Simulation numérique d'écoulements compressibles complexes par des méthodes de type Lagrange-projection : applications aux équations de Saint-Venant

Maxime Stauffert

NNT : 2018SACLV045

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université de Versailles Saint-Quentin-en-Yvelines

Laboratoires d'accueil : Laboratoire de mathématiques de Versailles, UMR 8100 CNRS et Maison de la Simulation, USR 3441 CEA-CNRS-INRIA-UPSUD-UVSQ

Spécialité de doctorat : Mathématiques appliquées

Maxime STAUFFERT

Simulation numérique d'écoulements compressibles complexes par des méthodes de type Lagrange-projection : applications aux équations de Saint-Venant

Date de soutenance : 5 octobre 2018

Après avis des rapporteurs : Christophe BERTHON (U. de Nantes)
Stéphane CLAIN (U. do Minho)

Jury de soutenance :

Christophe BERTHON	(U. de Nantes)	Rapporteur
Christophe CHALONS	(UVSQ)	Directeur de thèse
Stéphane CLAIN	(U. do Minho)	Rapporteur
Anaïs CRESTETTO	(U. de Nantes)	Examinatrice
Samuel KOKH	(CEA Saclay)	Co-directeur de thèse
Raphaël LOUBÈRE	(U. de Bordeaux)	Président du jury
Pascal TREMBLIN	(CEA Saclay)	Examinateur
Marie-Hélène VIGNAL	(U. de Toulouse)	Examinatrice

Remerciements

Me voici à la fin d'une aventure, un long fleuve, pas si tranquille, mais qui m'a permis de prendre conscience de tant de choses. J'ai rencontré et côtoyé beaucoup de personnes lors de cette thèse et je souhaite les remercier pour tout ce qu'ils m'ont apporté.

Tout d'abord, je tiens à remercier mon directeur de thèse Christophe Chalons. Merci de m'avoir fait découvrir plus en détails un sujet que je connaissais mal, tu as été un puits de connaissances théoriques. Tu as été très disponible tout au long de la thèse et bien que tu aies accumulé les responsabilités. Tu m'as également permis de rencontrer et de travailler avec plusieurs personnes de la communauté.

C'est naturellement vers Samuel Kokh, qui m'a encadré à la Maison de la Simulation, que se tournent maintenant mes remerciements. Tu m'as beaucoup soutenu dès les premiers entretiens à l'école doctorale pour obtenir le financement. Tu m'as également appris plein de techniques et de bonnes pratiques en programmation. J'aurais plaisir à reprendre notre projet de code hybride et de voir jusqu'où il est possible d'aller.

Je voudrais remercier Christophe Berthon et Stéphane Clain, pour avoir accepté de rapporter mon manuscrit. J'espère que sa lecture n'a pas été trop fastidieuse au bord de la piscine/mer cet été.

Je remercie également Anaïs Crestetto, Raphaël Loubère, Pascal Tremblin et Marie-Hélène Vignal pour avoir accepté de faire partie de mon jury et fait le déplacement depuis Bordeaux, Nantes, Toulouse et même Minho.

J'ai une attention particulière pour Marie-Hélène qui m'a toujours bien accueilli lors de mes séjours à Toulouse et qui m'avait proposé un sujet de post-doc, j'aurais apprécié travailler avec toi.

Un énorme merci à Raphaël, qui m'a reçu chez lui pour travailler sur notre collaboration. J'admire ton dosage très fin entre les mots pour rire et les remarques techniques. J'ai tellement appris de nos séances de travail et de débogages intensives !

Je tiens à remercier, dans un premier temps, tout le laboratoire de Mathématiques de Versailles (LMV). Merci d'abord à Catherine Donati-Martin et à Nadège Arnaud pour votre gentillesse. Merci également à Laure Frerejean pour ton caractère bien trempé et tes supers idées, notamment pour le CANUM. J'en profite pour remercier toute l'équipe EDP du LMV avec qui j'ai passé de très bons moments au Cap d'Agde.

Merci aux doctorants et post-doctorants que j'ai pu croiser à Versailles, notamment : Arsen, Benjamin, Hélène, Thomas, Salim, Sarah et Sybille. Patricio pour ton humour percutant et ta décontraction ; Ila pour ton grand sourire et tes supers présentations avec Alice et Bob ; Antoine

pour ta gentillesse, toutes les questions pratiques auxquelles tu as toujours su répondre et pour ton amour, tout comme le mien, des mots de félicitations ; Florian pour ton côté rustre qui contre balance une sensibilité humaine rare.

Merci beaucoup Seb pour le soutien inébranlable que tu m'as apporté à plusieurs moments pendant la thèse. Il est tellement appréciable de rencontrer quelqu'un d'aussi franc et en même temps d'aussi bienveillant que toi. Au plaisir de continuer à travailler avec toi et de faire des petites soirées sympa avec Lucie.

Je termine par Cami, avec qui j'ai tant partagé et discuté de tout et de rien. Merci pour tous tes bons gâteaux, je signe encore pour les goûter si jamais tu en as besoin. Me confier à toi m'a permis de comprendre plein de choses et de prendre des décisions importantes. Je te souhaite beaucoup de bonheur et d'accomplissement dans ta vie personnelle.

Dans un deuxième temps, je remercie ma deuxième maison d'accueil, la Maison de la Simulation (MdlS). Merci à Edouard Audit et à Valérie Belle pour votre gentillesse. Dans le désordre et pour toutes les discussions du repas de midi et des pauses café aussi intéressantes que déjantées, je tiens à remercier : Abel, Axelle, Cécile, Florence, Haithem, Julien D., Mathieu L., Maximilien, Michel, Olivier, Yacine. Pour des discussions plus sérieuses, que ce soit en mathématiques, en informatique ou sur les jeux vidéo, merci à Ksander et Arnaud. Merci également à Pierre pour nos échanges précis notamment sur les conditions de bords et la librairie PETSc. Bonne continuation à toi, Thomas, j'ai apprécié nos réflexions existentielles sur les schémas.

Un énorme merci à Mat et Pascal, qui m'ont apporté un soutien fort au moment où j'en avais le plus besoin. Vous avez su vous rendre très disponibles, être à l'écoute et partager vos expériences ainsi que vos avis. Ce fut un plaisir de découvrir un peu d'astrophysique à travers tes recherches, Pascal. Je te souhaite plein de bonheur en tant que nouveau papa. Bonne continuation à toi, Mat, avec ta grande petite famille.

Il n'y aurait pas eu autant de verres à Paris, d'escalade à Issy et de vacances en Ardèche sans Julien. Merci à toi pour ton entrain quotidien, tes blagues douteuses, ton autodérision et ta culture générale impressionnante.

Enfin, merci à Coco, avec qui j'ai tant parlé, notamment de développement personnel. J'ai pu me confier à toi ; tu m'as permis de tenir dans les moments durs et de faire attention à moi lorsque c'était important. Ta sensibilité me touche beaucoup et tes blagues salaces me font toujours bien rire. Je te souhaite plein de bonheur dans tous tes projets.

Continuons avec les amis, comme Momo et Wasim qui m'ont permis de penser à autre chose lorsque l'on grimpait.

Merci à Victor pour ton accueil à Toulouse et ta bonne humeur lors des différentes conférences où l'on s'est croisés. Bon courage à toi, David, pour ta dernière ligne droite, décalée d'un mois par rapport à moi. J'ai toujours apprécié notre partage d'expérience et j'espère que tu trouveras ta voie.

Une très grosse pensée à Charlotte et Estelle, que la distance, tant physique que professionnelle, n'a pas su éloigner. Je te souhaite plein de bonheur Chacha avec J-B, en espérant venir vous voir bientôt à Lyon. Bon courage à toi, Estelle ; tu as toujours été une passionnée dans tout ce que tu as fait et j'admire ta persévérance.

Et pour plus de fun et de légèreté, un énorme merci aux colocs de la villa des Roses. J'inclus

bien entendu Séverin, qui a certainement plus dormi que moi dans la chambre du bas. Merci pour tous les barbecues sans Antoine, les vacances à Amsterdam et en Bretagne. Ce fut une expérience de partage et de joie qui restera gravée dans ma mémoire. Bonne continuation à vous tous, au Canada, à Amsterdam, Madrid ou Nottingham, ou encore à Paris. J'espère vous réunir tous dans le Sud prochainement pour terraformer Mars ou Banguer !

Je termine enfin par remercier toute ma famille. Merci d'abord à la famille de Nat que j'apprends encore à connaître mais qui m'a déjà accueilli à bras ouverts. Vous êtes, à vous tous, une tribu pleine d'amour et de bienveillance les uns envers les autres ; cela fait vraiment plaisir à voir et à vivre.

Plein de bisous à Olivier, Sandrine et Romain. Merci beaucoup au Chalet, noyau dur de notre petite famille que j'aime tant : Mag, Jules et Paul. Une pensée émue à Kiki, qui n'y est pas vraiment pour rien là-dedans. Un énorme bisou à Mimi, que j'admire pour sa force, toujours aussi dévouée à nous tous. Un grand merci à mon frère, avec qui j'ai tellement de plaisir à partager de plus en plus de choses. Je te souhaite plein de bonheur avec Maïa et Leo. Je vous suis vraiment reconnaissant, mes parents ; vous avez été d'un soutien inébranlable tout au long de ma vie et encore plus pendant la thèse. Vous avez sauté dans des trains à chaque fois que j'en ai eu besoin et encore une fois pour ma soutenance. J'ai hâte de partager des moments plus sereins avec vous tous.

Je ne te remercierai jamais assez, petit chou, pour tout ce que tu as fait pour moi. Je ne suis plus vraiment le même depuis que l'on s'est rapprochés. Tu m'as libéré de tellement de freins insoupçonnés. Tu me rends le plus heureux des hommes et je n'ai qu'une seule envie, c'est celle de découvrir mini chou avec toi et de l'emmener voir plein de belles choses.

« Le bonheur est souvent la seule chose
qu'on puisse donner sans l'avoir
et c'est en le donnant qu'on l'acquiert. »
Voltaire

À mes petit et mini choux

Table des matières

Introduction générale

0.1 Contexte général

Cette thèse s'intéresse à la résolution numérique du système d'équations aux dérivées partielles de Saint-Venant (en anglais « shallow water equations »). Adhémar Barré de Saint-Venant a décrit le mouvement des cours d'eau dans une série d'articles publiés à la fin du 19ème siècle, parmi lesquels l'article historique [12]. Ces équations régissent plus généralement l'écoulement d'un fluide en eau peu profonde (d'où le mot « shallow » en anglais) au dessus d'une topographie donnée. Que ce soit en une ou deux dimensions d'espace, la hauteur d'eau, ou de manière équivalente la profondeur de l'eau, doit être petite par rapport aux dimensions du problème dans les directions horizontales. On rencontre ce système en géophysique par exemple pour décrire l'évolution de la surface libre d'un lac, les courants de marée ou encore les tsunamis.

Ces équations peuvent être dérivées du système de Navier-Stokes qui décrivent de manière plus générale l'évolution d'un fluide non-newtonien. Si on suppose que l'accélération verticale du fluide est nulle, on peut ainsi considérer que chaque variable peut être approchée par son intégrale sur la hauteur du fluide en chaque point du sol. En conservant le terme source correspondant à l'influence de la topographie, on obtient ainsi en deux dimensions d'espace

$$\begin{cases} \partial_t h + \nabla \cdot (h\mathbf{u}) & = 0, \\ \partial_t (h\mathbf{u}) + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla \dfrac{gh^2}{2} & = -gh\,\nabla z, \end{cases} \tag{0.1.1}$$

où $(\mathbf{x}, t) \in \mathbb{R}^2 \times \mathbb{R}^+ \mapsto h > 0$ correspond à la hauteur d'eau par rapport au sol (équivalent à l'intégrale de la densité sur la hauteur), $(\mathbf{x}, t) \in \mathbb{R}^2 \times \mathbb{R}^+ \mapsto \mathbf{u} = (u, v)^T \in \mathbb{R}^2$ est le vecteur vitesse, $g > 0$ la constante de gravitation et $\mathbf{x} \in \mathbb{R}^2 \mapsto z$ la fonction représentant la topographie.

On peut tout de suite remarquer que si cette dernière est constante (en espace puisqu'elle est indépendante du temps), on retombe sur le système d'Euler isentropique, avec la loi d'état sur la pression $p(h) = g\frac{h^2}{2}$. Cette proximité avec le système d'Euler sera utilisée à plusieurs reprises afin de tirer profit de méthodologies développées pour ce dernier, en leur apportant des modifications permettant de traiter l'influence de la topographie.

On peut montrer que le système 0.1.1 est hyperbolique, car la matrice associée aux termes conservatifs est diagonalisable avec des valeurs propres réelles distinctes. En effet, cette dernière peut s'écrire sous la forme $A(\mathbf{n}) = DF_x \cdot n_1 + DF_y \cdot n_2$, où $\mathbf{n} = (n_1, n_2)^T \in \mathbb{R}^2$ est un vecteur

unitaire quelconque et

$$DF_x = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{(hu)^2}{h^2} + gh & 2\frac{(hu)}{h} & 0 \\ -\frac{(hu)(hv)}{h^2} & \frac{(hv)}{h} & \frac{(hu)}{h} \end{pmatrix}, \qquad DF_y = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{(hu)(hv)}{h^2} & \frac{(hu)}{h} & \frac{(hv)}{h} \\ -\frac{(hv)^2}{h^2} + gh & 0 & 2\frac{(hv)}{h} \end{pmatrix}.$$

Ses valeurs propres sont $\{\mathbf{u}^T\mathbf{n} - c, \mathbf{u}^T\mathbf{n}, \mathbf{u}^T\mathbf{n} + c\}$, où $c = \sqrt{gh}$ est la vitesse des ondes de gravité (équivalente à la vitesse du son pour Euler). On peut trouver plus de détails sur cette propriété d'hyperbolicité dans la thèse de E. Audusse [1] pour le système de Saint-Venant ou dans le livre [17] sur les systèmes de lois de conservation de manière plus générale. Il est simplement important de noter que cette propriété justifie l'utilisation de méthodes dites Volumes Finis et Galerkin discontinues que l'on détaillera dans la Section 0.3 pour dériver les schémas numériques. En effet, celles-ci sont particulièrement adaptées au traitement de solutions discontinues des systèmes hyperboliques ainsi qu'à leur caractère conservatif.

L'énergie associée au système (0.1.1)

$$E = \frac{h|\mathbf{u}|^2}{2} + \frac{gh^2}{2} + ghz$$

satisfait l'inégalité d'entropie suivante :

$$\partial_t \mathcal{U} + \nabla \cdot \left[\left(\mathcal{U} + \frac{gh^2}{2} \right) \mathbf{u} \right] \leq 0,$$

où l'entropie $\mathcal{U}$ est choisie comme étant l'énergie E. On peut réécrire cette inégalité sous la forme non conservative suivante :

$$\partial_t \tilde{\mathcal{U}} + \nabla \cdot \left[\left(\tilde{\mathcal{U}} + \frac{gh^2}{2} \right) \mathbf{u} \right] \leq -gh\mathbf{u}\nabla z, \tag{0.1.2}$$

où l'entropie correspond cette fois-ci à l'énergie associée au système d'Euler, soit $\tilde{\mathcal{U}} = \frac{h|\mathbf{u}|^2}{2} + \frac{gh^2}{2}$. On s'attachera à démontrer des résultats discrets d'inégalité d'entropie pour les schémas numériques étudiés afin que les solutions numériques de ces schémas convergent vers une solution continue satisfaisant l'inégalité continue associée.

Les états stationnaires (solutions indépendantes du temps) du système (0.1.1) satisfont les deux équations suivantes :

$$\begin{cases} \nabla \cdot (h\mathbf{u}) & = 0, \\ \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla \dfrac{gh^2}{2} & = -gh\,\nabla z. \end{cases}$$

L'état dit du « lac au repos » est une famille particulière de ces solutions pour lesquelles il n'y a pas de mouvement, soit $\mathbf{u} = \mathbf{0}$. Par conséquent, la hauteur d'eau totale $H = h + z$ est constante, ce qui correspond à l'intuition physique. Lors de la construction des différentes méthodes numériques, la préservation des ces états au niveau discret tiendra une place importante dans l'étude mathématique ainsi que pour les résultats numériques. Il existe également d'autres

états stationnaires, avec des vitesses non-nulles, que l'on testera numériquement et que l'on caractérise facilement en 1D par

$$hu = \mathrm{C}_1 \qquad \text{et} \qquad \frac{u^2}{2} + g(z+h) = \mathrm{C}_2.$$

Bien que l'on ne conserve pas ces états exactement, il existe dans la littérature des méthodes qui y parviennent, on peut citer par exemple [3, 21].

0.2 Approche acoustique-transport

On présente dans cette section l'approche acoustique-transport (Lagrange-projection) avec la décomposition d'opérateur ainsi que la méthode de relaxation dans le cadre des équations de Saint-Venant (0.1.1). Il s'agit d'une adaptation de la méthodologie décrite dans la thèse de M. Girardin [16] et originellement proposée dans l'article [10].

On souhaite séparer l'opérateur de transport contenant les ondes à vitesse matière $\mathbf{u}$ des autres phénomènes physiques contenus dans les équations. En développant les dérivées spatiales dans le système complet, on obtient

$$\begin{cases} \partial_t h + h\,(\nabla\cdot\mathbf{u}) & + (\mathbf{u}\cdot\nabla)h & = 0, \\ \partial_t(h\mathbf{u}) + h\mathbf{u}\,(\nabla\cdot\mathbf{u}) + \nabla\left(g\dfrac{h^2}{2}\right) & + (\mathbf{u}\cdot\nabla)(h\mathbf{u}) & = -gh\,\nabla z. \end{cases}$$

On utilise une décomposition d'opérateur d'ordre 1 en temps pour obtenir les deux systèmes d'équations suivants. Tout d'abord celui que l'on appelle système acoustique, au vu de l'étude qui va suivre sur les valeurs propres qui lui sont associées, et qui s'écrit :

$$\begin{cases} \partial_t h + h\,(\nabla\cdot\mathbf{u}) & = 0, \\ \partial_t(h\mathbf{u}) + h\mathbf{u}\,(\nabla\cdot\mathbf{u}) + \nabla\left(g\dfrac{h^2}{2}\right) & = -gh\,\nabla z. \end{cases} \tag{0.2.1}$$

Puis le système de transport, qui s'écrit :

$$\begin{cases} \partial_t h + (\mathbf{u}\cdot\nabla)h & = 0, \\ \partial_t(h\mathbf{u}) + (\mathbf{u}\cdot\nabla)(h\mathbf{u}) & = 0. \end{cases}$$

En dimension 1 d'espace seulement et une fois discrétisé, le système de transport ci-dessus coïncide avec l'étape de projection dans les méthodes type Lagrange-projection classiques, dont on peut trouver une description dans le livre [17].

On opère le changement de variable $h \mapsto \tau = \frac{1}{h}$ dans le système acoustique (0.2.1), où τ est appelé le volume spécifique, pour arriver au système suivant dans lequel on rajoute l'équation sur la topographie

$$\begin{cases} \partial_t \tau - \tau(\mathbf{x},t)\,(\nabla\cdot\mathbf{u}) & = 0, \\ \partial_t \mathbf{u} + \tau(\mathbf{x},t)\,\nabla\left(\dfrac{g}{2\tau^2}\right) & = -g\,\nabla z, \\ \partial_t z & = 0. \end{cases}$$

Puis on approche les dérivées spatiales $\tau(\cdot,t)\,\nabla_{x_i}$ par $\tau(\cdot,t^n)\nabla_{x_i}$, où $i=1,2$ et t^n est le temps initial à la $n^{\text{ème}}$ itération, pour obtenir le système suivant :

$$\begin{cases} \partial_t\tau - \tau(\mathbf{x},t^n)\,(\nabla\cdot\mathbf{u}) & = 0, \\ \partial_t\mathbf{u} + \tau(\mathbf{x},t^n)\,\nabla\left(g\dfrac{h^2}{2}\right) & = -\dfrac{g}{\tau}\tau(\mathbf{x},t^n)\,\nabla z, \\ \partial_t z & = 0. \end{cases} \tag{0.2.2}$$

On remarque que dans le cadre 1D sans terme source et en considérant la variable de masse $m(x)$ telle que $\mathrm{d}m = \tau(x,t^n)\,\mathrm{d}x$, on obtient le système écrit en variables lagrangiennes :

$$\begin{cases} \partial_t\tau - \partial_m u & = 0, \\ \partial_t\mathbf{u} + \partial_m p(\tau) & = 0, \end{cases}$$

d'où l'assimilation du système acoustique avec la partie lagrangienne des méthodes de type Lagrange-projection classiques. On termine avec une relaxation de type Suliciu [24] de la pression dans le système acoustique afin de le rendre linéaire. On introduit une variable π, une linéarisation de la pression $p(\tau) = \frac{g}{2\tau^2}$, et on obtient le système relaxé suivant contenant une équation supplémentaire :

$$\begin{cases} \partial_t\tau - \tau(\mathbf{x},t^n)\,(\nabla\cdot\mathbf{u}) & = 0, \\ \partial_t\mathbf{u} + \tau(\mathbf{x},t^n)\,\nabla\pi & = -\dfrac{g}{\tau}\tau(\mathbf{x},t^n)\,\nabla z, \\ \partial_t\pi + a^2\tau(\mathbf{x},t^n)\,(\nabla\cdot\mathbf{u}) & = -\dfrac{\pi - \frac{g}{2\tau^2}}{\varepsilon}, \\ \partial_t z & = 0, \end{cases} \tag{0.2.3}$$

où ε est un temps caractéristique et a une constante qui doit satisfaire une condition de stabilité dite de Witham, $a > hc$, afin que le système ci-dessus soit une approximation visqueuse du système acoustique (0.2.2), voir par exemple le livre [5].

On approche donc le système acoustique (0.2.1) par le système relaxé associé (0.2.3) qui a pour propriétés d'être linéaire et hyperbolique lorsque l'on omet les termes sources, avec comme valeurs propres $\{-a,0,a\}$, où a est directement relié à c la vitesse des ondes de gravité. Comme ces ondes sont équivalentes aux ondes acoustiques pour le système d'Euler, on appelle par analogie le système (0.2.1) acoustique.

0.3 Méthodes numériques

Si on note $\mathbf{U}$ le vecteur des variables conservatives $(h,hu,z)^T \in \mathbb{R}^3$, on peut réécrire le système d'équations de Saint-Venant en une dimension d'espace sous la forme compacte :

$$\partial_t\mathbf{U} + \partial_x\mathbf{F}(\mathbf{U}) = \mathbf{S}(\mathbf{U}),$$

avec

$$\mathbf{F}(\mathbf{U}) = \begin{pmatrix} (hu) \\ \frac{(hu)^2}{h} + \frac{gh^2}{2} \\ 0 \end{pmatrix} \qquad \text{et} \qquad \mathbf{S}(\mathbf{U}) = \begin{pmatrix} 0 \\ -gh\,\partial_x z \\ 0 \end{pmatrix}.$$

Pour résoudre ce système, on utilise généralement des méthodes dites Volumes Finis (FV pour « Finite Volumes » en anglais). On peut trouver plusieurs livres de référence qui détaillent ce type de méthodes pour les systèmes hyperboliques, par exemple [17, 20, 25]. On discrétise le domaine physique $[0, L]$ en N mailles uniformes $\left(\left]x_{j-1/2}, x_{j+1/2}\right[\right)_{0\leq j\leq N-1}$, avec $x_{j+1/2} = (j+1)\Delta x$ les points des interfaces, $x_j = (j+\frac{1}{2})\Delta x$ les centres des mailles et $\Delta x = \frac{L}{N}$ le pas d'espace. En intégrant sur une maille $\left]x_{j-1/2}, x_{j+1/2}\right[$ entre deux pas de temps t^n et $t^{n+1} = t+\Delta t$, on obtient :

$$\int_{t^n}^{t^{n+1}} \int_{x_{j-1/2}}^{x_{j+1/2}} \partial_t \mathbf{U} \,\mathrm{d}x \,\mathrm{d}t + \int_{t^n}^{t^{n+1}} \int_{x_{j-1/2}}^{x_{j+1/2}} \partial_x \mathbf{F}(\mathbf{U}) \,\mathrm{d}x \,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{S}(\mathbf{U}) \,\mathrm{d}x \,\mathrm{d}t.$$

On intervertit les intégrales dans le premier membre pour obtenir :

$$\Delta x \left(\mathbf{U}_j^{n+1} - \mathbf{U}_j^n\right) + \int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{U})(x_{j+1/2}, t) \,\mathrm{d}t - \int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{U})(x_{j-1/2}, t) \,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{S}(\mathbf{U}) \,\mathrm{d}x \,\mathrm{d}t,$$

où

$$\mathbf{U}_j^n = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{U}(x, t^n) \,\mathrm{d}x$$

est la valeur moyenne de $\mathbf{U}$ au temps t^n sur la maille $\left]x_{j-1/2}, x_{j+1/2}\right[$. On approche les intégrales sur le temps $\int_{t^n}^{t^{n+1}} f(t) \,\mathrm{d}t$ grâce à une quadrature à gauche explicite, soit $\Delta t \, f(t^n)$, pour avoir :

$$\mathbf{U}_j^{n+1} - \mathbf{U}_j^n + \frac{\Delta t}{\Delta x} \left(\mathbf{F}(\mathbf{U})(x_{j+1/2}, t^n) - \mathbf{F}(\mathbf{U})(x_{j-1/2}, t^n)\right) = \frac{\Delta t}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{S}(\mathbf{U})(x, t^n) \,\mathrm{d}x.$$

Si on dénote par $\mathbf{F}_{j+1/2}^n$ le flux numérique, qui est une approximation du flux physique

$$\mathbf{F}(\mathbf{U})(x_{j+1/2}, t^n),$$

et par $\mathbf{S}_j^n$ une approximation de la valeur moyenne du terme source

$$\frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{S}(\mathbf{U})(x, t^n) \,\mathrm{d}x,$$

tous deux fonctions des valeurs moyennes des variables conservatives $\mathbf{U}_{0\leq k\leq N-1}^n$ au temps t^n, alors on définit un schéma FV explicite qui s'écrit sous la forme :

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x} \left(\mathbf{F}_{j+1/2}^n - \mathbf{F}_{j-1/2}^n\right) + \Delta t \, \mathbf{S}_j^n.$$

On peut généralement montrer que le schéma est stable, c'est-à-dire que les solutions restent bornées au cours de la simulation, sous une condition sur le pas de temps Δt, appelée condition de Courant-Friedrichs-Lewy. On peut par exemple regarder le livre [20] pour une définition de cette condition dans le cadre des méthodes FV. Si l'on veut s'affranchir d'une telle restriction, on peut se tourner vers un schéma implicite, qui sera beaucoup plus stable en échange de

solutions plus diffusées. Pour cela, on remplace le temps auquel on calcule les flux et le terme source au temps final t^{n+1}, de sorte que la mise à jour devient :

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{F}_{j+1/2}^{n+1} - \mathbf{F}_{j-1/2}^{n+1}\right) + \Delta t\, \mathbf{S}_j^{n+1}.$$

On doit alors résoudre une équation qui est soit linéaire, il faut dans ce cas utiliser une méthode de résolution standard comme la méthode itérative de Jacobi, dont on peut trouver une description dans le livre [7]. Soit non-linéaire, il faut alors trouver le 0 d'une fonction par exemple grâce à une méthode de Newton, décrite entre autre dans le livre [23]. Les efforts effectués sur le système acoustique dans la Section 0.2 prennent tout leur sens puisqu'ils permettent, en utilisant une méthode implicite autorisant un grand pas de temps, de garder un coût de calcul raisonnable car le système est linéaire.

Il existe différentes manière de définir les fonctions flux ci-dessus, on peut citer les méthodes avec flux amonts, de Lax-Friedrichs, de Osher ou de Roe. Une autre méthode classique est celle de Godunov, dont on peut obtenir plus de détails dans les livres [17, 20, 25]. On considère le problème de Riemann associé au système hyperbolique, c'est à-dire que l'on souhaite trouver la solution continue au problème

$$\begin{cases} \partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = \mathbf{S}(\mathbf{U}), \\ \mathbf{U}(x,0) = \begin{cases} \mathbf{U}_L, & \text{si } x \leq \frac{L}{2}, \\ \mathbf{U}_R, & \text{si } x > \frac{L}{2}. \end{cases} \end{cases}$$

On peut ainsi utiliser la solution exacte ou une solution approchée des problèmes de Riemann à chaque interface du maillage pour calculer les flux $\mathbf{F}_{j+1/2}^n$. Il est connu que cette méthode est d'ordre 1 en espace et en temps, c'est-à-dire que l'erreur de consistance entre le schéma numérique et l'équation continue s'écrit comme $O(\Delta t) + O(\Delta x)$. On remarque encore une fois l'intérêt d'avoir linéarisé le système acoustique puisque dans ce cas la solution exacte du problème de Riemann est facile à calculer.

Il y a différentes manières de monter en ordre en espace, c'est-à-dire d'améliorer la vitesse avec laquelle la solution numérique converge vers la solution continue lorsque l'on fait tendre la pas d'espace Δx vers 0. Dans le but d'obtenir de manière plus efficace une solution du système que l'on souhaite résoudre, on aimerait avoir des puissances plus élevées dans les erreurs de troncature en temps et en espace. C'est ce que l'on appelle méthodes d'ordres élevés, on peut citer par exemples les méthodes FV avec reconstruction des pentes à chaque interface. On calcule ces pentes grâce à un polynôme d'ordre suffisant construit à l'aide des valeurs moyennes sur plusieurs cellules. On peut trouver la description de méthodes VF d'ordre 2 dans le livre [17]. Il existe également les méthodes Galerkin discontinue (DG pour « Discontinuous Galerkin » en anglais) qui approchent la solution par un polynôme dans chaque maille, décrites par exemple dans l'article [9]. Leur avantage réside dans le fait que pour calculer les flux numériques à une interface, on n'utilise que des valeurs de chaque côté de l'interface, contrairement aux méthodes FV ordre élevé qui demandent de construire un polynôme à l'aide plusieurs mailles. Par conséquent, les calculs s'effectuent entre degrés de libertés plus localisés pour la méthodes DG, ce qui permettrait par exemple d'utiliser efficacement la parallélisation de ces calculs. En

revanche, la condition de stabilité sur le pas de temps Δt n'est pas théorique mais empirique en DG, avec des valeurs optimales données dans l'articles [9]. Enfin, pour toutes les méthodes ordres élevés, FV ou DG, il faut rajouter des limiteurs de pentes, c'est-à-dire qu'il faut modifier la solution finale obtenue pour éviter des phénomènes oscillatoires non physiques. Comme soulevé dans le livre [17] et l'article [9], on verra qu'il s'agit d'un enjeu majeur pour ces méthodes.

0.4 Écoulements à bas nombre de Froude

En adimensionnant le système d'équations de Saint-Venant (0.1.1) on peut faire apparaître une quantité caractéristique appelée le nombre de Froude (équivalent du nombre de Mach pour Euler). On considère par exemple pour la hauteur d'eau h qu'il existe une grandeur caractéristique h_0 lui correspondant, et on note $\tilde{h} = \frac{h}{h_0}$ la hauteur d'eau adimensionnée associée. On obtient ainsi le nouveau système :

$$\begin{cases} \partial_{\tilde{t}}\tilde{h} + \nabla_{\tilde{\mathbf{x}}} \cdot \left(\tilde{h}\tilde{\mathbf{u}}\right) & = 0, \\ \partial_{\tilde{t}}\left(\tilde{h}\tilde{\mathbf{u}}\right) + \nabla_{\tilde{\mathbf{x}}} \cdot \left(\tilde{h}\tilde{\mathbf{u}} \otimes \tilde{\mathbf{u}}\right) + \dfrac{1}{\mathrm{Fr}^2}\nabla_{\tilde{\mathbf{x}}}\dfrac{\tilde{h}^2}{2} & = -\dfrac{1}{\mathrm{Fr}^2}\tilde{h}\,\nabla_{\tilde{\mathbf{x}}}\tilde{z}, \end{cases}$$

où le nombre de Froude est défini par $\mathrm{Fr} = \frac{u_0}{c_0}$, soit le rapport entre la vitesse matérielle caractéristique et la vitesse caractéristique des ondes de gravité du problème. Pour un cours d'eau, la vitesse matérielle est celle à laquelle s'écoule la rivière. La vitesse de gravité, quant à elle, correspond à celle des petites ondes générées, par exemple, par une goutte tombée à la surface de l'eau.

Dans beaucoup de cas, comme celui d'un cours d'eau, le nombre de Froude est petit devant 1 et s'ensuit deux principales difficultés pour la simulation numérique lorsque ce nombre se rapproche de 0. La première difficulté concerne la stabilité et le pas de temps avec lequel on avance à chaque itération pour arriver à l'approximation de la solution au temps final. La contrainte de stabilité CFL s'exprime classiquement $\Delta t \leq \frac{\Delta x}{\max_j |u_j| + c_j}$ et lorsque la vitesse de gravité c croît localement alors que la vitesse matière u reste bornée, le pas de temps Δt est fortement contraint. Le schéma nécessite d'autant plus d'itérations que c est grand, ce qui implique un coût de calcul conséquent et une erreur accumulée à chaque itération importante. C'est un cas où le nombre de Froude est petit et qui a déjà motivé la décomposition d'opérateur Lagrange-projection présentée dans la section 0.2 afin d'obtenir un pas de temps indépendant de c. La deuxième difficulté concerne la consistance et la dépendance de l'erreur de troncature, reliée à l'erreur entre la solution approchée et la solution exacte, avec le nombre de Froude. On souhaite que cette erreur ne grandisse pas lorsque ce nombre est petit et ainsi garder une convergence indépendante de ce nombre vers la solution exacte de notre approximation lorsque l'on raffine le maillage. Les corrections apportées au schéma à cette fin sont issues des articles [13, 14]. Un schéma ayant une contrainte de stabilité sur le pas de temps et une erreur de troncature toutes deux indépendantes du nombre de Froude est appelé schéma tout régime. L'étude des schémas tout régime dans le cadre du système d'Euler qui nous intéresseront dans la suite de cette thèse a été réalisée dans la thèse de M. Girardin [16]. Une autre stratégie, particulièrement adaptée à ce type de système avec un paramètre tendant vers 0, consiste à construire un schéma préservant l'asymptotique (en anglais « asymptotic preserving »). Ces schémas ont

pour propriété de converger, lorsque le paramètre tend vers 0, vers un schéma consistant avec le système continu limite, généralement parabolique, obtenu grâce à un développement limité de chaque variable par rapport au paramètre. On ne cherche pas à construire ici un tel schéma mais on renvoie le lecteur vers les articles [19, 4, 18, 11, 15] et leurs références pour plus de détails.

Dans les sections suivantes, on résume en français les 5 chapitres présents dans cette thèse, dont les 4 premiers sont rédigés en anglais sous forme d'article.

0.5 Chapitre 1 : Un schéma de type Lagrange-projection à grands pas de temps et équilibre pour les équations de Saint-Venant à une dimension d'espace

Dans ce premier chapitre on s'intéresse à la résolution du système d'équations de Saint-Venant qui s'écrit en une dimension d'espace :

$$\begin{cases} \partial_t h + \partial_x(hu) & = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + g\dfrac{h^2}{2}\right) & = -gh\partial_x z, \end{cases} \tag{0.5.1}$$

où h est la hauteur d'eau au dessus de la topographie z et $g > 0$ la constante de gravitation.

On souhaite adapter au contexte Saint-Venant avec terme source le schéma Lagrange-projection décrit dans la Section 0.2 et dans les travaux liés à la thèse de M. Girardin [6] pour le système d'Euler complet. Il s'agit donc de développer un schéma implicite-explicite Volumes Finis d'ordre 1 en espace et en temps avec la même méthodologie et plusieurs objectifs détaillés ci-dessous.

Le premier objectif est de construire un schéma qui soit consistant avec le système ci-dessus et qui satisfait une version discrète d'une des inégalités d'entropie décrites dans la Section 0.1. Par consistant, on entend que la solution discrète converge vers la solution continue lorsque l'on fait tendre le pas de temps Δt et le pas d'espace Δx vers 0. On montrera que l'inégalité discrète vérifiée est celle associée à l'inégalité non conservative (0.1.2). On pourra noter que l'on retrouve bien sur le schéma pour le système d'Euler qui satisfait cette inégalité sans terme non conservatif lorsque la topographie est constante.

Comme deuxième objectif, on souhaite que le schéma conserve de manière exacte les états dits de « lacs au repos », que l'on a définis dans la Section 0.1. Pour obtenir ce résultat, on verra que le traitement du terme source dans la résolution du problème de Riemann associé au pas acoustique relaxé (0.2.3) jouera un rôle fondamental.

Un troisième objectif est d'assurer la stricte positivité de la hauteur d'eau h à chaque pas de temps. Cette condition d'admissibilité est nécessaire pour assurer que la solution trouvée ait un sens physique et que la simulation puisse être menée à terme quelques soient les conditions

initiales physiquement acceptables. En effet, on a vu dans la Section 0.2 que les valeurs propres du système acoustique sont $\{-a, 0, a\}$ avec a qui dépend de c, la vitesse des ondes de gravité. Cette dernière doit rester réelle positive, et ceci n'est plus vrai lorsque la hauteur d'eau devient strictement négative. De plus, l'utilisation de la variable de volume spécifique $\tau = \frac{1}{h}$ dans le pas acoustique interdit à la hauteur d'eau de devenir nulle. C'est notamment pour cette raison que les cas de transitions entres zones humides et seches sont compliqués à traiter avec l'utilisation de coordonnées lagrangiennes et ne seront pas évoqués dans la suite. Pour plus de détails sur les notions de stabilité pour les schémas Volumes Finis, on peut par exemple se référer au livre [5].

Enfin, on souhaite montrer que le schéma proposé est particulièrement bien conçu pour les écoulements à bas nombre de Froude. La condition CFL sur le pas de temps Δt fait intervenir l'inverse de la vitesse des ondes les plus rapides. À bas nombre de Froude, les ondes les plus rapides ont pour vitesse celle des ondes de gravité, à savoir c. Grâce au traitement implicite-explicite, la condition CFL est indépendante de ce nombre et le schéma reste stable pour des pas de temps longs liés uniquement à la vitesse matière u.

Pour conclure, on illustre les bonnes propriétés du schéma par plusieurs cas tests standards en une dimension d'espace. On fait notamment passer plusieurs cas tests d'écoulements stationnaires à vitesses non nulles. Bien que le schéma ne conserve pas strictement ces états stationnaires, on observe de bon résultats pour retrouver ces derniers.

Ce chapitre contient 3 annexes, la première concerne l'étude des espaces propres et des invariants de Riemann du système acoustique relaxé (0.2.3). On remarque notamment qu'il manque un invariant pour pouvoir trouver tous les états intermédiaires définissant la solution exacte du problème de Riemann associé. On présente dans la troisième annexes différentes approches pour pouvoir trouver une solution exacte qui permette d'obtenir les objectifs fixés ci-dessus. Enfin, la deuxième annexe contient la preuve détaillée de l'inégalité d'entropie discrète associée à l'inégalité continue (0.1.2).

0.6 Chapitre 2 : Un schéma Lagrange-projection tout régime et équilibre pour les équations de Saint-Venant sur maillages non-structurés

Dans ce deuxième chapitre on s'intéresse à la résolution du système d'équations de Saint-Venant (0.1.1). On étend à la dimension 2 les résultats du Chapitre 1, il s'agit donc de développer un schéma implicite-explicite Volumes Finis d'ordre 1 en espace et en temps sur maillages non structurés avec plusieurs objectifs détaillés ci-dessous.

Le premier objectif est de construire un schéma qui soit consistant avec le système ci-dessus et qui conserve de manière exacte les états dits de « lacs au repos » en 2D, soit $\mathbf{u} = \mathbf{0}$ et $H = h + z$ constant.

Un deuxième objectif est d'assurer la stricte positivité de la hauteur d'eau h à chaque pas de temps, toujours pour assurer la stabilité du schéma.

Comme troisième objectif, on souhaite que le schéma soit tout régime, à savoir que le pas de temps et les erreurs de consistance ne dépendent plus du nombre de Froude. On rappelle que ce dernier est défini avec les équations adimensionnées dans la Section 0.1. Ces deux propriétés sont indispensables pour obtenir une bonne approximation de la solution du problème tout en conservant un coût de calcul raisonnable lorsque le nombre de Froude est faible. On vérifiera que le schéma permet ainsi de traiter des problèmes avec plusieurs régimes d'écoulements vis-à-vis de ce nombre. Pour parvenir à ce résultat, les deux ingrédients principaux sont : l'approche acoustique-transport avec traitement implicite de l'étape acoustique, et la correction de la partie diffusive des flux de pressions inspirée des articles [13, 14].

Pour conclure, on illustre les bonnes propriétés du schéma par plusieurs cas tests en deux dimensions d'espace. On traite notamment un cas test quasi-stationnaire sur maillage non-structuré avec une topographie non triviale. Ensuite on propose un cas test de vortex en translation pour lequel on connaît la solution exacte sans topographie. On observe bien les effets dus aux erreurs de consistance sur le cas test du vortex lorsque l'on n'applique pas la correction des flux.

0.7 Chapitre 3 : Un schéma Lagrange-projection Galerkin discontinu équilibre pour les équation de Saint-Venant

Dans ce troisième chapitre on s'intéresse, comme dans le Chapitre 1, à la résolution du système d'équations de Saint-Venant (0.5.1) en une dimension d'espace. On propose une extension à l'ordre élevé en temps et en espace du schéma Lagrange-projection Volumes Finis (FV) ordre 1. On utilise une méthode Galerkin discontinue (DG) pour la discrétisation en espace et Runge-Kutta (RK) en temps. L'écriture de ce schéma pour le système d'Euler isentropique, soit Saint-Venant sans terme source, est donné dans l'Annexe 3.B. Une extension similaire a déjà été étudiée dans ce cadre par F. Renac [22]. On ajoute ici le traitement du terme source avec plusieurs objectifs détaillés ci-dessous.

Le premier objectif est de construire un schéma qui soit consistant avec le système de Saint-Venant à une dimension d'espace et qui satisfait une version discrète d'une des inégalités d'entropie décrites dans la Section 0.1. On montrera que l'inégalité discrète vérifiée est une version faible de celle associée à l'inégalité non conservative (0.1.2). On pourra d'ailleurs noter que l'on retombe bien à la fois sur le schéma DG pour le système d'Euler isentropique lorsque la topographie est constante, et sur le schéma FV pour le système de Saint-Venant lorsque l'on redescend à l'ordre 1.

Comme deuxième objectif, on souhaite que le schéma conserve de manière exacte les états dits de « lacs au repos », que l'on a défini dans la Section 0.1. On verra qu'il faut traiter de manière judicieuse l'intégrale approchant le terme source pour obtenir ce résultat quelque soit la régularité des conditions initiales.

Un troisième objectif est d'assurer la stricte positivité de la hauteur d'eau h à chaque pas de temps. Cette condition d'admissibilité va faire apparaître une condition de stabilité sur le pas de temps Δt, proche de la condition CFL empirique en DG.

Enfin, on souhaite montrer que le schéma proposé est particulièrement bien conçu pour les écoulements à bas nombre de Froude. De nouveau grâce au traitement implicite-explicite, la condition CFL est indépendante de ce nombre et le schéma reste stable pour des pas de temps longs liés uniquement à la vitesse matière u.

Pour conclure, on illustre les bonnes propriétés du schéma par plusieurs cas tests standards en une dimension d'espace. On commence par confirmer numériquement les résultats obtenus théoriquement sur la préservations des états stationnaires de « lacs au repos ». On fait également passer plusieurs cas tests d'écoulements stationnaires à vitesses non nulles, en comparant les solutions obtenues en fonction de l'ordre des méthodes utilisées. On remarque enfin les difficultés rencontrées pour le cas test « transcritique » avec choc pour lequel les limiteurs ne suffisent pas à atténuer les oscillations non physiques.

Ce chapitre est supplémenté de 3 annexes, dont la première contient la preuve détaillée de l'inégalité d'entropie discrète associée à l'inégalité continue (0.1.2). Une deuxième annexe est composée de l'article court sur l'écriture du schéma DG dans le cadre des équations d'Euler isentropique. Enfin la dernière annexe présente des courbes d'efficacité sur un cas test de type problème de Riemann dont on est capable de calculer la solution exacte. Ces courbes permettent d'illustrer l'intérêt de l'utilisation de ces méthodes ordres élevés pour obtenir de manière efficace une bonne approximation de la solution.

0.8 Chapitre 4 : Un schéma Lagrange-projection Galerkin discontinu pour les équation de Saint-Venant utilisant une détection de type MOOD

Dans ce quatrième chapitre on s'intéresse de nouveau à la résolution du système d'équations de Saint-Venant en une dimension d'espace, comme dans le Chapitre 1 et 3. On garde l'extension à l'ordre élevé en temps et en espace en réutilisant la méthode Galerkin discontinue (DG) proposée dans le Chapitre 3. Cependant, on présente une nouvelle méthodologie afin de corriger les problématiques liées aux méthodes d'ordres élevés et à leur instabilité.

Les résultats obtenus sur le schéma DG concernent principalement les valeurs moyennes des polynômes mis en jeu. Par exemple, sous la condition CFL proposée, la positivité de la hauteur d'eau h est assurée uniquement sur la valeur moyenne et non sur chaque degré de liberté. Ainsi, les corrections apportées à la solution dans le Chapitre 3, consistent à rapprocher les polynômes de leur valeur moyenne autant que nécessaire. De fait, ces corrections sont effectuées de manière globale dans chaque maille et font généralement descendre l'ordre de la méthode ainsi que la précision de la solution. Enfin, il est possible d'exhiber des cas dans lesquels la stabilité même du schéma ne peut être assurée par ce type de limitation.

La méthode MOOD (Multi-dimensionnal Optimal Order Detection [8]) est une technique de correction de la solution dite *a posteriori*. En effet, l'idée générale est de calculer à chaque pas de temps la solution sans aucune correction, de détecter les mailles dans lesquels il y a une solution non souhaitable, et de recalculer seulement ces mailles à l'aide d'une méthode plus robuste. On peut éventuellement utiliser comme méthode de substitution une méthode plus coûteuse, d'ordre moins élevé ou moins précise, mais qui doit avoir néanmoins comme caractéristique de fournir une solution admissible stable.

Dans ce contexte, on se propose de recalculer les mailles détectées par la méthode Volumes Finis (FV) du Chapitre 1 sur plusieurs sous-mailles. On souhaite ainsi conserver une précision raisonnable et un pas de temps compatible avec la méthode DG en jouant sur le nombre de sous-mailles, tout en jouissant des bonnes propriétés du schéma FV. On précisera tout d'abord les étapes clés nécessaires à la réalisation de cette méthodologie. On détaillera ensuite deux manières différentes de réaliser ces corrections en fonction de l'ordre des étapes permettant de mettre à jour la solution.

Pour conclure, on présente quelques résultats numériques comparant l'utilisation des limiteurs standards avec la méthode MOOD. Seule l'une des deux approches a pu être testée numériquement mais on espère prochainement obtenir de meilleurs résultats avec la deuxième.

0.9 Chapitre 5 : Implémentation

Dans ce cinquième chapitre on met en lumière différents aspects techniques liés à l'implémentation de deux codes de simulation, l'un pour la dimension 1 d'espace, l'autre pour la dimension 2.

En ce qui concerne la dimension 1 d'espace, on commence tout d'abord par l'écriture des schémas en 1D, dans l'ordre suivant : Volumes Finis (FV) explicite-explicite , FV implicite-explicite (IMEX) et Galerkin discontinu (DG) IMEX. On précise pour chacun d'eux la façon d'implémenter les deux sous-étapes associées à la méthodologie acoustique-transport, les structures à mettre en mémoire, ainsi que le traitement numérique des conditions de bords. On poursuit sur l'utilisation de la librairie PETSc [2] permettant de résoudre le système linéaire dans le cas du schéma DG IMEX. On termine enfin par quelques modifications apportées au code C++ afin de pouvoir utiliser le schéma DG combiné avec la méthode MOOD du Chapitre 4.

Pour la dimension 2, on présente l'assemblage de la matrice creuse par blocs apparaissant dans le système linéaire, que l'on résout de nouveau à l'aide de la librairie PETSc. On conclut en explicitant le traitement numérique du terme source.

Bibliographie

[1] Emmanuel Audusse. *Hyperbolic models and numerical analysis for shallow water flows.* Theses, Université Pierre et Marie Curie - Paris VI, September 2004.

[2] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Dave A. May, Lois Curfman McInnes, Richard Tran Mills, Todd Munson, Karl Rupp, Patrick Sanan, Barry F. Smith, Stefano Zampini, Hong Zhang, and Hong Zhang. PETSc Web page. `http://www.mcs.anl.gov/petsc`, 2018.

[3] Christophe Berthon and Christophe Chalons. A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations. *Mathematics of Computation*, 85(299) :1281–1307, 2016.

[4] Christophe Berthon and Rodolphe Turpault. Asymptotic preserving HLL schemes. *Numerical methods for partial differential equations*, 27(6) :1396–1422, 2011.

[5] François Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Springer Science & Business Media, 2004.

[6] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An all-regime Lagrange-projection like scheme for the gas dynamics equations on unstructured meshes. *Communications in Computational Physics*, 20(1) :188–233, 2016.

[7] Philippe G Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. masson, 1982.

[8] Stéphane Clain, Steven Diot, and Raphaël Loubère. A high-order finite volume method for systems of conservation laws—Multi-dimensional Optimal Order Detection (MOOD). *Journal of computational Physics*, 230(10) :4028–4050, 2011.

[9] Bernardo Cockburn and Chi-Wang Shu. Runge–Kutta discontinuous galerkin methods for convection-dominated problems. *Journal of scientific computing*, 16(3) :173–261, 2001.

[10] Frédéric Coquel, Quang Nguyen, Marie Postel, and Quang Tran. Entropy-satisfying relaxation method with large time-steps for Euler IBVPs. *Mathematics of Computation*, 79(271) :1493–1533, 2010.

[11] Floraine Cordier, Pierre Degond, and Anela Kumbaro. An asymptotic-preserving all-speed scheme for the Euler and Navier–Stokes equations. *Journal of Computational Physics*, 231(17) :5685–5704, 2012.

[12] AJC Barré de Saint-Venant. Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leurs lits. *Comptes Rendus des séances de l'Académie des Sciences*, 73 :237–240, 1871.

[13] Stéphane Dellacherie. Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. *Journal of Computational Physics*, 229(4) :978–1016, 2010.

[14] Stéphane Dellacherie, Jonathan Jung, Pascal Omnes, and P-A Raviart. Construction of modified Godunov-type schemes accurate at any Mach number for the compressible Euler system. *Mathematical Models and Methods in Applied Sciences*, 26(13) :2525–2615, 2016.

[15] Arnaud Duran, Fabien Marche, Rodolphe Turpault, and Christophe Berthon. Asymptotic preserving scheme for the shallow water equations with source terms on unstructured meshes. *Journal of Computational Physics*, 287 :184–206, 2015.

[16] Mathieu Girardin. *Asymptotic preserving and all-regime Lagrange-Projection like numerical schemes : application to two-phase flows in low mach regime*. Theses, Université Pierre et Marie Curie - Paris VI, December 2014.

[17] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, volume 118. Springer Science & Business Media, 1996.

[18] Jeffrey Haack, Shi Jin, and Jian-Guo Liu. An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations. *Communications in Computational Physics*, 12(4) :955–980, 2012.

[19] Shi Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing*, 21(2) :441–454, 1999.

[20] Randall J LeVeque. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge university press, 2002.

[21] Victor Michel-Dansac, Christophe Berthon, Stéphane Clain, and Françoise Foucher. A well-balanced scheme for the shallow-water equations with topography. *Computers & Mathematics with Applications*, 72(3) :568–593, 2016.

[22] Florent Renac. A robust high-order Lagrange-projection like scheme with large time steps for the isentropic Euler equations. *Numerische Mathematik*, 135(2) :493–519, 2017.

[23] François Rouvière. *Petit guide de calcul différentiel : à l'usage de la licence et de l'agrégation.* Cassini, 2009.

[24] I. Suliciu. On the thermodynamics of fluids with relaxation and phase transitions. Fluids with relaxation. *Internat. J. Engrg. Sci*, 36 :921–947, 1998.

[25] Eleuterio F Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics : A Practical Introduction.* Springer Science & Business Media, 2009.

Première partie

Schémas Volumes Finis Lagrange-projection d'ordre 1 pour les équations de Saint-Venant

Chapitre 1

Un schéma de type Lagrange-projection à grands pas de temps et équilibre pour les équations de Saint-Venant à une dimension d'espace

Ce chapitre a fait l'objet d'une publication dont les références sont : Christophe Chalons, Pierre Kestener, Samuel Kokh and Maxime Stauffert. A large time-step and well-balanced Lagrange-projection type scheme for the shallow water equations. *Communication in Mathematical Sciences*, 15(3) :765–788, 2017.

A large time-step and well-balanced Lagrange-projection type scheme for the shallow water equations in one space dimension

Abstract

This work focuses on the numerical approximation of the shallow water equations (SWE) using a Lagrange-projection type approach. We propose to extend to this context the recent implicit-explicit schemes developed in [16, 18] in the framework of compressible flows, with or without stiff source terms. These methods enable the use of time steps that are no longer constrained by the sound velocity thanks to an implicit treatment of the acoustic waves, and maintain accuracy in the subsonic regime thanks to an explicit treatment of the material waves. In the present setting, a particular attention will be also given to the discretization of the non-conservative terms in SWE and more specifically to the well-known well-balanced property. We prove that the proposed numerical strategy enjoys important non linear stability properties and we illustrate its behaviour past several relevant test cases.

1.1 Introduction

We are interested in the design of a numerical scheme for the well-known shallow water equations (SWE), given by

$$\begin{cases} \partial_t h + \partial_x (hu) = 0, & (1.1.1a) \\ \partial_t (hu) + \partial_x \left(hu^2 + g\dfrac{h^2}{2} \right) = -gh\partial_x z, & (1.1.1b) \end{cases}$$

where $z(x)$ denotes a given smooth topography and $g > 0$ is the gravity constant. The primitive variables are the water depth $h \geq 0$ and its velocity u, which both depend on the space and time variables, respectively $x \in \mathbb{R}$ and $t \in [0, \infty)$. At time $t = 0$, we assume that the initial water depth $h(x, t = 0) = h_0(x)$ and velocity $u(x, t = 0) = u_0(x)$ are given. In order to shorten the notations, we will use the following condensed form of Equation (1.1.1), namely

$$\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = \mathbf{S}(\mathbf{U}, z), \tag{1.1.2}$$

where $\mathbf{U} = (h, hu)^T$, $\mathbf{F}(\mathbf{U}) = (hu, hu^2 + gh^2/2)^T$ and $\mathbf{S}(\mathbf{U}, z) = (0, -gh\partial_x z)^T$. This system is supplemented with the validity of entropy inequalities which can be written either in a non-conservative form as follows,

$$\partial_t \mathcal{U}(\mathbf{U}) + \partial_x \mathcal{F}(\mathbf{U}) \leqslant -ghu\partial_x z, \tag{1.1.3}$$

with the non-conservative entropy $\mathcal{U}$ and the associated flux $\mathcal{F}$ defined by

$$\mathcal{U}(\mathbf{U}) = \frac{hu^2}{2} + \frac{gh^2}{2}, \quad \mathcal{F}(\mathbf{U}) = \left(\frac{u^2}{2} + gh\right) hu,$$

or in conservative form as follows,

$$\partial_t \tilde{\mathcal{U}}(\mathbf{U}, z) + \partial_x \tilde{\mathcal{F}}(\mathbf{U}, z) \leqslant 0, \tag{1.1.4}$$

where the conservative entropy $\tilde{\mathcal{U}}$ and the associated flux $\tilde{\mathcal{F}}$ now depend on z and are defined by,

$$\tilde{\mathcal{U}}(\mathbf{U}, z) = \mathcal{U}(\mathbf{U}) + ghz \quad \text{and} \quad \tilde{\mathcal{F}}(\mathbf{U}, z) = \mathcal{F}(\mathbf{U}) + ghuz.$$

A first objective is that the proposed numerical scheme should be consistent with Equation (1.1.2) and should satisfy a discrete form of (at least) one of these entropy inequalities.

The steady states of Equation (1.1.2) are governed by the ordinary differential system $\partial_x F(\mathbf{U}) = \mathbf{S}(\mathbf{U}, z)$, namely

$$hu = \text{constant}, \quad \frac{u^2}{2} + g(h + z) = \text{constant}.$$

In this paper, we will be more specifically interested in the so-called "lake at rest" steady solution defined by

$$h + z = \text{constant}, \quad u = 0. \tag{1.1.5}$$

As a second objective, we want the proposed numerical scheme to preserve discrete initial conditions matching condition (1.1.5), which corresponds to the very well-known well-balanced property (see for instance the recent book [28] for a review).

A third objective of the method is to ensure the positivity of the water height if the initial water height is positive.

Last but not least, we are especially interested in this work in subsonic or near low-Froude number flows. In this case, it turns out that the usual CFL time step limitation of Godunov-type numerical schemes is driven by the acoustic waves and can thus be very restrictive. We are thus interested in the design of a mixed implicit-explicit large time-step strategy following the lines of the pionneering work [23] and the more recent ones [14, 16, 17, 18]. By large time-step, we mean that the scheme should be stable under a CFL stability condition driven by the (slow) material waves, and not by the (fast) acoustic waves as it is customary in Godunov-type schemes. Numerical evidences will show a gain in efficiency.

There is a huge amount of works about the design of numerical schemes for the SWE, and most of the schemes intended to satisfy the first three properties above. To mention only a few of them, we refer the reader to the following well-known contributions [3, 29, 27, 30, 32, 24, 25, 1, 21, 35, 11, 10, 37, 36, 6, 7, 34, 2, 4, 5]. We also refer to the books [8] and [28] which provide additional references and very nice overviews.

The design of mixed implicit-explicit (IMEX) schemes based on a Lagrange-projection type approach which are stable under a CFL restriction driven by the slow material waves and not the acoustic waves has been given a first interest in the pionneering work [23] and was further developed for the computation of large friction or low-Mach regimes in [14, 16, 17, 18, 19] for single or two-phase flow models. It is the purpose of this paper to adapt these IMEX strategies to the shallow-water equations while preserving the first three properties above, namely the lake-at-rest well-balanced property, the positivity of the water height, and the validity of a discrete form of the entropy inequality. Another new large time step method for the shallow water flows in the low Froude number limit has been proposed in [7]. The strategy is also mixed implicit-explicit considering the fast acoustic waves and the slow transport waves respectively, but does not rely on the natural Lagrange-projection like decomposition proposed here. Note also that we focus here on subsonic or low Froude number flows, but we do not consider the low Froude number limit which is the purpose of a current work in progress. We also refer the reader to the recent contribution [38] which proves rigorously that the IMEX Lagrange projection scheme is AP for one-dimensional low-Mach isentropic Euler and low-Froude shallow water equations.

1.2 Operator splitting Lagrange-projection approach and relaxation procedure

In this section we adapt the so-called operator splitting Lagrange-projection strategy presented in [16] to the shallow water equations (1.1.1). This splitting involves a so-called Lagrange step system that accounts for the acoustic waves and topography variations for which we shall propose an approximation based on a Suliciu [33] relaxation approach using the notion of consistency in the integral sense [24, 25], and a so-called transport step accounting for the (slow) transport phenomenon.

Before describing the numerical method, we introduce classic notations pertaining to our discretization context. Space and time are discretized using a space step Δx and a time step Δt into a set of cells $[x_{j-1/2}, x_{j+1/2})$ and instants $t^{n+1} = n\Delta t$, where $x_{j+1/2} = j\Delta x$ and $x_j = (x_{j-1/2} + x_{j+1/2})/2$ are respectively the cell interfaces and cell centers, for $j \in \mathbb{Z}$ and $n \in \mathbb{N}$. For a given initial condition $x \mapsto \mathbf{U}^0(x)$, we consider a discrete initial data $\mathbf{U}_j^0$ defined by $\mathbf{U}_j^0 = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{U}^0(x)\,\mathrm{d}x$, for $j \in \mathbb{Z}$. The algorithm proposed in this paper aims at computing an approximation $\mathbf{U}_j^n$ of $\frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{U}(x,t^n)\,\mathrm{d}x$ where $x \to \mathbf{U}(x,t^n)$ is the exact solution of Equation (1.1.1) at time t^n.

