Aix*Marseille université
Socialement engagée

NNT/NL : 2021AIXM0187/001ED184

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 29 Mars 2021 par

Youssouf NASSERI

Analyse numérique de schémas volumes finis à mailles décalées pour certains systèmes hyperboliques issus de la mécanique des fluides

Discipline
Mathématiques

Spécialité
Mathématiques appliquées

École doctorale
184 Mathématique et Informatique

Laboratoire/Partenaires de recherche
Institut de Mathématiques de Marseille,
Institut de Rapiotrotection et Sûreté Nucléaire

INSTITUT DE MATHÉMATIQUES DE MARSEILLE

IRSN

Composition du jury

Enrique D. FERNÀNDEZ-NIETO Université de Séville	Rapporteur
Nicolas SEGUIN Université de Rennes 1	Rapporteur
Robert EYMARD Université de Marne la Vallée	Examinateur
Charlotte PERRIN Aix-Marseille Université	Examinatrice
Antonin NOVOTNY Université de Toulon	Examinateur
Raphaèle HERBIN Aix-Marseille Université	Directrice de thèse
Jean-Claude LATCHÉ IRSN Cadarache	co-directeur de thèse

Je soussigné, Youssouf Nasseri, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Raphaèle Herbin et Jean-Claude Latché, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le ···

signature

Résumé

Cette thèse s'inscrit dans la continuité de collaborations entre l'IRSN (Institut de Radioprotection et Sûreté Nucléaire) et l'I2M (Institut de Mathématiques de Marseille) sur le développement et l'analyse de schémas de discrétisation en temps et en espace pour la résolution numérique de certains problèmes de mécanique des fluides. La première partie de ce manuscrit concerne les équations de Saint-Venant. On propose une analyse d'un schéma numérique pour les équations Saint-Venant avec gradient de fond, avec un schéma de Heun en temps et un schéma MUSCL en espace pour des volumes finis sur grilles à mailles décalées (schéma MAC). La stabilité du schéma est démontrée, ainsi qu'un résultat de consistance "à la Lax" pour un opérateur général de convection non linéaire sur maillages décalées, qui s'applique à tous les systèmes de lois de conservation. Des tests numériques sont effectués pour établir la validité du schéma. On s'intéresse aussi aux mêmes équations, mais avec un terme source qui modélise la force de Coriolis pour la modélisation d'écoulements géostrophiques. La discrétisation MAC upwind est comparée à une discrétisation par éléments finis de type Rannacher-Turek avec une stabilisation qui permet de réduire la diffusion. Des résultats numériques permettent de comparer les deux schémas avec une résolution de type Godunov. Ensuite, on considère les équations de Saint-Venant en une dimension d'espace couplées avec une équation dite "d'Exner", qui modélise le transport de sédiment. Une régularisation de la loi de frottement permet d'obtenir un bilan d'énergie. Plusieurs formules de flux de sédiment déjà proposées dans la littérature sont étudiées. Les équations résultantes sont discrétisées par un schéma explicite par équation en temps et un schéma à mailles décalées en espace. Le tout est illustré par des résultats numériques.
La deuxième partie est consacrée à la résolution numérique d'un modèle de simulation de déflagration turbulente régi par les équations d'Euler réactif. La modélisation de la combustion est basée sur une approche phénoménologique : la propagation de la flamme est représentée par le transport de la fonction caractéristique de la zone brûlée, où la réaction chimique est complète ; en dehors de cette zone, l'atmosphère reste à l'état frais. Numériquement on adopte une approche de type pénalisation, c'est-à-dire en utilisant un taux de conversion fini avec un temps caractéristique tendant vers zéro avec les pas d'espace et de temps. Ici encore, le schéma numérique est à maillage décalé, et l'algorithme en temps consiste à résoudre d'abord les bilans de masse des espéces chimiques, puis, les bilans de masse, de quantité de mouvement et d'énergie du fluide. Des propriétés de stabilité sont démontrées, et on observe numériquement que la procédure de pénalisation converge. Une solution exacte pour le problème de la déflagration sphérique modélisée par les équations d'Euler réactif

est construite, dans le but d'obtenir une solution de référence pour les tests du code PREMICS d'incendie et sûreté nucléaire de l'IRSN.

Mots clés : Équations d'Euler, équations de Saint-Venant, équation d'exner, interpolation MUSCL, schéma de Heun, méthode de correction de pression, maillages décalés, analyse de consistance.

Abstract

This thesis is a continuation of collaborations between IRSN and I2M on the development and analysis of discretization schemes in time and space for the numerical resolution of certain fluid mechanics problems. The first part of this thesis concerns the shallow water equations. We propose an analysis of a numerical scheme for the shallow water equations with a gradient of the topography, based on a Heun scheme in time combined with a MUSCL scheme in space for finite volumes on staggered grids (MAC scheme). The stability of the scheme is proven, as well as a "Lax consistency" property. In addition, a lemma of consistency "in the sens of Lax" for a general operator of non-linear convection on staggered mesh grids is proved, which is applicable to all conservation law systems. Numerical tests are carried out to establish the validity of the scheme. We are also interested in the same equations, but with a source term that models the Coriolis force for modelling geostrophic flows. The MAC upwind discretization is compared to a Rannacher-Turek finite element discretization with a stabilization technique that reduces diffusion. Numerical results allow to compare the two schemes with a Godunov type solver. Then, the shallow water equations are considered in one dimension of space coupled with a so-called Exner equation, which models the sediment transport. A regularization of the friction law allows us to obtain an energy balance. Several sediment flow formulae already proposed in the literature are studied. The resulting equations are discretized by an explicit scheme equation by equation in time and by a staggered scheme in space. The whole is illustrated by numerical results.
The second part is devoted to the numerical resolution of a turbulent deflagration simulation model governed by reactive Euler equations. Combustion modelling is based on a phenomenological approach: flame propagation is represented by the transport of the characteristic function of the burnt zone, where the chemical reaction is complete; outside this zone, the atmosphere remains at fresh state. Numerically a penalty type approach is adopted, *i.e.* using a finite conversion rate with a characteristic time tending towards zero with space and time steps. Here again, the numerical scheme is with staggered meshes, and the time algorithm consists in solving first the mass balances of the chemical species, then the mass, momentum and energy balances. Stability properties are demonstrated, and it is numerically observed that the penalty procedure converges. An exact solution for the problem of spherical deflagration modelled by the reactive Euler equations is built, in order to obtain a reference solution for the tests of the IRSN's PREMICS fire and nuclear safety code.

Keywords: Euler equations, shallow water equations, Exner equation, MUSCL-like

interpolation, Heun scheme, pressure correction scheme, staggered discretization, numerical analysis.

Remerciements

Ce manuscrit ne verrait pas le jour sans le soutien sans faille de mes directeurs de thèse Raphaèle Herbin et Jean-Claude Latché, je leur en suis éternellement reconnaissant. Merci encore Raphaèle de m'avoir intégré dans l'équipe et Jean-Claude de m'avoir accueilli à l'IRSN, je suis heureux de faire partie de la longue chaîne de leurs étudiants de thèses. J'adresse ma gratitude au personnel de LIE de l'IRSN Cadarache pour leur collaboration.

J'ai l'immense joie d'adresser mes remerciements à Thierry Gallouët, pour l'intérêt qu'il a accordé à mes questions et pour sa disponibilité lors de nos réunions d'avancement ainsi que pour le rôle logistique qu'il a accompli au bénéfice de l'aboutissement de cette thèse.

Je désire remercier les membres du Jury, Enrique D. Fernàndez-Nieto et Nicolas Seguin d'avoir accepté de rapporter ma thèse ainsi que leur temps consacré à la lecture de ce manuscrit. Je remercie également Robert Eymard, Antonin Novotny et Charlotte Perrin d'avoir participer à l'évaluation de cette thèse.

Je voudrais aussi remercier les professeurs Sergey Gavrilyuk et Enrique D. Fernàndez-Nieto d'avoir soutenu mon projet de thèse examiné par la commission de l'école doctorale de Mathématiques et informatiques d'Aix-Marseille Université.

J'aimerais à cette occasion, exprimer ma reconnaissance à mes directeurs de projet lors du CEMRACS 19, Emmanuel Audusse, Arnaud Duran et Yohan Penel pour leur pédagogie et leur esprit d'équipe. Je remercie également mes co-équipiers Virgile Dubos et Noemie Gaveaux ainsi que les organisateurs et participants d'avoir rendu l'événement festif et convivial.

Je n'oublie pas mes collègues doctorants de l'I2M et en particulier Antoine de m'avoir fait connaître les formules de Cardan. Je les souhaite beaucoup de succès dans leur recherche.

Je souhaite remercier ma famille, mon père, mes tantes, mes oncles, mes frères, mes sœurs, ma femme et en particulier les membres de ma famille résidant en France pour leur soutien financier et leur aide au quotidien. Un grand merci à tous mes proches et amis ayant rendu mon séjour en France agréable et fructueux.

Table des matières

Introduction générale

La modélisation mathématique des phénomènes naturels liés aux tsunamis, inondations, courants océaniques et atmosphériques, avalanches ou encore aux problèmes de sécurité industrielle (explosions), s'appuie sur des modèles de mécanique des fluides qui font intervenir des équations mathématiques de type lois de conservation physiques couplées à des lois phénoménologiques. Les écoulements de fluides compressibles ou incompressibles sont régis par les équations de Navier-Stokes. Sous certaines hypothèses physiques, on peut déduire des équations de Navier-Stokes des modèles simplifiés, par exemple au moyen d'une réduction d'échelle ou en négligeant certains termes. Certains des modèles ainsi obtenus sont des systèmes d'équations hyperboliques non linéaires. Dans le cadre de cette thèse, nous nous intéressons à de tels systèmes, et plus précisément principalement aux équations de Saint-Venant pour les écoulements en eau peu profonde et aux équations d'Euler compressible pour des gaz non visqueux.

Les équations de Saint-Venant en deux dimensions d'espace avec topographie s'écrivent :

$$\partial_t h + \operatorname{div}(h\boldsymbol{u}) = 0 \qquad \text{in } \Omega \times (0,T), \tag{0.1a}$$

$$\partial_t (hu_i) + \operatorname{div}(h\boldsymbol{u}\, u_i) + \partial_i (\frac{1}{2} g h^2) + gh\, \partial_i z = 0 \; i = 1,2 \qquad \text{in } \Omega \times (0,T), \tag{0.1b}$$

où Ω est un ouvert borné de $\mathbb{R}^2$, t désigne la variable en temps, g la constante d'accélération et z la topographie, donnée supposée régulière, sauf en cas de spécification contraire. Le symbole ∂_i dénote la dérivée partielle en espace par rapport à la i-ème coordonnée d'espace x_i et ∂_t désigne la dérivée partielle par rapport à la variable t. Les inconnues principales sont la hauteur d'eau h, sensée être positive et le champ de vitesse $\boldsymbol{u} = (u_1, u_2)^T$. Les équations (0.1) sont souvent utilisées pour la modélisation des écoulements à surface libre de faible profondeur tels que les fleuves, les lacs ou les zones cotières.

On souhaite ensuite tenir compte de la force de Coriolis, qui est un terme d'inertie qui résulte du mouvement de rotation uniforme. Elle intervient comme un terme source dans les équations de Saint-Venant, qui s'écrivent alors, en considérant un fond plat,

c.à.d. $z = 0$:

$$\partial_t h + \mathrm{div}(h\boldsymbol{u}) = 0 \qquad \text{in } \Omega \times (0,T), \tag{0.2a}$$

$$\partial_t (h\boldsymbol{u}) + \mathrm{div}(h\boldsymbol{u} \otimes \boldsymbol{u}) + \nabla\left(\frac{1}{2} g h^2\right) = -\omega h \boldsymbol{u}^\perp \qquad \text{in } \Omega \times (0,T), \tag{0.2b}$$

où ω est la vitesse angulaire, $\boldsymbol{u} = (u_1, u_2)^T$ et $\boldsymbol{u}^\perp = (-u_2, u_1)^T$. Ces équations sont utilisées dans le cadre des mouvements des fluides atmosphériques ou océaniques en rotation.

On s'intéresse également à la linéarisation du système (0.2) autour d'un état constant $(h_0, \boldsymbol{u}_0)$ avec $\boldsymbol{u}_0 = 0$, qui produit le système suivant :

$$\partial_t h + h_0 \mathrm{div}(\boldsymbol{u}) = 0 \qquad \text{in } \Omega \times (0,T), \tag{0.3a}$$

$$\partial_t \boldsymbol{u} + g\, \nabla h = -\omega \boldsymbol{u}^\perp \qquad \text{in } \Omega \times (0,T). \tag{0.3b}$$

Un modèle de transport de sédiment par charriage en dimension un d'espace est ensuite étudié. Le modèle est constitué d'un système de deux équations de conservation : les équations de Saint-Venant vues précédemment pour la modélisation de l'écoulement fluide, et l'équation dite "d'Exner ", qui exprime la conservation du flux de sédiment, avec une loi phénoménologique qui donne l'expression de ce flux en fonction des inconnues.

$$\partial_t h + \partial_x (hu) = 0 \qquad \text{dans } \Omega \times (0,T), \tag{0.4a}$$

$$\partial_t (hu) + \partial_x \left(hu^2 + \frac{1}{2} g h^2\right) + g h \partial_x z = -\tau / \rho_w \qquad \text{dans } \Omega \times (0,T), \tag{0.4b}$$

$$\partial_t z + \frac{1}{1-\phi} \partial_x q_b = 0 \qquad \text{dans } \Omega \times (0,T), \tag{0.4c}$$

où z désigne la profondeur de la couche de sédiment dépendant de x et de t, ρ_w est la densité de l'eau, ϕ est la porosité (constante) appartenant à l'intervalle $[0,1)$ et τ est la contrainte de cisaillement, définie en fonction de h et u. Le terme q_b est le flux de sédiment qui dépend généralement de h et u. Les quantités τ et q_b peuvent dans certains cas dépendre aussi de z.
On s'intéresse également aux équations d'Euler compressibles couplées avec les équations de bilan de masse des espèces chimiques lors d'une combustion dans le cadre d'un modèle dit "relaxé". Les équations de Euler décrivant l'écoulement du fluide

s'écrivent :

$$\partial_t \rho + \mathrm{div}(\rho \boldsymbol{u}) = 0 \qquad \text{in } \Omega \times (0,T), \tag{0.5a}$$

$$\partial_t (\rho u_i) + \mathrm{div}(\rho u_i \boldsymbol{u}) + \partial_i p = 0 \quad i = 1,d, \qquad \text{in } \Omega \times (0,T), \tag{0.5b}$$

$$\partial_t (\rho E) + \mathrm{div}(\rho E \boldsymbol{u}) + \mathrm{div}(p\boldsymbol{u}) = 0 \qquad \text{in } \Omega \times (0,T), \tag{0.5c}$$

$$p = (\gamma - 1)\rho e_s, \qquad E = \frac{1}{2}|\boldsymbol{u}|^2 + e, \quad e = e_s + \sum_{i \in \mathscr{I}} y_i \Delta h^0_{f,i}\,, \tag{0.5d}$$

où ρ est la densité, p la pression, E l'énergie totale, e l'énergie interne, e_s l'entropie sensible du fluide, $\mathscr{I}$ est l'ensemble des espèces chimiques présentes dans la réaction distinguées par leurs masses volumiques $y_i, i \in \mathscr{I}$ et $\Delta h^0_{f,i}$ est la variation de l'enthalpie de formation de la $i-$éme espèce chimique. Le modèle relaxé consiste à écrire que les espèces chimiques satisfont une équation de la forme :

$$\partial_t (\rho y_i) + \mathrm{div}(\rho y_i \boldsymbol{u}) = \dot{\omega}_i, \quad \text{for } i \in \mathscr{I}, \qquad \text{in } \Omega \times (0,T), \tag{0.6}$$

où $\dot{\omega}_i$ désigne un terme réactif spécifique pour chacune des espèces présentes dans la réaction, qui dépend d'une fonction caractéristique G satisfaisant une équation de transport non linéaire, ainsi que des concentrations du fuel et de l'oxydant.

Dans cette thèse, nous proposons et nous analysons des méthodes numériques pour la résolution des systèmes (0.1) à (0.6). La discrétisation en espace mise en œuvre dans ce travail, s'inscrit dans la continuation des travaux Herbin, Latché et Nguyen 2013 ; Herbin, Latché et Nguyen 2018, Therme 2015, Gunawan 2015 ; Gunawan, Eymard et Pudjaprasetya 2015. Les premiers travaux Herbin, Latché et Nguyen 2013 ; Herbin, Latché et Nguyen 2018, Therme 2015 sont le fruit de la collaboration initiée il y a une quinzaine d'années par l'Institut de Mathématiques de Marseille et l'Institut de Radioprotection et Sûreté Nucléaire à Cadarache, qui a d'abord porté sur le développement de l'analyse de schémas de correction de pression pour les équations de Navier-Stokes et d'Euler compressible Gastaldo, Herbin et Latché 2010, Herbin, Kheriji et Latché 2014. Ces derniers schémas introduits par Chorin 1968 et Temam 1969 dans les années 60 pour les équations de Navier-Stokes incompressible, utilisent une discrétisation sur des mailles décalées, qui assure la stabilité du schéma grâce à la condition inf-sup. L'extension de ces schémas aux équations de Navier-Stokes et Euler compressible permettent la simulation numérique des écoulements fluides à tout nombre de Mach Kheriji 2011, Grapsas 2017, Grapsas, Herbin, Kheriji et al. 2016, Herbin, Latché et Saleh 2020, et c'est ce type de schéma que nous retenons pour la discrétisation des équations d'Euler réactif qui est l'objet du chapitre 4. Un schéma de type Euler explicite équation par équation a aussi été étudiée dans le cas des équations d'Euler isentropiques et Euler complet Herbin, Latché et Nguyen 2018. Ce schéma explicite a été repris par Gunawan 2015 pour les équations de Saint Venant, ainsi que pour le système de Saint-Venant Exner Gunawan, Eymard et Pudjaprasetya

2015 en une dimension d'espace. Ces travaux utilisent pour la discrétisation en espace des schémas à mailles décalées, dont fait partie le célèbre schéma Marker-and-Cell (MAC) introduit dans les années 60 HARLOW et WELSH 1965. Des travaux relatifs à ces méthodes ont été réalisés par les hydrologues ARAKAWA et LAMB 1981 pour la discrétisation des équations de Saint-Venant.

Dans les schémas à mailles décalées, les inconnues discrètes correspondant aux variables scalaires (densité ou hauteur par exemples) sont situées au centre des mailles tandis que les inconnues discrètes correspondant aux champs de vecteurs sont localisés au centre des faces en 3D (ou arêtes en 2D). Cette technique d'arrangement des inconnues discrètes diffère donc de la méthode standard de volumes finis, souvent qualifiée de co-localisée, où toutes les variables sont calculées au centre des mailles.

Dans cette famille de schémas, on peut noter les trois grandes techniques de discrétisation suivantes :

— La stratégie MAC déjà mentionnée, correspondant aussi au schéma "C-grid " d'Arakawa et Lamb ARAKAWA et LAMB 1981. Cette discrétisation nécessite des maillages rectangulaires en $2D$ ou en parallèpipèdes rectangles en $3D$ mais pas forcément uniformes. Pour cette première technique les variables scalaires sont situées au centre alors que les composantes normales des champs de vecteurs sont localisées aux faces orthogonales à la normale en question.

— Les éléments finis de RANNACHER et TUREK 1992 qui fonctionnent pour des maillages de quadrilatères en $2D$ ou d'hexaèdres en $3D$, rectangulaires ou non. Les inconnues sont alors toutes les composantes du champ de vecteur aux faces du maillage, et des champs scalaires au centre des mailles.

— Les éléments finis de CROUZEIX et P. RAVIART 1973 qui partagent la même approche que les éléments de Rannacher-Turek, la seule différence étant qu'on utilise maintenant des maillages simpliciaux.

Plusieurs types de discrétisation en temps ont été étudiées, selon le type d'applications. La plus simple est une méthode d'intégration de type Euler explicite équation par équation. C'est elle que nous appliquons pour les équations de Saint-Venant et d'Exner, en considérant également une montée à l'ordre 2 par un schéma de Heun dans le cas des équations de Saint-Venant. Cependant, comme toute méthode explicite, elle nécessite une restriction de pas de temps pour des raisons de stabilité. Dans le cas des équations d'Euler réactif, une méthode de correction de pression est donc mise en œuvre.

Ce travail s'articule en cinq chapitres : les trois premiers traitent des systèmes faisant intervenir les équations de Saint-Venant (0.1), (0.2) et (0.4), tandis que Les deux derniers chapitres s'intéressent à la résolution des équations d'Euler réactif (0.5). Nous donnons maintenant une description succincte des différents chapitres.

Dans le premier chapitre, on analyse des schémas numériques pour les équations de Saint-Venant construits à partir d'une discrétization MAC. L'étude menée par DOYEN et GUNAWAN 2014 ; GUNAWAN 2015 s'intéressait déjà à une telle discrétisation en une dimension d'espace, en effectuant d'une part une étude de la stabilité du schéma, et

d'autre part une analyse de consistance au sens de Lax par rapport à la solution faible entropique des équations. On étend ces résultats dans plusieurs directions. En premier lieu, le schéma et son analyse sont écrits en deux dimensions d'espace. En second lieu, pour améliorer la précision, les opérateurs de convection intervenant dans le bilan de masse et dans la quantité de mouvement sont discrétisés par un schéma de type de type MUSCL PIAR, BABIK, HERBIN et al. 2013. Enfin, après avoir étudié une discrétisation de type Euler explicite, toujours dans un souci d'amélioration de la précision, nous considérons une discrétisation en temps de Heun. On démontre que les schémas envisagés assurent, sous condition de CFL, la positivité de la hauteur d'eau ainsi que des solutions d'équilibre tels que le "lac au repos". On démontre aussi que, sous des conditions de CFL éventuellement plus strictes, ces schémas sont consistants au sens de Lax par rapport à la formulation faible des équations continues, au sens suivant : si la suite des solutions approchées est bornée dans L^∞ et tend vers vers une limite presque partout lorsque les pas de temps et d'espace tendent vers 0, alors la limite est solution faible des équations de Saint-Venant. Dans le cas du schéma d'Euler, on démontre également la consistance au sens de Lax pour l'entropie. Enfin, différents cas tests sont présentés pour évaluer la performance de ces schémas et en particulier mesurer l'efficacité du schéma de Heun par rapport au shéma d'Euler explicite.

Le second chapitre est consacré à la construction de schémas numériques pour les équations de Saint-Venant avec le terme source de Coriolis. L'objectif est de construire des schémas qui soient stables par rapport à la dissipation de l'énergie mécanique semi-discrète et linéairement bien équilibrés par rapport à la préservation de l'état d'équilibre géostrophique. Ce chapitre est constitué de deux parties : la première partie propose une adaptation du schéma MAC découplé d'ordre un en temps et en espace pour la résolution des équations (0.2) et (0.3). Une discrétisation en espace par les éléments finis de Rannacher-Turek (RT) est ensuite présentée. Nous procéderons à une méthode de stabilisation de schéma qui repose sur une technique de correction des flux numériques et du gradient discret. Les schémas non linéaires et linéaires stabilisés obtenus satisfont une dissipation de l'énergie mécanique semi-discrète qui leur est associée. Le schéma entièrement discret correspondant pour les équations non linéaires préserve la positivité de la hauteur d'eau grâce à une restriction locale du pas temps de type CFL. De plus, les schémas RT linéaires et non linéaires, préservent parfaitement l'équilibre géostrophique, état stable des équations linéaires. Des tests numériques sont éffectués pour comparer la précision des différents schémas avec ceux obtenus par un schéma habituel de type Godunov.

Dans le troisième chapitre, nous appliquons une version du schéma découplé présenté dans le chapitre 1 pour la résolution du système (0.4) couplant les équations de Saint-Venant en dimension un d'espace avec celle d'Exner, qui nécessite la donnée de deux lois de fermeture : une loi sur la contrainte τ et une définition pour le flux de sédiment q_b. Une étude de ce modèle avec le même schéma avait déjà été entreprise par GUNAWAN 2015 ; GUNAWAN, EYMARD et PUDJAPRASETYA 2015 en pre-

nant des expressions de flux classiques, qui ont l'inconvénient de ne pas respecter la conservation de la masse de sédiment ni la dissipation de l'énergie. Nous introduisons tout d'abord un terme algébrique de friction, obtenu par une technique de stabilisation du terme source τ. Nous reprenons ensuite un modèle obtenu récemment par développement des équations asymptotique des équations de Navier-Stokes FERNÀNDEZ-NIETO, MORALES DE LUNA, NARBONA-REINA et al. 2017 qui a l'avantage de respecter la conservation de la masse de sédiment et la dissipation de l'énergie. Nous illustrons les atouts et les limites et des différentes formules de flux par des tests numériques.

Le chapitre 4 est consacré à un modèle de simulation de déflagration turbulente. L'écoulement est régi par les équations d'Euler pour un mélange de composition variable, tandis que la modélisation de la combustion est basée sur une approche phénoménologique : la propagation de la flamme est représentée par le transport de la fonction caractéristique de la zone brûlée, où la réaction chimique est complète ; en dehors de cette zone, l'atmosphère reste à l'état frais. Le problème est approché numériquement par une technique de type pénalisation qui utilise un taux de conversion fini et un temps caractéristique tendant vers zéro avec les pas d'espace et de temps. Le schéma numérique fonctionne sur des maillages décalés, éventuellement non structurés. L'algorithme de la marche dans le temps est de type itératif : on résout dans un premier temps les bilans de masse des espèces chimiques, puis, dans un second temps, les bilans de masse, de quantité de mouvement et d'énergie. Pour cette dernière étape de l'algorithme, on utilise une technique de correction de la pression, et on résout une équation d'équilibre pour l'enthalpie dite “sensible" plutôt que le bilan énergétique total, avec des termes correctifs qui assure la consistance du schéma. On prouve que les solutions approchées satisfont les mêmes propriétés de stabilité que le problème continu : les fractions massiques des espèces chimiques sont maintenues dans l'intervalle $[0, 1]$, la densité et l'énergie interne sensible restent positives et l'intégrale sur le domaine de calcul d'une énergie totale discrète est conservée. De plus, nous montre que le schéma est en fait conservatif, c'est-à-dire que les solutions approchées satisfont une équation de conservation de bilan énergétique total discret consistante au sens de Lax-Wendroff. Enfin, on observe numériquement que la procédure de pénalisation converge, c'est-à-dire que le fait de faire tendre l'échelle de temps chimique vers zéro permet de converger vers la solution du problème continu limite (chimie infiniment rapide). Les tests montrent également que la précision du schéma dépend fortement de la discrétisation de l'opérateur de convection dans les bilans massiques des espèces chimiques.

On présente dans le dernier chapitre un algorithme pour calculer une solution de référence pour un écoulement gazeux induit par une flamme sphérique se dilatant à partir d'une source ponctuelle à une vitesse d'expansion constante, en supposant une réaction chimique instantanée. La solution exacte est auto-similaire et l'écoulement est divisé en trois zones : une zone intérieure composée de gaz brûlés au repos, une

zone intermédiaire où la solution est régulière et l'atmosphère initiale composée de gaz frais au repos. La zone intermédiaire est délimitée par le choc réactif (côté intérieur) et le choc dit précurseur (côté extérieur), pour lesquels les conditions de Rankine-Hugoniot sont écrites ; la solution dans cette zone est régie par deux équations différentielles ordinaires qui sont résolues numériquement. Nous montrons que, pour toute vitesse de choc précurseur admissible, la construction combinant cette résolution numérique avec l'exploitation des conditions de saut est unique, et donne des profils de pression, de densité et de vitesse décroissants dans la zone intermédiaire. En outre, la vitesse du choc réactif est supérieure à la vitesse du côté extérieur du choc, ce qui est cohérent avec le fait que la différence entre ces deux quantités est ce qu'on appelle la vitesse de la flamme, c'est-à-dire la vitesse (relative) à laquelle la réaction chimique progresse dans les gaz frais. Enfin, nous observons aussi numériquement que la fonction donnant la vitesse de la flamme en fonction de la vitesse du choc précurseur augmente ; cela permet d'intégrer la résolution dans une procédure de type Newton pour calculer le débit pour une vitesse de flamme donnée (au lieu d'une vitesse de choc précurseur donnée). L'algorithme numérique qui en résulte est appliqué au mélange stoechiométrique hydrogène-air.

Certains résultats de cette thèse ont déjà fait l'objet de publication : deux articles dans des congrès publiés, un article de congrès à paraître, un article en révision dans une revue et un article soumis.

HERBIN, LATCHÉ, NASSERI et al. 2019 *A decoupled staggered scheme for the shallow water equations*, Monografias Matematicas Garcia de Galdeano, 52 :1-16, 2019.

GALLOUËT, HERBIN, LATCHÉ et al. 2020 *A second order consistent MAC scheme for the shallow water equations on non uniform grids*, Proceedings of FVCA 2020, R. Kloefkorn ed., Springer, 2020.

GRAPSAS, HERBIN, LATCHÉ et al. 2020 *A staggered pressure correction numerical scheme to compute a travelling reactive interface in a partially premixed mixture*, à paraître dans Proceedings of the international conference of Numerical methods for Hyperbolic equations, NumHyp2019, Malaga, 2019, https ://hal.archives-ouvertes.fr/hal-02967051.

GRAPSAS, HERBIN, LATCHÉ et al. p. d. *Modelling of a spherical deflagration at constant speed*, en révision, https ://hal.archives-ouvertes.fr/hal-02967980.

AUDUSSE, DUBOS, DURAN et al. 2020 *Numerical approximation of the shallow water equations with Coriolis source term*, soumis au Proceedings and surveys for CEMRACS 2019, EDP scciences, SMAI 2020.

GALLOUËT, HERBIN, LATCHÉ et al. 2020 *A second order Heun scheme for the 2D shallow water equations on non uniform grids*, soumis.

1 First and second order MAC schemes for the two−dimensional shallow water equations

Sommaire

Abstract. A Lax-Wendroff type result of consistency is given for convection operators on staggered meshes. It is applied to a class of second order finite volume schemes developed to obtain approximate solutions of the shallow water equations with bathymetry. These schemes are based on staggered grids for the space discretization: scalar and vector unknowns are defined on different meshes. MUSCL-like interpolations for the discrete convection operators in the water height and momentum equations are performed in order to improve the precision of the scheme. The time discretization is performed either by a first order segregated forward Euler scheme in time or by the second order Heun scheme. Both schemes are shown to preserve the water height positivity under a CFL condition and an important state equilibrium known as the lake at rest. Using the above mentioned staggered Lax-Wendroff type results, these schemes are shown to be Lax-consistent with the weak formulation of the continuous equations; besides, the forward Euler scheme is shown to be consistent with a weak entropy inequality. Numerical results confirm the efficiency and accuracy of the schemes.

Keywords MAC discretization, Heun scheme, consistency analysis, shallow water equations, partial dam break problem.

1.1 Introduction

The shallow water equations form a hyperbolic system of two conservation equations (mass and momentum) which models the flow of an incompressible fluid, assuming that the mean vertical height of the fluid is small compared to the plane scale. It is widely used for the simulation of numerous geophysical phenomena, such as flow in rivers and coastal areas. For a fluid occupying the space-time domain $\Omega \times (0,T)$, where Ω is an open bounded subset of $\mathbb{R}^2$ and $T>0$, the shallow water equations with bathymetry solve the water height h and the (vector) velocity of the fluid $\boldsymbol{u}=(u_1,u_2)$ and read:

$$\partial_t h + \mathrm{div}(h\boldsymbol{u}) = 0 \qquad \text{in } \Omega\times(0,T), \tag{1.1a}$$

$$\partial_t(hu_i) + \mathrm{div}(h\boldsymbol{u}\, u_i) + \partial_i p + gh\partial_i z = 0,\ i=1,2 \qquad \text{in } \Omega\times(0,T), \tag{1.1b}$$

$$p = \frac{1}{2}gh^2 \qquad \text{in } \Omega\times(0,T), \tag{1.1c}$$

$$\boldsymbol{u}\cdot\boldsymbol{n} = 0 \qquad \text{on } \partial\Omega\times(0,T), \tag{1.1d}$$

$$h(\boldsymbol{x},0) = h_0,\ \boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0 \qquad \text{in } \Omega. \tag{1.1e}$$

where ∂_t is the partial time derivative, div denotes the spatial divergence operator, ∂_i stands to the partial space derivative, g is the standard gravity constant and z the (given) bathymetry, which is supposed to be regular in this paper. The initial conditions are $h_0 \in L^\infty(\Omega)$ and $\boldsymbol{u}_0 = (u_{0,1},u_{0,2}) \in L^\infty(\Omega,\mathbb{R}^2)$ with $h_0 \geq 0$. This system has therefore been intensively studied, both theoretically and numerically, so that it is impossible to give an exhaustive list of references. We refer to the books Tan 1992; Bouchut 2004 and to the more recent books or parts of books Audusse 2018; Castro, Morales de Luna, and Parés 2017; Xing 2017 and the references therein. We recall that it is wellknown that if no dry zone exists, the system is strictly hyperbolic. In all cases, the solution of the system may develop shocks, so that the finite volume method is often preferred for numerical simulations. Two main approaches are found: one is the colocated approach which is usually based on some approximate Riemann solver, see e.g. Bouchut 2004; Castro, Morales de Luna, and Parés 2017 and references therein; the other one is based on a staggered arrangement of the unknowns on the grid, which is quite classical in the hydraulic and ocean engineering community, see e.g. Arakawa and Lamb 1981; Bonaventura and Ringler 2005; Stelling and Duinmeijer 2003. These latter staggered schemes have been implemented with an upwind choice for the convection operators and a forward Euler time discretization and analysed in the case of one space dimension Doyen and Gunawan 2014; Gunawan 2015, following the works on the related barotropic Euler equations, see Herbin, Latché, and Nguyen 2018 and references therein. In particular, the weak consistency of the scheme is shown as well as a weak entropy consistency. Let us recall that if $(h,\boldsymbol{u})$ is a regular solution of (1.1), the following elastic potential energy balance and kinetic energy

balance are obtained by manipulations on the mass and momentum equations:

$$\partial_t(\frac{1}{2}gh^2) + \mathrm{div}(\frac{1}{2}gh^2\boldsymbol{u}) + \frac{1}{2}gh^2\mathrm{div}\boldsymbol{u} = 0 \tag{1.2}$$

$$\partial_t(\frac{1}{2}h|\boldsymbol{u}|^2) + \mathrm{div}(\frac{1}{2}h|\boldsymbol{u}|^2\boldsymbol{u}) + \boldsymbol{u}\cdot\nabla p + gh\boldsymbol{u}\cdot\nabla z = 0. \tag{1.3}$$

Summing these equations, we obtain an entropy balance equation: $\partial_t E + \mathrm{div}\Phi = 0$, where the entropy-entropy flux pair (E,Φ) is given by:

$$E = \frac{1}{2}h|\boldsymbol{u}|^2 + \frac{1}{2}gh^2 + ghz \text{ and } \Phi = (E + \frac{1}{2}gh^2)\boldsymbol{u}. \tag{1.4}$$

For non regular functions the above manipulations are no longer valid, and the entropy inequality $\partial_t E + \mathrm{div}\Phi \leq 0$ is satisfied in a distributional sense. The weak entropy consistency consists in showing that any possible limit of the scheme satisfies a weak form of the entropy inequality (1.4) given in (1.31) below.

In the case of two space dimensions, the consistency of the upwind scheme with respect to the weak formulation and to a weak entropy inequality is stated in Herbin, Latché, Nasseri, et al. 2019; a quasi-second order scheme in time and space using the second order Heun method in time dependent and a MUSCL-like interpolation in space was proposed in Gallouët, Herbin, Latché, et al. 2020.

Here, we analyse the former schemes both theoretically and numerically. The framework that is developed here includes three schemes : the first order scheme of Herbin, Latché, Nasseri, et al. 2019, the same scheme replacing the upwind choice in the numerical convection operator by a MUSCL-like procedure, and the quasi second order scheme proposed in Gallouët, Herbin, Latché, et al. 2020. Generic properties are shown to be preserved, such as the positivity of the water height and the preservation of the "lake at rest" steady state. The weak consistency of the schemes is proven thanks to a generalisation of Lax-Wendroff type result; this consistency result is interesting for its own sake and valid for general convection operators on general colocated or staggered grids in any space dimension. Furthermore, the two first schemes are shown to be entropy-weak consistent in the sense that a weak entropy inequality is satisfied by any possible limit of the scheme as the time and space steps tend to 0, under some CFL condition.

The remainder of the paper is organized as follows: In Section 2 we introduce the space and time discretization. The resulting approximate solutions have some discrete stability and well balance properties which are studied in Section 3.2. Furthermore, under some convergence and boundedness assumptions, the approximate solutions are shown in Section 4.6 to converge to a weak solution of (1.1). This proof of these results heavily relies on the general Lax-Wendroff consistency lemma which is given in the appendix 1.A. In Section 1.5 we consider the first order time discretization and show that any possible limit of the scheme satifies a weak entropy inequality, again using the consistency result of the appendix. Numerical results comparing the first

order scheme of Herbin, Latché, Nasseri, et al. 2019, the same scheme replacing the upwind choice in the numerical convection operator by a MUSCL-like procedure, and the quasi second order scheme proposed in Gallouët, Herbin, Latché, et al. 2020 are presented in Section 5.5. Finally, the appendix 1.A contains the general consistency result for a nonlinear convection operator on general meshes with a staggered arrangement of the unknowns, which generalizes the result obtained in Gallouët, Herbin, and Latché 2019, while the appendix 1.B contains some technical lemmas which were proved formerly and which are recalled for the sake of completeness.

1.2 Space and time discretization

1.2.1 Definitions and notations

We concentrate on the MAC discretization in space, see Harlow and Welsh 1965; Harlow and Amsden 1971 for some seminal papers and Gallouët, Herbin, Latché, and Mallem 2018 for the convergence analysis of the scheme applied to the incompressible Navier-Stokes equations. This scheme is also widely used by the hydrologist and known as the Arakawa scheme Arakawa and Lamb 1981.

Let Ω be a connected subset of $\mathbb{R}^2$ consisting in a union of rectangles whose edges are assumed to be orthogonal to the canonical basis vectors, denoted by $(\boldsymbol{e}^{(1)}, \boldsymbol{e}^{(2)})$.

Definition 1.1 (MAC discretization)**.** *A discretization $(\mathcal{M}, \mathcal{E})$ of Ω with a staggered rectangular grid (or MAC grid), is defined by:*

- *A primal mesh $\mathcal{M}$ which consists in a conforming structured, possibly non uniform, rectangular grid of Ω. A generic cell of this grid is denoted by K, and its mass center by $\boldsymbol{x}_K$. The scalar unknowns (water height and pressure) are associated to this mesh.*
- *A set $\mathcal{E}$ of all edges of the mesh, with $\mathcal{E} = \mathcal{E}_{\rm int} \cup \mathcal{E}_{\rm ext}$, where $\mathcal{E}_{\rm int}$ (resp. $\mathcal{E}_{\rm ext}$) are the edges of $\mathcal{E}$ that lie in the interior (resp. on the boundary) of the domain. The set of edges that are orthogonal to $\boldsymbol{e}^{(i)}$ is denoted by $\mathcal{E}^{(i)}$, for $i \in \{1,2\}$. We then have $\mathcal{E}^{(i)} = \mathcal{E}^{(i)}_{\rm int} \cup \mathcal{E}^{(i)}_{\rm ext}$, where $\mathcal{E}^{(i)}_{\rm int}$ (resp. $\mathcal{E}^{(i)}_{\rm ext}$) are the edges of $\mathcal{E}^{(i)}$ that lie in the interior (resp. on the boundary) of the domain.*

  For $\sigma \in \mathcal{E}_{\rm int}$, we write $\sigma = K|L$ if $\sigma = \partial K \cap \partial L$. A dual cell D_σ associated to an edge $\sigma \in \mathcal{E}$ is defined as follows:

  - *if $\sigma = K|L \in \mathcal{E}_{\rm int}$ then $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$, where $D_{K,\sigma}$ (resp. $D_{L,\sigma}$) is the half-part of K (resp. L) adjacent to σ (see Fig. 4.1);*
  - *if $\sigma \in \mathcal{E}_{\rm ext}$ is adjacent to the cell K, then $D_\sigma = D_{K,\sigma}$.*

  For each dimension $i = 1,2$, the domain Ω can also be split up in dual cells: $\Omega = \cup_{\sigma \in \mathcal{E}^{(i)}} \overline{D_\sigma}$, $i \in \{1,2\}$; the i^{th} grid is refered to as the i^{th} dual mesh; it is associated to the i^{th} velocity component, in a sense which is clarified below. The set of the edges of the i^{th} dual mesh is denoted by $\widetilde{\mathcal{E}}^{(i)}$ (note that these edges may be non-orthogonal to $\boldsymbol{e}^{(i)}$); the set $\widetilde{\mathcal{E}}^{(i)}$ is decomposed into the internal and

boundary edges: $\widetilde{\mathcal{E}}^{(i)} = \widetilde{\mathcal{E}}^{(i)}_{\rm int} \cup \widetilde{\mathcal{E}}^{(i)}_{\rm ext}$. *The dual edge separating two duals cells* D_σ *and* $D_{\sigma'}$ *is denoted by* $\epsilon = \sigma|\sigma'$. *We denote by* D_ϵ *the cell associated to a dual edge* $\epsilon \in \widetilde{\mathcal{E}}$ *defined as follows:*

- *if* $\epsilon = \sigma|\sigma' \in \widetilde{\mathcal{E}}_{int}$ *then* $D_\epsilon = D_{\sigma,\epsilon} \cup D_{\sigma',\epsilon}$, *where* $D_{\sigma,\epsilon}$ *(resp.* $D_{\sigma',\epsilon}$*) is the half-part of* D_σ *(resp.* $D_{\sigma'}$*) adjacent to* ϵ *(see Fig. 4.1);*
- *if* $\epsilon \in \widetilde{\mathcal{E}}_{ext}$ *is adjacent to the cell* D_σ, *then* $D_\epsilon = D_{\sigma,\epsilon}$.

In order to define the scheme, we need some additional notations. The set of edges of a primal cell K and of a dual cell D_σ are denoted by $\mathcal{E}(K) \subset \mathcal{E}$ and $\widetilde{\mathcal{E}}(D_\sigma)$ respectively; note that $\widetilde{\mathcal{E}}(D_\sigma) \subset \widetilde{\mathcal{E}}^{(i)}$ if $\sigma \in \mathcal{E}^{(i)}$. For $\sigma \in \mathcal{E}$, we denote by $\boldsymbol{x}_\sigma$ the mass center of σ. The vector $\boldsymbol{n}_{K,\sigma}$ stands for the unit normal vector to σ outward K. In some cases, we need to specify the orientation of various geometrical entities with respect to the axis:

- a primal cell K is denoted $K = [\overrightarrow{\sigma\sigma'}]$ if $\sigma, \sigma' \in \mathcal{E}^{(i)}(K)$ for some $i \in \{1,2\}$ are such that $(\boldsymbol{x}_{\sigma'} - \boldsymbol{x}_\sigma) \cdot \boldsymbol{e}^{(i)} > 0$;
- we write $\sigma = \overrightarrow{K|L}$ if $\sigma \in \mathcal{E}^{(i)}$, $\sigma = K|L$ and $\overrightarrow{\boldsymbol{x}_K \boldsymbol{x}_L} \cdot \boldsymbol{e}^{(i)} > 0$ for some $i \in \{1,2\}$;
- the dual edge ϵ separating D_σ and $D_{\sigma'}$ is written $\epsilon = \overrightarrow{\sigma|\sigma'}$ if $\overrightarrow{\boldsymbol{x}_\sigma \boldsymbol{x}_{\sigma'}} \cdot \boldsymbol{e}^{(i)} > 0$ for some $i \in \{1,2\}$.

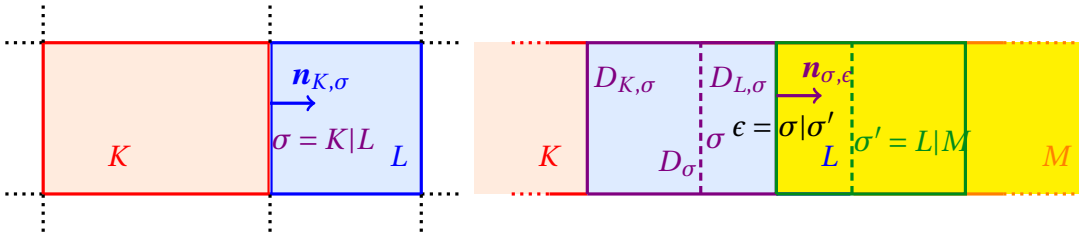


Figure 1.1 – Notations for the primal and dual meshes (in two space dimensions, for the first component of the velocity).

The size $\delta_{\mathcal{M}}$ of the mesh and its regularity $\theta_{\mathcal{M}}$ are defined by:

$$\delta_{\mathcal{M}} = \max_{K \in \mathcal{M}} \operatorname{diam}(K), \text{ and } \theta_{\mathcal{M}} = \max_{K \in \mathcal{M}} \max_{\sigma \in \mathcal{E}_K} \frac{|D_\sigma|}{|K|}, \tag{1.5}$$

where $|\cdot|$ stands for the one (or two) dimensional measure of a subset of $\mathbb{R}$ (or $\mathbb{R}^2$). Note that in the rectangular case that is considered here, the regularity parameter $\theta_{\mathcal{M}}$ is also equal to:

$$\theta_{\mathcal{M}} = \frac{1}{2}(1 + \max\Big\{\frac{|\sigma|}{|\sigma'|},\ (\sigma,\sigma') \in \mathcal{E}^{(i)2},\ i = 1,2\Big\}).$$

The discrete velocity unknowns are associated to the velocity cells and are denoted by $(u_{i,\sigma})_{\sigma \in \mathcal{E}^{(i)}}$, $i \in \{1,2\}$, while the discrete scalar unknowns (water height and pressure) are associated to the primal cells and are denoted respectively by $(h_K)_{K \in \mathcal{M}}$ and $(p_K)_{K \in \mathcal{M}}$.

Let us consider a uniform discretisation $0 = t_0 < t_1 < \cdots < t_N = T$ of the time interval $(0, T)$, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \cdots, N-1$ be the (constant, for the sake of simplicity) time step.

Here we present two schemes: a first order in time segregated scheme using the forward Euler scheme and the second order in time Heun scheme. Both schemes use a MUSCL-like technique for the computation of the numerical flux, see Piar, Babik, Herbin, et al. 2013, so that they are quasi second-order in space.

1.2.2 The segregated forward Euler scheme

We propose here a first order in time segregated discretisation and MAC discretization in space of the system (1.1); the scheme is written in compact form as follows:

$$\textbf{Initialisation:}\quad u^0_{i,\sigma} = \frac{1}{|D_\sigma|}\int_{D_\sigma} u_{i,0}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\ h^0 = \frac{1}{|K|}\int_K h_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\ p^0 = \frac{1}{2}g(h^0)^2. \tag{1.6a}$$

For $0 \le n \le N-1$: solve for h^{n+1}, p^{n+1} and $\boldsymbol{u}^{n+1} = (u_i^{n+1})_{i=1,2}$:

$$\eth_t h_K^{n+1} + \mathrm{div}_K(h^n \boldsymbol{u}^n) = 0, \qquad \forall K \in \mathscr{M} \tag{1.6b}$$

$$p_K^{n+1} = \frac{1}{2} g (h_K^{n+1})^2, \tag{1.6c}$$

$$\eth_t (h\, u_i)_\sigma^{n+1} + \mathrm{div}_{D_\sigma}(h^n \boldsymbol{u}^n u_i^n) + (\eth_i p^{n+1})_\sigma + g\, h_{\sigma,c}^{n+1}\ (\eth_i z)_\sigma = 0, \forall \sigma \in \mathscr{E}_{\mathrm{int}}^{(i)}, \tag{1.6d}$$

where the different discrete terms and operators introduced here are now defined.

Discrete time derivative - In the sequel, we shall denote by $\eth_t v^{n+1}$ the discrete forward time derivative of a given discrete function of time v, *i.e.*:

$$\eth_t v^{n+1} = \frac{v^{n+1} - v^n}{\delta t} \tag{1.7}$$

Discrete divergence and gradient operators - The discrete divergence operator on the primal mesh denoted by div_K is defined as follows:

$$\mathrm{div}_K(h\boldsymbol{u}) = \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma}, \text{ with } \boldsymbol{F}_\sigma = h_\sigma\, \boldsymbol{u}_\sigma, \tag{1.8}$$

with $\boldsymbol{u}_\sigma = u_{i,\sigma}\, \boldsymbol{e}^{(i)}$ for $\sigma \in \mathscr{E}^{(i)}$, $i \in \{1,2\}$ and h_σ is approximated by the MUSCL-like interpolation technique with respect to $\boldsymbol{u}_\sigma$; in the subsequent analysis, we do not

need to have an explicit formula for h_σ, but we need the following conditions to be satisfied:

$$\begin{aligned}
&\forall\, K\in\mathscr{M},\ \forall\sigma=K|L\in\mathscr{E}_{\rm int}(K),\\
&\quad -\exists\lambda_{K,\sigma}\in[0,1]: h_\sigma=\lambda_{K,\sigma}h_K+(1-\lambda_{K,\sigma})h_L \text{ if } \boldsymbol{F}_\sigma\cdot\boldsymbol{n}_{K,\sigma}\geq 0. \qquad (1.9)\\
&\quad -\exists\ \alpha_\sigma^K\in[0,1] \text{ and } M_\sigma^K\in\mathscr{M}\ :\ h_\sigma-h_K=\begin{cases}\alpha_\sigma^K(h_K-h_{M_\sigma^K}) & \text{if } \boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}\geq 0,\\ \alpha_\sigma^K(h_{M_\sigma^K}-h_K) & \text{otherwise.}\end{cases} \qquad (1.10)
\end{aligned}$$

By (1.9), h_σ is a convex combination of h_K and h_L, and if $\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}<0$, the cell M_σ^K in (1.10) can be chosen as L and $\alpha_{K,\sigma}$ as $1-\lambda_{K,\sigma}$. In the case of a discrete divergence free velocity field $\boldsymbol{u}$, this assumption ensures that h_K^{n+1} is a convex combination of the values h_K^n and $(h_M^n)_{M\in\mathscr{N}_m((K)}$, where $\mathscr{N}_m(K)$ denotes the set of cells M_σ^K satisfying (1.10), see Piar, Babik, Herbin, et al. 2013, Lemma 3.1, for any structured or unstructured mesh.

Note that if $K=[\sigma'\sigma]$ with $\sigma'=J|K$ and $\sigma=K|L$ and $\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}\geq 0$, the cell M_σ^K in Relation (1.10) can be chosen as the cell J and the value h_σ computed using the following limitation procedure:

$$h_\sigma-h_K=\frac{1}{2}\,\psi\big(h_L,h_K,h_J\big),\ \text{where}$$

$$\psi\big(h_L,h_K,h_J\big)=\begin{cases}\text{minmod}\big(\dfrac{h_L-h_J}{2},\zeta^+(h_L-h_K),\zeta^-(h_K-h_J)\big), & \text{if } (h_L-h_K)(h_K-h_J)>0,\\ 0, & \text{otherwise,}\end{cases}$$

where the limitation parameters ζ^+,ζ^- are such that $\zeta^+,\zeta^-\in[0,2]$. Observe that if $\zeta^+=\zeta^-=1$, the classical minmod limiter ($\text{minmod}\big(h_L-h_K,h_K-h_J\big)$) is recovered.

A local discrete derivative applied to a discrete scalar field ξ (with $\xi=p,h$ or z) is defined by:

$$(\eth_i\xi)_\sigma=\frac{|\sigma|}{|D_\sigma|}\,(\xi_L-\xi_K)\ \text{ for } \sigma=\overrightarrow{K|L}\in\mathscr{E}_{\rm int}^{(i)},\ i=1,2. \qquad (1.11)$$

The above defined discrete divergence and discrete derivatives satisfy the following div-grad duality relationship Gallouët, Herbin, Latché, and Mallem 2018, Lemma 2.4:

$$\sum_{K\in\mathscr{M}}|K|\xi_K\text{div}_K(h\boldsymbol{u})+\sum_{i=1}^{2}\sum_{\sigma\in\mathscr{E}_{\rm int}^{(i)}}|D_\sigma|h_\sigma u_{i,\sigma}\,(\eth_i\xi)_\sigma=0. \qquad (1.12)$$

Discrete water height for the bathymetry term – In equation (1.6d) the term $\eth_\sigma z$ denotes the discrete derivative (in the sense of (1.11)) of the piecewise constant function

$z_{\mathcal{M}} = \sum_{K\in\mathcal{M}} z(\boldsymbol{x}_K)\,\mathbb{1}_K$, that is:

$$\eth_\sigma z = \frac{|\sigma|}{|D_\sigma|}\left(z(\boldsymbol{x}_L) - z(\boldsymbol{x}_K)\right) \text{ for } \sigma = \overrightarrow{K|L} \in \mathscr{E}_{\rm int}. \tag{1.13}$$

The value $h_{\sigma,c}$ of the water height is defined so as to satisfy:

$$\eth_\sigma p + g\, h_{\sigma,c}\, \eth_\sigma z = 0 \text{ if } \eth_\sigma(h+z) = 0,\ \forall\, i = 1,2. \tag{1.14}$$

This requirement is fulfilled if $h_{\sigma,c}$ is centered:

$$h_{\sigma,c} = \begin{cases} \frac{1}{2}(h_K + h_L) & \text{for } \sigma = K|L \in \mathscr{E}_{\rm int}, \\ h_K & \text{for } \sigma \in \mathscr{E}_{\rm ext} \cap \mathscr{E}(K). \end{cases} \tag{1.15}$$

Indeed, if $h_{\sigma,c}$ is defined by (1.15), since $p = \frac{1}{2}gh^2$, one has from the definition of the discrete gradient (1.11), for $\sigma = K|L \in \mathscr{E}_{\rm int}^{(i)}$,

$$(\eth_i p)_\sigma + g\, h_{\sigma,c}\, (\eth_i z)_\sigma = \frac{1}{2} g \frac{|\sigma|}{|D_\sigma|}(h_K + h_L)\ (\eth_i(h+z))_\sigma$$

and therefore (1.14) holds, so that the “lake at rest" steady state is preserved, see Lemma 1.2 below.

Discrete convection operator – The term $(h\ u_i)_\sigma^{n+1}$ in the discrete time derivative in (1.6d) is defined by

$$(h\ u_i)_\sigma^{n+1} = h_{D_\sigma}^{n+1}\ u_{i,\sigma}^{n+1}, \tag{1.16a}$$

$$h_{D_\sigma} = \frac{1}{|D_\sigma|}\Big(|D_{K,\sigma}|\ h_K + |D_{L,\sigma}|\ h_L\Big), \text{ with } \sigma = K|L \in \mathscr{E}_{\rm int}, \tag{1.16b}$$

where D_σ, $D_{K,\sigma}$ and $D_{L,\sigma}$ are defined in Definition 1.1.

The discrete divergence operator on the dual mesh div_{D_σ} is given by:

$$\mathrm{div}_{D_\sigma}(hu_i\boldsymbol{u}) = \frac{1}{|D_\sigma|} \sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)} |\epsilon|\ \boldsymbol{G}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon}, \text{ with } \boldsymbol{G}_\epsilon = \boldsymbol{F}_\epsilon u_{i,\epsilon}, \tag{1.17}$$

where

— the flux $\boldsymbol{F}_\epsilon$ is computed from the primal numerical mass fluxes; following Herbin and Latché 2010 (see also Herbin, Latché, and Nguyen 2013, and Ansanay-Alex, Babik, Latché, et al. 2011 for an extension to triangular or quandrangular meshes using low order non-conforming finite element), it is defined as follows:

$$\text{for } \epsilon = \sigma|\sigma',\ \epsilon \subset K, \qquad \boldsymbol{F}_\epsilon = \frac{1}{2}\big(\boldsymbol{F}_\sigma + \boldsymbol{F}_{\sigma'}\big), \quad \epsilon \subset K, \ \text{(left on 1.2)} \tag{1.18a}$$

$$\text{for } \epsilon = \sigma|\sigma',\ \epsilon \not\subset K, \qquad \boldsymbol{F}_\epsilon = \frac{1}{|\epsilon|}\Big(\frac{1}{2}|\tau|\,\boldsymbol{F}_\tau + \frac{1}{2}|\tau'|\,\boldsymbol{F}_{\tau'}\Big), \ \text{(right on 1.2)}, \tag{1.18b}$$

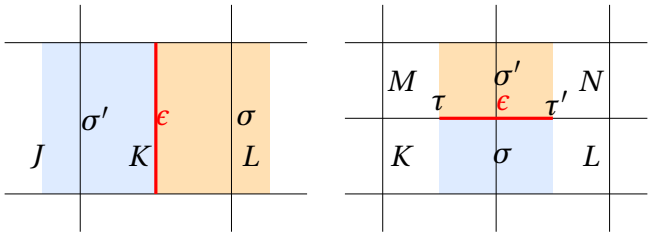


Figure 1.2 – Notation for the definition of the momentum flux on the dual mesh for the first component of the velocity- left: $\epsilon \subset K$ - right: $\epsilon \subset \tau \cup \tau'$.

— the value $u_{i,\epsilon}$ is expressed in terms of the unknowns $u_{i,\sigma}$, for $\sigma \in \mathscr{E}^{(i)}$ by a second order MUSCL-like interpolation scheme with respect to $\boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon}$ Piar, Babik, Herbin, et al. 2013; the values $u_{i,\sigma}$ satisfy the following property:

$$\begin{aligned}
&\forall\ \sigma \in \mathscr{E}^{(i)}_{\text{int}},\ i = 1,2,\ \forall\ \epsilon = \sigma|\sigma' \in \widetilde{\mathscr{E}}(D_\sigma),\\
&\quad u_{i,\epsilon} \text{ is a convex combination of } u_{i,\sigma} \text{ and } u_{i,\sigma'}:\\
&\quad \exists \mu_{\sigma,\epsilon} \in [0,1] : u_{i,\epsilon} = \mu_{\sigma,\epsilon} u_{i,\sigma} + (1-\mu_{\sigma,\epsilon}) u_{i,\sigma'} \qquad (1.19)\\
&\quad \exists\ \alpha^\sigma_\epsilon \in [0,1] \text{ and } \tau^\sigma_\epsilon \in \mathscr{E}^{(i)}_{\text{int}} : u_{i,\epsilon} - u_{i,\sigma} = \begin{cases} \alpha^\sigma_\epsilon (u_{i,\sigma} - u_{i,\tau^\sigma_\epsilon}) & \text{if } \boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} \geq 0,\\ \alpha^\sigma_\epsilon (u_{i,\tau^\sigma_\epsilon} - u_{i,\sigma}) & \text{otherwise.} \end{cases} \qquad (1.20)
\end{aligned}$$

Again note that in the case $\boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} < 0$, the edge τ^σ_ϵ may be chosen as σ'.
Let us emphasize that owing to the definitions (1.16b) and (1.18) the following discrete mass balance version on the dual mesh holds:

$$\frac{|D_\sigma|}{\delta t}(h^{n+1}_{D_\sigma} - h^n_{D_\sigma}) + \sum_{\epsilon \in \widetilde{\mathscr{E}}(D_\sigma)} |\epsilon|\, \boldsymbol{F}^n_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} = 0. \tag{1.21}$$

1.2.3 A second order in time Heun scheme

We retain here the quasi-second order space discretization which we just set up, but consider now a second order time discretization using the Heun (or Runge Kutta 2) scheme.

The initialization of the scheme is the same as that of the forward Euler scheme, see (1.6a), but the n-th step now reads:

Step n : For h^n and $\boldsymbol{u}^n = (u_i^n)_{i=1,2}$ known,

$$\hat{h}_K^{n+1} = h_K^n - \delta t \, \mathrm{div}_K(h^n \boldsymbol{u}^n), \qquad \forall K \in \mathscr{M} \tag{1.22a}$$

$$\hat{h}_{D_\sigma}^{n+1} \, \hat{u}_{i,\sigma}^{n+1} = h_{D_\sigma}^n \, u_{i,\sigma}^n - \delta t \, \mathscr{F}_{D_\sigma}(h^n, u_i^n), \qquad \forall \sigma \in \mathscr{E}_{\mathrm{int}}^{(i)} \tag{1.22b}$$

$$\tilde{h}_K^{n+1} = \hat{h}_K^{n+1} - \delta t \, \mathrm{div}_K(\hat{h}^{n+1} \hat{\boldsymbol{u}}^{n+1}), \qquad \forall K \in \mathscr{M} \tag{1.22c}$$

$$\tilde{h}_{D_\sigma}^{n+1} \, \tilde{u}_{i,\sigma}^{n+1} = \hat{h}_{D_\sigma}^{n+1} \, \hat{u}_{i,\sigma}^{n+1} - \delta t \, \mathscr{F}_{D_\sigma}(\hat{h}^{n+1}, \hat{u}_i^{n+1}), \qquad \forall \sigma \in \mathscr{E}_{\mathrm{int}}^{(i)} \tag{1.22d}$$

$$h_K^{n+1} = \frac{1}{2}(h_K^n + \tilde{h}_K^{n+1}), \qquad \forall K \in \mathscr{M} \tag{1.22e}$$

$$h_{D_\sigma}^{n+1} \, u_{i,\sigma}^{n+1} = \frac{1}{2}\big(h_{D_\sigma}^n \, u_{i,\sigma}^n + \tilde{h}_{D_\sigma}^{n+1} \, \tilde{u}_{i,\sigma}^{n+1}\big), \qquad \forall \sigma \in \mathscr{E}_{\mathrm{int}}^{(i)} \tag{1.22f}$$

where

$$\mathscr{F}_{D_\sigma}(h^n, u_i^n) = \mathrm{div}_{D_\sigma}(h^n \boldsymbol{u}^n u_i^n) + g \, h_{\sigma,c}^n \left((\eth_i h^n)_\sigma + (\eth_i z)_\sigma\right) \tag{1.23}$$

and the dual cell values $\hat{h}_{D_\sigma}^{n+1}$, $\tilde{h}_{D_\sigma}^{n+1}$ and $h_{D_\sigma}^{n+1}$ are computed from the corresponding cell values by the analogue of the formula (1.16b), so that they satisfy a dual mass balance of the type (1.21).

The steps (1.22c)-(1.22f) of the above scheme (1.22) may be replaced by the more compact form

$$\eth_t h_K^{n+1} = -\frac{1}{2}\left(\mathrm{div}_K(h^n \boldsymbol{u}^n) + \mathrm{div}_K(\hat{h}^{n+1} \hat{\boldsymbol{u}}^{n+1})\right), \qquad \forall K \in \mathscr{M} \tag{1.24a}$$

$$\eth_t(h_{D_\sigma} \, u_{i,\sigma})^{n+1} = -\frac{1}{2}\left(\mathscr{F}_{D_\sigma}(h^n, u_i^n) + \mathscr{F}_{D_\sigma}(\hat{h}^{n+1}, \hat{u}_i^{n+1})\right), \qquad \forall \sigma \in \mathscr{E}^{(i)}, \tag{1.24b}$$

where the dual cell value $h_{D_\sigma}^{n+1}$ is computed by the formula (1.16b) and hence satisfies a dual mass balance of the type (1.21).

1.3 Stability of the schemes

The positivity of the water height under a CFL like condition is ensured by both the schemes (1.6) and (1.22); it is a consequence of the property (1.10) of the MUSCL choice for the interface values. Indeed, the proof of the positivity in Piar, Babik, Herbin, et al. 2013, Lemma 3.1 remains valid even if the discrete velocity field is not divergence free, as is the case here.

Lemma 1.1 (Positivity of the water height)**.** *Let $n \in \{0, \cdots, N_t - 1\}$, let $(h_K^n)_{K \in \mathscr{M}} \subset \mathbb{R}_+^*$ and $(\boldsymbol{u}_\sigma^n)_{\sigma \in \mathscr{E}} \subset \mathbb{R}^d$ be given, and let h_K^{n+1} be computed by the forward Euler scheme, step*

(1.6b). *Then* $h_K^{n+1} > 0$, *for all* $K \in \mathcal{M}$ *under the following CFL condition,*

$$2\,\delta t \leq \frac{|K|}{\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,|\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}|}. \tag{1.25}$$

If (2.16) *is fullfilled and if furthermore*

$$2\,\delta t \leq \frac{|K|}{\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,|\hat{\boldsymbol{u}}_\sigma^{n+1} \cdot \boldsymbol{n}_{K,\sigma}|}, \tag{1.26}$$

then h_K^{n+1} *computed by the Heun scheme* (1.22) *is positive.*

Secondly, thanks to the choice (1.15) for the reconstruction of the water height, the property (1.14) holds, so that the so-called "lake at rest" steady state is preserved by both schemes.

Lemma 1.2 (Steady state "lake at rest")**.** *Let* $n \in \{0,\cdots,N_t-1\}$, $C \in \mathbb{R}_+$; *let* $(h_K^n)_{K\in\mathcal{M}} \subset \mathbb{R}$ *such that* $h_K^n + z_K = C$ *for all* $K \in \mathcal{M}$ *and* $\boldsymbol{u}_\sigma^n = 0$ *for* $\sigma \in \mathcal{E}$. *Then the solution* $(h_K^{n+1})_{K\in\mathcal{M}}$, $(\boldsymbol{u}_\sigma^{n+1})_{\sigma\in\mathcal{E}}$ *of the forward Euler scheme* (1.6) *(resp. Heun scheme* (1.22)*) satisfies* $h_K^{n+1} + z = C$ *for all* $K \in \mathcal{M}$ *and* $\boldsymbol{u}_\sigma^{n+1} = 0$ *for* $\sigma \in \mathcal{E}$.

As a consequence of the careful discretisation of the convection term, the segregated forward Euler scheme satisfies a discrete kinetic energy balance, as stated in the following lemma. The proof of this result is an easy adaptation of Herbin, Latché, and Nguyen 2018, Lemma 3.2.

Lemma 1.3 (Discrete kinetic energy balance, forward Euler scheme)**.** *A solution to the scheme* (1.6) *satisfies the following equality, for* $i = 1,2$, $\sigma \in \mathcal{E}^{(i)}$ *and* $0 \leq n \leq N-1$*:*

$$\begin{aligned}\frac{|D_\sigma|}{2\delta t}(h_{D_\sigma}^{n+1}(u_{i,\sigma}^{n+1})^2 - h_{D_\sigma}^n (u_{i,\sigma}^n)^2) + \frac{1}{2}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon|\,(u_{i,\epsilon}^n)^2\, \boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon} \\ + |D_\sigma|\, u_{i,\sigma}^{n+1}(\eth_i p^{n+1})_\sigma + |D_\sigma|\, g\, h_{\sigma,c}^{n+1}\, u_{i,\sigma}^{n+1}(\eth_i z)_\sigma z = -R_{i,\sigma}^{n+1},\end{aligned} \tag{1.27}$$

with

$$\begin{aligned}R_{i,\sigma}^{n+1} = \frac{1}{2\delta t}\,|D_\sigma|\,h_{D_\sigma}^{n+1}\left(u_{i,\sigma}^{n+1} - u_{i,\sigma}^n\right)^2 - \frac{1}{2}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon|\,\boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)^2 \\ + \sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon|\,\boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)\left(u_{i,\sigma}^{n+1} - u_{i,\sigma}^n\right).\end{aligned}$$

The scheme also satisfies the following potential energy balance.

Lemma 1.4 (Discrete potential balance, forward Euler scheme)**.** *Let, for* $K \in \mathcal{M}$ *and* $0 \leq n \leq N$ *the potential energy be defined by* $(E_p)_K^n = \frac{1}{2} g\,(h_K^n)^2 + g\,h_K^n z_K$. *A solution to*

the scheme (1.6) *satisfies the following equality, for* $K \in \mathscr{M}$ *and* $0 \le n \le N-1$*:*

$$\eth_t(E_p)_K^{n+1} + \mathrm{div}_K(p^n \boldsymbol{u}^n) + p_K^n \mathrm{div}_K(\boldsymbol{u}^n) + g z_K \mathrm{div}_K(h^n \boldsymbol{u}^n) = -R_K^{n+1}, \tag{1.28}$$

with

$$\eth_t(E_p)_K^{n+1} = \frac{1}{\delta t}\big((E_p)_K^{n+1} - (E_p)_K^n\big), \quad \mathrm{div}_K(p^n \boldsymbol{u}^n) = \sum_{\sigma \in \mathscr{E}(K)} |\sigma| \frac{1}{2} g (h_\sigma^n)^2 \, \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma}$$

and

$$|K| \, R_K^{n+1} = \frac{1}{2}\frac{|K|}{\delta t} g (h_K^{n+1} - h_K^n)^2 - \frac{1}{2} \sum_{\sigma \in \mathscr{E}(K)} |\sigma| \, g (h_\sigma^n - h_K^n)^2 \, \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma} + \sum_{\sigma \in \mathscr{E}(K)} |\sigma| g (h_K^{n+1} - h_K^n) \, h_\sigma^n \, \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}. \tag{1.29}$$

Proof. Applying Herbin, Latché, and Nguyen 2018, Lemma A1, (re-stated in Lemma 1.11 below for the sake of completeness), with $P = K$, $\psi : x \mapsto \frac{1}{2} g x^2$, $\rho_P = h_K^{n+1}$, $\rho_P^* = h_K^n$, $\eta = \sigma$, $\rho_\eta^* = h_\sigma^n$ and $V_\eta^* = |\sigma| \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}$, and $R_K^{n+1} = |K| \, r_K^{n+1}$, we get that

$$\frac{g}{2}\eth_t(h_K^{n+1})^2 + \mathrm{div}_K(p^n \boldsymbol{u}^n) + p_K^n \mathrm{div}_K(\boldsymbol{u}^n) = -\frac{g}{2\delta t}(h_K^{n+1} - h_K^n)^2 + \frac{g}{2}\frac{1}{|K|} \sum_{\sigma \in \mathscr{E}(K)} |\sigma| \, (h_\sigma^n - h_K^n)^2 \, \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma} - \frac{1}{|K|} \sum_{\sigma \in \mathscr{E}(K)} |\sigma| g (h_K^{n+1} - h_K^n) \, h_\sigma^n \, \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma},$$

Then, multiplying the discrete mass balance equation (1.6b) by $g z_K$ yields

$$\frac{1}{\delta t}\big((gzh)_K^{n+1} - (gzh)_K^n\big) + g z_K \frac{1}{|K|} \sum_{\sigma \in \mathscr{E}(K)^{(m)}} |\sigma| h_\sigma^n \, \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma} = 0$$

Summing the two above equations yields (1.29). □

Since the discrete kinetic and potential energies are computed on the dual and primal meshes respectively, the obtention of a discrete entropy inequality is not straightforward. In Herbin, Latché, Nasseri, et al. 2019, a kinetic energy inequality on the primal cell is obtained from the inequality (1.1d) to get a discrete local entropy inequality. Here, however we proceed otherwise, thanks to a general Lax-Wendroff Lemma for staggered grids (Lemma 1.8 in the section 1.A), which allows to handle each energy inequality on its respective mesh, without any reconstruction, see Section 1.5 below.

1.4 Weak consistency of the schemes

We now wish to prove the weak consistency of the scheme in the Lax-Wendroff sense, namely to prove that if a sequence of solutions is controlled in suitable norms and converges to a limit, this latter necessarily satisfies a weak formulation of the continuous problem.

The pair of functions $(\bar{h}, \bar{\boldsymbol{u}}) \in L^1(\Omega \times [0,T)) \times L^1(\Omega \times [0,T))^2$ is a weak solution to the continuous problem if it satisfies, for any $\varphi \in C_c^\infty\big(\Omega \times [0,T)\big)$ ($\boldsymbol{\varphi} \in C_c^\infty\big(\Omega \times [0,T)\big)^2$):

$$\int_0^T \int_\Omega \Big[\bar{h}\, \partial_t \varphi + \bar{h}\, \bar{u} \cdot \nabla \varphi \Big] \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega h_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0, \tag{1.30a}$$

$$\int_0^T \int_\Omega \Big[\bar{h}\, \bar{\boldsymbol{u}} \cdot \partial_t \boldsymbol{\varphi} + (\bar{h}\bar{\boldsymbol{u}} \otimes \bar{\boldsymbol{u}}) : \nabla \boldsymbol{\varphi} + \frac{1}{2}\, g\, \bar{h}^2 \mathrm{div} \boldsymbol{\varphi} + g\, \bar{h} \nabla z \boldsymbol{\varphi} \Big] \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega h_0(\boldsymbol{x})\, \boldsymbol{u}_0(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0. \tag{1.30b}$$

A weak solution of (1.30) is an entropy weak solution if for any nonnegative test function $\varphi \in C_c^\infty\big(\Omega \times [0,T), \mathbb{R}_+\big)$:

$$\int_0^T \int_\Omega \Big[\bar{E}\, \partial_t \varphi + \bar{\Phi} \cdot \nabla \varphi \Big] \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega E_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} \geq 0, \tag{1.31}$$

with

$$\bar{E} = \frac{1}{2}\bar{h}|\bar{\boldsymbol{u}}|^2 + \frac{1}{2} g \bar{h}^2 + g \bar{h} z \text{ and } \bar{\Phi} = (\bar{E} + \frac{1}{2} g \bar{h}^2)\bar{\boldsymbol{u}}.$$

Before stating the global weak consistency of the schemes (1.6) and (1.22), some definitions and assumptions are needed.

Let $(\mathscr{M}^{(m)}, \mathscr{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes in the sense of Definition 1.1 and let $(h^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be the associated sequence of solutions of the scheme (1.6)) defined almost everywhere on $(\Omega \times [0,T)$ by:

$$\begin{aligned} u_i^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1} \sum_{\sigma \in (\mathscr{E}^{(i)})^{(m)}} (u_i^{(m)})_\sigma^{n+1} \mathbb{1}\!\mathrm{l}_{D_\sigma}(\boldsymbol{x}) \mathbb{1}\!\mathrm{l}_{[t_n, t_{n+1})}(t), \text{ for } i \in \{1,2\} \\ h^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1} \sum_{K \in \mathscr{M}^{(m)}} (h^{(m)})_K^{n+1} \mathbb{1}\!\mathrm{l}_K(\boldsymbol{x}) \mathbb{1}\!\mathrm{l}_{[t_n, t_{n+1})}(t), \end{aligned}$$

where $\mathbb{1}\!\mathrm{l}_A$ is the characteristic function of a given set A, that is $\mathbb{1}\!\mathrm{l}_A(y) = 1$ if $y \in A$, $\mathbb{1}\!\mathrm{l}_A(y) = 0$ otherwise.

Assumed estimates - Some boundedness and compactness assumptions on the sequence of discrete solutions $(h^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ are needed in order to prove the Lax-Wendroff type consistency result. First of all we assume that $h^{(m)} > 0$, $\forall m \in \mathbb{N}$ which can be obtained under uniform versions of the CFL conditions (2.16) and (1.26), thanks to Lemma 1.1. Furthermore, we assume that:

– the water height $h^{(m)}$ and its inverse are uniformly bounded in $\mathrm{L}^\infty(\Omega\times(0,T))$, *i.e.* there exists $C^h_{\mathscr{M}}\in\mathbb{R}^*_+$ such that for $m\in\mathbb{N}$ and $0\le n<N^{(m)}$:

$$\frac{1}{C^h}<(h^{(m)})^n_K\le C^h,\quad \forall K\in\mathscr{M}^{(m)}, \tag{1.32}$$

– the velocity $\boldsymbol{u}^{(m)}$ is also uniformly bounded in $\mathrm{L}^\infty(\Omega\times(0,T))^2$, *i.e.* there exists $C^u\in\mathbb{R}^*_+$ such that

$$|(\boldsymbol{u}^{(m)})^n_\sigma|\le C^u,\quad \forall\sigma\in\mathscr{E}^{(m)}. \tag{1.33}$$

Theorem 1.1 (Weak consistency of the schemes)**.** *Let $(\mathscr{M}^{(m)},\mathscr{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes such that $\delta t^{(m)}$ and $\delta_{\mathscr{M}^{(m)}}\to 0$ as $m\to+\infty$; assume that there exists $\theta>0$ such that $\theta_{\mathscr{M}^{(m)}}\le\theta$ for any $m\in\mathbb{N}$ (with $\theta_{\mathscr{M}^{(m)}}$ defined by (1.5)).*

Let $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be a sequence of solutions to the scheme (1.6) satisfying (1.32) and (1.33) converging to $(\bar h,\bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$. Then $(\bar h,\bar{\boldsymbol{u}})$ satisfies the weak formulation (1.30) of the shallow water equations.

Similarly, if $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}},(\hat h^{(m)},\hat{\boldsymbol{u}}^{(m)})_{m\in\mathbb{N}}$ are sequences of solutions to the scheme (1.22) both uniformly bounded in the sense of (1.32) and (1.33) and converging to $(\bar h,\bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$, then the limit $(\bar h,\bar{\boldsymbol{u}})$ satisfies (1.30).

The proof of this theorem is the object of the following paragraphs; it relies on some general consistency lemmas which generalize the results of Gallouët, Herbin, and Latché 2019 to staggered meshes; these results are independent of the problem at hand and are given in the Section 1.A above. The proof of the consistency of the schemes is given in Section 1.4.1 for the forward Euler time discretization and in Section 1.4.2 for the Heun time discretization.

Note that because the convergence and boundedness of the approximate solutions are assumed, no CFL condition is required in Theorem 1.1. However, recall that a CFL condition is for instance already needed to show the positivity of the water height, see Lemma 1.1.

Finally, in Section 1.4.3, we give some conditions that imply the boundedness and convergence of the sequence $(\hat h^{(m)},\hat{\boldsymbol{u}}^{(m)})_{m\in\mathbb{N}}$ if the boundedness and convergence of the sequence $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ is assumed. One of this condition is a rather strong CFL-like condition.

1.4.1 Proof of consistency of the forward Euler MAC scheme

1.4.1.1 Consistency, mass equation

Under the assumptions of Theorem 1.1, the aim here is to prove that the limit $(\bar h,\bar{\boldsymbol{u}})$ of the scheme (1.6) satisfies the weak form of the mass equation (1.30a). In order to do so, we apply the consistency result of Lemma 1.8 in the section 1.A, with $U=(h,\boldsymbol{u})$,

$\beta(U) = h$, $\boldsymbol{f}(U) = h\boldsymbol{u}$, $\mathscr{P}^{(m)} = \mathscr{M}^{(m)}$, $\mathfrak{F}^{(m)} = \mathscr{E}^{(m)}$, and

$$\begin{aligned}\mathscr{C}^{(m)}_{\text{MASS}}(U^{(m)}):\quad & \Omega\times(0,T)\to\mathbb{R},\\ & (\boldsymbol{x},t)\mapsto \eth_t(h^{(m)})^{n+1}_K + \operatorname{div}_K((h^{(m)})^n\boldsymbol{u}^n) \text{ for } \boldsymbol{x}\in K \text{ and } t\in(t_n,t_{n+1})\end{aligned} \tag{1.34}$$

We first note that the assumptions (1.32) and (1.33) imply that (1.77) holds. Furthermore, the assumption of Theorem 1.1 that $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ is a sequence of solutions to the scheme (1.6) converging to $(\bar{h},\bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$ implies that (1.78) holds.

By the initialisation (1.6a) of the scheme, it is clear that

$$\sum_{K\in\mathscr{M}^{(m)}}\int_K |(h^{(m)})^0_K - h_0(\boldsymbol{x}))|d\boldsymbol{x} = 0,$$

so that the assumption (1.80) is satisfied.

Since for any $n\in[\![0,N_m-1]\!]$ and $K\in\mathscr{M}$, one has $\beta(U^{(m)}(\boldsymbol{x},t)) = h^n_K$ for any $(\boldsymbol{x},t)\in K\times[t_n t_{n+1})$ and $(\beta^{(m)})^n_K = h^n_K$,

$$\sum_{n=0}^{N_m-1}\sum_{K\in\mathscr{M}^{(m)}}\int_{t_n}^{t_{n+1}}\int_K |(h^{(m)})^n_K - h^{(m)}(\boldsymbol{x},t))|d\boldsymbol{x}\,dt = 0,$$

and therefore the assumption (1.81) is also clearly satisfied. Now $(\boldsymbol{F}^{(m)})^n_\sigma = h^n_\sigma\boldsymbol{u}^n_\sigma$ and, because the velocity components are piecewise constant on different grids,

$$\boldsymbol{f}(U^m(\boldsymbol{x},t)) = (f_1(U^m(\boldsymbol{x},t)), f_2(U^m(\boldsymbol{x},t))), \text{ with}$$
$$f_i(U^m(\boldsymbol{x},t)) = \begin{cases} h^n_K\, u^n_{i,\sigma} \text{ if } \boldsymbol{x}\in D_{K,\sigma}\\ h^n_K\, u^n_{i,\sigma'} \text{ if } \boldsymbol{x}\in D_{K,\sigma'},\end{cases} \quad \text{with } K=[\sigma\sigma'] \text{ and where } \sigma \text{ and } \sigma'\perp\boldsymbol{e}^{(i)}.$$

For $t\in[t_n t_{n+1})$ and $\boldsymbol{x}\in K=[\sigma\sigma']$ with $\sigma = K|L$,

$$\begin{aligned}\left|\left((\boldsymbol{F}^{(m)})^n_\sigma - \boldsymbol{f}(U^m(\boldsymbol{x},t)\right)\cdot\boldsymbol{n}_{K,\sigma}\right| &= \left|\left(h^n_\sigma\boldsymbol{u}^n_\sigma - h^n_K\boldsymbol{u}^n_\sigma + h^n_K\boldsymbol{u}^n_\sigma - h^n_K\boldsymbol{u}(\boldsymbol{x},t)\right)\cdot\boldsymbol{n}_{K,\sigma}\right|\\ &\le C^u\left|h_K - h_L\right| + C^h\left|\boldsymbol{u}_\sigma - \boldsymbol{u}_{\sigma'}\right|.\end{aligned}$$

Thanks to Lemma 1.12 (Gallouët, Herbin, and Latché 2019, Lemma 4.2, recalled in Lemma 1.12 in the appendix below) we have

$$\sum_{n=0}^{N_m-1}\sum_{K\in\mathscr{M}^{(m)}}\int_{t_n}^{t_{n+1}}\frac{\operatorname{diam}(K)}{|K|}\int_K|\sigma|\left|\left(h^n_\sigma\boldsymbol{u}^n_\sigma - h(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)\right)\cdot\boldsymbol{n}_{K,\sigma}\right|d\boldsymbol{x}\,dt\to 0 \text{ as } m\to+\infty.$$

so that the assumption (1.82) is also satisfied.

