)} f(u_K^m)). \end{aligned}$$

In the above system, we denote by  $K, L$  the control volumes, by  $|K|$  the measure of  $K$  (volume in 3D, area in 2D), by  $S$  the initial stencil of the scheme, defined as the set of pairs  $(K, L)$  having a common interface denoted  $\sigma_{K,L}$ , by  $n$  the time index, by  $\tau^n$  the time step ( $\tau^n = t^{n+1} - t^n$ ), by  $u_K^n$  the saturation in control volume  $K$  at time  $t^n$ , by  $s_K^{n+1}$  the quantity  $\frac{1}{\tau^n} \int_{t^n}^{t^{n+1}} \int_K s(\mathbf{x}, t) d\mathbf{x} dt$  and by  $c_K^{n+1}$  the quantity  $\frac{1}{\tau^n |K|} \int_{t^n}^{t^{n+1}} \int_K c(\mathbf{x}, t) d\mathbf{x} dt$ . The flux  $F_{K,L}^{n+1} = (F_{K,L}^{n+1})^{(+)} - (F_{L,K}^{n+1})^{(+)}$  is a generally implicit approximation of the flux  $\int_{\sigma_{K,L}} \mathbf{v} \cdot \mathbf{n}_{K,L} ds$  at the interface  $\sigma_{K,L}$  at time step  $n$  (where  $\mathbf{n}_{K,L}$  is the unit normal vector to  $\sigma_{K,L}$  oriented from  $K$  to  $L$ ), and, for all real  $a$ , the values  $a^{(+)}$  and  $a^{(-)}$  are non-negative and such that  $a^{(+)} - a^{(-)} = a$ . The value  $m$  is set to  $n$  in the case of the ‘‘IMPES’’ scheme (IMPlicit in Pressure and EXplicit in Saturation), and to  $n + 1$  for the implicit scheme.

We refer to [1, 3, 6, 10] for possible expressions of  $F_{K,L}^{n+1}$  allowing for the computation of an approximate pressure field; in this paper, we consider that these expressions are the data used in the definition of new fluxes dedicated to suppress Grid Orientation Effects. Hence Section 2 of this paper proposes a general method for defining a new stencil  $\hat{S}$  and new fluxes  $\hat{F}_{K,L}^{n+1}$  verifying at least the two following properties. We require that the flux continuity holds

$$\hat{F}_{K,L}^{n+1} + \hat{F}_{L,K}^{n+1} = 0, \quad \forall (K, L) \in \hat{S},$$

and that the balance in the control volumes is the same as that satisfied by the fluxes  $(F_{K,L}^{n+1})_{(K,L) \in S}$ :

$$\sum_{L,(K,L) \in \hat{S}} \hat{F}_{K,L}^{n+1} = \sum_{L,(K,L) \in S} F_{K,L}^{n+1}, \quad \forall K \in \mathcal{M}.$$

With these new fluxes and stencil, we write the following new scheme:

$$\begin{aligned} |K| (u_K^{n+1} - u_K^n) + \tau^n \sum_{L,(K,L) \in \hat{S}} &\left( f(u_K^m)(\hat{F}_{K,L}^{n+1})^{(+)} - f(u_L^m)(\hat{F}_{L,K}^{n+1})^{(+)} \right) = \\ &\tau^n ((s_K^{n+1})^{(+)} f(c_K^{n+1}) - (s_K^{n+1})^{(-)} f(u_K^m)). \end{aligned} \quad (3)$$

The present method is illustrated by the example of the design of a nine-point scheme, starting from a five-point scheme. This new scheme is mathematically analysed in Section 3 in the particular case where  $f(u) = u$  and  $k_1(u) + k_2(u)$  is constant, allowing for the fluxes to not depend on the time. Then numerical results show in Section 4 the efficiency of the method. A first test case, where Problem (1) has a radial analytical solution, allows for assessing the accuracy of the method, and a second test case which is a simple 3D case with three 2D layers shows the possibility to implement the scheme in industrial reservoir simulators. A short conclusion is finally proposed.

## 2 Mesh, stencils and fluxes

### 2.1 Construction of the new fluxes

This section is devoted to the method of construction of the new fluxes, using the old ones. Let us first precise the definition for the admissible discretizations which will be considered here.

**Definition 2.1** *We assume that  $\Omega \subset \mathbb{R}^d$ , with  $d \in \mathbb{N} \setminus \{0\}$  is a bounded open connected domain. We say that  $\mathcal{D} = (\mathcal{M}, \mathcal{F}, S)$  is an admissible discretization of  $\Omega$  if:*

1. *The set  $\mathcal{M}$  of the control volumes is such that all elements of  $\mathcal{M}$  are disjoint open connected subsets of  $\Omega$  with regular boundary, and such that  $\bar{\Omega} = \bigcup_{K \in \mathcal{M}} \bar{K}$ . The  $d$ -dimensional measure of  $K$  (resp.  $\Omega$ ) is denoted by  $|K|$  (resp.  $|\Omega|$ ) and the diameter of  $K$  is denoted  $h_K$ . We denote by  $h_{\mathcal{D}}$  the maximum value of  $(h_K)_{K \in \mathcal{M}}$ .*
2. *The interior faces of the mesh  $\sigma \in \mathcal{F}_{\text{int}}$  are obtained by  $\bar{K} \cap \bar{L} := \sigma_{K,L}$ , for all pairs of neighbouring control volumes  $K \in \mathcal{M}$  and  $L \in \mathcal{M}$ . They are assumed to be planar, with constant unit normal vector  $\mathbf{n}_{K,L}$  oriented from  $K$  to  $L$ . The exterior faces of the mesh  $\sigma \in \mathcal{F}_{\text{ext}}$  are obtained by  $\sigma = \bar{K} \cap \partial\Omega$ , for all control volumes  $K \in \mathcal{M}$ . The set of all the faces of the mesh  $\mathcal{F}$  is defined by  $\mathcal{F} = \mathcal{F}_{\text{int}} \cup \mathcal{F}_{\text{ext}}$ . The  $d-1$ -dimensional measure of  $\sigma \in \mathcal{F}$  is denoted by  $|\sigma|$ , assumed to be strictly positive. For all  $K \in \mathcal{M}$ , it is assumed that there exists a subset of  $\mathcal{F}$ , denoted by  $\mathcal{F}_K$ , such that  $\partial K = \bigcup_{\sigma \in \mathcal{F}_K} \sigma$ .*
3. *The stencil  $S$  is the set of all pairs  $(K, L)$  such that  $K \in \mathcal{M}$ ,  $L \in \mathcal{M} \setminus \{K\}$  and  $|\sigma_{K,L}| > 0$ .*

We then define

$$\theta_{\mathcal{D}} = \max_{K \in \mathcal{M}} \frac{h_K \sum_{\sigma \in \mathcal{F}_K} |\sigma|}{|K|}. \quad (4)$$

For an admissible discretization  $\mathcal{D} = (\mathcal{M}, \mathcal{F}, S)$ , we consider a real family  $(F_{K,L})_{(K,L) \in S}$ , which satisfies the following symmetry property:

$$F_{K,L} + F_{L,K} = 0, \quad \forall (K, L) \in S. \quad (5)$$

For any  $(K, L) \in S$ , we assume that is defined a non empty set  $\widehat{\mathcal{P}}_{K,L}$  (called the set of the paths from  $K$  to  $L$ ) such that

1. For all  $P \in \widehat{\mathcal{P}}_{K,L}$ , there exist  $m \in \mathbb{N} \setminus \{0\}$  and a set of  $m$  different control volumes  $\{K_1, \dots, K_m\} \subset \mathcal{M}$  with  $K_1 = K$  and  $K_m = L$  such that

$$P = \{(K_i, K_{i+1}), i = 1, \dots, m-1\}.$$

2. For any  $P = \{(K_i, K_{i+1}), i = 1, \dots, m-1\} \in \widehat{\mathcal{P}}_{K,L}$ , we denote by  $P^{\leftarrow}$  the inverse path from  $L$  to  $K$ , defined by  $P^{\leftarrow} = \{(K_{i+1}, K_i), i = 1, \dots, m-1\}$ . We assume that, for all  $(K, L) \in S$ ,  $\widehat{\mathcal{P}}_{L,K} = \{P^{\leftarrow}, P \in \widehat{\mathcal{P}}_{K,L}\}$ .

3. The new stencil  $\widehat{S} \subset \mathcal{M}^2$ , defined by

$$\widehat{S} = \bigcup_{(K,L) \in S, P \in \widehat{\mathcal{P}}_{K,L}} P \quad (6)$$

satisfies therefore that for all  $(K, L) \in \widehat{S}$ ,  $(L, K) \in \widehat{S}$ .

4. We denote by  $\theta_{\widehat{\mathcal{P}}}$  the maximum between:

- the ratio between the diameter of the reunion of all control volumes of a given path and the minimum diameter of the control volumes of the path,
- the number of elements in a path.

This may be written as

$$\theta_{\widehat{\mathcal{P}}} = \max\left\{\max\left(\frac{\text{diam}(\bigcup_{(I,J) \in P} (I \cup J))}{\min_{(I,J) \in P} (\min(h_I, h_J))}, \sharp P\right), (K, L) \in S, P \in \widehat{\mathcal{P}}_{K,L}\right\}. \quad (7)$$

For all  $(K, L) \in S$ , let  $(F_{K,L}^P)_{P \in \widehat{\mathcal{P}}_{K,L}}$  be a family such that

$$\forall (K, L) \in S, \quad \forall P \in \widehat{\mathcal{P}}_{K,L}, \quad F_{K,L}^P F_{K,L} \geq 0, \quad (8)$$

$$\forall (K, L) \in S, \quad \sum_{P \in \widehat{\mathcal{P}}_{K,L}} F_{K,L}^P = F_{K,L}, \quad (9)$$

and

$$\forall (K, L) \in S, \quad \forall P \in \widehat{\mathcal{P}}_{K,L}, \quad F_{L,K}^{P^-} = -F_{K,L}^P. \quad (10)$$

We define the families  $(\widetilde{F}_{K,L}, \widetilde{F}_{K,L}^{(+)})_{(K,L) \in \widehat{S}}$  by

$$\begin{aligned} \forall (I, J) \in \widehat{S}, \\ \widetilde{F}_{I,J}^{(+)} &= \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P \max(F_{K,L}^P, 0), \\ \widetilde{F}_{I,J} &= \widetilde{F}_{I,J}^{(+)} + \widetilde{F}_{J,I}^{(+)} = \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P |F_{K,L}^P|, \end{aligned} \quad (11)$$

where  $\xi_{I,J}^P$  is such that  $\xi_{I,J}^P = 1$  if  $(I, J) \in P$  and  $\xi_{I,J}^P = 0$  otherwise. We finally define, for a given  $\nu \in [0, 1]$ , the families  $(\widehat{F}_{I,J}, \widehat{F}_{I,J}^{(+)})_{(I,J) \in \widehat{S}}$  used in the new convection scheme (3) by

$$\forall (I, J) \in \widehat{S}, \quad \widehat{F}_{I,J}^{(+)} = G_\nu(\widetilde{F}_{I,J}^{(+)}, \widetilde{F}_{J,I}^{(+)}) \quad \text{and} \quad \widehat{F}_{I,J} = \widetilde{F}_{I,J}^{(+)} - \widetilde{F}_{J,I}^{(+)} = \widehat{F}_{I,J}^{(+)} - \widehat{F}_{J,I}^{(+)}, \quad (12)$$

where the function  $G_\nu$  is defined by

$$\forall \nu \in [0, 1], \forall (a, b) \in (\mathbb{R}^+)^2, \quad G_\nu(a, b) = \max(a - b, \frac{1}{2}(a - b + \nu(a + b)), 0). \quad (13)$$