1.2.1 Acoustic/transport operator decomposition

We describe here a procedure that allows to approximate the evolution of the system (1.1.1) over a time interval $[t^n, t^n + \Delta t)$. The guideline of the method consists in decoupling the terms responsible for the acoustic, the topography variations and the transport phenomena. In the sequel if $p^{\mathrm{EOS}} : 1/h \mapsto gh^2/2$ we shall note $p = p^{\mathrm{EOS}}(1/h)$, and define the sound velocity c of (1.1.1) by $c^2 = \frac{\mathrm{d}}{\mathrm{d}h}[(p^{\mathrm{EOS}})(1/h)] = gh$. By using the chain rule for the space derivatives we split

up the operators of system (1.1.1) so that it reads for smooth solutions

$$\begin{cases}\partial_t h + u\partial_x h + h\partial_x u = 0,\\ \partial_t(hu) + u\partial_x(hu) + hu\partial_x u + \partial_x p = -gh\partial_x z.\end{cases}$$

Consequently, we propose to approximate the solutions of (1.1.1) by approximating the solutions of the following two subsystems, namely

$$\begin{cases}\partial_t h + h\partial_x u = 0, & (1.2.1a)\\ \partial_t(hu) + hu\partial_x u + \partial_x p = -gh\partial_x z, & (1.2.1b)\end{cases}$$

and

$$\begin{cases}\partial_t h + u\partial_x h = 0, & (1.2.2a)\\ \partial_t(hu) + u\partial_x(hu) = 0, & (1.2.2b)\end{cases}$$

one after the other. System (1.2.1) deals with the acoustic effects and the topography variation, while system (1.2.2) involves the material transport. In the following we shall refer to system (1.2.1) as the acoustic or Lagrangian system and system (1.2.2) as the transport or projection system.

The overall algorithm can be described as follows: for a given discrete state $\mathbf{U}_j^n = (h, hu)_j^n$, $j \in \mathbb{Z}$ that describes the system at instant t^n, the update to the $\mathbf{U}_j^{n+1} = (h, hu)_j^{n+1}$ is a two-step process defined by

1. Update $\mathbf{U}_j^n$ to $\mathbf{U}_j^{n+1-}$ by approximating the solution of system (1.2.1),
2. Update $\mathbf{U}_j^{n+1-}$ to $\mathbf{U}_j^{n+1}$ by approximating the solution of system (1.2.2).

1.2.2 Relaxation approximation of the acoustic system

If we note $\tau = 1/h$ the specific volume, the acoustic system (1.2.1) takes the form

$$\begin{cases}\partial_t \tau - \tau\partial_x u = 0, & (1.2.3a)\\ \partial_t u + \tau\partial_x p^{\mathrm{EOS}} = -g\partial_x z, & (1.2.3b)\\ \partial_t z = 0. & (1.2.3c)\end{cases}$$

It is straightforward to check that the quasilinear system (1.2.3) is strictly hyperbolic over the space $\{(\tau, u, z)^T \in \mathbb{R}^3 \mid \tau > 0\}$ and the eigenstructure of the system is composed by three fields associated with the eigenvalues $\{-c, 0, c\}$. The wave associated with $\pm c$ (resp. 0) is genuinely nonlinear (resp. a stationary contact discontinuity). Let us underline that the material velocity u is not involved in the characteristic velocities of system (1.2.3) but only the sound velocity c.

For $t \in [t^n, t^n + \Delta t)$, we propose to approximate $\tau(x,t)\partial_x\cdot$ by $\tau(x,t^n)\partial_x\cdot$ and $\partial_x z = \tau(x,t)\partial_x z/\tau(x,t)$ by $\tau(x,t^n)\partial_x z/\tau(x,t)$ in system (1.2.3). If one introduces the mass variable m defined by $\mathrm{d}m(x) = \tau(x,t^n)^{-1}\mathrm{d}x$, up to a slight abuse of notations, system (1.2.3) can be recast into

$$\begin{cases}\partial_t \tau - \partial_m u = 0, & (1.2.4a)\\ \partial_t u + \partial_m p^{\mathrm{EOS}} = -g\partial_m z/\tau, & (1.2.4b)\\ \partial_t z = 0. & (1.2.4c)\end{cases}$$

Let us note that, when the topography is flat, system (1.2.4) is consistent with the usual form of the barotropic gas dynamics equations in Lagrangian coordinates with a mass space variable (see for instance [26]).

We carry on with the approximation process of the acoustic system (1.2.1) by using a Suliciu-type relaxation approximation of system (1.2.4), see [33]. We will see in the sequel that this strategy will allow us to design a simple and not expensive time implicit treatment of system (1.2.1) in order to remove the usual CFL restriction associated with the fast acoustic waves $\pm c$. The design principle of the so-called pressure relaxation methods is now very well-known, see for instance [33, 31, 22, 12, 15, 8, 9] and the references therein and consists in introducing a larger system with linearly degenerate characteristic fields so that the underlying Riemann problem is easy to solve. To do so, we introduce a new independent variable pressure Π that can be seen as a linearization of the pressure p^{EOS}. While the pressure p^{EOS} verifies $\partial_t p^{\mathrm{EOS}} + (c/\tau)^2 \partial_m u = 0$ when τ and u are smooth solutions of system (1.2.3), the surrogate pressure Π is evolved according to its own partial differential equation. Within the time interval $t \in [t^n, t^n + \Delta t)$, we propose to consider the following relaxation system

$$\begin{cases} \partial_t \tau - \partial_m u = 0, & (1.2.5a) \\ \partial_t u + \partial_m \Pi = -g \partial_m z/\tau, & (1.2.5b) \\ \partial_t \Pi + a^2 \partial_m u = \lambda(p^{\mathrm{EOS}}(\mathcal{T}) - \Pi), & (1.2.5c) \\ \partial_t z = 0, & (1.2.5d) \end{cases}$$

where $a > 0$ is a constant whose choice will be specified later on, $\lambda > 0$ is the relaxation parameter, and $\mathcal{T}$ obeys the well-defined (under appropriate conditions on a) implicit relation

$$\Pi = p^{\mathrm{EOS}}(\mathcal{T}) + a^2(\mathcal{T} - \tau).$$

System (1.2.5) is indeed an approximation of (1.2.4) in the sense that in the asymptotic regime $\lambda \to +\infty$ we have, at least formally that $\Pi \to p^{\mathrm{EOS}}$ and we recover (1.2.4), see [15] for a rigourous proof for both smooth and discontinuous solutions. Let us also briefly recall that this relaxation model can be endowed with a relaxation entropy defined by

$$\Sigma = \frac{u^2}{2} - \int^{\mathcal{T}} p^{\mathrm{EOS}}(\tau) d\tau + \frac{\Pi^2 - (p^{\mathrm{EOS}})^2(\mathcal{T})}{2\, a^2}, \tag{1.2.6}$$

which is such that $h\Sigma$ coincides with the entropy $\mathcal{U}$ at equilibrium $\mathcal{T} = \tau$. The chain rule and for smooth solutions easily satisfies

$$\partial_t \Sigma + \partial_m \Pi u = -\frac{\lambda}{a^2}(p^{\mathrm{EOS}}(\mathcal{T}) - \Pi)^2 - g\frac{u}{\tau}\partial_m z, \tag{1.2.7}$$

which is nothing but a relaxation and Lagrangian form of (1.1.3). Note that the first term of the right-hand side is negative so that the proposed relaxation process is entropy consistent in the sense of [20].

We adopt the classic method that allows to reach the $\lambda \to \infty$: at each time step, we enforce the equilibrium relation $\Pi_i^n = p^{\mathrm{EOS}}(\tau_i^n)$ and solve system (1.2.5) with $\lambda = 0$. In order to prevent this relaxation procedure from generating instabilities, it is now well established that a must be chosen sufficiently large in agreement with the Whitham subcharacteristic condition

$$a > \max\left(c(\tau)/\tau\right), \tag{1.2.8}$$

when τ spans the values of the solution of system (1.2.5) for $t \in [t^n, t^n + \Delta t)$ (see again the above references). For $\lambda = 0$, system (1.2.5) can take the compact form

$$\partial_t \mathbf{W} + \partial_m \mathbf{G}(\mathbf{W}) = \left(-\frac{g}{\tau}\partial_m z\right)\mathbf{E}_2, \tag{1.2.9}$$

where $\mathbf{W} = (\tau, u, \Pi, z)^T$, $\mathbf{G}(\mathbf{W}) = (-u, \Pi, a^2 u, 0)^T$, $\mathbf{E}_2 = (0, 1, 0, 0)^T$. Let us discuss a few properties of Equation (1.2.9). First, it can be easily proved that (1.2.9) is strictly hyperbolic and involves four linearly degenerate characteristic fields associated with the characteristic velocities $\{-a, 0, +a\}$ that are nothing but approximations of the eigenvalues of system (1.2.4). The jump relations involved with each field are detailed in Appendix 1.B. The non-conservative product that features in system (1.2.4) is well defined for smooth z under consideration here.

Before going any further, let us observe that system (1.2.9) can be recast into the following equivalent form

$$\begin{cases} \partial_t \tau - \partial_m u = 0, & (1.2.10a) \\ \partial_t \overrightarrow{w} + a\partial_m \overrightarrow{w} = -ag\partial_m z/\tau, & (1.2.10b) \\ \partial_t \overleftarrow{w} - a\partial_m \overleftarrow{w} = +ag\partial_m z/\tau, & (1.2.10c) \\ \partial_t z = 0, & (1.2.10d) \end{cases}$$

where the new variables $\overrightarrow{w}$ and $\overleftarrow{w}$ are defined by $\overrightarrow{w} = \Pi + au$, $\overleftarrow{w} = \Pi - au$. These quantities are nothing but the strong Riemann invariants associated with the characteristic speeds $\pm a$ of the relaxation system (1.2.10) when the topography terms are omitted. The closure relations for system (1.2.10) are naturally given by

$$u = \frac{\overrightarrow{w} - \overleftarrow{w}}{2a}, \quad \Pi = \frac{\overrightarrow{w} + \overleftarrow{w}}{2}.$$

This new formulation will be used in the sequel to study a time-implicit discretization of system (1.2.9).

We now need to propose a discretization strategy for system (1.2.9). Unfortunately, the classic relaxation solver strategy cannot be carried on here since the solution of the Riemann problem associated with system (1.2.9) cannot be defined easily. Indeed it is not possible to properly define the non-conservative term $\partial_m z/\tau$ with a piecewise constant initial value for z. However we will see in the next section that it is possible to derive an approximate Riemann solver for system (1.2.9) using a discretization of the non-conservative product that is consistent (in a sense to be specified later) with the smooth term $(g/\tau)\partial_m z$.

1.2.3 Approximate Riemann solver for the acoustic system

Let $\Delta m_L > 0$, $\Delta m_R > 0$ and suppose given a smooth function $m \mapsto z(m)$. If $\overline{m} \in \mathbb{R}$, we consider a piecewise initial data defined by

$$\mathbf{W}(m, t = 0) = \begin{cases} \mathbf{W}_L = (\tau_L, u_L, \Pi_L, z_L)^T, & \text{if } m < \overline{m}, \\ \mathbf{W}_R = (\tau_R, u_R, \Pi_R, z_R)^T, & \text{if } m > \overline{m}, \end{cases} \tag{1.2.11}$$

where Π_k and z_k, $k = L, R$ are defined by

$$\Pi_L = p^{\text{EOS}}(\tau_L), \quad \Pi_R = p^{\text{EOS}}(\tau_R),$$

$$z_L = \frac{1}{\Delta m_L}\int_{-\Delta m_L}^{0} z(\overline{m}+r)\mathrm{d}r, \quad z_R = \frac{1}{\Delta m_R}\int_{0}^{\Delta m_R} z(\overline{m}+r)\mathrm{d}r.$$

Note that Π_L and Π_R are at equilibrium. Let us now build an approximate Riemann solver for the relaxed acoustic system (1.2.9). We seek for a function $\mathbf{W}_{\mathrm{RP}}$ composed by four states separated by discontinuities as follows

$$\mathbf{W}_{\mathrm{RP}}\left(\frac{m-\overline{m}}{t};\mathbf{W}_L,\mathbf{W}_R\right) = \begin{cases} \mathbf{W}_L, & \text{if } \frac{m-\overline{m}}{t} < -a, \\ \mathbf{W}_L^*, & \text{if } -a < \frac{m-\overline{m}}{t} < 0, \\ \mathbf{W}_R^*, & \text{if } 0 < \frac{m-\overline{m}}{t} < a, \\ \mathbf{W}_R, & \text{if } a < \frac{m-\overline{m}}{t}, \end{cases} \tag{1.2.12}$$

where the intermediate states are such that the following three consistency properties a), b) and c) hold true (see [24, 25]):

a) $\mathbf{W}_{\mathrm{RP}}$ is consistent in the integral sense with the shallow water equations, more specifically in our context: if Δt is such that $a\Delta t \leq \min(\Delta m_L, \Delta m_R)/2$, then

$$\mathbf{G}(\mathbf{W}_R) - \mathbf{G}(\mathbf{W}_L) = -a(\mathbf{W}_L^* - \mathbf{W}_L) + a(\mathbf{W}_R - \mathbf{W}_R^*) - \frac{\Delta m_L + \Delta m_R}{2}\left\{\frac{g}{\tau}\partial_m z\right\}, \tag{1.2.13}$$

where $\left\{\frac{g}{\tau}\partial_m z\right\}$ is consistent with the influence of the source term, in the sense that

$$\lim_{\substack{\Delta m_L,\Delta m_R\to 0 \\ \mathbf{W}_L,\mathbf{W}_R\to(\overline{\tau},\overline{u},\overline{\Pi},z(\overline{m}))}} \left\{\frac{g}{\tau}\partial_m z\right\} = \frac{g}{\overline{\tau}}(\partial_m z)(\overline{m}); \tag{1.2.14}$$

b) In the case of constant bottom, i.e. $z_L = z_R$, $\mathbf{W}_{\mathrm{RP}}$ degenerates towards the classic solution of the Riemann problem of the acoustic relaxed system (1.2.9) for a flat bottom $\partial_m z = 0$;

c) If $\mathbf{W}_L$ and $\mathbf{W}_R$ verify the lake at rest condition

$$u_L = u_R = 0, \qquad 1/\tau_L + z_L = 1/\tau_R + z_R, \tag{1.2.15}$$

then $\mathbf{W}_L^* = \mathbf{W}_L$ and $\mathbf{W}_R^* = \mathbf{W}_R$.

Defining a proper function $\mathbf{W}_{\mathrm{RP}}$ thus simply boils down to proposing intermediate states $\mathbf{W}_L^*$ and $\mathbf{W}_R^*$ that comply with a), b) and c). We proceed as follows: first we impose that $\mathbf{W}_L$ and $\mathbf{W}_L^*$ (resp. $\mathbf{W}_R^*$ and $\mathbf{W}_R$) verify the jump conditions

$$\begin{cases} a(\mathbf{W}_L^* - \mathbf{W}_L) + \mathbf{G}(\mathbf{W}_L^*) - \mathbf{G}(\mathbf{W}_L) = 0, \\ -a(\mathbf{W}_R - \mathbf{W}_R^*) + \mathbf{G}(\mathbf{W}_R) - \mathbf{G}(\mathbf{W}_R^*) = 0. \end{cases} \tag{1.2.16}$$

This amounts to say that the discontinuity of velocity $\pm a$ of $\mathbf{W}_{\mathrm{RP}}$ behaves like the $(\pm)a$-wave of system (1.2.9) for a flat bottom. Similarly, across the discontinuity of velocity 0 we impose that

$$u_L^* = u_R^* =: u^*. \tag{1.2.17}$$

Relations (1.2.16) and (1.2.17) does not provide enough information to determine the intermediate states $\mathbf{W}_L^*$ and $\mathbf{W}_R^*$. Indeed, they provide us with only seven independent relations while we aim at defining eight quantities, namely the four components of each $\mathbf{W}_L^*$ and $\mathbf{W}_R^*$.

We choose to add another jump relation across the stationary discontinuity of $\mathbf{W}_{\mathrm{RP}}$ that complies with condition b): we impose that

$$\Pi_R^* - \Pi_L^* + \mathscr{M} = 0, \tag{1.2.18}$$

where $\mathscr{M}$ is a function to be specified such that $\mathscr{M} = 0$ if $z_L = z_R$. Relations (1.2.9), (1.2.17) and (1.2.18) allow to solve for $\mathbf{W}_L^*$ and $\mathbf{W}_R^*$ and we obtain

$$\begin{cases} \tau_L^* = \tau_L + \dfrac{1}{2a}\left(u_R - u_L\right) - \dfrac{1}{2a^2}\left(\Pi_R - \Pi_L\right) - \dfrac{\mathscr{M}}{2a^2}, \\ \tau_R^* = \tau_R + \dfrac{1}{2a}\left(u_R - u_L\right) + \dfrac{1}{2a^2}\left(\Pi_R - \Pi_L\right) + \dfrac{\mathscr{M}}{2a^2}, \\ u^* = u_R^* = u_L^* = \dfrac{u_R + u_L}{2} - \dfrac{1}{2a}\left(\Pi_R - \Pi_L\right) - \dfrac{\mathscr{M}}{2a}, \\ \Pi^* = \dfrac{\Pi_R + \Pi_L}{2} - \dfrac{a}{2}\left(u_R - u_L\right), \\ \Pi_L^* = \Pi^* + \dfrac{\mathscr{M}}{2}, \\ \Pi_R^* = \Pi^* - \dfrac{\mathscr{M}}{2}, \\ z_L^* = z_L, \\ z_R^* = z_R. \end{cases} \tag{1.2.19}$$

We now only need to determine $\mathscr{M}$ such that conditions a), b) and c) are satisfied. It is straightforward to see that the integral consistency requirement of condition a) implies by Equation (1.2.13) that

$$\mathscr{M} = \left\{\frac{g}{\tau}\partial_m z\right\}\frac{\Delta m_L + \Delta m_R}{2}. \tag{1.2.20}$$

A simple mean to comply with conditions a) and b) is to choose

$$\mathscr{M} = \frac{g}{\tau_\Delta(\mathbf{W}_L, \mathbf{W}_R)}(z_R - z_L), \tag{1.2.21}$$

where $\tau_\Delta(\mathbf{W}_L, \mathbf{W}_R)$ has to be chosen such that $\tau_\Delta(\mathbf{W}_L, \mathbf{W}_R) \to \overline{\tau}$ if $\tau_L, \tau_R \to \overline{\tau}$. At last, we need to ensure condition c): if we have condition (1.2.15), then $\mathbf{W}_L^* = \mathbf{W}_L$ and $\mathbf{W}_R = \mathbf{W}_R^*$ imply that

$$\frac{1}{\tau_\Delta(\mathbf{W}_L, \mathbf{W}_R)} = \frac{1}{2}\left(\frac{1}{\tau_L} + \frac{1}{\tau_R}\right). \tag{1.2.22}$$

As a conclusion, we choose to adopt Equation (1.2.22) as a definition of τ_Δ for any $\mathbf{W}_L$ and $\mathbf{W}_R$. This yields that

$$\left\{\frac{g}{\tau}\partial_m z\right\}(\mathbf{W}_L, \mathbf{W}_R, \Delta m_L, \Delta m_R) = \frac{g}{\tau_\Delta(\mathbf{W}_L, \mathbf{W}_R)}\frac{2}{\Delta m_L + \Delta m_R}(z_R - z_L). \tag{1.2.23}$$

It is then straightforward to check that the approximate Riemann solver defined by Equations (1.2.19) and (1.2.21) verifies the three conditions a), b) and c). We sum up in the following proposition the properties of our Riemann solver.

Proposition 1.2.1. *Consider the approximate Riemann solver* $\mathbf{W}_{\mathrm{RP}}$ *defined by Equations* (1.2.12), (1.2.19), (1.2.21) *and* (1.2.22). *Then one has properties:*

(i) $\mathbf{W}_{\mathrm{RP}}$ *is consistent in the integral sense with the shallow water equations* (1.1.1).

(ii) In the case of a constant bottom $z_L = z_R$, $\mathbf{W}_{\mathrm{RP}}$ *degenerates to a classic approximate Riemann solver for the barotropic Euler equations in Lagrange coordinates.*

(iii) If $\mathbf{W}_L$ *and* $\mathbf{W}_R$ *verify the lake at rest relation* (1.2.15), *then* $\mathbf{W}_k^* = \mathbf{W}_k$, $k = L, R$.

1.3 Numerical method

In this section, we now give the details of the two-step process proposed in Section 1.2.1 for solving the shallow water equations. Let us briefly recall that this two-step process is defined by

1. Update $\mathbf{U}_j^n$ to $\mathbf{U}_j^{n+1-}$ by approximating the solution of system (1.2.1),
2. Update $\mathbf{U}_j^{n+1-}$ to $\mathbf{U}_j^{n+1}$ by approximating the solution of system (1.2.2).

In the sequel we shall note $\Delta m_j = \Delta x_j h_j^n$, $\Delta m_{j+1/2} = (\Delta m_j + \Delta m_{j+1})/2$, and if we assume as given the approximate solution $\{\mathbf{U}_j^n\}_j$ at time t^n, we introduce the approximate solution $\{\mathbf{W}_j^n\}_j$ at equilibrium in the $\mathbf{W}$ variable with a clear and natural definition. We begin with a fully explicit discretization of the shallow water equations, which means that both steps of the process are solved with a time-explicit procedure, and we will go on with a mixed implicit-explicit strategy for which the solutions of system (1.2.1) are solved implicitly in time and the solutions of system (1.2.2) are solved explicitly. The latter strategy allows to get rid of the strong CFL restriction coming from the acoustic waves in the subsonic regime and corresponds to the very motivation of the present study.

1.3.1 Time-explicit discretization

Let us begin with the time-explicit discretization of the acoustic system (1.2.1), or equivalently (1.2.3).

Acoustic step

The acoustic update is achieved thanks to the proposed relaxation approximation and the corresponding approximate Riemann solver detailed in Section 1.2.3. More precisely, we propose to simply use a Godunov-type method based on this approximate Riemann solver. As it is customary and starting from the piecewise constant initial data defined by the sequence $\{\mathbf{W}_j^n\}_j$, it consists in averaging after a Δt-long time evolution, the juxtaposition of the approximate Riemann solutions defined locally at each interface $x_{j+1/2}$. Following the same lines as in [16] and [18], see also [24, 25, 13] and the references therein, this update procedure can be easily

expressed as follows after simple calculations,

$$\begin{cases} \tau_j^{n+1-} = \tau_j^n + \dfrac{\Delta t}{\Delta m_j}(u^*_{j+1/2} - u^*_{j+1/2}), & (1.3.1a) \\ u_j^{n+1-} = u_j^n - \dfrac{\Delta t}{\Delta m_j}(\Pi^*_{j+1/2} - \Pi^*_{j+1/2}) - \Delta t \left\{ \dfrac{g}{\tau}\partial_m z \right\}_j^n, & (1.3.1b) \\ \Pi_j^{n+1-} = \Pi_j^n - \dfrac{\Delta t}{\Delta m_j} a^2 (u^*_{j+1/2} - u^*_{j+1/2}). & (1.3.1c) \end{cases}$$

where $\Pi_j^n = p^{\mathrm{EOS}}(\tau_j^n)$ and

$$\begin{cases} u^*_{j+1/2} = \dfrac{1}{2}(u_j^n + u_{j+1}^n) - \dfrac{1}{2a}(\Pi_{j+1}^n - \Pi_j^n) - \dfrac{\Delta m_{j+1/2}}{2a}\left\{ \dfrac{g}{\tau}\partial_m z \right\}_{j+1/2}^n, & (1.3.2a) \\ \Pi^*_{j+1/2} = \dfrac{1}{2}(\Pi_j^n + \Pi_{j+1}^n) - \dfrac{a}{2}(u_{j+1}^n - u_j^n), & (1.3.2b) \\ \left\{ \dfrac{g}{\tau}\partial_m z \right\}_{j+1/2}^n = \left\{ \dfrac{g}{\tau}\partial_m z \right\} (\mathbf{W}_j^n, \mathbf{W}_j^n, \Delta m_j, \Delta m_{j+1}) \\ \qquad = \dfrac{g}{2}\left(\dfrac{1}{\tau_j^n} + \dfrac{1}{\tau_{j+1}^n} \right) \dfrac{z_{j+1} - z_j}{\Delta m_{j+1/2}}, & (1.3.2c) \\ \left\{ \dfrac{g}{\tau}\partial_m z \right\}_j^n = \dfrac{1}{2}\left(\dfrac{\Delta m_{j+1/2}}{\Delta m_j}\left\{ \dfrac{g}{\tau}\partial_m z \right\}_{j+1/2}^n + \dfrac{\Delta m_{j-1/2}}{\Delta m_j}\left\{ \dfrac{g}{\tau}\partial_m z \right\}_{j-1/2}^n \right). & (1.3.2d) \end{cases}$$

If we focus now on the conservative variable $\mathbf{U} = (h, hu)$, the discretization (1.3.1) yields the following formula for the update sequence $\{\mathbf{U}_j^{n+1-}\}_j$, namely

$$\begin{cases} L_j h^{n+1-} = h_j^n, & (1.3.3a) \\ L_j (hu)^{n+1-} = (hu)_j^n - \dfrac{\Delta t}{\Delta x_j}(\Pi^*_{j+1/2} - \Pi^*_{j-1/2}) - \Delta t\, h_j^n \left\{ \dfrac{g}{\tau}\partial_m z \right\}_j^n, & (1.3.3b) \\ L_j = 1 + \dfrac{\Delta t}{\Delta x_j}(u^*_{j+1/2} - u^*_{j-1/2}). & (1.3.3c) \end{cases}$$

Let us now continue with the discretization of the transport Equations (1.2.2).

Transport step

Denoting $\varphi \in \{h, hu\}$ and following again the same lines of [16] and [18], see again also [23], we use a standard time-explicit upwind discretization for the transport step by setting

$$\varphi_j^{n+1} = \varphi_j^{n+1-} - \frac{\Delta t}{\Delta x_j}(u^*_{j+1/2}\varphi_{j+1/2}^{n+1-} - u^*_{j-1/2}\varphi_{j-1/2}^{n+1-}) + \frac{\Delta t}{\Delta x_j}\varphi_j^{n+1-}(u^*_{j+1/2} - u^*_{j-1/2}), \qquad (1.3.4)$$

where

$$\varphi_{j+1/2}^{n+1-} = \begin{cases} \varphi_j^{n+1-}, \text{ if } u^*_{j+1/2} \geq 0, \\ \varphi_{j+1}^{n+1-}, \text{ if } u^*_{j+1/2} < 0. \end{cases}$$

Let us note that the transport update scheme (1.3.4) equivalently reads

$$\varphi_j^{n+1} = L_j \varphi_j^{n+1-} - \frac{\Delta t}{\Delta x_j}\left(u^*_{j+1/2}\varphi^{n+1-}_{j+1/2} - u^*_{j-1/2}\varphi^{n+1-}_{j-1/2}\right), \tag{1.3.5}$$

and that the interface value of the velocity $u^*_{j+1/2}$ coincides with the one proposed in the first step, which is actually crucial in order for the whole scheme to be conservative. The next statement gather the main properties satisfied by our explicit in time and two-step algorithm.

Overall Discretization

After injecting Equation (1.3.3b) into Equation (1.3.5) one obtains the complete update procedure from t^n to t^{n+1}. For the conservative variables it reads

$$\begin{cases} h_j^{n+1} = h_j^n - \dfrac{\Delta t}{\Delta x_j}\left(u^*_{j+1/2}h^{n+1-}_{j+1/2} - u^*_{j-1/2}h^{n+1-}_{j-1/2}\right), \\ (hu)_j^{n+1} = (hu)_j^n - \dfrac{\Delta t}{\Delta x_j}\left[u^*_{j+1/2}(hu)^{n+1-}_{j+1/2} + \Pi^*_{j+1/2} - u^*_{j-1/2}(hu)^{n+1-}_{j-1/2} - \Pi^*_{j-1/2}\right] \\ \qquad\qquad - \Delta t\, h_j^n \left\{\dfrac{g}{\tau}\partial_m z\right\}_j^n. \end{cases} \tag{1.3.6}$$

We sum up the properties of our explicit scheme (1.3.1)-(1.3.4) in the following proposition.

Proposition 1.3.1. *The fully explicit scheme* (1.3.1)-(1.3.4) *satisfies the following:*

(i) it is a conservative scheme for the water height h. It is also a conservative scheme for hu when the topography source term vanishes.

Under the Whitham subcharacteristic condition and the Courant-Friedrichs-Lewy (CFL) conditions,

$$\max_j \frac{\Delta t}{h_j^n \Delta x_j} \leq \frac{1}{2a} \quad \textit{and} \quad \max_j \frac{\Delta t}{\Delta x_j}\left((u^*_{j-\frac{1}{2}})^+ - (u^*_{j+\frac{1}{2}})^-\right) < 1, \tag{1.3.7}$$

it satisfies also:

(ii) the water height h_j^n is positive for all j and $n > 0$ provided that h_j^0 is positive for all j,

(iii) it is well-balanced, with respect to the lake at rest condition (1.1.5)*,*

(iv) it degenerates to the classic Lagrange-projection scheme when the bottom is flat.

Proof.

(i) This is a straightforward consequence of Equation (1.3.6).

(ii) Thanks to Equations (1.3.3b) and (1.3.3a), the CFL condition (1.3.7) ensures that $h_j^{n+1-} > 0$ for $j \in \mathbb{Z}$. The CFL condition (1.3.7) yields that h_j^{n+1} is a convex combination of $(h_k^{n+1-})_{k=j\pm1,j}$ and therefore $h_j^{n+1} > 0$.

(iii) Consider a discrete fluid state at instant t^n that matches the lake at rest condition, namely: $u_j^n = 0$, $h_j^n + z_j = h_{j+1}^n + z_{j+1}$, for all $j \in \mathbb{Z}$. Thanks to the condition c) verified by the approximated Riemann solver of the acoustic step, we know that $u_j^{n+1-} = 0$, $h_j^{n+1-} + z_j = h_{j+1}^{n+1-} + z_{j+1}$, for all $j \in \mathbb{Z}$. And thus, the transport step (1.3.4) boils down to $h_j^{n+1} = h_j^{n+1-}$ and $u_j^{n+1} = 0$.

(iv) This is consequence of condition b) imposed on the approximate Riemann solver for the acoustic step. □

Remark 1.3.1. *Following the theory proposed by Gallice in [24] and [25] for non conservative systems with source terms, it is also possible to prove that our time-explicit Godunov-type scheme based on the definition of a consistent approximate Riemann solver satisfies a discrete version of the non conservative entropy* (1.1.3) *under additional assumptions on the intermediate states and the propagation speed a. We refer for instance the reader to [2] and [4] for detailed calculations. Note that we are not able to prove at present that the scheme satisfies a discrete version of the conservative entropy inequality* (1.1.4).

1.3.2 Implicit in time Lagrange-projection method

Let us now consider the ultimate algorithm of this paper, which consists in considering a time-implicit scheme for the Lagrangian step and keeping unchanged the transport step. As we will see in the next theorem, this strategy will allow us to obtain a non linearly stable algorithm under a CFL restriction based on the material velocity u and not on the sound velocity c. In order to derive a time-implicit scheme for the Lagrangian step, we follow the following standard approach where the numerical fluxes are now evaluated at time t^{n+1-}, which gives here the same update formulas as in the explicit case which are

$$\begin{cases} \tau_j^{n+1-} = \tau_j^n + \dfrac{\Delta t}{\Delta m_j}(u^*_{j+1/2} - u^*_{j+1/2}), & (1.3.8a) \\ u_j^{n+1-} = u_j^n - \dfrac{\Delta t}{\Delta m_j}(\Pi^*_{j+1/2} - \Pi^*_{j+1/2}) - \Delta t \left\{ \dfrac{g}{\tau}\partial_m z \right\}_j^n, & (1.3.8b) \\ \Pi_j^{n+1-} = \Pi_j^n - \dfrac{\Delta t}{\Delta m_j} a^2 (u^*_{j+1/2} - u^*_{j+1/2}), & (1.3.8c) \end{cases}$$

but where the numerical fluxes now involve quantities at time t^{n+1-} apart from the term consistent with $\left\{ \frac{g}{\tau}\partial_m z \right\}$, which writes

$$\begin{cases} u^*_{j+1/2} = \dfrac{1}{2}(u_j^{n+1-} + u_{j+1}^{n+1-}) - \dfrac{1}{2a}(\Pi_{j+1}^{n+1-} - \Pi_j^{n+1-}) \\ \qquad\qquad - \dfrac{\Delta m_{j+1/2}}{2a} \left\{ \dfrac{g}{\tau}\partial_m z \right\}_{j+1/2}^n, & (1.3.9a) \\ \Pi^*_{j+1/2} = \dfrac{1}{2}(\Pi_j^{n+1-} + \Pi_{j+1}^{n+1-}) - \dfrac{a}{2}(u_{j+1}^{n+1-} - u_j^{n+1-}), & (1.3.9b) \end{cases}$$

with $\left\{ \frac{g}{\tau}\partial_m z \right\}_j^n$ and $\left\{ \frac{g}{\tau}\partial_m z \right\}_{j+1/2}^n$ given by Equations (1.3.2c) and (1.3.2d).

Let us observe that we suggest here to keep on evaluating the topography source term at time t^n. This choice is motivated by the fact that this implicit system to be solved turns out to be a linear system with a significantly reduced coupling of the variables. More precisely, it is interesting to see that it is equivalent to the following one written in characteristic variables,

namely

$$\begin{cases} \tau_j^{n+1-} = \tau_j^n + \dfrac{\Delta t}{\Delta m_j}\left(u^*_{j+1/2} - u^*_{j-1/2}\right), \\ \overrightarrow{w}_j^{n+1-} = \overrightarrow{w}_j^n - a\dfrac{\Delta t}{\Delta m_j}\left(\overrightarrow{w}_j^{n+1-} - \overrightarrow{w}_{j-1}^{n+1-}\right) - a\Delta t\dfrac{\Delta m_{j-1/2}}{\Delta m_j}\left\{\dfrac{g}{\tau}\partial_m z\right\}^n_{j-1/2}, \\ \overleftarrow{w}_j^{n+1-} = \overleftarrow{w}_j^n + a\dfrac{\Delta t}{\Delta m_j}\left(\overleftarrow{w}_{j+1}^{n+1-} - \overleftarrow{w}_j^{n+1-}\right) + a\Delta t\dfrac{\Delta m_{j+1/2}}{\Delta m_j}\left\{\dfrac{g}{\tau}\partial_m z\right\}^n_{j+1/2}, \\ z_j^{n+1} = z_j^n, \end{cases} \tag{1.3.10}$$

where of course $u^*_{j+1/2}$ means here $u^{*,n+1-}_{j+1/2}$ (the notation has been lightened for the sake of clarity). Notice that the coupling between the four variables is actually weak in scheme (1.3.10) since we can easily first solve the linear system given by the second and the third equations, which are nothing but

$$\begin{cases} \left(I_N + a\Delta t\, A_+^n\right)\overrightarrow{w}^{n+1-} = \overrightarrow{w}^n - a\Delta t\, b_+^n, \\ \left(I_N - a\Delta t\, A_-^n\right)\overleftarrow{w}^{n+1-} = \overleftarrow{w}^n + a\Delta t\, b_-^n, \end{cases}$$

where we have set

$$A_+^n = \begin{pmatrix} \frac{1}{\Delta m_1} & 0 & \cdots & 0 \\ \frac{-1}{\Delta m_2} & \frac{1}{\Delta m_2} & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & \frac{-1}{\Delta m_N} & \frac{1}{\Delta m_N} \end{pmatrix}, \qquad b_+^n = g\cdot\begin{pmatrix} \frac{1}{\Delta m_1}\frac{h_1^n+h_0^n}{2}(z_1 - z_0) \\ \frac{1}{\Delta m_2}\frac{h_2^n+h_1^n}{2}(z_2 - z_1) \\ \vdots \\ \frac{1}{\Delta m_N}\frac{h_N^n+h_{N-1}^n}{2}(z_N - z_{N-1}) \end{pmatrix},$$

and

$$A_-^n = \begin{pmatrix} \frac{-1}{\Delta m_1} & \frac{1}{\Delta m_1} & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \frac{-1}{\Delta m_{N-1}} & \frac{1}{\Delta m_{N-1}} \\ 0 & \vdots & 0 & \frac{-1}{\Delta m_N} \end{pmatrix}, \qquad b_-^n = g\cdot\begin{pmatrix} \frac{1}{\Delta m_1}\frac{h_2^n+h_1^n}{2}(z_2 - z_1) \\ \vdots \\ \frac{1}{\Delta m_{N-1}}\frac{h_N^n+h_{N-1}^n}{2}(z_N - z_{N-1}) \\ \frac{1}{\Delta m_N}\frac{h_{N+1}^n+h_N^n}{2}(z_{N+1} - z_N) \end{pmatrix}.$$

Let us of course notice that a few coefficients of the matrices A_+^n and A_-^n, and vectors b_+^n and b_-^n might be modified depending on the boundary conditions, but the purpose is to highlight that the characteristic variables $\overleftarrow{w}$ and $\overrightarrow{w}$ can be solved independently. Once this is done, the τ variable can be updated explicitly since $u^*_{j+1/2}$, or let us say $u^{*,n+1-}_{j+1/2}$, is explicitly known from the knowledge of $\overleftarrow{w}^{n+1-}$ and $\overrightarrow{w}^{n+1-}$ by the formulas

$$u_j^{n+1-} = \frac{1}{2a}(\overrightarrow{w}_j^{n+1-} - \overleftarrow{w}_j^{n+1-}), \quad \Pi_j^{n+1-} = \frac{1}{2}(\overrightarrow{w}_j^{n+1-} + \overleftarrow{w}_j^{n+1-}).$$

At last, notice that the matrices $\left(I_N + a\Delta t\, A_+^n\right)$ and $\left(I_N - a\Delta t\, A_-^n\right)$ are clearly triangular (and bidiagonal) with positive diagonal coefficients, so that the system (1.3.10) has a unique solution whatever the time step $\Delta t > 0$ is.

It is quite natural at this stage to wonder whether the proposed time-implicit treatment of the Lagrangian step is well-balanced, which was true for the time-explicit version and was the key property leading to the well-balanced property of the global Explicit-Explicit Lagrange-projection scheme in the previous section. It is the purpose of the next lemma.

Lemma 1.3.1. *Under the assumption of the lake at rest at the initial time, i.e.:*

$$\forall j \in \{1, \dots, N\}\,, \quad \begin{cases} u_j^0 = 0, \\ h_j^0 + z_j^0 = constant, \end{cases}$$

the implicit scheme for the Lagrangian step keeps this initial state unchanged, which means that the time-implicit Lagrangian step as well as the global Implicit-Explicit Lagrange-projection scheme is still well-balanced.

Proof. Under the assumption of the lake at rest, and thanks to the initialisation of the relaxation pressure, namely $\forall j \in \{1, \dots, N\}$, $\Pi_j^0 = \frac{g}{2}\left(h_j^0\right)^2$, we get

$$g\frac{h_j^0 + h_{j-1}^0}{2}(z_j - z_{j-1}) = -\frac{g}{2}\left((h_j^0)^2 - (h_{j-1}^0)^2\right) = -(\Pi_j^0 - \Pi_{j-1}^0).$$

Thus one can write

$$b_+^0 = -A_+^0 \Pi^0,$$

and, in this special case where $u^0 \equiv 0$,

$$\begin{aligned}\left(I_N + a\Delta t\, A_+^n\right)(\Pi + au)^{1-} &= \overrightarrow{w}^0 - a\Delta t\, b_+^0 \\ &= (\Pi + au)^0 + a\Delta t\, A_+^0 \Pi^0 = \left(I_N + a\Delta t\, A_+^n\right)(\Pi + au)^0\,,\end{aligned}$$

which finally yields to

$$\overrightarrow{w}^{1-} = (\Pi + au)^{1-} = (\Pi + au)^0 = \overrightarrow{w}^0.$$

Similarly one can prove that

$$\overleftarrow{w}^{1-} = (\Pi - au)^{1-} = (\Pi - au)^0 = \overleftarrow{w}^0,$$

so that

$$\begin{cases} u_j^{1-} = \dfrac{\overrightarrow{w}_j^{1-} - \overleftarrow{w}_j^{1-}}{2a} = \dfrac{\overrightarrow{w}_j^0 - \overleftarrow{w}_j^0}{2a} = u_j^0 = 0, \\ \Pi_j^{1-} = \dfrac{\overrightarrow{w}_j^{1-} + \overleftarrow{w}_j^{1-}}{2} = \dfrac{\overrightarrow{w}_j^0 + \overleftarrow{w}_j^0}{2} = \Pi_j^0 = \dfrac{g}{2}(h_j^0)^2, \end{cases}$$

and

$$u_{j+1/2}^{*,1-} = \frac{1}{2}(u_{j+1}^{1-} - u_j^{1-}) - \frac{1}{2a}(\Pi_{j+1}^{1-} - \Pi_j^{1-}) - \frac{g}{2a}\frac{h_{j+1}^0 + h_j^0}{2}(z_{j+1} - z_j) = 0,$$

for all j, and then

$$\tau_j^{1-} = \tau_j^0 + \frac{\Delta t}{\Delta m_j}\left(u_{j+1/2}^{*,1-} - u_{j-1/2}^{*,1-}\right) = \tau_j^0,$$

for all j. Finally we get that the lake is also at rest at the end of the Lagrangian step and, since the transport step is trivial because $u^{1^-} \equiv 0$, the global implicit-explicit scheme is well-balanced. □

Proposition 1.3.2. *Under the Whitham subcharacteristic condition and the CFL condition*

$$\max_j \frac{\Delta t}{\Delta x_j}\Big((u^*_{j-\frac{1}{2}})^+ - (u^*_{j+\frac{1}{2}})^-\Big) < 1, \tag{1.3.11}$$

the implicit-explicit scheme satisfies the following stability properties:

(i) it is a conservative scheme for the water height h. It is also a conservative scheme for hu when the topography source term vanishes,

(ii) the water height h^n_j is positive for all j and $n > 0$ provided that h^0_j is positive for all j,

(iii) it is well-balanced,

(iv) it satisfies a discrete entropy inequality,

(iv) and it gives the usual implicit-explicit Lagrange-projection scheme when the bottom is flat.

Proof. The properties are obtained in the same way as in the explicit case, except for the well-balanced property which has already been proved in the previous Lemma, and the validity of the entropy inequality which is proved in Appendix 1.B. Note that the strict inequality of the CFL condition ensures that the water height stays positive. □

1.4 Numerical results

The aim of this section is to illustrate the behaviour of our Lagrange-projection like strategies in one space dimension. We will also compare the results with the simple, well-balanced, positive and entropy-satisfying scheme recently proposed in [2] (the scheme will be referred to as the HLLACU scheme) and the very well-known hydrostatic reconstruction scheme [1] based on a classic HLL scheme and referred to as HRHLL in the following.

Let us first notice that two (classic) options will be considered in order to evaluate the artificial sound speed a involved in the acoustic step. Let $\kappa > 1$. The first one is based on a local definition of the Lagrangian sound speed in agreement with a local evaluation of the subcharacteristic condition, namely

$$a_{j+1/2} = \kappa \max\left(h^n_j\sqrt{gh^n_j}, h^n_{j+1}\sqrt{gh^n_{j+1}}\right), \quad \forall\, j, \tag{1.4.1}$$

while the second one considers an uniform estimate by setting

$$a_{j+1/2} = \kappa \max_i\left(h^n_i\sqrt{gh^n_i}\right), \quad \forall\, j. \tag{1.4.2}$$

In practice, we set $\kappa = 1.01$. For the sake of conciseness, the full-explicit scheme will be referred to as EXEX_{loc} (resp. $\text{EXEX}_{\text{glob}}$) and the semi-implicit scheme will be referred to as IMEX_{loc} (resp. $\text{IMEX}_{\text{glob}}$) when condition (1.4.1) (resp. (1.4.2)) is used.

Let us also mention for all the test cases, uniform space steps Δx will be considered and the time steps Δt will be chosen in agreement with the CFL conditions (1.3.7) and (1.3.11). More precisely, we will set (unless otherwise stated)

$$\Delta t = \frac{\Delta x}{2\max_j(\sqrt{gh^n_j}, |u^*_{j+1/2}|)}, \tag{1.4.3}$$

for the explicit schemes, and

$$\Delta t = \frac{\Delta x}{2\max_j(|u^*_{j+1/2}|)}, \tag{1.4.4}$$

for the implicit ones, where $u^*_{j+1/2}$ is calculated at time t^n for the sake of simplicity.

Before starting, let us finally mention that initial data matching the lake at rest condition (1.1.5) are preserved by construction by the $\mathrm{EXEX_k}$ and $\mathrm{IMEX_k}$ schemes, $k = \mathrm{loc}, \mathrm{glob}$. Therefore, such test cases will not be considered hereafter.

1.4.1 Dam break problem

We first consider the classic dam break. The space domain $[0, 1500]$ is divided into two parts with the same length and such that the water height is higher on the left side,

$$h(x, t = 0) = 20, \quad \text{if } x \leq 750, \quad h(x, t = 0) = 15, \quad \text{if } x > 750.$$

The velocity is set to be zero on both side at the initial time when the dam breaks and the water starts flowing. Importantly, the topography is not flat but given by the regularized two-step function

$$z(x) = \begin{cases} 4e^{2-\frac{150}{x-487.5}}, & \text{if } 487.5 < x <= 562.5, \\ 8 - 4e^{2-\frac{150}{637.5-x}}, & \text{if } 562.5 < x <= 637.5, \\ 8, & \text{if } 637.5 < x <= 862.5, \\ 8 - 4e^{2-\frac{150}{x-862.5}}, & \text{if } 862.5 < x <= 937.5, \\ 4e^{2-\frac{150}{1012.5-x}}, & \text{if } 937.5 < x <= 1012.5, \\ 0 & \text{otherwise.} \end{cases}$$

At last, the spatial domain is discretized over a 1500-cell grid and Neumann boundary conditions are used. Figures 1.4.1 and 1.4.2 show the solutions at final times $T = 10$ and $T = 50$ with different numerical strategies. The following comments are in order. We first observe that the implicit schemes are the most diffusive, which was clearly expected from the implicit treatment of the acoustic step. Note also that our Lagrangian-projection schemes are intrinsically made of two averaging steps, which is necessary to separate the acoustic and transport effects, but at the price of additional numerical diffusion compared to a direct Eulerian approach like the one proposed in the HLLACU scheme. We also observe that a local definition of parameter a is preferable to the global one in order to reduce the numerical diffusion. For this reason, we will only consider the local evaluation in the following test cases ($k = \mathrm{loc}$). As far as the time step is concerned, we observed for this test case that the averaged value (calculated from the time iterations needed to reach the final time $T = 50$) is about five times larger for the $\mathrm{IMEX_{loc}}$ than for the $\mathrm{EXEX_{loc}}$ schemes.

1.4.2 Propagation of perturbations

This test case focuses on the perturbation of a steady state solution by a pulse that splits into two opposite waves. More precisely, the space domain is reduced to the interval $[0, 2]$, the bottom topography is defined by $z(x) = 2 + 0.25(\cos(10\pi(x - 0.5)) + 1)$ if $1.4 < x < 1.6$, and 2 otherwise, and the initial state is such that $u(0, x) = 0$ and $h(0, x) = 3 - z(x) + \Delta h$ if

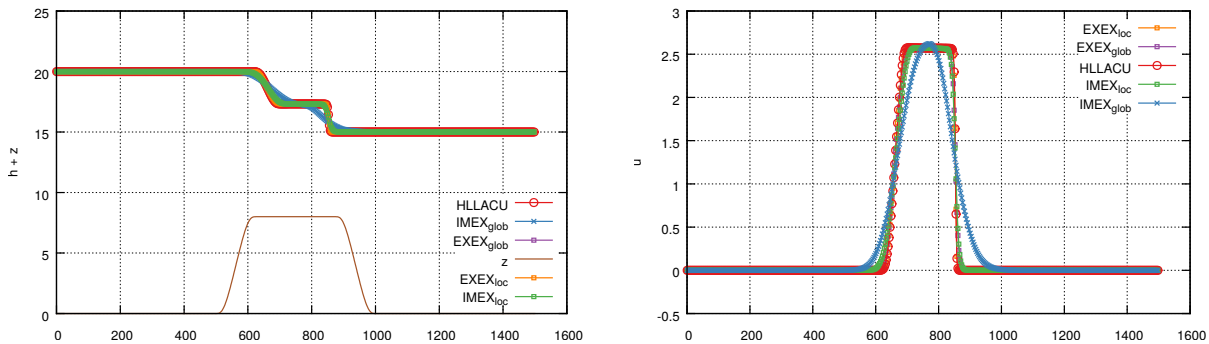

Figure 1.4.1 – Dam Break problem. Profile of $z + h$ and u at time $T = 10$.

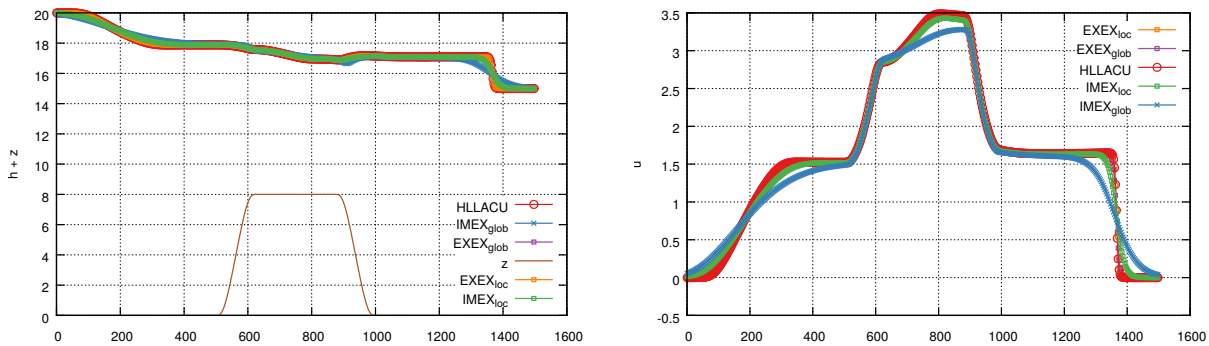

Figure 1.4.2 – Dam Break problem. Profile of $z + h$ and u at time $T = 50$.

$1.1 < x < 1.2$, and $3 - z(x)$ otherwise, where $\Delta h = 0.001$ is the height of the perturbation. The CFL parameter is set to 0.9 (instead of 1/2 in conditions (1.4.3) and (1.4.4)), the final time is $T = 0.2$, the space step equals $\Delta x = 1/500$ and Neumann boundary conditions are used.

It turns out that since the perturbation is small, the values of the velocity u keeps a small amplitude during the whole computation. As an immediate consequence, considering the natural CFL condition (1.4.4) gives very large time steps which naturally induces much numerical diffusion. In order to reduce the numerical diffusion and improve the overall accuracy of the numerical solution, the time step Δt_{imp} given by condition (1.4.4) was first limited to ten times the time step Δt_{exp} given by condition (1.4.3). In other words, we chose

$$\Delta t = \min(10\Delta t_{exp}, \Delta t_{imp})$$

for this test case. Figure 1.4.3 compares the numerical solutions given by the EXEX$_{\text{loc}}$, IMEX$_{\text{loc}}$ and HLLACU schemes. The implicit scheme is clearly more diffusive than the explicit ones. Note that the so-called reference solution is given by the solution of the HLLACU scheme on a 10000-cell grid. Figure 1.4.4 shows that same solutions but the implicit scheme is now run using the explicit CFL restriction (1.4.3). As expected, the numerical approximation is more accurate and the numerical diffusion is significantly reduced.

At last, Figure 1.4.5 shows the numerical solutions using a 10000-cell grid. The schemes converge to the same solution.

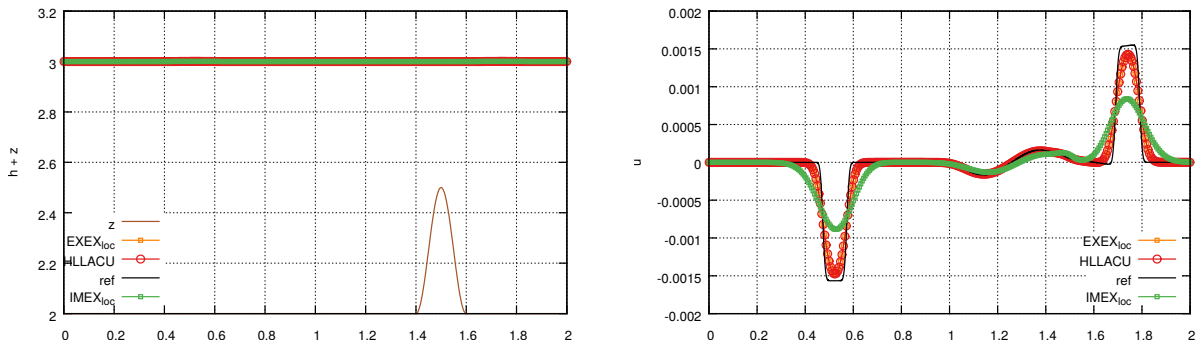

Figure 1.4.3 – Propagation of perturbations test at final time $T = 0.2$. On the left: total heights $h + z$, on the right: velocities u, with $\Delta x = 1/500$.

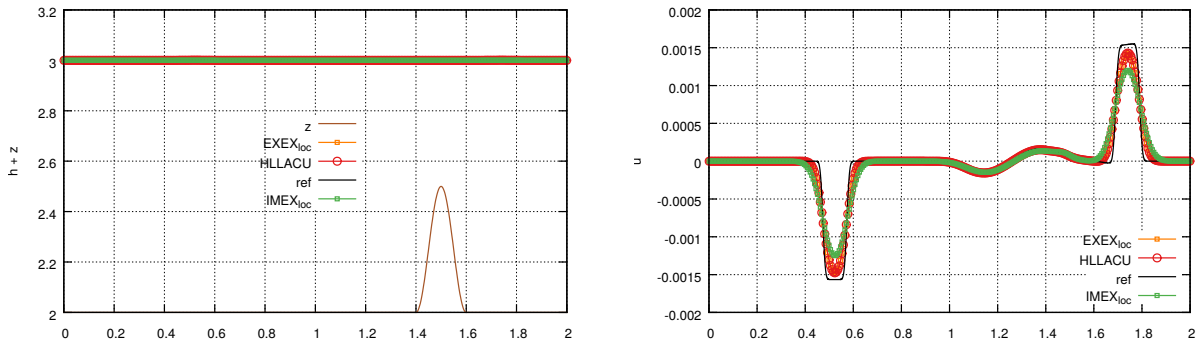

Figure 1.4.4 – Propagation of perturbations test at final time $T = 0.2$. On the left: total heights $h + z$, on the right: velocities u, with $\Delta x = 1/500$. Here, the implicit scheme is run using the explicit CFL restriction (1.4.3).

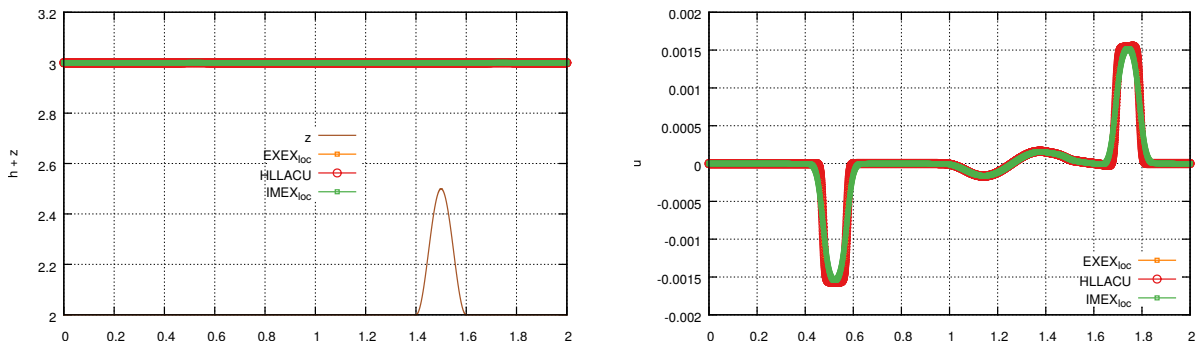

Figure 1.4.5 – Propagation of perturbations test at final time $T = 0.2$. On the left: total heights $h + z$, on the right: velocities u, with $\Delta x = 1/5000$.

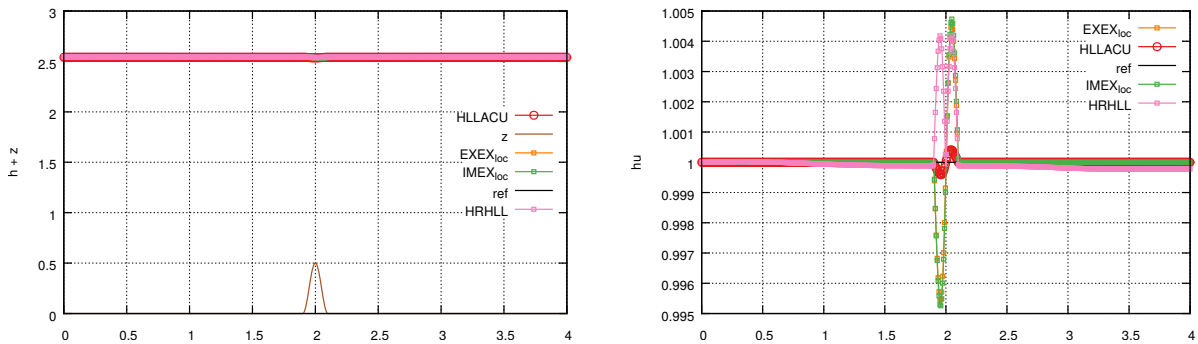

Figure 1.4.6 – Fluvial regime at time $T = 200$. On the left: total heights $h + z$, on the right: discharge hu.

1.4.3 Steady flow over a bump

The aim of this test case is to test the ability of the schemes to converge to some moving water equilibrium. Let us remind that the steady states are governed by the equations $hu = K_1$ and $\dfrac{u^2}{2} + g(h + z) = K_2$.

Fluvial regime

In this test case, we set $K_1 = 1$ and $K_2 = 25$, we denote $h_{eq}(x), u_{eq}(x)$ the values of h and u at this equilibrium. The domain is $[0, 4]$ and the bottom topography is defined by $z(x) = (\cos(10\pi(x - 1)) + 1)/4$ if $1.9 \leq x \leq 2.1$ and 0 elsewhere. The CFL parameter is equal to 0.5 and the space step to $\Delta x = 1/400$. The initial condition is chosen out of equilibrium and given by $h = h_{eq}$ and $u = 0$. The boundary conditions are set to be

$$\begin{cases} \partial_x h(x = 0) = 0, \\ (hu)(x = 0) = K_1, \end{cases} \quad \text{and} \quad \begin{cases} h(x = 4) = h_{eq}(x = 4), \\ \partial_x (hu)(x = 4) = 0. \end{cases}$$

Figure 1.4.6 shows the solution at the final time $t = 200$. We can observe that the solutions are close to the expected equilibrium, except near the mid domain where the momentum is not yet constant for the mesh size under consideration. The Lagrange-projection schemes give numerical solutions very close to the one obtained with the HRHLL scheme based on the hydrostatic reconstruction, while the ACU scheme is clearly more accurate. Note also that on this test case, the implicit CFL condition (1.4.4) allows to use time steps up to ten times larger than the explicit condition (1.4.3).

Transcritical regime without shock

In this test case, we set $K_1 = 3$, $K_2 = \frac{3}{2}(K_1 g)^{2/3} + \frac{g}{2}$. We used the same boundary conditions and started from the same initial condition as in the previous simulation. The solutions are shown at time $t = 10$ on Figure 1.4.7.

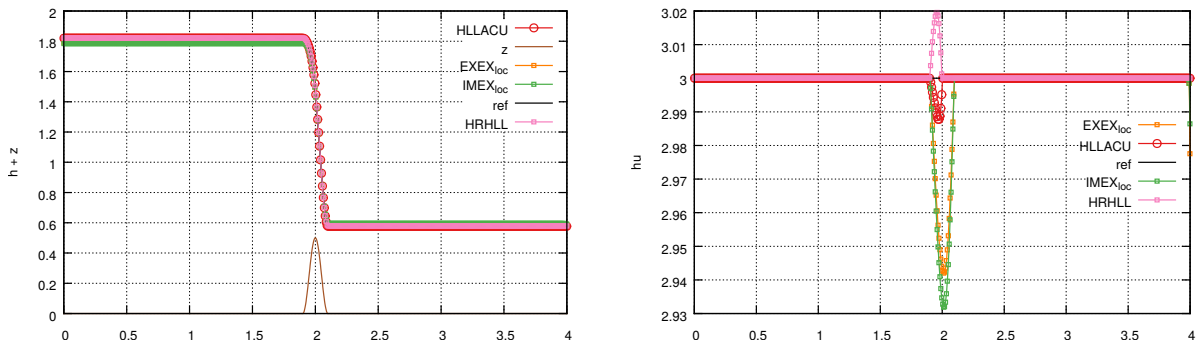

Figure 1.4.7 – Transcritical regime without shock. On the left: total heights $h+z$, on the right: discharge hu.