Hence, by Lemma 1.8,

$$\forall \varphi \in C_c^\infty\big(\Omega \times [0,T)\big),\ \int_0^T \int_\Omega \mathscr{C}_{\text{MASS}}^{(m)}(U^{(m)})\varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\,dt \to -\int_\Omega h_0(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}$$
$$-\int_0^T \int_\Omega \Big[\bar{h}(\boldsymbol{x},t)\,\partial_t\varphi(\boldsymbol{x},t) + \bar{h}(\boldsymbol{x},t)\,\bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)\Big]d\boldsymbol{x}\,dt \text{ as } m \to +\infty. \quad (1.35)$$

From (1.6b) and (1.35), we conclude that the limit $(\bar{h},\bar{\boldsymbol{u}})$ of the approximate solutions defined by the forward Euler scheme (1.6) satisfies (1.30a).

1.4.1.2 Consistency, momentum equation

Let $\boldsymbol{\varphi} = (\varphi_1,\cdots,\varphi_d) \in (C_c^\infty(\Omega\times[0,T)))^d$ be a test function and let $\varphi_{i,\sigma}^{n+1}$ denote the mean value of φ_i over $\sigma\times(t_n,t_{n+1})$. Multiplying the equation (1.6d) by $|D_\sigma|\varphi_{i,\sigma}^{n+1}$, summing the result over $\sigma\in\mathscr{E}^{(i)}$ and then summing over $n\in[\![0,N-1]\!]$ and $i=1,2$ yields:

$$\sum_{i=1}^{2} Q_{1,i}^{(m)} + Q_{2,i}^{(m)} + Q_{3,i}^{(m)} + Q_{4,i}^{(m)} = 0, \quad (1.36)$$

with (dropping the exponents (m) in the summations for the sake of simplicity)

$$Q_{1,i}^{(m)} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in(\mathscr{E}^{(m)})^{(i)}} |D_\sigma|\ \eth_t(h\,u_i)_\sigma^{n+1}, \quad (1.37)$$

$$Q_{2,i}^{(m)} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in(\mathscr{E}^{(m)})^{(i)}} |D_\sigma|\ \mathrm{div}_{D_\sigma}(h^n\boldsymbol{u}^n u_i^n)\varphi_{i,\sigma}^{n+1}, \quad (1.38)$$

$$Q_{3,i}^{(m)} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in(\mathscr{E}^{(m)})^{(i)}} |D_\sigma|\ (\eth_i p)_\sigma^{n+1}\ \varphi_{i,\sigma}^{n+1}, \quad (1.39)$$

$$Q_{4,i}^{(m)} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in(\mathscr{E}^{(m)})^{(i)}} |D_\sigma| g\, h_{\sigma,c}^{n+1}\ (\eth_i z)_\sigma\ \varphi_{i,\sigma}^{n+1}. \quad (1.40)$$

The nonlinear convection operator. In order to study the limit of the discrete non linear convection operator defined by $Q_i^{(m)} = Q_{1,i}^{(m)} + Q_{2,i}^{(m)}$, we apply Lemma 1.8 with $U=(h,\boldsymbol{u})$, $\beta(U)=hu_i$, $\boldsymbol{f}(U)=h\boldsymbol{u}u_i$, with $\mathscr{P}^{(m)}$ the set of dual cells associated with u_i (that is with the cells corresponding to the vertical edges for $i=1$ and the horizontal edges for $i=2$), with $\mathfrak{F}=(\widetilde{\mathscr{E}^{(m)}})^{(i)}$ and with the dual fluxes $(\boldsymbol{G})_\epsilon^n$ defined by (1.18). The

discrete non linear convection operator thus reads

$$[\mathscr{C}^{(m)}_{\mathrm{MOM}}(U^{(m)})]_i : \quad \Omega\times(0,T)\to\mathbb{R},\\ (\boldsymbol{x},t)\mapsto \eth_t(hu_i)^{n+1}_\sigma - \operatorname{div}_{D_\sigma}(h^n u_i \boldsymbol{u}^n) \text{ for } \boldsymbol{x}\in D_\sigma \text{ and } t\in(t_n,t_{n+1}) \tag{1.41}$$

(again dropping the exponents $^{(m)}$ for the sake of simplicity).

Again, by the initialisation of the scheme (1.6a) and by the definition of $(h\boldsymbol{u})^0_{i,\sigma}$ (see (1.16)), it is clear that

$$\sum_{\sigma\in(\widetilde{\mathscr{E}}^{(m)})^{(i)}}\int_{D_\sigma}|(h\boldsymbol{u})^0_{i,\sigma}-h_0(\boldsymbol{x})u_{i,0}(\boldsymbol{x})|d\boldsymbol{x}=0$$

$$\text{and}\quad \sum_{n=0}^{N_m-1}\sum_{\sigma\in\mathscr{E}^{(m)}}\int_{t_n}^{t_{n+1}}\int_{D_\sigma}|(h\boldsymbol{u})^n_{i,\sigma}-h(\boldsymbol{x},t)u_i(\boldsymbol{x},t)|d\boldsymbol{x}\,dt=0,\ i=1,2.$$

so that the assumptions (1.80) and (1.81) are satisfied.

In order to show that the assumption (1.82) is satisfied, we need to show that

$$\sum_{n=0}^{N_m-1}\sum_{\sigma\in\mathscr{E}^{(m)}}\int_{t_n}^{t_{n+1}}\frac{\operatorname{diam}(D_\sigma)}{|D_\sigma|}\int_{D_\sigma}|\epsilon|\Big|\sum_{\epsilon\in(\widetilde{\mathscr{E}}^{(m)})^{(i)}}\Big((\boldsymbol{G}^{(m)})^n_\epsilon-h(\boldsymbol{x},t)u_i(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)\Big)\cdot\boldsymbol{n}_{\sigma,\epsilon}\Big|d\boldsymbol{x}\,dt \\ \to 0 \text{ as } m\to+\infty. \tag{1.42}$$

Let us then estimate, for any $\epsilon\in(\widetilde{\mathscr{E}}^{(m)})^{(i)}$, $n\in[\![0,N_m-1]\!]$ and $\boldsymbol{x}\in D_\sigma$ the quantity Y^n_ϵ defined by:

$$Y^n_\epsilon(\boldsymbol{x})=\Big|\Big((\boldsymbol{G}^{(m)})^n_\epsilon-h(\boldsymbol{x},t)u_i(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)\Big)\cdot\boldsymbol{n}_{\sigma,\epsilon}\Big|.$$

Let L be the (primal) cell such that $\sigma=K|L$.

1. If $\epsilon=\sigma'|\sigma\subset K$, then $(\boldsymbol{G}^{(m)})^n_\epsilon$ is defined by (1.18a). By the triangular inequality and thanks to the assumptions (1.9), (1.19),(1.32), and (1.33), we get that

$$Y^n_\epsilon(\boldsymbol{x})\le\frac12(C^u)^2|h_K-h_L|+\frac12(C^u)^2|h_K-h_J|+C^hC^u|u_{\sigma,i}-u_{\sigma',i}|,\ \forall\boldsymbol{x}\in D_\sigma,$$

   where J is the (primal) cell such that $\sigma'=J|K$, see Figure 1.2, left.

2. If $\epsilon\subset K$, then $(\boldsymbol{G}^{(m)})^n_\epsilon$ is defined by (1.18b). Again by the triangular inequality and thanks to the assumptions (1.9), (1.19),(1.32), and (1.33), we get that

$$Y^n_\epsilon(\boldsymbol{x})\le\frac12(C^u)^2|h_K-h_M|+\frac12(C^u)^2|h_K-h_N|+C^hC^u|u_{\sigma,i}-u_{\sigma',i}|,\ \forall\boldsymbol{x}\in D_\sigma,$$

   where M and N are the two (primal) cells such that $\tau=K|M$ and $\tau'=L|N$, as depicted on Figure 1.2, right.

Now recall that the sequence of meshes is assumed to be regular in the sense that

$\theta^{(m)} \le \theta$ with $\theta^{(m)}$ defined by (1.5); therefore, since the sequences $h^{(m)}$ and $\boldsymbol{u}^{(m)}$ converge in L^1 as m tends to $+\infty$, we may again apply Lemma 1.12 below, to get that (1.42) holds. Hence, owing to Lemma 1.8, we get that

$$Q_i^{(m)} = Q_{1,i}^{(m)} + Q_{2,i}^{(m)} = \int_0^T \int_\Omega \mathscr{C}_{\text{MOM}}^{(m)}(U^{(m)})\ \varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\, dt \to$$
$$\int_0^T \int_\Omega \Big[\bar{h}(\boldsymbol{x},t)\,\bar{u}_i(\boldsymbol{x},t)\,\partial_t\varphi(\boldsymbol{x},t) + \bar{h}(\boldsymbol{x},t)\,\bar{u}_i(\boldsymbol{x},t)\,\bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)\Big] d\boldsymbol{x}\, dt$$
$$+ \int_\Omega h_0(\boldsymbol{x})\,u_{i,0}(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} \text{ as } m \to +\infty. \quad (1.43)$$

Pressure gradient and bathymetry Let us now study the terms $Q_{3,i}^{(m)}$ and $Q_{4,i}^{(m)}$ defined by (1.39) and (1.40). By the definition (1.11) of $\eth_\sigma p$ and by conservativity, we have (again dropping the exponents (m))

$$\sum_{i=1}^2 Q_{3,i}^{(m)} = -\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{K\in\mathscr{M}^{(m)}} \sum_{\sigma\in\mathscr{E}(K)} p_K^{n+1} \int_\sigma \boldsymbol{\varphi}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma}$$
$$= -\int_0^T \int_\Omega p^{(m)}(\boldsymbol{x},t)\ \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t$$

Since the sequence $(h^{(m)})m\in\mathbb{N}$ is bounded in $L^\infty(\Omega\times(0,T))$ and converges to $\bar{h}$ in $L^1(\Omega\times(0,T))$, the sequence $(p^{(m)})_{m\in\mathbb{N}}$ converges to $\bar{p} = \frac{1}{2}g\bar{h}^2$ in $L^1(\Omega\times(0,T))$ as $m\to+\infty$. Hence we get

$$\sum_{i=1}^2 Q_{3,i}^{(m)} \to \int_0^T \int_\Omega \bar{p}(\boldsymbol{x},t)\ \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty. \quad (1.44)$$

Let us now turn to the bathymetry term $Q_{4,i}^{(m)}$, which may be written

$$Q_{4,i}^{(m)} = \int_0^T \int_\Omega \widetilde{h}^{(m)}(\boldsymbol{x},t)\eth_i^{(m)} z(\boldsymbol{x})\widetilde{\varphi}_i^{(m)}(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t,$$

where

— the function $\widetilde{h}^{(m)} : \Omega\times(0,T)\to\mathbb{R}$ is defined by $\widetilde{h}(\boldsymbol{x},t) = h_{\sigma,c}^{n+1} = \frac{1}{2}(h_K^{n+1}+h_L^{n+1})$ for $\boldsymbol{x}\in D_\sigma$ and $t\in(t_n,t_{n+1})$; the sequence $(\widetilde{h}^{(m)})_{m\in\mathbb{N}}$ is therefore bounded in $L^\infty(\Omega\times(0,T))$ and converges to $\bar{h}$ in $L^1(\Omega\times(0,T))$;

— the function $\widetilde{\varphi}_i^{(m)} : \Omega\times(0,T)\to\mathbb{R}$ is defined by $\widetilde{\varphi}_i^{(m)}(\boldsymbol{x},t) = \varphi_\sigma^{n+1}$ for $\boldsymbol{x}\in D_\sigma$ and $t\in(t_n,t_{n+1})$; by the regularity of $\boldsymbol{\varphi}$, the sequence $(\widetilde{\varphi}_i^{(m)})_{m\in\mathbb{N}}$ converges to φ_i uniformly.

— by (1.13), the function $\eth_i^{(m)} z : \Omega \to \mathbb{R}$ is defined by

$$\eth_i^{(m)} z = \sum_{\substack{\sigma \in \mathscr{E}_{\rm int} \\ \sigma = K|L}} \frac{|\sigma|}{|D_\sigma|} \big(z(\boldsymbol{x}_L) - z(\boldsymbol{x}_K)\big) \mathbb{1}_{D_\sigma}$$

. Since z is a regular function, the sequence of functions $(\eth_i^{(m)} z)_{m\in\mathbb{N}}$ converges uniformly to the derivative $\partial_i z$ of z with respect to the i-th variable as $m \to +\infty$. Hence

$$Q_{4,i}^{(m)} \to \int_0^T \int_\Omega \bar{h}(\boldsymbol{x},t)\partial_i z(\boldsymbol{x})\varphi_i(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m \to \infty. \tag{1.45}$$

Limit of the momentum equation Passing to the limit in (1.36) as $m \to +\infty$, using (1.43), (1.44) and (1.45), we get that the limit $(\bar{h}, \bar{\boldsymbol{u}})$ of the approximate solutions defined by the forward Euler scheme (1.6) satisfies (1.30b), which concludes first part of the proof of Theorem 1.1.

1.4.2 Proof of the weak consistency of the Heun scheme

1.4.2.1 Mass balance

Under the assumptions of Theorem 1.1, the aim here is to prove that the limit $(\bar{h}, \bar{\boldsymbol{u}})$ of the scheme (1.22a)-(1.22f) satisfies the weak form of the mass equation (1.30a). In order to do so, we consider the equivalent mass equation (1.24a). Because of the structure of the scheme, we cannot use here Lemma 1.8 straightforwardly as in the case of the forward Euler scheme. We apply Lemma 1.9 with $U = (h, \boldsymbol{u})$, $\beta(U) = h$, $\boldsymbol{f}(U) = h\boldsymbol{u}$, $\mathscr{P}^{(m)} = \mathscr{M}^{(m)}, \mathfrak{F}^{(m)} = \mathscr{E}^{(m)}$ and then Lemma 1.10 twice: once with $U = (h, \boldsymbol{u})$, $\boldsymbol{f}(U) = h\boldsymbol{u}$, $\mathscr{P}^{(m)} = \mathscr{M}^{(m)}, \mathfrak{F}^{(m)} = \mathscr{E}^{(m)}$, and then with $U = (\hat{h}, \hat{\boldsymbol{u}})$, $\boldsymbol{f}(U) = \hat{h}\hat{\boldsymbol{u}}$. Thanks to the arguments developed in Section 1.4.1.1, it is easy to check that in each case, the assumptions of the lemmas are satisfied, so that we can conclude that $(\bar{h}, \bar{\boldsymbol{u}})$ satisfies (1.30a).

1.4.2.2 Momentum balance

Still under the assumptions of Theorem 1.1, we now prove that the limit $(\bar{h}, \bar{\boldsymbol{u}})$ of the scheme (1.22a)-(1.22f) satisfies the weak form of the mass equation (1.30b). Again we consider the equivalent momentum equation (1.24b). Multiplying the equation (1.24b) by $|D_\sigma|\varphi_{i,\sigma}^{n+1}$, summing the result over $\sigma \in \mathscr{E}^{(i)}$ and then summing over $n \in [\![0, N-1]\!]$ and $i = 1, 2$ yields:

$$\sum_{i=1}^{2} \left[Q_{1,i}^{(m)} + \frac{1}{2}(Q_{2,i}^{(m)} + \widehat{Q}_{2,i}^{(m)} + Q_{3,i}^{(m)} + \widehat{Q}_{3,i}^{(m)}) + Q_{4,i}^{(m)} + \widehat{Q}_{4,i}^{(m)}) \right] = 0, \tag{1.46}$$

where $Q_{1,i}^{(m)},\dots,Q_{4,i}^{(m)}$ are defined by (1.37)-(1.40), and $\widehat{Q}_{2,i}^{(m)},\widehat{Q}_{3,i}^{(m)},\widehat{Q}_{4,i}^{(m)}$ are defined by (1.38)-(1.40), replacing the unknowns h,p,u by $\hat{h},\hat{p},\hat{u}$.

Again, because of the structure of the scheme, we cannot use Lemma 1.8 directly: we use Lemma 1.9 for the time derivative term $Q_{1,i}^{(m)}$ and Lemma 1.10 for the terms $Q_{2,i}^{(m)}$ and $\widehat{Q}_{2,i}^{(m)}$, with $\mathscr{P}^{(m)}$ the set of dual cells associated with u_i (that is with the vertical edges for $i=1$ and the horizontal edges for $i=2$), with $\mathfrak{F}=(\widetilde{\mathscr{E}}^{(m)})^{(i)}$ and with the dual fluxes $(\boldsymbol{G})_\epsilon^n$ defined by (1.18). We first apply Lemma 1.9 with $U=(h,\boldsymbol{u})$, $\beta(U)=hu_i$, $\boldsymbol{f}(U)=h\boldsymbol{u}u_i$, and then Lemma 1.10, once with $U=(h,\boldsymbol{u})$, $\beta(U)=hu_i$, $\boldsymbol{f}(U)=h\boldsymbol{u}u_i$ and then with $U=(\hat{h},\hat{\boldsymbol{u}})$, $\beta(U)=\hat{h}\hat{u}_i$, $\boldsymbol{f}(U)=\hat{h}\hat{\boldsymbol{u}}\hat{u}_i$. Thanks to the arguments developed in Section 1.4.1.2, it is easy to check that in each case, the assumptions of the lemmas are satisfied, so that

$$\lim_{m\to+\infty}\Big[Q_{1,i}^{(m)}+\frac{1}{2}(Q_{2,i}^{(m)}+\widehat{Q}_{2,i}^{(m)})\Big]=\int_\Omega h_0(\boldsymbol{x})\,u_{i,0}(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}$$
$$+\int_0^T\int_\Omega\Big[\bar{h}(\boldsymbol{x},t)\,\bar{u}_i(\boldsymbol{x},t)\,\partial_t\varphi(\boldsymbol{x},t)+\bar{h}(\boldsymbol{x},t)\,\bar{u}_i(\boldsymbol{x},t)\,\bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)\Big]d\boldsymbol{x}\,dt. \quad (1.47)$$

The proof of convergence of the pressure gradient and bathymetry terms $Q_{3,i}^{(m)}$, $Q_{4,i}^{(m)}$, $\widehat{Q}_{3,i}^{(m)}$ and $\widehat{Q}_{4,i}^{(m)}$ follow the exact same lines as that of the terms $Q_{3,i}^{(m)}$ and $Q_{4,i}$ in Section 1.4.1.2. Hence

$$\lim_{m\to+\infty}\sum_{i=1}^{2}\frac{1}{2}\Big(Q_{3,i}^{(m)}+\widehat{Q}_{3,i}^{(m)}+Q_{4,i}^{(m)}+\widehat{Q}_{4,i}^{(m)}\Big)$$
$$=\int_0^T\int_\Omega\Big(\bar{p}(\boldsymbol{x},t)\,\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x},t)+\bar{h}(\boldsymbol{x},t)\nabla z(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x},t)\Big)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \quad (1.48)$$

Therefore, owing to (1.47) and (1.48), we may pass to the limit in (1.46) and conclude that $(\bar{h},\bar{\boldsymbol{u}})$ satisfies (1.30b). This concludes the proof of Theorem 1.1.

1.4.3 A sufficient condition for the convergence of the intermediate solutions

In Theorem 1.1, we assumed the boundedness and convergence of both sequences $(h^{(m)},u^{(m)})$ and $(\hat{h}^{(m)},\hat{u}^{(m)})$. In fact, under a restricted CFL condition, we may prove that the convergence and boundedness of the sequence $(h^{(m)},u^{(m)})$ implies the convergence and boundedness of the sequence $(\hat{h}^{(m)},\hat{u}^{(m)})$.

Lemma 1.5 (Bound on the intermediate step, Heun scheme)**.** *Let $n\in\{0,\cdots,N_t-1\}$, let $(h_K^n)_{K\in\mathcal{M}}\subset\mathbb{R}_+^*$ and $(\boldsymbol{u}_\sigma^n)_{\sigma\in\mathscr{E}}\subset\mathbb{R}^d$ be given. Assume that there exists $\zeta\in(0,1)$ such that the following restricted CFL-like condition holds (note that it is slightly more restrictive*

than (2.16)):

$$2\,\delta t \le \zeta \frac{|K|}{\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,|\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}|} \text{ for all } K\in\mathscr{M}. \tag{1.49}$$

Let $C_{\mathscr{M}}^{\delta t}$, $C_{\mathscr{M}}^{h}$ and $C_{\mathscr{M}}^{u}\in\mathbb{R}_+^*$ be such that

$$\delta t \le C_{\mathscr{M}}^{\delta t}\min_{\sigma\in\mathscr{E}}|\sigma|, \tag{1.50a}$$

$$\frac{1}{C_{\mathscr{M}}^{h}} \le h_K^n \le C_{\mathscr{M}}^{h}, \forall n\in\{0,\cdots,N_t-1\}, \forall K\in\mathscr{M}, \tag{1.50b}$$

$$\max_{\sigma\in\mathscr{E}}|\boldsymbol{u}_\sigma^n| \le C_{\mathscr{M}}^{u}, \forall n\in\{0,\cdots,N_t-1\}. \tag{1.50c}$$

Then the solutions $(\hat{h}_K^{n+1})_{K\in\mathscr{M}}$ $(\hat{\boldsymbol{u}}_\sigma^n)_{\sigma\in\mathscr{E}}$ of the Heun steps (1.22a)-(1.22b) satisfy:

$$\frac{1-\zeta}{C_{\mathscr{M}}^{h}} \le \hat{h}_K^{n+1} \le 2C_{\mathscr{M}}^{h}\ \forall K\in\mathscr{M}, \tag{1.51a}$$

$$|\hat{\boldsymbol{u}}_\sigma^{n+1}| \le C_{\mathscr{M}}^{u} + C_{\mathscr{M}}^{\delta t}\frac{(C_{\mathscr{M}}^{h})^2}{1-\zeta}\Big(4(C_{\mathscr{M}}^{u})^2 + g(C_{\mathscr{M}}^{h}+||z||_\infty)\Big),\ \forall\sigma\in\mathscr{E}. \tag{1.51b}$$

Proof. From (1.22a) and by the definition (1.8) of the discrete divergence, we have

$$\hat{h}_K^{n+1} = h_K^n - \sum_{\sigma\in\mathscr{E}(K)}(\omega_{K,\sigma}^n)^+ h_\sigma^n + \sum_{\sigma\in\mathscr{E}(K)}(\omega_{K,\sigma}^n)^- h_\sigma^n \text{ with } \omega_{K,\sigma}^n = \delta t\frac{|\sigma|}{|K|}\,\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}.$$

Owing to (1.10), there exists $\alpha_\sigma^K\in[0,1]$ and $M_\sigma^K\in\mathscr{M}$ such that $h_\sigma^n - h_K^n = \alpha_\sigma^K(h_K^n - h_{M_\sigma^K}^n)$ if $\omega_{K,\sigma}^n\ge 0$, and therefore

$$h_\sigma^n = h_K^n(1+\alpha_\sigma^K) - \alpha_\sigma^K h_{M_\sigma^K}^n.$$

Hence

$$\hat{h}_K^{n+1} = \Big(1 - \sum_{\sigma\in\mathscr{E}(K)}(\omega_{K,\sigma}^n)^+(1+\alpha_\sigma^K)\Big)h_K^n + \sum_{\sigma\in\mathscr{E}(K)}(\omega_{K,\sigma}^n)^- h_\sigma^n + \sum_{\sigma\in\mathscr{E}(K)}\alpha_\sigma^K(\omega_{K,\sigma}^n)^+ h_{M_\sigma^K}^n.$$

Therefore, thanks to the condition (1.49), we get (1.51a).

Let us now prove (1.51b); from (1.22b) we have

$$\begin{aligned}\hat{u}_{i,\sigma}^{n+1} &= \frac{1}{\hat{h}_{D_\sigma}^{n+1}}\Big(h_{D_\sigma}^n u_{i,\sigma}^n - \frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}u_{i,\epsilon}^n\Big) - \frac{\delta t|\sigma|}{|D_\sigma|}\frac{g h_{\sigma,c}^n}{\hat{h}_{D_\sigma}^{n+1}}\big(h_L^n - h_K^n + z_L - z_K\big)\\ &= \frac{1}{\hat{h}_{D_\sigma}^{n+1}}\Big[\Big(h_{D_\sigma}^n - \frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\Big)u_{i,\sigma}^n - \frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)\Big]\\ &\qquad - \frac{\delta t|\sigma|}{|D_\sigma|}\frac{g h_{\sigma,c}^n}{\hat{h}_{D_\sigma}^{n+1}}\big(h_L^n - h_K^n + z_L - z_K\big).\end{aligned}$$

Since the values $\hat{h}_{D_\sigma}^{n+1}$ and $\hat{h}_{D_\sigma}^{n}$ are computed by an equivalent formula to (1.17), they satisfy a discrete dual mass balance of the type (1.21), and therefore:

$$\hat{u}_{i,\sigma}^{n+1} = u_{i,\sigma}^{n} - \frac{1}{\hat{h}_{D_\sigma}^{n+1}}\left[\frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\tilde{\mathscr{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)\right] - \frac{\delta t|\sigma|}{|D_\sigma|}\frac{g h_{\sigma,c}^n}{\hat{h}_{D_\sigma}^{n+1}}\left(h_L^n - h_K^n + z_L - z_K\right).$$

Thanks to the CFL condition (1.50a) and to the bounds on $\boldsymbol{u}_\sigma^n$ and $\hat{h}_{D_\sigma}^{n+1}$ for all σ (recall that for $\sigma = K|L$, $\hat{h}_{D_\sigma}^{n+1}$ is a convex combination of $\hat{h}_K^{n+1}$ and $\hat{h}_L^{n+1}$),

$$\frac{1}{\hat{h}_{D_\sigma}^{n+1}}\left|\frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\tilde{\mathscr{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)\right| \le \frac{4C_{\mathscr{M}}^{\delta} t (C_{\mathscr{M}}^u)^2 (C_{\mathscr{M}}^h)^2}{1-\zeta}.$$

Furthermore, since $2h_{\sigma,c}^n = h_K^n + h_L^n$ and agina owing to (1.50a),

$$\begin{aligned}\frac{\delta t|\sigma|}{|D_\sigma|}\frac{g h_{\sigma,c}^n}{\hat{h}_{D_\sigma}^{n+1}}\left(h_L^n - h_K^n + z_L - z_K\right) &\le C_{\mathscr{M}}^{\delta t} g\frac{(C_{\mathscr{M}}^h)^2}{1-\zeta}\left(\max_{K\in\mathscr{M}}(h_K^n) + \max_{K\in\mathscr{M}}(z_K)\right)\\ &\le \frac{C_{\mathscr{M}}^{\delta t}(C_{\mathscr{M}}^h)^2 g}{1-\zeta}(C_{\mathscr{M}}^h + ||z||_\infty).\end{aligned}$$

Therefore,

$$|\hat{u}_{i,\sigma}^{n+1}| \le C_{\mathscr{M}}^u + \frac{4C_{\mathscr{M}}^{\delta} t (C_{\mathscr{M}}^u)^2 (C_{\mathscr{M}}^h)^2}{1-\zeta} + \frac{C_{\mathscr{M}}^{\delta t}(C_{\mathscr{M}}^h)^2 g}{1-\zeta}(C_{\mathscr{M}}^h + ||z||_\infty),$$

which concludes the proof that (1.51b) holds. □

Lemma 1.6 (L^1 convergence of the intermediate step, Heun scheme)**.** *Consider a sequence of meshes $(\mathscr{M}^{(m)}, \mathscr{E}^{(m)})_{m\in\mathbb{N}}$ such that $\delta t^{(m)}$ and $\delta_{\mathscr{M}^{(m)}} \to 0$ as $m \to +\infty$; assume that $(\mathscr{M}^{(m)}, \mathscr{E}^{(m)})_{m\in\mathbb{N}}$ is uniformly regular, in the sense that there exists $\theta > 0$ such that $\theta_{\mathscr{M}^{(m)}} \le \theta$ for any $m \in \mathbb{N}$ (with $\theta_{\mathscr{M}^{(m)}}$ defined by (1.5)).*

Let $(h^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be a sequence of solutions to the scheme (1.6) satisfying (1.32) and (1.33) converging to $(\bar{h}, \bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T)) \times L^1(\Omega\times(0,T))^2$.

Assume that there exists $\zeta \in (0,1)$ such that the following restricted CFL-like condition holds:

$$2\,\delta t^{(m)} \le \zeta \frac{|K|}{\sum_{\sigma\in\mathscr{E}(K)}|\sigma||(\boldsymbol{u}^{(m)})_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}|}, \ \forall K\in\mathscr{M}^{(m)}, \forall m\in\mathbb{N}, \tag{1.52}$$

and assume that there exists $C^{\delta t} \in \mathbb{R}_+^*$ not depending on m such that

$$\delta t^{(m)} \le C^{\delta t}\min_{\sigma\in\mathscr{E}^{(m)}}|\sigma|, \forall m\in\mathbb{N}. \tag{1.53}$$

Then there exists $\widehat{C}^u, \widehat{C}^h \in \mathbb{R}_+^*$ *such that*

$$\frac{1}{\widehat{C}^h} < (\hat{h}^{(m)})_K^n \leq \widehat{C}^h, \quad \forall K \in \mathscr{M}^{(m)}, \tag{1.54a}$$

$$|(\hat{\boldsymbol{u}}^{(m)})_\sigma^n| \leq \widehat{C}^u, \quad \forall \sigma \in \mathscr{E}^{(m)}. \tag{1.54b}$$

Furthermore, the sequence $\left(\hat{h}^{(m)}, \hat{\boldsymbol{u}}^{(m)}\right)_{m\in\mathbb{N}}$ *converges to* $(\bar{h}, \bar{\boldsymbol{u}})$ *in* $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$.

Proof. Under the above assumptions, the hypotheses (1.49) and (1.50) hold uniformly with respect to m, so that the bounds (1.54) are a direct consequence of Lemma 1.5.

Now, from equation (1.22a), we get that

$$\hat{h}_K^{n+1} - h_K^n = -\frac{\delta t}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\,(h_\sigma^n - h_K^n) - \frac{\delta t}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, h_K^n\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma},\ \forall K\in\mathscr{M}^{(m)}.$$

For $K\in\mathscr{M}^{(m)}$, let us denote by $\sigma_{K,i}$ and $\sigma'_{K,i}$ the edges of K in the direction $i\in\{1,2\}$, so that $K = [\overrightarrow{\sigma_{K,i}\sigma'_{K,i}}]$ for $i\in\{1,2\}$; noting that $\boldsymbol{n}_{K,\sigma_{K,i}} = -\boldsymbol{n}_{K,\sigma'_{K,i}}$ and that $|\sigma_{K,i}| = |\sigma'_{K,i}|$, and owing to (1.32), we get that

$$\Big|\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, h_K^n\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\Big| \leq C^h \sum_{i=1}^d |\sigma_{K,i}|\,|u_{i,\sigma_{K,i}}^n - u_{i,\sigma'_{K,i}}^n|,\ \forall K\in\mathscr{M}^{(m)}.$$

Since h_σ is a convex combination of h_K and h_L, with K and L such that $\sigma = K|L$, we get:

$$|\hat{h}_K^{n+1} - h_K^n| \leq \sum_{\substack{\sigma\in\mathscr{E}(K)\\ \sigma=K|L}} \frac{\delta t^{(m)}}{|K|}|\sigma|\,|\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}|\,|h_L^n - h_K^n| + C^h\sum_{i=1}^2 \delta t\frac{|\sigma|}{|K|}|u_{i,\sigma_{K,i}}^n - u_{i,\sigma'_{K,i}}^n|,\ \forall K\in\mathscr{M}^{(m)}.$$

Noting that (1.53) implies that $\dfrac{\delta t^{(m)}}{|K|}|\sigma| \leq 1$ and thanks to the condition (1.33), we thus get that there exists $C\in\mathbb{R}_+$ depending on C^h, C^u, $C^{\delta t}$ such that

$$|\hat{h}_K^{n+1} - h_K^n| \leq C\Big[\sum_{\substack{\sigma\in\mathscr{E}(K)\\ \sigma=K|L}} |h_L^n - h_K^n| + \sum_{i=1}^d |u_{i,\sigma_{K,i}}^n - u_{i,\sigma'_{K,i}}^n|\Big],\ \forall K\in\mathscr{M}^{(m)}.$$

Multiplying this latter inequality by $|K|\delta t^{(m)}$ and summing over $K\in\mathscr{M}^{(m)}$ and $n\in[\![0,N]\!]$, using the uniform regularity of the mesh and owing again to the convergence result on the space translates given in Lemma 1.12, we conclude that

$$\int_0^T\int_\Omega |\hat{h}^{(m)} - h^{(m)}|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \to 0 \text{ as } m\to+\infty.$$

Let us now turn to the intermediate velocities. Owing to (1.22b), (1.23) and since

$\widehat{u}$ satisfies a dual mass balance of the form (1.21), we have for $\sigma = K|L \in (\mathscr{E}_{\rm int}^{(i)})^{(m)}$, $i \in \{1,2\}$:

$$\hat{h}_{D_\sigma}^{n+1}(\hat{u}_{i,\sigma}^{n+1} - u_{i,\sigma}^n) = -(\hat{h}_{D_\sigma}^{n+1} - h_{D_\sigma}^n)u_{i,\sigma}^n - \frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)} |\epsilon| \boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon} u_{i,\epsilon}^n - \delta t\, g h_{\sigma,c}^n\big((\eth_\sigma h^n) + (\eth_\sigma z)\big)$$

$$- \sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)} \frac{\delta t\,|\epsilon|}{|D_\sigma|} \boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon}(u_{i,\epsilon}^n - u_{i,\sigma}^n) + \frac{\delta t|\sigma|}{|D_\sigma|} g h_{\sigma,c}^n\big(h_L^n - h_K^n + z_L - z_K\big).$$

Hence, owing to (1.32), (1.33), (1.53) and to the fact that for $\epsilon = \sigma|\sigma'$, $u_{i,\epsilon}^n$ is a convex combination of $u_{i,\sigma}^n$ and $u_{i,\sigma'}^n$, there exists $C \in \mathbb{R}_+$ depending only on C^h, C^u, $C^{\delta t}$ and g such that

$$|\hat{u}_{i,\sigma}^{n+1} - u_{i,\sigma}^n| \le C\Big[\sum_{\substack{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)\\ \epsilon=\sigma|\sigma'}} |u_{i,\sigma'}^n - u_{i,\sigma}^n| + |h_L^n - h_K^n| + |z_L - z_K|\Big], \text{ for } i = 1,2.$$

Multiplying this latter inequality by $|D_\sigma|\delta t^{(m)}$ and summing over $\sigma \in \mathscr{M}^{(m)}$ and $n \in [\![0, N]\!]$, using the uniform regularity of the mesh and again thanks to Lemma 1.12 we conclude that

$$\int_0^T \int_\Omega |\hat{u}_i^{(m)} - u_i^{(m)}|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \to 0 \text{ as } m \to +\infty, \text{ for } i = 1,2.$$

□

1.5 Weak entropy consistency of the forward Euler-MAC scheme

The weak consistency to the entropy inequality is only proved under additional assumptions as stated in the following theorem.

Theorem 1.2 (Weak entropy consistency of the forward Euler MAC scheme)**.** *Let $(\mathscr{M}^{(m)}, \mathscr{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes such that $\delta t^{(m)}$ and $\delta_{\mathscr{M}^{(m)}} \to 0$ as $m \to +\infty$; assume that there exists $\theta > 0$ such that $\theta_{\mathscr{M}^{(m)}} \le \theta$ for any $m \in \mathbb{N}$ (with $\theta_{\mathscr{M}^{(m)}}$ defined by (1.5)). Let $(h^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be a sequence of solutions to the scheme (1.6) converging to $(\bar{h}, \bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$, such that (1.32), (1.33) hold. Assume the following CFL-like condition:*

$$\delta t^{(m)} \le \frac{|D_\sigma| h_{D_\sigma}^{n+1}}{\displaystyle\sum_{\substack{\epsilon\in\mathscr{E}D_\sigma\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |\epsilon|\, \boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon}}. \tag{1.55}$$

Assume furthermore that

$$\exists C_{BVt} \in \mathbb{R}_+ : \sum_{K\in\mathcal{M}^{(m)}} |K||(h^{(m)})_K^{n+1} - (h^{(m)})_K^n| \leq C_{BVt} \ \forall m \in \mathbb{N}, \tag{1.56a}$$

$$\frac{\delta t^{(m)}}{\inf_{K\in\mathcal{M}^{(m)}} \operatorname{diam}(K)} \to 0 \text{ as } m \to +\infty., \tag{1.56b}$$

and that the coefficients $\lambda_{K,\sigma}$ *and* $\mu_{\sigma,\epsilon}$ *in* (1.9) *and* (1.19) *satisfy:*

$$\lambda_{K,\sigma} \in [\frac{1}{2}, 1] : \text{ if } \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \geq 0, \tag{1.57}$$

$$\mu_{\sigma,\epsilon} \in [\frac{1}{2}, 1] : \text{ if } \boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} \geq 0. \tag{1.58}$$

Then $(\bar{h}, \bar{\boldsymbol{u}})$ *satisfies the entropy inequality* (1.31).

Note also that the condition (1.9) implies that

$$\forall\ K \in \mathcal{M},\ \forall \sigma = K|L \in \mathcal{E}_{\text{int}}(K) \text{ with } K \text{ such that } \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \geq 0,$$
$$h_\sigma = \frac{1}{2}(h_K + h_L) + (\lambda_{K,\sigma} - \frac{1}{2})(h_K - h_L), \text{ with } \lambda_{K,\sigma} - \frac{1}{2} \geq 0. \tag{1.59}$$

Also note that the condition (1.57) is rather restrictive. Indeed, it is satisfied by the usual two slopes minmod limiter Godlewski and P.-A. Raviart 1996 only in the case of a uniform Cartesian mesh Piar, Babik, Herbin, et al. 2013, and it is not satisfied by the three slopes minmod limiter.

Proof. Let $\varphi \in C_c^\infty(\Omega \times [0,T), \mathbb{R}_+)$, and for a given discretization $(\mathcal{M}^{(m)}, \mathcal{E}^{(m)})$ let φ_K^n (resp. φ_σ^n) denote the mean value of φ on $K \times (t_n, t_{n+1})$ (resp. $D_\sigma \times (t_n, t_{n+1})$), for any $K \in \mathcal{M}^{(m)}$ (resp. $\sigma \in \mathcal{E}^{(m)}$) and $n \in [\![0, N_m - 1]\!]$. Let us multiply the discrete kinetic energy balance (1.27) by $\delta t \varphi_\sigma^{n+1}$ and sum over $\sigma \in \mathcal{E}^{(m)}$ and $i \in \{1,2\}$; let us then multiply the discrete potential energy balance (3.29) by $\delta t |K| \varphi_K^n$ and sum over $K \in \mathcal{M}^{(m)}$. Summing the two resulting equations and summing over $n \in [\![0, N_m - 1]\!]$, we get, owing to lemmas 3.1 and 1.4,

$$\int_0^T \int_\Omega \mathscr{C}_{\text{KIN}}^{(m)}(U^{(m)})\varphi(\boldsymbol{x},t)\, d\boldsymbol{x}\, dt + \int_0^T \int_\Omega \mathscr{C}_{\text{POT}}^{(m)}(U^{(m)})\varphi(\boldsymbol{x},t)\, d\boldsymbol{x}\, dt + \mathscr{P}^{(m)} + \mathscr{Z}^{(m)}$$
$$= -\mathscr{R}_k^{(m)} - \mathscr{R}_p^{(m)}, \tag{1.60}$$

with

$$\mathscr{C}^{(m)}_{\rm KIN}(U^{(m)})|_{D_\sigma} = \sum_{i=1}^{2}\Big((\eth_t E_{k,i})^{n+1}_\sigma + \sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)}|\epsilon|\,\frac{1}{2}(u^n_{i,\epsilon})^2\boldsymbol{F}^n_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}\Big) \text{ with } (E_{k,i})^n_\sigma = \frac{1}{2}h^n_{D_\sigma}(u^n_{i,\sigma})^2,$$

$$\mathscr{C}^{(m)}_{\rm POT}(U^{(m)})|_K = \frac{1}{2}g(\eth_t h^2_K)^n + {\rm div}_K(p^n\boldsymbol{u}^n),$$

$$\mathscr{P}^{(m)} = \sum_{n=0}^{N_m-1}\delta t^{(m)}\Big[\sum_{i=1}^{2}\sum_{\sigma\in(\mathscr{E}^{(i)})^{(m)}}|D_\sigma|u^{n+1}_{i,\sigma}(\eth_i p^{n+1})_\sigma\varphi^{n+1}_\sigma + \sum_{K\in\mathscr{M}^{(m)}}|K|p^n_K{\rm div}_K\boldsymbol{u}^n\varphi^{n+1}_K\Big],$$

$$\begin{aligned}\mathscr{Z}^{(m)} = +\sum_{n=0}^{N_m-1}\delta t^{(m)}\Big[&\sum_{i=1}^{2}\sum_{\sigma\in(\mathscr{E}^{(i)})^{(m)}}|D_\sigma|h^{n+1}_{\sigma,c}u^{n+1}_{i,\sigma}(\eth_i z)_\sigma\varphi^{n+1}_\sigma\\ &+\sum_{K\in\mathscr{M}^{(m)}}g\Big(z_K(\eth_t h_K)^n + g\,z_K{\rm div}_K(h^n\boldsymbol{u}^n)\Big)\varphi^{n+1}_K\Big],\end{aligned}$$

$$\begin{aligned}\mathscr{R}^{(m)}_k = \sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathscr{E}^{(i,m)}_{\rm int}}\Big[&\frac{1}{2}\frac{|D_\sigma|}{\delta t^{(m)}}h^{n+1}_{D_\sigma}(u^{n+1}_\sigma - u^n_\sigma)^2\\ &+\sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)}|\epsilon|\,\boldsymbol{F}^n_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}\Big(-\frac{1}{2}(u^n_{i,\epsilon}-u^n_{i,\sigma})^2 + (u^n_{i,\epsilon}-u^n_{i,\sigma})(u^{n+1}_{i,\sigma}-u^n_{i,\sigma})\Big)\Big]\varphi^{n+1}_\sigma\end{aligned}$$

$$\begin{aligned}\mathscr{R}^{(m)}_p \geq \sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{K\in\mathscr{M}^{(m)}}\Big[&-\frac{1}{2}g\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\,(h^n_\sigma - h^n_K)^2\,\boldsymbol{u}^n_\sigma\cdot\boldsymbol{n}_{K,\sigma}\\ &+\sum_{\sigma\in\mathscr{E}(K)}|\sigma|g(h^{n+1}_K - h^n_K)\,h^n_\sigma\,\boldsymbol{u}^n_\sigma\cdot\boldsymbol{n}_{K,\sigma}\Big]\varphi^{n+1}_K.\end{aligned}$$

Kinetic energy convection term

Let us check that the above defined convection operator $\mathscr{C}^{(m)}_{\rm KIN}$ satisfies the hypotheses (1.80)–(1.82) of Lax-Wendroff type consistency Lemma 1.8 given in the appendix which we apply here with $d=2$, $\mathscr{P}^{(m)}$ and $\mathfrak{F}^{(m)}$ the i-th dual mesh and its set of edges, $U=(h,\boldsymbol{u})$, $\beta(U)=E_{k,i}(U)=\frac{1}{2}hu_i^2$, for $i=1,2$.

Let us start with the assumption (1.80). For a given function $\psi\in L^1(\Omega)$, and any subset A of Ω we denote by $\langle\psi\rangle_A$ the mean value of ψ on A. By definition of the kinetic energy, we have $(E_{k,i})^0_\sigma = \frac{1}{2}h^0_{D_\sigma}|u^0_{i,\sigma}|^2 = \frac{1}{2}\langle h_0\rangle_{D_\sigma}(|\langle u_{i,0}\rangle_{D_\sigma}|)^2$ and $E_{k,i}(U_0) = E_{k,i}(h_0,\boldsymbol{u}_0) = \frac{1}{2}h_0u^2_{i,0}$. Therefore, owing to the assumptions (1.32)-(1.33) on the functions $h^{(m)}$ and $\boldsymbol{u}^{(m)}$ and to the fact that these sequences converge in L^1

$$\begin{aligned}\sum_{P\in\mathscr{P}^{(m)}}\int_P|(\beta^{(m)})^0_P - \beta(U_0(\boldsymbol{x}))|d\boldsymbol{x} &= \sum_{\sigma\in\mathscr{E}^{(m)}}\int_{D_\sigma}|(E_{k,i})^0_\sigma - E_{k,i}(h_0,\boldsymbol{u}_0)|d\boldsymbol{x}\\ &= \frac{1}{2}\sum_{\sigma\in\mathscr{E}^{(m)}}|D_\sigma|\Big|\langle h_0\rangle_{D_\sigma}\langle u_{i,0}\rangle^2_{D_\sigma} - \langle h_0u^2_{i,0}\rangle_{D_\sigma}\Big|\\ &\to 0 \text{ as } m\to+\infty.\end{aligned}$$

The assumption (1.80) is thus satisfied.

Let us then note that the assumption (1.81), which reads

$$\sum_{n=0}^{N_m-1}\sum_{\sigma\in\mathscr{E}^{(m)}}\int_{t_n}^{t_{n+1}}\int_{D_\sigma}|(E_{k,i}^{(m)})_\sigma^n - E_{k,i}(U^{(m)}(\boldsymbol{x},t))|d\boldsymbol{x}\,dt\to 0 \text{ as } m\to+\infty,$$

is satisfied, again thanks to the assumptions (1.32)-(1.33) on the functions $h^{(m)}$ and $\boldsymbol{u}^{(m)}$ and to the fact that these sequences converge in L^1.

Let us now turn to the assumption (1.82), which reads

$$\sum_{n=0}^{N_m-1}\sum_{\sigma\in\mathscr{E}^{(m)}}\int_{t_n}^{t_{n+1}}\frac{\mathrm{diam}(D_\sigma)}{|D_\sigma|}\int_{D_\sigma}\Big|\sum_{\epsilon\in\tilde{\mathscr{E}}^{(m)}}|\epsilon|\Big((\boldsymbol{G}^{(m)})_\epsilon^n-\boldsymbol{f}(U^m(\boldsymbol{x},t))\Big)\cdot\boldsymbol{n}_{\sigma,\epsilon}\Big|d\boldsymbol{x}\,dt \to 0 \text{ as } m\to+\infty,$$

with $(\boldsymbol{G}^{(m)})_\epsilon^n=\frac{1}{2}(u_{i,\epsilon}^n)^2\boldsymbol{F}_\epsilon^n$ and $\boldsymbol{f}(U)=\frac{1}{2}h|u|^2u_i$. This assumption is indeed satisfied since $\boldsymbol{F}_\epsilon^n$ is a convex combination of $h_\sigma\boldsymbol{u}_\sigma$ and $h_{\sigma'}\boldsymbol{u}_{\sigma'}$ for $\epsilon=\sigma|\sigma'$, and thanks to the boundedness and convergence assumptions on the sequences $(h^{(m)})_{m\in\mathbb{N}}$ and $(\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$.

By Lemma 1.8, we thus get that

$$\int_0^T\int_\Omega(\eth_t E_{k,i})_\sigma^{n+1}+\sum_{\epsilon\in\tilde{\mathscr{E}}(D_\sigma)}|\epsilon|\frac{1}{2}(u_{i,\epsilon}^n)^2\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\varphi(\boldsymbol{x},t)\,d\boldsymbol{x},\,dt\to-\int_\Omega E_{k,i}(U^{(0)}\varphi(x,0)\,\mathrm{d}\boldsymbol{x}$$
$$-\int_0^T\int_\Omega E_{k,i}(\bar{U})\partial_t\varphi+\frac{1}{2}E_{k,i}(\bar{U})\bar{u}_i\partial_i\varphi\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty,$$

with $E_{k,i}(\bar{U})=\frac{1}{2}g\bar{h}\bar{u}_i^2$. Summing over $i=1,2$, we get that

$$\int_0^T\int_\Omega\mathscr{C}_{\mathrm{KIN}}^{(m)}(U^{(m)})\varphi(\boldsymbol{x},t)\,d\boldsymbol{x}\,dt\to-\int_\Omega E_k(U^{(0)})\varphi(x,0)\,\mathrm{d}\boldsymbol{x}$$
$$-\int_0^T\int_\Omega\Big[E_k(\bar{U})\partial_t\varphi+\frac{1}{2}E_k(\bar{U})\boldsymbol{u}\cdot\nabla\varphi\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty, \quad (1.61)$$

with $E_k(\bar{U})=\frac{1}{2}g\bar{|\boldsymbol{u}|}^2$.

Potential energy convection terms

Let us now check that the above defined convection operator $\mathscr{C}_{\mathrm{POT}}^{(m)}$ satisfies the hypotheses (1.80)–(1.82) of Lemma 1.8 which we now apply with $d=2$, $\mathscr{P}^{(m)}$ and $\mathfrak{F}^{(m)}$ the primal mesh and its set of edges, $U=(h,\boldsymbol{u})$, $\beta(U)=\frac{1}{2}gh^2$ and $\boldsymbol{f}(U)=\frac{1}{2}gh^2\boldsymbol{u}$.

Indeed,

$$\sum_{K\in\mathscr{M}}\int_K\Big|\langle h(\cdot,0)^2\rangle_K-h(x,0)^2\Big|\,\mathrm{d}\boldsymbol{x}\to 0 \text{ as } m\to+\infty,$$

so that the hypothesis (1.80) is satisfied. Next,

$$\sum_{n=0}^{N_m-1}\int_{t_n}^{t_{n+1}}\sum_{K\in\mathcal{M}}\int_K\Big|(h_K^n)^2-h^2(\boldsymbol{x},t)\Big|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\to 0 \text{ as } m\to+\infty,$$

thanks to the boundedness and convergence assumptions on the sequence $(h^{(m)})_{m\in\mathbb{N}}$. so that the hypothesis (1.81) is satisfied. Finally, the left hand side of (1.82) reads

$$\begin{aligned}X_F&=\sum_{n=0}^{N_m-1}\int_{t_n}^{t_{n+1}}\sum_{K\in\mathcal{M}}\frac{\mathrm{diam}(K)}{|K|}\int_K\Big|\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\Big(\frac{1}{2}g(h_\sigma^n)^2\boldsymbol{u}_\sigma^n-\frac{1}{2}gh^2(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)\Big)\cdot\boldsymbol{n}_{K,\sigma}\Big|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\\&=\sum_{n=0}^{N_m-1}\int_{t_n}^{t_{n+1}}\sum_{K\in\mathcal{M}}\frac{\mathrm{diam}(K)}{|K|}\Big|\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\int_{D_{K,\sigma}}\Big(\frac{1}{2}g(h_\sigma^n)^2-\frac{1}{2}g(h_K^n)^2\Big)\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\Big|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\\&\to 0 \text{ as } m\to+\infty\end{aligned}$$

thanks to the fact that h_σ^n is a convex combination of h_K^n and h_L^n for $\sigma=K|L$, and thanks to the boundedness and convergence assumptions on the sequences $(h^{(m)})_{m\in\mathbb{N}}$ and $(\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$. Therefore, the assumption (1.82) is also satisfied.

Hence by Lemma 1.8,

$$\begin{aligned}\int_0^T\int_\Omega\mathscr{C}_{\mathrm{POT}}^{(m)}(U^{(m)})\varphi(\boldsymbol{x},t)\,d\boldsymbol{x}\,dt\to&-\frac{1}{2}\int_\Omega gh^2(\boldsymbol{x},0)\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}\\&-\int_0^T\int_\Omega\Big[\frac{1}{2}g\bar{h}^2\,\partial_t\varphi+\frac{1}{2}g\bar{h}^2\,\bar{\boldsymbol{u}}\cdot\nabla\varphi\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty.\end{aligned}\quad(1.62)$$

Pressure terms

Let us rewrite $\mathscr{P}^{(m)}$ as

$$\mathscr{P}^{(m)}=\sum_{n=0}^{N_m-1}\delta t^{(m)}\Big(\sum_{i=1}^2A_i^{n+1}+B^{n+1}\Big)-\delta t^{(m)}B^0,$$

with $A_i^n=\sum_{\sigma\in(\mathscr{E}_{\mathrm{int}}^{(i)})^{(m)},\,\sigma=K|L}|D_\sigma|u_{i,\sigma}^n(\eth_i p^n)_\sigma\varphi_\sigma^n$, and $B^n=\sum_{K\in\mathcal{M}^{(m)}}|K|p_K^n\mathrm{div}_K(\boldsymbol{u}^n)\varphi_K^n$.

By Lemma 1.7 below,

$$\sum_{i=1}^2A_i^{n+1}+B^{n+1}=\sum_{K\in\mathcal{M}^{(m)}}\sum_{\sigma\in\mathscr{E}(K)}|D_{K,\sigma}|p_K^{n+1}\boldsymbol{u}_\sigma^{n+1}\cdot\frac{|\sigma|(\varphi_K^{n+1}-\varphi_\sigma^{n+1})}{|D_{K,\sigma}|}\,\boldsymbol{n}_{K,\sigma}$$

On each subcell $D_{K,\sigma}$ the quantity $\dfrac{|\sigma|(\varphi_K^{n+1}-\varphi_\sigma^{n+1})}{|D_{K,\sigma}|}\,\boldsymbol{n}_{K,\sigma}$ is, up to higher order terms, a discrete differential quotient of φ between $\boldsymbol{x}_K$ and $\boldsymbol{x}_\sigma$, in the direction i if $\sigma\in\mathscr{E}^{(i)}$,

which uniformly converges to $\partial_i\varphi \boldsymbol{e}_i$ in the case of a rectangular grid, and therefore,

$$\sum_{n=0}^{N_m-1} \delta t^{(m)}(A^{n+1}+B^{n+1}) \to -\int_0^T \int_\Omega \bar{p}(\boldsymbol{x},t)\ \bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)d\boldsymbol{x}\,dt \text{ as } m\to+\infty.$$

Now, since we assume $u_0 \in L^1(\Omega)$,

$$\begin{aligned}\delta t^{(m)}|B^0| &= \delta t^{(m)}\Big|\sum_{K\in\mathscr{M}^{(m)}} |K| p_K^0 \mathrm{div}_K(\boldsymbol{u}^0)\varphi_K^0\Big| \\ &\le g\delta t^{(m)}\|h_0\|_\infty^2\|\varphi\|_\infty \sum_{K\in\mathscr{M}^{(m)}} |K||\mathrm{div}_K(\boldsymbol{u}^0)| \\ &\le 2g\frac{\delta t^{(m)}}{\inf_{K\in\mathscr{M}^{(m)}}\mathrm{diam}(K)}\|h_0\|_\infty^2\|\varphi\|_\infty \sum_{\sigma\in\mathscr{E}(K)} |\sigma| d_\sigma \|\boldsymbol{u}_0\|_\infty,\end{aligned}$$

so that, by the assumption (1.56b), $\sum_{n=0}^{N_m-1} \delta t^{(m)} B^0 \to 0$ as $m\to+\infty$; and therefore,

$$\mathscr{P}^{(m)} \to -\int_0^T \int_\Omega \bar{p}(\boldsymbol{x},t)\ \bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)d\boldsymbol{x}\,dt \text{ as } m\to+\infty. \tag{1.63}$$

In the above bound, we used the assumption (1.56b); this could be avoided if we assume $u_0 \in W^{1,1}(\Omega)$ or $u_0 \in L^1(0,T;BV(\Omega))$; indeed, in this case we have

$$|B^0| \le g\|h_0\|_\infty^2\|\varphi\|_\infty\|\boldsymbol{u}_0\|_{W^{1,1}(\Omega)}.$$

However, the assumption (1.56a) seems unavoidable to deal with the remainder term appearing in the discrete potential energy, see below.

Bathymetry terms

Let us introduce the following piecewise constant functions:

— $\widetilde{h}^{(m)}$ is the piecewise constant function equal to $h_{\sigma,c}^{n+1} = \frac{1}{2}(h_K^{n+1}+h_L^{n+1})$ on each set $D_\sigma\times(t_n,t_{n+1})$, for $\sigma = K|L \in \mathscr{E}_{\mathrm{int}}^{(m)}$ and $n\in[\![0,N_m-1]\!]$;

— $\nabla^{(m)}z^{(m)}$ is the piecewise constant function equal to $\dfrac{|\sigma|}{|D_\sigma|}(z_L-z_K)$ on each set D_σ, for $\sigma = K|L\in\mathscr{E}_{\mathrm{int}}^{(m)}$;

— $\widetilde{\varphi}^{(m)}$ is the piecewise constant function equal to φ_σ on each set on each set $D_\sigma\times(t_n,t_{n+1})$, for $\sigma = K|L\in\mathscr{E}_{\mathrm{int}}^{(m)}$ and $n\in[\![0,N_m-1]\!]$;

With these notations, we get that

$$\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{i=1}^{2} \sum_{\sigma\in(\mathscr{E}^{(i)})^{(m)}} |D_\sigma| h_{\sigma,c}^{n+1} u_{i,\sigma}^{n+1} (\eth_i z)_\sigma \, \varphi_\sigma^{n+1}$$
$$= \int_\Omega \widetilde{h}^{(m)}(\boldsymbol{x},t)\boldsymbol{u}^{(m)}(\boldsymbol{x},t)\cdot\nabla z^{(m)}(\boldsymbol{x})\,\widetilde{\varphi}^{(m)}(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t$$
$$\to \int_\Omega h(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)\cdot\nabla z(\boldsymbol{x})\,\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty, \quad (1.64)$$

thanks to the convergence assumptions on $h^{(m)}$ and $\boldsymbol{u}^{(m)}$ and owing to the strong convergence of the discrete gradient $\nabla^{(m)}$ (which would be only a weak convergence in the case of a non rectangular mesh, see Gallouët, Herbin, and Latché 2019, Lemma 3.1).

Now let

$$T_K^n = g\,\eth_t h_K^{n+1}\, z_K \text{ and } Z_K^n = \frac{1}{|K|} g \sum_{\sigma\in\mathscr{E}(K)} |\sigma| h_\sigma^n \, \boldsymbol{u}_{K,\sigma}^n\cdot\boldsymbol{n}_{K,\sigma}^n \, z_K.$$

Using a discrete summation by parts in time and thanks to the convergence assumption on $h^{(m)}$, we get that

$$\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{K\in\mathscr{M}^{(m)}} |K| T_K^n \to -\int_\Omega g z(\boldsymbol{x}) h(\boldsymbol{x},0)\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}$$
$$-\int_0^T\int_\Omega g z(\boldsymbol{x}) h(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty. \quad (1.65)$$

Using next a discrete summation by parts in space, we get

$$\sum_{K\in\mathscr{M}^{(m)}} |K| Z_K^n = \sum_{K\in\mathscr{M}^{(m)}} g z_K \varphi_K^{n+1} \sum_{\sigma\in\mathscr{E}(K)} |\sigma| h_\sigma^n \, \boldsymbol{u}_{K,\sigma}^n\cdot\boldsymbol{n}_{K,\sigma}^n \, z_K$$
$$= \sum_{\substack{\sigma\in\mathscr{E}_{\rm int}^{(m)}\\ \sigma=K|L}} |\sigma| h_\sigma^n \, \boldsymbol{u}_{K,\sigma}^n\cdot\boldsymbol{n}_{K,\sigma}^n \, (z_K\varphi_K^{n+1} - z_L\varphi_L^{n+1})$$
$$= -\sum_{\substack{\sigma\in\mathscr{E}_{\rm int}^{(m)}\\ \sigma=K|L}} |D_\sigma| h_\sigma^n \boldsymbol{u}_\sigma^n\cdot(\nabla^{(m)}(z\varphi))_\sigma^{n+1},$$

where $\nabla^{(m)}(z\varphi)$ is the piecewise constant discrete gradient defined by:

$$\forall \sigma = K|L \in \mathscr{E}_{\rm int}^{(m)},\ \forall n\in[\![0,N_m-1]\!],\ \forall(\boldsymbol{x},t)\in D_\sigma\times[t_n,t_{n+1}),$$
$$\nabla^{(m)}(z\varphi)^{n+1}(\boldsymbol{x},t) = (\nabla^{(m)}(z\varphi))_\sigma = ^{n+1} \frac{|\sigma|}{|D_\sigma|}(z_K\varphi_K^{n+1} - z_L\varphi_L^{n+1})\boldsymbol{n}_{K,\sigma},$$

which converges to $\nabla(z\varphi)$ uniformly in the case of a rectangular mesh, and weakly in

the case of a general mesh, see Gallouët, Herbin, and Latché 2019, Lemma 3.1.

Therefore, thanks to the convergence assumptions on h and $\boldsymbol{u}$,

$$\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{K\in\mathcal{M}^{(m)}}|K|Z_K^n\to-\int_0^T\int_\Omega g\bar{h}(\boldsymbol{x},t)\bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla(z\varphi)(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty. \tag{1.66}$$

Owing to (1.64), (1.65) and (1.66), we thus get that

$$\begin{aligned}\mathcal{Z}^{(m)}\to&-\int_\Omega gz(\boldsymbol{x})h(\boldsymbol{x},0)\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}-\int_0^T\int_\Omega gz(\boldsymbol{x})\bar{h}(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\\&-\int_0^T\int_\Omega g\bar{h}(\boldsymbol{x},t)z(\boldsymbol{x})\bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty.\end{aligned} \tag{1.67}$$

Remainder terms

The remainder term $\mathscr{R}_k^{(m)}$ in (1.60) satisfies

$$\mathscr{R}_k^{(m)}=\mathscr{R}_{k,1}^{(m)}+\mathscr{R}_{k,2}^{(m)}+\mathscr{R}_{k,3}^{(m)} \tag{1.68}$$

with

$$\mathscr{R}_{k,1}^{(m)}=\frac{1}{2}\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathscr{E}_{\mathrm{int}}^{(i)}}\frac{1}{\delta t}\,|D_\sigma|\,h_{D_\sigma}^{n+1}\left(u_{i,\sigma}^{n+1}-u_{i,\sigma}^{n}\right)^2\varphi_\sigma^{n+1},$$

$$\mathscr{R}_{k,2}^{(m)}=-\frac{1}{2}\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathscr{E}_{\mathrm{int}}^{(i)}}\sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n-u_{i,\sigma}^n\right)^2\varphi_\sigma^{n+1}$$

$$\mathscr{R}_{k,3}^{(m)}=\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathscr{E}_{\mathrm{int}}^{(i)}}\sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\big(u_{i,\epsilon}^n-u_{i,\sigma}^n\big)\big(u_{i,\sigma}^{n+1}-u_{i,\sigma}^n\big)\varphi_\sigma^{n+1}.$$

The term $\mathscr{R}_{k,3}^{(m)}$ satisfies

$$\mathscr{R}_{k,3}^{(m)}\geq\mathscr{R}_{k,3,1}^{(m)}+\mathscr{R}_{k,3,2}^{(m)} \tag{1.69}$$

with

$$\mathscr{R}_{k,3,1}^{(m)}=-\frac{1}{2}\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathscr{E}_{\mathrm{int}}^{(i)}}\sum_{\substack{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\big(u_{i,\sigma}^{n+1}-u_{i,\sigma}^n\big)^2\varphi_\sigma^{n+1},$$

$$\mathscr{R}_{k,3,2}^{(m)}=-\frac{1}{2}\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathscr{E}_{\mathrm{int}}^{(i)}}\sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\big(u_{i,\epsilon}^n-u_{i,\sigma}^n\big)^2\varphi_\sigma^{n+1}.$$

Thanks to the CFL condition (1.55), we get that

$$\mathscr{R}_{k,1}^{(m)}+\mathscr{R}_{k,3,1}^{(m)}\geq 0. \tag{1.70}$$

Let us now study the term

$$\widetilde{\mathscr{R}}_{k,2}^{(m)} = \mathscr{R}_{k,3,2}^{(m)} + \mathscr{R}_{k,2}^{(m)} = -\sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathscr{E}_{\rm int}^{(i)}} \sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)} |\epsilon|\, \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)^2\varphi_\sigma^{n+1},$$

which we decompose as: $\widetilde{\mathscr{R}}_{k,2}^{(m)} \geq \widetilde{\mathscr{R}}_{k,2,1}^{(m)} + \widetilde{\mathscr{R}}_{k,2,2}^{(m)}$, with

$$\widetilde{\mathscr{R}}_{k,2,1}^{(m)} = -\sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathscr{E}_{\rm int}^{(i)}} \sum_{\substack{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |\epsilon|\, \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)^2\varphi_\epsilon^{n+1},$$

$$\widetilde{\mathscr{R}}_{k,2,2}^{(m)} = -\sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathscr{E}_{\rm int}^{(i)}} \sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)} |\epsilon|\, \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)^2(\varphi_\epsilon^{n+1} - \varphi_\sigma^{n+1}),$$

and, by conservativity,

$$\begin{aligned}\widetilde{\mathscr{R}}_{k,2,1}^{(m)} &\geq \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\substack{\epsilon=\sigma|\sigma'\in\widetilde{\mathscr{E}}_{int}^{(i)}\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |\epsilon|\, \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\left[\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)^2 - \left(u_{i,\epsilon}^n - u_{i,\sigma'}^n\right)^2\right]\varphi_\epsilon^{n+1}\\ &\geq \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)}\|u_i\|_\infty \sum_{\substack{\epsilon=\sigma|\sigma'\in\widetilde{\mathscr{E}}_{int}^{(i)}\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |D_\epsilon|(2u_{i,\epsilon} - u_{i,\sigma} - u_{i,\sigma'})(u_{i,\sigma'} - u_{i,\sigma}).\end{aligned}$$

Therefore, thanks to (1.19) and (1.58),

$$\widetilde{\mathscr{R}}_{k,2,1}^{(m)} \geq \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\substack{\epsilon=\sigma|\sigma'\in\widetilde{\mathscr{E}}_{int}^{(i)}\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |\epsilon|\, \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}(2\mu_{\sigma,\epsilon}-1)\left(u_{i,\sigma}^n - u_{i,\sigma'}^n\right)^2\varphi_\epsilon^{n+1} \geq 0. \quad (1.71)$$

Let us then write that, thanks to the regularity of φ,

$$\begin{aligned}|\widetilde{\mathscr{R}}_{k,2,2}^{(m)}| &\leq C_\varphi \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathscr{E}_i nti} \sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)} |D_\epsilon|\,|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}|\left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)^2\\ &\leq C_\varphi\|h\|_\infty\|\boldsymbol{u}\|_\infty \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathscr{E}_{\rm int}^{(i)}} \sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)} |D_\epsilon|\, u_{i,\epsilon}^n - u_{i,\sigma}^n|\end{aligned}$$

so that, thanks to the L^1 convergence of $\boldsymbol{u}^{(m)}$ and to the regularity of the mesh, we may again apply Lemma 1.12 to obtain

$$|\widetilde{\mathscr{R}}_{k,2,2}^{(m)}| \to 0 \text{ as } m \to +\infty. \quad (1.72)$$

Therefore, owing to (1.68)-(1.72)

$$\lim_{m\to+\infty} \mathscr{R}_k^{(m)} \geq 0. \tag{1.73}$$

Let us now turn to the remainder $\mathscr{R}_p^{(m)}$. We have $\mathscr{R}_p^{(m)} \geq \mathscr{R}_{p,1}^{(m)} + \mathscr{R}_{p,2}^{(m)}$, with

$$\mathscr{R}_{p,1}^{(m)} = -\sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{K\in\mathscr{M}^{(m)}} g \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,(h_\sigma^n - h_K^n)^2\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\varphi_K^n,$$

$$\mathscr{R}_{p,2}^{(m)} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{K\in\mathscr{M}^{(m)}} \sum_{\sigma\in\mathscr{E}(K)} |\sigma| g(h_K^{n+1} - h_K^n)\, h_\sigma^n\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\varphi_K^n.$$

Note that if h_σ^n is the upwind choice for any $\sigma\in\mathscr{E}^{(m)}$, then $\mathscr{R}_{p,1}^{(m)} \geq 0$. In the general case, we may write that

$$\mathscr{R}_{p,1}^{(m)} = \mathscr{R}_{p,1,1}^{(m)} + \mathscr{R}_{p,1,2}^{(m)}$$

with

$$\mathscr{R}_{p,1,1}^{(m)} = -\sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{K\in\mathscr{M}^{(m)}} g \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,(h_\sigma^n - h_K^n)^2\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\varphi_\sigma^n$$

$$\mathscr{R}_{p,1,2}^{(m)} = -\sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{K\in\mathscr{M}^{(m)}} g \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,(h_\sigma^n - h_K^n)^2\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}(\varphi_K^n - \varphi_\sigma^n).$$

By conservativity,

$$\begin{aligned}\mathscr{R}_{p,1,1}^{(m)} &\geq -g\sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{\substack{\sigma=K|L\in\mathscr{E}\\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}>0}} |\sigma| \left[(h_\sigma^n - h_K^n)^2 - (h_\sigma^n - h_L^n)^2\right] \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\varphi_\sigma^n \\ &= -g\sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{\substack{\sigma=K|L\in\mathscr{E}\\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}>0}} |\sigma|\,(h_L^n - h_K^n)(2h_\sigma^n - h_K^n - h_L^n)\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\varphi_\sigma^n.\end{aligned}$$

Owing to the assumption (1.9), one has

$$(h_L^n - h_K^n)(2h_\sigma^n - h_K^n - h_L^n) = -2\lambda_{K,\sigma}(h_K^n - h_L^n)$$

and since by (1.57), $\lambda_{K,\sigma} \geq \frac{1}{2}$ if $\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma} > 0$,

$$\mathscr{R}_{p,1,1}^{(m)} = 2g\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\substack{\sigma=K|L\in\mathscr{E}\\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}>0}} |\sigma|(\lambda_{K,\sigma} - \frac{1}{2})\,(h_L^n - h_K^n)^2\, \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\varphi_\sigma^n \geq 0.$$

Now

$$|\mathscr{R}_{p,1,2}^{(m)}| \leq \sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{K\in\mathscr{M}^{(m)}} g\|h\|_\infty \|\boldsymbol{u}\|_\infty C_\varphi \sum_{\sigma\in\mathscr{E}(K)} |D_{K,\sigma}|\,|h_L^n - h_K^n| \to 0 \text{ as } m \to +\infty$$

so that

$$\lim_{m\to+\infty} \mathscr{R}_{p,1}^{(m)} \geq 0.$$

Let us now turn to $\mathscr{R}_{p,2}^{(m)}$. Since for all $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}(K)$ we have

$$|\sigma| \leq \frac{|K|}{\inf_{K\in\mathscr{M}} \operatorname{diam}(K)},$$

we get

$$|\mathscr{R}_{p,2}^{(m)}| \leq g\|h\|_\infty \|\boldsymbol{u}\|_\infty \|\varphi\|_\infty \sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{\inf_{K\in\mathscr{M}^{(m)}} \operatorname{diam}(K)} \sum_{n=0}^{N_m-1} \sum_{K\in\mathscr{M}^{(m)}} |K|\,|(h^{(m)})_K^{n+1} - (h^{(m)})_K^n|$$
$$\to 0 \text{ as } m \to +\infty,$$

thanks to the assumption (1.56). Hence

$$\lim_{m\to+\infty} \mathscr{R}_{p}^{(m)} \geq 0. \tag{1.74}$$

Conclusion of the proof

Owing to (1.73) and (1.74), passing to the limit in (1.60) as $m \to +\infty$ yields, together with (1.61), (1.62), (1.63) and (1.67), that the limit $(\bar{h}, \bar{u})$ satisfies the weak entropy inequality (1.31). □

The next lemma, used to pass to the limit in the pressure terms of the entropy is the discrete equivalent, on a staggered grid, of the formal equality

$$\int_\Omega (\boldsymbol{u}\cdot\nabla p\,\varphi + p\,\operatorname{div}\boldsymbol{u}\,\varphi)\,\mathrm{d}\boldsymbol{x} = -\int_\Omega p\,\boldsymbol{u}\cdot\nabla\varphi\,d\boldsymbol{x}.$$

Lemma 1.7 (Pressure terms)**.** *Let $(\mathscr{M},\mathscr{E})$ be a MAC discretization of Ω in the sense of Definition 1.1 ; Let $(p_K)_{K\in\mathscr{M}} \subset \mathbb{R}$ and $(\boldsymbol{u}_\sigma)_{\sigma\in\mathscr{E}} \subset \mathbb{R}^d$ be some discrete unknowns associated to $\mathscr{M}$ and $\mathscr{E}$ respectively. Let $\varphi \in C_c^\infty(\Omega)$, and let φ_K (resp. φ_σ) denote the*

mean value of φ on K (resp. D_σ), for any $K \in \mathcal{M}$ (resp. $\sigma \in \mathcal{E}^{(m)}$). Then

$$\sum_{i=1}^{2} \sum_{\substack{\sigma \in \mathcal{E}_{\rm int}^{(i)} \\ \sigma = K|L}} |D_\sigma|\, u_{i,\sigma}\, (\eth_i p)_\sigma\, \varphi_\sigma + \sum_{K \in \mathcal{M}} |K|\, p_K\, {\rm div}_K \boldsymbol{u}\, \varphi_K$$
$$= \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}(K)} |D_{K,\sigma}|\, p_K\, \boldsymbol{u}_\sigma \cdot \frac{|\sigma|(\varphi_K - \varphi_\sigma)}{|D_{K,\sigma}|} \boldsymbol{n}_{K,\sigma}.$$

Proof. Let us denote by A and B the first and second terms of the right hand side. Then, with the notations of Definition 1.1,

$$A = \sum_{K \in \mathcal{M}^{(m)}} \sum_{i=1}^{2} \sum_{\sigma \in \mathcal{E}^{(i)}(K)} |D_{K,\sigma}| u_{i,\sigma}\, (\eth_i p)_\sigma\, \varphi_\sigma$$
$$= \sum_{K \in \mathcal{M}^{(m)}} \sum_{\substack{\sigma \in \mathcal{E}(K) \\ \sigma = K|L}} |D_{K,\sigma}|\, \boldsymbol{u}_\sigma \cdot \frac{p_L - p_K}{|D_\sigma|} |\sigma|\, \varphi_\sigma\, \boldsymbol{n}_{K,\sigma}$$
$$= \sum_{K \in \mathcal{M}^{(m)}} \sum_{\substack{\sigma \in \mathcal{E}(K) \\ \sigma = K|L}} |D_{K,\sigma}|\, \boldsymbol{u}_\sigma \cdot \frac{p_\sigma - p_K}{|D_{K,\sigma}|} |\sigma|\, \varphi_\sigma\, \boldsymbol{n}_{K,\sigma},$$

where p_σ is defined by $\dfrac{p_\sigma - p_K}{|D_{K,\sigma}|} = \dfrac{p_L - p_K}{|D_\sigma|}$. By conservativity,

$$\sum_{K \in \mathcal{M}^{(m)}} \sum_{\substack{\sigma \in \mathcal{E}(K) \\ \sigma = K|L}} \boldsymbol{u}_\sigma \cdot p_\sigma |\sigma| \varphi_\sigma\, \boldsymbol{n}_{K,\sigma} = 0,$$

so that

$$A = -\sum_{K \in \mathcal{M}^{(m)}} \sum_{\substack{\sigma \in \mathcal{E}(K) \\ \sigma = K|L}} |D_{K,\sigma}| \boldsymbol{u}_\sigma \cdot \frac{p_K}{|D_{K,\sigma}|} |\sigma|\, \varphi_\sigma\, \boldsymbol{n}_{K,\sigma}$$

Now

$$B = \sum_{K \in \mathcal{M}^{(m)}} |K|\, p_K\, {\rm div}_K \boldsymbol{u}\, \varphi_K = \sum_{K \in \mathcal{M}^{(m)}} \sum_{\sigma \in \mathcal{E}(K)} |\sigma|\, p_K\, |D_{K,\sigma}|\, \boldsymbol{u}_\sigma \cdot \varphi_K\, \boldsymbol{n}_{K,\sigma}.$$

Adding the results for A and B concludes the proof. □

1.6 Numerical results

This section is devoted to numerical tests: we first check the order of convergence of the proposed scheme on a two-dimensional regular solution (Section 1.6.1); then we turn to one-dimensional and two-dimensional shock solutions on a plane topography (Sections 1.6.2 and 1.6.3); in Section 1.6.4, we address a two-dimensional dam-break problem in a closed computational domain with a variable topography, which, in

particular, shows tha ability of staggered scheme to "natively" cope with reflection boundary conditions; finally, we compute the motion of a liquid slug over a partly dry support (1.6.5).

In this section, we compare three schemes: the second-order scheme developed here, the scheme referred to in Section 1.2.2 as the segregated forward Euler scheme (combining a segregated forward Euler scheme in time and the proposed MUSCL-like discretization of the convection fluxes) and a first order scheme which still features the segregated forward Euler scheme in time but with first-order upwind convection fluxes. These schemes are referred to in the following as the *second-order*, *segregated* and *first-order* scheme respectively.

The schemes have been implemented within the CALIF^{3}S open-source software CALIF^{3}S n.d. of the French Institut de Sûreté et de Radioprotection Nucléaire (IRSN); this software is used for the following tests.