The function  $G_\nu$  is designed in order to minimise  $G_\nu(a, b) + G_\nu(b, a)$  (hence introducing the smallest additional numerical diffusion) under the constraints  $G_\nu(a, b) \geq 0$  (for monotonicity purposes),  $G_\nu(a, b) - G_\nu(b, a) = b - a$  (hence ensuring the conservativity) and  $G_\nu(a, b) + G_\nu(b, a) \geq \nu(a + b)$  (this property is using for controlling the fluxes  $(\widetilde{F}_{K,L})_{(K,L) \in \widehat{S}}$  from the weak BV inequality). Indeed, it is straightforward to check that the continuous function  $G_\nu(a, b)$  ensures the following property: if  $|a - b| > \nu(a + b)$ , we have  $G_\nu(a, b) = \max(a - b, 0)$  and  $G_\nu(b, a) = \max(b - a, 0)$ . Otherwise, we have  $G_\nu(a, b) = \frac{1}{2}(a - b + \nu(a + b))$  and  $G_\nu(b, a) = \frac{1}{2}(b - a + \nu(a + b))$ . Therefore we get

$$\begin{aligned} (G_\nu(a, b), G_\nu(b, a)) &= \operatorname{argmin}\{c + d, (c, d) \in (\mathbb{R}^+)^2, c - d = a - b, c + d \geq \nu(a + b)\}, \\ \forall (a, b) &\in (\mathbb{R}^+)^2, \forall \nu \in [0, 1]. \end{aligned} \quad (14)$$

We can then deduce that

$$\forall (I, J) \in \widehat{S}, \widehat{F}_{I,J} = \widehat{F}_{I,J}^{(+)} - \widehat{F}_{J,I}^{(+)} = \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P F_{K,L}^P. \quad (15)$$

*Remark 1* If the fluxes  $F_{KL}$  are computed using a MPFA scheme (i.e. there exist coefficients  $(a_{K,L}^M)_{M \in \mathcal{M}}$  such that  $F_{K,L} = \sum_{M \in \mathcal{M}} a_{K,L}^M p_M$  and  $\sum_{M \in \mathcal{M}} a_{K,L}^M = 0$ ), and if  $F_{K,L}^P = \omega_{K,L}^P F_{K,L}$  with  $\omega_{K,L}^P \geq 0$  and  $\sum_{P \in \widehat{\mathcal{P}}_{K,L}} \omega_{K,L}^P = 1$ , we get, using (15),  $\widehat{F}_{I,J} = \sum_{M \in \mathcal{M}} \widehat{a}_{I,J}^M p_M$  with

$$\widehat{a}_{I,J}^M = \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P \omega_{K,L}^P a_{K,L}^M,$$

and

$$\sum_{M \in \mathcal{M}} \widehat{a}_{I,J}^M = \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P \omega_{K,L}^P \sum_{M \in \mathcal{M}} a_{K,L}^M = 0.$$

Besides, if we let  $\nu = 0$ , the relation

$$\widehat{F}_{I,J}^{(+)} = \max(\widehat{F}_{I,J}, 0)$$

holds, which leads to a standard upstream weighting scheme coupled with a MPFA scheme for the pressure, which may be implemented in standard codes with a simple modification of the stencils and transmissivities. Note that the value  $\nu = 0$  is excluded in the mathematical analysis provided in Section 3, but that the numerical tests given in Section 4 show that this value seems to be efficient in practice. On the contrary, for  $\nu > 0$ , which is assumed in the mathematical analysis, the expression of the new fluxes cannot be obtained from a simple MPFA expression.

*Remark 2* If we let  $\mathcal{P}_{K,L} = \{P_0\}$  with  $P_0 = \{(K, L)\}$  (which leads to  $\widehat{S} = S$ ), the new fluxes are identical to the initial ones, independently of  $\nu$  chosen in  $[0, 1]$ .

Let us provide an example of application of this method.

## 2.2 Example: construction of a 9-point stencil scheme

We apply the method described in Section 2.1 to 2D structured quadrilateral meshes, which implies that the initial stencil  $S$  is the five-point stencil. For a given pair of neighbouring control volumes  $(K, L)$ , we define  $\widehat{\mathcal{P}}_{K,L}$  by  $\widehat{\mathcal{P}}_{K,L} = \{P_i, i = 0, \dots, 4\}$  with  $P_0 = \{(K, L)\}$  and  $P_i = \{(K, M_i), (M_i, L)\}$  for  $i = 1, 2, 3, 4$  (see Figure 1). Then we define  $(F_{K,L}^P)_{P \in \widehat{\mathcal{P}}_{K,L}}$  as follows. For a given  $\omega > 0$  (the value for  $\omega$  is chosen at 0.1 in the numerical examples), we take

$$\begin{cases} F_{K,L}^{P_0} = (1 - 4\omega)F_{K,L} \text{ for } P_0 = \{(K, L)\}, \\ F_{K,L}^{P_i} = \omega F_{K,L} \text{ for } P_i = \{(K, M_i), (M_i, L)\}, \forall i = 1, 2, 3, 4. \end{cases}$$

Then the new stencil  $\widehat{S}$  is the classical nine-point stencil (see Figure 1), defined by

$$\widehat{S} = S \cup \{(K, L) \in \mathcal{M}^2, \overline{K} \text{ and } \overline{L} \text{ have a common point}\}.$$

This method is illustrated by Figure 1, in which the double solid arrows represent the initial connectivity of the five-point stencil  $S$  and the double dashed arrows represent the new connectivity of the nine-point stencil  $\widehat{S}$ .

Assuming that this procedure has been applied to the whole mesh, let us give two examples of computation of  $\widehat{F}_{K,L}^{(+)}$  resulting from (11):

$$\begin{cases} \widetilde{F}_{K,L}^{(+)} = (1 - 4\omega) \max(F_{K,L}, 0) \\ \quad + \omega (\max(F_{K,M_1}, 0) + \max(F_{M_2,L}, 0) + \max(F_{K,M_3}, 0) + \max(F_{M_4,L}, 0)) \\ \widetilde{F}_{K,M_2}^{(+)} = \omega (\max(F_{K,L}, 0) + \max(F_{L,M_2}, 0) + \max(F_{K,M_1}, 0) + \max(F_{M_1,M_2}, 0)). \end{cases}$$

The values  $\widehat{F}_{K,L}^{(+)}$  are then obtained using (12).

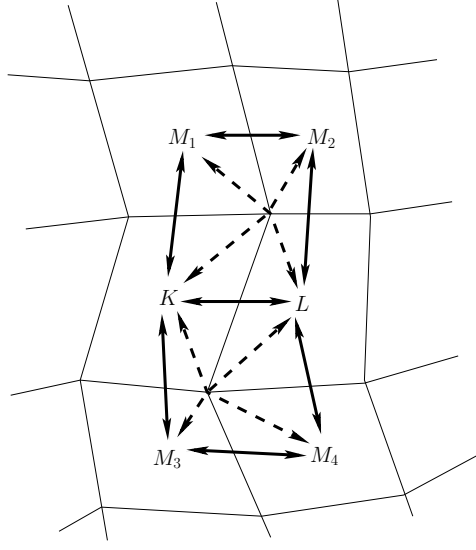


Figure 1: Five and nine point stencils on a structured quadrilateral mesh.

### 2.3 Properties of the new fluxes

We may now state the following result.

**Lemma 2.2 (New stencil and fluxes)** *Let  $\Omega \subset \mathbb{R}^d$ , with  $d \in \mathbb{N} \setminus \{0\}$  be a bounded open connected domain. Let  $\mathcal{D} = (\mathcal{M}, \mathcal{F}, S)$  be an admissible discretization of  $\Omega$  in the sense of Definition 2.1. Let  $(F_{K,L})_{(K,L) \in S}$  be such that (5) holds. Let  $(\widehat{\mathcal{P}}_{K,L})_{(K,L) \in S}$ ,  $\widehat{S}$  and  $(F_{K,L}^P)_{(K,L) \in S, P \in \widehat{\mathcal{P}}_{K,L}}$  such that (6)-(10) hold. Let  $(\widetilde{F}_{K,L}, \widetilde{F}_{K,L}^{(+)})_{(K,L) \in \widehat{S}}$  be defined by (11), let  $\nu \in [0, 1]$  be given and let  $(\widehat{F}_{I,J}, \widehat{F}_{I,J}^{(+)})_{(I,J) \in \widehat{S}}$  be defined by (12). Then the following properties hold:*

$$\forall (I, J) \in \widehat{S}, \nu \widetilde{F}_{I,J} \leq \widehat{F}_{I,J}^{(+)} + \widehat{F}_{J,I}^{(+)}, \quad (16)$$

$$\sum_{L, (K,L) \in \widehat{S}} \widehat{F}_{K,L} = \sum_{L, (K,L) \in S} F_{K,L}, \quad \forall K \in \mathcal{M}, \quad (17)$$

and

$$\sum_{(K,L) \in \widehat{S}} \max(h_K, h_L) |\widetilde{F}_{K,L}| \leq \theta_{\mathcal{P}}^2 \sum_{(K,L) \in S} \max(h_K, h_L) |F_{K,L}|. \quad (18)$$

PROOF. We get (16), using the properties (14) of the function  $G_\nu$  defined by (13). Let us turn to (17). For a given  $I \in \mathcal{M}$ , by reordering the sums, we can write that

$$\sum_{J, (I,J) \in \widehat{S}} \widehat{F}_{I,J} = \sum_{J, (I,J) \in \widehat{S}} \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P F_{K,L}^P = \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \chi_{I,P} F_{K,L}^P$$

where  $\chi_{I,P} = \sum_{J, (I,J) \in \widehat{S}} \xi_{I,J}^P$  is equal to 1 if there exists  $J \in \mathcal{M}$  such that  $(I, J) \in P$  (therefore

$I \neq L$ ), and to 0 otherwise. Note that, for  $(K, L) \in S$  with  $K \neq I$  and for  $P \in \widehat{\mathcal{P}}_{K,L}$  with  $\chi_{I,P} = 1$ , we have  $I \neq L$ ,  $(L, K) \in S$ ,  $P^- \in \widehat{\mathcal{P}}_{L,K}$  and  $\chi_{I,P^-} = 1$ . So, using (10), we obtain

$$\sum_{(K,L) \in S \text{ s.t. } K \neq I} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \chi_{I,P} F_{K,L}^P = 0.$$

