UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Victor MICHEL-DANSAC

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : Sciences et technologies de l'information et mathématiques

Discipline : Mathématiques et applications, section CNU 26
Unité de recherche : Laboratoire de Mathématiques Jean Leray (LMJL)

Soutenue le 29 septembre 2016

Development of high-order well-balanced schemes for geophysical flows

JURY

Président :	**M. Christophe CHALONS**, Professeur, Université de Versailles Saint-Quentin-en-Yvelines
Rapporteurs :	**M. Manuel J. CASTRO DÍAZ**, Professeur, Universidad de Málaga
	M. Jean-Paul VILA, Professeur, INSA Toulouse
Examinateurs :	**M. Stéphane CLAIN**, Professeur, Universidade do Minho
	M. Fabien MARCHE, Maître de conférences HDR, Université de Montpellier
Directeur de thèse :	**M. Christophe BERTHON**, Professeur, Université de Nantes
Co-encadrante de thèse :	**M^{me} Françoise FOUCHER**, Maître de conférences, École Centrale de Nantes

Contents

Remerciements

Comment commencer ce manuscrit autrement qu'en remerciant tous ceux et toutes celles sans qui son écriture n'aurait jamais pu avoir lieu ?

Tout d'abord, je tiens à remercier très chaleureusement mes encadrants de thèse, Christophe BERTHON et Françoise FOUCHER. C'est grâce à leurs conseils avisés et à leur patience que j'ai pu mener à bien cette thèse. Ils ont toujours su se montrer disponibles malgré un emploi du temps très surchargé, autant par des tâches administratives que de l'enseignement. Cette thèse n'aurait pas pu se passer aussi bien sans leur constante bienveillance. Merci.

Je remercie ensuite mes rapporteurs, Manuel J. CASTRO DÍAZ et Jean-Paul VILA, qui ont apporté d'importantes remarques m'ayant permis d'améliorer grandement mon manuscrit, en particulier en y rajoutant plusieurs cas-tests très pertinents.

Je remercie aussi les membres de mon jury, Stéphane CLAIN, Christophe CHALONS et Fabien MARCHE, d'avoir fait le déplacement pour ma soutenance. Je remercie tout particulièrement Stéphane de m'avoir accueilli deux fois à Braga et d'avoir toujours pris le temps, entre deux reuniões, de discuter de mes recherches.

Merci ensuite à Ana Paula, Anaïs, Annick, Brigitte, Katrin, Stéphanie et Valérie sans lesquelles je me serais sûrement perdu dans les méandres des ordres de mission et autres démarches administratives.

Je remercie également Éric et Saïd d'être toujours présents pour faire fonctionner le réseau, lax, et nos ordinateurs. J'en profite pour remercier Claude qui m'a sauvé en me dénichant un introuvable papier d'hydraulique. Merci aussi à Bertrand de la cafèt pour toutes les pauses café (j'en profite pour m'exuser de mon indécision chronique).

Toujours dans le laboratoire, je remercie toute l'équipe d'analyse numérique. Merci en particulier à Anaïs pour ses explications sur DG, à François pour tous ses encouragements, à Guy pour les cours d'OpenMP et pour sa patience quand on avait fait planter la frontale d'erdre, à Hélène pour les bières berlinoises, piriacaises et j'en oublie, à Marianne pour les repas partagés, toujours dans la bonne humeur, dans la froide et sombre salle commune du rez-de-chaussée, à Mazen pour cet excellent kebab berlinois, à Nicolas pour m'avoir bien dépatouillé sur des histoires d'entropie, de terme source et de champs caractéristiques, et enfin à Rodolphe pour son cours de M1 qui m'a donné envie de poursuivre sur la voie de l'analyse numérique.

Merci de plus à tous ceux et toutes celles rencontré(e)s ou retrouvé(e)s au hasard des conférences. Je pense, dans le désordre et en en oubliant beaucoup, à Markus, Nina, Flore, Franck,

François, Xavier, Laurent, Tong, Birte, Raphaël, Marie-Hélène, Matteo, Matthieu, Clémentine et Andrea. Je tiens à remercier tout particulièrement les organisateurs et les participants de SHARK-FV 2016. Il serait trop long de tous vous citer ici mais c'est grâce à vous tous que cette conférence m'a autant apporté, tant personnellement que mathématiquement.

Attaquons-nous maintenant aux doctorants du LMJL, passés, présents et à venir ! Il convient évidemment de commencer par mes cobureaux (et même colocataires, quoique brièvement pour Florian, désolé pour les ronflements !).

Dans l'ordre alphabétique, je commence par Florian. Merci pour tous ces moments partagés dans le bureau et en conférence (cf ton cahier de thèse... j'aurais dû en garder un peu pour les remerciements, c'était pas stratégique mon truc !). On se souviendra entre autres de la marque au plafond (qui a l'air de commencer à s'effacer toute seule), des séminaires que j'aurai sûrement loupés sans tes promptes interventions, surtout du haut des escaliers de Porto, en fait, et de tes vaches (près).

Quant à Moody, merci de ne pas avoir manqué ma soutenance (niark niark Florian) et d'avoir partagé à la fois mon bureau et mon appartement. Ces trois années de coburation (c'est sûrement un mot) et de colocation m'ont montré ta grande gentillesse et ton énorme générosité. Merci pour la Moodymobile et pour les soirées film. Par contre, pas merci d'avoir dégueulassé le tableau avec tes maths ! Pauvre tableau, il en a encore des séquelles... En plus, tu appuies fort, on en viendrait à se demander si tu n'as pas quelque chose contre les tableaux ☺.

J'en viens ensuite à tous les autres thésards, avec qui on trouve toujours matière à rigoler ou travailler, voire les deux en même temps !

Merci tout d'abord à ma grande sœur mathématique Céline pour ses encouragements et les dîners en compagnie d'Alex, Lucien et maintenant Martin. Merci aussi à mon grand frère Vivien, pour avoir partagé les arcanes de la thèse avec Christophe (lui donner du coca, ça l'apaise !) et pour avoir pris sous son aile un petit M2 perdu dans une conférence non loin d'Orléans.

Ensuite, dans l'ordre chronologique, merci à Gilberto pour les rappels à l'ordre sur la nourriture (et certaines choses merveilleuses, selon toi non comestibles) et sur les légumes (tu remarqueras que je n'inclus pas les légumes dans la nourriture). Merci à Thomas G. pour les discussions hardware et jeux vidéo (PoE, PoE et j'en oublie), et pour tant d'autres débats aussi intéressants les uns que les autres.

Merci à Antoine, dont le "mon canard" restera gravé dans les mémoires, pour cette visite guidée de Bruxelles (tu avais raison, on a bien mangé à Fritland !). Merci à Guogang pour cette soirée chez Virgile, j'espère n'avoir pas fait peur à ton fils de manière irréversible avec mes lacets ! Merci à Ilaria pour ton aide concernant le projet de recherche et les démarches pour la qualification. Merci à Virgile pour les bières partagées tant au Sur Mesure que lors de dîners chez Hélène, Marinette et toi.

Une mention spéciale pour Christophe, sans qui le laboratoire n'est plus le même. Merci pour ton amitié sincère, pour nous avoir accueillis à Aalborg, et pour toutes les discussions partagées dans le bureau après 18h et/ou chez Moody et toi après 20h.

Merci aux trois mousquetaires avec qui nous avons partagé autant d'années. Merci à Damien pour les bières au Michelet, et bon courage à Montréal (au moins tu auras froid, je suis jaloux) ! Merci à Pierre pour ta bonne humeur et tes fous rires communicatifs ; bonne chance dans les contrées lointaines du sud Loire, j'espère que tu parviendras à revenir nord Loire sous peu. Merci à Valentin pour nos comparaisons de PeiP en première année, et amuse-toi bien à Mayotte (sans pour autant manger de tortue) !

Courage à celles et à ceux dont la troisième année commence, et qui s'apprêtent à voir la lumière à la fin du tunnel. Merci à Guillaume, pour tes exposés encore plus clairs que ladite lumière. Merci à Johann, cobureau éphémère que l'on pardonnerait presque d'être Angevin. Merci à Noémie pour ces multiples déjeuners en salle commune et dehors, et pour ces dîners dégénérant en karaoké. Merci à Olivier de toujours répondre patiemment à mes questions mathématiques bizarres, pour les soirées et les bières (y compris à Aachen !). Merci d'avance à Thomas B. d'avoir manifesté (☺) un intérêt pour ma soutenance. Merci à Thomas W. pour ta bonne humeur, pour tes blagues, pour les parties de Citadelle et de Codenames, mais aussi pour le déménagement ! Merci à Victor VdR pour ta bonne humeur, pour tes blagues (tiens, ça ressemble à Thomas W. jusque là), et pour ta perte de voix costumée au Hellfest. ♪ We're not gonna take it... anymore ! ♪

Courage de même à ceux et celles qui entament leur deuxième année ou leur première année, et à ceux qui font leur thèse en quatre ans parce que leurs pays ont comme étrange coutume de faire passer le Master en même temps que la thèse. Merci à Hala pour ces instants de rire partagés à Aachen, surtout à la fin de la conférence où tu avais l'air vraiment fatiguée ! Merci à Radek de nous avoir fait goûter cette excellente et très rafraîchissante bière au pamplemousse. Merci à Vytauté pour son sourire constant.

J'en profite pour saluer les nouveaux et nouvelles : Caroline, Côme, Hélène, Solène et Zeinab, bienvenue dans le monde fantastique de la thèse, que vous avez la grande chance d'effectuer dans un laboratoire non moins fantastique !

Je n'oublie bien sûr pas les ATER et post-docs. Je pense particulièrement à Alberto (ce n'est que partie remise pour un dîner à un Amour de Pomme de Terre !), à Claire, qui a apporté sa chaleur toulousaine à la froide et sombre salle commune (je ne sais pas si je l'ai déjà mentionné mais cette salle commune est sombre et froide), à Guillem, qui nous a présenté le concept de soirée kiwi et donné plein d'idées de blagues intéressantes à faire (et de raisons de ne pas l'inviter en soirée 😄), à Nicolas, que je n'ai malheureusement pas eu trop le temps de connaître sauf au cours d'un déjeuner dans l'herbe, à Niccolò, un des seuls italiens qui ne manque pas de s'étouffer quand on parle de mettre de la raclette sur une pizza, à Victoria qui a su garder son calme et sa patience face à mes intrusions répétées dans son bureau, et à Zoé pour ta bonne humeur permanente, ta désinvolture communicative et pour cette danse sur RATM.

J'en viens enfin à ma famille et à mes amis, en commençant par des gros bisous à mon père et à ma mère qui ont toujours été là lorsque j'avais besoin d'eux. Merci aussi à mes grands-parents, toutes mes tantes et tous mes cousins/cousines pour d'excellents moments partagés à travers les années.

Merci à Alexandre, Alexis et Luc pour toutes ces années d'amitié. Alex, merci pour la coloc, merci pour toutes ces discussions et moments (voire des vacances entières, en fait)

partagés sur D2, sur WoW, sur PoE et bien d'autres jeux au cours du temps. En revanche, pas merci pour le chat défectueux ! Alexis, merci pour l'invitation à New York, pour les weekends à Paris et à Caen, pour Quickos et pour les innombrables parties de BG2. Luc, merci pour les virées à la Formathèque et à Brest au lycée, merci de savoir donner l'impression d'avoir joué à Dark Souls et à Starcraft sans pour autant avoir besoin de jouer à Dark Souls ou à Starcraft. Merci enfin à Alex et Luc pour les soirées Gigg's + Game Over !

Enfin, un très grand merci à Françoise et à Jean-Michel de m'avoir toujours accueilli dans des moments pas toujours faciles.

J'ai bien sûr gardé le plus grand merci de tous les mercis pour la fin. Merci à toi, ma Caroline, d'avoir égayé ma vie pendant ces (presque) deux dernières années. Merci de supporter mes bêtises, ma maladresse et mes définitions étranges du ménage et de ce qui est sale. Merci pour tous ces voyages passés et à venir, dont nous garderons toujours un souvenir magnifique. Merci aussi pour toutes nos discussions philosophiques et/ou mathématiques et/ou physiques, merci d'avoir supporté mes exposés de numérologue (plus qu'une dernière fois, c'est promis ! après ce ne sera un exposé différent). Merci pour tant d'autres choses qui me prendraient autant de place à écrire que le reste de la thèse. Merci, en tous cas, d'être présente à mes côtés.

Introduction

Version française

Les *équations de Saint-Venant* s'obtiennent à partir des équations de Navier-Stokes, en supposant que la dimension verticale est beaucoup plus petite que la dimension horizontale, et que la longueur d'onde des phénomènes modélisés est beaucoup plus grande que la profondeur de l'eau. Elles sont utilisées dans de nombreux domaines, comme la géophysique, l'océanographie ou l'évaluation des risques. Par exemple, le modèle de Saint-Venant est utilisé pour la simulation de *ruptures de barrage*, comme celle du barrage de Malpasset (voir [153]), qui s'est rompu en 1959 dans le Var, au sud de la France. Afin de mieux comprendre les conséquences d'une hypothétique rupture de barrage, le comportement de l'eau après la rupture doit être modélisé.

Une autre application directe des équations de Saint-Venant est l'étude de *tsunamis* ou d'*inondations*, comme par exemple à Madère (Portugal) ou à La Faute sur Mer (France) en 2010. D'autres travaux concernant la simulation et la prévention de tsunamis utilisent également les équations de Saint-Venant (voir [129, 9, 50]). Des glissements de terrain furent aussi modélisés en utilisant un modèle inspiré des équations de Saint-Venant (voir [104] par exemple).

Les équations de Saint-Venant en une dimension d'espace, munies des termes source de *topographie* et de *friction de Manning*, sont gouvernées par le système suivant (voir par exemple [122, 57]) :

$$\begin{cases} \partial_t h + \partial_x q = 0, \\ \partial_t q + \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -gh\partial_x Z - kq|q|h^{-\eta}, \end{cases} \tag{F1}$$

Dans (F1), $h(t,x)$, la hauteur d'eau, est positive ou nulle et $q(t,x)$, le débit de l'eau, a été moyenné sur la profondeur. De plus, g est la constante de gravité, $Z(x)$ est la fonction représentant la topographie, k est le coefficient de friction de Manning, et η est un paramètre, égal à $7/3$. Remarquons que, lorsque $Z = \mathrm{cst}$, la topographie est plate et le terme source de topographie s'annule, tandis que, lorsque $k = 0$, le terme source de friction devient nul.

Le but de ce manuscrit est de construire un schéma numérique adapté aux équations de Saint-Venant avec topographie et friction (F1). Notons que, lors de la simulation numérique de tsunamis, la préservation exacte d'un certain type de solutions est d'une importance cruciale. En effet, loin du tsunami, l'eau est au repos et sa surface ne doit pas être perturbée. La nécessité de cette propriété est particulièrement visible près de la côte, où, l'eau étant peu profonde, de petites perturbations de la hauteur d'eau deviennent relativement plus importantes, et viennent polluer l'approximation de la vitesse de pénétration du tsunami.

Par conséquent, un schéma numérique devrait assurer la préservation de solutions au repos, qui sont des cas particuliers de *solutions stationnaires*. Ces dernières sont obtenues lorsque h et q ne dépendent pas du temps, ce qui donne le système suivant :

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -gh\partial_x Z - kq|q|h^{-\eta}. \end{cases} \tag{F2}$$

La première équation de (F2) impose immédiatement un débit uniforme q. Comme les états stationnaires avec friction sont inconnus, une première partie de ce travail est d'étudier en détail la seconde équation, surtout dans le cas d'une topographie plate. Le but de cette étude est de comprendre au mieux les solutions stationnaires, afin d'aider à construire un schéma numérique capable de toutes les préserver.

La préservation numérique des états stationnaires des équations de Saint-Venant a été un important sujet de recherche au cours des deux dernières décennies. Nous devons à Bermudez et Vazquez [11], ainsi qu'à Greenberg et Leroux [87], les travaux pionniers dans ce domaine. Ces travaux portent sur la préservation des états stationnaires au repos. Dans ce deuxième article est introduite la propriété de *well-balance* d'un schéma, définie à l'origine pour qualifier un schéma capable de préserver ou capturer exactement les états stationnaires au repos. De tels schémas sont qualifiés de *schémas équilibre*. Ensuite, Gosse [82] étendit cette approche pour obtenir un schéma numérique pour les équations de Saint-Venant capable de préserver tous les états stationnaires, y compris ceux en mouvement, au prix d'une résolution approchée de l'équation non-linéaire les gouvernant. Ce travail fut ensuite simplifié par Audusse et al. [5], qui proposèrent la méthode de reconstruction hydrostatique, permettant de préserver les états stationnaires au repos sans avoir besoin de résoudre d'équation non-linéaire.

L'objectif principal de ce travail est de construire un schéma équilibre pour les équations (F1). Ce schéma doit être capable de capturer toutes les solutions stationnaires données par (F2). Le schéma numérique doit donc satisfaire les propriétés suivantes :

- préservation des états stationnaires donnés par (F2), y compris ceux où le débit est non nul ;
- préservation de la positivité de la hauteur d'eau ;
- capacité à approcher des transitions entres zones mouillées ($h \neq 0$) et sèches ($h = 0$).

De plus, la préservation des états stationnaires doit être réalisée sans résoudre d'équation non-linéaire, contrairement au schéma proposé par Gosse dans [82].

Un autre objectif de ce travail est de fournir deux extensions du schéma mentionné ci-dessus. La première extension concerne des géométries bidimensionnelles, primordiales pour pouvoir simuler des situations réelles, comme par exemple des inondations, des tsunamis ou des ruptures de barrage. La deuxième extension consiste à augmenter la précision du schéma, autrement dit à monter en ordre. La principale difficulté dans les deux cas est de recouvrer la propriété de préservation des états stationnaires vérifiée par le schéma unidimensionnel d'ordre un.

Plan du manuscrit

Premier chapitre : Les équations de Saint-Venant avec topographie et friction de Manning

Le premier chapitre de cette thèse est dédié à l'étude des équations de Saint-Venant et de ses termes sources de topographie et de friction de Manning. Ce système est gouverné par (F1). Ce chapitre contient à la fois des résultats connus (voir par exemple [80, 112]) et de nouveaux développements, qui portent surtout le terme de friction de Manning. Ces résultats seront largement utilisés lors de la dérivation d'un schéma numérique permettant une bonne approximation des solutions des équations de Saint-Venant.

Nous nous intéressons tout d'abord au système de Saint-Venant homogène, et nous rappelons plusieurs résultats connus, qui seront utiles pour étudier les effets des termes source. En particulier, nous exhibons les propriétés algébriques de ce système. Nous montrons que c'est un système hyperbolique de lois de conservation, et ce pour tout $h \geq 0$ et pour tout q. De plus, nous prouvons qu'il possède deux champs caractéristiques vraiment non-linéaires. Lors de l'étude d'un problème de Riemann pour les équations de Saint-Venant, chacun de ces champs caractéristiques est associé soit à une *onde de choc*, discontinue, soit à une *onde de détente*, continue. Nous exhibons plusieurs contraintes sur la solution exacte du problème de Riemann à la traversée de ces ondes. Dans le cas d'une onde de choc, la solution satisfait les relations de Rankine-Hugoniot, tandis que des quantités appelées invariants de Riemann sont constants à l'intérieur de l'éventail de l'onde de détente. Grâce à ces informations, nous obtenons ensuite la solution exacte du problème de Riemann. Plusieurs exemples de solutions exactes de problèmes de Riemann sont présentées, afin de mettre en lumière les propriétés du système de Saint-Venant homogène.

Par la suite, nous ajoutons les deux termes source au système de Saint-Venant. Nous présentons une nouvelle étude algébrique du système, qui prouve que l'hyperbolicité du système n'est pas perdue en présence des termes source, pourvu qu'une certaine condition soit satisfaite. Nous montrons aussi l'existence des deux mêmes champs caractéristiques vraiment non-linéaires. De plus, les termes source engendrent un champ caractéristique supplémentaire. Ce champ stationnaire est linéairement dégénéré, et il est associé à une *onde stationnaire*, c'est-à-dire une onde dont la vitesse caractéristique est nulle. Cette onde stationnaire est une discontinuité de contact, à la traversée de laquelle les invariants de Riemann sont constants. Cependant, cette onde stationnaire due aux termes source constitue une obstruction au calcul d'une solution exacte explicite du problème de Riemann.

En présence des termes source, nous avons donc une connaissance partielle de la structure du problème de Riemann. Nous essayons à présent d'exhiber les solutions stationnaires du système de Saint-Venant équipé de ses termes source. De telles solutions ne dépendent que de la variable d'espace, et satisfont donc un système d'équations différentielles ordinaires. Dans un souci de complétude et afin d'introduire plusieurs concepts essentiels par la suite, nous commençons par étudier les solutions stationnaires associées au seul terme source de topographie (voir [44]). Nous nous ramenons alors à étudier les zéros d'une fonction. Si une solution à ce problème existe, alors soit elle est unique, soit il y en a exactement deux. Dans ce deuxième cas, une des solutions est *subcritique*, tandis que la deuxième est *supercritique*.

Ensuite, nous étudions les solutions stationnaires régulières associées au seul terme source

de friction. Le problème de l'existence et de l'unicité de ces solutions se ramène encore à l'étude des zéros d'une fonction. En particulier, trois cas se présentent : soit il n'y a pas de solution, soit la solution est unique, soit deux solutions, l'une subcritique et l'autre supercritique, cohabitent. De plus, la hauteur d'eau critique, associée à la solution unique, est la même pour les deux termes source. Nous nous intéressons aussi à des solutions stationnaires discontinues, c'est-à-dire présentant des discontinuités admissibles. Les hauteurs d'eau de chaque côté d'une telle discontinuité doivent satisfaire à la fois les relations de Rankine-Hugoniot et une inégalité d'entropie. Nous donnons enfin quelques notions concernant les solutions stationnaires en présence des deux termes source de topographie et de friction.

Deuxième chapitre : Méthode des volumes finis

L'objectif du deuxième chapitre est d'introduire certaines notions essentielles à l'approximation numérique des équations de Saint-Venant, et plus largement de n'importe quel système hyperbolique de lois de conservation. Ces résultats sont bien connus, il n'y a pas de nouveauté dans ce chapitre. En revanche, il nous permet d'introduire des concepts et des notations qui seront utiles dans la suite du manuscrit.

Nous commençons par nous intéresser à la dérivation de schémas aux volumes finis en une dimension d'espace. De tels schémas sont utilisés pour approcher les solutions faibles de systèmes hyperboliques de lois de conservation. Après avoir introduit la discrétisation de l'espace en cellules et la discrétisation constante par morceaux de la solution du système, nous intégrons la loi de conservation afin d'exhiber le flux numérique, qui permet d'approcher l'intégrale en temps du flux physique. Plusieurs propriétés cruciales sont introduites : consistance, conservation et robustesse. Nous dérivons ensuite un schéma numérique aux volumes finis bien connu, le *schéma de Godunov*, introduit par Godunov en 1959 dans [81]. Ce schéma utilise la connaissance de la solution exacte du problème de Riemann associé à la loi de conservation afin d'obtenir un flux numérique. Cependant, connaître cette solution exacte est ardu, voire impossible, dans beaucoup de cas. Nous introduisons donc une autre technique, qui consiste à remplacer cette solution exacte par une solution approchée, obtenue par un solveur de Riemann approché. Cette méthode permet de définir les *schémas de type Godunov*, introduits au début des années 1980 par Roe (voir [135]) et Harten, Lax et van Leer (voir [90]). Un tel schéma numérique sera utilisé dans la suite du manuscrit afin d'approcher les solutions des équations de Saint-Venant, tout en respectant certaines propriétés essentielles.

Les schémas mentionnés ci-dessus sont d'ordre un en espace et en temps. Afin de les rendre plus précis et d'obtenir un ordre deux de convergence en espace, la méthode MUSCL a été proposée par van Leer dans [154]. Cette technique consiste à remplacer, dans chaque cellule, l'approximation constante par morceaux par une approximation linéaire par morceaux. Cette méthode peut être étendue pour obtenir un ordre supérieur à deux, en utilisant des reconstructions polynomiales de degré plus élevé. Cependant, cette technique introduit des instabilités, qui peuvent être corrigées par l'emploi d'un limiteur de pente.

Après avoir traité le cas d'une dimension d'espace, nous nous intéressons à des lois de conservation en *deux dimensions* d'espace. De la même façon que précédemment, l'espace est découpé en cellules, où la solution approchée est constante par morceaux. Le système de lois

de conservation est ensuite intégré sur les cellules afin d'obtenir un schéma aux volumes finis en deux dimensions d'espace. En particulier, ce schéma fait intervenir le flux numérique à chaque interface entre cellules. Nous démontrons aussi un résultat selon lequel ce schéma 2D peut s'écrire comme combinaison convexe de schémas 1D. Ce résultat permet de connaître aisément certaines propriétés du schéma 2D, pourvu qu'elles soient vérifiées par les schémas 1D.