1.6.1 A smooth solution

We begin here by checking the accuracy of the scheme on a known regular solution consisting in a travelling vortex. This solution is obtained through the following steps: we first derive a compact-support H^2 solution consisting in a standing vortex which becomes time-dependent by adding a constant velocity motion. The velocity field of the standing vortex and the pressure are sought under the form:

$$\hat{\boldsymbol{u}} = f(\xi)\begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}, \quad \hat{p} = \wp(\xi),$$

with $\xi = x_1^2 + x_2^2$. A simple derivation of these expressions yields:

$$\hat{\boldsymbol{u}} \cdot \nabla \hat{\boldsymbol{u}} = -f(\xi)^2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and

$$\nabla \hat{p} = 2\wp'(\xi)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Using the relation $p = \frac{1}{2}gh^2$, we thus obtain a stationary solution of the shallow water equations (1.1) with a topography $z = 0$ if $\wp$ satisfies $8g\wp = (F + c)^2$, where F is such that $F' = f^2$, $F(0) = 0$ and c is a positive real number. For the present numerical study, we choose $f(\xi) = 10\xi^2(1-\xi)^2$ if $\xi \in (0,1)$, $f = 0$ otherwise, which indeed yields an $H^2(\mathbb{R}^2)$ velocity field (note that as a consequence, the pressure and the water height are also regular), and $c = 1$. The problem is made unsteady by a time translation: given a constant vector field $\boldsymbol{a}$, the pressure p and the velocity $\boldsymbol{u}$ are deduced from the steady state solution $\hat{p}$ and $\hat{\boldsymbol{u}}$:

$$h(\boldsymbol{x}, t) = \hat{h}(\boldsymbol{x} - \boldsymbol{a}t), \qquad \boldsymbol{u}(\boldsymbol{x}, t) = \hat{\boldsymbol{u}}(\boldsymbol{x} - \boldsymbol{a}t) + \boldsymbol{a}.$$

The center of the vortex is initially located at $\boldsymbol{x}_0 = (0,0)^t$, the translation velocity $\boldsymbol{a}$ is set to $\boldsymbol{a} = (1,1)^t$, the computational domain is $\Omega = (-1.2, 2.)^2$ and the computation is run on the time interval $(0, 0.8)$. Computations are performed with successively

mesh	error(h)	ord(h)	error(u)	ord(u)
32×32	$3.61\,10^{-3}$	/	$2.93\,10^{-1}$	/
64×64	$1.15\,10^{-3}$	1.65	$1.14\,10^{-1}$	1.36
128×128	$2.58\,10^{-4}$	2.16	$4.06\,10^{-2}$	1.49
256×256	$5.85\,10^{-5}$	2.14	$1.49\,10^{-2}$	1.45
512×512	$1.53\,10^{-5}$	1.93	$4.67\,10^{-3}$	1.68

Table 1.1 – Measured numerical errors for the travelling vortex – Discrete L^1-norm of the difference between the numerical and exact solution at $t = 0.8$, for the height and the velocity, and corresponding order of convergence.

refined meshes with square cells, and the time step is $\delta t = \delta_{\mathcal{M}}/8$, and corresponds to a Courant (or CFL) number with respect to the celerity of the fastest waves close to $1/3$. The discrete L^1-norm of the difference between the exact solution and the solution obtained by the second-order scheme is given in Table 1.1. The observed order of convergence over the whole sequence is 2 for the water height and 1.5 for the velocity. Results with the first-order scheme are given in Table 1.2; one observes that the second-order scheme is much more accurate. Finally, the segregated scheme yields

mesh	error(h)	ord(h)	error(u)	ord(u)
32×32	$8.04\,10^{-3}$	/	$6.55\,10^{-1}$	/
64×64	$5.56\,10^{-3}$	0.53	$4.84\,10^{-1}$	0.44
128×128	$3.53\,10^{-3}$	0.66	$3.22\,10^{-1}$	0.59
256×256	$2.08\,10^{-3}$	0.76	$1.96\,10^{-1}$	0.72
512×512	$1.15\,10^{-3}$	0.85	$1.16\,10^{-1}$	0.76

Table 1.2 – Measured numerical errors for the travelling vortex with the first order scheme - Discrete L^1-norm of the difference between the numerical and exact solution at $t = 0.8$, for the height and the velocity, and corresponding order of convergence.

good results on coarse meshes (it is the most accurate scheme on the 32×32 mesh); unfortunately, when refining the mesh, oscillations appear, and the convergence is lost. This results confirms a behaviour already observed for the transport operator in Piar, Babik, Herbin, et al. 2013: for multi-dimensional problems, the smoothing produced by the Heun time-stepping seems to be necessary to compensate the oscillatory character of the MUSCL scheme (which, for the transport operator, does not lead,

of course, to violate the local maximum principle warranted by construction of the limitation process).

mesh	error(h)	ord(h)	error(u)	ord(u)
32×32	$2.06\,10^{-3}$	/	$2.33\,10^{-1}$	/
64×64	$1.37\,10^{-3}$		$1.18\,10^{-1}$	
128×128	$1.24\,10^{-3}$		$8.50\,10^{-2}$	
256×256	$1.26\,10^{-3}$		$6.16\,10^{-2}$	
512×512	$1.56\,10^{-3}$		$4.85\,10^{-2}$	

Table 1.3 – Measured numerical errors for the travelling vortex with the segregated scheme - Discrete L^1-norm of the difference between the numerical and exact solution at $t = 0.8$, for the height and the velocity.

1.6.2 A Riemann problem

We now turn to a one-dimensional shock solution, corresponding to a Riemann problem posed over $\Omega = (0,1)$. The initial height is $h = 1$ if $x < 0.5$ and $h = 0.2$ otherwise, and the topography z is set to zero over the computational domain; the fluid is initially at rest. The solution consists in a 1-rarefaction wave and a 2-shock.

We plot on Figure 1.4 and Figure 1.5 the results obtained a $t = 0.1$ with the second-order scheme, the segregated scheme and the first-order scheme. The space step is $\delta x = 1/200$ and the time step is chosen as $\delta t = \delta x/10$, which corresponds to a CFL number lower than 0.5 with respect to the waves celerity (the maximal speed of sound is close to 3 and the maximal velocity is close to 2). As expected, the first order scheme

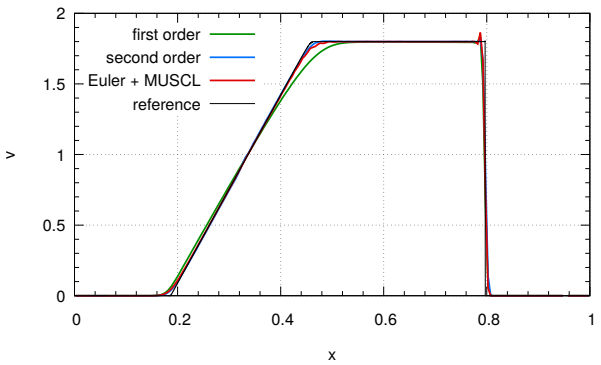


Figure 1.3 – Riemann problem: velocity.

is more diffusive than the other ones. As in the previous test, the segregated forward Euler scheme (with MUSCL fluxes) exhibits some oscillations, which are damped by

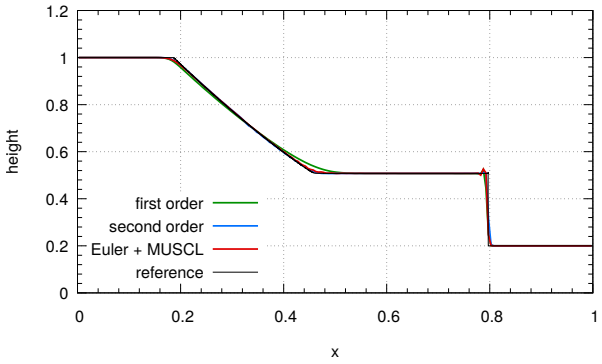


Figure 1.4 – Riemann problem: height

the Heun time discretization (see the Figure 1.5). In this test case, for both the second-order and the segregated scheme, the shock is captured with only one intermediate cell between the left and the right state.

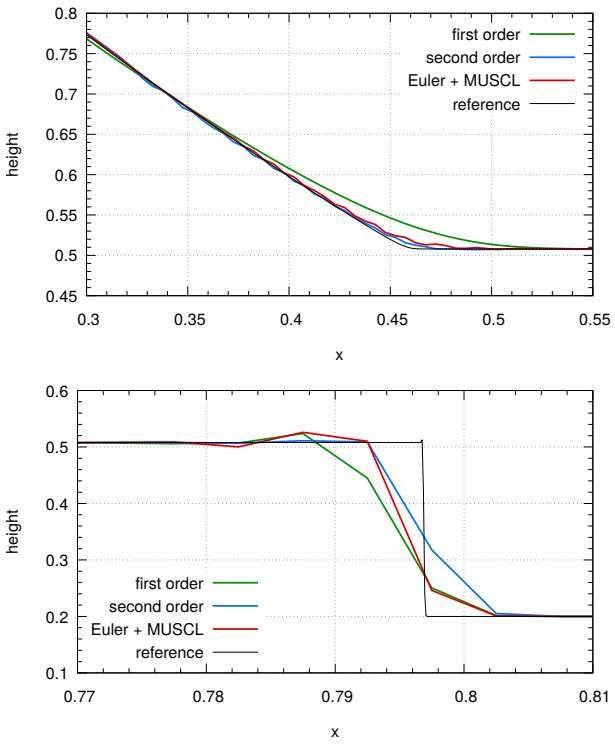


Figure 1.5 – Riemann problem. Details of the flow height.

1.6.3 A circular dam break problem

The objective of this test-case is to check the capability of the scheme to capture a multi-dimensional shock solution. The fluid is initially at rest and the height is given by:

$$h = 2.5 \text{ if } r < 2.5,\ h = 0.5 \text{ otherwise, with } r^2 = x_1^2 + x_2^2.$$

The computational domain is $\Omega = (-20,20) \times (-20,20)$ and the final time is $T = 4.7$.

We plot on Figure 1.6 the results obtained with a 800×800 uniform mesh, with the second-order scheme. The time-step is $\delta t = h_{\mathcal{M}}/10$ (with a maximal velocity in the range of 3.5 and a maximal speed of sound in the range of 5). In addition, to cure some oscillations (see Figure 1.8), we add a slight stabilization in the momentum balance equation which consists in adding to the dicrete momentum equation associated to an edge σ included in a cell K the following flux through a dual edge $\epsilon = D_\sigma | D_{\sigma'}$:

$$F_{\text{stab},\sigma,\epsilon} = \zeta\ h_K\ \text{diam}(K)^{d-1}\ (u_\sigma - u'_\sigma),$$

where ζ is a user-defined parameter. Here, $\zeta = 0.1$, which is significantly lower than the diffusion generated by the use of an upwind scheme in the momentum balance equation; indeed, the upwind scheme may be seen as the centered one complemented by a diffusion taking the same expression as $F_{\text{stab},\sigma,\epsilon}$ with $\zeta\ h_K$ replaced by $|F_{\sigma,\epsilon}|/2$. The interest of this stabilization stems from the fact that the numerical diffusion introduced in the present family of schemes depends on the material velocity (and not on the waves celerity as, for instance, in colocated schemes based on Riemann solvers), and is sometimes too low in the zones where the fluid is almost at rest Herbin, Latché, and Nguyen 2018. Note that, as a counterpart, the scheme does not become overdiffusive for low-Mach number flows. For the same computation, we give on Figure 1.7 the height and the radial velocity along the axis $x_2 = 0$ (*i.e.* the first component of the velocity) at different times.

This computation is also used as "reference computation" on Figure 1.8, where we compare the results obtained at $t = 3T/5$ with a 200×200 mesh with the second-order scheme, the second-order scheme with stabilization and the first-order scheme. This latter is significantly more diffusive, and we observe how the stabilization (even if added to the momentum balance only and not on the mass balance) damps the oscillations obtained with the second-order scheme for both the flow height and the velocity.

1.6.4 A so-called partial dam-break problem

We now turn to a test consisting in a partial dam-break problem with reflection phenomena, and with a non-flat bathymetry. In this test, the computational domain is $\Omega = (0,200) \times (0,200) \setminus \Omega_w$ with $\Omega_w = (95,105) \times (0,95) \cup (95,105) \times (170,200)$. The fluid is supposed to be initially at rest, the initial water height is $h = 10$ for $x_1 \leq 100$ and $h = 5 - 0.04\,(x_1 - 100)$ otherwise, and the bathymetry is $z = 0$ if $x_1 \leq 100$ and

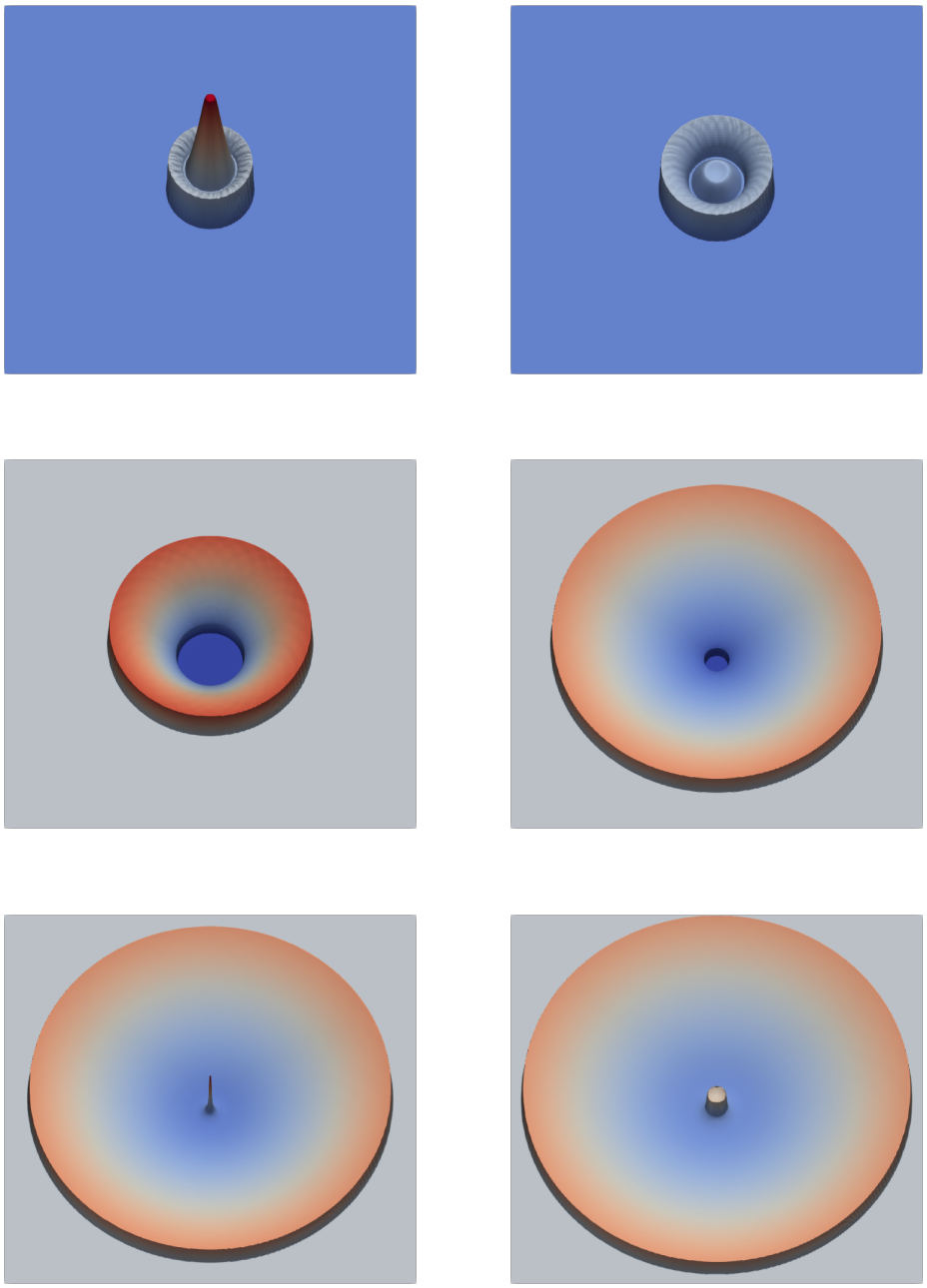


Figure 1.6 – Circular dam-break problem. Height obtained at $t = 0.38$, $t = 0.705$, $t = 1.88$, $t = 3.76$, $t = 4.28$ and $t = T = 4.7$ with the stabilized second-order scheme and a 800×800 mesh. The color range corresponds to the $(0.1, 2.5)$ interval for the first two plots, and to the $(0.1, 1)$ interval for the last four ones.

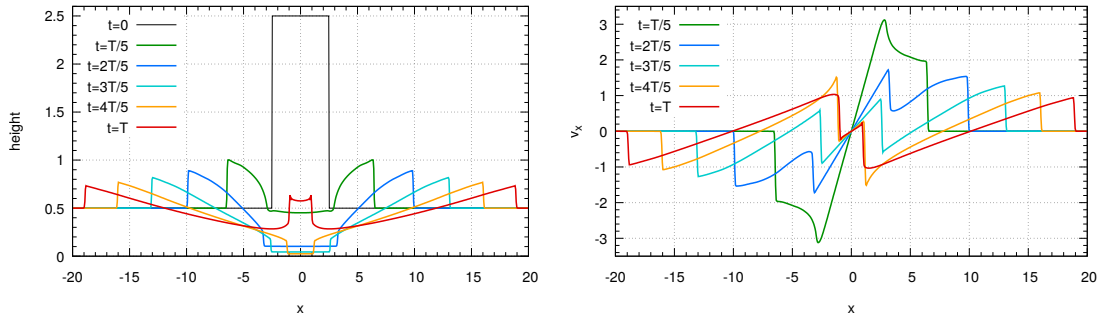


Figure 1.7 – Circular dam-break problem. Height and radial velocity obtained at different times along the line $x_2 = 0$ with the stabilized second-order scheme and a 800×800 mesh.

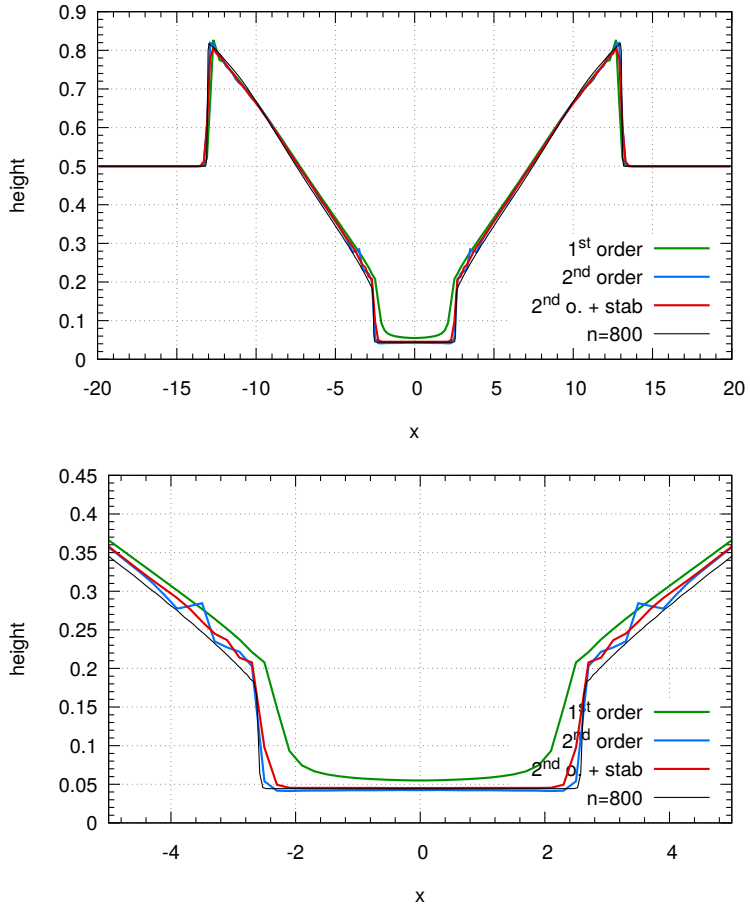


Figure 1.8 – Circular dam-break problem. Height obtained at $t = 3T/5$ with the first-order scheme and the second-order scheme with and without stabilization, with a 200×200 mesh.

$z = 0.04\,(x_1 - 100)$ otherwise. A zero normal velocity is prescribed at all the boundaries of the computational domain. The computation is performed with a mesh obtained from a 1000×1000 regular grid by removing the cells included in Ω_w. The time step is $\delta t = \delta_{\mathcal{M}}/40$ (the maximal speed of sound and the maximal velocity are both close to 10). A stabilization with $\zeta = 0.25$ (so two orders of magnitude lower than the artificial viscosity generated by the upwind scheme in high momentum zones) is added to damp oscillations appearing in the zones at rest, where no numerical diffusion is generated by our schemes. Results obtained at $t = 20$ with the first order in time and space and the present scheme are compared on Figure 1.9. One can observe that the second-order scheme is clearly less diffusive. In addition, these results illustrate the capacity of the staggered scheme to deal with reflection conditions by simply imposing the normal velocity to the boundary at zero.

1.6.5 Uniform circular motion in a paraboloid

We address in this section a classical test which admits a closed-form solution and corresponds to the uniform rotation of a drop of liquid on a paraboloid-shaped support. The solution is very regular (at a given time, the velocity field is constant and $h + z$ is affine outside the dry zones), and the essential interest of this test is to check whether the scheme is able to cope with dry zones, *i.e.* zones where the height is zero (in the continuous setting) or very close to zero, as we shall use numerically. The computational domain is $\Omega = (0, L) \times (0, L)$ and the topography is given by

$$z = -\frac{h_0}{a^2}\Big(a^2 - (x - \frac{L}{2})^2 - (y - \frac{L}{2})^2\Big),$$

with h_0 and a parameters which are given below. The height is:

$$h = \max(0, \bar{h}) \text{ with } \bar{h} = \eta\frac{h_0}{a^2}\Big(2\,(x - \frac{L}{2})\cos(\omega t) + 2(y - \frac{L}{2})\sin(\omega t) - \eta\Big) - z,$$

with η a parameter and ω (the angular rotation velocity of the drop) given by

$$\omega = \frac{(2gh_0)^{1/2}}{a}.$$

Finally, the velocity is

$$\boldsymbol{u} = \eta\,\omega \begin{bmatrix} -\sin(\omega t) \\ \cos(\omega)\, t \end{bmatrix}.$$

The computation is run up to $T = 6\pi/\omega$, so the drop is supposed to perform 3 turns and to lie at the final time at its initial position. The parameters are fixed here to $L = 4$, $h_0 = 0.1$, $a = 1$ and $\eta = 0.5$.

For numerical tests, we bound h from below by 10^{-8}, *i.e.* we set $h = \max(10^{-8}, \bar{h})$, in particular to avoid divisions by zero in the averaging steps of the Heun scheme

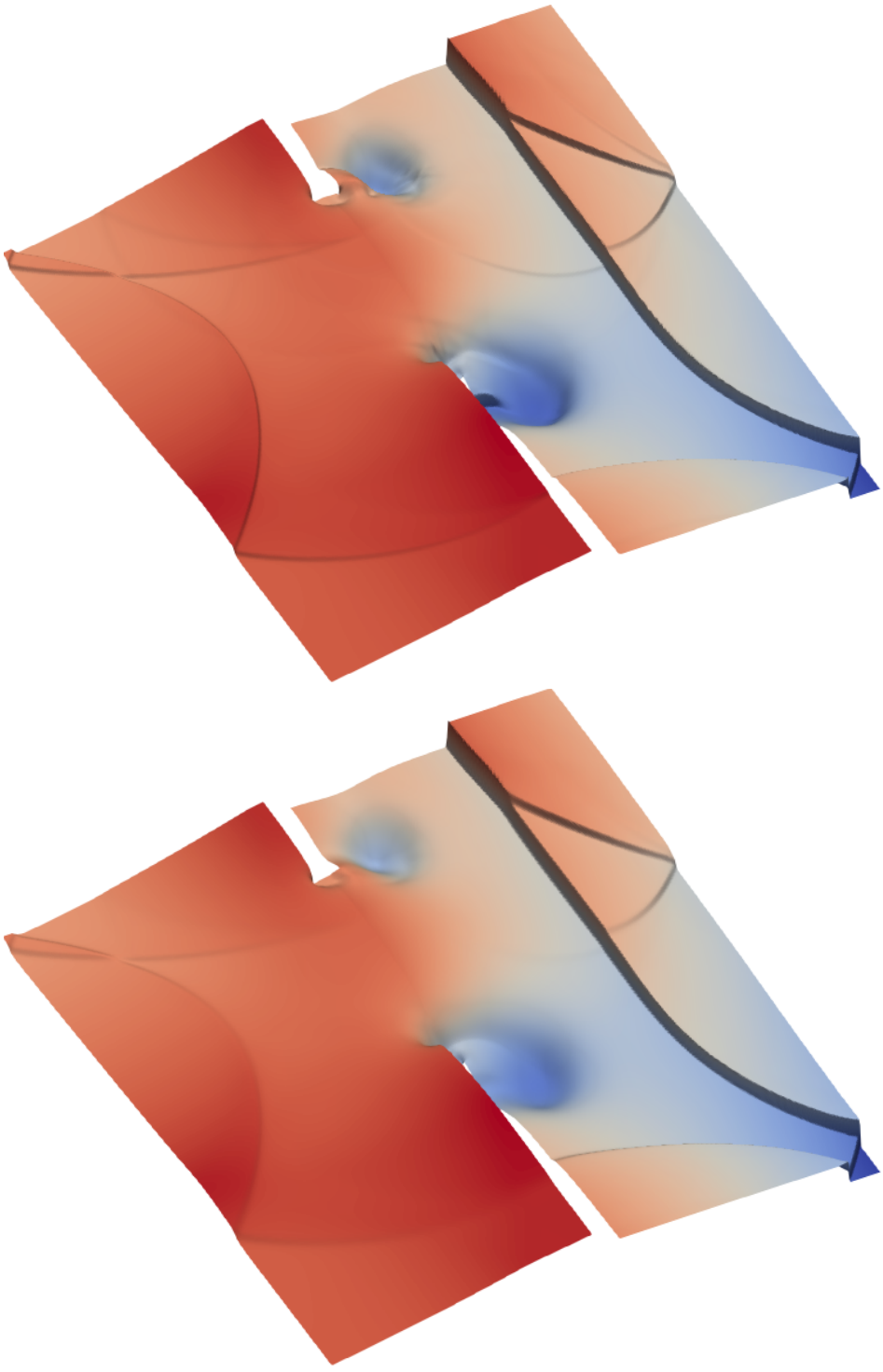


Figure 1.9 – Partial dam-break flow. Top: MUSCL scheme – Bottom: upwind scheme.

(Equations (1.22e) and (1.22f)). The computation are performed with a uniform 100 × 100 mesh, with $\delta t = \delta_{\mathscr{M}}/16$, without changing anything to the numerical fluxes to cope with dry zones. This is clearly dangerous, since a non-upwind approximation of the water height at a face separating two cells with a large ratio of water height may lead to a huge outflow mass flux in view of the cell mass inventory (or, in other words, a very large CFL number). This probably explains the rather small time step used here (the CFL number with respect to the celerity of the fastest waves is in the range of 1/8); the first-order scheme, which uses upwind fluxes, works with time steps four times larger. This problem would be probably cured by a more careful limitation of the mass fluxes outward an almost dry cell.

Results obtained with the first order, the segregated and the second order scheme at $t = 6\pi/\omega$ are plotted on Figure 1.11. All schemes give good results, which, for the first-order scheme, is probably due to the regularity of the solution. For the momentum, one observes that the second-order scheme is less accurate than the other ones; this seems to be due to the time-stepping procedure, which perhaps generates some diffusion at the interface between dry and wet zones, especially in the last averaging step, since the segregated scheme is the most accurate one (and superimposed to the exact solution on Figure 1.11).

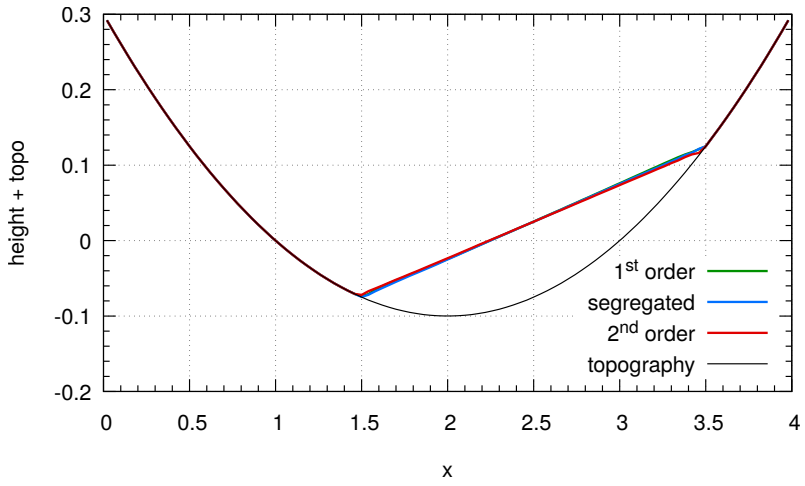


Figure 1.10 – Circular motion of a drop over a paraboloid-shaped topography. Sum of the height and the topography along the $y = L/2$ line at $t = 6\pi$.

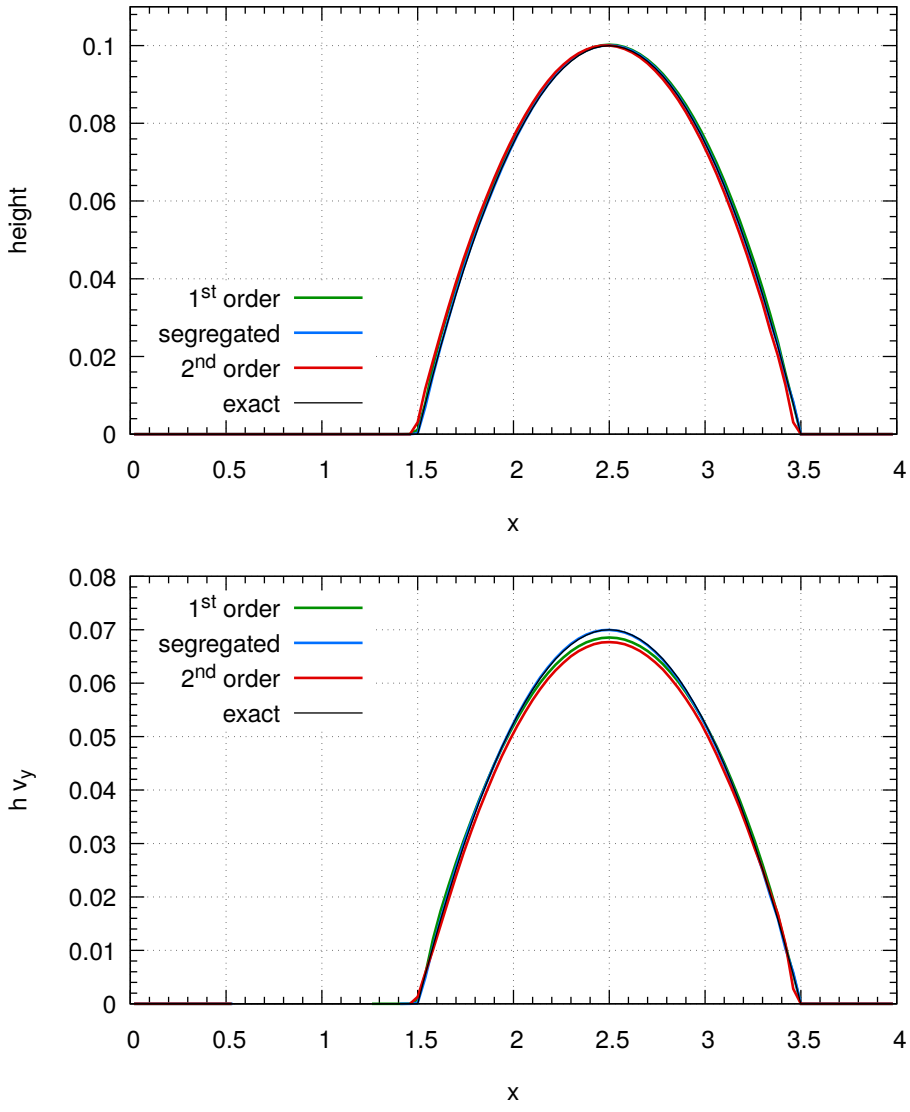


Figure 1.11 – Circular motion of a drop over a paraboloid-shaped topography. Height and second component of the momentum along the $y = L/2$ line at $t = 6\pi$.

Appendix

1.A Consistency results of numerical non linear convection fluxes on staggered meshes

We give here some general lemmas which generalise the Lax-Wendroff theorem to multidimensional staggered meshes, and which we state for any space dimension $d = 1, 2$ or 3. The well-known Lax-Wendroff theorem Lax and Wendroff 1960 states that, on uniform 1D grids, a flux-consistent and conservative cell-centered finite-volume scheme for a system of conservation laws is weakly consistent, in the sense that the limit of any a.e. convergent sequence of L^∞-bounded numerical solutions, obtained with a sequence of grids with mesh and time steps tending to zero, is a weak solution of the conservation law; it is also stated in a different form Leveque 2002, Section 12.10, with a BV bound assumption on the scheme It is generalised to non uniform 1D or Cartesian meshes in Eymard, Gallouët, and Herbin 2000, Theorem 21.2. In a recent work Ben-Artzi and J. Li 2019, the Lax-Wendroff theorem is extended to obtain some error estimates for higher order schemes on uniform 1D meshes. The case of general (and, in particular, unstructured) discretizations has been also been tackled over the past decades: Kroner, Rokyta, and Wierse 1996, Godlewski and P.-A. Raviart 1996, Section 4.2.2 Elling 2007, Gallouët, Herbin, and Latché 2019. In this latter work, the quasi-uniformity assumption that is required in Elling 2007 is relaxed, but while in Elling 2007 the flux is only required to be continuous, it is supposed to be Lipschitz continuous or at least “lip-diag". In all these works, the scheme is supposed to be colocated, in the sense that the discrete unknowns are associated to the cells of the mesh; these results may not be used directly on staggered meshes, and for instance, in Herbin, Latché, Minjeaud, et al. 2020, the consistency of an explicit staggered scheme for the full compressible Euler equations is proven recovering the kinetic energy inequality on the primal mesh.

The consistency result that we give here is valid for general polygonal or polyhedral grids with a colocated or staggered arrangement of the unknowns. The main new idea is that in the proof of consistency, rather than using a convergence result for the discrete gradient, which is only weak and demands some regularity on the mesh, we use the actual mean value of the gradient of the test function on each cell, which converges strongly to the gradient, and does not require any regularity of the mesh. As in Gallouët, Herbin, and Latché 2019, the proof also relies on the control of some residual terms, involving the difference between the numerical solution and a space or time translate of this latter, and we use the estimate on the translates given Gallouët,

Herbin, and Latché 2019, Lemma 4.2 to this purpose, which we recall in the appendix 1.B for the sake of completeness.

Let us suppose that:

$$\Omega \subset \mathbb{R}^d,\ d = 1,2,3,\ T \in (0,+\infty), \tag{1.75a}$$

$$p \in \mathbb{N}^*,\ \beta \in C^1(\mathbb{R}^p,\mathbb{R}),\ \boldsymbol{f} \in C^1(\mathbb{R}^p,\mathbb{R}^d),\ U \in L^\infty(\Omega \times (0,T),\mathbb{R}^p), \tag{1.75b}$$

and consider the conservative convection operator defined (in the distributional sense) by:

$$\begin{aligned} \mathscr{C}(U):\quad & \Omega \times (0,T) \to \mathbb{R}, \\ & (\boldsymbol{x},t) \mapsto \partial_t(\beta(U))(\boldsymbol{x},t) + \operatorname{div}(\boldsymbol{f}(U))(\boldsymbol{x},t). \end{aligned} \tag{1.76}$$

Lemma 1.8 (Weak consistency for a multi-dimensional conservative convection operator)**.** *Under the assumptions* (1.75)*, let* $(U^{(m)})_{m\in\mathbb{N}} \subset L^\infty(\Omega \times (0,T),\mathbb{R}^p)$ *be a sequence of functions such that:*

$$\exists\, C^u \in \mathbb{R}_+^*\ :\ \|U^{(m)}\|_\infty \le C^u\ \forall\, m \in \mathbb{N}, \tag{1.77}$$

$$\exists\, \bar{U} \in L^\infty(\Omega \times (0,T),\mathbb{R}^p) : \|U^{(m)} - \bar{U}\|_{L^1(\Omega\times(0,T),\mathbb{R}^p)} \to 0 \text{ as } m \to +\infty. \tag{1.78}$$

Let $(\mathscr{P}_m)_{m\in\mathbb{N}}$ *be a sequence of polygonal or polyhedral conforming mesh of* Ω *such that*

$$\delta(\mathscr{P}_m) = \max_{P\in\mathscr{P}_m} \operatorname{diam}(P) \to 0 \text{ as } m \to +\infty.$$

Let $\mathfrak{F}^{(m)}$ *denote the set of faces (or edges) of the mesh, and for a given polyhedron (or polygon)* $P \in \mathscr{P}^{(m)}$*, let* $\mathfrak{F}^{(m)}(P)$ *be the set of faces (or edges) of* P*. For* $m \in \mathbb{N}$*, let* $t_0^{(m)} = 0 < t_1^{(m)} < \ldots < t_{N_m}^{(m)} = T$ *be a discretization of* $(0,T)$ *with* $\delta t^{(m)} = t_{k+1}^{(m)} - t_k^{(m)} \to 0$ *as* $m \to +\infty$*, and consider the discrete convection operator*

$$\begin{aligned} \mathscr{C}^{(m)}(U^{(m)}):\quad & \Omega \times (0,T) \to \mathbb{R}, \\ & (\boldsymbol{x},t) \mapsto \eth_t(\beta^{(m)})_P^n + \frac{1}{|P|} \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| (\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta} \text{ for } \boldsymbol{x} \in P \text{ and } t \in (t_n, t_{n+1}) \end{aligned} \tag{1.79}$$

with $\eth_t(\beta^{(m)})_P^n = \frac{1}{\delta t}((\beta^{(m)})_P^{n+1} - (\beta^{(m)})_P^n)$ *and where the families* $\{(\beta^{(m)})_P^n, P \in \mathscr{P}^{(m)},\ n \in [\![0, N_m - 1]\!]\}$ *of real numbers and* $\{(\boldsymbol{F}^{(m)})_\zeta^n, \zeta \in \mathfrak{F}^{(m)},\ n \in [\![0, N_m - 1]\!]\}$ *of real vectors are*

such that

$$\sum_{P\in\mathscr{P}^{(m)}}\int_P |(\beta^{(m)})_P^0-\beta(U_0(\boldsymbol{x}))|d\boldsymbol{x}\to 0 \text{ as } m\to+\infty, \text{ with } U_0\in L^\infty(\Omega,\mathbb{R}^p), \tag{1.80}$$

$$\sum_{n=0}^{N_m-1}\sum_{P\in\mathscr{P}^{(m)}}\int_{t_n}^{t_{n+1}}\int_P |(\beta^{(m)})_P^n-\beta(U^{(m)}(\boldsymbol{x},t))|d\boldsymbol{x}\,dt\to 0 \text{ as } m\to+\infty, \tag{1.81}$$

$$\sum_{n=0}^{N_m-1}\sum_{P\in\mathscr{P}^{(m)}}\int_{t_n}^{t_{n+1}}\frac{\mathrm{diam}(P)}{|P|}\int_P\sum_{\zeta\in\mathfrak{F}^{(m)}}|\zeta|\left|\left((\boldsymbol{F}^{(m)})_\zeta^n-\boldsymbol{f}(U^m(\boldsymbol{x},t))\right)\cdot\boldsymbol{n}_{P,\zeta}\right|d\boldsymbol{x}\,dt\to 0 \text{ as } m\to+\infty. \tag{1.82}$$

Let $\varphi\in C_c^\infty(\Omega\times[0,t))$, then

$$\int_0^T\int_\Omega \mathscr{C}^{(m)}(U^{(m)})(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)\,d\boldsymbol{x}\,dt\to-\int_\Omega\beta(U_0(\boldsymbol{x}))\varphi(\boldsymbol{x},0)\,d\boldsymbol{x}$$
$$-\int_0^T\int_\Omega\Big(\beta(\bar U)(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t)+\boldsymbol{f}(\bar U)(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)\Big)d\boldsymbol{x}\,dt \text{ as } m\to+\infty. \tag{1.83}$$

Proof. The result of this lemma is the consequence of the two following lemmas, which prove respectively the convergence of the time derivative part and the space derivative part. Indeed, let us decompose

$$\int_0^T\int_\Omega \mathscr{C}^{(m)}(U^{(m)})(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)\,d\boldsymbol{x}\,dt=X_1^{(m)}+X_2^{(m)}, \text{ with} \tag{1.84}$$

$$X_1^{(m)}=\sum_{n=0}^{N_m-1}\delta t\sum_{P\in\mathscr{P}^{(m)}}|P|\eth_t^n\beta_P^{(m)}\,\varphi_P^n \tag{1.85}$$

$$X_2^{(m)}=\sum_{n=0}^{N_m-1}\delta t\sum_{P\in\mathscr{P}^{(m)}}\sum_{\zeta\in\mathfrak{F}^{(m)}(P)}|\zeta|(\boldsymbol{F}^{(m)})_\zeta^n\cdot\boldsymbol{n}_{P,\zeta}\varphi_P^n \tag{1.86}$$

where φ_P^n denotes the mean value of φ on $P\times(t_n,t_{n+1})$. Then, by Lemma 1.9 below,

$$X_1^{(m)}\to-\int_\Omega\beta(U_0(\boldsymbol{x}))\,d\boldsymbol{x}-\int_0^T\int_\Omega\beta(\bar U)(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t)d\boldsymbol{x}\,dt \text{ as } m\to+\infty,$$

and by Lemma 1.10 below,

$$X_2^{(m)}\to-\int_0^T\int_\Omega\boldsymbol{f}(\bar U)(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)d\boldsymbol{x}\,dt \text{ as } m\to+\infty,$$

which concludes the proof □

Lemma 1.9 (Weak consistency, time derivative)**.** *Under the assumptions and notations*

of Lemma 1.8,

$$\int_0^T \int_\Omega \eth_t(\beta^{(m)})_P^n \varphi(\boldsymbol{x},t)\, d\boldsymbol{x}\, dt \to -\int_\Omega \beta(U_0(\boldsymbol{x}))\varphi(\boldsymbol{x},0)\, d\boldsymbol{x} - \int_0^T \int_\Omega \beta(\bar{U})(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t) d\boldsymbol{x}\, dt \text{ as } m \to +\infty.$$

Proof. By definition of $\eth_t^n \beta_P^{(m)}(\boldsymbol{x},t)$ and thanks to a discrete integration by parts,

$$\begin{aligned} X_1^{(m)} &= \int_0^T \int_\Omega \eth_t(\beta^{(m)})_P^n \varphi(\boldsymbol{x},t)\, d\boldsymbol{x}\, dt \\ &= -\sum_{P\in\mathscr{P}^{(m)}} |P|\,(\beta^{(m)})_P^0 \varphi_P^0 - \sum_{n=1}^{N_m} \delta t \sum_{P\in\mathscr{P}^{(m)}} |P|\beta_P^{(m)}(\boldsymbol{x},t)\frac{1}{\delta t}\Big(\varphi_P^n - \varphi_P^{n-1}\Big). \end{aligned}$$

Thanks to the assumptions (1.77), (1.78) (1.80) and (1.81), we get that

$$\lim_{m\to+\infty} X_1^{(m)} = -\int_\Omega \beta(U_0)(\boldsymbol{x})\varphi(\boldsymbol{x},0)\, d\boldsymbol{x} - \int_0^T \int_\Omega \Big(\beta(\bar{U})(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t)\, dt\, d\boldsymbol{x}. \tag{1.87}$$

□

Lemma 1.10 (Weak consistency, space derivative)**.** *Under the assumptions and notations of Lemma 1.8,*

$$\int_0^T \int_\Omega \frac{1}{|P|} \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta|(\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta}\varphi(\boldsymbol{x},t)\, d\boldsymbol{x}\, dt \to -\int_0^T \int_\Omega \boldsymbol{f}(\bar{U})(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t) d\boldsymbol{x}\, dt \text{ as } m \to +\infty.$$

Proof. Let $X_2^{(m)}$ denote the left-hand-side of the above assertion. Since for a face ζ separating P and P', one has $\boldsymbol{n}_{P,\zeta} = -\boldsymbol{n}_{P',\zeta}$, we may rewrite $X_2^{(m)}$ as

$$\begin{aligned} X_2^{(m)} &= \int_0^T \int_\Omega \frac{1}{|P|} \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta|(\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta}\varphi(\boldsymbol{x},t)\, d\boldsymbol{x}\, dt \\ &= \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{P\in\mathscr{P}^{(m)}} A_P^n \text{ with } A_P^n = \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta|(\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta}\Big(\varphi_P^n - \varphi_\zeta^n\Big), \end{aligned}$$

where φ_P^n (resp. φ_ζ^n) denotes the mean value of φ over $P\times(t_n,t_{n+1})$ (resp. $\zeta\times(t_n,t_{n+1})$).

Now for any $\boldsymbol{x} \in P$, $t \in [t_n, t_{n+1})$, we can decompose A_P^n as

$$A_P^n = B_P^n(\boldsymbol{x},t) + R_P^n(\boldsymbol{x},t), \text{ with } B_P^n(\boldsymbol{x}) = \sum_{\zeta \in \mathfrak{F}^{(m)}(P)} |\zeta| \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t)) \cdot \boldsymbol{n}_{P,\zeta} \left(\varphi_P^n - \varphi_\zeta^n\right), \text{ and}$$

$$R_P^n(\boldsymbol{x},t) = \sum_{\zeta \in \mathfrak{F}^{(m)}(P)} |\zeta| (\boldsymbol{F}_\zeta^n - \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t))) \cdot \boldsymbol{n}_{P,\zeta} \left(\varphi_P^n - \varphi_\zeta^n\right) d\boldsymbol{x}.$$

Since $\sum_{\zeta \in \mathfrak{F}^{(m)}(P)} |\zeta| \boldsymbol{n}_{P,\zeta} = 0$, we have

$$B_P^n(\boldsymbol{x},t) = -\sum_{\zeta \in \mathfrak{F}^{(m)}(P)} |\zeta| \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t) \cdot \boldsymbol{n}_{P,\zeta} \varphi_\zeta^n = -|P| \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t) \cdot (\nabla\varphi)_P^n, \tag{1.88}$$

$$\text{with } (\nabla\varphi)_P^n = \frac{1}{|P|} \sum_{\zeta \in \mathfrak{F}^{(m)}(P)} |\zeta| \varphi_\zeta^n \boldsymbol{n}_{P,\zeta} = \frac{1}{|P|} \nabla\varphi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}.$$

Note that the piecewise function $\nabla^{(m)}\varphi : \Omega \times (0,T) \to \mathbb{R}^d$ defined by $\nabla^{(m)}\varphi(\boldsymbol{x},t) = (\nabla\varphi)_P^n$ for $(\boldsymbol{x},t) \in\in P \times (t_n, t_{n+1})$ converges uniformly to $\nabla\varphi$ in $L^\infty(\Omega \times (0,T))^d$. Integrating (1.88) over $\boldsymbol{x} \in P$,

$$\int_P B_P^n(\boldsymbol{x},t) d\boldsymbol{x} = |P| \int_P \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t)) \cdot \nabla^{(m)}\varphi(\boldsymbol{x},t)\, d\boldsymbol{x},$$

Since $A_P^n = \dfrac{1}{\delta t^{(m)}|P|} \Big(\displaystyle\int_{t_n}^{t_{n+1}} \int_P B_P^n(\boldsymbol{x},t) d\boldsymbol{x} dt + \int_{t_n}^{t_{n+1}} \int_P R_P^n(\boldsymbol{x},t) d\boldsymbol{x} dt \Big)$, we get

$$X_2^{(m)} = \sum_{n=0}^{N_m-1} \sum_{P \in \mathscr{P}^{(m)}} \Big(\int_{t_n}^{t_{n+1}} \int_P B_P^n(\boldsymbol{x},t) d\boldsymbol{x} dt + \int_{t_n}^{t_{n+1}} \frac{1}{|P|} \int_P R_P^n(\boldsymbol{x},t) d\boldsymbol{x} dt \Big)$$

$$= -\int_0^T \int_\Omega \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t)) \nabla^{(m)}\varphi(\boldsymbol{x},t)\, d\boldsymbol{x} dt + \sum_{n=0}^{N_m-1} \sum_{P \in \mathscr{P}^{(m)}} \int_{t_n}^{t_{n+1}} \frac{1}{|P|} \int_P R_P^n(\boldsymbol{x},t) d\boldsymbol{x} dt.$$

Owing to the boundedness and convergence assumptions on $U^{(m)}$ and to the uniform convergence of $\nabla^{(m)}\varphi$ to $\nabla\varphi$, the first term tends to

$$-\int_0^T \int_\Omega \boldsymbol{f}(U(\boldsymbol{x},t)) \nabla\varphi(\boldsymbol{x},t)\, d\boldsymbol{x} dt \text{ as } m \to +\infty.$$

Since $|\varphi_\zeta^n - \varphi_P^n| \le C_\varphi \mathrm{diam}(P)$, with C_φ depending only on φ, the second term tends to 0 thanks to the assumption (1.82). Therefore

$$\lim_{m \to +\infty} X_2^{(m)} \to -\int_0^T \int_\Omega \boldsymbol{f}(U(\boldsymbol{x},t)) \cdot \nabla\varphi(\boldsymbol{x},t)\, d\boldsymbol{x} dt. \tag{1.89}$$

□

1.B Former lemmas

1.B.1 A result on a finite volume convection operator

We begin with a property of the convection operator $\mathscr{C} : \rho \mapsto \partial_t(\rho) + \mathrm{div}(\rho \boldsymbol{u})$; at the continuous level, this property may be formally obtained as follows (see Herbin, Latché, and Nguyen 2018 for the detailed derivation). Let ψ be a regular function from $(0, +\infty)$ to $\mathbb{R}$; then:

$$\psi'(\rho)\,\mathscr{C}(\rho) = \partial_t\big(\psi(\rho)\big) + \mathrm{div}\big(\psi(\rho)\boldsymbol{u}\big) + \big(\rho\psi'(\rho) - \psi(\rho)\big)\,\mathrm{div}\boldsymbol{u}. \tag{1.90}$$

This computation is of course completely formal and only valid for regular functions ρ and $\boldsymbol{u}$. The following lemma states a discrete analogue to (1.90), and its proof follows the formal computation which we just described.

Lemma 1.11. *[On the discrete convection operator, Herbin, Latché, and Nguyen 2013, Lemma A1] Let P be a polygonal (resp. polyhedral) bounded set of $\mathbb{R}^2$ (resp. $\mathbb{R}^3$), and let $\mathscr{E}(P)$ be the set of its edges (resp. faces). Let ψ be a twice continuously differentiable function defined over $(0,+\infty)$. Let $\rho_P^* > 0$, $\rho_P > 0$, $\delta t > 0$; consider three families $(\rho_\eta^*)_{\eta\in\mathscr{E}(P)} \subset \mathbb{R}_+ \setminus \{0\}$, $(V_\eta^*)_{\eta\in\mathscr{E}(P)} \subset \mathbb{R}$ and $(F_\eta^*)_{\eta\in\mathscr{E}(P)} \subset \mathbb{R}$ such that*

$$\forall \eta \in \mathscr{E}(P), \qquad F_\eta^* = \rho_\eta^*\, V_\eta^*.$$

Let $R_{P,\delta t}$ be defined by:

$$\begin{aligned} R_{P,\delta t} = {} & \Big[\frac{|P|}{\delta t}(\rho_P - \rho_P^*) + \sum_{\eta\in\mathscr{E}(P)} F_\eta^*\Big]\,\psi'(\rho_P) \\ & - \Big[\frac{|P|}{\delta t}\,[\psi(\rho_P) - \psi(\rho_P^*)] + \sum_{\eta\in\mathscr{E}(P)} \psi(\rho_\eta^*)V_\eta^* + [\rho_P^*\psi'(\rho_P^*) - \psi(\rho_P^*)] \sum_{\eta\in\mathscr{E}(P)} V_\eta^*\Big]. \end{aligned}$$

Then this quantity may be expressed as follows:

$$\begin{aligned} R_{P,\delta t} = {} & \frac{1}{2}\frac{|P|}{\delta t}(\rho_P - \rho_P^*)^2\,\psi''(\overline{\rho}_P^{(1)}) - \frac{1}{2}\sum_{\eta\in\mathscr{E}(P)} V_\eta^*\,(\rho_P^* - \rho_\eta^*)^2\,\psi''(\overline{\rho}_\eta^*) \\ & + \sum_{\eta\in\mathscr{E}(P)} V_\eta^*\,\rho_\eta^*\,(\rho_P - \rho_P^*)\,\psi''(\overline{\rho}_P^{(2)}), \end{aligned}$$

where $\overline{\rho}_P^{(1)}$, $\overline{\rho}_P^{(2)} \in [\![\rho_P, \rho_P^*]\!]$ and $\forall \eta \in \mathscr{E}(P)$, $\overline{\rho}_\eta^* \in [\![\rho_P^*, \rho_\eta^*]\!]$. We recall that, for $a,\ b \in \mathbb{R}$, we denote by $[\![a,b]\!]$ the interval $[\![a,b]\!] = \{\theta a + (1-\theta)b,\ \theta \in [0,1]\}$.

1.B.2 A result on the space translates

Lemma 1.12 (Convergence of the space translates Gallouët, Herbin, and Latché 2019, Lemma 4.2)**.** *For a given mesh $\mathcal{M}$, let*

$$\theta_{\mathcal{M}} = \max_{K\in\mathcal{M}} \max_{\sigma\in\mathscr{E}_K} \frac{|D_\sigma|}{|K|}.$$

Let $\theta > 0$ and $(\mathcal{M}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes such that $\theta_{\mathcal{M}^{(m)}} \leq \theta$ for all $m \in \mathbb{N}$ and $\lim_{m\to+\infty} h_{\mathcal{M}^{(m)}} = 0$. We suppose that the number of faces of a cell $K \in \mathcal{M}^{(m)}$ is bounded by $\mathcal{N}_{\mathscr{E}}$, for any $m \in \mathbb{N}$. Let $\psi \in L^1(\Omega)$, let $\langle\psi\rangle_K$ denote the mean value of ψ on a cell K. Then,

$$\lim_{m\to+\infty} \sum_{\substack{\sigma\in\mathscr{E}_{\mathrm{int}} \\ \sigma=K|L}} |D_\sigma|\,|\langle\psi\rangle_K - \langle\psi\rangle_L| = 0. \tag{1.91}$$

2 Stabilized staggered schemes for the $2D$ shallow water equations with Coriolis source term on rectangular grid.

Sommaire

Abstract. We propose herein a class of staggered schemes designed for the shallow water equations with Coriolis source term based on a rectangular grid. A semi-implicit time discretization MAC scheme is first presented using an upwind scheme for the convection terms of the continuity and momentum equations. Then an extension of this scheme to a staggered scheme working on the Rannacher-Turek finite elements is thus proposed. This one enjoys some important numerical features, the positivity of the water height is ensured and the (linear) geostrophic equilibrium state is preserved as well as the "lake at rest" steady state. To improve the accuracy and reduce numerical diffusion produced by both upwind schemes, we follow a stabilization procedure consisting in introducing two consistent correction terms with the geostrophic equilibrium state, one for the mass flux and the other one for the discrete pressure gradient. This process is performed and analyzed for a staggered scheme based on the Rannacher-Turek elements for which the numerical mass flux is approximated by a sort of weighted average. The obtained stabilized non-linear and linear schemes satisfy a dissipation of the semi-discrete mechanical energy associated to them. Furthermore the corresponding fully discrete scheme for the non-linear equations is positivity-preserving under a local CFL like restriction and the geostrophic equilibrium state is perfectly preserved. Numerical simulations are presented to assess the stability and accuracy of the schemes when compared to a Godunov type scheme for some benchmark tests.

Keywords Shallow water equations, MAC discretization, Rannacher-Turek finite elements, Coriolis force, stabilization method.

2.1 Introduction

In this chapter, we build some numerical schemes for the shallow water equations with Coriolis (SWC) source term based principally on a staggered spatial discretization. The main objective we focus herein is to construct semi-discrete staggered schemes according to the space discretization that are stable with respect to a dissipation of the semi-discrete mechanical energy. Furthermore the corresponding fully discrete schemes for the non-linear SWC equations are expected to be robust with respect to the positivity-preserving of the water height. Last but not least the linearized schemes are expected also to be linearly well-balanced with respect to the preservation of the geostrophic equilibrium steady state which is the well known steady state of the linearized equations of SWC (see Audusse, Do, Omnes, et al. 2018).

Investigations of the SWC are realized on several times and recently advances are available in the literature essentially in the finite volume methods, using either a collocated (see Audusse, Klein, and Owinoh 2009, Audusse, Dellacherie, Do, et al. 2017, Audusse, Do, Omnes, et al. 2018, Beljadid, Mohammadian, and Qiblawey 2013, Bouchut, Lesommer, and Zeitlin 2014, Mousseau, Knoll, and Reisner 2002) or a staggered (see also Bonaventura and Ringler 2005, Ringler, Thuburn, Klemp, et al. 2010, Thuburn, Ringler, Skamarock, et al. 2009) discretization of principally unknowns, water height and the velocity field. Regarding the staggered discretization, the formally MAC approach is not straightforward to adopt and supplementary techniques are needed; for instance in the work of Gunawan 2015 a splitting-like technique is proposed. The problem is lying on the Coriolis term since its involves an exchange between the components of the velocity. To tackle this overcome a suitable reconstruction of the Coriolis force is performed which is consistent with the continuous source term. However with this manner the dissipation of the semi-discrete mechanical energy of the resulting scheme is probably lost whether for the non-linear or linear cases. An other issue to handle the constraints underlined above consists to use the Rannacher-Turek (RT) finite elements instead of the MAC discretization. Indeed for the RT finite elements both components of the velocity fields are evaluated on the same diamond cell of the mesh. Thus the main differences of both strategies are observed in the discretization of the momentum equations since the components of the velocity are computed in two different dual cells for the MAC grid. Numerical analysis shows that the RT semi-discrete (non-linear and linear) schemes ensure a dissipation of the semi-discrete mechanical energy. However the RT finite elements lead to some complications since five discrete equations are implemented, one for the discrete mass and four for the momentum, instead of three equations in the MAC setting.

In addition, the time discretization of the Coriolis force should be taken carefully in order to deal with the preservation of the geostrophic equilibrium state which is important for physical reasons and consists of a building block of the methods we are going to study in the sequel. Among the works of Audusse, Do, Omnes, et al. 2018

and the thesis of Gunawan 2015, they propose a θ-scheme method to discretize the Coriolis term since they established that an explicit time stepping is not stable. Here we through the road suggested by Gunawan 2015 consisting in a reformulation of the θ-scheme method. In fact, the interesting feature is the following: if a semi-discrete scheme is well-balanced with respect to the preservation of the equilibrium state then the θ-scheme we consider here allows to preserve this property at the fully discrete level for all θ being in $[0,1]$.

Omitting the source term, some numerical schemes resolving the SW equations ensure the energy dissipation and preserve obviously the lake at rest steady state. We refer to the recently work of Couderc, Duran, and Vila 2017, based on collocated unstructured meshes extended also in a staggered setting in Duran and Vila 2019; Duran, Vila, and Baraille 2017. We refer also to the paper of Parisot and Vila 2016 based on a regularized model where the advection velocity is modified with a pressure gradient in both mass and momentum equations. Recently in Berthon, Duran, Foucher, et al. 2019, an artificial numerical viscosity technique is mixed with the formally hydrostatic reconstruction method (see Audusse, Bouchut, Bristeau, et al. 2004) to get a fully discrete mechanical energy inequality. In all these works, the authors follow a stabilization procedure in order to deal with a dissipation of the discrete mechanical energy. We adopt the technique developed in Couderc, Duran, and Vila 2017; Duran, Vila, and Baraille 2017 using the correction terms introduced in Audusse, Do, Omnes, et al. 2018 for the mass flux and the discrete pressure gradient. In fact the way in which the mass flux correction is introduced allows to stabilize simultaneously the momentum flux since this latter is expressed in terms of the mass flux and the Coriolis force which is proportionally to the perpendicular of the mass flux. Such as corrections are precisely consistent with the geostrophic equilibrium state such that they vanish when we reach this equilibrium state in contrast to standard corrections expressed in terms of the gradient of the water height. The resulting schemes are depending only on two stabilization parameters which are chosen on the basis of existing results and/or numerical experiments, either as numerically or theoretically way by means of linear stability analysis. The linear stability study is left for future works for the sake of simplicity. Furthermore, the positivity-preserving stability is not easily achieved since the principal numerical mass flux is a sort of weighted average which yields a centered flux for an uniform rectangular grid. We show that the aid of additional regularization term in the mass flux allows to obtain a local CFL-like condition to maintain the positivity.

We first address the shallow water equations with Coriolis force posed on an open bounded domain Ω of $\mathbb{R}^2$ and for $T > 0$:

$$\partial_t h + \operatorname{div}(h\boldsymbol{u}) = 0 \qquad \text{in } \Omega \times (0,T), \tag{2.1a}$$

$$\partial_t (h\boldsymbol{u}) + \operatorname{div}(h\boldsymbol{u} \otimes \boldsymbol{u}) + gh\nabla(h+z) = -w\, h\, \boldsymbol{u}^{\perp} \qquad \text{in } \Omega \times (0,T), \tag{2.1b}$$

where t stands for the time, g is the standard gravity constant and z the topography, h

the water height, $\boldsymbol{u}$ the horizontal velocity, for $\boldsymbol{u} = (u, v)^T$, the perpendicular velocity is defined as $\boldsymbol{u}^\perp = (-v, u)^T$ and w the (constant) angular speed. In what follows, we consider a flat bottom i.e. $z = 0$.
A main property of this system is that for any regular solution $(h, \boldsymbol{u})$ of (2.1) where h is non-negative, must satisfy the following balance of the mechanical energy:

$$\partial_t E + \operatorname{div}((gh + \frac{1}{2}|\boldsymbol{u}|^2)h\,\boldsymbol{u}) = 0, \text{ with } E = \frac{1}{2}gh^2 + \frac{1}{2}h|\boldsymbol{u}|^2. \tag{2.2}$$

Furthermore, the shallow water equations (2.1) admit some steady states, solution of the following equations:

$$\begin{cases} \operatorname{div}(h\boldsymbol{u}) = 0, \\ \operatorname{div}(h\boldsymbol{u}\otimes\boldsymbol{u}) + gh\nabla(h+z) = -\omega\, h\,\boldsymbol{u}^\perp. \end{cases} \tag{2.3}$$

In fact the "lake at rest" state is a particular steady state characterized by $\boldsymbol{u} = 0$ and $h = h_0$. There exists other stationary solutions of (2.3) for instance in Audusse, Do, Omnes, et al. 2018; Mousseau, Knoll, and Reisner 2002 stationary vortex solutions are proposed.
Linearizing the system (2.1) around the state h_0 and $\boldsymbol{u}_0 = 0$, leads to the following linear equations:

$$\partial_t h + h_0 \operatorname{div}(\boldsymbol{u}) = 0 \qquad \text{in } \Omega\times(0,T), \tag{2.4a}$$

$$\partial_t \boldsymbol{u} + g\nabla h = -\omega\,\boldsymbol{u}^\perp \qquad \text{in } \Omega\times(0,T). \tag{2.4b}$$

The stationary solutions of these linear equations are characterized by:

$$\begin{cases} \operatorname{div}(\boldsymbol{u}) = 0 \\ g\nabla h + \omega\,\boldsymbol{u}^\perp = 0. \end{cases}$$

Note that the condition $g\nabla h + \omega\,\boldsymbol{u}^\perp = 0$ implies automatically the divergence free condition $\operatorname{div}(\boldsymbol{u}) = 0$, then the above relations boil down to the following geostrophic equilibrium steady state:

$$g\nabla h + \omega\,\boldsymbol{u}^\perp = 0. \tag{2.5}$$

The lake at rest state is a particular case of the geostrophic equilibrium state (2.5).
Furthermore the linear SWC equations satisfy the following mechanical energy balance equation:

$$\partial_t E + \operatorname{div}(gh\,\boldsymbol{u}) = 0, \text{ with } E = \frac{1}{2}\frac{g}{h_0}h^2 + \frac{1}{2}\,|\boldsymbol{u}|^2. \tag{2.6}$$

The remainder of this chapter is composed of three parts as follows: In Section 2.2, we propose and analyze two upwind schemes for the non-linear SWC equations based on a staggered spatial discretizations and thus deduce linear corresponding schemes for linear SWC equations. Then in Section 2.3 stabilization analysis is taken

for the non-linear schemes and thus the linear ones based on the RT finite elements. Finally numerical experiments are presented in Section 5.5 to compare these different schemes against classical Godunov type schemes.

2.2 Staggered schemes for the non-linear and linear equations

Here we propose two segregated fully discrete schemes which involve only explicit time stepping for the homogeneous system of (2.1) and work on a staggered discretization. We present first a MAC scheme and exclusively for this latter a reconstruction of the rotating velocity is performed to cope with the Coriolis force. A second staggered scheme based on the Rannacher-Turek finite elements is then proposed.

2.2.1 Mesh and notations

In order to define the scheme, we have to discretize the space domain Ω and introduce some notations. A discretization $(\mathscr{M},\mathscr{E})$ of Ω with a staggered rectangular grid with respect to the MAC grid or the Rannacher-Turek (RT) finite elements is defined as follows:

– A primal grid $\mathscr{M}$ which consists in a conforming structured partition of Ω in rectangles. A generic cell of this grid is denoted by K of volume $|K|$ and its mass center is denoted by $\boldsymbol{x}_K$.

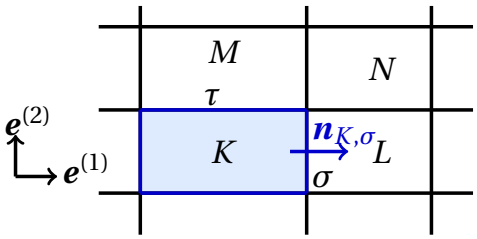


Figure 2.1 – Notations for control volume of the primal mesh, $(\boldsymbol{e}^{(1)},\boldsymbol{e}^{(2)})$ is the canonical basis of $\mathbb{R}^2$, $\boldsymbol{n}_{K,\sigma}$ is the unit vector to σ outward K.

– The set of all edges of the mesh $\mathscr{E}$, with $\mathscr{E}=\mathscr{E}_{\rm int}\cup\mathscr{E}_{\rm ext}$, where $\mathscr{E}_{\rm int}$ (resp. $\mathscr{E}_{\rm ext}$) are the edges of $\mathscr{E}$ that lie in the interior (resp. on the boundary) of the domain. We distinguish also by $\mathscr{E}^{(i)}$ the set of edges perpendicular to the unit vector $\boldsymbol{e}^{(i)}$. For $\sigma\in\mathscr{E}_{\rm int}$, we write $\sigma=K|L$ if $\sigma=\partial K\cap\partial L$. A dual cell D_σ associated to an edge $\sigma\in\mathscr{E}$ is defined as follows:

- **MAC grid** – If $\sigma=K|L\in\mathscr{E}_{\rm int}^{(i)}$, $i=1,2$, then D_σ is the union of the half-part of K and L denoted by $D_{K,\sigma}$ and $D_{L,\sigma}$ respectively adjacent to σ (see Figure 2.2 on the left) in such a way that $|D_\sigma|=|D_{K,\sigma}|+|D_{L,\sigma}|$, otherwise if

$\sigma \in \mathscr{E}_{\text{ext}}$ is adjacent to the cell K, then $D_\sigma = D_{K,\sigma}$. The set of the dual edge ϵ of D_σ associated to $\sigma \in \mathscr{E}^{(i)}$ is denoted by $\widetilde{\mathscr{E}}^{(i)}(D_\sigma)$;

- **RT finite elements** – If $\sigma = K|L \in \mathscr{E}_{\text{int}}$, then D_σ is the union of two diamond cells denoted also by $D_{K,\sigma}$ and $D_{L,\sigma}$ adjacent to σ (see Figure 2.2 on the right) with $|D_{K,\sigma}| = \frac{1}{4}|K|$ (resp $|D_{L,\sigma}| = \frac{1}{4}|L|$) such that $|D_\sigma| = |D_{K,\sigma}| + |D_{L,\sigma}|$. Otherwise if $\sigma \in \mathscr{E}_{\text{ext}}$ then $D_\sigma = D_{K,\sigma}$. The set of the dual edge ϵ of D_σ associated to $\sigma \in \mathscr{E}$ is denoted by $\widetilde{\mathscr{E}}(D_\sigma)$.

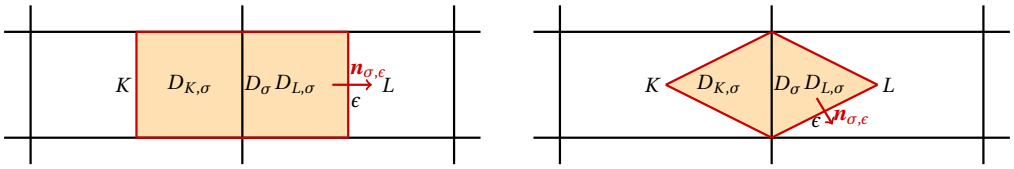


Figure 2.2 – Notations for control volumes of the dual mesh - left: MAC dual mesh, right: RT dual mesh associated to a vertical internal edge, $\boldsymbol{n}_{\sigma,\epsilon}$ is the unit vector to ϵ outward D_σ.

For both cases we shall denote by $\epsilon = \sigma|\sigma'$ the dual edged ϵ separating two dual or diamond cells D_σ and $D_{\sigma'}$.

The water height h is associated to the mesh $\mathscr{M}$ and its discrete equivalent h_K is defined by:

$$\forall\, K \in \mathscr{M}, \quad h_K = \frac{1}{|K|}\int_K h\,\mathrm{d}\boldsymbol{x}.$$

MAC grid – The components of the velocity (u_1, u_2) are evaluated separately on a dual cell D_σ as follows:

$$\forall\, \sigma \in \mathscr{E}_{\text{int}}^{(i)}, \quad u_{i,\sigma} = \frac{1}{|D_\sigma|}\int_{D_\sigma} u_i\,\mathrm{d}\boldsymbol{x},\ i = 1,2.$$

RT finite elements – The velocity field is defined on a diamond cell by:

$$\forall\, \sigma \in \mathscr{E}_{\text{int}}, \quad \boldsymbol{u}_\sigma = (u_{1,\sigma}, u_{2,\sigma}) \quad \text{with } u_{i,\sigma} = \frac{1}{|D_\sigma|}\int_{D_\sigma} u_i\,\mathrm{d}\boldsymbol{x},\ i = 1,2.$$

We are now ready to define the schemes starting by the non-linear MAC scheme.

2.2.2 Semi-implicit MAC schemes

On the basis of the discussion mentioned in the Introduction regarding the approximation of the Coriolis force, a carefully discretization is needed to cope with the exchange of the velocity components. In this context, for the discrete momentum equation that is built here-below, we perform a reconstruction of the Coriolis force

which we only require to be consistent with the continuous term $\boldsymbol{u}^{\perp}$. Then we propose a semi-implicit time discretization for the source term involving only a progressive time stepping.

Nonlinear semi-implicit MAC scheme – The fully discrete scheme for the SWC equations (2.1) reads:

$$\frac{1}{\delta t}\,(h_K^{n+1}-h_K^n)+\mathrm{div}_K(h^n\,\boldsymbol{u}^n)=0,\quad \forall\, K\in\mathscr{M}, \tag{2.7a}$$

$$\frac{1}{\delta t}\,(h_{D_\sigma}^{n+1}\,u_{1,\sigma}^{n+1}-h_{D_\sigma}^{n}\,u_{1,\sigma}^{n})+\mathrm{div}_{D_\sigma}(h^n\boldsymbol{u}^n\,u_1^n)+g\,h_{\sigma,c}^{n+1}\,(\eth_1 h^{n+1})_\sigma = \omega\,h_{D_\sigma}^{n+1}\,(u_{2,\sigma}^*)^n,\qquad \forall\,\sigma\in\mathscr{E}_{int}^{(1)}, \tag{2.7b}$$

$$\frac{1}{\delta t}\,(h_{D_\tau}^{n+1}\,u_{2,\tau}^{n+1}-h_{D_\tau}^{n}\,u_{2,\tau}^{n})+\mathrm{div}_{D_\tau}(h^n\boldsymbol{u}^n\,u_2^n)+g\,h_{\tau,c}^{n+1}\,(\eth_2 h^{n+1})_\tau = -\omega\,h_{D_\tau}^{n+1}\,(u_{1,\tau}^*)^{n+1},\qquad \forall\tau\in\mathscr{E}_{int}^{(2)}, \tag{2.7c}$$

where δt is the constant time step and where the definitions of the discrete terms and operators involving in this scheme are:

$$\begin{aligned}
\mathrm{div}_K(h^n\boldsymbol{u}^n) &= \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\,\boldsymbol{F}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma},\quad \text{with } \boldsymbol{F}_\sigma^n=h_\sigma^n\,u_{i,\sigma}^n\boldsymbol{e}^{(i)},\ \text{for}\ \sigma\in\mathscr{E}_{int}^{(i)},\\
h_{D_\sigma}^n &= \frac{1}{|D_\sigma|}(|D_{K,\sigma}|\,h_K^n+|D_{L,\sigma}|\,h_L^n),\quad h_{\sigma,c}^n=\frac{h_K^n+h_L^n}{2},\quad \text{for}\ \sigma=K|L\in\mathscr{E}_{int}^{(i)},\\
(\eth_i h^n)_\sigma &= \frac{|\sigma|}{|D_\sigma|}\,(h_L^n-h_K^n)\,\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{e}^{(i)},\qquad \text{for}\ \sigma=K|L\in\mathscr{E}_{int}^{(i)},\\
\mathrm{div}_{D_\sigma}(h^n\boldsymbol{u}^n\,u_i^n) &= \frac{1}{|D_\sigma|}\sum_{\sigma\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)}|\epsilon|\,u_{i,\epsilon}^n\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon},\qquad \text{for}\ \sigma\in\mathscr{E}_{int}^{(i)}.
\end{aligned}$$

The discrete terms h_σ^n and $u_{i,\epsilon}^n$ are approximated by the upwind scheme in terms of the principal unknowns $(h_K^n)_{K\in\mathscr{M}}$ and $(u_{i,\sigma}^n)_{\sigma\in\mathscr{E}_{\rm int}^{(i)}}$ respectively. The dual fluxes $\boldsymbol{F}_\epsilon^n$ are expressed in terms of the mass fluxes $\boldsymbol{F}_\sigma^n$ as follows:

$$\forall\,\epsilon=\sigma|\sigma':\quad |\epsilon|\boldsymbol{F}_\epsilon^n=\begin{cases}\frac{1}{2}|\sigma|\boldsymbol{F}_\sigma^n+\frac{1}{2}|\sigma'|\boldsymbol{F}_{\sigma'}^n, & \text{first case}\\ \frac{1}{2}|\tau_1|\boldsymbol{F}_{\tau_1}^n+\frac{1}{2}|\tau_2|\boldsymbol{F}_{\tau_2}^n, & \text{second case}\end{cases}$$

It remains to define the terms $u_{1,\tau}^*$ and $u_{2,\sigma}^*$ appearing in the equations (2.7b) and (2.7c). For the reasons outlined in the introduction these terms are defined in such a way to be consistent with their continuous counterpart. For this purpose we perform

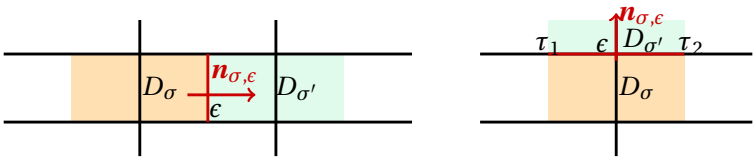


Figure 2.3 – Definitions of the dual fluxes: on the left-first case and on the right- second case

the following reconstruction:

$$\forall i,j\in\{1,2\},\ i\neq j: \quad u^*_{i,\sigma} = \frac{1}{2}(u^*_{i,K}+u^*_{i,L}), \quad \text{for } \sigma = K|L\in\mathscr{E}^{(j)}_{\text{int}}, \tag{2.8}$$
$$\text{with} \quad u^*_{i,K} = \frac{1}{2}\sum_{\tau\in\mathscr{E}^{(i)}(K)} u_{i,\tau}, \quad \text{for } K\in\mathscr{M}.$$

By construction this definition of $u^*_{i,\sigma}$ is consistent with u_i as expected which closes the definition of the MAC scheme (2.7).

Now, we recall briefly some discrete stability properties satisfied by this non-linear MAC scheme. Among them we have the robustness behavior with respect to the preservation of the water height positivity (see Proposition 2.1). More precisely we have the following result: if $h^n_K>0$ for all $K\in\mathscr{M}$ then $h^{n+1}_K>0$ under the following time step condition:

$$\Big(\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\,|\boldsymbol{u}^n_\sigma\cdot\boldsymbol{n}_{K,\sigma}|\Big)\delta t\le|K|$$

Furthermore this scheme holds the preservation (we refer also to the Proposition 2.1 below) of the "lake at rest state" which is characterized as follows:

$$\text{if}\begin{cases}u^n_{i,\sigma}=0, & \forall\sigma\in\mathscr{E}^{(i)},\ i=1,2\\ h^n_K=C, & \forall K\in\mathscr{M},\ \text{with } C>0\end{cases}\qquad\text{then}\begin{cases}u^{n+1}_{i,\sigma}=0, & \forall\sigma\in\mathscr{E}^{(i)},\ i=1,2\\ h^{n+1}_K=C, & \forall K\in\mathscr{M}\end{cases}$$

More important, a discrete mechanical energy may be derived from this nonlinear MAC scheme, since following Herbin, Latché, Nasseri, et al. 2019 we obtain the following discrete local energy balance:

$$\frac{|K|}{\delta t}\left[(E_k)^{n+1}_K+(E_p)^{n+1}_K-(E_k)^n_K-(E_p)^n_K\right]+\sum_{\sigma\in\mathscr{E}(K)}\left[G^n_{K,\sigma}+\frac{1}{2}|\sigma|\,g h^n_\sigma\,\boldsymbol{F}^n_\sigma\cdot\boldsymbol{n}_{K,\sigma}\right]$$
$$+\sum_{\sigma\in\mathscr{E}(K),\ \sigma=K|L}\frac{1}{2}|\sigma|\,((E_p)^n_K+(E_p)^n_L)\,\boldsymbol{u}^n_\sigma\cdot\boldsymbol{n}_{K,\sigma}=-(R_e)^{n+1}_K, \tag{2.9}$$

with

$$
\begin{aligned}
(E_p)_K^n &= \frac{1}{2} g\, (h_K^n)^2, \\
(E_k)_K^n &= \frac{1}{4\,|K|} \sum_{i=1}^{2} \sum_{\sigma \in \mathscr{E}^{(i)}(K)} |D_\sigma|\, h_{D_\sigma}^n (u_{i,\sigma}^n)^2, \\
G_{K,\sigma}^n &= \frac{1}{4} \sum_{i=1}^{2} \sum_{\sigma \in \mathscr{E}^{(i)}(K)} \sum_{\epsilon \in \widetilde{\mathscr{E}}^{(i)}(D_\sigma)} |\epsilon|\, (u_{i,\epsilon}^n)^2\, \boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon};
\end{aligned}
$$

and where the main residual $(R_e)_K^{n+1}$ is such that $(R_e)_K^{n+1} \geq (R_e^1)_K^{n+1} + (R_e^2)_K^{n+1} + (R_e^3)_K^{n+1}$ with

$$
\begin{aligned}
(R_e^1)_K^{n+1} &= g \sum_{\sigma \in \mathscr{E}(K)} |\sigma|\, (h_K^{n+1} - h_K^n)\, h_\sigma^n\, \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma} \\
(R_e^2)_K^{n+1} &= \sum_{\sigma \in \mathscr{E}(K),\ \sigma = K|L} \frac{1}{2} |\sigma| \Big[((E_p)_K^{n+1} - (E_p)_L^{n+1})\, \boldsymbol{u}_\sigma^{n+1} - ((E_p)_K^n - (E_p)_L^n)\, \boldsymbol{u}_\sigma^n \Big] \cdot \boldsymbol{n}_{K,\sigma}, \\
(R_e^3)_K^{n+1} &= \frac{\omega}{2} \Big[\sum_{\sigma \in \mathscr{E}^{(1)}(K)} |D_\sigma| h_{D_\sigma}^{n+1}\, (u_{2,\sigma}^*)^n u_{1,\sigma}^{n+1} - \sum_{\tau \in \mathscr{E}^{(2)}(K)} |D_\tau| h_{D_\tau}^{n+1}\, (u_{1,\tau}^*)^n u_{2,\tau}^{n+1} \Big].
\end{aligned}
$$

Remark that the main residual of this inequality may be not positive and even less does not disappear by global summation over $K \in \mathscr{M}$. However we emphasize that owing a consistency analysis we can show that this present scheme satisfies a weak energy inequality and we refer to the chapter 1 for more explanations.
Regarding the semi-discrete energy one has;

$$
\begin{aligned}
|K|\big((E_k)_K + (E_p)_K\big) + \sum_{\sigma \in \mathscr{E}(K)} \Big[G_{K,\sigma} + \frac{1}{2} |\sigma|\, g h_\sigma\, \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \Big] \\
+ \sum_{\sigma \in \mathscr{E}(K),\ \sigma = K|L} \frac{1}{2} |\sigma|\, ((E_p)_K + (E_p)_L)\, \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} = -(R_e)_K,
\end{aligned}
$$

with

$$
(R_e)_K \geq \frac{\omega}{2} \Big[\sum_{\sigma \in \mathscr{E}^{(1)}(K)} |D_\sigma| h_{D_\sigma}\, u_{2,\sigma}^* u_{1,\sigma} - \sum_{\tau \in \mathscr{E}^{(2)}(K)} |D_\tau| h_{D_\tau}\, u_{1,\tau}^* u_{2,\tau} \Big].
$$

Then summing the left and right hand sides of this inequality over $K \in \mathscr{M}$, the flux terms vanish and we get

$$
\sum_{K \in \mathscr{M}} |K|\big((E_k)_K + (E_p)_K\big) \leq -\frac{\omega}{2} \sum_{K \in \mathscr{M}} \Big[\sum_{\tau \in \mathscr{E}^{(2)}(K)} |D_\tau| h_{D_\tau}\, u_{1,\tau}^* u_{2,\tau} - \sum_{\sigma \in \mathscr{E}^{(1)}(K)} |D_\sigma| h_{D_\sigma}\, u_{2,\sigma}^* u_{1,\sigma} \Big].
$$

Here again, we can unfortunately make the same observation as above since the right hand side of this inequality may be not signed positively. In fact this non-dissipation of the semi-discrete energy is caused by the reconstructions $u_{2,\sigma}^*$ and $u_{1,\tau}^*$.

Now we turn out to the linear equations (2.4) for which a linear semi-implicit MAC scheme is obtained straightforwardly following a linearization of the non-linear scheme (2.7) around the state $(h_0, \boldsymbol{u}_0)$ with $\boldsymbol{u}_0 = 0$.