Therefore we can write, using (9),

$$\sum_{J,(I,J) \in \widehat{S}} \widehat{F}_{I,J} = \sum_{L,(I,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{I,L}} \chi_{I,P} F_{I,L}^P = \sum_{L,(I,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{I,L}} F_{I,L}^P = \sum_{L,(I,L) \in S} F_{I,L},$$

which proves (17). Finally, let us prove (18). Thanks to (11), reordering the sums and using (7) and (8), we obtain

$$\begin{aligned} \sum_{(I,J) \in \widehat{S}} \max(h_I, h_J) \widehat{F}_{I,J} &= \sum_{(I,J) \in \widehat{S}} \max(h_I, h_J) \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P |F_{K,L}^P| \\ &\leq \theta_{\widehat{\mathcal{P}}} \sum_{(I,J) \in \widehat{S}} \sum_{(K,L) \in S} \max(h_K, h_L) \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P |F_{K,L}^P| \\ &= \theta_{\widehat{\mathcal{P}}} \sum_{(K,L) \in S} \max(h_K, h_L) \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \#P |F_{K,L}^P| \\ &\leq \theta_{\widehat{\mathcal{P}}}^2 \sum_{(K,L) \in S} \max(h_K, h_L) \sum_{P \in \widehat{\mathcal{P}}_{K,L}} |F_{K,L}^P| \\ &= \theta_{\widehat{\mathcal{P}}}^2 \sum_{(K,L) \in S} \max(h_K, h_L) |F_{K,L}|. \end{aligned}$$

□

### 3 Convergence analysis in a simplified case

For the sake of the mathematical analysis, we only consider Problem (1) in the case where  $f(u) = u$  and where the function  $k_1(u) + k_2(u)$  is constant. Indeed, the analysis of Problem (1) in the case  $k_1(u) + k_2(u)$  not constant is an open problem, and the case of a general function  $f$  may be studied using the methods of [7]. Hence the mathematical study is focused on the convergence of the new approximate scheme for the following problem on  $\Omega \times (0, T)$ :

$$\operatorname{div} \mathbf{v} = s, \quad (19)$$

$$u_t + \operatorname{div}(u\mathbf{v}) = \max(s, 0)c + \min(s, 0)u \text{ in } \Omega \times (0, T), \quad (20)$$

together with the initial condition

$$u = u_{\text{ini}} \text{ in } \Omega, \quad (21)$$

under the following hypotheses, denoted (H) in this section:

**Definition 3.1** (Hypotheses (H))

1.  $\Omega$  is a bounded open connected subset of  $\mathbb{R}^d$ ,  $T > 0$  is the period of observation.
2. We assume that  $\mathbf{v} \in C^1(\overline{\Omega})$  is such that  $\mathbf{v} \cdot \mathbf{n}_{\partial\Omega} = 0$  on  $\partial\Omega$ . We denote by  $s = \operatorname{div} \mathbf{v}$ .
3. We assume that  $c \in L^\infty(\Omega \times (0, +\infty))$  and  $u_{\text{ini}} \in L^\infty(\Omega)$ , where the functions  $c$  and  $u_{\text{ini}}$  are essentially bounded by 0 and 1.

Then Problem (20)-(21) is considered in the following weak sense:

$$\int_0^{+\infty} \int_{\Omega} (u\varphi_t + u\mathbf{v} \cdot \nabla\varphi + (\max(s, 0)c + \min(s, 0)u)\varphi) d\mathbf{x} dt + \int_{\Omega} u_{\text{ini}}(\mathbf{x})\varphi(\mathbf{x}, 0) d\mathbf{x} = 0, \quad (22)$$

$\forall \varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}), \varphi = 0 \text{ in } \mathbb{R}^d \times [T, +\infty).$

### 3.1 Approximation by an upstream weighting scheme

We first extend the definition 2.1 to space-time discretizations.

**Definition 3.2** Let  $\Omega \subset \mathbb{R}^d$ , with  $d \in \mathbb{N} \setminus \{0\}$  be a bounded open connected domain and let  $T > 0$  be given. We say that  $\mathcal{D} = (\mathcal{M}, \mathcal{F}, S, N, (t^n)_{n=0, \dots, N})$  is an admissible time-space discretization of  $\Omega \times (0, T)$  if

1.  $(\mathcal{M}, \mathcal{F}, S)$  is an admissible discretization of  $\Omega$  in the sense of Definition 2.1,
2.  $N \in \mathbb{N} \setminus \{0\}$  and  $(t^n)_{n=0, \dots, N}$  is a real family such that  $t_0 = 0 < t^1 \dots < t^N = T$ .

We then denote  $\tau^n = t^{n+1} - t^n$  for  $n = 0, \dots, N-1$ . We continue to use  $\mathcal{D}$  as index for all quantities depending only on the space discretization.

Assuming Hypotheses (H), let  $\mathcal{D} = (\mathcal{M}, \mathcal{F}, S, N, (t^n)_{n=0, \dots, N})$  be an admissible time-space discretization of  $\Omega \times (0, T)$  in the sense of Definition 3.2. Let  $(F_{K,L})_{(K,L) \in S}$  be such that (5) hold. We then denote, for  $\sigma \in \mathcal{F}_K$  such that  $\sigma = \sigma_{K,L}$ ,  $F_{K,\sigma} = F_{K,L}$ , and for  $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{ext}}$ ,  $F_{K,\sigma} = 0$ . We assume that  $(F_{K,L})_{(K,L) \in S}$  satisfies the following discrete conservation property

$$\sum_{\sigma \in \mathcal{F}_K} F_{K,\sigma} = \sum_{L, (K,L) \in S} F_{K,L} = s_K, \quad \forall K \in \mathcal{M}, \quad (23)$$

where we denote

$$s_K = \int_K s(\mathbf{x}) d\mathbf{x}, \quad \forall K \in \mathcal{M}, \quad (24)$$

Let  $\widehat{S}$ ,  $(\widehat{\mathcal{P}}_{K,L})_{(K,L) \in S}$  and  $(F_{K,L}^P)_{(K,L) \in S, P \in \widehat{\mathcal{P}}_{K,L}}$  such that (6)-(10) hold. Let  $(\widetilde{F}_{K,L}, \widetilde{F}_{K,L}^{(+)})_{(K,L) \in \widehat{S}}$  be defined by (11), let  $\nu \in (0, 1]$  be given (the value  $\nu = 0$  is excluded, since some bounds in Lemma 3.4 and Theorem 3.5 are obtained with respect to  $1/\nu$ , see also Remark 1) and let  $(\widehat{F}_{I,J}, \widehat{F}_{I,J}^{(+)})_{(I,J) \in \widehat{S}}$  be defined by (12).

The implicit version of the upstream weighting scheme devoted for approximating (22) on  $[0, T]$  writes

$$\begin{aligned} & |K| \frac{u_K^{n+1} - u_K^n}{\tau^n} \\ & + \sum_{L \in \mathcal{M}} \left( \widehat{F}_{K,L}^{(+)} u_K^{n+1} - \widehat{F}_{L,K}^{(+)} u_L^{n+1} \right) + s_K^{(-)} u_K^{n+1} - s_K^{(+)} c_K^{n+1} = 0, \end{aligned} \quad (25)$$

$\forall n = 0, \dots, N-1, \quad \forall K \in \mathcal{M},$

letting  $\widehat{F}_{I,J}^{(+)} = 0$  for pairs of control volumes  $(I, J) \notin \widehat{S}$ , and where

$$c_K^{n+1} = \frac{1}{|K| \tau^n} \int_{t^n}^{t^{n+1}} \int_K c(\mathbf{x}, t) d\mathbf{x} dt, \quad \forall n = 0, \dots, N-1, \quad \forall K \in \mathcal{M}, \quad (26)$$

$$s_K^{(+)} = \int_K \max(s(\mathbf{x}), 0) d\mathbf{x}, \quad s_K^{(-)} = \int_K \max(-s(\mathbf{x}), 0) d\mathbf{x}, \quad \forall K \in \mathcal{M}, \quad (27)$$

and

$$u_K^0 = \frac{1}{|K|} \int_K u_{\text{ini}}(\mathbf{x}) d\mathbf{x}, \quad \forall K \in \mathcal{M}. \quad (28)$$

### 3.2 Estimates

In this section, one proves the existence and uniqueness of the discrete solution, as well as an  $L^\infty$  estimate and a weak-BV inequality.

**Lemma 3.3** ( $L^\infty$  estimate and existence, uniqueness of the discrete solution)

Under Hypotheses (H), let  $\mathcal{D} = (\mathcal{M}, \mathcal{F}, S, N, (t^n)_{n=0, \dots, N})$  be an admissible time-space discretization of  $\Omega \times (0, T)$  in the sense of Definition 3.2. Let  $(F_{K,L})_{(K,L) \in \mathcal{S}}$  be such that (5), (23) and (24) hold. Let  $(\widehat{\mathcal{P}}_{K,L})_{(K,L) \in \mathcal{S}}$ ,  $\widehat{S}$  and  $(F_{K,L}^P)_{(K,L) \in \mathcal{S}, P \in \widehat{\mathcal{P}}_{K,L}}$  such that (6)-(10) hold. Let  $(\widetilde{F}_{K,L}, \widetilde{F}_{K,L}^{(+)})_{(K,L) \in \widehat{\mathcal{S}}}$  be defined by (11), let  $\nu \in (0, 1]$  be given and let  $(\widehat{F}_{I,J}, \widehat{F}_{I,J}^{(+)})_{(I,J) \in \widehat{\mathcal{S}}}$  be defined by (12). Let  $(u_K^n)_{K \in \mathcal{M}, n=0, \dots, N}$  be such that (25)-(28) hold. Then

$$0 \leq u_K^n \leq 1, \quad \forall n = 0, \dots, N, \quad \forall K \in \mathcal{M}. \quad (29)$$

Moreover, there exists one and only one  $(u_K^n)_{K \in \mathcal{M}, n=0, \dots, N}$  such that (25)-(28) hold.

PROOF. We prove the lemma by induction. Using Definition (28) for  $u_K^0$ , we have  $0 \leq u_K^0 \leq 1$ , for all  $K \in \mathcal{M}$ . Let us assume that, for a given  $n = 0, \dots, N-1$ ,  $(u_K^n)_{K \in \mathcal{M}}$  is given with  $0 \leq u_K^n \leq 1$ , for all  $K \in \mathcal{M}$ . Let us prove that, for a given  $(u_K^{n+1})_{K \in \mathcal{M}}$  such that (25) holds, then  $0 \leq u_K^{n+1} \leq 1$ , for all  $K \in \mathcal{M}$ .