Enfin, nous introduisons un terme source dans la loi de conservation 2D, et nous dérivons un schéma numérique d'*ordre élevé* (c'est-à-dire d'ordre strictement supérieur à deux), qui se base sur une technique de reconstruction polynomiale introduite par Clain, Diot et Loubère (voir [46, 63, 65]). L'ordre élevé en temps est obtenu par l'utilisation de méthodes de type SSPRK (voir [84]). Comme dans le cas 1D, nous observons que cette reconstruction engendre des oscillations. Afin de s'en affranchir, nous suggérons d'utiliser la méthode MOOD. Cette méthode a elle aussi été introduite par Clain, Diot et Loubère ; elle consiste à baisser graduellement le degré de la reconstruction polynomiale dans les cellules où cela s'impose, jusqu'à ce que les oscillations disparaissent, et que les propriétés de robustesse du schéma 2D d'ordre un soient recouvrées.

Troisième chapitre : Un schéma équilibre pour les équations de Saint-Venant

Ce troisième chapitre est dédié à l'étude numérique des équations de Saint-Venant, dans le but de dériver un schéma numérique possédant certaines propriétés essentielles. Il doit être consistant, robuste, doit permettre d'approcher les interfaces entre zones mouillées et sèches, et il doit exactement préserver tous les états stationnaires des équations de Saint-Venant avec topographie et/ou friction de Manning.

Afin de s'assurer de la préservation des états stationnaires, on utilise un schéma de type Godunov, qui s'appuie sur la présence de l'onde stationnaire créée par les termes source, ainsi que sur une discrétisation pertinente de ceux-ci. Ce schéma est tout d'abord dérivé pour un terme source générique sur l'équation de conservation du débit, que l'on approche par une moyenne. Cette approximation est ensuite calculée pour les termes source individuels de topographie et de friction. Cependant, lorsque les deux termes source sont présents, la même méthode ne peut pas être appliquée puisque les états stationnaires sont gouvernés par une équation différentielle et ne peuvent pas être vus comme les zéros d'une certaine fonction. Par conséquent, tous les états stationnaires avec topographie et friction ne peuvent pas être préservés exactement ; seuls ceux provenant d'une certaine discrétisation de l'équation différentielle peuvent l'être. Nous donnons aussi une technique permettant d'assurer la robustesse du schéma, quel que soit le terme source (voir [7]). Enfin, nous étendons ce schéma pour prendre en compte des hauteurs d'eau nulles.

En revanche, le schéma que nous avons suggéré ne permet pas de donner une bonne approximation des transitions entre zones mouillées et zones sèches. En effet, le terme source de friction devient raide lorsque la hauteur d'eau diminue. Afin de remédier à ce problème sans modifier le pas de temps du schéma, nous proposons une méthode semi-implicite, qui consiste à traiter les parties flux et topographie de façon explicite, puis la partie friction de façon implicite. La propriété de préservation des états stationnaires est maintenue grâce à

une discrétisation pertinente de la hauteur d'eau.

La dernière étape de ce chapitre consiste à proposer des tests numériques permettant de valider les propriétés du schéma. Notons que, puisque nous avons dérivé un schéma équilibre, nous ne pouvons pas le valider avec les cas-tests usuels basés sur des solutions stationnaires ; en effet, de telles solutions sont exactement préservées. Nous effectuons tout d'abord des tests visant à vérifier que différents types de solutions stationnaires sont exactement préservées par le schéma : des solutions au repos, ainsi que des solutions stationnaires génériques pour les termes sources de topographie et/ou de friction, dont certains cas-tests bien connus venant de [86]. Des cas-tests de validation du schéma sont ensuite proposés. Ils permettent de vérifier les propriétés de consistance et de robustesse du schéma, ainsi que sa capacité à approcher les transitions entre zones mouillées et zones sèches. Nous proposons deux cas-tests tirés de [44], ainsi que plusieurs cas-tests de rupture de barrage, sur fond mouillé ou sec. En particulier, une rupture de barrage sur fond sec avec une topographie non plate permet de vérifier toutes les propriétés du schéma, y compris la préservation des solutions stationnaires au repos.

Quatrième chapitre : Extensions à deux dimension d'espace et à l'ordre élevé

Dans le chapitre précédent, nous avons dérivé un schéma numérique préservant tous les états stationnaires des équations de Saint-Venant munies des termes source de topographie et de friction de Manning. Le but de ce quatrième et dernier chapitre est d'étendre ce schéma pour prendre en compte des géométries bidimensionnelles et d'obtenir un ordre élevé d'approximation.

Tout d'abord, l'extension à *deux dimensions* d'espace est effectuée dans l'esprit de la combinaison convexe évoquée dans le deuxième chapitre. Certaines propriétés du schéma 1D sont ainsi conservées, comme la robustesse et le traitement semi-implicite de la friction. Cependant, la préservation des états stationnaires ne s'étend pas complètement en deux dimensions. En effet, les états stationnaires vraiment 2D sont régis par une équation aux dérivées partielles, et seuls les états stationnaires 1D par direction sont préservés.

La deuxième partie de ce chapitre concerne l'obtention d'un schéma d'*ordre élevé* à partir du schéma 2D proposé dans la première partie. Ce schéma est obtenu en suivant les idées énoncées dans le deuxième chapitre. La méthode MOOD est utilisée afin d'éliminer les oscillations induites par la reconstruction polynomiale. Cependant, la reconstruction modifie aussi les valeurs approchées aux interfaces, ce qui entraîne la perte de la propriété de préservation des états stationnaires. Afin de recouvrer cette propriété, nous suggérons une *combinaison convexe* entre le schéma d'ordre un et le schéma d'ordre élevé. Ce dernier est utilisé loin d'une solution stationnaire, tandis que le schéma d'ordre un est utilisé lorsque la solution approchée est assez proche d'une solution stationnaire. Par conséquent, le schéma obtenu est au moins d'ordre élevé, puisque le schéma d'ordre un est utilisé dans les zones stationnaires, où il est en fait exact (c'est-à-dire d'ordre infini).

Ensuite, nous évoquons l'implémentation de ce schéma. Nous choisissons de développer un code en Fortran, muni d'une parallélisation en OpenMP. Plusieurs fonctions de la bibliothèque LAPACK sont utilisées dans ce code, et ses fichiers de sortie sont au format `vtk`. Nous

étudions aussi l'efficacité de la parallélisation.

Enfin, nous proposons plusieurs cas-tests destinés à tester les propriétés du schéma bidimensionnel d'ordre élevé. Nous vérifions tout d'abord qu'il préserve bien les états stationnaires 1D par direction, et en particulier les états stationnaires au repos. Ensuite, nous proposons deux cas-tests destinés à vérifier l'ordre du schéma. S'ensuivent plusieurs simulations de validation numérique du schéma. Ces cas-tests sont des ruptures de barrage, sur fond mouillé ou sur fond sec. Ils permettent de mettre en évidence la contribution du terme source de friction, ainsi que la pertinence de la combinaison convexe. Finalement, deux simulations réelles sont proposées : celle du tsunami qui a frappé le Japon en 2011, et celle d'un tsunami sur une topographie urbaine.

English version

The *shallow-water equations* are derived from the Navier-Stokes equations, with the assumption that the vertical dimension is much smaller than the horizontal one, and that the wavelength of the phenomenon is much larger than the depth of the water. They are widely used in many fields, such as geophysics, oceanography or risk assessment. For instance, the shallow-water model is used in the simulation of *dam-breaks*, such as the Malpasset dam-break (see [153]), which took place in southern France in 1959. To better understand the consequences of a hypothetical dam failure, one has to model the behavior of water after the dam has failed.

Another direct application of the shallow-water equations is the study of *floods* or *tsunamis*. For instance, we mention the floods in La Faute sur Mer, in France, in 2010, and Madeira, in Portugal, also in 2010. Other work related to tsunami prevention and simulation also use the shallow-water equations (see [129, 9, 50]). A model inspired from the shallow-water equations was also used to perform landslide simulations (see [104] for instance).

The shallow-water equations in one space dimension with the *topography* and the *Manning friction* source terms read as follows (see [122, 57] for instance):

$$\begin{cases} \partial_t h + \partial_x q = 0, \\ \partial_t q + \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -gh\partial_x Z - kq|q|h^{-\eta}. \end{cases} \tag{E1}$$

In (E1), $h(t,x)$ is the nonnegative water height, $q(t,x)$ is the depth-averaged discharge of the water, g is the gravity constant, $Z(x)$ is the topography function representing the shape of the bottom, k is the Manning friction coefficient, and η is a parameter, equal to $7/3$. One can easily see that, when $Z = \mathrm{cst}$, the topography is flat and the topography source term vanishes, while, when $k = 0$, the friction source term vanishes.

The goal of this manuscript is to derive a numerical scheme suited to the shallow-water equations with topography and friction (E1). Let us remark that, in numerical simulations, for instance those involving tsunamis, the preservation a certain class of solutions is of prime importance. Indeed, away from the tsunami, the water is at rest and its surface should not be perturbed. This property is especially relevant next to the shore, since small perturbations in the water height are more detrimental to the solution in this area, and the approximation of the velocity of the tip of the tsunami is polluted by such perturbations.

As a consequence, a numerical scheme should ensure that the solutions at rest, which are noting but specific cases of *steady state solutions*, are exactly preserved. The steady state solutions are obtained by making the time derivatives in (E1) vanish, thus yielding the following system:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -gh\partial_x Z - kq|q|h^{-\eta}. \end{cases} \tag{E2}$$

The first equation immediately imposes a uniform discharge q. Since the steady states with friction are unknown, the first part of the present work is to perform an in-depth study of the second equation, especially in the case of a flat topography. The goal of this study is to

understand the steady state solutions as best as possible, in order to help build a relevant numerical scheme, able to preserve these steady states.

The numerical preservation of steady states for the shallow-water equations has been of prime importance during the last two decades. This work was pioneered by Bermudez and Vazquez [11] as well as Greenberg and Leroux [87], who tackled the preservation of steady states at rest. This second paper introduced the *well-balance* property of a scheme, originally defined as the ability of a scheme to exactly preserve and capture the steady states at rest. Next, Gosse [82] extended this approach to yield a well-balanced scheme for the shallow-water equations able to preserve all the steady states, including the moving ones, with the additional requirement of approximately solving the governing nonlinear equation. This work was later simplified by Audusse et al. [5], who proposed the so-called hydrostatic reconstruction, which allows the preservation of the steady states at rest without needing to solve a nonlinear equation.

As a consequence, the main objective of this work is to build a well-balanced scheme for the equations (E1). Here, the expression *well-balanced* describes a scheme that is able to exactly capture all the steady states (E2). Therefore, the numerical scheme needs to satisfy the following properties:

- well-balance, i.e. preservation of the steady states (E2), even the moving ones;
- robustness, i.e. preservation of the non-negativity of the water height;
- ability to approximate transitions between wet ($h \neq 0$) and dry ($h = 0$) areas.

In addition, the well-balance property must be satisfied without having to solve a nonlinear equation, unlike the scheme suggested by Gosse in [82].

Another objective of this work is the extension of the above scheme to two-dimensional geometries. Indeed, such an extension is primordial in order to consider real-life simulations, such as simulations of catastrophic events (for instance floods, tsunamis, dam-breaks). In addition, a high-order extension of the scheme must be considered. The main challenge of these two extensions is the recovery of the well-balance property.

Outline of the manuscript

Chapter 1: The shallow-water equations with topography and Manning friction

The first chapter is devoted to the study of the shallow-water system, supplemented with the source terms of topography and Manning friction, and governed by (E1). This chapter contains both known results (see for example [80, 112]) and new developments, especially concerning the Manning friction source term. These results will be heavily used when deriving a suitable numerical scheme to provide approximate solutions to the shallow-water equations.

We first consider the homogeneous shallow-water system, and we recall some well-known results, which will be instrumental when studying the effects of the source terms. In particular, we exhibit the algebraic properties of this system. It is shown to be a hyperbolic system of conservation laws for all $h \geq 0$ and all q. In addition, we prove that it possesses two genuinely nonlinear characteristic fields. When considering a Riemann problem for the shallow-water equations, each one of the characteristic fields is associated either to a discontinuous *shock*

wave or to a continuous *rarefaction wave*. Across these waves, several constraints are exhibited for the exact Riemann solution. Namely, the Rankine-Hugoniot relations are satisfied in the case of a shock wave, while the Riemann invariants are constant within the fan of a rarefaction wave. Equipped with the knowledge of these relations, the exact solution of the Riemann problem is derived. Several examples of Riemann problems, together with their exact solutions, are given to highlight the properties of the homogeneous shallow-water system.

Afterwards, we add both source terms to the shallow-water system. Another algebraic study of the system is then performed, which proves that it is still hyperbolic even in the presence of the source terms, under a specific condition. The same two genuinely nonlinear fields are also uncovered. Moreover, there is now an additional characteristic field, which corresponds to the source term. This stationary characteristic field is linearly degenerate and it is associated to a *stationary wave*, i.e. a wave with a zero characteristic velocity. This stationary wave is a contact discontinuity, through which the Riemann invariants are constant. However, the presence of the source terms, and thus that of this wave, does not allow computing an explicit solution to the Riemann problem anymore.

Equipped with some knowledge of the structure of the Riemann problem, we then turn to exhibiting several steady state solutions of the shallow-water system endowed with the source terms. Such solutions only depend on the space variable, and they satisfy a system of ordinary differential equations. For the sake of completeness and in order to introduce several key concepts, we first study the steady state solutions associated to the topography source term only (see [44]). We then show that this is equivalent to studying the zeros of a function. If a solution to this problem exists, then either it is unique or there are exactly two solutions. If two solutions exist, then one of them is *subcritical* and the other one is *supercritical*.

Subsequently, we study the smooth steady state solutions associated to the friction source term only. Studying the existence and the uniqueness of these solutions is again equivalent to finding the zeros of a function. In particular, there may be no solution, or there may be a unique solution, or there may be two solutions, a subcritical one and a supercritical one. In addition, the critical water height, associated to the unique solution, is the same for both source terms. We also study discontinuous steady states, i.e. steady states presenting admissible discontinuities. The water heights on each side of such discontinuities must satisfy both the Rankine-Hugoniot relations and an entropy inequality. Namely, their existence is tied to the direction of the steady water flow. We finally give a few words about the steady state solutions with both source terms of topography and friction.

Chapter 2: Finite volume methods

The objective of the second chapter of this manuscript is to introduce several essential notions related to the numerical approximation of the shallow-water equations, and more widely of any hyperbolic system of conservation laws. These notions are well-known, and there are no new results in this chapter. However, in the remainder of the manuscript, the concepts and notations introduced in this chapter will be heavily used.

We begin with the derivation of finite volume schemes in one space dimension. Such schemes are used to approximate weak solutions of hyperbolic systems of conservation laws.

After having introduced the discretization of the space domain in cells and the piecewise constant approximation of the solution, the conservation law is integrated in order to exhibit the numerical flux, which provides an approximation of the time integral of the physical flux. Several essential properties are introduced; namely, the consistency, the conservation and the robustness. We then derive a well-known finite volume scheme, *Godunov's scheme*, introduced by Godunov in 1959 in [81]. This scheme uses the knowledge of the exact solution to the Riemann problem associated to the conservation law in order to propose a numerical flux. However, knowing this exact solution is no easy task in the general case; it may even be impossible. We therefore introduce another method, which consists in replacing the exact Riemann solution with an approximate one, obtained thanks to an approximate Riemann solver. This technique allows defining the *Godunov-type schemes*, introduced at the beginning of the 1980s by Roe (see [135]) and Harten, Lax and van Leer (see [90]). Such a scheme will be used later in the manuscript in order to provide approximate solutions of the shallow-water equations, while retaining several essential properties.

The schemes mentioned above are first-order accurate in space and time. In order to improve the accuracy of such schemes and to obtain a second order of accuracy in space, we choose the MUSCL method, suggested by van Leer in [154]. This technique consists in replacing, in each cell, the piecewise constant approximation with a piecewise linear approximation. This method can also be extended to get a better order of convergence, by using a higher reconstruction degree. However, this technique also introduces instabilities, which may be corrected thanks to slope limiters.

After having tackled the case of one space dimension, we focus on conservation laws in *two space dimensions*. Similarly to the 1D case, the space domain is discretized with cells, and the approximate solution is assumed to be piecewise constant. The system of conservation laws is then integrated over the cells in order to obtain a finite volume scheme in two space dimensions. In particular, the numerical flux is used at each interface between cells. We also prove a result stating that this 2D scheme may be rewritten as a convex combination of 1D schemes. This result allows to immediately establish several properties of the 2D scheme, provided they are satisfied by the 1D scheme.

Finally, we add a source term to the 2D conservation law, and we derive a *high-order accurate* numerical scheme, i.e. a scheme of order strictly superior to two, based on a polynomial reconstruction technique introduced by Clain, Diot and Loubère (see [46, 63, 65]). The high order accuracy in time is obtained by using SSPRK methods (see [84]). As in the 1D case, we observe that this reconstruction procedure induces spurious oscillations. In order to ensure that such oscillations do not appear, we suggest using the MOOD method. This technique was also introduced by Clain, Diot and Loubère; it consists in gradually lowering the degree of the polynomial reconstruction in cells where this is needed, until the oscillations disappear, and until the robustness properties of the 2D schemes are recovered.

Chapter 3: A well-balanced scheme for the shallow-water equations

In this third chapter, we begin the numerical study of the shallow-water equations in order to derive a numerical scheme, which must satisfy several essential properties. It must be

consistent, robust, able to approximate the interfaces between wet and dry areas, and it must be well-balanced, i.e. it must exactly preserve all the steady state solutions of the shallow-water equations with topography and/or Manning friction.

In order to ensure the preservation of the steady states, we use a Godunov-type scheme, based on the stationary wave created by the source terms, and on a relevant discretization of the source terms. This scheme is first derived for a generic source term on the discharge equation, by introducing an approximation of the average of this source term. This approximation is then computed for the individual source terms of topography and friction. To that end, we rely on the fact that the steady states associated to the individual source terms can be seen as the zeros of a nonlinear function, and solving the nonlinear equations arising in this case is not necessary. However, when both source terms are present, the same method cannot be applied since the steady state solutions are now governed by a differential equation and cannot be seen as the zeros of a function. As a consequence, all the steady state solutions with topography and friction cannot be exactly preserved; the scheme is able to preserve only those obtained from a specific discretization of the differential equation. We also suggest a technique ensuring the robustness of the scheme, for any source term (see [7]). Finally, we extend this scheme to take vanishing water heights into account.

However, this scheme does not give a good approximation of the transitions between wet and dry areas. Indeed, the friction source term becomes stiff when the water height tends to zero. To address such an issue without modifying the time step of the scheme, we suggest a semi-implicit method. This technique consists in providing an explicit treatment of the flux and the topography, and an implicit treatment of the friction. The well-balance property is satisfied thanks to a relevant discretization of the water height.

The last part of this chapter consists in performing several numerical experiments, whose goal is to assess the properties of the scheme. Note that, since the scheme is well-balanced, we cannot perform the usual validation experiments involving steady state solutions; indeed, such solutions are exactly preserved. We first check the well-balance of the scheme. To that end, we try to preserve several types of steady state solutions: steady states at rest and moving steady states for the topography and/or the friction, including several well-known test cases from [86]. Several validation test cases are then carried out. They allow the verification of the consistency and robustness properties, as well as the ability of the scheme to approximate the transitions between wet and dry areas. We suggest two experiments from [44], as well as several dam-break simulations, either on a wet bed or on a dry bed. In particular, a dry dam-break problem with a non-flat topography allows assessing of all the properties satisfied by the scheme, including the well-balance.

Chapter 4: Two-dimensional and high-order extensions

In the previous chapter, we have derived a numerical scheme that preserves all the steady states of the shallow-water equations endowed with the topography and Manning friction source terms. The goal of this fourth and last chapter is to provide two-dimensional and high-order extensions of this scheme.

First, the extension in *two space dimensions* is carried out in the spirit of the convex com-

bination introduced in the second chapter. Several properties of the 1D scheme are thus conserved, such as the robustness and the semi-implicit treatment of the friction. However, the well-balance property is not fully extended to two dimensions. Indeed, the truly 2D steady states are governed by a partial differential equation, and only the 1D steady states are preserved: the scheme is said to be well-balanced by direction.

The second part of this chapter consists in providing a *high-order* extension of the 2D scheme. This extension is obtained by following the ideas presented in the second chapter. The MOOD method is used in order to eliminate the oscillations induced by the polynomial reconstruction. However, the reconstruction procedure also modifies the approximate solution at the interfaces, which leads to a loss of the well-balance property. In order to recover this property, we suggest a *convex combination* between the first-order well-balanced scheme and the high-order scheme. The former is used when the approximate solution is close to being a steady state, while the latter is favored when the approximate solution is far from a steady state. As a consequence, the scheme is at least high-order accurate, since the first-order scheme is used in areas where the solution is steady, that is to say where it is exact (i.e. where its order of accuracy is infinite).

Then, we discuss the implementation of this scheme. We elect to develop a Fortran code, which is supplemented with an OpenMP parallelization. Within this code, several routines from the LAPACK library are used; in addition, its output consists in `.vtk` files. The efficiency of the parallelization is also discussed.

Finally, we carry out several numerical experiments, whose purpose is to assess the properties of the 2D high-order well-balanced scheme. We first check the well-balance of the scheme on 1D steady states, and on a truly 2D steady state at rest. Then, two assessments of the high order of accuracy are performed. Afterwards, several dam-break validation test cases are carried out. Their purpose is to highlight the contribution of the friction source term, as well as the relevance of the convex combination procedure suggested to restore the well-balance property of the high-order scheme. Finally, two real-world simulations are suggested: the 2011 Great East Japan tsunami, in Tōhoku, Japan, and a tsunami on an urban topography.

Publication list

Published

V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography. *Comput. Math. Appl.*, 72(3):568-–593, 2016.

Preprint

V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography and Manning friction. *preprint available on HAL (HAL id: hal-01247813)*, December 2015.

In progress

V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A two-dimensional high-order well-balanced scheme for the shallow-water equations with topography and Manning friction.

Conference proceedings

C. Berthon, M. de Leffe, and V. Michel-Dansac. A conservative well-balanced hybrid SPH scheme for the shallow-water model. In *Finite volumes for complex applications. VII. Elliptic, parabolic and hyperbolic problems*, volume 78 of *Springer Proc. Math. Stat.*, pages 817–825. Springer, Cham, 2014.

Communication list

Talks

3. SHARK-FV 3, São Félix, Portugal, May 2016
2. 8th ICIAM, Beijing, China, August 2015
1. 3rd summer school of the GDR EGRIN, Piriac-sur-Mer, France, June 2015

Posters

3. HYP2016, Aachen, Germany, August 2016
2. 2nd summer school of the GDR EGRIN, Domaine de Chalès, France, July 2014
1. Finite Volumes for Complex Applications - FVCA VII, Berlin, Germany, June 2014

1

The shallow-water equations with topography and Manning friction

This chapter is dedicated to an introduction of the shallow-water system with topography and Manning friction. The shallow-water system has been introduced in 1871 by de Saint-Venant (see [10]) and is obtained by depth-integrating the Navier-Stokes equations, in the case where the wavelength of the modeled phenomena is much larger than the depth of the fluid. For instance, tsunami propagation in an ocean falls within this framework. In one space dimension, the shallow-water system with topography and Manning friction is given by:

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + \frac{1}{2}gh^2\right) = -gh\partial_x Z - kq|q|h^{-\eta}, \end{cases} \tag{1.1}$$

where $h(t,x) \geq 0$ is the *water height*, $u(t,x)$ is the *water velocity*, and $g = 9.81$ m.s^{-2} is the gravity constant. Both the height and the velocity depend on the time variable t and the space variable x. The *topography source term* $-gh\partial_x Z$ takes into account the geometry of the channel in which the water is flowing, thanks to the function $Z : \mathbb{R} \to \mathbb{R}$, which models the shape of the channel bottom, as displayed on Figure 1.1. The topography function is assumed to be smooth. The *Manning friction model*, introduced by Manning in [122], provides the source term $-kq|q|h^{-\eta}$, where $q = hu$ is the *discharge*. This source term models the friction of the channel bottom. The Manning coefficient k is used to determine the intensity of the friction: the higher k is, the more friction is exerted by the bottom on the water. The quantity η is a parameter, taken equal to $7/3$ in Manning's model.

