ESAIM: M2AN 54 (2020) 1917–1949
https://doi.org/10.1051/m2an/2020025

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

SOME MULTIPLE FLOW DIRECTION ALGORITHMS FOR OVERLAND FLOW ON GENERAL MESHES

Julien Coatléven*

Abstract. After recalling the most classical multiple flow direction algorithms (MFD), we establish their equivalence with a well chosen discretization of Manning–Strickler models for water flow. From this analogy, we derive a new MFD algorithm that remains valid on general, possibly non conforming meshes. We also derive a convergence theory for MFD algorithms based on the Manning–Strickler models. Numerical experiments illustrate the good behavior of the method even on distorted meshes.

Mathematics Subject Classification. 65N08, 65N12, 65N15.

Received September 17, 2019. Accepted April 9, 2020.

1. Introduction

Overland water flow plays a major role in hydrogeology at many time scales, from million of years for stratigraphic studies of interest for the oil and gas industry to several days or weeks for predicting river flooding and landslides. Intermediate time scales are also increasingly studied as they are of crucial interest for climatic forecasts, from glacier withdrawal to desert expansion. Of course, a huge literature exists in the corresponding fields, describing the physical phenomenons involved with an equally huge model diversity (Navier–Stokes, Stokes, shallow water, etc...). Due to the complexity underlying flow models, approximate, phenomenologically based models have been developed to increase computational speed. Among those models, a common approach that has given satisfactory results is the one based on the so-called multiple flow directions algorithms (MFD).

The idea underlying all MFD algorithms is that water routing at large space scales will be mainly governed by the topographic slope. Using digital elevation models that provide a meshed representation of the topography, those algorithms compute water flow by distributing water from the topographically higher discrete cells to the topographically lower ones, each distribution formula corresponding to a specific MFD algorithm. In most cases, the MFD algorithms are developed assuming that the mesh is a uniform cartesian grid, with the same space step in each direction. Historically, the first MFD algorithms were in fact single flow direction algorithms, where the distribution formula selected only one neighbour, generally the one with the steepest slope. The deficiencies of such simple algorithms are the origin of the development of MFD algorithms (see [26]). The most classical MFD algorithm [16, 22] uses directly the slope to distribute the flow, while models using powers of the slope were developed to concentrate the flow and limit diffusion effects due to the use of coarse meshes

Keywords and phrases. Multiple flow direction algorithm, overland flow, virtual element method, hybrid finite volume, general meshes.

IFP Énergies nouvelles, 1 et 4 avenue de Bois-Préau, Rueil-Malmaison 92852, France.
*Corresponding author: julien.coatleven@ifpen.fr

(see [18, 21, 25]). Some attempts have been made to apply MFD algorithms on more general meshes, with an emphasis on triangular ones (see [24,27]). We refer the reader to [9] for a comparison of some of those methods, and to the references in the aforementioned papers for a more complete overview of the tremendous literature on MFD algorithms, on which we will not try to be exhaustive.

All those algorithms proceed sequentially from the higher cells to the lower ones, however as was already noticed in [23], using an MFD algorithm is in fact equivalent to solving a linear system using a specific cell ordering. The key idea underlying the present paper comes from a deeper study of the linear system associated with one of the earliest MFD algorithm. Indeed, we will explain that this linear system is completely equivalent to a two point flux finite volume scheme (see [10]) applied to a stationary Manning–Strickler model for water flow. It was then clear that replacing the TPFA scheme that requires a strong orthogonality hypothesis on meshes to remain valid by more advanced flux approximation schemes would allow us to derive MFD algorithms adapted to general meshes. Moreover, this equivalence will also allow us to derive a theoretical framework within which we will be able to study the convergence properties of MFD algorithms. In our numerical experiments, we have considered two flux reconstructions inspired by two finite volume schemes: the hybrid finite volume (see [11,12]) and the virtual volume method (or conservative first order virtual element method, see [5]). Of course, other choices could have been made (for instance VAG finite volumes [13,14] or discontinuous Galerkin methods [7]), however those two schemes will be sufficient to illustrate our approach.

The paper will be organized as follows: after describing the data and meshes, we recall the most classical multiple flow direction algorithms, and reformulate them in a more algebraic fashion. Next, using this reformulation we explain how they are linked with the TPFA scheme for a family of Manning–Strickler models. We elaborate on this basis to overcome the mesh limitations induced by the TPFA scheme, introducing a new family of MFD algorithms that will remain valid on a huge class of meshes, including those with hanging nodes. We then study the convergence properties of all methods and conclude by some numerical illustrations.

2. Classical multiple flow direction algorithms

2.1. Mesh and data description

Let Ω be a bounded polyhedral connected domain of $\mathbb{R}^2$, whose boundary is denoted $\partial\Omega = \overline{\Omega}\backslash\Omega$. We recall the usual notations describing a mesh $\mathcal{M} = (\mathcal{T}, \mathcal{F})$ of Ω. $\mathcal{T}$ is a finite family of connected open disjoint polygonal subsets of Ω (the cells of the mesh), such that $\overline{\Omega} = \cup_{K\in\mathcal{T}}\overline{K}$. For any $K \in \mathcal{T}$, we denote by $|K|$ the measure of $|K|$, by $\partial K = \overline{K}\backslash K$ the boundary of K, by h_K its diameter and by $\boldsymbol{x}_K$ its barycenter. $\mathcal{F}$ is a finite family of disjoint subsets of hyperplanes of $\mathbb{R}^2$ included in $\overline{\Omega}$ (the faces of the mesh) such that, for all $\sigma \in \mathcal{F}$, its measure is denoted $|\sigma|$, its diameter h_σ and its barycenter $\boldsymbol{x}_\sigma$. For any $K \in \mathcal{T}$, there exists a subset $\mathcal{F}_K$ of $\mathcal{F}$ such that $\partial K = \cup_{\sigma\in\mathcal{F}_K}\overline{\sigma}$. Then, for any $\sigma \in \mathcal{F}$, we denote by $\mathcal{T}_\sigma = \{K \in \mathcal{T} \mid \sigma \in \mathcal{F}_K\}$. Next, for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{F}_K$, we denote by $\boldsymbol{n}_{K,\sigma}$ the unit normal vector to σ outward to K, and $d_{K,\sigma} = ||\boldsymbol{x}_\sigma - \boldsymbol{x}_K||$. The set of boundary faces is denoted $\mathcal{F}_{\text{ext}}$, while interior faces are denoted $\mathcal{F}_{\text{int}}$. Finally for any $\sigma \in \mathcal{F}_{\text{int}}$, we denote $d_{KL} = ||\boldsymbol{x}_K - \boldsymbol{x}_L||$ where $\mathcal{T}_\sigma = \{K, L\}$. We assume that there exists a subset $\mathcal{F}_{\text{ext,in}}$ of $\mathcal{F}_{\text{ext}}$ such that:

$$\overline{\partial\Omega}_{\text{in}} = \bigcup_{\sigma\in\mathcal{F}_{\text{ext,in}}} \overline{\sigma} \quad \text{where} \quad \partial\Omega_{\text{in}} = \{\boldsymbol{x} \in \partial\Omega \mid \nabla b \cdot n > 0\}$$

and we denote of course $\mathcal{F}_{\text{ext,out}} = \mathcal{F}_{\text{ext}}\backslash\mathcal{F}_{\text{ext,in}}$. As usual, $h = \max_{K\in\mathcal{T}} h_K$ will denote the mesh size. In what follows, we will assume that our mesh satisfies:

(A1) There exists a real number $\rho > 0$ and a matching simplicial submesh $\mathcal{ST}$ of $\mathcal{M}$ such that for any $T \in \mathcal{ST}$, $\rho h_T \leqslant r_T$ where r_T is the inradius of T, and for any $T \in \mathcal{T}$ and any $T \in \mathcal{ST}$ such that $T \subset K$, $\rho h_K \leqslant h_T$.

From [3,7], we know that assumption (A1) implies that for any $k \in \mathbb{N}$, there exists $C_{tr,k} > 0$ independent on h such that for any $K \in \mathcal{T}$, any $\sigma \in \mathcal{F}_K$ and any $p \in \mathbb{P}_k(K)$:

$$||p||_{L^2(\sigma)} \leqslant C_{tr,k} h_\sigma^{-1/2} ||p||_{L^2(K)}. \tag{2.1}$$

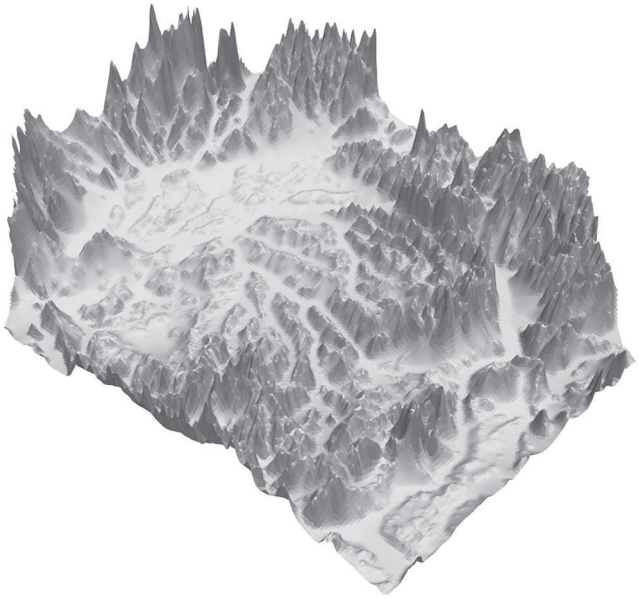

FIGURE 1. Example of real topography.

There also exists $C_{tr} > 0$ independent on h such that for any $v \in H^1(K)$:

$$||v||_{L^2(\sigma)} \leqslant C_{tr} \left(h_K^{-1} ||v||^2_{L^2(K)} + h_K ||\nabla v||^2_{L^2(K)} \right)^{1/2}. \tag{2.2}$$

Finally, assumption (A1) implies that for any integer k, there exists $C_{\mathrm{poly},k} > 0$ such that for any $K \in \mathcal{T}$ and any $v \in H^s(K)$ with $s \in \{1, \ldots, k+1\}$, there exists $p \in \mathbb{P}_k(K)$ such that

$$|v-p|_{H^m(K)} + h_K^{1/2} |v-p|_{H^m(\partial K)} \leqslant C_{\mathrm{poly},k} h_K^{s-m} |v|_{H^s(K)} \quad \text{for} \quad m \in \{0, \ldots, s-1\}. \tag{2.3}$$

In the estimates that will follow, the constant $C > 0$ will always denote by convention a quantity independent on the mesh size h, whose value can change from line to line. Also by convention for any $\sigma \in \mathcal{F}_{\mathrm{int}}$ we will denote $\mathcal{T}_\sigma = \{K, L\}$. In the same way, when considering a cell K and one of its interior faces $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\mathrm{int}}$, by convention the cell L will denote the other element of $\mathcal{T}_\sigma$.

In order to be able to establish error estimates, we assume that the topography, often called the basement in the geological community and thus usually denoted b, satisfies $b \in W^{2,\infty}(\Omega)$ and that there exists $\alpha > 0$ such that $-\Delta b \geqslant \alpha$ for almost every $\boldsymbol{x} \in \Omega$. However, from the practical point of view b is only known through some pointwise values, very often given as a digitized surface elevation, on which some interpolation and upscaling or downscaling processes have been applied (see Fig. 1). Thus, its Laplacian cannot be expected to be practically computable, and the above assumption should really be considered as an abstract technical requirement for establishing convergence estimates. Thus, our data will more realistically consists in some discrete mesh-based $(\boldsymbol{B}_K)_{K \in \mathcal{T}}$ representation of b, where each $\boldsymbol{B}_K$ can represent several pointwise values of b, complemented by a water source term $f \in L^\infty(\Omega)$, possibly also only known through a mesh-based representation.

2.2. The classical multiple flow direction algorithm

In the geological literature, multiple flow direction algorithms are considered as purely algorithmic ways of distributing water from one cell to another. Thus, they are generally described in a purely algorithmic fashion. Consequently, in this section to introduce the most classical MFD algorithms we will follow the presentation

generally found in the geological community, that is to say a purely algorithmic point of view. One of our first tasks will precisely consists in abstracting ourselves from this algorithmic setting. For any $K \in \mathcal{T}$, let b_K be a value for the topography associated with cell K. To fix notations, consider for the moment that

$$b_K = \frac{1}{|K|} \int_K b \quad \forall K \in \mathcal{T}$$

but other approximations can be considered, for instance $b_K = b(\boldsymbol{x}_K)$ for any $K \in \mathcal{T}$ if b is regular enough. Multiple flow direction algorithms are based on formulae to distribute the water flow from a cell to its neighbours that are topographically lower. The most classical distribution formula consists simply in distributing the flow proportionally to the ratio s_{KL}/s_K of the discrete slope s_{KL} between the high cell K and the low cell L regarding the total positive slope s_K of the high cell K, where the discrete slope s_{KL} is given by

$$s_{KL} = \frac{|\sigma|}{d_{KL}} (b_K - b_L)$$

and the total positive slope s_K of cell K is given by

$$s_K = \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{int}}, b_K \geqslant b_L} \frac{|\sigma|}{d_{KL}} (b_K - b_L).$$

Notice that in many cases of the literature, as the MFD algorithm is applied on a uniform cartesian mesh with the same space step in every direction, the face measure $|\sigma|$ is simply omitted (see for instance [16, 22]), with no impact on the ratio s_{KL}/s_K. To give a detailed description of the most classical multiple flow direction algorithms, we need to introduce some notations. Let $b_0 = \max\{b_K \mid K \in \mathcal{T}\}$, $\mathcal{T}_0 = \{K \in \mathcal{T} \mid b_K = b_0\}$ and $\hat{\mathcal{T}}_{-1} = \varnothing$. The set $\mathcal{T}_0$ thus denotes the set of cells with maximum topographic height. Then, for any $n \in \mathbb{N}$ we define the set $\hat{\mathcal{T}}_n$ of elements of $\mathcal{T}$ by setting:

$$\hat{\mathcal{T}}_n = \bigcup_{0 \leqslant i \leqslant n} \mathcal{T}_i \quad \text{where} \quad \mathcal{T}_i = \{K \in \mathcal{T} \mid b_K = b_i\} \quad \text{and} \quad b_i = \max\left\{b_K \mid K \in \mathcal{T} \backslash \hat{\mathcal{T}}_{i-1}\right\}.$$

By construction, as $\mathcal{T}$ is a finite set there exists $N_b > 0$ such that $\mathcal{T}_{N_b-1} \neq \varnothing$ and $\hat{\mathcal{T}}_{N_b-1} = \mathcal{T}$. Thus, for any $n \geqslant N_b$, we have $\mathcal{T}_n = \varnothing$ and $\hat{\mathcal{T}}_n = \hat{\mathcal{T}}_{N_b-1}$.

To model water sources, we define a family $(f_K)_{K \in \mathcal{T}}$ by setting:

$$f_K = \frac{1}{|K|} \int_K f \quad \forall K \in \mathcal{T}.$$

The classical MFD algorithm (reformulated from [16,22]) then reads as follows (see Fig. 2 for a visual illustration, where the width of the arrows is roughly following the amount of distributed water):

(i) For any $K \in \mathcal{T}$, the water influx $\widetilde{q}_K$ is initialized at 0.

(ii) For any $K \in \mathcal{T}_0$, the water influx $\widetilde{q}_K$ is given by $\widetilde{q}_K = |K| f_K$.

(iii) Loop for $n = 1 \ldots N_b - 1$.

(iii-1) For any $L \in \hat{\mathcal{T}}_{n-1}$, we distribute the entire water influx $\widetilde{q}_L$ to the neighbours that belong to $\mathcal{T}_n$ proportionally to the slope between the cell L and its neighbour, *i.e.*:

$$\widetilde{q}_K \longleftarrow \widetilde{q}_K + \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{int}}, b_K < b_L, L \in \hat{\mathcal{T}}_{n-1}} \frac{|\sigma| \widetilde{q}_L}{d_{KL} s_L} (b_L - b_K) \quad \text{for all } K \in \mathcal{T}_n$$

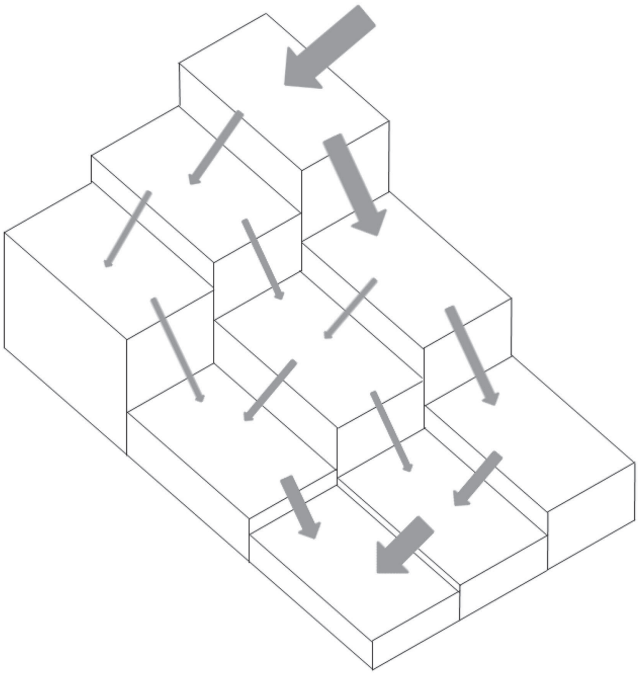

FIGURE 2. Basic principle of MFD algorithms: water is distributed to lower neighbouring cells proportionally to the slope.

where s_L is the total positive slope in cell L, and $\longleftarrow$ denotes the action of updating $\widetilde{q}_K$.