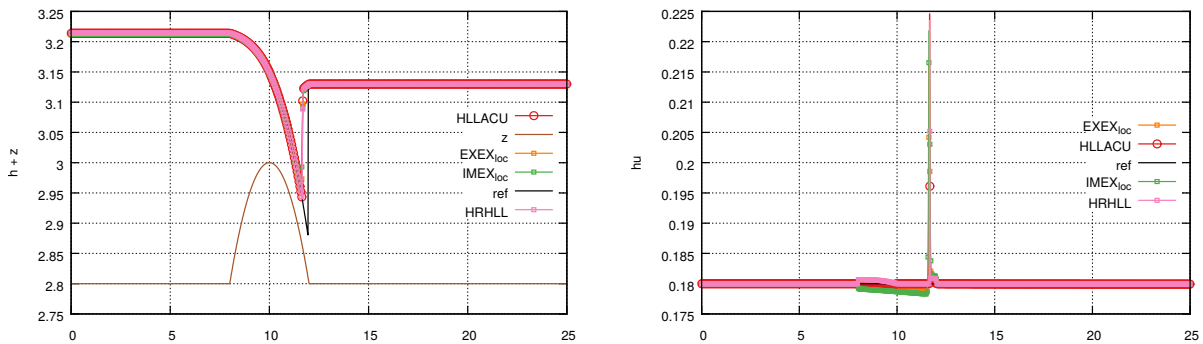

Figure 1.4.8 – Transcritical regime with shock at final time $T = 200$. On the left: total heights $h + z$, on the right: discharge hu.

Transcritical regime with shock

This test has been proposed by Castro et al. [11]. The parameters are described hereafter: the space domain is the interval $[0, 25]$, the bottom topography is defined by $z(x) = 3-0.005(x-10)^2$, if $8 < x < 12$, and 2.8 otherwise. The initial state is defined by $h(0, x) = 3.13 - z(x)$, $q(0, x) = 0.18$ and the boundary conditions are $q(t, 0) = 0.18$, $\partial_x q(t, 25) = 0$, $h(t, 25) = 0.33$ and $\partial_x h(t, 0) = 0$. The final time is set to $t = 200$, the space step to $\Delta x = 1/64$ and the CFL to 0.9.

We can see on the Figure 1.4.8 that we obtain similar results with the different schemes.

1.4.4 Non-unique solution to the Riemann problem

This aim of this test case is to consider a Riemann problem for which the entropy solution is not unique, in order to see whether the numerical schemes capture the same solution or not. The spatial domain is $[0, 1]$, the gravitational acceleration g is set to 2 and the CFL coefficient equals 0.9. Note however that considering the mixed implicit-explicit scheme, the time step Δt was restricted to three times the explicit time step, namely

$$\Delta t = \min(3\Delta t_{exp}, \Delta t_{imp})$$

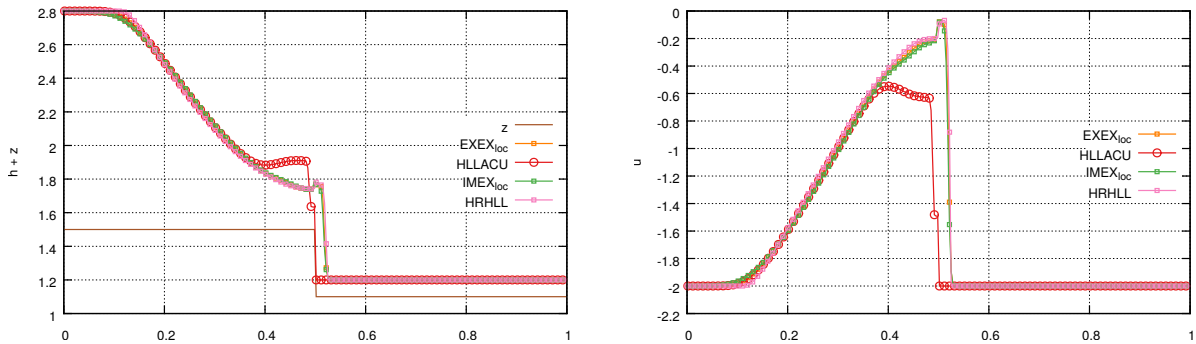

Figure 1.4.9 – Non-unique solution test case at final time $T = 0.1$. On the left: total heights $h + z$, on the right: velocities u.

where we have used the same notations as in the propagation of perturbations test case. The final time $T = 0.1$ and the space step is $\Delta x = 1/300$. The initial data is given by

$$(z, h, u)^T = \begin{cases} (1.5, 1.3, -2)^T & \text{if } x \leq 0.5, \\ (1.1, 0.1, -2)^T & \text{if } x > 0.5, \end{cases}$$

and we used Neumann boundary conditions. It is quite interesting to observe on Figure 1.4.9 that the methods proposed in the present paper and the hydrostatic scheme seem to converge to the same solution, while the HLLACU scheme capture a quite different solution. However, in such a case of multiple entropy solutions, a more general criterion is needed to get uniqueness and define what is the correct solution.

Conclusion

We have proposed a large time step and well-balanced scheme for the shallow-water equations and proved stability properties under a time step CFL restriction based on the material velocity u and not on the sound speed c as it is customary. The Lagrangian-projection decomposition proved to be efficient on a variety of test cases, but may be more diffusive than a direct Eulerian approach. We believe that the proposed implicit-explicit strategy is especially well adapted for subsonic flows but even more for large Froude numbers, which is our very motivation and the purpose of an ongoing work in several space dimensions. Works in progress also include a high-order accuracy extension using discontinuous Galerkin strategies for the space variable and Runge-Kunta techniques for the time variable.

Acknowledgement

This work was partially supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. The authors also thank S. Noelle and H. Zakerzadeh for useful and interesting discussions on the topic.

1.A Eigenstructure of the relaxed acoustic system

Considering smooth solutions, the homogeneous relaxed acoustic system (1.2.9) reads

$$\partial_t \mathbf{W} + A(\mathbf{W})\partial_m \mathbf{W} = 0, \quad A = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & g/\tau \\ 0 & a^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{1.A.1}$$

The matrix the eigenvalues of A are $\{-a, 0, +a\}$. A basis of right eigenvectors of A is

$$\mathbf{r}_0^{(1)} = (1,0,0,0)^T, \qquad \mathbf{r}_0^{(2)} = (0,0,-g,\tau)^T, \qquad \mathbf{r}_\pm = (1, \mp a, -a^2, 0)^T,$$

where $\mathbf{r}_0^{(1)}$ and $\mathbf{r}_0^{(1)}$ are associated with the double eigenvalue 0 and $\mathbf{r}_\pm$ is associated with $\pm a$. The system (1.A.1) is thus hyperbolic. All the characteristic fields of (1.A.1) are linearly degenerate.

The $(\pm a)$-field possesses three Riemann invariants

$$I_\pm^1 = \pi \mp au, \qquad I_\pm^2 = u \pm a\tau, \qquad I_\pm^3 = z.$$

As a consequence, the states $\mathbf{W}_L$ and $\mathbf{W}_R$ that can be connected by a $(\pm a)$-wave can be obtained thanks to the continuity of the $(\pm a)$-Riemann invariants, which amounts to verify the jump relations

$$\mp a(\mathbf{W}_R - \mathbf{W}_R) + \mathbf{G}(\mathbf{W}_R) - \mathbf{G}(\mathbf{W}_R) = 0. \tag{1.A.2}$$

Unfortunately, the eigenvalue 0 is of multiplicity 2 and the 0-field only has a single Riemann invariant

$$I_0 = u.$$

Therefore we can only state that if two states $\mathbf{W}_L$ and $\mathbf{W}_R$ are connected by a 0-wave then

$$u_R = u_L. \tag{1.A.3}$$

1.B Proof of the discrete entropy inequality of Proposition 1.3.2

The proof of the discrete entropy inequality follows exactly the same lines as the one proposed in [23] for the barotropic gas dynamics equations, but taking into account here the presence of the topography source term. It naturally leads to a non conservative version of the entropy inequality. A discrete and conservative entropy inequality for the proposed algorithm remains an open problem so far. Our result states as follows.

Lemma 1.B.1. *We have the following discrete form of the entropy inequality* (1.1.3) *for all* $j \in \mathbb{Z}$*, namely*

$$\mathcal{U}_j^{n+1} - \mathcal{U}_j^n + \frac{\Delta t}{\Delta x_j}\left(\mathcal{F}_{j+1/2}^{n+1-} - \mathcal{F}_{j-1/2}^{n+1-}\right) \leq -\Delta t\, g\left\{hu\partial_x z\right\}_j,$$

with the entropy numerical fluxes

$$\mathcal{F}_{j+1/2}^{n+1-} = \left(\Pi_{j+1/2}^* + \mathcal{U}_{j+1/2}^{n+1-}\right)\tilde{u}_{j+1/2}^*,$$

where

$$\mathcal{U}_{j+1/2}^{n+1-} = \begin{cases} \mathcal{U}_j^{n+1-} & \text{if } u_{j+1/2}^* \geq 0, \\ \mathcal{U}_j^{n+1-} & \text{if } u_{j+1/2}^* < 0, \end{cases}$$

and

$$\begin{cases} \tilde{u}_{j+1/2}^* = \dfrac{u_j^{n+1-} + u_{j+1}^{n+1-}}{2} - \dfrac{1}{2a}\left(\Pi_{j+1}^{n+1-} - \Pi_j^{n+1-}\right), \\ \Pi_{j+1/2}^* = \dfrac{\Pi_j^{n+1-} + \Pi_{j+1}^{n+1-}}{2} - \dfrac{a}{2}\left(u_{j+1}^{n+1-} - u_j^{n+1-}\right), \end{cases}$$

are consistent with $\mathcal{F}$, and the non conservative source term

$$\{hu\partial_x z\}_j = \frac{1}{2a\Delta x_j}\left[\frac{h_j^n + h_{j-1}^n}{2}\overrightarrow{w}_j^{n+1-}(z_j - z_{j-1}) - \frac{h_{j+1}^n + h_j^n}{2}\overleftarrow{w}_j^{n+1-}(z_{j+1} - z_j)\right]$$

is consistent with $hu\partial_x z$.

Proof. Let us first observe that smooth solutions of Equation (1.2.9) satisfy

$$\partial_t(\Pi^2 + a^2u^2) + 2a^2\partial_m \Pi u = -\frac{2a^2 gu}{\tau}\partial_m z. \tag{1.B.1}$$

In order to obtain a discrete version of this equality, let us define

$$\eta_j^{n+1-} := \frac{(\overleftarrow{w}_j^{n+1-})^2 + (\overrightarrow{w}_j^{n+1-})^2}{2} = (\Pi_j^{n+1-})^2 + a^2(u_j^{n+1-})^2,$$

and

$$q_{j+1/2}^{n+1-} := \frac{(\overrightarrow{w}_j^{n+1-})^2 - (\overleftarrow{w}_{j+1}^{n+1-})^2}{4a} = \Pi_{j+1/2}^*\,\tilde{u}_{j+1/2}^*.$$

The formulas (1.3.10) also read

$$\begin{cases} \tau_j^{n+1^-} - \tau_j^n = \dfrac{\Delta t}{\Delta m_j}\left[u_{j+1/2}^* - u_{j-1/2}^*\right], \\ \overleftarrow{w}_j^{n+1-} - \overleftarrow{w}_j^n = a\dfrac{\Delta t}{\Delta m_j}\left[\overleftarrow{w}_{j+1}^{n+1-} - \overleftarrow{w}_j^{n+1-} + g\dfrac{h_{j+1}^n + h_j^n}{2}(z_{j+1} - z_j)\right], \\ \overrightarrow{w}_j^{n+1-} - \overrightarrow{w}_j^n = -a\dfrac{\Delta t}{\Delta m_j}\left[\overrightarrow{w}_j^{n+1-} - \overrightarrow{w}_{j-1}^{n+1-} + g\dfrac{h_j^n + h_{j-1}^n}{2}(z_j - z_{j-1})\right], \end{cases}$$

while adding the third equation of (1.3.8) and a^2 times the first equation of (1.3.8) also gives $I_j^{n+1-} = I_j^n$ where $I = \Pi + a^2\tau$. Multiplying the second and the third equations above by $\overleftarrow{w}_j^{n+1-}$ and $\overrightarrow{w}_j^{n+1-}$ then gives

$$\begin{cases} I_j^{n+1-} = I_j^n, \\ \overleftarrow{w}_j^{n+1-}(\overleftarrow{w}_j^{n+1-} - \overleftarrow{w}_j^n) \\ \qquad = a\dfrac{\Delta t}{\Delta m_j}\left[\overleftarrow{w}_j^{n+1-}\left(\overleftarrow{w}_{j+1}^{n+1-} - \overleftarrow{w}_j^{n+1-}\right) + \overleftarrow{w}_j^{n+1-} g\dfrac{h_{j+1}^n + h_j^n}{2}(z_{j+1} - z_j)\right], \\ \overrightarrow{w}_j^{n+1-}(\overrightarrow{w}_j^{n+1-} - \overrightarrow{w}_j^n) \\ \qquad = -a\dfrac{\Delta t}{\Delta m_j}\left[\overrightarrow{w}_j^{n+1-}\left(\overrightarrow{w}_j^{n+1-} - \overrightarrow{w}_{j-1}^{n+1-}\right) + \overrightarrow{w}_j^{n+1-} g\dfrac{h_j^n + h_{j-1}^n}{2}(z_j - z_{j-1})\right], \end{cases}$$

that is to say, since

$$2b(b-a) = (b^2-a^2) + (b-a)^2 \quad \text{and} \quad 2b(a-b) = (a^2-b^2) - (b-a)^2,$$

$$\begin{cases} I_j^{n+1-} = I_j^n, \\ \left((\overleftarrow{w}_j^{n+1-})^2 - (\overleftarrow{w}_j^n)^2\right) - a\dfrac{\Delta t}{\Delta m_j}\left((\overleftarrow{w}_{j+1}^{n+1-})^2 - (\overleftarrow{w}_j^{n+1-})^2\right) = \\ \qquad -(\overleftarrow{w}_j^{n+1-} - \overleftarrow{w}_j^n)^2 + a\frac{\Delta t}{\Delta m_j}\left[-\left(\overleftarrow{w}_{j+1}^{n+1-} - \overleftarrow{w}_j^{n+1-}\right)^2 + 2\overleftarrow{w}_j^{n+1-} g\frac{h_{j+1}^n + h_j^n}{2}(z_{j+1}-z_j)\right], \\ \left((\overrightarrow{w}_j^{n+1-})^2 - (\overrightarrow{w}_j^n)^2\right) + a\dfrac{\Delta t}{\Delta m_j}\left((\overrightarrow{w}_j^{n+1-})^2 - (\overrightarrow{w}_{j-1}^{n+1-})^2\right) = \\ \qquad -(\overrightarrow{w}_j^{n+1-} - \overrightarrow{w}_j^n)^2 - a\frac{\Delta t}{\Delta m_j}\left[\left(\overrightarrow{w}_j^{n+1-} - \overrightarrow{w}_{j-1}^{n+1-}\right)^2 + 2\overrightarrow{w}_j^{n+1-} g\frac{h_j^n + h_{j-1}^n}{2}(z_j - z_{j-1})\right]. \end{cases}$$

Summing the last two equations, we immediately get the following discrete version of (1.B.1), namely

$$\eta_j^{n+1-} - \eta_j^n + 2a^2\frac{\Delta t}{\Delta m_j}(q_{j+1/2}^{n+1-} - q_{j-1/2}^{n+1-}) \le -\Delta t\, 2a^2\, g\, \tau_j^n \left\{hu\partial_x z\right\}_j.$$

The rest of the proof strictly follows the one proposed in [23]. It is given here for the sake of completeness. With this in mind, let us define the energy E such that $hE = \mathcal{U}$, which means

$$E = \frac{u^2}{2} + e(\tau) = \frac{u^2}{2} + e(\tau) + \frac{\Pi^2 - \Pi^2}{2a^2} = e(\tau) + \frac{\eta - \Pi^2}{2a^2},$$

where we have set $e(\tau) = \frac{g}{2\tau} = \frac{gh}{2}$. We clearly have

$$E_j^{n+1-} - E_j^n = e(\tau_j^{n+1-}) - e(\tau_j^n) + \frac{\eta_j^{n+1-} - \eta_j^n}{2a^2} - \frac{(\Pi_j^{n+1-})^2 - (\Pi_j^n)^2}{2a^2}$$

so that, since $a^2 - b^2 = (a-b)^2 + 2b(b-a)$, we have

$$E_j^{n+1-} - E_j^n = e(\tau_j^{n+1-}) - e(\tau_j^n) + \frac{\eta_j^{n+1-} - \eta_j^n}{2a^2} - \frac{(\Pi_j^{n+1-} - \Pi_j^n)^2}{2a^2} - \frac{\Pi_j^n(\Pi_j^{n+1-} - \Pi_j^n)}{a^2}.$$

But $I_j^{n+1-} = I_j^n$ gives $\Pi_j^{n+1-} - \Pi_j^n = -a^2(\tau_j^{n+1-} - \tau_j^n)$ so that

$$\begin{aligned} &E_j^{n+1-} - E_j^n + \frac{\Delta t}{\Delta m_j}(q_{j+1/2}^{n+1-} - q_{j-1/2}^{n+1-}) \\ &\qquad \le e(\tau_j^{n+1-}) - e(\tau_j^n) + \Pi_j^n(\tau_j^{n+1-} - \tau_j^n) - \frac{a^2}{2}(\tau_j^{n+1-} - \tau_j^n)^2 - \Delta t\, g\, \tau_j^n \left\{hu\partial_x z\right\}_j. \end{aligned}$$

Since the solution at time t^n is at equilibrium, we have $\Pi_j^n = p(\tau_j^n) = -e'(\tau_j^n) = \dfrac{g}{2}(h_j^n)^2$, so that a Taylor expansion gives

$$E_j^{n+1-} - E_j^n + \frac{\Delta t}{\Delta m_j}(q_{j+1/2}^{n+1-} - q_{j-1/2}^{n+1-}) \le \frac{(e''(\xi) - a^2)}{2}(\tau_j^{n+1-} - \tau_j^n)^2 - \Delta t\, g\, \tau_j^n \left\{hu\partial_x z\right\}_j,$$

and

$$\begin{aligned} &E_j^{n+1-} - E_j^n + \frac{\Delta t}{\Delta m_j}(q_{j+1/2}^{n+1-} - q_{j-1/2}^{n+1-}) \\ &\qquad \le \frac{(-p'(\xi) - a^2)}{2}(\tau_j^{n+1-} - \tau_j^n)^2 - \Delta t\, g\, \tau_j^n \left\{hu\partial_x z\right\}_j \le -\Delta t\, g\, \tau_j^n \left\{hu\partial_x z\right\}_j, \end{aligned}$$

by the Whitham subcharacteristic condition. This inequality is nothing but the expected entropy inequality but in Lagrangian coordinates. At this stage, it is very usual to combine the definition of the remap step (which, setting $X = h, hu$, gives X_j^{n+1} as a convex combination of X_{j-1}^{n+1-}, X_j^{n+1-} and X_{j+1}^{n+1-} under the transport CFL condition) together with the Jensen inequality for the convex mapping $(h, hu) \mapsto \mathcal{U}(h, hu)$, in order to get the expected entropy inequality in Eulerian coordinates, namely

$$\mathcal{U}_j^{n+1} - \mathcal{U}_j^n + \frac{\Delta t}{\Delta x_j}\left(\mathcal{F}_{j+1/2}^{n+1-} - \mathcal{F}_{j-1/2}^{n+1-}\right) \leq -\Delta t\, g\left\{hu\partial_x z\right\}_j .$$

We refer the reader to [23] for more details. □

1.C Fermeture du problème de Riemann associé à l'étape acoustique à l'aide des invariants

On propose dans cette annexe une autre méthode permettant d'obtenir la même solution exacte au problème de Riemann associé au système relaxé du pas acoustique (1.2.5). Dans la Section 1.2.3, on introduit un solveur de Riemann approché qui est représenté sur la Figure 1.C.1. Il y est également décrit une méthodologie permettant de définir les états intermédiaires de la solution $\mathbf{W}_{\mathrm{RP}}$ définit dans l'Equation (1.2.12). On souhaite en effet que la solution $\mathbf{W}_{\mathrm{RP}}$ vérifie 3 propriétés, que l'on rappelle ici brièvement mais qui sont détaillées dans la Section 1.2.3 :

a) $\mathbf{W}_{\mathrm{RP}}$ est consistent au sens intégral avec les équations de Saint-Venant,

b) lorsque la topographie est constante, on souhaite retrouver la solution $\mathbf{W}_{\mathrm{RP}}$ classique du problème de Riemann,

c) si les états initiaux $\mathbf{W}_L$ et $\mathbf{W}_R$ vérifient la condition du lac au repos, alors la solution reste identique à cet état au cours du temps.

Les états intermédiaires ($\mathbf{W}_L^*$ et $\mathbf{W}_R^*$) sont alors déterminés grâce aux relations de sauts (1.2.16) et (1.2.17). Notons qu'il manque une relation pour fermer le système, c'est pourquoi la relation (1.2.18) est proposée, où la variable $\mathscr{M}$ est déterminée grâce aux propriétés précédentes a), b) et c).

On revient dans cette annexe sur l'analyse des invariants de Riemann qui a permis d'obtenir cette dernière relation de saut (1.2.18). L'analyse du système acoustique relaxé (1.2.5) réalisée dans l'annexe 1.A nous donne seulement 7 invariants ($I_\pm^1$, $I_\pm^2$, $I_\pm^3$, I_0), au lieu de 8 nécessaires pour déterminer tous les états intermédiaires. On écrit la continuité de ces invariants afin d'obtenir le système d'équations suivant

$$\begin{cases} u_L^* - a\tau_L^* = u_L - a\tau_L, \\ \Pi_L^* + a^2\tau_L^* = \Pi_L + a^2\tau_L, \\ z_L^* = z_L, \end{cases} \qquad \left\{ u_L^* = u_R^*, \right. \qquad \begin{cases} u_R^* + a\tau_R^* = u_R + a\tau_R, \\ \Pi_R^* + a^2\tau_R^* = \Pi_R + a^2\tau_R, \\ z_R^* = z_R, \end{cases}$$

On peut alors exprimer les solutions de ce système en fonction d'une variable u^* :

$$\begin{cases} u_L^* = u^*, \\ \tau_L^* = \tau_L + \dfrac{1}{a}\left(u^* - u_L\right), \\ \Pi_L^* = \Pi_L - a\left(u^* - u_L\right), \\ z_L^* = z_L, \end{cases} \qquad \begin{cases} u_R^* = u^*, \\ \tau_R^* = \tau_R - \dfrac{1}{a}\left(u^* - u_R\right), \\ \Pi_R^* = \Pi_R + a\left(u^* - u_R\right), \\ z_R^* = z_R, \end{cases} \tag{1.C.1}$$

Les trois prochaines sections détaillent chacune une relation particulière que l'on a essayé d'imposer afin de fermer le système.

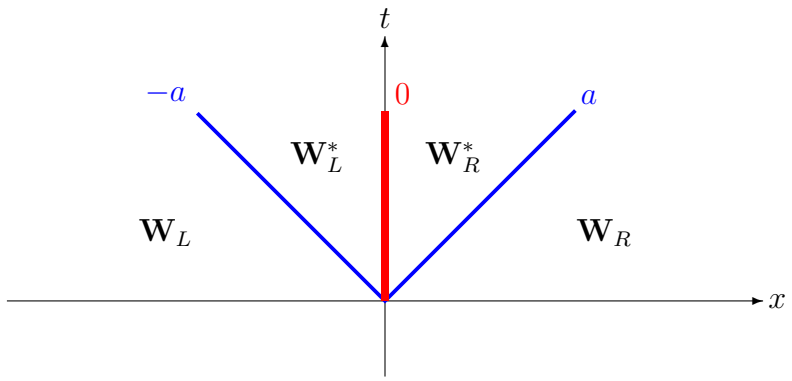

FIGURE 1.C.1 – Solveur de Riemann approché du système acoustique relaxé

1.C.1 Première approche : un invariant naturel

On ajoute tout d'abord l'invariant qui semble être le plus naturel afin d'obtenir la propriété équilibre c), à savoir l'invariant $I_0^2 = h + z = \frac{1}{\tau} + z$ qui discrétise cette propriété. On remarque tout de suite le caractère non-linéaire de la dernière équation $\frac{1}{\tau_L^*} + z_L^* = \frac{1}{\tau_R^*} + z_R^*$ qu'impose la continuité de cet invariant.

On peut même se retrouver dans un cas où le système n'a pas de solution, si $z_L = z_R$ et $a\left(\tau_L + \tau_R\right) = -\left(u_L + u_R\right)$ par exemple. En effet, on a dans ces conditions $\tau_L^* = \tau_R^*$ car $z_L^* = z_L$ et $z_R^* = z_R$, avec $\tau_L^* \neq 0$ et $\tau_R^* \neq 0$. Donc $u^* = \frac{u_L + u_R}{2} + \frac{a}{2}\left(\tau_R - \tau_L\right)$ d'après le système (1.C.1) et $\tau_L^* = \frac{\tau_L + \tau_R}{2} + \frac{1}{2a}\left(u_R - u_L\right) = 0$, ce qui est absurde.

Pour toutes ces raisons, on se propose de chercher un autre invariant permettant d'obtenir une solution quelques soient les conditions initiales admissibles et d'écrire un schéma plus simple tout en satisfaisant les trois propriétés souhaitées.

1.C.2 Deuxième approche : un invariant linéarisé

Afin de faire disparaître la non-linéarité de la première approche, on propose d'imposer une version linéarisée de la continuité de l'invariant naturel ci-dessus, à savoir l'équation

$$\frac{\tau_L^*}{\tau_L^2} + z_L = \frac{\tau_R^*}{\tau_R^2} + z_R.$$

On obtient après résolution du système linéaire l'unique solution suivante

$$\begin{cases} \tau_L^* = \tau_L + \dfrac{h_R^2/a}{h_R^2+h_L^2}\left(u_R-u_L\right) + \dfrac{1}{h_R^2+h_L^2}\left(h_R+z_R-h_L-z_L\right), \\ \tau_R^* = \tau_R + \dfrac{h_L^2/a}{h_R^2+h_L^2}\left(u_R-u_L\right) - \dfrac{1}{h_R^2+h_L^2}\left(h_R+z_R-h_L-z_L\right), \\ u^* = \dfrac{h_R^2u_R+h_L^2u_L}{h_R^2+h_L^2} + \dfrac{a}{h_R^2+h_L^2}\left(h_R+z_R-h_L-z_L\right), \\ \Pi_L^* = \Pi_L - \dfrac{ah_R^2}{h_R^2+h_L^2}\left(u_R-u_L\right) - \dfrac{a^2}{h_R^2+h_L^2}\left(h_R+z_R-h_L-z_L\right), \\ \Pi_R^* = \Pi_R - \dfrac{ah_L^2}{h_R^2+h_L^2}\left(u_R-u_L\right) + \dfrac{a^2}{h_R^2+h_L^2}\left(h_R+z_R-h_L-z_L\right). \end{cases}$$

Si on suppose que l'on a initialement un état stationnaire de type lac au repos, soient $u_L = u_R = 0$ et $h_L + z_L = h_R + z_R$, alors il est assez aisé de vérifier que l'on obtient comme solution

$$\begin{cases} \tau_L^* = \tau_L, \\ \tau_R^* = \tau_R, \\ u^* = 0, \\ \Pi_L^* = \Pi_L, \\ \Pi_R^* = \Pi_R, \end{cases}$$

ce qui nous assure la propriété équilibre (c) du pas acoustique.

Malheureusement, le schéma ainsi construit ne satisfait pas la deuxième propriété (b) puisqu'il ne dégénère pas vers le schéma standard lorsque la topographie est constante. En effet, si l'on suppose $z_L = z_R$, nous obtenons les formules suivantes pour les inconnues

$$\begin{cases} \tau_L^* = \tau_L + \dfrac{h_R^2/a}{h_R^2+h_L^2}\left(u_R-u_L\right), \\ \tau_R^* = \tau_R + \dfrac{h_L^2/a}{h_R^2+h_L^2}\left(u_R-u_L\right), \\ u^* = \dfrac{h_R^2u_R+h_L^2u_L}{h_R^2+h_L^2}, \\ \Pi_L^* = \Pi_L - \dfrac{ah_R^2}{h_R^2+h_L^2}\left(u_R-u_L\right), \\ \Pi_R^* = \Pi_R - \dfrac{ah_L^2}{h_R^2+h_L^2}\left(u_R-u_L\right), \end{cases}$$

qui ne correspondent pas aux états intermédiaires souhaités, à savoir

$$\begin{cases} \tau_L^* = \tau_L + \dfrac{1}{2a}\left(u_R-u_L\right) - \dfrac{1}{2a^2}\left(\Pi_R-\Pi_L\right), \\ \tau_R^* = \tau_R + \dfrac{1}{2a}\left(u_R-u_L\right) + \dfrac{1}{2a^2}\left(\Pi_R-\Pi_L\right), \\ u^* = \dfrac{u_R+u_L}{2} - \dfrac{1}{2a}\left(\Pi_R-\Pi_L\right), \\ \Pi_L^* = \Pi_R^* = \dfrac{\Pi_R+\Pi_L}{2} - \dfrac{a}{2}\left(u_R-u_L\right). \end{cases}$$

1.C.3 Dernière approche : un invariant perturbé

On peut remarquer qu'avec la deuxième approche, on n'obtient pas la relation $\Pi_L^* = \Pi_R^*$ attendue lorsque la topographie est constante. Pour cette raison, on propose de perturber le dernier invariant associé au système d'Euler isentropique, à savoir $I_0^2 = \Pi$. On lui ajoute une fonction dépendant linéairement de la topographie z, ce qui nous assure la propriété b), et il nous reste à trouver une constante adaptée permettant d'obtenir les deux autres propriétés a) et c). Il parait assez naturel, pour des raisons de dimensions physiques, de faire apparaître la constante de gravitation g dans cette fonction linéaire, et donc une constante $\tilde{h}$ qui représenterait une hauteur d'eau, ce qui donne comme dernier invariant $I_0^2 = \Pi + g\tilde{h}z$.

Lorsque l'on résout le problème de Riemann avec pour dernière équation

$$\Pi_R^* - \Pi_L^* = -g\tilde{h}\left(z_R^* - z_L^*\right), \tag{1.C.2}$$

on obtient

$$\begin{cases} \tau_L^* = \tau_L + \dfrac{1}{2a}\left(u_R - u_L\right) - \dfrac{1}{2a^2}\left(\Pi_R - \Pi_L\right) - \dfrac{g\tilde{h}}{2a^2}\left(z_R - z_L\right), \\ \tau_R^* = \tau_R + \dfrac{1}{2a}\left(u_R - u_L\right) + \dfrac{1}{2a^2}\left(\Pi_R - \Pi_L\right) + \dfrac{g\tilde{h}}{2a^2}\left(z_R - z_L\right), \\ u^* = \dfrac{u_R + u_L}{2} - \dfrac{1}{2a}\left(\Pi_R - \Pi_L\right) - \dfrac{g\tilde{h}}{2a}\left(z_R - z_L\right), \\ \Pi_L^* = \dfrac{\Pi_R + \Pi_L}{2} - \dfrac{a}{2}\left(u_R - u_L\right) + \dfrac{g\tilde{h}}{2}\left(z_R - z_L\right), \\ \Pi_R^* = \dfrac{\Pi_R + \Pi_L}{2} - \dfrac{a}{2}\left(u_R - u_L\right) - \dfrac{g\tilde{h}}{2}\left(z_R - z_L\right). \end{cases} \tag{1.C.3}$$

Pour la propriété équilibre (c), on prend tout d'abord comme conditions initiales :

$$\begin{cases} u_L = u_R = 0, \\ \dfrac{1}{\tau_L} + z_L = \dfrac{1}{\tau_R} + z_R, \end{cases}$$

en n'oubliant pas que la relaxation instantanée précédant l'étape acoustique impose :

$$\begin{cases} \Pi_L = \dfrac{g}{2\tau_L^2}, \\ \Pi_R = \dfrac{g}{2\tau_R^2}. \end{cases}$$

On obtient :

$$\begin{aligned} \tau_R^* &= \tau_R + \frac{1}{2a}\left(u_R - u_L\right) + \frac{1}{2a^2}\left(\Pi_R - \Pi_L\right) + \frac{g\tilde{h}}{2a^2}\left(z_R - z_L\right) \\ &= \tau_R + \frac{g\left(h_R - h_L\right)}{2a^2}\left(\frac{h_R + h_L}{2} - \tilde{h}\right). \end{aligned}$$

Afin que le schéma soit équilibre, il est nécessaire que $\frac{1}{\tau_R^*} + z_R^* = \frac{1}{\tau_R} + z_R$. Comme $z_R^* = z_R$ alors nécessairement $\tau_R^* = \tau_R$, soit :

$$\tilde{h} = \frac{h_R + h_L}{2}.$$

On vérifie que dans ce cas, les états intermédiaires deviennent

$$\begin{cases} \tau_L^* = \tau_L, \\ \tau_R^* = \tau_R, \\ u_L^* = u_R^* = 0, \\ \Pi_L^* = \dfrac{g}{2\tau_L^2}, \\ \Pi_R^* = \dfrac{g}{2\tau_R^2}. \end{cases}$$

Donc l'état d'équilibre est préservé, par conséquent la troisième propriété (c) est satisfaite.

On peut à présent remarquer que si l'on pose

$$\mathscr{M} = g\tilde{h}\left(z_R - z_L\right) = g\frac{h_L + h_R}{2}\left(z_R - z_L\right),$$

la dernière équation (1.C.2) que l'on impose correspond à la relation de saut (1.2.18), les états intermédiaires (1.C.3) correspondent à ceux donnés à l'Equation (1.2.19) et la variable $\mathscr{M}$ satisfait bien les relations (1.2.20) et (1.2.23). On a par voie de conséquence la consistance au sens intégral (1.2.13) et donc la propriété a).

Ainsi on a trouvé une dernière relation à imposer permettant d'obtenir les trois propriétés a), b) et c), sous la forme :

$$\Pi_R^* - \Pi_L^* = -\mathscr{M} = -\Delta x\ \{gh\partial_x z\}.$$

On a également montré que l'on est bien retombé sur le même schéma que celui proposé dans la Section 1.2.3.

Bibliographie

[1] Emmanuel Audusse, François Bouchut, Marie-Odile Bristeau, Rupert Klein, and Benoît Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing*, 25(6) :2050–2065, 2004.

[2] Emmanuel Audusse, Christophe Chalons, and Philippe Ung. A very simple well-balanced positive and entropy-satisfying scheme for the shallow-water equations. *Commun. Math. Sci*, 13(5) :1317–1332, 2015.

[3] Alfredo Bermudez and Ma Elena Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8) :1049–1071, 1994.

[4] Christophe Berthon and Christophe Chalons. A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations. *Mathematics of Computation*, 85(299) :1281–1307, 2016.

[5] Christophe Berthon, Christophe Chalons, Sélim Cornet, and Gianmarco Sperone. Fully well-balanced, positive and simple approximate Riemann solver for shallow water equations. *Bulletin of the Brazilian Mathematical Society, New Series*, 47(1) :117–130, 2016.

[6] Christophe Berthon and Françoise Foucher. Efficient well-balanced hydrostatic upwind schemes for shallow-water equations. *Journal of Computational Physics*, 231(15) :4993–5015, 2012.

[7] Georgij Bispen, Koottungal Revi Arun, Mária Lukáčová-Medvid'ová, and Sebastian Noelle. IMEX large time step finite volume methods for low Froude number shallow water flows. *Communications in Computational Physics*, 16(2) :307–347, 2014.

[8] François Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Springer Science & Business Media, 2004.

[9] François Bouchut. A reduced stability condition for nonlinear relaxation to conservation laws. *Journal of Hyperbolic Differential Equations*, 1(01) :149–170, 2004.

[10] François Bouchut and Tomas Morales De Luna. A subsonic-well-balanced reconstruction scheme for shallow water flows. *SIAM Journal on Numerical Analysis*, 48(5) :1733–1758, 2010.

[11] Manuel J Castro, Alberto Pardo Milanés, and Carlos Parés. Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique. *Mathematical Models and Methods in Applied Sciences*, 17(12) :2055–2113, 2007.

[12] Christophe Chalons and Frédéric Coquel. Navier-Stokes equations with several independent pressure laws and explicit predictor-corrector schemes. *Numerische Mathematik*, 101(3) :451–478, 2005.

[13] Christophe Chalons, Frédéric Coquel, Edwige Godlewski, Pierre-Arnaud Raviart, and Nicolas Seguin. Godunov-type schemes for hyperbolic systems with parameter-dependent source : the case of Euler system with friction. *Mathematical Models and Methods in Applied Sciences*, 20(11) :2109–2166, 2010.

[14] Christophe Chalons, Frédéric Coquel, Samuel Kokh, and Nicole Spillane. Large time-step numerical scheme for the seven-equation model of compressible two-phase flows. In *Finite Volumes for Complex Applications VI Problems & Perspectives*, pages 225–233. Springer, 2011.

[15] Christophe Chalons and Jean-François Coulombel. Relaxation approximation of the Euler equations. *Journal of Mathematical Analysis and Applications*, 348(2) :872–893, 2008.

[16] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. Large time step and asymptotic preserving numerical schemes for the gas dynamics equations with source terms. *SIAM Journal on Scientific Computing*, 35(6) :A2874–A2902, 2013.

[17] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. Operator-splitting based AP schemes for the 1D and 2D gas dynamics equations with stiff sources. *AIMS Series on Applied Mathematics*, 8 :607–614, 2014.

[18] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An all-regime Lagrange-projection like scheme for the gas dynamics equations on unstructured meshes. *Communications in Computational Physics*, 20(1) :188–233, 2016.

[19] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An all-regime Lagrange-projection like scheme for 2D homogeneous models for two-phase flows on unstructured meshes. *Journal of Computational Physics*, 335 :885–904, 2017.

[20] Gui-Qiang Chen, C David Levermore, and Tai-Ping Liu. Hyperbolic conservation laws with stiff relaxation terms and entropy. *Communications on Pure and Applied Mathematics*, 47(6) :787–830, 1994.

[21] Ashwin Chinnayya, Alain-Yves LeRoux, and Nicolas Seguin. A well-balanced numerical scheme for the approximation of the shallow-water equations with topography : the resonance phenomenon. *Int. J. Finite Volumes*, 1 :1–33, 2004.

[22] Frédéric Coquel, Edwige Godlewski, Benoıt Perthame, Arun In, and Paul Rascle. Some new Godunov and relaxation methods for two-phase flow problems. In *Godunov methods*, pages 179–188. Springer, 2001.

[23] Frédéric Coquel, Quang Nguyen, Marie Postel, and Quang Tran. Entropy-satisfying relaxation method with large time-steps for Euler IBVPs. *Mathematics of Computation*, 79(271) :1493–1533, 2010.

[24] Gérard Gallice. Solveurs simples positifs et entropiques pour les systemes hyperboliques avec terme source. *Comptes Rendus Mathématique*, 334(8) :713–716, 2002.

[25] Gérard Gallice. Positive and entropy stable godunov-type schemes for gas dynamics and MHD equations in Lagrangian or Eulerian coordinates. *Numerische Mathematik*, 94(4) :673–713, 2003.

[26] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, volume 118. Springer Science & Business Media, 1996.

[27] Laurent Gosse. A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. *Computers & Mathematics with Applications*, 39(9-10) :135–159, 2000.

[28] Laurent Gosse. *Computing qualitatively correct approximations of balance laws*, volume 2. Springer, 2013.

[29] Joshua M Greenberg and Alain-Yves LeRoux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM Journal on Numerical Analysis*, 33(1) :1–16, 1996.

[30] Shi Jin. A steady-state capturing method for hyperbolic systems with geometrical source terms. *ESAIM : Mathematical Modelling and Numerical Analysis*, 35(4) :631–645, 2001.

[31] Shi Jin and Zhouping Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Communications on pure and applied mathematics*, 48(3) :235–276, 1995.

[32] Benoît Perthame and Chiara Simeoni. A kinetic scheme for the saint-venant system with a source term. *Calcolo*, 38(4) :201–231, 2001.

[33] I. Suliciu. On the thermodynamics of fluids with relaxation and phase transitions. Fluids with relaxation. *Internat. J. Engrg. Sci*, 36 :921–947, 1998.

[34] Yulong Xing. Exactly well-balanced discontinuous Galerkin methods for the shallow water equations with moving water equilibrium. *Journal of Computational Physics*, 257 :536–553, 2014.

[35] Yulong Xing and Chi-Wang Shu. High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. *Journal of Computational Physics*, 214(2) :567–598, 2006.

[36] Yulong Xing, Chi-Wang Shu, and Sebastian Noelle. On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations. *Journal of scientific computing*, 48(1-3) :339–349, 2011.

[37] Yulong Xing, Xiangxiong Zhang, and Chi-Wang Shu. Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations. *Advances in Water Resources*, 33(12) :1476–1493, 2010.

[38] Hamed Zakerzadeh. On the Mach-uniformity of the Lagrange-projection scheme. *ESAIM : Mathematical Modelling and Numerical Analysis*, 51(4) :1343–1366, 2017.

Chapitre 2

Un schéma Lagrange-projection tout régime et équilibre pour les équations de Saint-Venant sur maillages non-structurés

Ce chapitre a été rédigé sous la forme d'un article en vue d'une soumission sous les références : Christophe Chalons, Samuel Kokh and Maxime Stauffert. An all-regime and well-balanced Lagrange-projection type scheme for the shallow water equations on unstructured meshes.

An all-regime and well-balanced Lagrange-projection type scheme for the shallow water equations on unstructured meshes

Abstract

In this work, we focus on the numerical approximation of the shallow water equations (SWE). Our aim is to propose a well-balanced, all-regime and positive scheme. By well-balanced, it is meant that the scheme is able to preserve the so-called lake at rest smooth equilibrium solutions. By all-regime, we mean that the scheme is able to deal with all flow regimes, including the low-Froude regime which is known to be challenging when using usual Godunov-type finite volume schemes. At last, the scheme should be positive which means that the water height stays positive for all time. Our approach is based on a Lagrange-projection decomposition which allows to naturally decouple the acoustic and transport terms. Numerical experiments on unstructured meshes illustrate the good behaviour of the scheme.

2.1 Introduction

We are interested in the numerical approximation of the shallow water equations (SWE)

$$\begin{cases} \partial_t h + \nabla \cdot (h\mathbf{u}) = 0, & (2.1.1a) \\ \partial_t (h\mathbf{u}) + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla \dfrac{gh^2}{2} = -gh\nabla z, & (2.1.1b) \end{cases}$$

where $\mathbf{x} \in \mathbb{R}^2 \mapsto z$ denotes a given smooth topography and $g > 0$ is the gravity constant. Both the water depth h and the velocity $\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2$ depend on the space and time variables, namely $\mathbf{x} \in \mathbb{R}^2$ and $t \in [0, \infty)$. We assume that the initial water depth $h(\mathbf{x}, t = 0) = h_0(\mathbf{x})$ and velocity $u(\mathbf{x}, t = 0) = u_0(\mathbf{x})$ are given.

Let us briefly recall that system (2.1.1) is strictly hyperbolic over the phase space $\Omega = \{(h, h\mathbf{u}^T) \in \mathbb{R}^3 \mid h > 0\}$. Moreover, if $\mathbf{n} \in \mathbb{R}^2$ is an arbitrary unit vector, the eigenstructure of system (2.1.1) is composed by two genuinely nonlinear characteristic fields associated with the eigenvalues $\{\mathbf{u}^T\mathbf{n} - c, \mathbf{u}^T\mathbf{n} + c\}$, where $c =: \sqrt{gh}$ is the sound speed, and a linearly degenerated

field associated with the eigenvalues $\mathbf{u}^T\mathbf{n}$. We recall also that the regions where $(\mathbf{u}^T\mathbf{n})^2 < c^2$ (resp. $(\mathbf{u}^T\mathbf{n})^2 > c^2$) are called subcritical or subsonic (resp. supercritical or supersonic).

We are interested in this work at developing a scheme that satisfy the well-balanced property. This is to say that we want our scheme to strictly preserve the "lake at rest" steady solutions, which are the states satisfying

$$h + z = \text{constant}, \qquad \mathbf{u} = \mathbf{0}.$$

For a review on this so-called C-property and the way to design schemes that satisfy it, one can look at the pioneering work [1] or the book [3] and more recently [13]. An article on a Lagrange-projection type scheme [9], like the one we propose in this work but in one dimension, has been published recently and studies the treatment of the source term in order to preserve the "lake at rest" solutions.

The methodology we use in this work is well suited for subsonic or near low-Froude number flows. We use an implicit-explicit strategy that allows to keep a stable scheme under a CFL time step limitation which is driven only by (slow) material waves and not by (fast) acoustic waves. The Lagrange-projection [12] scheme is designed following the pioneering work [10]. More recent works have been published over this schemes in the framework of Euler system for the computation of large friction or low-Mach regimes in [4], [5, 6, 7, 8] for single or two-phase flow models. Regarding the *all-regime* property of the scheme with respect of the Froude number, we follow more specifically the work [7]. Indeed, we transpose in our framework the correction over the pressure terms in the flux of the acoustic operator, proposed for example by [11].

The SWE has been largely studied and one can find nice overviews and references in the books [3] and [13]. The scheme proposed in [9] in one dimension has been studied in the framework of SWE and more specifically its behaviour for low-Froude number flows in [17]. A different implicit-explicit methodology in two dimension context has been proposed by [2].

In section 2, we study the dimensionless system associated to the SWE (2.1.1) and its asymptotic limit in low Froude regimes. In section 3, we present the Lagrange-projection like acoustic / transport decomposition associated to system (2.1.1). In section 4, we present the schemes, the finite volume scheme in 1D, the study of its truncation error in low Froude regimes and the proposed correction, and finally the extension towards 2D schemes on unstructured meshes. At last, we show some numerical results in 2D to verify the well-balanced property and illustrate the behaviour of the scheme in different regimes, especially in the low Froude one.

2.2 Low Froude limit for continuous equations

2.2.1 Dimensionless shallow water equations

In this section, we briefly introduce the dimensionless SWE. These equations will be useful to study the low-Froude asymptotic behaviour of the solutions of system (2.1.1). With this in mind, we define the following dimensionless quantities

$$\tilde{t} = \frac{t}{T}, \qquad \tilde{\mathbf{x}} = \frac{\mathbf{x}}{L}, \qquad \tilde{h} = \frac{h}{h_0}, \qquad \tilde{\mathbf{u}} = \frac{\mathbf{u}}{u_0}, \qquad \tilde{z} = \frac{z}{z_0},$$

where T, L, h_0, u_0 and z_0 are respectively reference time, length, water height, velocity and topography such that

$$u_0 = \frac{L}{T} \quad \text{and} \quad z_0 = h_0.$$

Defining the Froude number Fr by

$$\mathrm{Fr} = \frac{u_0}{c_0},$$

where $c_0 = \sqrt{gh_0}$ is the reference sound speed, easy calculations then give the dimensionless SWE

$$\begin{cases} \partial_{\tilde{t}}\tilde{h} + \nabla_{\tilde{\mathbf{x}}} \cdot \left(\tilde{h}\tilde{\mathbf{u}}\right) = 0, & (2.2.1a) \\ \partial_{\tilde{t}}\left(\tilde{h}\tilde{\mathbf{u}}\right) + \nabla_{\tilde{\mathbf{x}}} \cdot \left(\tilde{h}\tilde{\mathbf{u}} \otimes \tilde{\mathbf{u}}\right) + \dfrac{1}{\mathrm{Fr}^2}\nabla_{\tilde{\mathbf{x}}}\dfrac{\tilde{h}^2}{2} = -\dfrac{1}{\mathrm{Fr}^2}\tilde{h}\nabla_{\tilde{\mathbf{x}}}\tilde{z}. & (2.2.1b) \end{cases}$$

2.2.2 Asymptotic equations in low Froude limit

In this section, we give the asymptotic behaviour of the solutions of the SWE equations in the low Froude limit and several space dimensions. If we omit the tilde notation for the sake of readability in system (2.2.1) and, if we introduce the dimensionless pressure function $p(h) = \frac{h^2}{2}$, we get

$$\begin{cases} \partial_t h + \nabla \cdot (h\mathbf{u}) = 0, & (2.2.2a) \\ \partial_t (h\mathbf{u}) + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + \dfrac{1}{\mathrm{Fr}^2}\nabla p = -\dfrac{1}{\mathrm{Fr}^2}h\nabla z. & (2.2.2b) \end{cases}$$

Let us assume that h and z admit the following expansions in powers of the Froude number:

$$h = h^{(0)} + h^{(1)}\mathrm{Fr} + h^{(2)}\mathrm{Fr}^2 + \mathcal{O}(\mathrm{Fr}^3) \qquad \text{and} \qquad \mathbf{u} = \mathbf{u}^{(0)} + \mathbf{u}^{(1)}\mathrm{Fr} + \mathbf{u}^{(2)}\mathrm{Fr}^2 + \mathcal{O}(\mathrm{Fr}^3),$$

which gives in particular

$$p = p^{(0)} + p^{(1)}\mathrm{Fr} + p^{(2)}\mathrm{Fr}^2 + \mathcal{O}(\mathrm{Fr}^3) = p(h^{(0)}) + h^{(1)}p'(h^{(0)})\mathrm{Fr} + \mathcal{O}(\mathrm{Fr}^2).$$

The governing equations give at order -2 and -1 with respect to the Froude number that

$$\begin{aligned} &\nabla p^{(0)} + h^{(0)}\nabla z = 0 \Rightarrow \nabla h^{(0)} = -\nabla z \Rightarrow h^{(0)} + z = H(t), \\ &\nabla p^{(1)} + h^{(1)}\nabla z = 0 \Rightarrow h^{(0)}\nabla h^{(1)} = 0 \Rightarrow \nabla h^{(1)} = 0 \Rightarrow h^{(1)} = h^{(1)}(t). \end{aligned}$$

The asymptotic behavior is then given by

$$\begin{cases} \partial_t h^{(0)} + \nabla \cdot (h^{(0)}\mathbf{u}^{(0)}) = 0, \\ \partial_t(h^{(0)}\mathbf{u}^{(0)}) + \nabla \cdot (h^{(0)}\mathbf{u}^{(0)} \otimes \mathbf{u}^{(0)}) + \nabla p^{(2)} = -h^{(2)}\nabla z. \end{cases}$$

Now if we impose one of the following velocity boundary conditions

$$\left(\int_\Omega \nabla \cdot \mathbf{u}\,\mathrm{d}\Omega = 0 \text{ and } \int_\Omega \nabla \cdot (z\mathbf{u})\,\mathrm{d}\Omega = 0\right) \qquad \text{or} \qquad \left(\int_\Omega \nabla \cdot (h\mathbf{u})\,\mathrm{d}\Omega = 0\right),$$

integrating in space the first equation gives

$$\begin{aligned}0 = \int_\Omega \left(\partial_t h^{(0)} + \nabla\cdot(h^{(0)}\mathbf{u}^{(0)})\right)\,\mathrm{d}\Omega &= \int_\Omega \partial_t(H-z)\,\mathrm{d}\Omega + \int_\Omega \nabla\cdot((H-z)\,\mathbf{u}^{(0)})\,\mathrm{d}\Omega \\ &= \int_\Omega \partial_t H\,\mathrm{d}\Omega + H\int_\Omega \nabla\cdot\mathbf{u}^{(0)}\,\mathrm{d}\Omega - \int_\Omega \nabla\cdot(z\mathbf{u}^{(0)})\,\mathrm{d}\Omega \\ &= \int_\Omega H'(t)\,\mathrm{d}\Omega = |\Omega|\partial_t(h^{(0)}+z) = |\Omega|\partial_t h^{(0)}\end{aligned}$$

thus $\partial_t h^{(0)} = 0$ and $h^{(0)}+z = H$ is constant both in space and time. This leads to $\nabla\cdot(h^{(0)}\mathbf{u}^{(0)}) = 0$ and therefore

$$\nabla\cdot\mathbf{u}^{(0)} = \nabla\cdot(\frac{z}{H}\mathbf{u}^{(0)}),$$

while the evolution of $\mathbf{u}^{(0)}$ is given by

$$\left(1-\frac{z}{H}\right)\partial_t\mathbf{u}^{(0)} + \nabla\cdot(\mathbf{u}^{(0)}\otimes\mathbf{u}^{(0)}) + \frac{1}{H}\nabla p^{(2)} = \nabla\cdot(\frac{z}{H}\mathbf{u}^{(0)}\otimes\mathbf{u}^{(0)}) - h^{(2)}\nabla\frac{z}{H}.$$

Notice that when the topography is flat and set to 0, the three equations

$$\begin{cases} h^{(0)} + z = H = \text{cste} \\ \nabla\cdot\mathbf{u}^{(0)} = \nabla\cdot(\frac{z}{H}\mathbf{u}^{(0)}) \\ \left(1-\frac{z}{H}\right)\partial_t\mathbf{u}^{(0)} + \nabla\cdot(\mathbf{u}^{(0)}\otimes\mathbf{u}^{(0)}) + \frac{1}{H}\nabla p^{(2)} = \nabla\cdot(\frac{z}{H}\mathbf{u}^{(0)}\otimes\mathbf{u}^{(0)}) - h^{(2)}\nabla\frac{z}{H} \end{cases}$$

degenerate towards the well-known incompressible Euler equations

$$\begin{cases} h^{(0)} = \text{cste} \\ \nabla\cdot\mathbf{u}^{(0)} = 0 \\ \partial_t\mathbf{u}^{(0)} + \nabla\cdot(\mathbf{u}^{(0)}\otimes\mathbf{u}^{(0)}) + \frac{1}{h^{(0)}}\nabla p^{(2)} = 0. \end{cases}$$

2.3 An acoustic/transport operator decomposition

Let us first introduce notations related to our discretization context. We suppose that the computational domain $\Omega \subset \mathbb{R}^2$ is covered by N polygonal cells $(\Omega_j)_{1\le j\le N}$. We consider Γ, a face of the cell j, and we suppose the following admissibility assumptions are satisfied:

— either there exists a single $1 \le k \le N$ such that $\Gamma = \overline{\Omega_j}\cap\overline{\Omega_k} \neq \emptyset$. In this case we note $\Gamma = \Gamma_{jk}$ and Γ_{jk} can either be a vertex or a single face of the mesh,

— either $\Gamma \subset \partial\Omega$ and we suppose that there exists a single $k > N$ that will help to index ghost values for boundary conditions and we shall note $\Gamma = \Gamma_{jk}$.

For $1 \le j \le N$, we note $\mathcal{N}(j)$ the set of indices k such that Γ_{jk} is a face of Ω_j and if $k \in \mathcal{N}(j)$ we set $\mathbf{n}_{jk}$ to be the unit normal vector to Γ_{jk} pointing out of Ω_j.

We can now turn onto the acoustic / transport decomposition of the system (2.1.1). If we develop the spatial derivatives and isolate the transport terms $(\mathbf{u}\cdot\nabla)\varphi$, where $\varphi = h, h\mathbf{u}$, we can use an order 1 splitting operator with respect to time to obtain on one hand the acoustic step

$$\partial_t h + h\nabla\cdot(\mathbf{u}) = 0, \quad \partial_t(h\mathbf{u}) + h\mathbf{u}(\nabla\cdot\mathbf{u}) + \nabla p = -gh\nabla z, \tag{2.3.1}$$

and on the other hand the transport step

$$\partial_t h + (\mathbf{u}\cdot\nabla)h = 0, \quad \partial_t(h\mathbf{u}) + (\mathbf{u}\cdot\nabla)(h\mathbf{u}) = 0. \tag{2.3.2}$$

With these notations, the Lagrange-projection algorithm is defined as follows: for a given discrete state $(h, h\mathbf{u})_j^n$, $j \in \mathbb{Z}$, defining $(h, h\mathbf{u})_j^{n+1}$ is a two-step process defined as follows

1. Update $(h, h\mathbf{u})_j^n$ to $(h, h\mathbf{u})_j^{n+1-}$ by approximating the solution of system (2.3.1),
2. Update $(h, h\mathbf{u})_j^{n+1-}$ to $(h, h\mathbf{u})_j^{n+1}$ by approximating the solution of system (2.3.2).

Before entering the details of these two steps in the following section, let us notice that if we denote $\tau = 1/h$, by simple manipulations system (2.3.1) can be recast into:

$$\partial_t \tau - \tau(\mathbf{x},t)\nabla\cdot\mathbf{u} = 0, \quad \partial_t \mathbf{u} + \tau(\mathbf{x},t)\nabla p = -\tau(\mathbf{x},t)\frac{g}{\tau}\nabla z.$$

Following [10], we will choose to approximate the solution of system (2.3.1) thanks to a Suliciu-relaxation process. More precisely we will solve

$$\partial_t\tau - \tau(\mathbf{x},t)\nabla\cdot\mathbf{u} = 0, \quad \partial_t\mathbf{u} + \tau(\mathbf{x},t)\nabla\Pi = -\tau(\mathbf{x},t)\frac{g}{\tau}\nabla z, \quad \partial_t\Pi + \tau(\mathbf{x},t)a^2\nabla\cdot\mathbf{u} = \lambda(p^{\mathrm{EOS}}(\tau) - \Pi), \tag{2.3.3}$$

with $p = p^{\mathrm{EOS}}(\tau) = g/(2\tau^2)$, in the regime $\lambda \to +\infty$. The parameter a is a constant that is chosen in agreement with stability constraints. Over the the time interval $[t^n, t^n + \Delta t)$, we can account for the limit $\lambda \to +\infty$ by setting $\Pi(\mathbf{x}, t^n) = p^{\mathrm{EOS}}(\tau(\mathbf{x}, t^n))$, and then solving the relaxed system with $\lambda = 0$. We add another approximation by supposing that over $[t^n, t^n + \Delta t)$ it is reasonable to replace $\tau(\mathbf{x},t)\partial_{x_r}$ by $\tau(\mathbf{x},t^n)\partial_{x_r}$, $r = 1, 2$. Finally, we will define our approximation of the acoustic system (2.3.1) by solving

$$\partial_t\tau - \tau(\mathbf{x},t^n)\nabla\cdot\mathbf{u} = 0, \quad \partial_t\mathbf{u} + \tau(\mathbf{x},t^n)\nabla\Pi = -\tau(\mathbf{x},t^n)\frac{g}{\tau}\nabla z, \quad \partial_t\Pi + \tau(\mathbf{x},t^n)a^2\nabla\cdot\mathbf{u} = 0, \tag{2.3.4}$$

over $[t^n, t^n + \Delta t)$, with $\Pi(\mathbf{x}, t^n) = p^{\mathrm{EOS}}(\tau(\mathbf{x}, t^n))$.

Note that the system (2.3.4) is rotational invariant. This will allow us in the following at defining every flux for the two dimensional problem in the referential of each face. In this last referential, the problem will be reduced to a quasi-one dimensional problem (the transversal velocity v will be involved) that we study in the beginning of next section. One can also notice that the eigenstructure of the system (2.3.4) in the phase space $\left\{(h, h\mathbf{u}^T, \Pi, z) \in \mathbb{R}^5, h > 0, z > 0\right\}$ is very simple since it has three eigenvalues $\{-a, 0, a\}$ all associated to linearly degenerated characteristic fields.

2.4 Finite volume approximation

In this paragraph, we present the first-order finite volume scheme associated with the acoustic / transport decomposition of section 2.3.

2.4.1 A well-balanced Lagrange-projection finite volume scheme in 1D

We start by considering one-dimensional problems and briefly recall the method proposed in [9]. In this case the Saint-Venant equations read

$$\begin{cases} \partial_t h + \partial_x(hu_1) = 0, \\ \partial_t(hu_1) + \partial_x\left(hu_1^2 + g\dfrac{h^2}{2}\right) = -gh\partial_x z, \\ \partial_t(hu_2) + \partial_x(hu_1u_2) = 0. \end{cases}$$

The system associated with the acoustic step reads

$$\begin{cases} \partial_t\tau - \tau(x,t^n)\partial_x u_1 = 0, \\ \partial_t u_1 + \tau(x,t^n)\partial_x\Pi = -\tau(x,t^n)\dfrac{g}{\tau}\partial_x z, \\ \partial_t u_2 = 0, \\ \partial_t\Pi + \tau(x,t^n)a^2\partial_x u_1 = 0, \end{cases}$$

and the system that accounts for transport boils down to

$$\partial_t\varphi + u_1\partial_x\varphi = 0, \qquad \varphi \in \{h, u_1, u_2\}.$$

We suppose given a strictly increasing sequence $x_{j+1/2} \in \mathbb{R}$, for $j \in \mathbb{Z}$ and we consider the set of cells $\Omega_j = [x_{j-1/2}, x_{j+1/2})$. The local space step is defined by $\Delta x_j = x_{j+1/2} - x_{j-1/2}$. We note $\Delta t > 0$ the space step and set $t^n = n\Delta t$ for $n \in \mathbb{N}$.