Linear semi-implicit MAC scheme – The linear scheme reads:

$$\frac{1}{\delta t}\,(h_K^{n+1} - h_K^n) + h_0 \mathrm{div}_K(\boldsymbol{u}^n) = 0, \qquad \forall\, K \in \mathscr{M}, \tag{2.10a}$$

$$\frac{1}{\delta t}\,(u_{1,\sigma}^{n+1} - u_{1,\sigma}^n) + g\,(\eth_1 h^{n+1})_\sigma = \omega\,(u_{2,\sigma}^*)^n, \quad \text{for } \sigma = K|L \in \mathscr{E}_{int}^{(1)}, \tag{2.10b}$$

$$\frac{1}{\delta t}\,(u_{2,\tau}^{n+1} - u_{2,\tau}^n) + g\,(\eth_2 h^{n+1})_\tau = -\omega\,(u_{1,\tau}^*)^{n+1}, \quad \text{for } \tau = K|M \in \mathscr{E}_{int}^{(2)}, \tag{2.10c}$$

with

$$\mathrm{div}_K(\boldsymbol{u}^n) = \frac{1}{|K|}\Big[\sum_{\sigma\in\mathscr{E}^{(1)}(K)} |\sigma|\, u_{1,\sigma}\, \boldsymbol{e}^{(1)}\cdot\boldsymbol{n}_{K,\sigma} + \sum_{\tau\in\mathscr{E}^{(2)}(K)} |\tau|\, u_{2,\tau}\, \boldsymbol{e}^{(2)}\cdot\boldsymbol{n}_{K,\tau}\Big],$$

where the quantities $(\eth_1 h^{n+1})_\sigma$, $(\eth_2 h^{n+1})_\tau$, $(u_{2,\sigma}^*)^n$ and $(u_{1,\tau}^*)^{n+1}$ are defined as previously.

In order to establish the well-balanced behavior of this linear scheme with respect to the preservation of the geostrophic equilibrium state, we introduce the following discrete counterpart of the continuous equilibrium state (2.5):

$$\forall\, 0 \le n \le N-1 : \left|\begin{array}{ll} \mathrm{div}_K(\boldsymbol{u}^n) = 0, & \text{for } K \in \mathscr{M}, \\ g\,(\eth_1 h^n)_\sigma - \omega\,(u_{2,\sigma}^*)^n = 0, & \text{for } \sigma = K|L \in \mathscr{E}_{int}^{(1)}, \\ g\,(\eth_2 h^n)_\tau + \omega\,(u_{1,\tau}^*)^n = 0, & \text{for } \tau = K|M \in \mathscr{E}_{int}^{(2)} \end{array}\right. \tag{2.11}$$

For now up, it is straightforward to see that all solution (h^n, u_1^n, u_2^n) of the linear scheme (2.10) satisfying the discrete relations (2.11) yields the following steady solution:

$$\forall\, 0 \le n \le N-1 : \left|\begin{array}{ll} h_K^{n+1} = h_K^n, & \forall K \in \mathscr{M}, \\ u_{1,\sigma}^{n+1} = u_{1,\sigma}^n, & \forall \sigma \in \mathscr{E}_{int}^{(1)}, \\ u_{2,\tau}^{n+1} = u_{2,\tau}^n, & \forall \tau \in \mathscr{E}_{int}^{(2)} \end{array}\right.$$

However the last two relations of (2.11) do not imply the first one of divergence free, which is in total disagreement with the continuous setting. Indeed one can show that if (h^n, u_1^n, u_2^n) satisfies both last equations of (2.10) then we have

$$\mathrm{div}_K(\boldsymbol{u}^n) + \mathrm{div}_L(\boldsymbol{u}^n) + \mathrm{div}_M(\boldsymbol{u}^n) + \mathrm{div}_N(\boldsymbol{u}^n) = 0,$$

which does not imply the divergence free condition and where N is a neighboring cell of L and M (see Figure 2.1).

In what follows, the attempt is to discretize the velocity equations on an associated diamond cell of the RT elements instead of the MAC grid. In particular for the upwind RT scheme we are designing follows, we no longer need a reconstruction for the perpendicular velocity $\boldsymbol{u}^{\perp}$. This quantity is approximated by a θ-scheme type method as suggested in Gunawan 2015.

2.2.3 Semi-implicit RT schemes

We begin by describing the non-linear scheme for the equations (2.1) and then the corresponding linearized version.

Nonlinear semi-implicit RT scheme – The semi-implicit RT scheme is written in vector form as follows:

$$\frac{1}{\delta t}\,(h_K^{n+1}-h_K^n)+\mathrm{div}_K(h^n\,\boldsymbol{u}^n)=0,\quad \forall\, K\in\mathscr{M}, \tag{2.12a}$$

$$\frac{1}{\delta t}\,(h_{D_\sigma}^{n+1}\boldsymbol{u}_\sigma^{n+1}-h_{D_\sigma}^n\,\boldsymbol{u}_\sigma^n)+\mathrm{div}_{D_\sigma}(h^n\boldsymbol{u}^n\,\boldsymbol{u}^n)+g\,h_{\sigma,c}^{n+1}\,(\nabla h^{n+1})_\sigma = -\omega\,(h\boldsymbol{u}^{\perp})_{D_\sigma}^{n,\theta},\qquad \forall\,\sigma\in\mathscr{E}_{int}, \tag{2.12b}$$

where

$$\begin{aligned}
\mathrm{div}_K(h\boldsymbol{u}) &= \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\,\boldsymbol{F}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}, && \text{with } \boldsymbol{F}_\sigma^n=h_\sigma^n\,\boldsymbol{u}_\sigma^n,\\
(\nabla h)_\sigma &= \frac{|\sigma|}{|D_\sigma|}\,(h_L-h_K)\,\boldsymbol{n}_{K,\sigma}, && \text{for } \sigma=K|L\in\mathscr{E}_{\mathrm{int}},\\
\mathrm{div}_{D_\sigma}(h\boldsymbol{u}\,\boldsymbol{u}) &= \frac{1}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)}|\epsilon|\,\boldsymbol{u}_\epsilon\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}, && \text{for } \sigma\in\mathscr{E}_{\mathrm{int}}.
\end{aligned}$$

The terms $h_{\sigma,c}$, h_{D_σ}, h_σ and $\boldsymbol{u}_\epsilon$ are approximated the same way as in the MAC scheme. While the quantities $\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}$ are defined as a linear combination of $\boldsymbol{F}_\sigma$ as follows:

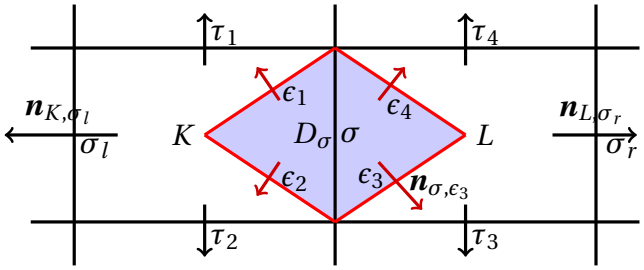


Figure 2.4 – Definition of the momentum flux $\boldsymbol{F}_\epsilon$ outward the dual cell D_σ.

$\forall \epsilon \in \widetilde{\mathscr{E}}(D_\sigma)$ such that $\epsilon \subset K$, we have

$$|\epsilon|\ \boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} = \lambda_\sigma\ |\sigma|\ \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma} + \lambda_{\sigma_l}\ |\sigma_l|\ \boldsymbol{F}_{\sigma_l} \cdot \boldsymbol{n}_{K,\sigma_l} + \lambda_{\tau_1}\ |\tau_1|\ \boldsymbol{F}_{\tau_1} \cdot \boldsymbol{n}_{K,\tau_1} + \lambda_{\tau_2}\ |\tau_2|\ \boldsymbol{F}_{\tau_2} \cdot \boldsymbol{n}_{K,\tau_2}, \tag{2.13}$$

where the coefficients λ_σ are now defined. Using the notations introducing in Figure 2.13, we have $\widetilde{\mathscr{E}}(D_\sigma) = \{\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$ and thus

$$\sum_{\epsilon \in \widetilde{\mathscr{E}}(D_\sigma)} |\epsilon|\ \boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} = |\epsilon_1|\ \boldsymbol{F}_{\epsilon_1} \cdot \boldsymbol{n}_{\sigma,\epsilon_1} + |\epsilon_2|\ \boldsymbol{F}_{\epsilon_2} \cdot \boldsymbol{n}_{\sigma,\epsilon_2} + |\epsilon_3|\ \boldsymbol{F}_{\epsilon_3} \cdot \boldsymbol{n}_{\sigma,\epsilon_3} + |\epsilon_4|\ \boldsymbol{F}_{\epsilon_4} \cdot \boldsymbol{n}_{\sigma,\epsilon_4}.$$

Following Ansanay-Alex, Babik, Latché, et al. 2011, the fluxes $\boldsymbol{F}_\epsilon$ are defined as a linear combination of $\boldsymbol{F}_\sigma$ by:

$$|\epsilon_1|\ \boldsymbol{F}_{\epsilon_1} \cdot \boldsymbol{n}_{\sigma,\epsilon_1} = \frac{1}{8}\big(-3\ |\sigma|\ \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma} + |\sigma_l|\ \boldsymbol{F}_{\sigma_l} \cdot \boldsymbol{n}_{K,\sigma_l} + 3\ |\tau_1|\ \boldsymbol{F}_{\tau_1} \cdot \boldsymbol{n}_{K,\tau_1} - |\tau_2|\ \boldsymbol{F}_{\tau_2} \cdot \boldsymbol{n}_{K,\tau_2}\big),$$

$$|\epsilon_2|\ \boldsymbol{F}_{\epsilon_2} \cdot \boldsymbol{n}_{\sigma,\epsilon_2} = \frac{1}{8}\big(-3\ |\sigma|\ \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma} + |\sigma_l|\ \boldsymbol{F}_{\sigma_l} \cdot \boldsymbol{n}_{K,\sigma_l} - |\tau_1|\ \boldsymbol{F}_{\tau_1} \cdot \boldsymbol{n}_{K,\tau_1} + 3\ |\tau_2|\ \boldsymbol{F}_{\tau_2} \cdot \boldsymbol{n}_{K,\tau_2}\big),$$

$$|\epsilon_3|\ \boldsymbol{F}_{\epsilon_3} \cdot \boldsymbol{n}_{\sigma,\epsilon_3} = \frac{1}{8}\big(-3\ |\sigma|\ \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{L,\sigma} + |\sigma_r|\ \boldsymbol{F}_{\sigma_r} \cdot \boldsymbol{n}_{L,\sigma_r} + 3\ |\tau_3|\ \boldsymbol{F}_{\tau_3} \cdot \boldsymbol{n}_{L,\tau_3} - |\tau_4|\ \boldsymbol{F}_{\tau_4} \cdot \boldsymbol{n}_{L,\tau_4}\big),$$

$$|\epsilon_4|\ \boldsymbol{F}_{\epsilon_4} \cdot \boldsymbol{n}_{\sigma,\epsilon_4} = \frac{1}{8}\big(-3\ |\sigma|\ \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{L,\sigma} + |\sigma_r|\ \boldsymbol{F}_{\sigma_r} \cdot \boldsymbol{n}_{L,\sigma_r} - |\tau_3|\ \boldsymbol{F}_{\tau_3} \cdot \boldsymbol{n}_{L,\tau_3} + 3\ |\tau_4|\ \boldsymbol{F}_{\tau_4} \cdot \boldsymbol{n}_{L,\tau_4}\big).$$

Then the rotation term $h\boldsymbol{u}^\perp$ is approximated by $(h\boldsymbol{u})^{n,\theta}_{D_\sigma}$ which is computed by:

$$(h\boldsymbol{u})^{n,\theta}_{D_\sigma} = \big(h^n_{D_\sigma}\boldsymbol{u}^n_\sigma + \boldsymbol{\theta}(h^{n+1}_{D_\sigma}\boldsymbol{u}^{n+1}_\sigma - h^n_{D_\sigma}\boldsymbol{u}^n_\sigma)\big)^\perp,\ \forall\ \sigma \in \mathscr{E}_{int}, \tag{2.14}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ with $\theta_1, \theta_2 \in [0,1]$ and

$$\big(h^n_{D_\sigma}\boldsymbol{u}^n_\sigma + \boldsymbol{\theta}(h^{n+1}_{D_\sigma}\boldsymbol{u}^{n+1}_\sigma - h^n_{D_\sigma}\boldsymbol{u}^n_\sigma)\big)^\perp = \begin{bmatrix} -h^n_{D_\sigma}u^n_{2,\sigma} - \theta_2(h^{n+1}_{D_\sigma}u^{n+1}_{2,\sigma} - h^n_{D_\sigma}u^n_{2,\sigma}) \\ \\ h^n_{D_\sigma}u^n_{1,\sigma} + \theta_1(h^{n+1}_{D_\sigma}u^{n+1}_{1,\sigma} - h^n_{D_\sigma}u^n_{1,\sigma}) \end{bmatrix}.$$

In fact, we note that the value $\boldsymbol{\theta} = (0,0)$, corresponds to an explicit time integration of the Coriolis term while $\boldsymbol{\theta} = (1,1)$ involves a full implicit method. By a simple algebraic computation, the components of the velocity $u^{n+1}_{1,\sigma}$ and $u^{n+1}_{2,\sigma}$ are explicitly solved by:

$$\begin{aligned}(1+\delta t^2\omega^2\ \theta_1\theta_2)h^{n+1}_{D_\sigma}u^{n+1}_{1,\sigma} = (1-\delta t^2\omega^2(1-\theta_1)\theta_2)h^n_{D_\sigma}u^n_{1,\sigma} + \delta t\omega\ h^n_{D_\sigma}u^n_{2,\sigma} \\ -\delta t\big(\mathrm{div}_{D_\sigma}(h^n\boldsymbol{u}^n\ \ u^n_1) + \delta t\omega\theta_2\ \mathrm{div}_{D_\sigma}(h^n\boldsymbol{u}^n\ \ u^n_2)\big) \\ -\delta t g\ h^{n+1}_{\sigma,c}\ \big((\eth_1 h^{n+1})_\sigma + \delta t\omega\theta_2\ (\eth_2 h^{n+1})_\sigma\big);\end{aligned}$$

and

$$(1+\delta t^2\omega^2\,\theta_1\theta_2)h_{D_\sigma}^{n+1}u_{2,\sigma}^{n+1} = (1-\delta t^2\omega^2(1-\theta_2)\theta_1)h_{D_\sigma}^{n}u_{2,\sigma}^{n} - \delta t\omega\; h_{D_\sigma}^{n}u_{1,\sigma}^{n} - \delta t\big(\mathrm{div}_{D_\sigma}(h^n\boldsymbol{u}^n\;\; u_2^n) - \delta t\omega\theta_1\;\mathrm{div}_{D_\sigma}(h^n\boldsymbol{u}^n\;\; u_1^n)\big) - \delta t g\; h_{\sigma,c}^{n+1}\;\big((\eth_2 h^{n+1})_\sigma - \delta t\omega\theta_1\;(\eth_1 h^{n+1})_\sigma\big).$$

Remark 2.1. *Let us note here that when we consider an uniform rectangular grid, the discrete terms $h_{\sigma,c}$ and h_{D_σ} are equal, indeed in this particular case $|K|$ remains constant for all $K\in\mathscr{M}$.*
In addition, in this same configuration the i^{th} component of the discrete pressure gradient defined in the MAC and RT environments are in fact different since we have

$$\textit{MAC:}\qquad (\eth_i\; h)_\sigma = \frac{|\sigma|}{|K|}\;(h_L - h_K)\,\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{e}^{(i)},\;\; \textit{for}\;\sigma = K|L\in\mathscr{E}_{\mathrm{int}}^{(i)},$$

$$\textit{RT:}\qquad (\nabla h)_\sigma = \big((\eth_1\; h)_\sigma, (\eth_2\; h)_\sigma\big)^T\;\; \textit{with}\;\; (\eth_i\; h)_\sigma = \frac{2|\sigma|}{|K|}\;(h_L - h_K)\,\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{e}^{(i)}.$$

Beyond this disagreement, both quantities are perfectly consistent with the pressure gradient on their associated dual or diamond mesh and are defined in order to satisfy the following div-grad relationship:

$$\sum_{K\in\mathscr{M}}|K|\;\mathrm{div}_K(h\;\boldsymbol{u}) + \sum_{i=1}^{2}\sum_{\sigma\in\mathscr{E}_{\mathrm{int}}^{(i)}}|D_\sigma|\;u_{i,\sigma}\;(\eth_i\; h)_\sigma = 0. \tag{2.15}$$

2.2.3.1 Stability of the RT scheme

We investigate the discrete properties satisfied by the semi-implicit RT scheme (2.12). The first lemma ensures the preservation of the positivity of the water height under a CFL like condition and the "lake at rest" steady state.

Proposition 2.1 (Preservation of the positivity and the lake at rest steady state)**.** *Let $n\in\{0,\cdots,N_t-1\}$, let $(h_K^n,\;\boldsymbol{u}_\sigma^n)_{K\in\mathscr{M},\,\sigma\in\mathscr{E}}$ be given such that $h_K^n\geq 0$, for all $K\in\mathscr{M}$ and let h_K^{n+1} computed by the scheme (2.12). Then $h_K^{n+1}>0$, for all $K\in\mathscr{M}$ under the following CFL condition,*

$$\delta t \leq \frac{|K|}{\displaystyle\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\,|\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}|}. \tag{2.16}$$

Furthermore if $\boldsymbol{u}^n=0$ and $h^n=C$ with $C\in\mathbb{R}_+$, then $\boldsymbol{u}^{n+1}=0$ and $h^{n+1}=C$.

Proof. By the definition of h_σ^n, one has

$$h_\sigma^n\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma} = h_K^n\;(\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma})^+ + h_L^n\;(\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma})^-,\;\;\text{for}\;\sigma=K|L.$$

Thus the discrete mass equation (2.12a) yields:

$$h_K^{n+1} = h_K^n - \frac{\delta t}{|K|} \sum_{\sigma\in\mathscr{E}(K)} |\sigma| \, (\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma})^+ \, h_K^n + \frac{\delta t}{|K|} \sum_{\sigma\in\mathscr{E}(K)} |\sigma| \, (\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma})^- \, h_L^n,$$

Since $|\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}| \geq (\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma})^+$ we obtain that

$$h_K^{n+1} \geq \left[1 - \frac{\delta t}{|K|} \sum_{\sigma\in\mathscr{E}(K)} |\sigma| \, (\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma})^+\right] h_K^n \geq \left[1 - \frac{\delta t}{|K|} \sum_{\sigma\in\mathscr{E}(K)} |\sigma| \, |\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}|\right] h_K^n.$$

Since (by assumption) $h_K^n > 0$, then $h_K^{n+1} > 0$ provided that the time step δt satisfies (2.16). Next if $\boldsymbol{u}^n = 0$ and $h^n = C$ then we get immediately that $h^{n+1} = h^n = C$ and $\boldsymbol{u}^{n+1} = \boldsymbol{u}^n = 0$ which concludes the proof. □

Before studying the well balance property, we define the discrete geostrophic equilibrium state as follows:

$$\forall\, \sigma \in \mathscr{E}_{\rm int},\ \forall\, n \in \{0, \cdots, N-1\}: \quad g\,(\nabla h^n)_\sigma + \omega\,(\boldsymbol{u}_\sigma^n)^\perp = 0. \tag{2.17}$$

Let us note that the relation (2.17) automatically implies the divergence free condition $\mathrm{div}_K(\boldsymbol{u}^n) = 0$ as in the continuous problem. Indeed it is straightforward to see that (2.17) leads to:

$$\omega\, \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma} = g\,(\nabla h^n)_\sigma^\perp \cdot \boldsymbol{n}_{K,\sigma} = \frac{|\sigma|}{|D_\sigma|}\,(h_L - h_K)\,\boldsymbol{n}_{K,\sigma}^\perp \cdot \boldsymbol{n}_{K,\sigma} = 0,\ \forall\, \sigma = K|L \in \mathscr{E}_{\rm int}.$$

In addition, in the case of an uniform rectangular grid, the scheme (2.12) holds the preservation of the discrete geostrophic equilibrium state. This result is proven in the same way as the Proposition 2.3.

Proposition 2.2 (Preservation of the geostrophic equilibrium state)**.** *Let $n \in \{0, \cdots, N_t - 1\}$, let consider $(h_K^{n+1},\ \boldsymbol{u}_\sigma^{n+1})_{K\in\mathscr{M},\ \sigma\in\mathscr{E}}$ computed from the scheme (2.12) such that $h_K^n > 0$, for all $K \in \mathscr{M}$ and satisfying*

$$g\,(\nabla h^n)_\sigma + w\,(\boldsymbol{u}_\sigma^n)^\perp = 0, \quad \forall\, \sigma \in \mathscr{E}_{\rm int}.$$

Assume additionally that the meshing $\mathscr{M}$ is uniform that means the space step $\delta_{\mathscr{M}}$ is kept constant. Then $h_K^{n+1} = h_K^n$, for all $K \in \mathscr{M}$ and $\boldsymbol{u}_\sigma^{n+1} = \boldsymbol{u}_\sigma^n$, for all $\sigma \in \mathscr{E}_{\rm int}$.

Next we focus on the discrete mechanical energy derived from the current scheme that results of the sum of the discrete potential and kinetic energies.

Lemma 2.1 (Potential energy balance)**.** *Let $n \in \{0, \cdots, N-1\}$. A discrete solution to the*

semi-implicit scheme (2.12) *satisfies the following equation:*

$$\frac{1}{2}\frac{1}{\delta t}\, g\left((h_K^{n+1})^2-(h_K^n)^2\right)+\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\, g\, h_{\sigma,c}^{n+1}\, \boldsymbol{F}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}$$
$$-\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}\frac{1}{2}|D_\sigma|\, g\, (\nabla h^{n+1})_\sigma\cdot\boldsymbol{F}_\sigma^n=-\frac{1}{2}\frac{1}{\delta t}\, g\, (h_K^{n+1}-h_K^n)^2. \quad (2.18)$$

Sketch of proof. The result is obtained multiplying the discrete water height equation (2.12a) by $\frac{g}{h_0}\, h_K^{n+1}$ and using the following identities:

$$(h_K^{n+1}-h_K^n)h_K^{n+1}=\frac{1}{2}\left((h_K^{n+1})^2-(h_K^n)^2\right)+\frac{1}{2}\,(h_K^{n+1}-h_K^n)^2$$

and

$$h_K^{n+1}=\frac{h_K^{n+1}+h_L^{n+1}}{2}-\frac{h_L^{n+1}-h_K^{n+1}}{2}.$$

As a consequence of the careful discretization of the non-linear convection term, the scheme (2.12) satisfies a discrete kinetic energy balance, as stated in the following Lemma. The proof of this result is an easy adaptation of Herbin, Latché, and Nguyen 2018, Lemma 3.2.

Lemma 2.2 (Discrete kinetic balance)**.** *A solution to the scheme* (2.12b) *satisfies the following equality, for* $\sigma\in\mathscr{E}_{\rm int}$ *and* $0\le n\le N-1$*:*

$$\frac{1}{2\delta t}(h_{D_\sigma}^{n+1}|\boldsymbol{u}_\sigma^{n+1}|^2-h_{D_\sigma}^n|\boldsymbol{u}_\sigma^n|^2)+\frac{1}{2}\sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)}|\epsilon|\,|\boldsymbol{u}_\epsilon^n|^2\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}$$
$$+|D_\sigma|\, g\, h_{\sigma,c}^{n+1}\,(\nabla h^{n+1})_\sigma\cdot\boldsymbol{u}_\sigma^{n+1}+|D_\sigma|\,\omega h_{D_\sigma}^n\,((1-\boldsymbol{\theta})\boldsymbol{u}_\sigma^n)^\perp\cdot\boldsymbol{u}_\sigma^{n+1}=-\boldsymbol{R}_\sigma^{n+1}, \quad (2.19)$$

with

$$\boldsymbol{R}_\sigma^{n+1}=\frac{1}{2\delta t}\,|D_\sigma|\, h_{D_\sigma}^{n+1}\,(\boldsymbol{u}_\sigma^{n+1}-\boldsymbol{u}_\sigma^n)^2+\frac{1}{2}\sum_{\epsilon=\sigma|\sigma\in\widetilde{\mathscr{E}}}|\epsilon|\,(\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon})^-\,(\boldsymbol{u}_{\sigma'}^n-\boldsymbol{u}_\sigma^n)^2$$
$$-\sum_{\epsilon=\sigma|\sigma\in\widetilde{\mathscr{E}}}|\epsilon|\,(\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon})^-\,(\boldsymbol{u}_{\sigma'}^n-\boldsymbol{u}_\sigma^n)\cdot(\boldsymbol{u}_\sigma^{n+1}-\boldsymbol{u}_\sigma^n).$$

Furthermore the residual $\boldsymbol{R}_\sigma^{n+1}$ *is non-negative under the following CFL like condition:*

$$\delta t\le\frac{|D_\sigma|\, h_{D_\sigma}^{n+1}}{\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,(\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon})^-}.$$

Then let denote by E^n the discrete mechanical energy, this quantity being globally

defined by:

$$E^n = \sum_{K\in\mathcal{M}} |K| \frac{1}{2} g\,(h_K^n)^2 + \sum_{\sigma\in\mathscr{E}_{\rm int}} |D_\sigma| \frac{1}{2} h_{D_\sigma}^{n+1}\,|\boldsymbol{u}_\sigma^n|^2. \tag{2.20}$$

Then gathering the kinetic (2.19) and potential (2.18) energies, we get the following discrete mechanical energy inequality:

$$\frac{1}{\delta t}(E^{n+1}-E^n) \le \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathscr{E}(K)} |D_\sigma| \frac{g}{2} (\nabla h^{n+1})_\sigma \cdot \boldsymbol{F}_\sigma^n - \sum_{\sigma\in\mathscr{E}_{\rm int}} |D_\sigma|\, g\, h_{\sigma,c}^{n+1}\, (\nabla h^{n+1})_\sigma \cdot \boldsymbol{u}_\sigma^{n+1} - \sum_{\sigma\in\mathscr{E}_{\rm int}} |D_\sigma|\, \omega\, h_{D_\sigma}^n\, ((\boldsymbol{\theta}-1)\boldsymbol{u}_\sigma^n)^\perp \cdot \boldsymbol{u}_\sigma^{n+1}. \tag{2.21}$$

However we are not able to show that the term of the right hand side of this inequality (2.21) is negative, hence this discrete energy may not be dissipated. This instability is not occurred when we consider the semi-discrete mechanical energy inequality with respect to the space discretization, given by:

$$\frac{d}{dt}E \le R \qquad \text{with } R = \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathscr{E}(K)} \frac{1}{2}|D_\sigma|\, g(\nabla h)_\sigma \cdot \boldsymbol{F}_\sigma - \sum_{\sigma\in\mathscr{E}_{\rm int}} |D_\sigma|\, g\, h_{\sigma,c}\, (\nabla h)_\sigma \cdot \boldsymbol{u}_\sigma.$$

Thanks to the upwind approximation of h_σ and the centered interpolation of $h_{\sigma,c}$, the residual R is negative. Indeed, we have that:

$$R = -\sum_{\sigma=K|L\in\mathscr{E}_{\rm int}} \frac{1}{2}|\sigma|\, g(h_L-h_K)^2 \left[(\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma})^+ + (\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma})^-\right] \le 0,$$

since both quantities $(\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma})^+$ and $(\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma})^-$ are non-negative by definition. Thus the semi-discrete mechanical energy E is decreasing.

Here-below we present the linearized semi-implicit scheme.

2.2.3.2 Linear semi-implicit RT scheme

Following a linearization procedure for the scheme (2.12) around the state $(h_0, \boldsymbol{u}_0 = 0)$, we get the corresponding linear scheme which reads:

$$\frac{1}{\delta t}\,(h_K^{n+1}-h_K^n) + h_0 \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma} = 0, \quad \forall\, K\in\mathcal{M}, \tag{2.22a}$$

$$\frac{1}{\delta t}\,(\boldsymbol{u}_\sigma^{n+1}-\boldsymbol{u}_\sigma^n) + g\,(\nabla h^{n+1})_\sigma = -\omega\ (\boldsymbol{u}_\sigma^n + \boldsymbol{\theta}(\boldsymbol{u}_\sigma^{n+1}-\boldsymbol{u}_\sigma^n))^\perp, \forall\,\sigma\in\mathscr{E}_{int} \tag{2.22b}$$

where

$$\left(\boldsymbol{u}_\sigma^n + \boldsymbol{\theta}(\boldsymbol{u}_\sigma^{n+1}-\boldsymbol{u}_\sigma^n)\right)^\perp = \begin{bmatrix} -(u_{2,\sigma}^n + \theta_2(u_{2,\sigma}^{n+1}-u_{2,\sigma}^n)) \\ u_{1,\sigma}^n + \theta_1(u_{1,\sigma}^{n+1}-u_{1,\sigma}^n) \end{bmatrix} \tag{2.23}$$

As in the non-linear scheme (2.12), the discrete velocity equation (2.22b) is explicitly solved as follows:

$$u_{1,\sigma}^{n+1} = \frac{\left(1-\delta t^2\omega^2\,(1-\theta_1)\theta_2\right)\,u_{1,\sigma}^n - \delta t g\left((\eth_1 h^{n+1})_\sigma + \delta t\omega\theta_2\,(\eth_2 h^{n+1})_\sigma\right) + \delta t\omega\,u_{2,\sigma}^n}{1+\delta t^2\omega^2\,\theta_1\theta_2},$$

$$u_{2,\sigma}^{n+1} = \frac{\left(1-\delta t^2\omega^2\,(1-\theta_2)\theta_1\right)\,u_{2,\sigma}^n - \delta t g\left((\eth_2 h^{n+1})_\sigma - \delta t\omega\theta_1\,(\eth_1 h^{n+1})_\sigma\right) - \delta t\omega\,u_{1,\sigma}^n}{1+\delta t^2\omega^2\,\theta_1\theta_2}.$$

where the discrete divergence and gradient operators involved in the scheme (2.22) are defined in the previous section.

This linear scheme inherits some discrete properties from the non-linear one (2.12) that are enumerated below.

Preservation of the geostrophic equilibrium state – The discrete solution $(h^n, \boldsymbol{u}^n)$ computed from the linear scheme (2.22) preserves the discrete geostrophic equilibrium state.

Proposition 2.3 (Preservation of the geostrophic equilibrium state, linear scheme)**.** *Let $n \in \{0,\cdots,N_t-1\}$ and let $(h_K^n,\ \boldsymbol{u}_\sigma^n)_{K\in\mathscr{M},\,\sigma\in\mathscr{E}}$ the discrete solution of (2.22) such that $g\,(\nabla h^n)_\sigma + w\,(\boldsymbol{u}_\sigma^n)^\perp = 0, \quad \forall\,\sigma\in\mathscr{E}_{\rm int}$. Then $h_K^{n+1} = h_K^n$, for all $K\in\mathscr{M}$ and $\boldsymbol{u}_\sigma^{n+1} = \boldsymbol{u}_\sigma^n$, for all $\sigma\in\mathscr{E}_{\rm int}$.*

Proof. Since the discrete equilibrium state $g\,(\nabla h^n)_\sigma + w\,(\boldsymbol{u}_\sigma^n)^\perp = 0$ implies that $\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma} = 0$, then $\mathrm{div}_K(\boldsymbol{u}^n) = 0$ and thus $h_K^{n+1} = h_K^n$ for all $K\in\mathscr{M}$. Hence the discrete velocity equations boils down to:

$$\frac{u_{1,\sigma}^{n+1}-u_{1,\sigma}^n}{\delta t} = -\frac{\delta t\omega\,\theta_2(g(\eth_2 h^n)_\sigma + \omega\,u_{1,\sigma}^n)}{1+\delta t^2\omega^2\,\theta_1\theta_2} - \frac{g(\eth_1 h^n)_\sigma - \omega\,u_{2,\sigma}^n}{1+\delta t^2\omega^2\,\theta_1\theta_2},$$

$$\frac{u_{2,\sigma}^{n+1}-u_{2,\sigma}^n}{\delta t} = \frac{\delta t\omega\,\theta_1(g(\eth_1 h^n)_\sigma - \omega\,u_{2,\sigma}^n)}{1+\delta t^2\omega^2\,\theta_1\theta_2} - \frac{g(\eth_2 h^n)_\sigma + \omega\,u_{1,\sigma}^n}{1+\delta t^2\omega^2\,\theta_1\theta_2}.$$

Then the relation $g\,(\nabla h^n)_\sigma + w\,(\boldsymbol{u}_\sigma^n)^\perp = 0$ leads to:

$$\frac{u_{1,\sigma}^{n+1}-u_{1,\sigma}^n}{\delta t} = 0 \ \text{ and } \ \frac{u_{2,\sigma}^{n+1}-u_{2,\sigma}^n}{\delta t} = 0,\ \forall\,\sigma\in\mathscr{E}_{\rm int};$$

which concludes the proof. □

Discrete mechanical energy – A discrete mechanical energy inequality may derived from the linear scheme (2.22)resulting of the addition of the potential and kinetic energies. This latter one is quickly obtained by taking the scalar product of the discrete velocity (2.22b) with $\frac{1}{2}(\boldsymbol{u}_\sigma^{n+1}+\boldsymbol{u}_\sigma^n)$:

$$\frac{1}{2}\frac{1}{\delta t}\left(|\boldsymbol{u}_\sigma^{n+1}|^2 - |\boldsymbol{u}_\sigma^n|^2\right) + g\frac{1}{2}\,(\nabla h^{n+1})_\sigma\cdot(\boldsymbol{u}_\sigma^{n+1}+\boldsymbol{u}_\sigma^n) = \frac{\omega}{2}\,((2\boldsymbol{\theta}-1)\boldsymbol{u}_\sigma^n)^\perp\cdot\boldsymbol{u}_\sigma^{n+1}.$$

The discrete potential satisfies:

$$\frac{1}{2}\frac{1}{\delta t}\frac{g}{h_0}\left((h_K^{n+1})^2-(h_K^n)^2\right)+\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\,g h_{\sigma,c}^{n+1}\,\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}$$
$$-\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}\frac{1}{2}|D_\sigma|\,g(\nabla h^{n+1})_\sigma\cdot\boldsymbol{u}_\sigma^n=-\frac{1}{2}\frac{1}{\delta t}\frac{g}{h_0}\,(h_K^{n+1}-h_K^n)^2.$$

Then multiplying the potential equation by $|K|$ (resp the kinetic by $|D_\sigma|$) and summing the obtained result over $K\in\mathscr{M}$ (resp $\sigma\in\mathscr{E}_{\rm int}$), we get

$$\frac{1}{\delta t}(E^{n+1}-E^n)\le\sum_{K\in\mathscr{M}}\sum_{\sigma\in\mathscr{E}(K)}|D_\sigma|\,\frac{g}{2}\,(\nabla h^{n+1})_\sigma\cdot\boldsymbol{u}_\sigma^n-\sum_{\sigma\in\mathscr{E}_{\rm int}}\frac{1}{2}|D_\sigma|\,g(\nabla h^{n+1})_\sigma\cdot(\boldsymbol{u}_\sigma^{n+1}+\boldsymbol{u}_\sigma^n)$$
$$-\sum_{\sigma\in\mathscr{E}_{\rm int}}|D_\sigma|\,\frac{\omega}{2}\,((2\boldsymbol{\theta}-1)\boldsymbol{u}_\sigma^{n+1})^\perp\cdot\boldsymbol{u}_\sigma^n,$$

with

$$E^n=\sum_{K\in\mathscr{M}}|K|\,\frac{1}{2}\frac{g}{h_0}\,(h_K^n)^2+\sum_{\sigma\in\mathscr{E}_{\rm int}}|D_\sigma|\,\frac{1}{2}\,|\boldsymbol{u}_\sigma^n|^2.$$

It is legitimate to emphasize that a linear RT scheme based on the Crank-Nicholson time integration method consists in a good alternative to the linear semi-implicit scheme 2.22. As shown in the Appendix 2.B, the linear Crank-Nicholson time discretization scheme holds the preservation of the geostrophic equilibrium as well as the conservation of the fully discrete mechanical energy. Despite its good properties, the extension to the non-linear case is much more complex to solve in the sense that its leads to a non-linear resolution system and even less one probably loses control of the mechanical energy. For the sake of simplicity, this track is purely discarded and we will proceed to a stabilization technique to design non-linear and linear schemes in the following section.

The semi-implicit MAC and RT schemes are known to be more diffusive since the first order upwind method is used for both the mass and momentum fluxes. In order to reduce the numerical diffusion produced by the upwind MAC and RT schemes, we follow a stabilization procedure mimicking the technique developed in Couderc, Duran, and Vila 2017 and Audusse, Do, Omnes, et al. 2018. The content of the next section is part of a project of the CEMRACS 19 event around the topic **numerical approximation of the shallow water equations with Coriolis source term**. A first paper of this project is submitted Audusse, Dubos, Duran, et al. 2020 and an other undertaken work will appear in the future. This work was carried out in conjunction with **E. Auddusse, V. Dubos, A. Duran, N. Gaveau and Y. Penel**. For the sake of compactness only staggered schemes working on the RT finite elements are studied in the sequel while regarding collocated schemes an analogue technique are under review.

2.3 Stabilization methods for non-linear and linear schemes

In this section, we address to a modified version of the original problem (2.1) where two corrections are introduced following the technique developed in Audusse, Do, Omnes, et al. 2018; Couderc, Duran, and Vila 2017, as follows:

$$\partial_t h + \operatorname{div}(h\,\boldsymbol{u} - \Lambda) = 0, \tag{2.24a}$$

$$\partial_t (h\boldsymbol{u}) + \operatorname{div}((h\boldsymbol{u} - \Lambda) \otimes \boldsymbol{u}) + gh\nabla h - \nabla\pi = -w\,(h\,\boldsymbol{u} - \Lambda)^{\perp}. \tag{2.24b}$$

The quantities Λ and π are two artificial corrections of the mass flow and the pressure gradient respectively that are motivated by the following arguments. Firstly both quantities are expected to vanish when we reach the geostrophic equilibrium state where the relation $g\nabla h + w\,\boldsymbol{u}^{\perp} = 0$ holds. Second point, Λ and π are defined to dissipate the mechanical energy of the modified equations (2.24).
Multiplying the equations (2.24a) and (2.24b) respectively by gh and $\boldsymbol{u}$, we get after some algebraic computations:

$$\partial_t E + \operatorname{div}\Big((gh + \frac{1}{2}|\boldsymbol{u}|^2)\,(h\boldsymbol{u} - \Lambda)\Big) + \operatorname{div}(h\;\boldsymbol{u}\;\pi) = -\Big(\Lambda \cdot (g\nabla h + \omega\,\boldsymbol{u}^{\perp}) + \pi\operatorname{div}(\boldsymbol{u})\Big),$$

with $E = \frac{1}{2}\frac{g}{h_0}h^2 + \frac{1}{2}\,|\boldsymbol{u}|^2$. Thus is straightforward to see that if $\Lambda \approx g\nabla h + w\boldsymbol{u}^{\perp}$ and $\pi \approx \operatorname{div}(\boldsymbol{u})$, then

$$\partial_t E + \operatorname{div}\Big((gh + \frac{1}{2}|\boldsymbol{u}|^2)\,(h\boldsymbol{u} - \Lambda)\Big) + \operatorname{div}(h\;\boldsymbol{u}\;\pi) \leq 0.$$

Then the modified linear equations are obtained following a linearization of the system (2.24) around the state $(h_0, \boldsymbol{u}_0 = 0)$:

$$\partial_t h + h_0\,\operatorname{div}(\boldsymbol{u} - \Lambda) = 0, \tag{2.25a}$$

$$\partial_t(\boldsymbol{u}) + \nabla\Big(gh - \pi\Big) = -w\,(\boldsymbol{u} - \Lambda)^{\perp}, \tag{2.25b}$$

where $\Lambda \approx g\nabla h + w\,\boldsymbol{u}^{\perp}$ and $\pi \approx \operatorname{div}(\boldsymbol{u})$. As a consequence of these definitions, the linear and modified linear equations hold the same steady state solution and furthermore these quantities yield the following mechanical energy inequality:

$$\partial_t(\frac{1}{h_0}gh^2 + \frac{1}{2}|\boldsymbol{u}|^2) + \operatorname{div}(gh(\boldsymbol{u} - \Lambda)) + \operatorname{div}(\boldsymbol{u}\;\pi) \leq 0.$$

The novelty of this approach is observed in the source term where the mass flux correction is introduced which is crucial to guarantee a dissipation of the semi-discrete mechanical energy.

In this current section we wish to design semi-discrete schemes with respect to the space variable for the non-linear system (2.24) which ensure a dissipation of the semi-

discrete mechanical energy and which preserve the geostrophic equilibrium state. In this purpose, we propose and analyze semi-discrete and fully discrete schemes, starting by the semi-discrete schemes and closing the numerical resolution following a segregated time integration mimicking the semi-implicit RT scheme (2.12).

2.3.1 Semi-discrete staggered schemes and stability analysis

The semi-discrete scheme according to a staggered spatial discretization for the non-linear equations (2.24) reads:

$$\frac{d}{dt}\, h_K + \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, \boldsymbol{F}_\sigma\cdot\boldsymbol{n}_{K,\sigma} = 0, \;\text{ with }\; \boldsymbol{F}_\sigma = h_{D_\sigma}\boldsymbol{u}_\sigma - \Lambda_\sigma, \quad \forall\, K\in\mathscr{M}, \tag{2.26a}$$

$$\frac{d}{dt}\,(h_{D_\sigma}\boldsymbol{u}_\sigma) + \frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)} |\epsilon|\, \boldsymbol{u}_\epsilon\, \boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon} + g h_{D_\sigma}(\nabla h)_\sigma - (\nabla\pi)_\sigma$$
$$= -\omega(h_{D_\sigma}\boldsymbol{u}_\sigma - \Lambda_\sigma)^\perp, \forall\, \sigma\in\mathscr{E}_{int}, \tag{2.26b}$$

where the quantity Λ_σ is a numerical diffusion defined on the dual mesh by:

$$\Lambda_\sigma = \gamma\,\delta t\; h_{D_\sigma}(g\,(\nabla h)_\sigma + \omega\,\boldsymbol{u}_\sigma^\perp), \quad \gamma\geq 0;$$

and h_{D_σ} is given by

$$h_{D_\sigma} = \frac{|D_{K,\sigma}|}{|D_\sigma|}h_K + \frac{|D_{L,\sigma}|}{|D_\sigma|}h_L, \quad \forall\,\sigma = K|L\in\mathscr{E}_{\text{int}}.$$

We recall that the semi-discrete gradient $\nabla(\cdot)_\sigma$ applied to π (resp. h) is defined by:

$$(\nabla\pi)_\sigma = \frac{|\sigma|}{|D_\sigma|}(\pi_L - \pi_K)\boldsymbol{n}_{K,\sigma}, \;\text{ for }\sigma = K|L\in\mathscr{E}_{\text{int}},$$

where π_K is also a correction term defined in the primal mesh as:

$$\pi_K = \nu\,\delta t\; g\; h_K\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\;\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}, \qquad \nu\geq 0.$$

Finally the discrete velocity $\boldsymbol{u}_\epsilon$ is upwinding with respect to the flow $\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}$ given by:

$$|\epsilon|\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon} = \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\lambda_\sigma\;\boldsymbol{F}_\sigma\cdot\boldsymbol{n}_{K,\sigma},$$

where the coefficients λ_σ are given in the Table 2.1. The above definitions of h_{D_σ} and $\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}$ satisfy the following balance:

$$\frac{d}{dt}\, h_{D_\sigma} + \frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)} |\epsilon|\;\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon} = 0. \tag{2.27}$$

Let notice here that the main part of the mass flux $\boldsymbol{F}_\sigma$ that means $h_{D_\sigma}\boldsymbol{u}_\sigma$ is not usual but is necessary in order to deal with a dissipation of the semi-discrete mechanical energy as we are going to tackle below. This kind of approximation is proposed by Duran and Vila 2019 and yields to a centered mass flux in the case of an uniform rectangular grid.

Semi-discrete energy balance – The semi-discrete scheme described herebefore satisfies a semi-discrete energy balance equation which is collected from the semi-discrete potential and kinetic energies that we handle in Lemmas 2.3, 2.4 below.

A semi-discrete potential energy is derived from the scheme (2.26) as stated in the following lemma.

Lemma 2.3 (Semi-discrete potential energy balance)**.** *A semi-discrete solution of the scheme* (2.26) *satisfies the following equation:*

$$\frac{d}{dt}\left(\frac{1}{2}g\,h_K^2\right)+\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|g\,h_{\sigma,c}\,h_{D_\sigma}\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}-\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|g\,h_{\sigma,c}\,\Lambda_\sigma\cdot\boldsymbol{n}_{K,\sigma}$$
$$-\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|D_\sigma|\,\frac{1}{2}g\,(\nabla h)_\sigma\cdot h_{D_\sigma}\boldsymbol{u}_\sigma+\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|D_\sigma|\,\frac{1}{2}g\,(\nabla h)_\sigma\cdot\Lambda_\sigma=0. \quad (2.28)$$

Proof. Multiplying equation (2.26a) by gh_k, we get

$$\left(\frac{d}{dt}\,h_K+\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|\,\boldsymbol{F}_\sigma\cdot\boldsymbol{n}_{K,\sigma}\right)g\,h_K$$
$$=\frac{d}{dt}\left(\frac{1}{2}g\,h_K^2\right)+\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|g\,h_K\,h_{D_\sigma}\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}-\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|g\,h_K\,\Lambda_\sigma\cdot\boldsymbol{n}_{K,\sigma}.$$

Then using the identity $h_K=\frac{1}{2}(h_K+h_L)-\frac{1}{2}(h_L-h_K)$, we get the result, which concludes the proof. □

Next we derive a semi-discrete kinetic energy balance as follows.

Lemma 2.4 (Semi-discrete kinetic inequality)**.** *A solution of the scheme* (2.26) *satisfies the following equality, for* $\sigma\in\mathscr{E}_{\rm int}$*:*

$$\frac{d}{dt}\left(\frac{1}{2}h_{D_\sigma}\,|\boldsymbol{u}_\sigma|^2\right)+\frac{1}{|D_\sigma|}\frac{1}{2}\sum_{\epsilon\in\tilde{\mathscr{E}}(D_\sigma)}|\epsilon|\,|\boldsymbol{u}_\epsilon|^2\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}$$
$$+g\,(\nabla h)_\sigma\cdot h_{D_\sigma}\boldsymbol{u}_\sigma+\omega\,\boldsymbol{u}_\sigma^\perp\cdot\Lambda_\sigma-\nabla(\pi)_\sigma\cdot\boldsymbol{u}_\sigma\le 0. \quad (2.29)$$

Proof. Taking the scalar product of the velocity equation (2.26b) with $\boldsymbol{u}_\sigma$, $\forall\,\sigma\in\mathscr{E}_{int}$,

we have

$$\frac{d}{dt}\left(h_{D_\sigma}\boldsymbol{u}_\sigma\right)\cdot\boldsymbol{u}_\sigma+\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,\boldsymbol{u}_\epsilon\cdot\boldsymbol{u}_\sigma\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}+\Big(gh_{D_\sigma}(\nabla h)_\sigma-(\nabla\pi)_\sigma-\omega\Lambda_\sigma^\perp\Big)\cdot\boldsymbol{u}_\sigma=0. \tag{2.30}$$

Then on one hand we get $\frac{d}{dt}\left(h_{D_\sigma}\boldsymbol{u}_\sigma\right)\cdot\boldsymbol{u}_\sigma=\dfrac{d}{dt}\big(\frac{1}{2}h_{D_\sigma}\,|\boldsymbol{u}_\sigma|^2\big)+\frac{1}{2}|\boldsymbol{u}_\sigma|^2\dfrac{d}{dt}h_{D_\sigma}$. On other hand, using the identity $\boldsymbol{u}_\epsilon\cdot\boldsymbol{u}_\sigma=\frac{1}{2}|\boldsymbol{u}_\epsilon|^2+\frac{1}{2}|\boldsymbol{u}_\sigma|^2-\frac{1}{2}(\boldsymbol{u}_\sigma-\boldsymbol{u}_\sigma)^2$ we obtain

$$\begin{aligned}\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,\boldsymbol{u}_\epsilon\cdot\boldsymbol{u}_\sigma\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}=\frac{1}{2}\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,|\boldsymbol{u}_\epsilon|^2\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}+\frac{1}{2}\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,|\boldsymbol{u}_\sigma|^2\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}\\-\frac{1}{2}\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,(\boldsymbol{u}_\epsilon-\boldsymbol{u}_\sigma)^2\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}\end{aligned}$$

Next thanks to the semi-discrete balance (2.27), we get

$$\begin{aligned}&\frac{d}{dt}\left(h_{D_\sigma}\boldsymbol{u}_\sigma\right)\cdot\boldsymbol{u}_\sigma+\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,\boldsymbol{u}_\epsilon\cdot\boldsymbol{u}_\sigma\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}\\&\qquad=\frac{d}{dt}\big(\frac{1}{2}h_{D_\sigma}\,|\boldsymbol{u}_\sigma|^2\big)+\frac{1}{2}\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,|\boldsymbol{u}_\epsilon|^2\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}\\&\qquad\qquad-\frac{1}{2}\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,(\boldsymbol{u}_\epsilon-\boldsymbol{u}_\sigma)^2\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}.\end{aligned}$$

Finally by the upwind approximation of the velocity $\boldsymbol{u}_\sigma$, the equation (2.30) yields

$$\begin{aligned}&\frac{d}{dt}\big(\frac{1}{2}h_{D_\sigma}\,|\boldsymbol{u}_\sigma|^2\big)+\frac{1}{2}\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}|\epsilon|\,|\boldsymbol{u}_\epsilon|^2\,\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon}+gh_{D_\sigma}\boldsymbol{u}_\sigma\cdot(\nabla h)_\sigma-(\nabla\pi)_\sigma\cdot\boldsymbol{u}_\sigma+\omega\Lambda_\sigma\cdot\boldsymbol{u}_\sigma^\perp\\&\qquad=-\frac{1}{2}\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma),\ \epsilon=\sigma|\sigma'}|\epsilon|\,(\boldsymbol{u}_{\sigma'}-\boldsymbol{u}_\sigma)^2\,(\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon})^-\le 0,\end{aligned}$$

since $(\boldsymbol{F}_\epsilon\cdot\boldsymbol{n}_{\sigma,\epsilon})^-$ is positive, which finishes the proof. □

For now up, we can write the semi-discrete mechanical energy balance which is the sum of the global potential and kinetic energies.

Lemma 2.5 (Semi-discrete mechanical energy inequality)**.** *A solution of* (2.26) *satisfies the following inequality:*

$$\frac{d}{dt}E\le 0,\quad \textit{with } E=\sum_{K\in\mathscr{M}}|K|\,\frac{1}{2}gh_K^2+\sum_{\sigma\in\mathscr{E}_{\rm int}}|D_\sigma|\,\frac{1}{2}h_{D_\sigma}|\boldsymbol{u}_\sigma|^2.$$

Proof. The semi-discrete energy E holds:

$$\frac{d}{dt}E\le R_1+R_2,$$

where the residual terms are:

$$\begin{aligned} R_1 &= -\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}(K)} |D_\sigma|\,\frac{1}{2}g\,(\nabla h)_\sigma\cdot\Lambda_\sigma - \sum_{\sigma\in\mathcal{E}_{\rm int}} |D_\sigma|\,\omega\,\boldsymbol{u}_\sigma^\perp\cdot\Lambda_\sigma \\ R_2 &= \sum_{\sigma\in\mathcal{E}_{\rm int}} |D_\sigma|\,(\nabla\pi)_\sigma\cdot\boldsymbol{u}_\sigma. \end{aligned}$$

Then it remains to show that these residuals are negative. Following Duran and Vila 2019, Lemma 5, one has

$$\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}(K)} |D_\sigma|\,\frac{1}{2}g\,(\nabla h)_\sigma\cdot\Lambda_\sigma = \sum_{\sigma\in\mathcal{E}_{\rm int}} |D_\sigma|\,g\,(\nabla h)_\sigma\cdot\Lambda_\sigma$$

Thanks to the definition of Λ_σ, the first term R_1 rewrites:

$$R_1 = -\sum_{\sigma\in\mathcal{E}_{\rm int}} |D_\sigma|\big[g\,(\nabla h)_\sigma + \omega\,\boldsymbol{u}_\sigma^\perp\big]\cdot\Lambda_\sigma \le 0.$$

The last term R_2 gives:

$$\begin{aligned} R_2 &= \frac{1}{2}\sum_{\sigma=K|L,\ \sigma\in\mathcal{E}_{\rm int}} |\sigma|\,(\nabla\pi)_\sigma\cdot\boldsymbol{u}_\sigma = -\frac{1}{2}\sum_{K\in\mathcal{M}} |K|\,\pi_K\,{\rm div}_K(\boldsymbol{u}) \\ &= -\frac{1}{2}\sum_{K\in\mathcal{M}} |K|\,\pi_K\left(\frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}\right)\le 0, \end{aligned}$$

thanks to the definition of π_K and the div-grad duality relationship which concludes the proof. □

Let us focus on the geostrophic equilibrium state that the linear version of the semi-discrete scheme (2.26) is expected to preserve.

Semi-discrete linear scheme – The linearized semi-discrete scheme of (2.26) reads:

$$\begin{aligned} &\frac{d}{dt}\,h_K + \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,h_0\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma} - \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,\Lambda_\sigma\cdot\boldsymbol{n}_{K,\sigma} = 0, \quad \forall\, K\in\mathcal{M}, \\ &\frac{d}{dt}\,\boldsymbol{u}_\sigma + g\,(\nabla h)_\sigma + \omega\,\boldsymbol{u}_\sigma^\perp = (\nabla\pi)_\sigma - \omega\,\Lambda_\sigma^\perp, \quad \forall\,\sigma\in\mathcal{E}_{int}, \end{aligned}$$

where the numerical diffusions Λ and π are given by:

$$\begin{aligned} \forall\,\sigma\in\mathcal{E}_{\rm int}, \quad \Lambda_\sigma &= \gamma\,\delta t\,(g\,(\nabla h)_\sigma + \omega\,\boldsymbol{u}_\sigma^\perp), \quad \gamma\ge 0, \\ \forall\, K\in\mathcal{M}, \quad \pi_K &= \nu\,\delta t\,g\,\frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}, \quad \nu\ge 0. \end{aligned}$$

This linear scheme preserves exactly the geostrophic equilibrium state as expected

since when we reach the equilibrium state (where the semi-discrete relationship $g\,(\nabla h)_\sigma + \omega\,\boldsymbol{u}_\sigma^\perp = 0$, holds) the numerical mass flux and pressure gradient corrections Λ_σ and π_K vanish as well as the normal velocity $\boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma}$ for all $\sigma \in \mathscr{E}_{\rm int}$ and $K \in \mathscr{M}$. Hence we get

$$\frac{d}{dt}\, h_K = 0, \quad \forall\, K \in \mathscr{M},$$
$$\frac{d}{dt}\, \boldsymbol{u}_\sigma = 0, \quad \forall\, \sigma \in \mathscr{E}_{int}.$$

In fact we can show that a semi-discrete mechanical energy derived from this semi-discrete linear scheme is decreasing following the same manipulations used in the non-linear case.

Below we complete the numerical resolution of the modified systems (2.24) and (2.25) by performing a time discretization of the semi-discrete unknowns h_K and $\boldsymbol{u}_\sigma$.

2.3.2 Stabilized staggered schemes

Owing a segregated time integration of the semi-discrete scheme (2.26), we get a fully discrete scheme for the non-linear equations (2.24) which reads:

$$\frac{1}{\delta t}\,(h_K^{n+1} - h_K^n) + \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,\boldsymbol{F}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma} = 0, \quad \forall\, K \in \mathscr{M}, \tag{2.33a}$$

$$\frac{1}{\delta t}\,(h_{D_\sigma}^{n+1}\boldsymbol{u}_\sigma^{n+1} - h_{D_\sigma}^n\,\boldsymbol{u}_\sigma^n) + \frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)} |\epsilon|\,\boldsymbol{u}_\epsilon^n\,\boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon} + g h_{D_\sigma}^{n+1}\,(\nabla h^{n+1})_\sigma - (\nabla \pi^n)_\sigma$$
$$= -\omega\,(h\boldsymbol{u}^\perp)_{D_\sigma}^{n,\theta} + \omega\,(\Lambda^\perp)_\sigma^n, \quad \forall\, \sigma \in \mathscr{E}_{int}, \tag{2.33b}$$

where the various discrete terms and operators according to the space discretization are defined in the semi-discrete scheme above and the Coriolis term $(h\boldsymbol{u}^\perp)_{D_\sigma}^{n,\theta}$ is given by (2.14).

Let us now assess the robustness and other preservation properties of this scheme.

2.3.2.1 Stability of the scheme

The positivity of the water height h^n is guaranteed by means of a suitable CFL control as stated in the following.

Proposition 2.4 (Preservation of the positivity)**.** *Let $(h_K^{n+1},\ \boldsymbol{u}_\sigma^{n+1})_{K\in\mathscr{M},\,\sigma\in\mathscr{E}}$ a discrete solution of the scheme (2.33) such that $h_K^n > 0$, for $n \in \{0,\cdots,N_t-1\}$ and for all $K \in \mathscr{M}$. Then $h_K^{n+1} > 0$, for all $K \in \mathscr{M}$ under the following CFL condition,*

$$\frac{\delta t}{\delta_{\mathscr{M}}}\left(|\boldsymbol{u}_\sigma^n| + \sqrt{\gamma}\,\sqrt{\omega\,|\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}| + g\,|(\nabla h^n)_\sigma \cdot \boldsymbol{n}_{K,\sigma}|}\right) \le \frac{\xi}{1+\xi}\,\frac{\max_{K\in\mathscr{M}}(h_K)}{\min_{K\in\mathscr{M}}(h_K^n)}. \tag{2.34}$$

where $\xi \in (0,1]$.

Following the same arguments developed in the previous section regarding the semi-implicit RT scheme (2.33), the current scheme preserves the (linear) geostrophic equilibrium state irrespectively of the meshing at hand whether uniform or not.

Proposition 2.5 (Preservation of the geostrophic equilibrium, non-linear scheme)**.** *Let* $(h_K^{n+1},\ \boldsymbol{u}_\sigma^{n+1})_{K\in\mathscr{M},\,\sigma\in\mathscr{E}}$, *for* $n \in \{0,\cdots,N_t-1\}$ *given by the scheme* (2.33) *such that* $g(\nabla h^n)_\sigma + \omega(\boldsymbol{u}_\sigma^n)^\perp = 0$. *Then* $\boldsymbol{u}^{n+1} = \boldsymbol{u}^n$ *and* $h^{n+1} = h^n$.

2.3.2.2 Linear stabilized RT scheme

Linearizing the current non-linear stabilized scheme around the state $(h_0, \boldsymbol{u}_0 = 0)$ we get the following linear scheme for the equations (2.24):

$$\frac{1}{\delta t}\,(h_K^{n+1} - h_K^n) + \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, h_0\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma} - \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, h_0\ \Lambda_\sigma^n\cdot\boldsymbol{n}_{K,\sigma} = 0,$$

$$\frac{1}{\delta t}\,(\boldsymbol{u}_\sigma^{n+1} - \boldsymbol{u}_\sigma^n) + g(\nabla h^{n+1})_\sigma - (\nabla\pi^n)_\sigma = -\omega\left(\boldsymbol{u}_\sigma^n + \boldsymbol{\theta}\ (\boldsymbol{u}_\sigma^{n+1} - \boldsymbol{u}_\sigma^n)\right)^\perp + \omega\ (\Lambda_\sigma^n)^\perp,$$

with

$$\begin{aligned}
\forall\ K\in\mathscr{M},\quad \pi_K^n &= \nu\,\delta t\ g\ \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma},\quad \nu\geq 0,\\
\forall\ \sigma\in\mathscr{E}_{\text{int}},\quad \Lambda_\sigma^n &= \gamma\,\delta t\left(g\ (\nabla h^n)_\sigma + \omega\ (\boldsymbol{u}_\sigma^n)^\perp\right),\quad \gamma\geq 0,
\end{aligned}$$

where the quantity $\left(\boldsymbol{u}_\sigma^n + \boldsymbol{\theta}\ (\boldsymbol{u}_\sigma^{n+1} - \boldsymbol{u}_\sigma^n)\right)^\perp$ is given by (2.23).

As in the semi-discrete linear scheme presented above, the geostrophic equilibrium state is preserved by the present scheme. Indeed if $g\ (\nabla h^n)_\sigma + \omega\ (\boldsymbol{u}_\sigma^n)^\perp = 0,\quad \forall\ \sigma\in\mathscr{E}_{\text{int}}$, then

$$g\ (\nabla h^{n+1})_\sigma + \omega\ (\boldsymbol{u}_\sigma^{n+1})^\perp = 0,\quad \forall\ \sigma\in\mathscr{E}_{\text{int}},\ \forall\ n\in\{0,\cdots,N_t-1\}.$$

Noting also that omitting the correction terms i.e. $\nu = \gamma = 0$, the above linear scheme degenerates to the semi-implicit linear RT scheme (2.22).

However a linear stability analysis should be performed in order to control the stabilization parameters which remains yet to be define. This step is very important and needs a thorough study that is in fact more complex to tackle. The main difficulty lying on the staggered arrangement of the unknowns involving five discrete equations. To get an idea of this, we refer to the papers Couderc, Duran, and Vila 2017; Audusse, Do, Omnes, et al. 2018 where this type of analysis is made in the framework of collocated schemes. For the sake of simplicity, the diffusion coefficients γ and ν are chosen on the basis of numerical experiments for the following numerical investigations.

2.4 Numerical simulations

Here we present some experiments to validate the numerical theory presented in the previous sections, starting by highlighting the efficiency of the linear schemes. Attention is particularly paid on the staggered discretization techniques to show the good features of the MAC schemes compared to the RT schemes in both linear and non-linear cases.

2.4.1 Numerical results for the linear schemes

For the sake of clarity, the linear schemes we compare here-below are rewritten in the standard Cartesian notation $(i\,j)$ as follows:

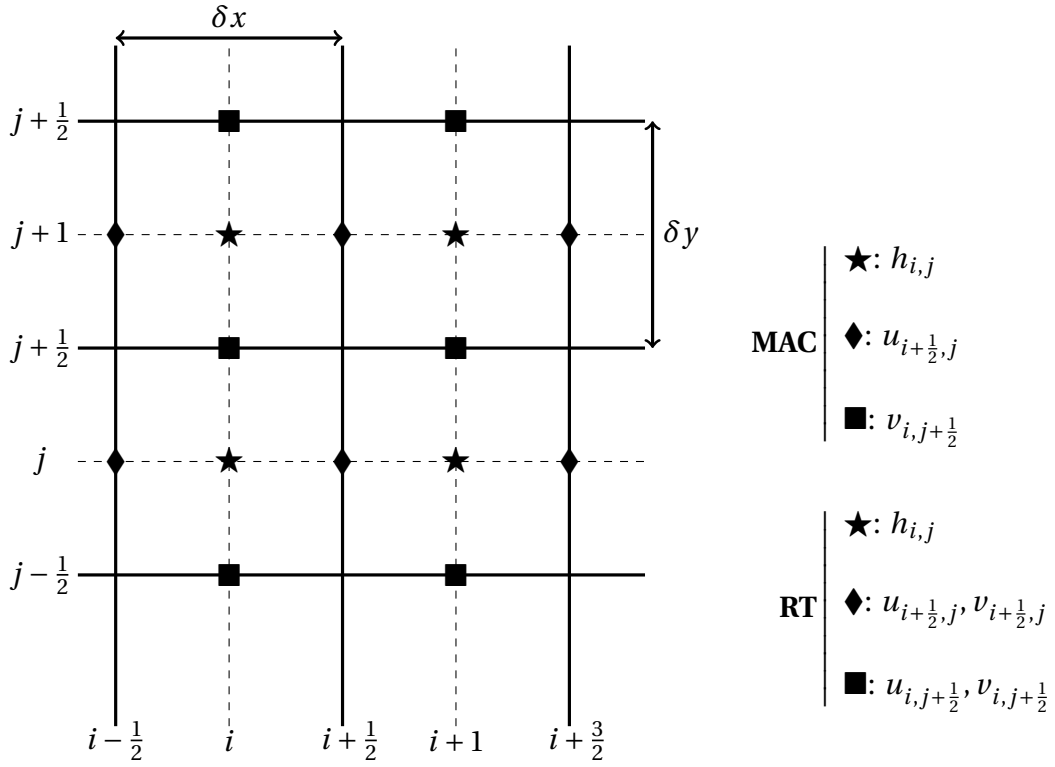


Figure 2.5 – Uniform Cartesian grid.

Linear semi-implicit MAC scheme:

$$\frac{1}{\delta t}\,(h_{i,j}^{n+1}-h_{i,j}^{n})+\frac{h_0}{\delta x}\,(u_{i+\frac{1}{2},j}^{n}-u_{i-\frac{1}{2},j}^{n})+\frac{h_0}{\delta y}\,(v_{i,j+\frac{1}{2}}^{n}-v_{i,j-\frac{1}{2}}^{n})=0,$$

$$\frac{1}{\delta t}\,(u_{i+\frac{1}{2},j}^{n+1}-u_{i+\frac{1}{2},j}^{n})+g\frac{1}{\delta x}(h_{i+1,j}^{n+1}-h_{i,j}^{n+1})=\omega\,\frac{1}{4}\big(v_{i,j+\frac{1}{2}}^{n}+v_{i+1,j+\frac{1}{2}}^{n}+v_{i,j-\frac{1}{2}}^{n}+v_{i+1,j-\frac{1}{2}}^{n}\big),$$

$$\frac{1}{\delta t}\,(v_{i,j+\frac{1}{2}}^{n+1}-v_{i,j+\frac{1}{2}}^{n})+g\frac{1}{\delta y}(h_{i,j+1}^{n+1}-h_{i,j}^{n+1})=-\omega\,\frac{1}{4}\big(u_{i+\frac{1}{2},j}^{n+1}+u_{i+\frac{1}{2},j+1}^{n+1}+u_{i-\frac{1}{2},j}^{n+1}+u_{i-\frac{1}{2},j+1}^{n+1}\big).$$

Linear semi-implicit RT scheme (LRS):

$$
\begin{aligned}
&\frac{1}{\delta t}\,(h^{n+1}_{i,j}-h^{n}_{i,j})+\frac{h_0}{\delta x}\,(u^{n}_{i+\frac12,j}-u^{n}_{i-\frac12,j})+\frac{h_0}{\delta y}\,(v^{n}_{i,j+\frac12}-v^{n}_{i,j-\frac12})=0,\\
&\frac{1}{\delta t}\,(u^{n+1}_{i+\frac12,j}-u^{n}_{i+\frac12,j})+g\frac{2}{\delta x}(h^{n+1}_{i+1,j}-h^{n+1}_{i,j})=\omega\; v^{n}_{i+\frac12,j}+\omega\theta_2(v^{n+1}_{i+\frac12,j}-v^{n}_{i+\frac12,j}),\\
&\frac{1}{\delta t}\,(v^{n+1}_{i+\frac12,j}-v^{n}_{i+\frac12,j})=-\omega\; u^{n}_{i+\frac12,j}-\omega\theta_1(u^{n+1}_{i+\frac12,j}-u^{n}_{i+\frac12,j}),\\
&\frac{1}{\delta t}\,(u^{n+1}_{i,j+\frac12}-u^{n}_{i,j+\frac12})=\omega\; v^{n}_{i,j+\frac12}+\omega\theta_4(v^{n+1}_{i,j+\frac12}-v^{n}_{i,j+\frac12}),\\
&\frac{1}{\delta t}\,(v^{n+1}_{i,j+\frac12}-v^{n}_{i,j+\frac12})+g\frac{2}{\delta y}(h^{n+1}_{i,j+1}-h^{n+1}_{i,j})=-\omega\; u^{n}_{i,j+\frac12}-\omega\theta_3(u^{n+1}_{i,j+\frac12}-u^{n}_{i,j+\frac12}).
\end{aligned}
$$

Linear stabilized semi-implicit scheme (LSS):

$$
\begin{aligned}
&\frac{1}{\delta t}\,(h^{n+1}_{i,j}-h^{n}_{i,j})+\frac{h_0}{\delta x}\,(F^{n}_{i+\frac12,j}-F^{n}_{i-\frac12,j})+\frac{h_0}{\delta y}\,(F^{n}_{i,j+\frac12}-F^{n}_{i,j-\frac12})=0,\\
&\frac{1}{\delta t}\,(u^{n+1}_{i+\frac12,j}-u^{n}_{i-\frac12,j})+g\;\eth(h^{n+1})_{i+\frac12,j}-\eth(\pi^{n})_{i+\frac12,j}=\omega\; v^{n}_{i+\frac12,j}\\
&\qquad\qquad+\omega\theta_2(v^{n+1}_{i+\frac12,j}-v^{n}_{i+\frac12,j})-\omega^2\,\gamma\delta t\; u^{n}_{i+\frac12,j},\\
&\frac{1}{\delta t}\,(v^{n+1}_{i+\frac12,j}-v^{n}_{i-\frac12,j})=-\omega\; u^{n}_{i+\frac12,j}-\omega\theta_1(u^{n+1}_{i+\frac12,j}-u^{n}_{i+\frac12,j})+\omega\;\gamma\delta t\;(g\;\eth(h^{n})_{i+\frac12,j}-\omega v^{n}_{i+\frac12,j}),\\
&\frac{1}{\delta t}\,(u^{n+1}_{i,j+\frac12}-u^{n}_{i,j-\frac12})=\omega\; v^{n}_{i,j+\frac12}+\omega\theta_4(v^{n+1}_{i,j+\frac12}-v^{n}_{i,j+\frac12})-\omega\;\gamma\delta t\;(g\;\eth(h^{n})_{i,j+\frac12}+\omega u^{n}_{i,j+\frac12}),\\
&\frac{1}{\delta t}\,(v^{n+1}_{i,j+\frac12}-v^{n}_{i,j-\frac12})+g\;\eth(h^{n+1})_{i,j+\frac12}-\eth(\pi^{n})_{i,j+\frac12}=-\omega\; u^{n}_{i,j+\frac12}\\
&\qquad\qquad-\omega\theta_3(u^{n+1}_{i,j+\frac12}-u^{n}_{i,j+\frac12})-\omega^2\gamma\delta t\; v^{n+1}_{i,j+\frac12},
\end{aligned}
$$

with

$$
\begin{aligned}
F^{n}_{i+\frac12,j} &= u^{n}_{i+\frac12,j}-\gamma\delta t\;(g\;\eth(h^{n})_{i+\frac12,j}-\omega\; v^{n}_{i+\frac12,j}), \qquad \eth(h^{n})_{i+\frac12,j}=\frac{2}{\delta x}(h^{n}_{i+1,j}-h^{n}_{i,j}),\\
F^{n}_{i,j+\frac12} &= v^{n}_{i,j+\frac12}-\gamma\delta t\;(g\;\eth(h^{n})_{i,j+\frac12}+\omega\; u^{n}_{i,j+\frac12}), \qquad \eth(h^{n})_{i,j+\frac12}=\frac{2}{\delta y}(h^{n}_{i,j+1}-h^{n}_{i,j}),\\
\pi^{n}_{i,j} &= \nu\delta t g\Big(\frac{2}{\delta x}\,(u^{n}_{i+\frac12,j}-u^{n}_{i-\frac12,j})+\frac{2}{\delta y}\,(v^{n}_{i+\frac12,j}-v^{n}_{i-\frac12,j})\Big). \qquad (2.36)
\end{aligned}
$$

Omitting the correction terms involved in the LSS scheme, this latter matches the LRS scheme; hence we unify the denomination LSS for these schemes. For the sake of simplicity we take $\theta_1=\theta_3=1$ and $\theta_2=\theta_4=0$ for all numerical results presented in the sequel.

We are going to compare these linear staggered schemes against a Godunov type

scheme considered by Audusse, Do, Omnes, et al. 2018 which reads:

Linear Godunov type scheme:

$$\frac{1}{\delta t}(h_{i,j}^{n+1}-h_{i,j}^{n})+\frac{h_0}{\delta x}(F_{i+\frac{1}{2},j}^{n}-F_{i-\frac{1}{2},j}^{n})+\frac{h_0}{\delta y}(F_{i,j+\frac{1}{2}}^{n}-F_{i,j-\frac{1}{2}}^{n})=0$$
$$\frac{1}{\delta t}(u_{i,j}^{n+1}-u_{i,j}^{n})+g\,\frac{h_{i+1,j}^{n}-h_{i-1,j}^{n}}{2\delta x}-g\frac{\zeta_u}{2}\frac{u_{i+1,j}^{n}-2u_{i-1,j}^{n}+u_{i-1,j}^{n}}{\delta x}=\omega\, v_{i,j}^{n},$$
$$\frac{1}{\delta t}(v_{i,j}^{n+1}-v_{i,j}^{n})+g\,\frac{h_{i,j+1}^{n}-h_{i,j-1}^{n}}{2\delta y}-g\frac{\zeta_v}{2}\frac{u_{i+1,j}^{n}-2u_{i-1,j}^{n}+u_{i-1,j}^{n}}{\delta x}=-\omega\, u_{i,j}^{n+1},$$

with

$$\begin{aligned}F_{i+\frac{1}{2},j}^{n} &= \frac{1}{2}(u_{i+1,j}^{n}+u_{i,j}^{n})-\zeta_h\,\frac{1}{2}(h_{i+1,j}^{n}-h_{i,j}^{n}),\\ F_{i,j+\frac{1}{2}}^{n} &= \frac{1}{2}(v_{i,j+1}^{n}+v_{i,j}^{n})-\zeta_h\,\frac{1}{2}(h_{i,j+1}^{n}-h_{i,j}^{n}).\end{aligned}$$

In the sequel the parameter ζ_h is fixed to 1 while ζ_u and ζ_v are set to zero.

2.4.1.1 Gravity wave test case

In this first test, we compare the above linear schemes with an exact solution of equations (2.4) without Coriolis force effects *i.e.* $\omega = 0$. The exact solution considered here consists in a gravity wave defined by:

$$\begin{aligned}h(\boldsymbol{x},t) &= h_0 + H\,\sin(\mathbf{k}\cdot\boldsymbol{x}-\omega_0 t),\\ u_1(\boldsymbol{x},t) &= U\,\sin(\mathbf{k}\cdot\boldsymbol{x}-\omega_0 t)\\ u_2(\boldsymbol{x},t) &= V\,\sin(\mathbf{k}\cdot\boldsymbol{x}-\omega_0 t)\end{aligned}$$

where $\mathbf{k}\cdot\boldsymbol{x} = k_1 x + k_2 y$ with $k_1 = \frac{2\pi}{L_x}$, $k_2 = \frac{2\pi}{L_y}$, $\omega_0 = \sqrt{g h_0 (k_1^2+k_2^2)}$, $U = gH\frac{k_1}{\omega_0}$, $V = gH\frac{k_2}{\omega_0}$, $H = 0.01$, $h_0 = 1$, $g = 9.81$. We consider the domain $(-L_x, L_x)\times(-L_y, L_y)$ with $Lx = L_y = 25$ and the computations are running for 200×200 cells. Then only periodic boundary conditions are used in the following test cases. The Figure 2.6 shows a horizontal cut off the initial water height.

Then the goal is to highlight the inaccuracy of the linear staggered schemes based on the RT elements for the linear shallow water equations without source term. For this purpose we assess the behavior of the numerical waves computed from the LSS, MAC and Godunov type schemes during one period of time $T_p = \frac{2\pi}{\omega}$ which is precisely the period of the analytic solution. The results are plotted in Figure 3.4 for varying time.

One should observe that the MAC and Godunov numerical wave's are travelling in the same speed with the analytic wave with a smaller amplitude for the Godunov wave's. While the numerical wave produced by the LSS scheme is moving faster than the one from the MAC scheme. This latter has its own mode of propagation

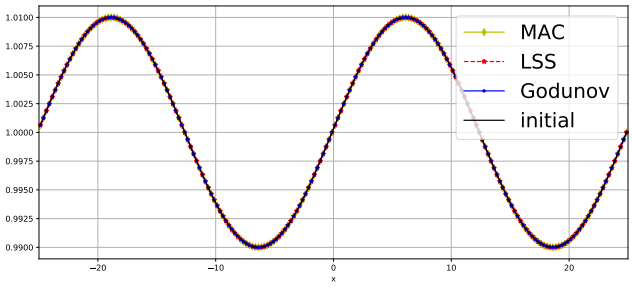
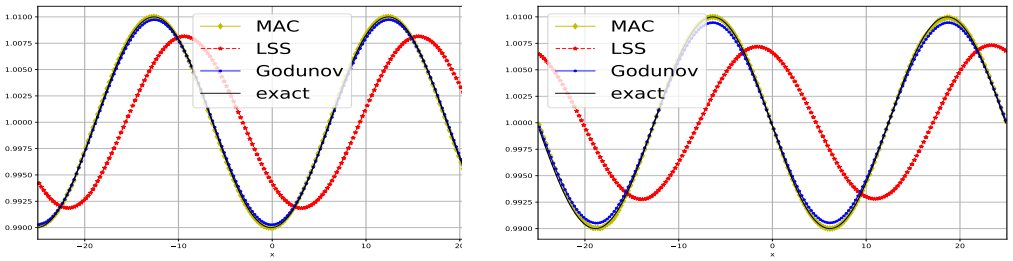
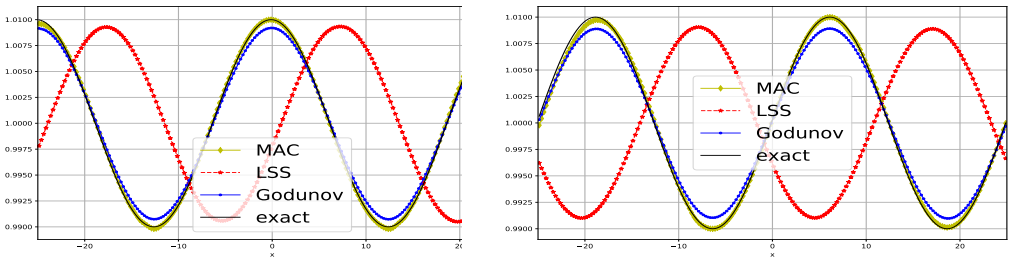


Figure 2.6 – Gravity wave: profile of the water height at time $T = 0$



(a) $T = T_p/4$ (b) $T = T_p/2$



(c) $T = 3T_p/4$ (d) $T = T_p$

Figure 2.7 – Gravity wave – Inaccuracy of the linear RT scheme: horizontal cut off of the water height during one period of time.

different from other schemes with a short period of time which is approximately close to $5.65Tp/8$ as shown in the Figure 2.8 where we compare the numerical solution at $T = 5.65Tp/8$ with the initial condition.This behavior is linked to the fact that the discrete pressure gradient defined by the LSS scheme is twice compared to the MAC scheme's. This latter gives satisfactory results compared to the Godunov type scheme while the LSS one remains inaccuracy for this present test case. This allows to say that the MAC scheme is a better approximation for the linear shallow water equation than the RT scheme.