The system (1.1) can be rewritten under the following condensed form:

$$\partial_t W + \partial_x F(W) = \mathfrak{S}(W),$$

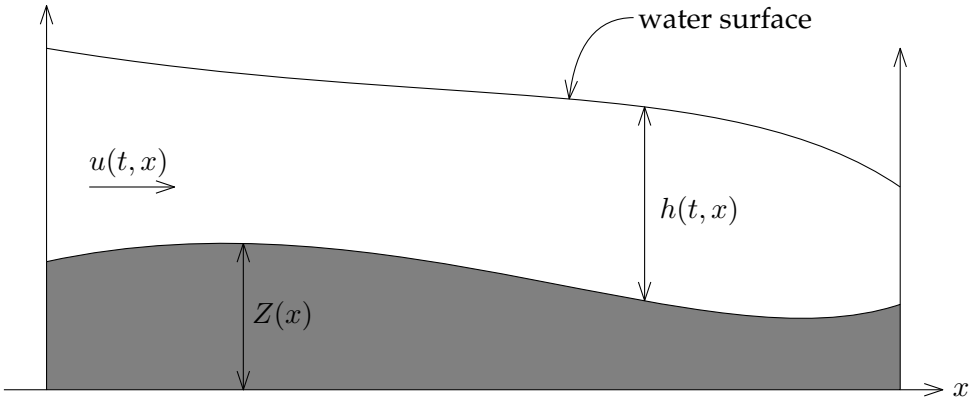

Figure 1.1 – The 1D shallow-water equations with a non-flat bottom. The gray area is the topography.

where we have set:

$$W = \begin{pmatrix} h \\ hu \end{pmatrix} \quad ; \quad F(W) = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \end{pmatrix} \quad ; \quad \mathfrak{S}(W) = \begin{pmatrix} 0 \\ -gh\partial_x Z - kq|q|h^{-\eta} \end{pmatrix}, \tag{1.2}$$

where W lives in the *admissible states space* Ω, to be defined later.

Note that the case $h = 0$ corresponds to a dry area. Such areas naturally appear, for instance while considering the penetration of a wave on a beach or the breaking of a dam. We here make the important remark that the shallow-water system (1.1) can be extended for vanishing water heights. Since the velocity is given by $q = hu$, a definition of the velocity has to be provided for vanishing water heights. To address this issue, the following assumption is made.

Assumption. The velocity vanishes as soon as the water height does.

This assumption makes physical sense: if there is no water, then the water does not move. We remark that the friction source term $-kq|q|h^{-\eta}$ also requires a special treatment when h tends to zero. Throughout the manuscript, the following assumption is made.

Assumption. The friction source term vanishes as soon as the water height does.

This assumption is motivated by the fact that a vanishing water height means that the bottom is no longer able to exert a friction force on the water. Therefore, the admissible states space Ω is defined as follows for the shallow-water equations (1.1):

$$\Omega = \left\{ W = {}^t(h,q) \in \mathbb{R}^2 \; ; \; h \geq 0, \, q \in \mathbb{R} \right\}. \tag{1.3}$$

In addition, the homogeneous shallow-water equations, obtained by making the source terms vanish in (1.1), admit an *entropy pair* (see [112, 5] for instance). The notion of entropy (see for instance [107, 80, 108]) is used to determine the physical admissibility of a weak solution of the system (1.1), i.e. a solution which satisfies (1.1) in the space of distributions. An entropy pair is made of a convex function $s \in \mathcal{C}^2(\Omega)$, the *entropy*, and a function $G \in \mathcal{C}^2(\Omega)$,

the *entropy flux*, such that the following identity holds:

$$\nabla_W F(W) \nabla_W s(W) = \nabla_W G(W).$$

In the context of the homogeneous shallow-water equations, the entropy pair is given by:

$$s(W) = \frac{1}{2} h u^2 + \frac{1}{2} g h^2 \quad \text{and} \quad G(W) = hu\left(\frac{u^2}{2} + gh\right). \tag{1.4}$$

A weak solution W of the homogeneous shallow-water system is called *entropy-satisfying* if it satisfies the following entropy inequality:

$$\partial_t s(W) + \partial_x G(W) \leq 0. \tag{1.5}$$

Regarding the inhomogeneous shallow-water system (1.1), an entropy inequality is also exhibited (see for instance [5, 12]). With the same entropy pair (1.4), the following entropy inequality holds for a weak entropy-satisfying solution of (1.1):

$$\partial_t s(W) + \partial_x G(W) \leq -ghu\partial_x Z - kq^2|q|h^{-\eta-1}. \tag{1.6}$$

Since the topography is a smooth function, the entropy inequality (1.6) can be rewritten as follows:

$$\partial_t \tilde{s}(W, Z) + \partial_x \widetilde{G}(W, Z) \leq -kq^2|q|h^{-\eta-1}, \tag{1.7}$$

where the entropy pair $(\tilde{s}, \widetilde{G})$ is given by:

$$\tilde{s}(W, Z) = s(W) + hgZ \quad \text{and} \quad \widetilde{G}(W, Z) = G(W) + hugZ. \tag{1.8}$$

Equipped with these general properties of the shallow-water equations, the goal of this chapter is to provide some particular solutions of the shallow-water equations with topography and Manning friction (1.1).

Namely, the structure of the solutions of a Riemann problem is studied in Section 1.1. First, the Riemann problem for the homogeneous shallow-water equations is discussed and several examples are given. Afterwards, we determine the structure of the Riemann solution for the inhomogeneous shallow-water system (1.1).

Then, the steady state solutions are exhibited in Section 1.2. The steady states are a specific class of solutions for which the time derivative vanishes. These solutions are non-trivial because of the presence of the source terms. First, steady state solutions for the source term of topography only are highlighted. Then, we exhibit steady state solutions for the Manning friction source term only. Finally, a few words are given on the steady state solutions associated to both source terms of topography and friction.

1.1 Properties of the shallow-water equations

This section is devoted to highlighting some essential properties of the shallow-water equations in one space dimension; it is organized as follows.

First, we consider in Section 1.1.1 the homogeneous system, obtained by making the source terms vanish in (1.1). We exhibit the eigenvalues of the Jacobian matrix of its flux function. These eigenvalues are associated to waves that appear when considering a Riemann problem. The nature of these waves is then discussed by focusing on the characteristic fields.

Equipped with this algebraic study of the shallow-water system, Section 1.1.2 is then devoted to exhibiting the solution of the Riemann problem for the homogeneous shallow-water equations. This process has been described in [112] (see also [150] for the case of a generic system of conservation laws, and see [98] for the Euler equations with two different equations of state). First, the general form of the solution is established. Then, examples are provided.

Finally, the source terms of topography and Manning friction are added to the system in Section 1.1.3. The equations (1.1) are studied in the presence of these source terms, and a stationary wave is exhibited.

1.1.1 The homogeneous system

The goal of this section is to show that the homogeneous shallow-water system is hyperbolic, and to provide some insight on the characteristic fields that this system induces. We consider a smooth solution of the following homogeneous system, obtained from (1.1):

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + \dfrac{1}{2}gh^2\right) = 0. \end{cases} \tag{1.9}$$

In addition, we assume that $h > 0$. All the computations in this section will be made with respect to the primitive variables $U = {}^t(h, u)$. This choice is made for the sake of simplicity, since the properties of the system are independent of the choice of the variables. As a consequence, in this section, the primitive variables U lie in the following restricted admissible states space:

$$\Omega_U = \left\{U = {}^t(h, u) \in \mathbb{R}^2 \ ; \ h > 0, \ u \in \mathbb{R}\right\}.$$

The primitive variables

We begin by rewriting the shallow-water system (1.9) with the primitive variables $U = {}^t(h, u)$. The goal here is to exhibit the eigenvectors of the Jacobian matrix of the flux function. For a smooth solution, the system (1.9) reads:

$$\begin{cases} \partial_t h + u\partial_x h + h\partial_x u = 0, \\ \partial_t u + g\partial_x h + u\partial_x u = 0, \end{cases}$$

We can therefore cast the shallow-water system into the following nonconservative form:

$$\partial_t U + A(U)\partial_x U = 0, \tag{1.10}$$

where $A(U)$ represents a matrix similar to the Jacobian matrix of the physical flux function. The matrix $A(U)$ is given by:

$$A(U) = \begin{pmatrix} u & h \\ g & u \end{pmatrix}.$$

Hyperbolicity of the system

The next step in the study of the shallow-water system consists in computing the eigenvalues of the matrix $A(U)$. If $A(U)$ is diagonalizable in $\mathbb{R}$, then the shallow-water system is hyperbolic. After straightforward computations, we get the following expressions for the two eigenvalues of the matrix $A(U)$:

$$\lambda_-(U) = u - c \quad \text{and} \quad \lambda_+(U) = u + c, \tag{1.11}$$

where we have introduced the sound speed c, defined by

$$c = \sqrt{gh}. \tag{1.12}$$

Since $h > 0$, we have $\lambda_-(U) \in \mathbb{R}$ and $\lambda_+(U) \in \mathbb{R}$. In addition, the eigenvalues of $A(U)$ satisfy $\lambda_-(U) < \lambda_+(U)$. Therefore, for $h > 0$, the shallow-water system is strictly hyperbolic, since its eigenvalues are real and distinct.

Nature of the characteristic fields

Next, the nature of the characteristic fields associated to the hyperbolic problem (1.10) is studied. This study involves the computation of the eigenvectors $R_-(U)$ and $R_+(U)$ of the Jacobian matrix $A(U)$. The nature of the characteristic fields is given by the following definition.

Definition 1.1. Let $C_\pm(U) := \nabla_U \lambda_\pm(U) \cdot R_\pm(U)$. The following three cases arise:

1. if $C_\pm(U) \neq 0$ for all $U \in \Omega_U$, then the characteristic field associated to the eigenvalue $\lambda_\pm(U)$ is *Genuinely NonLinear* (GNL);
2. if $C_\pm(U) = 0$ for all $U \in \Omega_U$, then the characteristic field associated to $\lambda_\pm(U)$ is *Linearly Degenerate* (LD);
3. otherwise, we cannot conclude on the nature of the characteristic field.

We now determine the nature of the characteristic fields of the shallow-water equations. The eigenvectors associated to $\lambda_\pm(U)$ are given by:

$$R_\pm(U) = \begin{pmatrix} \pm h \\ \sqrt{gh} \end{pmatrix}.$$

As a consequence, the quantity $C_\pm(U)$ satisfies:

$$C_\pm(U) = \frac{3}{2}\sqrt{gh}.$$

Since $h > 0$, we have proven that $C_\pm(U) \neq 0$ for all $U \in \Omega_U$, and that both fields are GNL.

Riemann invariants

To conclude the study of the algebraic properties of the homogeneous shallow-water equations, we turn to computing the *Riemann invariants*. These quantities are constant in specific cases, described in the next section. They are functions $\Phi(U)$, governed by the following equation:

$$\nabla_U \Phi(U) \cdot R(U) = 0. \tag{1.13}$$

In the present context, since both components of the eigenvectors are nonzero, (1.13) can be rewritten as follows:

$$\frac{dU^1}{R^1_\pm(U)} = \frac{dU^2}{R^2_\pm(U)}, \tag{1.14}$$

where $R^1_\pm(U) \neq 0$ and $R^2_\pm(U) \neq 0$ are the two components of the eigenvector $R_\pm(U)$, and where U^1 and U^2 are the two components of the vector U of the primitive variables. The above equation is therefore equivalent to:

$$\frac{dh}{\pm h} = \frac{du}{\sqrt{gh}},$$

which yields, after straightforward computations, the following Riemann invariant for the field associated to $\lambda_\pm(U)$:

$$u \mp 2\sqrt{gh}. \tag{1.15}$$

1.1.2 Riemann problem

Now, we consider a Riemann problem for the shallow-water equations. It is a Cauchy problem with discontinuous initial data, as follows:

$$\begin{cases} \partial_t W + \partial_x F(W) = 0, \\ W(0,x) = \begin{cases} W_L & \text{if } x < 0, \\ W_R & \text{if } x > 0, \end{cases} \end{cases} \tag{1.16}$$

where $W \in \Omega$ and $F(W)$ are given by (1.2). The initial data is made of two constant states W_L and W_R, respectively defined by $W_L = {}^t(h_L, q_L)$ and $W_R = {}^t(h_R, q_R)$. We assume that $h_L \neq h_R$, or $q_L \neq q_R$, or both $h_L \neq h_R$ and $q_L \neq q_R$. Otherwise, the initial condition is constant, and it stays solution to (1.16) for all $t > 0$. Introducing the left and right velocities u_L and u_R, the discharges q_L and q_R satisfy $q_L = h_L u_L$ and $q_R = h_R u_R$.

The configuration of the exact solution of the Riemann problem is displayed on (1.2). In particular, this Riemann solution is self-similar, i.e. it only depends on x/t instead of the individual variables x and t.

We know that the shallow-water system is hyperbolic and admits two GNL characteristic fields. Therefore, the exact solution to the Riemann problem (1.16) possesses two waves, the first one associated to the eigenvalue λ_-, and the second one associated to λ_+. These two waves will henceforth be referred to as the *1-wave* and the *2-wave*. Since both fields are GNL, each of these two waves may either be a *shock wave* or a *rarefaction wave*. On the one hand, a shock wave connects two constant states with a single jump discontinuity, and the Rankine-

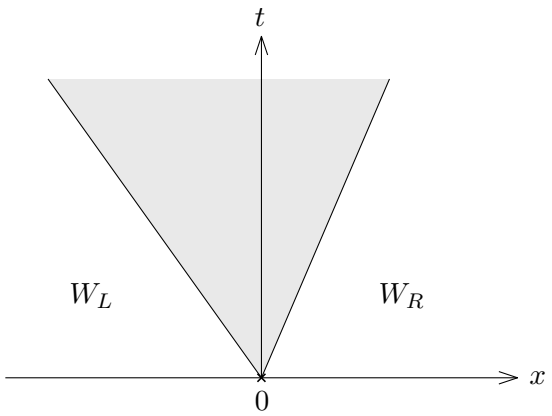

Figure 1.2 – Riemann problem configuration. The gray area represents the area where the solution of the Riemann problem (2.19) lies.

Hugoniot relations are satisfied (see [79, 150] for instance). These relations are proven in Appendix A in a more general setting, and they are given by

$$\sigma(W_R - W_L) = F(W_R) - F(W_L), \tag{1.17}$$

where σ is the velocity of the discontinuity. In the context of the shallow-water equations, the Rankine-Hugoniot relations read (see [112] for instance):

$$\begin{cases} \sigma[h] = [q], \\ \sigma[q] = \left[\dfrac{q^2}{h} + \dfrac{1}{2}gh^2\right], \end{cases} \tag{1.18}$$

where $[X] = X_R - X_L$ represents the jump of the quantity X across the discontinuity. On the other hand, a rarefaction wave connects the two constant states with a continuous function. Within a rarefaction wave, the Riemann invariants (1.15) are constant. Between these two waves, the Riemann solution is constant, and is denoted W_*. The structure of such a solution is displayed on Figure 1.3.

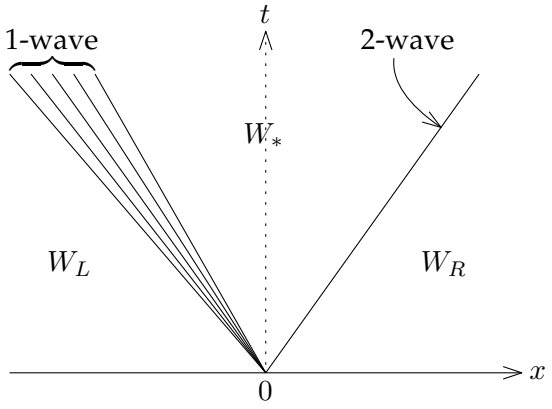

Figure 1.3 – Riemann problem for the shallow-water equations, in the case where the 1-wave is a rarefaction wave and the 2-wave is a shock wave.

The goal of the present section is to provide the solution to the Riemann problem (1.16) with respect to the values of W_L and W_R. The following lemmas state sufficient conditions for

the 1-wave or the 2-wave to be either rarefaction waves or entropy-satisfying shock waves. The entropy condition we impose on the shock wave is a sufficient condition for the entropy inequality (1.5) to be satisfied (see [107, 79] for instance).

Lemma 1.2. *The 1-wave of the Riemann problem (1.16) is a* rarefaction wave *if the Riemann invariants (1.15) are constant across the wave and if the eigenvalues are ordered, as follows:*

$$\begin{cases} \lambda_-(U_L) \leq \lambda_-(U_*), \\ u_L + 2\sqrt{gh_L} = u_* + 2\sqrt{gh_*}. \end{cases}$$

Similarly, the 2-wave is a rarefaction wave if:

$$\begin{cases} \lambda_+(U_*) \leq \lambda_+(U_R), \\ u_* - 2\sqrt{gh_*} = u_R - 2\sqrt{gh_R}. \end{cases}$$

Lemma 1.3. *The 1-wave of the Riemann problem (1.16) is an* entropy-satisfying shock wave *with velocity σ if it satisfies the Lax entropy condition and the Rankine-Hugoniot relations (1.18). Namely, the following relations have to be satisfied:*

$$\begin{cases} \lambda(U_*) \leq \sigma \leq \lambda(U_L), \\ \sigma(h_* - h_L) = h_*u_* - h_Lu_L, \\ \sigma(h_*u_* - h_Lu_L) = h_*u_*^2 + \dfrac{1}{2}gh_*^2 - h_Lu_L^2 - \dfrac{1}{2}gh_L^2. \end{cases}$$

Similarly, the 2-wave is an entropy-satisfying shock wave if:

$$\begin{cases} \lambda(U_R) \leq \sigma \leq \lambda(U_*), \\ \sigma(h_R - h_*) = h_Ru_R - h_*u_*, \\ \sigma(h_Ru_R - h_*u_*) = h_Ru_R^2 + \dfrac{1}{2}gh_R^2 - h_*u_*^2 - \dfrac{1}{2}gh_*^2. \end{cases}$$

Thanks to these lemmas, the following result holds.

Proposition 1.4. *The natures of the waves of the Riemann problem (1.16) are given as follows.*

- *The 1-wave is:*
  - *a rarefaction wave if $h_* \leq h_L$, $u_* \geq u_L$ and $u_* = u_L - 2\sqrt{g}\big(\sqrt{h_*} - \sqrt{h_L}\big)$;*
  - *a shock wave if $h_* \geq h_L$, $u_* \leq u_L$ and $u_* = u_L - \sqrt{\dfrac{g}{2}\left(\dfrac{1}{h_L} + \dfrac{1}{h_*}\right)}(h_* - h_L)$.*
- *The 2-wave is:*
  - *a rarefaction wave if $h_* \leq h_R$, $u_* \leq u_R$ and $u_* = u_R - 2\sqrt{g}\big(\sqrt{h_R} - \sqrt{h_*}\big)$;*
  - *a shock wave if $h_* \geq h_R$, $u_* \geq u_R$ and $u_* = u_R - \sqrt{\dfrac{g}{2}\left(\dfrac{1}{h_*} + \dfrac{1}{h_R}\right)}(h_R - h_*)$.*

Proof. The proof of this result relies on using Lemma 1.2 and Lemma 1.3 to determine the necessary conditions for the nature of each wave. The computations involved are straightforward but quite tedious, and we do not write them here; the reader is referred to [112] for instance. □

The unknown intermediate state $W_* = {}^t(h_*, h_*u_*)$ can now be computed by arguing Proposition 1.4 to exhibit the two relevant relations linking W_* to the known states W_L and W_R. However, this computation cannot be done analytically in the general case, and a root-finding algorithm, such as Newton's method, is required. In addition, the knowledge of this value does not provide enough information to get the full Riemann problem solution. Indeed, the velocities of the shock wave and the rarefaction wave still have to be determined. We also need to provide a value of both h and u within the fan of a rarefaction wave.

The velocity of the shock wave is given by the Rankine-Hugoniot conditions (1.18). Therefore, if the 1-wave is a shock wave, then its velocity σ_1 is given as follows:

$$\sigma_1 = \frac{q_* - q_L}{h_* - h_L}. \tag{1.19}$$

Similarly, if the 2-wave is a shock wave, then its velocity σ_2 is given by:

$$\sigma_2 = \frac{q_R - q_*}{h_R - h_*}. \tag{1.20}$$

Regarding the rarefaction waves, we introduce the notion of *head* and *tail* of the wave. The head of a rarefaction wave is the part of the fan that travels the fastest, while its tail is the part that travels the slowest. If the 1-wave is a rarefaction wave, recall from Lemma 1.2 that $\lambda_-(U_L) \leq \lambda_-(U_*)$. In this case, the head of the wave travels at the velocity $\lambda_-(U_L)$ and its tail travels at $\lambda_-(U_*)$. Similarly, if the 2-wave is a rarefaction wave, then we have $\lambda_+(U_*) \leq \lambda_+(U_R)$: the head of the wave travels at $\lambda_+(U_R)$, while its tail travels at $\lambda_+(U_*)$. See Figure 1.4 for a Riemann problem where the 1-wave is a rarefaction wave and the 2-wave is a shock wave.

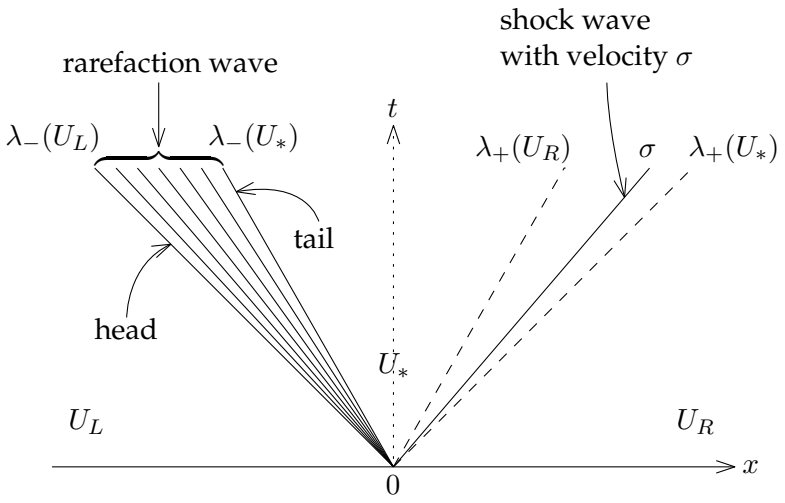

Figure 1.4 – Riemann problem for the shallow-water equations, in the case where the 1-wave is a rarefaction wave and the 2-wave is a shock wave. The wave speeds are displayed.

To achieve the full determination of the Riemann problem solution, we conclude by computing the value of the solution within the fan formed by the rarefaction wave. This solution is self-similar, i.e. it only depends on $\xi := x/t$. Note that, in the fan, the information $U = {}^t(h, u)$ travels with the speed $\lambda_-(U)$ for a 1-wave and $\lambda_+(U)$ for a 2-wave. Therefore, within the fan, we have $\xi = x/t = \lambda_\pm(U)$. In addition, recall that the Riemann invariants (1.15) are constant

in this fan. The combination of these two statements allow to uniquely determine the value of U within the fan. If the 1-wave is a rarefaction wave, the value of $U(\xi)$ within the fan, denoted by $U_1 = {}^t(h_1, u_1)$, reads as follows:

$$\begin{cases} h_1(\xi) = \dfrac{1}{9g}\Big(u_L + 2\sqrt{gh_L} - \xi\Big)^2, \\ u_1(\xi) = \xi + \sqrt{gh_1(\xi)}. \end{cases} \tag{1.21}$$

Similarly, for a 2-wave, the value of $U_2(\xi)$ is defined by:

$$\begin{cases} h_2(\xi) = \dfrac{1}{9g}\Big(u_R - 2\sqrt{gh_R} - \xi\Big)^2, \\ u_2(\xi) = \xi - \sqrt{gh_2(\xi)}. \end{cases} \tag{1.22}$$

We conclude this section by presenting four examples of exact Riemann solutions. The first solution is made of two rarefaction waves, i.e. both the 1-wave and the 2-wave are rarefaction waves. The second one is made of two shock waves, while the third one is a dam-break solution, with the 1-wave being a rarefaction wave and the 2-wave being a shock wave. The fourth and last one deals with the degenerate case of a dry dam-break, where the water height of either the left or right state is zero.

Two-rarefaction case

The initial data of the first example is given as follows:

$$U_L = \begin{pmatrix} h_L \\ u_L \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \end{pmatrix} \quad \text{and} \quad U_R = \begin{pmatrix} h_R \\ u_R \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \tag{1.23}$$

Physically speaking, this initial condition represents a body of water with uniform height, but with two streams of water moving away from one another. Hence, the exact solution consists in two rarefaction waves, linking the left and right states to an intermediate state (computed using Proposition 1.4). Note that, in this specific case of two rarefaction waves, the intermediate state U_* can be computed exactly, to get:

$$\begin{cases} h_* = \dfrac{1}{16g}\Big(u_L - u_R + 2\Big(\sqrt{gh_L} + \sqrt{gh_R}\Big)\Big)^2, \\ u_* = \dfrac{1}{2}(u_L + u_R) - \sqrt{g}\Big(\sqrt{h_R} - \sqrt{h_L}\Big). \end{cases}$$

Equipped with this intermediate state, the exact solution of the Riemann problem (1.16) with the initial condition (1.23) is given as follows, with $\xi = x/t$:

$$U_{\mathcal{R}}(\xi; U_L, U_R) = \begin{cases} U_L & \text{if } \xi < \lambda_-(U_L), \\ U_1(\xi) & \text{if } \lambda_-(U_L) < \xi < \lambda_-(U_*), \\ U_* & \text{if } \lambda_-(U_*) < \xi < \lambda_+(U_*), \\ U_2(\xi) & \text{if } \lambda_+(U_*) < \xi < \lambda_+(U_R), \\ U_R & \text{if } \xi > \lambda_+(U_R), \end{cases} \tag{1.24}$$

where $U_1(\xi)$ and $U_2(\xi)$ are respectively given by (1.21) and (1.22), and the eigenvalues $\lambda_\pm(U)$ are defined by (1.11). The exact height and velocity are displayed on Figure 1.5 for $t = 0.1$s and $x \in [-1, 1]$. In addition, we display the water height and the velocity in the (x, t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$ on Figure 1.6 and Figure 1.7, respectively.