The following discretization strategy was presented in [9]: the acoustic step (2.3.4) is approximated by

$$\begin{cases} \tau_j^{n+1-} = \tau_j^n - \tau_j^n\dfrac{\Delta t}{\Delta x_j}\left(u^\sharp_{j+1/2} - u^\sharp_{j-1/2}\right) & \text{(2.4.1a)} \\ (u_1)_j^{n+1-} = (u_1)_j^n - \tau_j^n\dfrac{\Delta t}{\Delta x_j}\left(\Pi^{L,\sharp}_{j+1/2} - \Pi^{R,\sharp}_{j-1/2}\right) & \text{(2.4.1b)} \\ (u_2)_j^{n+1-} = (u_2)_j^n & \text{(2.4.1c)} \\ \Pi_j^{n+1-} = \Pi_j^n - \tau_j^n\dfrac{\Delta t}{\Delta x_j}a^2\left(u^\sharp_{j+1/2} - u^\sharp_{j-1/2}\right) & \text{(2.4.1d)} \end{cases}$$

where the numerical fluxes $u^\sharp_{j+1/2}$, $\Pi^{L,\sharp}_{j+1/2}$ and $\Pi^{R,\sharp}_{j-1/2}$ are defined by

$$\begin{aligned} u^\sharp_{j+1/2} &= u_\Delta(\mathbf{U}^\sharp_j, \mathbf{U}^n_j, \mathbf{U}^\sharp_{j+1}, \mathbf{U}^n_{j+1}), \\ \Pi^{R,\sharp}_{j+1/2} &= \Pi^R_\Delta(\mathbf{U}^\sharp_j, \mathbf{U}^n_j, \mathbf{U}^\sharp_{j+1}, \mathbf{U}^n_{j+1}), \\ \Pi^{L,\sharp}_{j+1/2} &= \Pi^L_\Delta(\mathbf{U}^\sharp_j, \mathbf{U}^n_j, \mathbf{U}^\sharp_{j+1}, \mathbf{U}^n_{j+1}), \end{aligned}$$

where $\mathbf{U}$ is the state $\begin{pmatrix} h \\ h\mathbf{u} \\ \Pi \\ z \end{pmatrix}$ and with

$$\{gh\Delta z\}_\Delta(\mathbf{U}^n_L, \mathbf{U}^n_R) = g\frac{h^n_L + h^n_R}{2}(z_R - z_L)$$

$$\Pi_\Delta(\mathbf{U}_L^\sharp, \mathbf{U}_R^\sharp) = \frac{\Pi_L^\sharp + \Pi_R^\sharp}{2} - a\frac{(u_1)_R^\sharp - (u_1)_L^\sharp}{2}$$

$$u_\Delta(\mathbf{U}_L^\sharp, \mathbf{U}_L^n, \mathbf{U}_R^\sharp, \mathbf{U}_R^n) = \frac{(u_1)_L^\sharp + (u_1)_R^\sharp}{2} - \frac{\Pi_R^\sharp - \Pi_L^\sharp}{2a} - \frac{1}{2a}\{gh\Delta z\}_\Delta(\mathbf{U}_L^n, \mathbf{U}_R^n)$$

$$\Pi_\Delta^L(\mathbf{U}_L^\sharp, \mathbf{U}_L^n, \mathbf{U}_R^\sharp, \mathbf{U}_R^n) = \Pi_\Delta(\mathbf{U}_L^\sharp, \mathbf{U}_R^\sharp) + \frac{1}{2}\{gh\Delta z\}_\Delta(\mathbf{U}_L^n, \mathbf{U}_R^n)$$

$$\Pi_\Delta^R(\mathbf{U}_L^\sharp, \mathbf{U}_L^n, \mathbf{U}_R^\sharp, \mathbf{U}_R^n) = \Pi_\Delta(\mathbf{U}_L^\sharp, \mathbf{U}_R^\sharp) - \frac{1}{2}\{gh\Delta z\}_\Delta(\mathbf{U}_L^n, \mathbf{U}_R^n)$$

If one chooses $\sharp = n$ (resp. $\sharp = n+1^-$) the system (2.4.1) provides a time-explicit (resp. time-implicit) discretization of the acoustic system (2.3.4). The approximation of the transport step is performed thanks to an upwind scheme for $\varphi \in \{h, hu_1, hu_2\}$

$$\varphi_j^{n+1} = \varphi_j^n - \frac{\Delta t}{\Delta x_j}\left(u_{j+1/2}^\sharp \varphi_{j+1/2}^{n+1-} - u_{j-1/2}^\sharp \varphi_{j-1/2}^{n+1-}\right) - \frac{\Delta t}{\Delta x_j}\varphi_j^{n+1-}\left(u_{j+1/2}^\sharp - u_{j-1/2}^\sharp\right), \tag{2.4.2}$$

where

$$\varphi_{j+1/2}^{n+1-} = \begin{cases} \varphi_j^{n+1-}, & \text{if } u_{j+1/2}^\sharp \le 0, \\ \varphi_{j+1}^{n+1-}, & \text{if } u_{j+1/2}^\sharp > 0. \end{cases}$$

Note that $\{gh\Delta z\}_\Delta$ that accounts for the gravity source term is always evaluated at time t^n, even for the time implicit scheme.

In the above formulas, the parameter a is an approximation of the Lagrangian sound speed $hc = h\sqrt{gh}$ and must satisfy the sub-characteristic condition $a > hc$ which ensure that the relaxed system (2.3.3) is a viscous approximation of the acoustic step (2.3.1). In order to limit the numerical diffusion we take a local approximation of the Lagrangian sound speed at every interface, given by $a_{j+1/2} = \kappa \max(h_j\sqrt{gh_j}, h_{j+1}\sqrt{gh_{j+1}})$, where $\kappa > 1$.

For detailed properties of the numerical scheme (2.4.1)-(2.4.2) we refer the reader to [9], nevertheless let us recall that: the overall discretization is conservative in the usual sense of finite volumes methods with respect to (h, hu_1, hu_2). Moreover, the scheme is also well-balanced for lake at rest conditions: if $\mathbf{u}_j^n = 0$ and $h_j^n + z_j^n = h_{j+1}^n + z_{j+1}^n$ for all $j \in \mathbb{Z}$, then $h_j^{n+1} = h_j^n$ and $\mathbf{u}_j^{n+1} = \mathbf{u}_j^n$, $j \in \mathbb{Z}$. At last, the time-implicit scheme is stable under a condition which does not depend either on the acoustic system or the sound speed c, but which only depends on the transport step and its material velocity $\mathbf{u}$.

2.4.2 Truncation error in the low-Froude regime

In this paragraph, we consider the dimensionless shallow-water equation and we motivate a correction of the above scheme in order to make it efficient in low-Froude regimes. The correction is similar to the one in [7] for low-Mach regimes.

In the following we will say that the flow is in the low Froude regime if $\mathrm{Fr} \ll 1$ and $\partial_x p + h\partial_x z = \mathcal{O}(\mathrm{Fr}^2)$. Regarding the dimensionless equations (2.2.1), we can observe that, in this regime, the variations of the discharge hu remain of order 1.

We can express the fluxes given in the previous section, using the dimensionless quantities, which leads to

$$u_{j+1/2}^n = \frac{1}{2}(u_j^n + u_{j+1}^n) - \frac{1}{2a\mathrm{Fr}}\left(\Pi_{j+1}^n - \Pi_j^n + \frac{h_j^n + h_{j+1}^n}{2}(z_{j+1} - z_j)\right),$$

$$\Pi^{L,n}_{j+1/2} = \frac{\Pi^n_j}{\mathrm{Fr}^2} + \frac{1}{2\mathrm{Fr}^2}\left(\Pi^n_{j+1} - \Pi^n_j + \frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j)\right) - \frac{a}{2\mathrm{Fr}}(u^n_{j+1} - u^n_j),$$
$$\Pi^{R,n}_{j+1/2} = \frac{\Pi^n_{j+1}}{\mathrm{Fr}^2} - \frac{1}{2\mathrm{Fr}^2}\left(\Pi^n_{j+1} - \Pi^n_j + \frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j)\right) - \frac{a}{2\mathrm{Fr}}(u^n_{j+1} - u^n_j),$$

if one focuses on the time-explicit scheme for the sake of simplicity. Here again, we have omitted the tilde notations, and we have rescaled the sound velocity a, dividing it by $a_0 = h_0 c_0$.

If we compute the truncation errors in the fluxes above, using the fact that

$$\Pi^n_{j+1} - \Pi^n_j + \frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j) = \mathcal{O}(\mathrm{Fr}^2 \Delta x),$$

we obtain:

$$u^n_{j+1/2} = \frac{1}{2}(u^n_j + u^n_{j+1}) + \mathcal{O}(\mathrm{Fr}\Delta x),$$
$$\Pi^{L,n}_{j+1/2} = \frac{\Pi^n_j}{\mathrm{Fr}^2} + \frac{1}{2\mathrm{Fr}^2}\left(\Pi^n_{j+1} - \Pi^n_j + \frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j)\right) + \mathcal{O}(\frac{\Delta x}{\mathrm{Fr}}),$$
$$\Pi^{R,n}_{j+1/2} = \frac{\Pi^n_{j+1}}{\mathrm{Fr}^2} - \frac{1}{2\mathrm{Fr}^2}\left(\Pi^n_{j+1} - \Pi^n_j + \frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j)\right) + \mathcal{O}(\frac{\Delta x}{\mathrm{Fr}}).$$

In order to avoid large errors in the numerical diffusion terms when the Froude tends to zero, we propose to correct the fluxes $\Pi^{L,n}_{j+1/2}$ and $\Pi^{R,n}_{j+1/2}$ by

$$\Pi^{L,n,\theta}_{j+1/2} = \Pi^{n,\theta}_{j+1/2} + \frac{1}{2\mathrm{Fr}^2}\frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j), \quad \Pi^{R,n,\theta}_{j+1/2} = \Pi^{n,\theta}_{j+1/2} - \frac{1}{2\mathrm{Fr}^2}\frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j),$$

where

$$\Pi^{n,\theta}_{j+1/2} = \frac{1}{2\mathrm{Fr}^2}(\Pi^n_j + \Pi^n_{j+1}) - \theta_{j+1/2}\frac{a}{2\mathrm{Fr}}(u^n_{j+1} - u^n_j),$$

which amounts to reduce the numerical diffusion on the pressure gradient in the low Froude regime. Indeed, the truncation errors become on pressure fluxes:

$$\Pi^{L,n,\theta}_{j+1/2} = \frac{\Pi^n_j}{\mathrm{Fr}^2} + \frac{1}{2\mathrm{Fr}^2}\left(\Pi^n_{j+1} - \Pi^n_j + \frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j)\right) + \mathcal{O}(\frac{\theta_{j+1/2}\Delta x}{\mathrm{Fr}}),$$
$$\Pi^{R,n,\theta}_{j+1/2} = \frac{\Pi^n_{j+1}}{\mathrm{Fr}^2} - \frac{1}{2\mathrm{Fr}^2}\left(\Pi^n_{j+1} - \Pi^n_j + \frac{h^n_j + h^n_{j+1}}{2}(z_{j+1} - z_j)\right) + \mathcal{O}(\frac{\theta_{j+1/2}\Delta x}{\mathrm{Fr}}),$$

and as long as we take $\theta_{j+1/2} = \mathcal{O}(\mathrm{Fr})$, we get the uniform consistency of the global scheme with respect to the Froude number. In practice, we will set for both full-explicit and implicit-explicit schemes the explicit correction $\theta_{j+1/2} = \min\left(\frac{|u^n_{j+1/2}|}{\max(c_j, c_{j+1})}, 1\right)$.

2.4.3 The Lagrange-projection scheme on 2D unstructured meshes

We now extend the Lagrange-projection scheme to two dimensions framework following the work of [7]. Let $\mathbf{n} \in \mathbb{R}^2$ be a unit vector and $\mathbf{U}^T = (h, h\mathbf{u}^T)$, we define

$$R_{\mathbf{n}} = \begin{bmatrix} n_1 & n_2 \\ -n_2 & n_1 \end{bmatrix}, \qquad T_{\mathbf{n}}\mathbf{U} = \begin{bmatrix} h \\ h(R_{\mathbf{n}}\mathbf{u}) \\ \Pi \\ z \end{bmatrix}.$$

Following the lines of [12], we extend the one-dimensional scheme of [9] that was recalled in section 2.4.1 by taking advantage of the rotational invariance of the acoustic system (2.3.4). It gives

$$\begin{cases} \mathbf{u}_j^{n+1-} = \mathbf{u}_j^n - \tau_j^n \Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}\, \Pi_{jk}^{\sharp,\theta} \mathbf{n}_{jk}, & (2.4.3a) \\ \Pi_j^{n+1-} = \Pi_j^n - \tau_j^n \Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}\, (a_{jk})^2 u_{jk}^\sharp, & (2.4.3b) \\ \tau_j^{n+1-} = \tau_j^n + \tau_j^n \Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}\, u_{jk}^\sharp = L_j^\sharp \tau_j^n, & (2.4.3c) \end{cases}$$

where $\sigma_{jk} = \frac{|\Gamma_{jk}|}{|\Omega_j|}$, and the fluxes are

$$u_{jk}^\sharp = u_\Delta(T_{\mathbf{n}_{jk}}\mathbf{U}_j^\sharp, T_{\mathbf{n}_{jk}}\mathbf{U}_j^n, T_{\mathbf{n}_{jk}}\mathbf{U}_k^\sharp, T_{\mathbf{n}_{jk}}\mathbf{U}_k^n), \quad \Pi_{jk}^{\sharp,\theta} = \Pi_\Delta^{L,\theta}(T_{\mathbf{n}_{jk}}\mathbf{U}_j^\sharp, T_{\mathbf{n}_{jk}}\mathbf{U}_j^n, T_{\mathbf{n}_{jk}}\mathbf{U}_k^\sharp, T_{\mathbf{n}_{jk}}\mathbf{U}_k^n),$$

that is to say

$$\begin{aligned} u_{jk}^\sharp &= \frac{1}{2}\mathbf{n}_{jk}^T(\mathbf{u}_j^\sharp + \mathbf{u}_k^\sharp) - \frac{1}{2a_{jk}}(\Pi_k^\sharp - \Pi_j^\sharp) - \frac{1}{2a_{jk}}\left\{gh\Delta z\right\}_{jk}^n, \\ \Pi_{jk}^{\sharp,\theta} &= \frac{1}{2}(\Pi_j^\sharp + \Pi_k^\sharp) - \frac{a_{jk}\theta_{jk}}{2}\mathbf{n}_{jk}^T(\mathbf{u}_k^\sharp - \mathbf{u}_j^\sharp) + \frac{1}{2}\left\{gh\Delta z\right\}_{jk}^n, \end{aligned}$$

with

$$\begin{aligned} a_{jk} &\geq \max[(hc)_j^n, (hc)_k^n], \\ \{gh\Delta z\}_{jk}^n &= g\frac{h_j^n + h_k^n}{2}(z_k - z_j). \end{aligned}$$

The treatment of the source term is accounted for by the terms $\Pi_{jk}^{\sharp,\theta}$ since the fluxes $\Pi_{jk}^{\sharp,\theta}\mathbf{n}_{jk}$ in Equation (2.4.3a) are not symmetric, indeed $\Pi_{jk}^{\sharp,\theta}\mathbf{n}_{jk} \neq -\Pi_{kj}^{\sharp,\theta}\mathbf{n}_{kj}$. We remark that, as in one dimension, we first solve the linear system over $(\mathbf{u}, \Pi)$ resulting of the equations 2.4.3a and 2.4.3b for all j, before treating the Equation 2.4.3c over τ as an explicit upload.

As far as the transport step is concerned and in order to discretize the system (2.3.2), we use an explicit scheme between times t^{n+1-} and t^{n+1}, where the fluxes are chosen upwind with respect to $u_{jk}^\sharp$. If $\varphi \in \{h, h\mathbf{u}\}$, the scheme for the transport step reads

$$\varphi_j^{n+1} = \varphi_j^{n+1-} - \Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}\varphi_{jk}^{n+1-}u_{jk}^\sharp + \Delta t\varphi_j^{n+1-} \sum_{k\in\mathcal{N}(j)} \sigma_{jk}u_{jk}^\sharp, \tag{2.4.4}$$

where

$$\varphi_{jk}^{n+1-} = \begin{cases} \varphi_j^{n+1-}, & \text{if } u_{jk}^\sharp \geq 0, \\ \varphi_k^{n+1-}, & \text{if } u_{jk}^\sharp < 0. \end{cases}$$

Note that one can rewrite the transport step (2.4.4) as follows

$$\varphi_j^{n+1} = L_j^\sharp\varphi_j^{n+1-} - \Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}u_{jk}^\sharp\varphi_{jk}^{n+1-},$$

where $L_j^\sharp$ comes from the Equation (2.4.3c). Therefore, replacing the quantities φ_j^{n+1-} using system (2.4.3) gives the following update formulas which take into account the acoustic and transport steps together:

$$\begin{cases} h_j^{n+1} = h_j^n - \Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk} h_{jk}^{n+1-} u_{jk}^\sharp, & (2.4.5a) \\ (h\mathbf{u})_j^{n+1} = (h\mathbf{u})_j^n - \Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk} \left((h\mathbf{u})_{jk}^{n+1-} u_{jk}^\sharp + \Pi_{jk}^{\sharp,\theta} \mathbf{n}_{jk} \right) & (2.4.5b) \\ \qquad - \frac{1}{2}\Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk} \left\{ gh\Delta z \right\}_{jk}^n \mathbf{n}_{jk}, & (2.4.5c) \end{cases}$$

where the quantities $u_{jk}^\sharp$ and $\Pi_{jk}^{\sharp,\theta}$ are computed with $\mathbf{u}^\sharp$ and $\Pi^\sharp$, and the quantities h_{jk}^{n+1-} and $(h\mathbf{u})_{jk}^{n+1-}$ with τ^{n+1-} and $\mathbf{u}^{n+1-}$ from system (2.4.3).

Let us now state some stability and well-balanced properties of this scheme. We begin with the time-explicit scheme.

Proposition 2.4.1. *Let us assume that the time step Δt is such that the CFL conditions associated to the acoustic step*

$$\Delta t \max_{1\le j\le N} \left(\tau_j^n \max_{k\in\mathcal{N}(j)} \sigma_{jk} a_{jk} \right) \le \frac{1}{2},$$

and to the transport step

$$\Delta t \max_{1\le j\le N} \left(\sum_{k\in\mathcal{N}(j), u_{jk}^n<0} \sigma_{jk} |u_{jk}^n| \right) \le 1,$$

hold true. Therefore,

1. *the time-explicit scheme is conservative with respect to the water height h, and with respect to $h\mathbf{u}$ if the topography is flat,*
2. *the water height h_j^n is positive for all j and $n > 0$ provided that h_j^0 is positive for all j,*
3. *if $\theta = O(\mathrm{Fr})$, then the truncation error of the numerical scheme is uniform with respect to $\mathrm{Fr} \ll 1$.*

Proof. For the first property, one can sum over all the cells the global scheme (2.4.5) and prove the conservativity. The second property comes from the fact that L_j^n in Equation (2.4.3c) is positive under the transport CFL condition and that the transport step correspond to a convex combination of states at time t^{n+1-} under the transport CFL condition. Finally, the last property has been studied in one dimension framework in section 2.4.2. □

Let now turn to the time-implicit scheme.

Proposition 2.4.2. *Under the CFL condition associated to the transport step*

$$\Delta t \max_{1\le j\le N} \left(\sum_{k\in\mathcal{N}(j), u_{jk}^{n+1-}<0} \sigma_{jk} |u_{jk}^{n+1-}| \right) \le 1,$$

the implicit-explicit scheme on 2D unstructured mesh defined by system (2.4.3) *and* (2.4.4) *with $\sharp = n+1-$ satisfies the same properties as the full explicit one.*

Proof. All the clues given for the time-explicit scheme are again trues for the time-implicit one. □

Remark 2.4.1. *It is interesting to notice that the scheme degenerates towards the scheme proposed by [7] adapted to the framework of isentropic Euler system when the bottom is flat ($z = \text{cste}$),*

At last, we show that the schemes are well-balanced for the lake at rest.

Proposition 2.4.3. *Under a lake at rest initial condition ($\forall j$, $h_j^0 + z_j = H = cste$ and $\mathbf{u}_j^0 = \mathbf{0}$), the full explicit scheme on 2D unstructured mesh is well-balanced ($\forall j, n$, $h_j^n + z_j = cste$ and $\mathbf{u}_j^n = \mathbf{0}$).*

Proof. With this initial condition, we have the following fluxes computations:

$$\begin{aligned}
\{gh\Delta z\}_{jk}^0 &= g\frac{h_j^0 + h_k^0}{2}\left((H - h_k^0) - (H - h_j^0)\right) = \Pi_j^0 - \Pi_k^0,\\
u_{jk}^0 &= -\frac{1}{2a_{jk}}\left(\Pi_k^0 - \Pi_j^0\right) - \frac{1}{2a_{jk}}\{gh\Delta z\}_{jk}^0 = 0,\\
\Pi_{jk}^{0,\theta} &= \frac{1}{2}\left(\Pi_j^0 + \Pi_k^0\right) + \frac{1}{2}\{gh\Delta z\}_{jk}^0 = \Pi_j^0.
\end{aligned}$$

Injecting those fluxes in the acoustic step gives:

$$\begin{cases}
\mathbf{u}_j^{1-} = \mathbf{u}_j^0 - \tau_j^0\Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}\,\Pi_j^0\mathbf{n}_{jk} = -\tau_j^0\Pi_j^0\Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}\,\mathbf{n}_{jk} = \mathbf{0},\\
\Pi_j^{1-} = \Pi_j^0,\\
\tau_j^{1-} = \tau_j^0.
\end{cases}$$

Finally, since $u_{jk}^0 = 0$, $\forall j, k$, the transport step is trivial and the variables h and $h\mathbf{u}$ are unchanged at time t^1. Thus the lake is still at rest at the end of the first step. □

Proposition 2.4.4. *Under a lake at rest initial condition ($\forall j$, $h_j^0 + z_j = H = cste$ and $\mathbf{u}_j^0 = \mathbf{0}$), the implicit-explicit scheme on 2D unstructured mesh is well-balanced ($\forall j, n$, $h_j^n + z_j = cste$ and $\mathbf{u}_j^n = \mathbf{0}$).*

Proof. With the same calculus as in explicit case, we can verify that the vector $(\mathbf{u}_j^{1-}, \Pi_j^{1-}) = (\mathbf{0}, \Pi_j^0)$ is the only solution of the coupled system over $(\mathbf{u}, \Pi)$. Indeed we have

$$\begin{aligned}
\{gh\Delta z\}_{jk}^0 &= g\frac{h_j^0 + h_k^0}{2}\left((H - h_k^0) - (H - h_j^0)\right) = \Pi_j^0 - \Pi_k^0,\\
u_{jk}^{1-} &= -\frac{1}{2a_{jk}}\left(\Pi_k^{1-} - \Pi_j^{1-}\right) - \frac{1}{2a_{jk}}\{gh\Delta z\}_{jk}^0 = 0,\\
\Pi_{jk}^{1^-,\theta} &= \frac{1}{2}\left(\Pi_j^{1-} + \Pi_k^{1-}\right) + \frac{1}{2}\{gh\Delta z\}_{jk}^0 = \Pi_j^0.
\end{aligned}$$

and

$$\begin{cases}
\mathbf{u}_j^{1-} = \mathbf{u}_j^0 - \tau_j^0\Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}\,\Pi_{jk}^{1^-,\theta}\mathbf{n}_{jk} = -\tau_j^0\Pi_j^0\Delta t \sum_{k\in\mathcal{N}(j)} \sigma_{jk}\,\mathbf{n}_{jk} = \mathbf{0},\\
\Pi_j^{1-} = \Pi_j^0.
\end{cases}$$

Thus we have $\tau_j^{1-} = \tau_j^0$ and we can conclude as before for the transport step since $u_{jk}^{1-} = 0$, $\forall j, k$. □

2.5 Numerical experiments

We now present several test cases that aim at testing our scheme against classical flow configuration on unstructured meshes and also in the low Froude regime. In the following, the EXEX scheme refer to the full explicit scheme with time step:

$$\Delta t_{\text{EXEX}} = \frac{\text{coeffCFL}}{2\max_j \left(\frac{\sum_{k\in\mathcal{N}(j)} |\Gamma_{jk}|}{|\Omega_j|} \max_{k\in\mathcal{N}(j)} \max(v_{jk}^{\text{Acou}}, v_{jk}^{\text{Trans}}) \right)},$$

where $v_{jk}^{\text{Acou}} = \tau_j a_{jk}$, $a_{jk} = 1.01 \max(h_j c_j, h_k c_k)$, $c_j = \sqrt{g h_j}$ and $v_{jk}^{\text{Trans}} = |u_{jk}^n|$. The IMEX scheme will refer to the implicit-explicit scheme with time step:

$$\Delta t_{\text{IMEX}} = \frac{\text{coeffCFL}}{2\max_j \left(\frac{\sum_{k\in\mathcal{N}(j)} |\Gamma_{jk}|}{|\Omega_j|} \max_{k\in\mathcal{N}(j)} (v_{jk}^{\text{Trans}}) \right)}$$

where $v_{jk}^{\text{Trans}} = |u_{jk}^{n+1-}|$. Thus, the time step of the IMEX scheme is not constrained by the acoustic waves. Apart from the traveling vortex test case with flat bottom, in which we compare results with and without low Froude correction, we will always take θ as follows:

$$\theta_{jk} = \frac{\max\left(\|\mathbf{u}_j\|, \|\mathbf{u}_k\|\right)}{\max\left(c_j, c_k\right)},$$

so that θ approximate a local Froude number on every face.

2.5.1 Test of the well-balanced property

In order to test the well-balanced property of the scheme, we consider the following lake at rest initial condition:

$$\begin{aligned} h(x,y,0) &= H - z(x,y), \\ \mathbf{u}(x,y,0) &= \mathbf{0}, \end{aligned}$$

where $H = 0.5$ is constant and the topography z is a smooth bump defined by

$$z(x,y) = 0.3 \times \begin{cases} 0.5\exp(2 - \frac{0.1}{x-0.325}), & \text{if } 0.325 < x \leq 0.375, \\ 1 - 0.5\exp(2 - \frac{0.1}{0.425-x}), & \text{if } 0.375 < x < 0.425, \\ 1, & \text{if } 0.425 \leq x \leq 0.575, \\ 1 - 0.5\exp(2 - \frac{0.1}{x-0.575}), & \text{if } 0.575 < x < 0.625, \\ 0.5\exp(2 - \frac{0.1}{0.675-x}), & \text{if } 0.625 \leq x < 0.675, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5.1}$$

The physical domain $[0,1] \times [0,1]$ is discretized over a 20 000-cell triangular mesh. We impose absorbing boundary conditions and we observe the solution at final time $T_f = 0.1$.

In Figure 2.5.1 we have represented the water height h. For both EXEX and IMEX schemes, the errors between the numerical and the exact solution, which is also the initial stationary condition, are machine epsilon and we have not reported them here.

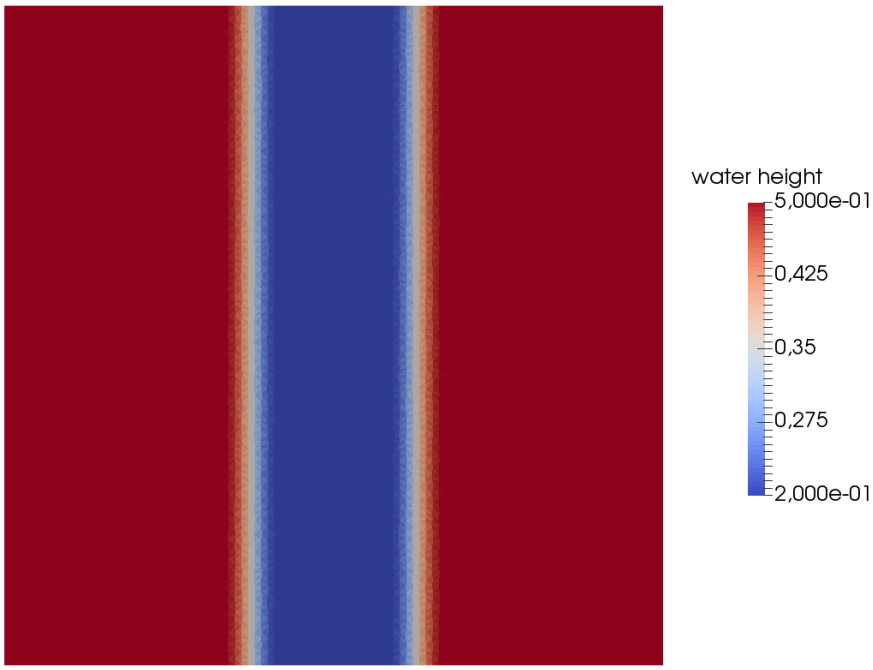

Figure 2.5.1 – Well-balanced property test: mapping of the water height h.

2.5.2 Planar dam break test problem

We are interested in the behaviour of our schemes with regard to the propagation of a rarefaction wave and a shock wave. We use the same mesh, boundary conditions and T_f value as in section 2.5.1. The topography is also kept identical to the one given in (2.5.1), the velocity initialized to zero and the initial total water height $H = h + z$ is defined as follows:

$$H(x, y, 0) = \begin{cases} 0.5 & \text{if } x \leq 0.5, \\ 1 & \text{otherwise.} \end{cases}$$

In Figure 2.5.2 we present the results for both the EXEX and IMEX schemes. We have performed a cut of the solution along the $y = 0.5$ axis and compared it with the one computed by a genuine 1D code with a 200-cell uniform grid. We can observe that for both EXEX and IMEX the results of the 2D simulations are agrees with the 1D results although they were computed with an unstructured mesh. It is worth noting that the 2D simulation manages to fairly preserve the planar structure of the approximate solutions.

2.5.3 Near steady-state flow

This test case, which has been taken from [16] and given in [14], involves the evolution of a perturbed 2D steady-state initial condition. We use two triangular meshes that discretize the physical domain $[0, 2] \times [0, 1]$, the coarse one has $40 \cdot 10^3$ cells and the refined one $360 \cdot 10^3$ cells. We impose periodic boundary conditions along the x-direction and absorbing ones along the y-direction, and we observe the solution at final time $T_f = 0.48$. Initially the water height is

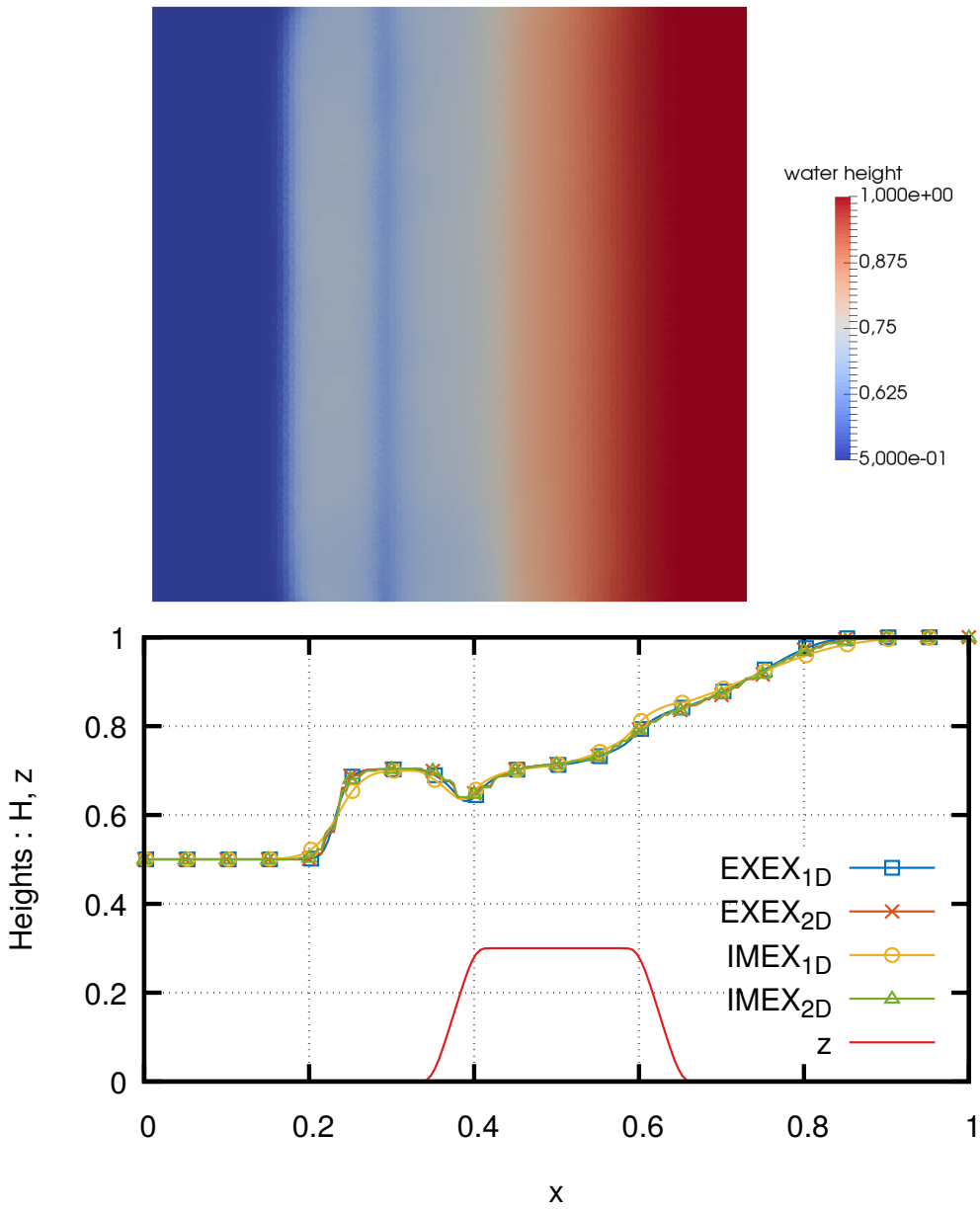

Figure 2.5.2 – Dam break test case at $T_f = 0.1$: mapping of the total water height with the IMEX scheme (top) and profile of H and z along the $y = 0.5$ axis obtained with both the EXEX and IMEX with 2D and 1D simulations (bottom).

given by $h(x, y, 0) = H(x, y, 0) - z(x, y)$ and the velocity by $\mathbf{u}(x, y, 0) = \mathbf{0}$, where

$$H(x, y, 0) = 1 + \begin{cases} 0.01 & \text{if } 0.05 \leq x \leq 0.15, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$z(x, y) = 0.8 \exp\left(-5(x - 0.9)^2 - 50(y - 0.5)^2\right).$$

Our schemes are more diffusive than the fifth order WENO scheme tested in [16], nevertheless we can observe in Figure 2.5.3 that the EXEX scheme manages to capture small features of the flow relative to the total water height H. With the IMEX scheme, these small features are destroyed by the numerical diffusion as we can see in Figure 2.5.4. The numbers of iterations and the CPU times needed to perform the three simulations are given in the Table 2.1.

	Nb time steps	CPU time
EXEX coarse mesh	3047	787
EXEX refined mesh	9170	14900
IMEX coarse mesh	306	1233

Table 2.1 – Near steady-state flow. Numbers of iterations and CPU times for every simulation.

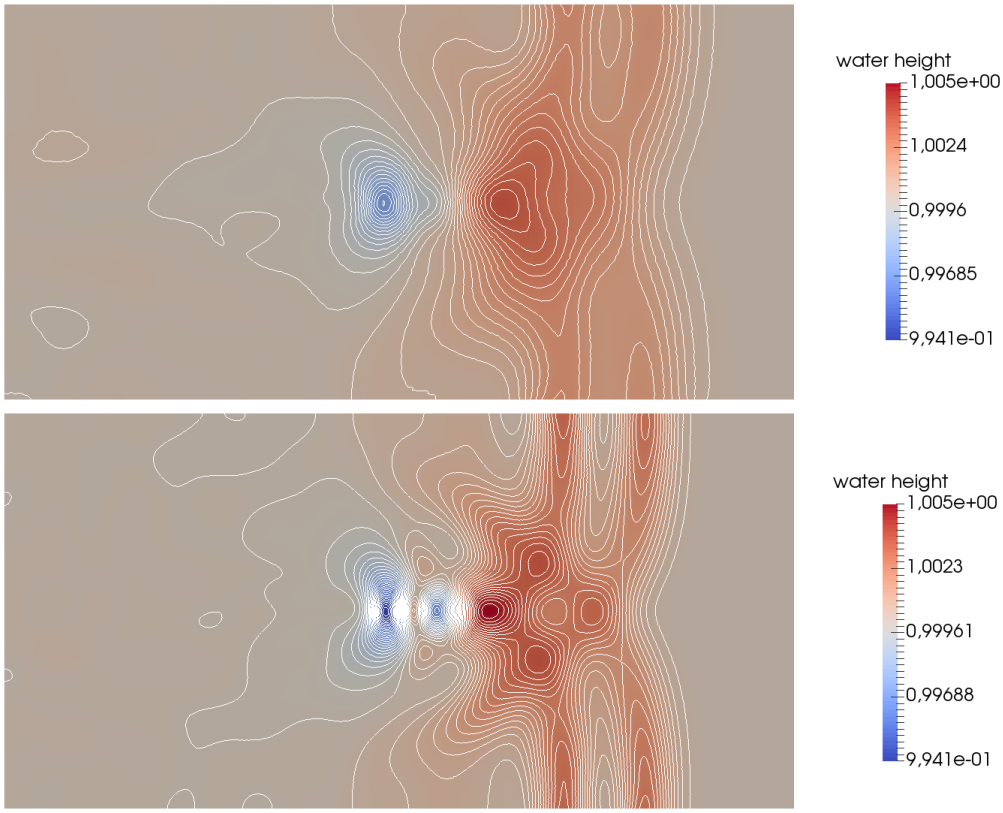

Figure 2.5.3 – Near steady-state flow. Mapping and isolines (from 0.9941 to 1.0051) of $H = h+z$ computed with the EXEX scheme at $T_f = 0.48$ obtained with a coarse mesh (top) and a refined one (bottom).

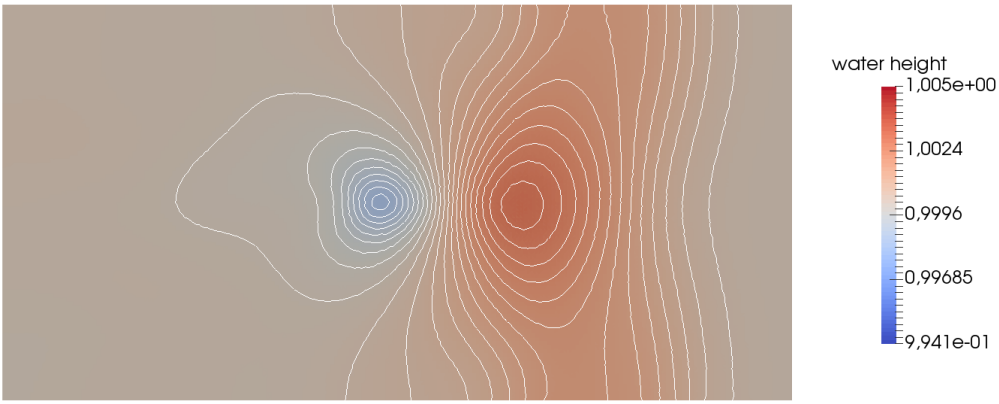

Figure 2.5.4 – Near steady-state flow. Mapping and isolines (from 0.9941 to 1.0051) of $H = h+z$ computed with the IMEX scheme at $T_f = 0.48$ with a coarse mesh.

2.5.4 Traveling vortex with flat bottom

In order to challenge our schemes with low Froude regimes, we consider a traveling vortex as in [2]. The exact solution of this test is detailed in [15]. For this test case we consider a flat bottom and we use a structured mesh of 160×160 cells that discretizes the physical domain $[0,1] \times [0,1]$. The boundary conditions imposed are periodic along the x-direction and absorbing boundaries along the y-direction. The initial conditions are given by:

$$h(x,y,0) = 110 + \begin{cases} \frac{\Gamma^2}{g\omega^2}\left(k(\omega r_c) - k(\pi)\right) & \text{if } \omega r_c \leq \pi, \\ 0 & \text{otherwise,} \end{cases}$$

$$u(x,y,0) = 0.6 + \begin{cases} \Gamma\left(1+\cos(\omega r_c)\right)(0.5-y) & \text{if } \omega r_c \leq \pi, \\ 0 & \text{otherwise,} \end{cases}$$

$$v(x,y,0) = 0 + \begin{cases} \Gamma\left(1+\cos(\omega r_c)\right)(x-0.5) & \text{if } \omega r_c \leq \pi, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$r_c = \|\mathbf{x} - (0.5, 0.5)\|, \quad \Gamma = 15.0, \quad \omega = 4\pi,$$

and

$$k(r) = 2\cos(r) + 2r\sin(r) + \frac{1}{8}\cos(2r) + \frac{r}{4}\sin(2r) + \frac{3}{4}r^2.$$

Due to the periodic boundary conditions, the exact solution is periodic of period $T = \frac{5}{3}$ and given at any time $t > 0$ by:

$$\begin{aligned} h(x,y,t) &= h(x - t/T, y, 0), \\ u(x,y,t) &= u(x - t/T, y, 0), \\ v(x,y,t) &= v(x - t/T, y, 0). \end{aligned}$$

We present the results of both the EXEX and IMEX schemes, with ($\theta = O(\mathrm{Fr})$) and without correction ($\theta = 1$) using $\epsilon = 0.05$. The mapping of the velocity magnitude is displayed in Figure 2.5.5 and we can observe that the accuracy of the solution is really improved by the low-Froude correction. Furthermore, the accuracy of the solution between the EXEX and the IMEX scheme with low-Froude correction is comparable whereas it took about 100 times less time steps and 6 times less CPU time computation to reach the final time with the IMEX than with the EXEX scheme as we can see in Table 2.2. Finally, the mapping of the Froude number is not given here, but is similar to the one of the velocity magnitude, with a range of values from $2.0 \cdot 10^{-3}$ to $1.1 \cdot 10^{-2}$.

	Nb time steps	CPU time
EXEX	60264	2109
IMEX	689	378

Table 2.2 – Traveling vortex test case with flat bottom. Numbers of iterations and CPU times for each scheme with low-Froude correction.

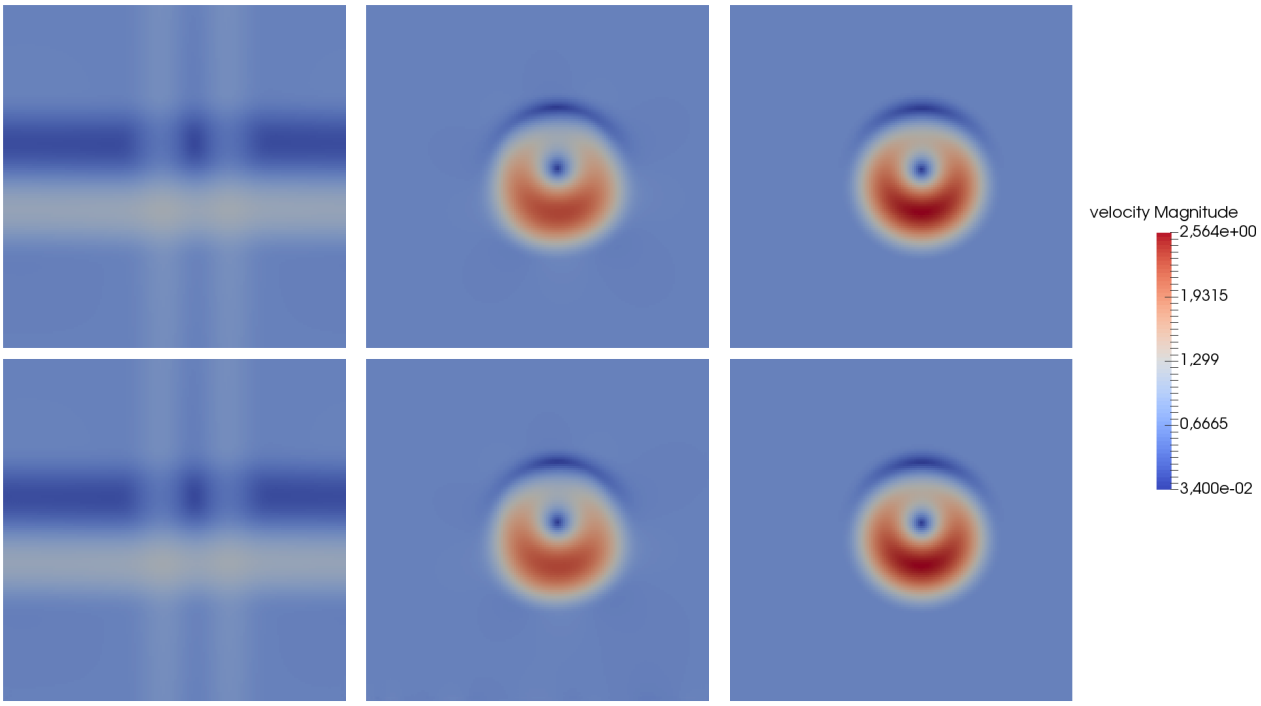

Figure 2.5.5 – Traveling vortex test case with flat bottom. Mapping of the velocity magnitude at $T_f = 0.1$ obtained with the EXEX scheme (top) and the IMEX scheme (bottom). We used the values $\theta = 1$ (left) and $\theta = O(\mathrm{Fr})$ (center). The right column displays the exact solution.

2.5.5 Traveling vortex with non-flat bottom

We extend the physical domain of the traveling vortex test above to the rectangle $[0, 2] \times [0, 1]$. The boundary conditions and initial conditions for h and u are the same as in section 2.5.4. However we consider here a topography defined by $z(x, y) = 10 \exp\left(-5(x-1)^2 - 50(y-0.5)^2\right)$ following the idea of [2].

We do not have exact analytical solution because of the non-flat bottom but we still can compare in figure 2.5.6 the results between EXEX and IMEX schemes, with or without low Froude correction. Here again, the vortex structure of the flow is completely destroyed by numerical diffusion without low-Froude corrections $\theta = O(\mathrm{Fr})$, with both schemes. The mapping of the Froude number is not given here, but is similar to the one of the velocity magnitude, with a range of values from $1.6 \cdot 10^{-3}$ to $1.1 \cdot 10^{-2}$. Finally, we can remark that the EXEX scheme took about 20 times more iterations and 10 times more CPU times than the IMEX scheme, both with low-Froude correction, as we can see in Table 2.3.

	Nb time steps	CPU time
EXEX	60264	15187
IMEX	2733	1607

Table 2.3 – Traveling vortex test case with flat bottom. Numbers of iterations and CPU times for each scheme with low-Froude correction.

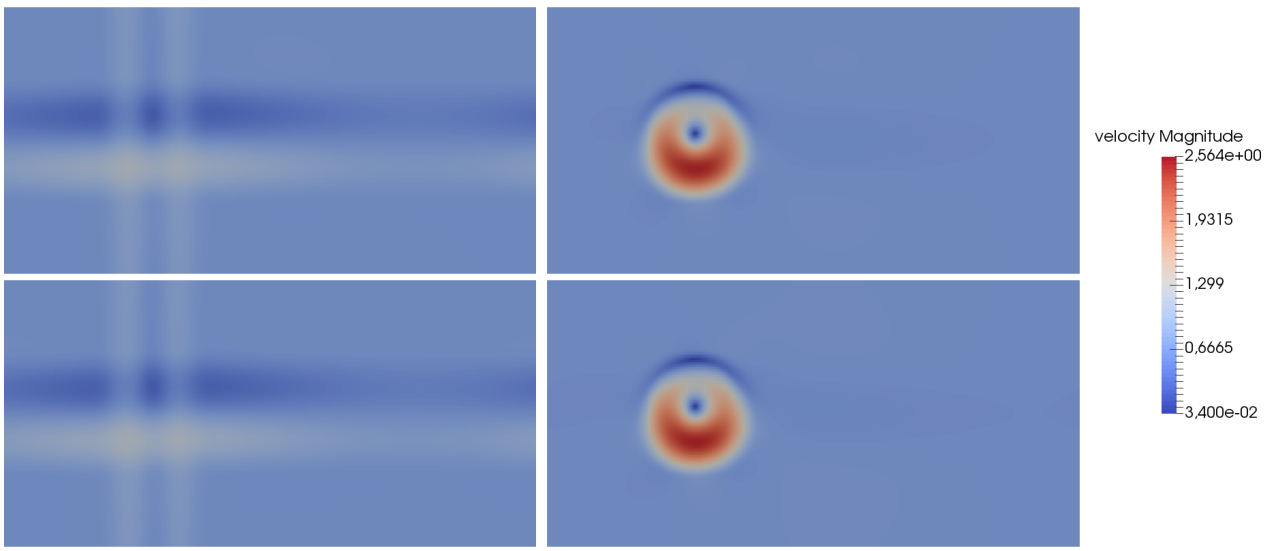

Figure 2.5.6 – Traveling vortex test case with non-flat bottom. Mapping of the velocity magnitude at instant $T_f = 0.1$ obtained with the EXEX scheme (top) and IMEX scheme (bot) with $\theta = 1$ (left) and $\theta = O(\mathrm{Fr})$ (right).

Conclusion

We proposed a large time step, well-balanced scheme for the shallow-water equations in two dimensions for unstructured meshes. We studied the truncation of the scheme with respect to the Froude number Fr and gave a correction in accordance to the source term. Finally, we obtained a conservative scheme with a uniform truncation error with respect to the Froude number for one-dimensional flows, and extended it to two-dimensional simulations.

Moreover, we showed that the time-implicit scheme has good numerical results for flows from low to high Froude values since its CFL condition is based on (slow) material waves only, and its acoustic system is linear so that the numerical resolution can be implemented efficiently.

Further developments shall include extensions to high-order methods in multiple-dimensions, following for example what has already been achieved with Finite Volume or discontinuous Galerkin methods in 1D. The method proposed in this paper can also be extended to more general compressible flows involving source terms.

Bibliographie

[1] Alfredo Bermudez and Ma Elena Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8) :1049–1071, 1994.

[2] Georgij Bispen, Koottungal Revi Arun, Mária Lukáčová-Medvid'ová, and Sebastian Noelle. Imex large time step finite volume methods for low Froude number shallow water flows. *Communications in Computational Physics*, 16(2) :307–347, 2014.

[3] François Bouchut. *Nonlinear stability of finite Volume Methods for hyperbolic conservation laws : And Well-Balanced schemes for sources.* Springer Science & Business Media, 2004.

[4] Christophe Chalons, Frédéric Coquel, Samuel Kokh, and Nicole Spillane. Large time-step numerical scheme for the seven-equation model of compressible two-phase flows. In *Finite Volumes for Complex Applications VI Problems & Perspectives*, pages 225–233. Springer, 2011.

[5] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. Large time step and asymptotic preserving numerical schemes for the gas dynamics equations with source terms. *SIAM Journal on Scientific Computing*, 35(6) :A2874–A2902, 2013.

[6] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. Operator-splitting based AP schemes for the 1D and 2D gas dynamics equations with stiff sources. *AIMS Series on Applied Mathematics*, 8 :607–614, 2014.

[7] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An all-regime Lagrange-projection like scheme for the gas dynamics equations on unstructured meshes. *Communications in Computational Physics*, 20(1) :188–233, 2016.

[8] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An all-regime Lagrange-projection like scheme for 2D homogeneous models for two-phase flows on unstructured meshes. *Journal of Computational Physics*, 335 :885–904, 2017.

[9] Christophe Chalons, Pierre Kestener, Samuel Kokh, and Maxime Stauffert. A large time-step and well-balanced Lagrange-projection type scheme for the shallow water equations. *Communication in Mathematical Sciences*, 15(3) :765–788, 2017.

[10] Frédéric Coquel, Quang Nguyen, Marie Postel, and Quang Tran. Entropy-satisfying relaxation method with large time-steps for Euler IBVPs. *Mathematics of Computation*, 79(271) :1493–1533, 2010.

[11] Stéphane Dellacherie. Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. *Journal of Computational Physics*, 229(4) :978–1016, 2010.

[12] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, volume 118. Springer Science & Business Media, 1996.

[13] Laurent Gosse. *Computing qualitatively correct approximations of balance laws*, volume 2. Springer, 2013.

[14] Randall J LeVeque. Balancing source terms and flux gradients in high-resolution Godunov methods : the quasi-steady wave-propagation algorithm. *Journal of Computational Physics*, 146(1) :346–365, 1998.

[15] Mario Ricchiuto and Andreas Bollermann. Stabilized residual distribution for shallow water simulations. *Journal of Computational Physics*, 228(4) :1071–1115, 2009.

[16] Yulong Xing and Chi-Wang Shu. High order finite difference WENO schemes with the exact conservation property for the shallow water equations. *Journal of Computational Physics*, 208(1) :206–227, 2005.

[17] Hamed Zakerzadeh. On the Mach-uniformity of the Lagrange-projection scheme. *ESAIM : Mathematical Modelling and Numerical Analysis*, 51(4) :1343–1366, 2017.

Deuxième partie

Extension à des schémas d'ordres élevés Galerkin discontinus

Chapitre 3

Un schéma Lagrange-projection Galerkin discontinu équilibre pour les équation de Saint-Venant

Le corps de ce chapitre est disponible sous la forme d'un *preprint* dont les références sont : Christophe Chalons and Maxime Stauffert. A well-balanced discontinuous-Galerkin Lagrange-projection scheme for the shallow water equations. 2017. <hal-01612292>

L'annexe 3.B a fait l'objet d'une publication comme acte : Christophe Chalons and Maxime Stauffert. A high-order discontinuous Galerkin Lagrange-projection scheme for the barotropic Euler equations. In *International Conference on Finite Volumes for Complex Applications* (pp. 63-70). Springer, 2017.

A well-balanced Discontinuous-Galerkin Lagrange-projection scheme for the shallow water equations

Abstract

This work considers the shallow water equations (SWE) and proposes a high order conservative scheme based on a Lagrange-projection decomposition. The high order in space and time are achieved using discontinuous Galerkin (DG) and Runge-Kutta (RK) strategies. The use of a Lagrange-projection decomposition enables the use of time steps that are not constrained by the sound speed thanks to an implicit treatment of the acoustic waves (Lagrange step), while the transport waves (projection step) are treated explicitly. We prove that our scheme satisfies the well-balanced property as well as non linear stability properties. Numerical evidences are also given.

3.1 Introduction

We are interested in the shallow water equations

$$\begin{cases} \partial_t h + \partial_x (hu) = 0, \\ \partial_t (hu) + \partial_x \left(hu^2 + p \right) = -gh\partial_x z, \end{cases} \tag{3.1.1}$$

where $h > 0$ is the water height, u the velocity, z the topography height and $p = gh^2/2$ is the pressure term where $g > 0$ is the gravity constant. The unknowns depend on the space and time variables x and t, with $x \in \mathbb{R}$ and $t \in [0, \infty)$. At time $t = 0$, the model is supplemented with a given initial data $h(x, t = 0) = h_0(x)$ and $u(x, t = 0) = u_0(x)$. The entropy inequality associated with the system (3.1.1) can be written either in a non conservative form as follows:

$$\partial_t hE\left(w\right) + \partial_x \mathcal{F}\left(w\right) \leqslant -ghu\partial_x z, \tag{3.1.2}$$

where $w = (h, hu)^T$ and, setting $e(h) = gh/2$,

$$hE\left(w\right) = h\frac{u^2}{2} + he(h), \quad \mathcal{F}\left(w\right) = \left(\frac{u^2}{2} + gh\right) hu, \tag{3.1.3}$$

or in conservative form as follows:

$$\partial_t h\tilde{E}\left(w\,,z\right)+\partial_x\tilde{\mathcal{F}}\left(w\,,z\right)\leqslant 0, \tag{3.1.4}$$

with the *conservative entropy* $\tilde{E}$ and the associated flux $\tilde{\mathcal{F}}$ defined by,

$$h\tilde{E}\left(w\,,z\right)=hE\left(w\right)+ghz \quad \text{and} \quad \tilde{\mathcal{F}}\left(w\,,z\right)=\mathcal{F}\left(w\right)+ghuz. \tag{3.1.5}$$

Note that the proposed numerical scheme will satisfy a discrete form of the non conservative entropy inequality (3.1.2).

The aim of this paper is to propose a high order discretization of system (3.1.1) based on a Lagrange-projection decomposition and using discontinuous Galerkin (DG) [9, 16] and Runge-Kutta (RK) [14] strategies for the space and time variables respectively. We will also pay a particular attention to the well-known well-balanced property, see for instance [2, 11], the references therein and the large litterature on this topic.

The proposed strategy can be understood as a natural extension to the present setting of the first-order well-balanced Lagrange-projection scheme developed in [7] for the shallow water equations, and of the high order Lagrange-projection scheme introduced in [8] (see also [13] for a similar approach) for the barotropic gas dynamics equations. The Lagrange-projection (or equivalently Lagrange-remap) decomposition naturally decouples the acoustic and transport terms of system (3.1.1). It proved to be useful and very efficient when considering subsonic or low-Mach number flows. In this case, the CFL restriction of Godunov-type schemes is driven by the acoustic waves and can be very restrictive. As we will see, the Lagrange-projection strategy allows for a very natural implicit-explicit scheme with a CFL restriction based on the (slow) transport waves and not on the (fast) acoustic waves, see the pionneering paper [10]. Note that the low-Mach (or low-Froude in the present setting of shallow water equations) limit using the same techniques as in [4, 5, 6] will not be considered in the present paper but is the topic of current research. Here, we focus on the design of a high order well-balanced implicit explicit scheme in a Lagrange-projection framework.

The outline of the paper is as follows. We first briefly recall the Lagrange-projection decomposition and the underlying first-order finite volume scheme in the next section. We then formulate this scheme using a discontinous Galerkin framework in Section 3.3. The stability and well-balanced properties are collected in Section 3.4. At last, Section 3.5 illustrates the behaviours of our schemes on several test cases.

3.2 Lagrange-projection decomposition and first-order relaxation scheme

In this section, we briefly present the Lagrange-projection decomposition considered in this paper as well as the corresponding first-order finite volume scheme based on a relaxation approach. Note from now on that the latter will provide a natural linearization of the pressure term p, which will be helpful in order to define an efficient implicit discretization of the Lagrangian step.

3.2.1 Operator splitting decomposition and relaxation approximation

Using the chain rule for the space derivatives of system (3.1.1), the Lagrange-projection decomposition consists in first solving

$$\begin{cases} \partial_t h + h\partial_x u = 0, \\ \partial_t (hu) + hu\partial_x u + \partial_x p = -gh\partial_x z, \end{cases} \tag{3.2.1}$$

which gives in Lagrangian coordinates $\tau\partial_x = \partial_m$, with $\tau = 1/h$,

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m p = -\dfrac{g}{\tau}\partial_m z, \end{cases} \tag{3.2.2}$$

and then the transport system

$$\begin{cases} \partial_t h + u\partial_x h = 0, \\ \partial_t (hu) + u\partial_x (hu) = 0. \end{cases} \tag{3.2.3}$$

It is interesting to note that the strictly hyperbolic system (3.2.2) with eigenvalues $\pm hc$ where the sound speed c equals $\sqrt{p'(h)}$ accounts for the acoustic waves of system (3.1.1) (or equivalently (3.2.1)). On the other hand, the hyperbolic system (3.2.3) with eigenvalues u accounts for the transport waves of system (3.1.1).

In the following, the Lagrangian system (3.2.2) will be treated considering the following relaxation approximation [12, 15, 2]

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m \Pi = -\dfrac{g}{\tau}\partial_m z, \\ \partial_t \Pi + a^2 \partial_m u = \lambda \left(p - \Pi\right). \end{cases} \tag{3.2.4}$$

Here, the new variable Π represents a linearization of the real pressure p, the constant parameter a is a linearization of the Lagrangian sound speed hc such that the sub-characteristic condition $a > hc$ is satisfied, and the relaxation parameter λ allows to recover $\Pi = p$ and the original system (3.2.2) in the asymptotic regime $\lambda \to \infty$. As usual, the relaxation system will be solved using a splitting strategy which consists in first setting $\Pi = p$ at initial time (which is formally equivalent to considering $\lambda \to \infty$), and then solving the relaxation system (3.2.4) with $\lambda = 0$.

3.2.2 First-order well-balanced Lagrange-projection scheme

In this paragraph, we briefly recall the first-order finite volume scheme given in [7] and associated with the above Lagrange-projection decomposition and relaxation approximation. Space and time will be discretized using constant space step Δx and time step Δt. We will consider a set of cells $\kappa_j = [x_{j-1/2}, x_{j+1/2})$ and instants $t^n = n\Delta t$, where $x_{j+1/2} = j\Delta x$ and $x_j = (x_{j-1/2} + x_{j+1/2})/2$ are respectively the cell interfaces and cell centers, for $j \in \mathbb{Z}$ and $n \in \mathbb{N}$.