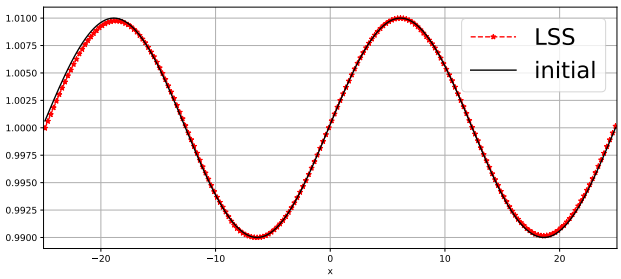


Figure 2.8 – Gravity wave: Horizontal cut off the water height computed from the LSS scheme at time $T = 5.65Tp/8$

2.4.1.2 Well balance test case

The aim of this test is to investigate the preservation of the geostrophic equilibrium state $g\nabla h + \omega \boldsymbol{u}^{\perp}$ in the discrete setting. The computational domain is the squared $[-L, L]^2$ with $L = 0.5$. This benchmark is available in Audusse, Do, Omnes, et al. 2018 and consists in a stationary vortex solution of the linear shallow water equations (2.4). The initial water height is defined as a Gaussian function by

$$h_0(x, y) = 1 - \exp\big(-(\frac{3}{L})^2(x^2 + y^2)\big)$$

Then the initialization of the components of the velocity u_0 and v_0 are deduced from the continuous geostrophic equilibrium state and are given by:

$$\begin{cases} u_0(x, y) = -\frac{g}{\omega}(\frac{3}{L})^2\, 2y \exp\big(-(\frac{3}{L})^2(x^2 + y^2)\big), \\ v_0(x, y) = \frac{g}{\omega}(\frac{3}{L})^2\, 2x \exp\big(-(\frac{3}{L})^2(x^2 + y^2)\big). \end{cases} \tag{2.37}$$

Since the unknowns are assessed in staggered way, the initialization of the velocity components are defined by (2.37) respectively on the primal and dual meshes. The time step is fixed to $\delta t = \delta x/5$ and the computations use 50×50 cells. Below we present the obtained results at time $T = 0.015$ and $T = 5$.
Obviously the MAC scheme gives good results compared to the other schemes and its numerical solution is in good agreement with the initial solution. This better behavior encourages the reconstruction technique undertaken in the linear MAC scheme to approximate the rotation velocity $\boldsymbol{u}^{\perp}$. However from these illustrations we can see that the LSS scheme is no longer able to maintain the equilibrium geostrophic steady state which contradicts the numerical theory presented in the previous sections. We think that the problem is lying in the initialization step of the components of the velocity for the RT scheme. Indeed by means of a suitable initialization of the velocity field this scheme holds the preservation of this equilibrium state. To do that, the terms u^0 and v^0 are initialized such a way to mimic the continuous equilibrium equation, more

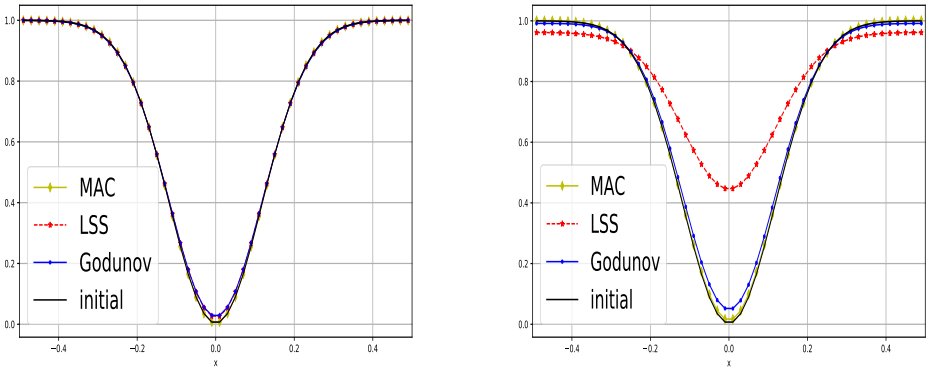


Figure 2.9 – Well balance test – Horizontal cut off the watert height obtained by the differents schemes at time $T = 0.015$ on the left and $T = 5$ on the right.

precisely both quantities are computed from the discrete gradient of h^0 as follows:

$$\begin{cases} u^0_{i+\frac{1}{2},j} = 0, \\ v^0_{i+\frac{1}{2},j} = -\frac{g}{\omega}\,\frac{2}{\delta x}(h^0_{i+1,j} - h^0_{i,j}), \end{cases} \quad \text{and} \quad \begin{cases} u^0_{i,j+\frac{1}{2}} = -\frac{g}{\omega}\,\frac{2}{\delta y}(h^0_{i,j+1} - h^0_{i,j}), \\ v^n_{i,j+\frac{1}{2}} = 0; \end{cases} \tag{2.38}$$

where we recall here:

$$h^0_{i,j} = \frac{1}{\delta x\,\delta y}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} h_0(x,y)\,dy\,dx.$$

This reconstruction way of the velocity ensures the preservation of the linear equilibrium at hand at the initialization level which was not be the case in the previous computations. Then implementing (2.38) in the RT scheme yields the following results obtained at time $T = 5$ and $T = 50$:
Figure 2.10 shows that after a computational time $T = 50$ the numerical solution of the MAC scheme remains close to the initial state unlike to the Godunov type scheme. This latter is less accuracy and completely loses the well-balanced behavior since its numerical solution rapidly moves away from the initial state as shown in Figure 2.10 on the right. We note here the good agreement of the LSS scheme that is none other than the RT scheme since no stabilization effects are added. The fact that LSS scheme matches perfectly the analytic solution is strongly depending on the way the velocity field is initialized (2.38). In other words, in order to deal with the preservation of the linear geostrophic equilibrium state a carefully initialization of the velocity is required for the LSS scheme.

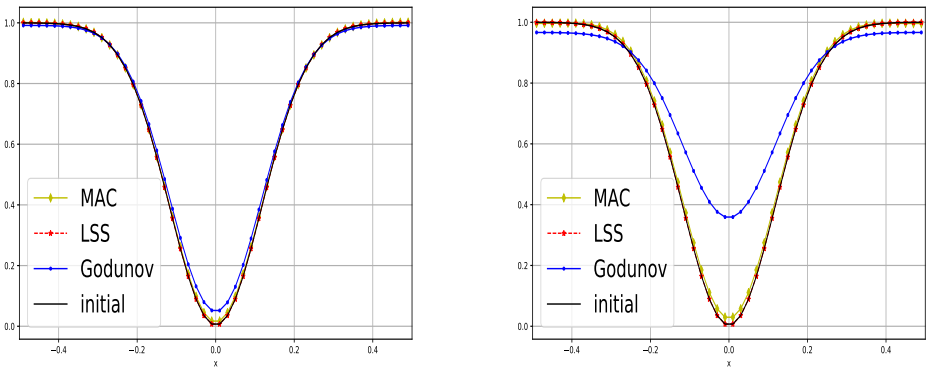


Figure 2.10 – Well balance test – Preservation of the geostrophic equilibrium state: horizontal cut off the water height at time $T = 5$ on the left and $T = 50$ on the right

2.4.1.3 Circular dam break test case

The present test case is also proposed by Audusse, Do, Omnes, et al. 2018 to highlight the performance of the schemes against discontinuity data. The problem is posed on the squared $[-5,5]^2$ and the initial water height is given by:

$$h_0(x,y) = \begin{cases} 2, & \text{if } x^2+y^2 \le 1, \\ 1, & \text{otherwise}, \end{cases}$$

while the velocity field is initialized to zero. After a short time of simulation $T = 0.25$ strong instabilities occur for both MAC and non stabilized LSS schemes, see Figure 2.11. Then adding a correction term in the discrete velocity equation consisting in $\pi_{i,j}$ given by (2.36) with a slight stabilization coefficient ν equals to 2, allows to stabilize the LSS and MAC schemes as shown in Figure 2.11 on the right.

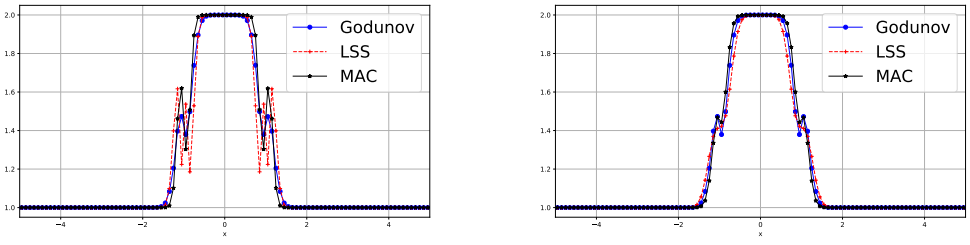
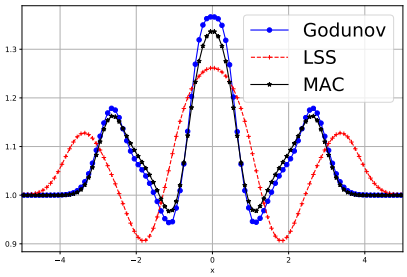
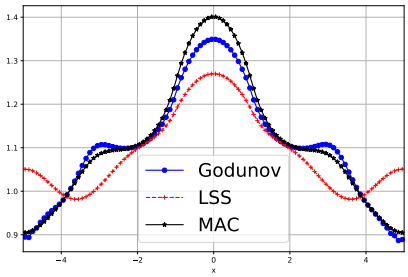


Figure 2.11 – Circular dam break: Horizontal cut off the water height at time $T = 0.25$: left whihout correction, right with correction.

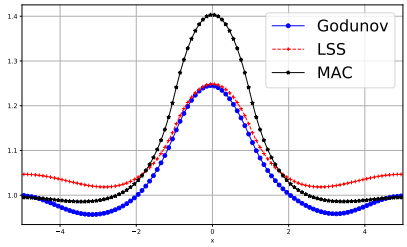
This manner of stabilization improves considerably the stability of the linear LSS and MAC schemes. Through the Figures 2.12c, 2.12d, 2.12e, 2.12f we observe that



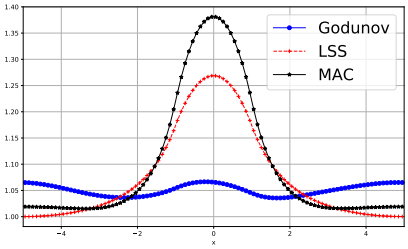
(a) $T = 2$



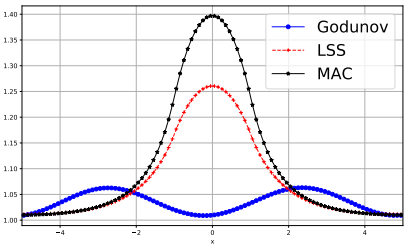
(b) $T = 5$



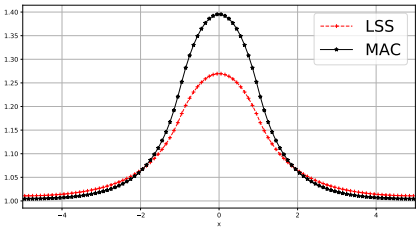
(c) $T = 25$



(d) $T = 50$



(e) $T = 100$



(f) $T = 200$

Figure 2.12 – Circular dam break test: Long time behavior of the schemes for varying simulation time

the corrected MAC scheme seems to reach a specific steady state for a large time of simulation $T \geq 100$. Finally, it should be noted that the difference in altitude between the solutions of the MAC and RT schemes is mainly due to the discretization of the pressure gradient as mentioned in Remark 2.1. Indeed the discrete pressure gradient for the RT scheme is a vector for which one of its components is zero while for the MAC scheme this quantity yields a single component defined on a dual cell following the normal or tangential directions.

2.4.2 Numerical results for the non-linear schemes

Here we turn to the non-linear staggered schemes that are the semi-implicit upwind MAC and RT schemes on one hand and the stabilized staggered (NSS) schemes on the other hand presented in Sections 2.2 and 2.3 respectively. As far as the following computations are posed only on squared domains, the NSS scheme consists in a centered stabilized numerical mass flux and an upwind scheme for the momentum flux. Before presenting the numerical results, we start by describing these schemes in the formally Cartesian basis (i, j).

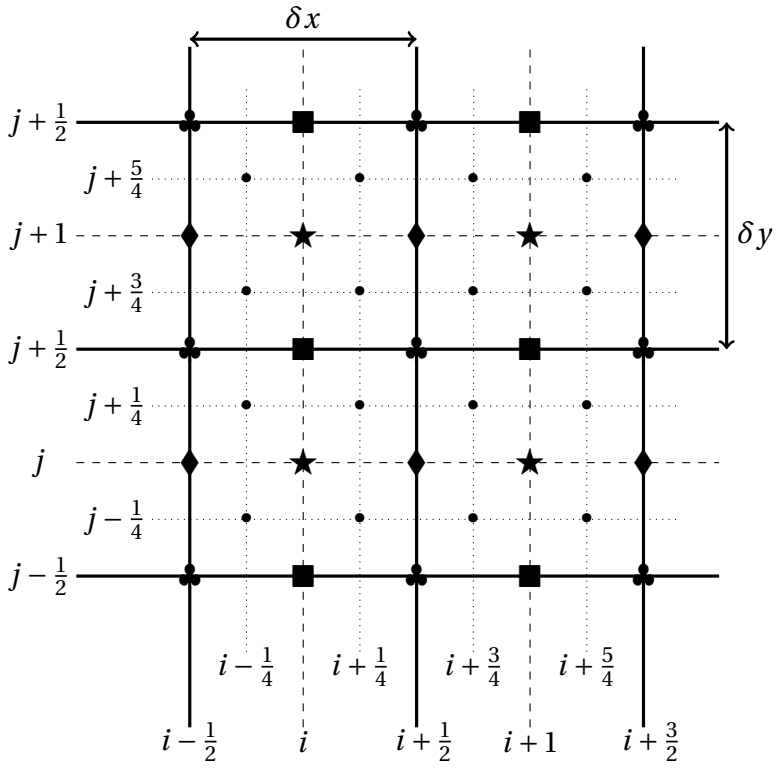


Figure 2.13 – Uniform Cartesian grid.

MAC

★: $h_{i,j}$

♦: $u_{i+\frac{1}{2},j}$

■: $v_{i,j+\frac{1}{2}}$

♣: $F_{i+\frac{1}{2},j+\frac{1}{2}}$

RT

★: $h_{i,j}$

♦: $u_{i+\frac{1}{2},j}, v_{i+\frac{1}{2},j}$

■: $u_{i,j+\frac{1}{2}}, v_{i,j+\frac{1}{2}}$

•: $F_{i+\frac{1}{4},j+\frac{1}{4}}$

Nonlinear semi-implicit MAC scheme:

$$\frac{1}{\delta t}\,(h_{i,j}^{n+1}-h_{i,j}^{n})+\frac{1}{\delta x}\,(F_{i+\frac{1}{2},j}^{n}-F_{i-\frac{1}{2},j}^{n})+\frac{1}{\delta y}\,(F_{i,j+\frac{1}{2}}^{n}-F_{i,j-\frac{1}{2}}^{n})=0,$$

$$\frac{1}{\delta t}\left((hu)_{i+\frac{1}{2},j}^{n+1}-(hu)_{i+\frac{1}{2},j}^{n}\right)+\frac{1}{\delta x}\left((uF)_{i+1,j}^{n}-(uF)_{i,j}^{n}\right)+\frac{1}{\delta y}\left((uF)_{i+\frac{1}{2},j+\frac{1}{2}}^{n}-(uF)_{i+\frac{1}{2},j-\frac{1}{2}}^{n}\right)$$
$$+\,g\,h_{i+\frac{1}{2},j}^{n+1}\,\eth(h^{n+1})_{i+\frac{1}{2},j}=\omega\,\frac{1}{4}h_{i+\frac{1}{2},j}^{n+1}\left(v_{i,j+\frac{1}{2}}^{n}+v_{i+1,j+\frac{1}{2}}^{n}+v_{i,j-\frac{1}{2}}^{n}+v_{i+1,j-\frac{1}{2}}^{n}\right),$$

$$\frac{1}{\delta t}\left((hv)_{i,j+\frac{1}{2}}^{n+1}-(hv)_{i,j+\frac{1}{2}}^{n}\right)+\frac{1}{\delta x}\left((vF)_{i,j+1}^{n}-(vF)_{i,j}^{n}\right)+\frac{1}{\delta y}\left((vF)_{i+\frac{1}{2},j+\frac{1}{2}}^{n}-(vF)_{i-\frac{1}{2},j+\frac{1}{2}}^{n}\right)\Big)$$
$$+\,g\,h_{i,j+\frac{1}{2}}^{n+1}\,\eth(h^{n+1})_{i,j+\frac{1}{2}}=-\omega\,\frac{1}{4}h_{i,j+\frac{1}{2}}^{n+1}\left(u_{i+\frac{1}{2},j}^{n+1}+u_{i+\frac{1}{2},j+1}^{n+1}+u_{i-\frac{1}{2},j}^{n+1}+u_{i-\frac{1}{2},j+1}^{n+1}\right),$$

with

$$\begin{aligned}
F_{i+\frac{1}{2},j}^{n} &= h_{i,j}^{n}(u_{i+\frac{1}{2},j}^{n})^{+}-h_{i+1,j}^{n}(u_{i+\frac{1}{2},j}^{n})^{-},\\
F_{i,j+\frac{1}{2}}^{n} &= h_{i,j}^{n}(v_{i,j+\frac{1}{2}}^{n})^{+}-h_{i,j+1}^{n}(v_{i,j+\frac{1}{2}}^{n})^{-},\\
(hu)_{i+\frac{1}{2},j}^{n} &= \frac{1}{2}(h_{i,j}^{n}+h_{i+1,j}^{n})u_{i+\frac{1}{2},j}^{n},\\
(hu)_{i,j+\frac{1}{2}}^{n} &= \frac{1}{2}(h_{i,j}^{n}+h_{i,j+1}^{n})u_{i,j+\frac{1}{2}}^{n},\\
\eth(h^{n})_{i+\frac{1}{2},j} &= \frac{1}{\delta x}(h_{i+1,j}^{n}-h_{i,j}^{n}),\\
\eth(h^{n})_{i,j+\frac{1}{2}} &= \frac{1}{\delta y}(h_{i,j+1}^{n}-h_{i,j}^{n}),
\end{aligned}$$

and the dual fluxes are given by:

$$\begin{aligned}
(uF)_{i,j}^{n} &= u_{i-\frac{1}{2},j}^{n}\,(F_{i,j}^{n})^{+}-u_{i,j+\frac{1}{2}}^{n}\,(F_{i+\frac{1}{2},j+\frac{1}{2}}^{n})^{-},\\
(vF)_{i,j}^{n} &= v_{i,j-\frac{1}{2}}^{n}\,(F_{i,j}^{n})^{+}-v_{i,j+\frac{1}{2}}^{n}\,(F_{i,j}^{n})^{-},\\
(uF)_{i+\frac{1}{2},j+\frac{1}{2}}^{n} &= u_{i+\frac{1}{2},j}^{n}\,(F_{i+\frac{1}{2},j+\frac{1}{2}}^{n})^{+}-u_{i+\frac{1}{2},j+1}^{n}\,(F_{i+\frac{1}{2},j+\frac{1}{2}}^{n})^{-},\\
(uF)_{i+\frac{1}{2},j+\frac{1}{2}}^{n} &= v_{i,j+\frac{1}{2}}^{n}\,(F_{i+\frac{1}{2},j+\frac{1}{2}}^{n})^{+}-v_{i+1,j+\frac{1}{2}}^{n}\,(F_{i+\frac{1}{2},j+\frac{1}{2}}^{n})^{-},
\end{aligned}$$

where we recall that the notations $()^{+}$ and $()^{-}$ stand for the positive and negative parts respectively defined by $(a)^{+}=\max(a,0)$ and $(a)^{-}=-\min(a,0)$ for any $a\in\mathbb{R}$.

Nonlinear semi-implicit RT scheme:

$$\frac{1}{\delta t}\,(h_{i,j}^{n+1}-h_{i,j}^{n})+\frac{1}{\delta x}\,(F_{i+\frac{1}{2},j}^{n}-F_{i-\frac{1}{2},j}^{n})+\frac{1}{\delta y}\,(F_{i,j+\frac{1}{2}}^{n}-F_{i,j-\frac{1}{2}}^{n})=0,$$

$$\frac{1}{\delta t}\left((hu)_{i+\frac{1}{2},j}^{n+1}-(hu)_{i+\frac{1}{2},j}^{n}\right)+\frac{2}{\delta x}\left((uF)_{i+\frac{3}{4},j-\frac{1}{4}}^{n}+(uF)_{i+\frac{1}{4},j+\frac{1}{4}}^{n}-(uF)_{i+\frac{3}{4},j+\frac{3}{4}}^{n}-(uF)_{i+\frac{1}{4},j-\frac{1}{4}}^{n}\right)$$
$$+\,g h_{i+\frac{1}{2},j}^{n+1}\,\eth(h^{n+1})_{i+\frac{1}{2},j}=\omega\,(hv)_{i+\frac{1}{2},j}^{n},$$

$$\frac{1}{\delta t}\left((hv)_{i+\frac{1}{2},j}^{n+1}-(hv)_{i+\frac{1}{2},j}^{n}\right)+\frac{2}{\delta x}\left((vF)_{i+\frac{3}{4},j-\frac{1}{4}}^{n}+(vF)_{i+\frac{1}{4},j+\frac{1}{4}}^{n}-(vF)_{i+\frac{3}{4},j+\frac{3}{4}}^{n}-(vF)_{i+\frac{1}{4},j-\frac{1}{4}}^{n}\right)$$
$$=-\omega(hu)_{i+\frac{1}{2},j}^{n+1},$$

$$\frac{1}{\delta t}\left((hu)_{i,j+\frac{1}{2}}^{n+1}-(hu)_{i,j+\frac{1}{2}}^{n}\right)+\frac{2}{\delta y}\left(-(uF)_{i-\frac{1}{4},j+\frac{3}{4}}^{n}-(uF)_{i+\frac{1}{4},j+\frac{1}{4}}^{n}+(uF)_{i+\frac{1}{4},j+\frac{3}{4}}^{n}+(uF)_{i-\frac{1}{4},j+\frac{1}{4}}^{n}\right)$$
$$=\omega\,(hv)_{i,j+\frac{1}{2}}^{n},$$

$$\frac{1}{\delta t}\left((hv)_{i,j+\frac{1}{2}}^{n+1}-(hv)_{i,j-\frac{1}{2}}^{n}\right)+\frac{2}{\delta y}\left(-(vF)_{i-\frac{1}{4},j+\frac{3}{4}}^{n}-(vF)_{i+\frac{1}{4},j+\frac{1}{4}}^{n}+(vF)_{i+\frac{1}{4},j+\frac{3}{4}}^{n}+(vF)_{i-\frac{1}{4},j+\frac{1}{4}}^{n}\right)$$
$$+\,g h_{i,j+\frac{1}{2}}^{n+1}\,\eth(h^{n+1})_{i,j+\frac{1}{2}}=-\omega(hu)_{i,j+\frac{1}{2}}^{n+1},$$

with

$$\begin{aligned}
F_{i+\frac{1}{2},j}^{n} &= (hu)_{i+\frac{1}{2},j}^{n}, & \eth(h^{n})_{i+\frac{1}{2},j}&=\frac{2}{\delta x}(h_{i+1,j}^{n}-h_{i,j}^{n}),\\
F_{i,j+\frac{1}{2}}^{n} &= (hv)_{i,j+\frac{1}{2}}^{n}, & \eth(h^{n})_{i,j+\frac{1}{2}}&=\frac{2}{\delta y}(h_{i,j+1}^{n}-h_{i,j}^{n}).
\end{aligned}$$

and the dual fluxes are given by:

$$\begin{aligned}
(uF)_{i+\frac{1}{4},j+\frac{1}{4}}^{n} &= u_{i+\frac{1}{2},j}^{n}\,(F_{i+\frac{1}{4},j+\frac{1}{4}}^{n})^{+}-u_{i,j+\frac{1}{2}}^{n}\,(F_{i+\frac{1}{4},j+\frac{1}{4}}^{n})^{-},\\
(uF)_{i+\frac{1}{4},j-\frac{1}{4}}^{n} &= u_{i,j-\frac{1}{2},j}^{n}\,(F_{i+\frac{1}{4},j-\frac{1}{4}}^{n})^{+}-u_{i+\frac{1}{2},j}^{n}\,(F_{i+\frac{1}{4},j-\frac{1}{4}}^{n})^{-},\\
(uF)_{i-\frac{1}{4},j-\frac{1}{4}}^{n} &= u_{i-\frac{1}{2},j}^{n}\,(F_{i-\frac{1}{4},j-\frac{1}{4}}^{n})^{+}-u_{i,j-\frac{1}{2}}^{n}\,(F_{i-\frac{1}{4},j-\frac{1}{4}}^{n})^{-},\\
(uF)_{i-\frac{1}{4},j+\frac{1}{4}}^{n} &= u_{i,j+\frac{1}{2},j}^{n}\,(F_{i-\frac{1}{4},j+\frac{1}{4}}^{n})^{+}-u_{i-\frac{1}{2},j}^{n}\,(F_{i-\frac{1}{4},j+\frac{1}{4}}^{n})^{-};
\end{aligned}$$

where $F_{k+\frac{1}{4},l+\frac{1}{4}}=\lambda_{i+\frac{1}{2},j}\,F_{i+\frac{1}{2},j}-\lambda_{i-\frac{1}{2},j}\,F_{i-\frac{1}{2},j}+\lambda_{i,j+\frac{1}{2}}\,F_{i,j+\frac{1}{2}}-\lambda_{i,j-\frac{1}{2}}\,F_{i,j-\frac{1}{2}}$ and the scalars λ for the different fluxes are summarized in the Table 2.1:

$F_{k+\frac{1}{4},l+\frac{1}{4}}$	$\lambda_{i-\frac{1}{2},j}$	$\lambda_{i+\frac{1}{2},j}$	$\lambda_{i,j+\frac{1}{2}}$	$\lambda_{i,j-\frac{1}{2}}$
$F_{i+\frac{1}{4},j+\frac{1}{4}}$	1/8	−3/8	3/8	−1/8
$F_{i-\frac{1}{4},j-\frac{1}{4}}$	−3/8	1/8	−1/8	3/8
$F_{i-\frac{1}{4},j+\frac{1}{4}}$	3/8	−1/8	−3/8	1/8
$F_{i+\frac{1}{4},j-\frac{1}{4}}$	−1/8	3/8	1/8	−3/8

Table 2.1 – Definition of the coefficients λ...

Non-linear staggered stabilized (NSS) scheme:

$$\frac{1}{\delta t}\,(h^{n+1}_{i,j}-h^{n}_{i,j})+\frac{1}{\delta x}\,(F^{n}_{i+\frac{1}{2},j}-F^{n}_{i-\frac{1}{2},j})+\frac{1}{\delta y}\,(F^{n}_{i,j+\frac{1}{2}}-F^{n}_{i,j-\frac{1}{2}})=0,$$

$$\frac{1}{\delta t}\left((hu)^{n+1}_{i+\frac{1}{2},j}-(hu)^{n}_{i-\frac{1}{2},j}\right)+\frac{2}{\delta x}\left((uF)^{n}_{i+\frac{3}{4},j-\frac{1}{4}}+(uF)^{n}_{i+\frac{1}{4},j+\frac{1}{4}}-(uF)^{n}_{i+\frac{3}{4},j+\frac{3}{4}}-(uF)^{n}_{i+\frac{1}{4},j-\frac{1}{4}}\right)$$
$$+\,gh^{n+1}_{i+\frac{1}{2},j}\,\eth(h^{n+1})_{i+\frac{1}{2},j}-\eth(\pi^{n})_{i+\frac{1}{2},j}=\omega\,(hv)^{n}_{i+\frac{1}{2},j}-\omega^{2}\gamma\delta t\,(hu)^{n}_{i+\frac{1}{2},j},$$

$$\frac{1}{\delta t}\left((hv)^{n+1}_{i+\frac{1}{2},j}-(hv)^{n}_{i-\frac{1}{2},j}\right)+\frac{2}{\delta x}\left((vF)^{n}_{i+\frac{3}{4},j-\frac{1}{4}}+(vF)^{n}_{i+\frac{1}{4},j+\frac{1}{4}}-(vF)^{n}_{i+\frac{3}{4},j+\frac{3}{4}}-(vF)^{n}_{i+\frac{1}{4},j-\frac{1}{4}}\right)$$
$$=-\omega(hu)^{n+1}_{i+\frac{1}{2},j}+\omega\,\gamma\delta t\,h^{n}_{i+\frac{1}{2},j}(g\,\eth(h^{n})_{i+\frac{1}{2},j}-\omega v^{n}_{i+\frac{1}{2},j}),$$

$$\frac{1}{\delta t}\left((hu)^{n+1}_{i,j+\frac{1}{2}}-(hu)^{n}_{i,j-\frac{1}{2}}\right)+\frac{2}{\delta y}\left(-(uF)^{n}_{i-\frac{1}{4},j+\frac{3}{4}}-(uF)^{n}_{i+\frac{1}{4},j+\frac{1}{4}}+(uF)^{n}_{i+\frac{1}{4},j+\frac{3}{4}}+(uF)^{n}_{i-\frac{1}{4},j+\frac{1}{4}}\right)$$
$$=\omega\,(hv)^{n}_{i,j+\frac{1}{2}}-\omega\,\gamma\delta t\,h^{n}_{i,j+\frac{1}{2}}(g\,\eth(h^{n+1})_{i,j+\frac{1}{2}}+\omega u^{n}_{i,j+\frac{1}{2}}),$$

$$\frac{1}{\delta t}\left((hv)^{n+1}_{i,j+\frac{1}{2}}-(hv)^{n}_{i,j-\frac{1}{2}}\right)+\frac{2}{\delta y}\left(-(vF)^{n}_{i-\frac{1}{4},j+\frac{3}{4}}-(vF)^{n}_{i+\frac{1}{4},j+\frac{1}{4}}+(vF)^{n}_{i+\frac{1}{4},j+\frac{3}{4}}+(vF)^{n}_{i-\frac{1}{4},j+\frac{1}{4}}\right)$$
$$+\,gh^{n+1}_{i,j+\frac{1}{2}}\,\eth(h^{n+1})_{i,j+\frac{1}{2}}-\eth(\pi^{n})_{i,j+\frac{1}{2}}=-\omega(hu)^{n+1}_{i,j+\frac{1}{2}}-\omega^{2}\gamma\delta t\,(hv)^{n}_{i,j+\frac{1}{2}},$$

with

$$\begin{aligned}
F^{n}_{i+\frac{1}{2},j} &= (hu)^{n}_{i+\frac{1}{2},j}-\gamma\delta t\,h^{n}_{i+\frac{1}{2},j}\,(g\,\eth(h^{n})_{i+\frac{1}{2},j}-\omega\,v^{n}_{i+\frac{1}{2},j}),\\
F^{n}_{i,j+\frac{1}{2}} &= (hv)^{n}_{i,j+\frac{1}{2}}-\gamma\delta t\,h^{n}_{i,j+\frac{1}{2}}\,(g\,\eth(h^{n})_{i,j+\frac{1}{2}}+\omega\,u^{n}_{i,j+\frac{1}{2}}),\\
\eth(h^{n})_{i,j+\frac{1}{2}} &= \frac{2}{\delta y}(h^{n}_{i,j+1}-h^{n}_{i,j}),\qquad \eth(h^{n})_{i+\frac{1}{2},j}=\frac{2}{\delta x}(h^{n}_{i+1,j}-h^{n}_{i,j}),\\
\pi^{n}_{i,j} &= \nu\delta t g\,h^{n}_{i,j}\left(\frac{1}{\delta x}\,(u^{n}_{i+\frac{1}{2},j}-u^{n}_{i-\frac{1}{2},j})+\frac{1}{\delta y}\,(v^{n}_{i,j+\frac{1}{2}}-v^{n}_{i,j-\frac{1}{2}})\right).
\end{aligned}$$

These non-linear schemes are in turn compared to a Godunov type scheme equipped with a HLLC solver for the non-linear shallow water equations which a short descrip-

tion is given in Appendix 2.A.

As we saw previously in the linear circular dam break test a correction is needed to stabilize the upwind MAC and RT schemes. Doing so, the following stabilization quantities depending on $\pi_{i,j}^n$ (defined above) are judiciously injected in the discrete velocity equations for the circular dam break test case (second test below):

$$\frac{1}{\delta x}(\pi_{i+1,j}^n - \pi_{i,j}^n) \text{ and } \frac{1}{\delta y}(\pi_{i,j+1}^n - \pi_{i,j}^n).$$

2.4.2.1 Geostrophic adjustment test case

Here again we consider a stationary vortex $(h_0, \boldsymbol{u}_0)$ of the problem (2.1) for which the water height and the tangential velocity are initialized as follows:

$$\begin{cases} h_0(x,y) = \bar{h}(r) \\ \boldsymbol{u}_0(x,y) = \bar{u}(r)\, \boldsymbol{e}_\theta \end{cases}$$

with

$$r = x^2 + y^2, \quad \theta = \arctan\left(\frac{y}{x}\right) \text{ and } \boldsymbol{e}_\theta = \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \end{pmatrix}$$

Substituting $(h_0, \boldsymbol{u}_0)$ in the equations (2.1) we get after some basic algebraic computation:

$$g\nabla h_0 + \omega\, \boldsymbol{u}_0^\perp = -\frac{\bar{u}^2}{r^2}\, \boldsymbol{e}_\theta^\perp.$$

We remark that if $\bar{u}^2$ is close to zero, then the stationary solution $(h_0, \boldsymbol{u}_0)$ satisfies the geostrophic equilibrium state: $g\nabla h_0 + \omega\, \boldsymbol{u}_0^\perp = 0$, more precisely this state is a solution of the linear equations (2.4). In order to highlight this point of view, the setup we consider here is proposed in Audusse, Do, Omnes, et al. 2018 where the function $\bar{u}$ is given by

$$\bar{u}(r) = \epsilon\left(5r\mathbb{1}_{(0,\frac{1}{5})}(r) + (2-5r)\mathbb{1}_{(\frac{1}{5},\frac{2}{5})}(r)\right) \text{ and } \max_r \bar{u}(r) = \epsilon, \text{ with } \epsilon \geq 0.$$

Then $\bar{h}$ is the solution of the following ODE:

$$\begin{cases} g\,\bar{h}'(r) = \omega\,\bar{u}(r) + \frac{\bar{u}^2}{r} \\ \bar{h}(0) = 1. \end{cases}$$

The computational domain is the square $[-0.5, 0.5]^2$ using $Nx = Ny = 50$ cells and the time step is fixed to $dt = dx/5$.
Let first illustrate that the upwind RT and the NSS schemes behave quite similarly when we nullify the stabilization terms. The following results justify more this point of view: It can be seen in Figure 2.14 that the corrections added in the NSS scheme do

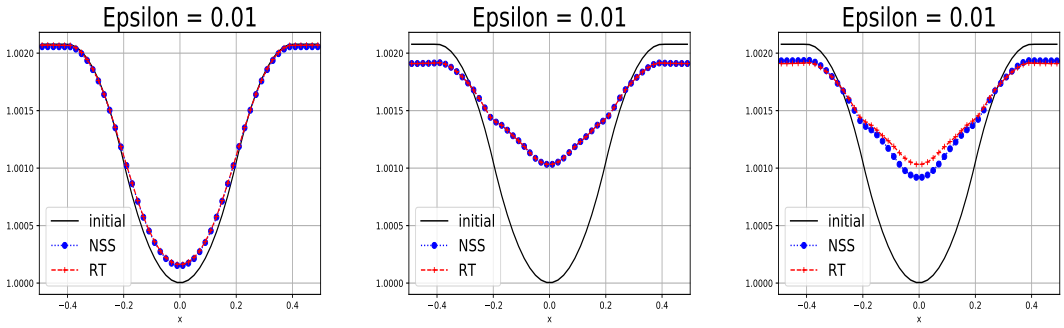


Figure 2.14 – Geostrophic adjustment test: Horizontal cut off the water height computed from the RT and NSS schemes at time $T = 5$ (left) and $T = 50$ (center) without correction terms and $T = 50$ (right) with correction.

not have a considerable effect with regard to the upwind RT scheme. We think that the problem remains potentially at the initialization of the velocity components for the RT upwind and NSS schemes as mentioned previously in the linear well balance test.

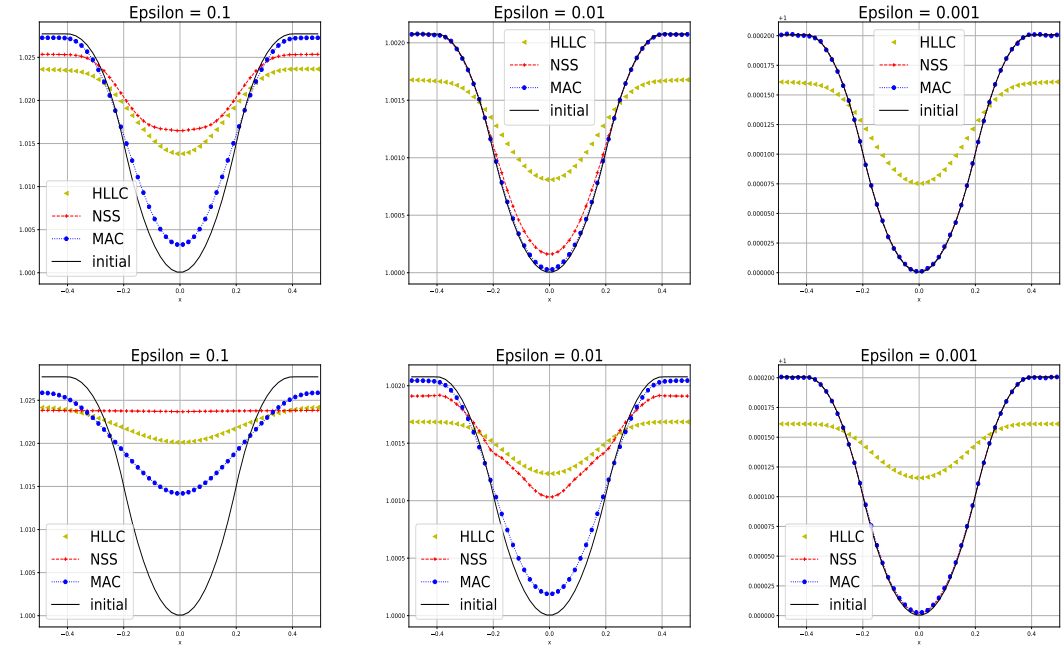


Figure 2.15 – Geostrophic adjustment: Horizontal cut off the water height resulting at time $T = 5$ on the top line and $T = 50$ on the bottom

Next we consider only the NSS scheme and the Figure 2.15 plots the results for several values of ϵ. Through these various figures we can see that the staggered MAC and NSS

schemes are very close to the initial condition for some values of ϵ sufficiently small which correspond to the low Froude regime unlike the HLLC scheme. One should also observe the good behavior in long time of the MAC and HLLC schemes for highly unstable areas where the NSS scheme becomes constant in spite of its good discrete properties. However only the MAC scheme has the capacity to remain very close to the initial state for a very large simulation time, contrary to the RT scheme which moves away from it more and more quickly.

The Figure 2.16 below shows the following relative error type in L^2 norm of the HLLC, MAC and NSS schemes in term of ϵ at time $T = 5$:

$$\frac{||h(T,\cdot) - h_0||_{L^2(\Omega)}}{||h_0 - \underline{h}||_{L^2(\Omega)}}, \quad \text{with } \underline{h} = \max_{x\in\Omega}(h_0(x))$$

Here we see that the relative error increases in term of ϵ that means that the numerical

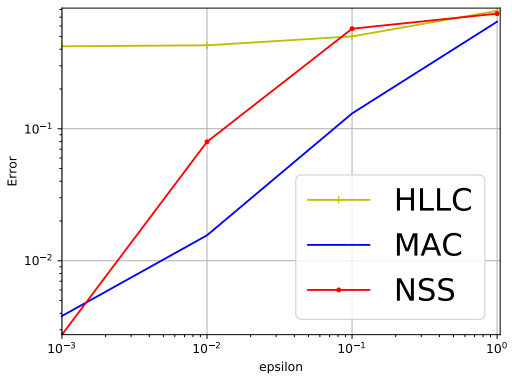


Figure 2.16 – Geostrophic adjustment: Relative error in L^2 norm in term of ϵ between the initial and numerical solution at time $T = 5$

solutions move away from the geostrophic equilibrium ($g\nabla h + \omega \boldsymbol{u}^{\perp} = 0$), for higher epsilon values. Basically this feature is rightful since the geostrophic equilibrium state is only satisfied for the linear SWC equations (2.4). However for $\epsilon \leq 10^{-3}$ which correspond to the low Froude regime, the non-linear SWC equations degenerate to the linear equations and thus the numerical solution of the MAC and NSS schemes remain very close to the initial state. This explains why the relative errors of the MAC and NSS schemes are of order 10^{-2} and 10^{-3} respectively for $\epsilon = 10^{-3}$. This accuracy justifies the well behavior of the corresponding linear MAC and NSS schemes with respect to the preservation of the geostrophic equilibrium state as shown in Figure 2.15.

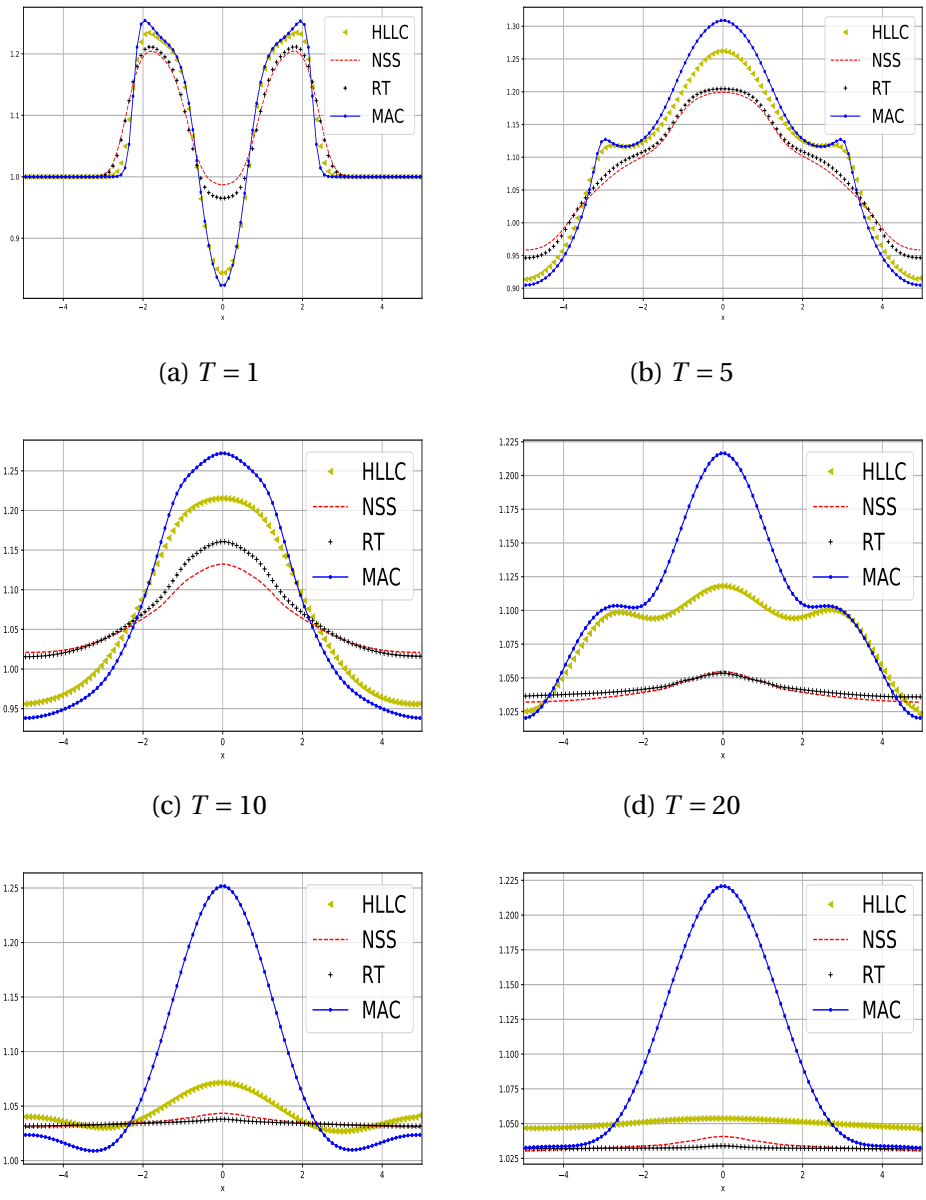
2.4.2.2 Circular dam break test case

In this last test we consider the same circular dam break test presented in the linear case where the computational domain is the square $[-5,5]^2$ and the initial water height is given by:

$$h_0(x,y) = \begin{cases} 2, & \text{if } x^2+y^2 \leq 1, \\ 1, & \text{otherwise}, \end{cases}$$

while the components of the velocity are initialized to zero. Here both upwind MAC and RT schemes are corrected by a discrete gradient of π with a stabilization coefficient $\nu = 2$. Regarding the NSS scheme a correction of order 2 is only operated on the numerical mass flux.

Below we plot in Figures 2.17 and 2.18 the obtained results at different final times. Firstly for a short time of simulation $T = 1$ the schemes behave identically with a notable difference in altitude as shown in Figure 2.12c. Then we observe a huge difference between the water height computed by the MAC scheme compared to the others schemes including HLLC scheme as was the case for the linear circular dam break test (see Figure 2.12). This difference can be explained by the fact that the size of a dual (MAC) cell (or primal cell) is twice the size of a diamond (RT) cell which implies a scaling of the velocity components by a factor $\frac{1}{2}$. Hence the discrete height is thus impacted by the intermediate of the mass flux expressed in term of the discrete normal velocity.
All these results enhance on one hand the robustness of the upwind MAC scheme corrected by a controlled numerical diffusion added in the discrete momentum equation. On the other hand this test case evidences once again the non efficiency of the staggered schemes based on the RT finite elements whether corrected upwind or centered stabilized non-linear schemes.



(a) $T = 1$ (b) $T = 5$

(c) $T = 10$ (d) $T = 20$

(e) $T = 50$ (f) $T = 100$

Figure 2.17 – Circular dam break – Long time behavior: horizontal cut off the water height for the different schemes

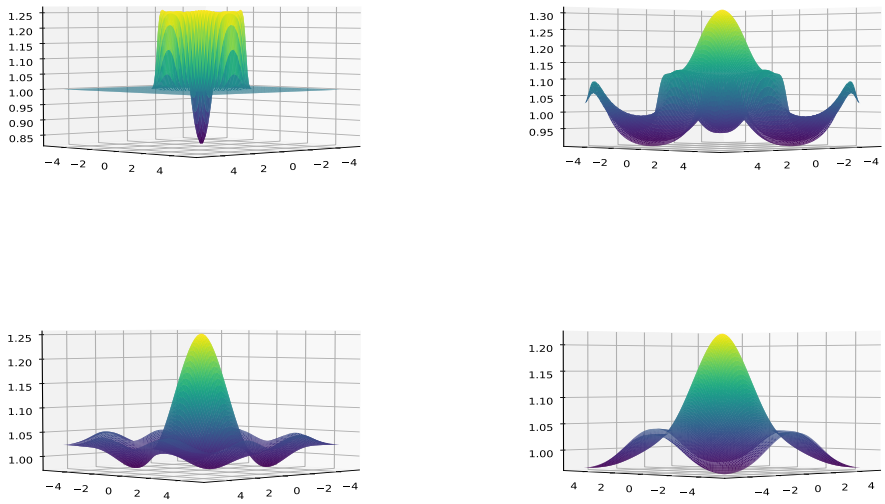


Figure 2.18 – Circular dam break: $3D$ plot of the water height resulting of the MAC scheme at time $T = 1$, $T = 5$, $T = 50$ and $T = 100$ respectively

Appendix

2.A HLLC scheme for the non-linear shallow water equations with Coriolis force

We begin by rewriting the shallow water equations (2.1) in a conservative form as follows

$$\partial_t U + \partial_x F(U) + \partial_y G(U) = S(U), \tag{2.42}$$

where the conservative variable U and the vector functions F, G and S are given by

$$U = \begin{bmatrix} h \\ hu \\ hv \end{bmatrix}, \quad F(U) = \begin{bmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{bmatrix}, \quad G(U) = \begin{bmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \end{bmatrix}, \qquad S(U) = \begin{bmatrix} 0 \\ \omega hv \\ -\omega hu \end{bmatrix}.$$

Then the scheme for the non-homogeneous system (2.42) reads:

$$\frac{U_{i,j}^{n+1} - U_{i,j}^{n}}{\delta t} + \frac{F_{i+\frac{1}{2},j}^{n} - F_{i-\frac{1}{2},j}^{n}}{\delta x} + \frac{G_{i,j+\frac{1}{2}}^{n} - G_{i,j-\frac{1}{2}}^{n}}{\delta y} = S_{i,j}^{n,n+1}, \tag{2.43}$$

with

$$U_{i,j}^{n} = \begin{bmatrix} h_{i,j}^{n} \\ (hu)_{i,j}^{n} \\ (hv)_{i,j}^{n} \end{bmatrix}, \quad S_{i,j}^{n,n+1} = \begin{bmatrix} 0 \\ \omega(hv)_{i,j}^{n} \\ -\omega(hu)_{i,j}^{n+1} \end{bmatrix}.$$

It remains to define the numerical fluxes $F_{i+\frac{1}{2},j}$ and $G_{i,j+\frac{1}{2}}$ in order to close the algorithm. Both quantities are approximately defined by the so-called HLLC (see in Toro 1997) solver that can be reformulated as follows:

$$F_{i+\frac{1}{2},j} = \begin{cases} F_{i,j} & \text{if } 0 \le \lambda_{i,j} \\ F_{i,j}^{*} & \text{if } \lambda_{i,j} \le 0 \le \lambda_{i+\frac{1}{2},j}^{*} \\ F_{i+1,j}^{*} & \text{if } \lambda_{i+\frac{1}{2},j}^{*} \le 0 \le \lambda_{i+1,j} \\ F_{i+1,j} & \text{if } 0 \ge \lambda_{i+1,j}, \end{cases}$$

with $F^*_{i,j} = F_{i,j} + \lambda_{i,j}(U^*_{i,j} - U_{i,j})$ where the intermediate variable $U^*_{i,j}$ is computed by:

$$U^*_{i,j} = \begin{bmatrix} h^*_{i,j} \\ h^*_{i,j}\,\lambda^*_{i+\frac{1}{2},j} \\ h^*_{i,j}\,v_{i,j} \end{bmatrix} \quad \text{with} \quad h^*_{i,j} = \frac{\lambda_{i,j} - u_{i,j}}{\lambda_{i,j} - \lambda^*_{i+\frac{1}{2},j}}\, h_{i,j}.$$

Finally the wave speeds $\lambda_{i,j}$, $\lambda^*_{i+\frac{1}{2},j}$ and $\lambda_{i+1,j}$ are given by:

$$\begin{aligned}
\lambda_{i,j} &= \min\big(u_{i,j} - \sqrt{g h_{i,j}},\ u_{i+1,j} - \sqrt{g h_{i+1,j}}\big) \\
\lambda_{i+1,j} &= \max\big(u_{i,j} + \sqrt{g h_{i,j}},\ u_{i+1,j} + \sqrt{g h_{i+1,j}}\big) \\
\lambda^*_{i+\frac{1}{2},j} &= \frac{[\lambda(hu)]_{i+\frac{1}{2},j} - [hu^2 + \frac{1}{2}gh^2]_{i+\frac{1}{2},j}}{[\lambda h]_{i+\frac{1}{2},j} - [hu]_{i+\frac{1}{2},j}},
\end{aligned}$$

where the notation $[\bullet]_{i+\frac{1}{2},j}$ stands to the following jump

$$[\bullet]_{i+\frac{1}{2},j} = [\bullet]_{i+1,j} - [\bullet]_{i,j}.$$

The definition of the flux $G_{i,j+\frac{1}{2}}$ mimics the one of $F_{i,j+\frac{1}{2}}$ by simply changing the role of the indexes which closes the description of the HLLC scheme.

2.B Crank-Nicholson scheme for the linear shallow water equations with Coriolis force based on the Rannacher-Turek elements

The Crank-Nicholson scheme is well known to be unconditionally stable and second order in time. The linear staggered scheme described below is performed with the Crank-Nicholson method respect to the time integration instead of the segregated discretization presented in the above sections and works on the RT finite elements for the space discretization. The scheme reads:

$$\frac{1}{\delta t}\,(h_K^{n+1} - h_K^n) + h_0\,\frac{1}{2}\big(\mathrm{div}_K(\boldsymbol{u}^n) + \mathrm{div}_K(\boldsymbol{u}^{n+1})\big) = 0, \quad \forall\, K \in \mathscr{M}, \tag{2.44a}$$

$$\frac{1}{\delta t}\,(\boldsymbol{u}_\sigma^{n+1} - \boldsymbol{u}_\sigma^n) + g\,\frac{1}{2}\,\big((\nabla h^n)_\sigma + (\nabla h^{n+1})_\sigma\big) = -\omega\,\frac{1}{2}\,\big(\boldsymbol{u}_\sigma^n + \boldsymbol{u}_\sigma^{n+1}\big)^\perp, \forall\, \sigma \in \mathscr{E}_{int}, \tag{2.44b}$$

where the discrete operators appearing here are previously defined in Section 2.2.3. The scheme (2.44) involves a linear system which can be written in vector form as

follows:

$$A\,U^{n+1} = B\,U^n, \text{ with } U^n = U(h^n, \boldsymbol{u}^n), \tag{2.45}$$

where A and B are two matrices given by:

$$A\,U^n = \begin{bmatrix} h_K^n + \delta t\; h_0\; \frac{1}{2}\; \mathrm{div}_K(\boldsymbol{u}^n) \\ \boldsymbol{u}_\sigma^n + g\; \frac{1}{2}\; (\nabla h^n)_\sigma + \omega\; \frac{1}{2}\; (\boldsymbol{u}_\sigma^n)^\perp \end{bmatrix} \quad \text{and} \quad U^n = \begin{bmatrix} h_K^n - \delta t\; h_0\; \frac{1}{2}\; \mathrm{div}_K(\boldsymbol{u}^n) \\ \boldsymbol{u}_\sigma^n - g\; \frac{1}{2}\; (\nabla h^n)_\sigma - \omega\; \frac{1}{2}\; (\boldsymbol{u}_\sigma^n)^\perp \end{bmatrix}.$$

The well-posedness of the linear system (2.45) and other stability properties are treated in the following proposition.

Proposition 2.6 ("Well posedness and well balance" for the Crank-Nicholson scheme)**.** *Let $n \in \{0, \cdots, N-1\}$ and let given a pair of discrete functions $(h^n, \boldsymbol{u}^n)$. Then the scheme (2.44) admits one and only one discrete solution $(h^{n+1}, \boldsymbol{u}^{n+1})$. Furthermore if $(h^n, \boldsymbol{u}^n)$ verifies (2.17), then $(h^{n+1}, \boldsymbol{u}^{n+1})$ satisfies (2.17) too.*

Proof. Firstly we prove that the linear system (2.45) admits a unique solution that means the matrix A is invertible. So let us suppose that $A\,U^n = 0$, then we have

$$\begin{cases} h_K^n + \delta t\; h_0\; \frac{1}{2}\; \mathrm{div}_K(\boldsymbol{u}^n) = 0 \\ \boldsymbol{u}_\sigma^n + \delta t\; g\; \frac{1}{2}\; (\nabla h^n)_\sigma + \omega\; \frac{1}{2}\; (\boldsymbol{u}_\sigma^n)^\perp = 0. \end{cases}$$

Then multiplying the first equation by $|K|\,\frac{g}{h_0} h_K^n$ and the second by $|D_\sigma|\,\boldsymbol{u}_\sigma^n$ and summing the result over $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}_{\mathrm{int}}$ respectively, we get:

$$\sum_{K\in\mathscr{M}} |K|\,\frac{g}{h_0}\,(h_K^n)^2 + \sum_{\sigma\in\mathscr{E}_{\mathrm{int}}} |D_\sigma|\,|\boldsymbol{u}_\sigma^n|^2 = -g\,\frac{1}{2}\Big(\sum_{K\in\mathscr{M}} |K|\,h_K^n \mathrm{div}_K(\boldsymbol{u}^n) + \sum_{\sigma\in\mathscr{E}_{\mathrm{int}}} |D_\sigma|\,(\nabla h^n)_\sigma \cdot \boldsymbol{u}_\sigma^n\Big).$$

Thanks to the div-grad relationschip, the right hand side term vanishes and we obtain $h_K^n = 0$ for all $K \in \mathscr{M}$ and $u_\sigma^n = 0$ for all $\sigma \in \mathscr{E}_{\mathrm{int}}$. Hence $U^n = 0$ and thus A is invertible. Then for the well balance property, it is straightforward to see that if $(h^n, \boldsymbol{u}^n)$ verifies (2.17), then the discrete steady state $h^{n+1} = h^n$ and $\boldsymbol{u}^{n+1} = \boldsymbol{u}^n$ is the only solution of the scheme (2.44), which concludes the proof. □

Unlike the linear RT and LSS schemes presented in the previous sections, the Crank-Nicholson scheme conserves exactly the discrete energy as stated in the following.

Proposition 2.7 (Discrete mechanic energy balance)**.** *Let $n \in \{0, \cdots, N-1\}$. The discrete solution of the Crank-Nicholson scheme (2.44) ensures the following discrete energy conservation:*

$$E^{n+1} = E^n, \quad \forall\; n \in \{0, \cdots, N-1\}. \tag{2.46}$$

Proof. Let us derive first the potential energy balance of the scheme which is obtained

by multiplying (2.44a) by $\frac{g}{h_0}\frac{1}{2}\,(h_K^{n+1}+h_K^n)$ to get:

$$\frac{1}{2}\frac{1}{\delta t}\frac{g}{h_0}\left((h_K^{n+1})^2-(h_K^n)^2\right)+\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|\sigma|g\left(\frac{h_K^n+h_L^n}{2}+\frac{h_K^{n+1}+h_L^{n+1}}{2}\right)\frac{\boldsymbol{u}_\sigma^n+\boldsymbol{u}_\sigma^{n+1}}{2}\cdot\boldsymbol{n}_{K,\sigma}$$
$$-\frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)}|D_\sigma|\,g\,\frac{(\nabla h^n)_\sigma+(\nabla h^{n+1})_\sigma}{2}\cdot\frac{\boldsymbol{u}_\sigma^n+\boldsymbol{u}_\sigma^{n+1}}{2}=0.$$

In the same way multiplying the velocity equation (2.44b) by $\frac{\boldsymbol{u}_\sigma^n+\boldsymbol{u}_\sigma^{n+1}}{2}$ we get the following discrete kinetic balance:

$$\frac{1}{2}\frac{1}{\delta t}\left(|\boldsymbol{u}_\sigma^{n+1}|^2-|\boldsymbol{u}_\sigma^n|^2\right)+g\,\frac{(\nabla h^n)_\sigma+(\nabla h^{n+1})_\sigma}{2}\cdot\frac{\boldsymbol{u}_\sigma^n+\boldsymbol{u}_\sigma^{n+1}}{2}=0.$$

Finally, doing the following computations yield the result:

$$\sum_{K\in\mathscr{M}}|K|\,\frac{1}{2}\frac{1}{\delta t}\frac{g}{h_0}\left((h_K^{n+1})^2-(h_K^n)^2\right)+\sum_{\sigma\in\mathscr{E}_{\rm int}}|D_\sigma|\,\frac{1}{2}\frac{1}{\delta t}\left(|\boldsymbol{u}_\sigma^{n+1}|^2-|\boldsymbol{u}_\sigma^n|^2\right)=0,$$

which concludes the proof. □

3 A staggered scheme for the one-dimensional shallow water flow and sediment transport with a stabilized friction term

Sommaire

Abstract. In this chapter, we extend a staggered scheme designed for the numerical simulation of the shallow water flow to the coupled shallow water-Exner equations. The scheme is based on the first order upwind method for the convection mass and momentum parts and centered discrete gradient applied to the water height and the sediment depth. The numerical scheme yields a completely algebraic algorithm following a decoupled time stepping thanks to an easy adaptive explicit-implicit time integration for the momentum and Exner equations. For the coupling of the water flow and sediment transport model, we investigate the influence of the shear stress source term and the sediment transport flux. Compared to some standard shear stress formulae, a source term accounts for the bed elevation and the bedload sediment transport is implemented in the momentum equation. This algebraic term allows a dissipation of the associated energy and preserves the hyperbolicity of the original system. Then regarding the sediment transport flux, two kinds of bedload formulae are considered, the classical formulae like Meyer-Peter & Müller formula and a corrected formula. This latter results of an asymptotic analysis from the incompressible Navier-Stokes equations which depends on the depth of the movable sediment layer. Finally, the efficiency of the decoupled scheme is numerically investigated as well as the influence of the sediment transport flux.

Keywords shallow water equations, Exner equation, sediment transport, staggered scheme.

3.1 Introduction

Let Ω be a bounded domain of $\mathbb{R}$ with and let $T > 0$. We consider the shallow water equations with time dependent topography given by:

$$\partial_t h + \partial_x(hu) = 0 \qquad \text{in } \Omega \times (0,T), \tag{3.1a}$$

$$\partial_t(hu) + \partial_x(hu^2 + p) + gh\partial_x z = -\tau/\rho_w \qquad \text{in } \Omega \times (0,T), \tag{3.1b}$$

$$p = \frac{1}{2}gh^2 \qquad \text{in } \Omega \times (0,T). \tag{3.1c}$$

where t stands for the time, g is the gravity constant, z the sediment depth, ρ_w is the density of the water, h the water height and u the velocity of the flow. The source term τ appearing in the right hand-side of (3.1b) is the shear stress given by $\tau = \rho_w g h S_f$ where S_f is a friction term which depends only on h and u. The system (3.1) is coupled with the Exner equation which describes the evolution of the sediment layer in the following way:

$$\partial_t z + \frac{1}{1-\phi}\partial_x q_b = 0, \tag{3.2}$$

where $\phi \in [0,1)$ is a constant porosity and q_b is the sediment discharge or the bedload transport flux depending also on h and u and sometimes on S_f. This coupled system is

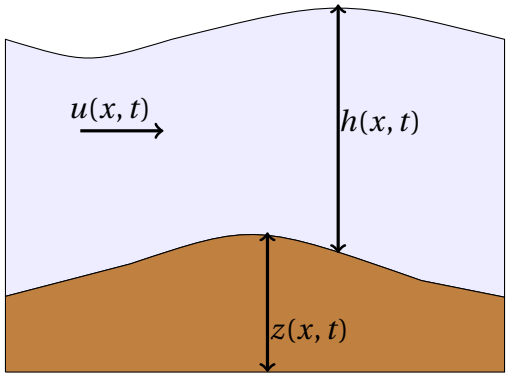


Figure 3.1 – Configuration of the water flow and sediment transport: $z(t,x)$ the sediment depth and $h(t,x) + z(t,x)$ the free surface.

used to model the interaction between the shallow water flow and sediment transport process. An other formulation of the Exner equation is also presented in several works, based on a decomposition of the sediment layer in two sub-layers as depicted in Figure 3.2 below: an erodible movable layer of dimension z_m, in direct exchange with the water column and a fixed layer known as the bedrock, of depth z_b constant with respect to the time variable as that $\partial_t z(x,t) = \partial_t z_m(x,t)$.

There exist several formulae of the sediment transport flux q_b, for instance the Grass 1981, Meyer-Peter and Muller 1948 bedload transport discharges and other related formula (see Ashida and Michiue 1972; Nielson 1992; Van Rijn 1984; Fernandez Luque

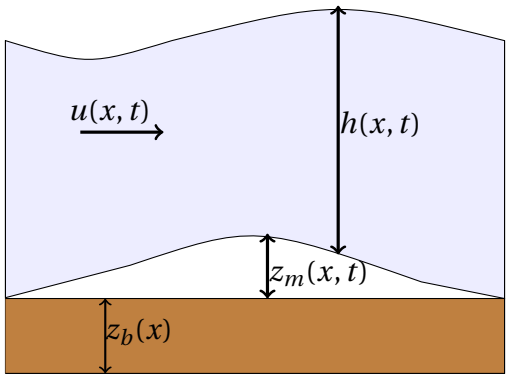


Figure 3.2 – Splitting of the sediment layer.

and Van Beek 1976; Einstein 1942). The two most used definitions of S_f, defined in an empirically way, are the Darcy-Weisbach friction law and the Manning friction law. For some classical models the quantities S_f and q_b are related and the hyperbolicity of the system depends on the friction law. In fact it is well known that the Grass bedload formula makes the system (3.1-3.2) strictly hyperbolic while for other bedload formula like Meyer-Peter & Müller (and other related formula) the system (3.1-3.2) is hyperbolic for S_f given by the Darcy-Weisbach friction law and conditionally hyperbolic for the Manning friction law (see Audusse 2018, Castro Dìaz, Fernàndez-Nieto, and Ferreiro 2008, Fernàndez-Nieto, Lucas, Morales De Luna, et al. 2014).

The system (3.1-3.2) suffers from two major disadvantages: firstly for most friction laws, these equations do not have an energy balance or dissipation which constitutes a crucial point for a hyperbolic system. Secondly the sediment flux q_b does not take into account the evolution of the sediment depth and hence the Exner equation may fail. Indeed the sediment transport process is preceded by a sediment deposition and followed by the entertainment; thus an equilibrium regime is observed when the deposition rate is equal to the entertainment rate. Then the sediment flux q_b may account for the quantity of deposited material, which is not the case for all classical formulae. In the sequel we introduce a regularized friction law which is shown to ensure energy dissipation, this answers the first above mentioned disadvantage. As to the second one, we introduce a corrected sediment flux q_b, mainly we focus on recent advances for shallow water-Exner models proposed in Boittin 2019, Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017. Both works ensure a dissipation of the associated energy and treat simultaneously the deposition and entertainment of the sediment. In the sequel we adopt the model developed in Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017, in which the authors propose a linear and quadratic bedload formulae of q_b resulting in a quasi-uniform regime where the erosion rate equals to the deposit one.

For the numerical setting, we use a decoupled approach which computes first the water flow and then solves the Exner equation. This strategy yields a simple algorithm

which is very easy to implement using a decoupled time discretization and a staggered arrangement of the discrete unknowns. Furthermore, this scheme is shown in Gunawan, Eymard, and Pudjaprasetya 2015 to be more accurate; the efficiency of the decoupled approach is proven by the numerical results which are in good agreement with experimental data. This scheme is based on the standard upwind scheme both for the mass and momentum numerical fluxes. The preservation of the positivity of the water height is ensured by means of a classical restriction on the time step.
The objective here is to extend this scheme to the new Saint-Venant-Exner coupling model of Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017 which involves a correction of the sediment transport flux which takes into account the depth of the movable layer (see also Boittin 2019). To this purpose a decoupled time integration scheme is built in the spirit of the technique presented in Gunawan, Eymard, and Pudjaprasetya 2015; Gunawan 2015 for the computation of the fluid flow. Thus for the computation of the Exner equation, we perform an adaptive explicit-implicit time discretization. This scheme involves only explicit time stepping with respect to the sediment depth in the computation of the bedload transport flux and full implicit time step for the water height and the velocity.

The remainder of this chapter has the following structure: in Section 3.2, we study a coupled shallow water flow and sediment transport system using a stabilization source term which does not require the precise the definition of q_b. Physical limitations of the classical bedload formulae are treated in Section 3.3 and a formally definition of the sediment flux is thus considered. In Section 3.4, we investigate the numerical approximation of the coupled system. Finally we present some numerical tests and comparison with the literature in Section 3.5.

3.2 A stabilized friction term

Since the source term $\tau/\rho_w = ghS_f$ is defined empirically by a friction law for S_f, we investigate here a stabilized shear stress formula for τ fulfilling the following requirements:

- the preservation of the lake at rest steady state, which implies that τ is expected to vanish when $u = 0$ for the resulting system.
- the dissipation of the energy of the equations (3.1-3.2).
- the preservation of the hyperbolicity of the original system (3.1-3.2) in other words the regularized friction we are seeking here must maintain the hyperbolicity of the equations (3.1-3.2) for a standard friction law.

We do not need here to specify the expression of the bedload transport q_b and we take $\phi = 0$ for the sake of simplicity. Let us consider a regular solution (h, u, z) of the system (3.1-3.2) and let us introduce a regular function ψ which depends only on h i.e. $\psi = \psi(h)$. We denote by r the ratio of the density of the water ρ_w with respect to that of the sediment ρ_s *i.e.* $r = \rho_w/\rho_s$.

Multiplying the mass equation (3.1a) by $r\psi(h)$, we obtain after some algebraic manipulations, the following balance equation

$$\partial_t(rh\psi(h)) + \partial_x(rh\psi(h)u) + rh^2\psi'(h)\partial_x u = 0. \tag{3.3}$$

Then, multiplying the Exner equation by $g(rh+z)$, we find

$$\partial_t(\frac{1}{2}gz^2) + rgh\partial_t z + g(rh+z)\partial_x q_b = 0.$$

Thanks to the mass equation (3.1a), one has $rgh\partial_t z = \partial_t(grhz) + rgz\partial_x(hu)$. Thus the following equation holds

$$\partial_t(\frac{1}{2}gz^2 + rghz) + g(rh+z)\partial_x q_b + rgz\partial_x(hu) = 0.$$

Thus adding this equation with (3.3) we get the following equality:

$$\partial_t(rh\psi(h)+\frac{1}{2}gz^2+rghz)+\partial_x(rh\psi(h)u)+rh^2\psi'(h)\partial_x u+g(rh+z)\partial_x q_b+rgz\partial_x(hu) = 0. \tag{3.4}$$

Next, taking the scalar product of the momentum equation (3.1b) by ru, using twice the mass equation (3.1a), we obtain the kinetic energy balance

$$\partial_t(E_k) + \partial_x(E_k u) + ru\partial_x p + rghu\partial_x z = -rghS_f\ u, \tag{3.5}$$

with $E_k = \frac{1}{2}rhu^2$.
Summing up (3.4) and (3.5), we get

$$\begin{aligned}\partial_t(rh\psi(h)+\frac{1}{2}gz^2+rghz+E_k)+\partial_x\big((rh\psi(h)+\frac{1}{2}rh|u|^2+grhz+rp)u\big)+\partial_x(g(rh+z)q_b)&\\ = -rghS_f u + gq_b\partial_x(rh+z) - r(h^2\psi'(h)-p)\partial_x u.&\end{aligned} \tag{3.6}$$

Let us define $\psi(h)$ so as to satisfy the relation: $h^2\psi'(h) - p = 0$, i.e. $\psi'(h) = p/h^2 = \frac{1}{2}g$ and $\psi(h) = \frac{1}{2}gh$.
In this case, the equation (3.6) boils bown to

$$\begin{aligned}\partial_t(E_p+E_k) + \partial_x\big((gh^2 + \frac{1}{2}h|u|^2 + ghz)ru\big) + \partial_x(g(rh+z)q_b)&\\ = -\big(rghuS_f - gq_b\partial_x(rh+z)\big),&\end{aligned} \tag{3.7}$$

with $E_p = \frac{1}{2}rgh^2 + \frac{1}{2}gz^2 + rghz$.
In order to get an energy balance equation (or inequality), it is important that the right-hand side of (3.7) be negative, that means:

$$rghuS_f - gq_b\partial_x(rh+z) \geq 0. \tag{3.8}$$

Note that the condition (3.8) cannot hold for a classical friction law of the form $f(h)|u|u$ with $f(h)$ a non-negative function depending on h. In order for (3.8) to hold, we choose

$$S_f = \begin{cases} \partial_x(rh+z)\dfrac{q_b}{rhu}, & \text{if } u \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.9}$$

Then it is straightforward to observe that the definition (3.9) fulfills (3.8) whatever the law chosen for q_b, since q_b always vanishes when $u = 0$ for classical bedload formulae (see for instance (3.16), (3.17), (3.21)). Hence the energy dissipation is ensured. Moreover the first above mentioned requirement holds since S_f vanishes when the fluid is at rest *i.e.* $u = 0$.

The shallow water Exner system thus reads:

$$\partial_t h + \partial_x(hu) = 0 \qquad \text{in } \Omega \times (0,T), \tag{3.10a}$$

$$\partial_t(hu) + \partial_x(hu^2 + \frac{1}{2}gh^2) + gh\partial_x z = -ghS_f \qquad \text{in } \Omega \times (0,T), \tag{3.10b}$$

$$\partial_t z + \partial_x(q_b) = 0 \qquad \text{in } \Omega \times (0,T), \tag{3.10c}$$

where the friction term is defined by (3.9) and the sediment transport discharge q_b is specified in the sequel.

Let us now turn to the hyperbolicity of the modified shallow water-Exner system (3.10).
In the case $u \neq 0$, the system (3.10) can be written in vector form as follows:

$$\partial_t U + \partial_x(F(U)) + B(U)\partial_x U = 0, \tag{3.11}$$

where

$$U = \begin{bmatrix} h \\ q \\ z \end{bmatrix}, \qquad F(U) = \begin{bmatrix} q \\ \frac{q^2}{h} + \frac{1}{2}gh^2 \\ q_b \end{bmatrix}, \qquad B(U) = \begin{bmatrix} 0 & 0 & 0 \\ gh\frac{q_b}{q} & 0 & gh(1+\frac{q_b}{rq}) \\ 0 & 0 & 0 \end{bmatrix}$$

with $q = hu$. The equation (3.11) is equivalent to:

$$\partial_t U + A(U)\partial_x U = 0, \quad \text{with } A(U) = \begin{bmatrix} 0 & 1 & 0 \\ gh\frac{q_b}{q} + gh - \frac{q^2}{h^2} & 2\frac{q}{h} & gh(1+\frac{q_b}{rq}) \\ \partial_h q_b & \partial_q q_b & 0 \end{bmatrix} \tag{3.12}$$

In order to compute the eigenvalues of this matrix, it is better to adopt the variable

$V=(u,2c,z)^t$ with $c^2=gh$ instead of U. Thus the system (3.12) becomes:

$$\partial_t V + M(V)\partial_x V = 0, \quad \text{with } M(V) = \begin{bmatrix} u & c\frac{q_b}{q}+\frac{gh}{c} & g(1+\frac{q_b}{rq}) \\ c & u & 0 \\ \frac{c^2}{g}\partial_q q_b & \frac{c}{g}(\partial_h q_b + u\partial_q q_b) & 0 \end{bmatrix} \tag{3.13}$$

It is well known that the matrices $M(V)$ and $A(U)$ have the same eigenvalues, so the hyperbolicity of the system (3.13) implies that of the system (3.12).
For the sake of simplicity, we restrict the study for bedload formulae which satisfy the following conditions:

$$\partial_q q_b > 0, \quad \frac{q_b}{q} > 0 \tag{3.14a}$$

$$\partial_h q_b + u\partial_q q_b = 0. \tag{3.14b}$$

Then under the compatibility conditions (3.14), we establish in the Proposition 3.1 the nature of the eigenvalues of $M(V)$. Notice that, the conditions (3.14) are the assumptions kind considered in Fernàndez-Nieto, Lucas, Morales De Luna, et al. 2014 in order to deal with the hyperbolicity of the shallow water Exner system under some classical bedload formulae.

The following result states the hyperbolicity of the system (3.10) for some bedload formulae.

Proposition 3.1. *Let (h,q,z) be a solution of the equations (3.11) and consider a bedload formula of q_b depending only on h and u and such that the compatibility conditions (3.14) hold. Then if $u\neq 0$, the matrix $M(V)$ admits three distinct real eigenvalues.*

Proof. The characteristic polynomial of the matrix $M(V)$ is given by:

$$P_M(\lambda) = \lambda\Big((u-\lambda)^2 - (c^2\frac{q_b}{q}+gh)\Big) + c^2(1+\frac{q_b}{rq})\Big((u-\lambda)\partial_q q_b - (\partial_h q_b + u\partial_q q_b)\Big). \tag{3.15}$$

Owing to the conditions (3.14), two cases are possible:

- First case: If $u>0$ then we have

$$P_M(0) = c^2 u(1+\tfrac{q_b}{rq})\partial_q q_b > 0, \qquad P_M(u) = -u(c^2\frac{q_b}{q}+gh) < 0,$$
$$\lim_{\lambda\to-\infty} P_M(\lambda) = -\infty \qquad \text{and} \qquad \lim_{\lambda\to+\infty} P_M(\lambda) = +\infty.$$

- Second case: If $u<0$ then we get

$$P_M(0)=c^2u(1+\tfrac{q_b}{rq})\partial_q q_b<0, \qquad P_M(u)=-u(c^2\frac{q_b}{q}+gh)>0,$$
$$\lim_{\lambda\to-\infty}P_M(\lambda)=-\infty \qquad \text{and} \lim_{\lambda\to+\infty}P_M(\lambda)=+\infty.$$

In each case, P_M admits three distinct real roots $\lambda_1,\lambda_2,\lambda_3$ satisfying:

$$\lambda_1\le 0<\lambda_2<u<\lambda_3 \text{ or } \lambda_1<u<\lambda_2\le 0<\lambda_3,$$

which concludes the proof. □

An important property is that if $q_b=0$ and thus $\partial_t z=0$, then the characteristic polynomial $P_M(\lambda)$ is reduced to:

$$P_M(\lambda)=\lambda(u-\lambda-c)(u-\lambda+c);$$

then we recover the eigenvalues of the classical shallow water system that are 0, $u-c$ and $u+c$.

Now we examine the compatibility conditions (3.14) with respect to the hyperbolicity of the system (3.10), for some classical bedload formulae.

Grass bedload formula The Grass formula proposed in Grass 1981 for the sediment transport discharge q_b is given by:

$$q_b=A_g u|u|^{m-1}, \text{ for } 1\le m\le 4 \text{ and } 0<A_g\le 1, \tag{3.16}$$

From this definition, one can quickly check that the compatibility conditions (3.14) are satisfied, since we have

$$\partial_h q_b+u\partial_q q_b=-m\frac{q_b}{h}+mu\frac{q_b}{q}=0, \quad \partial_q q_b=m\frac{q_b}{hu}>0, \text{ and } \frac{q_b}{q}>0.$$

Meyer-Peter & Müller bedload formula In this case q_b is defined as in Meyer-Peter and Muller 1948 by:

$$q_b=8\Gamma\, sgn(\tau)\,(\theta-\theta_c)_+^{\frac{3}{2}}, \tag{3.17}$$

with

$$\theta=\frac{|\tau|}{\rho_w(1/r-1)gd_s} \text{ and } \Gamma=\sqrt{(1/r-1)gd_s^3},$$

where d_s is the main diameter of the grain sediment, θ is a non-dimensional shear stress called also Shields parameter and θ_c is a non-dimensional critical shear stress.

Thus, it depends on the shear stress τ given by:

$$\tau = \rho_w g h S_f, \tag{3.18}$$

where the friction term S_f is explicitly defined by one of the following laws:

$$\textbf{Darcy-Weisbach friction law: } S_f = \frac{f|u|u}{8g}, \tag{3.19}$$

where f the Darcy-Weisbach coefficient or

$$\textbf{Manning friction law: } \qquad S_f = \frac{n_m^2|u|u}{h^{4/3}}, \tag{3.20}$$

where n_m is the Manning coefficient.
Attention is made first to justify that the definition (3.17) fulfills the compatibility conditions (3.14). Let us start by considering the Darcy-Weisbach friction law, then we have

$$\partial_h q_b = -8\Gamma \frac{sgn(\tau)}{h}\left(2(\theta-\theta_c)^{\frac{3}{2}} + 3\theta(\theta-\theta_c)^{\frac{1}{2}}\right)$$

and

$$\partial_q q_b = 8\Gamma \frac{sgn(\tau)}{q}\left(2(\theta-\theta_c)^{\frac{3}{2}} + 3\theta(\theta-\theta_c)^{\frac{1}{2}}\right).$$

Thus we get

$$\partial_q q_b > 0 < \frac{q_b}{q} \text{ and } \partial_h q_b + u\partial_q q_b = 0.$$

Then the system (3.11) associated to (3.19) and (3.17) remains strictly hyperbolic. Unfortunately the conditions (3.14) may fail if we use the Manning friction law and thus the hyperbolicity of the system (3.11) associated to the Meyer-Peter & Müller bedload formula (3.17) using the Manning friction law (3.20) and needs a supplementary condition.

Ashida & Michue bedload formula The Ashida and Michiue 1972 bedload formula Ashida and Michiue 1972 reads:

$$q_b = 17\Gamma\, sgn(\tau)\,(\theta-\theta_c)_+\,(\sqrt{\theta}-\sqrt{\theta_c}), \tag{3.21}$$

where the various parameters appearing here are defined as in the Meyer-Peter&Müller formula. Similarly, under the Darcy-Weisbach friction law (3.19), we can prove that the compatibility conditions (3.14) are fulfilled.

3.3 Limitation of the classical bedload transport

As mentioned in the introduction, even though the classical formulae are easy to implement, they do not guarantee the sediment conservation. Indeed since the sediment flux q_b depends only on the water height and the velocity flow, then q_b may be different from zero even though there is no sediment deposition *i.e.* $z = 0$.
To circumvent this problem, a class of shallow water flow and sediment transport models involving a correction of the bedload transport formula which takes into account the bed elevation, has recently been developed Fowler, Kopteva, and Oakley 2007; Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017; Boittin 2019. Following the work of Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017, the standard shear stress (3.18) is replaced by the following effective shear stress:

$$\tau_{eff} = \frac{\tau}{\rho_w} - \frac{g d_s \nu}{r} \partial_x (rh + z) \quad \text{with} \quad \frac{\tau}{\rho_w} = g h S_f, \tag{3.22}$$

where the physical parameter ν is defined by:

$$\nu = \frac{\theta_c}{tan(\delta)},$$

where δ is a friction angle, chosen empirically. The corrected bedload formula proposed by Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017 which we refer in the sequel as the FMNZ formula, reads:

$$q_b = k \Gamma \, sgn(\tau_{eff}) \, (\theta_{eff} - \theta_c)_+^{3/2}, \tag{3.23}$$

where k is a positive real number and θ_{eff} the effective Shields parameter given by:

$$\theta_{eff} = \frac{|\tau_{eff}|}{(1/r - 1)\, d_s \, g}. \tag{3.24}$$

This formula can be seen as a generalization of the Meyer-Peter & Müler one defined by (3.17) and both formulae of q_b will coincide if $k = 8$ and if we neglect all gravitational effects. Indeed if $\nu = 0$, then the effective shear stress is reduced to: $\tau_{eff} = g h S_f$.
Note also that the angle δ plays a relevant role in the sediment transport process and constitutes a supplementary criterion for the solid transport. In order to emphasize the particularity of the FMNZ formula, we consider a specific regime where the water flow is at rest, so $u = 0$ and $h = z + C$ with $C \in \mathbb{R}_+$. Then after some basic manipulations, the sediment flux boils down to

$$q_b = -k \Gamma \theta_c^{3/2} \, sgn(\partial_x z) \, \Big(\frac{|\partial_x z|}{\tan(\delta)} - 1 \Big)_+^{3/2}.$$

In this case, it is obvious to see that the transport of the material may occur according to the profile of z and the friction angle δ. Indeed if the repose angle of the sediment

is smaller than the friction angle the material moves; otherwise the sediment is not transported, see Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017 for more details. We evidence numerically this behavior in test 3.5.2 below.