Let us multiply (25) by  $\tau^n$ . We get,

$$|K| (u_K^{n+1} - u_K^n) + \tau^n \sum_{L \in \mathcal{M}} \widehat{F}_{K,L}^{(+)} u_K^{n+1} - \tau^n \sum_{L \in \mathcal{M}} \widehat{F}_{L,K}^{(+)} u_L^{n+1} + \tau^n s_K^{(-)} u_K^{n+1} - \tau^n s_K^{(+)} c_K^{n+1} = 0, \quad \forall K \in \mathcal{M}, \quad (30)$$

Using (17) and (23), we have,

$$\sum_{L \in \mathcal{M}} (\widehat{F}_{K,L}^{(+)} - \widehat{F}_{L,K}^{(+)}) + s_K^{(-)} - s_K^{(+)} = 0, \quad \forall K \in \mathcal{M} \quad (31)$$

Multiplying (31) by  $\tau^n u_K^{n+1}$  and subtracting from (30), we obtain

$$|K| (u_K^{n+1} - u_K^n) + \tau^n \sum_{L \in \mathcal{M}} \widehat{F}_{L,K}^{(+)} (u_K^{n+1} - u_L^{n+1}) + \tau^n s_K^{(+)} (u_K^{n+1} - c_K^{n+1}) = 0, \quad \forall K \in \mathcal{M}. \quad (32)$$

Let  $K_1$  denote some cell where the maximum of  $(u_K^{n+1})_{K \in \mathcal{M}}$  is reached ( $K_1$  is not necessarily unique). We suppose that  $u_{K_1}^{n+1} > 1$ , thus,

- $|K_1| (u_{K_1}^{n+1} - u_{K_1}^n) > 0$ ,
- $\widehat{F}_{L,K_1}^{(+)} (u_{K_1}^{n+1} - u_L^{n+1}) \geq 0$  by using (H),
- $s_{K_1}^{(+)} (u_{K_1}^{n+1} - c_{K_1}^{n+1}) \geq 0$  by using (26) and (H),

This is in contradiction with (32) for  $K = K_1$ , which proves that, for all  $K \in \mathcal{M}$ ,  $u_K^{n+1} \leq 1$ . By using a similar argument on the minimum of  $u_K^{n+1}$ , we prove that  $u_K^{n+1} \geq 0$ , for all  $K \in \mathcal{M}$ , hence concluding that  $0 \leq u_K^{n+1} \leq 1$ , for all  $K \in \mathcal{M}$ .

We now remark that, for  $U = (u_K^{n+1})_{K \in \mathcal{M}}$  satisfying (25), then  $U$  is solution to a linear system under the form  $AU = B$ , where  $A$  is a square matrix. Let  $U$  be such that  $AU = 0$ . Since the arguments used above remain true if  $u_K^n = 0$  and  $c_K^{n+1} = 0$ , for all  $K \in \mathcal{M}$ , we conclude that all components of  $U$  are bounded which implies  $U = 0$ . Hence the matrix  $A$  is invertible, which proves that there exists one and only one  $(u_K^{n+1})_{K \in \mathcal{M}}$  satisfying (25). This concludes the proof.  $\square$

**Lemma 3.4** (Weak BV-inequality)

Under Hypotheses (H), let  $\mathcal{D} = (\mathcal{M}, \mathcal{F}, S, N, (t^n)_{n=0, \dots, N})$  be an admissible time-space discretization of  $\Omega \times (0, T)$  in the sense of Definition 3.2. Let  $(F_{K,L})_{(K,L) \in \mathcal{S}}$  be such that (5), (23) and (24) hold. Let  $(\widehat{\mathcal{P}}_{K,L})_{(K,L) \in \mathcal{S}}$ ,  $\widehat{S}$  and  $(F_{K,L}^P)_{(K,L) \in \mathcal{S}, P \in \widehat{\mathcal{P}}_{K,L}}$  such that (6)-(10) hold. Let  $(\widetilde{F}_{K,L}, \widetilde{F}_{K,L}^{(+)})_{(K,L) \in \widehat{\mathcal{S}}}$  be defined by (11), let  $\nu \in (0, 1]$  be given and let  $(\widehat{F}_{I,J}, \widehat{F}_{I,J}^{(+)})_{(I,J) \in \widehat{\mathcal{S}}}$  be defined by (12). Let  $(u_K^n)_{K \in \mathcal{M}, n=0, \dots, N}$  be such that (25)-(28) hold.

Then there exists  $C_{\text{BV}} > 0$ , only depending on  $\Omega$ ,  $s$  and  $T$  such that :

$$\sum_{n=0}^{N-1} \tau^n \sum_{(K,L) \in \widehat{\mathcal{S}}} \widetilde{F}_{K,L} (u_K^{n+1} - u_L^{n+1})^2 \leq \frac{C_{\text{BV}}}{\nu}. \quad (33)$$

PROOF.

Let us multiply (32) by  $u_K^{n+1}$ , sum on  $n = 0, \dots, N-1$  and  $K \in \mathcal{M}$ . We get  $T_1 + T_2 = 0$  with

$$T_1 = \sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}} |K| (u_K^{n+1} - u_K^n) u_K^{n+1}$$

and

$$T_2 = \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} \left( \sum_{L \in \mathcal{M}} \widehat{F}_{L,K}^{(+)} (u_K^{n+1} - u_L^{n+1}) u_K^{n+1} + s_K^{(+)} (u_K^{n+1} - c_K^{n+1}) u_K^{n+1} \right).$$

Using the relation

$$(u_K^{n+1} - u_K^n) u_K^{n+1} = \frac{1}{2} (u_K^{n+1})^2 + \frac{1}{2} (u_K^{n+1} - u_K^n)^2 - \frac{1}{2} (u_K^n)^2, \quad (34)$$

we can rewrite  $T_1$  as  $T_1 = T_3 + T_4$  with

$$T_3 = \frac{1}{2} \sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}} |K| (u_K^{n+1} - u_K^n)^2$$

and

$$T_4 = \frac{1}{2} \sum_{K \in \mathcal{M}} |K| ((u_K^N)^2 - (u_K^0)^2)$$

Similarly, we can rewrite  $T_2$  as  $T_2 = T_5 + T_6$  with

$$T_5 = \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} \left( \sum_{L \in \mathcal{M}} \widehat{F}_{L,K}^{(+)} (u_K^{n+1} - u_L^{n+1})^2 + s_K^{(+)} (u_K^{n+1} - c_K^{n+1})^2 \right)$$

and

$$T_6 = \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} \left( \sum_{L \in \mathcal{M}} \widehat{F}_{L,K}^{(+)} ((u_K^{n+1})^2 - (u_L^{n+1})^2) + s_K^{(+)} ((u_K^{n+1})^2 - (c_K^{n+1})^2) \right).$$

Gathering by faces and using (16), we get

$$\begin{aligned} \sum_{K \in \mathcal{M}} \sum_{L \in \mathcal{M}} \widehat{F}_{L,K}^{(+)} (u_K^{n+1} - u_L^{n+1})^2 &= \frac{1}{2} \sum_{(K,L) \in \widehat{\mathcal{S}}} (\widehat{F}_{K,L}^{(+)} + \widehat{F}_{L,K}^{(+)}) (u_K^{n+1} - u_L^{n+1})^2 \\ &\geq \frac{\nu}{2} \sum_{(K,L) \in \widehat{\mathcal{S}}} \widetilde{F}_{K,L} (u_K^{n+1} - u_L^{n+1})^2. \end{aligned}$$

Taking this relation into account in the expression of  $T_5$ , we obtain

$$T_5 \geq \frac{\nu}{4} \sum_{n=0}^{N-1} \tau^n \left( \sum_{\{K,L\} \subset \mathcal{M}} \widetilde{F}_{K,L} (u_K^{n+1} - u_L^{n+1})^2 + \sum_{K \in \mathcal{M}} s_K^{(+)} (u_K^{n+1} - c_K^{n+1})^2 \right).$$

Again gathering by faces, we now write

$$\begin{aligned}
& \sum_{K \in \mathcal{M}} \sum_{L \in \mathcal{M}} \widehat{F}_{L,K}^{(+)} ((u_K^{n+1})^2 - (u_L^{n+1})^2) \\
&= \sum_{\{K,L\} \subset \mathcal{M}} (\widehat{F}_{L,K}^{(+)} ((u_K^{n+1})^2 - (u_L^{n+1})^2) + \widehat{F}_{K,L}^{(+)} ((u_L^{n+1})^2 - (u_K^{n+1})^2)) \\
&= - \sum_{\{K,L\} \subset \mathcal{M}} \widehat{F}_{K,L} ((u_K^{n+1})^2 - (u_L^{n+1})^2) \\
&= - \sum_{K \in \mathcal{M}} \sum_{L \in \mathcal{M}} \widehat{F}_{K,L} (u_K^{n+1})^2 = - \sum_{K \in \mathcal{M}} s_K (u_K^{n+1})^2,
\end{aligned}$$

which leads to

$$T_6 = \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} \left( s_K^{(-)} (u_K^{n+1})^2 - s_K^{(+)} (c_K^{n+1})^2 \right).$$

Gathering the above relations, we get  $T_3 + T_4 + T_5 + T_6 = 0$  with

$$T_3 \geq 0,$$

$$T_4 \geq -\frac{1}{2} \sum_{K \in \mathcal{M}} |K| (u_K^0)^2 \geq -\frac{1}{2} |\Omega|,$$

$$T_5 \geq \frac{\nu}{4} \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} \widetilde{F}_{K,L} (u_K^{n+1} - u_L^{n+1})^2,$$

and

$$T_6 \geq -\frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} s_K^{(+)} (c_K^{n+1})^2 \geq -\frac{1}{2} T \sum_{K \in \mathcal{M}} s_K^{(+)}.$$

We then obtain

$$\sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} \widetilde{F}_{K,L} (u_K^{n+1} - u_L^{n+1})^2 \leq \frac{2}{\nu} \left( |\Omega| + T \int_{\Omega} \max(s(\mathbf{x}), 0) d\mathbf{x} \right),$$

which concludes the proof.  $\square$

### 3.3 Convergence study

It is now possible to give a convergence proof for the scheme in the linear case. This proof could be extended to the nonlinear scalar hyperbolic case by following the methods proposed in [7], based on the convergence to the unique entropy process solution. Let us also note that, referring to Remark 2, the present mathematical analysis applies (with  $\nu > 0$ ) to an upstream weighting scheme written with the initial fluxes.

**Theorem 3.5** *Under Hypotheses (H), let  $\mathcal{D} = (\mathcal{M}, \mathcal{F}, S, N, (t^n)_{n=0, \dots, N})$  be an admissible time-space discretization of  $\Omega \times (0, T)$  in the sense of Definition 3.2. Let  $(F_{K,L})_{(K,L) \in S}$  be such that (5), (23) and (24) hold. Let  $(\widehat{P}_{K,L})_{(K,L) \in S}$ ,  $\widehat{S}$  and  $(F_{K,L}^P)_{(K,L) \in S, P \in \widehat{P}_{K,L}}$  such that (6)-(10) hold. Let  $(\widetilde{F}_{K,L}, \widetilde{F}_{K,L}^{(+)})_{(K,L) \in \widehat{S}}$  be defined by (11), let  $\nu \in (0, 1]$  be given and let  $(\widehat{F}_{I,J}, \widehat{F}_{I,J}^{(+)})_{(I,J) \in \widehat{S}}$  be defined by (12). Let  $(u_K^n)_{K \in \mathcal{M}, n=0, \dots, N}$  be such that (25)-(28) hold and let  $u_{\mathcal{D}}$  be the function defined by*

$$u_{\mathcal{D}}(x, t) = u_K^{n+1}, \text{ for a.e. } (x, t) \in K \times (t^n, t^{n+1}), \forall n = 0, \dots, N-1, \forall K \in \mathcal{M}. \quad (35)$$

We assume that

$$\lim_{h_{\mathcal{D}} \rightarrow 0} \sum_{(K,L) \in S} \frac{\max(h_K, h_L)}{|\sigma_{K,L}|} \left( F_{K,L} - \int_{\sigma_{K,L}} \mathbf{v} \cdot \mathbf{n}_{K,L} ds \right)^2 = 0. \quad (36)$$

Then, as  $h_{\mathcal{D}} \rightarrow 0$  and  $\max \tau^n \rightarrow 0$  while  $\nu$  remains fixed,  $\theta_{\mathcal{D}}$  and  $\theta_{\widehat{\mathcal{P}}}$  remain bounded,  $u_{\mathcal{D}}$  converges for the weak- $\star$  topology of  $L^\infty(\Omega \times (0, T))$  to the unique function  $u \in L^\infty(\Omega \times (0, T))$  satisfying (22).