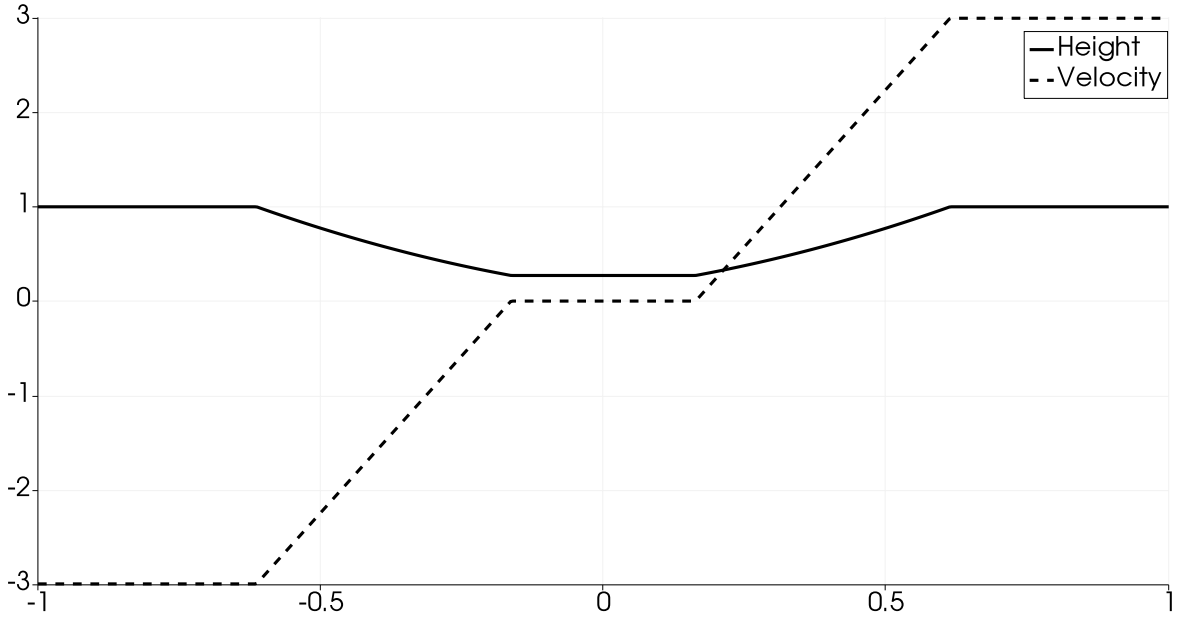

Figure 1.5 – Exact solution (1.24) of the Riemann problem (1.16) – (1.23) at time $t = 0.1$s. This solution is made of two rarefaction waves.

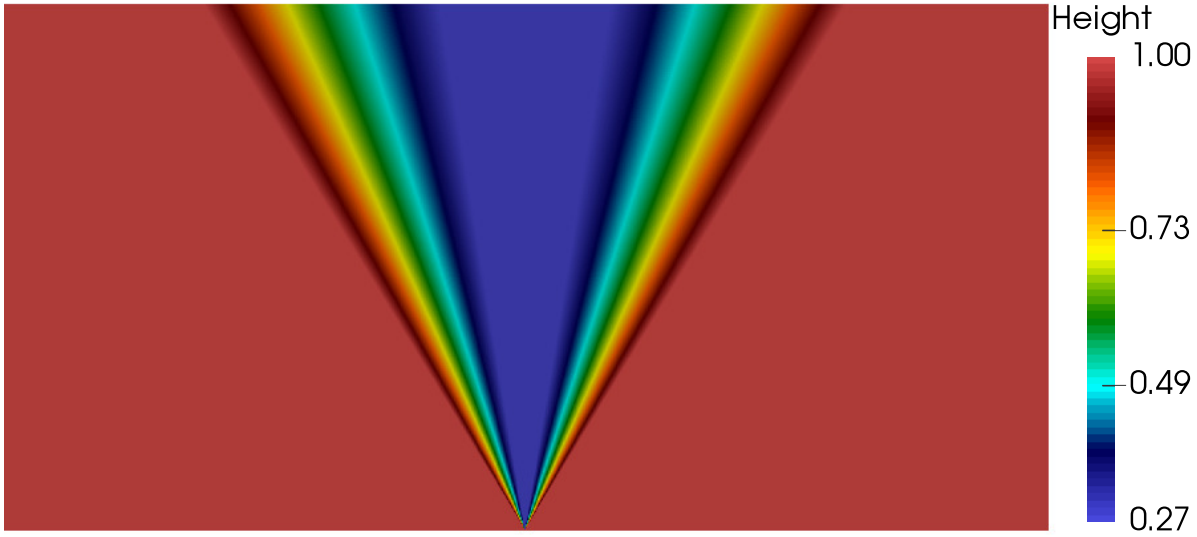

Figure 1.6 – Exact solution (1.24) of the dam-break problem (1.16) – (1.23). Representation of the water height in two space dimensions, in the (x, t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$.

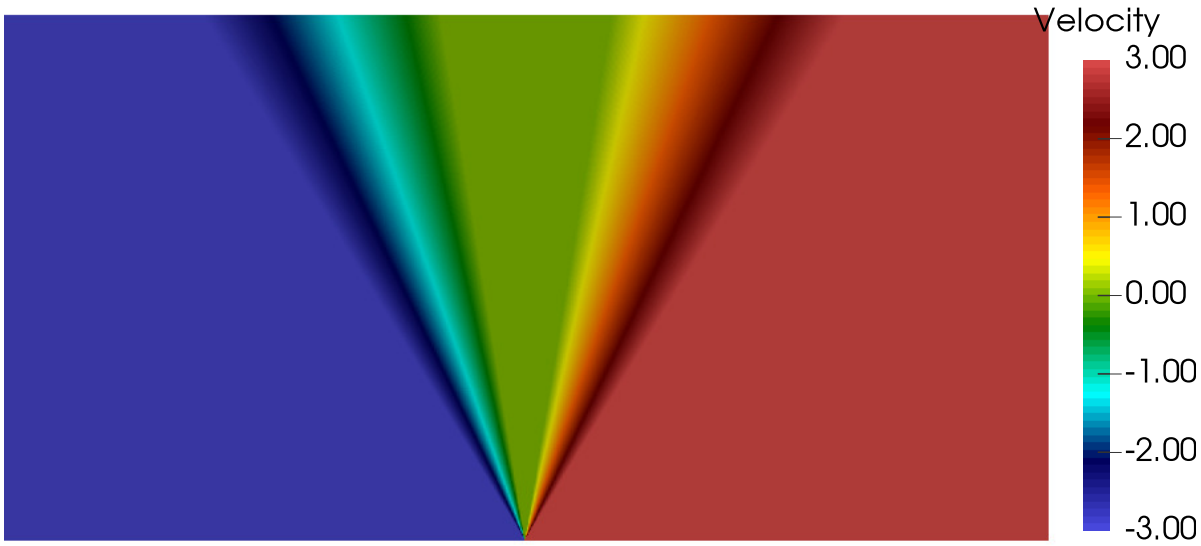

Figure 1.7 – Exact solution (1.24) of the dam-break problem (1.16) – (1.23). Representation of the velocity in two space dimensions, in the (x,t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$.

Two-shock case

Concerning the second example, we take the following initial data:

$$U_L = \begin{pmatrix} h_L \\ u_L \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{and} \quad U_R = \begin{pmatrix} h_R \\ u_R \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}. \tag{1.25}$$

Such initial data produces two discontinuities, since the two streams of water are coming into contact with each other. The exact solution hence involves two shock waves, linking the left state and the right state to the intermediate state U_*, whose computation uses Proposition 1.4. This exact solution is hence given by:

$$U_{\mathcal{R}}(\xi; U_L, U_R) = \begin{cases} U_L & \text{if } \xi < \sigma_1, \\ U_* & \text{if } \sigma_1 < \xi < \sigma_2, \\ U_R & \text{if } \xi > \sigma_2, \end{cases} \tag{1.26}$$

where σ_1 and σ_2 are respectively given by (1.19) and (1.20). This exact solution is displayed on Figure 1.8 for $t = 0.1$s and $x \in [-1, 1]$. The exact solution in the (x,t)-plane, for $t \in [0, 0.1]$ and $x \in [-1, 1]$, is depicted on Figure 1.9 (water height) and on Figure 1.10 (velocity).

Wet dam-break

The third example is a *wet dam-break*. Initially, a large quantity of water is held by a dam to form an artificial lake. At $t = 0$s, the dam breaks, thus liberating the water and making it flow downstream, where a smaller (but nonzero) quantity of water is present. Before the dam breaks, the water is at rest; it starts moving as soon as the dam breaks. Therefore, the

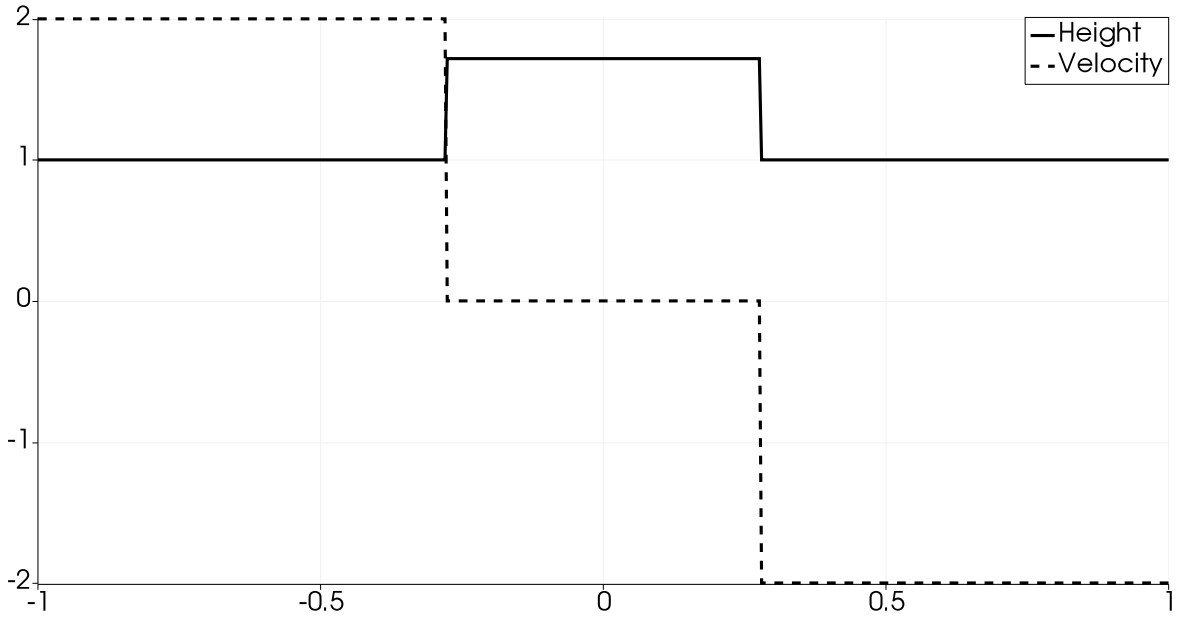

Figure 1.8 – Exact solution (1.26) of the Riemann problem (1.16) – (1.25) at time $t = 0.1$s. This solution is made of two shock waves.

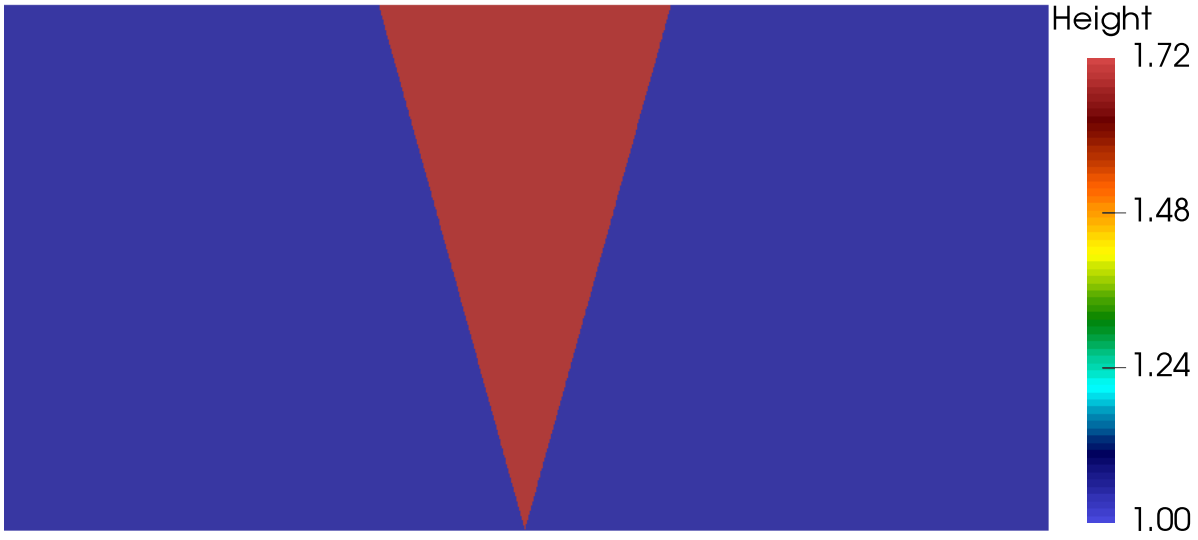

Figure 1.9 – Exact solution (1.26) of the dam-break problem (1.16) – (1.25). Representation of the water height in two space dimensions, in the (x, t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$.

following initial data corresponds to a wet dam-break situation:

$$U_L = \begin{pmatrix} h_L \\ u_L \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \end{pmatrix} \quad \text{and} \quad U_R = \begin{pmatrix} h_R \\ u_R \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{1.27}$$

Applying Proposition 1.4, we show that the solution of the Riemann problem (1.16) with the initial data (1.27) is made of a rarefaction wave traveling left and a shock wave traveling right,

Figure 1.10 – Exact solution (1.26) of the dam-break problem (1.16) – (1.25). Representation of the velocity in two space dimensions, in the (x,t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$.

as follows:

$$U_{\mathcal{R}}(\xi; U_L, U_R) = \begin{cases} U_L & \text{if } \xi < \lambda_-(U_L), \\ U_1(\xi) & \text{if } \lambda_-(U_L) < \xi < \lambda_-(U_*), \\ U_* & \text{if } \lambda_-(U_*) < \xi < \sigma_2, \\ U_R & \text{if } \xi > \sigma_2. \end{cases} \tag{1.28}$$

This structure corresponds to the ones sketched on Figure 1.3 and Figure 1.4. The exact solution is displayed on Figure 1.11 for $t = 0.1$s and $x \in [-1, 1]$. In addition, Figure 1.12 and Figure 1.13 respectively display the exact height and the exact velocity in the (x,t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$. This figure may be compared to Figure 1.3 and Figure 1.4.

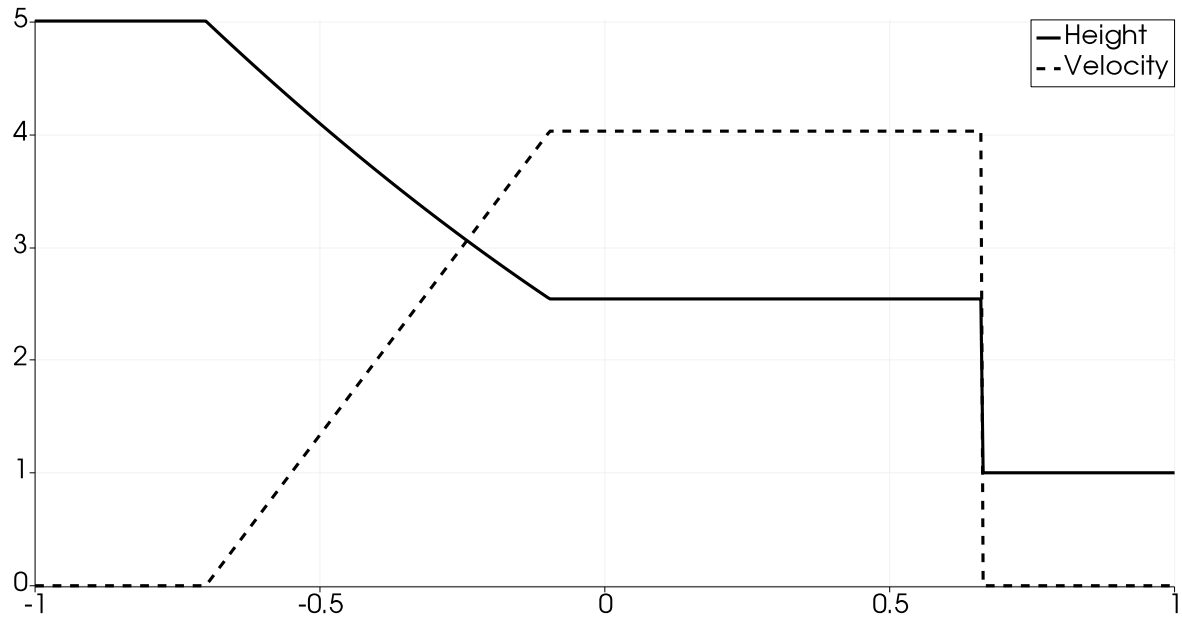

Figure 1.11 – Exact solution (1.28) of the dam-break problem (1.16) – (1.27) at time $t = 0.1$s. This 1-wave is a rarefaction wave and the 2-wave is a shock wave.

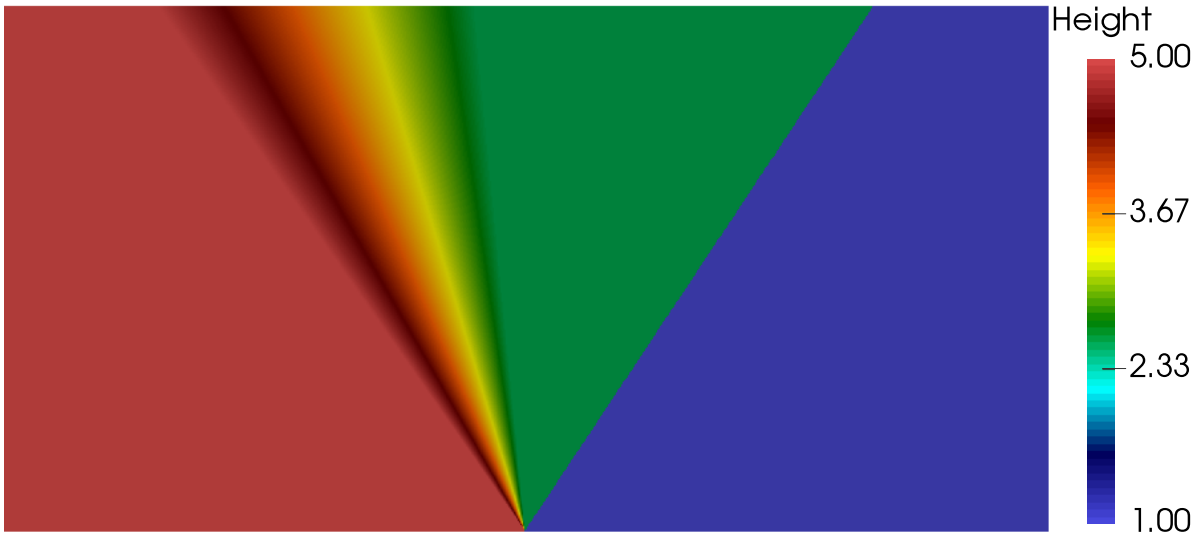

Figure 1.12 – Exact solution (1.28) of the dam-break problem (1.16) – (1.27). Representation of the water height in two space dimensions, in the (x, t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$.

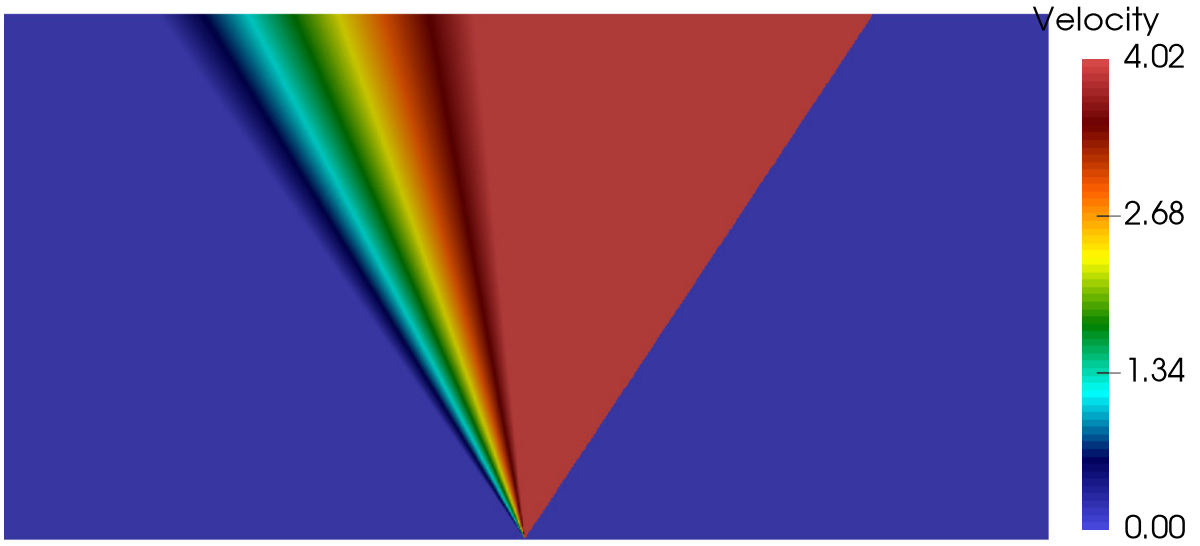

Figure 1.13 – Exact solution (1.28) of the dam-break problem (1.16) – (1.27). Representation of the velocity in two space dimensions, in the (x, t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$.

Dry dam-break

For the fourth and last experiment, we turn to a *dry dam-break*. This experiment consists in considering the degenerate case of a vanishing water height. To achieve a dry dam-break, we take the following initial conditions for the Riemann problem (1.16):

$$U_L = \begin{pmatrix} h_L \\ u_L \end{pmatrix} = \begin{pmatrix} 1.5 \\ 0 \end{pmatrix} \quad \text{and} \quad U_R = \begin{pmatrix} h_R \\ u_R \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{1.29}$$

Such a Riemann problem is solved by taking the limit of the solution to the wet dam-break (i.e. the previous example) when h_R tends to 0. This process is explained in [57], where the authors exhibit Ritter's solution [134]. From [57], the exact solution of the Riemann problem

(1.16) with the initial data (1.29) reads as follows:

$$U_{\mathcal{R}}(\xi; U_L, U_R) = \begin{cases} U_L & \text{if } \xi < -\sqrt{gh_L}, \\ \widehat{U}(\xi) & \text{if } -\sqrt{gh_L} < \xi < 2\sqrt{gh_L}, \\ U_R & \text{if } \xi > 2\sqrt{gh_L}, \end{cases} \tag{1.30}$$

where the intermediate state $\widehat{U}$ is given by:

$$\widehat{U}(\xi) = \begin{pmatrix} \dfrac{4}{9g}\left(\sqrt{gh_L} - \dfrac{\xi}{2}\right)^2 \\ \dfrac{2}{3}\left(\sqrt{gh_L} + \xi\right) \end{pmatrix}.$$

In this case, the 1-wave is a rarefaction wave and the 2-wave is a shock wave. In addition, h_* vanishes, while $u_* \neq 0$. Since $h_* = h_R = 0$, the intermediate water height and the right water height are identical. Hence, for the water height, the 2-wave is a shock wave between the same water heights, and is therefore not visible. On the contrary, $u_* \neq u_R$, so both a rarefaction wave and a shock wave are visible on the water velocity. Finally, one can show that the shock wave travels at the same velocity as the tail of the rarefaction wave. As a consequence, the constant intermediate state U_* is never actually used, as shown by the expression (1.30) of the exact solution.