3.4 Numerical approximation

The shallow water-Exner equations are frequently discretized in space by collocated finite volumes, using principally a Riemann solver type see *e.g.* Castro Dìaz, Fernàndez-Nieto, and Ferreiro 2008, Fernàndez-Nieto, Lucas, Morales De Luna, et al. 2014, Gunawan and Lhébrard 2015, Berthon, Boutin, and Turpault 2015, Audusse 2018. Here we consider the decoupled staggered scheme proposed in Gunawan, Eymard, and Pudjaprasetya 2015. The novelty here is that we implement the stabilized friction given by (3.9) and the corrected FMNZ formula (3.23).

3.4.1 A decoupled staggered scheme

In order to design a numerical scheme for the equations (3.1-3.2), we discretize the time and space intervals and then we introduce the discrete unknowns.

Uniform discretization of the time interval $(0,T)$**:** Let us consider a partition $0 = t_0 < t_1 < \cdots < t_{N_t} = T$ of the time interval $(0,T)$, which we suppose to be uniform for the sake of simplicity, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \cdots, N_t - 1$ be the (constant) time step.

Uniform discretization of the space interval $(0,L)$**:** Let $0 = x_{\frac{1}{2}} < x_{\frac{1}{2}} < \cdots < x_{N_x+\frac{1}{2}} = L$, such that

$$x_{i+\frac{1}{2}} = x_{i-\frac{1}{2}} + \delta x,\ i = 1, \cdots, N_x - 1 \ \text{ and } \ x_i = \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2},\ i = 1, \cdots, N_x.$$

The discrete unknowns corresponding to the velocity u, the water height h and the sediment depth z are denoted by $u_{i+\frac{1}{2}}^{n+1}$, h_i^{n+1} and z_i^{n+1} respectively.

The numerical scheme for the equations (3.10) reads:

Initialization level:

$$h_i^0 = \frac{1}{\delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} h_0(x)\,dx,$$
$$u_{i+\frac{1}{2}}^0 = \frac{1}{\delta x}\int_{x_i}^{x_{i+1}} u_0(x)\,dx,$$
$$z_i^0 = \frac{1}{\delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} z_0(x)\,dx.$$

Current level: Compute h^{n+1}, u^{n+1} and z^{n+1}:

$$h_i^{n+1} = h_i^n - \frac{\delta t}{\delta x}\left((hu)_{i+\frac{1}{2}}^n - (hu)_{i-\frac{1}{2}}^n\right), \qquad \forall\, i = 1, \cdots, N_x, \tag{3.25a}$$

$$h_{i+\frac{1}{2}}^{n+1}\, u_{i+\frac{1}{2}}^{n+1} = h_{i+\frac{1}{2}}^n\, u_{i+\frac{1}{2}}^n - \frac{\delta t}{\delta x}\left((hu\, u)_{i+1}^n - (hu\, u)_i^n + g h_{i+\frac{1}{2}}^{n+1}(h_{i+1}^{n+1} - h_i^{n+1} + z_{i+1}^n - z_i^n)\right)$$
$$- \delta t\, g h_{i+\frac{1}{2}}^{n+1}\, S_{f,i+\frac{1}{2}}^{n,n+1}, \quad \forall\, i = 1, \cdots, N_x - 1, \tag{3.25b}$$

$$z_i^{n+1} = z_i^n - \frac{1}{1-\phi}\frac{\delta t}{\delta x}(q_{b,i+\frac{1}{2}}^{n+1} - q_{b,i-\frac{1}{2}}^{n+1}), \qquad \forall\, i = 1, \cdots, N_x, \tag{3.25c}$$

where the various discrete terms and operators are defined below.

Discrete water flow quantities – the numerical fluxes are approximated by the upwind technique as follows:

– Mass flux:

$$(hu)_{i+\frac{1}{2}} = \begin{cases} u_{i+\frac{1}{2}}\, h_i, & \text{if } u_{i+\frac{1}{2}} \geq 0, \\ u_{i+\frac{1}{2}}\, h_{i+1}, & \text{otherwise}, \end{cases}$$

– Momentum flux:

$$(hu\, u)_{i+1} = \begin{cases} u_{i+\frac{1}{2}}\, (hu)_{i+1} & \text{if } (hu)_{i+1} \geq 0, \\ u_{i+\frac{3}{2}}\, (hu)_{i+1} & \text{otherwise}, \end{cases}$$

where the term $(hu)_{i+1}$ is interpolated as in Gunawan, Eymard, and Pudjaprasetya 2015; Herbin, Latché, and Nguyen 2018 by:

$$(hu)_{i+1} = \frac{1}{2}\left[h_{i+\frac{1}{2}}\, u_{i+\frac{1}{2}} + h_{i+\frac{3}{2}}\, u_{i+\frac{3}{2}}\right], \quad \text{with} \quad h_{i+\frac{1}{2}} = \frac{1}{2}(h_i + h_{i+1}).$$

Note that a discrete dual mass balance holds, which reads:

$$h_{i+\frac{1}{2}}^{n+1} = h_{i-\frac{1}{2}}^n - \frac{\delta t}{\delta x}\left((hu)_{i+1}^n - (hu)_i^n\right), \qquad \forall\, i = 1, \cdots, N_x. \tag{3.26}$$

Discrete friction law – We implement here the Manning and stabilized friction laws :

– Manning friction:

$$S_{f,i+\frac{1}{2}}^{n,n+1} = n_m^2\, \frac{|u_{i+\frac{1}{2}}^n|\, u_{i+\frac{1}{2}}^{n+1}}{(h_{i+\frac{1}{2}}^{n+1})^{4/3}}$$

– Stabilized friction:

$$S_{f,i+\frac{1}{2}}^{n,n+1} = \begin{cases} \dfrac{r(h_{i+1}^{n+1} - h_i^{n+1}) + z_{i+1}^n - z_i^n}{\delta x} \dfrac{q_{b,i+\frac{1}{2}}^{n,n+1}}{r h_{i+\frac{1}{2}}^{n+1} u_{i+\frac{1}{2}}^n}, & \text{if } u_{i+\frac{1}{2}}^n \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

where $q_{b,i+\frac{1}{2}}^{n,n+1} = q_b(h_i^{n+1}, u_{i+\frac{1}{2}}^n)$

Discrete sediment flux– Three different formulae are are tested for the sediment flux:

– Grass formula:

$$q_{b,i+\frac{1}{2}}^{n+1} = A_g u_{i+\frac{1}{2}}^{n+1} |u_{i+\frac{1}{2}}^{n+1}|^{m-1}.$$

– Meyer-Peter & Müller formula:

$$q_{b,i+\frac{1}{2}}^{n+1} = 8\,\Gamma\, sgn(u_{i+\frac{1}{2}}^{n+1}) \left(\theta_{i+\frac{1}{2}}^{n+1} - \theta_c\right)_+^{3/2}$$

with

$$\theta_{i+\frac{1}{2}}^{n+1} = \frac{n_m^2}{(\frac{1}{r} - 1) d_s} \frac{|u_{i+\frac{1}{2}}^{n+1}|^2}{(h_{i+\frac{1}{2}}^{n+1})^{2/3}}.$$

– FMNZ formula:

$$q_{b,i+\frac{1}{2}}^{n,n+1} = k\,\Gamma\, sgn(\tau_{eff,\, i+\frac{1}{2}}^{n,n+1}) \left(\theta_{eff,\, i+\frac{1}{2}}^{n,n+1} - \theta_c\right)_+^{3/2}$$

with

$$\theta_{eff,\, i+\frac{1}{2}}^{n,n+1} = \frac{\left|\tau_{eff,\, i+\frac{1}{2}}^{n,n+1}\right|}{(1/r - 1) g d_s}$$

and

$$\tau_{eff,\, i+\frac{1}{2}}^{n,n+1} = g n_m^2 \frac{|u_{i+\frac{1}{2}}^{n+1}| u_{i+\frac{1}{2}}^{n+1}}{(h_{i+\frac{1}{2}}^{n+1})^{1/3}} - \frac{g d_s \nu}{r} \frac{r(h_{i+1}^{n+1} - h_i^{n+1}) + z_{i+1}^n - z_i^n}{\delta x}.$$

3.4.2 Stability of the numerical scheme

The decoupled scheme (3.25) is known to maintain the positivity of the water height under the following CFL type condition:

$$\left(\sum_{i=1}^{N_x - 1} |u_{i+\frac{1}{2}}^n|\right) \delta t \leq \delta x.$$

Furthermore the well-balanced behavior with respect to the preservation of the lake at rest steady state holds. Indeed,

$$\text{if } \begin{cases} u^n_{i+\frac{1}{2}} = 0, \\ h^n_i + z^n_i = C, \text{ with } C \in \mathbb{R}_+, \end{cases} \qquad \text{then} \qquad \begin{cases} u^{n+1}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}}, \\ h^{n+1}_i = h^n_i, \\ z^{n+1}_i = z^n_i. \end{cases}$$

The purpose hereafter is to show that discrete potential and kinetic energies hold for this proposed scheme. The following result states that the scheme satisfies a discrete counterpart of the kinetic energy.

Lemma 3.1 (Discrete kinetic energy)**.** *A solution to the scheme* (3.25) *satisfies the following equality, for* $n \in \{0, \cdots, N_t - 1\}$ *and* $i \in \{1, \cdots, N_x - 1\}$*:*

$$\begin{aligned} \frac{1}{\delta t}\frac{1}{2} r\big(h^{n+1}_{i+\frac{1}{2}}(u^{n+1}_{i+\frac{1}{2}})^2 - h^n_{i+\frac{1}{2}}(u^n_{i+\frac{1}{2}})^2\big) + \frac{1}{\delta x} r\,\frac{1}{2}((hu\,u^2)^n_{i+1} - (hu\,u^2)^n_i) \\ + g r h^{n+1}_{i+\frac{1}{2}}\, u^{n+1}_{i+\frac{1}{2}}\, \frac{1}{\delta x}\,(h^{n+1}_{i+1} + z^n_{i+1} - h^{n+1}_i - z^n_i) \\ = -r g h^{n+1}_{i+\frac{1}{2}}\, S^{n,n+1}_{i+\frac{1}{2}}\, u^{n+1}_{i+\frac{1}{2}} - R^{n+1}_{i+\frac{1}{2}}, \end{aligned} \tag{3.27}$$

where $(hu\,u^2)^n_i = (hu)^n_i\,(u^n_{i-\frac{1}{2}})^2$ *if* $(hu)^n_i \geq 0$ *or* $(hu\,u^2)^n_i = (hu)^n_i\,(u^n_{i+\frac{1}{2}})^2$ *otherwise and* $R^{n+1}_{i+\frac{1}{2}} \geq 0$ *under the CFL like restriction:*

$$\forall\ 1 \leq i \leq N_x - 1, \qquad \delta t \leq \frac{\delta x\, h^{n+1}_{i+\frac{1}{2}}}{((hu)^n_{i+1})^- + ((hu)^n_i)^-}. \tag{3.28}$$

Sketch of proof. The computations mimic the technique used in the continuous problem see (3.5); they consist in multiplying the discrete momentum equation by $r\,u^{n+1}_{i+\frac{1}{2}}$ and using the discrete mass balance on the dual cells (3.26). The remainder of the proof is an easy adaptation of Herbin, Latché, and Nguyen 2018, Lemma 3.2.

A discrete potential energy is also satisfied by the scheme.

Lemma 3.2 (Discrete potential energy)**.** *A solution to the scheme* (3.25) *satisfies the following inequality, for* $i \in \{1, \cdots, N_x\}$ *and* $n \in \{0, \cdots, N_t - 1\}$*:*

$$\begin{aligned} \frac{1}{\delta t}\big((E_p)^{n+1}_i - (E_p)^n_i\big) + \frac{1}{\delta x}\Big[\frac{1}{2} r g((h^2 u)^n_{i+\frac{1}{2}} - (h^2 u)^n_{i-\frac{1}{2}}) + \frac{1}{2} r g (h^n_i)^2 (u^n_{i+\frac{1}{2}} - u^n_{i-\frac{1}{2}})\Big] \\ + \frac{g}{1-\phi}\frac{1}{\delta x}\,((q_b)^{n,n+1}_{i+\frac{1}{2}} - (q_b)^{n,n+1}_{i+\frac{1}{2}})(r h^{n+1}_i + z^{n+1}_i) \\ + r g z^n_i\,\frac{1}{\delta x}((hu)^n_{i+\frac{1}{2}} - (hu)^n_{i-\frac{1}{2}}) \leq -r g \frac{1}{\delta x}\,((hu)^n_{i+\frac{1}{2}} - (hu)^n_{i-\frac{1}{2}})(h^{n+1}_i - h^n_i), \end{aligned} \tag{3.29}$$

with $(E_p)_i = \frac{1}{2} r g h_i^2 + \frac{1}{2} g z_i^2 + r g h_i z_i$ *and* $(h^2u)^n_{i+\frac{1}{2}} = h_i^2 u^n_{i+\frac{1}{2}}$ *if* $u^n_{i+\frac{1}{2}} \geq 0$ *or* $(h^2u)^n_{i+\frac{1}{2}} = h^2_{i+1} u^n_{i+\frac{1}{2}}$ *otherwise.*

Proof. Multiplying the discrete mass equation (3.25a) by $r g h_i^{n+1}$ and following the proof of Lemma 1.4 yields

$$\frac{1}{\delta t}\frac{1}{2} g r \left((h_i^{n+1})^2 - (h_i^n)^2\right) + \frac{1}{\delta x}\frac{1}{2} r g ((h^2u)^n_{i+\frac{1}{2}} - (h^2u)^n_{i-\frac{1}{2}}) + \frac{1}{2} r g (h_i^n)^2 \frac{1}{\delta x}(u^n_{i+\frac{1}{2}} - u^n_{i-\frac{1}{2}}) = -(R_h)_i^{n+1},$$

with

$$(R_h)_i^{n+1} = \frac{1}{\delta t}\frac{1}{2} r g (h_i^{n+1} - h_i^n)^2 + \frac{rg}{\delta x}\left[(h^n_{i+1} - h_i^n)^2 ((hu)^n_{i+\frac{1}{2}}))^- + (h^n_{i-1} - h_i^n)^2 ((hu)^n_{i-\frac{1}{2}}))^-\right] + \frac{rg}{\delta x}\left((hu)^n_{i+\frac{1}{2}} - (hu)^n_{i-\frac{1}{2}}\right)(h_i^{n+1} - h_i^n)$$

Since $((hu)^n_{i-\frac{1}{2}}))^- \geq 0$, we get

$$(R_h)_i^{n+1} \geq r g \frac{1}{\delta x}\left((hu)^n_{i+\frac{1}{2}} - (hu)^n_{i-\frac{1}{2}}\right)(h_i^{n+1} - h_i^n).$$

Then, multiplying the discrete Exner equation (3.25c) by $g(r h_i^{n+1} + z_i^{n+1})$ we get:

$$\frac{1}{\delta t}\frac{1}{2} g\left((z_i^{n+1})^2 - (z_i^n)^2\right) + \frac{1}{\delta t}(z_i^{n+1} - z_i^n) r g h_i^{n+1} + \frac{1}{1-\phi} g (r h_i^{n+1} + z_i^{n+1}) \frac{1}{\delta x}\left((q_b)^{n,n+1}_{i+\frac{1}{2}} - (q_b)^{n,n+1}_{i+\frac{1}{2}}\right) = -\frac{1}{\delta t}\frac{1}{2} g (z_i^{n+1} - z_i^n)^2,$$

where we use the identity $2ab = a^2 + b^2 - (a-b)^2$.
Then the second term of the left hand-side can be written as follows:

$$\frac{1}{\delta t}(z_i^{n+1} - z_i^n) r g h_i^{n+1} = \frac{1}{\delta t} r g \left(z_i^{n+1} h_i^{n+1} - z_i^n h_i^n\right) - \frac{1}{\delta t}(h_i^{n+1} - h_i^n) r g z_i^n.$$

Thus thanks to the discrete mass equation, we find

$$-\frac{1}{\delta t}(h_i^{n+1} - h_i^n) r g z_i^n = r g z_i^n \frac{1}{\delta x}((hu)^n_{i+\frac{1}{2}} - (hu)^n_{i-\frac{1}{2}}).$$

Finally reordering all the terms we get the desired result which concludes the proof. □

At this stage, following the technique introduced in Herbin, Latché, Nasseri, et al. 2019, a discrete local energy defined on a primal cell i is obtained from a discrete potential and a discrete local kinetic energy on the cell i. We denote by $(E_k)_i$ the local

kinetic energy on the cell $]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[$ defined by:

$$(E_k)_i = \frac{1}{2} r \Big(\frac{1}{2} h_{i+\frac{1}{2}} (u^2)_{i+\frac{1}{2}} + \frac{1}{2} h_{i-\frac{1}{2}} (u^2)_{i-\frac{1}{2}} \Big).$$

Thanks to (3.27) and under the condition (3.28), $(E_k)_i$ satisfies the following discrete inequality:

$$\begin{aligned}
\frac{1}{\delta t}((E_k)_i^{n+1} - (E_k)_i^n) + \frac{1}{\delta x}\frac{1}{4} r((hu\,u^2)_{i+1}^n - (hu\,u^2)_i^n) + \frac{1}{\delta x}\frac{1}{4} r((hu\,u^2)_i^n - (hu\,u^2)_{i-1}^n) \\
+ \frac{1}{\delta x}\frac{1}{2} rg h_{i+\frac{1}{2}}^{n+1} u_{i+\frac{1}{2}}^{n+1} (h_{i+1}^{n+1} + z_{i+1}^n - h_i^{n+1} - z_i^n) \\
+ \frac{1}{\delta x}\frac{1}{2} rg h_{i-\frac{1}{2}}^{n+1} u_{i-\frac{1}{2}}^{n+1} (h_i^{n+1} + z_i^n - h_{i-1}^{n+1} - z_{i-1}^n) \\
\leq -\frac{1}{2} rg (h_{i+\frac{1}{2}}^{n+1} S_{i+\frac{1}{2}}^{n,n+1} u_{i+\frac{1}{2}}^{n+1} + h_{i-\frac{1}{2}}^{n+1} S_{i-\frac{1}{2}}^{n,n+1} u_{i-\frac{1}{2}}^{n+1}).
\end{aligned}$$

Then gathering this inequality with the equation (3.29), we obtain the following discrete energy inequality:

$$\begin{aligned}
\frac{\delta x}{\delta t}\Big((E_p)_i^{n+1} + (E_k)_i^{n+1} - (E_p)_i^n - (E_k)_i^n\Big) + \sum_{j\in\{1,-1\}} jr\,\frac{1}{2}\Big(\frac{1}{2}(hu\,u^2)_{i+j}^n - \frac{1}{2}(hu\,u^2)_i^n\Big) \\
+ \sum_{j\in\{1,-1\}} j\frac{1}{2} rg\Big((h^2u)_{i+\frac{j}{2}}^n + \frac{1}{2}((h_{i+j}^n)^2 + (h_i^n)^2 u_{i+\frac{j}{2}}^n + (z_{i+j}^n + z_i^n)\,(hu)_{i+\frac{j}{2}}^n\Big) \\
+ \frac{1}{1-\phi}\sum_{j\in\{1,-1\}} jg\frac{1}{2}(rh_{i+j}^{n+1} + z_{i+j}^{n+1} + rh_i^{n+1} + z_i^{n+1})\,(q_b)_{i+\frac{j}{2}}^{n,n+1} \leq -R_i^{n+1},
\end{aligned}$$

with

$$\begin{aligned}
R_i^{n+1} \geq \sum_{j\in\{1,-1\}} j\,\frac{1}{2} gr\Big(h_{i+\frac{j}{2}}^{n+1} u_{i+\frac{j}{2}}^{n+1}(h_{i+j}^{n+1} - h_i^{n+1}) - (hu)_{i+\frac{j}{2}}^n (h_{i+j}^n - h_i^n)\Big) \\
+ \sum_{j\in\{1,-1\}} j\,\frac{1}{2} gr(h_{i+\frac{j}{2}}^{n+1} u_{i+\frac{j}{2}}^{n+1} - (hu)_{i+\frac{j}{2}}^n)(z_{i+j}^n - z_i^n) + gr\sum_{j\in\{1,-1\}} j\,(hu)_{i+\frac{j}{2}}^n (h_i^{n+1} - h_i^n) \\
-\frac{1}{1-\phi}\sum_{j\in\{1,-1\}} j\frac{1}{2} g(rh_{i+j}^{n+1} + z_{i+j}^{n+1} - rh_i^{n+1} - z_i^{n+1})(q_b)_{i+\frac{j}{2}}^{n,n+1} + \delta x \sum_{j\in\{1,-1\}} \frac{1}{2} rg h_{i+\frac{j}{2}}^{n+1} S_{i+\frac{j}{2}}^{n,n+1} u_{i+\frac{j}{2}}^{n+1},
\end{aligned}$$

and where the identity $2a_i = (a_{i+j} + a_i) - (a_{i+j} - a_i)$ is used several times for any discrete scalar unknown $(a_i)_{1,\cdots,N_x}$.
Unfortunately the current definition of the discrete stabilized friction $S_{i+\frac{j}{2}}^{n,n+1}$ does not satisfy the following inequality:

$$\delta x\, rh_{i+\frac{j}{2}}^{n+1} S_{i+\frac{j}{2}}^{n,n+1} u_{i+\frac{j}{2}}^{n+1} - \frac{1}{1-\phi}(rh_{i+j}^{n+1} + z_{i+j}^{n+1} - rh_i^{n+1} - z_i^{n+1})\,(q_b)_{i+\frac{j}{2}}^{n,n+1} \geq 0;$$

because of the segregated time discretization taken on the term S_f.

3.5 Numerical experiments

Numerical experiments are now performed to test the performance of the decoupled scheme (3.25) and verify the behavior of the algebraic model. We evidence numerically the limitations of the classical formulae with respect to the FMNZ formula.

3.5.1 Test 1: transcritical steady state

We consider in this first test a steady state solution of the shallow water equation posed on the domain $[0,10]$. The initial data are defined as in Fernàndez-Nieto, Lucas, Morales De Luna, et al. 2014, by:

$$\begin{cases} (hu)(x,0) = 0.6 \\ z(x,0) = 0.1 + 0.1\ e^{-(x-5)^2}, \\ h(x,0) + z(x,0) = 0.4 \end{cases}$$

We set $(hu)(x=0,0) = 0.6$ for the left boundary conditions and $h(x=10,0) = 0.4$ on the right, the other boundaries are free.

Firstly we let the water flow at this steady state without friction effect, and with no evolution of the sediment layer taken into account. Hence only the shallow water equations are solved. After a simulation time $T \geq 15$ the numerical solution reaches an equilibrium regime known as the transcritical steady state where the wave speed is equal to the flow velocity. The result is plotted in Figure 3.3 below: After this regime we

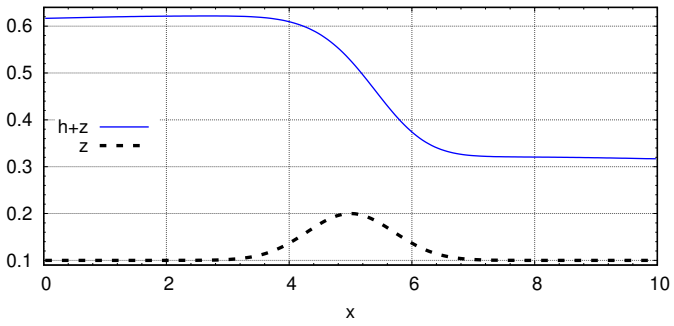
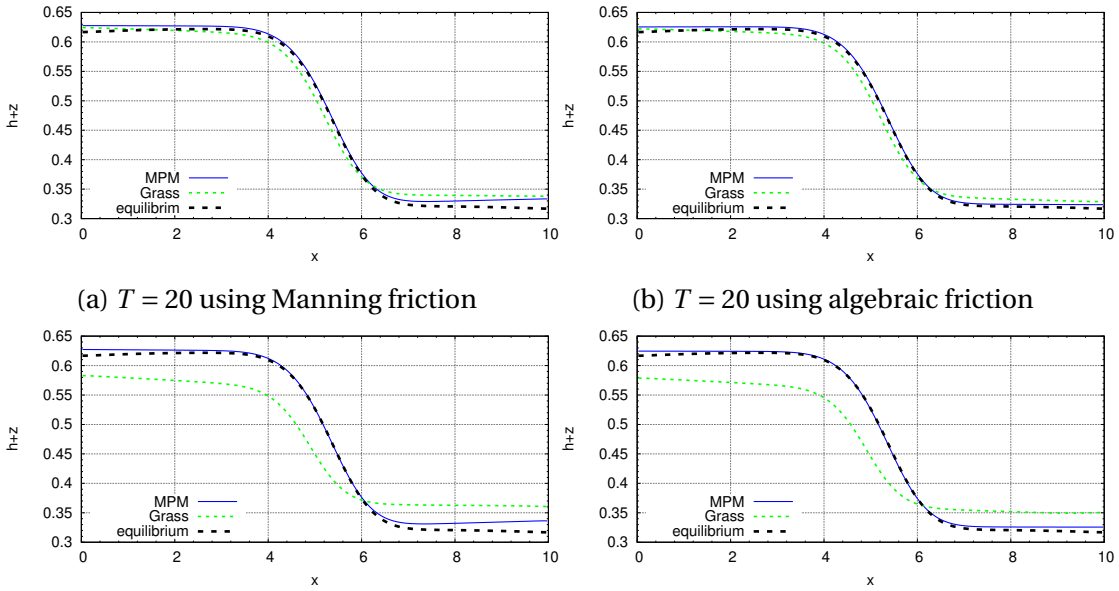


Figure 3.3 – Transcritical steady state: Free surface at time $T = 15$ with $\delta t = \delta x/5$ using 200 cells.

activate the evolution of the sediment bed solving the Exner equation with the Grass and Meyer-Peter & Müller formulae to compute sediment bedload q_b. The physical

parameters are chosen $r = 0.34$, $d_s = 0.001$, $\theta_c = 0.047$, $n_m = 0.01$, $Ag = 0.005$ and $m = 3$. The purpose is to compare the numerical solution obtained using the stabilized friction term S_f against the one computed by the Manning friction law. Figures 3.4 and 3.5 illustrate respectively the evolution of the free surface and sediment depth after the equilibrium state taking into account the kind of sediment transport flux at time $T = 20$ and $T = 30$.



(a) $T = 20$ using Manning friction

(b) $T = 20$ using algebraic friction

(c) $T = 30$ using Manning friction

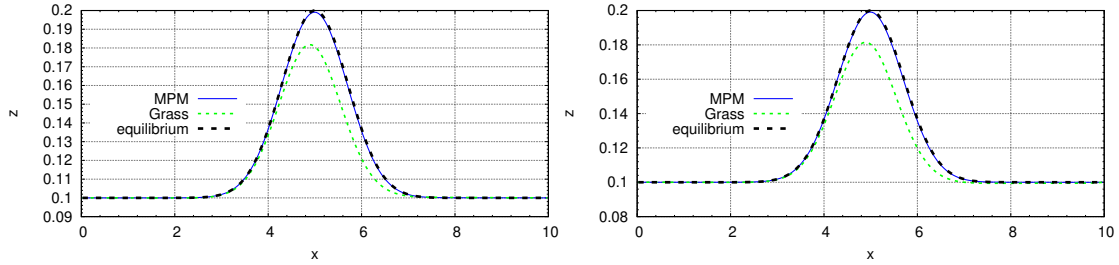
(d) $T = 30$ using algebraic friction

Figure 3.4 – Classical models: Free surface after the steady state regime with friction effect.

The figures 3.4 and 3.5 show very similar results obtained with the Manning friction term compared to the results computed from the stabilized friction. This good agreement encourages the use of a regularized friction term. One can also observe that the erosion rate caused by the Grass formula is more important and grows significantly with respect to the time compared to the Meyer-Peter & Müller formula since for the Grass formula the sediment transport process begins at the same time as the water flow.

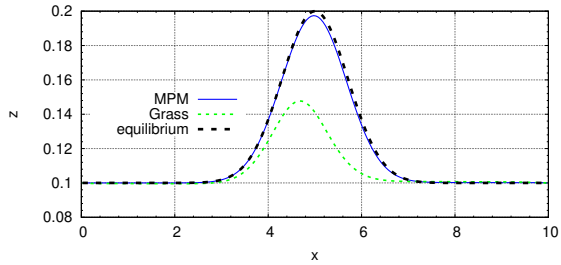
3.5.2 Test 2: inaccuracy of the classical bedload formulae

The attempt of this test is to confirm that the FMNZ formula is relevant and more efficient than the classical formulae to simulate a sediment transport motion even with a different discretization scheme than that of Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017. To further understand the limitation of the usual bedload transport, we keep the water flow at rest that means we set $u = 0$ and $h_0(x) = 0.2 - z_0(x)$

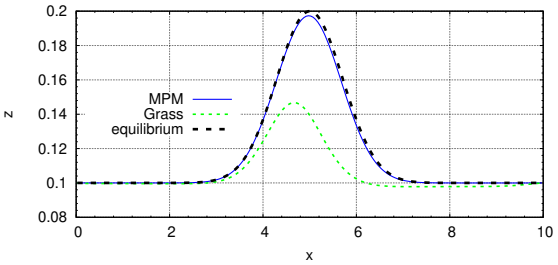


(a) $T = 20$ using Manning friction

(b) $T = 20$ using algebraic friction



(c) $T = 30$ using Manning friction



(d) $T = 30$ using algebraic friction

Figure 3.5 – Classical models: Evolution of the sediment depth after the steady state regime with friction effect.

where the profile of $z_0(x)$ is defined on the domain $[0,1]$ by:

$$z_0(x) = \begin{cases} x - \frac{1}{3} & \text{if } \frac{1}{3} \leq x \leq \frac{1}{2}, \\ \frac{2}{3} - x & \text{if } \frac{1}{2} \leq x \leq \frac{2}{3}, \\ 0 & \text{otherwise} \end{cases}$$

The computation uses the physical parameters $d_s = 0.001$, $r = 0.34$, $\theta_c = 0.047$, $n_m = 0.01$ and $k = 10$ and the Meyer-Peter & Müller (MPM) and FMNZ formulae for varying friction angles δ. Figure 3.6 shows the evolution of the sediment depth when the water flow is at rest computed from the MPM and FMNZ formulae for the angles $\delta = 50$ and $\delta = 30$. As one should observe in Figure 3.6, the MPM formula refers to the initial legend, leads to a fixed bed when the water flow is at rest unlike the FMNZ formula. For this latter, it appears that for a friction angle smaller than 45 which corresponds to the repose angle of the initial profile, the sediment is transported while for a value of δ larger than 45 the sediment bed does not move.

3.5.3 Test 3: adapted boundary conditions test case

The objective of the present test is to highlight the influence of the boundary conditions used in the computation of the water flow. We consider the initial data used by Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017 where the sediment

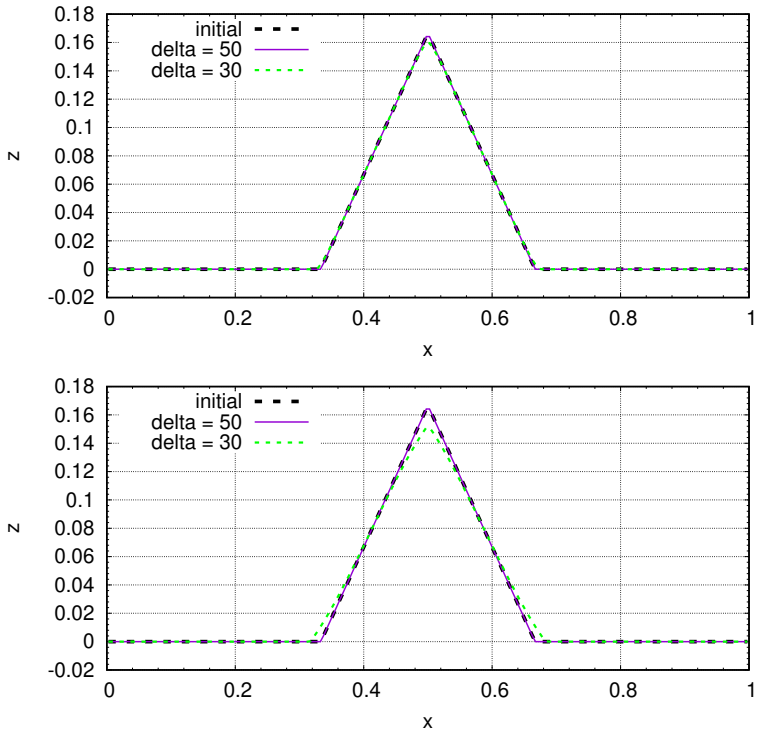


Figure 3.6 – Water flow at rest: Evolution of the sediment depth at time $T = 5$ on the top and $T = 50$ on the bottom using 200 cells and with $\delta t = \delta x/5$

depth is given by:

$$z_0(x) = \begin{cases} 0.1 + 0.1\left(1 + \cos\left(\frac{x-0.4}{0.2}\pi\right)\right) & \text{if } x \in [0.2, 0.6] \\ 0.1 & \text{otherwse} \end{cases}$$

The others initial conditions are $h_0(x) = 1.1 - z_0(x)$ and $(hu)(x,0) = 1.4$ for $x \in [0,1]$. The right boundary conditions are kept at $h(1,t) = 1$, $u(1,t) = 1.4/h(1,t)$ and for the left boundary $(hu)(0,t) = 1.4$. The friction angle is set to $\delta = 45$ while the others physical parameters remain unchanged. According to the friction term the obtained results at time $T = 200$ are shown in Figures 3.7 and 3.8 above. Both figures illustrate more the influence of the boundary conditions for the computation of the water height and sediment depth. These results show also that both Manning and stabilized frictions produce similar features by means of a suitable boundary conditions for a regular sediment bed. This observation leads us to think that the kind of friction source term does not play a relevant role in the sediment transport process.

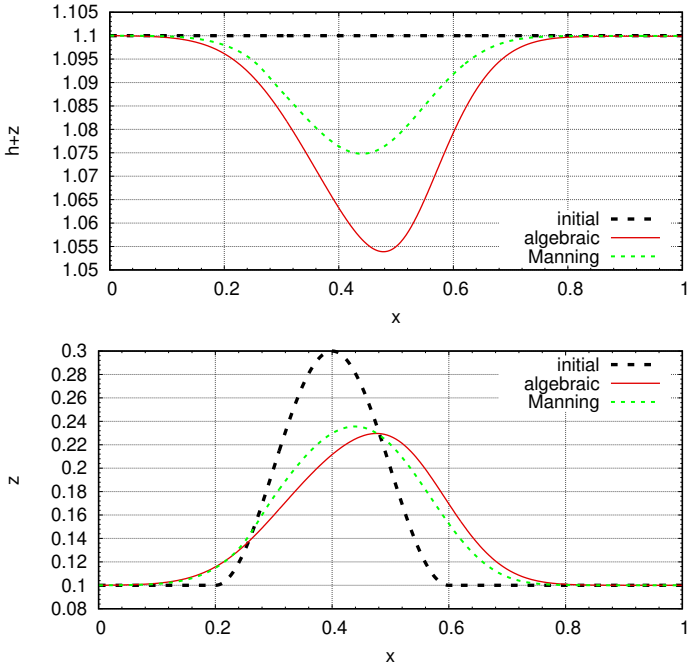


Figure 3.7 – Free boundary conditions: on the top the free surface and sediment depth on the bottom using 800 cells and with $\delta t = \delta x/5$.

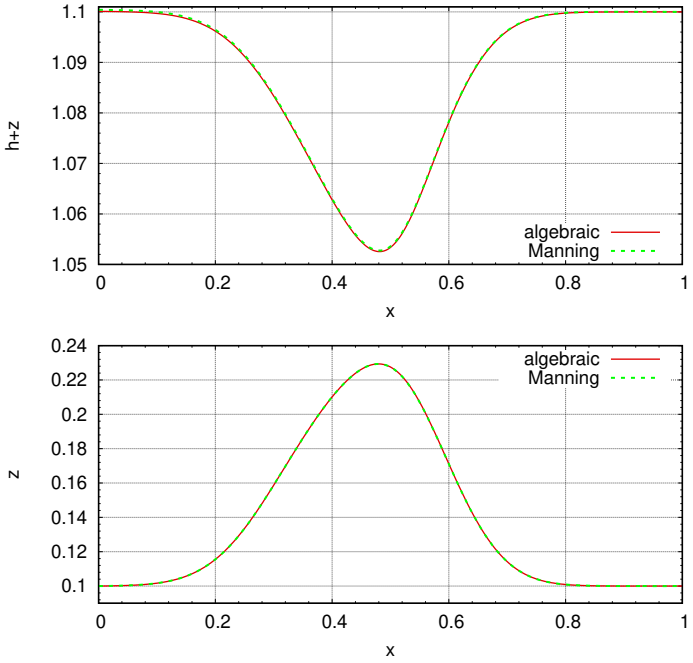


Figure 3.8 – Adapted boundary conditions: on the top the free surface and sediment depth on the bottom using 800 cells and with $\delta t = \delta x/5$.

3.5.4 Test 4: discontinuity movable bed

Now we consider also the test case proposed in Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017 which consists in a discontinuity movable bed where the sediment depth and the height are initialized on the domain $[0,10]$ as follows:

$$\begin{cases} (hu)(x,0) = 1.5 \\ h(x,0) = 1 - z(x,0) \\ z(x,0) = \begin{cases} 0.2 & \text{if } 4 \le x \le 6 \\ 0.1 & \text{otherwise} \end{cases} \end{cases}$$

The computation uses 800 cells with a constant time set $\delta t = \delta x/5$ and with the same

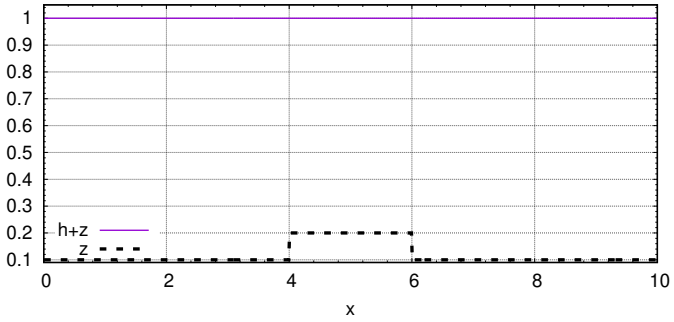


Figure 3.9 – Profile of the free surface and sediment depth at time $T = 0$

physical parameters given above. The boundary conditions are $(hu)(0,t) = 1.5$ for the left and $h(10,t) = 1$, $u(10,t) = 1.5/h(10,t)$ for the right. At time $T = 2000$ and for varying the friction angles we get the following results:
In Figure 3.10 (resp Figure 3.11) one can remark that the erosion rate is more important for smaller values of the friction angle δ. We see also that the friction terms Manning or stabilized do not have a notable difference for the sediment depth as for the case of regular bed shown in Figure 3.8. While for the free surface the Manning friction produces a slightly high height that the algebraic one tries to stabilize.

Below we consider the same configuration varying only the parameter k. The goal is to show the correspondance between the MPM formula of the sediment transport and the FMNZ formula for certain parameters. We can see in Figure 3.11 that the results computed from the MPM formula match perfectly the ones obtained by the FMNZ formula for $\delta = 89$ and $k = 8$ as emphasized in Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017.

Finally we perform a short comparison between the result obtained by the present staggered scheme with an algebraic friction term against to the result presented in Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017. Therein the authors

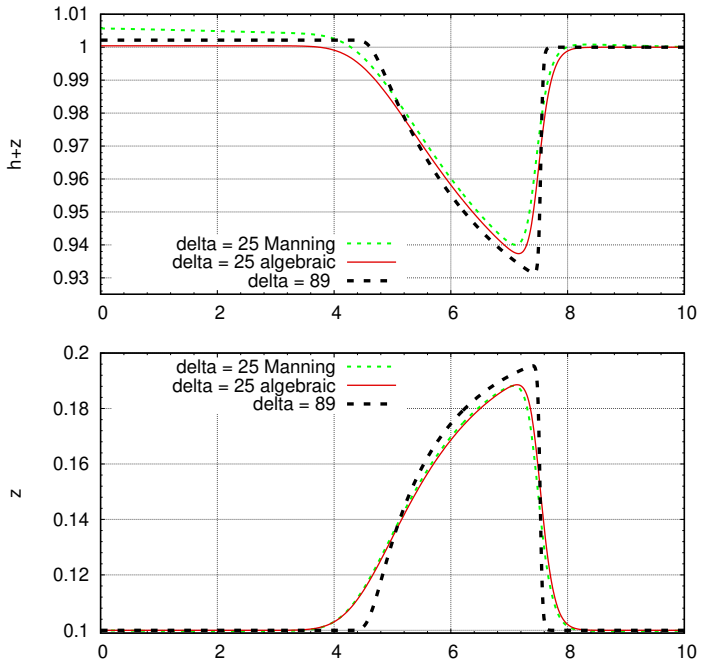


Figure 3.10 – Movable discontinuity bed: Free surface on the top and sediment depth on the bottom for $k = 10$

propose a stabilized friction type term resulting of a formal deduction of the Saint-Venant-Exner model. In addition they use an approximate Riemann type solver developed in Castro Dìaz, Fernàndez-Nieto, and Ferreiro 2008 for the numerical approximation. Doing so, a screen capture from the reference paper Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017 showing the evolution of the sediment layer is presented here. In Figure 3.12 below we show the results for both staggered and approximate Riemann solver strategies by means of a zoom of the regions of high variations of the sediment depth. As one can observe both figures illustrate:

- the equivalence between "classical" definition which refers to the MPM formula and FMNZ formula for the parameters $\delta = 89$ and $k = 8$.
- the front shock progress in the downstream for the friction angle $\delta = 89$ and the commonly behavior for the others values of δ.

Note that in these simulations the friction terms (algebraic or formal deduction) yield similar results. However the sediment transport flux formula plays a relevant role and yields sediment profiles which are more realistic: see Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017 for a comparison with experimental data. We note also the improvement of the decoupled staggered scheme for this complex test case.

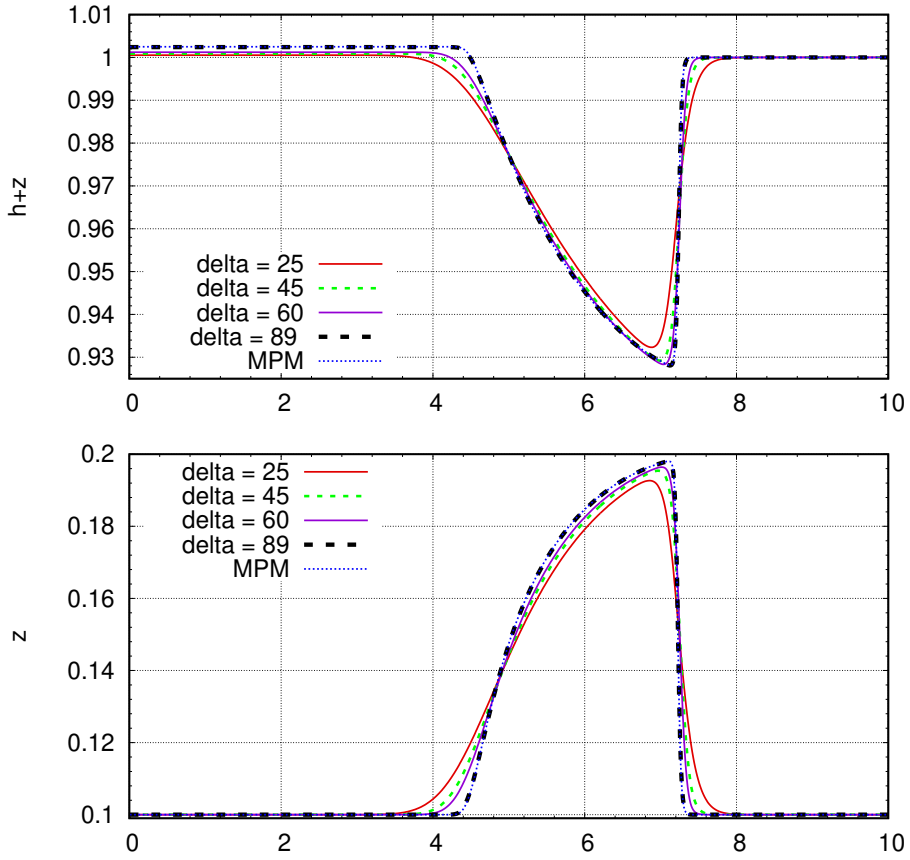


Figure 3.11 – Link between MPM&FMNZ: Free surface on the top and sediment depth on the bottom resulting of an algebraic friction with $k = 8$

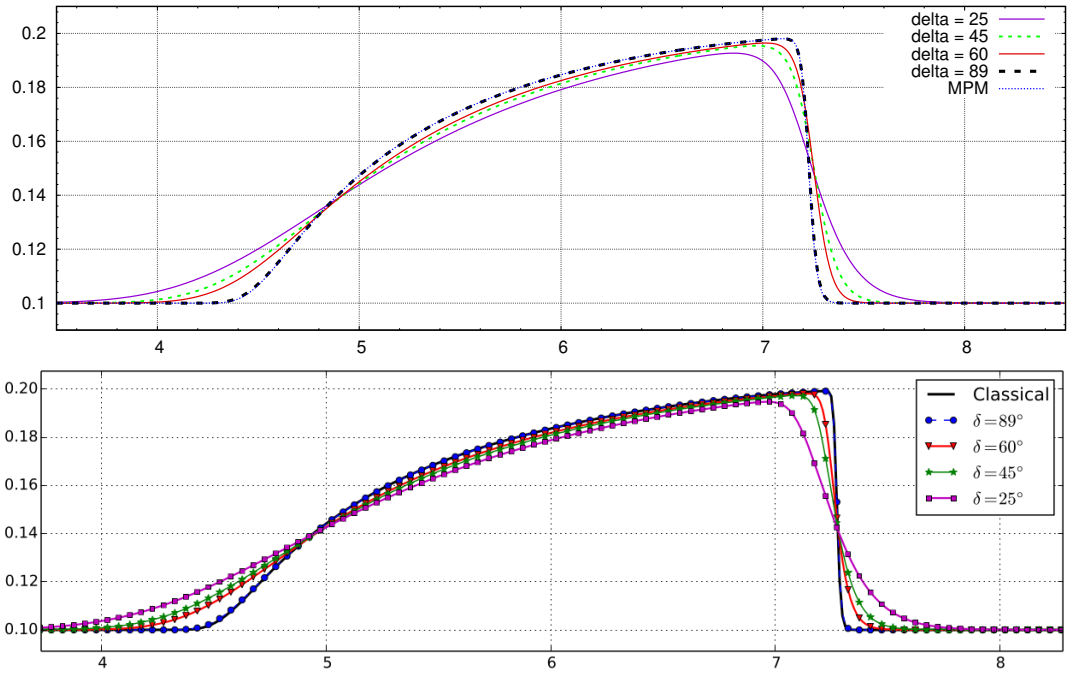


Figure 3.12 – Comparison with the literature: Zoom in the zones of high variation of the sediment depth. On the top: results of the staggered scheme with an algebraic friction; below: results from FMNZ with a Riemann type solver and an alternative corrected friction term on the bottom. The legend "classical" refers to the "MPM" formula.

4 A staggered pressure correction numerical scheme to compute a travelling reactive interface in a partially premixed mixture

Sommaire

Abstract. We address in this chapter a model for the simulation of turbulent deflagrations in industrial applications. The flow is governed by the Euler equations for a variable composition mixture and the combustion modelling is based on a phenomenological approach: the flame propagation is represented by the transport of the characteristic function of the burnt zone, where the chemical reaction is complete; outside this zone, the atmosphere remains in its fresh state. Numerically, we approximate this problem by a penalization-like approach, *i.e.* using a finite conversion rate with a characteristic time tending to zero with the space and time steps. The numerical scheme works on staggered, possibly unstructured, meshes. The time-marching algorithm is of segregated type, and consists in solving in a first step the chemical species mass balances and then, in a second step, mass, momentum and energy balances. For this latter stage of the algorithm, we use a pressure correction technique, and solve a balance equation for the so-called sensible enthalpy instead of the total energy balance, with corrective terms for consistency. The scheme is shown to satisfy the same stability properties as the continuous problem: the chemical species mass fractions are kept in the $[0,1]$ interval, the density and the sensible internal energy stay positive and the integral over the computational domain of a discrete total energy is conserved. In addition, we show that the scheme is in fact conservative, *i.e.* that its solution satisfy a conservative discrete total energy balance equation, with space and time discretizations which are unusual but consistent in the Lax-Wendroff sense. Finally, we observe numerically that the penalization procedure converges, *i.e.* that making the chemical time scale tend to zero allows to converge to the solution of the target (infinitely fast chemistry) continuous problem. Tests also evidence that the scheme accuracy dramatically depends on the discretization of the convection operator in the chemical species mass balances.

Keywords Finite-volume scheme, staggered discretization, pressure correction, compressible flows, reactive flows.

4.1 Problem position

In this paper, we study a numerical scheme for the computation of large scale turbulent deflagrations occurring in a partially premixed atmosphere. In usual situations, such a physical phenomena is driven by the progress in the atmosphere of a shell-shaped thin zone, where the chemical reaction occurs and which thus separates the burnt area from fresh gases; this zone is called the flame brush. The onset of the chemical reaction is due to the temperature elevation, so the displacement of the flame brush is driven by the heat transfers inside and in the neighbour of this zone. Modelling of deflagrations still remains a challenge, since the flame brush has a very complex structure (sometimes presented as fractal in the literature), due to thermo-convective instabilities or turbulence Poinsot and Veynante 2005; Peters 2000. Whatever the modelling strategy, the problem thus needs a multiscale approach, since the local flame brush structure is out of reach of the computations aimed at simulating the flow dynamics at the observation scale, *i.e.* the whole reactive atmosphere scale. A possible way to completely circumvent this problem is to perform an explicit computation of the flame brush location, solving a transport-like equation for a characteristic function of the burnt zone; such an approach transfers the modelling difficulty to the evaluation of the flame brush velocity (or, more precisely speaking, to the relative velocity of the flame brush with respect to the fresh gases), by an adequate closure relation, and the resulting model is generally referred to as a Turbulent Flame velocity Closure (TFC) model Zimont 2000. The transport equation for the characteristic function of the burnt zone is called in this context the G-equation, its unknown being denoted by G Peters 2000. Such a modelling is implemented in the in-house software P^{2}REMICS (for Partially PREMIxed Combustion Solver) developed, on the basis of the software components library CALIF^{3}S (for Components Adaptative Library For Fluid Flow Simulations, see CALIF^{3}S n.d.) at the French Institut de Radioprotection et Sûreté Nucléaire (IRSN) for safety evaluation purposes; this is the context of the work presented in the present paper.

Usually, TFC models apply to perfectly premixed flows (*i.e.* flows with constant initial composition), and the chemical state of the flow is governed by the value of G only: $G \in [0,1]$, for $G \geq 0.5$, the mixture is supposed to be in its fresh (initial) state and $G < 0.5$ is supposed to correspond to the burnt state; in both cases, the composition of the gas is known (it is equal to the initial value in the fresh zones, and to the state resulting from a complete chemical reaction in the burnt zone).

However, for partially premixed turbulent flows (*i.e.* flows with non-constant initial composition), the situation is longer complex, since the composition of the mixture can no more be deduced from the value of G. An extension for this situation, in the inviscid case, is proposed in Beccantini and Studer 2010. The line followed to formulate this model is to write transport equations for the chemical species initially present in the flow, as if no chemical reaction occured, and then to compute the actual composition in the burnt zone (*i.e.* the part of the physical space where $G < 0.5$) as

the chemical equilibrium composition, thus supposing an infinitely fast reaction. This model is referred to in the following as the *"asymptotic model"*, and is recalled in the first part of Section 4.2.

We propose here an alternate extension, which consists in keeping the classical reactive formulation of the chemical species mass balance, but evaluating the reaction term as a function of G: it is set to zero in the fresh zone ($G \geq 0.5$), and to a finite (but possibly large) value in the burnt zone ($G < 0.5$). This model is referred to as the *"relaxed model"*; it is in fact more general, as it may be readily extended to cope with diffusion terms, while the "asymptotic model" cannot (to this purpose, a balance for the actual mass fractions is necessary). We then build a numerical scheme, based on a staggered discretization of the unknowns, for the solution of the relaxed model; this algorithm is of fractional step type, and employs a pressure correction technique for hydrodynamics. The balance energy solved by the scheme is the so-called (non conservative) sensible enthalpy balance, with corrective terms in order to ensure the weak consistency (in the Lax-Wendroff senses) of the scheme. It enjoys the same stability properties as the continuous model: positivity of the density and, thanks to the choice of the enthalpy balance, the internal energy, conservation of the total energy, chemical species mass fractions lying in the interval $[0,1]$. In addition, it is shown to be in fact conservative: indeed, its solutions satisfy a discrete conservative total energy balance whose time and space discretization is non-standard, but weakly consistent with its continuous counterpart. This algorithm is an extension to the reactive case of the numerical scheme for compressible Navier-Stokes equations described and tested in Grapsas, Herbin, Kheriji, et al. 2016.

As the reaction term gets stiffer, the relaxed model should boil down to the asymptotic one, for which a closed form of the solution of Riemann problems is available. Numerical tests are performed which show that indeed this is the case. In addition, we observe that the accuracy of the scheme (for this kind of application) is highly dependent on the numerical diffusion introduced by the scheme in the mass balance equation for the chemical species, comparing the results for three approximations of the convection operator in these equations: the standard upwind scheme, a MUSCL-like scheme introduced in Piar, Babik, Herbin, et al. 2013 and a first order scheme designed to reduce diffusion proposed in Després and Lagoutière 2002.

The presentation is structured as follows. We first introduce the asymptotic and the relaxed models in Section 4.2. Then we give an overview of the content of this paper in Section 4.3, writing the scheme in the time semi-discrete setting and stating its stability and consistency property. The fully discrete setting is given in two steps, first describing the space discretization (Section 4.4) and then the scheme itself (Section 4.5). The conservativity of the scheme is shown in Section 4.6. Finally, numerical experiments are presented in Section 5.5.

4.2 The physical models

We begin with the description of the asymptotic model introduced in Beccantini and Studer 2010 and then turn to the relaxed model proposed in the present work.

The asymptotic model - For the sake of simplicity, only four chemical species are supposed to be present in the flow, namely the fuel (denoted by F), the oxydant (O), the product (P) of the reaction, and a neutral gas (N). A one-step irreversible total chemical reaction is considered, which is written:

$$\nu_F F + \nu_O O + N \rightarrow \nu_P P + N,$$

where ν_F, ν_O and ν_P are the molar stoichiometric coefficients of the reaction. We denote by $\mathscr{I}$ the set of the subscripts used to refer to the chemical species in the flow, so $\mathscr{I} = \{F, O, N, P\}$ and the set of mass fractions of the chemical species in the flow reads $\{y_i,\ i \in \mathscr{I}\}$ (*i.e.* $\{y_F,\ y_O,\ y_N,\ y_P\}$). We now define the auxiliary unknowns $\{\tilde{y}_i,\ i \in \mathscr{I}\}$ as the result of the (inert) transport by the flow of the initial state, which means that the $\{\tilde{y}_i,\ i \in \mathscr{I}\}$ are the solutions to the following system of equation:

$$\partial_t(\rho\tilde{y}_i) + \mathrm{div}(\rho\tilde{y}_i\boldsymbol{u}) = 0, \quad \tilde{y}_i(\boldsymbol{x},0) = y_{i,0}(\boldsymbol{x}) \qquad \text{for } i \in \mathscr{I}, \tag{4.1}$$

where ρ stands for the fluid density, $\boldsymbol{u}$ for the velocity, and $y_{i,0}(\boldsymbol{x})$ is the initial mass fraction of the chemical species i in the flow. These equations are supposed to be posed over a bounded domain Ω of $\mathbb{R}^d$, $d \in \{1,2,3\}$ and a finite time interval $(0,T)$. The initial conditions are supposed to verify $\sum_{i\in\mathscr{I}} y_{i,0} = 1$ everywhere in Ω, and this property is assumed to be valid for any $t \in (0,T)$, which is equivalent with the mixture mass balance, given below. The characteristic function G is supposed to obey the following equation:

$$\partial_t(\rho G) + \mathrm{div}(\rho G\boldsymbol{u}) + \rho_u u_f |\nabla G| = 0, \tag{4.2}$$

associated to the initial conditions $G = 0$ at the location where the flame starts and $G = 1$ elsewhere. The quantity ρ_u is a constant density, which, from a physical point of view, stands for a characteristic value for the unburnt gases density. The chemical mass fractions are now computed as:

$$\left|\begin{array}{l} \text{if } G > 0.5, \quad y_i = \tilde{y}_i \quad \text{for } i \in \mathscr{I}, \\ \text{if } G \le 0.5, \quad y_F = \nu_F W_F \tilde{z}^+,\ y_O = \nu_O W_O \tilde{z}^-,\ y_N = \tilde{y}_N, \\ \qquad\qquad \text{with } \tilde{z} = \dfrac{1}{\nu_F W_F}\tilde{y}_F - \dfrac{1}{\nu_O W_O}\tilde{y}_O. \end{array}\right. \tag{4.3}$$

In these relation, $\tilde{z}^+$ and $\tilde{z}^-$ stand for the positive and negative part of $\tilde{z}$, respectively, *i.e.* $\tilde{z}^+ = \max(\tilde{z},0)$ and $\tilde{z}^- = -\min(\tilde{z},0)$, and, for $i \in \mathscr{I}$, W_i is the molar mass of the chemical species i. The physical meaning of Relation (4.3) is that the chemical reaction is supposed to be infinitely fast, and thus that the flow composition is stuck to the chemical equilibrium composition in the so-called burnt zone, which explains

why the model is qualified as “asymptotic". The product mass fraction is given by $y_P = 1 - (y_F + y_O + y_N)$. The flow is governed by the Euler equations:

$$\partial_t \rho + \mathrm{div}(\rho \boldsymbol{u}) = 0, \tag{4.4a}$$

$$\partial_t (\rho u_i) + \mathrm{div}(\rho u_i \boldsymbol{u}) + \partial_i p = 0, \quad i = 1, d, \tag{4.4b}$$

$$\partial_t (\rho E) + \mathrm{div}(\rho E \boldsymbol{u}) + \mathrm{div}(p \boldsymbol{u}) = 0, \tag{4.4c}$$

$$p = (\gamma - 1)\, \rho e_s, \qquad E = \frac{1}{2}|\boldsymbol{u}|^2 + e, \quad e = e_s + \sum_{i \in \mathscr{I}} y_i \Delta h^0_{f,i} \tag{4.4d}$$

where p stands for the pressure, E for the total energy, e for the internal energy, e_s for the so-called sensible internal energy and, for $i \in \mathscr{I}$, $\Delta h^0_{f,i}$ is the formation enthalpy of the chemical species i. The equation of state (4.4d) supposes that the fluid is a perfect mixture of ideal gases, with the same iso-pressure to iso-volume specific heat ratio $\gamma > 1$. This set of equations is complemented by homogeneous Neumann boundary conditions for the velocity:

$$\boldsymbol{u} \cdot \boldsymbol{n} = 0 \quad \text{a.e. on } \partial\Omega, \tag{4.5}$$

where $\partial\Omega$ stands for the boundary of Ω and $\boldsymbol{n}$ its outward normal vector.

The “relaxed" model – This model retains the original form governing equations for reactive flows: a a transport/reaction equation is written for each of the chemical species mass fractions; the value of G controls the reaction rate $\dot{\omega}$, which is set to zero when $G \geq 0.5$, and takes non-zero (and possibly large) values otherwise. The unknowns $\{y_i,\ i \in \mathscr{I}\}$ are thus now solution to the following balance equations:

$$\partial_t (\rho y_i) + \mathrm{div}(\rho y_i \boldsymbol{u}) = \dot{\omega}_i, \quad \tilde{y}_i(\boldsymbol{x}, 0) = y_{i,0}(\boldsymbol{x}) \qquad \text{for } i \in \mathscr{I}, \tag{4.6}$$

where the reactive term $\dot{\omega}_i$ is given by:

$$\dot{\omega}_i = \frac{1}{\varepsilon}\, \zeta_i\, \nu_i W_i\, \dot{\omega}, \text{ with } \dot{\omega} = \eta(y_F, y_O)\ (G - 0.5)^-$$
$$\text{and } \eta(y_F, y_O) = \min\left(\frac{y_F}{\nu_F W_F}, \frac{y_O}{\nu_O W_O}\right), \tag{4.7}$$

with $\zeta_F = \zeta_O = -1$, $\zeta_P = 1$ and $\zeta_N = 0$. Note that, since $\nu_F W_F + \nu_O W_0 = \nu_P W_P$, we have $\sum_{i \in \mathscr{I}} \dot{\omega}_i = 0$, which, summing on $i \in \mathscr{I}$ the species mass balance, allows to recover the equivalence between the mass balance and the fact that $\sum_{i \in \mathscr{I}} y_i = 1$. The factor $\eta(y_F, y_O)$ is a cut-off function, which prevents the chemical species mass fractions from taking negative values (and, consequently, values greater than 1, since their sum is equal to 1).

The rest of the model is left unchanged.

4.3 General description of the scheme and main results

Time semi-discrete algorithm Instead of the total energy balance equation, the scheme solves a balance equation for the sensible enthalpy $h_s = e_s + p/\rho$, which is formally derived as follows. The first step is to establish the kinetic energy balance formally and subtract from (4.4c) to obtain a balance equation for the internal energy. Thanks to the mass balance equation, for any regular function ψ

$$\partial_t(\rho\psi) + \mathrm{div}(\rho\psi\boldsymbol{u}) = \rho\partial_t\psi + \rho\boldsymbol{u}\cdot\nabla\psi.$$

Using twice this identity and then the momentum balance equation, we have for $1 \le i \le d$:

$$\frac{1}{2}\partial_t(\rho u_i^2) + \frac{1}{2}\mathrm{div}(\rho u_i^2\,\boldsymbol{u}) = \rho\, u_i\partial_t u_i + \rho u_i\boldsymbol{u}\cdot\nabla u_i \;= u_i\big[\partial_t(\rho u_i) + \mathrm{div}(\rho u_i\boldsymbol{u})\big] = -u_i\partial_i p,$$

and, summing for $i = 1$ to d, we obtain the kinetic energy balance:

$$\frac{1}{2}\partial_t(\rho|\boldsymbol{u}|^2) + \frac{1}{2}\mathrm{div}(\rho|\boldsymbol{u}|^2\boldsymbol{u}) = \boldsymbol{u}\cdot\big[\partial_t(\rho\boldsymbol{u}) + \mathrm{div}(\rho\boldsymbol{u}\otimes\boldsymbol{u})\big] = -\boldsymbol{u}\cdot\nabla p.$$

Substituting the expression of the total energy in (4.4c), yields

$$\partial_t(\rho e) + \mathrm{div}(\rho e\boldsymbol{u}) + \frac{1}{2}\partial_t(\rho|\boldsymbol{u}|^2) + \frac{1}{2}\mathrm{div}(\rho|\boldsymbol{u}|^2) + \boldsymbol{u}\cdot\nabla p + p\mathrm{div}(\boldsymbol{u}) = 0,$$

which, using the kinetic energy balance, gives the total internal energy balance:

$$\partial_t(\rho e) + \mathrm{div}(\rho e\boldsymbol{u}) + p\mathrm{div}(\boldsymbol{u}) = 0. \tag{4.8}$$

Using the linearity of the mass balance of the chemical species i, for any $i \in \mathscr{I}$, we derive the reactive energy balance:

$$\partial_t\big[\rho\big(\sum_{i\in\mathscr{I}}\Delta h^0_{f,i}y_i\big)\big] + \mathrm{div}\big[\rho\big(\sum_{i\in\mathscr{I}}\Delta h^0_{f,i}y_i\big)\boldsymbol{u}\big] = \sum_{i\in\mathscr{I}}\Delta h^0_{f,i}\dot{\omega}_i = -\dot{\omega}_\theta. \tag{4.9}$$

Subtracting (4.9) from (4.8) yields the sensible internal energy balance:

$$\partial_t(\rho e_s) + \mathrm{div}(\rho e_s\boldsymbol{u}) + p\mathrm{div}(\boldsymbol{u}) = \dot{\omega}_\theta. \tag{4.10}$$

Finally, using the relation between the sensible energy and the sensible enthalpy, we obtain the sensible enthalpy balance:

$$\partial_t(\rho h_s) + \mathrm{div}(\rho h_s\boldsymbol{u}) - \partial_t p - \boldsymbol{u}\cdot\nabla p = \dot{\omega}_\theta. \tag{4.11}$$

The numerical resolution of the mathematical model is realized by a fractional step

algorithm, which implements a pressure correction technique for hydrodynamics in order to separate the resolution of the momentum balance from the other equations of the Euler system. Supposing that the time interval $(0,T)$ is split in N sub-intervals, of constant length $\delta t = T/N$, the semi-discrete algorithm is given by:

Reactive step:

$$G^{n+1}: \quad \frac{1}{\delta t}(\rho^n G^{n+1} - \rho^{n-1} G^n) + \operatorname{div}(\rho^n G^k \boldsymbol{u}^n) + \rho_u u_f |\nabla G^{n+1}| = 0, \tag{4.12a}$$

$$Y_N^{n+1}: \quad \frac{1}{\delta t}(\rho^n y_N^{n+1} - \rho^{n-1} y_N^n) + \operatorname{div}(\rho^n y_N^k \boldsymbol{u}^n) = 0. \tag{4.12b}$$

$$z^{n+1}: \quad \frac{1}{\delta t}(\rho^n z^{n+1} - \rho^{n-1} z^n) + \operatorname{div}(\rho^n z^k \boldsymbol{u}^n) = 0. \tag{4.12c}$$

$$Y_F^{n+1}: \quad \frac{1}{\delta t}(\rho^n y_F^{n+1} - \rho^{n-1} y_F^n) + \operatorname{div}(\rho^n y_F^k \boldsymbol{u}^n) = -\frac{1}{\varepsilon} \nu_F W_F \dot{\omega}(y_F^{n+1}, z^{n+1}), \tag{4.12d}$$

$$Y_P^{n+1}: \quad y_F^{n+1} + y_O^{n+1} + y_N^{n+1} + y_P^{n+1} = 1. \tag{4.12e}$$

(4.12f)

Euler step:

$$\tilde{\boldsymbol{u}}^{n+1}: \quad \frac{1}{\delta t}(\rho^n \tilde{u}_i^{n+1} - \rho^{n-1} u_i^n) + \operatorname{div}(\rho^n \tilde{u}_i^{n+1} \boldsymbol{u}^n) + \left(\frac{\rho^n}{\rho^{n-1}}\right)^{1/2} \partial_i p^n = 0, \quad i = 1, \dots, d, \tag{4.12g}$$

$$\begin{matrix} \boldsymbol{u}^{n+1} \\ \rho^{n+1} \\ h_s^{n+1} \\ p^{n+1} \end{matrix} : \quad \left| \begin{array}{l} \dfrac{1}{\delta t}\rho^n (u_i^{n+1} - \tilde{u}_i^{n+1}) + \partial_i p^{n+1} - \sqrt{\dfrac{\rho^n}{\rho^{n-1}}} \partial_i p^n = 0, i \in [\![1, d]\!], \\ \dfrac{1}{\delta t}(\rho^{n+1} - \rho^n) + \operatorname{div}(\rho^{n+1} \boldsymbol{u}^{n+1}) = 0, \\ \dfrac{1}{\delta t}(\rho^{n+1} h_s^{n+1} - \rho^n h_s^n) + \operatorname{div}(\rho^{n+1} h_s^{n+1} \boldsymbol{u}^{n+1}) \\ \qquad -\dfrac{1}{\delta t}(p^{n+1} - p^n) - u^{n+1} \cdot \nabla p^{n+1} = \dot{\omega}_\theta^{n+1} + S^{n+1}, \\ p^{n+1} = \dfrac{\gamma - 1}{\gamma} \rho^{n+1} h_s^{n+1}. \end{array} \right. \tag{4.12h}$$

Equations (4.12a)-(4.12h) are solved successively, and the unknown for each equation is specified before each equation. In the convection term of the equations of the reactive step, the index k may take the value n (so the scheme is explicit) or $n+1$ (so the scheme is implicit). The unknown z is an affine combination of y_F and y_O, defined

so that the reactive term cancels:

$$z = \frac{1}{\nu_F W_F} y_F - \frac{1}{\nu_O W_O} y_O. \tag{4.13}$$

Thus the value of y_O^{n+1} is deduced from y_F^{n+1} and z^{n+1}, which allows to express $\dot{\omega}$ in (4.12d) as a function of y_F^{n+1} and z^{n+1}, instead of y_F^{n+1} and y_O^{n+1} as suggested by Relation (4.7). In addition, we have:

$$\begin{aligned}\eta(y_F^{n+1}, y_O^{n+1}) &= \min\left(\frac{y_F^{n+1}}{\nu_F W_F}, \frac{y_O^{n+1}}{\nu_O W_O}\right) \\ &= \begin{cases} \dfrac{1}{\nu_F W_F} y_F^{n+1} & \text{if } z^{n+1} \le 0, \\ \dfrac{1}{\nu_O W_O} y_O^{n+1} = \dfrac{1}{\nu_F W_F} y_F^{n+1} - z^{n+1} & \text{otherwise.} \end{cases}\end{aligned}$$

Hence, because of the specific form of the function η, the right hand side of (4.12d) boils down to an affine term, even if η vanishes when y_F or y_O vanishes, and the scheme is fully implicit in time with respect to the reaction term. This is the motivation for the choice of the form of η. It is fundamental to remark that Equations (4.12b)-(4.12e) are equivalent to the following system:

$$\frac{1}{\delta t}(\rho^n y_i^{n+1} - \rho^{n-1} y_i^n) + \operatorname{div}(\rho^n y_i^k \boldsymbol{u}^n) = \frac{1}{\varepsilon} \zeta_i \nu_i W_i \dot{\omega}(y_F^{n+1}, y_O^{n+1}), i \in \mathscr{I}, \tag{4.14}$$

where we recall that $\zeta_F = \zeta_O = -1$, $\zeta_P = 1$ and $\zeta_N = 0$. Indeed, dividing the fuel mass balance equation (4.12d) by $\nu_F W_F$, substracting Equation (4.12c) and finally multiplying by $\nu_O W_O$ yields the desired mass balance equation for the oxydant chemical species. Finally, we suppose that the product mass balance holds:

$$\frac{1}{\delta t}(\rho^n y_P^{n+1} - \rho^{n-1} y_P^n) + \operatorname{div}(\rho^n y_P^k \boldsymbol{u}^n) = \frac{1}{\varepsilon} \nu_P W_P \dot{\omega}(y_F^{n+1}, y_O^{n+1}). \tag{4.15}$$

Since the sum of the chemical reaction terms vanishes, we have for $\Sigma = y_F + y_O + y_P + y_N$, summing all the chemical species mass balances,

$$\frac{1}{\delta t}(\rho^n \Sigma^{n+1} - \rho^{n-1} \Sigma^n) + \operatorname{div}(\rho^n \Sigma^k \boldsymbol{u}^n) = 0, \tag{4.16}$$

and this equation may equivalently replace the product mass balance equation (4.15). Thanks to the mixture balance, we see that, provided that Σ^n satisfies $\Sigma^n = 1$ everywhere in Ω, the solution to Equation (4.16) is $\Sigma^{n+1} = 1$ everywhere in Ω. Since the initialization yields $\Sigma^0 = 1$, this last equality is indeed true, and (4.15) is equivalent to (4.12e). Finally, note that, when the chemical step is performed, the mass balance at step $n+1$ is not yet solved; hence the (unusual) backward time shift for the densities and for the mass fluxes in the equations of this step.

Equations (4.12g)-(4.12h) implement a pressure correction technique, where the correction step couples the velocity correction equation, the mass balance and the sensible enthalpy balance. This coupling ensures that the pressure and velocity are kept constant through the contact discontinuity associated to compositional non-reactive Euler equations (precisely speaking, the usual contact discontinuity, already present in 1D equations, but not slip lines); for this property to hold, it is necessary that all chemical species share the same heat capacity ratio γ. The term S_K^{n+1} in the sensible enthalpy balance equation is a corrective term which is necessary for consistency; schematically speaking, it compensates the numerical dissipation which appears in a discrete kinetic energy balance that is obtained from the discrete momentum balance. Its expression is given in Section 4.5, and its derivation is explained in Section 4.6, where the conservativity of the scheme is discussed.

Space discretization The space dicretization is performed by a finite volume technique, using a staggered arrangement of the unknowns (the scalar variables are approximated at the cell centers and the velocity components at the face centers), using either a MAC scheme (for structured discretizations) or the degrees of freedom of low-order non-conforming finite elements: Crouzeix-Raviart Crouzeix and P. Raviart 1973 for simplicial cells and Rannacher-Turek Rannacher and Turek 1992 for quadrangles ($d = 2$) or hexahedra ($d = 3$). For the Euler equations (*i.e.* Steps (4.12g)- (4.12h)), upwinding is performed by building positivity-preserving convection operators, in the spirit of the so-called Flux-Splitting methods, and only first-order upwinding is implemented. The pressure gradient is built as the transpose (with respect to the L^2 inner product) of the natural velocity divergence operator. For the balance equations for the other scalar unknowns, the time discretization is implicit when first-order upwinding is used in the convection operator (in other words, $k = n+1$ in (4.12a)-(4.12d)) or explicit ($k = n$ in (4.12a)-(4.12d)) when a higher order (of MUSCL type, *cf.* Appendix 4.A) flux or an anti-diffusive flux (*cf.* Appendix 4.B) is used.

Properties of the scheme First, the positivity of the density is ensured by construction of the discrete mass balance equation, *i.e.* by the use of a first order upwind scheme. In addition, the physical bounds of the mass fractions are preserved thanks to the following (rather standard) arguments: first, building a discrete convection operator which vanishes when the convected unknown is constant thanks to the discrete mass balance equation ensures a positivity-preservation property Larrouturou 1991, under a CFL condition if an explicit time approximation is used; second, the discretization of the chemical reaction rate ensures either that it vanishes when the unknown of the equation vanishes (for y_F and y_O), or that it is non-negative (for y_P). Consequently, mass fractions are non-negative and, since their sum is equal to 1 (see above), they are also bounded by 1.

The positivity of the sensible energy stems from two essential arguments: first, a discrete analog of the internal energy equation (4.8) may be obtained from the discrete

sensible enthalpy balance, by mimicking the continuous computation; second, this discrete relation may be shown to have only positive solutions, once again thanks to the consistency of the discrete convection operator and the mass balance. This holds provided that the equation is exothermic ($\dot{\omega}_\theta \geq 0$) and thanks to the non-negativity of S^{n+1} (see below).

In order to calculate correct shocks, it is crucial for the scheme to be consistent with the following weak formulation of the problem:

$$
\begin{aligned}
&\forall \phi \in C_c^\infty(\Omega \times [0,T)\big), \\
&\int_0^T \int_\Omega \big[\rho \partial_t \phi + \rho \boldsymbol{u} \cdot \nabla \phi\big] \mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega \rho_0(\boldsymbol{x})\phi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} = 0, \\
&\int_0^T \int_\Omega \big[\rho u_i \partial_t \phi + (\rho \boldsymbol{u} u_i) \cdot \nabla \phi + p \partial_i \phi\big] \mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\
&\qquad\qquad + \int_\Omega \rho_0(\boldsymbol{x})(u_i)_0(\boldsymbol{x})\phi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} = 0,\ 1 \leq i \leq d, \\
&\int_0^T \int_\Omega \big[\rho E \partial_t \phi + (\rho E + p)\boldsymbol{u} \cdot \nabla \phi\big] \mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega \rho_0(\boldsymbol{x}) E_0(\boldsymbol{x})\phi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} = 0, \\
&\int_0^T \int_\Omega \big[\rho y_i \partial_t \phi + \rho y_i \boldsymbol{u} \cdot \nabla \phi\big] \mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_0^T \int_\Omega \rho_0(\boldsymbol{x}) y_{i,0}(\boldsymbol{x})\phi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} = \\
&\qquad\qquad - \int_0^T \int_\Omega \dot{\omega}_i \phi \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t,\ 1 \leq i \leq d, \\
&p = (\gamma - 1)\rho e_s.
\end{aligned}
\tag{4.17}
$$

Remark that this system features the total energy balance equation and not the sensible enthalpy balance equation, which is actually solved here. However, we show in Section 4.6 that the solutions of the scheme satisfy a discrete total energy balance, with a time and space dicretization which is unusual but allows however to prove the consistency in the Lax-Wendroff sense. Finally, the integral of the total energy over the domain is conserved, which yields a stability result for the scheme (irrespectively of the time and space step, for this relation; recall however that the overall stability of the scheme needs a CFL condition if an explicit version of the convection operator for chemical species is used).

4.4 Meshes and unknowns

Let the computational domain Ω be an open polygonal subset of $\mathbb{R}^d$, $1 \leq d \leq 3$, with boundary $\partial\Omega$ and let $\mathscr{M}$ be a decomposition of Ω, supposed to be regular in the usual sense of the finite element literature (*e.g.* Ciarlet 1991). The cells may be:

- for a general domain Ω, either convex quadrilaterals ($d = 2$) or hexahedra ($d = 3$) or simplices, both type of cells being possibly combined in a same mesh,
- for a domain the boundaries of which are hyperplanes normal to a coordinate axis, rectangles ($d = 2$) or rectangular parallelepipeds ($d = 3$) (the faces of which, of course, are then also necessarily normal to a coordinate axis).

By $\mathscr{E}$ and $\mathscr{E}(K)$ we denote the set of all $(d-1)$-faces σ of the mesh and of the element $K \in \mathscr{M}$ respectively. The set of faces included in the boundary of Ω is denoted by $\mathscr{E}_{\text{ext}}$ and the set of internal edges (*i.e.* $\mathscr{E} \setminus \mathscr{E}_{\text{ext}}$) is denoted by $\mathscr{E}_{\text{int}}$; a face $\sigma \in \mathscr{E}_{\text{int}}$ separating the cells K and L is denoted by $\sigma = K|L$. The outward normal vector to a face σ of K is denoted by $\boldsymbol{n}_{K,\sigma}$. For $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}$, we denote by $|K|$ the measure of K and by $|\sigma|$ the $(d-1)$-measure of the face σ. For any $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}(K)$, we denote by $d_{K,\sigma}$ the Euclidean distance between the center x_K of the mesh and the edge σ. For any $\sigma \in \mathscr{E}$, we define $d_\sigma = d_{K,\sigma} + d_{L,\sigma}$, if $\sigma \in \mathscr{E}_{\text{int}}$ and $d_\sigma = d_{K,\sigma}$ if $\sigma \in \mathscr{E}_{\text{ext}}$. The size of the mesh is denoted by h. For $1 \le i \le d$, we denote by $\mathscr{E}^{(i)} \subset \mathscr{E}$ and $\mathscr{E}^{(i)}_{\text{ext}} \subset \mathscr{E}_{\text{ext}}$ the subset of the faces of $\mathscr{E}$ and $\mathscr{E}_{\text{ext}}$ respectively which are perpendicular to the i^{th} unit vector of the canonical basis of $\mathbb{R}^d$.

The space discretization is staggered, using either the Marker-And Cell (MAC) scheme Harlow and Welsh 1965; Harlow and Amsden 1971, or nonconforming low-order finite element approximations, namely the Rannacher and Turek (RT) element Rannacher and Turek 1992 for quadrilateral or hexahedric meshes, or the lowest degree Crouzeix-Raviart (CR) element Crouzeix and P. Raviart 1973 for simplicial meshes.

For all these space discretizations, the degrees of freedom for the pressure, the density, the enthalpy, the mixture, fuel and neutral gas mass fractions and the flame indicator are associated to the cells of the mesh $\mathscr{M}$ and are denoted by:

$$\{p_K,\ \rho_K,\ h_K,\ y_{F,K},\ y_{N,K},\ z_K,\ G_K,\ K \in \mathscr{M}\}.$$

Let us then turn to the degrees of freedom for the velocity (*i.e.* the discrete velocity unknowns).

- **Rannacher-Turek** or **Crouzeix-Raviart** discretizations – The degrees of freedom for the velocity components are located at the center of the faces of the mesh, and we choose the version of the element where they represent the average of the velocity through a face. The set of degrees of freedom reads:

$$\{\boldsymbol{u}_\sigma,\ \sigma \in \mathscr{E}\},\ \text{of components } \{u_{i,\sigma},\ \sigma \in \mathscr{E},\ 1 \le i \le d\}.$$

- **MAC** discretization – The degrees of freedom for the i^{th} component of the velocity are defined at the centre of the faces of $\mathscr{E}^{(i)}$, so the whole set of discrete velocity unknowns reads:

$$\left\{u_{i,\sigma},\ \sigma \in \mathscr{E}^{(i)},\ 1 \le i \le d\right\}.$$

For the definition of the schemes, we need a dual mesh which is defined as follows.