*Remark 3* Condition (36) is naturally satisfied if  $F_{K,L} = \int_{\sigma_{K,L}} \mathbf{v} \cdot \mathbf{n}_{K,L} ds$ . More interestingly, it also holds if  $F_{K,L}$  is obtained using a finite volume scheme for the approximation of Problem  $\operatorname{div} \mathbf{v} = s$  with  $\mathbf{v} = -\Lambda \nabla p$  and Neumann boundary conditions (see [7], pp. 996–1012).

**PROOF.** In order to prove Theorem 3.5, we consider a sequence  $(\mathcal{D}_m)_{m \in \mathbb{N}}$  of admissible time-space discretizations, such that  $h_{\mathcal{D}_m}$  (denoted by  $h_m$  in the following) and  $\max_n(\tau_m^n)$  tend to zero as  $m \rightarrow \infty$ . We assume that, for each  $m$ , the families implicitly indexed by  $m$ :  $(F_{K,L})_{(K,L) \in \mathcal{S}}$ ,  $\widehat{S}$ ,  $(\widehat{\mathcal{P}}_{K,L})_{(K,L) \in \mathcal{S}}$  and  $(F_{K,L}^P)_{(K,L) \in \mathcal{S}, P \in \widehat{\mathcal{P}}_{K,L}}$  satisfy the hypotheses of the theorem with the same value  $\nu \in (0, 1)$ , while  $\theta_{\mathcal{D}_m}$  and  $\theta_{\widehat{\mathcal{P}}}$  remain bounded as  $m$  tends to  $\infty$ . We denote  $u_m = u_{\mathcal{D}_m}$  for all  $m \in \mathbb{N}$ .

Let us prove the convergence of the sequence  $(u_m)_{m \in \mathbb{N}}$  to the weak solution  $u$  of Problem (22) for the weak- $\star$  topology of  $L^\infty(\Omega \times (0, T))$ , for all  $T > 0$ . The classical argument of the uniqueness of this limit suffices for concluding the proof of the theorem.

We first notice that, thanks to Lemma 3.3, we get the existence of a subsequence, again noted  $(u_m)_{m \in \mathbb{N}}$ , which converges to some function  $u \in L^\infty(\Omega \times (0, T))$  for the weak- $\star$  topology of  $L^\infty(\Omega \times (0, T))$  as  $m \rightarrow +\infty$ . The aim of this proof is to show that  $u$  satisfies (22).

Let  $\varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R})$  be such that  $\varphi = 0$  in  $\mathbb{R}^d \times [T, +\infty)$ . In this proof, we denote by  $C_\varphi$  an  $L^\infty$  bound of first and second derivatives of  $\varphi$ . Let  $m \in \mathbb{N}$ . In the following, we drop some indices  $m$ , using the notations  $\mathcal{D} = \mathcal{D}_m$ . We define  $\varphi_K^n$  by

$$\varphi_K^{n+1} = \frac{1}{|K|} \int_K \varphi(\mathbf{x}, t^n) d\mathbf{x} dt, \quad \forall K \in \mathcal{M}, \quad \forall n = 0, \dots, N.$$

Let us multiply (32) by  $\varphi_K^{n+1}$ , sum over  $K \in \mathcal{M}$  and  $n = 0, \dots, N-1$ . We obtain  $T_7^{(m)} + T_8^{(m)} = 0$ , with

$$T_7^{(m)} = \sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}} |K| (u_K^{n+1} - u_K^n) \varphi_K^{n+1}$$

and

$$T_8^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} \left( \sum_{L \in \mathcal{M}} \widehat{F}_{L,K}^{(+)} (u_K^{n+1} - u_L^{n+1}) \varphi_K^{n+1} + s_K^{(+)} (u_K^{n+1} - c_K^{n+1}) \varphi_K^{n+1} \right).$$

Let us study  $T_7^{(m)}$ . Thanks to  $\varphi(\mathbf{x}, t^N) = \varphi(\mathbf{x}, T) = 0$ , we have

$$\begin{aligned} T_7^{(m)} &= - \sum_{n=1}^N \sum_{K \in \mathcal{M}} |K| u_K^n (\varphi_K^{n+1} - \varphi_K^n) - \sum_{K \in \mathcal{M}} |K| u_K^0 \varphi_K^1 \\ &= - \sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}} u_K^{n+1} \int_{t^n}^{t^{n+1}} \int_K \varphi_t(\mathbf{x}, t) d\mathbf{x} dt - \sum_{K \in \mathcal{M}} u_K^0 \int_K \varphi(\mathbf{x}, 0) d\mathbf{x} \\ &= - \int_0^T \int_\Omega u_m(\mathbf{x}, t) \varphi_t(\mathbf{x}, t) d\mathbf{x} dt - \sum_{K \in \mathcal{M}} u_K^0 \int_K \varphi(\mathbf{x}, 0) d\mathbf{x}. \end{aligned}$$

Using the weak- $\star$  convergence of  $(u_m)_{m \in \mathbb{N}}$  to  $u$ , we deduce that

$$\lim_{m \rightarrow +\infty} T_7^{(m)} = - \int_0^{+\infty} \int_\Omega u(\mathbf{x}, t) \varphi_t(\mathbf{x}, t) d\mathbf{x} dt - \int_\Omega u_{\text{ini}}(\mathbf{x}) \varphi(\mathbf{x}, 0) d\mathbf{x} = 0.$$

Let us now prove the convergence of the sequence  $(T_8^{(m)})_{m \in \mathbb{N}}$  to  $T_9$  defined by

$$T_9 = - \int_0^{+\infty} \int_\Omega (u \mathbf{v} \cdot \nabla \varphi + (\max(s, 0)c + \min(s, 0)u) \varphi) d\mathbf{x} dt.$$

To this purpose, let us define  $T_{10}^{(m)}$  by

$$T_{10}^{(m)} = - \int_0^{+\infty} \int_{\Omega} (u_m \mathbf{v} \cdot \nabla \varphi + (\max(s, 0)c_m + \min(s, 0)u_m)\varphi) \, d\mathbf{x}dt,$$

with

$$c_m(x, t) = c_K^{n+1} \text{ for a.e. } (x, t) \in K \times (t^n, t^{n+1}), \forall n = 0, \dots, N-1, \forall K \in \mathcal{M},$$

where  $c_K^{n+1}$  is defined by (26). Using the weak- $\star$  convergence of  $(u_m)_{m \in \mathbb{N}}$  to  $u$ , we deduce that

$$\lim_{m \rightarrow +\infty} T_{10}^{(m)} = T_9.$$

We then define  $T_{11}^{(m)}$  by

$$T_{11}^{(m)} = - \sum_{n=0}^{N-1} \tau^n \int_{\Omega} (u_m^{n+1} \mathbf{v} \cdot \nabla \varphi(\mathbf{x}, t^n) + (\max(s, 0)c_m^{n+1} + \min(s, 0)u_m^{n+1})\varphi(\mathbf{x}, t^n)) \, d\mathbf{x}.$$

Since  $|\nabla \varphi(\mathbf{x}, t^n) - \nabla \varphi(\mathbf{x}, t)| \leq C_{\varphi} \max(\tau^n)$  and  $|\varphi(\mathbf{x}, t^n) - \varphi(\mathbf{x}, t)| \leq C_{\varphi} \max(\tau^n)$ , we get that

$$\lim_{m \rightarrow \infty} |T_{10}^{(m)} - T_{11}^{(m)}| = 0. \quad (37)$$

Let us prove that  $\lim_{m \rightarrow \infty} |T_{11}^{(m)} - T_8^{(m)}| = 0$ . Using  $\mathbf{v} \cdot \nabla \varphi = \operatorname{div}(\varphi \mathbf{v}) - (\max(s, 0) + \min(s, 0))\varphi$ , we may write  $T_{11}^{(m)}$  under the form

$$T_{11}^{(m)} = - \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} \left( u_K^{n+1} \sum_{\sigma \in \mathcal{F}_K} \int_{\sigma} \varphi(\mathbf{x}, t^n) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}_{K, \sigma} ds(\mathbf{x}) + (c_K^{n+1} - u_K^{n+1}) \int_K \max(s(\mathbf{x}), 0) \varphi(\mathbf{x}, t^n) d\mathbf{x} \right).$$

Let us now define  $T_{12}^{(m)}$  by

$$T_{12}^{(m)} = - \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} \left( u_K^{n+1} \sum_{\sigma \in \mathcal{F}_K} \varphi_{\sigma}^{n+1} \int_{\sigma} \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}_{K, \sigma} ds(\mathbf{x}) dt + s_K^{(+)} (c_K^{n+1} - u_K^{n+1}) \varphi_K^{n+1} \right),$$

with

$$\varphi_{\sigma}^{n+1} = \frac{1}{|\sigma|} \int_{\sigma} \varphi(\mathbf{x}, t^n) ds(\mathbf{x}).$$

We have, for all  $n \in [0, N]$  and all  $K \in \mathcal{M}$ ,

$$\begin{aligned} & \int_{\sigma} \varphi(\mathbf{x}, t^n) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}_{K, \sigma} ds(\mathbf{x}) - \varphi_{\sigma}^{n+1} \int_{\sigma} \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}_{K, \sigma} ds(\mathbf{x}) \\ &= \int_{\sigma} (\varphi(\mathbf{x}, t^n) - \varphi_{\sigma}^{n+1}) (\mathbf{v}(\mathbf{x}) - \mathbf{v}_{\sigma}) \cdot \mathbf{n}_{K, \sigma} ds(\mathbf{x}), \end{aligned}$$

where we set

$$\mathbf{v}_{\sigma} = \frac{1}{|\sigma|} \int_{\sigma} \mathbf{v}(\mathbf{x}) ds(\mathbf{x}).$$

Using the regularity properties of  $\mathbf{v}$  and  $\varphi$  (we denote by  $C_{\mathbf{v}}$  a bound for the derivatives of  $\mathbf{v}$ ), we get that

$$\left| \int_{\sigma} \varphi(\mathbf{x}, t^n) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}_{K, \sigma} ds(\mathbf{x}) - \varphi_{\sigma}^{n+1} \int_{\sigma} \varphi(\mathbf{x}, t) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}_{K, \sigma} ds(\mathbf{x}) \right| \leq C_{\varphi} C_{\mathbf{v}} h_K^2 |\sigma|.$$

Since we may write

$$\left| \int_K \max(s(\mathbf{x}), 0) \varphi(\mathbf{x}, t^n) d\mathbf{x} - s_K^{(+)} \varphi_K^{n+1} \right| \leq C_\varphi h_K s_K^{(+)},$$

we get from (4)

$$\left| T_{11}^{(m)} - T_{12}^{(m)} \right| \leq h_m T C_\varphi \left( C_v \theta_D |\Omega| + 2 \int_\Omega \max(s(\mathbf{x}), 0) d\mathbf{x} \right).$$