The exact solution is displayed on Figure 1.14. We also display the exact solution in the (x, t)-plane, on Figure 1.15 and on Figure 1.16, for $t \in [0, 0.1]$ and $x \in [-1, 1]$. To compute the velocity in dry areas, we have assumed that it vanished as soon as the water height vanished.

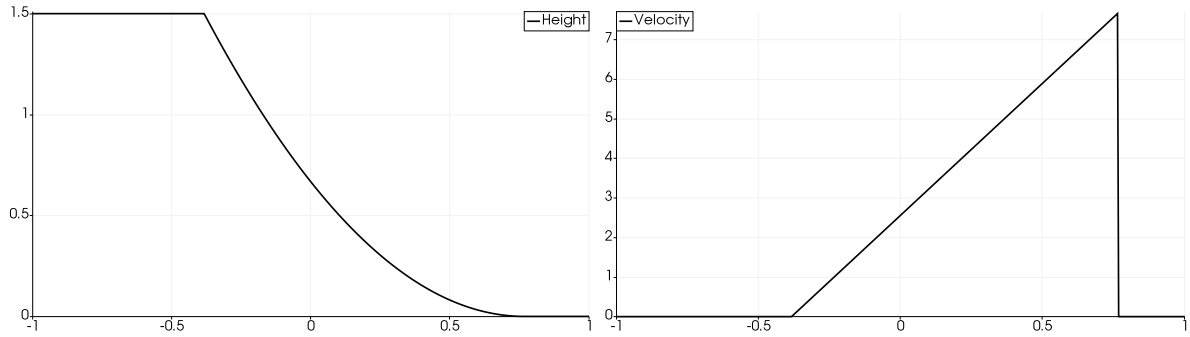

Figure 1.14 – Exact water height (left panel) and exact velocity (right panel) (1.30) of the dam-break problem (1.16) – (1.29) at time $t = 0.1$s. The 1-wave is a rarefaction wave and the 2-wave is a shock wave (not visible for the water height).

1.1.3 Algebraic study of the inhomogeneous system

We now turn to studying the shallow-water model equipped with two source terms, the topography source term and the Manning friction source term. The system is governed by

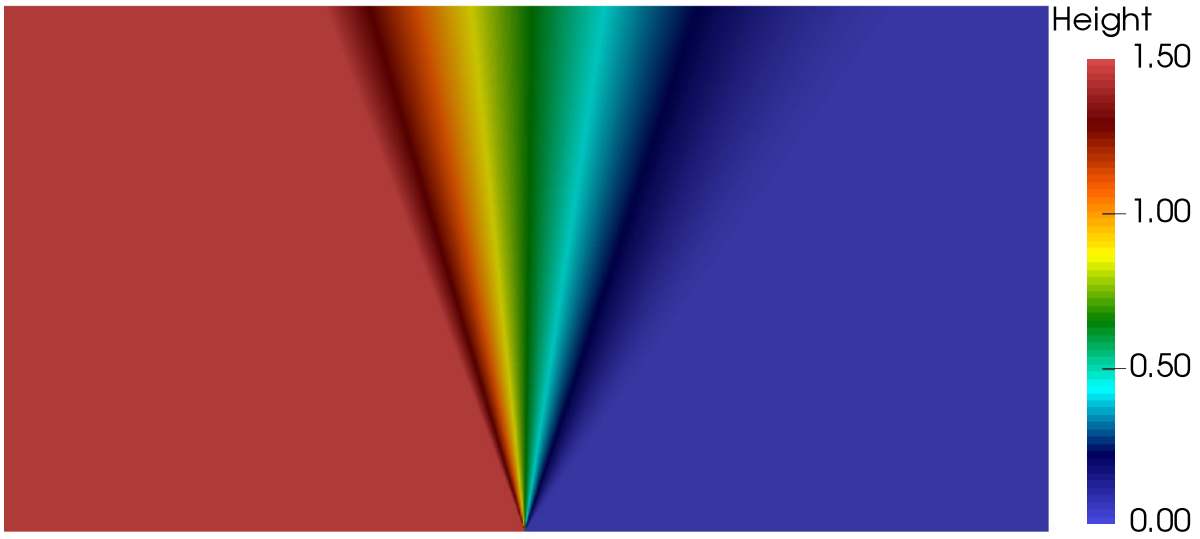

Figure 1.15 – Exact solution (1.30) of the dam-break problem (1.16) – (1.29). Representation of the water height in two space dimensions, in the (x,t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$.

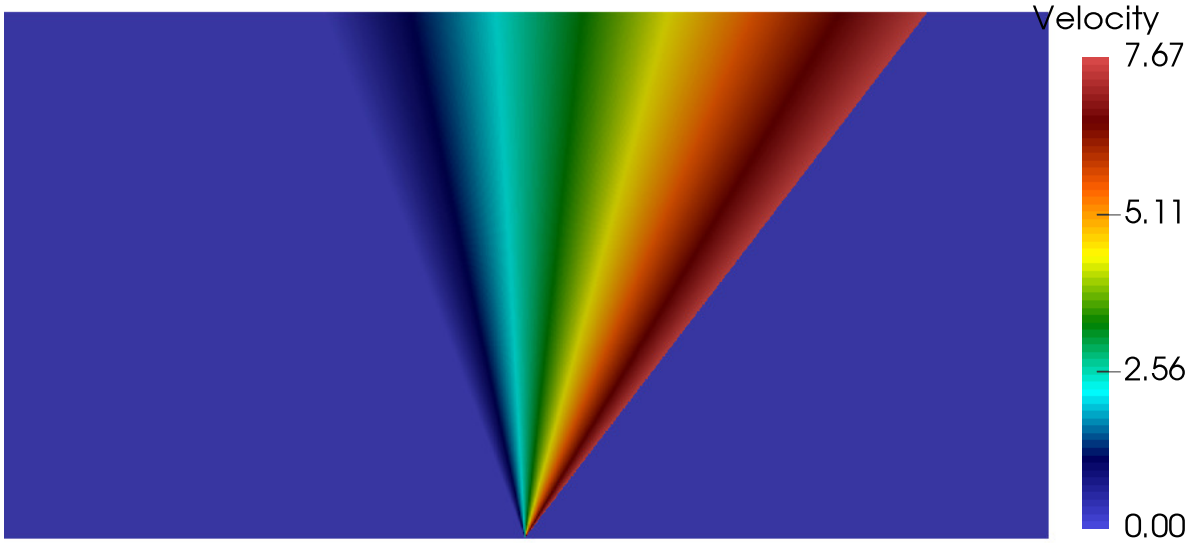

Figure 1.16 – Exact solution (1.30) of the dam-break problem (1.16) – (1.29). Representation of the velocity in two space dimensions, in the (x,t)-plane for $t \in [0, 0.1]$ and $x \in [-1, 1]$.

the equations (1.1), as follows:

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + \frac{1}{2}gh^2\right) = -gh\partial_x Z - k\,u|u|h^{2-\eta}. \end{cases} \tag{1.31}$$

We recall that the smooth function Z represents the shape of the bottom topography and depends only on the space variable x.

Now, we exhibit the eigenvalues of the Jacobian matrix of the system and the nature of the characteristic fields, to determine the consequences of the source terms presence in the equations (see [109, 110, 12] for instance).

A change of variables

To perform the algebraic study of the system, we introduce the function Y such that $Y(x) := x$. Therefore, the function Y satisfies:

$$\partial_x Y = 1 \quad \text{and} \quad \partial_t Y = 0.$$

Regarding the topography function, we also note that

$$\partial_t Z = 0.$$

As a consequence, the shallow-water system with source terms (1.31) rewrites as follows:

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + \dfrac{1}{2}gh^2\right) = -gh\partial_x Z - k\,u|u|h^{2-\eta}\partial_x Y, \\ \partial_t Z = 0, \\ \partial_t Y = 0. \end{cases} \tag{1.32}$$

For smooth solutions and positive water heights (i.e. $h > 0$), the system (1.32) reads:

$$\begin{cases} \partial_t h + u\partial_x h + h\partial_x u = 0, \\ \partial_t u + g\partial_x h + u\partial_x u + g\partial_x Z + k\,u|u|h^{1-\eta}\partial_x Y = 0, \\ \partial_t Z = 0, \\ \partial_t Y = 0. \end{cases} \tag{1.33}$$

Hence, the shallow-water equations (1.31) rewrite under the condensed form

$$\partial_t U + A(U)\partial_x U = 0,$$

where U and $A(U)$ are given by:

$$U = \begin{pmatrix} h \\ u \\ Z \\ Y \end{pmatrix} \quad \text{and} \quad A(U) = \begin{pmatrix} u & h & 0 & 0 \\ g & u & g & k\,u|u|h^{1-\eta} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Hyperbolicity of the system

Straightforward computations show that the matrix $A(U)$ possesses the following eigenvalues:

$$\lambda_-(U) = u - \sqrt{gh} \quad ; \quad \lambda_t(U) = 0 \quad ; \quad \lambda_f(U) = 0 \quad ; \quad \lambda_+(U) = u + \sqrt{gh}. \tag{1.34}$$

Note that 0 is a double eigenvalue and that all four eigenvalues lie in $\mathbb{R}$. To conclude on the hyperbolicity of the system (1.33), we compute the eigenvectors of $A(U)$. The eigenvectors associated to the eigenvalues $\lambda_\pm(U)$ are denoted by $R_\pm(U)$. There are two additional

eigenvectors associated to $\lambda_t(U)$ and $\lambda_f(U)$. They are denoted by $R_t(U)$ and $R_f(U)$, and they respectively correspond to the contributions of the topography and the friction. The eigenvectors are given by:

$$R_\pm(U) = \begin{pmatrix} \pm h \\ \sqrt{gh} \\ 0 \\ 0 \end{pmatrix} \quad , \quad R_t(U) = \begin{pmatrix} -gh \\ gu \\ gh - u^2 \\ 0 \end{pmatrix} \quad \text{and} \quad R_f(U) = \begin{pmatrix} -ku|u|h^{2-\eta} \\ ku^2|u|h^{1-\eta} \\ 0 \\ gh - u^2 \end{pmatrix}. \tag{1.35}$$

These eigenvectors form a basis of $\mathbb{R}^4$ if $u \neq \pm\sqrt{gh}$. Hence, the Jacobian matrix $A(U)$ is diagonalizable in $\mathbb{R}$ if $u \neq \pm\sqrt{gh}$. Therefore, under this condition, the system (1.33) is hyperbolic.

Nature of the characteristic fields

Equipped with the hyperbolicity of the system, the next step in the study of its algebraic properties is the determination of the nature of its characteristic fields.

The eigenvectors $R_\pm(U)$ associated to the eigenvalues $\lambda_\pm(U)$ are defined by (1.35). Using Definition 1.1, the nature of the field associated to $\lambda_\pm(U)$ is given by:

$$\forall U \in \Omega_U,\ \nabla_U \lambda_\pm(U) \cdot R_\pm(U) = \frac{3}{2}\sqrt{gh} \neq 0.$$

Therefore, the characteristic fields associated to $\lambda_\pm(U)$ are GNL. As a consequence, the waves associated to these fields will either be rarefaction waves or shock waves.

For the eigenvalues $\lambda_t(U)$ and $\lambda_f(U)$, note that $\nabla_U \lambda_t(U) = 0$ and $\nabla_U \lambda_f(U) = 0$. Hence, from Definition 1.1, the characteristic fields associated to these eigenvalues are Linearly Degenerate (LD). A linearly degenerate field connects the left and right states with a *contact discontinuity*. This type of discontinuity is governed by the Rankine-Hugoniot conditions, like a shock wave. In addition, Riemann invariants also apply to contact discontinuities: these quantities are studied in the next paragraph.

Riemann invariants

From now on, the waves associated to the eigenvalues $\lambda_-(U)$ and $\lambda_+(U)$ will respectively be labeled the *1-wave* and the *2-wave*. In addition, the waves associated to $\lambda_t(U)$ and $\lambda_f(U)$ will be respectively called the *t-wave* and the *f-wave*. The goal of this paragraph is to study the invariant quantities across the four waves.

First, we study Riemann invariants associated to $\lambda_\pm(U)$. To ensure that these waves do not have a vanishing velocity and that the system is hyperbolic, we assume that $u \neq \pm\sqrt{gh}$. The following result states these Riemann invariants.

Lemma 1.5. *The Riemann invariants for the 1-wave are:*

$$u + 2\sqrt{gh} \quad ; \quad Z \quad ; \quad Y.$$

In addition, the Riemann invariants for the 2-wave are:

$$u - 2\sqrt{gh} \quad ; \quad Z \quad ; \quad Y.$$

Proof. Recall the expression (1.35) of the associated eigenvectors. Since the third and fourth components of $R_\pm(U)$ are zero, the definition (1.13) of the Riemann invariants ensures that both quantities Z and Y are Riemann invariants for the 1-wave and the 2-wave. Then, the third Riemann invariant is governed by (1.14). Straightforward computations prove that this third Riemann invariant is $u + 2\sqrt{gh}$ for the 1-wave and $u - 2\sqrt{gh}$ for the 2-wave. The proof is thus achieved. □

Second, we turn to the t-wave, associated to the eigenvalue $\lambda_t(U) = 0$. For this wave, the Riemann invariants are given by the following result.

Lemma 1.6. *The Riemann invariants for the t-wave, i.e. the wave associated to the topography source term, are given by:*

$$hu \quad ; \quad \frac{u^2}{2} + g(h + Z) \quad ; \quad Y. \tag{1.36}$$

Proof. Recall the expression (1.35) of the eigenvector $R_t(U)$ associated to $\lambda_t(U)$. Note that the fourth component $R_t^4(U)$ of this eigenvector is zero. Therefore, after (1.13), Y is a Riemann invariant for this wave. Now, we determine the other two Riemann invariants. They satisfy:

$$\frac{dU^1}{R_t^1(U)} = \frac{dU^2}{R_t^2(U)} = \frac{dU^3}{R_t^3(U)},$$

or, equivalently,

$$\frac{dh}{-gh} = \frac{du}{gu} = \frac{dZ}{gh - u^2}. \tag{1.37}$$

The first equality of (1.37) rewrites as follows:

$$d(hu) = 0.$$

Hence, the discharge $q = hu$ is a Riemann invariant for t-wave.

A third Riemann invariant is now determined. The second equality of (1.37) rewrites:

$$\left(g - \frac{q^2}{h^3}\right) dh + g dZ = 0.$$

Using the constant discharge, we show that the following quantity is a Riemann invariant for the t-wave:

$$\frac{u^2}{2} + g(h + Z).$$

All three Riemann invariants have been determined, and the proof is achieved. □

Third, we focus on the f-wave, associated to $\lambda_f(U)$. The following result gives the Riemann invariants for this wave.

Lemma 1.7. *The Riemann invariants for the f-wave, i.e. the wave associated to the friction source term, are given by:*

$$q \quad ; \quad g\frac{h^{\eta+2}}{\eta+2} - q^2\frac{h^{\eta-1}}{\eta-1} + kq|q|Y \quad ; \quad Z. \tag{1.38}$$

Proof. The eigenvector $R_f(U)$, associated to the f-wave, is given by (1.35). Note that its third component $R_f^3(U)$ is zero. As a consequence, arguing (1.13) yields that Z is a Riemann invariant for the f-wave. The other Riemann invariants satisfy:

$$\frac{dU^1}{R_f^1(U)} = \frac{dU^2}{R_f^2(U)} = \frac{dU^4}{R_f^4(U)}.$$

The above equalities rewrite as follows:

$$\frac{dh}{-h} = \frac{du}{u} = \frac{ku|u|h^{1-\eta}}{gh-u^2}dY. \tag{1.39}$$

The first equality of (1.39) yields

$$d(hu) = 0.$$

Therefore, a Riemann invariant for the f-wave is the discharge $q = hu$.

Using this Riemann invariant, the second equality of (1.39) rewrites as follows:

$$(gh^{\eta+1} - q^2h^{\eta-2})dh + kq|q|dY = 0.$$

Hence, the last Riemann invariant for this wave is the following:

$$g\frac{h^{\eta+2}}{\eta+2} - q^2\frac{h^{\eta-1}}{\eta-1} + kq|q|Y.$$

The three Riemann invariants for the f-wave have thus been determined, which completes the proof. □

As a consequence of Lemma 1.6 and Lemma 1.7, we notice that the discharge is constant across the stationary contact discontinuity associated to the double eigenvalue 0, since it is a Riemann invariant for both the t-wave and the f-wave.

1.2 Steady state solutions

In the previous section, we have exhibited the algebraic properties of the inhomogeneous shallow-water equations. Now, in this section, we study the steady state solutions, which are specific solutions of the shallow-water system with the source terms of topography and Manning friction whose time derivative is zero.

Recall that the inhomogeneous shallow-water system is governed by (1.1). As a consequence, a solution $W = {}^t(h, q)$ of the shallow-water equations with both topography and

Manning friction is a steady state solution if is satisfies the following identities:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -gh\partial_x Z - kq|q|h^{-\eta}. \end{cases} \tag{1.40}$$

The first equation of (1.40) immediately yields that $q = \mathrm{cst}$. This constant value is denoted, throughout the whole manuscript, by q_0. This very important remark greatly simplifies the study of the steady states, since only the second equation of (1.40) is not trivial.

Therefore, to exhibit the steady states, the second equation of (1.40) is studied in the following three cases:

1. first, in Section 1.2.1, we consider a vanishing friction contribution, by taking $k = 0$;

2. second, in Section 1.2.2, we consider a vanishing topography contribution, by enforcing a flat topography $Z = \mathrm{cst}$, thus ensuring that $\partial_x Z = 0$;

3. third, in Section 1.2.3, we give some comments on the steady state solutions with both friction and topography.

1.2.1 Topography steady states

In this section, we focus on the well-known steady state solutions of the shallow-water equations with topography only (see for instance [44]). Such steady states are obtained by neglecting the friction contribution in (1.40), i.e. by taking $k = 0$. As a consequence, these solutions are governed by the following set of equations:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -gh\partial_x Z. \end{cases} \tag{1.41}$$

Hence, as expected, the steady discharge is constant, and its value is denoted by q_0. Therefore, for $q_0 \in \mathbb{R}$, the topography steady states are completely described by the following equation, which links the unknown water height h to the known topography Z:

$$\partial_x \left(\frac{q_0^2}{h} + \frac{1}{2} g h^2 \right) + gh\partial_x Z = 0. \tag{1.42}$$

The focus of this section is the study of the equation (1.42). This study is well-known, but it is recalled here to introduce several techniques which will be instrumental to the study of the steady state solutions for the friction source term. We begin by deriving smooth steady states for a nonzero water height. Afterwards, we give a word on smooth steady states with a dry area, i.e. an area where $h = 0$. Finally, the case of a discontinuous steady state solution is briefly discussed.

1.2.1.1 Smooth steady states

The smooth steady states with $q_0 = 0$ consist in the *lake at rest* steady state, given by:

$$h + Z = \text{cst}\,. \tag{1.43}$$

This steady state solution is well-known, and it has been widely studied (see for instance [87, 26, 74, 19, 7]).

We now assume that $q_0 \neq 0$ and study the equation (1.42) in order to exhibit the *moving steady states*. For smooth $h > 0$ and smooth Z, (1.42) rewrites as follows:

$$\partial_x \left(\frac{q_0^2}{2h^2} + g(h + Z) \right) = 0, \tag{1.44}$$

which is nothing but a statement of Bernoulli's principle.

We make here an interesting remark. Note that $q_0 = hu$, where both h and u depend on x. As a consequence, the equation (1.44) rewrites:

$$\frac{u^2}{2} + g(h + Z) = \text{cst}\,.$$

This uniform quantity is usually called the *total head* (see [86] for instance); we denote it by $\mathcal{E}$. This equation describes the moving steady state solutions of the shallow-water equations with topography. The same equation, as well as the uniformity of the discharge, described the Riemann invariants (1.36) across the contact discontinuity associated to the topography source term. As a consequence, Bernoulli's principle governs both the steady state solutions and the Riemann invariants. We also remark that the total head $\mathcal{E}$ is closely related to the entropy flux $\widetilde{G}$ defined by (1.8). Indeed, we have $\widehat{G} = q\mathcal{E}$.

We now study the equation (1.44) with respect to h in order to obtain a characterization of the steady water height h associated to a uniform discharge q_0 and a given topography function Z. Such a study is present in [44] in the context of Riemann invariants, and in [123] for steady states. Throughout the rest of this section, h is assumed to be a positive function of x.

Let $x_0 \in \mathbb{R}$. We denote by h_0 and Z_0 the respective values $h(x_0)$ and $Z(x_0)$ of the water height and the topography at the point x_0. Integrating the equation (1.44) on $[x_0, x]$ immediately yields:

$$\frac{q_0^2}{2h^2} + g(h + Z) - \frac{q_0^2}{2h_0^2} - g(h_0 + Z_0) = 0, \tag{1.45}$$

where $h = h(x)$ and $Z = Z(x)$. Recall that the smooth function Z is assumed to be known. Hence, the knowledge of the water height h_0 at the point x_0 and of the uniform discharge q_0 is enough to determine the steady water height h at any point x, provided the equation (1.45) admits at least one solution. Determining the existence and uniqueness of a solution to (1.45), as well as the properties of such a solution, is therefore the focus of the remainder of this section.

For the sake of simplicity in the notations, we now introduce the function ξ, defined as

follows:

$$\xi(h; Z, q_0, h_0, Z_0) := \frac{q_0^2}{2h^2} + g(h + Z) - \frac{q_0^2}{2h_0^2} - g(h_0 + Z_0), \tag{1.46}$$

such that (1.45) rewrites

$$\xi(h; Z, q_0, h_0, Z_0) = 0. \tag{1.47}$$

The function ξ thus depends on the unknown steady water height h at point x, as well as on the parameters Z, q_0, h_0, Z_0, which are assumed to be known.

The study of solutions to (1.47) is now performed. As a first step, we seek the variations of the function $h \mapsto \xi(h; Z, q_0, h_0, Z_0)$. To that end, we differentiate ξ with respect to h, as follows:

$$\frac{\partial \xi}{\partial h}(h; Z, q_0, h_0, Z_0) = -\frac{q_0^2}{h^3} + g. \tag{1.48}$$

As a consequence, the derivative of ξ with respect to h vanishes for $h = h_c$, with

$$h_c = \left(\frac{q_0^2}{g}\right)^{1/3}. \tag{1.49}$$

From (1.48), we deduce the following result.

Lemma 1.8. *With h_c defined by (1.49), the function ξ defined by (1.46) satisfies the following properties:*

- *$h \mapsto \xi(h; Z, q_0, h_0, Z_0)$ is a strictly decreasing function for $h < h_c$;*
- *$h \mapsto \xi(h; Z, q_0, h_0, Z_0)$ is a strictly increasing function for $h > h_c$.*

The function ξ therefore admits a single extremum, located at $h = h_c$. Moreover, this extremum is a minimum.

Now, we determine the existence and uniqueness of solutions to the equation (1.47) (or, equivalently, (1.45)). To that end, we study the sign of ξ. After straightforward computations, the following result is proven.

Lemma 1.9. *The function ξ defined by (1.46) admits the following limits:*

- $\lim_{h \to 0^+} \xi(h; Z, q_0, h_0, Z_0) = +\infty$;
- $\lim_{h \to +\infty} \xi(h; Z, q_0, h_0, Z_0) = +\infty$.

In addition, the following evaluation of ξ at h_c (defined by (1.49)) is verified:

$$\xi_c(Z) := \xi(h_c; Z, q_0, h_0, Z_0) = \frac{q_0^2}{2}\left(\frac{3}{h_c^2} - \frac{1}{h_0^2}\right) + g(Z - Z_0 - h_0).$$

Before combining Lemma 1.8 and Lemma 1.9 to prove a result on the existence of roots of the function ξ, we give examples of situations that may be encountered. For these examples, the quantities take the following values:

- $q_0 = \sqrt{g}$, so that $h_c = 1$;
- $h_0 = h_c = 1$;
- $Z_0 = 0.75$.

As a consequence, Lemma 1.9 yields, after straightforward computations:

$$\xi_c(Z) = g(Z - Z_0).$$

From Lemma 1.8, the function ξ reaches its minimum for $h = h_c$. The above equality shows that $\xi_c(Z) \geq 0$ if and only if $Z \geq Z_0$. Therefore, the number of zeros of the function ξ is tied to the sign of $\xi_c(Z)$. To highlight this property, we display on Figure 1.17 the function ξ for $Z \in \{0.7, 0.75, 0.8\}$.