- **Rannacher-Turek** or **Crouzeix-Raviart** discretizations – For the RT or CR discretizations, the dual mesh is the same for all the velocity components. When $K \in \mathscr{M}$ is a simplex, a rectangle or a rectangular cuboid, for $\sigma \in \mathscr{E}(K)$, we define

$D_{K,\sigma}$ as the cone with basis σ and with vertex the mass center of K (see Figure 4.1). We thus obtain a partition of K in m sub-volumes, where m is the number of faces of the mesh, each sub-volume having the same measure $|D_{K,\sigma}| = |K|/m$. We extend this definition to general quadrangles and hexahedra, by supposing that we have built a partition still of equal-volume sub-cells, and with the same connectivities; note that this is of course always possible, but that such a volume $D_{K,\sigma}$ may be no longer a cone; indeed, if K is far from a parallelogram, it may not be possible to build a cone having σ as basis, the opposite vertex lying in K and a volume equal to $|K|/m$ (note that these dual cells do not need to be constructed in the implementation of the scheme, only their volume is needed). The volume $D_{K,\sigma}$ is referred to as the half-diamond cell associated to K and σ. For $\sigma \in \mathscr{E}_{\rm int}$, $\sigma = K|L$, we now define the diamond cell D_σ associated to σ by $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$; for an external face $\sigma \in \mathscr{E}_{\rm ext} \cap \mathscr{E}(K)$, D_σ is just the same volume as $D_{K,\sigma}$.

- **MAC** discretization – For the MAC scheme, the dual mesh depends on the component of the velocity. For each component, the MAC dual mesh only differs from the RT or CR dual mesh by the choice of the half-diamond cell, which, for $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}(K)$, is now the rectangle or rectangular parallelepiped of basis σ and of measure $|D_{K,\sigma}| = |K|/2$.

We denote by $|D_\sigma|$ the measure of the dual cell D_σ, and by $\epsilon = D_\sigma|D_{\sigma'}$ the dual face separating two diamond cells D_σ and $D_{\sigma'}$.

In order to be able to write a unique expression of the discrete equations for both MAC and CR/RT schemes, we introduce the set of faces $\mathscr{E}^{(i)}_{\mathscr{S}}$ associated with the degrees of freedom of each component of the velocity ($\mathscr{S}$ stands for "scheme"):

$$\mathscr{E}^{(i)}_{\mathscr{S}} = \left| \begin{array}{l} \mathscr{E}^{(i)} \setminus \mathscr{E}^{(i)}_{\rm ext} \text{ for the MAC scheme,} \\ \mathscr{E} \setminus \mathscr{E}^{(i)}_{\rm ext} \text{ for the CR or RT schemes.} \end{array} \right.$$

Similarly, we unify the notation for the set of dual faces for both schemes by defining:

$$\widetilde{\mathscr{E}}^{(i)}_{\mathscr{S}} = \left| \begin{array}{l} \widetilde{\mathscr{E}}^{(i)} \setminus \widetilde{\mathscr{E}}^{(i)}_{\rm ext} \text{ for the MAC scheme,} \\ \widetilde{\mathscr{E}} \setminus \widetilde{\mathscr{E}}^{(i)}_{\rm ext} \text{ for the CR or RT schemes,} \end{array} \right.$$

where the symbol $\tilde{}$ refers to the dual mesh; for instance, $\widetilde{\mathscr{E}}^{(i)}$ is thus the set of faces of the dual mesh associated with the i^{th} component of the velocity, and $\widetilde{\mathscr{E}}^{(i)}_{\rm ext}$ stands for the subset of these dual faces included in the boundary. Note that, for the MAC scheme, the faces of $\widetilde{\mathscr{E}}^{(i)}$ are perpendicular to a unit vector of the canonical basis of $\mathbb{R}^d$, but not necessarily to the i^{th} one.

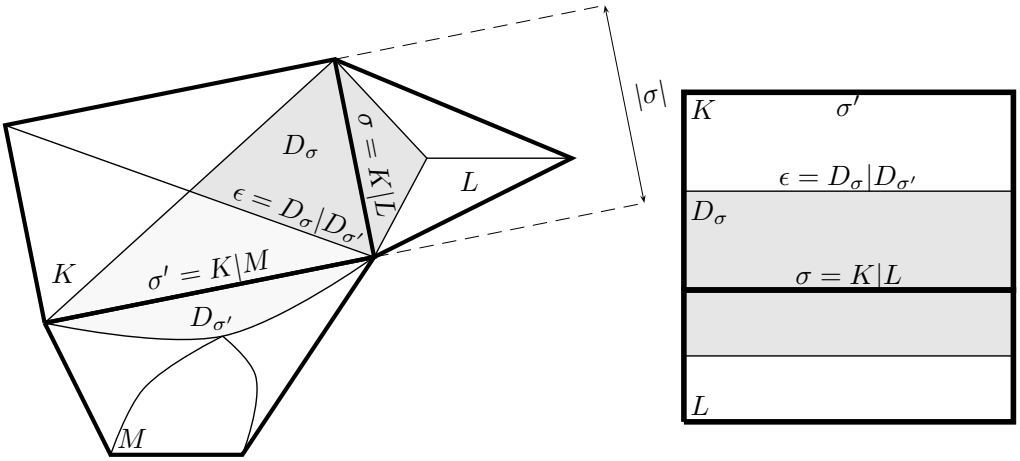


Figure 4.1 – Primal and dual meshes for the Rannacher-Turek and Crouzeix-Raviart elements.

4.5 The scheme

In this section, we give the fully discrete form of the scheme. Even if it corresponds to the reverse order with respect to the semi-discrete scheme given in (4.12), we begin with the hydrodynamics (Section 4.5.1) and then turn to the mass balance step for chemical species and the transport of the characteristic function for the burnt zone (Section 4.5.2). This choice is due to the fact that the definition of the convection operators for scalar variables necessitates to introduce the discretization of the mixture mass balance equation first.

4.5.1 Euler step

For $0 \leq n < N$, the step $n+1$ of the algorithm for the resolution of the Euler equations reads:

Pressure gradient scaling step – Solve for $(\widetilde{\nabla p})^{n+1}$:

$$\forall \sigma \in \mathscr{E}, \qquad (\widetilde{\nabla p}^{n+1})_\sigma = \left(\frac{\rho^n_{D_\sigma}}{\rho^{n-1}_{D_\sigma}}\right)^{1/2} (\nabla p^n)_\sigma. \tag{4.18a}$$

Prediction step – Solve for $\tilde{\boldsymbol{u}}^{n+1}$:

For $1 \leq i \leq d$, $\forall \sigma \in \mathscr{E}^{(i)}_{\mathscr{D}}$,

$$\frac{1}{\delta t}(\rho^n_{D_\sigma} \tilde{u}^{n+1}_{i,\sigma} - \rho^{n-1}_{D_\sigma} u^n_{i,\sigma}) + \mathrm{div}_{D_\sigma}(\rho^n \tilde{u}^{n+1}_i \boldsymbol{u}^n) + (\widetilde{\nabla p})^{n+1}_{i,\sigma} = 0. \tag{4.18b}$$

Correction step – Solve for ρ^{n+1}, p^{n+1} and $\boldsymbol{u}^{n+1}$:

For $1 \leq i \leq d$, $\forall \sigma \in \mathscr{E}^{(i)}_{\mathscr{D}}$,

$$\frac{1}{\delta t}\rho^n_{D_\sigma}(u^{n+1}_{i,\sigma} - \tilde{u}^{n+1}_{i,\sigma}) + (\nabla p)^{n+1}_{i,\sigma} - (\widetilde{\nabla p})^{n+1}_{i,\sigma} = 0, \tag{4.18c}$$

$$\forall K \in \mathscr{M}, \ \frac{1}{\delta t}(\rho^{n+1}_K - \rho^n_K) + \mathrm{div}_K(\rho \boldsymbol{u})^{n+1} = 0, \tag{4.18d}$$

$$\begin{aligned} \forall K \in \mathscr{M}, \ &\frac{1}{\delta t}\left[\rho^{n+1}_K (h_s)^{n+1}_K - \rho^n_K (h_s)^n_K\right] + \mathrm{div}_K(\rho h_s \boldsymbol{u})^{n+1} \\ &- \frac{1}{\delta t}(p^{n+1}_K - p^n_K) - \left(u \cdot \nabla p\right)^{n+1}_K = (\dot{\omega}_\theta)^{n+1}_K + S^{n+1}_K, \end{aligned} \tag{4.18e}$$

$$\forall K \in \mathscr{M}, \qquad p^{n+1}_K = \frac{\gamma - 1}{\gamma}(h_s)^{n+1}_K \rho^{n+1}_K. \tag{4.18f}$$

The initial approximations for ρ^{-1}, h^0_s and $\boldsymbol{u}^0$ are given by the mean values of the initial conditions over the primal and dual cells:

$$\forall K \in \mathscr{M}, \quad \rho^{-1}_K = \frac{1}{|K|}\int_K \rho_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad \text{and} \quad (h_s)^0_K = \frac{1}{|K|}\int_K (h_s)_0(\boldsymbol{x}),$$

$$\forall \sigma \in \mathscr{E}^{(i)}_{\mathscr{D}}, \ 1 \leq i \leq d, \quad u^0_{i,\sigma} = \frac{1}{|D_\sigma|}\int_{D\sigma} (\boldsymbol{u}_0(\boldsymbol{x}))_i \mathrm{d}\boldsymbol{x}.$$

Then, ρ^0 is computed by the mass balance equation (4.18d) and p^0 is computed by the equation of state (4.18f).

We now define each of the discrete operators featured in System (4.18).

Mass balance equation Equation (4.18d) is a finite volume discretisation of the mass balance (4.4a) over the primal mesh. For a discrete density field ρ and a discrete

velocity field $\boldsymbol{u}$, the discrete divergence is defined by:

$$\mathrm{div}_K(\rho \boldsymbol{u}) = \frac{1}{|K|} \sum_{\sigma \in \mathscr{E}(K)} F_{K,\sigma}, \quad F_{K,\sigma} = |\sigma|\, \rho_\sigma \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma},$$

where $\boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma}$ is an approximation of the normal velocity to the face σ outward K. The definition of this latter quantity depends on the discretization: in the MAC case, $\boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} = u_{i,\sigma}\, \boldsymbol{e}^{(i)} \cdot \boldsymbol{n}_{K,\sigma}$ for a face σ of K perpendicular to $\boldsymbol{e}^{(i)}$, with $\boldsymbol{e}^{(i)}$ the i-th vector of the orthonormal basis of $\mathbb{R}^d$, and, in the CR and RT cases, $\boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} = \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma}$ for any face σ of K. The density at the face $\sigma = K|L$ is approximated by the upwind technique, so $\rho_\sigma = \rho_K$ if $\boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \geq 0$ and $\rho_\sigma = \rho_L$ otherwise. Since we assume that the normal velocity vanishes on the boundary faces, the definition is complete.

Convection operators associated to the primal mesh We may now define the discrete convection operator of any discrete field z defined on the primal cell by

$$\mathrm{div}_K(\rho z \boldsymbol{u}) = \frac{1}{|K|} \sum_{\sigma \in \mathscr{E}(K)} F_{K,\sigma}\, z_\sigma,$$

where z_σ is the upwind approximation with respect to the mass flux $F_{K,\sigma}$ at the face σ.

Momentum balance equation and pressure gradient scaling We now turn to the discrete momentum balance (4.18b). For the MAC discretization, but also for the RT and CR discretizations, the time derivative and convection terms are approximated in (4.18b) by a finite volume technique over the dual cells, so the convection term reads:

$$\mathrm{div}_{D_\sigma}(\rho \tilde{u}_i \boldsymbol{u}) = \mathrm{div}_{D_\sigma}\big(\tilde{u}_i(\rho \boldsymbol{u})\big) = \frac{1}{|D_\sigma|} \sum_{\epsilon \in \widetilde{\mathscr{E}}(D_\sigma)} F_{\sigma,\epsilon}\, \tilde{u}_{i,\epsilon},$$

where $F_{\sigma,\epsilon}$ stands for a mass flux through the dual face ϵ, and $\tilde{u}_{i,\epsilon}$ is a centered approximation of the i^{th} component of the velocity $\tilde{\boldsymbol{u}}$ on ϵ. The density at the dual cell ρ_{D_σ} is obtained by a weighted average of the density in the neighbour cells: $|D_\sigma|\, \rho_{D_\sigma} = |D_{K,\sigma}|\, \rho_K + |D_{L,\sigma}|\, \rho_L$ for $\sigma = K|L \in \mathscr{E}_{\mathrm{int}}$, and $\rho_{D_\sigma} = \rho_K$ for an external face of a cell K. The mass fluxes $(F_{\sigma,\epsilon})_{\epsilon \in \widetilde{\mathscr{E}}(D_\sigma)}$ are evaluated as linear combinations, with constant coefficients, of the primal mass fluxes at the neighbouring faces, in such a way that the following discrete mass balance over the dual cells is implied by the discrete mass balance (4.18d):

$$\forall \sigma \in \mathscr{E} \text{ and } n \in \mathbb{N}, \qquad \frac{|D_\sigma|}{\delta t} (\rho_{D_\sigma}^{n+1} - \rho_{D_\sigma}^{n}) + \sum_{\epsilon \in \widetilde{\mathscr{E}}(D_\sigma)} F_{\sigma,\epsilon}^{n+1} = 0. \tag{4.19}$$

This relation is critical to derive a discrete kinetic energy balance (see Section 4.6 below). The computation of the dual mass fluxes is such that the flux through a dual face lying on the boundary, which is then also a primal face, is the same as the primal

flux, that is zero. For the expression of these densities and fluxes, we refer to Gastaldo, Herbin, Kheriji, et al. 2011; Herbin, Kheriji, and Latché 2014; Herbin and Latché 2010. Since the mass balance is not yet solved at the velocity prediction stage, they have to be built from the mass balance at the previous time step: hence the backward time shift for the densities in the time-derivative term.

The term $(\nabla p)_{i,\sigma}$ stands for the i-th component of the discrete pressure gradient at the face σ. This gradient operator is built as the transpose of the discrete operator for the divergence of the velocity, *i.e.* in such a way that the following duality relation with respect to the L^2 inner product holds:

$$\sum_{K\in\mathcal{M}} |K| p_K \mathrm{div}_K(\boldsymbol{u}) + \sum_{i=1}^{d} \sum_{\sigma\in\mathcal{E}_{\mathscr{D}}^{(i)}} |D_\sigma| u_{i,\sigma} (\nabla p)_{i,\sigma} = 0.$$

This leads to the following expression:

$$\forall \sigma = K|L \in \mathcal{E}_{\mathrm{int}}, \qquad (\nabla p)_{i,\sigma} = \frac{|\sigma|}{|D_\sigma|}(p_L - p_K)\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{e}^{(i)}.$$

The scaling of the pressure gradient (4.18a) is necessary for the solution to the scheme to satisfy a local discrete (finite volume) kinetic energy balance Grapsas, Herbin, Kheriji, et al. 2016, Lemma 4.1.

Sensible enthalpy equation The equation is discretised in such a way that the present enthalpy formulation is strictly equivalent to the internal energy formulation of the energy balance equation used in Grapsas, Herbin, Kheriji, et al. 2016. Consequently, the term $-\big(u\cdot\nabla p\big)_K$ reads:

$$-\big(u\cdot\nabla p\big)_K = \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}\,(p_K - p_\sigma),$$

where p_σ is the upwind approximation of p at the face σ with respect to $\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}$. The reaction heat, $(\dot\omega_\theta)_K$, is written in the following way:

$$(\dot\omega_\theta)_K = -\sum_{i=1}^{N_s} \Delta h^0_{f,i}(\dot\omega_i)_K = \Big(\nu_F W_F \Delta h^0_{f,F} + \nu_O W_O \Delta h^0_{f,O} - \nu_P W_P \Delta h^0_{f,P}\Big)\dot\omega_K.$$

The definition of $\dot\omega_K$ is given in Section 4.5.2, and the definition of the corrective term S_K^{n+1} is given in Section 4.6 (see Equation (4.30) and Remark 4.3 below).

4.5.2 Chemistry step

For $0 \leq n < N$, the step $n+1$ for the solution of the transport of the characteristic function of the burnt zone and the chemical species mass balance equations reads:

Computation of the burnt zone characteristic function – Solve for G^{n+1}:

$$\forall K \in \mathscr{M},\ \frac{1}{\delta t}(\rho_K^n G_K^{n+1} - \rho_K^{n-1} G_K^n) + \mathrm{div}_K(\rho^n G^{n+1} \boldsymbol{u}^n) + (\rho_u^n u_f^n |\nabla G|)_K = 0. \tag{4.20a}$$

Computation of the variable z – Solve for z^{n+1}:

$$\forall K \in \mathscr{M},\ \frac{1}{\delta t}(\rho_K^n z_K^{n+1} - \rho_K^{n-1} z_K^n) + \mathrm{div}_K(\rho^n z^{n+1} \boldsymbol{u}^n) = 0. \tag{4.20b}$$

Neutral gas mass fraction computation – Solve for y_N^{n+1}:

$$\forall K \in \mathscr{M},\ \frac{1}{\delta t}\left[\rho_K^n (y_N)_K^{n+1} - \rho_K^{n-1} (y_N)_K^n\right] + \mathrm{div}_K(\rho^n y_N^{n+1} \boldsymbol{u}^n) = 0. \tag{4.20c}$$

Fuel mass fraction computation – Solve for y_F^{n+1}:

$$\forall K \in \mathscr{M},\ \frac{1}{\delta t}\left[\rho_K^n (y_F)_K^{n+1} - \rho_K^{n-1} (y_F)_K^n\right] + \mathrm{div}_K(\rho^n y_F^{n+1} \boldsymbol{u}^n) = -\frac{1}{\varepsilon} \nu_F W_F \dot{\omega}_K^{n+1}. \tag{4.20d}$$

Product mass fraction computation – Compute y_P^{n+1} given by:

$$\forall K \in \mathscr{M},\ (y_P)_K^{n+1} = 1 - (y_F)_K^{n+1} - (y_O)_K^{n+1} - (y_N)_K^{n+1}. \tag{4.20e}$$

The initial value of the chemical variables is the mean value of the initial condition over the primal cells and $\forall K \in \mathscr{M}$ we define:

$$G_K^0 = \frac{1}{|K|}\int_K G_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad z_K^0 = \frac{1}{|K|}\int_K z_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad (y_i)_K^0 = \frac{1}{|K|}\int_K (y_i)_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\ i = N, F,$$

where the reduced variable z is the linear combination of y_F and y_O given by Equation (4.13). According to the developments of Section 4.3, the chemical reaction term reads $\dot{\omega}_K^{n+1} = \eta((y_F)_K^{n+1}, z_K^{n+1})\ (G_K^{n+1} - 0.5)^-$ with

$$\eta((y_F)_K^{n+1}, z_K^{n+1}) = \begin{cases} \dfrac{1}{\nu_F W_F}(y_F)_K^{n+1} & \text{if } z^{n+1} \leq 0, \\ \dfrac{1}{\nu_F W_F}(y_F)_K^{n+1} - z_K^{n+1} & \text{otherwise,} \end{cases}$$

and the chemical species mass fractions satisfy the following system, which is equivalent to (4.20b)-(4.20e):

$$\frac{1}{\delta t}(\rho_K^n (y_i)_K^{n+1} - \rho_K^{n-1} (y_i)_K^n) + \mathrm{div}_K(\rho^n y_i^k \boldsymbol{u}^n) = \frac{1}{\varepsilon}\zeta_i \nu_i W_i \dot{\omega}_K^{n+1}, \text{ for } i \in \mathscr{I} \text{ and } K \in \mathscr{M}. \tag{4.21}$$

At the continuous level, the last term of equation (4.20a) may be written:

$$\rho_u\, u_f\, |\nabla G| = \boldsymbol{a}\cdot\nabla G = \mathrm{div}(G\,\boldsymbol{a}) - G\,\mathrm{div}(\boldsymbol{a}), \quad \text{with } \boldsymbol{a} = \rho_u\, u_f\, \frac{\nabla G}{|\nabla G|}.$$

Using an upwind finite volume discretization of both divergence terms in this relation, we get:

$$|K|\ (\rho_u^n\, u_f^n\, |\nabla G|)_K = \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\ (G_\sigma^{n+1} - G_K^{n+1})\ \boldsymbol{a}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma},$$

where G_σ^{n+1} stands for the upwind approximation of G^{n+1} on σ with respect to $\boldsymbol{a}^n\cdot\boldsymbol{n}_{K,\sigma}$. The flame velocity on σ, $\boldsymbol{a}_\sigma^n$, is evaluated as

$$\boldsymbol{a}_\sigma^n = (\rho_u\, u_f)_\sigma^n\, \frac{(\nabla G)_\sigma^n}{|(\nabla G)_\sigma^n|},$$

where $(\rho_u\, u_f)_\sigma^n$ stands for an approximation of the product $\rho_u\, u_f$ on the face σ at t^n (this product is often constant in applications), and the gradient of G on $\sigma = K|L$ is computed as:

$$(\nabla G)_\sigma = \frac{1}{|K\cup L|}\Big[\sum_{\tau\in\mathscr{E}(K)} |\tau|\ \hat{G}_\tau\ \boldsymbol{n}_{K,\tau} + \sum_{\tau\in\mathscr{E}(L)} |\tau|\ \hat{G}_\tau\ \boldsymbol{n}_{L,\tau}\Big],$$

where $\hat{G}_\tau$ is a second order approximation of G at the center of the face τ.

4.6 Scheme conservativity

Let the discrete sensible internal energy be defined by $p_K^n = (\gamma-1)\,\rho_K^n (e_s)_K^n$ for $K\in\mathscr{M}$ and $0\le n\le N$. In view of the equation of state (4.18f), this definition implies $\rho_K^n (h_s)_K^n = \rho_K^n (e_s)_K^n + p_K^n$, for $K\in\mathscr{M}$ and $0\le n\le N$. The following lemma states that the discrete solutions satisfy a local internal energy balance.

Lemma 4.1 (Discrete internal energy balance)**.**
A solution to (4.18)-(4.20) *satisfies the following equality, for any* $K\in\mathscr{M}$ *and* $0\le n<N$*:*

$$\frac{1}{\delta t}\big[(\rho e)_K^{n+1} - (\rho e)_K^n\big] + \widetilde{\mathrm{div}}_K(\rho e\boldsymbol{u})^{n+1} + p_K^{n+1}\mathrm{div}_K(\boldsymbol{u})^{n+1} = S_K^{n+1}, \tag{4.22}$$

where

$$(\rho e)_K^{n+1} = \rho_K^{n+1}(e_s)_K^{n+1} + \rho_K^n \sum_{i\in\mathscr{I}} \Delta h_{f,i}^0 (y_i)_K^{n+1},$$

$$\widetilde{\mathrm{div}}_K(\rho e\boldsymbol{u})^{n+1} = \mathrm{div}_K\Big[(\rho e_s)^{n+1}\boldsymbol{u}^{n+1} + \rho^n\big[\sum_{i\in\mathscr{I}} \Delta h_{f,i}^0 y_i^{n+1}\big]\boldsymbol{u}^n\Big].$$

Proof. We begin by deriving a local sensible internal energy balance, starting from the sensible enthalpy balance (4.18e) and mimicking the previously given formal passage

between these two equations at the continuous level (*i.e.* the passage from Equation (4.11) to Equation (4.10)). To this purpose, let us write (4.18e) as $T_1 + T_2 = T_3$ with

$$\begin{aligned} T_1 &= \frac{1}{\delta t}\left[\rho_K^{n+1}(h_s)_K^{n+1} - \rho_K^n (h_s)_K^n\right] - \frac{1}{\delta t}(p_K^{n+1} - p_K^n), \\ T_2 &= \mathrm{div}_K(\rho h_s \boldsymbol{u})^{n+1} - \left(\boldsymbol{u}\cdot\nabla p\right)_K^{n+1}, \\ T_3 &= (\dot{\omega}_\theta)_K^{n+1} + S_K^{n+1}. \end{aligned}$$

Using $\rho_K^\ell (h_s)_K^\ell = \rho_K^\ell (e_s)_K^\ell + p_K^\ell$ for $\ell = n$ and $\ell = n+1$, we easily get

$$T_1 = \frac{1}{\delta t}\left[\rho_K^{n+1}(e_s)_K^{n+1} - \rho_K^n (e_s)_K^n\right].$$

The term T_2 reads:

$$|K|\, T_2 = \sum_{\sigma\in\mathscr{E}(K)} |\sigma| \left[\rho_\sigma^{n+1}(h_s)_\sigma^{n+1} - p_\sigma^{n+1} + p_K^{n+1}\right] \boldsymbol{u}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma}.$$

If $\boldsymbol{u}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma} > 0$, by definition, $\rho_\sigma^{n+1}(h_s)_\sigma^{n+1} = \rho_K^{n+1}(h_s)_K^{n+1}$ and $p_\sigma^{n+1} = p_K^{n+1}$; otherwise, thanks to the assumptions on the boundary conditions, σ is an internal face and, denoting by L the adjacent cell to K such that $\sigma = K|L$, $\rho_\sigma^{n+1}(h_s)_\sigma^{n+1} = \rho_L^{n+1}(h_s)_L^{n+1}$ and $p_\sigma^{n+1} = p_L^{n+1}$. In both cases, denoting by $(e_s)_\sigma^{n+1}$ the upwind choice for $(e_s)^{n+1}$ at the face σ, we get

$$\rho_\sigma^{n+1}(h_s)_\sigma^{n+1} - p_\sigma^{n+1} = \rho_\sigma^{n+1}(e_s)_\sigma^{n+1},$$

so, finally

$$|K|\, T_2 = \sum_{\sigma\in\mathscr{E}(K)} F_{K,\sigma}^{n+1}(e_s)_\sigma^{n+1} + p_K^{n+1} \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, \boldsymbol{u}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma}.$$

We thus get the following sensible internal energy balance:

$$\begin{aligned} \frac{|K|}{\delta t}\left[\rho_K^{n+1}(e_s)_K^{n+1} - \rho_K^n (e_s)_K^n\right] + \sum_{\sigma\in\mathscr{E}(K)} F_{K,\sigma}^{n+1}(e_s)_\sigma^{n+1} \\ + p_K^{n+1} \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, \boldsymbol{u}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma} = |K|\left[(\dot{\omega}_\theta)_K^{n+1} + S_K^{n+1}\right], \end{aligned} \tag{4.23}$$

or, using the discrete differential operator formalism,

$$\frac{1}{\delta t}\left[\rho_K^{n+1}(e_s)_K^{n+1} - \rho_K^n (e_s)_K^n\right] + \mathrm{div}_K(\rho e_s \boldsymbol{u})^{n+1} + p_K^{n+1}\mathrm{div}_K \boldsymbol{u}^{n+1} = (\dot{\omega}_\theta)_K^{n+1} + S_K^{n+1}. \tag{4.24}$$

We now derive from this relation a discrete (sensible and chemical) internal energy balance. Multiplying the mass fraction balance equations by the corresponding for-

mation enthalpy $(\Delta h^0_{f,i})_{i\in\mathscr{I}}$ and summing over $i\in\mathscr{I}$ yields:

$$\frac{1}{\delta t}\sum_{i\in\mathscr{I}}\Delta h^0_{f,i}\big[\rho^n_K(y_i)^{n+1}_K-\rho^{n+1}_K(y_i)^n_K\big]+\sum_{\sigma\in\mathscr{E}(K)}F^n_{K,\sigma}\sum_{i\in\mathscr{I}}\Delta h^0_{f,i}\,(y_i)^{n+1}_\sigma$$
$$=\sum_{i\in\mathscr{I}}\Delta h^0_{f,i}\,(\dot\omega_i)^{n+1}_K=-(\dot\omega_\theta)^{n+1}_K.$$

Adding this relation to (4.23) yields the balance equation which we are looking for. □

Remark 4.1 (Positivity of the sensible internal energy)**.** *Equation* (4.24) *implies that the sensible internal energy remains positive, provided that the right-hand side is non-negative, which is true if* $\dot\omega_\theta\geq 0$*, i.e. if the chemical reaction is exothermic. The proof of this property may be found in Grapsas, Herbin, Kheriji, et al. 2016, Lemma 4.3, and relies on two arguments: first, the convection operator may be recast as a discrete positivity-preserving transport operator thanks to the mass balance, and, second, the pressure* p^{n+1}_K *vanishes when* e^{n+1}_K*, by the equation of state.*

The following local discrete kinetic energy balance holds on the dual mesh (see Grapsas, Herbin, Kheriji, et al. 2016, Lemma 4.1 for a proof).

Lemma 4.2 (Discrete kinetic energy balance on the dual mesh)**.**
A solution to (4.18)-(4.20) *satisfies the following equality, for* $1\leq i\leq d$, $\sigma\in\mathscr{E}^{(i)}_{\mathscr{D}}$ *and* $0\leq n<N$*:*

$$\frac{|D_\sigma|}{\delta t}\big[(e_k)^{n+1}_{i,\sigma}-(e_k)^n_{i,\sigma}\big]+\sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)}F^n_{\sigma,\epsilon}(e_k)^{n+1}_{\epsilon,i}+|D_\sigma|(\nabla p)^{n+1}_{i,\sigma}u^{n+1}_{i,\sigma}=-R^{n+1}_{i,\sigma},\tag{4.25}$$

where

$$(e_k)^{n+1}_{i,\sigma}=\frac{1}{2}\rho^n_{D_\sigma}(u^{n+1}_{i,\sigma})^2+\delta t^2\frac{|D_\sigma|}{2\rho^n_{D_\sigma}}\big((\nabla p)^{n+1}_{i,\sigma}\big)^2,$$
$$(e_k)^{n+1}_{\epsilon,i}=\frac{1}{2}\tilde u^{n+1}_{i,\sigma}\tilde u^{n+1}_{i,\sigma'},$$
$$R^{n+1}_{i,\sigma}=\frac{|D_\sigma|\,\rho^{n-1}_{D_\sigma}}{2\delta t}(\tilde u^{n+1}_{i,\sigma}-u^n_{i,\sigma})^2.$$

We now derive a kinetic energy balance equation on the primal cells from Relation (4.25). For the sake of clarity, we make a separate exposition for the Rannacher-Turek case and the MAC case. The case of simplicial discretizations, with the degrees of freedom of the Crouzeix-Raviart element, is an easy extension of the Rannacher-Turek case.

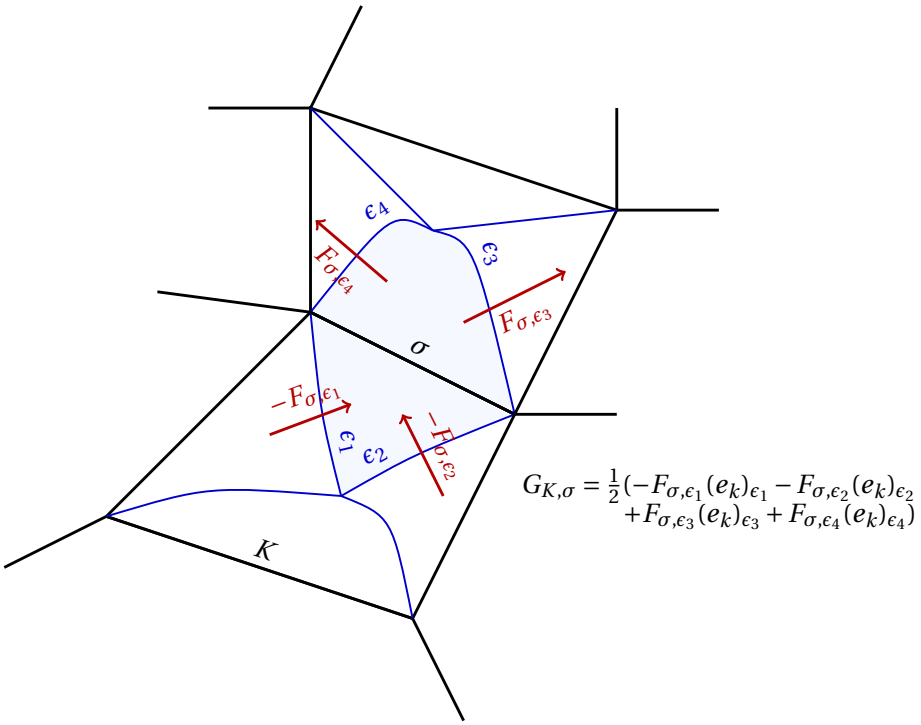


Figure 4.2 – From fluxes at dual faces to fluxes at primal faces, for the Rannacher-Turek discretization.

The Rannacher-Turek case Since the dual meshes are the same for all the velocity components in this case, we may sum up Equation (4.25) over $i = 1,\dots d$ to obtain, for $\sigma \in \mathscr{E}$ and $0 \le n < N$:

$$\frac{|D_\sigma|}{\delta t}\left[(e_k)_\sigma^{n+1} - (e_k)_\sigma^n\right] + \sum_{\epsilon\in\widetilde{\mathscr{E}}(D_\sigma)} F_{\sigma,\epsilon}^n (e_k)_\epsilon^{n+1} + |D_\sigma|(\nabla p)_\sigma^{n+1}\cdot \boldsymbol{u}_\sigma^{n+1} = -R_\sigma^{n+1}, \tag{4.26}$$

with

$$(e_k)_\sigma^\ell = \sum_{i=1}^d (e_k)_{i,\sigma}^\ell, \text{ for } \ell = n \text{ or } \ell = n+1, \quad (e_k)_\epsilon^{n+1} = \sum_{i=1}^d (e_k)_{i,\epsilon}^{n+1}, \text{ and } R_\sigma^{n+1} = \sum_{i=1}^d R_{i,\sigma}^{n+1}.$$

For $K \in \mathscr{M}$, let us define a kinetic energy associated to K and the flux $G_{K,\sigma}^{n+1}$ as follows (see Figure 4.2):

$$(e_k)_K^\ell = \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}(K)} |D_\sigma|\,(e_k)_\sigma^\ell,\ \ell = n \text{ or } \ell = n+1,$$
$$G_{K,\sigma}^{n+1} = -\frac{1}{2}\sum_{\epsilon\in\mathscr{E}(D_\sigma),\epsilon\subset K} F_{\sigma,\epsilon}^n\,(e_k)_\epsilon^{n+1} + \frac{1}{2}\sum_{\epsilon\in\mathscr{E}(D_\sigma),\epsilon\not\subset K} F_{\sigma,\epsilon}^n\,(e_k)_\epsilon^{n+1}.$$

We easily check that the fluxes $G_{K,\sigma}^{n+1}$ are conservative, in the sense that, for $\sigma = K|L$, $G_{K,\sigma}^{n+1} = -G_{L,\sigma}^{n+1}$. Let us now divide Equation (4.26) by 2 and sum over the faces of K. A reordering of the summations, using the conservativity of the mass fluxes through the dual edges and the expression of the discrete pressure gradient, yields:

$$\frac{|K|}{\delta t}\left[(e_k)_K^{n+1} - (e_k)_K^n\right] + \sum_{\sigma\in\mathscr{E}(K)} G_{K,\sigma}^{n+1} + \sum_{\sigma=K|L} |\sigma|\,(p_L^{n+1} - p_K^{n+1})\,\boldsymbol{u}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma} = -R_K^{n+1},$$
$$\text{with } R_K^{n+1} = \frac{1}{2}\sum_{\sigma\in\mathscr{E}(K)} R_\sigma^{n+1}. \tag{4.27}$$

The MAC case Let $1 \le i \le d$, let $K \in \mathscr{M}$, let us denote by σ and σ' the two faces of $\mathscr{E}^{(i)}(K)$, and let us define:

$$(e_k)_{i,K}^{\ell} = \frac{1}{|K|}\left[|D_\sigma|\,(e_k)_{i,\sigma}^{\ell} + |D_\sigma|\,(e_k)_{i,\sigma}^{\ell}\right], \text{ for } \ell = n \text{ or } \ell = n+1.$$

Case of primal faces parallel to the dual faces. Let $\tau = \sigma$ or $\tau = \sigma'$, let ϵ_1 and ϵ_2 be the two faces of D_τ perpendicular to $\boldsymbol{e}^{(i)}$, and let ϵ' be included in K (see Figure 4.3). Then we define

$$G_{i,K,\tau}^{n+1} = \frac{1}{2}\left[F_{\tau,\epsilon_1}(e_k)_{i,\epsilon_1}^{n+1} - F_{\tau,\epsilon_2}(e_k)_{i,\epsilon_2}^{n+1}\right].$$

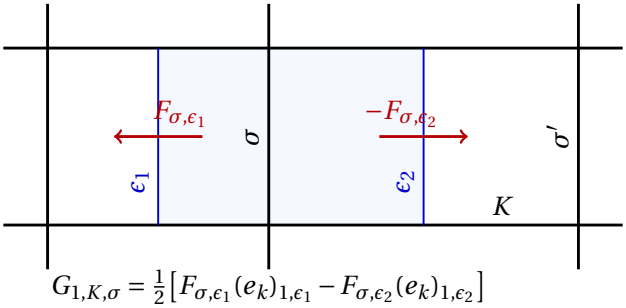


Figure 4.3 – From fluxes at dual faces to fluxes at primal faces, for the MAC discretization, primal faces parallel to the dual edges, first component of the velocity.

Case of primal faces orthogonal to the dual faces. For $\tau \in \mathscr{E}(K)\setminus\{\sigma,\sigma'\}$, let ϵ and ϵ' be such that $\tau \subset (\bar{\epsilon}\cup\bar{\epsilon}')$ with ϵ a face of D_σ and ϵ' a face of $D_{\sigma'}$ (see Figure 4.4).

Then we define

$$G_{i,K,\tau}^{n+1} = F_{\sigma,\epsilon}(e_k)_{i,\epsilon}^{n+1} - F_{\sigma',\epsilon'}(e_k)_{i,\epsilon'}^{n+1}.$$

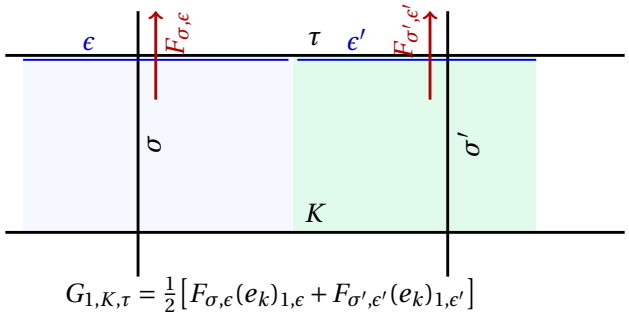


Figure 4.4 – From fluxes at dual faces to fluxes at primal faces, for the MAC discretization, primal faces orthogonal to the dual edges, first component of the velocity.

Summing Equation (4.25) written for σ and for σ' and dividing the result by 2 yields:

$$\frac{|K|}{\delta t}\left[(e_k)_{i,K}^{n+1} - (e_k)_{i,K}^{n}\right] + \sum_{\sigma\in\mathscr{E}(K)} G_{i,K,\sigma}^{n+1} + \sum_{\substack{\sigma\in\mathscr{E}^{(i)}(K)\\ \sigma=K|L}} |\sigma|\,(p_L^{n+1} - p_K^{n+1})\boldsymbol{u}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma} = -\frac{1}{2}\left(R_{i,\sigma}^{n+1} + R_{i,\sigma'}^{n+1}\right). \quad (4.28)$$

Now let $(e_k)_K^\ell = \sum_{i=1}^{d}(e_k)_{i,K}^\ell$, for $\ell = n$ or $\ell = n+1$, and $G_{K,\sigma}^{n+1} = \sum_{i=1}^{d} G_{i,K,\sigma}^{n+1}$, for $\sigma\in\mathscr{E}(K)$. Since only one equation is written for a given face σ of the mesh (for the velocity component i with i such that the normal vector to σ is parallel to $\boldsymbol{e}^{(i)}$), we may define in the MAC case $R_\sigma^{n+1} = R_{i,\sigma}^{n+1}$. Summing Equation 4.28 over the space dimension, we finally get

$$\frac{|K|}{\delta t}\left[(e_k)_K^{n+1} - (e_k)_K^{n}\right] + \sum_{\sigma\in\mathscr{E}(K)} G_{K,\sigma}^{n+1} + \sum_{\sigma=K|L} |\sigma|\,(p_L^{n+1} - p_K^{n+1})\,\boldsymbol{u}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma} = -R_K^{n+1},$$
$$\text{with } R_K^{n+1} = \frac{1}{2}\sum_{\sigma\in\mathscr{E}(K)} R_\sigma^{n+1}, \quad (4.29)$$

which is formally the same equation as Relation (4.27) (although with a different definition of all the terms in the equation except the pressure gradient).

Remark 4.2 (On the definition of the cell kinetic energy)**.** *Note that, both in the Rannacher-Turek and the MAC case, the cell kinetic energy is not a convex combination of the face kinetic energies, since, on a non-uniform mesh, the equalities* $|K| = \frac{1}{2}\sum_{\sigma\in\mathscr{E}(K)}|D_\sigma|$ *(Rannacher Turek case) and* $|K| = \frac{1}{2}\sum_{\sigma\in\mathscr{E}^{(i)}(K)}|D_\sigma|$ *(MAC case) do not hold in general. Consequently, the cell kinetic energy may for instance oscillate from cell*

to cell while the face kinetic energy does not. Nevertheless, the discrete time derivative of the cell kinetic energy is consistent in the Lax-Wendroff sense.

Equations (4.27) and (4.29) suggest a choice for the term S_K^{n+1}, the purpose of which is to compensate the numerical dissipation terms appearing in the kinetic energy balance:

$$S_K^{n+1} = R_K^{n+1}, \text{ for } K \in \mathcal{M} \text{ and } 0 \leq n < N. \tag{4.30}$$

This expression yields a conservative scheme, in the sense that the discrete solutions satisfy a discrete total energy balance without any remainder term (see Equation (4.4c) below); as a consequence, the scheme can be proven to be consistent in the Lax-Wendroff sense. However, different definitions are possible (and this latitude may be useful in explicit variants of the scheme, to ensure the positivity of S_K^{n+1}, see Remark 4.3 below.

We are now in position to state a total energy balance for the scheme.

Theorem 4.1 (Discrete total energy and stability of the scheme)**.**
A solution to (4.18)-(4.20) *satisfies the following equality, for any* $K \in \mathcal{M}$ *and* $0 \leq n < N$*:*

$$\frac{1}{\delta t}\left[(\rho E)_K^{n+1} - (\rho E)_K^n\right] + \widetilde{\mathrm{div}}_K((\rho E + p)\boldsymbol{u})^{n+1} = 0, \tag{4.31}$$

where

$$(\rho E)_K^\ell = (e_k)_K^\ell + \rho_K^\ell (e_s)_K^\ell + \rho_K^{\ell-1} \sum_{i \in \mathscr{I}} \Delta h_{f,i}^0 (y_i)_K^\ell, \text{ for } \ell = n \text{ and } \ell = n+1,$$

$$\widetilde{\mathrm{div}}_K((\rho E + p)\,\boldsymbol{u})^{n+1} = \mathrm{div}_K\Big((\rho e_s)^{n+1}\boldsymbol{u}^{n+1} + \rho^n\big[\sum_{i \in \mathscr{I}} \Delta h_{f,i}^0 y_i^{n+1}\big]\boldsymbol{u}^n\Big) + \frac{1}{|K|}\sum_{\sigma = K|L} |\sigma|\,(p_K^{n+1} + p_L^{n+1})\,\boldsymbol{u}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma} + \frac{1}{|K|}\sum_{\sigma \in \mathscr{E}(K)} G_{K,\sigma}^{n+1}.$$

Let us suppose that e_s^0, ρ^0 *and* ρ^{-1} *are positive. Then, a solution to* (4.18)-(4.20) *satisfies* $\rho^{n+1} > 0$, $e^{n+1} > 0$ *and the following stability result:*

$$E^n = E^0,$$

where, for $0 \leq n \leq N$,

$$E^n = \sum_{K \in \mathcal{M}} |K|(\rho e)_K^n + \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma \in \mathscr{E}_{\mathscr{I}}^{(i)}} |D_\sigma|(u_{i,\sigma}^n)^2 + \delta t^2 \sum_{\sigma \in \mathscr{E}_{\mathrm{int}}} \frac{|D_\sigma|}{\rho_{D_\sigma}^{n-1}} |(\nabla p)_\sigma^n|^2.$$

Proof. The discrete total energy balance equation (4.31) is obtained by summing the internal energy balance (4.22) and the kinetic energy balance, *i.e.* Equation (4.27) in the Rannacher-Turek case and Equation (4.29) for the MAC scheme, and remarking that the numerical dissipation terms in the kinetic energy balance R_K^{n+1}

exactly compensate with the corrective terms S_K^{n+1} in the internal energy balance. Then the stability result is obtained by summation over the time steps. □

Remark 4.3 (Consistency of the scheme)**.** *The consistency in the Lax-Wendroff sense follows from the conservativity of the scheme (for all balance equations) so, in particular, from the fact that the discrete solutions satisfy the discrete total energy balance* (4.31)*, thanks to standard (but technical) arguments.*
Note however that the consistency of the scheme does not require a strict conservativity, and in particular, variants for the choice (4.30) *of the compensation term in the sensible enthalpy balance are possible; indeed, what is really needed is only that the difference between the dissipation in the kinetic energy balance and its compensation tend to zero in a distributional sense. In practice, this allows a different redistribution of the face residuals to the neighbour primal cells, and this can help to preserve the non-negativity of the compensation term for explicit versions of the scheme.*

4.7 Numerical tests

At the continuous level, the boundedness of the chemical mass fractions formally implies that, when $\varepsilon \to 0$, the relaxed model converges to the asymptotic one. Indeed, integrating any of the reactive species mass balance equations with respect to time and space, we observe that $||\dot{\omega}||_{L^1(\Omega\times(0,T))}$ tends to zero as ε, and thus two separate zones appear: a zone characterized by $G < 0.5$ where the reaction is complete, and a zone corresponding to $G \geq 0.5$, where no reaction has occured.

A closed form of the solution of the Riemann problem for the asymptotic model is available Beccantini and Studer 2010. In order to perform numerical tests, a Riemann problem with initial conditions such that the analytic solution has the profile presented in Figure 4.5 is chosen.

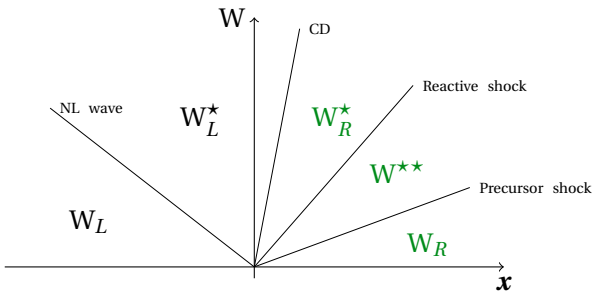


Figure 4.5 – The analytic solution of the numerical test configuration.

Moreover, the selected configuration imposes zero amplitude for the contact discontinuity and the left non linear wave, thus the solution consists of three different constant states: W_R^*, W^{**} and W_R. The right state corresponds to a stoichiometric mixture of

hydrogen and air (so the molar fractions of Hydrogen, Oxygen and Nitrogen are 2/7, 1/7 and 4/7 respectively) at rest, at the pressure $p = 9.9\,10^4$ Pa and the temperature $T = 283°$ K. The velocity is supposed to be zero in the left state, which is sufficient to determine the solution. Physically, speaking, supposing that the initial discontinuity lies at $x = 0$, this situation corresponds to the left part of a (symmetrical) constant velocity plane deflagration starting at $x = 0$.. The flame velocity is $u_f = 63$ m/s and the formation enthalpies are zero except for the product (*i.e.* steam), with $\Delta h^0_{f,O} = -13.255\,10^6$ J (Kg K)$^{-1}$. The quantity ρ_u is the analytical density in the intermediate state (so the total velocity of the flame brush is equal to the sum of u_f and the material velocity on the right side of the reactive shock, see Beccantini and Studer 2010). The computation is initialized by the analytical solution at $t = 0.002$ and the final time is $t = 0.005$. The computational domain is the interval $(0, 4.5)$.

The numerical tests performed aim at checking the convergence of the scheme to such a solution, which in fact may result from two different properties: the convergence of the relaxed model to the asymptotic model when ε tends to zero, and the convergence of the scheme towards a numerical solution when the time and space steps tend to zero. To this purpose, we choose ε proportional to the space step and make it tend to zero, with a constant CFL number. We test the scheme behaviour with three different discretizations of the convection operator in the chemical mass species balances: the standard upwind scheme, a MUSCL-like discretization which is an extension to variable density flows of the scheme proposed in Piar, Babik, Herbin, et al. 2013 and is described in Appendix 4.A, and a first-order anti-diffusive scheme which is an adaptation to our setting of the scheme proposed in Després and Lagoutière 2002; we detail it in Appendix 4.B for the sake of completeness. Results obtained at $t = 0.005$ with the upwind scheme, the MUSCL-like scheme and the anti-diffusive scheme, for increasingly refined meshes, are shown on Figure 4.6, Figure 4.7 and Figure 4.8 respectively, together with the analytical solution. The expected convergence is indeed observed but, with the upwind discretization, the rate of convergence is poor. This seems to be due to the interaction between the numerical diffusion of the upwind scheme, which artificially introduces unburnt reactive masses to the burnt zone, and the stiffness of the reaction term. As expected in such a case, the results are significantly improved by the use of a less diffusive scheme for the chemical species balance equations. Indeed, passing from the upwind to the MUSCL-like and to the anti-diffusive discretization improves the accuracy of the scheme, as may be observed in Figure 4.9, where the results obtained by the three discretizations for a regular mesh composed of 500 cells are plotted together with the continuous solution. This observation is comforted by the measure, in L^1-norm, of the difference between the discrete and continuous solutions, see Table 4.1. For every mesh and variable, the anti-diffusive scheme is the most accurate and the upwind one the least. The calculated order of convergence is close to 0.5 for the upwind scheme, and to 1 for the MUSCL-like and anti-diffusive schemes.

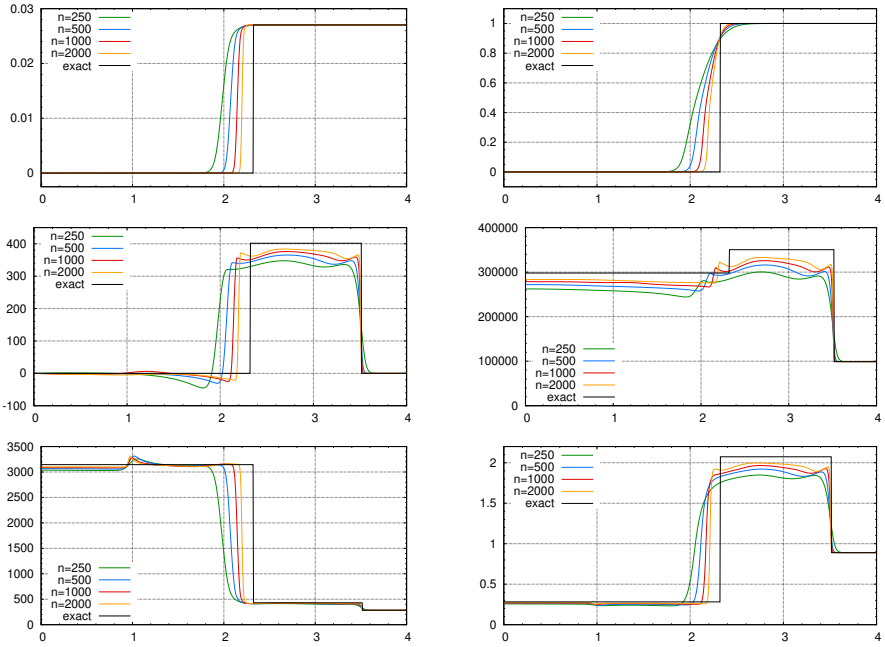


Figure 4.6 – Upwind scheme – From top left to bottom right, fuel mass fraction, G, velocity, pressure, temperature and density at $t = 0.005$, as a function of the space variable.

h	$\|\|p - p_{ex}\|\|_{L^1} \times 10^{-4}$			$\|\|\boldsymbol{u} - \boldsymbol{u}_{ex}\|\|_{L^1} \times 10^{-2}$			$\|\|\rho - \rho_{ex}\|\|_{L^1} \times 10$		
h_0	16.5	7.26	4.59	2.17	1.56	1.07	7.69	3.71	2.74
$\frac{h_0}{2}$	12.5	3.88	2.43	1.64	0.787	0.579	6.16	2.23	1.65
$\frac{h_0}{4}$	9.66	2.05	1.38	1.23	0.471	0.371	4.73	1.26	0.913
$\frac{h_0}{8}$	7.58	1.17	0.708	0.958	0.263	0.175	3.63	0.691	0.476
$\frac{h_0}{20}$	5.78	0.673	0.375	0.728	0.160	0.103	2.77	0.382	0.267
$\frac{h_0}{40}$	4.31	0.414	0.194	0.543	0.0786	0.0458	2.03	0.201	0.134

Table 4.1 – L^1 norm of the error between the discrete and continuous solutions for the various schemes - Black : upwind scheme, blue: MUSCL scheme, orange: anti-diffusive scheme; $h_0 = 4.5/250$ is the size of the least refined mesh.

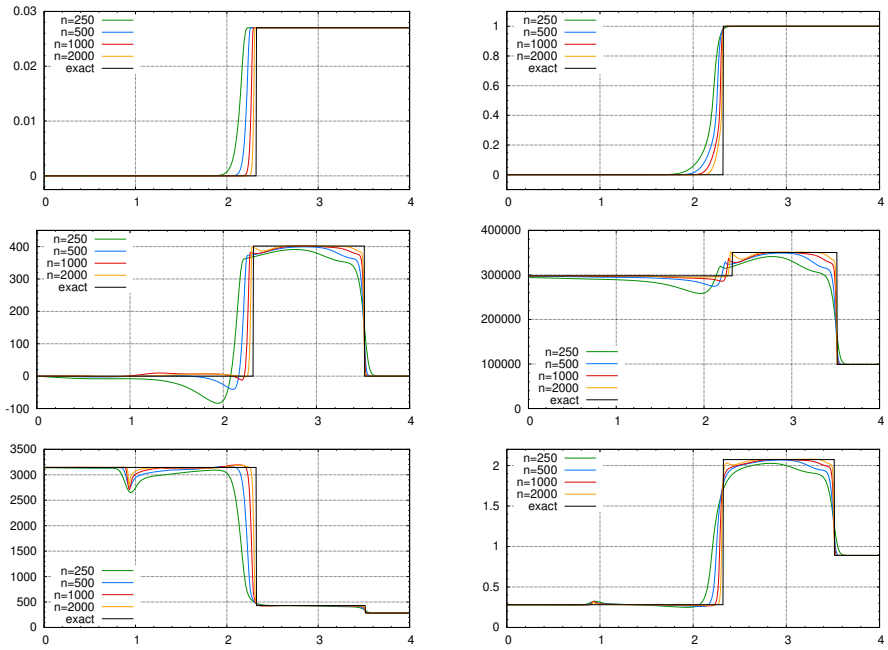


Figure 4.7 – MUSCL scheme – From top left to bottom right, fuel mass fraction, G, velocity, pressure, temperature and density at $t = 0.005$, as a function of the space variable.

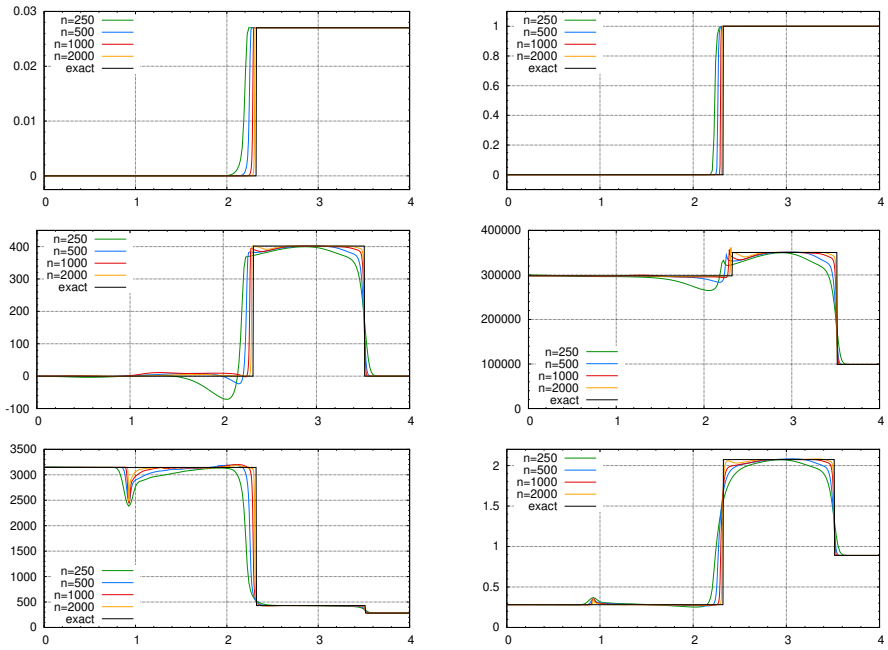


Figure 4.8 – Anti-diffusive scheme – From top left to bottom right, fuel mass fraction, G, velocity, pressure, temperature and density at $t = 0.005$, as a function of the space variable.

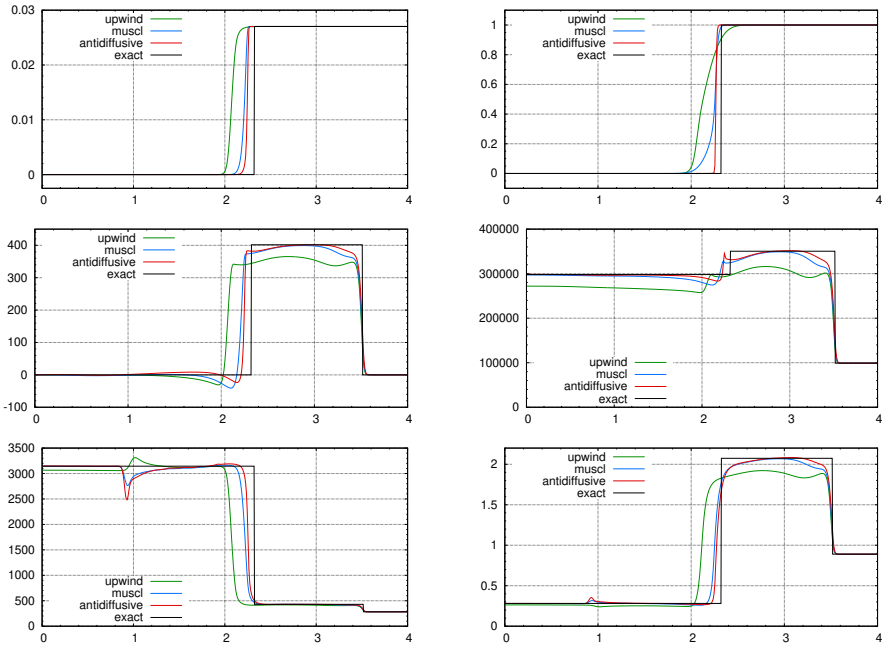


Figure 4.9 – Comparison of the solutions obtained with the upwind, MUSCL and anti-diffusive scheme – From top to bottom, fuel mass fraction, G, velocity, pressure, temperature and density at $t = 0.005$, as a function of the space variable. Results obtained with a regular mesh composed of $n = 500$ cells.

Appendix

4.A The MUSCL interpolation scheme

The MUSCL discretization of the convection operators of the chemical species balance and G-equation closely follows the technique proposed in Piar, Babik, Herbin, et al. 2013. To present this discretization, we consider the following system of equations:

$$\partial_t \rho + \operatorname{div}(\rho \boldsymbol{u}) = 0,$$
$$\partial_t (\rho y) + \operatorname{div}(\rho \boldsymbol{u} y) = 0.$$

We suppose for short that this system is complemented by impermeability boundary conditions, *i.e.* that the normal velocity, both at the continuous and the discrete level, vanishes on the boundary of the computational domain.

The discretization of the above system reads:

$$\forall K \in \mathcal{M}, \quad \frac{\rho_K^{n+1} - \rho_K^n}{\delta t} + \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1} = 0,$$
$$\frac{\rho_K^{n+1} y_K^{n+1} - \rho_K^n y_K^n}{\delta t} + \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1} y_\sigma^n = 0.$$

For any $\sigma \in \mathcal{E}$, the procedure consists in three steps:

- calculate a tentative value for y_σ as a linear interpolate of nearby values,
- calculate an interval for y_σ which guarantees the stability of the scheme,
- project the tentative value y_σ on this stability interval.

For the tentative value of y_σ, let us choose some real coefficients $(\alpha_K^\sigma)_{K \in \mathcal{M}}$ such that

$$\boldsymbol{x}_\sigma = \sum_{K \in \mathcal{M}} \alpha_K^\sigma \boldsymbol{x}_K, \qquad \sum_{K \in \mathcal{M}} \alpha_K^\sigma = 1.$$

The coefficients used in this interpolation are chosen in such a way that as few as possible cells, to be picked up in the closest cells to σ, take part. For example, for $\sigma = K|L$ and if $\boldsymbol{x}_K$, $\boldsymbol{x}_\sigma$, $\boldsymbol{x}_L$ are aligned, only two non-zero coefficients exist in the family $(\alpha_K^\sigma)_{K \in \mathcal{M}}$, namely α_K^σ and α_L^σ. Then, these coefficients are used to calculate the tentative value of y_σ by

$$y_\sigma = \sum_{K \in \mathcal{M}} \alpha_K^\sigma y_K.$$

The construction of the stability interval must be such that the following property holds:

$$\forall K \in \mathcal{M},\ \forall \sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\text{int}},\ \exists \beta_K^\sigma \in [0,1] \text{ and } M_K^\sigma \in \mathcal{M} \text{ such that}$$

$$y_\sigma^n - y_K^n = \begin{cases} \beta_K^\sigma (y_K^n - y_{M_K^\sigma}^n), \text{ if } F_{K,\sigma}^{n+1} \geq 0, \\ \beta_K^\sigma (y_{M_K^\sigma}^n - y_K^n), \text{ otherwise.} \end{cases} \tag{4.32}$$

Indeed, under this latter hypothesis and a CFL condition, the scheme preserves the initial bounds of y.

Remark 4.4. *Note that, in Assumption (4.32), only internal faces are considered, since the fluxes through external faces are supposed to vanish. However, the present discussion may easily be generalized to cope with convection fluxes entering the domain.*

Definition 4.1. *The so-called CFL number reads for any* $0 \leq n \leq N$*:*

$$\text{CFL}^{n+1} = \max_{K \in \mathcal{M}} \left\{ \frac{\delta t}{\rho_K^{n+1} |K|} \sum_{\sigma \in \mathcal{E}(K)} \left| F_{K,\sigma}^{n+1} \right| \right\}.$$

Lemma 4.3. *Let us suppose that* $\text{CFL}^{n+1} \leq 1$*. For* $K \in \mathcal{M}$*, let us note by* $\mathcal{V}(K)$ *the union of the set of cells* M_K^σ*,* $\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\text{int}}$ *such that (4.32) holds. Then* $\forall K \in \mathcal{M}$*, the value of* y_K^{n+1} *is a convex combination of* $\{y_K^n, (y_M^n)_{M \in \mathcal{V}(K)}\}$*.*

Proof. The discrete mass balance equation yields:

$$\rho_K^n = \rho_K^{n+1} + \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1}.$$

Replacing this expression of ρ_K^n in the discrete balance equation of y and using the relations provided by (4.32), we obtain:

$$\begin{aligned}
\rho_K^{n+1} y_K^{n+1} &= \rho_K^n y_K^n - \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1} y_\sigma^n \\
&= \rho_K^{n+1} y_K^n - \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1} (y_\sigma^n - y_K^n) \\
&= \rho_K^{n+1} y_K^n - \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} \left(F_{K,\sigma}^{n+1}\right)^+ (y_\sigma^n - y_K^n) + \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} \left(F_{K,\sigma}^{n+1}\right)^- (y_\sigma^n - y_K^n) \\
&= \rho_K^{n+1} y_K^n - \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} \left(F_{K,\sigma}^{n+1}\right)^+ \beta_K^\sigma (y_K^n - y_{M_K^\sigma}^n) \\
&\qquad + \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} \left(F_{K,\sigma}^{n+1}\right)^- \beta_K^\sigma (y_{M_K^\sigma}^n - y_K^n).
\end{aligned}$$

This relation yields

$$y_K^{n+1} = y_K^n\Big(1 - \frac{\delta t}{\rho_K^{n+1}|K|}\sum_{\sigma\in\mathscr{E}(K)}\beta_K^\sigma\big|F_{K,\sigma}^{n+1}\big|\Big) + \frac{\delta t}{|K|}\sum_{\sigma\in\mathscr{E}(K)} y_{M_K^\sigma}^n\beta_K^\sigma\big|F_{K,\sigma}^{n+1}\big|,$$

which concludes the proof under the hypothesis that CFL ≤ 1. □

We now need to reformulate (4.32) in order to construct the stability interval. Let $\sigma\in\mathscr{E}$, let us denote by V^- and V^+ the upstream and downstream cell separated by σ, and by $\mathscr{V}_\sigma(V^-)$ and $\mathscr{V}_\sigma(V^+)$ two sets of neighbouring cells of V^- and V^+ respectively, and let us suppose:

$$\text{(H1)} - \exists M\in\mathscr{V}_\sigma(V^+) \text{ s.t. } u_\sigma^n \in |[u_M^n, u_M^n + \frac{\zeta^+}{2}(u_{V^+}^n - u_M^n)]|,$$

$$\text{(H2)} - \exists M\in\mathscr{V}_\sigma(V^-) \text{ s.t. } u_\sigma^n \in |[u_{V^-}^n, u_{V^-}^n + \frac{\zeta^-}{2}(u_{V^-}^n - u_M^n)]|,$$

where for a, $b\in\mathbb{R}$, we denote by $|[a,b]|$ the interval $\{\alpha a + (1-\alpha)b,\ \alpha\in[0,1]\}$, and ζ^+ and ζ^- are two numerical parameters lying in the interval $[0,2]$.

Conditions (H1)-(H2) and (4.32) are linked in the following way: let $K\in\mathscr{M}$ and $\sigma\in\mathscr{E}(K)$. If $F_{K,\sigma}^n\leq 0$, *i.e.* K is the downstream cell for σ, denoted above by V^+, since $\zeta^+\in[0,2]$, condition (H1) yields that there exists $M\in\mathscr{M}$ such that $u_\sigma^n\in|[u_K^n,u_M^n]|$, which is (4.32). Otherwise, *i.e.* if $F_{K,\sigma}^n\geq 0$ and K is the upstream cell for σ, denoted above by V^-, condition (H2) yields that there exists $M\in\mathscr{M}$ such that $y_\sigma^n\in|[y_K^n, 2y_K^n - y_M^n]|$, so $y_\sigma^n - y_K^n\in|[0, y_K^n - y_M^n]|$, which is once again (4.32).

Remark 4.5. *For $\sigma\in\mathscr{E}$, if $V^-\in\mathscr{V}_\sigma(V^+)$, the upstream choice $y_\sigma^n = y_{V^-}^n$ always satisfies the conditions (H1)-(H2), and is the only one to satisfy them if we choose $\zeta^- = \zeta^+ = 0$.*

Remark 4.6 (1D case)**.** *Let us take the example of an interface σ separating K_i and K_{i+1} in a 1D case (see Figure 4.10 for the notations), with a uniform meshing and a positive advection velocity, so that $V^- = K_i$ and $V^+ = K_{i+1}$. In 1D, a natural choice is $\mathscr{V}_\sigma(K_i) = \{K_{i-1}\}$ and $\mathscr{V}_\sigma(K_{i+1}) = \{K_i\}$. On Figure 4.10, we sketch: on the left, the admissible interval given by (H1) with $\zeta^+ = 1$ (green) and $\zeta^+ = 2$ (orange); on the right, the admissible interval given by (H2) with $\zeta^- = 1$ (green) and $\zeta^- = 2$ (orange). The parameters ζ^- and ζ^+ may be seen as limiting the admissible slope between $(\boldsymbol{x}_i, y_i^n)$ and $(\boldsymbol{x}_\sigma, y_\sigma^n)$ (with $\boldsymbol{x}_i$ the abscissa of the mass centre of K_i and $\boldsymbol{x}_\sigma$ the abscissa of σ), with respect to a left and right slope, respectively. For $\zeta^- = \zeta^+ = 1$, one recognizes the usual minmod limiter (*e.g.* Godlewski and P.-A. Raviart 1996, Chapter III). Note that, since, on the example depicted on Figure 4.10, the discrete function y^n has an extremum in K_i, the combination of the conditions (H1) and (H2) imposes that, as usual, the only admissible value for y_σ^n is the upwind one.*

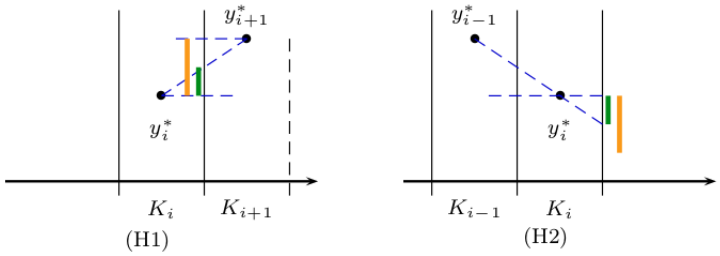


Figure 4.10 – Conditions (H1) and (H2) in 1D.

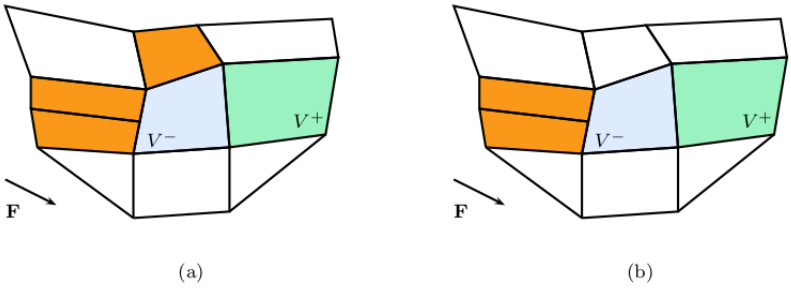


Figure 4.11 – Notations for the definition of the limitation process. In orange, control volumes of the set $\mathcal{V}_\sigma(V^-)$ for $\sigma = V^-|V^+$, with a constant advection field **F**: upwind cells (a) or opposite cells (b).

Finally, we need to specify the choice of the sets $\mathcal{V}_\sigma(V^-)$ and $\mathcal{V}_\sigma(V^+)$. Here, we just set $\mathcal{V}_\sigma(V^+) = \{V^-\}$; such a choice guarantees that at least the upstream choice is in the intersection of the intervals defined by (H1) and (H2), as explained in Remark 4.6. The set $\mathcal{V}_\sigma(V^-)$ may be defined in two different ways (*cf.* Figure 4.11):

– as the "upstream cells" to V^-, *i.e.*

$$\mathcal{V}_\sigma(V^-) = \{L \in \mathcal{M},\ L \text{ shares a face } \sigma \text{ with } V^- \text{ and } F_{V^-,\sigma} \leq 0\},$$

– when this makes sense (*i.e.* with a mesh obtained by Q_1 mappings from the $(0,1)^d$ reference element), the opposite cells to σ in V^- are chosen. Note that for a structured mesh, this choice allows to recover the usual minmod limiter.

4.B An anti-diffusive scheme

The scheme proposed in Després and Lagoutière 2002 by of B. Després and F. Lagoutière for the constant velocity advection problem presents some interesting properties in one space dimension (and may be extended to structured multi-dimensional meshes using alternate directions techniques); in particular, it notably limits the numerical diffusion. We extend here this scheme to work with unstructured meshes for

which the "opposite cell to a face" (in the sense introduced in the previous section) may be defined and with a variable density. With the same notations as in the previous section, for $\sigma \in \mathscr{E}_{\text{int}}$, $\sigma = K|L$ with $F_{K,\sigma}^{n+1} \geq 0$,

- the tentative value for y_σ is chosen as the downwind value, *i.e.* $y_\sigma^n = y_L^n$,
- Then we project y_σ^n on the interval

$$I_\sigma = \left[y_K^n,\, y_K^n + \frac{1-\nu}{\nu}(y_K - y_M) \right], \quad \nu = \frac{|F_{K,\sigma}^{n+1}|\,\delta t}{\rho_K^{n+1}\,|K|},$$

  where $M \in \mathscr{M}$ is the control volume which stands at the opposite side of K with respect to L.

The original scheme presented in Després and Lagoutière 2002 is recovered by this formulation for the one-dimensional constant velocity convection equation. In addition, by arguments similar to those of the previous section, the discretization proposed here may be shown to satisfy a discrete maximum principle.

5 Modelling of a spherical deflagration at constant speed

Sommaire

Abstract. We build in this paper a numerical solution procedure to compute the flow induced by a spherical flame expanding from a point source at a constant expansion velocity, with an instantaneous chemical reaction. The solution is supposed to be self-similar and the flow is split in three zones: an inner zone composed of burnt gases at rest, an intermediate zone where the solution is regular and the initial atmosphere composed of fresh gases at rest. The intermediate zone is bounded by the reactive shock (inner side) and the so-called precursor shock (outer side), for which Rankine-Hugoniot conditions are written; the solution in this zone is governed by two ordinary differential equations which are solved numerically. We show that, for any admissible precursor shock speed, the construction combining this numerical resolution with the exploitation of jump conditions is unique, and yields decreasing pressure, density and velocity profiles in the intermediate zone. In addition, the reactive shock speed is larger than the velocity on the outer side of the shock, which is consistent with the fact that the difference of these two quantities is the so-called flame velocity, *i.e.* the (relative) velocity at which the chemical reaction progresses in the fresh gases. Finally, we also observe numerically that the function giving the flame velocity as a function of the precursor shock speed is increasing; this allows to embed the resolution in a Newton-like procedure to compute the flow for a given flame speed (instead of for a given precursor shock speed). The resulting numerical algorithm is applied to stoichiometric hydrogen-air mixtures.

keywords spherical flames, reactive Euler equations, Riemann problems.

5.1 Problem position

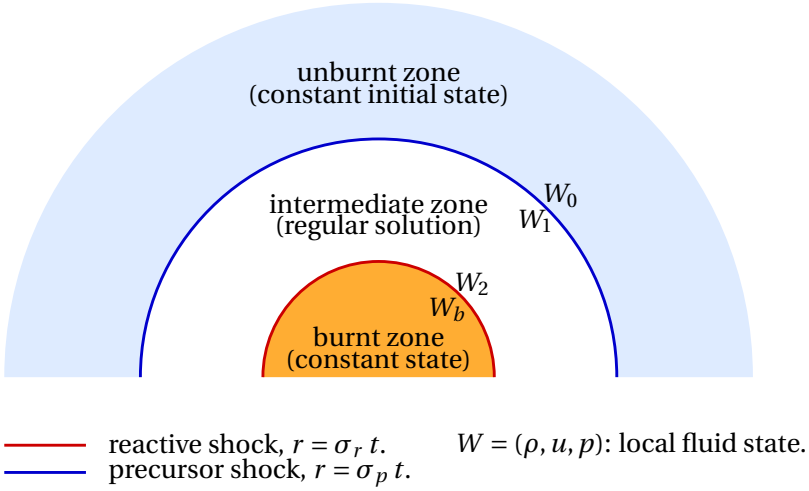


Figure 5.1 – Structure of the solution.

We address the flame propagation in a reactive infinite atmosphere of initial constant composition. The ignition is supposed to occur at a single point (chosen to be the origin of $\mathbb{R}^3$) and the flow is supposed to satisfy a spherical symmetry property: the density ρ, the pressure p, the internal energy e and the entropy s only depend on the distance r to the origin and the velocity reads $\boldsymbol{u} = u\boldsymbol{r}/r$, where $\boldsymbol{r}$ stands for the position vector. The flame is supposed to be infinitely thin and to move at a constant speed. The flow is governed by the Euler equations, and we seek a solution with the following structure:

— the solution is self-similar, *i.e.* the quantities ρ, p, e, s and u are functions of the variable $x = r/t$ only.

— the flow is split in three zones, referred to as the inner, intermediate and outer zones. The inner zone stands for the burnt zone while, in the other two zones, the gas is supposed to be in its initial (referred to as fresh or unburnt) composition. Burnt and fresh gases differ by the expression of the total energy:

$$E = \frac{1}{2}u^2 + e - \zeta_b\, Q, \quad \zeta_b = 1 \text{ in the burnt zone, } \zeta_b = 0 \text{ in the fresh zone,} \tag{5.1}$$

with $Q > 0$ the chemical heat reaction. Both burnt and unburnt gases are considered as ideal gases, possibly with different heat capacity ratios:

$$p = (\gamma - 1)\rho e, \quad \gamma = \gamma_b \text{ for burnt gases, } \gamma = \gamma_u \text{ for unburnt gases.}$$

— In the burnt zone, the solution is supposed to be constant; this constant state is denoted by $W_b = (\rho_b, u_b, p_b)$. For symmetry reasons, the fluid is at rest in this

zone, *i.e.* $u_b = 0$.

— The burnt and intermediate zones are separated by a shock, which coincides with the flame front. This shock is called the reactive shock, and travels at a constant speed σ_r. The outer state of the shock is denoted by $W_2 = (\rho_2, u_2, p_2)$. Note that the usual Rankine-Hugoniot jump conditions apply at the reactive shock, up to the fact that the expression of the total energy in the inner and outer states differ (see Equation (5.1)).

— The intermediate and the outer zone are separated by a 3-shock, referred to as the precursor shock, and travelling at a velocity denoted by σ_p. We denote by $W_1 = (\rho_1, u_1, p_1)$ the inner state of the precursor shock, and, since the usual jump conditions for the Euler equations apply, we have $\sigma_p \geq u_1$. In the outer zone, conditions are constant and equal to the initial condition $W_0 = (\rho_0, u_0, p_0)$; the fluid is at rest, *i.e.* $u_0 = 0$.

— In the intermediate zones, the states W_2 and W_1 are supposed to be linked by a regular solution.

In addition, for physical reasons, we expect that

$$u_2 > 0 \text{ and } \sigma_r = u_2 + u_f \text{ with } u_f > 0. \tag{5.2}$$

Indeed, the velocity u_f is the velocity at which the chemical reaction progresses in the fresh gases; these are pushed away from the origin by the expansion of the burnt gases, and therefore $u_2 > 0$.

The aim of this paper is to build a numerical procedure to compute a solution with the above described structure. More precisely speaking, we present the two following developments:

— First, for a given precursor shock speed σ_p, we derive a solution with the desired structure in a constructive way (and this construction yields a unique solution), and propose a simple numerical scheme to compute it. Moreover, the constructed solution is such that the inequalities (5.2) are satisfied (in fact, we obtain that $\sigma_r - u_2 > 0$ since we seek and find a solution such that $x - u(x) > 0$ in the whole intermediate zone), and thus yields a physically meaningful flame velocity u_f.

— As a by-product, we numerically obtain the velocity u_f as a function of σ_p, *i.e.* we construct a function $\widetilde{\mathcal{G}}$ such that $u_f = \widetilde{\mathcal{G}}(\sigma_p)$, and observe that this function $\widetilde{\mathcal{G}}$ is strictly increasing, which was expected from physical reasons (the faster the combustion, the stronger the generated shock-wave). It is thus easy to build an iteration to compute the flow associated to a given u_f, which is generally the problem of physical interest.

Finally, this process is implemented in the free software CALIF^{3}S n.d. developped at the french Institut de Radioprotection et de Sûreté Nucléaire (IRSN) and applied to obtain solutions as a function of u_f for a stoichiometric mixture of hydrogen and air.

The derivation of a solution for the same problem may be found in Kuhl, Kamel, and

Oppenheim 1973; however, the techniques used in this latter paper are different (the solution is performed in the phase space), and the uniqueness of the construction together with the proof of the decreasing properties of the solution are not explicit. The developments in Kuhl, Kamel, and Oppenheim 1973 are built upon techniques developed for non-reactive problems in Sedov 1945; Taylor 1946. Approximate solutions in closed form are given in Guirano, Bash, and Lee 1976; Cambray and Deshaies 1977, and extensions to accelerating flames may be found in Strehlow, Luckritz, Adamczyk, et al. 1979; Deshaies and Leyer 1981. Finally, the complete solution of the plane case (*i.e.* the one-dimensional case in cartesian coordinates), in closed form, is given in Beccantini and Studer 2010.

The remainder of the this paper is organized as follows: In Section 5.2 we present the Euler equations in spherical coordinates. Then in Section 5.3 we describe the solution for a given precursor shock speed followed by the solution for a given flame speed in Section 5.4. Finally in Section 5.5 we perform an application for a flame propagating in a stoichiometric mixture of hydrogen and air.