(iii-2) For any $K \in \mathcal{T}_n$, the water influx is complemented by the local sources by setting

$$\widetilde{q}_K \longleftarrow \widetilde{q}_K + |K| f_K.$$

(iv) End loop for $n = 1 \dots N_b - 1$.

The MFD algorithm admits a reformulation as a linear system that will play a key role in the remaining of the paper. To our knowledge, only [23] mentioned this reformulation, although without exhibiting an explicit formula. This link seemed to have been most of the time simply overlooked by the geological community:

Theorem 2.1. *The MFD algorithm is equivalent to solving the following linear system for the unknown* $(\widetilde{q}_K)_{K\in\mathcal{T}}$

$$\widetilde{q}_K - \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm int}, b_K<b_L} \frac{|\sigma|\widetilde{q}_L}{d_{KL}s_L}(b_L - b_K) = |K| f_K \quad \forall K \in \mathcal{T} \tag{2.4}$$

where

$$s_K = \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm int}, b_K\geqslant b_L} \frac{|\sigma|}{d_{KL}}(b_K - b_L)$$

using an ordering for the cells of $\mathcal{T}$ *based on decreasing topography* b_K *and a lower triangular solver.*

Proof. First, notice that as cells $K \in \mathcal{T}_n$ are updated only at step n, their expression is complete past this step and given by:

$$\widetilde{q}_K = \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm int}, b_K<b_L, L\in\hat{\mathcal{T}}_{n-1}} \frac{|\sigma|\widetilde{q}_L}{d_{KL}s_L}(b_L - b_K) + |K| f_K \quad \text{for all } K \in \mathcal{T}\backslash\mathcal{T}_n.$$

However, by construction of the sets $\hat{\mathcal{T}}_i$, any $L \in \mathcal{T}$ such that $b_L > b_K$ for $K \in \mathcal{T}_n$, will satisfy $L \in \hat{\mathcal{T}}_{n-1}$. Thus, the above expression can be simplified in:

$$\widetilde{q}_K = \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{int}}, b_K < b_L} \frac{|\sigma| \widetilde{q}_L}{d_{KL} s_L} (b_L - b_K) + |K| f_K \quad \text{for all } K \in \mathcal{T} \backslash \mathcal{T}_n$$

where the set $\hat{\mathcal{T}}_{n-1}$ no longer appears. As for any $K \in \mathcal{T}$, there exists $0 \leqslant n \leqslant N_b - 1$ such that $K \in \mathcal{T}_n$, and noticing that $s_L > 0$ as soon as there exists $\sigma \in \mathcal{F}_L \cap \mathcal{F}_{\text{int}}$, $\mathcal{T}_\sigma = \{K, L\}$ such that $b_L > b_K$, (2.4) follows by induction. Finally, starting from (2.4), if one chooses an ordering for the cells of $\mathcal{T}$ based on decreasing topography b_K, it is clear that the above system becomes a lower triangular one for $\widetilde{q}_K$. It should be obvious at this point that the associated lower triangular solver then coincides exactly with the classical MFD algorithm. □

A great advantage of considering the above system rather than its algorithmic counterpart is that it makes clear that triangular solvers are not the only possible linear solvers, which was indeed the main point of [23]. A wide range of linear solvers can be used to solve (2.4), and most importantly parallel solvers than can considerably speed up the solving process on meshes with a huge cell number.

Remark 2.2. Many variants of this classical MFD algorithm exist, which at least to the authors knowledge mainly consist in modifying the way the influx is distributed from an upper cell to its lower neighbours: powers of the slope instead of the slope, probabilistic repartitions, repartitions using a specified number of neighbours for a given mesh structure, etc... With the exception of distribution formulae that use powers of the slope instead of the slope itself, proceeding as we have done for the most classical MFD algorithm it is not difficult to show that all those variants can be rewritten under the form:

$$\widetilde{q}_K - \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{int}}, b_K < b_L} \frac{\lambda_\sigma |\sigma| \widetilde{q}_L}{d_{KL} s_L} (b_L - b_K) = |K| f_K \quad \forall K \in \mathcal{T}$$

with this time

$$s_K = \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{int}}, b_K \geqslant b_L} \frac{\lambda_\sigma |\sigma|}{d_{KL}} (b_K - b_L) .$$

The coefficient λ_σ associated with each face will take different values depending on the way one want to modify the influx repartition. Notice that we have chosen to not consider distribution formulae based on powers of the slope as they only had some technical difficulties with no fundamental differences in the analysis.

3. On the link with Manning–Strickler models

Consider the following classical transient Manning–Strickler model:

$$\left| \begin{array}{ll} \dfrac{\partial u}{\partial t} - \operatorname{div}(\lambda u^m \nabla b) = f & \text{in} \quad \Omega \\ -u^m \nabla b \cdot \boldsymbol{n} = 0 & \text{on} \quad \partial\Omega_{\text{in}} \end{array} \right.$$

where u is the water height, m a parameter and $\boldsymbol{n}$ denotes the outward normal to Ω. The coefficient $\lambda \in L^\infty(\Omega)$ is the inverse of the Manning–Strickler coefficient expressing soil roughness, and is assumed to satisfy $0 < \lambda_- \leqslant \lambda \leqslant \lambda_+ < +\infty$ almost everywhere in Ω. Formally, the steady state associated with the above system is the following stationary Manning–Strickler model for overland flow:

$$\left| \begin{array}{ll} -\operatorname{div}(\lambda u^m \nabla b) = f & \text{in} \quad \Omega \\ -\lambda u^m \nabla b \cdot n = 0 & \text{on} \quad \partial\Omega_{\text{in}}. \end{array} \right. \tag{3.1}$$

Denoting $u_{\lambda,m} = \lambda u^m$, $\boldsymbol{\beta} = -\nabla b \in W^{1,\infty}(\Omega)$ and $\mu = -\Delta b \geqslant \alpha > 0 \in L^\infty(\Omega)$, remark that (3.1) can be rewritten:

$$\begin{vmatrix} \boldsymbol{\beta} \cdot \nabla u_{\lambda,m} + \mu u_{\lambda,m} = f & \text{in} & \Omega \\ u_{\lambda,m} = 0 & \text{on} & \partial\Omega_{\text{in}}. \end{vmatrix} \tag{3.2}$$

Stationary transport problems of the form (3.2) have of course received a lot of attention in the existing literature, in particular as they correspond to a popular model for neutron transport. Their well-posedness has been considered for instance in [1, 2], or more recently in [8, 15, 17]. Thus, from the results of [8, 15] and the regularity hypotheses on b, f, and λ, we know that there exists a unique $u \in L^\infty(\Omega)$ solution of (3.2). As the water height u only appears through its m-th power u^m, in the remaining of the paper we will with a slight abuse of notations directly use u instead of u^m.

We are now going to explain that the classical MFD algorithm coincides with a well chosen discretization of the stationary Manning–Strickler model. Assume that the mesh is orthogonal, *i.e.* there exists a family of centroids $(\overline{\boldsymbol{x}}_K)_{K\in\mathcal{T}}$ such that:

$$\overline{\boldsymbol{x}}_K \in \mathring{K} \quad \forall K \in \mathcal{T} \quad \text{and} \quad \frac{\overline{\boldsymbol{x}}_L - \overline{\boldsymbol{x}}_K}{||\overline{\boldsymbol{x}}_L - \overline{\boldsymbol{x}}_K||} = \boldsymbol{n}_{K,\sigma} \quad \text{for} \quad \sigma \in \mathcal{F}_{\text{int}},\ \sigma = \{K, L\}$$

and let us denote $\overline{d}_{K,\sigma}$ the distance of $\overline{\boldsymbol{x}}_K$ to the hyperplane containing σ for any $\sigma \in \mathcal{F}_K$ and any $K \in \mathcal{T}$. Then, one can use a two-point finite volume scheme to discretize the steady-state Manning model. Assume that $b_K = b(\overline{\boldsymbol{x}}_K)$ or at least a second order approximation of it. Denoting u_K for any $K \in \mathcal{T}$ the discrete water height unknown, if one further assumes that $b_\sigma = b_K$ for any $\sigma \in \mathcal{F}_{\text{ext}}$ and $K \in \mathcal{T}_\sigma$ which is generally what is done in practical applications of the MFD algorithm, for any $K \in \mathcal{T}$ we get:

$$\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}}} \tau_{KL}\widetilde{u}_\sigma^{\text{up}}(b_K - b_L) = |K| f_K$$

where $\widetilde{u}_\sigma^{\text{up}} = u_K$ if $b_K \geqslant b_L$ and $\widetilde{u}_\sigma^{\text{up}} = u_L$ if $b_K < b_L$ and the transmissivity τ_{KL} is given for instance by the harmonic mean:

$$\tau_{KL} = \frac{|\sigma|\lambda_K\lambda_L}{\lambda_K\overline{d}_{L,\sigma} + \lambda_L\overline{d}_{K,\sigma}} \quad \text{where} \quad \lambda_K = \frac{1}{|K|}\int_K \lambda.$$

Immediately we deduce the following equivalence result, which despite its simplicity is the cornerstone of the present paper:

Theorem 3.1. *The TPFA scheme for Manning Strickler's model is equivalent to the MFD algorithm.*

Proof. Gathering the faces by upwinding kind, we get:

$$\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}},b_K\geqslant b_L} \tau_{KL}u_K(b_K - b_L) - \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}},b_K<b_L} \tau_{KL}u_L(b_L - b_K) = |K| f_K. \tag{3.3}$$

Setting

$$s_K = \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}},b_K\geqslant b_L} \tau_{KL}(b_K - b_L)$$

and noticing that $s_L > 0$ as soon as there exists $\sigma \in \mathcal{F}_L \cap \mathcal{F}_{\text{int}}$ such that $b_L > b_K$, we see that equation (3.3) can be rewritten:

$$s_K u_K - \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}},b_K<b_L} \tau_{KL}u_L(b_L - b_K) = |K| f_K.$$

Defining the water influx by $\widetilde{q}_K = s_K u_K$, we thus obtain:

$$\widetilde{q}_K - \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}},b_K<b_L} \tau_{KL}\frac{\widetilde{q}_L}{s_L}(b_L - b_K) = |K| f_K. \tag{3.4}$$

If we take $\lambda = 1$, immediately we see that (3.4) is exactly the MFD equation (2.4). □

Surprisingly, this result seems to be absent of the MFD literature. The main reason is probably that formulations of MFD algorithms as linear systems are equally difficult to find. From this equivalence between the classical MFD and the two-point flux approximation (TPFA) of the classical Manning–Strickler model, some useful observations can be made. Probably the most surprising one is that the MFD unknown $(\widetilde{q}_K)_{K\in\mathcal{T}}$ can be used instead of u_K to solve the two-point approximation of the Manning–Strickler model. Moreover, existence and uniqueness of solutions of the MFD problem (3.4) are now an immediate consequence of its equivalence with the MFD algorithm without requiring any hypothesis on the topography. Next, let us mention that when modeling overland flow, the quantity of interest is not the water height but the water discharge, *i.e.* the norm $||\lambda h^m \nabla b||$ of the water flux vector. This is the reason why no water height explicitly appears in MFD algorithms. However, a crucial consequence of our equivalence result is that the usual unknown $\widetilde{q}_K$ of the MFD algorithm, while an excellent choice from the algebraic perspective, is probably not the good quantity to represent the water discharge. Indeed, using $\widetilde{q}_K = s_K u_K$ and the consistency of the two-point formula we see that it approximates:

$$\widetilde{q}_K \approx \sum_{\sigma\in\mathcal{F}_K} \int_\sigma u \left(-\lambda \nabla b \cdot \boldsymbol{n}_{K,\sigma}\right)^+$$

where $v^+ = \max(0, v)$. We cannot expect such a $\widetilde{q}_K$ to be an approximation of the norm of $\lambda u \nabla b$, as it approximates the accumulated influx in a cell which is a mesh dependent quantity. Thus, no kind of convergence let alone approximation properties can be expected for such a quantity in general. To effectively compute a discrete water discharge q_K for each cell $K \in \mathcal{T}$, we reconstruct cellwise the water flux vector by setting:

$$\boldsymbol{Q}_K = \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm int}, b_K > b_L} \frac{\tau_{KL}\widetilde{q}_K}{|K| s_K} \left(b_K - b_L\right)\left(\boldsymbol{x}_\sigma - \boldsymbol{x}_K\right) - \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm int}, b_K < b_L} \frac{\tau_{KL}\widetilde{q}_L}{|K| s_L} \left(b_L - b_K\right)\left(\boldsymbol{x}_\sigma - \boldsymbol{x}_K\right).$$

The consistent water discharge is immediately given by $q_K = ||\boldsymbol{Q}_K||$. We will illustrate in the numerical section the convergence deficiency of $\widetilde{q}_K$, and how q_K has a much better behavior. Obviously, for any $K \in \mathcal{T}$ such that $s_K \neq 0$, we can compute an equivalent positive water height by setting $u_K = \widetilde{q}_K / s_K$. Cells where $s_K = 0$ are cells where all discrete fluxes are ingoing fluxes, thus it is clear that one should either set $u_K = +\infty$ or $u_K = 0$ depending if water effectively reaches the cell or not. From the definition of $\boldsymbol{Q}_K$, we see that the value of u_K for such cells will have no influence on the water discharge and its asymptotic convergence. It is thus clear that one can always define u_K by setting:

$$u_K = \begin{cases} \dfrac{\widetilde{q}_K}{s_K} & \text{if} \quad s_K > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Remark 3.2. Notice that using the harmonic mean is not mandatory if one can use an approximation of λ directly on the faces, leading to the transmissivity $\tau_{KL} = |\sigma|\lambda_\sigma / d_{KL}$ in which case (3.4) will correspond to one of the many variants of the classical MFD.

4. Some conservative and consistent flux reconstructions on general meshes

From the equivalence between the MFD algorithms and the TPFA scheme for the stationary Manning–Strickler model, an obvious way of generalizing MFD algorithms to general meshes consists in replacing the TPFA flux reconstruction formula by more advanced flux reconstruction techniques that are still valid on general meshes. We follow the idea of [4]: we consider a gradient reconstruction in each cell that will consist in a consistent part plus a stabilization part. However, contrary to [4] where the stabilization was used to obtain coercivity, here we will use this stabilization to enforce conservativity.

4.1. A conservative flux reconstruction formula

For any cell $K \in \mathcal{T}$, let us denote $\boldsymbol{B}_K = \boldsymbol{\mathcal{D}}_K(b)$ the local values associated with the topography b, and $X_K = \mathbb{R}^{\mathcal{N}_K}$ the set of local values, with $\mathcal{N}_K$ the number of local values. The operator $\boldsymbol{\mathcal{D}}_K : H^1(K) \longmapsto X_K$ will of course depend on the considered reconstruction formula, however to simplify notations we consider that the value b_K at $\boldsymbol{x}_K$ always belongs to the set of local values. Those local values represent all the knowledge we have in practice of the field b, and once again its Laplacian cannot be expected to be computable. For typical applications such as stratigraphic modelling it consists in cell values complemented by vertex or face values, thus conditioning the schemes we can effectively use. We assume that we are given a gradient reconstruction operator $\nabla_K : X_K \longmapsto \mathbb{R}^2$, and we define a reconstruction formula in each cell through the operator $\Pi_K : X_K \longmapsto \mathbb{P}_1(K)$, where $\mathbb{P}_1(K)$ is the set of first order polynomials on K and:

$$\Pi_K\left(\boldsymbol{B}_K\right)(\boldsymbol{x}) = b_K + \nabla_K\left(\boldsymbol{B}_K\right) \cdot (\boldsymbol{x} - \boldsymbol{x}_K)$$

and thus $\nabla \Pi_K\left(\boldsymbol{B}_K\right) = \nabla_K\left(\boldsymbol{B}_K\right)$. Such gradient reconstructions are the usual basis of modern finite volume methods (see [4, 5, 12]), and many formulae can be found in the literature. Using those reconstructions, it is tempting to define the flux operator $F_{K,\sigma}\left(\boldsymbol{B}_K\right)$ by setting:

$$F_{K,\sigma}\left(\boldsymbol{B}_K\right) = -|\sigma| \lambda_K \nabla_K\left(\boldsymbol{B}_K\right) \cdot \boldsymbol{n}_{K,\sigma}.$$

As consistency of the flux will of course rely on the consistency of the gradient reconstruction operators, the above definition will indeed lead to consistent fluxes, however to construct our generalized MFD algorithms conservativity will play a crucial role. This means that we will require the conservativity of the flux:

$$\sum_{K \in \mathcal{T}_\sigma} F_{K,\sigma}\left(\boldsymbol{B}_K\right) = 0$$

which cannot be satisfied in general (except by the two-point fluxes) with such a simple formula. This is the reason why, following the ideas of [4], we introduce a stabilized gradient operator $\nabla_{K,\sigma} : X_K \times \mathbb{R} \longmapsto \mathbb{R}$ by setting:

$$\nabla_{K,\sigma}\left(\boldsymbol{B}_K, \rho_\sigma\right) = \nabla_K\left(\boldsymbol{B}_K\right) + \frac{\eta}{d_{K,\sigma}}\left(\left(\rho_\sigma - b_K\right) - \nabla_K\left(\boldsymbol{B}_K\right) \cdot \left(\boldsymbol{x}_\sigma - \boldsymbol{x}_K\right)\right) \boldsymbol{n}_{K,\sigma}$$

where $\eta > 0$ is a stabilization parameter. Then, for any $K \in \mathcal{T}$ and any $\sigma \in \mathcal{F}_K$, we define an intermediate flux operator $\hat{F}_{K,\sigma} : X_K \times \mathbb{R} \longmapsto \mathbb{R}$ by setting:

$$\hat{F}_{K,\sigma}\left(\boldsymbol{B}_K, \rho_\sigma\right) = -|\sigma| \lambda_K \nabla_{K,\sigma}\left(\boldsymbol{B}_K, \rho_\sigma\right) \cdot \boldsymbol{n}_{K,\sigma}.$$