We then define  $T_{13}^{(m)}$  (recalling that  $F_{K,\sigma} = F_{K,L}$  for  $\sigma = \sigma_{K,L}$  else  $F_{K,\sigma} = 0$  for  $\sigma \in \mathcal{F}_{\text{ext}}$ ) by

$$T_{13}^{(m)} = - \sum_{n=0}^{N-1} \tau^n \sum_{K \in \mathcal{M}} \left( u_K^{n+1} \sum_{\sigma \in \mathcal{F}_K} \varphi_\sigma^{n+1} F_{K,\sigma} + s_K^{(+)} (c_K^{n+1} - u_K^{n+1}) \varphi_K^{n+1} \right).$$

We have

$$T_{13}^{(m)} - T_{12}^{(m)} = - \sum_{n=0}^{N-1} \tau^n \sum_K u_K^{n+1} \sum_{\sigma \in \mathcal{F}_K} \varphi_\sigma^{n+1} (F_{K,\sigma} - |\sigma| \mathbf{v}_\sigma \cdot \mathbf{n}_{K,\sigma}).$$

Using (23) (which implies  $\sum_{\sigma \in \mathcal{F}_K} (F_{K,\sigma} - |\sigma| \mathbf{v}_\sigma \cdot \mathbf{n}_{K,\sigma}) = 0$ ), we get

$$T_{13}^{(m)} - T_{12}^{(m)} = - \sum_{n=0}^{N-1} \tau^n \sum_K u_K^{n+1} \sum_{\sigma \in \mathcal{F}_K} (\varphi_\sigma^{n+1} - \varphi_K^{n+1}) (F_{K,\sigma} - |\sigma| \mathbf{v}_\sigma \cdot \mathbf{n}_{K,\sigma}),$$

which leads to

$$(T_{13}^{(m)} - T_{12}^{(m)})^2 \leq C_\varphi^2 \left( \sum_{n=0}^{N-1} \tau^n \sum_K \sum_{\sigma \in \mathcal{F}_K} |\sigma| h_K \right) \left( \sum_{n=0}^{N-1} \tau^n \sum_K \sum_{\sigma \in \mathcal{F}_K} \frac{h_K}{|\sigma|} (F_{K,\sigma} - |\sigma| \mathbf{v}_\sigma \cdot \mathbf{n}_{K,\sigma})^2 \right),$$

hence providing

$$(T_{13}^{(m)} - T_{12}^{(m)})^2 \leq C_\varphi^2 T^2 \theta_D |\Omega| \sum_K \sum_{\sigma \in \mathcal{F}_K} \frac{h_K}{|\sigma|} (F_{K,\sigma} - |\sigma| \mathbf{v}_\sigma \cdot \mathbf{n}_{K,\sigma})^2,$$

which tends to zero thanks to (36). Gathering by pairs of control volumes (each one appears once in the summation), we have

$$T_8^{(m)} - T_{13}^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1}) (\widehat{F}_{L,K}^{(+)} \varphi_K^{n+1} - \widehat{F}_{K,L}^{(+)} \varphi_L^{n+1} + F_{K,L} \varphi_{K,L}^{n+1}),$$

setting  $\varphi_{K,L}^{n+1} = \varphi_{\sigma_{K,L}}^{n+1}$  if  $(K,L) \in S$  else  $\varphi_{K,L}^{n+1} = \frac{1}{2}(\varphi_K^{n+1} + \varphi_L^{n+1})$  (recall that  $F_{K,L} = 0$  if  $(K,L) \notin S$ ). Let us prove that  $\lim_{m \rightarrow \infty} |T_8^{(m)} - T_{13}^{(m)}| = 0$ , result which completes our proof. Since  $\widehat{F}_{K,L}^{(+)} - \widehat{F}_{L,K}^{(+)} = \widehat{F}_{K,L}$ , we get  $T_8^{(m)} - T_{13}^{(m)} = T_{14}^{(m)} + T_{15}^{(m)}$  with

$$T_{14}^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1}) (\widehat{F}_{L,K}^{(+)} (\varphi_K^{n+1} - \varphi_{K,L}^{n+1}) - \widehat{F}_{K,L}^{(+)} (\varphi_L^{n+1} - \varphi_{K,L}^{n+1})),$$

and

$$T_{15}^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1}) (F_{K,L} - \widehat{F}_{K,L}) \varphi_{K,L}^{n+1}.$$

We may write

$$\begin{aligned} |T_{14}^{(m)}| &\leq C_\varphi \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} |u_K^{n+1} - u_L^{n+1}| \max(h_K, h_L) (\widehat{F}_{K,L}^{(+)} + \widehat{F}_{L,K}^{(+)}) \\ &\leq C_\varphi \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} |u_K^{n+1} - u_L^{n+1}| \max(h_K, h_L) \widetilde{F}_{K,L}. \end{aligned}$$

Turning to the study of  $T_{15}^{(m)}$ , we get  $T_{15}^{(m)} = T_{16}^{(m)} - T_{17}^{(m)}$  with

$$T_{16}^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1}) F_{K,L} \varphi_{K,L}^{n+1},$$

and

$$T_{17}^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1}) \widehat{F}_{K,L} \varphi_{K,L}^{n+1}.$$

Remarking that, for  $P \in \widehat{\mathcal{P}}_{K,L}$ , we have

$$\sum_{(I,J) \in P} (u_I^{n+1} - u_J^{n+1}) = (u_K^{n+1} - u_L^{n+1}),$$

and that

$$\sum_{P \in \widehat{\mathcal{P}}_{K,L}} F_{K,L}^P = F_{K,L},$$

we get that

$$T_{16}^{(m)} = \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{(K,L) \in \mathcal{S}} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \sum_{(I,J) \in P} (u_I^{n+1} - u_J^{n+1}) \varphi_{K,L}^{n+1} F_{K,L}^P.$$

Besides, we have

$$\begin{aligned} T_{17}^{(m)} &= \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{(I,J) \in \widehat{\mathcal{S}}} (u_I^{n+1} - u_J^{n+1}) \widehat{F}_{I,J} \varphi_{I,J}^{n+1} \\ &= \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{(I,J) \in \widehat{\mathcal{S}}} (u_I^{n+1} - u_J^{n+1}) \sum_{(K,L) \in \mathcal{S}} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P F_{K,L}^P \varphi_{I,J}^{n+1}, \end{aligned}$$

which leads, thanks to  $\xi_{I,J}^P = 1$  if  $(I, J) \in P$  else  $\xi_{I,J}^P = 0$ , to

$$T_{17}^{(m)} = \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{(K,L) \in \mathcal{S}} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \sum_{(I,J) \in P} (u_I^{n+1} - u_J^{n+1}) \varphi_{I,J}^{n+1} F_{K,L}^P.$$

Hence

$$T_{15}^{(m)} = \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{(K,L) \in \mathcal{S}} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \sum_{(I,J) \in P} (u_I^{n+1} - u_J^{n+1}) (\varphi_{I,J}^{n+1} - \varphi_{K,L}^{n+1}) F_{K,L}^P.$$

We have  $|\varphi_{I,J}^{n+1} - \varphi_{K,L}^{n+1}| \leq C_\varphi \theta_{\widehat{\mathcal{P}}} \max(h_I, h_J)$ . Therefore we get

$$|T_{15}^{(m)}| \leq \frac{C_\varphi}{2} \theta_{\widehat{\mathcal{P}}} \sum_{n=0}^{N-1} \tau^n \sum_{(K,L) \in \mathcal{S}} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \sum_{(I,J) \in P} |u_I^{n+1} - u_J^{n+1}| \max(h_I, h_J) |F_{K,L}^P|,$$

which may also be rewritten as

$$|T_{15}^{(m)}| \leq \frac{C_\varphi}{2} \theta_{\hat{p}} \sum_{n=0}^{N-1} \tau^n \sum_{(K,L) \in \hat{S}} |u_K^{n+1} - u_L^{n+1}| \max(h_K, h_L) \tilde{F}_{K,L}.$$

Hence we get, setting  $C_1 = C_\varphi + C_\varphi \theta_{\hat{p}}$

$$|T_8^{(m)} - T_{13}^{(m)}| \leq C_1 \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} |u_K^{n+1} - u_L^{n+1}| \max(h_K, h_L) \tilde{F}_{K,L}.$$

Thanks to the Cauchy-Schwarz inequality and defining  $T_{18}$  by

$$T_{18}^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} \max(h_K, h_L) \tilde{F}_{K,L},$$

we have, thanks to Lemma 3.4,

$$(T_8^{(m)} - T_{13}^{(m)})^2 \leq C_1^2 T_{18}^{(m)} \left( h_m \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1})^2 \tilde{F}_{K,L} \right) \leq C_1^2 T_{18}^{(m)} h_m \frac{C_{BV}}{\nu}.$$

It now suffices to show that  $T_{18}^{(m)}$  remains bounded. Using (18), we have

$$T_{18}^{(m)} \leq \theta_{\hat{p}}^2 \sum_{\{K,L\} \subset \mathcal{M}} \max(h_K, h_L) |F_{K,L}|.$$

We then remark that

$$\left( \sum_{(K,L) \in S} \max(h_K, h_L) |F_{K,L}| \right)^2 \leq \left( \sum_{(K,L) \in S} \max(h_K, h_L) |\sigma_{K,L}| \right) \left( \sum_{(K,L) \in S} \frac{\max(h_K, h_L)}{|\sigma_{K,L}|} (F_{K,L})^2 \right).$$

The term  $\sum_{(K,L) \in S} \max(h_K, h_L) |\sigma_{K,L}|$  is bounded by  $2\theta_{\mathcal{D}} |\Omega|$ , and the term  $\sum_{(K,L) \in S} \frac{\max(h_K, h_L)}{|\sigma_{K,L}|} (F_{K,L})^2$  remains bounded thanks to (36) and to the bound  $2\theta_{\mathcal{D}} |\Omega| \|\mathbf{v}\|_\infty^2$  on  $\sum_{(K,L) \in S} \max(h_K, h_L) |\sigma_{K,L}| (\mathbf{v}_{\sigma_{K,L}} \cdot \mathbf{n}_{K,L})^2$ . This achieves the proof that

$$\lim_{m \rightarrow \infty} (T_8^{(m)} - T_{13}^{(m)}) = 0.$$

Gathering all the above results completes the proof that

$$\lim_{m \rightarrow +\infty} T_8^{(m)} = T_9,$$

and therefore the proof that  $u$  satisfies (22).  $\square$

## 4 Numerical results

### 4.1 A 2D case with radial symmetry

Let us consider Problem (1) on  $\Omega = (0, 1)^2$  with the following data:

$$k_1(u) = u^2, \quad k_2(u) = \frac{1}{\mu} (1 - u)^2, \quad f(u) = \frac{k_1(u)}{k_1(u) + k_2(u)},$$

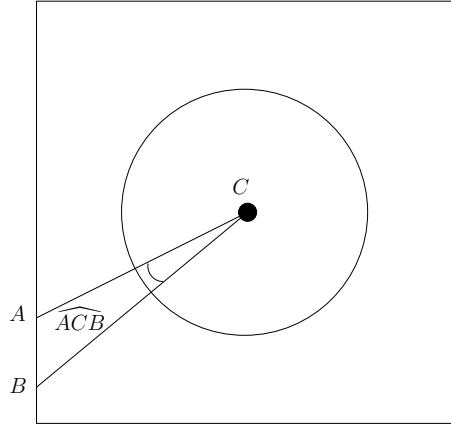


Figure 2: Geometry of the radial circular test.