Following [7] and using standard notations, the Lagrangian step is discretized by

$$\begin{cases} \tau_j^{n+1-} = L_j^n \tau_j^n, \\ L_j^\alpha (hu)_j^{n+1-} = (hu)_j^n - \dfrac{\Delta t}{\Delta x}(\Pi_{j+1/2}^{*,\alpha} - \Pi_{j-1/2}^{*,\alpha}) - \Delta t \left\{ gh\partial_x z \right\}_j^n, \\ L_j^\alpha (h\Pi)_j^{n+1-} = (h\Pi)_j^n + a^2 \dfrac{\Delta t}{\Delta x}(u_{j+1/2}^{*,\alpha} - u_{j-1/2}^{*,\alpha}), \end{cases}$$

with

$$L_j^\alpha = 1 + \frac{\Delta t}{\Delta x}(u_{j+1/2}^{*,\alpha} - u_{j-1/2}^{*,\alpha}),$$

$$u_{j+1/2}^{*,\alpha} = \frac{1}{2}(u_j^\alpha + u_{j+1}^\alpha) - \frac{1}{2a}(\Pi_{j+1}^\alpha - \Pi_j^\alpha) - \frac{\Delta x}{2a}\{gh\partial_x z\}_{j+1/2},$$

$$\Pi_{j+1/2}^{*,\alpha} = \frac{1}{2}(\Pi_j^\alpha + \Pi_{j+1}^\alpha) - \frac{a}{2}(u_{j+1}^\alpha - u_j^\alpha),$$

and

$$\{gh\partial_x z\}_j^n = \frac{1}{2}(\{gh\partial_x z\}_{j-1/2}^n + \{gh\partial_x z\}_{j+1/2}^n),$$

$$\{gh\partial_x z\}_{j+1/2}^n = g\frac{h_j^n + h_{j+1}^n}{2}\frac{z_{j+1} - z_j}{\Delta x}.$$

In the above formulas, $\Pi_j^n = p_j^n = p(h_j^n)$ for all j and α refers to the time index and equals n (respectively $n+1-$) if the scheme is taken to be explicit (resp. implicit) in time. Note that the source term is always taken at time t^n, even for the time implicit scheme. In the following, we will be especially interested in the choice $\alpha = n+1-$ in order to get rid of the usual acoustic CFL restriction.

As far as the transport step is concerned, a natural discretization of system (3.2.3) reads

$$\begin{cases} h_j^{n+1} = h_j^{n+1-} - \dfrac{\Delta t}{\Delta x}\left[\left(u_{j+1/2}^{*,\alpha}\right)_- \left(h_{j+1}^{n+1-} - h_j^{n+1-}\right) + \left(u_{j-1/2}^{*,\alpha}\right)_+ \left(h_j^{n+1-} - h_{j-1}^{n+1-}\right)\right], \\ (hu)_j^{n+1} = (hu)_j^{n+1-} - \dfrac{\Delta t}{\Delta x}\Bigg[\left(u_{j+1/2}^{*,\alpha}\right)_- \left((hu)_{j+1}^{n+1-} - (hu)_j^{n+1-}\right) \\ \qquad\qquad + \left(u_{j-1/2}^{*,\alpha}\right)_+ \left((hu)_j^{n+1-} - (hu)_{j-1}^{n+1-}\right)\Bigg], \end{cases} \tag{3.2.5}$$

where $\left(u_{j\pm1/2}^{*,\alpha}\right)_+ = \max(u_{j\pm1/2}^{*,\alpha}, 0)$ and $\left(u_{j\pm1/2}^{*,\alpha}\right)_- = \min(u_{j\pm1/2}^{*,\alpha}, 0)$ for all j.

At last, we underline that easy calculations give that the whole scheme is conservative in the usual sense of finite volume methods and writes

$$\begin{cases} h_j^{n+1} = h_j^n - \dfrac{\Delta t}{\Delta x}\left(h_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} - h_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha}\right), \\ (hu)_j^{n+1} = (hu)_j^n - \dfrac{\Delta t}{\Delta x}\left((hu)_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} + \Pi_{j+1/2}^{*,\alpha} - (hu)_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha} - \Pi_{j-1/2}^{*,\alpha}\right) \\ \qquad\qquad - \Delta t\{gh\partial_x z\}_j^n, \end{cases} \tag{3.2.6}$$

with, for $X = h, hu$,

$$X_{j+1/2}^{*,n+1-} = \begin{cases} X_j^{n+1-}, & \text{if } u_{j+1/2}^{*,\alpha} \geq 0, \\ X_{j+1}^{n+1-}, & \text{if } u_{j+1/2}^{*,\alpha} \leq 0. \end{cases}$$

We refer the reader to [7] for more details and the non linear stability properties satisfied by this first order finite volume numerical scheme. Let us just underline that it satisfies the well-balanced property for the lake at rest. More precisely, if the discrete fluid state at time t^n matches the lake at rest conditions $u_j^n = 0$ and $h_j^n + z_j^n = h_{j+1}^n + z_{j+1}^n$ for all j, then $h_j^{n+1} = h_j^n$ and $u_j^{n+1} = u_j^n$.

The aim of the next section is to develop a high order and well-balanced extension of this scheme using discontinuous Galerkin and Runge-Kutta techniques for the space and time variables respectively.

3.3 Discontinuous Galerkin discretization

We begin this section by introducing the notations of the DG discretization. Recall that the DG approach considers that the approximate solution at each time t^n is defined on each cell κ_j by a polynomial in space of order less or equal than p for a given integer $p \geq 1$ ($p = 0$ corresponds to the usual first-order and piecewise constant finite volume scheme). With this in mind, we consider the $(p+1)$ Lagrange polynomials $\{\ell_i\}_{i=0,\dots,p}$ associated with the Gauss-Lobatto quadrature points in $[-1, 1]$. More precisely, denoting $-1 = s_0 < s_1 < \cdots < s_p = 1$ the $p+1$ Gauss-Lobatto quadrature points, ℓ_i is defined by the relations $\ell_i(s_k) = \delta_{i,k}$ for $k = 0, ..., p$, where δ is the Kronecker symbol. Then, in each cell κ_j, we define the shifted Lagrange polynomials $\Phi_{i,j}$ by $\Phi_{i,j}(x) = \ell_i\Big(\frac{2}{\Delta x}(x - x_j)\Big)$ and we take $\{\Phi_{i,j}\}_{i=0,\dots,p}$ as a basis for polynomials or order less or equal than p on κ_j. Note that for all j

$$\sum_{i=0}^{p} \Phi_{i,j}(x) = 1. \tag{3.3.1}$$

If we denote by $X_{\Delta x}$ the DG approximation of X, we thus have $X_{\Delta x}(x,t) = \sum_{k=0}^{p} X_{k,j}(t)\, \Phi_{k,j}(x)$ for all $x \in \kappa_j$, where the coefficients $X_{k,j}$ depend on the time t and correspond to the value of X at the shifted Gauss-Lobatto quadrature points $x_{k,j} = x_j + \frac{\Delta x}{2} s_k$.

Before entering the details of the numerical approximation, let us briefly recall that the Gauss-Lobatto quadrature formula for evaluating the space integral of a given function $f : \kappa_j \times \mathbb{R}^+ \to \mathbb{R}$ writes

$$\int_{\kappa_j} f(x,t)\,\mathrm{d}x \approx \frac{\Delta x}{2} \sum_{k=0}^{p} \omega_k f(x_{k,j}, t), \tag{3.3.2}$$

where ω_k are the weights of the Gauss-Lobatto quadrature. It is well-known that this formula is exact as soon as f is a polynomial of order less or equal than $(2p-1)$ with respect to x on κ_j. In particular, we have

$$\sum_{k=0}^{p} \frac{\omega_k}{2} = 1. \tag{3.3.3}$$

Note also that the integral $\int_{\kappa_j} \Phi_{i,j}(x)\Phi_{k,j}(x)\,\mathrm{d}x$ will be therefore approximated by $\frac{\Delta x}{2}\omega_i \delta_{i,k}$ in the following. At last, note that the piecewise constant case $p = 0$ can be also considered in this framework provided that we set $s_0 = 0$, $\Phi_{0,j} = 1$ and $\omega_0 = 2$.

3.3.1 The acoustic step

We begin with the acoustic step (3.2.4) with $\lambda = 0$.

Time discretization ($t^n \to t^{n+1}$)

Multiplying the three equations by $\Phi_{i,j}$, integrating over κ_j, and considering the piecewise polynomial approximations $X_{\Delta x}$ for $X = \tau, u, \Pi$ easily leads to

$$\begin{cases} \dfrac{\Delta x}{2}\omega_i \partial_t \tau_{i,j}(t) - \displaystyle\int_{\kappa_j} \Phi_{i,j}(x)\partial_m u(x,t)\,\mathrm{d}x = 0, \\ \dfrac{\Delta x}{2}\omega_i \partial_t u_{i,j}(t) + \displaystyle\int_{\kappa_j} \Phi_{i,j}(x)\partial_m \Pi(x,t)\,\mathrm{d}x = -\int_{\kappa_j} \Phi_{i,j}(x)\frac{g}{\tau(x,t)}\partial_m z(x)\,\mathrm{d}x, \\ \dfrac{\Delta x}{2}\omega_i \partial_t \Pi_{i,j}(t) + a^2 \displaystyle\int_{\kappa_j} \Phi_{i,j}(x)\partial_m u(x,t)\,\mathrm{d}x = 0, \end{cases}$$

that we discretize in time by

$$\begin{cases} \tau_{i,j}^{n+1-} = \tau_{i,j}^n + \dfrac{2\Delta t}{\omega_i \Delta x}\displaystyle\int_{\kappa_j} \Phi_{i,j}(x)\partial_m u(x,t^\alpha)\,\mathrm{d}x, \\ u_{i,j}^{n+1-} = u_{i,j}^n - \dfrac{2\Delta t}{\omega_i \Delta x}\left(\displaystyle\int_{\kappa_j} \Phi_{i,j}(x)\partial_m \Pi(x,t^\alpha)\,\mathrm{d}x - \int_{\kappa_j} \Phi_{i,j}(x)\frac{g}{\tau(x,t^n)}\partial_m z(x)\,\mathrm{d}x\right), \\ \Pi_{i,j}^{n+1-} = \Pi_{i,j}^n - a^2\dfrac{2\Delta t}{\omega_i \Delta x}\displaystyle\int_{\kappa_j} \Phi_{i,j}(x)\partial_m u(x,t^\alpha)\,\mathrm{d}x, \end{cases} \tag{3.3.4}$$

where $\alpha = n$ or $\alpha = n+1-$ depending on whether the time discretization is taken to be explicit or implicit. Again, we are especially interested in this work in the case $\alpha = n+1-$.

Volume integrals and flux calculations

We first aim at approximating the integrals

$$\int_{\kappa_j} \Phi_{i,j}(x)\partial_m X(x,t^\alpha)\,\mathrm{d}x,$$

with $X = u, \Pi$. Observe that

$$\int_{\kappa_j} \Phi_{i,j}(x)\partial_m X(x,t^\alpha)\,\mathrm{d}x \approx \frac{\Delta x}{2}\omega_i \tau_{i,j}^n \partial_x X(x_{i,j},t^\alpha)\,\mathrm{d}x = \tau_{i,j}^n \int_{\kappa_j} \Phi_{i,j}(x)\partial_x X(x,t^\alpha)\,\mathrm{d}x,$$

the last equality is indeed exact since X and Φ are polynomials of order less or equal than p, so that $\Phi_{i,j}\partial_x X(\cdot,t)$ is of order less or equal than $(2p-1)$. The objective is now to use one integration by part to move the derivative from X to Φ, and to use the numerical fluxes to evaluate the interfacial terms, which gives

$$\int_{\kappa_j} \Phi_{i,j}(x)\partial_x X(x,t^\alpha)\,\mathrm{d}x = \delta_{i,p} X_{j+1/2}^{*,\alpha} - \delta_{i,0} X_{j-1/2}^{*,\alpha} - \frac{\Delta x}{2}\sum_{k=0}^{p} \omega_k X_{k,j}^\alpha \partial_x \Phi_{i,j}(x_{k,j}).$$

Note that in the above formula and the following ones, the star quantities $X_{j+1/2}^{*,\alpha}$ with $X = u, \Pi$ will be defined using a similar definition as in Section 3.2, namely

$$u_{j+1/2}^{*,\alpha} = \frac{1}{2}(u_{p,j}^\alpha + u_{0,j+1}^\alpha) - \frac{1}{2a}(\Pi_{0,j+1}^\alpha - \Pi_{p,j}^\alpha) - \frac{\Delta x}{2a}\{gh\partial_x z\}_{j+1/2}, \tag{3.3.5}$$

$$\Pi^{*,\alpha}_{j+1/2} = \frac{1}{2}(\Pi^{\alpha}_{p,j} + \Pi^{\alpha}_{0,j+1}) - \frac{a}{2}(u^{\alpha}_{0,j+1} - u^{\alpha}_{p,j}), \tag{3.3.6}$$

with

$$\{gh\partial_x z\}^n_{j+1/2} = g\frac{h^n_{p,j} + h^n_{0,j+1}}{2}\frac{z_{0,j+1} - z_{p,j}}{\Delta x}. \tag{3.3.7}$$

In particular, our DG scheme will naturally degenerate towards the first-order scheme when $p = 0$.
As far as the source term integral is concerned, it is not possible to move the derivative from z to Φ only. Therefore and according to Equation (3.3.2), we simply write

$$\int_{\kappa_j} \Phi_{i,j}(x)\frac{g}{\tau(x,t^n)}\partial_m z(x)\,\mathrm{d}x \approx \tau^n_{i,j}\frac{\Delta x}{2}\left(\delta_{i,p}\{gh\partial_x z\}^n_{j+1/2} + \delta_{i,0}\{gh\partial_x z\}^n_{j-1/2} + \omega_i\{gh\partial_x z\}^n_{i,j}\right),$$

where $\{gh\partial_x z\}^n_{i,j}$ will be given later on when we prove the well-balanced properties. Finally, we obtain from system (3.3.4) and for the acoustic step the following update formulas

$$\begin{cases} L^{\alpha}_{i,j}h^{n+1-}_{i,j} = h^n_{i,j}, \\ L^{\alpha}_{i,j}(hu)^{n+1-}_{i,j} = (hu)^n_{i,j} - \dfrac{2\Delta t}{\omega_i\Delta x}\left[\delta_{i,p}\Pi^{*,\alpha}_{j+1/2} - \delta_{i,0}\Pi^{*,\alpha}_{j-1/2} - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p}\omega_k\Pi^{\alpha}_{k,j}\partial_x\Phi_{i,j}(x_{k,j})\right] \\ \qquad\qquad - \Delta t\left[\dfrac{\delta_{i,p}}{\omega_p}\{gh\partial_x z\}_{j+1/2} + \dfrac{\delta_{i,0}}{\omega_0}\{gh\partial_x z\}_{j-1/2} + \{gh\partial_x z\}^n_{i,j}\right], \\ L^{\alpha}_{i,j}(h\Pi)^{n+1-}_{i,j} = (h\Pi)^n_{i,j} - a^2\dfrac{2\Delta t}{\omega_i\Delta x}\left[\delta_{i,p}u^{*,\alpha}_{j+1/2} - \delta_{i,0}u^{*,\alpha}_{j-1/2} - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p}\omega_k u^{\alpha}_{k,j}\partial_x\Phi_{i,j}(x_{k,j})\right], \end{cases} \tag{3.3.8}$$

with

$$L^{\alpha}_{i,j} = 1 + \frac{2\Delta t}{\omega_i\Delta x}\left[\delta_{i,p}u^{*,\alpha}_{j+1/2} - \delta_{i,0}u^{*,\alpha}_{j-1/2} - \frac{\Delta x}{2}\sum_{k=0}^{p}\omega_k u^{\alpha}_{k,j}\partial_x\Phi_{i,j}(x_{k,j})\right]. \tag{3.3.9}$$

3.3.2 The transport step

We continue with the transport step (3.2.3).

Time discretization ($t^n \to t^{n+1}$)

Along the lines of the acoustic step, we are led to set

$$\begin{cases} h^{n+1}_{i,j} = h^{n+1-}_{i,j} - \dfrac{2\Delta t}{\omega_i\Delta x}\displaystyle\int_{\kappa_j}\Phi_{i,j}(x)u(x,t^{\alpha})\partial_x h(x,t^{n+1-})\,\mathrm{d}x, \\ (hu)^{n+1}_{i,j} = (hu)^{n+1-}_{i,j} - \dfrac{2\Delta t}{\omega_i\Delta x}\displaystyle\int_{\kappa_j}\Phi_{i,j}(x)u(x,t^{\alpha})\partial_x(hu)(x,t^{n+1-})\,\mathrm{d}x. \end{cases} \tag{3.3.10}$$

Note that this transport step is always treated explicitly in time.

Volume integrals and flux calculations

We want to evaluate the integrals

$$\int_{\kappa_j} \Phi_{i,j}(x) u(x,t^\alpha) \partial_x X(x,t^{n+1-})\,\mathrm{d}x$$

with $X = h, hu$. Using the equality $u\partial_x X = \partial_x(Xu) - X\partial_x u$ and pulling out X from the integral, with the same process as with τ before, leads to

$$\begin{aligned}
&\int_{\kappa_j} \Phi_{i,j}(x) u(x,t^\alpha) \partial_x X(x,t^{n+1-})\,\mathrm{d}x \\
&\qquad = \int_{\kappa_j} \Phi_{i,j}(x) \partial_x \left[X(x,t^{n+1-})u(x,t^\alpha)\right]\,\mathrm{d}x - \int_{\kappa_j} \Phi_{i,j}(x) X(x,t^{n+1-}) \partial_x u(x,t^\alpha)\,\mathrm{d}x \\
&\qquad \approx \int_{\kappa_j} \Phi_{i,j}(x) \partial_x \left[X(x,t^{n+1-})u(x,t^\alpha)\right]\,\mathrm{d}x - X_{i,j}^{n+1-} \int_{\kappa_j} \Phi_{i,j}(x) \partial_x u(x,t^\alpha)\,\mathrm{d}x.
\end{aligned}$$

We approximate the first integral as follows

$$\begin{aligned}
\int_{\kappa_j} \Phi_{i,j}(x) \partial_x \left[X(x,t^{n+1-})u(x,t^\alpha)\right]\,\mathrm{d}x \approx{}& \delta_{i,p} X_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} - \delta_{i,0} X_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha} \\
&- \frac{\Delta x}{2} \sum_{k=0}^{p} \omega_k X_{k,j}^{n+1-} u_{k,j}^{\alpha} \partial_x \Phi_{i,j}(x_{k,j}).
\end{aligned}$$

where we set the upwind flux as in one dimension

$$X_{j+1/2}^{*,\alpha} = \begin{cases} X_{p,j}^{\alpha}, & \text{if } u_{j+1/2}^{*,\alpha} \geq 0, \\ X_{0,j+1}^{\alpha}, & \text{if } u_{j+1/2}^{*,\alpha} \leq 0, \end{cases} \qquad X = h, hu. \tag{3.3.11}$$

And we use the same formula as in the acoustic step for the second integral:

$$\int_{\kappa_j} \Phi_{i,j} \partial_x u(x,t^\alpha)\,\mathrm{d}x = \delta_{i,p} u_{j+1/2}^{*,\alpha} - \delta_{i,0} u_{j-1/2}^{*,\alpha} - \frac{\Delta x}{2} \sum_{k=0}^{p} \omega_k u_{k,j}^{\alpha} \partial_x \Phi_{i,j}(x_{k,j}).$$

At last and from system (3.3.10), we obtain for the transport step

$$\begin{cases}
h_{i,j}^{n+1} = L_{i,j}^{\alpha} h_{i,j}^{n+1-} - \dfrac{2\Delta t}{\omega_i \Delta x} \Bigg[\delta_{i,p} h_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} - \delta_{i,0} h_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha} \\
\qquad\qquad - \dfrac{\Delta x}{2} \displaystyle\sum_{k=0}^{p} \omega_k h_{k,j}^{n+1-} u_{k,j}^{\alpha} \partial_x \Phi_{i,j}(x_{k,j}) \Bigg], \\
(hu)_{i,j}^{n+1} = L_{i,j}^{\alpha} (hu)_{i,j}^{n+1-} - \dfrac{2\Delta t}{\omega_i \Delta x} \Bigg[\delta_{i,p} (hu)_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} - \delta_{i,0} (hu)_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha} \\
\qquad\qquad - \dfrac{\Delta x}{2} \displaystyle\sum_{k=0}^{p} \omega_k (hu)_{k,j}^{n+1-} u_{k,j}^{\alpha} \partial_x \Phi_{i,j}(x_{k,j}) \Bigg].
\end{cases} \tag{3.3.12}$$

3.3.3 The whole scheme for the nodal and mean values

Gathering steps (3.3.8) and (3.3.12), it is easy to see that the whole Lagrange-projection scheme writes

$$\begin{cases} h_{i,j}^{n+1} = h_{i,j}^n - \dfrac{2\Delta t}{\omega_i \Delta x}\Bigg[\delta_{i,p} h_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} - \delta_{i,0} h_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha} \\ \qquad\qquad - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} \omega_k h_{k,j}^{n+1-} u_{k,j}^{\alpha} \partial_x \Phi_{i,j}(x_{k,j})\Bigg], \\ (hu)_{i,j}^{n+1} = (hu)_{i,j}^n - \dfrac{2\Delta t}{\omega_i \Delta x}\left[\delta_{i,p}\Pi_{j+1/2}^{*,\alpha} - \delta_{i,0}\Pi_{j-1/2}^{*,\alpha} - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p}\omega_k \Pi_{k,j}^{n+1-}\partial_x\Phi_{i,j}(x_{k,j})\right] \\ \qquad - \dfrac{2\Delta t}{\omega_i \Delta x}\Bigg[\delta_{i,p}(hu)_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} - \delta_{i,0}(hu)_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha} \\ \qquad\qquad - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p}\omega_k (hu)_{k,j}^{n+1-} u_{k,j}^{\alpha}\partial_x\Phi_{i,j}(x_{k,j})\Bigg] \\ \qquad - \Delta t\left[\dfrac{\delta_{i,p}}{\omega_p}\{gh\partial_x z\}_{j+1/2} + \dfrac{\delta_{i,0}}{\omega_0}\{gh\partial_x z\}_{j-1/2} + \{gh\partial_x z\}_{i,j}^n\right]. \end{cases} \tag{3.3.13}$$

In particular, the first-order scheme (3.2.6) is recovered when $p = 0$. On the other hand and in order to state the stability properties satisfied by this scheme, we will be interested in the evolution of the mean values $\overline{X}_j^{n+1}$ for $X = h, hu$, which are naturally defined for all n an j by

$$\overline{X}_j^{n+1} = \frac{1}{\Delta x}\int_{\kappa_j} X(x, t^n)\,\mathrm{d}x = \sum_{i=0}^{p}\frac{\omega_i}{2} X_{i,j}^n.$$

Multiplying system (3.3.13) by $\omega_i/2$ and summing over i leads to

$$\begin{cases} \overline{h}_j^{n+1} = \overline{h}_j^n - \dfrac{\Delta t}{\Delta x}\left[h_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} - h_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha}\right], \\ \overline{(hu)}_j^{n+1} = \overline{(hu)}_j^n - \dfrac{\Delta t}{\Delta x}\left[\Pi_{j+1/2}^{*,\alpha} + (hu)_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,\alpha} - \Pi_{j-1/2}^{*,\alpha} - (hu)_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,\alpha}\right] \\ \qquad - \Delta t\left[\dfrac{\{gh\partial_x z\}_{j-1/2}^n + \{gh\partial_x z\}_{j+1/2}^n}{2} + \displaystyle\sum_{i=0}^{p}\frac{\omega_i}{2}\{gh\partial_x z\}_{i,j}^n\right], \end{cases} \tag{3.3.14}$$

recalling that equality (3.3.1) holds true, so that $\sum_{i=0}^{p}\Phi_{i,j}'(x) = 0$. Note that by Equation (3.3.2), the last term of system (3.3.14) approximates a volume integral of the source term such that

$$\frac{1}{\Delta x}\int_{\kappa_j} gh\partial_x z\,\mathrm{d}x \approx \sum_{i=0}^{p}\frac{\omega_i}{2}\{gh\partial_x z\}_{i,j}^n.$$

3.4 Stability and well-balanced properties

This section aims at giving the stability and well-balanced properties of our schemes, and to discuss the use of limiters.

3.4.1 Positivity properties and discrete entropy inequality

Equipped with systems (3.3.13) and (3.3.14), we now aim at proving some stability properties of the scheme under some suitable CFL condition. We are especially interested in the case $\alpha = n+1-$ even though the first two lemmas stay valid for $\alpha = n$.

Lemma 3.4.1. *Under the CFL condition*

$$\frac{\Delta t}{\Delta x} \max_j \max_i \frac{1}{w_i} \left(\int_{\kappa_j} u(x,t^\alpha) \partial_x \Phi_{i,j}(x)\, \mathrm{d}x - \delta_{i,p} \left(u^{*,\alpha}_{j+1/2} \right)_- + \delta_{i,0} \left(u^{*,\alpha}_{j-1/2} \right)_+ \right) < \frac{1}{2}, \tag{3.4.1}$$

the mean values $\overline{X}_j^{n+1}$, with $X = h, hu$, are convex combinations of the nodal values $X_{i,j}^{n+1-}$ for $i = 0, ..., p$, $X_{0,j+1}^{n+1-}$ and $X_{p,j-1}^{n+1-}$. More precisely, we have

$$\overline{X}_j^{n+1} = \sum_{i=0}^{p} \left(\frac{w_i}{2} - \frac{\Delta t}{\Delta x} \left[\int_{\kappa_j} u(x,t^\alpha) \partial_x \Phi_{i,j}(x) dx - \delta_{i,p} \left(u^{*,\alpha}_{j+1/2} \right)_- + \delta_{i,0} \left(u^{*,\alpha}_{j-1/2} \right)_+ \right] \right) X_{i,j}^{n+1-} \\ - \frac{\Delta t}{\Delta x} \left(u^{*,\alpha}_{j+1/2} \right)_- X_{0,j+1}^{n+1-} + \frac{\Delta t}{\Delta x} \left(u^{*,\alpha}_{j-1/2} \right)_+ X_{p,j-1}^{n+1-}, \tag{3.4.2}$$

where

$$\int_{\kappa_j} u(x,t^\alpha) \partial_x \Phi_{i,j}(x) dx = \frac{\Delta x}{2} \sum_{k=0}^{p} \omega_k u^\alpha_{k,j} \partial_x \Phi_{i,j}(x_{k,j}).$$

Proof. Let us multiply system (3.3.12) by ω_i and sum over i. Setting $X = h, hu$, we immediately get

$$\overline{X}_j^{n+1} = \sum_{i=0}^{p} \frac{\omega_i}{2} L^\alpha_{i,j} X_{i,j}^{n+1-} - \frac{\Delta t}{\Delta x} \left[X^{*,n+1-}_{j+1/2} u^{*,\alpha}_{j+1/2} - X^{*,n+1-}_{j-1/2} u^{*,\alpha}_{j-1/2} \right].$$

Recall that $L^\alpha_{i,j}$ is given by Equation (3.3.9) or equivalently

$$L^\alpha_{i,j} = \frac{2}{\omega_i} \left(\frac{\omega_i}{2} - \frac{\Delta t}{\Delta x} \left[\int_{\kappa_j} u^\alpha\, \partial_x \Phi_{i,j} - \delta_{i,p} \left(u^{*,\alpha}_{j+1/2} \right)_- + \delta_{i,0} \left(u^{*,\alpha}_{j-1/2} \right)_+ \right] \right. \\ \left. + \delta_{i,p} \frac{\Delta t}{\Delta x} \left(u^{*,\alpha}_{j+1/2} \right)_+ - \delta_{i,0} \frac{\Delta t}{\Delta x} \left(u^{*,\alpha}_{j-1/2} \right)_- \right). \tag{3.4.3}$$

Using the definition (3.3.11) of $X^{*,n+1-}_{j\pm1/2}$ easily gives Equation (3.4.2). On the other hand, the coefficients in the Equation (3.4.2) are non negative under the CFL condition (3.4.1) and since $\left(u^{*,\alpha}_{j+1/2} \right)_- \leq 0$ and $\left(u^{*,\alpha}_{j-1/2} \right)_+ \geq 0$. Finally, the sum of these coefficients equals one thanks to equalities (3.3.1) and (3.3.3). □

As a consequence of this lemma, we can easily prove the following result which is concerned with the positivity of the water heights.

Lemma 3.4.2. *Under the CFL condition* (3.4.1)*, the quantities $L^\alpha_{i,j}$ are positive. Thus, if the water heights are positive at time t^n, that is to say if $h^n_{i,j} > 0$ for all i and j, then the water heights are also positive at the fictitious time t^{n+1-}, that is to say $h^{n+1-}_{i,j} > 0$ for all i and j, and the mean values are positive at time t^{n+1}, namely $\overline{h}^{n+1}_j > 0$ for all j.*

Proof. Under the CFL condition (3.4.1) and by Equation (3.4.3), we have $L^{\alpha}_{i,j} > 0$. Since $L^{\alpha}_{i,j} h^{n+1-}_{i,j} = h^n_{i,j}$, it is thus clear that $h^{n+1-}_{i,j} > 0$ provided that $h^n_{i,j} > 0$. Finally, the mean values are positive by convex combination, which concludes the proof. □

We now state that the proposed implicit explicit scheme satisfies a discrete entropy inequality. The proof is given in Appendix 3.A.

Theorem 3.4.1. *Under the CFL condition* (3.4.1)*, the implicit explicit scheme* (3.3.8) *and* (3.3.12) *(or equivalently* (3.3.13)*) with* $\alpha = n+1-$ *satisfies the following in-cell discrete non conservative entropy inequality which is consistent with inequality* (3.1.2)*, namely for all* j

$$\begin{aligned}(hE)(\overline{\boldsymbol{U}}^{n+1}_j) - \overline{(hE)}^n_j \\ + \frac{\Delta t}{\Delta x}\left[\left(\pi^{*,n+1-}_{j+1/2} + (hE)^{*,n+1-}_{j+1/2}\right) u^{*,n+1-}_{j+1/2} - \left(\pi^{*,n+1-}_{j-1/2} + (hE)^{*,n+1-}_{j-1/2}\right) u^{*,n+1-}_{j-1/2}\right] \\ \leq -\Delta t \, \{ghu\partial_x z\}^{n+1-}_j \,, \end{aligned} \tag{3.4.4}$$

where

$$\{ghu\partial_x z\}^{n+1-}_j = \sum_{i=0}^{p} \frac{\omega_i}{2} u^{n+1-}_{i,j} \{gh\partial_x z\}^n_{i,j} - \frac{1}{2a} \overleftarrow{W}^{n+1-}_{p,j} \{gh\partial_x z\}^n_{j+1/2} + \frac{1}{2a} \overrightarrow{W}^{n+1-}_{0,j} \{gh\partial_x z\}^n_{j-1/2} \,,$$

$$(hE)^{n+1-}_{j+1/2} = \begin{cases} (hE)^{n+1-}_{p,j}, & \text{if } u^*_{j+1/2} \geq 0, \\ (hE)^{n+1-}_{0,j+1}, & \text{if } u^*_{j+1/2} \leq 0, \end{cases}$$

and

$$\overrightarrow{W} = \Pi + au \quad \text{and} \quad \overleftarrow{W} = \Pi - au.$$

3.4.2 Well-balanced properties

We now give the well-balanced properties satisfied by our scheme. In the rest of this section, we assume that the initial condition satisfies the so-called lake at rest conditions, namely

$$u^0_{i,j} = 0 \quad \text{and} \quad h^0_{i,j} + z(x_{i,j}) = K,$$

for all i and j and for a given constant K. We first state conditional well-balanced properties associated with the simple source term definition (3.4.5), and then give a new definition (3.4.6) which leads to unconditional well-balanced properties. These properties will be illustrated in the numerical section.

Conditional well-balanced properties

The conditional well-balanced properties are valid for the simple source term definition:

$$\{gh\partial_x z\}^n_{i,j} = gh^n_{i,j} \partial_x z(x_{i,j}). \tag{3.4.5}$$

Our first result is concerned with the explicit explicit scheme, that is to say $\alpha = n$ in the Lagrangian step.

Proposition 3.4.1. — *(i) Let us assume that the initial water height h^0 and the topography z are smooth polynomial functions of order less or equal than p. Then, the explicit explicit scheme ($\alpha = 0$) satisfies the well-balanced property for the mean values, that is to say*

$$\overline{u}_j^1 = 0 \quad \textit{and} \quad \overline{h}_j^1 + \overline{z}_j = K$$

for all j.
(ii) Let us assume that the initial water height h^0 and the topography z are smooth polynomial functions of order less or equal than $p/2$. Then, the explicit explicit scheme satisfies the well-balanced property for the nodal values, that is to say

$$u_{i,j}^1 = 0 \quad \textit{and} \quad h_{i,j}^1 + z(x_{i,j}) = K$$

for all j.

Proof. Let us begin with the mean values. By definition (3.3.7) and the lake at rest initial condition, we have

$$\{gh\partial_x z\}_{j+1/2}^0 = g\frac{h_{p,j}^0 + h_{0,j+1}^0}{2}\frac{(-h_{0,j+1}^0) - (-h_{p,j}^0)}{\Delta x} = \frac{\Pi_{p,j}^0 - \Pi_{0,j+1}^0}{\Delta x},$$

thus we can simplify the expressions of the flux (3.3.5) and (3.3.6)

$$u_{j+1/2}^{*,0} = 0, \qquad \text{and} \qquad \Pi_{j+1/2}^{*,0} = \frac{\Pi_{p,j}^0 + \Pi_{0,j+1}^0}{2}.$$

Therefore, system (3.3.14) gives

$$\begin{cases} \overline{h}_j^1 = \overline{h}_j^0, \\ \overline{(hu)}_j^1 = \overline{(hu)}_j^0 - \dfrac{\Delta t}{\Delta x}\left[\Pi_{p,j}^0 - \Pi_{0,j}^0 + \Delta x \displaystyle\sum_{i=0}^{p} \frac{\omega_i}{2}\{gh\partial_x z\}_{i,j}^0\right]. \end{cases}$$

However, since $gh\partial_x z$ is a polynomial of order less than or equal to $2p-1$, we have

$$\Delta x \sum_{i=0}^{p} \frac{\omega_i}{2}\{gh\partial_x z\}_{i,j}^0 = \int_{\kappa_j} gh^0\partial_x z \,\mathrm{d}x = -\int_{\kappa_j} gh^0\partial_x h^0 \,\mathrm{d}x = 0 - \int_{\kappa_j} \partial_x \Pi^0 \,\mathrm{d}x = \Pi_{0,j}^0 - \Pi_{p,j}^0,$$

so that $\overline{h}_j^1 = \overline{h}_j^0$ and $\overline{(hu)}_j^1 = \overline{(hu)}_j^0$, which concludes the proof of (i). Let us now turn to the well-balanced property on the nodal values (ii). By system (3.3.13), the same arguments as above give

$$\begin{cases} h_{i,j}^1 = h_{i,j}^0, \\ (hu)_{i,j}^1 = (hu)_{i,j}^0 - \dfrac{2\Delta t}{\omega_i \Delta x}\left[\delta_{i,p}\Pi_{p,j}^0 - \delta_{i,0}\Pi_{0,j}^0 - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p}\omega_k \Pi_{k,j}^0 \partial_x \Phi_{i,j}(x_{k,j}) + \dfrac{\Delta x}{2}\omega_i\{gh^0\partial_x z\}_{i,j}^0\right]. \end{cases}$$

On the one hand, $gh\partial_x z$ is a polynomial of order less than or equal to $p-1$, thus we have that

$$\frac{\Delta x}{2}\omega_i\{gh^0\partial_x z\}_{i,j}^0 = \int_{\kappa_j} \Phi_{i,j} gh^0\partial_x z \,\mathrm{d}x.$$

On the other hand,

$$\begin{aligned}\frac{\Delta x}{2}\sum_{k=0}^{p}\omega_k\Pi^0_{k,j}\partial_x\Phi_{i,j}(x_{k,j}) - \frac{\Delta x}{2}\omega_i\{gh^0\partial_x z\}^0_{i,j} &= \int_{\kappa_j}\Pi^0_{k,j}\partial_x\Phi_{i,j}\,\mathrm{d}x - \int_{\kappa_j}\Phi_{i,j}gh^0\partial_x z\,\mathrm{d}x\\ &= \delta_{i,p}\Pi^0_{p,j} - \delta_{i,0}\Pi^0_{0,j} - \int_{\kappa_j}\Phi_{i,j}\partial_x\Pi^0_{k,j}\,\mathrm{d}x - \int_{\kappa_j}\Phi_{i,j}gh^0\partial_x z\,\mathrm{d}x\\ &= \delta_{i,p}\Pi^0_{p,j} - \delta_{i,0}\Pi^0_{0,j} - \int_{\kappa_j}\Phi_{i,j}\partial_x\Pi^0_{k,j}\,\mathrm{d}x + \int_{\kappa_j}\Phi_{i,j}\partial_x\Pi^0_{k,j}\,\mathrm{d}x\\ &= \delta_{i,p}\Pi^0_{p,j} - \delta_{i,0}\Pi^0_{0,j},\end{aligned}$$

which gives $(hu)^1_{i,j} = (hu)^0_{i,j}$ and concludes the proof. □

The next result concern the implicit-explicit scheme.

Proposition 3.4.2. *Let us assume that the initial water height h^0 and the topography z are smooth polynomial functions of order less or equal than $p/2$. Then, the implicit explicit scheme ($\alpha = 1-$) satisfies the well-balanced property for the nodal values, that is to say*

$$u^1_{i,j} = 0 \quad \textit{and} \quad h^1_{i,j} + z(x_{i,j}) = K$$

for all j.

Proof. We first aim at checking that $h^{1-}_{i,j} = h^0_{i,j}$, $u^{1-}_{i,j} = u^0_{i,j} = 0$ and $\Pi^{1-}_{i,j} = \Pi^0_{i,j}$ is the unique solution of system (3.3.8) with $\alpha = n+1-$ (and $n = 0$). From definition (3.3.9), it is clear that $L^{1-}_{i,j} = 1$ and system (3.3.8) becomes

$$\begin{cases} h^0_{i,j} = L^{1-}_{i,j}h^{1-}_{i,j} = h^0_{i,j},\\ 0 = L^{1-}_{i,j}(hu)^{1-}_{i,j} = 0 - \dfrac{2\Delta t}{\omega_i\Delta x}\left[\delta_{i,p}\Pi^n_{p,j} - \delta_{i,0}\Pi^n_{0,j} - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p}\omega_k\Pi^n_{k,j}\partial_x\Phi_{i,j}(x_{k,j}) \right.\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. + \dfrac{\Delta x}{2}\omega_i\{gh^n\partial_x z\}^n_{i,j}\right],\\ (h\Pi)^0_{i,j} = L^{1-}_{i,j}(h\Pi)^{1-}_{i,j} = (h\Pi)^0_{i,j} - 0.\end{cases}$$

The same arguments as in the proof of the previous Proposition also apply here to show that the second equality actually holds true. Then, the projection step is trivial since $L^1_{i,j} = 1$ and all the velocities are zero, which concludes the proof. □

Unconditional well-balanced properties

As clearly stated in the above results, the well-balanced properties are subject to a restriction on the shapes of the initial water height and topography. It is actually possible to get rid of these restriction by simply changing the definition of the volume integral (3.4.5), as it was suggested to us by M. J. Castro. More precisely, if suffices to set

$$\{gh\partial_x z\}^n_{i,j} = gh^n_{i,j}\partial_x(h^n + z)(x_{i,j}) - \partial_x\Pi^n(x_{i,j}). \tag{3.4.6}$$

Proof. From the proofs with the conditional well-balanced properties above, we can keep the same clues. For example, for the implicit-explicit scheme, we obtain again for the second equation of system (3.3.8):

$$0 = L^{1-}_{i,j}(hu)^{1-}_{i,j} = 0 - \frac{2\Delta t}{\omega_i \Delta x}\left[\delta_{i,p}\Pi^n_{p,j} - \delta_{i,0}\Pi^n_{0,j} - \frac{\Delta x}{2}\sum_{k=0}^{p}\omega_k \Pi^n_{k,j}\partial_x \Phi_{i,j}(x_{k,j}) + \frac{\Delta x}{2}\omega_i\{gh^n\partial_x z\}^n_{i,j}\right],$$

The only computation we have to detail is the following one, true without any more assumption over h and z than $h + z =$ constant:

$$\begin{aligned}
\frac{\Delta x}{2}\sum_{k=0}^{p}\omega_k \Pi^0_{k,j}\partial_x \Phi_{i,j}(x_{k,j}) - \frac{\Delta x}{2}\omega_i\{gh^0\partial_x z\}^0_{i,j} \\
&= \int_{\kappa_j}\Pi^0_{k,j}\partial_x\Phi_{i,j}\,\mathrm{d}x - \frac{\Delta x}{2}\omega_i g h^0_{i,j}\partial_x(h+z)(x^0_{i,j}) + \frac{\Delta x}{2}\omega_i\partial_x\Pi(x^0_{i,j}) \\
&= \delta_{i,p}\Pi^0_{p,j} - \delta_{i,0}\Pi^0_{0,j} - \int_{\kappa_j}\Phi_{i,j}\partial_x\Pi^0_{k,j}\,\mathrm{d}x + \int_{\kappa_j}\Phi_{i,j}\partial_x\Pi^0_{k,j}\,\mathrm{d}x \\
&= \delta_{i,p}\Pi^0_{p,j} - \delta_{i,0}\Pi^0_{0,j}.
\end{aligned}$$

We have proven that any lake at rest is strictly preserved with this source term treatment, by the explicit-explicit and the implicit-explicit scheme. □

3.4.3 Positivity and generalized slope limiters

This section describes the limiters used to stabilize the proposed Discontinuous-Galerkin approach.

Positivity limiters

We have already proved the positivity of the nodal values $h^{n+1-}_{i,j}$ at the end of the Lagrange step and the mean values $\overline{h}^{n+1}_j$ at the end of the projection step. Therefore and similarly to [13], we use a positivity limiter to ensure that the nodal values $h^{n+1}_{i,j}$ are positive at the end of the projection step. More precisely, we suggest to replace $h^{n+1}_{i,j}$ by $\theta_j h^{n+1}_{i,j} + (1-\theta_j)\,\overline{h}^{n+1}_j$, where the coefficients θ_j are taken to be

$$\theta_j = \min\left(1, \frac{\overline{h}^{n+1}_j - \varepsilon}{\overline{h}^{n+1}_j - \min_i h^{n+1}_{i,j}}\right).$$

This formula ensures that if h is less than a given threshold $\varepsilon > 0$, the nodal value of the corresponding cell is replaced by the positive mean value. In practice, ε is taken to be equal to 10^{-10}.

Generalized slope limiters in conservative variables

In order to avoid non physical oscillations, we also use the generalized slope limiters introduced in [9]. More precisely, considering the *minmod* function $m(a,b,c) = s \cdot \min(|a|,|b|,|c|)$ if $s = \text{sign}(a) = \text{sign}(b) = \text{sign}(c)$ and 0 otherwise, the increments

$$\Delta_+ \overline{X}_j^{n+1} = \overline{X}_{j+1}^{n+1} - \overline{X}_j^{n+1}, \quad \Delta_- \overline{X}_j^{n+1} = \overline{X}_j^{n+1} - \overline{X}_{j-1}^{n+1},$$

and the values

$$\begin{cases} X_{j+1/2}^{-,n+1} = \overline{X}_j^{n+1} + m\Big(X_{p,j}^{n+1} - \overline{X}_j^{n+1}, \Delta_+ \overline{X}_j^{n+1}, \Delta_- \overline{X}_j^{n+1}\Big), \\ X_{j-1/2}^{+,n+1} = \overline{X}_j^{n+1} - m\Big(\overline{X}_j^{n+1} - X_{0,j}^{n+1}, \Delta_+ \overline{X}_j^{n+1}, \Delta_- \overline{X}_j^{n+1}\Big), \end{cases}$$

the new states at time t^{n+1} are defined by

$$\begin{cases} X_{i,j}^{n+1} & \text{if } X_{j+1/2}^{-,n+1} = X_{p,j}^{n+1} \text{ and } X_{j-1/2}^{+,n+1} = X_{0,j}^{n+1}, \\ \overline{X}_j^{n+1} + \frac{2}{\Delta x}(x_{i,j} - x_j) \cdot m\Big(\partial_x X^{n+1}(x_j), \Delta_+ \overline{X}_j^{n+1}, \Delta_- \overline{X}_j^{n+1}\Big) & \text{otherwise.} \end{cases}$$

Generalized slope limiters in characteristic variables

As proposed in [16], one can also use the slope limiters written in characteristic variables. For each cell, instead of limiting on conservative variables $X = h$ or hu, we use the above general slope limiter on characteristic variables $\tilde{X} = W_1$ or W_2, where $W = R^{-1}\begin{pmatrix} h & hu \end{pmatrix}^T$. The matrix R is the one diagonalizing the Jacobian of the physical flux evaluated at the local mean values, namely

$$R^{-1} \frac{\partial f\big(\overline{U}\big)}{\partial U} R = \begin{pmatrix} \overline{u} + \overline{c} & 0 \\ 0 & \overline{u} - \overline{c} \end{pmatrix}$$

with $\overline{u} = \overline{q}/\overline{h}$, $\overline{c} = \sqrt{g\overline{h}}$,

$$\frac{\partial f\big(\overline{U}\big)}{\partial U} = \begin{pmatrix} 0 & 1 \\ g\overline{h} - \overline{u}^2 & 2\overline{u} \end{pmatrix} \qquad \text{and} \qquad R = \begin{pmatrix} 1 & 1 \\ \overline{u} + \overline{c} & \overline{u} - \overline{c} \end{pmatrix}.$$

We also take profit of the methodology proposed in [16] in order to keep the well-balanced property. We thus apply the general slope limiter on the local characteristic variables associated to $(h + z, q)$, instead of (h, q), on cells where the positivity limiter is not activated.

3.4.4 Time discretization

To conclude the description of our numerical strategy, let us briefly mention that the high order time discretization is obtained as in [13] using the Strong-Stability-Preserving Runge-Kutta (RKSSP) approach, see [9]. For that, we consider the two steps of our Lagrange-projection scheme as a single step which allows to define the solution at time t^{n+1} from the solution at time t^n. When we use a spatial discretization of order $p+1$, that is to say using base polynomials of order p, we combine with a RKSSP time discretization of the corresponding order $p+1$. In practice, we have implemented methods up to the order 3 and when we need to test greater spatial orders we constrain the time steps so that the truncature error related to the time is of needed spatial order.

3.5 Numerical results

The aim of this section is to illustrate the behaviour of our explicit-explicit and implicit-explicit Lagrange-projection schemes, respectively denoted EXEX_p and IMEX_p, where p refers to the polynomial order of the DG approach. Recall that the sound speed is given by $c = \sqrt{p'(h)}$ with $p(h) = gh^2/2$ so that the parameter a is chosen locally at each interface according to

$$a_{j+1/2} = \kappa \max\left(h_j^n \sqrt{gh_j^n}, h_{j+1}^n \sqrt{gh_{j+1}^n}\right)$$

with $\kappa = 1.01$ and $g = 9.81$. We set $\Delta t = \min(\Delta t_{\text{Lag}}, \Delta t_{\text{Tra}})$ for the EXEX_p schemes and $\Delta t = \Delta t_{\text{Tra}}$ for the IMEX_p schemes where

$$\Delta t_{\text{Lag}} = \frac{\Delta x}{(2p+1)\max_j\left(2a_{j+1/2}\max(\tau_{p,j}, \tau_{0,j+1})\right)}$$

is the DG time-step restriction associated with the Lagrangian step, while the transport step CFL restriction is taken from condition (3.4.1) to define Δt_{Tra}.

3.5.1 Comparison between limiters

In this test case, we first compare the results given by the general slope limiters applied to local conservative variables and characteristic variables. For that, we consider the simpler case of a constant topography, meaning that the source term is not taken into account. The space domain $[0, 1500]$ is divided into two parts with the same length and such that the total water height on the right-hand side is small compared to the left-hand side,

$$h(x, t = 0) = \begin{cases} 20, & \text{if } x \leq 750, \\ 1, & \text{if } x > 750. \end{cases}$$

The initial velocity is set to be zero on both sides and the final time is $T = 20$. At last, the spatial domain is discretized over a 500-cell uniform grid and absorbing (Neumann) boundary conditions are used.

We can observe on Figure 3.5.1 that the results of the EXEX_2 scheme with general slope limiters applied to local characteristic variables gives spurious oscillations in the rarefaction wave of a magnitude which is smaller than the one with general slope limiters applied to local conservative variables.

From now on, we will always use the general slope limiters applied to the characteristic variables.

3.5.2 Well-balanced property

The aim of this test case is to illustrate the theoretical results of Sections 3.4.2 and 3.4.2 on the well-balanced properties of the EXEX and IMEX schemes, and both nodal and mean values. The initial velocity is set to be zero and the total water height is constant and equal to 15. The topography z is given by

$$z(x) = \begin{cases} 4, & \text{if } x \leq 750, \\ 2 + \frac{10}{750}(x - 750), & \text{if } x > 750. \end{cases}$$

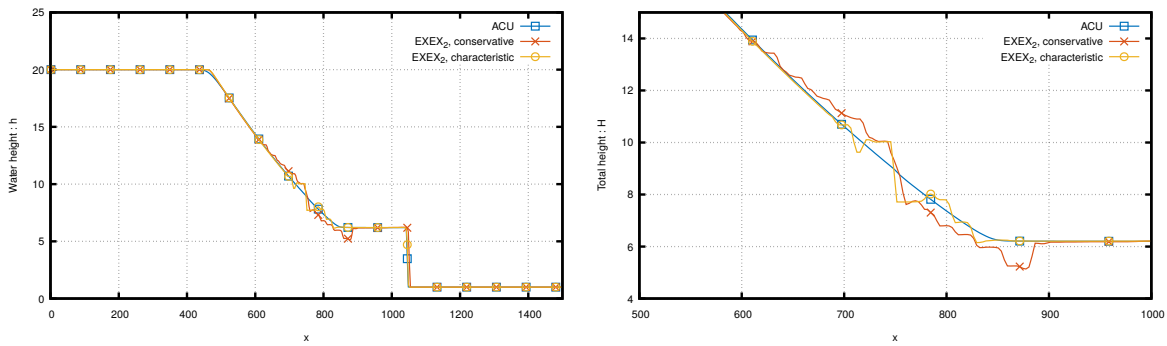

Figure 3.5.1 – Comparison between general slope limiters applied on local conservative variables and characteristic variables at time $T = 20$, with a zoom around the oscillations on the right.

and represented on Figure 3.5.2. It is clearly discontinuous, constant on the left-hand side of the physical domain, and first-order polynomial of degree $r = 1$ on the right-hand side. As stated in Section 3.4.2, one can see in Table 3.1 that when considering the source term definition (3.4.5) the EXEX_1 scheme is exact for the mean values but not for nodal values, while the IMEX_1 scheme is not exact for both mean and nodal values. On the contrary, when $p \geq 2r = 2$, both EXEX_p and IMEX_p are exact for the mean and nodal values. Considering now the source term definition (3.4.6), one can recover the unconditional well-balanced property as clearly seen in Table 3.2 and according to Section 3.4.2.

In order to emphasize those results, we have run the same test case with now initial polynomials of degree $r = 2$ on the right-hand side. More precisely, the topography is now given by

$$z(x) = \begin{cases} 4, & \text{if } x \leq 750, \\ 2 + \frac{10}{750^2}\left(x - 750\right)^2, & \text{if } x > 750. \end{cases}$$

and is represented on Figure 3.5.3. The measure of the well-balanced properties are respectively given in Table 3.3 (Table 3.4) for source term definition (3.4.5) (definition (3.4.6)).

Note that in these test cases, we have used no limiters, although we observed that the use of general slope limiters compatible with the well-balanced property actually improves the results of the EXEX_p schemes when $p \leq 2r$.

3.5.3 Manufactured smooth solution

This test case is taken from [13] and allows us to test the experimental order of accuracy (EOA) of our schemes, especially on the transport step. The space domain is $[0, 1]$, the boundary conditions are periodic and the initial conditions are $h_0(x) = 1 + 0.2\sin(2\pi x)$ and $u_0(x) = 1$. We solve system (3.1.1) with a source term such that the exact solution is $h(x,t) = 1 + 0.2\sin\left(2\pi(x - t)\right)$ and $u(x,t) = 1$,which just means that we impose $u_{i,j}^{n+1^-} = 1$ and $\Pi_{i,j}^{n+1^-} = \Pi_{i,j}^n$, so that the Acoustic step is trivial. For that reason we only present in this special case the results of EXEX_p schemes.

The EOA are reported in Table 3.5. We can observe that we have (at least) the correct $p + 1$ EOA.

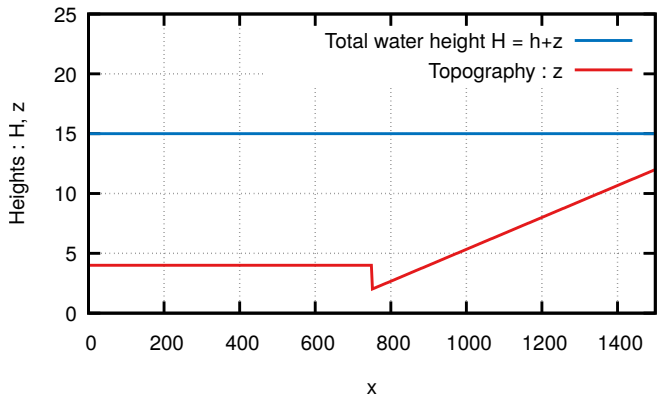

Figure 3.5.2 – Well-balanced property, first-order polynomial topography and initial water height.

$r = 1$		$T = \Delta t$, mean values		$T = \Delta t$, nodal values	
500-cell grid		$\lVert \overline{h+z} - 15 \rVert_\infty / 15$	$\lVert \overline{q}/\overline{h} \rVert_\infty$	$\lVert h+z-15 \rVert_\infty / 15$	$\lVert q/h \rVert_\infty$
EXEX	$p = 0$	9.88 E−17	0.00 E−17	9.88 E−17	0.00 E−17
	$p = 1$	**9.88 E−17**	**4.49 E−16**	9.87 E−17	**9.75 E−6**
	$p = 2$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17
	$p = 3$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17
IMEX	$p = 0$	9.88 E−17	0.00 E−17	9.88 E−17	0.00 E−17
	$p = 1$	**4.91 E−7**	**9.92 E−6**	8.37 E−7	**4.86 E−5**
	$p = 2$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17
	$p = 3$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17

Table 3.1 – Measure of the well-balanced property associated with initial h and z given on Figure 3.5.2 and source term definition (3.4.5).

$r = 1$		$T = \Delta t$, mean values		$T = \Delta t$, nodal values	
500-cell grid		$\left\Vert \overline{h+z}-15 \right\Vert_\infty / 15$	$\left\Vert \overline{q}/\overline{h} \right\Vert_\infty$	$\Vert h+z-15 \Vert_\infty / 15$	$\left\Vert q/h \right\Vert_\infty$
EXEX	$p = 0$	9.88 E−17	0.00 E−17	9.88 E−17	0.00 E−17
	$p = 1$	**9.88 E−17**	**0.00 E−17**	9.87 E−17	**0.00 E−17**
	$p = 2$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17
	$p = 3$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17
IMEX	$p = 0$	9.88 E−17	0.00 E−17	9.88 E−17	0.00 E−17
	$p = 1$	**9.88 E−17**	**0.00 E−17**	9.87 E−17	**0.00 E−17**
	$p = 2$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17
	$p = 3$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17

Table 3.2 – Measure of the well-balanced property associated with initial h and z given on Figure 3.5.2 and source term definition (3.4.6).

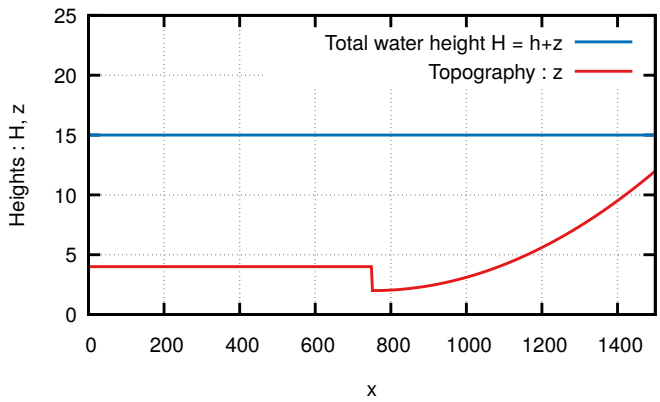

Figure 3.5.3 – Well-balanced property, second-order polynomial topography and initial water height.

$r=2$		$T=\Delta t$, mean values		$T=\Delta t$, nodal values	
500-cell grid		$\left\Vert\overline{h+z}-15\right\Vert_\infty/15$	$\left\Vert\overline{q}/\overline{h}\right\Vert_\infty$	$\Vert h+z-15\Vert_\infty/15$	$\Vert q/h\Vert_\infty$
EXEX	$p=0$	9.87 E−17	0.00 E−17	9.87 E−17	0.00 E−17
	$p=1$	9.87 E−17	4.47 E−16	9.87 E−17	**3.88 E−5**
	$p=2$	**1.97 E−16**	**3.05 E−16**	9.87 E−17	**4.67 E−8**
	$p=3$	**1.97 E−16**	**2.63 E−16**	9.87 E−17	**1.34 E−11**
	$p=4$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17
IMEX	$p=0$	9.87 E−17	0.00 E−17	9.87 E−17	0.00 E−17
	$p=1$	**1.97 E−6**	**3.89 E−5**	2.67 E−6	**1.93 E−4**
	$p=2$	**4.70 E−10**	**9.74 E−9**	6.83 E−9	**1.72 E−7**
	$p=3$	**3.45 E−14**	**6.86 E−13**	1.78 E−12	**3.97 E−11**
	$p=4$	1.98 E−16	0.00 E−17	9.87 E−17	0.00 E−17

Table 3.3 – Measure of the well-balanced property associated with initial h and z given on Figure 3.5.3 and source term definition (3.4.5).

$r=2$		$T=\Delta t$, mean values		$T=\Delta t$, nodal values	
500-cell grid		$\left\Vert\overline{h+z}-15\right\Vert_\infty/15$	$\left\Vert\overline{q}/\overline{h}\right\Vert_\infty$	$\Vert h+z-15\Vert_\infty/15$	$\Vert q/h\Vert_\infty$
EXEX	$p=0$	9.87 E−17	0.00 E−17	9.87 E−17	0.00 E−17
	$p=1$	9.87 E−17	0.00 E−17	9.87 E−17	**0.00 E−17**
	$p=2$	**1.97 E−16**	**0.00 E−17**	9.87 E−17	**0.00 E−17**
	$p=3$	**1.97 E−16**	**0.00 E−17**	9.87 E−17	**0.00 E−17**
	$p=4$	1.97 E−16	0.00 E−17	9.87 E−17	0.00 E−17
IMEX	$p=0$	9.87 E−17	0.00 E−17	9.87 E−17	0.00 E−17
	$p=1$	**9.87 E−17**	**0.00 E−17**	9.87 E−17	**0.00 E−17**
	$p=2$	**1.97 E−16**	**0.00 E−17**	9.87 E−17	**0.00 E−17**
	$p=3$	**1.97 E−16**	**0.00 E−17**	9.87 E−17	**0.00 E−17**
	$p=4$	1.97 E−16	0.00 E−17	9.87 E−17	0.00 E−17

Table 3.4 – Measure of the well-balanced property associated with initial h and z given on Figure 3.5.3 and source term definition (3.4.6).

Δx	$p=0$		$p=1$		$p=2$	
	L^1-error	EOA	L^1-error	EOA	L^1-error	EOA
1/64	5.79E-02		6.46E-04		9.30E-08	
1/128	3.34E-02	0.792118	1.62E-04	1.996037	6.48E-09	3.843292
1/256	1.80E-02	0.889836	4.05E-05	1.998163	4.23E-10	3.936130
1/512	9.39E-03	0.943063	1.01E-05	1.999183	2.72E-11	3.959514

Table 3.5 – EOA for the manufactured smooth solution at time $T = 0.5$, L^1-error $= \|h-h^{\mathrm{Ex}}\|_1 / \|h^{\mathrm{Ex}}\|_1$.