Obviously, for any $\sigma \in \mathcal{F}_{\text{int}}$, we would like to define $\hat{b}_\sigma$ as the solution of

$$\sum_{K \in \mathcal{T}_\sigma} \hat{F}_{K,\sigma}\left(\boldsymbol{B}_K, \hat{b}_\sigma\right) = 0. \tag{4.1}$$

Fortunately, as λ is positive almost everywhere we immediately deduce that such a $\hat{b}_\sigma$ is always defined and is given by:

$$\hat{b}_\sigma = \frac{1}{\left(\sum_{M \in \mathcal{T}_\sigma} \frac{\eta \lambda_M}{d_{M,\sigma}}\right)} \left(\sum_{K \in \mathcal{T}_\sigma} \frac{\eta \lambda_K}{d_{K,\sigma}} b_K + \sum_{K \in \mathcal{T}_\sigma} \nabla_K\left(\boldsymbol{B}_K\right) \cdot \left(\frac{\eta \lambda_K}{d_{K,\sigma}} \left(\boldsymbol{x}_\sigma - \boldsymbol{x}_K\right) - \lambda_K \boldsymbol{n}_{K,\sigma} \right) \right). \tag{4.2}$$

Thus, for any $\sigma \in \mathcal{F}_{\text{int}}$, it is legitimate to define the conservative flux operator $F_{K,\sigma} : \times_{L \in \mathcal{T}_\sigma} X_L \longmapsto \mathbb{R}$ by setting:

$$F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) = \hat{F}_{K,\sigma}\left(\boldsymbol{B}_K, \hat{b}_\sigma\right) \tag{4.3}$$

where $\hat{b}_\sigma$ is the unique solution of (4.1), while for any $\sigma \in \mathcal{F}_{\text{ext}}$, we simply set:

$$F_{K,\sigma}(\boldsymbol{B}_\sigma) = -|\sigma|\lambda_K \nabla_K(\boldsymbol{B}_K)\cdot \boldsymbol{n}_{K,\sigma}. \tag{4.4}$$

Remark 4.1. It is tempting to recover conservativity using the following simpler formula for $\hat{F}_{K,\sigma}(\boldsymbol{B}_\sigma)$:

$$\hat{F}_{K,\sigma}(\boldsymbol{B}_\sigma) = \frac{1}{2}\left(F_{K,\sigma}(\boldsymbol{B}_K) - F_{L,\sigma}(\boldsymbol{B}_L)\right).$$

The main drawback of this alternative formula can be understood in cases where λ is discontinuous, where this would obviously lead to a very coarse approximation of the flux along the lines of discontinuity of λ. Provided those discontinuities are resolved by the mesh, our formula for $\hat{F}_{K,\sigma}(\boldsymbol{B}_\sigma)$ will instead behave as the usual harmonic mean. It is thus a more robust and versatile choice.

4.2. Consistency properties of the flux

Denoting $B(\boldsymbol{x},\xi)$ the ball centered at $\boldsymbol{x}$ of radius ξ and $B_\Omega(\boldsymbol{x},\xi) = B(\boldsymbol{x},\xi)\cap\overline{\Omega}$, we recall the usual definition of strong consistency for gradient reconstruction operators:

Definition 4.2 (Strong consistency)**.** The gradient reconstruction operator ∇_K is strongly consistent if and only if there exists $C>0$ independent on h such that for any $\varphi \in C^2\left(\overline{B_\Omega(\boldsymbol{x}_K,\xi)}\right)$, every $0<\xi\leqslant 2h_K$ and every $\boldsymbol{x}\in B_\Omega(\boldsymbol{x}_K,\xi)$:

$$|\nabla\varphi(\boldsymbol{x}) - \nabla_K(\boldsymbol{\mathcal{D}}_K(\varphi))| \leqslant C\xi||\varphi||_{W^{2,\infty}(B_\Omega(\boldsymbol{x}_K,\xi)\cap\Omega)}.$$

As an immediate consequence of the above definition, we have, using Taylor's expansion and the density of $C^\infty\left(\overline{B_\Omega(\boldsymbol{x}_K,\xi)}\right)$ in $W^{2,\infty}(B_\Omega(\boldsymbol{x},\xi))$ that for any $\varphi\in W^{2,\infty}(\Omega)$ there exists $C>0$ depending on φ but not on h such that for every $h_K\leqslant\xi\leqslant 2h_K$, any $K\in\mathcal{T}$ and almost every $\boldsymbol{x}\in B_\Omega(\boldsymbol{x}_K,\xi)$:

$$|\nabla\varphi(\boldsymbol{x}) - \nabla_K(\boldsymbol{\mathcal{D}}_K(\varphi))| \leqslant C\xi||\varphi||_{W^{2,\infty}(B_\Omega(\boldsymbol{x}_K,\xi))}$$

and

$$|\varphi(\boldsymbol{x}) - \Pi_K(\boldsymbol{\mathcal{D}}_K(\varphi))(\boldsymbol{x})| \leqslant C\xi^2||\varphi||_{W^{2,\infty}(B_\Omega(\boldsymbol{x}_K,\xi))}.$$

Another immediate consequence is the consistency of our discrete fluxes:

Proposition 4.3. *Assume that assumption (A1) is satisfied and that the gradient reconstruction operator is strongly consistent. Then, there exists $C>0$ independent on h such that for any $K\in\mathcal{T}$, any $\sigma\in\mathcal{F}_K$, any $b\in W^{2,\infty}(\Omega)$ and any $\lambda\in W^{1,\infty}(\Omega)$, we have:*

$$\left|\left|\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right|\right|_{L^\infty(B_\Omega(\boldsymbol{x}_K,2h_K))} \leqslant Ch_K||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{2,\infty}(\Omega)}. \tag{4.5}$$

Proof. By density of $C^\infty\left(\overline{\Omega}\right)$ in $W^{1,\infty}(\Omega)$ and $W^{2,\infty}(\Omega)$, it is clear that it suffices to establish the result for $\lambda\in C^\infty\left(\overline{\Omega}\right)$ and $b\in C^\infty\left(\overline{\Omega}\right)$. Assuming such regularity, we start by establishing that there exists $C>0$ independent on h such that for any $\sigma\in\mathcal{F}_{\text{int}}$:

$$|\hat{b}_\sigma - b(\boldsymbol{x}_\sigma)| \leqslant Ch_K^2||b||_{W^{2,\infty}(\Omega)} \quad\text{and}\quad |\hat{b}_\sigma - \Pi_K(\boldsymbol{B}_K)(\boldsymbol{x}_\sigma)| \leqslant Ch_K^2||b||_{W^{2,\infty}(\Omega)} \quad \forall K\in\mathcal{T}_\sigma. \tag{4.6}$$

As by construction, we have $\Pi_K(\boldsymbol{B}_K)(\boldsymbol{x}_\sigma) = b_K + \nabla_K(\boldsymbol{B}_K)\cdot(\boldsymbol{x}_\sigma - \boldsymbol{x}_K)$, slightly rewriting (4.2) we get:

$$\hat{b}_\sigma - b(\boldsymbol{x}_\sigma) = \frac{1}{\left(\sum_{M\in\mathcal{T}_\sigma}\frac{\eta\lambda_M}{d_{M,\sigma}}\right)}\left(\sum_{L\in\mathcal{T}_\sigma}\frac{\eta\lambda_L}{d_{L,\sigma}}\left(\Pi_L(\boldsymbol{B}_L)(\boldsymbol{x}_\sigma) - b(\boldsymbol{x}_\sigma)\right) - \sum_{L\in\mathcal{T}_\sigma}\lambda_L\nabla_L(\boldsymbol{B}_L)\cdot\boldsymbol{n}_{L,\sigma}\right).$$

Immediately, using the strong consistency of the gradient operator we get that for all $K \in \mathcal{T}_\sigma$:

$$|\Pi_K(\boldsymbol{B}_K)(\boldsymbol{x}_\sigma) - b(\boldsymbol{x}_\sigma)| \leqslant Ch_K^2 ||b||_{W^{2,\infty}(B_\Omega(\boldsymbol{x}_K,h_K))}$$

and thus for all $K \in \mathcal{T}_\sigma$:

$$\frac{1}{\left(\sum_{M\in\mathcal{T}_\sigma}\frac{\eta\lambda_M}{d_{M,\sigma}}\right)}\sum_{L\in\mathcal{T}_\sigma}\frac{\eta\lambda_L}{d_{L,\sigma}}|\Pi_L(\boldsymbol{B}_L)(\boldsymbol{x}_\sigma) - b(\boldsymbol{x}_\sigma)| \leqslant C\frac{\sum_{L\in\mathcal{T}_\sigma}\frac{\eta\lambda_L}{d_{L,\sigma}}h_L^2}{\left(\sum_{M\in\mathcal{T}_\sigma}\frac{\eta\lambda_M}{d_{M,\sigma}}\right)}||b||_{W^{2,\infty}(B_\Omega(\boldsymbol{x}_K,h_K))} \leqslant Ch_K^2||b||_{W^{2,\infty}(\Omega)}$$

as under assumption (A1), we know (see [3,7]) that there exists $C > 0$ such that for any $(K, M) \in \mathcal{T}^2$ such that $\mathcal{F}_K \cap \mathcal{F}_M \neq \varnothing$, we have $h_M/h_K \leqslant C$. Moreover, as the gradient operator is strongly consistent, we know that:

$$|\lambda_K \nabla_K(\boldsymbol{B}_K) - \lambda(\boldsymbol{x}_\sigma)\nabla b(\boldsymbol{x}_\sigma)| \leqslant Ch\left(||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{1,\infty}(\Omega)} + ||\lambda||_{L^\infty(\Omega)}||b||_{W^{2,\infty}(\Omega)}\right).$$

Using:

$$\sum_{K\in\mathcal{T}_\sigma}\lambda(\boldsymbol{x}_\sigma)\nabla b(\boldsymbol{x}_\sigma)\cdot\boldsymbol{n}_{K,\sigma} = 0$$

we get:

$$\left|\sum_{K\in\mathcal{T}_\sigma}\lambda_K\nabla_K(\boldsymbol{B}_K)\cdot\boldsymbol{n}_{K,\sigma}\right| \leqslant Ch\left(||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{1,\infty}(\Omega)} + ||\lambda||_{L^\infty(\Omega)}||b||_{W^{2,\infty}(\Omega)}\right).$$

Meanwhile, we clearly get using again assumption (A1):

$$\frac{1}{\left(\sum_{M\in\mathcal{T}_\sigma}\frac{\eta\lambda_M}{d_{M,\sigma}}\right)} \leqslant \frac{1}{\eta\lambda_-}\frac{1}{\left(\sum_{M\in\mathcal{T}_\sigma}\frac{1}{d_{M,\sigma}}\right)} \leqslant \frac{1}{\eta\lambda_-\operatorname{card}(\mathcal{T}_\sigma)}\max_{M\in\mathcal{T}_\sigma}d_{M,\sigma} \leqslant \frac{1}{\eta\lambda_-}\max_{M\in\mathcal{T}_\sigma}h_M \leqslant \frac{Ch_K}{\eta\lambda_-}$$

which finally establishes that:

$$|\hat{b}_\sigma - b(\boldsymbol{x}_\sigma)| \leqslant Ch_K^2||b||_{W^{2,\infty}(\Omega)}$$

and (4.6) immediately follows using the triangular inequality and the strong consistency of the gradient operator. Then, for any $\boldsymbol{x} \in B_\Omega(\boldsymbol{x}_K, 2h_K)$ and any $\sigma \in \mathcal{F}_{\text{int}}$:

$$\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda(\boldsymbol{x})\nabla b(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma} = \lambda_K(\nabla b(\boldsymbol{x}) - \nabla_K(\boldsymbol{B}_K))\cdot\boldsymbol{n}_{K,\sigma} + (\lambda(\boldsymbol{x}) - \lambda_K)\nabla b(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}$$
$$- \frac{\eta\lambda_K}{d_{K,\sigma}}\left(\hat{b}_\sigma - \Pi_K(\boldsymbol{B}_K)(\boldsymbol{x}_\sigma)\right).$$

Immediately, as the gradient reconstruction operator is strongly consistent we get:

$$|\lambda_K(\nabla b(\boldsymbol{x}) - \nabla_K(\boldsymbol{B}_K))\cdot\boldsymbol{n}_{K,\sigma}| \leqslant C\lambda_+ h_K||b||_{W^{2,\infty}(\Omega)}$$

while Taylor's expansion immediately gives:

$$|(\lambda(\boldsymbol{x}) - \lambda_K)\nabla b(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}| \leqslant Ch_K||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{1,\infty}}.$$

Under assumption (A1), we have $h_K/d_{K,\sigma} \leqslant h_K/\rho_K \leqslant 1/\rho$ which combined with (4.6) gives the desired result for $\sigma \in \mathcal{F}_{\text{int}}$. For $\sigma \in \mathcal{F}_{\text{ext}}$, we directly get:

$$\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda(\boldsymbol{x})\nabla b(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma} = \lambda_K(\nabla b(\boldsymbol{x}) - \nabla_K(\boldsymbol{B}_K))\cdot\boldsymbol{n}_{K,\sigma} + (\lambda(\boldsymbol{x}) - \lambda_K)\nabla b(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}$$

and thus applying the same estimates than in the case of interior faces concludes the proof of (4.5). □

5. A multiple flow direction algorithm on general meshes

5.1. The generalized MFD algorithm

Given an approximation $(\boldsymbol{B}_K)_{K\in\mathcal{T}}$ of the topography and $(f_K)_{K\in\mathcal{T}}$ of the source term as data, and using both upwinding for the water height and our discrete fluxes, the most natural discrete formulation of the Manning–Strickler model consists in finding $\left((u_K)_{K\in\mathcal{T}}, (u_\sigma)_{\sigma\in\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}}}\right)$ such that:

$$\left|\begin{array}{ll} \displaystyle\sum_{\sigma\in\mathcal{F}_K} u_\sigma^{\mathrm{up}} F_{K,\sigma}(\boldsymbol{B}_\sigma) = |K| f_K & \forall K\in\mathcal{T} \\ u_\sigma = 0 & \forall \sigma\in\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}} \end{array}\right. \tag{5.1}$$

where the upwinding formula v_σ^{up} for any set of cell values $(v_K)_{K\in\mathcal{T}}$ is defined by:

$$v_\sigma^{\mathrm{up}} = \left|\begin{array}{lll} v_K & \text{if} & F_{K,\sigma}(\boldsymbol{B}_\sigma) \geqslant 0 \\ v_L & \text{if} & F_{K,\sigma}(\boldsymbol{B}_\sigma) < 0 \end{array}\right. \quad \forall \sigma\in\mathcal{F}_{\mathrm{int}},\ \mathcal{T}_\sigma = \{K,L\}$$

and

$$v_\sigma^{\mathrm{up}} = \left|\begin{array}{lll} v_K & \text{if} & F_{K,\sigma}(\boldsymbol{B}_\sigma) \geqslant 0 \\ 0 & \text{if} & F_{K,\sigma}(\boldsymbol{B}_\sigma) < 0 \end{array}\right. \quad \forall \sigma\in\mathcal{F}_{\mathrm{ext}},\ \mathcal{T}_\sigma = \{K\}$$

and where we have defined the influx boundary associated with the discrete fluxes by setting

$$\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}} = \{\sigma\in\mathcal{F}_{\mathrm{ext}} \mid F_{K,\sigma}(\boldsymbol{B}_\sigma) < 0,\ \mathcal{T}_\sigma = \{K\}\}.$$

By construction of the fluxes, for any $\sigma\in\mathcal{F}_{\mathrm{int}}$ we have that $F_{L,\sigma}(\boldsymbol{B}_\sigma) = -F_{K,\sigma}(\boldsymbol{B}_\sigma)$. Mimicking what we have done for the TPFA scheme, we gather faces by upwinding kind and use the fact that $u_\sigma^{\mathrm{up}} = 0$ for all $\sigma\in\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}}$, leading to:

$$\left(\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)\geqslant 0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\right) u_K - \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{int}}, F_{L,\sigma}(\boldsymbol{B}_\sigma)>0} u_L F_{L,\sigma}(\boldsymbol{B}_\sigma) = |K| f_K.$$

Defining:

$$s_K = \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)\geqslant 0} F_{K,\sigma}(\boldsymbol{B}_\sigma) \quad \text{and} \quad \widetilde{q}_K = s_K u_K$$

and noticing that $s_L > 0$ as soon as there exists $\sigma\in\mathcal{F}_L$ such that $F_{L,\sigma}(\boldsymbol{B}_\sigma) > 0$, we see that the above equation can be rewritten:

$$\widetilde{q}_K - \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{int}}, F_{L,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{\widetilde{q}_L}{s_L} F_{L,\sigma}(\boldsymbol{B}_\sigma) = |K| f_K.$$