A source term is imposed at each linear segment  $[A, B]$  of the boundary, equal to  $-\widehat{ACB}/(2\pi)$ , where  $\widehat{ACB}$  is the angle between the segments  $[C, A]$  and  $[C, B]$  and  $C$  has coordinates  $(\frac{1}{2}, \frac{1}{2})$  (see Figure 2). A punctual source term, equal to 1, is imposed at the point  $C$ . Then there exists a unique entropy weak analytical solution  $(p, u)$  only depending at each time on the distance  $r$  between  $\boldsymbol{x}$  and  $C$ , called the Buckley-Leverett solution in the framework of oil engineering (recall that, for a nonlinear problem without an entropy criterion, there exists an infinity of weak solutions in the general 1D case; in this 2D case, there may exist weak solutions without radial symmetry):

$$\begin{aligned} \bar{u} &= 1/\sqrt{\mu+1}, \quad \bar{v} = (1 + \sqrt{\mu+1})/2, \\ u(r, t) &= 0 \text{ for } \pi r^2 > \bar{v}t, \\ u(r, t) &= (f')^{(-1)}(\pi r^2/t) \text{ for } \pi r^2 < \bar{v}t, \\ p(r, t) &= \int_{r_0}^r \frac{1}{2\pi s(k_1(u(s, t)) + k_2(u(s, t)))} ds + p_0, \end{aligned}$$

where the value of the pressure is fixed at  $p_0$  at the distance  $r_0$  to point  $C$ .

We first consider the case  $\mu = 10$ . For the above solution, a circular discontinuity with height  $\bar{u}$  is located at the circle with centre  $C$  and radius  $R(t) = \sqrt{\bar{v}t}/\pi$  (for  $t = 0.2$  and  $\mu = 10$ , we have  $R(t) \simeq 0.37$ ). We use an IMPES scheme in a prototype running under SCILAB environment. At each time step, we use a standard 5-point scheme for solving the pressure equation, providing the values  $F_{K,L}^{n+1}$ . We then compute the new fluxes  $\widehat{F}_{K,L}^{n+1}$ , following the method described in Section 2.2. The strategy for determining the time step is based on a desired maximum variation of saturation between two time steps. This desired variation has been set to 0.2 for all run, except the modified scheme with  $\nu = 1$ , where we had to set this variation to 0.1 for stability reasons. The mesh is composed of  $41^2$  squares with side  $h = 1/41$  for all runs. The results are shown in Figures 3 to 6.

We observe in Figure 3 a very small dependence of the pressure isovalue lines on the GOE in this case. The profiles using the initial or the modified scheme show both a high radial symmetry for the approximated pressures (see Figure 4). Although the analytical pressure tends to infinity as  $r \rightarrow 0$ , the approximate pressures are greater, for the distances to  $C$  considered here, than the analytical ones, due to the fact that the approximate saturations are lower than the analytical ones in the neighbourhood of the point  $C$ . On the contrary, we see in Figure 5 a significant dependence of the isovalue lines on the GOE. We first remark that the value  $\omega = 0.1$  leads to much better results than  $\omega = 0$  (initial scheme). We also remark that the value  $\nu = 1$  leads to a solution where the numerical diffusion is slightly more important than with  $\nu = 0.1$  or  $\nu = 0$  (let us remark that the solutions obtained with  $\nu = 0.1$  and  $\nu = 0$  cannot be graphically distinguished, which enhances the possibility to use  $\nu = 0$  in practical cases).

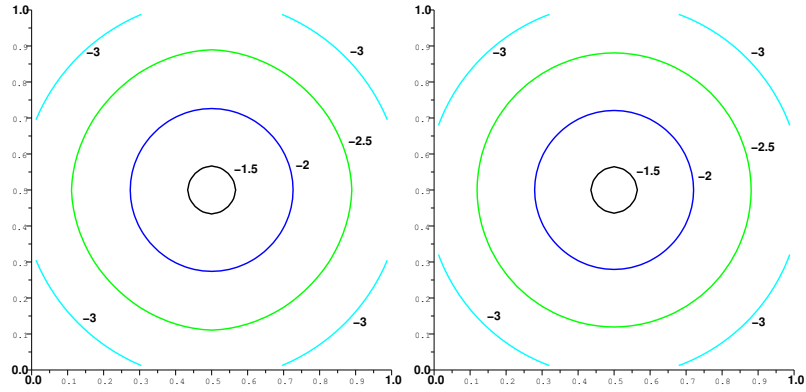


Figure 3: Contours of pressures at  $t = 0.2$  ( $\mu = 10$ ) with the initial (left) and modified scheme (right) with  $\nu = 0.1$ ,  $\omega = 0.1$ .

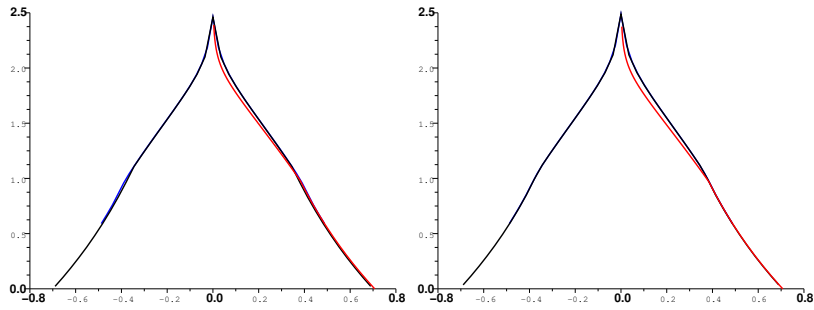


Figure 4: Profiles of pressures ( $p_0$  set to 0 at distance  $r_0 = \sqrt{2}/2$ ) at  $t = 0.2$  ( $\mu = 10$ ) with the initial scheme (left) and the modified scheme (right) with  $\nu = 0.1$ ,  $\omega = 0.1$ : analytical solution (red), profile along median axis (blue), diagonal profile (black).

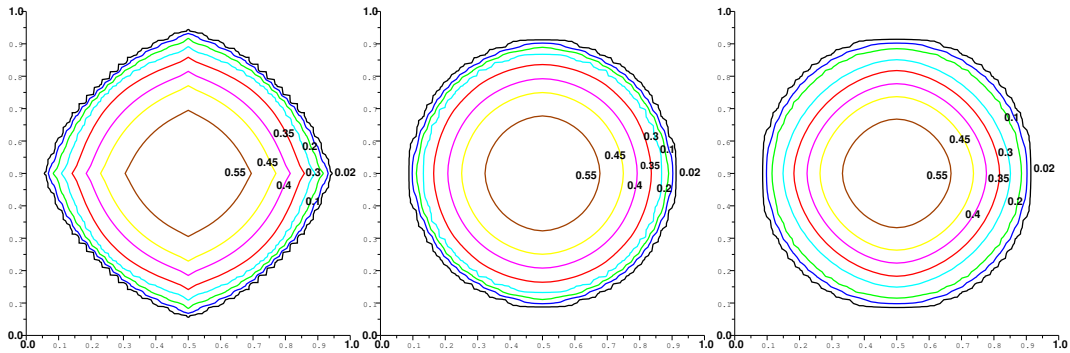


Figure 5: Contours of saturations at  $t = 0.2$  ( $\mu = 10$ ) with the initial (left), modified scheme (middle) with  $\nu = 0.1$ ,  $\omega = 0.1$ , modified scheme (right) with  $\nu = 1$   $\omega = 0.1$ .

We have then considered the case  $\mu = 200$ . The same observations as above hold, with a higher discrepancy between the initial and the modified scheme (see Figures 7 and 8). In order to get stable results, we had to set the desired variation of saturation to 0.05.

## 4.2 A 3D test case with three layers

The numerical tests presented here are inspired by [9]. The domain is defined by

$$\Omega = [-0.5, 0.5] \times [-0.5, 0.5] \times [-0.15, 0.15].$$

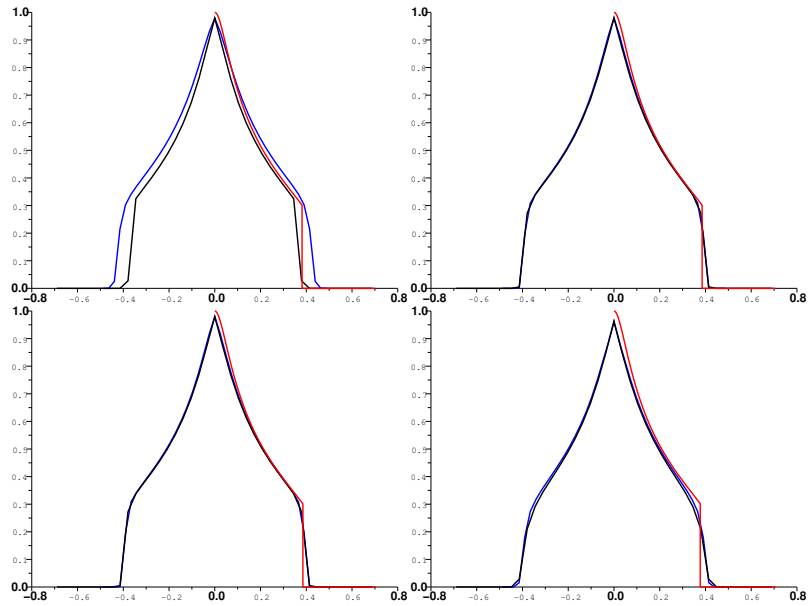


Figure 6: Profiles of saturations at  $t = 0.2$  ( $\mu = 10$ ) : analytical solution (red), profile along median axis (blue), diagonal profile (black); initial scheme (top left),  $\nu = 0$   $\omega = 0.1$  (top right),  $\nu = 0.1$   $\omega = 0.1$  (bottom left),  $\nu = 1$   $\omega = 0.1$  (bottom right).

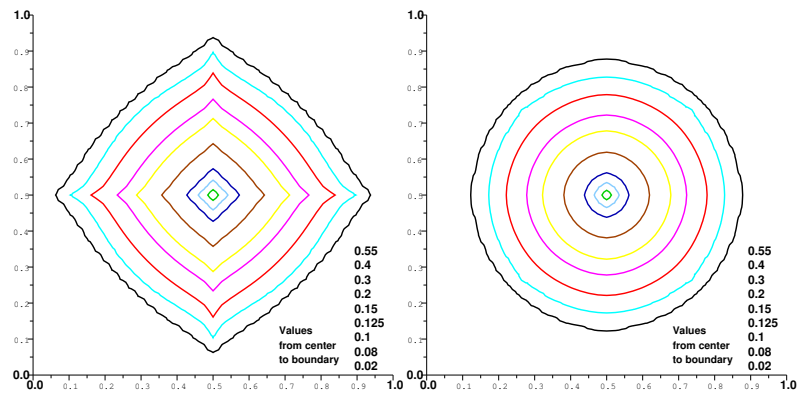


Figure 7: Contours of saturations at  $t = 0.05$  ( $\mu = 200$ ) with the initial (left) and the modified scheme (right) with  $\nu = 0.1$ ,  $\omega = 0.1$ .