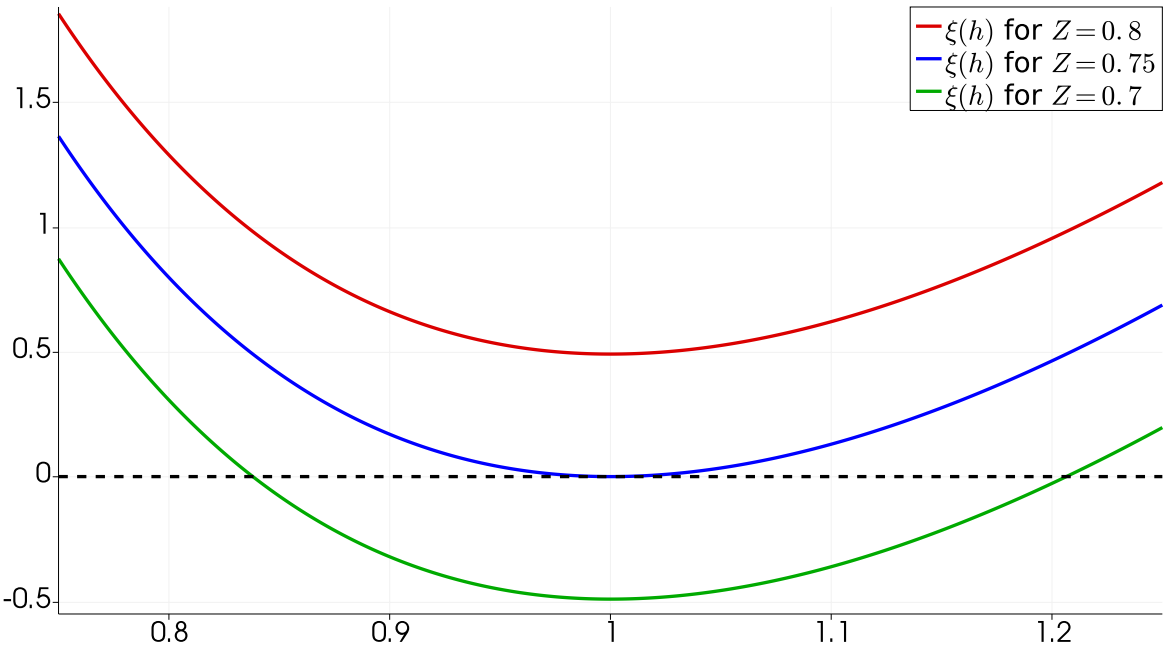

Figure 1.17 – Sketches of $\xi(h; Z, \sqrt{g}, 1, 0.75)$ for $h \in [0.75, 1.25]$ and for different values of Z. Red curve: $Z = 0.8$, no zero for ξ. Blue curve: $Z = 0.75$, unique zero for ξ. Green curve: $Z = 0.7$, two distinct zeros for ξ.

Equipped with Lemma 1.8 and Lemma 1.9, the properties inferred from the example presented in Figure 1.17 are summarized in the following result.

Proposition 1.10. *Assume $h > 0$ and $q_0 \neq 0$. Then, $h_c > 0$ according to (1.49), and following properties hold.*

(i) If $\xi_c(Z) > 0$, then there is no solution to the equation (1.47).

(ii) If $\xi_c(Z) = 0$, then the equation (1.47) admits a unique solution. This solution, $h = h_c$, is a double root of the function $h \mapsto \xi(h; Z, h_0, q_0, Z_0)$.

(iii) If $\xi_c(Z) < 0$, then the equation (1.47) admits two distinct solutions, $h_{sup} \in (0, h_c)$ and $h_{sub} \in (h_c, +\infty)$.

Proof. The proof of this result relies on using the properties of ξ we have obtained above. From Lemma 1.8, the function ξ admits a unique minimum, reached for $h = h_c$. Moreover, from Lemma 1.9, the function ξ tends to infinity as h tends to 0^+ or infinity. Therefore, the number of zeros of ξ depends on the sign of its minimum value $\xi_c(Z) = \xi(h_c; Z, q_0, h_0, Z_0)$.

Equipped with these results, the proofs of (i), (ii) and (iii) are obvious. The proof is thus achieved. □

Note that the three assertions of Proposition 1.10 respectively correspond to the red, blue and green curves of Figure 1.17.

In the third assertion of Proposition 1.10, we have labeled the two solutions as h_{sup} and h_{sub}. These denominations are connected to the Froude number, defined below.

Definition 1.11. The *Froude number* is a dimensionless quantity defined by:

$$\mathrm{Fr} = \frac{|u|}{c}, \tag{1.50}$$

where u is the velocity of the water and $c = \sqrt{gh}$ is the sound speed. On the one hand, the flow is called *supercritical* (or *torrential*) if $\mathrm{Fr} > 1$, which corresponds to a large water velocity and/or a small water height. On the other hand, it is called *subcritical* (or *fluvial*) if $\mathrm{Fr} < 1$, i.e. for a small water velocity and/or a large water height.

In the current context of a steady state solution, the Froude number reads:

$$\mathrm{Fr} = \frac{|q_0|}{\sqrt{gh^3}} = \left(\frac{h_c}{h}\right)^{3/2}. \tag{1.51}$$

As a consequence, $h < h_c$ corresponds to a supercritical flow, while $h > h_c$ corresponds to a subcritical flow. The quantity h_c is hence called the *critical height*. This remark is the basis for the notations h_{sup} and h_{sub} introduced in Proposition 1.10 to label the two roots of the function ξ, since $h_{sup} \in (0, h_c)$ is a supercritical solution and $h_{sub} \in (h_c, +\infty)$ is a subcritical solution.

Equipped with Proposition 1.10 and the above remark, we state the following corollary of Proposition 1.10.

Corollary 1.12. *Assume $h > 0$ and $q_0 \neq 0$. Thus, $h_c > 0$ according to (1.49). Let Z_c be a critical topography value, given by:*

$$Z_c = Z_0 + h_0 + \frac{h_c}{2}\left(\frac{h_c^2}{h_0^2} - 3\right).$$

The following properties, concerning the solutions of (1.47), hold.

(i) If $Z > Z_c$, then there is no solution to the equation (1.47).

(ii) If $Z = Z_c$, then the equation (1.47) admits a unique solution. This solution, $h = h_c$, is a double root of the function $h \mapsto \xi(h; Z, h_0, q_0, Z_0)$.

(iii) If $Z < Z_c$, then the equation (1.47) admits two distinct solutions, $h_{sup} \in (0, h_c)$ and $h_{sub} \in (h_c, +\infty)$.

Proof. This result directly follows from noting that $\xi_c(Z) > 0$ if and only if $Z > Z_c$. Thus, the proof is achieved by invoking Proposition 1.10. □

Remark 1.13. Assume that $h > 0$ and $h \neq h_c$. From (1.42), the derivative of h with respect to x reads

$$\partial_x h = \frac{h^3 \partial_x Z}{h_c^3 - h^3}.$$

Therefore, if the solution is subcritical, i.e. $h > h_c$, then the sign of $\partial_x h$ is the opposite of the sign of $\partial_x Z$, whereas the sign of $\partial_x h$ is that of $\partial_x Z$ if the solution is supercritical. These

results are in accordance with the subcritical and supercritical experiments presented in [86] for instance.

Both Proposition 1.10 and Corollary 1.12, as well as Remark 1.13, are illustrated with the following example. This example consists in exhibiting a solution to (1.47) in the context of Figure 1.17, i.e. with $q_0 = \sqrt{g}$, $h_0 = 1$ and $Z(x_0) = 0.75$. Note that, with this value of q_0, we have $h_0 = h_c = 1$, and Corollary 1.12 yields $Z_c = Z_0 = 0.75$. Hence, a steady state will exist if and only if $Z(x) \leq 0.75$. In order to check this property, we take the following topography function:

$$Z(x) = \frac{1}{4} + \cos^2\Big(\pi(x - x_0) + \frac{\pi}{4}\Big).$$

With $x_0 = 0.75$, this topography function satisfies the required property that $Z(x_0) = 0.75$. Equipped with the topography function, we now solve (1.47), using Newton's method, for $x \in [0.65, 1.35]$. The results are displayed on Figure 1.18.

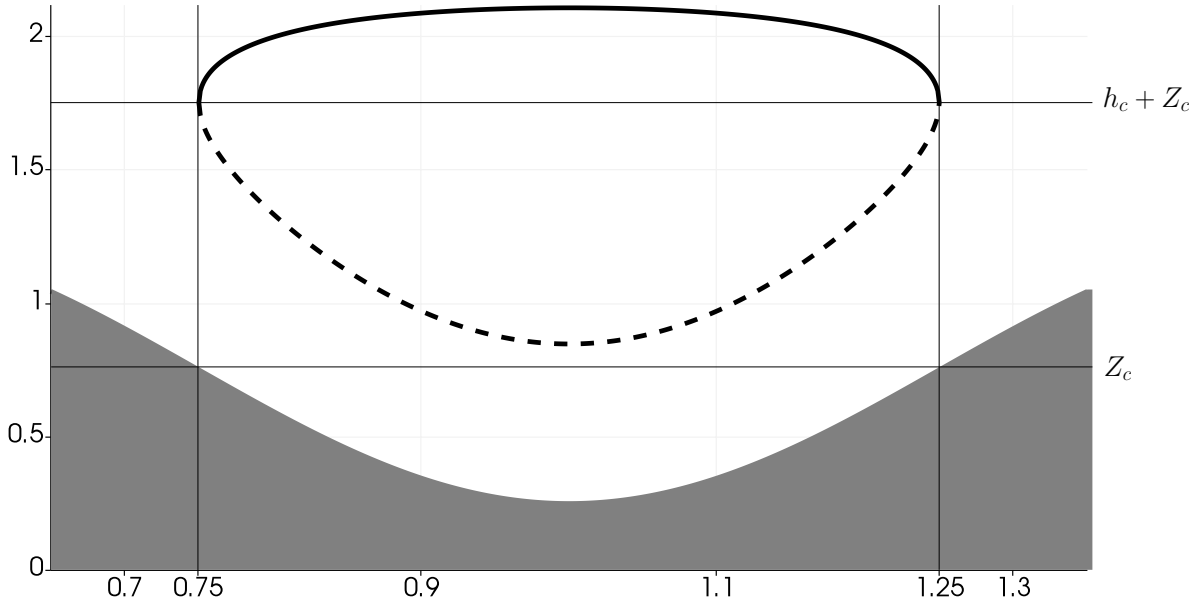

Figure 1.18 – Solutions $h(x)$ of (1.47) (where they exist). Full line: subcritical solution. Dotted line: supercritical solution. Gray area: topography.

For $x \in [0.65, 0.75) \cup (1.25, 1.35]$, we have $Z(x) > Z_c$. Hence, after Proposition 1.10 and Corollary 1.12, the equation (1.47) does not admit a solution on this domain. The results presented on Figure 1.18 are in good agreement with this conclusion. In addition, the topography is decreasing for $x < 1$, and increasing for $x > 1$. As expected, both the supercritical and subcritical solutions exhibit the behavior predicted by Remark 1.13.

We have thus completed the study, for the topography source term only, of smooth steady state solutions with positive water heights. More examples of steady state solutions will be provided in Chapter 3.

1.2.1.2 Case of a dry area

To complete the determination of smooth steady state solutions for the topography source terms, we now turn to the study of steady states involving dry areas, i.e. areas where the

water height is zero. The following result characterizes a steady state solution where a dry area is present.

Proposition 1.14. *As soon as a dry area is involved, smooth steady states must be at rest.*

Proof. We begin by defining the kinetic energy in a wet area where $h > 0$, as follows:

$$E = \frac{1}{2}\frac{q^2}{h}.$$

Since we inject a bounded quantity of energy at the initial time, the kinetic energy has to be bounded, i.e.

$$\|E\|_\infty < +\infty.$$

The above formula yields

$$\lim_{h\to 0} \frac{q^2}{h} = \lim_{h\to 0} E < +\infty.$$

As a consequence, we necessarily have $q = \mathcal{O}(\sqrt{h})$ when h tends to 0^+. Thus, we immediately obtain that, if there is some $x_D \in \mathbb{R}$ such that $h(x_D) = 0$, then $q(x_D) = 0$. Now, recall from (1.41) that, for a steady state, $\partial_x q = 0$. Therefore, for all $x \in \mathbb{R}$, $q(x) = q(x_D) = 0$, i.e. the water is at rest. We conclude that, as soon as a smooth steady state solution involves a dry area, this steady state must be at rest. □

This situation of a steady state with a dry area is displayed on Figure 1.19.

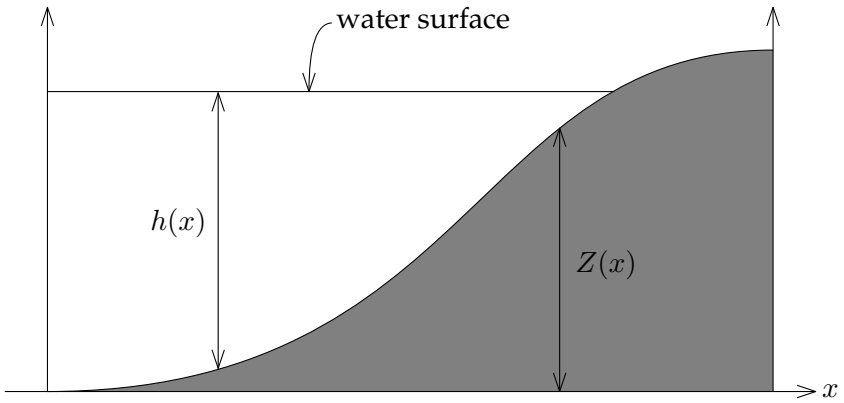

Figure 1.19 – Steady state solution with a dry area. The gray area is the topography.

1.2.1.3 Discontinuous steady state solutions

Finally, we give a few words on discontinuous steady state solutions. Such solutions are piecewise smooth functions W, whose smooth pieces verify the Bernoulli relation (1.44), and whose discontinuities satisfy the Rankine-Hugoniot relations (see Appendix A), as well as en entropy condition. We assume that the topography function Z is also piecewise smooth.

In the present context, the Rankine-Hugoniot relations (A.2) read:

$$\begin{cases} \sigma[h] = [q], \\ \sigma[q] = \left[\dfrac{q^2}{h} + \dfrac{1}{2}gh^2\right]. \end{cases}$$

However, note that the discontinuities necessarily have to be stationary for a steady solution. Therefore, their velocity σ, present in the Rankine-Hugoniot relations, vanishes. As a consequence, the Rankine-Hugoniot relations rewrite as follows:

$$\begin{cases} [q] = 0, \\ \left[\dfrac{q^2}{h} + \dfrac{1}{2}gh^2\right] = 0. \end{cases}$$

In particular, the discharge is constant across the discontinuity.

The entropy inequality, in the presence of the topography source term only, is given by (1.7) (see [5, 12] for instance). In the current context of a steady state solution, the entropy inequality (1.7) reads:

$$\partial_x \widetilde{G}(W, Z) \leq 0,$$

with $\widetilde{G}(W, Z)$ given by (1.8).

As a consequence, for a piecewise smooth topography function Z, a piecewise smooth steady state solution $W = {}^t(h, q)$ satisfies the following properties (see [12]):

- the discharge $q = q_0$ is uniform throughout the domain;
- if h and Z are smooth, then the Bernoulli relation $\partial_x\left(\dfrac{q_0^2}{2h^2} + g(h + Z)\right) = 0$ is verified;
- across a discontinuity of h or Z, the water height satisfies the following two relations:
  - the jump relation $\left[\dfrac{q_0^2}{h} + \dfrac{1}{2}gh^2\right] = 0$;
  - the discrete entropy inequality $\left[q_0\left(\dfrac{q_0^2}{2h^2} + g(h + Z)\right)\right] \leq 0$.

Such an approach has been used in [86] for the shallow-water equations with topography, to define the *transcritical flow with shock* steady state solution.

1.2.2 Friction steady states

After having exhibited steady state solutions for the topography source term in the previous section, we now turn to steady state solutions for the Manning friction source term only. To that end, we take $k \neq 0$ and a flat topography, i.e. $\partial_x Z = 0$. As a consequence, the steady state solutions are governed by the following set of equations:

$$\begin{cases} \partial_x q = 0, \\ \partial_x\left(\dfrac{q^2}{h} + \dfrac{1}{2}gh^2\right) = -kq|q|h^{-\eta}. \end{cases} \tag{1.52}$$

From the first equation of (1.52), we recover that the discharge must be uniform, and we label this uniform discharge q_0, as usual. Equipped with this notation, only the second equation governs the water height for steady state solutions, as follows:

$$\partial_x\left(\frac{q_0^2}{h} + \frac{1}{2}gh^2\right) = -kq_0|q_0|h^{-\eta}. \tag{1.53}$$

Note that, if $q_0 = 0$, the friction source term vanishes, and (1.53) implies that $h = \text{cst}$. This behavior is to be expected. Indeed, if the water is at rest, there is no bottom friction. Therefore, from now on, we only consider steady states with $q_0 \neq 0$, i.e. moving steady states. First, smooth steady states are studied. Then, we suggest a way to define steady states with a jump discontinuity.

1.2.2.1 Study of smooth steady states

The goal of this section is to study the equation (1.53) with $k \neq 0$, $q_0 \neq 0$ and assuming a positive smooth water height $h(x) > 0$. In this case, the equation (1.53) rewrites as follows:

$$\left(-q_0^2 h^{\eta-2} + g h^{\eta+1}\right)\partial_x h + k q_0 |q_0| = 0, \tag{1.54}$$

which in turn yields:

$$\partial_x \left(-q_0^2 \frac{h^{\eta-1}}{\eta-1} + g \frac{h^{\eta+2}}{\eta+2} + k q_0 |q_0| x \right) = 0. \tag{1.55}$$

Note that we once again recover the Riemann invariant given by (1.38). Now, we set $x_0 \in \mathbb{R}$, and we introduce the notation $h_0 = h(x_0)$. The relation (1.55) is then integrated on (x_0, x), to get:

$$\frac{-q_0^2}{\eta-1}\left(h^{\eta-1} - h_0^{\eta-1}\right) + \frac{g}{\eta+2}\left(h^{\eta+2} - h_0^{\eta+2}\right) + k q_0 |q_0| (x - x_0) = 0. \tag{1.56}$$

The solutions h of (1.56) represent the water height for steady state solutions. To shorten the notations, we define

$$\chi(h; x, q_0, x_0, h_0) := \frac{-q_0^2}{\eta-1}\left(h^{\eta-1} - h_0^{\eta-1}\right) + \frac{g}{\eta+2}\left(h^{\eta+2} - h_0^{\eta+2}\right) + k q_0 |q_0| (x - x_0), \tag{1.57}$$

such that (1.56) rewrites

$$\chi(h; x, q_0, x_0, h_0) = 0. \tag{1.58}$$

The goal is now to find zeros of the function $\chi(h; x, q_0, x_0, h_0)$. First, we compute the derivative of χ with respect to h, as follows:

$$\frac{\partial \chi}{\partial h}(h; x, q_0, x_0, h_0) = -q_0^2 h^{\eta-2} + g h^{\eta+1}.$$

Hence, since $h > 0$, the derivative of χ with respect to h vanishes for $h = h_c$, with h_c defined by (1.49). Note that the same critical height h_c appears in the previous section, in the case where only the topography source term was present. The following result therefore holds.

Lemma 1.15. *With h_c defined by (1.49), the function χ defined by (1.57) satisfies the following properties:*

- *$h \mapsto \chi(h; x, q_0, x_0, h_0)$ is a strictly decreasing function for $h < h_c$;*
- *$h \mapsto \chi(h; x, q_0, x_0, h_0)$ is a strictly increasing function for $h > h_c$.*

The function χ therefore admits a single extremum, located at $h = h_c$. Moreover, this extremum is a minimum.

Equipped with the variations of χ, we now turn to determining the sign of this function to give existence results for solutions to (1.58) (or, equivalently, to (1.56)). First, we note that the following limit obviously holds:

$$\lim_{h\to+\infty} \chi(h; x, q_0, x_0, h_0) = +\infty. \tag{1.59}$$

Then, let $\chi_\ell(x)$ denote the evaluation $\chi(0; x, q_0, x_0, h_0)$ of χ for $h = 0$. Concerning $\chi_\ell(x)$, the following sequence of equalities holds:

$$\begin{aligned} \chi_\ell(x) := \chi(0; x, q_0, x_0, h_0) &= h_0^{\eta-1}\left(\frac{q_0^2}{\eta-1} - \frac{gh_0^3}{\eta+2}\right) + kq_0|q_0|(x-x_0) \\ &= gh_0^{\eta-1}\left(\frac{h_c^3}{\eta-1} - \frac{h_0^3}{\eta+2}\right) + kq_0|q_0|(x-x_0), \end{aligned} \tag{1.60}$$

where h_c is defined by (1.49). Finally, we denote by $\chi_c(x)$ the value of χ for $h = h_c$, to get:

$$\begin{aligned} \chi_c(x) := \chi(h_c; x, q_0, x_0, h_0) &= \chi_\ell(x) - h_c^{\eta-1}\left(\frac{q_0^2}{\eta-1} - \frac{gh_c^3}{\eta+2}\right) \\ &= \chi_\ell(x) - gh_c^{\eta-1}\left(\frac{h_c^3}{\eta-1} - \frac{h_c^3}{\eta+2}\right) \\ &= \chi_\ell(x) - \frac{3gh_c^{\eta+2}}{(\eta-1)(\eta+2)}. \end{aligned} \tag{1.61}$$

We remark that $\chi_c(x) < \chi_\ell(x)$. This was expected from Lemma 1.15, since χ is a strictly decreasing function on $(0, h_c)$. The following result summarizes the equations (1.59), (1.60) and (1.61).

Lemma 1.16. *With h_c given by (1.49), the function χ defined by (1.57) admits the following evaluations:*

- $\chi_\ell(x) := \chi(0; x, q_0, x_0, h_0) = gh_0^{\eta-1}\left(\frac{h_c^3}{\eta-1} - \frac{h_0^3}{\eta+2}\right) + kq_0|q_0|(x-x_0)$,
- $\chi_c(x) := \chi(h_c; x, q_0, x_0, h_0) = \chi_\ell(x) - \frac{3gh_c^{\eta+2}}{(\eta-1)(\eta+2)} < \chi_\ell(x)$,

as well as the following limit

$$\lim_{h\to+\infty} \chi(h; x, q_0, x_0, h_0) = +\infty.$$

To better understand Lemma 1.15 and Lemma 1.16, we now present sketches of the function χ for a specific set of variables. We take the following values:

- $q_0 = -\sqrt{g}/8$, so that $h_c = 0.25$;
- $h_0 = h_c = 0.25$;
- $x_0 = 0.75$.

As a consequence, Lemma 1.16 yields:

$$\chi_\ell(x) = \frac{3g}{(\eta-1)(\eta+2)}(0.25)^{\eta+2} - \frac{kg}{64}(x-x_0) \quad \text{and} \quad \chi_c(x) = -\frac{kg}{64}(x-x_0).$$

Recall from Lemma 1.15 that the function χ reaches its minimum for $h = h_c$. The number of zeros of he function χ is therefore related to the signs of χ_ℓ and χ_c. This property is highlighted by Figure 1.20, where χ is displayed for $x \in \{0.7, 0.75, 0.8, 0.85\}$ and $k = 1$.

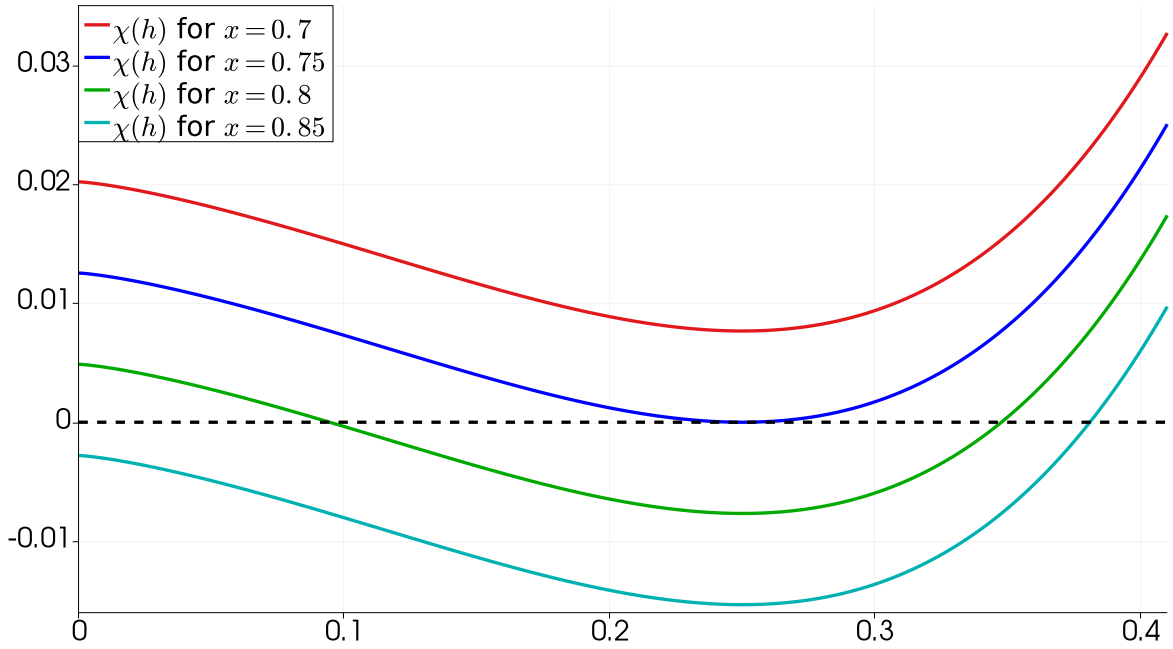

Figure 1.20 – Sketches of $\chi(h; x, -\sqrt{g}/8, 0.75, 0.25)$ for $h \in [0, 0.41]$ and for different values of x. Red curve: $x = 0.7$, no zero for χ. Blue curve: $x = 0.75$, unique zero for χ. Green curve: $x = 0.8$, two distinct zeros for χ. Cyan curve: $x = 0.85$, unique zero for χ.