For any $\sigma\in\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}}$, we set:

$$s_\sigma = -\sum_{K\in\mathcal{T}_\sigma, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma) \quad \text{and} \quad \widetilde{q}_\sigma = s_\sigma u_\sigma.$$

Gathering all those results, we obtain:

$$\left|\begin{array}{ll} \widetilde{q}_K - \displaystyle\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{int}}, F_{L,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{\widetilde{q}_L}{s_L} F_{L,\sigma}(\boldsymbol{B}_\sigma) = |K| f_K & \forall K\in\mathcal{T} \\ \widetilde{q}_\sigma = 0 & \forall \sigma\in\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}}. \end{array}\right. \tag{5.2}$$

Notice that the above system is set for the unknowns $\left(\left(\widetilde{q}_K\right)_{K\in\mathcal{T}}, \left(\widetilde{q}_\sigma\right)_{\sigma\in\mathcal{F}^{\mathcal{D}}_{\text{ext,in}}}\right)$, thus it has the same number of unknowns than the original system (5.1). We still reconstruct a water flux vector cellwise by setting:

$$\boldsymbol{Q}_K = \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{\widetilde{q}_K}{|K|s_K} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\boldsymbol{x}_\sigma - \boldsymbol{x}_K\right) + \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{\widetilde{q}_L}{|K|s_L} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\boldsymbol{x}_\sigma - \boldsymbol{x}_K\right) \quad (5.3)$$

the consistent water discharge being given by $q_K = ||\boldsymbol{Q}_K||$. Finally, for any cell $K \in \mathcal{T}$, we set u_K as we have done for the TPFA case:

$$u_K = \begin{vmatrix} \dfrac{\widetilde{q}_K}{s_K} & \text{if} \quad s_K \neq 0 \\ 0 & \text{otherwise} \end{vmatrix} \quad (5.4)$$

while for any $\sigma \in \mathcal{F}^{\mathcal{D}}_{\text{ext,in}}$, $u_\sigma = 0$. Notice that this new system allows to correctly define the numerical method, while the system (5.1) does not uniquely define the water height on cells where $s_K = 0$. Contrary to system (3.4), system (5.2) admits no obvious renumbering of mesh elements that makes the system triangular. However, as it is anyway necessary to enable parallelism, we say that the generalized MFD algorithm will consists in solving the linear system (5.2) using any linear solver of the literature. Notice that as we are using the unknown $\widetilde{q}$ rather than u, the discrete system (5.2) takes the form of a perturbation of the identity matrix. Thus, we expect its condition number to be relatively good (depending of course of the coefficient λ) and in any case much better than if had tried to solve (5.2) directly.

5.2. On existence and uniqueness for the generalized MFD algorithm

Existence and uniqueness of a solution to the above system are much less obvious than in the case of the TPFA scheme. To explain this fact, let us denote

$$\mathcal{T}^* = \{K \in \mathcal{T} \mid s_K > 0\}.$$

Using the definition of $\mathcal{T}^*$, summing (5.2) over $K \in \mathcal{T}$, after a straightforward manipulation of the last sum we get:

$$\sum_{K\in\mathcal{T}\backslash\mathcal{T}^*} \widetilde{q}_K + \sum_{K\in\mathcal{T}^*} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{\widetilde{q}_K}{s_K} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) - \sum_{L\in\mathcal{T}^*} \sum_{\sigma\in\mathcal{F}_L\cap\mathcal{F}_{\text{int}}, F_{L,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{\widetilde{q}_L}{s_L} F_{L,\sigma}\left(\boldsymbol{B}_\sigma\right) = \sum_{K\in\mathcal{T}} |K| f_K$$

and consequently:

$$\sum_{K\in\mathcal{T}\backslash\mathcal{T}^*} \widetilde{q}_K + \sum_{K\in\mathcal{T}^*} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{\widetilde{q}_K}{s_K} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) = \sum_{K\in\mathcal{T}} |K| f_K. \quad (5.5)$$

Thus, the existence of a solution for any second member $(f_K)_{K\in\mathcal{T}}$ requires that:

$$\mathcal{A} = (\mathcal{T}\backslash\mathcal{T}^*) \cup \mathcal{T}^{**} \neq \varnothing \quad \text{where} \quad \mathcal{T}^{**} = \{K \in \mathcal{T}^* \mid \{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{ext}} \mid F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) > 0\} \neq \varnothing\}. \quad (5.6)$$

In the case of the TPFA scheme, condition (5.6) is necessarily satisfied. If condition (5.6) is not satisfied, then (5.2) can have a solution only if

$$\sum_{K\in\mathcal{T}} |K| f_K = 0.$$

From this observation, it seems clear that establishing an exhaustive existence and uniqueness theory for system (5.2) is a more complex task than what one could expect. For the continuous problem, well-posedness is directly linked to the properties of the topography b and in particular to its Laplacian. Unsurprisingly, well-posedness of (5.2) will be directly linked to the properties of the quantity that plays the role of the topography b at the

discrete level, that is to say the flux family $(F_{K,\sigma}(\boldsymbol{B}_\sigma))_{K\in\mathcal{T},\sigma\in\mathcal{F}_K}$. Without any further assumption on the mesh, far from the asymptotic regime consistency alone cannot be expected to ensure well-posedness for coarse meshes. This means that well-posedness will in general be controlled by the structure of the discrete flux family rather than by the properties of b. As in practice the Laplacian of b cannot be computed in general, it is anyway better to have an existence result that uses only computable quantities. In fact, not only the necessary condition (5.6) must be satisfied, but it also requires that there does not exist what we will call a flux cycle. For any subset $\mathcal{C}$ of $\mathcal{T}$, we denote:

$$\mathcal{F}_{\text{int}}^{\mathcal{C}} = \{\sigma \in \mathcal{F}_{\text{int}} \mid \mathcal{T}_\sigma \subset \mathcal{C}\}.$$

We say that a set of cells $\mathcal{C} \subset \mathcal{T}$ form a flux cycle if and only if for any $K \in \mathcal{C}$

$$F_{K,\sigma}(\boldsymbol{B}_\sigma) \leqslant 0 \quad \forall \sigma \in \mathcal{F}_K \backslash \left(\mathcal{F}_K \cap \mathcal{F}_{\text{int}}^{\mathcal{C}}\right)$$

$$\text{and} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}}^{\mathcal{C}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}(\boldsymbol{B}_\sigma) > 0 \quad \text{and} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}}^{\mathcal{C}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma) < 0.$$

The first condition implies that water can enter the cycle but not leave it, while the second condition implies that water will run through the cycle without stopping anywhere. Whether such configurations can exist on coarse meshes for consistent flux families is a difficult question. However, from the physical point of view flux cycles are completely unrealistic as they represent regions where water will at some point go from a topographically low region to a topographically higher one. Moreover, under the positivity hypothesis on the Laplacian of the topography, we know that for the exact fluxes we have:

$$\sum_{\sigma\in\mathcal{F}_K} \int_\sigma -\lambda \nabla b \cdot \boldsymbol{n}_{K,\sigma} > 0$$

and thus no flux cycle can exist. Thus, the presence of flux cycles should be considered as anomalous and a sign that a too coarse mesh has been used. A natural sufficient condition of existence and uniqueness for (5.2) should then be one that is satisfied by the continuous fluxes and immediately implies that no flux cycle exist.

Proposition 5.1. *We say that there exists a flowing path from $K \in \mathcal{T}\backslash\mathcal{A}$ to $\mathcal{A}$ if there exists $\widetilde{K} \in \mathcal{A}$ such that*

$$\exists\, n_K \in \mathbb{N},\ n_K > 0 \quad \text{and} \quad (\sigma_i)_{0\leqslant i\leqslant n_K-1} \quad \text{such that} \quad \sigma_i \in \mathcal{F}_{\text{int}} \quad \forall 0 \leqslant i \leqslant n_K - 1$$

$$\text{and} \quad \mathcal{T}_{\sigma_i} = \{K_i, K_{i+1}\} \quad \forall 0 \leqslant i \leqslant n_K - 1 \quad \text{where} \quad K = K_0 \quad \widetilde{K} = K_{n_K} \quad \text{and} \quad F_{K_i,\sigma_i}(\boldsymbol{B}_{\sigma_i}) > 0.$$

If $\mathcal{A} \neq \varnothing$ and for any cell $K \in \mathcal{T}\backslash\mathcal{A}$ there exists a flowing path to $\mathcal{A}$, then (5.2) is well-posed.

Proof. As system (5.2) is linear, it suffices to establish uniqueness to ensure the existence of solutions. Thus, assume that $\widetilde{\boldsymbol{q}} = \left((\widetilde{q}_K)_{K\in\mathcal{T}}, (\widetilde{q}_\sigma)_{\sigma\in\mathcal{F}_{\text{ext,in}}^{\mathcal{D}}}\right)$ is solution of (5.2) with zero right-hand side. Then, taking the modulus of (5.2) and summing over $K \in \mathcal{T}$ we get:

$$\sum_{K\in\mathcal{T}} |\widetilde{q}_K| \leqslant \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}}, F_{L,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{|\widetilde{q}_L|}{s_L} F_{L,\sigma}(\boldsymbol{B}_\sigma). \tag{5.7}$$

Again, we have:

$$\sum_{K\in\mathcal{T}} |\widetilde{q}_K| = \sum_{K\in\mathcal{T}\backslash\mathcal{T}^*} |\widetilde{q}_K| + \sum_{K\in\mathcal{T}^*} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{|\widetilde{q}_K|}{s_K} F_{K,\sigma}(\boldsymbol{B}_\sigma) + \sum_{K\in\mathcal{T}^*} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{|\widetilde{q}_K|}{s_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)$$

while:

$$\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}}, F_{L,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{|\widetilde{q}_L|}{s_L} F_{L,\sigma}(\boldsymbol{B}_\sigma) = \sum_{L\in\mathcal{T}^*} \sum_{\sigma\in\mathcal{F}_L\cap\mathcal{F}_{\text{int}}, F_{L,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{|\widetilde{q}_L|}{s_L} F_{L,\sigma}(\boldsymbol{B}_\sigma).$$

Injecting this into (5.7), we obtain:

$$\sum_{K\in\mathcal{T}\backslash\mathcal{T}^*}|\widetilde{q}_K|+\sum_{K\in\mathcal{T}^*}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}\frac{|\widetilde{q}_K|}{s_K}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\leqslant 0$$

which immediately implies that $\widetilde{q}_K=0$ for any $K\in\mathcal{A}$. Next, let us denote $\mathcal{T}_0=\mathcal{T}$ and $\mathcal{A}_0=\mathcal{A}$, and define $\mathcal{T}_i$ for any $i\geqslant 1$ by setting:

$$\mathcal{T}_i=\mathcal{T}\backslash\bigcup_{0\leqslant k\leqslant i-1}\mathcal{A}_k\quad\text{where}\quad\mathcal{A}_i=(\mathcal{T}_i\backslash\mathcal{T}_i^*)\cup\mathcal{T}_i^{**}$$

with

$$\mathcal{T}_i^*=\{K\in\mathcal{T}_i\mid s_K>0\}\quad\text{and}\quad\mathcal{T}_i^{**}=\left\{K\in\mathcal{T}_i^*\mid\left\{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext}^i\mid F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)>0\right\}\neq\varnothing\right\}$$

and for any $i\geqslant 1$:

$$\mathcal{F}_i=\{\sigma\in\mathcal{F}\mid\mathcal{T}_\sigma\cap\mathcal{T}_i\neq\varnothing\}\quad\text{and}\quad\mathcal{F}_{\rm int}^i=\{\sigma\in\mathcal{F}_i\cap\mathcal{F}_{\rm int}\mid\mathcal{T}_\sigma\subset\mathcal{T}_i\}\quad\text{and}\quad\mathcal{F}_{\rm ext}^i=\mathcal{F}_i\backslash\mathcal{F}_{\rm int}^i.$$

Clearly, as $\mathcal{A}\neq\varnothing$ we have $\mathcal{T}_1\subsetneq\mathcal{T}_0$ or $\mathcal{T}_1=\varnothing$. We now proceed by induction. Assume that for $n\geqslant 1$, we have established that

$$\widetilde{q}_K=0\quad\text{for all }K\in\bigcup_{0\leqslant k\leqslant n-1}\mathcal{A}_k\quad\text{and}\quad\mathcal{T}_i\subsetneq\mathcal{T}_{i-1}\quad\text{or}\quad\mathcal{T}_i=\varnothing\quad\forall 1\leqslant i\leqslant n-1.$$

If $\mathcal{T}_n=\varnothing$, then there is nothing to prove. Otherwise, as for any $K\in\mathcal{T}_n$, there exists a flow path to $\mathcal{A}$, then $\mathcal{T}_n^{**}\neq\varnothing$. Thus, summing over $\mathcal{T}_n$ and proceeding as above we obtain that:

$$\sum_{K\in\mathcal{T}_n\backslash\mathcal{T}_n^*}|\widetilde{q}_K|+\sum_{K\in\mathcal{T}_n^*}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext}^n,F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}\frac{|\widetilde{q}_K|}{s_K}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\leqslant 0.$$

Thus, we obtain that $\widetilde{q}_K=0$ for any $K\in\mathcal{A}_n$ and $\mathcal{A}_n\neq\varnothing$ which implies $\mathcal{T}_{n+1}\subsetneq\mathcal{T}_n$. Then, it is clear that the sequence $(\mathcal{T}_i)_{i\geqslant 0}$ is strictly decreasing in the sense that it satisfies $\mathcal{T}_i\subsetneq\mathcal{T}_{i-1}$ or $\mathcal{T}_i=\varnothing$ because of the flow path assumption. Thus, there exists $n_{\mathcal{T}}>0$ such that $\mathcal{T}_{n_{\mathcal{T}}}=\varnothing$ and thus $\widetilde{q}=0$, which concludes the proof. □

6. Convergence analysis

The main idea underlying our convergence analysis is that for all consistent fluxes, if the topography b is regular enough numerical fluxes will converge to the continuous ones. The major originality of the present analysis regarding for instance finite volumes or discontinuous Galerkin methods for steady-advection diffusion problems of the form (3.2) is that here the coefficients $\boldsymbol{\beta}$ and μ are approximated through discrete differential operators applied to the topography b. Thus, this additional approximation has to be handled carefully to recover the correct convergence estimates.

To establish precise error estimates we first need to choose a discrete norm. To this end, we define

$$\omega_{K,\sigma}=\frac{1}{2}\left(|F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)|+|\sigma|\,||\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}||_{L^\infty(B(\boldsymbol{x}_K,2h_K))}\right)$$

and set for $v=(v_K)_{K\in\mathcal{T}}$:

$$||v||_h^2=\frac{\alpha\lambda_-}{2}\sum_{K\in\mathcal{T}}|K|v_K^2+\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}\omega_{K,\sigma}|v_K|^2+\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\omega_{K,\sigma}(v_K-v_\sigma^{\rm up})^2.$$

In the above expression, one would more naturally expect to find $F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)$ instead of $\omega_{K,\sigma}$. However, their behavior is difficult to control and in particular, it is complex to estimate the way this fluxes stay away from zero

when the continuous fluxes $\int_\sigma \lambda \nabla b \cdot \boldsymbol{n}_{K,\sigma}$ cancel creating technical difficulties, which $\omega_{K,\sigma}$ avoids. Immediately, we get that:

$$|\omega_{K,\sigma} - |F_{K,\sigma}(\boldsymbol{B}_\sigma)|\,| = \frac{1}{2}\big|\,|F_{K,\sigma}(\boldsymbol{B}_\sigma) - |\sigma|\,||\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}||_{L^\infty(B(\boldsymbol{x}_K,2h_K))}\big|$$

and thus:

$$|\omega_{K,\sigma} - |F_{K,\sigma}(\boldsymbol{B}_\sigma)|\,| \leqslant \frac{1}{2}R_{K,\sigma} \tag{6.1}$$

where for any $K \in \mathcal{T}$ and any $\sigma \in \mathcal{F}_K$ we have denoted:

$$R_{K,\sigma} = \max\left(\left|F_{K,\sigma}(\boldsymbol{B}_\sigma) + \int_\sigma \lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right|, \big||F_{K,\sigma}(\boldsymbol{B}_\sigma)| - |\sigma|\,||\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}||_{L^\infty(B(\boldsymbol{x}_K,2h_K))}\big|\right). \tag{6.2}$$

Notice that from (4.5), we get that there exists $C > 0$ such that for any $K \in \mathcal{T}$ and any $\sigma \in \mathcal{F}_K$, we have:

$$R_{K,\sigma} \leqslant C|\sigma|h_K||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{2,\infty}(\Omega)}. \tag{6.3}$$

The main objective of this section is to establish the following explicit L^2 error estimate:

Theorem 6.1. *Assume that the gradient operator underlying the fluxes is consistent. Further assume that the mesh satisfies assumption (A1), that $\lambda \in W^{1,\infty}(\Omega)$ and that the solution $\overline{u}$ of (3.1) belongs to $H^1(\Omega)$. Let*

$$\gamma(h) = \max\left(\sup_{K\in\mathcal{T},\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\frac{R_{K,\sigma}}{\omega_{K,\sigma}}, \sup_{K\in\mathcal{T},\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext},F_{K,\sigma}(\boldsymbol{B}_\sigma)\neq 0}\frac{R_{K,\sigma}}{\omega_{K,\sigma}}\right)$$

with $R_{K,\sigma}$ defined by (6.2). Then, there exists $\mathcal{I}_h(\overline{u}) \in L^2(\Omega)$ such that for any $K \in \mathcal{T}$, $\mathcal{I}_h(\overline{u})|_K = \mathcal{I}_K(\overline{u})$ is constant on K and:

$$|\overline{u} - \mathcal{I}_K(\overline{u})|^2_{L^2(K)} + h_K^{1/2}|\overline{u} - \mathcal{I}_K(\overline{u})|^2_{L^2(\partial K)} \leqslant C_{\rm poly}h_K|\overline{u}|_{H^1(K)}.$$

Moreover if $\gamma(h) \to 0$ when $h \to 0$, then there exists $h_0 > 0$ and $C > 0$ depending on $\overline{u}$, λ, b and the mesh parameters such that for any $h \leqslant h_0$:

$$||u - \mathcal{I}_h(\overline{u})||_h \leqslant C\max(h,\gamma(h))^{1/2}||\overline{u}||_{H^1(\Omega)} \tag{6.4}$$

where u is reconstructed by (5.4) from the solution of (5.2).