3.5.4 Dam break problem

In this test case, we consider a classic dam break problem. We take the same kind of initial values as in the very first test case above, with velocity set to zero and total water height set to

$$H(x,t=0) = h(x,t=0) + z(x) = \begin{cases} 20, & \text{if } x \leq 750 \\ 15, & \text{if } x > 750. \end{cases}$$

However, the topography is not flat but given by the regularized two-step function

$$z(x) = \begin{cases} 4e^{2-\frac{150}{x-487.5}}, & \text{if } 487.5 < x <= 562.5, \\ 8 - 4e^{2-\frac{150}{637.5-x}}, & \text{if } 562.5 < x <= 637.5, \\ 8, & \text{if } 637.5 < x <= 862.5, \\ 8 - 4e^{2-\frac{150}{x-862.5}}, & \text{if } 862.5 < x <= 937.5, \\ 4e^{2-\frac{150}{1012.5-x}}, & \text{if } 937.5 < x <= 1012.5, \\ 0 & \text{otherwise.} \end{cases}$$

At last, the spatial domain is discretized over a 1500-cell grid and we keep the absorbing boundary conditions.

We can observe the results on Figure 3.5.4 for EXEX$_p$ schemes on the left column and IMEX$_p$ schemes on the right one. The top graphs represents the topography and total water heights at final time $T = 50$. A zoom of the shock moving towards the right boundary is given in the middle graphs. We can see here the expected diffusivity of the IMEX$_p$ schemes although the slope becomes stiffer as the order grows. This last remark can be also done for the EXEX$_p$ schemes. Finally the last graphs represents the velocities. We also show on these graphs the results given by the so-denoted ACU scheme derived in [1].

3.5.5 Propagation of perturbations

This test case focuses on the perturbation of a steady state solution by a pulse that splits into two opposite waves. More precisely, the space domain is reduced to the interval $[0,2]$, the bottom topography is defined by $z(x) = 2 + 0.25(\cos(10\pi(x-0.5)) + 1)$ if $1.4 < x < 1.6$,

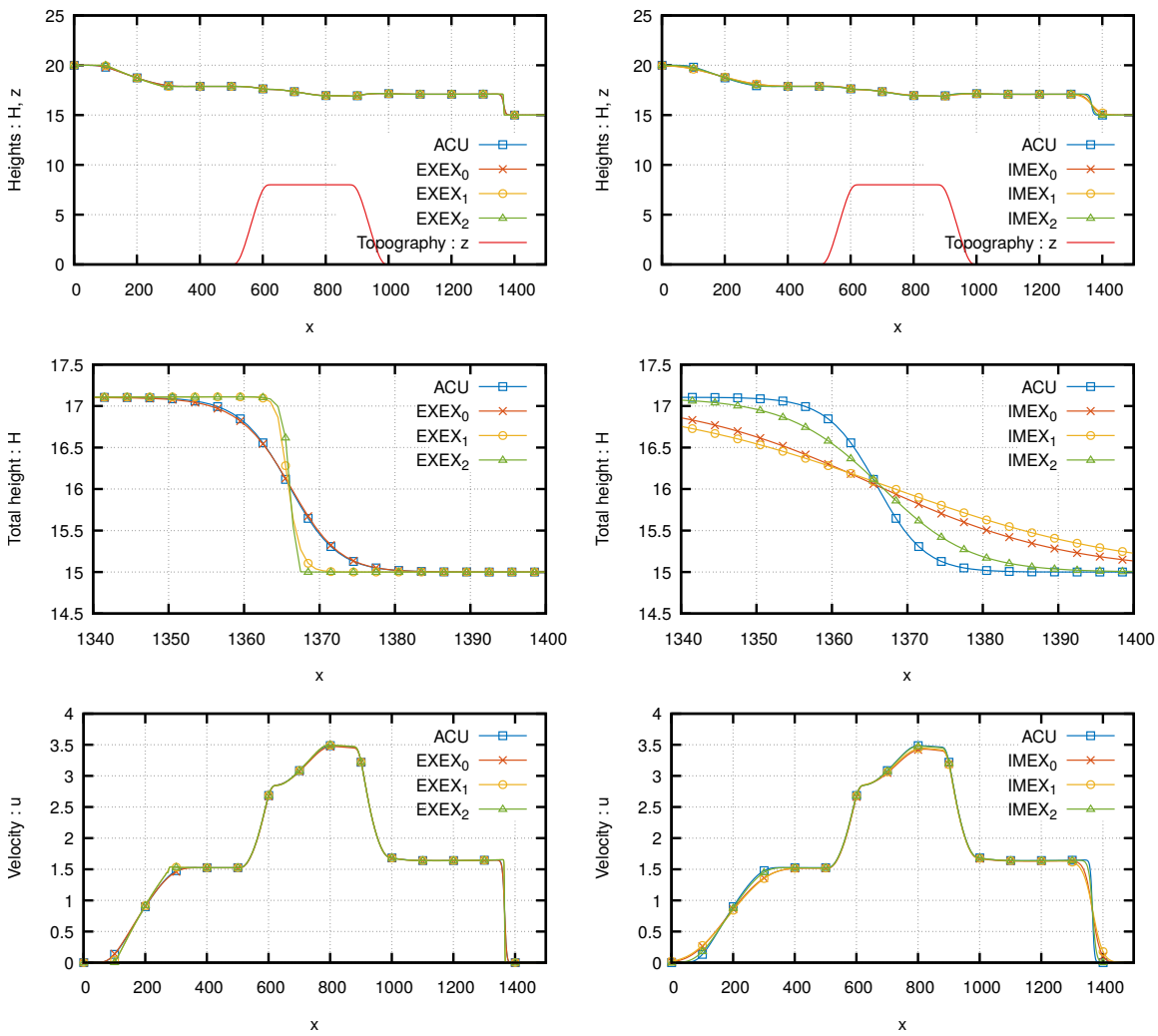

Figure 3.5.4 – Dam break problem, topography and total water heights (top and middle) and velocity (bottom) at time $T = 50$, EXEX_p (left), IMEX_p (right).

and 2 otherwise, and the initial state is such that $u(0,x) = 0$ and $h(0,x) = 3 - z(x) + \Delta h$ if $1.1 < x < 1.2$, and $3 - z(x)$ otherwise, where $\Delta h = 0.001$ is the height of the perturbation. The CFL parameter is set to 0.9, the final time is $T = 0.2$, the space step equals $\Delta x = 1/500$ and Neumann boundary conditions are used.

It turns out that since the perturbation is small, the values of the velocity u keeps a small amplitude during the whole computation. As an immediate consequence, considering the natural implicit-explicit CFL condition gives very large time steps which naturally induces much numerical diffusion. In order to reduce the numerical diffusion and improve the overall accuracy of the numerical solution, the time step is taken as $\Delta t = \min\left(10\Delta t_{\mathrm{Lag}}, \Delta t_{\mathrm{Tra}}\right)$ for the IMEX_p schemes.

Figure 3.5.5 compares the numerical solutions given by the EXEX_p, IMEX_p and ACU schemes. The implicit-explicit schemes are clearly more diffusive than the full explicit ones. Thought in both explicit and implicit schemes we denote an improvement when p gets larger.

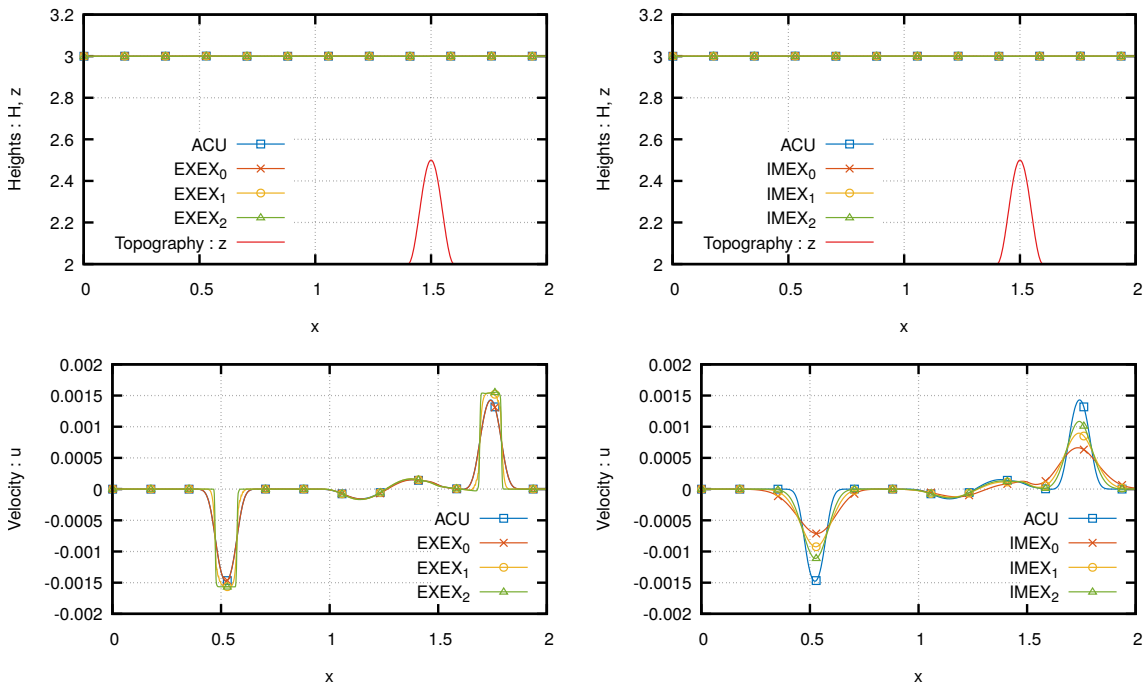

Figure 3.5.5 – Propagation of perturbation problem, topography and total water heights (top) and velocity (bottom) at time $T = 0.2$, EXEX$_p$ (left), IMEX$_p$ (right).

3.5.6 Fluvial regime

The aim of this test case is to test the ability of the schemes to converge to some moving water equilibrium. Let us remind that the steady states are governed by the equations $hu = K_1$ and $\dfrac{u^2}{2} + g(h+z) = K_2$, and we denote $h_{eq}(x), u_{eq}(x)$ the values of h and u at this equilibrium. In this fluvial case we set $K_1 = 1$ and $K_2 = 25$. The domain is $[0, 4]$ and the bottom topography is defined by $z(x) = (\cos(10\pi(x-1)) + 1)/4$ if $1.9 \leq x \leq 2.1$ and 0 elsewhere. The CFL parameter is equal to 0.5 and the space step to $\Delta x = 1/400$. The initial condition is chosen out of equilibrium and given by $h = h_{eq}$ and $u = 0$. The boundary conditions are set to be

$$\begin{cases} \partial_x h(x=0) = 0, \\ (hu)(x=0) = K_1, \end{cases} \quad \text{and} \quad \begin{cases} h(x=4) = h_{eq}(x=4), \\ \partial_x(hu)(x=4) = 0. \end{cases}$$

Figure 3.5.6 shows the solution at the final time $t = 50$. We can observe that the solutions are close to the expected equilibrium, except near the mid domain where the momentum is not yet constant for the mesh size under consideration. Finally, the solutions are more accurate for both EXEX and IMEX schemes when the order grows.

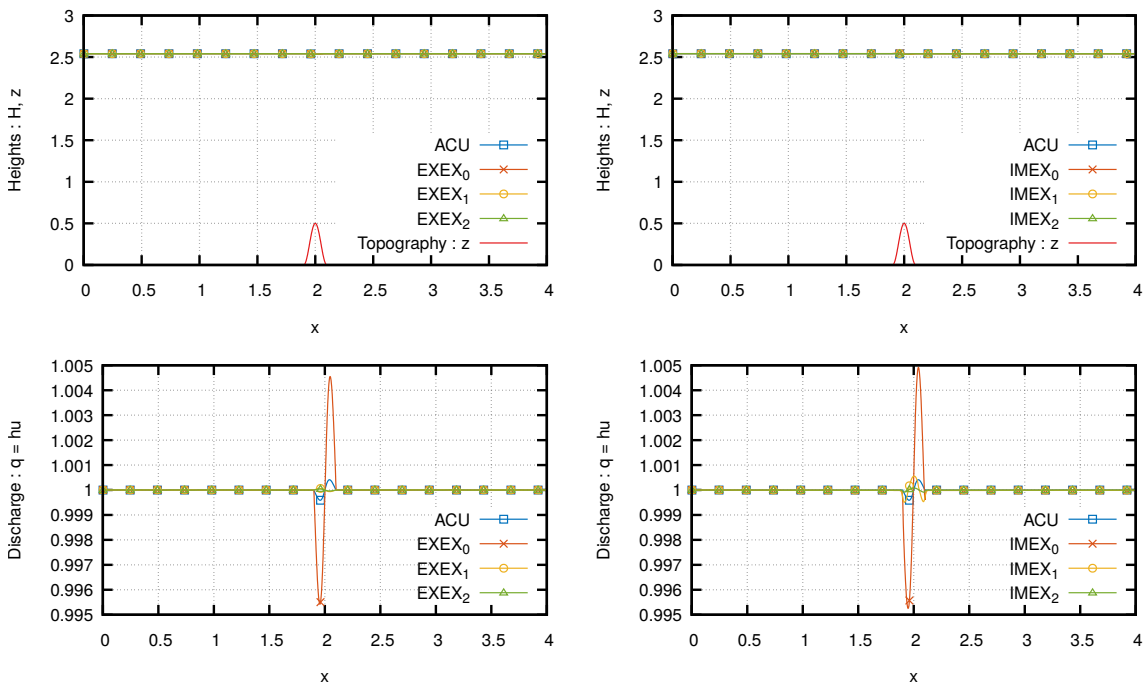

Figure 3.5.6 – Fluvial regime at time $T = 200$. On top : total heights $h + z$, on bottom : discharge hu, respectively for EXEX$_p$ (IMEX$_p$) on the left (right).

3.5.7 Transcritical regime without shock

In this test case, we take the same framework as in the fluvial test case above but we set $K_1 = 3$, $K_2 = \frac{3}{2}(K_1 g)^{2/3} + \frac{g}{2}$. The solutions of IMEX schemes are shown at time $T = 10$ on Figure 3.5.7 and here again the accuracy gets better when the order grows. The solutions of EXEX schemes are not given since the limiters are not able to subside the spurious oscillations on this test case.

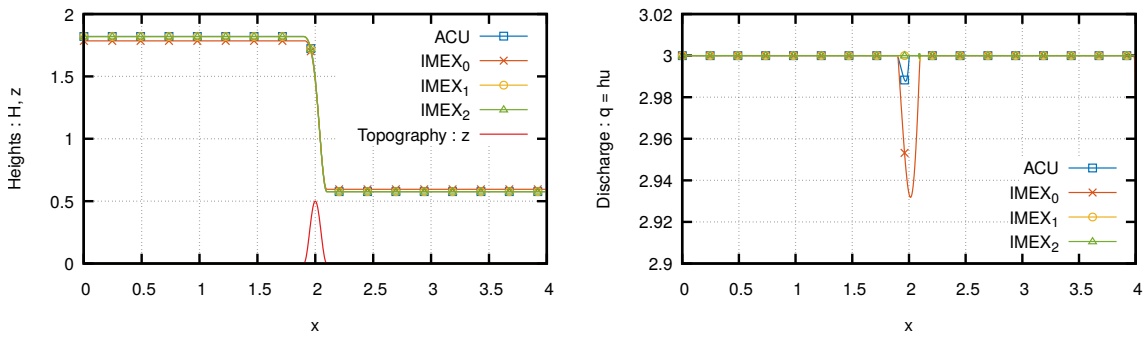

Figure 3.5.7 – Transcritical regime without shock. On the left : total heights $h+z$, on the right : discharge hu.

3.5.8 Transcritical regime with shock

This test has been proposed by Castro et al. [3]. The parameters are described hereafter: the space domain is the interval $[0, 25]$, the bottom topography is defined by $z(x) = 3 - 0.005(x - 10)^2$, if $8 < x < 12$, and 2.8 otherwise. The initial state is defined by $h(0, x) = 3.13 - z(x)$, $q(0, x) = 0.18$ and the boundary conditions are $q(t, 0) = 0.18$, $\partial_x q(t, 25) = 0$, $h(t, 25) = 0.33$ and $\partial_x h(t, 0) = 0$. The final time is set to $T = 200$, the space step to $\Delta x = 1/64$ and the CFL to 0.9. We can see on Figure 3.5.8 that the total water height is properly computed while the limiters are not able to reasonably control the spurious oscillations (note however that an overshoot is already present with the first-order finite volume ACU scheme).

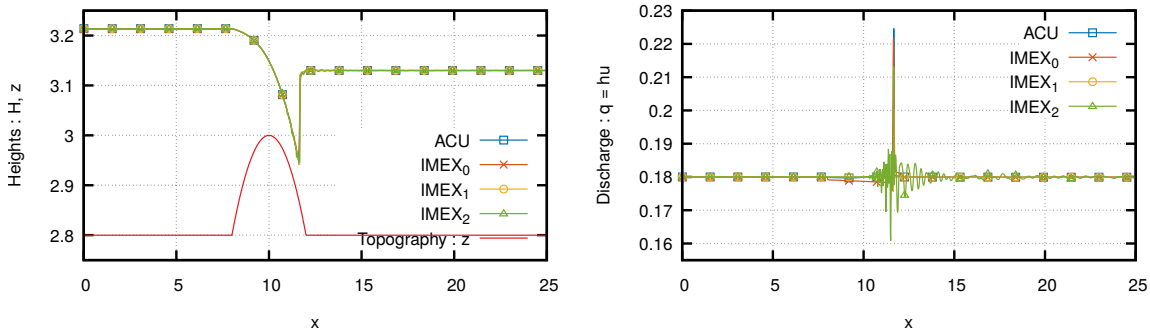

Figure 3.5.8 – Transcritical regime with shock at final time $T = 200$. On the left : total heights $h + z$, on the right : discharge hu.

Acknowledgments

The authors are very grateful to P. Kestener, S. Kokh and F. Renac for stimulating discussions, and the "Maison de la Simulation" for providing excellent working conditions to the second author. The authors are also very grateful to M. J. Castro for providing us with the definition (3.4.6) of the source term volume integral leading to the validity of the unconditional well-balanced property with no restriction on the shape of the initial water height and topography (see Section 3.4.2). The first author was partially funded by ANR Achylles (grant ANR-14-CE25-0001-03).

3.A Proof of the discrete entropy inequality

We adapt the proof given in [7] for the first order finite volume implicit explicit Lagrange-projection scheme to the present discontinuous Galerkin setting.

3.A.1 The acoustic step

Let us first introduce the characteristic variables

$$\overrightarrow{W} = \Pi + au \quad \text{and} \quad \overleftarrow{W} = \Pi - au.$$

Thanks to system (3.3.8) and using an integration by part to equivalently replace the quadrature formula of $\int \overrightarrow{W}\partial_x\Phi dx$ by the one of $\int \partial_x(\overrightarrow{W}\Phi)dx - \int \Phi\partial_x\overrightarrow{W}dx$, we easily get

$$\begin{aligned}
\overrightarrow{W}_{i,j}^{n+1-} = \overrightarrow{W}_{i,j}^{n} - a\frac{2\Delta t}{\omega_i\Delta x}\tau_{i,j}^n \Bigg[& \delta_{i,p}\left(\overrightarrow{W}_{p,j}^{n+1-} - \frac{\Delta x}{2}\left\{gh\partial_x z\right\}_{j+1/2}\right) \\
& - \delta_{i,0}\left(\overrightarrow{W}_{p,j-1}^{n+1-} - \frac{\Delta x}{2}\left\{gh\partial_x z\right\}_{j-1/2}\right) \\
& - \frac{\Delta x}{2}\sum_{k=0}^{p}\omega_k\overrightarrow{W}_{k,j}^{n+1-}\partial_x\phi_{i,j}(x_{k,j})\Bigg] \\
- a\,\Delta t\,\tau_{i,j}^n & \left[\frac{\delta_{i,p}}{\omega_p}\{gh\partial_x z\}_{j+1/2}^n + \frac{\delta_{i,0}}{\omega_0}\{gh\partial_x z\}_{j-1/2}^n + \{gh\partial_x z\}_{i,j}^n\right] \\
= \overrightarrow{W}_{i,j}^{n} - a\frac{2\Delta t}{\omega_i\Delta x}\tau_{i,j}^n \Bigg[& \frac{\Delta x}{2}\omega_i\left(\partial_x\overrightarrow{W}|_{i,j}^{n+1-} + \{gh\partial_x z\}_{i,j}^n\right) \\
& + \delta_{i,0}\left(\overrightarrow{W}_{0,j}^{n+1-} - \overrightarrow{W}_{p,j-1}^{n+1-} + \Delta x\left\{gh\partial_x z\right\}_{j-1/2}^n\right)\Bigg],
\end{aligned}$$

and similarly

$$\begin{aligned}
\overleftarrow{W}_{i,j}^{n+1-} = \overleftarrow{W}_{i,j}^{n} + a\frac{2\Delta t}{\omega_i\Delta x}\tau_{i,j}^n \Bigg[& \frac{\Delta x}{2}\omega_i\left(\partial_x\overleftarrow{W}|_{i,j}^{n+1-} + \{gh\partial_x z\}_{i,j}^n\right) \\
& + \delta_{i,p}\left(\overleftarrow{W}_{0,j+1}^{n+1-} - \overleftarrow{W}_{p,j}^{n+1-} + \Delta x\left\{gh\partial_x z\right\}_{j+1/2}^n\right)\Bigg].
\end{aligned}$$

We then multiply the first equation by $\overrightarrow{W}_{i,j}^{n+1-}$ and the second one by $\overleftarrow{W}_{i,j}^{n+1-}$ to obtain

$$\begin{aligned}
\overrightarrow{W}_{i,j}^{n+1-}\left(\overrightarrow{W}_{i,j}^{n+1-} - \overrightarrow{W}_{i,j}^{n}\right) = -a\frac{2\Delta t}{\omega_i\Delta x}\tau_{i,j}^n \Bigg[& \frac{\Delta x}{2}\omega_i\left(\partial_x\frac{\overrightarrow{W}^2}{2}|_{i,j}^{n+1-} + \overrightarrow{W}_{i,j}^{n+1-}\{gh\partial_x z\}_{i,j}^n\right) \\
& + \delta_{i,0}\overrightarrow{W}_{0,j}^{n+1-}\left(\overrightarrow{W}_{0,j}^{n+1-} - \overrightarrow{W}_{p,j-1}^{n+1-} + \Delta x\left\{gh\partial_x z\right\}_{j-1/2}^n\right)\Bigg],
\end{aligned}$$

and

$$\begin{aligned}
\overleftarrow{W}_{i,j}^{n+1-}\left(\overleftarrow{W}_{i,j}^{n+1-} - \overleftarrow{W}_{i,j}^{n}\right) = a\frac{2\Delta t}{\omega_i\Delta x}\tau_{i,j}^n \Bigg[& \frac{\Delta x}{2}\omega_i\left(\partial_x\frac{\overleftarrow{W}^2}{2}|_{i,j}^{n+1-} + \overleftarrow{W}_{i,j}^{n+1-}\{gh\partial_x z\}_{i,j}^n\right) \\
& + \delta_{i,p}\overleftarrow{W}_{p,j}^{n+1-}\left(\overleftarrow{W}_{0,j+1}^{n+1-} - \overleftarrow{W}_{p,j}^{n+1-} + \Delta x\left\{gh\partial_x z\right\}_{j+1/2}^n\right)\Bigg].
\end{aligned}$$

Using the identities $2b(b-a) = (b^2-a^2) + (a-b)^2$ and $2a(b-a) = (b^2-a^2) - (b-a)^2$, we easily get

$$(\overrightarrow{W}_{i,j}^{n+1-})^2 - (\overrightarrow{W}_{i,j}^{n})^2 + a\frac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \left[\frac{\Delta x}{2}\omega_i \partial_x \overrightarrow{W}^2|_{i,j}^{n+1-} + \delta_{i,0}\left((\overrightarrow{W}_{0,j}^{n+1-})^2 - (\overrightarrow{W}_{p,j-1}^{n+1-})^2\right)\right]$$
$$\leq -2a\frac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \left[\frac{\Delta x}{2}\omega_i \overrightarrow{W}_{i,j}^{n+1-}\{gh\partial_x z\}_{i,j}^n + \delta_{i,0}\overrightarrow{W}_{0,j}^{n+1-}\Delta x\,\{gh\partial_x z\}_{j-1/2}^n\right],$$

and

$$(\overleftarrow{W}_{i,j}^{n+1-})^2 - (\overleftarrow{W}_{i,j}^{n})^2 - a\frac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \left[\frac{\Delta x}{2}\omega_i \partial_x \overleftarrow{W}^2|_{i,j}^{n+1-} + \delta_{i,p}\left((\overleftarrow{W}_{0,j+1}^{n+1-})^2 - (\overleftarrow{W}_{p,j}^{n+1-})^2\right)\right]$$
$$\leq 2a\frac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \left[\frac{\Delta x}{2}\omega_i \overleftarrow{W}_{i,j}^{n+1-}\{gh\partial_x z\}_{i,j}^n + \delta_{i,p}\overleftarrow{W}_{p,j}^{n+1-}\Delta x\,\{gh\partial_x z\}_{j+1/2}^n\right].$$

Note now that

$$\frac{\overrightarrow{W}^2 + \overleftarrow{W}^2}{2} = \Pi^2 + a^2u^2 \quad \text{and} \quad \frac{\overrightarrow{W}^2 - \overleftarrow{W}^2}{2} = 2au\Pi.$$

Therefore, setting

$$\eta = \frac{\overrightarrow{W}^2 + \overleftarrow{W}^2}{2}$$

and summing the two inequalities above leads to

$$\eta_{i,j}^{n+1-} - \eta_{i,j}^n + 2a^2\frac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \Big[\frac{\Delta x}{2}\omega_i \partial_x(\Pi u)|_{i,j}^{n+1-} - \frac{\delta_{i,p}}{4a}\left((\overleftarrow{W}_{0,j+1}^{n+1-})^2 - (\overleftarrow{W}_{p,j}^{n+1-})^2\right)$$
$$+ \frac{\delta_{i,0}}{4a}\left((\overrightarrow{W}_{0,j}^{n+1-})^2 + (\overrightarrow{W}_{p,j-1}^{n+1-})^2\right)\Big]$$
$$\leq -2a^2\frac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \Big[\frac{\Delta x}{2}\omega_i u_{i,j}^{n+1-}\{gh\partial_x z\}_{i,j}^n - \frac{\delta_{i,p}}{2a}\overleftarrow{W}_{p,j}^{n+1-}\Delta x\,\{gh\partial_x z\}_{j+1/2}^n$$
$$+ \frac{\delta_{i,0}}{2a}\overrightarrow{W}_{0,j}^{n+1-}\Delta x\,\{gh\partial_x z\}_{j-1/2}^n\Big]. \quad (3.A.1)$$

With a little abuse in the notations, let us now consider the internal energy e and the pressure p as functions of $\tau = 1/h$, so that $e(\tau) = g/2\tau$ and $e'(\tau) = -p(\tau)$, while the total energy E is still given by $E = u^2/2 + e$. Since

$$E - \frac{\eta}{2a^2} = e + \frac{\Pi^2}{2a^2}$$

and

$$\Pi_{i,j}^{n+1-} - \Pi_{i,j}^n = -a^2(\tau_{i,j}^{n+1-} - \tau_{i,j}^n)$$

by system (3.3.4), we thus have

$$E_{i,j}^{n+1-} - E_{i,j}^n - \frac{\eta_{i,j}^{n+1-} - \eta_{i,j}^n}{2a^2} = e(\tau_{i,j}^{n+1-}) - e(\tau_{i,j}^n) - \frac{(\Pi_{i,j}^{n+1-} - \Pi_{i,j}^n)^2}{2a^2} - \frac{\Pi_{i,j}^n(\Pi_{i,j}^{n+1-} - \Pi_{i,j}^n)}{a^2}$$
$$= e(\tau_{i,j}^{n+1-}) - e(\tau_{i,j}^n) + \Pi_{i,j}^n(\tau_{i,j}^{n+1-} - \tau_{i,j}^n) - \frac{a^2}{2}(\tau_{i,j}^{n+1-} - \tau_{i,j}^n)^2$$
$$= \frac{(e''(\xi) - a^2)}{2}(\tau_{i,j}^{n+1-} - \tau_{i,j}^n)^2$$

for some ξ in between $\tau_{i,j}^n$ and $\tau_{i,j}^{n+1-}$. Note that $e''(\xi) = -p'(\xi)$ so that under the sub-characteristic condition

$$a > \max_j \max_{\tau \in \mathcal{I}(\tau_{i,j}^n, \tau_{i,j}^{n+1-})} \sqrt{-p'(\tau)},$$

we have

$$E_{i,j}^{n+1-} - E_{i,j}^n - \frac{\eta_{i,j}^{n+1-} - \eta_{i,j}^n}{2a^2} \leq 0.$$

Multiplying this inequality by $h_{i,j}^n = L_{i,j}^{n+1-} h_{i,j}^{n+1-}$, we get

$$(hE)_{i,j}^{n+1-} - (hE)_{i,j}^n + \left(L_{i,j}^{n+1-} - 1\right)(hE)_{i,j}^{n+1-} - h_{i,j}^n \frac{\eta_{i,j}^{n+1-} - \eta_{i,j}^n}{2a^2} \leq 0.$$

We are now ready to establish the energy mean value inequality. Multiplying the last inequality by $\omega_i/2$, summing over i and using inequality (3.A.1) leads to

$$\begin{aligned}\overline{(hE)}_j^{n+1-} - \overline{(hE)}_j^n + \sum_{i=0}^p \frac{\omega_i}{2}\left(L_{i,j}^{n+1-} - 1\right)(hE)_{i,j}^{n+1-} + \frac{\Delta t}{\Delta x}\Bigg[\int_{\kappa_j} \partial_x(\Pi u)(x, t^{n+1-})dx \\ - \frac{1}{4a}\left((\overleftarrow{W}_{0,j+1}^{n+1-})^2 - (\overleftarrow{W}_{p,j}^{n+1-})^2\right) + \frac{1}{4a}\left((\overrightarrow{W}_{0,j}^{n+1-})^2 - (\overrightarrow{W}_{p,j-1}^{n+1-})^2\right)\Bigg] \\ \leq -\Delta t\ \{ghu\partial_x z\}_j^{n+1-},\end{aligned}$$

where we have set

$$\{ghu\partial_x z\}_j^{n+1-} = \sum_{i=0}^p \frac{\omega_i}{2} u_{i,j}^{n+1-}\{gh\partial_x z\}_{i,j}^n - \frac{1}{2a}\overleftarrow{W}_{p,j}^{n+1-}\{gh\partial_x z\}_{j+1/2}^n + \frac{1}{2a}\overrightarrow{W}_{0,j}^{n+1-}\{gh\partial_x z\}_{j-1/2}^n.$$

Finally, since

$$(\Pi u)_{0,j}^{n+1-} = \frac{(\overrightarrow{W}_{0,j}^{n+1-})^2 - (\overleftarrow{W}_{0,j}^{n+1-})^2}{4a}, \quad (\Pi u)_{p,j}^{n+1-} = \frac{(\overrightarrow{W}_{p,j}^{n+1-})^2 - (\overleftarrow{W}_{p,j}^{n+1-})^2}{4a},$$

and

$$\frac{(\overrightarrow{W}_{p,j}^{n+1-})^2 - (\overleftarrow{W}_{0,j+1}^{n+1-})^2}{4a} = \pi_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,n+1-},$$

one easily gets

$$\begin{aligned}\overline{(hE)}_j^{n+1-} - \overline{(hE)}_j^n + \frac{\Delta t}{\Delta x}\left[\pi_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,n+1-} - \pi_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,n+1-}\right] + \sum_{i=0}^p \frac{\omega_i}{2}\left(L_{i,j}^{n+1-} - 1\right)(hE)_{i,j}^{n+1-} \\ \leq -\Delta t\ \{ghu\partial_x z\}_j^{n+1-}. \quad (3.A.2)\end{aligned}$$

3.A.2 The transport step

It has already been shown that under the CFL condition (3.4.1), $\overline{X}_j^{n+1}$ is a convex combination of $X_{i,j}^{n+1-}$, $X_{p,j-1}^{n+1-}$ and $X_{0,j+1}^{n+1-}$ for $X = h, hu$. Since the function $(h, hu) \mapsto (hE)(h, hu)$

is a convex function, the Jensen inequality implies

$$(hE)(\overline{\boldsymbol{U}}_j^{n+1}) \leq \sum_{i=0}^{p} \left(\frac{w_i}{2} - \frac{\Delta t}{\Delta x} \left[\int_{\kappa_j} u^\alpha \partial_x \phi_{i,j} - \delta_{i,p} \left(u_{j+1/2}^{*,\alpha} \right)_- + \delta_{i,0} \left(u_{j-1/2}^{*,\alpha} \right)_+ \right] \right) (hE)_{i,j}^{n+1-} \\ - \delta_{i,p} \frac{\Delta t}{\Delta x} \left(u_{j+1/2}^{*,\alpha} \right)_- (hE)_{0,j+1}^{n+1-} + \delta_{i,0} \frac{\Delta t}{\Delta x} \left(u_{j-1/2}^{*,\alpha} \right)_+ (hE)_{p,j-1}^{n+1-}.$$

We can rewrite this inequality as follows,

$$(hE)(\overline{\boldsymbol{U}}_j^{n+1}) - \overline{(hE)}_j^{n+1-} + \frac{\Delta t}{\Delta x} \left[(hE)_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,n+1-} - (hE)_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,n+1-} \right] \\ \leq \sum_{i=0}^{p} \frac{\omega_i}{2} \left(L_{i,j}^{n+1-} - 1 \right) (hE)_{i,j}^{n+1-}, \quad (3.A.3)$$

where we have set

$$(hE)_{j+1/2}^{n+1-} = \begin{cases} (hE)_{p,j}^{n+1-}, & \text{if } u_{j+1/2}^* \geq 0, \\ (hE)_{0,j+1}^{n+1-}, & \text{otherwise.} \end{cases}$$

Finally, combining inequalities (3.A.2) and (3.A.3) we obtain the expected entropy inequality

$$(hE)(\overline{\boldsymbol{U}}_j^{n+1}) - \overline{(hE)}_j^{n} \\ + \frac{\Delta t}{\Delta x} \left[\left(\pi_{j+1/2}^{*,n+1-} + (hE)_{j+1/2}^{*,n+1-} \right) u_{j+1/2}^{*,n+1-} - \left(\pi_{j-1/2}^{*,n+1-} + (hE)_{j-1/2}^{*,n+1-} \right) u_{j-1/2}^{*,n+1-} \right] \\ \leq -\Delta t \, \{ghu\partial_x z\}_j^{n+1-}.$$

3.B A high-order discontinuous Galerkin Lagrange - projection scheme for the barotropic Euler equations

Abstract

This work considers the barotropic Euler equations and proposes a high-order conservative scheme based on a Lagrange-projection decomposition. The high-order in space and time are achieved using discontinuous Galerkin (DG) and Runge-Kutta (RK) strategies. The use of a Lagrange-projection decomposition enables the use of time steps that are not constrained by the sound speed thanks to an implicit treatment of the acoustic waves (Lagrange step), while the transport waves (projection step) are treated explicitly. We compare our DG discretization with the recent one [13] and state that it satisfies important non linear stability properties. The behaviour of our scheme is illustrated by several test cases.

3.B.1 Introduction

We are interested in the gas dynamics equations in Eulerian coordinates

$$\begin{cases} \partial_t \rho + \partial_x (\rho u) = 0, \\ \partial_t (\rho u) + \partial_x \left(\rho u^2 + p \right) = 0, \end{cases} \quad (3.B.1)$$

where $\rho > 0$ is the density, u the velocity and $p = p(\rho)$ is the pressure of the fluid such that $p'(\rho) > 0$. In the numerical experiments, we will choose $p(\rho) = g\rho^2/2$ where $g > 0$ is the gravity constant so that the model can also be understood as the shallow water equations with flat topography (in this case, ρ stands for the water depth). The unknowns depend on the space and time variables x and t, with $x \in \mathbb{R}$ and $t \in [0, \infty)$. At time $t = 0$, the model is supplemented with a given initial data $\rho(x, t = 0) = \rho_0(x)$ and $u(x, t = 0) = u_0(x)$.

The aim of this paper is to propose a high-order discretization based on a Lagrange-projection decomposition of the governing equations and using a discontinuous Galerkin (DG) [9, 16] strategy for the space variable.

The Lagrange-projection (or equivalently Lagrange-remap) decomposition is interesting since it allows to naturally decouple the acoustic and transport terms of the model. It proved to be useful and very efficient when considering subsonic or low-Mach number flows. In this case, the CFL restriction of Godunov-type schemes is driven by the acoustic waves and can be very restrictive. The Lagrange-projection strategy allows for a very natural implicit-explicit scheme with a CFL restriction based on the (slow) transport waves and not on the (fast) acoustic waves. We refer for instance the reader to [10, 4, 5], to the recent contribution [7], and to the references therein. Note that the later contribution considers the shallow water equations with non flat topography and that the corresponding (implicit-explicit) Lagrange-projection scheme is well-balanced but only first-order accurate. It is the purpose of this contribution to extend the first-order Lagrange-projection schemes of the above references to high-order of accuracy in both space and time. The proposed approach is quite close to the one recently developed in [13], but as we will see, the corresponding projection step turns out to be different.

3.B.2 Lagrange-projection decomposition and finite-volume scheme

In this section, we briefly present the Lagrange-projection decomposition considered in this paper and the corresponding first-order finite volume scheme.

Operator splitting decomposition and relaxation approximation

Using the chain rule for the space derivatives of (3.B.1), the Lagrange-projection decomposition consists in first solving

$$\begin{cases} \partial_t \rho + \rho \partial_x u = 0, \\ \partial_t(\rho u) + \rho u \partial_x u + \partial_x p = 0, \end{cases} \tag{3.B.2}$$

which gives in Lagrangian coordinates $\tau \partial_x = \partial_m$, with $\tau = 1/\rho$,

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m p = 0, \end{cases} \tag{3.B.3}$$

and then the transport system

$$\begin{cases} \partial_t \rho + u \partial_x \rho = 0, \\ \partial_t(\rho u) + u \partial_x(\rho u) = 0. \end{cases} \tag{3.B.4}$$

The numerical approximation of systems (3.B.3) and (3.B.4) will be given in the next sections but let us notice from now on that the Lagrangian system (3.B.3) will be treated

considering the following relaxation approximation [12, 15],

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m \Pi = 0, \\ \partial_t \Pi + a^2 \partial_m u = \lambda \left(p - \Pi\right). \end{cases} \tag{3.B.5}$$

Here, the new variable Π represents a linearization of the real pressure p, the constant parameter a is a linearization of the Lagrangian sound speed ρc such that the sub-characteristic condition $a > \rho c$, $c = \sqrt{p'(\rho)}$, is satisfied, and the relaxation parameter λ allows to recover $\Pi = p$ and the original system (3.B.3) in the asymptotic regime $\lambda \to \infty$. As usual, the relaxation system will be solved using a splitting strategy which consists in first setting $\Pi = p$ at initial time (which is formally equivalent to considering $\lambda \to \infty$ in system (3.B.5)), and then solving the relaxation system (3.B.5) with $\lambda = 0$.

First-order numerical scheme

The first-order finite volume scheme associated with the above decomposition and relaxation approximation is classical and given for instance in [5]. Nevertheless, it will be recovered in the DG extension proposed in the next section by setting the degree of all polynomials p to 0. Space and time will be discretized using a space step Δx and a time step Δt. We will consider a set of cells $\kappa_j = [x_{j-1/2}, x_{j+1/2})$ and instants $t^n = n\Delta t$, where $x_{j+1/2} = j\Delta x$ and $x_j = (x_{j-1/2} + x_{j+1/2})/2$ are respectively the cell interfaces and cell centers, for $j \in \mathbb{Z}$ and $n \in \mathbb{N}$.

3.B.3 Discontinuous Galerkin discretization

We begin this section by introducing the notations of the DG discretization. Recall that the DG approach considers that the approximate solution at each time t^n is defined on each cell κ_j by a polynomial in space of order less or equal than p for a given integer $p \geq 1$ ($p = 0$ corresponds to the usual first-order and piecewise constant finite volume scheme). With this in mind, we consider the $(p+1)$ Lagrange polynomials $\{\ell_i\}_{i=0,\dots,p}$ associated with the Gauss-Lobatto quadrature points in $[-1, 1]$. More precisely, denoting $-1 = s_0 < s_1 < \cdots < s_p = 1$ the $p+1$ Gauss-Lobatto quadrature points, ℓ_i is defined by the relations $\ell_i(s_k) = \delta_{i,k}$ for $k = 0, ..., p$, where δ is the Kronecker symbol. Then, in each cell κ_j, we define the shifted Lagrange polynomials $\Phi_{i,j}$ by $\Phi_{i,j}(x) = \ell_i\Big(\frac{2}{\Delta x}(x - x_j)\Big)$ and we take $\{\Phi_{i,j}\}_{i=0,\dots,p}$ as a basis for polynomials of order less or equal than p on κ_j. If we denote by $X_{\Delta x}$ the DG approximation of X, we thus have

$$X_{\Delta x}(x,t) = \sum_{k=0}^{p} X_{k,j}(t)\, \Phi_{k,j}(x), \quad \forall x \in \kappa_j,$$

where the coefficients $X_{k,j}$ depend on time and correspond to the value of the numerical solution at the shifted Gauss-Lobatto quadrature points $x_{k,j} = x_j + \frac{\Delta x}{2} s_k$.

Before entering the details of the numerical approximation, let us briefly recall that the Gauss-Lobatto quadrature formula for $f : \kappa_j \times \mathbb{R}^+ \to \mathbb{R}$ writes

$$\int_{\kappa_j} f(x,t)\,\mathrm{d}x \approx \frac{\Delta x}{2} \sum_{k=0}^{p} \omega_k f(x_{k,j}, t),$$

where ω_k are the weights of the Gauss-Lobatto quadrature. It is well-known that this formula is exact as soon as f is a polynomial of order less or equal than $(2p-1)$ with respect to x on κ_j. Just note that the integral

$$\int_{\kappa_j} \Phi_{i,j}(x)\Phi_{k,j}(x)\,\mathrm{d}x$$

will be therefore approximated by $\frac{\Delta x}{2}\omega_i\delta_{i,k}$ in the following. At last, note that the piecewise constant case $p=0$ can be also considered in this framework provided that we set $s_0=0$, $\Phi_{0,j}=1$ and $\omega_0=2$.

Time discretization $(t^n \to t^{n+1})$

We begin with the acoustic step (3.B.5) with $\lambda=0$. Multiplying the three equations by $\Phi_{i,j}$, integrating over κ_j, and considering the piecewise polynomial approximations $X_{\Delta x}$ for $X=\tau,u,\Pi$ easily leads to

$$\begin{cases} \dfrac{\Delta x}{2}\omega_i\partial_t\tau_{i,j}(t) - \displaystyle\int_{\kappa_j}\Phi_{i,j}(x)\partial_m u(x,t)\,\mathrm{d}x = 0, \\ \dfrac{\Delta x}{2}\omega_i\partial_t u_{i,j}(t) + \displaystyle\int_{\kappa_j}\Phi_{i,j}(x)\partial_m \Pi(x,t)\,\mathrm{d}x = 0, \\ \dfrac{\Delta x}{2}\omega_i\partial_t \Pi_{i,j}(t) + a^2\displaystyle\int_{\kappa_j}\Phi_{i,j}(x)\partial_m u(x,t)\,\mathrm{d}x = 0, \end{cases}$$

that we discretize in time by

$$\begin{cases} \tau_{i,j}^{n+1^-} = \tau_{i,j}^n + \dfrac{2\Delta t}{\omega_i\Delta x}\displaystyle\int_{\kappa_j}\Phi_{i,j}(x)\partial_m u(x,t^\alpha)\,\mathrm{d}x, \\ u_{i,j}^{n+1^-} = u_{i,j}^n - \dfrac{2\Delta t}{\omega_i\Delta x}\displaystyle\int_{\kappa_j}\Phi_{i,j}(x)\partial_m \Pi(x,t^\alpha)\,\mathrm{d}x, \\ \Pi_{i,j}^{n+1^-} = \Pi_{i,j}^n - a^2\dfrac{2\Delta t}{\omega_i\Delta x}\displaystyle\int_{\kappa_j}\Phi_{i,j}(x)\partial_m u(x,t^\alpha)\,\mathrm{d}x, \end{cases} \tag{3.B.6}$$

where the superscript $n+1^-$ formally represents the fictitious time t^{n+1^-}, and $\alpha=n$ or $\alpha=n+1^-$ if the time discretization is taken to be explicit or implicit.

As far as the transport step is concerned, the same process of reasoning leads to

$$\begin{cases} \rho_{i,j}^{n+1} = \rho_{i,j}^{n+1^-} - \dfrac{2\Delta t}{\omega_i\Delta x}\displaystyle\int_{\kappa_j}\Phi_{i,j}(x)u(x,t^\alpha)\partial_x\rho(x,t^{n+1^-})\,\mathrm{d}x, \\ (\rho u)_{i,j}^{n+1} = (\rho u)_{i,j}^{n+1^-} - \dfrac{2\Delta t}{\omega_i\Delta x}\displaystyle\int_{\kappa_j}\Phi_{i,j}(x)u(x,t^\alpha)\partial_x(\rho u)(x,t^{n+1^-})\,\mathrm{d}x. \end{cases} \tag{3.B.7}$$

Note that this transport step is always treated explicitly in time.

Volume integrals and flux calculations

Considering the acoustic step, we aim at approximating the integrals

$$\int_{\kappa_j}\Phi_{i,j}(x)\partial_m X(x,t^\alpha)\,\mathrm{d}x,$$

with $X = u, \Pi$. Observe that

$$\int_{\kappa_j} \Phi_{i,j}(x)\partial_m X(x,t^\alpha)\,\mathrm{d}x \approx \frac{\Delta x}{2}\omega_i \tau_{i,j}^n \partial_x X(x_{i,j},t^\alpha)\,\mathrm{d}x = \tau_{i,j}^n \int_{\kappa_j} \Phi_{i,j}(x)\partial_x X(x,t^\alpha)\,\mathrm{d}x,$$

the last equality is indeed exact since X and Φ are polynomials of order less or equal than p, so that $\Phi_{i,j}\partial_x X(\cdot,t)$ is of order less or equal than $(2p-1)$. The objective is now to use one integration by part to move the derivative from X to Φ, and to use the numerical fluxes to evaluate the interfacial terms, which gives

$$\int_{\kappa_j} \Phi_{i,j}(x)\partial_x X(x,t^\alpha)\,\mathrm{d}x \approx \delta_{i,p} X_{j+1/2}^{*,\alpha} - \delta_{i,0} X_{j-1/2}^{*,\alpha} - \frac{\Delta x}{2}\sum_{k=0}^{p} \omega_k X_{k,j}^\alpha \partial_x \Phi_{i,j}(x_{k,j}).$$

Again, we refer the reader to [5] for the expressions of the star quantities in the above formula and the following ones, which are nothing but the numerical fluxes of the first-order finite volume scheme. At last, from system (3.B.6) we obtain the acoustic step

$$\begin{cases} \tau_{i,j}^{n+1^-} = \tau_{i,j}^n + \dfrac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \left[\delta_{i,p} u_{j+1/2}^{*,\alpha} - \delta_{i,0} u_{j-1/2}^{*,\alpha} - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} \omega_k u_{k,j}^\alpha \partial_x \Phi_{i,j}(x_{k,j})\right] \\ \qquad\quad = L_{i,j}^\alpha \tau_{i,j}^n, \\ u_{i,j}^{n+1^-} = u_{i,j}^n - \dfrac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \left[\delta_{i,p} \Pi_{j+1/2}^{*,\alpha} - \delta_{i,0} \Pi_{j-1/2}^{*,\alpha} - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} \omega_k \Pi_{k,j}^\alpha \partial_x \Phi_{i,j}(x_{k,j})\right], \\ \Pi_{i,j}^{n+1^-} = \Pi_{i,j}^n - a^2 \dfrac{2\Delta t}{\omega_i \Delta x}\tau_{i,j}^n \left[\delta_{i,p} u_{j+1/2}^{*,\alpha} - \delta_{i,0} u_{j-1/2}^{*,\alpha} - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} \omega_k u_{k,j}^\alpha \partial_x \Phi_{i,j}(x_{k,j})\right], \end{cases} \tag{3.B.8}$$

with $L_{i,j}^\alpha = 1 + \frac{2\Delta t}{\omega_i \Delta x}\left[\delta_{i,p} u_{j+1/2}^{*,\alpha} - \delta_{i,0} u_{j-1/2}^{*,\alpha} - \frac{\Delta x}{2}\sum_{k=0}^{p} \omega_k u_{k,j}^\alpha \partial_x \Phi_{i,j}(x_{k,j})\right]$.

Regarding the transport step, we want to evaluate the integrals

$$\int_{\kappa_j} \Phi_{i,j}(x) u(x,t^\alpha)\partial_x X(x,t^{n+1^-})\,\mathrm{d}x$$

with $X = \rho, \rho u$. The same process as before leads to

$$\begin{aligned} \int_{\kappa_j} \Phi_{i,j}(x) u(x,t^\alpha)\partial_x X(x,t^{n+1^-})\,\mathrm{d}x &= \delta_{i,p} X_{j+1/2}^{*,n+1^-} u_{j+1/2}^{*,\alpha} - \delta_{i,0} X_{j-1/2}^{*,n+1^-} u_{j-1/2}^{*,\alpha} \\ &\quad - \int_{\kappa_j} (Xu)\partial_x \Phi_{i,j}\,\mathrm{d}x - X_{i,j}^{n+1^-} \int_{\kappa_j} \Phi_{i,j}(x)\partial_x u(x,t^\alpha)\,\mathrm{d}x, \end{aligned}$$

where we take

$$\int_{\kappa_j} \Phi_{i,j}\partial_x u(x,t^\alpha)\,\mathrm{d}x = \delta_{i,p} u_{j+1/2}^{*,\alpha} - \delta_{i,0} u_{j-1/2}^{*,\alpha} - \frac{\Delta x}{2}\sum_{k=0}^{p} \omega_k u_{k,j}^\alpha \partial_x \Phi_{i,j}(x_{k,j})$$

and

$$\int_{\kappa_j} (Xu)\partial_x \Phi_{i,j}\,\mathrm{d}x \approx \frac{\Delta x}{2}\sum_{k=0}^{p} \omega_k X_{k,j}^{n+1^-} u_{k,j}^\alpha \partial_x \Phi_{i,j}(x_{k,j}).$$

Conservativity property and mean values

Easy calculations not reported here show that the whole Lagrange-projection scheme can be written as follows

$$\begin{cases} \rho_{i,j}^{n+1} = \rho_{i,j}^{n} - \dfrac{2\Delta t}{\omega_i \Delta x}\Bigg[\delta_{i,p}\rho_{j+1/2}^{*,n+1^-} u_{j+1/2}^{*,\alpha} - \delta_{i,0}\rho_{j-1/2}^{*,n+1^-} u_{j-1/2}^{*,\alpha} \\ \qquad\qquad - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} \omega_k \rho_{k,j}^{n+1^-} u_{k,j}^{\alpha} \partial_x \Phi_{i,j}(x_{k,j})\Bigg], \\ (\rho u)_{i,j}^{n+1} = (\rho u)_{i,j}^{n} - \dfrac{2\Delta t}{\omega_i \Delta x}\left[\delta_{i,p}\Pi_{j+1/2}^{*,\alpha} - \delta_{i,0}\Pi_{j-1/2}^{*,\alpha} - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} \omega_k \Pi_{k,j}^{n+1^-} \partial_x \Phi_{i,j}(x_{k,j})\right] \\ \qquad - \dfrac{2\Delta t}{\omega_i \Delta x}\Bigg[\delta_{i,p}(\rho u)_{j+1/2}^{*,n+1^-} u_{j+1/2}^{*,\alpha} - \delta_{i,0}(\rho u)_{j-1/2}^{*,n+1^-} u_{j-1/2}^{*,\alpha} \\ \qquad\qquad - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} \omega_k (\rho u)_{k,j}^{n+1^-} u_{k,j}^{\alpha} \partial_x \Phi_{i,j}(x_{k,j})\Bigg], \end{cases}$$

while the mean values

$$\overline{X}_j^{n+1} = \frac{1}{\Delta x}\int_{\kappa_j} X(x, t^{n+1})\,\mathrm{d}x = \sum_{i=0}^{p} \frac{\omega_i}{2} X_{i,j}^{n+1}$$

with $X = \rho, \rho u$ obey the conservative formulas

$$\begin{cases} \overline{\rho}_j^{n+1} = \overline{\rho}_j^{n} - \dfrac{\Delta t}{\Delta x}\left[\rho_{j+1/2}^{*,n+1^-} u_{j+1/2}^{*,\alpha} - \rho_{j-1/2}^{*,n+1^-} u_{j-1/2}^{*,\alpha}\right], \\ \overline{(\rho u)}_j^{n+1} = \overline{(\rho u)}_j^{n} - \dfrac{\Delta t}{\Delta x}\Big[\Pi_{j+1/2}^{*,\alpha} + (\rho u)_{j+1/2}^{*,n+1^-} u_{j+1/2}^{*,\alpha} \\ \qquad\qquad - \Pi_{j-1/2}^{*,\alpha} - (\rho u)_{j-1/2}^{*,n+1^-} u_{j-1/2}^{*,\alpha}\Big]. \end{cases} \tag{3.B.9}$$

Additional nonlinear stability properties can be proved for both the implicit and explicit schemes ($\alpha = n$ and $\alpha = n+1^-$). In particular, we have been able to prove the positivity of the nodal densities $\rho_{i,j}^{n+1^-}$ at time t^{n+1^-} and of the mean densities $\overline{\rho}_j^{n+1}$ at time t^{n+1}, but also the validity of a discrete entropy inequality for the mean energy following the same lines as in [13].

Comparison with the double integration by part used in [13]

The present scheme turns out to be very close to the one recently proposed in [13], and it shares the same stability properties. However, the overall process in [13] is based on double integrations by part leading to the use of both numerical and exact fluxes at the interfaces, instead of only numerical fluxes in our approach. Interestingly, we observed that both schemes are strictly equivalent if one considers the mean values, but the nodal values turn out to be different because of the transport step. These little differences are due to the use of quadrature formulas to integrate the polynomials $Xu\partial_x\Phi_{i,j}$. In this case, the numerical integrations are not exact since polynomials $Xu\partial_x\Phi_{i,j}$ are of order $3p-1 > 2p-1$.

Positivity and generalized slope limiters

We have already stated the positivity of the nodal values $\rho_{i,j}^{n+1^-}$ at the end of the acoustic step and of the mean values $\overline{\rho}_j^{n+1}$ at the end of the transport step. Similarly to [13], we suggest to use a positivity limiter to ensure that $\rho_{i,j}^{n+1} > 0$. More precisely, we replace $\rho_{i,j}^{n+1}$ by $\theta_j \rho_{i,j}^{n+1} + (1-\theta_j)\,\overline{\rho}_j^{n+1}$, where the coefficients θ_j are taken to be $\theta_j = \min\left(1, \frac{\overline{\rho}_j^{n+1}-\varepsilon}{\overline{\rho}_j^{n+1}-\min_i \rho_{i,j}^{n+1}}\right)$. This formula ensures that if ρ is less than the threshold ε, the nodal values of the corresponding cell are corrected, using the positive mean value, towards values greater than ε. In general we set the parameter ε to $1.0\mathrm{e}^{-10}$. Note that in the forthcoming numerical experiments, the positivity limiter is not active. In order to avoid non physical oscillations, we also use the generalized slope limiters introduced in [9]. More precisely, considering the *minmod* function

$$m(a,b,c) = \begin{cases} s \cdot \min(|a|,|b|,|c|), & \text{if } s = \operatorname{sign}(a) = \operatorname{sign}(b) = \operatorname{sign}(c), \\ 0, & \text{otherwise,} \end{cases}$$

the increments $\Delta_+ \overline{X}_j^{n+1} = \overline{X}_{j+1}^{n+1} - \overline{X}_j^{n+1}$, $\Delta_- \overline{X}_j^{n+1} = \overline{X}_j^{n+1} - \overline{X}_{j-1}^{n+1}$, and the values

$$\begin{cases} X_{j+1/2}^{-,n+1} = \overline{X}_j^{n+1} + m\Big(X_{p,j}^{n+1} - \overline{X}_j^{n+1}, \Delta_+ \overline{X}_j^{n+1}, \Delta_- \overline{X}_j^{n+1}\Big), \\ X_{j-1/2}^{+,n+1} = \overline{X}_j^{n+1} - m\Big(\overline{X}_j^{n+1} - X_{0,j}^{n+1}, \Delta_+ \overline{X}_j^{n+1}, \Delta_- \overline{X}_j^{n+1}\Big), \end{cases}$$

the new states at time t^{n+1} are defined by

$$\begin{cases} X_{i,j}^{n+1}, & \text{if } X_{j+1/2}^{-,n+1} = X_{p,j}^{n+1} \text{ and } X_{j-1/2}^{+,n+1} = X_{0,j}^{n+1}, \\ \overline{X}_j^{n+1} + \frac{2}{\Delta x}\,(x_{i,j} - x_j) \cdot m\Big(\partial_x X^{n+1}(x_j), \Delta_+ \overline{X}_j^{n+1}, \Delta_- \overline{X}_j^{n+1}\Big), & \text{otherwise.} \end{cases}$$

3.B.4 Numerical results

The aim of this section is to compare our explicit-explicit EXEX_p and implicit-explicit IMEX_p Lagrange-projection schemes, where p refers to the polynomial order of the DG approach. The time integrations are performed using Strong Stability Preserving Runge-Kutta methods described in [9]. Recall that $p(\rho) = g\rho^2/2$ so that the parameter a is chosen locally at each interface according to

$$a_{j+1/2} = \kappa \max\left(\rho_j^n \sqrt{g\rho_j^n}, \rho_{j+1}^n \sqrt{g\rho_{j+1}^n}\right)$$

with $\kappa = 1.01$ and $g = 9.81$. We set $\Delta t = \min(\Delta t_{\mathrm{Lag}}, \Delta t_{\mathrm{Tra}})$ for the EXEX_p schemes and $\Delta t = \Delta t_{\mathrm{Tra}}$ for the IMEX_p schemes where

$$\Delta t_{\mathrm{Lag}} = \frac{\Delta x}{2p+1} \min_j \Big(2 a_{j+1/2} \min(\tau_{p,j}, \tau_{0,j+1})\Big)$$

is the DG time-step restriction associated with the Lagrangian step, while the transport step CFL restriction reads

$$\Delta t_{\mathrm{Tra}} = \Delta x \min_{i,j} \frac{2}{\omega_i} \left(\int_{\kappa_j} u^\alpha \partial_x \Phi_{i,j}\, \mathrm{d}x - \delta_p u_{j+1/2}^{*,\alpha,-} + \delta_0 u_{j-1/2}^{*,\alpha,+} \right).$$

Δx	$p=0$		$p=1$		$p=2$	
	$\|\rho_{\Delta x}-\rho\|_1$	order	$\|\rho_{\Delta x}-\rho\|_1$	order	$\|\rho_{\Delta x}-\rho\|_1$	order
1/512	9.3986E-03	0.9432	1.0196E-05	1.9996	1.3457E-08	2.9907
1/1024	4.7945E-03	0.9710	2.5493E-06	1.9998	1.6849E-09	2.9977
1/2048	2.4217E-03	0.9854	6.3736E-07	1.9999	2.1070E-10	2.9994

Table 3.6 – EOA for the manufactured smooth solution at time $T = 0.5$

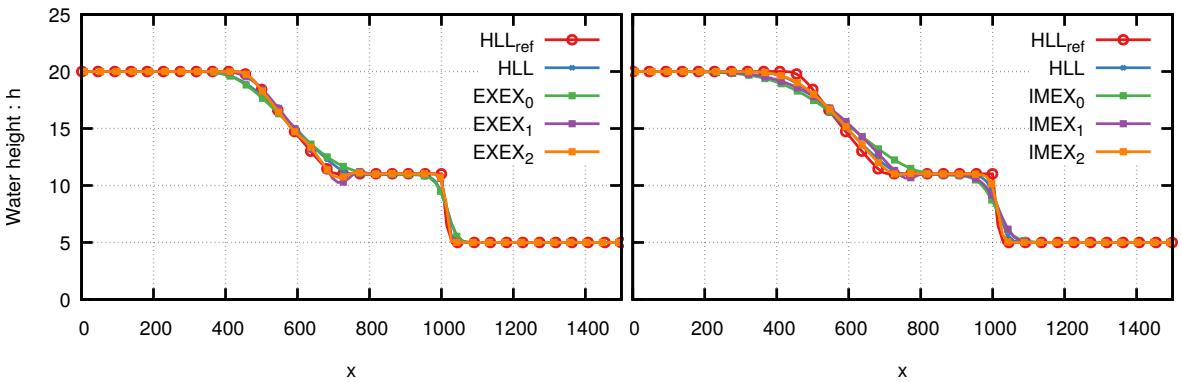

Figure 3.B.1 – Dam Break problem, water height at time $T = 10$, EXEX$_p$ (left), IMEX$_p$ (right)

Manufactured smooth solution

This preliminary test case is taken from [13] and allows us to test the experimental order of accuracy (EOA) of the schemes, especially on the transport step. The space domain is $[0, 1]$, the boundary conditions are periodic and the initial conditions are $\rho_0(x) = 1 + 0.2\sin(2\pi x)$ and $u_0(x) = 1$. We solve system (3.B.1) with a source term such that the exact solution is $\rho(x,t) = 1 + 0.2\sin(2\pi(x-t))$ and $u(x,t) = 1$, which just means that we impose $u_{i,j}^{n+1^-} = 1$ and $\Pi_{i,j}^{n+1^-} = \Pi_{i,j}^n$, so that the Acoustic step is trivial. Note that we use in this special case the EXEX$_p$ schemes. The EOA are reported in Table 3.6.