The next result summarizes the conditions for the function χ to possess one or more zeros, and thus for one or more solutions of (1.58) to exist.

Proposition 1.17. *Assume $h > 0$ and $q_0 \neq 0$. Thus, $h_c > 0$ according to (1.49), and the following assertions hold.*

(i) If $\chi_c(x) > 0$, then there is no solution to the equation (1.58).

(ii) If $\chi_c(x) = 0$, then the equation (1.58) admits a unique solution. This solution, $h = h_c$, is a double root of the function $h \mapsto \chi(h; x, q_0, x_0, h_0)$.

(iii) If $\chi_c(x) < 0$ and $\chi_\ell(x) > 0$, then the equation (1.58) admits two distinct solutions, $h_{sup} \in (0, h_c)$ and $h_{sub} \in (h_c, +\infty)$.

(iv) If $\chi_c(x) < 0$ and $\chi_\ell(x) \leq 0$, then the equation (1.58) admits a unique solution $h_{sub} \in (h_c, +\infty)$.

Proof. The proof of this result relies on using Lemma 1.15 and Lemma 1.16.

From Lemma 1.15, the function χ reaches its unique minimum χ_c for $h = h_c$. The proofs of (i) and (ii) are therefore immediate.

If $\chi_c < 0$, there is at least one zero of χ located in $(h_c, +\infty)$. The number of zeros of χ now depends on the value of $\chi_\ell(x) = \chi(0; x, q_0, x_0, h_0)$, given by Lemma 1.16.

– On the one hand, if $\chi_\ell > 0$, then the function χ admits another zero, located in $(0, h_c)$.

– On the other hand, if $\chi_\ell < 0$, then the function χ does not admit another zero.

– Finally, if $\chi_\ell = 0$, then $\chi(0; x, q_0, x_0, h_0) = 0$, i.e. $h = 0$ is a solution to (1.58). However, this solution is not admissible since we have assumed $h \neq 0$ in order to proceed with the previous computations.

The assertions (iii) and (iv) are thus proven, which concludes the proof. □

Note that the four assertions of Proposition 1.17 respectively correspond to the red, blue, green and cyan curves of Figure 1.20.

The same remark as in the topography case can be made here. Indeed, the Froude number is still defined by (1.50) in the general case and by (1.51) in the current case of a steady state solution. We have again labeled the two solutions of (1.58) as h_{sup} and h_{sub}, since they respectively correspond to a supercritical flow and a subcritical flow.

We now determine conditions on x for the existence of solutions to (1.58) (or, equivalently, of zeros of the function χ).

Corollary 1.18. *Assume $h > 0$ and $q_0 \neq 0$. Thus, $h_c > 0$ according to (1.49). We define the following limit value of the position:*

$$x_u = x_0 + \frac{h_0^{\eta-1}}{k\mu_0}\left(\frac{1}{\eta+2}\frac{h_0^3}{h_c^3} - \frac{1}{\eta-1}\right),$$

where $\mu_0 = \mathrm{sgn}(q_0)$ denotes the sign of q_0, i.e. the direction of the steady water flow. We also define the following critical position:

$$\begin{aligned} x_c &= x_0 + \frac{1}{(\eta-1)(\eta+2)}\frac{1}{k\mu_0}\frac{1}{h_c^3}\left((\eta-1)h_0^{\eta+2} - (\eta+2)h_c^3 h_0^{\eta-1} + 3h_c^{\eta+2}\right) \\ &= x_u + \frac{1}{k\mu_0}\frac{3h_c^{\eta-1}}{(\eta-1)(\eta+2)}. \end{aligned}$$

Equipped with x_u and x_c, the following properties hold.

(i) If $\mu_0 x > \mu_0 x_c$, then there is no solution to the equation (1.58).

(ii) If $\mu_0 x = \mu_0 x_c$, then the equation (1.58) admits a unique solution. This solution, $h = h_c$, is a double root of the function $h \mapsto \chi(h; x, q_0, x_0, h_0)$.

(iii) If $\mu_0 x < \mu_0 x_c$ and $\mu_0 x > \mu_0 x_u$, then the equation (1.58) admits two distinct solutions, $h_{sup} \in (0, h_c)$ and $h_{sub} \in (h_c, +\infty)$.

(iv) If $\mu_0 x < \mu_0 x_c$ and $\mu_0 x \leq \mu_0 x_u$, then the equation (1.58) admits a unique solution $h_{sub} \in (h_c, +\infty)$.

Proof. This result is a direct consequence of Proposition 1.17. Indeed, note the following relations, which are obtained by performing straightforward but tedious computations:

- $\chi_c(x) > 0$ if and only if $\mu_0 x > \mu_0 x_c$;
- $\chi_\ell(x) > 0$ if and only if $\mu_0 x > \mu_0 x_u$.

Arguing Proposition 1.17 then proves Corollary 1.18. □

Remark 1.19. The identity (1.54), characterizing the smooth steady states, rewrites:

$$gh^{\eta-2}\left(-\frac{q_0^2}{g} + h^3\right)\partial_x h = -kq_0|q_0|.$$

Now, recall the definition (1.49) of h_c, and assume that $h > 0$ and $h \neq h_c$. The above relation rewrites as follows:

$$\partial_x h = \frac{q_0}{h_c^3 - h^3} \frac{k|q_0|}{gh^{\eta-2}}.$$

Therefore, the sign of $\partial_x h$ is that of $q_0(h_c - h)$. For instance, if $q_0 < 0$, then the subcritical solution ($h > h_c$) is increasing, while the supercritical solution ($h < h_c$) is decreasing. As a consequence, since both subcritical solution and supercritical solution are strictly monotonic, they are bijective on their respective domains.

We conclude this section on smooth steady states for the friction source term with an example, to illustrate Proposition 1.17, Corollary 1.18 and Remark 1.19. For this example, we follow the steady state presented on Figure 1.20 and we take $q_0 = -\sqrt{g}/8 < 0$. Therefore, we have $h_c = 0.25$ and we take $h_0 = h_c = 0.25$. As a consequence, $x_c = x_0$ from Corollary 1.18, and we set $x_0 = 0.75$. In addition, we have $x_u > x_0$. The solutions $h(x)$ of (1.58), obtained with Newton's method, are displayed on Figure 1.21 for $x \in [0.7, 2x_u - x_c]$ and $k = 1$.

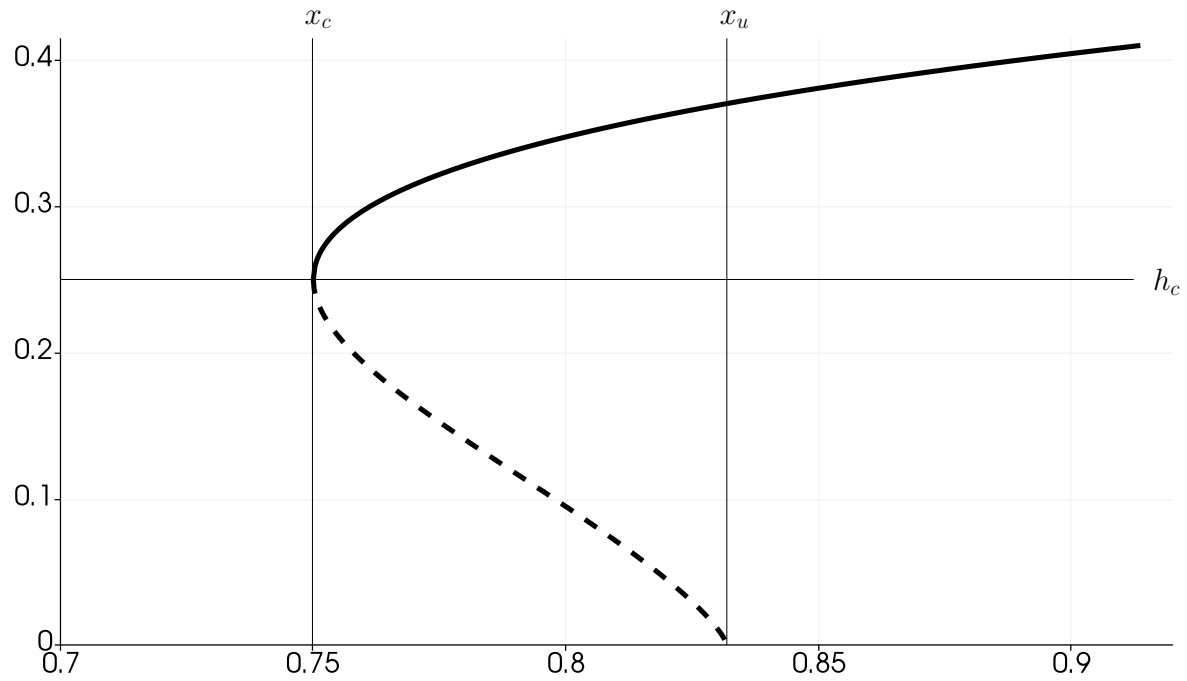

Figure 1.21 – Solutions $h(x)$ of (1.58) (where they exist). Full line: subcritical solution. Dotted line: supercritical solution.

We observe on Figure 1.21 that the solutions of the equation (1.58) indeed follow the pattern of existence and uniqueness predicted by Proposition 1.17 and Corollary 1.18. In addition, the conclusions of Remark 1.19 are verified: here, $q_0 < 0$, and we observe that the subcritical solution is indeed increasing, while the supercritical solution decreases. This example concludes this section on friction-only smooth steady states.

1.2.2.2 Considerations on discontinuous steady states

We have obtained the general form of smooth steady states for the shallow-water equations with friction and flat topography, respectively given by the subcritical and supercritical

solutions. Let us underline that there is no steady state solution as long as $q_0(x - x_0) > 0$. However, there exists an infinity of subcritical and supercritical smooth steady states that are solution to (1.58), provided the initial conditions x_0 and h_0 are chosen differently. Now, we remark that, from these smooth steady states, it is possible to define discontinuous solutions of (1.52). The remainder of this section is devoted to defining such non-smooth steady states. In order to exhibit relevant discontinuous steady states, we need to find *admissible discontinuities* connecting any two smooth solutions, i.e. discontinuities that verify the Rankine-Hugoniot relations and the entropy inequality. The Rankine-Hugoniot conditions have been presented in Appendix A, and the entropy inequality is given by (1.6).

We consider a steady state solution h, subcritical or supercritical, which we try to link to another steady state solution $\tilde{h}$, obtained by solving (1.56) with a different initial condition. In addition, we assume that $\tilde{h} \neq h$, to avoid the degenerate case of a discontinuity linking the same two states. For the sake of simplicity in the forthcoming developments, we choose that $h(x_0) = h_0 = h_c$, i.e. h is solution to $\chi(h; x, q_0, x_0, h_c) = 0$, with χ defined by (1.57). In addition, we assume that $\tilde{h}(\tilde{x}_0) = \tilde{h}_0 = h_c$, i.e. $\tilde{h}$ is solution to $\chi(\tilde{h}; x, q_0, \tilde{x}_0, h_c) = 0$.

We consider an admissible stationary discontinuity. As a first step, we consider the conditions imposed by the Rankine-Hugoniot relations (A.2) For a stationary discontinuity, we get (see also Section 1.2.1.3):

$$\begin{cases} [q] = 0, & \text{(1.62a)} \\ \left[\dfrac{q^2}{h} + \dfrac{1}{2}gh^2\right] = 0, & \text{(1.62b)} \end{cases}$$

where the notation $[X]$ denotes the jump of the quantity X across the discontinuity. Arguing the Rankine-Hugoniot relations yields the following result.

Lemma 1.20. *Consider two steady state solutions ${}^t(h, q_0)$ and ${}^t(\tilde{h}, q_0)$, subcritical or supercritical. Assume that these solutions are connected with an admissible discontinuity. As a consequence, $\tilde{h}$ may be viewed as a function of h, and the following relation holds:*

$$\tilde{h}(h) = \frac{h}{2}\left(\sqrt{1 + 8\frac{h_c^3}{h^3}} - 1\right). \tag{1.63}$$

Proof. Since q is equal to the constant q_0, the relation (1.62a) obviously holds. Concerning (1.62b), we have

$$\left(\frac{q_0^2}{\tilde{h}} + \frac{1}{2}g\tilde{h}^2\right) - \left(\frac{q_0^2}{h} + \frac{1}{2}gh^2\right) = 0,$$

and we immediately obtain

$$\frac{1}{2}g\tilde{h}^3 - \left(\frac{q_0^2}{h} + \frac{1}{2}gh^2\right)\tilde{h} + q_0^2 = 0.$$

Introducing a factorization by $(\tilde{h} - h)$ and arguing the definition (1.49) of h_c, we get

$$(\tilde{h} - h)\left(\tilde{h} - \frac{h}{2}\left(\sqrt{1 + 8\frac{h_c^3}{h^3}} - 1\right)\right)\left(\tilde{h} + \frac{h}{2}\left(\sqrt{1 + 8\frac{h_c^3}{h^3}} + 1\right)\right) = 0.$$

Since $\tilde{h}$ must be positive and $\tilde{h} \neq h$, it is clear that the admissible discontinuity connects h to

$$\tilde{h}(h) = \frac{h}{2}\left(\sqrt{1+8\frac{h_c^3}{h^3}} - 1\right).$$

The proof is therefore completed. □

As a second step, we focus on the entropy. The entropy inequality (1.6) rewrites as follows for a flat topography:

$$\partial_t s(W) + \partial_x G(W) \leq -kq^2|q|h^{-\eta-1},$$

with s and G given by (1.4). In the current context of a steady state solution involving a stationary discontinuity, this entropy inequality reads:

$$\partial_x G(W) \leq -kq_0^2|q_0|h^{-\eta-1}.$$

Note that, since we have assumed a nonzero steady discharge q_0, we have $-kq_0^2|q_0|h^{-\eta-1} < 0$. As a consequence, to recover the above entropy inequality, it is sufficient to take h and $\tilde{h}$ such that the jump of G between these two states is negative, as follows:

$$[G] < 0.$$

As a consequence, arguing the definition (1.4) of the entropy flux G, we get that h and $\tilde{h}$ are connected with an admissible discontinuity if and only if

$$q_0\left[\frac{q_0^2}{2h^2} + gh\right] < 0. \tag{1.64}$$

The study of this equation allows the statement of the following result.

Lemma 1.21. *Consider two steady state solutions ${}^t(h, q_0)$ and ${}^t(\tilde{h}, q_0)$, subcritical or supercritical. Assume that these solutions are connected with an admissible discontinuity. We define the following quantity:*

$$\hat{h}_+ = \frac{h_c^3}{4h^2}\left(1 + \sqrt{1+8\frac{h^3}{h_c^3}}\right). \tag{1.65}$$

The following two assertions hold:

(i) if $q_0 > 0$, then $\tilde{h} \in (\min(h, \hat{h}_+), \max(h, \hat{h}_+))$;

(ii) if $q_0 < 0$, then $\tilde{h} \notin (\min(h, \hat{h}_+), \max(h, \hat{h}_+))$.

Proof. We start by proving (i). As a consequence, we assume $q_0 > 0$. In this case, the discrete entropy inequality $[G] < 0$, or equivalently (1.64), reads:

$$\frac{q_0^2}{2\tilde{h}^2} + g\tilde{h} < \frac{q_0^2}{2h^2} + gh. \tag{1.66}$$

Introducing a factorization by $(\tilde{h}-h)$, the above inequality rewrites as follows:

$$(\tilde{h}-h)\left(-\frac{h_c^3}{2}\frac{\tilde{h}+h}{\tilde{h}^2h^2}+1\right)>0.$$

Since $\tilde{h}^2h^2>0$, the above estimation reads:

$$(\tilde{h}-h)(2h^2\tilde{h}^2-h_c^3\tilde{h}-h_c^3h)<0.$$

As a consequence, we have $[G]<0$ if and only if:

$$(\tilde{h}-h)(\tilde{h}-\hat{h}_-)(\tilde{h}-\hat{h}_+)<0,$$

where we have set

$$\hat{h}_\pm=\frac{h_c^3}{4h^2}\left(1\pm\sqrt{1+8\frac{h^3}{h_c^3}}\right).$$

We immediately note that $\hat{h}_-<0$. Therefore, $\tilde{h}-\hat{h}_-$ is always negative, and the discontinuity satisfies the entropy inequality if and only if:

$$(\tilde{h}-h)(\tilde{h}-\hat{h}_+)<0. \tag{1.67}$$

As a consequence, we have $\tilde{h}\in(\min(h,\hat{h}_+),\max(h,\hat{h}_+))$, which achieves the proof of (i).

Regarding the proof of (ii), we assume that $q_0<0$. The inequality (1.66) now reads:

$$\frac{q_0^2}{2\tilde{h}^2}+g\tilde{h}>\frac{q_0^2}{2h^2}+gh.$$

We note that the direction of the inequality has been reversed. As a consequence, for $q_0<0$, the relation (1.67) rewrites:

$$(\tilde{h}-h)(\tilde{h}-\hat{h}_+)>0.$$

Therefore, the assertion $\tilde{h}\notin(\min(h,\hat{h}_+),\max(h,\hat{h}_+))$ is established. The proof of (ii) is thus concluded, which completes the proof of Lemma 1.21. □

We have therefore uncovered several relations defining an admissible discontinuity linking h to $\tilde{h}$. Namely, Lemma 1.20 states the condition obtained from the Rankine-Hugoniot relations, which gives an expression of $\tilde{h}$ with respect to h. Lemma 1.21 states the necessary location of $\tilde{h}$ with respect to h and $\hat{h}_+$, defined by (1.65), to satisfy the entropy conditions. Note that Lemma 1.21 involves a comparison between $\tilde{h}$ and $\hat{h}_+$. The expression (1.63) allows us to state the following result.

Lemma 1.22. *Define $\tilde{h}$ and $\hat{h}_+$ by (1.63) and (1.65), as follows:*

$$\tilde{h}=\frac{h}{2}\left(\sqrt{1+8\frac{h_c^3}{h^3}}-1\right)\quad\text{and}\quad\hat{h}_+=\frac{h_c^3}{4h^2}\left(1+\sqrt{1+8\frac{h^3}{h_c^3}}\right). \tag{1.68}$$

For all $h \in \mathbb{R}_+^* \setminus \{h_c\}$, we have:

$$\tilde{h} < \hat{h}_+.$$

Proof. Let us prove that $\tilde{h} < \hat{h}_+$, with $\tilde{h}$ and $\hat{h}_+$ defined by (1.68). As a consequence, we have to show that

$$\frac{h_c}{h}\left(\frac{h_c^2}{h^2} + \sqrt{\frac{h_c^4}{h^4} + 8\frac{h_c}{h}}\right) - 2\left(\sqrt{1 + 8\frac{h_c^3}{h^3}} - 1\right) > 0.$$

Introducing $\alpha = h_c/h$, the above inequality holds if and only if $f(\alpha) > 0$, where the function f is defined by:

$$f(\alpha) = \alpha\Big(\alpha^2 + \sqrt{\alpha^4 + 8\alpha}\Big) - 2\Big(\sqrt{1 + 8\alpha^3} - 1\Big).$$

We now study the sign of the function f for $\alpha \in (0, +\infty)$. We remark that $f(\alpha)$ may be rewritten as follows, after straightforward computations:

$$f(\alpha) = \frac{\alpha^3}{\sqrt{1 + 8\alpha^3} + 1}\left[\left(1 + \sqrt{1 + \frac{8}{\alpha^3}}\right)\left(1 + \sqrt{1 + 8\alpha^3}\right) - 16\right].$$

Let $\tau(\alpha) := 1 + \sqrt{1 + 8\alpha^3}$. Equipped with this notation, $f(\alpha) > 0$ is equivalent to the following inequality:

$$\tau\left(\frac{1}{\alpha}\right)\tau(\alpha) - 16 > 0. \tag{1.69}$$

After straightforward computations, we get:

$$\tau\left(\frac{1}{\alpha}\right)\tau(\alpha) = 1 + \sqrt{1 + 8\alpha^3} + \sqrt{1 + \frac{8}{\alpha^3}} + \sqrt{65 + 8\left(\alpha^3 + \frac{1}{\alpha^3}\right)}. \tag{1.70}$$

In order to study a lower bound of the above expression, we introduce two following two notations:

$$\tau_1(\alpha) = \sqrt{1 + 8\alpha^3} + \sqrt{1 + \frac{8}{\alpha^3}} \quad \text{and} \quad \tau_2(\alpha) = \sqrt{65 + 8\left(\alpha^3 + \frac{1}{\alpha^3}\right)}.$$

Note that $\tau(1/\alpha)\tau(\alpha) = 1 + \tau_1(\alpha) + \tau_2(\alpha)$.

— We first study the variations of τ_1. The following formula gives the derivative of τ_1:

$$\tau_1'(\alpha) = \frac{12}{\alpha^4}\left(\frac{\alpha^6}{\sqrt{1 + 8\alpha^3}} - \frac{1}{\sqrt{1 + 8\alpha^{-3}}}\right).$$

Straightforward computations show that $\tau_1'(\alpha) > 0$ if and only if $\alpha > 1$. As a consequence, the function τ_1 reaches its unique minimum for $\alpha = 1$, and we get $\tau_1(\alpha) > \tau_1(1)$. Therefore, we obtain the following lower bound for τ_1:

$$\forall \alpha \in (0, +\infty),\ \tau_1(\alpha) \geq 6. \tag{1.71}$$

In (1.71), the equality case corresponds to $\alpha = 1$.

— We then study the variations of the function τ_2. Since the derivative of τ_2 is given by

$$\tau_2'(\alpha) = \frac{24\alpha^{-4}}{\sqrt{65 + 8(\alpha^3 + \alpha^{-3})}}\big(\alpha^6 - 1\big),$$

we immediately obtain that τ_2 also reaches its unique minimum for $\alpha = 1$. The following lower bound therefore holds:

$$\forall \alpha \in (0, +\infty),\ \tau_2(\alpha) \geq 9. \tag{1.72}$$

The equality case in (1.72) again corresponds to $\alpha = 1$.

As a consequence, from (1.70), we get that $\tau(1/\alpha)\tau(\alpha) \geq 16$, with the equality case corresponding to $\alpha = 1$. Therefore, (1.69) holds for all $\alpha \in \mathbb{R}_+^* \setminus \{1\}$. Hence, we conclude that, for all $h \in \mathbb{R}_+^* \setminus \{h_c\}$, we have $\tilde{h} < \hat{h}_+$. The proof is thus achieved. □

Remark 1.23. The case $h = h_c$ leads, by application of Lemma 1.20, to $\tilde{h} = h_c = h$. As a consequence, we do not consider this case, as it is the degenerate case of a discontinuity connecting the same two states, and we take $h \in \mathbb{R}_+^* \setminus \{h_c\}$. Lemma 1.22 therefore holds for all values of h under consideration.

Equipped with Lemma 1.22, which introduces a comparison between $\tilde{h}$ and $\hat{h}_+$, we can eliminate several cases from Lemma 1.21. Indeed, we replace the estimations uncovered in Lemma 1.21 by the following two assertions.

(1a) If $q_0 > 0$, then $\tilde{h} > h$.

(1b) If $q_0 < 0$, then $\tilde{h} < h$.

Now, remark that the expression of $\tilde{h}$ from Lemma 1.20 immediately yields the following estimations, with h linked to $\tilde{h}$ through a discontinuity satisfying the Rankine-Hugoniot conditions.

(2a) We have $\tilde{h} > h$ if and only if $h < h_c$.

(2b) We have $\tilde{h} < h$ if and only if $h > h_c$.

As a consequence, the following result holds.