Notice that the existence of $\mathcal{I}_h(\overline{u})$ in the above result is an immediate consequence of assumption (A1) and (2.3). As a corollary, we will then be able to establish an explicit error estimate for the water discharge:

Corollary 6.2. *Under the assumptions of Proposition 6.1, there exists $h_0 > 0$ and $C > 0$ depending on $\overline{u}$, λ, b and the mesh parameters such that for any $h \leqslant h_0$:*

$$\left(\sum_{K\in\mathcal{T}}||-\lambda\overline{u}\nabla b - \boldsymbol{Q}_K||^2_{L^2(K)^2}\right)^{1/2} \leqslant C\max(h,\gamma(h))^{1/2}(1+\max(h,\gamma(h))^{1/2})||\overline{u}||_{H^1(\Omega)}. \tag{6.5}$$

The error estimates use the surprising factor $\gamma(h)$ instead of h as one could legitimately expect. This comes from the fact that we use approximate fluxes, which is equivalent in some sense to using a quadrature formula for the coefficients of (3.2). As the natural norm for the discrete problem uses those approximate fluxes, it seems unfortunately unavoidable that this impacts the error estimate. In the case where for any $K \in \mathcal{T}$ and any $\sigma \in \mathcal{F}_K$ that is indeed present in the norm $||\cdot||_h$, the $\omega_{K,\sigma}$'s are bounded by below by a constant independent on h, which will happen most of the time in practice, then $\gamma(h)$ will behave as h and we retrieve a convergence at rate $h^{1/2}$. Controlling the lower bound for $\omega_{K,\sigma}$ is the reason why we have used a supremum on a set containing

more than $\overline{K}$ in the definition of $\omega_{K,\sigma}$. If this supremum is zero, most gradient reconstruction operators and in particular those described in our numerical experiments will provide a discrete gradient that aligns with the continuous one and the discrete flux will be zero too, avoiding most of the problematic cases for the asymptotic behavior of $\omega_{K,\sigma}$. In the case of exact fluxes it is known that for the water height u the optimal convergence rate to $\overline{u}$ is $h^{1/2}$, even if superconvergence at rate h is very often observed in practice, as revealed by [6,19] (see also [7]). Thus, as additional flux consistency terms could only decrease the convergence rate, it seems clear that our estimate for the water height cannot be expected to be improved in terms of order of convergence.

The proof of Theorem 6.1 being rather lengthy, we will try to decompose it as much as possible. Let us notice that in the special case where ∇b is constant, the discrete flux are exact. Then our problem corresponds to a piecewise constant version of the discontinuous Galerkin method described in [20], and we can in principle follow the steps of their proof. Denoting $||v||_h^2 = \sum_{K\in\mathcal{T}} ||v||_{K,h}^2$ with obvious notations, the first step of their proof is to establish a local error identity for $||u_K - v_K||_{K,h}$ for any v_K in $\mathbb{R}$, using the exactness of the flux. Then, taking $v_K = \mathcal{I}_K(\overline{u})$ and summing over $K \in \mathcal{T}$, they obtain a global error estimate where the residual terms only involve $\overline{u} - \mathcal{I}(\overline{u})$, which can be straightforwardly estimated using polynomial approximation results.

The general guideline of our proof is the same, however we have to endure some additional technicalities due to the non exactness of the flux. We first start by establishing the equivalent of the local error identity of [20], but keeping our approximate flux in the result. Thus, an additional residual term standing for the consistency error of the flux will appear on the right hand side of our identity. Then, to match the definition of the $||\cdot||_h$ norm, each time our estimates will involve the sum of the flux over the full set $\mathcal{F}_K$, we will replace it by the Laplacian of b, and put the difference in a residual term. For an isolated flux term, we will do the same using this time either $\omega_{K,\sigma}$ if the term is involved in the $||\cdot||_h$ norm or by the continuous flux if it is directly a residual term. Thus, when summing the local error identities over $K \in \mathcal{T}$ to obtain the global error estimate, residual terms involving the discrete counterpart of the Laplacian will be handled through a technical result described in the following subsection, while other residual terms will be estimated directly using the strong consistency of the flux. The reason for doing so is to obtain an error estimate involving only the $||\cdot||_h$ norm on the left-hand side and residual terms involving either $\overline{u} - \mathcal{I}(\overline{u})$ or flux consistency error terms on the right-hand side. Finally, each residual term will be estimated using polynomial approximation results and the consistency of the flux.

6.1. Local error identity and approximation results for the Laplacian

Let us now establish an abstract local error identity, following the lines of [20].

Lemma 6.3. *For any $K \in \mathcal{T}$, any $v_K \in \mathbb{R}$ and any $\eta = (\eta_\sigma)_{\sigma\in\mathcal{F}_K} \in L^2(\partial K)$ constant on each $\sigma \in \mathcal{F}_K$, we have:*

$$\begin{aligned}
&\frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\,(u_K - v_K)^2 + \frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\,(u_\sigma^{\mathrm{up}} - \eta_\sigma)^2 \\
&\quad - \frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\,((u_K - v_K) - (u_\sigma^{\mathrm{up}} - \eta_\sigma))^2 + \frac{1}{2}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)\,(u_K - v_K)^2 \\
&\quad = \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}(\boldsymbol{B}_\sigma)\,(\overline{u} - v_K)\,(u_K - v_K) \\
&\quad\quad + \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}(\boldsymbol{B}_\sigma)\,(\overline{u} - \eta_\sigma)\,(u_K - v_K) \\
&\quad\quad - \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \left(\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right)\overline{u}\,(u_K - v_K)
\end{aligned} \tag{6.6}$$

where $\overline{u}$ is the solution of the continuous problem (3.1).

Proof. This proof is a direct adaptation of its counterpart in [20]. For any $w_K \in \mathbb{R}$ and any $\xi = (\xi_\sigma)_{\sigma\in\mathcal{F}_K} \in L^2(\partial K)$ constant on each $\sigma \in \mathcal{F}_K$, consider:

$$\mathcal{I}_K = -\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)(w_K-\xi_\sigma)w_K + \sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)w_K^2.$$

Using the identity $a(a-b) = \frac{1}{2}\left(a^2-b^2+(a-b)^2\right)$, we obtain:

$$\begin{aligned}\mathcal{I}_K = &\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)w_K^2 - \frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)w_K^2 + \frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\xi_\sigma^2\\ &-\frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)(w_K-\xi_\sigma)^2\end{aligned}$$

and thus:

$$\begin{aligned}\mathcal{I}_K = &\frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}(\boldsymbol{B}_\sigma)w_K^2 + \frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\xi_\sigma^2\\ &-\frac{1}{2}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)(w_K-\xi_\sigma)^2 + \frac{1}{2}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)w_K^2.\end{aligned}$$

On the other hand, from (5.1) multiplying by w_K and using the definition of u_σ^{up} we have by construction:

$$-\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)(u_K-u_\sigma^{\mathrm{up}})w_K + \sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)u_K w_K = \int_K f w_K.$$

Let us now set $w_K = u_K - v_K$ and $\xi_\sigma = u_\sigma^{\mathrm{up}} - \eta_\sigma$ for all $\sigma \in \mathcal{F}_K$. We obtain:

$$\mathcal{I}_K = -\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)(\eta_\sigma-v_K)w_K + \int_K f w_K - \sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)v_K w_K.$$

Using $-\mathrm{div}\,(\overline{u}\lambda\nabla b) = f$, this leads to:

$$\mathcal{I}_K = -\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)(\eta_\sigma-v_K)w_K + \int_K -\mathrm{div}\,(\overline{u}\lambda\nabla b)w_K - \sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)v_K w_K.$$

Integrating by parts, we get:

$$\begin{aligned}\int_K -\mathrm{div}\,(\overline{u}\lambda\nabla b)w_K &= -\sum_{\sigma\in\mathcal{F}_K}\int_\sigma \lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\overline{u}w_K\\ &= \sum_{\sigma\in\mathcal{F}_K}\frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}(\boldsymbol{B}_\sigma)\overline{u}w_K - \sum_{\sigma\in\mathcal{F}_K}\int_\sigma\left(\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma)+\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right)\overline{u}w_K.\end{aligned}$$

Then:

$$\begin{aligned}\mathcal{I}_K = &-\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}(\boldsymbol{B}_\sigma)(\eta_\sigma-v_K)w_K + \sum_{\sigma\in\mathcal{F}_K}\frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}(\boldsymbol{B}_\sigma)(\overline{u}-v_K)w_K\\ &-\sum_{\sigma\in\mathcal{F}_K}\int_\sigma\left(\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma)+\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right)\overline{u}w_K\\ = &\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}(\boldsymbol{B}_\sigma)(\overline{u}-\eta_\sigma)w_K + \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}\frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}(\boldsymbol{B}_\sigma)(\overline{u}-v_K)w_K\\ &-\sum_{\sigma\in\mathcal{F}_K}\int_\sigma\left(\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma)+\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right)\overline{u}w_K\end{aligned}$$

which finally establishes the desired identity. □

In the above local error identity, one can see that the left hand side strongly looks like the K-term of the $||\cdot||_h$ norm. However, to match the $||\cdot||_h$ norm, we need to replace the sum of the fluxes over $\mathcal{F}_K$ by the Laplacian of b, as well as isolated flux by $\omega_{K,\sigma}$. If this last operation straightforwardly leads to a residual term because of (6.1), however the following lemma precise what kind of link we can expect between the discrete Laplacian and the true Laplacian:

Lemma 6.4. *For any $(w_K)_{K\in\mathcal{T}}$, one has:*

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)\, w_K = -\sum_{K\in\mathcal{T}}\int_K \lambda\Delta b w_K + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\int_\sigma \left(\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right) \times (w_K - w_\sigma^{\mathrm{up}}) + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}\int_\sigma \left(\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right) w_K.$$

Proof. Let us first notice that:

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)\, w_\sigma^{\mathrm{up}} = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\, w_\sigma^{\mathrm{up}}.$$

Indeed, using $F_{K,\sigma}(\boldsymbol{B}_\sigma) + F_{L,\sigma}(\boldsymbol{B}_\sigma) = 0$ for any $\sigma\in\mathcal{F}_{\mathrm{int}}$, $\mathcal{T}_\sigma = \{K, L\}$ we have

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{int}}} F_{K,\sigma}(\boldsymbol{B}_\sigma)\, w_\sigma^{\mathrm{up}} = 0$$

while for $\sigma\in\mathcal{F}_{\mathrm{ext}}$ such that $F_{K,\sigma}(\boldsymbol{B}_\sigma) > 0$, $\mathcal{T}_\sigma = \{K\}$, $w_\sigma^{\mathrm{up}} = 0$ immediately implies that:

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\, w_\sigma^{\mathrm{up}} = 0.$$

Thus, we have:

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)\, w_K = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\,(w_K - w_\sigma^{\mathrm{up}}) + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}(\boldsymbol{B}_\sigma)\, w_K$$

using the definition of w_σ^{up}. In the same way:

$$\sum_{K\in\mathcal{T}}\int_K -\lambda\Delta b w_K = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K}\int_\sigma -\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma} w_K = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\int_\sigma -\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}(w_K - w_\sigma^{\mathrm{up}}) + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}\int_\sigma -\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma} w_K.$$

Combining those expressions gives the expected result. □

6.2. Global error estimate

From the local error identity, and the link between the discrete and continuous Laplacians, we deduce a global estimate for the discrete norm of the error:

Lemma 6.5. *Using the hypothesis and notations of Proposition 6.1, we have:*

$$
\begin{aligned}
||u-\mathcal{I}_h(\overline{u})||_h^2 \leqslant & \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm int},F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\overline{u}-\mathcal{I}_\sigma^{\rm up}(\overline{u})\right)\left((u_\sigma^{\rm up}-\mathcal{I}_\sigma^{\rm up}(\overline{u}))-(u_K-\mathcal{I}_K(\overline{u}))\right)\\
&+\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}\frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\overline{u}-\mathcal{I}_K(\overline{u})\right)\left(u_K-\mathcal{I}_K(\overline{u})\right)\\
&+\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext},F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\overline{u}-\mathcal{I}_\sigma^{\rm up}(\overline{u})\right)\left((u_K-\mathcal{I}_K(\overline{u}))-(u_\sigma^{\rm up}-\mathcal{I}_\sigma^{\rm up}(\overline{u}))\right)\\
&-\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K}\int_\sigma\left(\frac{1}{|\sigma|}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)+\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right)\overline{u}\left(u_K-\mathcal{I}_K(\overline{u})\right)\\
&-\frac{1}{2}\left(\sum_{K\in\mathcal{T}}\int_K\lambda\Delta b|u_K-\mathcal{I}_K(\overline{u})|^2+\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)|u_K-\mathcal{I}_K(\overline{u})|^2\right)\\
&-\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\left(|F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)|-\omega_{K,\sigma}\right)\left((u_K-\mathcal{I}_K(\overline{u}))-(u_\sigma^{\rm up}-\mathcal{I}_\sigma^{\rm up}(\overline{u}))\right)^2\\
&-\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}\left(|F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)|-\omega_{K,\sigma}\right)|u_K-\mathcal{I}_K(\overline{u})|^2.
\end{aligned}
$$

Proof. We start from (6.6) with $v_K=\mathcal{I}_K(\overline{u})$ and $\eta_\sigma=\mathcal{I}_\sigma^{\rm up}(\overline{u})$. Summing over $K\in\mathcal{T}$ the second term of the left hand side of (6.6) leads to:

$$
\begin{aligned}
&\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(u_\sigma^{\rm up}-\mathcal{I}_\sigma^{\rm up}(\overline{u})\right)^2\\
&\quad=-\frac{1}{2}\sum_{L\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_L\cap\mathcal{F}_{\rm int},F_{L,\sigma}(\boldsymbol{B}_\sigma)>0}F_{L,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(u_L-\mathcal{I}_L(\overline{u})\right)^2\\
&\qquad+\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext},F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(u_\sigma-\mathcal{I}_\sigma(\overline{u})\right)^2\\
&\quad=-\frac{1}{2}\sum_{L\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_L\cap\mathcal{F}_{\rm int},F_{L,\sigma}(\boldsymbol{B}_\sigma)>0}F_{L,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(u_L-\mathcal{I}_L(\overline{u})\right)^2
\end{aligned}
$$

as both u_σ and $\mathcal{I}_\sigma(\overline{u})$ cancels for $\sigma\in\mathcal{F}_{\rm ext,in}^{\mathcal{D}}$. Thus, summing over $K\in\mathcal{T}$ the first two terms of the left hand side of (6.6), we obtain:

$$
\begin{aligned}
&\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(h_{m,K}-\mathcal{I}_K(\overline{u})\right)^2+\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(u_\sigma^{\rm up}-\mathcal{I}_\sigma^{\rm up}(\overline{u})\right)^2\\
&\quad=\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(h_{m,K}-\mathcal{I}_K(\overline{u})\right)^2.
\end{aligned}
$$

Consequently, summing over $K \in \mathcal{T}$ the entire left hand side of (6.6) leads to:

$$\begin{aligned}\sum_{K\in\mathcal{T}} \mathcal{I}_K = &\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)|u_K - \mathcal{I}_K(\overline{u})|^2 \\ &- \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left((u_K - \mathcal{I}_K(\overline{u})) - (u_\sigma^{\mathrm{up}} - \mathcal{I}_\sigma^{\mathrm{up}}(\overline{u}))\right)^2 \\ &+ \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)|u_K - \mathcal{I}_K(\overline{u})|^2 = F_1 + F_2 + F_3.\end{aligned}$$