5.2 Euler equations in spherical coordinates

The Euler equations read in Cartesian coordinates:

$$\partial_t \bar{\rho} + \operatorname{div}(\bar{\rho}\bar{\boldsymbol{u}}) = 0, \tag{5.3a}$$

$$\partial_t (\bar{\rho}\bar{\boldsymbol{u}}) + \operatorname{div}(\bar{\rho}\bar{\boldsymbol{u}} \otimes \bar{\boldsymbol{u}}) + \nabla \bar{p} = 0, \tag{5.3b}$$

$$\partial_t (\bar{\rho}\bar{E}) + \operatorname{div}(\bar{\rho}\bar{E}\bar{\boldsymbol{u}} + \bar{p}\bar{\boldsymbol{u}}) = 0, \tag{5.3c}$$

where $\bar{\rho} = \bar{\rho}(t,\boldsymbol{x}) \in \mathbb{R}$ the density, $\bar{\boldsymbol{u}} = \bar{\boldsymbol{u}}(t,\boldsymbol{x}) \in \mathbb{R}^3$ the velocity, $\bar{p} = \bar{p}(t,\boldsymbol{x}) \in \mathbb{R}$ the pressure and $\bar{E} = \bar{E}(t,\boldsymbol{x}) \in \mathbb{R}$ the total energy for all $t \in \mathbb{R}$ and $\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$. This system is closed by the equation of state, which for a perfect gas, is given by

$$\bar{E} = \frac{1}{2}|\bar{\boldsymbol{u}}|^2 + \bar{e}, \text{ with } \bar{p} = (\gamma - 1)\bar{\rho}\bar{e}, \tag{5.4}$$

where $\bar{e} = \bar{e}(t,\boldsymbol{x})$ the internal energy and $\gamma > 1$ the heat capacity ratio. We suppose that the flow satisfies a spherical symmetry assumption, so the solution of equations (5.3)-(5.4) may be recast as:

$$\bar{\rho}(t,\boldsymbol{x}) = \rho(t,r),\ \bar{p}(t,\boldsymbol{x}) = p(t,r),\ \bar{E}(t,\boldsymbol{x}) = E(t,r) \text{ and } \bar{\boldsymbol{u}}(t,\boldsymbol{x}) = u(t,r)\frac{\boldsymbol{x}}{r}, \tag{5.5}$$

with $r = |\boldsymbol{x}|$ and where $(\rho, u, p, E)(t,r) \in \mathbb{R}^4$ are scalar functions, *i.e.* $(\rho, u, p, E) \in \mathbb{R}^4$. The aim of this section is to derive the system of equations satisfied by (ρ, u, p, E). We suppose first that these functions are regular, so we obtain the strong form of the so-called Euler equations in spherical coordinates; then we turn to the weak form, valid for discontinuous solutions.

5.2.1 Regular solutions

Let us use the notation $\partial_i = \dfrac{\partial}{\partial x_i}$. We begin by deriving the following three identities:

(*a*) $\partial_i r = \dfrac{x_i}{r}$, for $i = 1, 2, 3$.

(*b*) If $f = f(r)$, $\mathrm{div}(f\bar{\boldsymbol{u}}) = \dfrac{1}{r^2}\partial_r(r^2 f u)$.

(*c*) If $f = f(r)$, $\nabla f = \partial_r f \dfrac{\boldsymbol{x}}{r}$.

The first item is a straightforward consequence of the definition $r = |\boldsymbol{x}|$. For Item (*b*), we have, thanks to (*a*),

$$\begin{aligned}
\mathrm{div}(f\bar{\boldsymbol{u}}) &= \sum_{i=1}^{3} \partial_i (f u \frac{x_i}{r}) \\
&= f u \sum_{i=1}^{3} \partial_i (\frac{x_i}{r}) + \sum_{i=1}^{3} \frac{x_i}{r} \partial_i (f u) \\
&= f u \sum_{i=1}^{3} (\frac{1}{r} - \frac{x_i^2}{r^3}) + \sum_{i=1}^{3} \frac{x_i}{r} \partial_i r \, \partial_r (f u) \\
&= \frac{1}{r} f u \sum_{i=1}^{3} (1 - \frac{x_i^2}{r^2}) + \partial_r (f u) \sum_{i=1}^{3} \frac{x_i^2}{r^2} \\
&= \frac{2}{r} f u + \partial_r (f u) = \frac{1}{r^2} \partial_r (r^2 f u).
\end{aligned}$$

Item (*c*) is an immediate consequence of (*a*).

We are now in position to state the following lemma.

Lemma 5.1. *Suppose that* $(\bar{\rho}, \bar{\boldsymbol{u}}, \bar{p}, \bar{E})$ *is solution of* (5.3)*; then* (ρ, u, p, E) *satisfies:*

$$\partial_t (r^2 \rho) + \partial_r (r^2 \rho u) = 0, \tag{5.6a}$$

$$\partial_t (r^2 \rho u) + \partial_r (r^2 (\rho u^2 + p)) = 2 r p, \tag{5.6b}$$

$$\partial_t (r^2 \rho E) + \partial_r (r^2 (\rho u E + p u)) = 0, \tag{5.6c}$$

with $E = \frac{1}{2} u^2 + e$ *and* $p = (\gamma - 1)\rho e$.

Proof. The mass balance equation (5.6a) is a straightforward consequence of Item (*b*). For the momentum balance equation, we first remark that, for any function $f(r)$, we have:

$$\mathrm{div}(f x_i \bar{\boldsymbol{u}}) = \frac{x_i}{r^2} \partial_r (r^2 f u) + \frac{x_i}{r} f u.$$

Indeed, this relation follows from the development $\mathrm{div}(f x_i \bar{\boldsymbol{u}}) = x_i \mathrm{div}(f \bar{\boldsymbol{u}}) + f \bar{\boldsymbol{u}} \cdot \nabla x_i$

thanks to Item (b). Applying this identity with $f = \dfrac{\rho u}{r}$, we thus obtain

$$\operatorname{div}(\rho u_i \boldsymbol{u}) = \operatorname{div}(\frac{\rho u}{r} x_i \boldsymbol{u}) = \frac{x_i}{r^2}\left(\partial_r(\rho u^2 r) + \rho u^2\right)$$
$$= \frac{x_i}{r^2}\left(r\partial_r(\rho u^2) + 2\rho u^2\right) = \frac{x_i}{r^3}\partial_r(r^2\rho u^2).$$

Thanks to this relation and using (c) for the pressure gradient, for $i = 1,\ 2,\ 3$, the $i-$th component of the momentum equation reads:

$$\frac{x_i}{r}\left[\partial_t(\rho u) + \frac{1}{r^2}\partial_r(r^2\rho u^2) + \partial_r p\right] = 0.$$

The three components of this vector balance equation thus boil down to the single relation:

$$\partial_t(\rho u) + \frac{1}{r^2}\partial_r(r^2\rho u^2) + \partial_r p = 0.$$

Multiplying by r^2, we obtain the conservative form (5.6b). For the energy balance equation, we first note that $\bar{E} = \frac{1}{2}|\bar{\boldsymbol{u}}|^2 + \bar{e} = \frac{1}{2}u^2 + e = E$. Then, using once again (b), we get

$$\operatorname{div}(\bar{\rho}\bar{E}\bar{\boldsymbol{u}}) = \frac{1}{r^2}\partial_r(\rho E u r^2) \text{ and } \operatorname{div}(\bar{p}\bar{\boldsymbol{u}}) = \frac{1}{r^2}\partial_r(p u r^2),$$

and (5.6c) follows by adding the time derivative of $(\bar{\rho}\bar{E})$. □

We can apply the same process for the entropy balance equation. In the Cartesian system of coordinates, this relation reads, for regular solutions:

$$\partial_t(\bar{\rho}\bar{s}) + \operatorname{div}(\bar{\rho}\bar{\boldsymbol{u}}\bar{s}) = 0, \quad \text{with } \bar{s} = \frac{\bar{p}}{\bar{\rho}^\gamma}, \tag{5.7}$$

with $\bar{s} = \bar{s}(t,\boldsymbol{x})$. By this latter expression, under spherical symmetry assumption, there exists a function $s(t,r)$ such that $\bar{s}(t,\boldsymbol{x}) = s(t,r)$. This latter function satisfies

$$s = \frac{p}{\rho^\gamma},$$

and a straightforward application of Identity (b) with $f = \rho s$ yields, from (5.7):

$$\partial_t(r^2\rho s) + \partial_r(r^2\rho s u) = 0. \tag{5.8}$$

5.2.2 Weak solutions

Let us now treat the case of no classical solution of (5.3) and for the sake of simplicity we only consider the mass balance equation for a velocity $\bar{\boldsymbol{u}}$ given by (5.5). So $\bar{\rho}$ is

weak solution of $\partial_t\bar{\rho}+\operatorname{div}(\bar{\rho}\bar{\boldsymbol{u}})=0$, if $\forall\bar{\varphi}\in C_c^\infty(\mathbb{R}_+\times\mathbb{R}^d,\mathbb{R})$, we have

$$\int_{\mathbb{R}_+}\int_{\mathbb{R}}\bar{\rho}\partial_t\bar{\varphi}+\bar{\rho}\bar{\boldsymbol{u}}\cdot\nabla\bar{\varphi}\,d\boldsymbol{x}\,dt+\int_{\mathbb{R}^d}\bar{\rho}_0(\boldsymbol{x})\bar{\varphi}(0,\boldsymbol{x})\,d\boldsymbol{x}=0, \tag{5.9}$$

with $\bar{\rho}_0(\boldsymbol{x})=\bar{\rho}(0,\boldsymbol{x})$.

Lemma 5.2. *Let Ω a open set of $\mathbb{R}_+^*$ and let consider $\bar{\rho}$, given by (5.5), a weak solution of (5.9) in the form*

$$\rho(t,r)=\begin{cases}\rho_L & \text{if } r\in\Omega^-\\ \rho_R & \text{if } r\in\Omega^+\end{cases}\quad \text{with } u(t,r)=\begin{cases}u_L & \text{if } r\in\Omega^-\\ u_R & \text{if } r\in\Omega^+\end{cases} \tag{5.10}$$

where $\Omega^-=\{r\in\Omega, r\le t\sigma\}$ and $\Omega^+=\{r\in\Omega, r>t\sigma\}$ with (ρ_R,u_R) and (ρ_L,u_L) are constant in the domains $\mathbb{R}_+\times\Omega^+$ and $\mathbb{R}_+\times\Omega^-$ respectively.
Let $\bar{\varphi}\in C_c^\infty(\mathbb{R}_+\times\mathbb{R}^d,\mathbb{R})$ and $\varphi\in C_c^\infty(\mathbb{R}_+\times\mathbb{R}_+^*,\mathbb{R})$ such that $\bar{\varphi}(t,\boldsymbol{x})=\varphi(t,r)$, then ρ satisfies the following weak formulation :

$$\int_{\mathbb{R}_+}\int_{\mathbb{R}_+^*}(r^2\rho\partial_t\varphi+r^2\rho u\partial_r\varphi)\,dr\,dt+\int_{\mathbb{R}_+^*}r^2\rho_0(r)\varphi(0,r)\,dr=0. \tag{5.11}$$

Furthermore σ checks the relationship:

$$\sigma(\rho_R-\rho_L)=(\rho_Ru_R-\rho_Lu_L). \tag{5.12}$$

Proof. Let $\bar{\varphi}\in C_c^\infty(\mathbb{R}_+\times\mathbb{R}^d,\mathbb{R})$ and let introduce the function φ such that $\bar{\varphi}(t,\boldsymbol{x})=\varphi(t,r)$, then $\varphi\in C_c^\infty(\mathbb{R}_+\times\mathbb{R}_+^*,\mathbb{R})$. So we have that

$$\begin{aligned}
&\int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\bar{\rho}(\partial_t\bar{\varphi}+\bar{\rho}\bar{\boldsymbol{u}}\cdot\nabla\bar{\varphi})\,d\boldsymbol{x}\,dt\\
&\qquad=\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}(\rho(t,r)\partial_t\varphi(t,r)+\rho(t,r)u(t,r)\frac{\boldsymbol{x}}{r}\cdot\nabla\varphi(t,r))r^2\,dr\,dt\\
&\qquad=\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}(r^2\rho(t,r)\partial_t\varphi(t,r)+r^2\rho(t,r)u(t,r)\frac{\boldsymbol{x}}{r}\cdot\frac{\boldsymbol{x}}{r}\partial_r\varphi(t,r))\,dr\,dt\\
&\qquad\qquad=\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}(r^2\rho(t,r)\partial_t\varphi(t,r)+r^2\rho(t,r)u(t,r)\partial_r\varphi(t,r))\,dr\,dt.
\end{aligned}$$

We have also that

$$\int_{\mathbb{R}^d}\bar{\rho}_0(\boldsymbol{x})\bar{\varphi}(0,\boldsymbol{x})\,d\boldsymbol{x}=\int_{\mathbb{R}_+}r^2\rho_0(r)\varphi(0,r)\,dr$$

Thus (5.9) reads:

$$\int_{\mathbb{R}_+}\int_{\mathbb{R}_+} (r^2\rho(t,r)\partial_t\varphi(t,r) + r^2\rho(t,r)u(t,r)\partial_r\varphi(t,r))\,dr\,dt \\ + \int_{\mathbb{R}_+} r^2\rho_0(r)\varphi(0,r)\,dr = 0,\ \forall\varphi\in C_c^\infty(\mathbb{R}_+\times\mathbb{R}_+,\mathbb{R}).$$

According to the definition (5.10) and decomposing the integral into space on Ω^+ and Ω^-, we obtain that

$$\int_{\mathbb{R}_+}\int_{\Omega^+} r^2\rho_R\partial_t\varphi + r^2\rho_R u_R\partial_r\varphi\,dr\,dt + \int_{\mathbb{R}_+}\int_{\Omega^-} r^2\rho_L\partial_t\varphi + r^2\rho_L u_L\partial_r\varphi\,dr\,dt = 0.$$

Thus as in the scalar case, we can show that, σ satisfies the following Rankine Hugoniot relationship:

$$\sigma(r^2(\rho_R-\rho_L)) = (r^2(\rho_R u_R - \rho_L u_L)).$$

The lemma is thus proved. □

5.3 Solution for a given precursor shock speed

We first propose a constructive derivation of the solution for a given precursor shock speed, and then state the numerical scheme to compute it.

5.3.1 Derivation of the solution

Since the fluid state in the outer zone W_0 is given, we may equivalently use hereafter either σ_p or the precursor shock Mach number defined by $M_p = \sigma_p/c_0$, with $c_0 = (\gamma_u p_0/\rho_0)^{1/2}$ the speed of sound in the outer zone. We recall that, for entropy condition reasons, $M_p > 1$.

Left state of a shock as a function of the right state and the shock velocity
In this section, we recall a classical computation which consists in determining the left state of a shock as a function of the right state and the shock velocity.

Lemma 5.3. *Let $W_1 = (\rho_1, u_1, p_1)$ be the left state of a shock travelling at the given speed σ. Let $W_R = (\rho_R, u_R, p_R)$ be the given right state, which is supposed to satisfy $u_R = 0$. Let c_R be the speed of sound in the right state,*i.e. *$c_R^2 = \gamma p_R/c_R$, and let M be the Mach*

number associated to the incident shock, defined by $M = \sigma / c_R$. Then W_1 is given by:

$$\rho_1 = \frac{\gamma+1}{\gamma - 1 + \dfrac{2}{M^2}}\,\rho_R, \tag{5.13a}$$

$$u_1 = (1 - \frac{\rho_R}{\rho_1})\,\sigma, \tag{5.13b}$$

$$p_1 = p_R + (1 - \frac{\rho_R}{\rho_1})\rho_R \sigma^2. \tag{5.13c}$$

Proof. We first change the coordinate system, in such a way that the shock is steady in the new coordinate system. The density and the pressure are left unchanged, while the velocity is now $u_1 - \sigma$ and $-\sigma$ in the left and right state respectively. In this coordinate system, the Rankine-Hugoniot conditions imply that the jump of the fluxes vanishes, which reads for the Euler equations:

$$\rho_1 (u_1 - \sigma) = \rho_R (-\sigma), \tag{5.14a}$$

$$\rho_1 (u_1 - \sigma)^2 + p_1 = \rho_R (-\sigma)^2 + p_R, \tag{5.14b}$$

$$\frac{1}{2}\rho_1 (u_1 - \sigma)^3 + \rho_1 e_1 (u_1 - \sigma) + p_1 (u_1 - \sigma) = \frac{1}{2}\rho_R (-\sigma)^3 + \rho_R e_R (-\sigma) + p_R (-\sigma). \tag{5.14c}$$

This system must be complemented by the equation of state $p = (\gamma - 1)\rho e$ and $p_R = (\gamma - 1)\rho_R e_R$. Thanks to this relation, we may recast (5.14c) as:

$$\rho_1 (\sigma - u_1)^3 + \xi\, p_1 (\sigma - u_1) = \rho_R \sigma^3 + \xi\, \sigma\, p_R, \quad \xi = \frac{2\gamma}{\gamma - 1}. \tag{5.15}$$

Relation (5.14b) reads

$$\frac{1}{\rho_1}\left[\rho_1 (u_1 - \sigma)\right]^2 + p_1 = \rho_R (-\sigma)^2 + p_R,$$

so, using (5.14a):

$$\frac{1}{\rho_1}(\rho_R \sigma)^2 + p_1 = \rho_R \sigma^2 + p_R.$$

We thus obtain p_1 as a function of known quantities (*i.e.* σ and the right state) and ρ_1 only:

$$p_1 = p_R + (1 - \frac{\rho_R}{\rho_1})\,\sigma^2. \tag{5.16}$$

We now notice that substituting, in the jump condition associated to the energy balance (5.15), this expression for p_1 and $(\rho_R/\rho_1)\,\sigma$ for $(u_1 - \sigma)$, thanks once again to

Equation (5.14a), we get an equation for ρ only:

$$\rho_1 \left(\frac{\rho_R}{\rho_1}\sigma\right)^3 + \xi\left(p_R + (1-\frac{\rho_R}{\rho_1})\sigma^2\right)\left(\frac{\rho_R}{\rho_1}\sigma\right) = \rho_R\sigma^3 + \xi\sigma.$$

If $\sigma = 0$, the first jump condition implies $u_1 = 0$ (excluding $\rho_1 = 0$), then the second one yields $p_1 = P_R$ and the third one is automatically satisfied: the considered discontinuity is a (stationary) contact. In such a case, the right state remains partially undermined by the jump conditions: ρ_1 and e_1 may take any value satisfying $(\gamma-1)\rho_1 e_1 = p_R$. If we only consider a shock, $\sigma \neq 0$ and the last relation may be simplified by σ. Reordering, we get:

$$\rho_R\sigma^2\left((\frac{\rho_R}{\rho_1})^2 - 1\right) + \xi\, p_R(\frac{\rho_R}{\rho_1} - 1) - \xi\rho_R\sigma^2\frac{\rho_R}{\rho_1}(\frac{\rho_R}{\rho_1} - 1) = 0.$$

The case $\rho_1 = \rho_R$ has no interest: it yields, by the first jump condition, $u_1 = 0$, and the second one implies that $p_1 = p_R$, which means *in fine* that $W_1 = W_R$, *i.e.* that there is no discontinuity at all. We may thus simplify by $1 - \rho_R/\rho_1$, to obtain a linear equation for the ratio ρ_R/ρ_1. Solving this latter equation, we obtain:

$$\rho_1 = \frac{\gamma+1}{\gamma - 1 + \dfrac{2}{M^2}}\,\rho_R, \quad \text{with } M = \frac{\sigma}{c_R},\ c_R^2 = \gamma\frac{p_R}{\rho_R}.$$

We thus obtain (5.13a). Relation (5.13b) is a straightforward consequence of (5.14a) and (5.13c) was already proven (Relation (5.16) below). The proof is thus complete. □

For a 3-shock, entropy conditions requires that $\sigma > c_R$, which is equivalent to $M > 1$. We thus have $\rho_1 > p_R$, $p_1 > p_R$ and $0 < u_1 < \sigma$.

It is worth noting that it is now easy to relax the assumption $u_R = 0$ of Lemma 5.3. Indeed, for the general case, we may work in the system of coordinates in translation at the velocity u_R with respect to the initial one: in the new coordinate system, the right state is now at rest and Lemma 5.3 applies, replacing σ by $\sigma - u_R$ and u_1 by $u_1 - u_R$ in the definition of the Mach number M and in the system of equations (5.13).

A relation satisfied by the right state of a shock when the left state is at rest
Here we perform a technical computation motivated by the following arguments. Let us suppose that a shock travelling at the speed σ_r separates a left state denoted by $W_b = (\rho_b, u_b, p_b)$ and a right state denoted by $W_2 = (\rho_2, u_2, p_2)$, and that $u_b = 0$. The Rankine-Hugoniot conditions yield 3 independent equations, and thus constitute a system in which ρ_b and p_b may be eliminated, to obtain a relation linking W_2 and σ_r only. It is this relation that we now derive, supposing that the following specific

constitutive relations hold for both states:

$$E_b = \frac{1}{2}u_b^2 + e_b + Q,\ p_b = (\gamma_b - 1)\,\rho_b\, e_b, \tag{5.17a}$$

$$E_2 = \frac{1}{2}u_2^2 + e_2,\ p_2 = (\gamma_u - 1)\,\rho_2\, e_2. \tag{5.17b}$$

Note that eliminating ρ_b and p_b consists in establishing an expression of these quantities (and thus of W_b, since $u_b = 0$) as a function of W_2 and σ_r. All of these relations, *i.e.* the equation linking W_2 and σ_r and the expression of W_b as a function of these variables, are gathered in the following lemma.

Lemma 5.4 (Some conditions at the reactive shock)**.** *The state W_2 and the shock speed σ_r satisfy the following relation*

$$\frac{1}{2}u_2^2 + \frac{1}{\gamma_b - 1}u_2\sigma_r + \Big(\frac{\gamma_u}{\gamma_u - 1} - \frac{\gamma_b}{\gamma_b - 1}\frac{\sigma_r}{\sigma_r - u_2}\Big)\frac{p_2}{\rho_2} + Q = 0. \tag{5.18}$$

In addition, W_b is given as a function of W_2 and σ_r by:

$$\rho_b = \rho_2\Big(\frac{\sigma_r - u_2}{\sigma_r}\Big), \quad p_b = p_2 - \rho_2 u_2(\sigma_r - u_2). \tag{5.19}$$

Proof. Using the standard change of coordinates to work in the coordinate system in which the shock is at rest (see *e.g.* Godlewski and P.-A. Raviart 1996), the Rankine-Hugoniot relationships (which boil down to the fact that the jump of the fluxes vanishes) through the shock for the mass and momentum balance equations read respectively:

$$\rho_b\,\sigma_r = \rho_2\,(\sigma_r - u_2), \quad \rho_b\,\sigma_r^2 + p_b = \rho_2(\sigma_r - u_2)^2 + p_2.$$

These two relations readily yields the expression (5.19) of W_b as a function of W_2 and σ_r that we are looking for:

$$\rho_b = \rho_2\,\Big(\frac{\sigma_r - u_2}{\sigma_r}\Big) \text{ and } p_b = p_2 - \rho_2\,u_2\,(\sigma_r - u_2).$$

Let us now write the Rankine-Hugoniot condition for the conservation equation of the total energy:

$$\rho_b\,\sigma_r\,E_b + \sigma_r\,p_b = \rho_2\,(\sigma_r - u_2)\,E_2 + (\sigma_r - u_2)\,p_2. \tag{5.20}$$

We may divide the left-hand side of his relation by $\rho_b\sigma_r$ and the right-hand side by $\rho_2(\sigma_r - u_2)$ (since these two expressions are equal by the jump condition associated to the mass balance equation), to obtain:

$$E_b + \frac{p_b}{\rho_b} = E_2 + \frac{p_2}{\rho_2}.$$

Using the constitutive relations (5.17) and the expression (5.19) of ρ_b and p_b in this equation yields (5.18) and thus concludes the proof.

□

To complete the derivation of the solution, we must now show that the following program makes sense: starting from $x = \sigma_p$, solve the Euler equations for $x \leq \sigma_p$ until the point $x = \sigma_r$ where Equation (5.18) is verified. The solution at this point is equal to W_2 and Equations (5.19) yield the burnt state W_b. Let us now embark on this development.

Governing equations in the intermediate zone Since we suppose that the solution is regular in this zone, we may replace the total energy balance in the Euler equations by the entropy equation, which, under the spherical symmetry assumption, yields the following system:

$$\partial_t(r^2\rho) + \partial_r(r^2\rho u) = 0, \tag{5.21a}$$
$$\partial_t(r^2\rho u) + \partial_r(r^2(\rho u^2 + p)) = 2rp, \tag{5.21b}$$
$$\partial_t(r^2\rho s) + \partial_r(r^2\rho s u) = 0. \tag{5.21c}$$

The mass balance equation (5.21a) may be developed to obtain:

$$\partial_t\rho + u\partial_r\rho + \rho\partial_r u + \frac{2}{r}\rho u = 0. \tag{5.22}$$

In addition, thanks to the mass balance equation and for a regular function $f = f(t,r)$, we have:

$$\partial_t(r^2\rho f) + \partial_r(r^2\rho f u) = \rho\,(\partial_t f + \rho u\partial_r f).$$

Using this identity in the momentum and entropy balances, *i.e.* Equation (5.21b) and (5.21c) respectively, we get:

$$\begin{aligned} &\partial_t u + u\partial_r u + \frac{1}{\rho}\partial_r p = 0, \\ &\partial_t s + u\partial_r s = 0. \end{aligned} \tag{5.23}$$

We now use the fact that, if a regular function $\varphi(t,r)$ only depends on $x = r/t$, which means that there exists $\tilde{\varphi} : \mathbb{R} \to \mathbb{R}$ such that $\tilde{\varphi}(x) = \varphi(t,r)$, we have

$$\partial_t\varphi(t,r) = -\frac{r}{t^2}\tilde{\varphi}(x) \text{ and } \partial_r\varphi(t,r) = \frac{1}{t}\tilde{\varphi}'(x).$$

Since we look for a self-similar solution, we may apply this identity to (5.22) and (5.23). Keeping the same notation for functions of the pair (t,r) and x for short, we obtain

the following system:

$$\frac{-x+u}{\rho}\,\rho'(x)+u'(x)+\frac{2u(x)}{x}=0, \tag{5.24a}$$

$$(-x+u(x))\,u'(x)+\frac{1}{\rho}(x)\,p'(x)=0, \tag{5.24b}$$

$$(u(x)-x)\,s'(x)=0. \tag{5.24c}$$

Let us now suppose that $u<x$ in the intermediate zone. Note that the fact that the precursor shock is a 3-shock implies that $u_1<\sigma_p$, so the assumed inequality is true in the outer boundary of the intermediate zone, and the assumption amounts to suppose that the intermediate zone ends (more precisely speaking, may be made to end in the construction of the solution) before $u=x$ occurs, which will be checked further. The last relation thus implies that the entropy remains constant over the zone:

$$s=\frac{p}{\rho^{\gamma_u}}=s_1=\frac{p_1}{\rho_1^{\gamma_u}}, \tag{5.25}$$

and this is a known value thanks to (5.13). We thus have $p'=\gamma_u\,s_1\,\rho^{\gamma_u-1}\rho'$; using $c^2=\gamma_u\,p/\rho=\gamma_u\,s_1\,\rho^{\gamma_u-1}$, we thus get $p'=c^2\,\rho'$. Substituting this expression in (5.24a)-(5.24b) and solving for ρ' and u', we get:

$$\rho'(x)=-\frac{2u(x)(u(x)-x)}{x\big((u(x)-x)^2-c(x)^2\big)}\,\rho(x), \tag{5.26a}$$

$$u'(x)=\frac{2c(x)^2}{x\big((u(x)-x)^2-c(x)^2\big)}\,u(x). \tag{5.26b}$$

This system of coupled ODEs is complemented by initial conditions, which consist in the data of the velocity and the density at the precursor shock, *i.e.* at the outer boundary of the intermediate zone $x=\sigma_p$:

$$\rho(\sigma_p)=\rho_1,\,u(\sigma_p)=u_1. \tag{5.27}$$

Existence, uniqueness and properties of the solution We begin by proving an *a priori* property of the solution, namely the fact that $\rho(x)$ and $u(x)$ are necessarily decreasing functions in the intermediate zone. To this end, we will invoke the following easy lemma, which is a consequence of the mean value theorem.

Lemma 5.5. *Let h be a continuously differentiable real function, let us suppose that there exists $a>0$ such that $h(a)>0$, and that h satisfies the property $h'(x)\leq 0$ if $h(x)>0$. Then $h(x)\geq h(a)$, for all $x\leq a$.*

From the expression (5.13), we know that $0<u_1<\sigma_p$ and $\rho_1>0$. Let us now

introduce σ_ℓ as the largest real number in $[0,\sigma_p)$ such that, for $x \in [\sigma_\ell,\sigma_p]$, $0 \le u \le x$ and $\rho \ge 0$. Note that such a closed interval exists by the continuity (assumed in this zone) of ρ and u. We are now in position to state the following result.

Lemma 5.6 (Variations of the solution)**.** *Let us suppose that the pair* (ρ,u) *satisfies* (5.26)*. Then* ρ *and* u *are two decreasing functions over* $[\sigma_\ell,\sigma_p]$*. Consequently,* $\rho \ge \rho_1 > 0$ *and* $u \ge u_1 > 0$ *over* $[\sigma_\ell,\sigma_p]$*.*

Proof. Let us consider the function $h : \mathbb{R}_+ \in \mathbb{R}$ defined by $h(x) = u(x) + c(x) - x$. First, we remark that, by the Lax entropy condition, we have $h(\sigma_p) = u_1 + c_1 - \sigma_p > 0$ (and this property may be checked using the expressions (5.13) of W_1). Second, if $h(x) > 0$, since by assumption $u(x) \le x$, $\rho \ge 0$ and $c \ge 0$, we have:

$$(u(x)-x)^2 - c(x)^2 = (u(x)-x-c(x))\,(u(x)-x+c(x)) \le 0.$$

Equations (5.26a) and (5.26b) thus readily imply that $\rho' \le 0$ and $u' \le 0$ since, still by assumption, $u \ge 0$ and $\rho \ge 0$. The function h is thus the sum of three non-increasing functions, and is hence non-increasing itself. Lemma 5.5 applies, and yields $h(x) = u(x)-x-c(x) \ge h(\sigma_p) > 0$ over the whole interval $[\sigma_\ell,\sigma_p]$, which in turn implies $\rho' \le 0$ and $u' \le 0$. We thus have $\rho \ge \rho_1 > 0$ and $u \ge u_1 > 0$, which finally yields $\rho' < 0$ and $u' < 0$ over $[\sigma_\ell,\sigma_p]$.

□

Note that the inequality $h(x) \ge h(\sigma_p) > 0$ derived in this proof implies that the denominator in Equations (5.26a) and (5.26b) does not vanish in the interval $[\sigma_\ell,\sigma_p]$. The right-hand side of System (5.26) is thus a C^∞ function of ρ, u and x, and the existence and uniqueness of a solution follows by the Cauchy-Lipschitz theorem. This result is stated in the following lemma.

Lemma 5.7 (Existence and uniqueness of the solution)**.** *There exists one and only one solution* (ρ,u) *of System* (5.26)-(5.27) *over the interval* $[\sigma_\ell,\sigma_p]$*.*

In addition, Lemma 5.6 allows to characterize σ_ℓ. Indeed, since $u \ge u_1 > 0$ and $\rho \ge \rho_1 > 0$ over $[\sigma_\ell,\sigma_p]$, by definition of σ_ℓ, either $\sigma_\ell = 0$ or $u(\sigma_\ell) = \sigma_\ell$. Since $u \ge u_1 > 0$ and $u(x) \ge x$ over $[\sigma_\ell,\sigma_p]$, the first option cannot hold, and we get

$$u(\sigma_\ell) = \sigma_\ell. \tag{5.28}$$

To complete the construction of a solution, it now remains to show the existence of a real number σ_r, *i.e.* the fact that there exists $x \in (\sigma_\ell,\sigma_p)$ such that $W(x) = \big(\rho(x),u(x),p(x)\big)$ satisfies the condition $\mathscr{F}_r(x) = 0$, where $\mathscr{F}_r$ is defined by:

$$\mathscr{F}_r(x) := \frac{1}{2}u(x)^2 + \frac{1}{\gamma_b - 1}\,x\,u(x) + \Big(\frac{\gamma_u}{\gamma_u - 1} - \frac{\gamma_b}{\gamma_b - 1}\frac{x}{x-u(x)}\Big)\frac{p(x)}{\rho(x)} + Q. \tag{5.29}$$

The existence of σ_r is stated in the following lemma.

Lemma 5.8 (Existence of σ_r)**.** *The function $\mathscr{F}_r$ is defined and continuously differentiable on the interal $(\sigma_\ell, \sigma_p]$, and $\lim_{x\to\sigma_\ell^+} \mathscr{F}_r(x) = -\infty$. In addition, $\mathscr{F}_r(\sigma_p) > 0$ when $\gamma_u = \gamma_b$ or when the reaction heat Q is large enough, according to Equation (5.31) below. Consequently, under one of these conditions, the set $\mathscr{S}_r = \{x \in (\sigma_\ell, \sigma_p)$ such that $\mathscr{F}_r(x) = 0\}$ is a non-empty closed subset of (σ_ℓ, σ_p) which admits a maximal element σ_r.*

Proof. When x tends to σ_ℓ, we have seen that $u(\sigma_\ell)$ tends to σ_ℓ and thus $\mathscr{F}_r$ tends to $-\infty$. When $x = \sigma_p$, if $\gamma_u = \gamma_b$, we observe that:

$$\mathscr{F}_r(\sigma_p) = Q > 0.$$

This relation is a consequence of the fact that, with $Q = 0$, $\mathscr{F}_r(\sigma) = 0$ is the relation satisfied by one adjacent state of a shock travelling at the speed σ when the velocity in the other adjacent state is zero, which is precisely the case of the state W_1 with $\sigma = \sigma_p$. It may also be checked by injecting the relations (5.13) in the definition of $\mathscr{F}_r$. When $\gamma_u \neq \gamma_b$, we thus get:

$$\mathscr{F}_r(\sigma_p) = \Big[\frac{1}{\gamma_b - 1} - \frac{1}{\gamma_u - 1}\Big]\sigma_p\, u_1 - \Big[\frac{\gamma_b}{\gamma_b - 1} - \frac{\gamma_u}{\gamma_u - 1}\Big]\,\frac{\sigma_p}{\sigma_p - u_1}\,\frac{p_1}{\rho_1} + Q. \tag{5.30}$$

Let us recast Relations (5.13) as

$$\rho_1 = \frac{1}{\alpha}\,\rho_0, \quad u_1 = (1-\alpha)\,\sigma_p, \quad p_1 = p_0 + (1-\alpha)\,\rho_0\,\sigma_p^2, \quad \text{with } \alpha = \frac{\gamma_u - 1 + \dfrac{2}{M_p^2}}{\gamma_u + 1}.$$

Using these relations in (5.30), we get:

$$\mathscr{F}_r(\sigma_p) = \Big[\frac{1-(1-\alpha)\,\gamma_b}{\gamma_b - 1} - \frac{1-(1-\alpha)\,\gamma_u}{\gamma_u - 1}\Big](1-\alpha)\,c_0^2\,M_p^2 - \Big[\frac{\gamma_b}{\gamma_b - 1} - \frac{\gamma_u}{\gamma_u - 1}\Big]\,\frac{p_0}{\rho_0} + Q. \tag{5.31}$$

Thanks to the Lax entropy conditions, $M_p > 1$, and, for $M_p \in (1, +\infty)$, the function $\alpha(M_p)$ decreases from 1 to $(\gamma_u - 1)/(\gamma_u + 1)$. Depending on the values of γ_b and γ_u, the quantity $1 - (1-\alpha)\,\gamma_b$ may become negative when $M_p \to +\infty$ for admissible values of α, and thus the first term may tends to $-\infty$; however, for a given M_p, this term is finite. The second term may be negative (still according to the values of γ_b and γ_u) but does not depend on M_p. Hence, for any given M_p, for Q large enough, the condition $F_r(\sigma_p) > 0$ is satisfied. □

In addition, when $\gamma_u = \gamma_b$, we are able to prove that $\mathscr{F}_r(\sigma_p)$ is an increasing function over (σ_ℓ, σ_p), and therefore the set $\mathscr{S}_r$ contains a single point; this result is stated in the following Lemma.

Lemma 5.9 (Variations of $\mathscr{F}_r$ when $\gamma_b = \gamma_u$)**.** *The function $\mathscr{F}_r$ defined by (5.32) is increasing over (σ_ℓ, σ_p) if the heat capacity ratios γ_b and γ_u are equal.*

Proof. Let us denote by γ the common heat capacity ratio, and recall the expression of $\mathscr{F}_r$:

$$\mathscr{F}_r(x) = \frac{1}{2}u(x)^2 + \frac{1}{\gamma-1}x\,u(x) - \frac{\gamma}{\gamma-1}\frac{u(x)}{x-u(x)}\frac{p(x)}{\rho(x)} + Q. \tag{5.32}$$

Using $\gamma p/\rho = c^2$, we have

$$\mathscr{F}_r'(x) = u(x)u'(x) + \frac{x}{\gamma-1}u'(x) + \frac{1}{\gamma-1}u(x) - \frac{1}{\gamma-1}\frac{xu'(x)-u(x)}{(x-u(x))^2}c^2(x) - \frac{1}{\gamma-1}\frac{u(x)}{x-u(x)}(c^2)'(x),$$

with

$$(c^2)' = s_1\gamma(\rho^{\gamma-1})' = s_1\gamma(\gamma-1)\,\rho^{\gamma-2}\rho' = (\gamma-1)\,c^2\,\frac{\rho'}{\rho}.$$

So we get that $\mathscr{F}_r'(x) = T_1(x) + T_2(x) + T_3(x)$ with

$$\begin{aligned}
T_1(x) &= \frac{1}{\gamma-1}\Big(1 + \frac{c^2(x)}{(x-u(x))^2}\Big)u(x),\\
T_2(x) &= \Big(u(x) + \frac{x}{\gamma-1} - \frac{1}{\gamma-1}\frac{xc^2(x)}{(x-u(x))^2}\Big)u'(x),\\
T_3(x) &= -\frac{u(x)\,c^2(x)}{(x-u(x))\,\rho}\rho'.
\end{aligned}$$

Since $u > 0$ by Lemma 5.6, $T_1 > 0$. In addition, in the proof of the same lemma 5.6, we showed that $u(x) + c(x) - x > 0$, so that $c(x) > x - u(x)$ and

$$u(x) + \frac{x}{\gamma-1}\Big(1 - \frac{c^2(x)}{(x-u(x))^2}\Big) \le u(x).$$

Since $u' \le 0$, this implies that $T_2 \ge uu'$. Replacing ρ' and u' by their expressions given in (5.26), we obtain that $uu' + T_3 = 0$, which concludes the proof. □

Finally, note that $\sigma_r > \sigma_\ell$; since u is a decreasing function, this yields that $\sigma_r - u(\sigma_r) > 0$. As mentioned in (5.2), this was expected, from a physical point of view, since this quantity is nothing else that the flame velocity u_f.

5.3.2 Numerical approximation of the solution in the intermediate zone

The problem tackled in this section is twofold: first, we need to solve numerically the system of ODEs (5.26)-(5.27), and second, to determine the speed of the reactive

shock σ_r. To this purpose, we solve (5.26)-(5.27) by an explicit Euler scheme, starting at $N \in \mathbb{N}$ and $x^N = \sigma_p$ and, for indices n decreasing from N, performing steps of $-\delta x$, with $\delta x = \sigma_p / N$; at each new step n associated to $x^n = n\delta x$, we obtain W^n and we evaluate the function $\mathscr{F}_r$, until we obtain $\mathscr{F}_r(x^n) \leq 0$. Here, the algorithm stops and we know that the computed approximation σ_r^{app} of σ_r satisfies $x^n < \sigma_r^{\text{app}} < x^{n+1}$; for δx small enough, x^{n+1} may thus be considered as a reasonable approximation of σ_r; this is indeed the way it is computed in the numerical experiments described below.

The scheme thus reads:

$$
\begin{aligned}
&\text{for } n = N,\ u^N = u_1,\ \rho^N = \rho_1,\\
&\\
&\text{for } n = N-1 \text{ to } 0 \text{ and while } \mathscr{F}_r(x^{n+1}) > 0,\\
&\qquad (c^{n+1})^2 = \gamma\, s_1\, (\rho^{n+1})^{\gamma_u - 1},\\
&\qquad \rho^n = \rho^{n+1} + \delta x\, \frac{2\, u^{n+1}\, (u^{n+1} - x^{n+1})}{x^{n+1}\left((u^{n+1} - x^{n+1})^2 - (c^{n+1})^2\right)}\, \rho^{n+1},\\
&\qquad u^n = u^{n+1} - \delta x\, \frac{2\, (c^{n+1})^2}{x^{n+1}\left((u^{n+1} - x^{n+1})^2 - (c^{n+1})^2\right)}\, u^{n+1}.
\end{aligned}
\tag{5.33}
$$

Then, for any valid value of $n \leq N$, the pressure is given by

$$p^n = s_1\, (\rho^n)^{\gamma_u}.$$

Since the algorithm stops as soon as $\mathscr{F}_r(x^{n+1})$ becomes negative, from the expression of this latter function, we have $u^n < x^n$ in all the performed steps n. We thus have $u^n > 0$, $\rho^n > 0$, $u^n \geq u^{n+1}$ and $\rho^n \geq \rho^{n+1}$ at all steps.

5.4 Solution for a given flame speed

The construction performed in the previous section shows that, to any precursor shock velocity σ_p greater than the speed of sound c_0 in the outer zone of the fresh atmosphere, we are able to associate a positive flame velocity u_f given by $u_f = \sigma_r - u_2$. In addition, even if we have no proof, physical arguments suggest that u_f is an increasing function of σ_p (or equivalently of the Mach number $M = \sigma_p / c_0$, considering a family of problems with the same initial atmosphere and thus c_0 as a fixed parameter); this behaviour is confirmed by numerical experiments (see Section 5.5). Computing the flow for a given u_f, which is in fact usually the engineering problem to be tackled, amounts to invert the function $u_f = \widetilde{\mathscr{G}}(M)$, and this equation for M thus should have one and only one solution, at least for reasonable values of u_f. To compute this solution, we define $\mathscr{G}$ by

$$\mathscr{G}(M) := \widetilde{\mathscr{G}}(M) - u_f \tag{5.34}$$

and search for M such that $\mathcal{G}(M) = 0$ with the following iterative algorithm depending on the parameters M_0, δ and ϵ:

initialization:	let M_0 be given, and compute $\mathcal{G}(M_0)$,
	let $M_1 = M_0 + \delta$, and compute $\mathcal{G}(M_1)$
current iteration:	For $k \geq 2$, let $M_k = M_{k-1} - \dfrac{M_{k-1} - M_{k-2}}{\mathcal{G}(M_{k-1}) - \mathcal{G}(M_{k-2})} \mathcal{G}(M_{k-1})$, and compute $\mathcal{G}(M_k)$.
stopping criteria:	stop when $\mathcal{G}(M_k) \leq \epsilon$.

This algorithm is used in the following section with $M_0 = 1.0001$, $\delta = 0.001$ and $\epsilon = 10^{-5}$. Convergence is obtained for all cases, provided that the number of cells used in the numerical computation of the solution in the intermediate zone is large enough; otherwise, the error on σ_r is too large and the prescribed tolerance threshold for the value of $\mathcal{G}$ cannot be reached.

5.5 Application to hydrogen deflagrations

We now apply the developed procedure for a flame propagating in a stoichiometric mixture of hydrogen and air. We consider a unique total and irreversible chemical reaction, which reads:

$$2\,H_2 + O_2 \longrightarrow 2\,H_2O$$

Supposing that air is composed of 1/5 of oxygen and 4/5 nitrogen (molar or volume proportions), the molar fractions of hydrogen, oxygen and nitrogen in the considered stoichiometric mixture are thus equal to 2/7, 1/7 and 4/7 respectively. The mass fractions of these constituents are thus easily deduced from these values:

$$y_{H_2} = \frac{2\,W_{H_2}}{W_t}, \quad y_{O_2} = \frac{W_{O_2}}{W_t}, \quad y_{N_2} = \frac{4\,W_{N_2}}{W_t}, \quad W_t = 2W_{H_2} + W_{O_2} + 4\,W_{N_2}$$

where $W_{H_2} = 0.002$ Kg, $W_{O_2} = 0.032$ Kg and $W_{N_2} = 0.028$ Kg stand for the molar mass of the hydrogen, oxygen and nitrogen molecules respectively. Since H_2 and O_2 are pure substances (and thus their formation enthalpy is equal to zero), the chemical reaction heat reads:

$$Q = y_{H_2O}\,\Delta H_0^f = (y_{H_2} + y_{O_2})\,\Delta H_0^f,$$

where $\Delta H_0^f = 1.3255\,10^7$ J/Kg stands for the formation enthalpy of steam. The initial pressure is $p = 10^5$ Pa, the initial temperature is $T = 283\,^\circ$K, the initial density is given by the Boyle-Mariotte law and the heat capacity ratio is $\gamma_u = \gamma_b = 1.4$.

We plot on Figures 5.2 and 5.3 the density, velocity and pressure profiles obtained

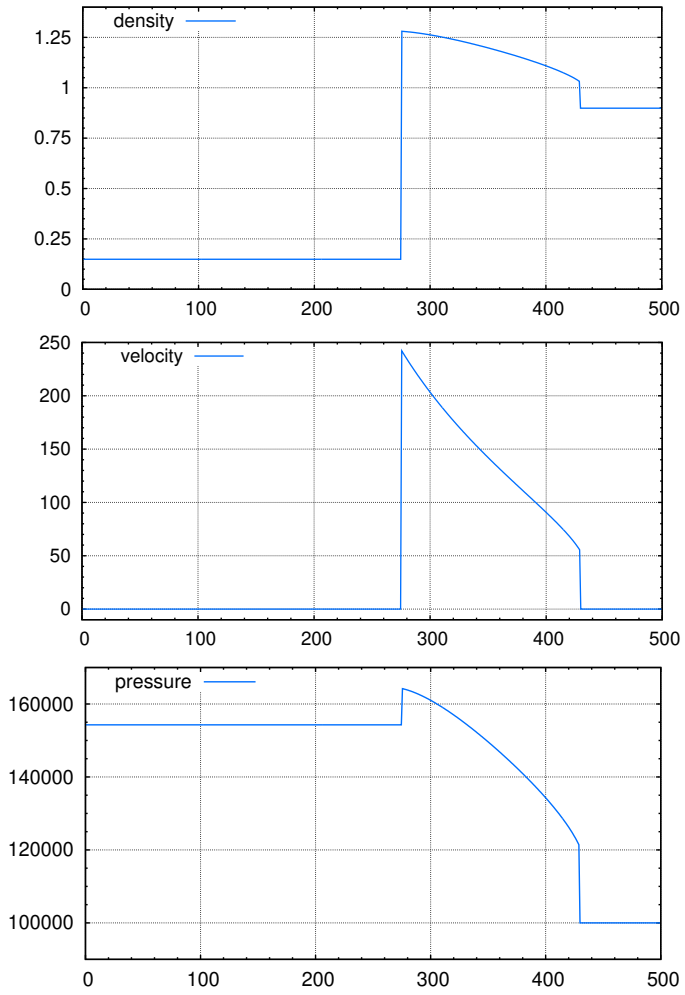


Figure 5.2 – Density ($\mathrm{kg\,m^{-3}}$), velocity ($\mathrm{m\,s^{-1}}$) and pressure (Pa) profiles obtained for a flame velocity of $u_f = 32\,\mathrm{m/s}$.

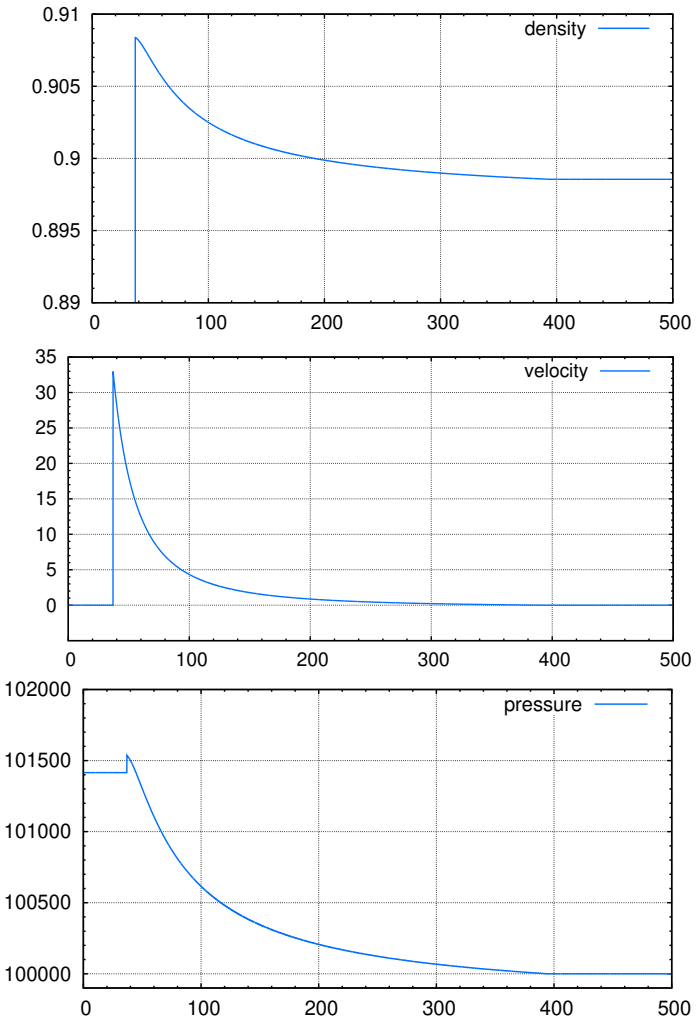


Figure 5.3 – Density ($\mathrm{kg\,m^{-3}}$), velocity ($\mathrm{m\,s^{-1}}$) and pressure (Pa) profiles obtained for a flame velocity of $u_f = 4\,\mathrm{m/s}$.

for $u_f = 32\,\mathrm{m/s}$ and $u_f = 4\,\mathrm{m/s}$ respectively. The solution in the intermediate zone is obtained with a regular mesh, splitting the interval between 5 m and the position of the precursor shock (which is unknown up to the last solution step of the algorithm) in 5000 equal subintervals. Then we show on Figures 5.4-5.6 the evolution of the states W_1, W_2 and W_b as a function of the flame velocity. The temperature in the burnt state, not shown here, is close to $T = 3050\,^\circ\mathrm{K}$ for all the values of the flame velocity u_f. We observe that the precursor shock is of very weak amplitude for low values of the flame velocity; in fact, it becomes visible only when u_f reaches 20 m/s. For $u_f = 4$, the computed velocity at state W_1 is lower than $10^{-6}\,\mathrm{m/s}$, while it reaches values greater than 30 m/s at State W_2. Since the ordinary differential equation governing the velocity in the intermediate zone (5.26b) is of the form

$$u' = f(\rho, u)\, u,$$

one may anticipate such a low value as initial (right) condition to lead to severe accuracy problems. In this respect, the computed value which seems to be the most affected is the velocity at state W_2: the convergence value seems to be close to 33.00 m/s, we obtain $u_2 \simeq 34.5\,\mathrm{m/s}$ with $n = 5\,10^3$ cells and $u_2 \in (32.95\,\mathrm{m/s},\ 33\,\mathrm{m/s})$ for $n = 8\,10^4$, $n = 16\,10^4$, $n = 32\,10^4$ and $n = 64\,10^4$. As expected, convergence is easier when the precursor shock has a significant amplitude: $u_2 \simeq 243.0\,\mathrm{m/s}$ for $n = 5000$, for a convergence value in the range of 243.8 m/s.

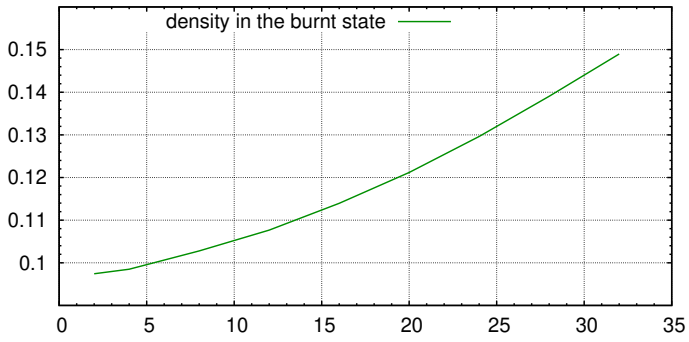


Figure 5.4 – Density in the burnt zone as a function of the flame velocity.

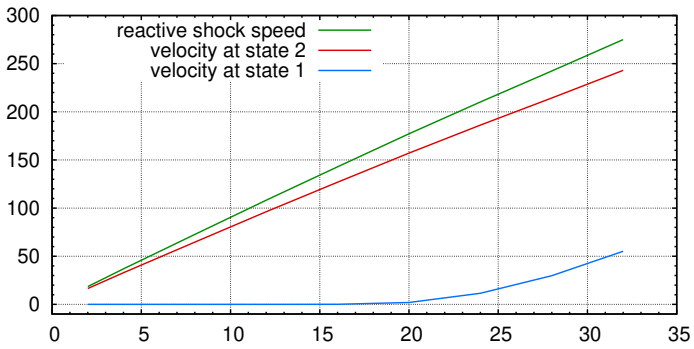


Figure 5.5 – Velocity at state 1 and state 2 and speed of the precursor shock as a function of the flame velocity.

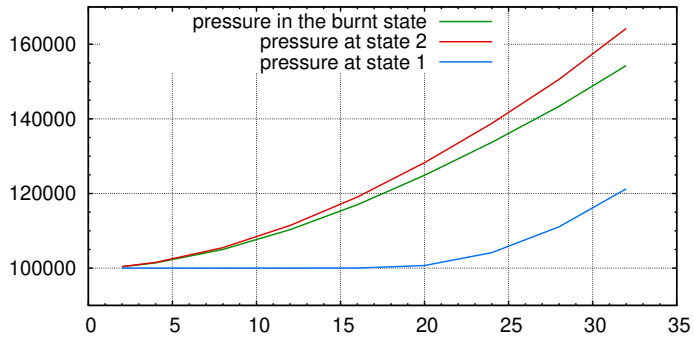


Figure 5.6 – Pressure at state 1, at state 2 and in the burnt zone as a function of the flame velocity.

Conclusion

The aim of this Phd thesis was to contribute to the development of efficient numerical scheme for the simulation of incompressible and compressible flows using staggered grids. Several time integration methods are studied in this thesis for different fluid flows which are enforced by numerical analysis and simulations.

The first chapter presented an extension of the Lax-Wendroff consistency Theorem in the case of a staggered discretization for the $2D$ shallow water equations with regular topography. Indeed, the consistency of the MAC schemes are shown for both the first order segregated and second order Heun schemes with a MUSCL-like interpolation for the numerical fluxes (mass and momentum). Furthermore the consistency of the entropy inequality is only proven for the first order scheme thanks to a BV-norm bounded assumption. Numerical investigations show the better accuracy of the Heun scheme with respect to the first order scheme. Indeed, the order of convergence of this latter is close to 1 for both the height and the velocity while the Heun scheme yields an order 2 for the height and 1.5 for the velocity. It is shown numerically that the Heun scheme has a smoothing effect which damps numerical oscillations occurring for a shock wave; these ocillations are furthermore reduced by a regularization term which is introduced in the discrete momentum equation to stabilize the scheme.

In chapter 2, formal upwind and stabilized centered schemes based on a staggered discretization are studied for the non-linear shallow water equations with a Coriolis source term. A theoretical study shows that the schemes based on the RT finite elements enjoy important discrete properties such as the decay of the semi-discrete mechanical energy and the preservation of the (linear) geostrophic equilibrium. However the upwind and stabilized centered RT schemes produce results that are not as good as those obtained by the MAC schemes.The MAC schemes are largely stable and accurate compared to RT and HLLC schemes despite their discrete properties that do not meet the requirements set that are the preservation of the geostrophic equilibrium for the linear schemes and the dissipation of the semi-discrete mechanical energy for the non-linear one. In conclusion, this small comparative study allows us to argue that: a staggered scheme working on the RT finite elements is less efficient on rectangular grids.
In this context, a future work is planed in the purpose to develop staggered schemes for the non-linear SWC equations. Conduct an advanced stabilization study for the MAC scheme while trying to decrease the mechanical energy computed from the semi-discrete scheme.

We showed in Chapter 3 the efficiency of the decoupled staggered scheme for the computation of the sediment transport process. We evidence numerically the relevant role played by a correction of the sediment transport flux when classical formulae fail. Furthermore we show the compatibility of the stabilized friction term which allows the modified shallow-water-Exner equations to dissipate the mechanical energy and which produces satisfactory results compared to the formal Saint-Venant-Exner model proposed by Fernàndez-Nieto, Morales De Luna, Narbona-Reina, et al. 2017. The numerical results validate on one hand the decoupled staggered approach and on the other hand the improvement of the sediment flux and shear stress corrections to model shallow water flow and sediment transport.
An interesting issue is then to extend the present work in the framework of two dimensional water flow and sediment process as performed by S. Li and Duffy 2011 where the authors implement a fully coupled approach based on a Roe approximate Riemann solver.

Chapter 4 is devoted to a pressure correction time discretization scheme for the numerical approximation of the reactive Euler equations. This scheme works on general staggered grids and solves the so-called sensible enthalpy instead of the total energy for the Euler equations. The approximation of the convection terms is improved thanks to a MUSCL-like interpolation and an anti-diffusive scheme. The consistency and the robustness of the scheme are shown as well as the efficiency of the algorithm for the simulation of a plane deflagration.
An important perspective is to validate this staggered pressure correction scheme in the case of three dimensional flow using the reference solution constructed in Chapter 5. This track is undertaken thanks to the CALIF^{3}S free software developed by the IRSN.

Bibliographie

[Ans+11] G. ANSANAY-ALEX, F. BABIK, J.-C. LATCHÉ et al. « An L^2-stable approximation of the Navier-Stokes convection operator for low-order non-conforming finite elements ». In : *International Journal for Numerical Methods in Fluids* 66 (2011), p. 555-580 (cf. p. 26, 84).

[AL81] A. ARAKAWA et V.R. LAMB. « A Potential Enstrophy and Energy Conserving Scheme for the Shallow Water Equations ». In : *Monthly Weather Review* 109 (1981), p. 18-36 (cf. p. 14, 20, 22).

[AM72] K. ASHIDA et M. MICHIUE. « Study on hydraulic resistance and bedload transport rate in alluvial streams ». In : *JSCE Tokyo* 206 (1972), p. 58-69 (cf. p. 122, 129).

[Aud18] E. AUDUSSE. « Autour du système de Saint-Venant : Méthodes numériques pour le transport sédimentaire, les fluides en rotation et les équations primitives ». Habilitation à diriger des recherches. Université Paris 13 Villeraneuse, nov. 2018. URL : https://hal.archives-ouvertes.fr/tel-02005164 (cf. p. 20, 123, 131).

[Aud+04] E. AUDUSSE, F. BOUCHUT, M.-O. BRISTEAU et al. « A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows ». In : *Siam J. Sci. Comput* 25 (2004), p. 2050-2065 (cf. p. 75).

[Aud+17] E. AUDUSSE, S. DELLACHERIE, M.H. DO et al. « Godunov type schemes for the linear wave equation with Coriolis source term ». In : *ESAIM Proc. Surv.* 58 (2017), p. 1-26 (cf. p. 74).

[Aud+18] E. AUDUSSE, M.H. DO, P. OMNES et al. « Analysis of modified Godunov type schemes for the two-dimensional linear wave equation with Coriolis source term on Cartesian meshes ». In : *Journal of Computational Physics* 373 (2018), p. 91-129 (cf. p. 74-76, 90, 91, 97, 100, 102, 104, 110).

[Aud+20] E. AUDUSSE, V. DUBOS, A. DURAN et al. « Numerical approximation of the shallow water equation with Coriolis source term ». In : *ESAIM Proc. Surv., submitted* 373 (2020), p. 91-129 (cf. p. 17, 90).

[AKO09] E. AUDUSSE, R. KLEIN et A. OWINOH. « Conservative discretization of Coriolis force in a finite volume framework ». In : *Journal of Computational Physics* 228 (2009), p. 2934-2950 (cf. p. 74).

[BS10] A. BECCANTINI et E. STUDER. « The reactive Riemann problem for thermally perfect gases at all combustion regimes ». In : *International Journal for Numerical Methods in Fluids* 64 (2010), p. 269-313 (cf. p. 148, 150, 171, 172, 186).

[BMQ13] A. BELJADID, A. MOHAMMADIAN et H.M. QIBLAWEY. « An unstructured finite volume method for large-scale shallow flows using the fourth-order Adams scheme ». In : *Comput. Fluids* 88 (2013), p. 579-589 (cf. p. 74).

[BL19] M. BEN-ARTZI et J. LI. *Consistency and Convergence of Finite Volume Approximations to Nonlinear Hyperbolic Balance Laws*. 2019. arXiv : 1902.09047 [math.NA]. URL : https://arxiv.org/abs/1902.09047 (cf. p. 65).

[BBT15] C. BERTHON, B. BOUTIN et R. TURPAULT. « Shock profiles for the Shallow-water Exner models ». In : *Advances in Applied Mechanics* 7 (2015), p. 267-294. URL : https://hal.archives-ouvertes.fr/hal-01202866 (cf. p. 131).

[Ber+19] C. BERTHON, A. DURAN, F. FOUCHER et al. « Improvement of the hydrostatic reconstruction scheme to get fully discrete entropy inequalities ». In : *Journal of scientific computing, Springer verlag* 11 (2019), p. 11-11 (cf. p. 75).

[Boi19] L. BOITTIN. « Modeling, analysis and simulation of two geophysical flows. Sediment transport and variable density flows ». Phd thesis. Sorbonne Université, avr. 2019. URL : https://tel.archives-ouvertes.fr/tel-02935869 (cf. p. 123, 124, 130).

[BR05] L. BONAVENTURA et T.D. RINGLER. « Analysis of Discrete Shallow-Water Models on Geodesic Delaunay Grids with C-Type Staggering ». In : *Monthly Weather Review* 133.8 (2005), p. 2351-2373 (cf. p. 20, 74).

[Bou04] F. BOUCHUT. *Nonlinear Stability of finite volume methods for hyperbolic conservation laws*. Birkhauser, 2004 (cf. p. 20).

[BLZ14] F. BOUCHUT, J. LESOMMER et V. ZEITLIN. *Frontal geostrophic adjustment and nonlinear wave phenomena in one-dimensional rotating shallow water. Part 2. High-resolution numerical simulations*. T. 514. 2014, p. 35-63 (cf. p. 74).

[CAL] CALIF^{3}S. *A software components library for the computation of fluid flows*. https://gforge.irsn.fr/gf/project/califs (cf. p. 54, 148, 185).

[CD77] P. CAMBRAY et B. DESHAIES. « Ecoulement engendré par un piston sphérique : solution analytique approchée ». In : *Acta Astronautica* 5 (1977), p. 611-617 (cf. p. 186).

[CFF08] M. J. CASTRO DÌAZ, E.D. FERNÀNDEZ-NIETO et A.M. FERREIRO. « Sediment transport models in Shallow Water equations and numerical approach by high order finite volume methods ». In : *M2AN* 1500012 (2008), p. 133-142 (cf. p. 123, 131, 143).

[CMP17] M. J. CASTRO, T. MORALES DE LUNA et C. PARÉS. « Well-balanced schemes and path-conservative numerical methods ». In : *Handbook of numerical methods for hyperbolic problems*. T. 18. Handb. Numer. Anal. Elsevier/North-Holland, Amsterdam, 2017, p. 131-175 (cf. p. 20).

[Cho68] A.J. CHORIN. « Numerical solution of the Navier-Stokes equations ». In : *Mathematics of Computation* 22 (1968), p. 745-762 (cf. p. 13).

[Cia91] P.G. CIARLET. « Basic Error Estimates for Elliptic Problems ». In : *Handbook of Numerical Analysis, Volume II*. Sous la dir. de P. CIARLET et J.L. LIONS. North Holland, 1991, p. 17-351 (cf. p. 156).

[CDV17] F. COUDERC, A. DURAN et J.-P. VILA. « An explicit asymptotic preserving low Froude scheme for the multilayer shallow water model with density stratification ». In : *Journal of Computational Physics* 343 (2017), p. 235-270 (cf. p. 75, 90, 91, 97).

[CR73] M. CROUZEIX et P.A. RAVIART. « Conforming and nonconforming finite element methods for solving the stationary Stokes equations ». In : *RAIRO Série Rouge* 7 (1973), p. 33-75 (cf. p. 14, 155, 157).

[DL81] B. DESHAIES et J.C. LEYER. « Flow induced by unconfined spherical accelerating flames ». In : *Combustion and Flame* 40 (1981), p. 141-153 (cf. p. 186).

[DL02] B. DESPRÉS et F. LAGOUTIÈRE. « Contact discontinuity capturing scheme for linear advection and compressible gas dynamics ». In : *Journal of Scientific Computing* 16 (2002), p. 479-524 (cf. p. 149, 172, 180, 181).

[DG14] D. DOYEN et H.P. GUNAWAN. « An explicit staggered finite volume scheme for the shallow water equations ». In : *Finite volumes for complex applications. VII. Methods and theoretical aspects*. T. 77. Springer Proc. Math. Stat. Springer, Cham, 2014, p. 227-235 (cf. p. 14, 20).

[DV19] A. DURAN et J.-P. VILA. « Energy-stable staggered schemes for the Shallow Water equations ». In : *Journal of Computational Physics* 343 (2019), p. 235-270 (cf. p. 75, 93, 95).

[DVB17] A. DURAN, J.-P. VILA et R. BARAILLE. « Semi-implicit staggered mesh scheme for the multi-layer shallow water system ». In : *C. R. Acad. Sci. Paris* 355 (2017), p. 1298-1306 (cf. p. 75).

[Ein42] H.A. EINSTEIN. « Formulas for the transportation of bed load ». In : *Transactions of the American Society of Civil Engineers* 107.1 (1942), p. 561-575. URL : https://cedb.asce.org/CEDBsearch/record.jsp?dockey=0288730 (cf. p. 123).

[Ell07] V. ELLING. « A Lax-Wendroff type theorem for unstructured quasi-uniform grids ». In : *Mathematics of Computation* 76 (2007), p. 251-272 (cf. p. 65).

[EGH00] R. EYMARD, T. GALLOUËT et R. HERBIN. « Finite Volume Methods ». In : *Handbook of Numerical Analysis, Volume VII*. Sous la dir. de P. CIARLET et J.L. LIONS. North Holland, 2000, p. 713-1020. URL : https://hal.archives-ouvertes.fr/ (cf. p. 65).

[FV76] R. FERNANDEZ LUQUE et R. VAN BEEK. « Erosion And Transport Of Bed-Load Sediment ». In : *Journal of Hydraulic Research* 14.2 (1976), p. 127-144. DOI : 10.1080/00221687609499677. eprint : https://doi.org/10.1080/00221687609499677. URL : https://doi.org/10.1080/00221687609499677 (cf. p. 122).

[Fer+14] E.D. FERNÀNDEZ-NIETO, C. LUCAS, T. MORALES DE LUNA et al. « On the influence of the thickness of the sediment moving layer in the definition of the bedload transport formula in Exner systems ». In : *Computers and Fluids* 91 (mar. 2014), p. 87-106. DOI : 10.1016/j.compfluid.2013.11.031. URL : https://hal.archives-ouvertes.fr/hal-00821659 (cf. p. 123, 127, 131, 137).

[Fer+17] E.D. FERNÀNDEZ-NIETO, T. MORALES DE LUNA, G. NARBONA-REINA et al. « Formal deduction of the Saint-Venant-Exner model including arbitrarily sloping sediment beds and associated energy ». In : *ESAIM : M2AN* 51.1 (2017), p. 115-145. DOI : 10.1051/m2an/2016018. URL : https://doi.org/10.1051/m2an/2016018 (cf. p. 16, 123, 124, 130, 131, 138, 139, 142, 143, 206).

[FKO07] A.C. FOWLER, N. KOPTEVA et C. OAKLEY. « The formation of river channels ». In : *SIAM J. Appl. Math* 67 (2007), p. 1016-1040 (cf. p. 130).

[GHL19] T. GALLOUËT, R. HERBIN et J.-C. LATCHÉ. « On the weak consistency of finite volume schemes for conservation laws on general meshes ». In : *under revision* (2019). URL : https://hal.archives-ouvertes.fr/hal-02055794 (cf. p. 22, 32, 33, 48, 49, 65, 71).

[Gal+18] T. GALLOUËT, R. HERBIN, J.-C. LATCHÉ et K. MALLEM. « Convergence of the marker-and-cell scheme for the incompressible Navier-Stokes equations on non-uniform grids ». In : *Foundations of Computational Mathematics* 18 (2018), p. 249-289 (cf. p. 22, 25).

[Gal+20a] T. GALLOUËT, R. HERBIN, J.-C. LATCHÉ et al. « A second order consistent MAC scheme for the shallow water equations on non uniform grids ». In : *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*. Springer, juin 2020, p. 123-131 (cf. p. 17, 21, 22).

[Gal+20b] T. GALLOUËT, R. HERBIN, J.-C. LATCHÉ et al. « Weak consistency of non linear convection operators on staggered meshes. Application to a quasi-second order staggered scheme for the two-dimensional shallow water equations ». In : *submitted* (2020). URL : https://hal.archives-ouvertes.fr/hal-02940981 (cf. p. 17).

[Gas+11] L. GASTALDO, R. HERBIN, W. KHERIJI et al. « Staggered discretizations, pressure correction schemes and all speed barotropic flows ». In : *Finite Volumes for Complex Applications VI - Problems & Perspectives - Prague, Czech Republic*. T. 2. Springer, 2011, p. 39-56 (cf. p. 162).

[GHL10] L. GASTALDO, R. HERBIN et J.-C. LATCHÉ. « An unconditionally stable finite element-finite volume pressure correction scheme for the drift-flux model ». en. In : *ESAIM : Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique* 44.2 (2010), p. 251-287. DOI : 10.1051/m2an/2010002. URL : http://www.numdam.org/item/M2AN_2010__44_2_251_0 (cf. p. 13).

[GR96] E. GODLEWSKI et P.-A. RAVIART. « Numerical approximation of hyperbolic systems of conservation laws ». In : *Springer*. Applied Mathematical Sciences, New York, 1996 (cf. p. 43, 65, 179, 193).

[Gra17] D. GRAPSAS. « Staggered fractional step numerical schemes for models for reactive flows ». Phd thesis. Aix-Marseille Université, 2017 (cf. p. 13).

[Gra+16] D. GRAPSAS, R. HERBIN, W. KHERIJI et al. « An unconditionally stable Finite Element-Finite Volume pressure correction scheme for the compressible Navier-Stokes equations ». In : *SMAI Journal of Computational Mathematics* 2 (2016), p. 51-97 (cf. p. 13, 149, 162, 166).

[Gra+] D. GRAPSAS, R. HERBIN, J.-C. LATCHÉ et al. « Modelling of spherical deflagration at constant speed ». In : *under revision* (). URL : https://hal.archives-ouvertes.fr/hal-02967980 (cf. p. 17).

[Gra+20] D. GRAPSAS, R. HERBIN, J.-C. LATCHÉ et al. « A staggered pressure correction numerical scheme to compute a travelling reactive interface in a partially premixed mixture ». In : (juin 2020), p. 123-131. URL : https://hal.archives-ouvertes.fr/hal-02967051 (cf. p. 17).

[Gra81] A. GRASS. *Sediment transport by waves and currents*. London : University College, London, Dept. of Civil Engineering, 1981 (cf. p. 122, 128).

[GBL76] C.M. GUIRANO, G.G. BASH et J.H. LEE. « Pressure waves generated by spherical flames ». In : *Combustion and Flame* 27 (1976), p. 341-351 (cf. p. 186).

[Gun15] H.P. GUNAWAN. « Numerical simulation of shallow water equations and related models ». Thèse de doct. Université Paris-Est et Institut Teknologi Bandung, 2015 (cf. p. 13-15, 20, 74, 75, 83, 124).

[GEP15] H.P. GUNAWAN, R. EYMARD et S.R. PUDJAPRASETYA. « Staggered scheme for the Exner - shallow water equations ». In : *Computational Geosciences* 19 (2015), p. 1197-1206 (cf. p. 13, 15, 124, 131, 132).

[GL15] H.P. GUNAWAN et X. LHÉBRARD. « Hydrostatic relaxation scheme for the 1D shallow water - Exner equations in bedload transport ». In : *Computers Fluids* 121 (2015), p. 44-50 (cf. p. 131).

[HA71] F.H. HARLOW et A.A. AMSDEN. « A Numerical Fluid Dynamics Calculation Method for All Flow Speeds ». In : *Journal of Computational Physics* 8 (1971), p. 197-213 (cf. p. 22, 157).

[HW65] F.H. HARLOW et J.E. WELSH. « Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface ». In : *Physics of Fluids* 8, 2 (1965), p. 2182-2189 (cf. p. 14, 22, 157).

[HKL14] R. HERBIN, W. KHERIJI et J.-C. LATCHÉ. « On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations ». In : *Mathematical Modelling and Numerical Analysis* 48 (2014), p. 1807-1857 (cf. p. 13, 162).

[HL10] R. HERBIN et J.-C. LATCHÉ. « Kinetic energy control in the MAC discretization of the compressible Navier-Stokes equations ». In : *International Journal of Finites Volumes,* 7 (2010) (cf. p. 26, 162).

[Her+20] R. HERBIN, J.-C. LATCHÉ, S. MINJEAUD et al. « Conservativity and weak consistency of a class of staggered finite volume methods for the Euler equations ». In : *Mathematics of Computation, on line* (2020). arXiv : 1910.03998 [math.NA] (cf. p. 65).

[Her+19] R. HERBIN, J.-C. LATCHÉ, Y. NASSERI et al. « A decoupled staggered scheme for the shallow water equations ». In : *Monografías Matemáticas García de Galdeano* 52 (2019), p. 1-16 (cf. p. 17, 21, 22, 30, 80, 135).

[HLN13] R. HERBIN, J.-C. LATCHÉ et T.T. NGUYEN. « Explicit staggered schemes for the compressible Euler equations ». In : *ESAIM : Proceedings* 40 (2013), p. 83-102 (cf. p. 13, 26, 70).

[HLN18] R. HERBIN, J.-C. LATCHÉ et T.T. NGUYEN. « Consistent segregated staggered schemes with explicit steps for the isentropic and full Euler equations ». In : *ESAIM : Mathematical Modelling and Numerical Analysis* 52 (2018), p. 893-944 (cf. p. 13, 20, 29, 30, 58, 70, 87, 132, 134).

[HLS20] R. HERBIN, J.-C. LATCHÉ et K. SALEH. « Low Mach number limit of some staggered schemes for compressible barotropic flows ». In : *to appear in Math. Comp.* (2020) (cf. p. 13).

[Khe11] W. KHERIJI. « Méthodes de correction de pression pour les équations de Navier-Stokes compressibles ». Phd thesis. Université Aix-Marseille 1, 2011 (cf. p. 13).

[KRW96] D. KRONER, M. ROKYTA et M. WIERSE. « A Lax-Wendroff type theorem for upwind finite volume schemes in 2-D ». In : *East-West Journal of Numerical Mathematics* 4 (1996), p. 279-292 (cf. p. 65).

[KKO73] A.L. KUHL, M.M. KAMEL et A.K. OPPENHEIM. « Pressure waves generated by steady flames ». In : *Symposium (International) on Combustion* 14.1 (1973). Fourteenth Symposium (International) on Combustion, p. 1201-1215. ISSN : 0082-0784. DOI : https://doi.org/10.1016/S0082-0784(73)80108-0. URL : http://www.sciencedirect.com/science/article/pii/S0082078473801080 (cf. p. 185, 186).

[Lar91] B. LARROUTUROU. « How to Preserve the Mass Fractions Positivity when Computing Compressible Multi-Component Flows ». In : *Journal of Computational Physics* 95 (1991), p. 59-84 (cf. p. 155).

[LW60] P.D. LAX et B. WENDROFF. « Systems of conservation laws ». In : *Communications in Pure and Applied Mathematics* 13 (1960), p. 217-237 (cf. p. 65).

[Lev02] R.J. LEVEQUE. *Finite Volume Methods for Hyperbolic Problems*. Cambridge texts in applied mathematics. Cambridge University Press, 2002 (cf. p. 65).

[LD11] S. LI et C.J. DUFFY. « Fully coupled approach to modeling shallow water flow, sediment transport, and bed evolution in rivers ». In : *Water Ressouces Research* 47 (2011), p. 1-20. URL : https://doi.org/10.1029/2010WR009751 (cf. p. 206).

[MM48] E. MEYER-PETER et R. MULLER. « Formulas for bed-load transport ». In : *Proceedings of 2nd meeting of the International Association for Hydraulic Structures Research*. (Stockholm). International Association for Hydraulic Structures Research. Juin 1948, p. 39-64. URL : http://resolver.tudelft.nl/uuid:4fda9b61-be28-4703-ab06-43cdc2a21bd7 (cf. p. 122, 128).

[MKR02] V.A. MOUSSEAU, D.A. KNOLL et J.M. REISNER. « An Implicit Nonlinearly Consistent Method for the Two-Dimensional Shallow-Water Equations with Coriolis Force ». In : *American Meteorological Society* (2002) (cf. p. 74, 76).

[Nie92] P. NIELSON. « Coastal boundary layers and sediment transport ». In : *World Scientific Publishing, Singapore. Advanced Series on Ocean Engineering* 4 (1992) (cf. p. 122).

[PV16] M. PARISOT et J.-P. VILA. « Centered-potential regularization for the advection upstream splitting method ». In : *SIAM J. Numer. Anal* 52 (2016), p. 3083-3104 (cf. p. 75).

[Pet00] N. PETERS. *Turbulent Combustion*. Cambridge Monographs of Mechanics. Cambridge University Press, 2000 (cf. p. 148).

[Pia+13] L. PIAR, F. BABIK, R. HERBIN et al. « A formally second order cell centered scheme for convection-diffusion equations on general grids ». In : *International Journal for Numerical Methods in Fluids* 71 (2013), p. 873-890 (cf. p. 15, 24, 25, 27, 28, 43, 55, 149, 172, 177).

[PV05] T. POINSOT et D. VEYNANTE. *Theoretical and Numerical Combustion*. Editions R.T Edwards Inc., 2005 (cf. p. 148).

[RT92] R. RANNACHER et S. TUREK. « Simple Nonconforming Quadrilateral Stokes Element ». In : *Numerical Methods for Partial Differential Equations* 8 (1992), p. 97-111 (cf. p. 14, 155, 157).

[Rin+10] T.D. RINGLER, J. THUBURN, J.B. KLEMP et al. « A unified approach to energy conservation and potential vorticity dynamics for arbitrarily- structured C-grids ». In : *J. Comput. Phys.* 229 (2010), p. 3065-3090 (cf. p. 74).

[Sed45] L.I. SEDOV. « On certain unsteady motions of compressible fluid ». In : *Prikladnaya Matematika i Mekhanika* 9 (1945), p. 293-311 (cf. p. 186).

[SD03] G.S. STELLING et S.P.A. DUINMEIJER. « A staggered conservative scheme for every Froude number in rapidly varied shallow water flows ». In : *International Journal for Numerical Methods in Fluids* 43 (2003), p. 1329-1354 (cf. p. 20).

[Str+79] R.A. STREHLOW, R.T. LUCKRITZ, A.A. ADAMCZYK et al. « The blast wave generated by spherical flames ». In : *Combustion and Flame* 35 (1979), p. 297-310 (cf. p. 186).

[Tan92] W.-Y. TAN. *Shallow water hydrodynamics : Mathematical theory and numerical solution for a two-dimensional system of shallow-water equations*. Elsevier, 1992 (cf. p. 20).

[Tay46] G.I. TAYLOR. « The air wave surrounding an expanding sphere ». In : *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186 (1946), p. 273-292 (cf. p. 186).

[Tem69] R. TEMAM. « Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires II ». In : *Arch. Rat. Mech. Anal.* 33 (1969), p. 377-385 (cf. p. 13).

[The15] N. THERME. « Schémas numériques pour la simulation de l'explosion ». Phd thesis. Aix-Marseille Université, 2015 (cf. p. 13).

[Thu+09] J. THUBURN, T.D. RINGLER, W.C. SKAMAROCK et al. « Numerical representation of geostrophic modes on arbitrarily structured C-grids ». In : *J. Comput. Phys.* 228 (2009), p. 8321-8335 (cf. p. 74).

[Tor97] E.F. TORO. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer, 1997 (cf. p. 116).

[Van84] L.C. VAN RIJN. « Sediment transport (**I**) : bed load transport ». In : *J. Hydraul. Div. Proc. ASCE* 110 (1984), p. 1431-1456 (cf. p. 122).

[Xin17] Y. XING. « Numerical methods for the nonlinear shallow water equations ». In : *Handbook of numerical methods for hyperbolic problems*. T. 18. Handb. Numer. Anal. Elsevier/North-Holland, Amsterdam, 2017, p. 361-384 (cf. p. 20).

[Zim00] V.L. ZIMONT. « Gas premixed combustion at high turbulence. Turbulent flame closure combustion model ». In : *Experimental Thermal and Fluid Science* 21 (2000), p. 179-186 (cf. p. 148).