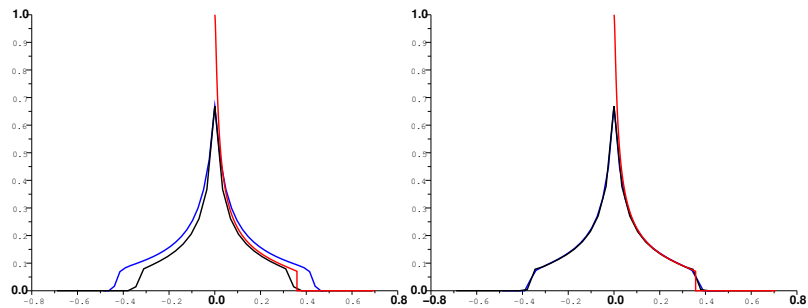


Figure 8: Profiles of saturations at  $t = 0.05$  ( $\mu = 200$ ) : analytical solution (red), profile along median axis (blue), diagonal profile (black); initial scheme (left),  $\nu = 0.1$   $\omega = 0.1$  (right).

The permeability  $\Lambda(\mathbf{x})$ ,  $\mathbf{x} \in \Omega$  is equal to 1 if the distance from  $\mathbf{x}$  to the vertical axis  $Oz$  is lower than 0.48, and to  $10^{-3}$  otherwise (see Figure 9), which ensures the confinement of the flow in the cylinder with axis  $Oz$  and radius 0.48. The density ratio is equal to 0.8. We use Corey-type relative permeability,  $k_1(u) = u^4$  and  $k_2 = (1 - u)^2/100$ . At the initial state, the reservoir is assumed to be saturated by the oil phase. Water is injected at the origin by an injection well. Two production wells, denoted by  $P_1$  and  $P_2$ , are respectively located at the points  $(-0.3\cos\frac{\pi}{3}, -0.3\sin\frac{\pi}{3}, 0)$  and  $(0.3\cos\frac{\pi}{3}, -0.3\sin\frac{\pi}{3}, 0)$

A prototype of an industrial code written in FORTRAN, based on an implicit scheme, is used for obtaining numerical results with two Cartesian grids, the second one deduced from the first one by a rotation of angle  $\theta = \frac{\pi}{6}$  with axis  $Oz$ . The number of cells in each direction  $(x, y, z)$  are  $N_x = N_y = 51$  and  $N_z = 3$  (which means that the three wells are numerically taken into account as source terms in the middle layer of the mesh).

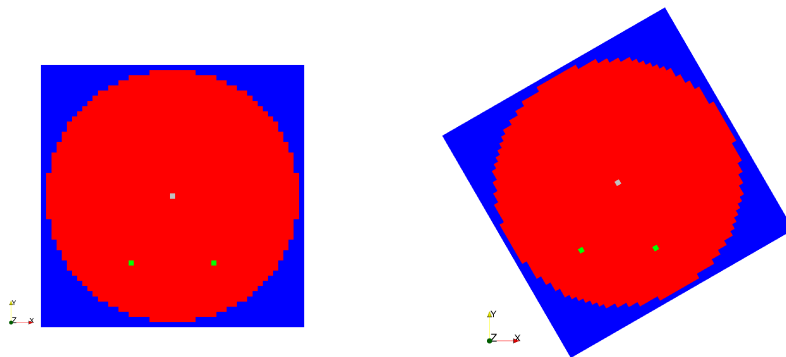
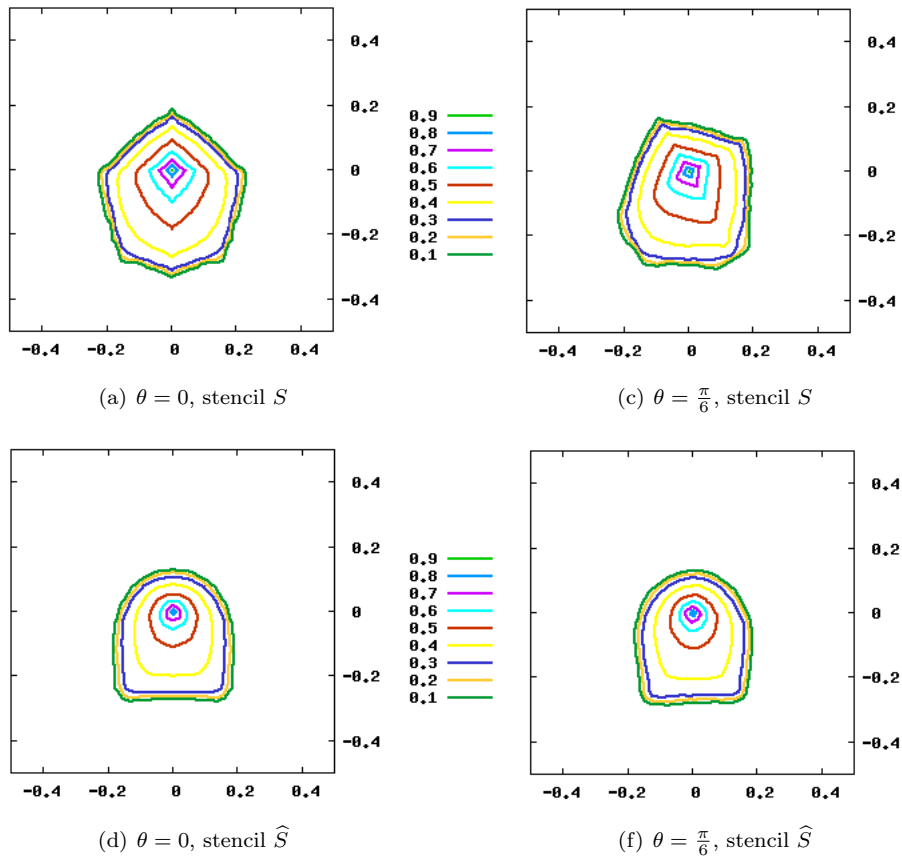


Figure 9: The two meshes used. In red, the highest permeability zone, in blue the lower permeability zone. Squares indicate wells.

At each time step, we use a the MPFA L-scheme [2] for solving the pressure equation, providing the values  $F_{K,L}$ . Then the method described in section 2.2 is used for the definition of new stencils, selecting  $\omega = 0.1$  for all faces which are inscribed in the cylinder. The parameter  $\nu$  is taken equal to 0, allowing to implement the scheme in standard industrial codes by only modifying the stencil of the MPFA scheme (see Remark 1).

The same value for the time step is used for all the computations, which are stopped once a given quantity of water has been injected. Note that, in the mesh depicted on the right part of Figure 9, the line  $(P_2, O)$  becomes the  $Oy$  axis of the mesh.

We see on Figure 10 the resulting contours of the saturation. We observe that the results obtained using the method described in this paper look very similar in the two grids, whereas the ones obtained using the initial five-point stencil are strongly distorted by the Grid Orientation Effect.

Figure 10: Water saturation contours  $u$  at the same time.

## 5 Conclusion

In this paper we have considered the nonlinear system of PDEs resulting from the conservation equations of two incompressible immiscible phases flowing within a porous medium. This system, which may also be seen as the coupling of a diffusion equation with respect to the pressure and a convection equation with respect to the saturation, is shown in practical cases of mobility contrast, to lead to the apparition of the so-called Grid Orientation Effect (GOE). We propose a new procedure to overcome this phenomenon, based on the modification of the stencil of the discrete version of the convection equation, without modifying the pressure equation.

Some theoretical results (such as the  $L^\infty$  estimate and the convergence of the new scheme by using the weak BV-inequality) are obtained in a simplified case and some numerical results, including the comparison with an analytical solution, show the efficiency and the accuracy of the method in the non-simplified one.

For some values of the parameters of the method, we obtain a natural version of the nine-point schemes defined some decades ago on regular grids, whose advantage is to apply on the structured but not regular grids used in reservoir simulation, in association with Multi-Point Flux Approximation finite volume schemes. In this case, it may be immediately implemented in standard industrial codes by a simple modification of the stencils.

## References

- [1] Aavatsmark, I., Eigestad, G.T.: Numerical Convergence of the MPFA O-method and U-method for General Quadrilateral Grids. *Int. J. Numer. Meth. Fluids* **51**, 939–961 (2006)
- [2] Aavatsmark, I., Eigestad, G.T., Heimsund, B.-O., Mallison, B.T., Nordbotten, J.M., Oian, E.: A New Finite-Volume Approach to Efficient Discretization on Challenging Grids. *SPE J.* **15(3)**, 658–669 (2010)
- [3] Agelas, L., Guichard, C., Masson, R.: Convergence of Finite Volume MPFA O-type Schemes for Heterogeneous Anisotropic Diffusion Problems. *IJVF* **7(2)**, (2010)
- [4] Aziz, K., Ramesh, A.B., Woo, P.T.: Fourth SPE Comparative Solution Project: Comparison of Steam Injection Simulators. *J. Pet. Tech.* **39**, 1576–1584 (1987)
- [5] Corre, B., Eymard, R., Quettier, L.: Applications of a Thermal Simulator to Field Cases, *SPE ATCE*, (1984)
- [6] Dawson, C., Sun, S., Wheeler, M.F.: Compatible Algorithms for Coupled Flow and Transport. *Comput. Meth. Appl. Mech. Eng.* **193**, 2565–2580 (2004)
- [7] Eymard, R., Gallouët, T., Herbin, R.: The Finite Volume Method. *Handbook of Numerical Analysis, Ph. Ciarlet J.L. Lions eds*, **7**, 715–1022 (2000).
- [8] Eymard, R., Sonier, F.: Mathematical and Numerical Properties of Control-Volume Finite-Element Scheme for Reservoir Simulation. *SPE Reservoir Eng.* **9**, 283–289 (1994)
- [9] Keilegavlen, E., Kozdon, J., Mallison, B.T.: Monotone Multi-dimensional Upstream Weighting on General Grids. *Proc. ECMOR XII*, Oxford, (2010)
- [10] Lipnikov, K., Moulton, J.D., Svyatskiy, D.: A Multilevel Multiscale Mimetic (M3) Method for Two-phase Flows in Porous Media. *J. Comput. Phys.* **14**, 6727–6753 (2008)
- [11] D.K. Ponting. Corner Point Geometry in Reservoir Simulation. *In Clarendon Press, editor, Proc. ECMOR I*, 45–65, Cambridge, (1989)
- [12] Vinsome, P., Au, A.: One Approach to the Grid Orientation Problem in Reservoir Simulation. *Old SPE J.* **21**, 160–161 (1981)
- [13] Yanosik, J.L., McCracken, T.A.: A Nine-point, Finite-Difference Reservoir Simulator for Realistic Prediction of Adverse Mobility Ratio Displacements. *Old SPE J.* **19**, 253–262 (1979)