The first term of the above expression rewrites:

$$\begin{aligned}F_1 = &\frac{1}{2}\sum_{K\in\mathcal{T}}\int_K -\lambda\Delta b|u_K - \mathcal{I}_K(\overline{u})|^2 \\ &+ \frac{1}{2}\left(\sum_{K\in\mathcal{T}}\int_K \lambda\Delta b|u_K - \mathcal{I}_K(\overline{u})|^2 + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)|u_K - \mathcal{I}_K(\overline{u})|^2\right) = L_0 + L_1\end{aligned}$$

while the second term and third term give:

$$\begin{aligned}F_2 + F_3 = &\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \omega_{K,\sigma}\left((u_K - \mathcal{I}_K(\overline{u})) - (u_\sigma^{\mathrm{up}} - \mathcal{I}_\sigma^{\mathrm{up}}(\overline{u}))\right)^2 \\ &+ \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \omega_{K,\sigma}|u_K - \mathcal{I}_K(\overline{u})|^2 \\ &+ \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \left(|F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)| - \omega_{K,\sigma}\right)\left((u_K - \mathcal{I}_K(\overline{u})) - (u_\sigma^{\mathrm{up}} - \mathcal{I}_\sigma^{\mathrm{up}}(\overline{u}))\right)^2 \\ &+ \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \left(|F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)| - \omega_{K,\sigma}\right)|u_K - \mathcal{I}_K(\overline{u})|^2 = L_2 + L_3 + L_4 + L_5\end{aligned}$$

with obvious notations. Immediately, using the hypothesis on λ and b we have:

$$L_0 \geqslant \frac{\alpha\lambda_-}{2}\sum_{K\in\mathcal{T}}|K|\left(u_K - \mathcal{I}_K(\overline{u})\right)^2.$$

Then, noticing that the second member of the sum over $K \in \mathcal{T}$ of (6.6) can be decomposed into three terms $T_1 + T_2 + T_3$ with obvious notations and combining the above results, we have that $\sum_{K\in\mathcal{T}} \mathcal{I}_K = T_1 + T_2 + T_3$ is equivalent to $||u - \mathcal{I}_h(\overline{u})||_h^2 \leqslant T_1 + T_2 + T_3 - L_1 - L_4 - L_5$ as $L_0 + L_2 + L_3$ exactly gives the square of the $||\cdot||_h$ norm. Using the fact that both u_σ and $\mathcal{I}_\sigma(\overline{u})$ cancels for $\sigma \in \mathcal{F}_{\mathrm{ext,in}}^{\mathcal{D}}$, we obtain:

$$\begin{aligned}T_1 + T_2 = &\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{int}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\overline{u} - \mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})\right)\left((u_\sigma^{\mathrm{up}} - \mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})) - (u_K - \mathcal{I}_K(\overline{u}))\right) \\ &+ \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\overline{u} - \mathcal{I}_K(\overline{u})\right)\left(u_K - \mathcal{I}_K(\overline{u})\right) \\ &+ \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{1}{|\sigma|}\int_\sigma F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\overline{u} - \mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})\right)\left((u_K - \mathcal{I}_K(\overline{u})) - (u_\sigma^{\mathrm{up}} - \mathcal{I}_\sigma^{\mathrm{up}}(\overline{u}))\right)\end{aligned}$$

which concludes the proof. □

6.3. Estimation of the residual terms

Establishing Proposition 6.1 now just consists in estimating each term appearing in the second member of the above global estimate. From Lemma 6.5, we see that $||u-\mathcal{I}_h(\overline{u})||_h^2 \leqslant \sum_{i=1}^{7} T_i$ with obvious notations. The first three terms account for the polynomial approximation error, while the four other terms account for the consistency error of the flux. We now estimate those two families of residual terms separately.

Proof of Proposition 6.1: terms accounting for polynomial approximation error. Using Cauchy–Schwarz inequality and the fact that $(u_\sigma^{\mathrm{up}}-\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u}))-(u_K-\mathcal{I}_K(\overline{u}))$ is constant on σ gives:

$$|T_1|^2 \leqslant \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{int}},F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \left(\int_\sigma \frac{1}{|\sigma|}|F_{K,\sigma}(\boldsymbol{B}_\sigma)|\,||\overline{u}-\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})|^2\right) ||u-\mathcal{I}_h(\overline{u})||_h^2.$$

As the gradient operator underlying the fluxes is consistent, we know from Proposition 4.5 that there exists $C>0$ such that:

$$\begin{aligned}\left|\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma)\right| &\leqslant ||\frac{1}{|\sigma|}F_{K,\sigma}(\boldsymbol{B}_\sigma)-\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}||_{L^\infty(\sigma)}+||\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}||_{L^\infty(\sigma)}\\ &\leqslant Ch||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{2,\infty}(\Omega)}+||\lambda||_{L^\infty(\Omega)}||b||_{W^{2,\infty}(\Omega)}.\end{aligned}$$

As soon as $h\leqslant h_0$ for some fixed h_0 this leads to the existence of some $C>0$ independent on h such that

$$|T_1|^2 \leqslant C||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{2,\infty}} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{int}},F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \left(\int_\sigma |\overline{u}-\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})|^2\right) ||u-\mathcal{I}_h(\overline{u})||_h^2.$$

Then, using the definition of $\mathcal{I}_h(\overline{u})$, we get

$$\int_\sigma |\overline{u}-\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})|^2 \leqslant Ch_{K_\sigma^{\mathrm{up}}}|\overline{u}|^2_{H^1(K_\sigma^{\mathrm{up}})}$$

where, if $\sigma\in\mathcal{F}_{\mathrm{int}}$, $K_\sigma^{\mathrm{up}}=K$ if $F_{K,\sigma}(\boldsymbol{B}_\sigma)\geqslant 0$ and $K_\sigma^{\mathrm{up}}=L$ if $F_{K,\sigma}(\boldsymbol{B}_\sigma)<0$, and $K_\sigma^{\mathrm{up}}=K$ if $\sigma\in\mathcal{F}_{\mathrm{ext}}$. From [7], we know that assumption (A1) implies that $\mathrm{card}(\mathcal{F}_K)$ is bounded by a constant $\rho_{\mathcal{F}}$ depending on the mesh parameters but independent on h, and thus there exists $C>0$ such that $|T_1|^2\leqslant 2hC||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{2,\infty}}|\overline{u}|^2_{H^1(\Omega)}||u-\mathcal{I}_h(\overline{u})||_h^2$. For T_2, obviously we have using Cauchy–Schwarz inequality:

$$|T_2|^2 \leqslant \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \left(\int_\sigma \frac{1}{|\sigma|}|F_{K,\sigma}(\boldsymbol{B}_\sigma)|\,||\overline{u}-\mathcal{I}_K(\overline{u})|^2\right) ||u-\mathcal{I}_h(\overline{u})||_h^2$$

and thus, proceeding the same way than for T_1, we obtain $|T_2|^2\leqslant 2hC||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{2,\infty}}|\overline{u}|^2_{H^1(\Omega)}||u-\mathcal{I}_h(\overline{u})||_h^2$. The situation is more delicate for T_3. Indeed, $\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})=0$ for $\sigma\in\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}}$, while $\overline{u}=0$ on $\sigma\in\mathcal{F}_{\mathrm{ext,in}}$, thus $\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})$ is not directly a good approximation of $\overline{u}$. For any $\sigma\in\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}}\cap\mathcal{F}_{\mathrm{ext,in}}$, $\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})=\overline{u}=0$ and thus the contribution to T_3 is zero. Then, it just remains to consider the case where $\sigma\in\mathcal{F}^{\mathcal{D}}_{\mathrm{ext,in}}$ and $\sigma\notin\mathcal{F}_{\mathrm{ext,in}}$. In this case we have $-\int_\sigma\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\geqslant 0$ while $F_{K,\sigma}(\boldsymbol{B}_\sigma)<0$. However, by definition of the residual $\left|F_{K,\sigma}(\boldsymbol{B}_\sigma)+\int_\sigma\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right|\leqslant R_{K,\sigma}$ which immediately implies that $\left|\int_\sigma\lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right|\leqslant R_{K,\sigma}$ and $|F_{K,\sigma}(\boldsymbol{B}_\sigma)|\leqslant R_{K,\sigma}$. Consequently, we get using Cauchy–Schwarz inequality:

$$\begin{aligned}|T_3| \leqslant &\left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}},F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{1}{|\sigma|}R_{K,\sigma}\int_\sigma \overline{u}^2\right)^{1/2}\\ &\times\left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}},F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} R_{K,\sigma}\left((u_K-\mathcal{I}_K(\overline{u}))-(u_\sigma^{\mathrm{up}}-\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u}))\right)^2\right)^{1/2}.\end{aligned}$$

Using the definition of $\gamma(h)$ and the trace inequality on Ω, this leads to:

$$T_3 \leqslant \gamma(h)^{1/2} \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{1}{|\sigma|} R_{K,\sigma} \int_\sigma \overline{u}^2 \right)^{1/2} ||u-\mathcal{I}_h(\overline{u})||_h \leqslant Ch^{1/2}\gamma(h)^{1/2}||\overline{u}||_{H^1(\Omega)}||u-\mathcal{I}_h(\overline{u})||_h.$$

□

It finally remains to estimate the terms accounting for the consistency error on the fluxes.

Proof of Proposition 6.1: terms accounting for flux consistency error. For T_4, let us first remark that:

$$\begin{aligned} T_4 = & -\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{int}}} \int_\sigma \left(\frac{1}{|\sigma|} F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda \nabla b \cdot \boldsymbol{n}_{K,\sigma} \right) \overline{u} \left((u_K - \mathcal{I}_K(\overline{u})) - (u_\sigma^{\text{up}} - \mathcal{I}_\sigma^{\text{up}}(\overline{u})) \right) \\ & -\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{ext}}} \int_\sigma \left(\frac{1}{|\sigma|} F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda \nabla b \cdot \boldsymbol{n}_{K,\sigma} \right) \overline{u} \left(u_K - \mathcal{I}_K(\overline{u}) \right). \end{aligned}$$

Using the fact that both u_σ and $\mathcal{I}_\sigma(\overline{u})$ cancel for $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{ext}}$ and $F_{K,\sigma}(\boldsymbol{B}_\sigma) < 0$, this rewrites:

$$\begin{aligned} T_4 = & -\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \int_\sigma \left(\frac{1}{|\sigma|} F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda \nabla b \cdot \boldsymbol{n}_{K,\sigma} \right) \overline{u} \left((u_K - \mathcal{I}_K(\overline{u})) - (u_\sigma^{\text{up}} - \mathcal{I}_\sigma^{\text{up}}(\overline{u})) \right) \\ & -\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} \int_\sigma \left(\frac{1}{|\sigma|} F_{K,\sigma}(\boldsymbol{B}_\sigma) + \lambda \nabla b \cdot \boldsymbol{n}_{K,\sigma} \right) \overline{u} \left(u_K - \mathcal{I}_K(\overline{u}) \right) = T_{4,1} + T_{4,2}. \end{aligned}$$

Cauchy–Schwarz inequality then leads to:

$$\begin{aligned} |T_{4,1}| \leqslant & \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{1}{|\sigma|} R_{K,\sigma} \int_\sigma \overline{u}^2 \right)^{1/2} \\ & \times \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} R_{K,\sigma} \left((u_K - \mathcal{I}_K(\overline{u})) - (u_\sigma^{\text{up}} - \mathcal{I}_\sigma^{\text{up}}(\overline{u}))^2 \right) \right)^{1/2}. \end{aligned}$$

Using (6.3), (2.2) and the definition of $\gamma(h)$, we easily get:

$$|T_{4,1}| \leqslant C\gamma(h)^{1/2} \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} h_K \int_\sigma \overline{u}^2 \right)^{1/2} ||u-\mathcal{I}_h(\overline{u})||_h \leqslant C\gamma(h)^{1/2}||\overline{u}||_{H^1(\Omega)}||u-\mathcal{I}_h(\overline{u})||_h.$$

In the same way, Cauchy–Schwarz inequality and this time the trace inequality on Ω directly gives:

$$|T_{4,2}| \leqslant C\gamma(h)^{1/2} \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} h_K \int_\sigma \overline{u}^2 \right)^{1/2} ||u-\mathcal{I}_h(\overline{u})||_h \leqslant C\gamma(h)^{1/2}h^{1/2}||\overline{u}||_{H^1(\Omega)}||u-\mathcal{I}_h(\overline{u})||_h.$$

From Lemma 6.4, we deduce that:

$$\begin{aligned} |T_5| \leqslant & \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} R_{K,\sigma} \left((u_K - \mathcal{I}_K(\overline{u}))^2 - (u_\sigma^{\text{up}} - \mathcal{I}_\sigma^{\text{up}}(\overline{u}))^2 \right) \\ & + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\text{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} R_{K,\sigma} \left(u_K - \mathcal{I}_K(\overline{u}) \right)^2. \end{aligned}$$

Using $(a^2-b^2)=(a-b)(a+b)$ and Cauchy–Schwarz inequality, we obtain:

$$|T_5| \leqslant \gamma(h)^{1/2}||u-\mathcal{I}_h(\overline{u})||_h \left(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} R_{K,\sigma}\left((u_K-\mathcal{I}_K(\overline{u}))+(u_\sigma^{\mathrm{up}}-\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u}))\right)^2\right)^{1/2}$$
$$+\gamma(h)^{1/2}||u-\mathcal{I}_h(\overline{u})||_h \left(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} R_{K,\sigma}\left(u_K-\mathcal{I}_K(\overline{u})\right)^2\right)^{1/2}.$$

Using (6.3) and assumption (A1), we immediately obtain that:

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} R_{K,\sigma}\left(u_K-\mathcal{I}_K(\overline{u})\right)^2 \leqslant C\rho_{\mathcal{F}}\sum_{K\in\mathcal{T}}|K|\left(u_K-\mathcal{I}_K(\overline{u})\right)^2$$

and in the same way, using the fact that $u_\sigma^{\mathrm{up}}=$ and $\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})=0$ on any $\sigma\in\mathcal{F}_{\mathrm{ext}}$ such that $F_{K,\sigma}(\boldsymbol{B}_\sigma)<0$ and the commensurability of h_K and h_L if $\mathcal{F}_K\cap\mathcal{F}_K\neq\varnothing$ under assumption (A1):

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} R_{K,\sigma}\left(u_\sigma^{\mathrm{up}}-\mathcal{I}_\sigma^{\mathrm{up}}(\overline{u})\right)^2 \leqslant C\sum_{L\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_L,F_{L,\sigma}(\boldsymbol{B}_\sigma)>0}|\sigma|h_L\left(h_L-\mathcal{I}_L^{\mathrm{up}}(\overline{u})\right)^2$$
$$\leqslant C\rho_{\mathcal{F}}\sum_{K\in\mathcal{T}}|K|\left(u_K-\mathcal{I}_K(\overline{u})\right)^2$$

and also:

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} R_{K,\sigma}\left(u_K-\mathcal{I}_K(\overline{u})\right)^2 \leqslant C\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}|\sigma|h_K\left(u_K-\mathcal{I}_K(\overline{u})\right)^2$$
$$\leqslant C\rho_{\mathcal{F}}\sum_{K\in\mathcal{T}}|K|\left(u_K-\mathcal{I}_K(\overline{u})\right)^2.$$

Gathering those intermediate results, we get that $|T_5|\leqslant C\gamma(h)^{1/2}||u-\mathcal{I}_h(\overline{u})||_h^2$. Finally, we have by definition of $\gamma(h)$ and $||\cdot||_h$ that $|T_6+T_7|\leqslant C\gamma(h)^{1/2}||u-\mathcal{I}_h(\overline{u})||_h^2$. Then, using the hypothesis that $\gamma(h)$ goes to zero when h goes to zero, for any $C>0$ independent on h there exists $h_0>0$ such that $1-C\max(h,\gamma(h))^{1/2}>1/2$, and the result then immediately follows. □

6.4. Error estimate for the water discharge

To establish Corollary 6.2, we will need an auxiliary discrete stability result:

Lemma 6.6. *The following estimates holds:*

$$\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K}F_{K,\sigma}(\boldsymbol{B}_\sigma)\,u_K^2+\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}},F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}F_{K,\sigma}(\boldsymbol{B}_\sigma)\,u_K^2$$
$$-\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\left(u_K-u_\sigma^{\mathrm{up}}\right)^2F_{K,\sigma}(\boldsymbol{B}_\sigma)=\sum_{K\in\mathcal{T}}|K|f_Ku_K. \tag{6.7}$$

Proof. To establish (6.7), using the definition of u_σ^{up}, we obtain multiplying (5.1) by u_K and summing over $K\in\mathcal{T}$:

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)>0}F_{K,\sigma}(\boldsymbol{B}_\sigma)\,u_K^2+\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K,F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}F_{K,\sigma}(\boldsymbol{B}_\sigma)\,u_\sigma^{\mathrm{up}}u_K=\sum_{K\in\mathcal{T}}|K|f_Ku_K.$$

Remark that $u_\sigma^{\mathrm{up}} u_K = \frac{1}{2}\left(u_K^2 + u_\sigma^{up^2}\right) - \frac{1}{2}\left(u_K - u_\sigma^{\mathrm{up}}\right)^2$ which immediately leads to

$$\sum_{K\in\mathcal{T}} |K| f_K u_K = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) u_K^2 + \frac{1}{2}\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) u_K^2$$
$$+ \frac{1}{2}\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) u_\sigma^{up^2} - \frac{1}{2}\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(u_K - u_\sigma^{\mathrm{up}}\right)^2.$$

Using the fact that:

$$\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) u_\sigma^{up^2} = -\sum_{L\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_L\cap\mathcal{F}_{\mathrm{int}}, F_{L,\sigma}(\boldsymbol{B}_\sigma)>0} F_{L,\sigma}\left(\boldsymbol{B}_\sigma\right) u_L^2$$

we immediately get (6.7) as:

$$\frac{1}{2}\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) u_K^2 + \frac{1}{2}\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) u_\sigma^{up^2}$$
$$= \frac{1}{2}\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\mathrm{ext}}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right) u_K^2.$$