Dam break problem

In this test case, we take $\rho_0(x) = 20$ if $x \in [0, 750[$, $\rho_0(x) = 10$ if $x \in]750, 1500]$, and $u_0 = 0$ everywhere. The solutions given by the EXEX$_p$ and IMEX$_p$ schemes with $p = 0$, 1 and 2 are shown on Figure 3.B.1 using a 100-cell mesh, and compared with the classical first-order HLL scheme over a 100-cell mesh and a reference 1000-cell refined mesh. Note that the slope limiters allow to reduce spurious oscillations, but there is still a little undershoot for the EXEX$_1$ scheme.

Acknowledgements

The authors are very grateful to P. Kestener, S. Kokh and F. Renac for stimulating dis-

cussions, and the "Maison de la Simulation" for providing excellent working conditions to the second author.

3.C Courbes d'efficacité

Il n'est pas aisé d'obtenir des courbes d'erreur attestant du bon ordre de convergence pour ces schémas DG. En effet, plusieurs problèmes se posent à nous. Tout d'abord il faut un cas test régulier avec des conditions de bords suffisamment simple. On verra dans le Chapitre 5 que le traitement des conditions de bords pour ces schémas à 2 sous pas de temps pose des difficultés. Il n'est pas clair théoriquement que l'on puisse obtenir l'ordre souhaité en temps puisque nous avons recours à une décomposition d'opérateurs d'ordre 1 pour séparer l'opérateur acoustique de celui de transport. Bien que l'on ait réussi à obtenir des ordres de convergence théoriques dans un cas de transport pur, il semble que l'on doive faire converger en temps pour pouvoir obtenir un ordre correspondant à celui attendu en espace. À défaut d'obtenir pour le moment des courbes d'erreur attestant du bon ordre de convergence pour des cas tests réguliers mais non triviaux, on présente ici des courbes d'efficacité montrant l'intérêt d'utiliser des méthodes d'ordres élevés.

Le cas test considéré est celui d'un problème de Riemann, dont on est capable de calculer une solution exacte à chaque instant, ce qui nous permet de calculer précisément l'erreur. On considère une topographie plate et on prend initialement une hauteur d'eau h constante égale à 1. Le domaine physique est le segment $]-1, 1[$ et vitesse u est initialisée de telle sorte que deux ondes de raréfactions se forment :

$$u = \begin{cases} -\sqrt{g}, & \text{si } x < 0, \\ +\sqrt{g}, & \text{si } x \geq 0. \end{cases}$$

On impose des conditions de bords absorbant, mais on arrête la simulation avant que les ondes les atteignent, ils ne jouent de fait aucun rôle. On représente la solution exacte au temps final $t_F = 0.1$ sur la Figure 3.C.1.

On a tracé pour différents schémas le temps de calcul en fonction de l'erreur L1 relative sur la hauteur d'eau h par rapport à la solution exacte sur la Figure 3.C.2. Les schémas ACU sont des schémas Volumes Finis HLL issue de l'article [1], P0 pour l'ordre 1 et P1 pour l'ordre 2. Les schémas EX_k et IM_k sont les schémas DG d'ordre k de type Lagrange-projection respectivement explicite-explicite et implicite-explicite.

Conformément à l'intuition, à erreur fixée, les schémas explicites sont plus rapides que les schémas implicite-explicite pour atteindre cette erreur. À part pour le schéma IM_2 qui n'est pas meilleur que le schéma IM_1, il est plus rapide d'obtenir une erreur donnée lorsque l'on monte en ordre.

Bibliographie

[1] Emmanuel Audusse, Chalons Chalons, and Philippe Ung. A very simple well-balanced positive and entropy-satisfying scheme for the shallow-water equations. *Commun. Math. Sci*, 13(5) :1317–1332, 2015.

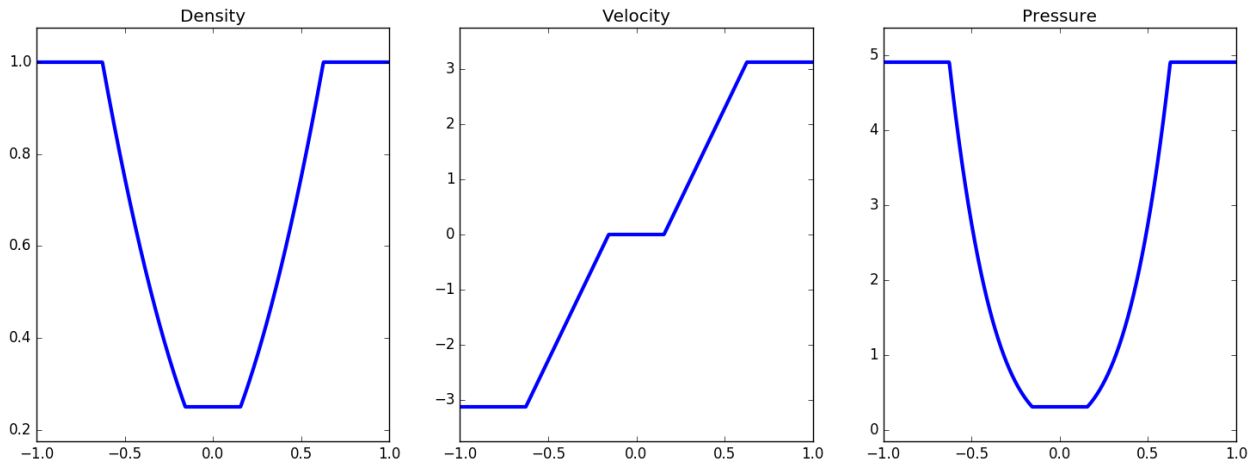

FIGURE 3.C.1 – Solution exacte du problème de Riemann au temps $t_F = 0.1$

[2] François Bouchut. Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources. *Frontiers in Mathematics series, Birkhäuser*, 2004.

[3] Manuel J Castro, Alberto Pardo Milanés, and Carlos Parés. Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique. *Mathematical Models and Methods in Applied Sciences*, 17(12) :2055–2113, 2007.

[4] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. Large time step and asymptotic preserving numerical schemes for the gas dynamics equations with source terms. *SIAM Journal on Scientific Computing*, 35(6) :A2874–A2902, 2013.

[5] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An all-regime Lagrange-projection like scheme for the gas dynamics equations on unstructured meshes. *Communications in Computational Physics*, 20(1) :188–233, 2016.

[6] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An all-regime Lagrange-projection like scheme for 2D homogeneous models for two-phase flows on unstructured meshes. *Journal of Computational Physics*, 335 :885–904, 2017.

[7] Christophe Chalons, Pierre Kestener, Samuel Kokh, and Maxime Stauffert. A large time-step and well-balanced Lagrange-projection type scheme for the shallow water equations. *Communication in Mathematical Sciences*, 15(3) :765–788, 2017.

[8] Christophe Chalons and Maxime Stauffert. A high-order discontinuous Galerkin Lagrange-projection scheme for the barotropic Euler equations. pages 63–70. Springer, 2017.

[9] Bernardo Cockburn and Chi-Wang Shu. Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of scientific computing*, 16(3) :173–261, 2001.

[10] Frédéric Coquel, Quang Nguyen, Marie Postel, and Quang Tran. Entropy-satisfying relaxation method with large time-steps for Euler IBVPs. *Mathematics of Computation*, 79(271) :1493–1533, 2010.

[11] Laurent Gosse. *Computing qualitatively correct approximations of balance laws*, volume 2. Springer, 2013.

[12] Shi Jin and Zhouping Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Communications on pure and applied mathematics*, 48(3) :235–276, 1995.

[13] Florent Renac. A robust high-order Lagrange-projection like scheme with large time steps for the isentropic Euler equations. *Numerische Mathematik*, 135(2) :493–519, 2017.

[14] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of computational physics*, 77(2) :439–471, 1988.

[15] I. Suliciu. On the thermodynamics of fluids with relaxation and phase transitions. Fluids with relaxation. *Internat. J. Engrg. Sci*, 36 :921–947, 1998.

[16] Yulong Xing, Xiangxiong Zhang, and Chi-Wang Shu. Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations. *Advances in Water Resources*, 33(12) :1476–1493, 2010.

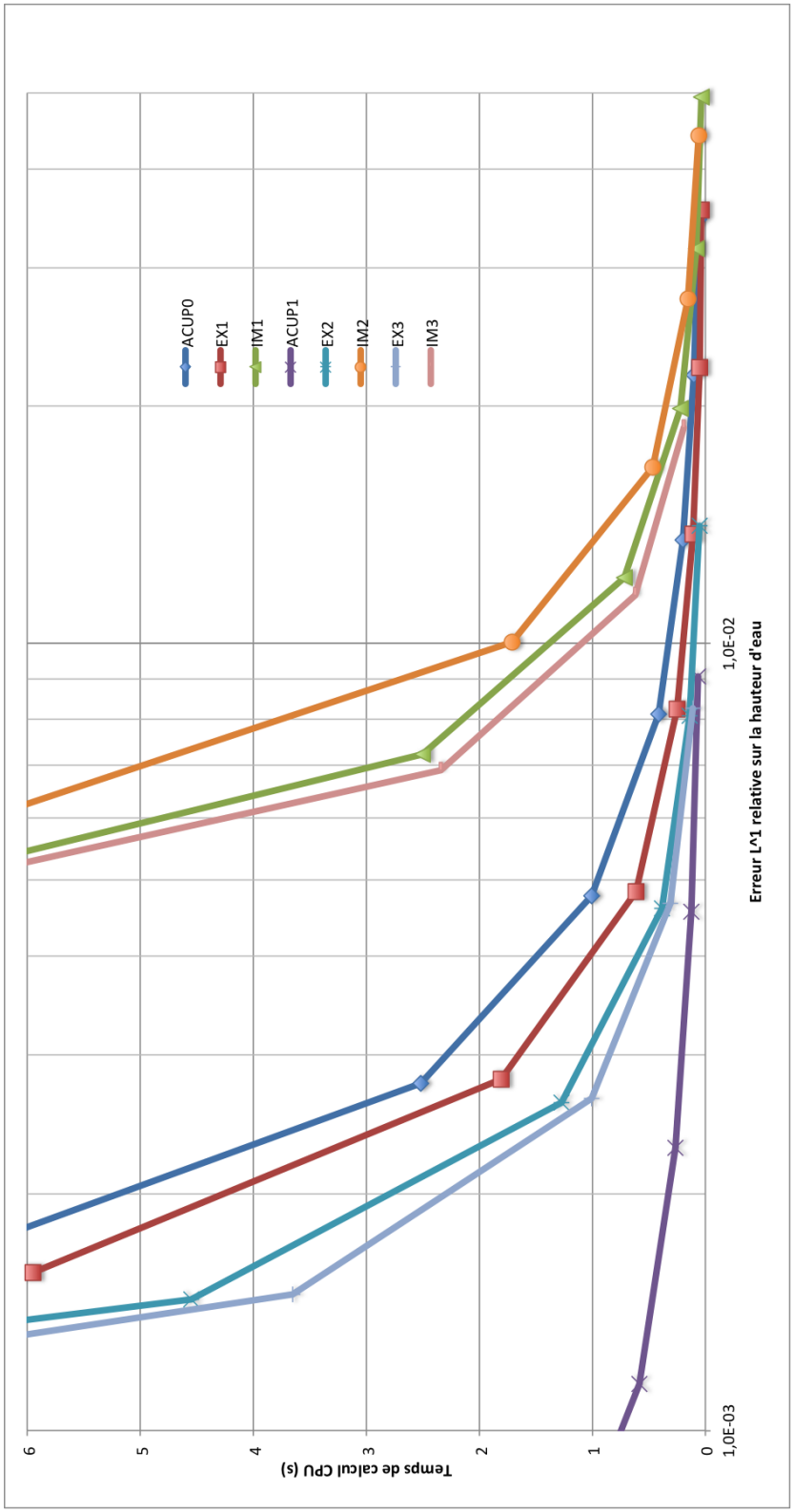

FIGURE 3.C.2 – Courbes d'efficacité pour le cas test du problème de Riemann. Temps de calcul CPU (secondes) en fonction de l'erreur L^1 relative sur la hauteur d'eau h.

Chapitre 4

Un schéma Lagrange-projection Galerkin discontinu équilibre pour les équation de Saint-Venant utilisant une méthode de limitation *a posteriori* de type MOOD

Ce chapitre a été rédigé en collaboration avec Christophe Chalons et Raphaël Loubère à l'Institut Mathématiques de Bordeaux.

A well-balanced discontinuous Galerkin Lagrange-projection scheme for the shallow water equations using Multi-dimensional Optimal Order Detection (MOOD)

4.1 Introduction

We are interested in the shallow water equations (SWE) system in 1D which writes :

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + g\dfrac{h^2}{2}\right) = -gh\partial_x z, \end{cases} \tag{4.1.1}$$

where $(x,t) \in \mathbb{R} \times [0,+\infty) \mapsto h > 0$ is the height of water over a given smooth topography $x \mapsto z > 0$, $(x,t) \mapsto u$ the velocity and g the gravity constant. At the initial time $t = 0$, we assume that the initial conditions over the water height $h(x,0) = h_0(x)$ and the velocity $u(x,0) = u_0(x)$ are given.

The aim of this chapter is to propose a scheme that coupled the acoustic-transport discontinuous Galerkin (DG) methodology, presented in Chapter 3, with the Multi-dimensional Optimal Order Detection strategy, first given in [4]. We first give a brief presentation of the acoustic-transport decomposition we use in the following. Then, we recall the DG scheme, without the use of MOOD *a posteriori* limitation. Afterwards, we present the MOOD technique and its use in this specific framework adding a detection loop to the DG scheme. Finally, we give preliminary results with numerical test cases that support the improvement of this approach.

4.2 Acoustic-transport decomposition

We follow the same methodology as in Section 3.2.1 or, with more precision, the one given in Section 1.2. We write here only the few equations needed to understand the acoustic-transport decomposition used for the DG scheme in the next section.

4.2.1 Operator decomposition

Using the chain rule for the space derivatives of (4.1.1), the Lagrange-projection decomposition consists in first solving

$$\begin{cases} \partial_t h + h\partial_x u = 0, \\ \partial_t(hu) + hu\partial_x u + \partial_x \dfrac{gh^2}{2} = -gh\partial_x z, \end{cases} \tag{4.2.1}$$

which gives in Lagrangian coordinates $\tau\partial_x = \partial_m$, with $\tau = \frac{1}{h}$,

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m \dfrac{g}{2\tau^2} = -\dfrac{g}{\tau}\partial_m z, \end{cases} \tag{4.2.2}$$

and then the transport system

$$\begin{cases} \partial_t h + u\partial_x h = 0, \\ \partial_t(hu) + u\partial_x(hu) = 0. \end{cases} \tag{4.2.3}$$

4.2.2 Relaxation on acoustic step

We use a Suliciu type relaxation approximation, see [10], in order to linearize the pressure term in the acoustic system (4.2.2). Indeed, we introduce a new variable Π and we write the relaxed acoustic system as follow

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m \Pi = -\dfrac{g}{\tau}\partial_m z, \\ \partial_t \Pi + a^2 \partial_m u = \lambda \left(\dfrac{g}{2\tau^2} - \Pi \right), \end{cases} \tag{4.2.4}$$

where λ is the relaxation parameter and $a > 0$ a constant that must satisfy the following Witham sub-characteristic condition: $a > hc := h\sqrt{gh}$.

Finally, we do the classical instantaneous relaxation, which consist in first defining Π at each time step by $\Pi_j^n = g\frac{(h_j^n)^2}{2}$, for all j indexing the cells, and then solving the relaxed system (4.2.4) with $\lambda = 0$.

4.3 DG scheme

We describe in this section the two-step DG scheme. We recall that the two-step process is defined by

1. Update $\mathbb{U}_j^n$ to $\mathbb{U}_j^{n+1-}$ by approximating the solution of (4.2.4),
2. Update $\mathbb{U}_j^{n+1-}$ to $\mathbb{U}_j^{n+1}$ by approximating the solution of (4.2.3).

We use the same discretization and notations as in the Section 3.3. More precisely, if we set $p > 0$ the degree of the polynomials representing the solution in each cell, we can denote

by $-1 = s_0 < s_1 < \cdots < s_p = 1$ the $(p+1)$ Gauss-Lobatto quadrature points and $(\ell_i)_{0\leq i\leq p}$ the associated Lagrange interpolation polynomials such that $\ell_i(s_k) = \delta_{i,k}$, where $\delta_{i,k} = 1$ if $i = k$ and 0 otherwise. We now that the quadrature is such that for all polynomial P of order less than $2p-1$, we have the following equality

$$\int_{-1}^{1} P(s)\,\mathrm{d}s = \sum_{k=0}^{p} w_k P(s_k),$$

where the weights are defined by $w_k = \int_{-1}^{1} \ell_k(s)\,\mathrm{d}s$. Now we can discretize the physical domain by N cells $\omega_j = \left[x_{j-1/2}, x_{j+1/2}\right]$, $0 \leq j < N$, of uniform size Δx, and define the basis polynomials over ω_j as the Lagrange interpolation polynomials of the points $x_{i,j} = x_j + \frac{\Delta x}{2} s_i$, which is to say the polynomials $\Phi_{i,j}(x) = \ell_i\left(\frac{2}{\Delta x}(x - x_j)\right)$, $\forall x \in \omega_j$. Finally, every variable, τ, u, Π for the acoustic system, and $h, (hu)$ for the transport one, are piecewise polynomials of order p in each cell. If we denote by φ one of this variable, $\varphi_j = \sum_{k=0}^{p} \varphi_{k,j}\Phi_{k,j}$ represents the polynomial on the cell ω_j, where $\varphi_{k,j}$ is the value of φ on the point $x_{k,j}$.

4.3.1 DG discretization

Acoustic step

The first step is the acoustic step, which is here explicit and writes

$$\begin{cases} \tau_{i,j}^{n+1-} = \tau_{i,j}^n + \dfrac{2\Delta t}{w_i \Delta x}\tau_{i,j}^n \left[\delta_{i,p} u_{j+1/2}^* - \delta_{i,0} u_{j-1/2}^* - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} w_k u_{k,j}^n \partial_x \Phi_{i,j}(x_{k,j})\right] \\ \qquad\quad = L_{i,j}^n \tau_{i,j}^n, \\ u_{i,j}^{n+1-} = u_{i,j}^n - \dfrac{2\Delta t}{w_i \Delta x}\tau_{i,j}^n \left[\delta_{i,p} \Pi_{j+1/2}^* - \delta_{i,0} \Pi_{j-1/2}^* - \dfrac{\Delta x}{2}\displaystyle\sum_{k=0}^{p} w_k \Pi_{k,j}^n \partial_x \Phi_{i,j}(x_{k,j})\right] \\ \qquad\qquad - \dfrac{2\Delta t}{w_i \Delta x}\tau_{i,j}^n \displaystyle\int_{\Omega_j} \Phi_{i,j}(x) \dfrac{g}{\tau(x,t^n)} \partial_m z(x)\,\mathrm{d}x, \end{cases} \tag{4.3.1}$$

with

$$L_{i,j}^n = 1 + \frac{2\Delta t}{w_i \Delta x}\left[\delta_{i,p} u_{j+1/2}^* - \delta_{i,0} u_{j-1/2}^* - \frac{\Delta x}{2}\sum_{k=0}^{p} w_k u_{k,j}^n \partial_x \Phi_{i,j}(x_{k,j})\right],$$

and the numerical fluxes

$$\begin{aligned} a_{j+1/2} &\geq \max\left(h_{p,j}\sqrt{gh_{p,j}}, h_{0,j+1}\sqrt{gh_{0,j+1}}\right), \\ \{gh\Delta z\}_{j+1/2} &= g\frac{h_{p,j}^n + h_{0,j+1}^n}{2}\left(z_{0,j+1} - z_{p,j}\right), \\ u_{j+1/2}^* &= \frac{u_{p,j}^n + u_{0,j+1}^n}{2} - \frac{1}{2a_{j+1/2}}\left(\Pi_{0,j+1}^n - \Pi_{0,j}^n\right) - \frac{1}{2a_{j+1/2}}\{gh\Delta z\}_{j+1/2}, \\ \Pi_{j+1/2}^* &= \frac{\Pi_{p,j}^n + \Pi_{0,j+1}^n}{2} - \frac{a_{j+1/2}}{2}\left(u_{0,j+1}^n - u_{0,j}^n\right). \end{aligned}$$

Source term treatment

The above volume integral treating the source term is approximated as follow

$$\int_{\Omega_j} \Phi_{i,j}(x) \frac{g}{\tau(x,t^n)} \partial_m z(x)\,\mathrm{d}x = \frac{\delta_{i,p}}{2}\{gh\partial_x z\}_{j+1/2} + \frac{\delta_{i,0}}{2}\{gh\partial_x z\}_{j-1/2} + \frac{w_i \Delta x}{2}\{gh\partial_x z\}^n_{i,j},$$

where

$$\{gh\partial_x z\}^n_{i,j} = gh^n_{i,j}\partial_x(h^n + z)(x_{i,j}) - g\partial_x \Pi^n(x_{i,j}).$$

Transport step

The second step, which is the transport step, writes

$$\begin{aligned}\varphi^{n+1}_{i,j} = L^n_{i,j}\varphi^{n+1-}_{i,j} - \frac{2\Delta t}{\omega_i \Delta x}\Bigg[&\delta_{i,p}\varphi^{*,n+1-}_{j+1/2} u^*_{j+1/2} - \delta_{i,0}\varphi^{*,n+1-}_{j-1/2} u^*_{j-1/2} \\ &- \frac{\Delta x}{2}\sum_{k=0}^{p} \omega_k \varphi^{n+1-}_{k,j} u^n_{k,j} \partial_x \Phi_{i,j}(x_{k,j})\Bigg],\end{aligned} \tag{4.3.2}$$

where $\varphi = h$ or (hu), and the upwind fluxes are

$$\varphi^{*,n+1-}_{j+1/2} = \begin{cases} \varphi^{n+1-}_{p,j}, & \text{if } u^*_{j+1/2} \geq 0, \\ \varphi^{n+1-}_{0,j+1}, & \text{if } u^*_{j+1/2} < 0. \end{cases}$$

4.3.2 Properties

We summarize here the most important properties that we have been able to prove on the DG scheme.

The scheme is conservative for the water height h, in the sense that the mean values satisfy a FV type scheme, given in 3.3.14, which is conservative. It is also conservative for the discharge hu when the bottom is flat.

It also conserves exactly the lake at rest states, which are characterized by $u = 0$ and $H = h + z$ constant over the domain.

Under a particular CFL condition, given in 3.4.1, the mean values at time t^{n+1} of h stay positives if the nodal values at time t^n were positives.

Moreover, under the same CFL condition, the scheme satisfy a weak discrete entropy inequality, given in 3.4.4.

4.3.3 Limiters and limitations

Here we use the same positivity limiter and generalized slope limiter on characteristic variables as used in 3.4.3.

The difficulty with this kind of limiters is that it smooths the solution all over the cell towards the mean value. First it means that we can lose the order of accuracy, and moreover the precision if the cell is large, which is often the case when the order of the polynomials is large, greater than 4 or 5 for instance. It also mixes cell polynomials together, which can increase the

problems described above. Finally, it is not possible with *a priori* limiting techniques to predict some situations for non-linear schemes and non-linear equations, for instance when positivity must be ensured for variables obtained as a non-linear combination of some state variables (pressure in Euler equations).

We recall the difficulty we encountered with this limiter in Chapter 3 with a strong dam break test case. We had represented in Figure 3.5.1 the results for the limiters on both the conservative and the characteristic variables. Even though we had found that the solution was less oscillatory with the limiters applied on the characteristic variables, it seems that we are not able to eliminate the non-physical oscillations in the rarefaction wave. In the next section, we will compare the results with the MOOD *a posteriori* limitation technique after introducing it for this framework.

4.4 *a posteriori* MOOD subcell limiting

4.4.1 General idea

We have seen that the limiting strategy proposed for the DG scheme is not efficient for some test cases. Indeed, the limiters basically reduce the order of the polynomials to 1 on detected cells, which induces a loss of the accuracy of the solution, while they do not seem to be able to discard all spurious oscillations. The solution will also be dissipated when the cell is wrongly detected, which is not the case with the MOOD loop.

The MOOD limiting strategy is an *a posteriori* limiting first introduced in [4] for FV schemes and later studied in the DG framework in [8] for instance. Some improvements of the detectors have been given in [7] and we can refer to [2] and [5] for applications of this methodology respectively to the Euler system with an entropy preserving scheme and to the shallow water equations system we are interested in. The MOOD strategy consists in computing a candidate solution with a high order unlimited DG scheme, and using *troubled cell detectors* over this candidate solution. If some cells discredit the candidate as a good solution, locally to those cells we re-compute a numerical solution using a more *robust scheme*, generally with acknowledging a smaller order of accuracy.

In this work the robust scheme is the first order Finite Volume (FV) Lagrange-Projection type scheme [3], towards which the DG scheme degenerates when the polynomial degree $p = 0$. This scheme ensures important properties such as the preservation of positive water height, the well-balanced property and the discrete entropy inequality. In order to have a reasonable truncation error, we divide each cell into N_s sub-cells and perform the FV scheme onto this sub-cell refined mesh. Doing so the 1st order FV scheme acts on a smaller characteristics length of size $\delta x = \frac{\Delta x}{N_s}$, reducing the truncation error. This strategy also permits to correct only locally within the detected large cells and uses the fact that the time step for the DG scheme is lower than the one for the FV scheme. It has been discussed in [8] that one good choice for the number of sub-cells is $N_s = 2p + 1$ in terms of accuracy and to avoid time subcycling. Indeed with $2p + 1$ sub-cells, the DG timestep on cells matches the FV timestep on sub-cells [8].

4.4.2 Projection and reconstruction

Projection: from cell based DG polynomial to sub-cell based FV data

Starting from the DG polynomial representation defined by $p+1$ components for all x in the cell Ω of size Δx and on a basis $(\Phi_i)_{i=0,\dots,p}$ as

$$\varphi(x) = \sum_{i=0}^{p} \varphi_i \Phi_i(x),$$

we project the polynomial $\varphi = (\varphi_0, \dots, \varphi_p)$ onto a set of $2p+1$ uniform sub-cells ω_k, $k = 0, \dots, 2p$ of size $\delta x = \frac{\Delta x}{2p+1}$, as

$$\widetilde{\varphi}_k = \frac{1}{\delta x} \int_{\omega_k} \varphi(x)\,\mathrm{d}x = \frac{1}{\delta x} \sum_{i=0}^{p} \varphi_i \left(\int_{\omega_k} \Phi_i(x)\,\mathrm{d}x \right).$$

Let us denote the previous integrals of basis functions over sub-cells as $P_{k,i} = \frac{1}{\delta x} \int_{\omega_k} \Phi_i(x)\,\mathrm{d}x$. These coefficients over all sub-cells and basis polynomials form a rectangular $(2p+1) \times (p+1)$ matrix $\mathcal{P} = \{P_{k,i}\}_{0 \le k \le 2p,\, 0 \le i \le p}$. Therefore the vector of piecewise constant sub-cell values $\widetilde{\varphi} = (\widetilde{\varphi}_0, \dots, \widetilde{\varphi}_{2p})$ is obtained by a matrix-vector multiplication:

$$\widetilde{\varphi} = \mathcal{P}\varphi.$$

Note that the projection matrix $\mathcal{P}$ can be precomputed and stored. Moreover the projection is conservative by construction. In other words the mean value of the DG polynomial $\bar{\varphi} = \frac{1}{\Delta x} \int_\Omega \varphi(x)\,\mathrm{d}x$ is preserved on sub-cells because

$$\sum_{k=0}^{2p} \delta x\, \widetilde{\varphi}_k = \sum_{k=0}^{2p} \int_{\omega_k} \varphi(x)\,\mathrm{d}x = \int_\Omega \varphi(x)\,\mathrm{d}x = \Delta x\, \bar{\varphi}.$$

In the next subsection let us describe the procedure reconstructing cell DG polynomials from sub-cell FV data.

Reconstruction: from sub-cell FV data to cell DG polynomials

Starting from the FV representation $\widetilde{\varphi} = (\widetilde{\varphi}_0, \dots, \widetilde{\varphi}_{2p})$, we are looking for a polynomial $\varphi(x) = \sum_{i=0}^{p} \varphi_i \Phi_i(x)$ such that its projection onto the sub-cells is $\widetilde{\varphi}$. Because the number of sub-cells is larger than the number of components, $2p+1 > p+1$, a unique solution does not generally exist. In this work we search for the unique solution in the least square sense, and add a linear constraint that ensures the conservation of the mean value over the cell Ω for both the FV representation $\widetilde{\varphi}$ and the DG one φ. Formally we ought to minimize the cost functional

$$J(\varphi, \lambda) = \underset{(\varphi,\lambda)}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{k=0}^{2p} \left| \frac{1}{\delta x} \int_{\omega_k} \varphi(x)\,\mathrm{d}x - \widetilde{\varphi}_k \right|^2 + \lambda \left(\frac{1}{\Delta x} \int_\Omega \varphi(x)\,\mathrm{d}x - \frac{1}{2p+1} \sum_{k=0}^{2p} \widetilde{\varphi}_k \right) \right].$$

This least-square problem can be recast into the following linear system

$$\begin{pmatrix} \mathcal{P}^\top \mathcal{P} & \frac{1}{2} W^\top \\ \frac{1}{2} W & 0 \end{pmatrix} \begin{pmatrix} \varphi \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathcal{P}^\top \\ \frac{1}{2p+1} \mathbf{1}_{2p+1} \end{pmatrix} \widetilde{\varphi}, \tag{4.4.1}$$

where $W = (w_0, \dots, w_p)$ and $w_i = \dfrac{2}{\Delta x}\displaystyle\int_\Omega \Phi_i(x)\,\mathrm{d}x$, and, $\mathbf{1}_{2p+1} = (1, \dots, 1)$. Thus we can define the rectangular reconstruction matrix $\mathcal{R}$ of size $(p+1)\times(2p+1)$ as

$$\mathcal{R} = \begin{pmatrix} \mathrm{Id}_{p+1} & \mathbf{0}_{p+1}^\top \end{pmatrix} \begin{pmatrix} \mathcal{P}^\top\mathcal{P} & \frac{1}{2}W^\top \\ \frac{1}{2}W & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{P}^\top \\ \frac{1}{2p+1}\mathbf{1}_{2p+1} \end{pmatrix}.$$

Since by construction $\frac{1}{2p+1}\mathbf{1}_{2p+1}\mathcal{P} = \frac{1}{2}W$, we can deduce the following properties

$$\mathcal{R}\mathcal{P} = \mathrm{Id}_{p+1}, \qquad \mathcal{P}\mathcal{R} \neq \mathrm{Id}_{2p+1}, \qquad \text{and} \qquad \frac{1}{2}W\mathcal{R} = \frac{1}{2p+1}\mathbf{1}_{2p+1}.$$

Some precomputations can be performed in advance to solve efficiently all linear systems over time, for example by computing and storing initialy the LU decomposition of the $(p+2)\times(p+2)$ matrix to be inverted in (4.4.1).

As a summary we have two representations of the same numerical solution at time t^n: the cell DG polynomial $\mathbb{U}^n$ and the sub-cell FV mean values $\mathbb{V}^n$.

4.4.3 Detection procedure

From the DG candidate solution $\mathbb{U}^*$ at time t^{n+1} we want to detect the problematic/bad/troubled cells which need to be recomputed with a more robust scheme, in this work with the FV scheme on the sub-cells. We have made the choice to detect over the FV solution $\mathbb{V}^*$ and not the DG one $\mathbb{U}^*$ because we expect the overall scheme to maintain the admissibility of the solutions. Only the robust FV scheme maintains such property. Let us present two different detectors used in combination: the first one detects non-admissible FV solution by checking the positivity of water height, and, the second one tracks down spurious oscillations by refraining new extrema from exceeding pre-defined thresholds.

Positivity of water height

When the water height $\widetilde{h}_{k,j}$ of the kth sub-cell $\omega_{k,j}$ of the cell Ω_j is smaller than a threshold ϵ, we tag the cell Ω_j for recomputation. The positivity of the water height is mandatory for the solutions to be in the phase space. Moreover it ensures that the schemes are able to compute a meaningful numerical solution for the next time step.

Spurious oscillations

Although the physical variables h, hu do not satisfy a continuous maximum principle, one idea of [4] to detect spurious oscillations was to test if the FV variables remains in some local bounds. However, after some small numerical tests, it seemed more relevant to tag the cell if there was some local extrema in the solution at time t^{n+1} in order to recompute it. More precisely, for $\varphi = h$, hu or u, we consider that the solution can contain spurious oscillations and must be recomputed with a robust scheme if a new extrema is created. That is if

$$\widetilde{\varphi}^*_{k-1,j} \leq \widetilde{\varphi}^*_{k,j} + \varepsilon \qquad \text{and} \qquad \widetilde{\varphi}^*_{k,j} \leq \widetilde{\varphi}^*_{k+1,j} + \varepsilon$$

or

$$\widetilde{\varphi}^*_{k-1,j} \geq \widetilde{\varphi}^*_{k,j} - \varepsilon \qquad \text{and} \qquad \widetilde{\varphi}^*_{k,j} \geq \widetilde{\varphi}^*_{k+1,j} - \varepsilon$$

where ε is a small parameter.
In order to avoid tagging cells due to floating point errors when the variables are really flat, we usually forbid the detection if $\left(M^{n+1}_{k,j} - m^{n+1}_{k,j}\right)$ is smaller than a parameter which depends only of the spatial discretization Δx, where the bounds are defined by

$$m^{n+1}_{k,j} = \min\left(\widetilde{\varphi}^{n+1}_{k-1,j}, \widetilde{\varphi}^{n+1}_{k,j}, \widetilde{\varphi}^{n+1}_{k+1,j}\right) \qquad \text{and} \qquad M^{n+1}_{k,j} = \max\left(\widetilde{\varphi}^{n+1}_{k-1,j}, \widetilde{\varphi}^{n+1}_{k,j}, \widetilde{\varphi}^{n+1}_{k+1,j}\right).$$

In other words, if the maximal observed jump in the vicinity is genuinely small compared to the local characteristics length, then we accept the cell as is.

4.4.4 MOOD loop

The MOOD loop is the machinery employed to get a valid solution $\mathbb{U}^{n+1}$ starting from valid cell DG polynomial $\mathbb{U}^n$ at t^n. The first step consists in computing the unlimited DG solution at t^{n+1} denoted $\mathbb{U}^*$. This solution could be oscillatory in presence of steep gradients, could have negative water height, or could have undefined or unrepresentable values (NaN, Inf). There are at least two natural approaches that can emerge from the use of a MOOD loop in conjunction with a DG scheme.

Natural but incompatible MOOD loop

The first one, which seems the most natural, considers the cell DG polynomial representation as our starting data at t^n. The DG scheme produces the unlimited DG solution $\mathbb{U}^*$ at t^{n+1} which is further projected onto the sub-cells to get the representation $\mathbb{V}^*$. The detectors split the cells into good ones, which can exit the loop, and, bad ones which demand recomputation, see Figure 4.4.1. The cell DG solution in bad cells is discarded and, starting from the sub-cell projection $\mathbb{V}^n$ of $\mathbb{U}^n$ at t^n we apply the FV scheme to update the sub-cells of the bad cells and get $\mathbb{V}^{n+1}$. Those updated and valid sub-cells are sent for reconstruction. Finaly, the reconstruction procedure produces a valid cell DG polynomial $\mathbb{U}^{n+1}$ for each bad cell flagged by the detectors. As such we start from cell DG polynomials $\mathbb{U}^n$ and end with $\mathbb{U}^{n+1}$ constituted of good cells from $\mathbb{U}^*$ and reconstructed cell from the FV sub-cells $\mathbb{V}^{n+1}$.
This methodology seems to work well since we know that $\mathcal{RP} = \mathrm{Id}_{p+1}$, and thus we do not modify the DG representation $\mathbb{U}$ on non-detected cells. On the other hand, there remains an incompatibility because the FV scheme can ensure a valid solution $\mathbb{V}^{n+1}$ only if the data $\mathbb{V}^n$ are valid ones. Unfortunately the projection procedure does not ensure such validity by construction. Therefore, in this MOOD loop, when bad cells are detected, the projection $\mathcal{P}(\mathbb{U}^n)$ does not necessarily ensures valid sub-cell data $\mathbb{V}^n$. This can be seen from a different perspective as $\mathcal{PR} \neq \mathrm{Id}_{2p+1}$. Because this may lead to numerical difficulties, we have adopted another compatible MOOD loop.

Compatible MOOD loop

The second approach consists in considering a procedure which computes FV solutions $\mathbb{V}^n$ on the sub-cell mesh made of $2p+1$ sub-cells for any cell. The idea is that whenever possible we

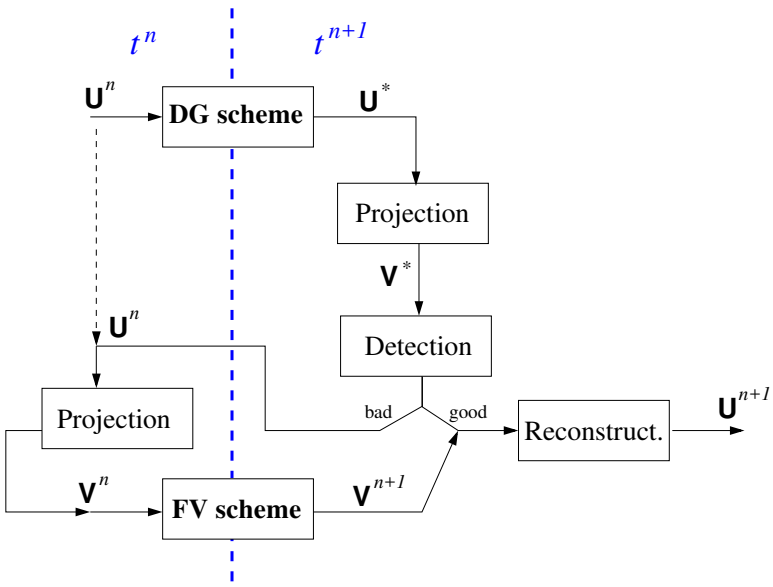

Figure 4.4.1 – Illustration of an update using the most natural MOOD loop. This solution is however invalid as there is no more exact equivalence between FV $\mathbb{V}$ data and DG $\mathbb{U}$ polynomials.

accelerate the solution update and increase the order of accuracy by using the DG representation $\mathbb{U}$ and the DG scheme, see Figure 4.4.2.
Here again we benefit from the identity: $\mathcal{RP} = \mathrm{Id}_{p+1}$ and avoid the recomputation of DG polynomials on valid cells. As such, by construction, the FV representation at any time is admissible, because any possible non-admissible candidate FV solution $\mathbb{V}^*$ is flagged by the positivity detector when appropriate. Hence, for such bad cells, the robust FV scheme would employ valid FV data at t^n to produce (by construction) a valid FV solution at t^{n+1}. This seemingly re-ordering of operations between the two MOOD loops is important to ensure that the scheme always produce an acceptable solution for both representations, DG and FV.

4.4.5 Robust Finite Volume scheme

Sub-cell based FV scheme

Dividing each cell Ω_j into $2p+1$ sub-cells $\omega_{k,j}$, with $k = 0, \ldots, 2p$, we denote by $\widetilde{\varphi}^n_{k,j}$ the approximation of the quantity $\dfrac{1}{\delta x}\displaystyle\int_{\omega_{k,j}} \varphi(x, t^n)\,\mathrm{d}x$, where $\delta x = |\omega_{k,j}| = \frac{\Delta x}{2p+1}$, which is nothing but the true mean value of solution φ over the sub-cell $\omega_{k,j}$ at time t^n.

We use the same FV scheme as the well-balanced one proposed in 1D in Chapter 1. The explicit FV acoustic step writes in this context

$$\begin{cases} \widetilde{\tau}^{n+1^-}_{k,j} = \widetilde{\tau}^n_{k,j} + \dfrac{\Delta t}{\delta x}\widetilde{\tau}^n_{k,j}\left(\widetilde{u}^*_{k+1/2,j} - \widetilde{u}^*_{k-1/2,j}\right) =: \widetilde{L}_{k,j}\widetilde{\tau}^n_{k,j}, \\ \widetilde{u}^{n+1^-}_{k,j} = \widetilde{u}^n_{k,j} - \dfrac{\Delta t}{\delta x}\widetilde{\tau}^n_{k,j}\left(\widetilde{\Pi}^{*,L}_{k+1/2,j} - \widetilde{\Pi}^{*,R}_{k-1/2,j}\right), \end{cases} \tag{4.4.2}$$

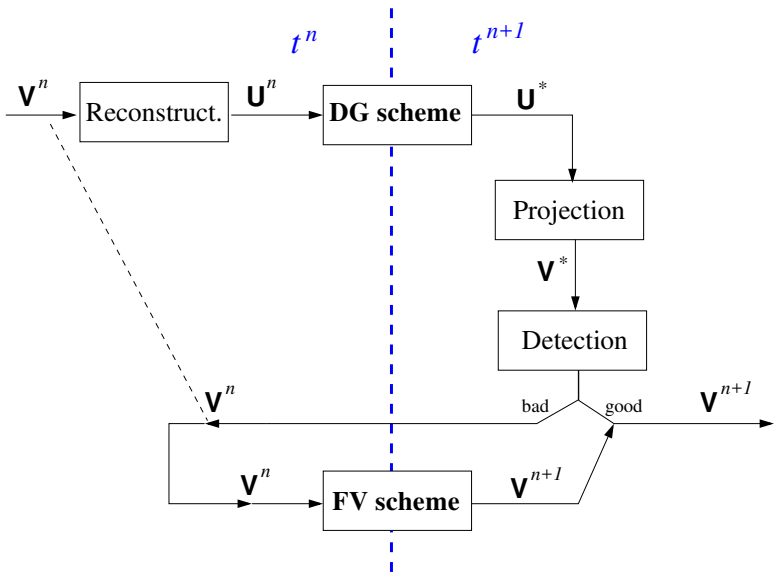

Figure 4.4.2 – Illustration of an update using a more robust MOOD loop. Contrarily to figure 4.4.1

where the fluxes are defined by

$$\begin{aligned}
\widetilde{u}^*_{k+1/2,j} &= \frac{\widetilde{u}^n_{k,j} + \widetilde{u}^n_{k+1,j}}{2} - \frac{1}{2\widetilde{a}_{k+1/2,j}}\left(\widetilde{\Pi}_{k+1,j} - \widetilde{\Pi}_{k,j}\right) - \frac{1}{2\widetilde{a}_{k+1/2,j}}\left\{g\widetilde{h}\Delta\widetilde{z}\right\}_{k+1/2,j}, \\
\widetilde{\Pi}^{*,L/R}_{k+1/2,j} &= \widetilde{\Pi}^*_{k+1/2,j} \pm \frac{1}{2}\left\{g\widetilde{h}\Delta\widetilde{z}\right\}_{k+1/2,j}, \\
\widetilde{\Pi}^*_{k+1/2,j} &= \frac{\widetilde{\Pi}^n_{k,j} + \widetilde{\Pi}^n_{k+1,j}}{2} - \frac{\widetilde{a}_{k+1/2,j}}{2}\left(\widetilde{u}_{k+1,j} - \widetilde{u}_{k,j}\right),
\end{aligned}$$

where the wave speed and source term are given by

$$\begin{aligned}
\widetilde{a}_{k+1/2,j} &\geq \max\left(\widetilde{h}^n_{k,j}\sqrt{g\widetilde{h}^n_{k,j}}, \widetilde{h}^n_{k+1,j}\sqrt{g\widetilde{h}^n_{k+1,j}}\right), \\
\left\{g\widetilde{h}\Delta\widetilde{z}\right\}_{k+1/2,j} &= g\frac{\widetilde{h}^n_{k,j} + \widetilde{h}^n_{k+1,j}}{2}\left(\widetilde{z}_{k+1,j} - \widetilde{z}_{k,j}\right),
\end{aligned}$$

and we have used the convention $\widetilde{\varphi}^n_{-1,j} = \widetilde{\varphi}^n_{2p,j-1}$ and $\widetilde{\varphi}^n_{2p+1,j} = \widetilde{\varphi}^n_{0,j+1}$.

The upwind FV transport step writes

$$\begin{cases}
\widetilde{h}^{n+1}_{k,j} = \widetilde{L}_{k,j}\widetilde{h}^{n+1^-}_{k,j} - \dfrac{\Delta t}{\delta x}\left(\widetilde{h}^{*,n+1^-}_{k+1/2,j}\widetilde{u}^*_{k+1/2,j} - \widetilde{h}^{*,n+1^-}_{k-1/2,j}\widetilde{u}^*_{k-1/2,j}\right) \\
\widetilde{(hu)}^{n+1}_{k,j} = \widetilde{L}_{k,j}\widetilde{(hu)}^{n+1^-}_{k,j} - \dfrac{\Delta t}{\delta x}\left(\widetilde{(hu)}^{*,n+1^-}_{k+1/2,j}\widetilde{u}^*_{k+1/2,j} - \widetilde{(hu)}^{*,n+1^-}_{k-1/2,j}\widetilde{u}^*_{k-1/2,j}\right)
\end{cases} \tag{4.4.3}$$

where

$$\widetilde{\varphi}^{*,n+1^-}_{k+1/2,j} = \begin{cases} \widetilde{\varphi}^{*,n+1^-}_{k,j}, & \text{if } \widetilde{u}^*_{k+1/2,j} \geq 0, \\ \widetilde{\varphi}^{*,n+1^-}_{k+1,j}, & \text{otherwise.} \end{cases}$$

Global conservation

If we want to maintain a global conservation of the water height h, and the discharge hu when the topography is flat, special attention is needed for the fluxes on the mixed interfaces where the solutions on left and right sides have not been updated with the same scheme. Let us suppose, without loss of generality, that the candidate DG solution $\mathbb{U}_j^*$ at $t^{n+1} = t^n + \Delta t$ on cell Ω_j has passed the detectors and becomes the updated solution $\mathbb{U}_j^{n+1}$. Let us moreover assume that the DG solution $\mathbb{U}_{j+1}^*$ on the neighbour cell Ω_{j+1} failed and demands recomputation by the FV scheme. Then the final DG solution $\mathbb{U}_{j+1}^{n+1}$ will be the reconstruction of the FV representation $\mathbb{V}_{j+1}^{n+1}$ which is a sub-cell FV update of the FV state $\mathbb{V}_{j+1}^n$ at previous time t^n. *A priori* a problem of conservation seems to arise without modifying the DG quantities and fluxes on cell Ω_j, leading to the unpleasant and time consuming situation of recomputing a previously valid cell.

We can write the global update over the mean values on cells Ω_j by the DG scheme

$$\begin{cases} \overline{h}_j^{n+1} = \dfrac{1}{\Delta x}\displaystyle\int_{\Omega_j} h_j^{n+1}(x)\,\mathrm{d}x = \sum_{i=0}^{p} \frac{w_i}{2} h_{i,j}^{n+1} = \overline{h}_j^n - \frac{\Delta t}{\Delta x}\left(h_{j+1/2}^{*,n+1^-} u_{j+1/2}^* - h_{j-1/2}^{*,n+1^-} u_{j-1/2}^*\right), \\ \overline{(hu)}_j^{n+1} = \overline{(hu)}_j^n - \dfrac{\Delta t}{\Delta x}\left((hu)_{j+1/2}^{*,n+1^-} u_{j+1/2}^* + \Pi_{j+1/2}^{*,L} - (hu)_{j-1/2}^{*,n+1^-} u_{j-1/2}^* - \Pi_{j-1/2}^{*,R}\right), \end{cases}$$

and on Ω_{j+1} by the FV scheme

$$\begin{cases} \overline{\widetilde{h}}_{j+1}^{n+1} = \dfrac{1}{2p+1}\displaystyle\sum_{k=0}^{2p} \widetilde{h}_{k,j+1}^{n+1} = \overline{\widetilde{h}}_{j+1}^n - \frac{\Delta t}{\Delta x}\left(\widetilde{h}_{2p+1/2,j+1}^{*,n+1^-} \widetilde{u}_{2p+1/2,j+1}^* - \widetilde{h}_{-1/2,j+1}^{*,n+1^-} \widetilde{u}_{-1/2,j+1}^*\right), \\ \overline{\widetilde{(hu)}}_{j+1}^{n+1} = \overline{\widetilde{(hu)}}_{j+1}^n - \dfrac{\Delta t}{\Delta x}\Bigg(\widetilde{(hu)}_{2p+1/2,j+1}^{*,n+1^-} \widetilde{u}_{2p+1/2,j+1}^* + \widetilde{\Pi}_{2p+1/2,j+1}^{*,L} \\ \qquad\qquad - \widetilde{(hu)}_{-1/2,j+1}^{*,n+1^-} \widetilde{u}_{-1/2,j+1}^* - \widetilde{\Pi}_{-1/2,j+1}^{*,R}\Bigg). \end{cases}$$

However thanks to the property $\frac{1}{2p+1}\mathbf{1}_{2p+1}\mathcal{P} = \frac{1}{2}W$, we have that $\overline{\widetilde{h}}_{j+1}^n = \overline{h}_{j+1}^n$ when we project the solution at time t^n. We also know that the reconstruction procedure satisfies $\frac{1}{2}W\mathcal{R} = \frac{1}{2p+1}\mathbf{1}_{2p+1}$, thus we deduce that $\overline{(hu)}_{j+1}^{n+1} = \overline{\widetilde{(hu)}}_{j+1}^{n+1}$ when we reconstruct the solution at time t^{n+1}. Finally, if we define the FV fluxes on the right side of the interface between the cells Ω_j and Ω_{j+1} as the DG fluxes that would have been imposed if the cell Ω_{j+1} would not have been detected as bad, we obtain the global conservation of the MOOD scheme, for both approaches. More precisely for the fluxes, we set

$$\widetilde{\varphi}_{-1/2,j+1}^* := \varphi_{j+1/2}^*,$$

where $\varphi = u$, Π^R, h and (hu) respectively.

4.5 Numerical results

This section is dedicated to the verification of this DG numerical method with sub-cell FV limiter based on an *a posteriori* MOOD loop for the shallow water equations. We only show the

results for the natural loop as the compatible one is not completely debugged yet. We compare first the numerical solution of a dam break problem with a bump of the topography, showing that the scheme combined with the MOOD loop gives standard results. Then, we present a case in which the classical limiters have some trouble to reduce the spurious oscillations and we illustrates some good improvement with the MOOD *a posteriori* limitation.

In the following, we call EX_k the explicit-explicit DG scheme with polynomials of order k, MO_k the same DG schemes combined with MOOD loop. In order to compare them with a reference order 1 FV scheme in an Eulerian approach, we use the well-balanced scheme proposed in [1] and denoted in the following by ACU.

The high order time discretization used in this work is the same as in Chapter 3, that is to say strong stability preserving Runge-Kutta (RK) methods, described for instance in [6]. For every tested scheme, we use the RK method with the same theoretical corresponding order: 1 for ACU and $k+1$ for both EX_k and MO_k. In the case of the MO_k schemes, this means that the MOOD loops are done for every sub-step of the RK method. We have not been able yet to compare the results with another option which is to consider that the MOOD loop includes the RK methods for both the DG and the FV schemes. We refer to [9] for the use of another time discretization method, called ADER, which permits arbitrary high order methods.

We take as the time step the FV one, that is to say $\Delta t = \min\left(\Delta t_{\mathrm{lag}}, \Delta t_{\mathrm{tra}}\right)$ where

$$\widetilde{a}_{k+1/2,j} = 1.01 \cdot \max\left(\widetilde{h}^n_{k,j}\sqrt{g\widetilde{h}^n_{k,j}}, \widetilde{h}^n_{k+1,j}\sqrt{g\widetilde{h}^n_{k+1,j}}\right),$$

$$\Delta t_{\mathrm{lag}} = \frac{\delta x}{\max_j \max_k \left(2\widetilde{a}_{k+1/2,j} \max(\widetilde{\tau}_{k,j}, \widetilde{\tau}_{k+1,j})\right)},$$

and

$$\Delta t_{\mathrm{tra}} = \frac{\delta x}{\max_j \max_k \left(|\widetilde{u}^*_{k+1/2,j}|\right)}.$$

4.5.1 Dam break with topography

This first test case is the same as the one presented in Section 1.4.1. We consider a break of a dam at the initial time and we can observe one rarefaction wave going towards the left hand side, and a shock wave towards the right hand side. The physical domain is $[0, 1500]$, the velocity u is set to zero everywhere at $t = 0$, and the total water height $H = h + z$ to

$$H(x, t=0) = \begin{cases} 20, & \text{if } x \leq 750, \\ 15, & \text{if } x > 750. \end{cases}$$

We recall the regularized bump function for the topography:

$$z(x) = \begin{cases} 4e^{2-\frac{150}{x-487.5}}, & \text{if } 487.5 < x <= 562.5, \\ 8 - 4e^{2-\frac{150}{637.5-x}}, & \text{if } 562.5 < x <= 637.5, \\ 8, & \text{if } 637.5 < x <= 862.5, \\ 8 - 4e^{2-\frac{150}{x-862.5}}, & \text{if } 862.5 < x <= 937.5, \\ 4e^{2-\frac{150}{1012.5-x}}, & \text{if } 937.5 < x <= 1012.5, \\ 0 & \text{otherwise.} \end{cases}$$

We discretize the domain over a 500-cell grid and we use Neumann boundary conditions. The total water height at time $T = 50$ is represented in Figure 4.5.1. We can observe that the slopes of the shock in the zoomed figure are quite the same for the EX_1 (EX_2) and the MO_1 (MO_2) schemes. Even if the computation times are, as expected, longer for the MO_k schemes, we see in Table 4.1 that the number of iterations are similar with respect to the EX_k ones. Finally, there are for instance 120 cells over the 500 which are detected for the MO_2 scheme at the last iteration. Thus, more work has to be done in order to improve the detectors and only detect the needed cells.

	ACU	EX0	EX1	MO1	EX2	MO2
Nb iterations	534	525	1573	1573	2622	2621
CPU times	0.517	0.526	1.548	3.256	3.468	9.363

Table 4.1 – Dam break with topography. Numbers of iterations and CPU times for every simulation.

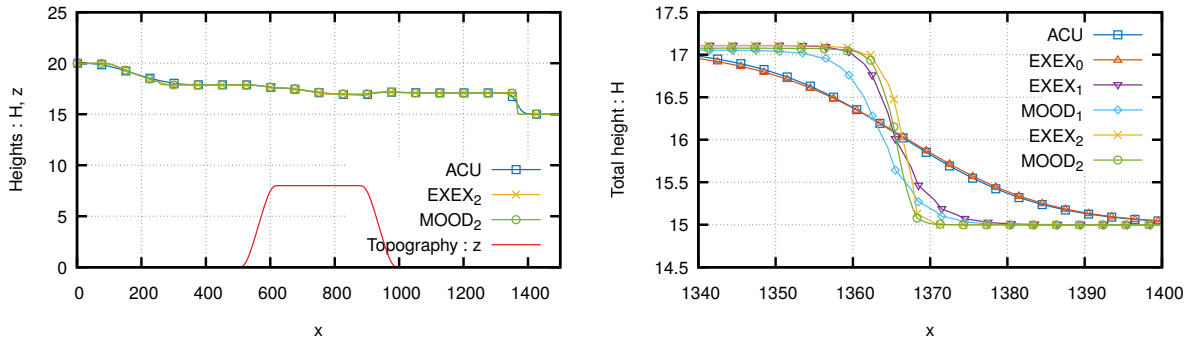

Figure 4.5.1 – Dam break with topography. Profile of $H = z + h$ at time $T = 50$, over the domain (left) and zoomed on the shock wave (right).

4.5.2 Strong dam break with flat bottom

We take the same test case, however we remove the topography and we set initially a lower water height on the right hand side:

$$H(x, t = 0) = h(x, t = 0) = \begin{cases} 20, & \text{if } x \leq 750, \\ 1, & \text{if } x > 750. \end{cases}$$

We observe the same jumps in the rarefaction wave at time $T = 20$ in the Figure 4.5.2 as in the Figure 3.5.1. Note that here we have used only limiters on characteristic variables for the EX_1 and EX_2 schemes. We can see that both MO_1 and MO_2 give better solutions in the zoomed area. The same remarks as for the previous test case can be done regarding the

number of iteration and computation times reported in the Table 4.2. This time only 45 cells are detected at the last iteration, however we generally want to get less than 5% of detected cells.

	ACU	EX0	EX1	MO1	EX2	MO2
Nb iterations	296	223	645	651	1158	1062
CPU times	0.481	0.465	0.940	1.364	1.980	3.331

Table 4.2 – Strong dam break. Numbers of iterations and CPU times for every simulation.

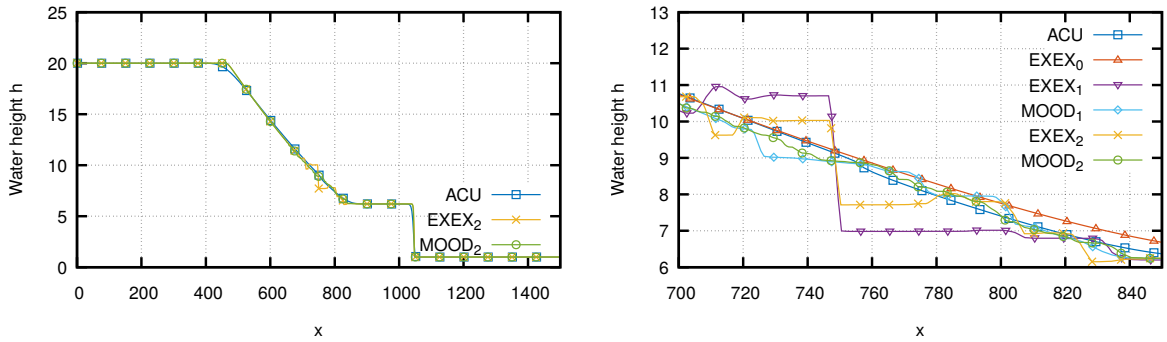

Figure 4.5.2 – Strong dam break. Profile of h at time $T = 20$, over the domain (left) and zoomed on the rarefaction wave (right).

4.6 Conclusions and perspectives

In this chapter we have proposed a theoretical combination of the MOOD loop with the DG scheme studied in Chapter 3. We detailed the different steps needed to construct the new scheme, rising two different approaches. Finally, we presented the first numerical results we have with one of the two approaches. We found the improvement of the methodology promising.

However, we stressed the requirement of further numerical tests, with the comparison of the two approaches. In particular, it seems important to improve the detection criteria in order to obtain a systematic method which tags only the relevant bad cells. Finally, long term perspectives could be an optimization of the code, an extension to 2 and 3 dimension of space, and an extension to other systems of partial differential equations such that Euler with source terms.

Bibliographie

[1] Emmanuel Audusse, Christophe Chalons, and Philippe Ung. A very simple well-balanced positive and entropy-satisfying scheme for the shallow-water equations. *Commun. Math. Sci*, 13(5) :1317–1332, 2015.

[2] Christophe Berthon and Vivien Desveaux. An entropy preserving MOOD scheme for the Euler equations. *International Journal on Finite Volumes*, pages 1–39, 2014.

[3] Christophe Chalons, Pierre Kestener, Samuel Kokh, and Maxime Stauffert. A large time-step and well-balanced Lagrange-Projection type scheme for the shallow-water equations. *Communication in Mathematical Sciences*, 15(3) :765–788, 2017.

[4] Stéphane Clain, Steven Diot, and Raphaël Loubère. A high-order finite volume method for systems of conservation laws—Multi-dimensional Optimal Order Detection (MOOD). *Journal of computational Physics*, 230(10) :4028–4050, 2011.

[5] Stéphane Clain and J Figueiredo. The MOOD method for the non-conservative shallow-water system. *Computers & Fluids*, 145 :99–128, 2017.

[6] Bernardo Cockburn and Chi-Wang Shu. Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of scientific computing*, 16(3) :173–261, 2001.

[7] Steven Diot, Stéphane Clain, and Raphaël Loubère. Improved detection criteria for the multi-dimensional optimal order detection (MOOD) on unstructured meshes with very high-order polynomials. *Computers & Fluids*, 64 :43–63, 2012.