□

Proof of Corollary 6.2. Let us denote $(\boldsymbol{e}_i)_{0\leqslant i\leqslant 1}$ the canonical basis of $\mathbb{R}^2$. Assume that $h \leqslant h_0$, where h_0 is defined in Proposition 6.1 and consider:

$$\hat{\boldsymbol{Q}}_K = \sum_{\sigma\in\mathcal{F}_K} \frac{u_K}{|K|} F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\left(\boldsymbol{x}_\sigma - \boldsymbol{x}_K\right).$$

We begin by establishing that

$$\left(\sum_{K\in\mathcal{T}} ||-\lambda\overline{u}\nabla b - \hat{\boldsymbol{Q}}_K||^2_{L^2(K)^2}\right)^{1/2} \leqslant C\max(h,\gamma(h))^{1/2}||\overline{u}||_{H^1(\Omega)}. \tag{6.8}$$

Indeed, we have:

$$\int_K \lambda_K \partial_\xi b = \int_K \lambda_K \nabla b\nabla\left(x_i - x_{K,i}\right) = -\int_K \lambda_K \Delta b\left(x_i - x_{K,i}\right) + \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \lambda_K \nabla b\cdot\boldsymbol{n}_{K,\sigma}\left(x_i - x_{K,i}\right).$$

Using again the consistency estimate of Proposition 4.5 for the fluxes, we get:

$$\left|\sum_{\sigma\in\mathcal{F}_K}\int_\sigma\left(-\lambda_K\nabla b\cdot\boldsymbol{n}_{K,\sigma} - \frac{1}{|\sigma|}F_{K,\sigma}\left(\boldsymbol{B}_\sigma\right)\right)\left(x_i - x_{K,i}\right)\right| \leqslant C|\sigma|\mathrm{card}\left(\mathcal{F}_K\right)h_K^2||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{2,\infty}(\Omega)}.$$

And thus, using assumption (A1) we get that:

$$\left|-u_K\lambda_K\frac{1}{|K|}\int_K \nabla b - \hat{\boldsymbol{Q}}_K\right| \leqslant Ch_K||\lambda||_{W^{1,\infty}(\Omega)}||b||_{W^{2,\infty}(\Omega)}|u_K|.$$

Using the triangle inequality:

$$||-\lambda\overline{u}\nabla b - \hat{\boldsymbol{Q}}_K||^2_{L^2(K)^2} \leqslant ||\left(\lambda-\lambda_K\right)\overline{u}\nabla b||^2_{L^2(K)^2} + ||\lambda_K\left(\overline{u}-u_K\right)\nabla b||^2_{L^2(K)^2}$$
$$+ ||\lambda_K u_K\left(\nabla b - \frac{1}{|K|}\int_K\nabla b\right)||^2_{L^2(K)^2} + ||-\lambda_K u_K\frac{1}{|K|}\int_K\nabla b - \hat{\boldsymbol{Q}}_K||^2_{L^2(K)^2}.$$

Poincare–Wirtinger's inequality applied component by component (see [7]) gives:

$$||\nabla b - \frac{1}{|K|}\int_K \nabla b||_{L^2(K)^2} \leqslant C_P h_K ||b||_{H^2(K)}$$

where $C_P > 0$ is independent on h under assumption (A1), and (6.8) follows from Proposition 6.1 and the above estimate. To conclude, it suffices to remark that by definition:

$$\begin{aligned}|K|\left(\hat{\boldsymbol{Q}}_K - \boldsymbol{Q}_K\right) = &\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm int}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} (u_K - u_L)(\boldsymbol{x}_\sigma - \boldsymbol{x}_K) F_{K,\sigma}(\boldsymbol{B}_\sigma)\\ &+ \sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} u_K(\boldsymbol{x}_\sigma - \boldsymbol{x}_K) F_{K,\sigma}(\boldsymbol{B}_\sigma).\end{aligned}$$

Thus applying Cauchy–Schwarz inequality, we get:

$$\begin{aligned}\sum_{K\in\mathcal{T}} |K|\left|\hat{\boldsymbol{Q}}_K - \boldsymbol{Q}_K\right|^2 &\leqslant 2\left(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm int}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{||\boldsymbol{x}_\sigma - \boldsymbol{x}_K||^2}{|K|}|F_{K,\sigma}(\boldsymbol{B}_\sigma)|\right.\\ &\left. + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext}, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} \frac{||\boldsymbol{x}_\sigma - \boldsymbol{x}_K||^2}{|K|}|F_{K,\sigma}(\boldsymbol{B}_\sigma)|\right)||u||_{h,*}\\ &\leqslant C\left(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0}\frac{|\sigma|h_K^2}{|K|}\right) \leqslant Ch|\Omega|||u||_{h,*}\end{aligned}$$

using the bound on the flux from the proof of Proposition 6.1, and where:

$$||u||_{h,*}^2 = \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K\cap\mathcal{F}_{\rm ext}, F_{K,\sigma}(\boldsymbol{B}_\sigma)>0} F_{K,\sigma}(\boldsymbol{B}_\sigma)|u_K|^2 - \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K, F_{K,\sigma}(\boldsymbol{B}_\sigma)<0} F_{K,\sigma}(\boldsymbol{B}_\sigma)(u_K - u_\sigma^{\rm up})^2.$$

From (6.7) we know that, still denoting u the cellwise constant function of $L^2(\Omega)$ taking the value u_K in cell K:

$$||u||_{h,*}^2 \leqslant ||f||_{L^2(\Omega)}||u||_{L^2(\Omega)} - \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)u_K^2.$$

However, one has

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)u_K^2 = \sum_{K\in\mathcal{T}}\left(\int_K \Delta b + \sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)\right)u_K^2 - \sum_{K\in\mathcal{T}}\int_K \Delta b u_K^2.$$

Using Stokes' formula, the consistency of the flux (4.5) and assumption (A1), we get:

$$\left|\int_K \Delta b + \sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma)\right| = \left|\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(\boldsymbol{B}_\sigma) + \int_\sigma \lambda\nabla b\cdot\boldsymbol{n}_{K,\sigma}\right| \leqslant C\sum_{\sigma\in\mathcal{F}_K}|\sigma|h_K \leqslant C|K|.$$

Thus, we have $||u||_{h,*}^2 \leqslant ||f||_{L^2(\Omega)}||u||_{L^2(\Omega)} + C||u||_{L^2(\Omega)}^2$, and using the triangle inequality $||u||_{L^2(\Omega)} \leqslant ||u - \overline{u}||_{L^2(\Omega)} + ||\overline{u}||_{L^2(\Omega)}$, we get that:

$$||u||_{h,*}^2 \leqslant C\left(1 + \max(h,\gamma(h))^{1/2}\right)\left(||f||_{L^2(\Omega)} + \left(1 + \max(h,\gamma(h))^{1/2}\right)||\overline{u}||_{H^1(\Omega)}\right)||\overline{u}||_{H^1(\Omega)}$$

which concludes the proof. □

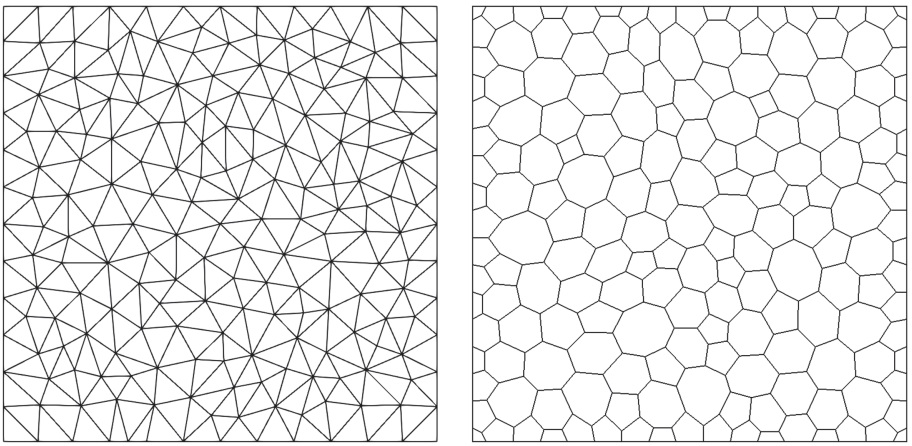

FIGURE 3. Example of meshes for the *2dDelaunay* and *2dDualDelaunay* mesh sequences.

7. Numerical exploration

To assess the asymptotic behavior of the method, we use a very simple analytical solution on which the positivity of the Laplacian is guaranteed (contrary to typical geological applications). Let us consider the square domain $\Omega =]0, L[\times]0, L[$ associated with the topography $b(x, y) = b_0 - \delta\left((x - x_0)^2 + (y - y_0)^2\right)$ where $x_0 = y_0 = L/2$, for which $-\Delta b = 4\delta$. Consider the water height $\overline{u}$ defined by $\overline{u}(x, y) = u_0 \exp\left(-\alpha\left((x - x_0)^2 + (y - y_0)^2\right)\right)$. Defining $\lambda(x, y) = 1$, we obtain $-\mathrm{div}(\lambda\overline{u}\nabla b) = f$ with $f(x, y) = 4\delta\overline{u}(x, y)\left(1 - \alpha\left((x - x_0)^2 + (y - y_0)^2\right)\right)$. The corresponding water discharge $q = ||\lambda\overline{u}\nabla b||$ is given by

$$q = 2\delta\overline{u}(x, y)\left((x - x_0)^2 + (y - y_0)^2\right)^{1/2}.$$

In practice we take $u_0 = 1$, $\alpha = 2$, $\delta = 1/2$, $b_0 = 2$ and $L = 1$. We consider seven types of mesh sequences. The first sequence consists in classical Delaunay meshes (*2dDelaunay*). The second one (*2dDualDelaunay*) is obtained by considering the dual meshes of a sequence of Delaunay meshes (Fig. 3). The third sequence (*2dVoronoi*) is made of Voronoi meshes, possessing the mesh orthogonality property. The fourth sequence (*2dKershawBox*) is a sequence of Kershaw meshes, while the fifth one (*2dCheckerBoardBox*) is a sequence of checkerboard meshes. These two sequences have only quadrangular cells which are distorted for the sequence *2dKershawBox*, while the sequence *2dCheckerBoardBox* allows to test the behavior of the method in presence of non conformities. The sixth sequence, named *2dSquareCart*, is simply a uniform cartesian mesh with square cells, while the seventh one named *2dRectCart* is a uniform cartesian mesh with rectangular cells. These two last sequences are mainly intended to serve as reference. For the generalized MFD algorithm, we consider two consistent gradient reconstructions coming from finite volumes: the first one is the hybrid finite volume gradient operator of [12], that uses faces values for the topography b, while the second one is the vertex based finite volume gradient operator (VVM) of [5].

As it is the quantity of practical interest, we will express our convergence results directly in terms of the water discharge q instead of using the water height. Let us begin by some results for sequences *2dSquareCart*, *2dRectCart* and *2dVoronoi*. Being the only ones that satisfy the mesh orthogonality requirement (see Figs. 4 and 5), we expect that the TPFA scheme will also converge on those sequences. The corresponding results are displayed on Figures 6 and 7, while the associated approximate convergence orders for each scheme are given in Table 1.

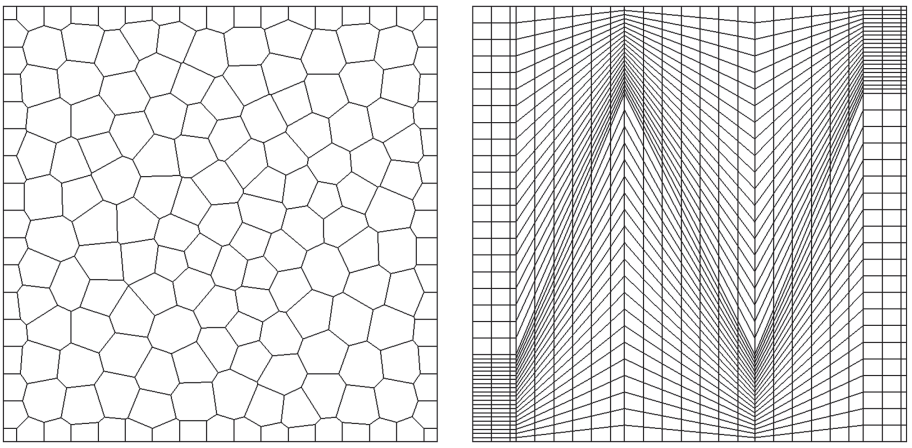

FIGURE 4. Example of meshes for the *2dVoronoi* and *2dKershawBox* mesh sequences.

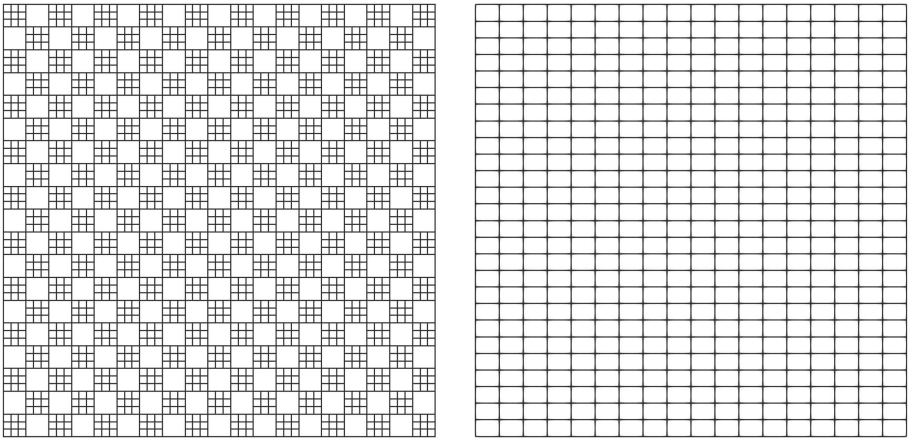

FIGURE 5. Example of meshes for the *2dCheckerBoardBox* and *2dRectCart* mesh sequences.

Clearly, on the basic cartesian cases *2dSquareCart* and *2dRectCart*, all the schemes give identical results. A closer look at our generalized flux formula immediately reveals that this is perfectly normal as the flux $F_{K,\sigma}$ degenerate into the two-point flux formula on those cartesian meshes where the barycenter of each internal face is exactly the intersection point between the face and the segment joining the cell centers on each side of the face (see [12]). Moreover, Table 1 reveals that all the methods are superconvergent on the three orthogonal mesh sequences. In [20] it is indeed established in the case of exact fluxes that the approximation of the water height superconverges at least on rectangular meshes, which explains that this observed superconvergence is probably not anomalous. Next, we turn to the other mesh sequences, for which we cannot expect convergence for the TPFA scheme. The corresponding results are displayed on Figures 8 and 9, while the associated approximate convergence orders for each scheme are given in Table 2.

As expected, the TPFA scheme does not converge anymore, however the generalized MFD algorithm is converging for both the hybrid and virtual volume (VVM) gradients. The results for those two variants are in fact relatively close to each other. In particular, the behavior of the method on sequence *2dCheckerBoardBox* confirms its ability to deal with non conformities and thus local mesh refinement. Remark that superconvergence

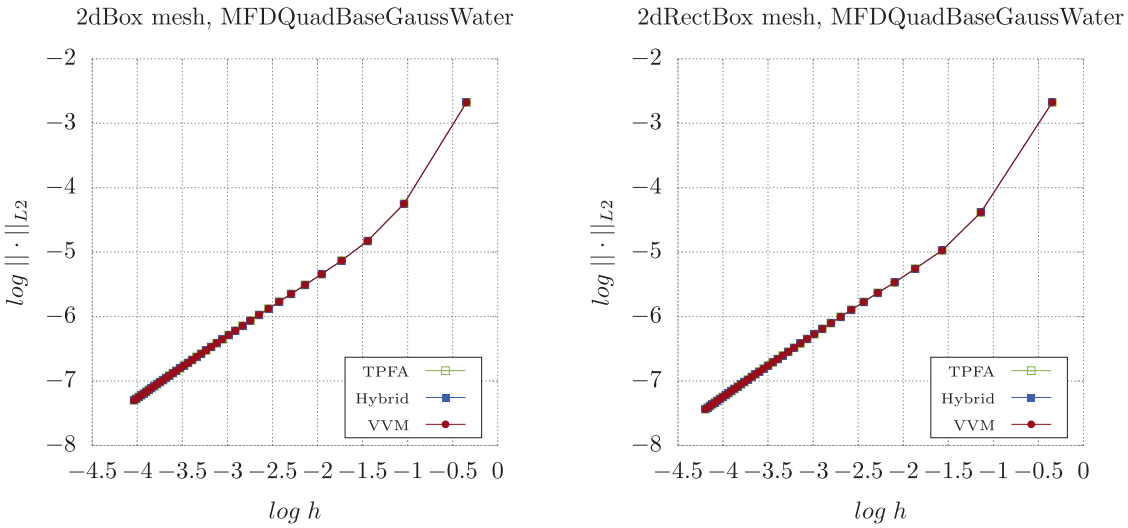

FIGURE 6. Convergence curves for sequences *2dSquareCart* and *2dRectCart*.

TABLE 1. Approximate orders of convergence for the three schemes on orthogonal meshes.

	2dSquareCart	*2dRectCart*	*2dVoronoi*
TPFA	0.943	0.942	1.147
Hybrid	0.943	0.942	1.061
VVM	0.943	0.942	1.061

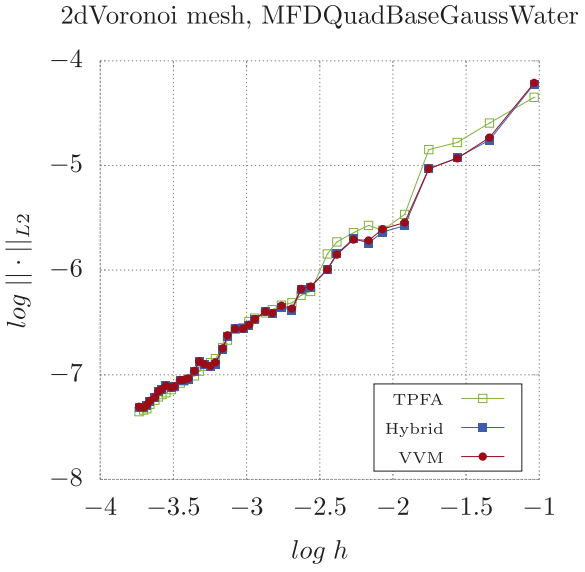

FIGURE 7. Convergence curves for sequence *2dVoronoi*.

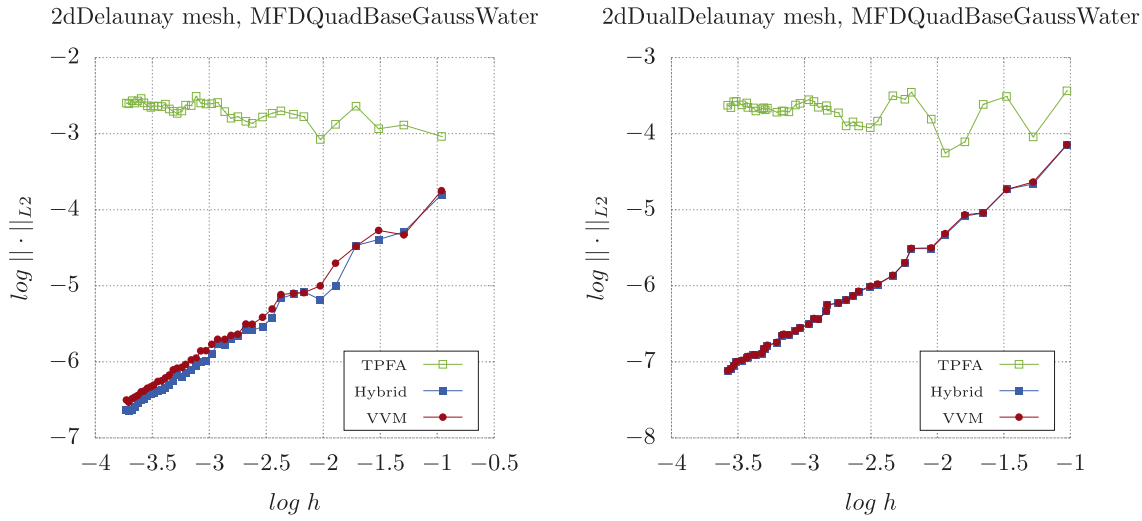

FIGURE 8. Convergence curves for sequences *2dDelaunay* and *2dDualDelaunay*.

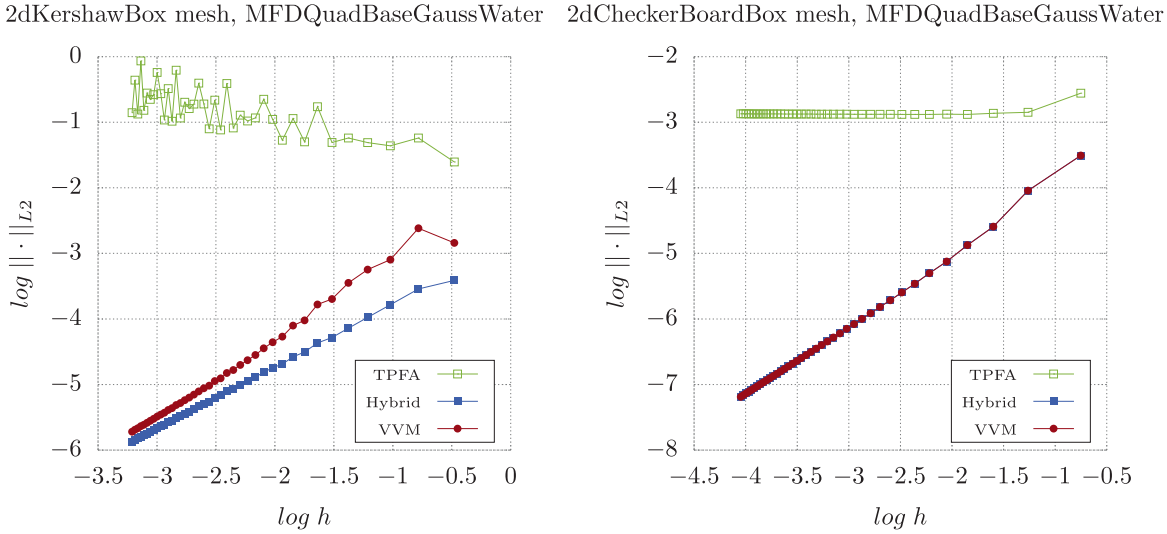

FIGURE 9. Convergence curves for sequences *2dKershawBox* and *2dCheckerBoardBox*.

TABLE 2. Approximate orders of convergence for the three schemes on general meshes.

	2dDelaunay	*2dDualDelaunay*	*2dKershawBox*	*2dCheckerBoardBox*
Hybrid	0.991	1.081	0.941	1.032
VVM	0.971	1.084	1.213	1.032

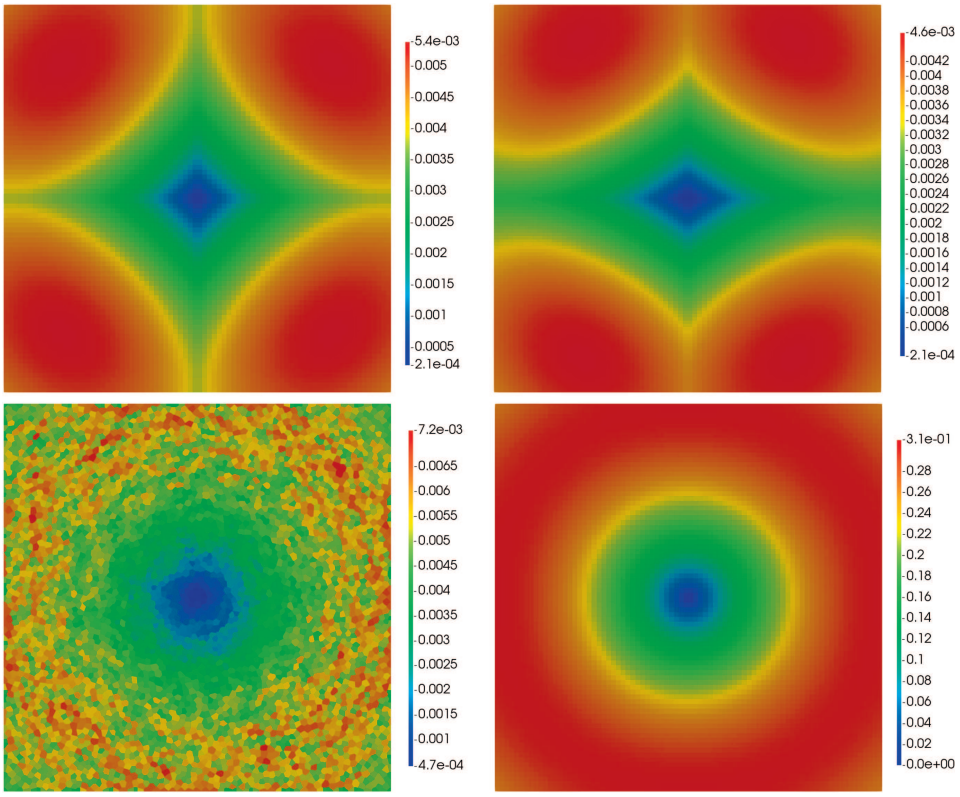

FIGURE 10. Comparison between influx and water discharge for the TPFA scheme on orthogonal meshes (exact water discharge is bottom right).

is also achieved again on those test cases. To conclude, on Figure 10, we represent on the bottom-right the correct water discharge, while the top-left figure is the water influx $\widetilde{q}_K$ for sequence *2dSquareCart*, the top-right figure is the water influx for sequence *2dRectCart* and the bottom-left figure is the water influx for sequence *2dVoronoi*. These very basic examples illustrate our point on the water influx $\widetilde{q}_K$: it is clearly a mesh dependent quantity, that cannot be expected to reproduce the behavior of the discharge correctly, even on those very simple cases. If one does not consider the Manning–Strickler framework, one of the tricky features of $\widetilde{q}_K$ is that on very common cartesian mesh sequences such as *2dSquareCart* and *2dRectCart* it has a relatively nice qualitative behavior. It even seems to converge, although to a mesh sequence dependent quantity. This explains why it has been used inadvertently in the hydrogeology community, despite of its mesh dependent nature: on the widely used cartesian meshes, it has an acceptable behavior that cannot be easily discarded without a reference continuous model, in particular for complex topographies. However, on more general meshes, the situation is very different, even for the TPFA scheme as the case of sequence *2dVoronoi* reveals. We display on Figure 11 the water influx $\widetilde{q}_K$ for the four other mesh sequences, in the case of the hybrid gradient. It is completely obvious on those sequences that $\widetilde{q}_K$ should not be considered as a physically relevant quantity.

8. CONCLUSION

We have established the equivalence between a classical family of multiple flow direction algorithms and the TPFA scheme applied to a family of stationary Manning–Strickler models. From this equivalence we have proposed an improvement of the derivation of the discrete water discharge based on a consistent flux reconstruction

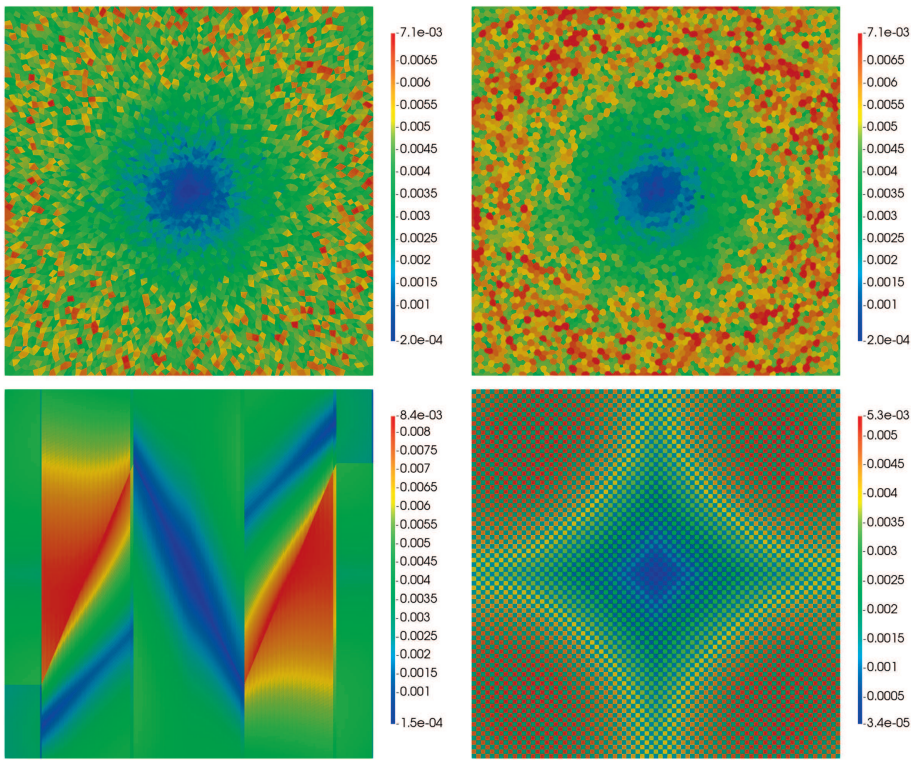

FIGURE 11. Water influx for hybrid gradient on meshes *2dDelaunay*, *2dDualDelaunay*, *2dKershawBox* and *2dCheckerBoardBox*.

rather than directly using the water influx intermediate unknown that is not only non convergent in general, but does not approximate the water discharge in any case. Then, using more advanced flux reconstruction schemes, we have proposed a multiple flow direction algorithm adapted to general meshes. A convergence theory that covers both cases was developed, and numerical experiments illustrate the good behavior of the method. Remark that the extension to the case of Manning–Strickler tensors rather than coefficients is straightforward, as would be the extension of the present analysis to multiple flow direction models that use powers of the topographic slope.

Acknowledgements. The author would like to thank G. Enchéry and L. Agélas for our numerous discussions on the contents of this article.

REFERENCES

[1] C. Bardos, Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; application à l'équation de transport. *Ann. Sci. Ec. Norm. Sup. Ser. 4* **3** (1970) 185–233.
[2] H. Beirão Da Veiga, Existence results in Sobolev spaces for a stationary transport equation. *Ricerche Mat. Suppl. XXXVI* (1987) 173–184.
[3] S. Brenner and R. Scott, The Mathematical Theory of Finite Element Methods, 3rd ed. Springer (2008).
[4] J. Coatléven, Semi hybrid method for heterogeneous and anisotropic diffusion problems on general meshes. *ESAIM: M2AN* **49** (2015) 1063–1084.
[5] J. Coatléven, A virtual volume method for heterogeneous and anisotropic diffusion-reaction problems on general meshes. *ESAIM: M2AN* **51** (2017) 797–824.
[6] B. Cockburn, B. Dong and J. Guzmán, Optimal convergence of the original DG method for the transport-reaction equation on special meshes. *SIAM J. Numer. Anal.* **46** (2008) 1250–1265.

[7] D.A. Di Pietro and A. Ern, Mathematical Aspects of Discontinuous Galerkin Methods. Springer (2012).
[8] R.J. DiPerna and P.L. Lions, Ordinary differential equations, transport theory and Sobolev spaces. *Invent. Math.* **98** (1989) 511–547.
[9] R.H. Erskine, T.R. Green, J.A. Ramirez and L.H. MacDonald, Comparison of grid-based algorithms for computing upslope contributing area. *Water Resour. Res.* **42** (2006) W09416.
[10] R. Eymard, T. Gallouët and R. Herbin, Finite volume methods, edited by P.G. Ciarlet and J.-L. Lions. In: *Handbook of Numerical Analysis: Techniques of Scientific Computing, Part III. North-Holland, Amsterdam* (2000) 713–1020.
[11] R. Eymard, T. Gallouët and R. Herbin, A new finite volume scheme for anisotropic diffusion problems on general grids: convergence analysis. *C. R. Math. Acad. Sci. Paris* **344** (2007) 403–406.
[12] R. Eymard, T. Gallouët and R. Herbin, Discretisation of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilisation and hybrid interfaces. *IMA J. Numer. Anal.* **30** (2010) 1009–1043.
[13] R. Eymard, C. Guichard and R. Herbin, Benchmark 3D: the vag scheme. In: Vol. 2 of *Springer Proceedings in Mathematics, FVCA6, Prague* (2011) 213–222.
[14] R. Eymard, C. Guichard and R. Herbin, Small-stencil 3D schemes for diffusive flows in porous media. *ESAIM: M2AN* **46** (2011) 265–290.
[15] E. Fernández-Cara, F. Guillén and R.R. Ortega, Mathematical modeling and analysis of visco-elastic fluids of the oldroyd kind, edited by P.G. Ciarlet and J.L. Lions. In: Vol. VIII of *Handbook of Numerical Analysis: Numerical Methods for Fluids, Part 2.* North-Holland, Amsterdam (2002) 543–661.
[16] T.G. Freeman, Calculating catchment area with divergent flow based on a regular grid. *Comput. Geosci.* **17** (1991) 413–422.
[17] V. Girault and L. Tartar, l^p and $w^{1,p}$ regularity of the solution of a steady transport equation. *C. R. Acad. Sci. Paris, Ser. I* **348** (2010) 885–890.
[18] P. Holmgren, Multiple flow direction algorithms for runoff modelling in grid based elevation models: an empirical evaluation. *Hydrol. Process.* **8** (1994) 327–334.
[19] C. Johnson and J. Pitkäranta, An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comput.* **46** (1986) 1–26.
[20] P. Lesaint and P.A. Raviart, On a finite element method for solving the neutron transport equation. *Publ. Math. Inf. Rennes* **S4** (1974) 1–40.
[21] C. Qin, A.-X. Zhu, T. Pei, B. Li, C. Zhou and L. Yang, An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. *Int. J. Geog. Inf. Sci.* **21** (2007) 443–458.
[22] P. Quinn, K. Beven, P. Chevallier and O. Planchon, The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrol. Process.* **5** (1991) 59–79.
[23] A. Richardson, C.N. Hill and J.T. Perron, IDA: an implicit, parallelizable method for calculating drainage area. *Water Resour. Res.* **50** (2013) 4110–4130.
[24] J. Seibert and B.L. McGlynn, A new triangular multiple flow direction algorithm for computing upslope areas from gridded digital elevation models. *Water Resour. Res.* **43** (2007) W04501.
[25] D.G. Tarboton, A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resour. Res.* **33** (1997) 309–319.
[26] D.M. Wolock and G.J. McCabe Jr., Comparison of single and multiple flow direction algorithms for computing topographic parameters in topmodel. *Water Resour. Res.* **31** (1995) 1315–1324.
[27] Q. Zhou, P. Pilesjö and Y. Chen, Estimating surface flow paths on a digital elevation model using a triangular facet network. *Water Resour. Res.* **47** (2011) W07522.