[8] Michael Dumbser, Olindo Zanotti, Raphaël Loubère, and Steven Diot. A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *Journal of Computational Physics*, 278 :47–75, 2014.

[9] Raphaël Loubere, Michael Dumbser, and Steven Diot. A new family of high order unstructured MOOD and ADER finite volume schemes for multidimensional systems of hyperbolic conservation laws. *Communications in Computational Physics*, 16(3) :718–763, 2014.

[10] I. Suliciu. On the thermodynamics of fluids with relaxation and phase transitions. Fluids with relaxation. *Internat. J. Engrg. Sci*, 36 :921–947, 1998.

Troisième partie

Implémentation

Chapitre 5

Implémentation

5.1 Code de simulation à une dimension d'espace

On veut résoudre numériquement les équations de Saint-Venant (0.5.1) avec terme source de topographie à une dimension d'espace. On implémente à l'aide du langage de programmation compilé orienté objet C++ les schémas Volumes Finis ordre 1 décrits dans le Chapitre 1 et les schémas Galerkin discontinus d'ordres élevés décrits dans les Chapitres 3 et 4. Le fichier principal est décrit dans le Listing 5.1.

— L'objet `pb` de la classe `PB` contient tous les paramètres du cas test.

— Les objets `stateOld` et `stateNew` sont deux jeux de valeurs des variables conservatives (h, hu, z), le premier au temps t^n et le deuxième au temps t^{n+1}.

— Le choix de la méthode Runge-Kutta en temps se fait grâce aux classes filles de la classe mère `RK`, à savoir `RK1`, `RK2` ou `RK3`. Chaque classe fille de `RK` contient la méthode `TimeStep()` qui rempli l'état `stateNew` grâce aux valeurs de `stateOld`.

— Le choix du solveur se fait grâce aux classes filles de la classe mère `SOLVER`, à savoir `EXEXFV` pour le schéma Volumes Finis explicite-explicite de la Section 5.1.1, `IMEXFV` pour la version implicite-explicite de la Section 5.1.2, et `IMEXDG` pour le schéma Galerkin discontinu implicite-explicite de la Section 5.1.3. Chaque classe fille de `SOLVER` contient les méthodes `FillGhosts()`, qui rempli les mailles fictives de l'état `stateOld`, et `computeDt()` qui calcule le plus grand pas de temps que l'on peut utiliser en accord avec la condition CFL du solveur considéré. Elle contient également la méthode `SolverStep()` qui rempli les états intermédiaires de chaque sous pas de la méthode RK utilisée, et qui est donc appelée dans la méthode `TimeStep` des classes filles de `RK`.

Listing 5.1 – main.cpp

```
1 PB pb("/repertoire_cas_test");
2 SOLVER solver(pb); // SOLVER = EXEXFV, IMEXFV, IMEXDG, ...
3 RK rk(solver, pb); // RK = RK1, RK2 ou RK3
4
5 STATE stateOld(pb); // (h,hu,z)^n
6 STATE stateNew(pb); // (h,hu,z)^{n+1}
7 double t(0.0);
8
9 while (t < pb.m_tf) {
```

```
    SwapStates(stateOld, stateNew); // (h,hu,z)^n <-> (h,hu,z)^{n+1}
    solver.FillGhosts(stateOld);

    double dt(solver.computeDt(stateOld, pb));
    rk.TimeStep(stateNew, stateOld, pb, dt);
    t += dt;
}
```

Dans les prochaines sections, nous allons détailler uniquement les parties importantes des fonctions `FillGhosts()` et `SolverStep()`.

5.1.1 Schéma Volumes Finis d'ordre 1 explicite-explicite

Étapes principales du code

Avant de détailler les étapes clés du code, rappelons les équations qui décrivent le schéma numérique. Le schéma global s'écrit

$$\begin{cases} h_j^{n+1} = h_j^n - \dfrac{\Delta t}{\Delta x}\left(h_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,n} - h_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,n}\right), \\ (hu)_j^{n+1} = (hu)_j^n - \dfrac{\Delta t}{\Delta x}\left((hu)_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,n} + \Pi_{j+1/2}^{*,n} - (hu)_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,n} - \Pi_{j-1/2}^{*,n}\right) \\ \qquad\qquad - \dfrac{\Delta t}{\Delta x}\left(\dfrac{1}{2}\left\{gh\Delta z\right\}_{j+1/2}^n + \dfrac{1}{2}\left\{gh\Delta z\right\}_{j-1/2}^n\right). \end{cases} \tag{5.1.1}$$

avec $a_{j+1/2} = \max\left(h_j^n\sqrt{gh_j^n}, h_{j+1}^n\sqrt{gh_{j+1}^n}\right)$,

$$\begin{aligned} \left\{gh\Delta z\right\}_{j+1/2}^n &= g\frac{h_j^n + h_{j+1}^n}{2}\left(z_{j+1} - z_j\right), \\ u_{j+1/2}^{*,n} &= \frac{u_j^n + u_{j+1}^n}{2} - \frac{1}{2a_{j+1/2}}\left(\Pi_{j+1}^n - \Pi_j^n\right) - \frac{1}{2a_{j+1/2}}\left\{gh\Delta z\right\}_{j+1/2}^n, \\ \Pi_{j+1/2}^{*,n} &= \frac{\Pi_j^n + \Pi_{j+1}^n}{2} - \frac{a_{j+1/2}}{2}\left(u_{j+1}^n - u_j^n\right), \end{aligned}$$

et

$$\varphi_{j+1/2}^{*,n+1-} = \begin{cases} \varphi_j^{n+1-}, & \text{si } u_{j+1/2}^{*,n} \geq 0, \\ \varphi_{j+1}^{n+1-}, & \text{sinon,} \end{cases}$$

où $\varphi = h, (hu)$. Les quantités au temps fictif t^{n+1-} sont calculées grâce au pas acoustique suivant

$$\begin{cases} L_j = 1 + \dfrac{\Delta t}{\Delta x}\left(u_{j+1/2}^{*,n} - u_{j-1/2}^{*,n}\right), \\ L_j h^{n+1-} = h_j^n, \\ L_j (hu)^{n+1-} = (hu)_j^n - \dfrac{\Delta t}{\Delta x}\left(\Pi_{j+1/2}^{*,n} - \Pi_{j-1/2}^{*,n}\right) \\ \qquad\qquad - \dfrac{\Delta t}{\Delta x}\left(\dfrac{1}{2}\left\{gh\Delta z\right\}_{j+1/2}^n + \dfrac{1}{2}\left\{gh\Delta z\right\}_{j-1/2}^n\right). \end{cases} \tag{5.1.2}$$

On peut noter que le schéma ainsi écrit est la combinaison de deux sous pas à 3 points et est donc globalement un schéma à 5 points.

On programme alors schématiquement la mise à jour entre les temps t^n et t^{n+1} dans la fonction `SolverStep()` comme suit :

— `AcousticStep()` est une fonction qui parcourt toutes les mailles et qui calcule les états φ^{n+1-}, avec $\varphi = h, (hu)$, grâce au système (5.1.2). Comme nous sommes en 1D, lorsque l'on parcourt les mailles, on atteint également toutes les interfaces et on peut ainsi stocker les valeurs aux interfaces suivantes : $\{gh\Delta z\}^n$, $u^{*,n}$, $\Pi^{*,n}$ et $\varphi^{*,n+1-}$.

— `GlobalStep()` est une fonction qui parcourt toutes les mailles et qui met à jour les états $\varphi = h, (hu)$, grâce au système (5.1.1) et aux valeurs aux interfaces ci-dessus.

Traitement du bord

Afin de prendre en considération les conditions de bord, il est assez classique d'augmenter le domaine physique en un domaine de calcul légèrement plus grand avec des mailles fictives. Étant donné que notre schéma est un schéma à 5 points, on a besoin de rajouter deux mailles fictives de chaque côté du domaine physique en 1D, comme représenté sur la Figure 5.1.1.

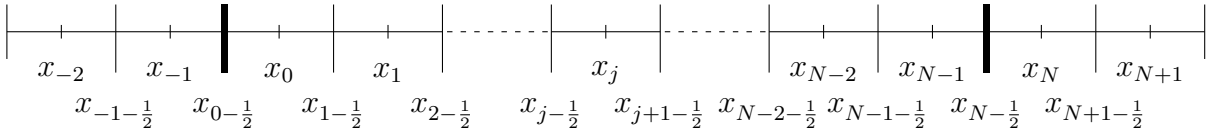

FIGURE 5.1.1 – Représentation des mailles fantômes ainsi que des indices de maille et d'interface du maillage en 1D.

Une autre possibilité aurait été de n'imposer qu'une seule maille fantôme de chaque côté du domaine physique et de les remplir à la fois au temps t^n pour le pas acoustique, et au temps fictif t^{n+1-} pour le pas transport. Cette méthode demande de connaître la bonne valeur à mettre dans les mailles fantômes à un temps fictif pour des quantités qui n'ont pas nécessairement de sens physique.

Conditions de bords périodiques. Si on somme sur toutes les mailles l'équation (5.1.1), on obtient pour la valeur moyenne de la hauteur d'eau :

$$\overline{h}^{n+1} = \frac{1}{N}\sum_{j=0}^{N-1} h_j^{n+1} = \overline{h}^n - \frac{\Delta t}{\Delta x}\left(h^{*,n+1-}_{N-\frac{1}{2}} u^{*,n}_{N-\frac{1}{2}} - h^{*,n+1-}_{0-\frac{1}{2}} u^{*,n}_{0-\frac{1}{2}}\right).$$

Pour avoir un schéma conservatif ($\overline{h}^{n+1} = \overline{h}^n$), il suffit que $u^{*,n}_{N-\frac{1}{2}} = u^{*,n}_{0-\frac{1}{2}}$ et $h^{*,n+1-}_{N-\frac{1}{2}} = h^{*,n+1-}_{0-\frac{1}{2}}$. La première égalité peut être assurée en prenant $(h, hu)^n_{-1} = (h, hu)^n_{N-1}$ et $(h, hu)^n_N = (h, hu)^n_0$. À partir de ces valeurs, il est assez facile de vérifier grâce au système (5.1.2) qu'en imposant $(h, hu)^n_{-2} = (h, hu)^n_{N-2}$ et $(h, hu)^n_{N+1} = (h, hu)^n_1$ on obtient $h^{n+1-}_{-1} = h^{n+1-}_{N-1}$ et $h^{n+1-}_N = h^{n+1-}_0$, d'où la deuxième égalité. En résumé, on a copié les deux premières mailles du domaine physique dans les deux mailles fantômes à l'extrémité droite et inversement les deux dernières mailles dans les deux mailles fantômes à l'extrémité gauche comme représenté dans la Figure 5.1.2.

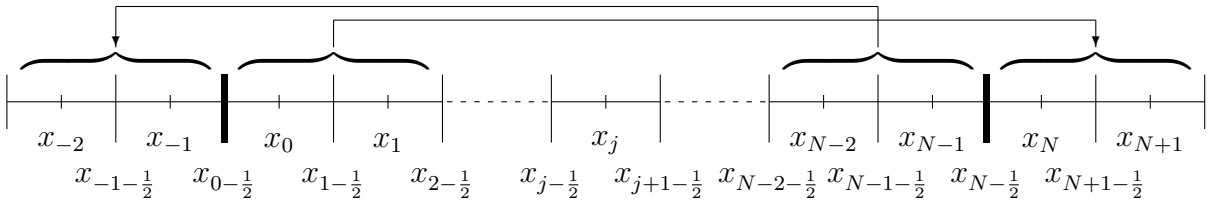

FIGURE 5.1.2 – Remplissage des mailles fantômes pour le schéma explicite-explicite avec conditions de bords périodiques.

Conditions de bord réfléchissant. Concernant les conditions de bord réfléchissant, on souhaite avoir $u^{*,n}_{0-\frac{1}{2}} = 0$, c'est pourquoi on prend naturellement $(h, hu)^n_{-1} = (h, -hu)^n_0$ qui assure l'égalité précédente. Les valeurs imposées sur la première maille fantôme, $j = -2$, ne sont pas très importantes ici puisqu'elles n'influent pas sur la valeur de $u^{*,n}_{0-\frac{1}{2}}$. On propose tout de même d'imposer $(h, hu)^n_{-2} = (h, -hu)^n_1$. Le bord droit est traité de façon identique, à savoir $(h, hu)^n_N = (h, -hu)^n_{N-1}$ et $(h, hu)^n_{N+1} = (h, -hu)^n_{N-2}$.

Conditions de bord absorbant. On propose deux possibilités pour la condition de bord absorbant : celle qui a été implémentée et une autre qui n'a pas été testée numériquement. On souhaite obtenir $u^{*,n}_{0-\frac{1}{2}} \approx u(x_{0-\frac{1}{2}}, t^n)$ et $\varphi^{*,n+1-}_{0-\frac{1}{2}} \approx \varphi(x_{0-\frac{1}{2}}, t^{n+1-})$, avec $\varphi = h, hu$. Ce qui a été implémenté est assez simple et correspond à un prolongement constant de l'état dans la première maille $j = 0$, soit $(h, hu)^n_{-1} = (h, hu)^n_0$ et $(h, hu)^n_{-2} = (h, hu)^n_0$.

Une autre possibilité est de prolonger en une droite les valeurs en x_0 et x_1, soit $(h, u)^n_{-1} = 2(h, u)^n_0 - (h, u)^n_1$ et $(h, u)^n_{-2} = 3(h, u)^n_0 - 2(h, u)^n_1$. En effet, le polynôme associé s'écrit

$$\frac{\varphi_1 - \varphi_0}{\Delta x}(X - x_0) + \varphi_0,$$

où $\varphi = h, u, z$. Donc on obtient bien les valeurs ci-dessus en $x_{-1} = x_0 - \Delta x$ et $x_{-2} = x_0 - 2\Delta x$.

Enfin, le bord droit est traité de façon identique, à savoir

$$(h, hu)^n_N = (h, hu)^n_{N-1} \quad \text{et} \quad (h, hu)^n_{N+1} = (h, hu)^n_{N-1},$$

ou

$$(h, u, z)^n_N = 2(h, u, z)^n_{N-1} - (h, u, z)^n_{N-2} \quad \text{et} \quad (h, u, z)^n_{N+1} = 3(h, u, z)^n_{N-1} - 2(h, u, z)^n_{N-2}.$$

Conditions de bord de Dirichlet. Enfin, pour les condition de bord de type Dirichlet, on choisit d'imposer les valeurs de toutes les variables sur les deux mailles fantômes $j = -1, -2$, au temps t^n, aux valeurs que l'on aimerait imposer au bord du domaine physique. Plus précisément, si on souhaite imposer $\varphi^{\text{Dir}}_{0-1/2}(t^n)$ en $x_{0-1/2}$ au temps t^n, on prend $\varphi^n_{-1} = \varphi^{\text{Dir}}_{0-1/2}(t^n)$ et $\varphi^n_{-2} = \varphi^{\text{Dir}}_{0-1/2}(t^n)$. On procède de encore une fois de la même manière sur le bord droit.

5.1.2 Schéma Volumes Finis d'ordre 1 implicite-explicite

Étapes principales du code

On reprend comme pour le schéma explicite-explicite ci-dessus les équations qui décrivent le schéma implicite-explicite. Le schéma global s'écrit

$$\begin{cases} h_j^{n+1} = h_j^n - \dfrac{\Delta t}{\Delta x}\left(h_{j+1/2}^{*,n+1-}u_{j+1/2}^{*,n+1-} - h_{j-1/2}^{*,n+1-}u_{j-1/2}^{*,n+1-}\right), \\ (hu)_j^{n+1} = (hu)_j^n - \dfrac{\Delta t}{\Delta x}\left((hu)_{j+1/2}^{*,n+1-}u_{j+1/2}^{*,n+1-} + \Pi_{j+1/2}^{*,n+1-} - (hu)_{j-1/2}^{*,n+1-}u_{j-1/2}^{*,n+1-} - \Pi_{j-1/2}^{*,n+1-}\right) \\ \qquad\qquad - \dfrac{\Delta t}{\Delta x}\left(\dfrac{1}{2}\left\{gh\Delta z\right\}_{j+1/2}^n + \dfrac{1}{2}\left\{gh\Delta z\right\}_{j-1/2}^n\right). \end{cases} \tag{5.1.3}$$

avec

$$\begin{aligned} u_{j+1/2}^{*,n+1-} &= \frac{u_j^{n+1-} + u_{j+1}^{n+1-}}{2} - \frac{1}{2a_{j+1/2}}\left(\Pi_{j+1}^{n+1-} - \Pi_j^{n+1-}\right) - \frac{1}{2a_{j+1/2}}\left\{gh\Delta z\right\}_{j+1/2}^n, \\ \Pi_{j+1/2}^{*,n+1-} &= \frac{\Pi_j^{n+1-} + \Pi_{j+1}^{n+1-}}{2} - \frac{a_{j+1/2}}{2}\left(u_{j+1}^{n+1-} - u_j^{n+1-}\right), \end{aligned}$$

et

$$\varphi_{j+1/2}^{*,n+1-} = \begin{cases} \varphi_j^{n+1-}, & \text{si } u_{j+1/2}^{*,n+1-} \geq 0, \\ \varphi_{j+1}^{n+1-}, & \text{sinon}, \end{cases}$$

où $\varphi = h, (hu)$. On peut remarquer que la principale différence avec le schéma explicite-explicite provient du fait que les flux u^* et Π^* sont calculés à l'aide de quantités au temps fictif t^{n+1-} au lieu de t^n. Les quantités au temps fictif t^{n+1-} sont elles calculées grâce au pas acoustique suivant : on résout d'abord le système linéaire en u et Π

$$\begin{cases} u_j^{n+1-} = u_j^n - \dfrac{\Delta t}{\Delta x}\tau_j^n\left(\Pi_{j+1/2}^{*,n+1-} - \Pi_{j-1/2}^{*,n+1-}\right) \\ \qquad\qquad - \dfrac{\Delta t}{\Delta x}\tau_j^n\left(\dfrac{1}{2}\left\{gh\Delta z\right\}_{j+1/2}^n + \dfrac{1}{2}\left\{gh\Delta z\right\}_{j-1/2}^n\right), \\ \Pi_j^{n+1-} = \Pi_j^n - \dfrac{\Delta t}{\Delta x}\tau_j^n a_j^2\left(u_{j+1/2}^{*,n+1-} - u_{j-1/2}^{*,n+1-}\right), \end{cases} \tag{5.1.4}$$

où $a_j = \frac{a_{j-1/2}+a_{j+1/2}}{2}$, puis on met à jour les variables conservatives comme suit

$$\begin{cases} L_j = 1 + \dfrac{\Delta t}{\Delta x}\left(u_{j+1/2}^{*,n+1-} - u_{j-1/2}^{*,n+1-}\right), \\ L_j h^{n+1-} = h_j^n, \\ L_j (hu)^{n+1-} = (hu)_j^n - \dfrac{\Delta t}{\Delta x}\left(\Pi_{j+1/2}^{*,n+1-} - \Pi_{j-1/2}^{*,n+1-}\right) \\ \qquad\qquad - \dfrac{\Delta t}{\Delta x}\left(\dfrac{1}{2}\left\{gh\Delta z\right\}_{j+1/2}^n + \dfrac{1}{2}\left\{gh\Delta z\right\}_{j-1/2}^n\right). \end{cases} \tag{5.1.5}$$

On programme alors schématiquement la mise à jour entre les temps t^n et t^{n+1} dans la fonction `SolverStep()`comme suit :

— `AcousticStep()` est une fonction qui met à jour les variables conservatives pour le pas acoustique seulement et qui se décompose en trois sous fonctions. `BuildLinearSystem()` qui rempli la matrice et le second membre du système linéaire (5.1.4). `FillGhostsMat()` qui ajoute les conditions de bord au système et que l'on détaille dans la section suivante. Enfin, `SolveLinearSystem()` qui résout le système à la librairie PETSc[1], puis parcourt toutes les mailles et qui calcule les états φ^{n+1-}, avec $\varphi = h, (hu)$, grâce au système (5.1.5). Comme nous sommes en 1D, lorsque l'on parcourt les mailles, on atteint également toutes les interfaces et on peut ainsi stocker les valeurs aux interfaces suivantes : $u^{*,n+1-}$, $\Pi^{*,n+1-}$, $\{gh\Delta z\}^n$ et $\varphi^{*,n+1-}$.

— `GlobalStep()` est une fonction qui parcourt toutes les mailles et qui met à jour les états $\varphi = h, (hu)$, grâce au système (5.1.3) et aux valeurs aux interfaces ci-dessus.

Traitement du bord

Il y a une difficulté supplémentaire concernant les conditions de bords pour ce schéma, à savoir que le premier pas acoustique est traité implicitement. En plus d'initialiser à chaque pas de temps t^n les variables des mailles fictives comme pour le schéma explicite-explicite, nous allons ajouter au système linéaire sur (u, Π) du pas acoustique les conditions de bords au temps fictif t^{n+1-}.

Conditions de bords périodiques. Cette fois-ci, lorsque l'on somme sur toutes les mailles l'équation (5.1.3), on obtient pour la valeur moyenne de la hauteur d'eau :

$$\overline{h}^{n+1} = \frac{1}{N}\sum_{j=0}^{N-1} h_j^{n+1} = \overline{h}^n - \frac{\Delta t}{\Delta x}\left(h_{N-\frac{1}{2}}^{*,n+1-} u_{N-\frac{1}{2}}^{*,n+1-} - h_{0-\frac{1}{2}}^{*,n+1-} u_{0-\frac{1}{2}}^{*,n+1-}\right).$$

Pour avoir un schéma conservatif ($\overline{h}^{n+1} = \overline{h}^n$), il suffit que $u_{N-\frac{1}{2}}^{*,n+1-} = u_{0-\frac{1}{2}}^{*,n+1-}$ et $h_{N-\frac{1}{2}}^{*,n+1-} = h_{0-\frac{1}{2}}^{*,n+1-}$. On rajoute 8 lignes à la matrice du système à inverser, dans lesquelles on impose :

$$\begin{aligned} (u,\Pi)_{-1}^{n+1-} &= (u,\Pi)_{N-1}^{n+1-}, & (u,\Pi)_{-2}^{n+1-} &= (u,\Pi)_{N-2}^{n+1-}, \\ (u,\Pi)_{N}^{n+1-} &= (u,\Pi)_{0}^{n+1-}, & (u,\Pi)_{N+1}^{n+1-} &= (u,\Pi)_{1}^{n+1-}. \end{aligned}$$

On a alors clairement $u_{N-\frac{1}{2}}^{*,n+1-} = u_{0-\frac{1}{2}}^{*,n+1-}$ et on peut vérifier à l'aide du système (5.1.5) que $h_{N-\frac{1}{2}}^{*,n+1-} = h_{0-\frac{1}{2}}^{*,n+1-}$.

Conditions de bord réfléchissant. On rajoute 4 lignes à la matrice du système à inverser, dans lesquelles on impose : $(u,\Pi)_{-1}^{n+1-} = (-u,\Pi)_{0}^{n+1-}$ et $(u,\Pi)_{-2}^{n+1-} = (-u,\Pi)_{1}^{n+1-}$. On a clairement $u_{0-\frac{1}{2}}^{*,n+1-} = 0$ comme souhaité. De la même manière pour le bord droit : $(u,\Pi)_{N}^{n+1-} = (-u,\Pi)_{N-1}^{n+1-}$ et $(u,\Pi)_{N+1}^{n+1-} = (-u,\Pi)_{N-2}^{n+1-}$.

Conditions de bord absorbant. On rajoute 4 lignes à la matrice du système à inverser, dans lesquelles on impose : $(u,\Pi)_{-1}^{n+1-} = (u,\Pi)_{0}^{n+1-}$ et $(u,\Pi)_{-2}^{n+1-} = (u,\Pi)_{0}^{n+1-}$. On a clairement $u_{0-\frac{1}{2}}^{*,n+1-} = u_0^{n+1-}$ comme souhaité. On fait de même pour le bord droit.

Conditions de bord de Dirichlet. On rajoute 4 lignes à la matrice du système à inverser, dans lesquelles on impose : $(u,\Pi)_{-1}^{n+1-} = (u,\Pi)_{-2}^{n+1-} = (u,\Pi)_{0-1/2}^{\mathrm{Dir}}(t^n)$. Contrairement aux cas précédents pour lesquels on n'ajoutait que des zéros au second membre, ici ce dernier contient les valeurs à imposer au bord. On fait de même pour le bord droit.

5.1.3 Schémas Galerkin discontinus d'ordres élevés

Étapes principales du code

On détaille ici l'implémentation de l'extension à l'ordre élevé du schéma implicite-explicite uniquement, le cas explicite-explicite étant plus simple et ne demandant pas de point technique supplémentaire à détailler.

On prends $p > 0$ un entier, on note alors $(s_0 = -1, s_1, \ldots, s_{p-1}, s_p = 1)$ les $p+1$ points de quadrature de Gauss-Lobatto, ℓ_i le polynôme de Lagrange tel que $\ell_i(s_k) = \delta_{i,k}$, où $\delta_{i,k} = 1$ si $i = k$ et 0 sinon, et $w_i = \int_{-1}^{1} \ell_i(s)\,\mathrm{d}s$ le $i^{\text{ème}}$ poids. Chaque variable φ est discrétisée en un polynôme de degré p sur la maille arbitraire $\kappa_j = \left[x_{j-1/2}, x_{j+1/2}\right]$, représenté par ses $p+1$ degrés de liberté $(\varphi_{0,j}, \ldots, \varphi_{p,j})$. Le schéma global s'écrit alors

$$\left\{\begin{aligned}
h_{i,j}^{n+1} &= h_{i,j}^{n} - \frac{2\Delta t}{w_i \Delta x}\Bigg[\delta_{i,p} h_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,n+1-} - \delta_{i,0} h_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,n+1-} \\
&\qquad\qquad - \sum_{k=0}^{p} w_k h_{k,j}^{n+1-} u_{k,j}^{n+1-} \ell_i'(s_k)\Bigg], \\
(hu)_{i,j}^{n+1} &= (hu)_{i,j}^{n} - \frac{2\Delta t}{w_i \Delta x}\left[\delta_{i,p}\Pi_{j+1/2}^{*,n+1-} - \delta_{i,0}\Pi_{j-1/2}^{*,n+1-} - \sum_{k=0}^{p} w_k \Pi_{k,j}^{n+1-} \ell_i'(s_k)\right] \\
&\quad - \frac{2\Delta t}{w_i \Delta x}\Bigg[\delta_{i,p}(hu)_{j+1/2}^{*,n+1-} u_{j+1/2}^{*,n+1-} - \delta_{i,0}(hu)_{j-1/2}^{*,n+1-} u_{j-1/2}^{*,n+1-} \\
&\qquad\qquad - \sum_{k=0}^{p} w_k (hu)_{k,j}^{n+1-} u_{k,j}^{n+1-} \ell_i'(s_k)\Bigg] \\
&\quad - \frac{2\Delta t}{w_i \Delta x}\left[\frac{\delta_{i,p}}{2}\left\{gh\Delta z\right\}_{j+\frac{1}{2}}^{n} + \frac{\delta_{i,0}}{2}\left\{gh\Delta z\right\}_{j-\frac{1}{2}}^{n}\right] \\
&\quad - \Delta t \sum_{k=0}^{p}\left[gh_{i,j}^{n}\left(h_{k,j}^{n} + z_{k,j}\right) - \Pi_{k,j}^{n}\right]\ell_k'(s_i),
\end{aligned}\right. \tag{5.1.6}$$

où les flux dépendent cette fois-ci du dernier degré de liberté de la maille de gauche et du premier de celle de droite, à savoir

$$\begin{aligned}
\left\{gh\Delta z\right\}_{j+1/2}^{n} &= g\frac{h_{p,j}^{n} + h_{0,j+1}^{n}}{2}\left(z_{0,j+1} - z_{p,j}\right), \\
u_{j+1/2}^{*,n+1-} &= \frac{u_{p,j}^{n+1-} + u_{0,j+1}^{n+1-}}{2} - \frac{1}{2a_{j+1/2}}\left(\Pi_{0,j+1}^{n+1-} - \Pi_{p,j}^{n+1-}\right) - \frac{1}{2a_{j+1/2}}\left\{gh\Delta z\right\}_{j+1/2}^{n}, \\
\Pi_{j+1/2}^{*,n+1-} &= \frac{\Pi_{p,j}^{n+1-} + \Pi_{0,j+1}^{n+1-}}{2} - \frac{a_{j+1/2}}{2}\left(u_{0,j+1}^{n+1-} - u_{p,j}^{n+1-}\right),
\end{aligned}$$

et $a_{j+1/2} = \max\left(h^n_{p,j}\sqrt{gh^n_{p,j}}, h^n_{0,j+1}\sqrt{gh^n_{0,j+1}}\right)$. Comme pour le schéma Volumes Finis implicite-explicite, le pas acoustique se décompose en la résolution d'un système linéaire, qui s'écrit ici

$$\begin{cases} u^{n+1-}_{i,j} = u^n_{i,j} - \dfrac{2\Delta t}{w_i\Delta x}\tau^n_{i,j}\left[\delta_{i,p}\Pi^{*,n+1-}_{j+1/2} - \delta_{i,0}\Pi^{*,n+1-}_{j-1/2} - \displaystyle\sum_{k=0}^{p} w_k\Pi^{n+1-}_{k,j}\ell'_i(s_k)\right] \\ \qquad\qquad - \dfrac{2\Delta t}{w_i\Delta x}\tau^n_{i,j}\left[\dfrac{\delta_{i,p}}{2}\left\{gh\Delta z\right\}^n_{j+\frac{1}{2}} + \dfrac{\delta_{i,0}}{2}\left\{gh\Delta z\right\}^n_{j-\frac{1}{2}}\right] \\ \qquad\qquad - \Delta t\,\tau^n_{i,j}\displaystyle\sum_{k=0}^{p}\left[gh^n_{i,j}\left(h^n_{k,j} + z_{k,j}\right) - \Pi^n_{k,j}\right]\ell'_k(s_i), \\ \Pi^{n+1-}_{i,j} = \Pi^n_{i,j} - \dfrac{2\Delta t}{w_i\Delta x}\tau^n_{i,j}a_j^2\left[\delta_{i,p}u^{*,n+1-}_{j+1/2} - \delta_{i,0}u^{*,n+1-}_{j-1/2} - \displaystyle\sum_{k=0}^{p} w_k u^{n+1-}_{k,j}\ell'_i(s_k)\right], \end{cases} \tag{5.1.7}$$

puis en la mise à jour des variables conservatives comme suit

$$\begin{cases} L_{i,j} = 1 + \dfrac{2\Delta t}{w_i\Delta x}\left[\delta_{i,p}u^{*,n+1-}_{j+1/2} - \delta_{i,0}u^{*,n+1-}_{j-1/2} - \displaystyle\sum_{k=0}^{p} w_k u^{n+1-}_{k,j}\ell'_i(s_k)\right], \\ L_{i,j}h^{n+1-}_{i,j} = h^n_{i,j} \\ L_{i,j}(hu)^{n+1-}_{i,j} = (hu)^n_{i,j} - \dfrac{2\Delta t}{w_i\Delta x}\left[\delta_{i,p}\Pi^{*,n+1-}_{j+1/2} - \delta_{i,0}\Pi^{*,n+1-}_{j-1/2} - \displaystyle\sum_{k=0}^{p} w_k\Pi^{n+1-}_{k,j}\ell'_i(s_k)\right] \\ \qquad\qquad - \dfrac{2\Delta t}{w_i\Delta x}\left[\dfrac{\delta_{i,p}}{2}\left\{gh\Delta z\right\}^n_{j+\frac{1}{2}} + \dfrac{\delta_{i,0}}{2}\left\{gh\Delta z\right\}^n_{j-\frac{1}{2}}\right] \\ \qquad\qquad - \Delta t\displaystyle\sum_{k=0}^{p}\left[gh^n_{i,j}\left(h^n_{k,j} + z_{k,j}\right) - \Pi^n_{k,j}\right]\ell'_k(s_i). \end{cases} \tag{5.1.8}$$

La fonction `SolverStep()` est la même que celle du schéma Volumes Finis 5.1.2, mise à part que le système est plus grand et que chaque boucle sur les mailles contient en plus une boucle sur les $p+1$ degrés de liberté.

Remarque 5.1.1. *On peut calculer une fois pour toutes au début de la simulation les points s_i et les poids w_i de la quadrature de Gauss-Lobatto, ainsi que la matrice des dérivées des polynômes de Lagrange en ces points, à savoir $M_{i,j} = \ell'_j(s_i)$.*

Traitement du bord

On explicite enfin le traitement des conditions de bord pour les schémas Galerkin discontinus. Habituellement, pour un schéma DG à un seul pas avec des points de Gauss-Lobatto, il suffit de donner les valeurs du premier degré de liberté de l'autre côté de l'interface de bord du domaine physique (par exemple $\varphi_{p,-1}$ pour l'interface de gauche). En effet, grâce à ces valeurs on peut calculer les états étoilés du problème de Riemann de l'interface de bord et donc évaluer les flux entrants et sortants aux bords du domaine. Ici, à cause des deux sous pas de temps, on a besoin de donner les valeurs de tous les degrés de liberté de la première maille de l'autre

côté de l'interface de bord, en plus d'un degré dans la maille suivante. (Par exemple $\varphi^n_{i,-1}$, avec $i = 0, \ldots, p$, et également $\varphi^n_{p,-2}$ pour l'interface de gauche.) Dans la pratique, on remplira tous les degrés de liberté des quatre mailles fantômes même s'ils n'interviennent pas tous dans la solution au pas de temps suivant. On a représenté ces degrés de liberté pour les premières mailles et avec $p = 4$ dans la Figure 5.1.3.

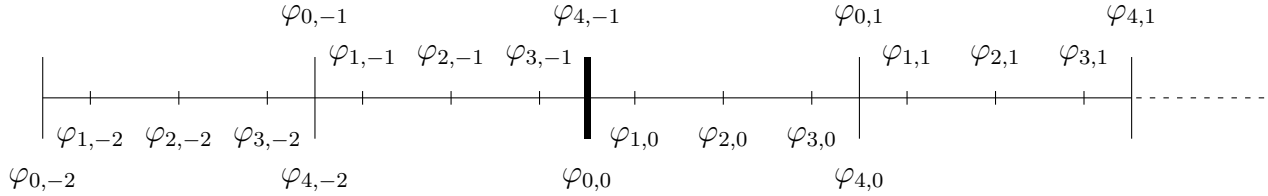

FIGURE 5.1.3 – Degrés de liberté dans les premières mailles du maillage avec $p = 4$, degré des polynômes de base.

Conditions de bords périodiques. On impose

$$(h, hu)^n_{i,-2} = (h, hu)^n_{i,N-2}, \qquad (h, hu)^n_{i,-1} = (h, hu)^n_{i,N-1},$$
$$(h, hu)^n_{i,N} = (h, hu)^n_{i,0}, \qquad (h, hu)^n_{i,N+1} = (h, hu)^n_{i,1},$$

ainsi que

$$(u, \Pi)^{n+1-}_{i,-2} = (u, \Pi)^{n+1-}_{i,N-2}, \qquad (u, \Pi)^{n+1-}_{i,-1} = (u, \Pi)^{n+1-}_{i,N-1},$$
$$(u, \Pi)^{n+1-}_{i,N} = (u, \Pi)^{n+1-}_{i,0}, \qquad (u, \Pi)^{n+1-}_{i,N+1} = (u, \Pi)^{n+1-}_{i,1},$$

avec $i = 0, \ldots, p$. Comme pour les schémas Volumes Finis, on copie les polynômes solutions des deux premières mailles $j = 0$ et $j = 1$ dans les mailles fictives $j = N$ et $j = N + 1$ et inversement des deux dernières mailles $j = N - 2$ et $j = N - 1$ dans les mailles fictives $j = -2$ et $j = -1$. On vérifie alors que

$$h^{*,n+1-}_{N-\frac{1}{2}} = h^{*,n+1-}_{0-\frac{1}{2}}, \qquad u^{*,n+1-}_{N-\frac{1}{2}} = u^{*,n+1-}_{0-\frac{1}{2}},$$

et on obtient bien un schéma conservatif. En effet, si on regarde l'évolution de la valeur moyenne pour la hauteur d'eau, on a encore

$$\begin{aligned}\overline{h}^{n+1} &= \frac{1}{N}\sum_{j=0}^{N-1}\frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} h_j^{n+1}(x)\,\mathrm{d}x = \frac{1}{N}\sum_{j=0}^{N-1}\sum_{i=0}^{p}\frac{w_i}{2}h_{i,j}^{n+1}\\ &= \overline{h}^n - \frac{\Delta t}{\Delta x}\left(h^{*,n+1-}_{N-\frac{1}{2}}u^{*,n+1-}_{N-\frac{1}{2}} - h^{*,n+1-}_{0-\frac{1}{2}}u^{*,n+1-}_{0-\frac{1}{2}}\right)\end{aligned}$$

grâce à l'égalité $\sum_{i=0}^{p} \ell_i'(s_k) = 0$, $\forall k \in \{0, \ldots, p\}$.

Conditions de bord réfléchissant. On impose

$$(h, hu)^n_{i,-1} = (h, -hu)^n_{p-i,0}, \qquad (h, hu)^n_{i,-2} = (h, -hu)^n_{p-i,1},$$
$$(u, \Pi)^{n+1-}_{i,-1} = (-u, \Pi)^{n+1-}_{p-i,0}, \qquad (u, \Pi)^{n+1-}_{i,-2} = (-u, \Pi)^{n+1-}_{p-i,1},$$

avec $i = 0, \ldots, p$. Remarquons que l'on a inversé l'ordre des degrés de liberté pour la symétrie. On vérifie que l'on a bien $u^{*,n+1-}_{0-\frac{1}{2}} = 0$. On procède de même pour le bord droit, à savoir

$$(h, hu)^n_{i,N} = (h, -hu)^n_{p-i,N-1}, \qquad (h, hu)^n_{i,N+1} = (h, -hu)^n_{p-i,N-2},$$
$$(u, \Pi)^{n+1-}_{i,N} = (-u, \Pi)^{n+1-}_{p-i,N-1}, \qquad (u, \Pi)^{n+1-}_{i,N+1} = (-u, \Pi)^{n+1-}_{p-i,N-2}.$$

Conditions de bord absorbant. On impose

$$(h, hu)^n_{i,-1} = (h, hu)^n_{0,0}, \qquad (h, hu)^n_{i,-2} = (h, hu)^n_{0,0},$$
$$(u, \Pi)^{n+1-}_{i,-1} = (u, \Pi)^{n+1-}_{0,0}, \qquad (u, \Pi)^{n+1-}_{i,-2} = (u, \Pi)^{n+1-}_{0,0},$$

avec $i = 0, \ldots, p$. Remarquons que les polynômes sont ainsi constants dans les mailles fantômes. On vérifie que l'on a bien $u^{*,n+1-}_{0-\frac{1}{2}} = u^{n+1-}_{0,0}$. On fait de même pour le bord droit.

Conditions de bord de Dirichlet. On impose

$$(h, hu)^n_{i,-1} = (h, hu)^n_{i,-2} = (h, hu)^{\mathrm{Dir}}_{0-1/2}(t^n), \qquad (u, \Pi)^{n+1-}_{i,-1} = (u, \Pi)^{n+1-}_{i,-2} = (u, \Pi)^{\mathrm{Dir}}_{0-1/2}(t^n),$$

avec $i = 0, \ldots, p$. Remarquons que les polynômes sont encore constants dans les mailles fantômes. On fait de même sur le bord droit, avec des conditions différentes sur les variables.

5.1.4 Résolution du système linéaire à l'aide de Petsc en 1D

Système sous forme de résidus explicites

Si on réécrit le système linéaire sous la forme simplifiée

$$\left(\mathrm{I}_{2(p+1)(N+2)} + \frac{\Delta t}{\Delta x} A\right) x^{n+1-} = x^n - \frac{\Delta t}{\Delta x} b,$$

avec $x = (u_{-2}, \Pi_{-2}, u_{-1}, \Pi_{-1}, \ldots, u_{N+1}, \Pi_{N+1})$, et $\varphi_j = (\varphi_{0,j}, \ldots, \varphi_{p,j})$, $\varphi = u, \Pi$, on peut alors montrer qu'il est équivalent au système avec résidus explicites suivant

$$\left(\mathrm{I}_{2(p+1)(N+2)} + \frac{\Delta t}{\Delta x} A\right) (x^{n+1-} - x^n) = -\frac{\Delta t}{\Delta x} (Ax^n + b). \tag{5.1.9}$$

La version implémentée correspond à cette dernière expression du système à résoudre. En effet, sous cette forme, on peut alors tester si la solution au temps t^n est une solution stationnaire, soit $Ax^n + b = 0$ à erreurs machines près. Lorsque c'est le cas, on peut alors imposer $x^{n+1-} = x^n$ et éviter des erreurs trop grandes lors de la résolution du système qui nous écarteraient de cette solution stationnaire. Les erreurs peuvent être évaluées très précisément avec le nombre d'opérations et l'ordre de grandeur des quantités mises en jeu.

Méthodes C++ pour résoudre le système

Lors de la construction de l'objet `solver` de la classe `IMEXDG`, on initialise et alloue la matrice et vecteurs nécessaires. On utilise les fonctions PETSc `MatCreate()`, `MatSetSize()`, `MatSetType()` et `MatSetPreallocation()` pour créer une matrice `m_A` de taille globale

$$(2(p+1)(N+4))^2 .$$

Elle est creuse par blocs de taille $(p+1)^2$, et chaque ligne de blocs contient 6 blocs non nuls. Le solveur et le pré-conditionneur fournis par PETSc et utilisés à chaque étape pour résoudre le système sont respectivement GMRES [3] et bloc Jacobi. On impose une précision relative de 2.22×10^{-15} pour cette résolution. On a enfin besoin de 3 vecteurs `m_x`, `m_v` et `m_y` créés avec les bonnes dimensions par rapport à la matrice `m_A` gâce à la fonction `MatCreateVecs()`.

L'étape acoustique se décompose comme suit :

```
IMEXDG::AcousticStep(STATE& stateOld, double dt) {
    this->BuildLinearSystem(stateOld, dt);
    this->FillGhostsMat();
    this->SolveLinearSystem(dt);
}
```

Les fonctions `BuildLinearSystem()` et `FillGhostsMat()` remplissent la matrice `m_A` comme la matrice A dans l'Equation (5.1.9).

Enfin la fonction qui résout le système a été implémentée de manière schématique comme suit :

```
SolveLinearSystem(double dx, double dt) {
    m_y = m_A*m_v + m_y; // m_y = A*x^n+b

    // test WB :
    int isLakeAtRest = (m_y == 0) ? 1 : 0;

    if (isLakeAtRest == 1) {
        // steady state :
        m_x = 0;
    } else {
        // solve the system :
        m_A = dt/dx * m_A;          // m_A = dt/dx * A
        m_A = I + m_A;              // m_A = I + dt/dx * A
        m_y = -dt/dx * m_y;         // m_y = -dt/dx * (A*x^n+b)
        m_x = m_A^{-1} * m_y;       // m_x = x^{n+1-} - x^n
    }

    m_x = m_x + m_v;
    // m_x = x^n si isLakeAtRest == 1
    // m_x = x^{n+1-} sinon
}
```

5.1.5 Détails sur le code C++ pour la boucle MOOD

On commence par l'approche naturelle, on va décrire la fonction `SolverStep()`, de l'objet `MOODv1` de la classe `SOLVER`, mettant à jour les données au temps t^{n+1}, stockées dans la structure `stateNew`, grâce aux données au temps t^n, sockées dans la structure `stateOld`. Ces deux jeux de données sont stockées avec un nombre de degrés de liberté correspondant à la discrétisation Galerkin discontinue (DG). L'objet `MOODv1` contient des attributs pouvant stocker les représentation Volumes Finis (FV) aux deux instants t^{n+1} et t^n : `m_stateNFV` et `m_stateOFV`. Il contient également deux attributs correspondant aux deux solveurs FV et DF : `m_solverFV` et `m_solverDG`. Il contient enfin un attribut `m_detect` pour savoir quelles mailles ont été détectées et un attribut `m_flux` permettant de stocker la valeur des flux de la méthode DG pour adapter les flux aux interfaces des mailles détectées afin de garder un schéma globalement conservatif.

La fonction `SolverStep` s'écrit alors :

```
1  MOODv1::SolverStep(STATE& stateNew, STATE& stateOld, double dt) {
2      // Projection de DG stateOld vers FV m_stateOFV
3      Projection(m_stateOFV, stateOld);
4
5      // Solver DG
6      m_solverDG.SolverStep(stateNew, stateOld, m_flux, dt);
7
8      // Projection de DG stateNew vers FV m_stateNFV
9      Projection(m_stateNFV, stateNew);
10
11     // On note les cellules detectees dans m_detect
12     Detectors(&m_detect, m_stateNFV, m_stateOFV);
13
14     // Solver FV seulement sur les mailles detectees
15     m_solverFV.SolverStep(m_stateNFV, m_stateOFV, m_flux,
16                                                   m_detect, dt);
17
18     // Reconstruction de FV m_stateNFV vers DG stateNew
19     // seulement sur les mailles detectees
20     Reconstruct(stateNew, m_stateNFV, &m_detect);
21 }
```

Pour l'approche dite compatible, on change de point de vue puisque cette fois la fonction `SolverStep()` de l'objet `MOODv2` met à jour les représentations FV : `stateNFV` et `stateOFV`. Ainsi, les attributs pour stocker les données sont les représentations DG, appelées `m_stateNDG` et `m_stateODG`. La fonction `SolverStep()` n'est pas très éloignée mais s'écrit dans ce cas :

```
1  MOODv2::SolverStep(STATE& stateNew, STATE& stateOld, double dt) {
2      // m_stateNDG devient m_stateODG et inversement
3      SwapStates(m_stateNDG, m_stateODG);
4
5      // Solver DG
6      m_solverDG.SolverStep(m_stateNDG, m_stateODG, m_flux, dt);
```

```
7
8       // Projection de DG m_stateNDG vers FV stateNew
9       Projection(stateNew, m_stateNDG);
10
11      // On note les cellules detectees dans m_detect
12      Detectors(&m_detect, stateNew, stateOld);
13
14      // Solver FV seulement sur les mailles detectees
15      m_solverFV.SolverStep(stateNew, stateOld, m_flux, m_detect, dt);
16
17      // Reconstruction de FV stateNew vers DG m_stateNDG
18      // seulement sur les mailles detectees
19      Reconstruct(m_stateNDG, stateNew, &m_detect);
20  }
```

On peut remarquer dans cette implémentation des deux approches que l'on retrouve l'incompatibilité pour la première et la compatibilité pour la deuxième. En effet, dans l'approche naturelle, les données au temps t^{n+1} au début du pas $t^{n+1} \to t^{n+2}$ sont obtenues grâce à l'opérateur $\mathcal{P} \circ \mathcal{R} \neq \mathrm{Id}$, appliqué sur les données au temps t^{n+1} à la fin du pas $t^n \to t^{n+1}$. En revanche, dans l'approche compatible, les données `stateNew` au temps t^{n+1} à la fin du pas $t^n \to t^{n+1}$ sont directement redonnées comme entrée `stateOld` au temps t^{n+1} au début du pas $t^{n+1} \to t^{n+2}$.

5.2 Code de simulation à deux dimensions sur maillages non-structurés

5.2.1 Écriture de la matrice du système linéaire en 2D

Si on écrit le système linéaire à résoudre en 2D sans terme source dans la partie acoustique, on obtient :

$$\begin{cases} \mathbf{u}_K^{n+1-} = \mathbf{u}_K^n - \Delta t \sum\limits_{L \in N(K)} \dfrac{|\Gamma_{KL}|}{|\Omega_K|} \Pi_{KL}^{n+1-} \mathbf{n}_{KL}, \\ \Pi_K^{n+1-} = \Pi_K^n - \Delta t \sum\limits_{L \in N(K)} \dfrac{|\Gamma_{KL}|}{|\Omega_K|} a_{KL}^2 u_{KL}^{n+1-}, \end{cases}$$

avec $N(K)$ l'ensemble des indices des mailles voisines de la maille K, $|\Gamma_{KL}|$ la longueur de l'interface entre les mailles K et L, $|\Omega_K|$ la surface de la maille K et $\mathbf{n}_{KL}$ la normale à l'interface Γ_{KL} allant de K vers L. Les flux numériques sont définis comme suit :

$$\begin{aligned} a_{KL} &\geq \max\left(h_K\sqrt{gh_K}, h_L\sqrt{gh_L}\right), \\ \Pi_{KL}^{n+1-} &= \frac{\Pi_K^{n+1-} + \Pi_L^{n+1-}}{2} - \frac{a_{KL}}{2}\left(u_L^{n+1-} - u_K^{n+1-}\right)\mathbf{n}_{KL}^T, \\ \mathbf{u}_{KL}^{n+1-} &= \frac{\mathbf{u}_K^{n+1-} + \mathbf{u}_L^{n+1-}}{2}\mathbf{n}_{KL}^T - \frac{1}{2a_{KL}}\left(\Pi_L^{n+1-} - \Pi_K^{n+1-}\right). \end{aligned}$$

En développant les flux dans le système ci-dessus, on montre qu'il peut s'écrire sous la forme condensée :

$$\left[\mathrm{I}_3 + \Delta t \sum_{L\in N(K)} D_{KL}\right] \cdot \begin{pmatrix}\mathbf{u}\\ \Pi\end{pmatrix}_K^{n+1-} + \Delta t \sum_{L\in N(K)} E_{KL} \cdot \begin{pmatrix}\mathbf{u}\\ \Pi\end{pmatrix}_L^{n+1-} = \begin{pmatrix}\mathbf{u}\\ \Pi\end{pmatrix}_K^{n},$$

avec

$$D_{KL} = \frac{|\Gamma_{KL}|}{|\Omega_K|}\begin{pmatrix} \frac{a_{KL}}{2}\,\mathbf{n}_{KL}\otimes\mathbf{n}_{KL} & \frac{1}{2}\,\mathbf{n}_{KL} \\ \frac{a_{KL}^2}{2}\,\mathbf{n}_{KL}^T & \frac{a_{KL}}{2}\end{pmatrix},$$

$$E_{KL} = \frac{|\Gamma_{KL}|}{|\Omega_K|}\begin{pmatrix} -\frac{a_{KL}}{2}\,\mathbf{n}_{KL}\otimes\mathbf{n}_{KL} & \frac{1}{2}\,\mathbf{n}_{KL} \\ \frac{a_{KL}^2}{2}\,\mathbf{n}_{KL}^T & -\frac{a_{KL}}{2}\end{pmatrix}.$$

En parcourant toutes les interfaces, on peut ainsi remplir la matrice globale, qui est creuse par blocs de taille 3×3. Pour une interface Γ_{KL}, on vient ajouter les matrices D_{KL} sur la ligne de blocs K et la colonne K, la matrice D_{LK} en (L,L), E_{KL} en (K,L) et E_{LK} en (L,K). Enfin, si on suppose par exemple que l'interface Γ_{KL} est une interface de bord avec K une maille interne et L une maille fictive, il faut remplacer dans la formule ci-dessus $\begin{pmatrix}\mathbf{u}\\ \Pi\end{pmatrix}_L^{n+1-}$ par les valeurs correspondantes et ajouter la matrice E_{KL} au bon endroit dans la matrice, ou soustraire à la ligne K du second membre le terme $\Delta t\, E_{KL}\cdot\begin{pmatrix}\mathbf{u}\\ \Pi\end{pmatrix}_L^{n+1-}$.

5.2.2 Résolution du système linéaire à l'aide de Petsc en 2D

Les simulations du Chapitre 2 ont été réalisées à l'aide d'un code en langage C qui a été une base de celui utilisé dans la thèse de M. Girardin [2]. La travail a consister à apporter les modifications nécessaires afin de tenir compte du terme source dans ce code 2D pour maillages non-structurés. On a également rendu possible l'utilisation de maillages structurés avec conditions de bords périodiques pour les cas tests de vortex en translation sans et avec topographie.

On rappelle que le schéma en deux dimensions d'espace pour les équations de Saint-Venant s'écrit comme dans la Section précédente avec des flux légèrement modifiés :

$$\Pi_{KL}^{n+1-} = \frac{\Pi_K^{n+1-}+\Pi_L^{n+1-}}{2} - \frac{a_{KL}}{2}\left(u_L^{n+1-}-u_K^{n+1-}\right)\mathbf{n}_{KL}^T + \frac{1}{2}g\frac{h_K^n+h_L^n}{2}\left(z_L-z_K\right),$$
$$\mathbf{u}_{KL}^{n+1-} = \frac{\mathbf{u}_K^{n+1-}+\mathbf{u}_L^{n+1-}}{2}\mathbf{n}_{KL}^T - \frac{1}{2a_{KL}}\left(\Pi_L^{n+1-}-\Pi_K^{n+1-}\right) - \frac{1}{2a_{KL}}g\frac{h_K^n+h_L^n}{2}\left(z_L-z_K\right).$$

Il a donc fallu apporter des modifications à plusieurs étapes clés du code que l'on précise ici :

— lors du calcul de la CFL qui nécessite $\mathbf{u}_{KL}^n$,
— lors de l'assemblage du second membre dans lequel interviennent le traitement (explicite) du terme source,
— lors du calcul des flux Π_{KL}^{n+1-} et $\mathbf{u}_{KL}^{n+1-}$ après résolution du système.

Bibliographie

[1] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Dave A. May, Lois Curfman McInnes, Richard Tran Mills, Todd Munson, Karl Rupp, Patrick Sanan, Barry F. Smith, Stefano Zampini, Hong Zhang, and Hong Zhang. PETSc Web page. `http://www.mcs.anl.gov/petsc`, 2018.

[2] Mathieu Girardin. *Asymptotic preserving and all-regime Lagrange-Projection like numerical schemes : application to two-phase flows in low mach regime.* Theses, Université Pierre et Marie Curie - Paris VI, December 2014.

[3] Youcef Saad and Martin H Schultz. Gmres : A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, 7(3) :856–869, 1986.

Conclusions et perspectives

Dans ces travaux, on a proposé plusieurs extensions d'un schéma de type Lagrange-projection. Ce schéma permet de réaliser une décomposition naturelle des phénomènes physiques contenus dans les équations auxquelles on s'intéresse. Il est donc particulièrement adapté à des systèmes dans lesquels il y a un paramètre qui peut être très petit localement, comme le nombre de Mach pour Euler ou de Froude pour Saint-Venant. On a réussi à écrire une version tout régime volumes finis pour les équations de Saint-Venant en faisant attention au traitement du terme source afin de pouvoir préserver certains états stationnaires. Dans le cadre de la dimension 1 d'espace, on a également proposé une extension Galerkin discontinue avec des résultats prometteurs au vue du gain en efficacité lorsque l'on augmente les degrés des polynômes. On a également étudié certaines possibilités de combinaison avec une méthodologie innovante de limitations, MOOD, qui se démocratise dans l'utilisation de méthodes d'ordres élevés afin de palier certaines de leurs mauvaises propriétés. Une part importante du travail réalisé a consisté à implémenter ces schémas implicites-explicites dans deux codes : un pour les schémas DG avec la méthodologie MOOD en 1D, et un pour les schémas FV sur maillages non-structurés en 2D.

On est loin d'avoir résolu toutes les difficultés soulevées par la montée en ordre grâce aux méthodes DG. Une extension naturelle du travail effectué serait de proposer et implémenter une version 2D du schéma. Au préalable, il serait intéressant de chercher des façons de combiner les deux sous pas, acoustique et transport, avec une méthode en temps adaptée, de façon à obtenir les ordres de convergence théoriques. Il semble que l'erreur de troncature en temps ne provienne pas de la décomposition d'opérateur puisque l'étape transport s'écrit comme une compensation entre l'étape acoustique et le système conservatif global. Cependant, le traitement à l'ordre 1 de la partie acoustique à chaque sous pas Runge-Kutta du schéma global pourrait en être le responsable. Il reste à trouver comment monter en ordre sur cette première étape sans briser la conservativité ou les bonnes propriétés sur la stabilité du schéma.

La collaboration entamée avec Raphaël Loubère pour le couplage des schémas Lagrange-projection avec la stratégie MOOD mérite d'être poursuivie. La première étape serait de rendre possible la simulation à l'aide du schéma combinée avec la deuxième approche, dite compatible. Il y a également une recherche importante à réaliser dans le choix de détecteurs adaptés à ces schémas et aux équations de Saint-Venant. On souhaite trouver un bon compromis entre une détection nécessaire des cellules dans lesquelles la solution n'est pas satisfaisante, et suffisante pour ne pas détecter trop de cellules inutilement et augmenter démesurément le coût de calcul. Comme extension à plus long terme, le couplage des méthodes Lagrange-projection implicite-explicite avec la stratégie MOOD semble être délicate. Il faut à la fois savoir résoudre un système linéaire avec des mailles de natures différentes et dans un temps raisonnable lors du recalcule

de la solution.

L'application de cette stratégie acoustique-transport à d'autres systèmes contenant un paramètre pouvant être petit constitue également de nouvelles ouvertures. T. Padioleau à la Maison de la Simulation s'intéresse notamment à les appliquer à Euler avec gravité. On peut penser au système de la magnétohydrodynamique (MHD) dans lequel apparaît plusieurs nombres adimensionnés, on peut mentionner le nombre de Mach et le nombre d'Alfvén.

Enfin, il serait intéressant d'étudier l'utilisation de ce type de méthodes, avec décomposition d'opérateurs et implicite-explicite, dans un cadre de calcul haute performance. Les questions de répartitions des données et de résolution de systèmes linéaires de manières parallèles, couplées à des techniques de raffinement de maillage, restent ouvertes. La Maison de la Simulation développe un code, CanoP, permettant d'utiliser un maillage en octree qui peut être raffiné ou dé-raffiné au cours de la simulation. Cet outil offre ainsi la possibilité de tester des méthodes explicites avec une gestion parallélisée des données et des calculs, tout en augmentant la précision de la solution en raffinant le maillage uniquement dans les zones contenant des phénomènes physiques difficiles à approcher. Cependant le couplage avec une librairie telle que PETSc pour pouvoir utiliser des méthodes implicites n'a pas encore été réalisé. De même, l'extension vers des méthodes ordres élevés est compliquée car, soit on augmente le stencil de la méthode pour les méthodes VF, soit on a plusieurs degrés de liberté par maille, et notamment sur les bords des mailles, pour les méthodes DG. Dans les deux cas, la recherche des voisins sur un maillage non conforme et avec des tailles de mailles différentes peut s'avérer compliqué.

université PARIS-SACLAY
École doctorale de mathématiques Hadamard (EDMH)

Titre : Simulation numérique d'écoulements compressibles complexes par des méthodes de type Lagrange-projection : applications aux équations de Saint-Venant.

Mots Clefs : Équations de Saint-Venant, décomposition de type Lagrange-projection, schéma numérique implicite-explicite, grands pas de temps, bas nombre de Froude, propriété équilibre.

Résumé : On étudie dans le cadre de la thèse une famille de schémas numériques permettant de résoudre les équations de Saint-Venant. Ces schémas utilisent une décomposition d'opérateur de type Lagrange-projection afin de séparer les ondes de gravité et les ondes de transport. Un traitement implicite du système acoustique (relié aux ondes de gravité) permet aux schémas de rester stable avec de grands pas de temps. La correction des flux de pression rend possible l'obtention d'une solution approchée précise quel que soit le régime d'écoulement vis-à-vis du nombre de Froude. Une attention toute particulière est portée sur le traitement du terme source qui permet la prise en compte de l'influence de la topographie. On obtient notamment la propriété dite équilibre permettant de conserver exactement certains états stationnaires, appelés état du « lac au repos ». Des versions 1D et 2D sur maillages non-structurés de ces méthodes ont été étudiées et implémentées dans un cadre volumes finis. Enfin, une extension vers des méthodes ordres élevés Galerkin discontinue a été proposée en 1D avec des limiteurs classiques ainsi que combinée avec une boucle MOOD de limitation *a posteriori*.

Title : Numerical simulation of complex compressible flows by Lagrange-projection type methods : applications to shallow water equations.

Keys words : Shallow water equations, Lagrange-projection type splitting, implicit-explicit numerical scheme, large time steps, low Froude, well-balanced property.

Abstract : In this thesis we study a family of numerical schemes solving the shallow water equations system. These schemes use a Lagrange-projection like splitting operator technique in order to separate the gravity waves and the transport waves. An implicit-explicit treatment of the acoustic system (linked to the gravity waves) allows the schemes to stay stable with large time step. The correction of the pressure fluxes enables the obtain of a precise approximation solution whatever the regime flow is with respect to the Froude number. A particular attention has been paid over the source term treatment which permits to take the topography into account. We especially obtain the so-called well-balanced property giving the exact conservation of some steady states, namely the "lake at rest" state. 1D and 2D versions of this methods have been studied and implemented in the finite volumes framework. Finally, a high order discontinuous Galerkin extension has been proposed in 1D with classical limiters along with a combined MOOD loop *a posteriori* limiting strategy.