Proposition 1.24. *Let $q_0 \in \mathbb{R}^*$, and define h_c by (1.49). Let $h \in \mathbb{R}_+^* \setminus \{h_c\}$. We wish to build a discontinuous steady state solution, i.e. we seek $\tilde{h}$ connected to h by an admissible discontinuity. The quantity $\tilde{h}$ is given by (1.63). The Rankine-Hugoniot conditions and the entropy inequality yield the following assertions.*

- *If $q_0 > 0$, then the only possible solution that can be connected to another one with an admissible discontinuity is a supercritical solution, i.e. $h < h_c$. In that case, we have $\tilde{h} > h$.*
- *If $q_0 < 0$, then the only possible solution that can be connected to another one with an admissible discontinuity is a subcritical solution, i.e. $h > h_c$. In that case, we have $\tilde{h} < h$.*

As a consequence, on the one hand, if we start with a subcritical solution (such that $h > h_c$), then the admissible discontinuity connects h to $\tilde{h} < h$. Therefore, the discontinuity may connect the subcritical solution h to another subcritical solution with $\tilde{h} < h$, or to a

supercritical solution. On the other hand, if we wish to link a supercritical solution (satisfying $h < h_c$) to $\tilde{h}$, then we necessarily have $\tilde{h} > h$. Hence, this supercritical solution may be connected to either another supercritical solution $\tilde{h} > h$, or to a subcritical solution.

1.2.3 Topography and friction steady states

In the previous sections, we have studied steady state solutions of the shallow-water equations endowed with either the topography source term or the Manning friction source term. The goal of this section is to provide some insight on steady state solutions when both source terms are present. Such solutions are governed by the following equations:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -gh\partial_x Z - kq|q|h^{-\eta}. \end{cases}$$

As expected, the discharge q is uniform for steady states. As usual, this uniform discharge is denoted by q_0. Then, equipped with q_0, the steady water height $h(x)$ is a solution of the following ordinary differential equation:

$$\partial_x \left(\frac{q_0^2}{h} + \frac{1}{2} g h^2 \right) = -gh\partial_x Z - kq_0|q_0|h^{-\eta}. \tag{1.73}$$

We now exhibit several steady state solutions obtained in specific cases. Namely, we exhibit two specific solutions of (1.73), with $h = \mathrm{cst}$ and $h + Z = \mathrm{cst}$. Then, we give a word on the solutions of (1.73) in the general case.

Uniform water height

First, with $q_0 \neq 0$, we consider a uniform water height $h(x) = h_0 \neq 0$. This case has been studied in [42], where the authors propose a scheme able to capture this steady state. With $h(x) = h_0$, (1.73) becomes

$$gh_0\partial_x Z + kq_0|q_0|h_0^{-\eta} = 0.$$

Recall the notation $\mu_0 = \mathrm{sgn}(q_0)$. For the steady state with $h = h_0$, we therefore get:

$$\partial_x Z = -\frac{kq_0|q_0|}{gh_0^{\eta+1}}. \tag{1.74}$$

As a consequence, the topography function has to be affine, of slope given by (1.74).

Uniform free surface

Second, we no longer assume that the water height is constant. Instead, we take steady state solution with a constant free surface H_0, i.e. a steady solution such that $h(x)+Z(x) = H_0$. Using $Z(x) = H_0 - h(x)$, (1.73) becomes

$$\partial_x \left(\frac{q_0^2}{h} + \frac{1}{2} g h^2 \right) = gh\partial_x h - kq_0|q_0|h^{-\eta}.$$

Now, we assume that the solution h of the above equation is a smooth function. Therefore, it satisfies:

$$\frac{q_0^2}{h^2}\partial_x h = kq_0|q_0|h^{-\eta}.$$

Using $\mu_0 = \mathrm{sgn}(q_0)$ yields:

$$h^{\eta-2}\partial_x h = k\mu_0.$$

Let $x_0 \in \mathbb{R}$. We assume that the water height $h_0 = h(t, x_0)$ at x_0 is known. The above identity is then integrated over $[x_0, x]$, to get:

$$h^{\eta-1} = h_0^{\eta-1} + (\eta-1)k\mu_0(x - x_0). \tag{1.75}$$

The water height h must be nonnegative. As a consequence, there is no solution if the right-hand side of (1.75) is negative. Hence, we assume that this right-hand side is positive, i.e. that

$$h_0^{\eta-1} + (\eta-1)k\mu_0(x - x_0) > 0. \tag{1.76}$$

Therefore, recalling that $Z(x) = H_0 - h(x)$, the water height and topography are given by:

$$\begin{aligned} h(x) &= \Big(h_0^{\eta-1} + (\eta-1)k\mu_0(x - x_0)\Big)^{\frac{1}{\eta-1}}, \\ Z(x) &= H_0 - \Big(h_0^{\eta-1} + (\eta-1)k\mu_0(x - x_0)\Big)^{\frac{1}{\eta-1}}. \end{aligned} \tag{1.77}$$

The general case

Third, we turn to the general case, where the steady states are solutions of (1.73) without simplifications. We make the very important remark that (1.73) cannot be rewritten under the algebraic form $\zeta(h) = 0$, contrary to the individual cases of the topography and friction, where the steady states were respectively governed by (1.47) and (1.58). Therefore, an analogous study cannot be applied in the case where both source terms are present. In order to exhibit solutions to (1.73), numerical methods have to be used. Namely, a discretization of this equation has to be provided.

2

Finite volume methods

The goal of this chapter is to present a state of the art regarding finite volume methods applied to hyperbolic conservation laws and balance laws, in one or two space dimensions. The techniques discussed in this chapter are all well-known, but we recall them here since they will be widely used in the remainder of this manuscript. For more information on systems of conservation laws, the reader is referred to [79, 141, 142] for instance, but this reference list is not exhaustive. This chapter provides a general setting for the numerical approximation of the shallow-water system (1.1).

Systems of conservation laws are governed, in one space dimension, by the following initial value problem:

$$\begin{cases} \partial_t W(t,x) + \partial_x F(W(t,x)) = 0, \\ W(0,x) = W_0(x). \end{cases} \tag{2.1}$$

In (2.1), $W : \mathbb{R}_+ \times \mathbb{R} \to \Omega$ is the *vector of conserved variables*, whose values lie within the *admissible states space* $\Omega \subset \mathbb{R}^N$, supposed to be convex. In addition, we assume that the space Ω is invariant, i.e.

$$\text{if, } \forall x \in \mathbb{R},\ W_0(x) \in \Omega,\ \text{ then } \forall (t,x) \in \mathbb{R}_+^* \times \mathbb{R},\ W(t,x) \in \Omega. \tag{2.2}$$

The variable t represents the time, while x is the space variable. The function $F : \Omega \to \mathbb{R}^N$ is called the *physical flux function*, and is assumed to be smooth. We assume that the system (2.1) is hyperbolic. The initial condition W_0 is a potentially discontinuous function of x.

It is a well-known fact that, even with smooth initial data, solutions of (2.1) may present *discontinuities* in finite time if the flux function is nonlinear. To address this issue, from now on, we focus on weak solutions of the problem. In addition, it has been proven that, if a weak solution admits a discontinuity, then this discontinuity satisfies the Rankine-Hugoniot conditions (1.17) (see Appendix A for a proof of this result in a more general setting).

From the numerical point of view, it is crucial that the properties satisfied by the equations

also be satisfied by the numerical scheme, in order to provide a good approximation of the solutions to (2.1). In particular, a scheme is called *robust* if the invariance of the admissible states space (2.2) is preserved at the discrete level. More detail on these discrete properties will be given in the next section, in the context of a finite volume scheme.

Over the course of this chapter, *Riemann problems* will naturally appear while dealing with the numerical approximation of (2.1). A Riemann problem is a Cauchy problem with discontinuous initial data, see (1.16). More details on Riemann problems are present in [80, 79] for instance. Since the conservation law $\partial_t W + \partial_x F(W) = 0$ is hyperbolic, the exact solution of the Riemann problem (1.16) is made of waves, traveling at finite velocities and separating constant intermediate states. The nature of these waves is linked to the nature of the characteristic fields associated to the eigenvalues of the Jacobian matrix of F (see Definition 1.1). We assume that each field is either linearly degenerate or genuinely nonlinear.

This framework fits the shallow-water equations (1.1) (see Chapter 1). This system can be cast under the general form of a balance law, as follows:

$$\begin{cases} \partial_t W + \partial_x F(W) = \mathfrak{S}(W), \\ W(0,x) = W_0(x), \end{cases} \tag{2.3}$$

where $\mathfrak{S}(W)$ is a source term and the homogeneous system obtained from making the source term vanish in (2.3) is hyperbolic. In the context of the shallow-water equations with topography and Manning friction, $\mathfrak{S}(W)$ is made of the two source terms. The presence of the LD fields causes the Riemann problem for the shallow-water equations with source terms to be much harder to solve explicitly than in the homogeneous case. As a consequence, numerical methods have to be applied.

For the balance law (2.3), the *steady state solutions* (or *steady states*) are defined by making the time derivative $\partial_t W$ vanish, as follows:

$$\partial_x F(W) = \mathfrak{S}(W). \tag{2.4}$$

In the context of the shallow-water equations, the steady state solutions have been studied in Section 1.2. Examples of steady state solutions have been exhibited in the cases of the topography and the friction source terms. However, since they are governed by very nonlinear equations, usual numerical schemes do not exactly preserve such solutions, and special treatments have to be introduced. From the numerical point of view, a scheme will be called *well-balanced* if it exactly preserves all the steady states at the discrete level.

Well-balanced schemes have been extensively studied in the last two decades. First, in [11], Bermudez and Vazquez introduced the C-property to describe schemes preserving the steady states at rest of the shallow-water equations. The term of well-balanced scheme was then coined by Greenberg and LeRoux in their seminal work [87], where the authors defined a well-balanced scheme as able to exactly preserve solutions at rest. Afterwards, Goutal and Maurel proposed a review [86] of several steady state solutions for the shallow-water water equations with topography, and of several schemes to approximate such steady states. Then, still within the framework of the shallow-water equations with topography, Gosse proposed a scheme in [82], based on an approximate resolution of Bernoulli's equation, to exactly capture

all the steady states of the equations. His approach was later simplified by Audusse et al. in [5], who focused on the steady states at rest (i.e. where the velocity vanishes) to propose the well-known hydrostatic reconstruction.

A lot of other first-order schemes were derived to preserve the steady state solutions of the shallow-water water equations with the topography source term. In particular, we mention work dealing with robust schemes that preserve the steady states at rest, see for instance [100, 133, 26, 44, 105, 28, 19, 23, 7] for 1D meshes, and see for instance [6, 29, 69, 166] for unstructured 2D geometries. Several schemes that also preserve the moving steady states, but are not robust, have then been derived (see for instance [33, 72]). Later on, in [12, 13], the authors suggest a robust and entropy-satisfying scheme that preserves the moving steady states. The advantages of schemes that preserve all the steady states, including the moving ones, have been highlighted in [160]. Finally, we mention schemes preserving the steady states associated to the friction source term in specific cases (see [115, 42]), and the Coriolis force source term (see [119, 43]).

Some work involving high-order techniques to capture the steady states exactly or with a high order of approximation has also been proposed. For instance, in [127, 128, 158], the authors suggest a WENO approach, while the authors of [161, 157] focus on discontinuous Galerkin methods (see also the review article [159]). Other high-order methods, using the steady state solutions, have been developed in [32, 74, 136, 35]. In addition, in [47], a scheme that preserves the lake at rest on unstructured meshes has been derived.

We also present a non-exhaustive list of well-balanced schemes for other systems. For instance, we mention well-balanced scheme for the Euler equations with gravity (see [37, 163, 60, 101, 39, 61, 41]), the equations of chemotaxis (see [125, 15]), a two-layer shallow-water model (see [106, 67]), the Ripa model (see [152, 62], the equations of hemodynamics (see [56]), and the shallow-water equations with pollutant transport (see [70]).

The goal of this manuscript is to provide a numerical scheme that is consistent and robust. In addition, it has to be well-balanced, i.e. able to *preserve* and *capture* the steady state solutions exhibited in Section 1.2. These two terms of preservation and capture are defined as follows.

- Let W_0 be an initial condition at rest, i.e. satisfying (2.4). A numerical scheme is said to preserve the steady states if the approximate solution obtained from W_0 stays stationary, i.e. satisfies (2.4).
- Let W_0 be an unstationary initial condition, i.e. that does not satisfy (2.4). Assume that, in finite time, the solution $W(t,x)$ of (2.3) becomes stationary, i.e. satisfies (2.4), after a transient state. A numerical scheme is said to capture the steady states if the approximate solution obtained from W_0 also becomes stationary.

As a consequence, this chapter is dedicated to providing some numerical methods that will be the basis of the work presented later in this manuscript. The numerical methods are here presented in the general setting of a system of conservation laws, and will be applied to the shallow-water system later on. The chapter is organized as follows.

Section 2.1 focuses on the first-order finite volume approximation of one-dimensional (1D) hyperbolic conservation laws. To that end, we begin by providing a finite volume discretization of the equations. After discretizing the space domain into cells, the conservation law is in-

tegrated over some cell, in order to exhibit the main ingredients of any finite volume method. Namely, the approximated solution is piecewise constant on the cells, and the interactions between cells are represented by a numerical flux function. Then, we introduce Godunov's method. This method uses the exact solution of a Riemann problem for the conservation law to approximate the result of the interaction between two contiguous cells. Afterwards, we present Godunov-type methods, an extension of Godunov's method. Godunov-type methods do not require the exact solution of a Riemann problem. Instead, they use an approximation of this solution. Thanks to this approximation, these methods are more versatile, as they can be applied to systems for which the exact Riemann solution is not known. The derivation of the HLL scheme, a notable Godunov-type scheme whose approximate Riemann solution is heavily used later in this manuscript, is also presented.

The second section of this chapter, Section 2.2, presents second-order spatial accuracy techniques in one space dimension. Such techniques are used to improve the spatial accuracy of the scheme for smooth and non-smooth solutions. Namely, we introduce the MUSCL technique, which consists in providing a piecewise linear approximation of the solution in each cell, instead of piecewise constant. However, robustness properties satisfied by the first-order scheme can be lost because of the linear reconstruction, and oscillations may appear in the approximate solution. In order to recover such properties and eliminate the oscillations, slope limiters are designed to make sure the slope of the reconstruction is small enough.

The remainder of the chapter is devoted to the approximation of conservation laws and balance laws in two space dimensions. We begin this study in Section 2.3 by deriving two-dimensional (2D) first-order finite volume schemes for 2D conservation laws. As before, we first provide a finite volume discretization of the space domain and of the equations. To that end, we introduce a polygonal mesh of the 2D domain, and we integrate the equations over this mesh. Then, we prove that this 2D scheme can be rewritten as a convex combination of 1D schemes. Thanks to this convex combination, the 2D scheme is immediately shown to satisfy some robustness properties verified by the 1D scheme.

Finally, Section 2.4 deals with high-order schemes in two space dimensions. High-order schemes are based on suitable polynomial reconstructions (for instance, the MUSCL technique is based on a linear reconstruction). We begin by defining such a reconstruction. Afterwards, the polynomial reconstruction is used to derive a high-order finite volume scheme for a 2D balance law. First, we present a scheme that is high-order accurate in space. Second, a high-order time discretization is provided in. Third, the Multidimensional Optimal Order Detection (MOOD) method is presented. The MOOD method consists in enforcing some properties by choosing to lower the degree of the reconstruction in areas where the properties are not satisfied.

2.1 One-dimensional first-order finite volume schemes for hyperbolic problems

We turn to the numerical approximation of the solutions of hyperbolic systems of conservation laws (2.1). Several challenges arise, such as the ability to approximate discontin-

uous solutions. To address this issue and to provide a suitable numerical approximation, we introduce the framework of *finite volume schemes*. First, in Section 2.1.1, we focus on the discretization of the equations and of the space domain. Then, in Section 2.1.2, we present Godunov's scheme. This scheme involves the exact solution of a Riemann problem. Finally, in Section 2.1.3, extensions of Godunov's scheme, the Godunov-type schemes, are discussed. These schemes use an approximation of the Riemann solution instead of the exact solution. For the remainder of the section, the reader is referred to [92, 93, 111, 150] for instance, but this list in non-exhaustive.

2.1.1 Finite volume discretization

The first step of providing a finite volume discretization of the equation (2.1) consists in discretizing the space domain $\mathbb{R}$. Let us consider a discretization made of *cells* c_i. For the sake of simplicity, we assume that all cells have the same length Δx. We denote by x_i the x-coordinate of the center of the cell c_i. In addition, we denote by $x_{i+\frac{1}{2}}$ the x-coordinate of the interface between cells c_i and c_{i+1}. These notations are displayed on Figure 2.1.

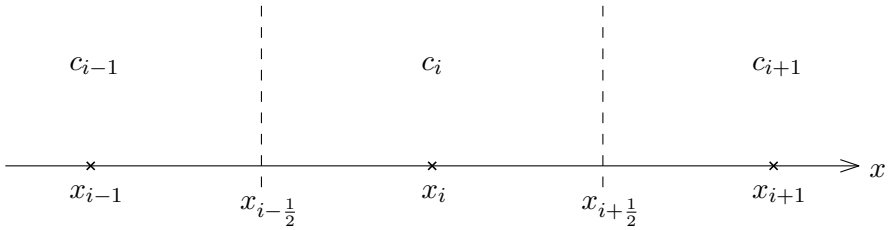

Figure 2.1 – Discretization of the one-dimensional space domain $\mathbb{R}$.

We adopt a straightforward time discretization. To discretize the time domain $\mathbb{R}_+$, we set $t^{n+1} = t^n + \Delta t^n$, with $t^0 = 0$. Note that the time step Δt^n depends on the current time t^n. For the sake of simplicity in the notations, we do not explicitly write this dependence. Instead, the time step is labeled by Δt.

Equipped with the time and space discretization, we turn to discretizing the solution of (2.1). For $x \in c_i$ and $t \in [t^n, t^{n+1})$, we choose to approximate the exact solution $W(t,x)$ of (2.1) by a constant value W_i^n. Actually, this value corresponds to the average of the exact solution $W(t,x)$ at time t^n over the cell c_i, as follows:

$$W_i^n \simeq \frac{1}{\Delta x} \int_{c_i} W(t^n, x)\, dx. \tag{2.5}$$

This approximation is initialized by taking the average of the initial condition over each cell:

$$\forall i \in \mathbb{Z}, \quad W_i^0 = \frac{1}{\Delta x} \int_{c_i} W_0(x) dx.$$

The goal of a numerical scheme is, knowing W_i^n for all $i \in \mathbb{Z}$, to give a value to the updated approximation W_i^{n+1} for all $i \in \mathbb{Z}$. In order to provide such an approximation, let us write the average of the conservation law (2.1) over the rectangle formed by the cell $c_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$

and the time interval $[t^n, t^{n+1})$:

$$\frac{1}{\Delta x}\frac{1}{\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{t^n}^{t^{n+1}} \partial_t W(t,x)\,dt\,dx + \frac{1}{\Delta x}\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \partial_x F(W(t,x))\,dx\,dt = 0.$$

Using (2.5) and performing straightforward computations in the above identity yields:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left(\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} F(W(t,x_{i+\frac{1}{2}}))\,dt - \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} F(W(t,x_{i-\frac{1}{2}}))\,dt\right). \tag{2.6}$$

The main issue with the expression (2.6) of the updated state W_i^{n+1} is that the time integrals of the physical flux function are difficult to evaluate in practice. Indeed, the flux function F may be strongly nonlinear (in the case of the shallow-water equations or the Euler equations for instance). To address this issue, we introduce the *numerical flux*. It is an approximation of the time integral of the physical flux, as follows:

$$\mathcal{F}_{i+\frac{1}{2}}^n \simeq \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} F(W(t,x_{i+\frac{1}{2}}))\,dt.$$

The value of this numerical flux depends on the value of the physical flux F at the interface $x_{i+\frac{1}{2}}$ between the cells c_i and c_{i+1}, which respectively contain the constant values W_i^n and W_{i+1}^n. Therefore, the numerical flux can be viewed as a function $\mathcal{F}$ such that $\mathcal{F}_{i+\frac{1}{2}}^n = \mathcal{F}(W_i^n, W_{i+1}^n)$. Examples of such functions will be provided in the next two subsections, devoted respectively to Godunov's scheme and to Godunov-type schemes. Equipped with the numerical flux function $\mathcal{F}$, we state the final expression of the 1D first-order finite volume scheme:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left(\mathcal{F}_{i+\frac{1}{2}}^n - \mathcal{F}_{i-\frac{1}{2}}^n\right). \tag{2.7}$$

Such a scheme is easily shown to satisfy an essential property, the conservation property. Indeed, the scheme is said to be *conservative* if the following equality holds:

$$\forall n \in \mathbb{N},\ \sum_{i\in\mathbb{Z}} W_i^{n+1}\Delta x = \sum_{i\in\mathbb{Z}} W_i^n \Delta x. \tag{2.8}$$

The scheme (2.7) indeed satisfies this property, since the sum over $\mathbb{Z}$ of the difference $\mathcal{F}_{i+\frac{1}{2}}^n - \mathcal{F}_{i-\frac{1}{2}}^n$ vanishes for all $n \in \mathbb{N}$. This property is required for the scheme to capture the correct shock waves.

Another property that we require the scheme to satisfy is the consistency property. The numerical flux function, and therefore the scheme, is said to be *consistent* with (2.1) if it satisfies

$$\forall W \in \Omega,\ \ \mathcal{F}(W,W) = F(W). \tag{2.9}$$

An important ingredient in designing a numerical flux is to make sure that this property is verified. Otherwise, the scheme approximates the wrong equations.

A third important property is the robustness. The scheme (2.7) is *robust* if we have, for all

$n \in \mathbb{N}$, the following discrete analogue of (2.2):

$$\text{if, } \forall i \in \mathbb{Z},\ W_i^n \in \Omega, \text{ then } \forall i \in \mathbb{Z},\ W_i^{n+1} \in \Omega.$$

This property ensures that the physical admissibility of the initial condition is preserved by the scheme. A non-robust scheme may therefore yield an approximate solution that is not physically admissible. This property is another important ingredient in the design of a numerical flux.

Equipped with the finite volume scheme (2.7), we now introduce Godunov's method and Godunov-type methods. These methods offer a way of defining the numerical flux function.

2.1.2 Godunov's scheme

Let us recall that the approximate solution at time t^n is made of piecewise constant values W_i^n on each cell c_i. Therefore, in a neighborhood of each interface $x_{i+\frac{1}{2}}$, the conservation law (2.1) reads as follows:

$$\begin{cases} \partial_t W + \partial_x F(W) = 0, \\ W(t^n, x) = \begin{cases} W_i^n \text{ if } x < x_{i+\frac{1}{2}}, \\ W_{i+1}^n \text{ if } x > x_{i+\frac{1}{2}}. \end{cases} \end{cases} \tag{2.10}$$

The initial value problem (2.10) is nothing but a Riemann problem (2.19). We suppose that the exact solution $W_{\mathcal{R}}$ of the Riemann problem (2.10) is known (see [81, 150] for instance). This solution is self-similar and depends on W_i^n and W_{i+1}^n. For $t \in (0, \Delta t)$ and $x \in [x_i, x_{i+1}]$, we adopt the following notation for the exact Riemann solution:

$$W_{\mathcal{R}}^{i+\frac{1}{2}}\left(\frac{x - x_{i+\frac{1}{2}}}{t}\right) = W_{\mathcal{R}}\left(\frac{x - x_{i+\frac{1}{2}}}{t}; W_i^n, W_{i+1}^n\right).$$

Since the system is hyperbolic, the velocity of the waves originating from the Riemann problem is finite. Let us denote by $\lambda_{i+\frac{1}{2}}^- < \lambda_{i+\frac{1}{2}}^+$ the smallest and greatest wave velocities, respectively. Within the fan formed by the extremal wave speeds $\lambda_{i+\frac{1}{2}}^-$ and $\lambda_{i+\frac{1}{2}}^+$ lies the exact solution of the Riemann problem (2.10). This situation is illustrated by Figure 2.2.

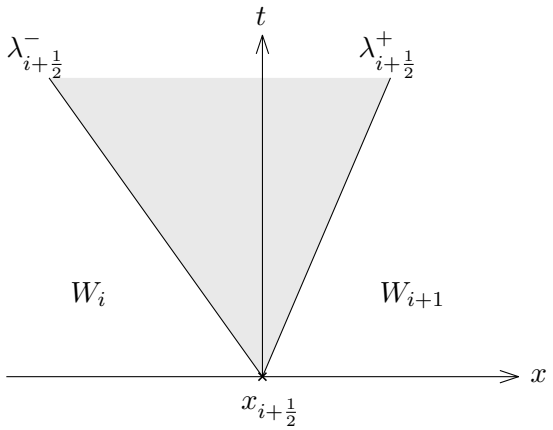

Figure 2.2 – Riemann problem configuration. The gray area represents the area where the solution of the Riemann problem (2.10) lies.

As mentioned above, since the equation (2.1) is hyperbolic, the information propagated by the equation travels at finite speed. Let us therefore emphasize that that the consecutive Riemann problem solutions do not interact as long as t is small enough. We can thus give a sufficient condition on the time step Δt to ensure that it is small enough to prevent interactions between the waves from two consecutive Riemann problems. An illustration of such a sufficient condition is presented on Figure 2.3.

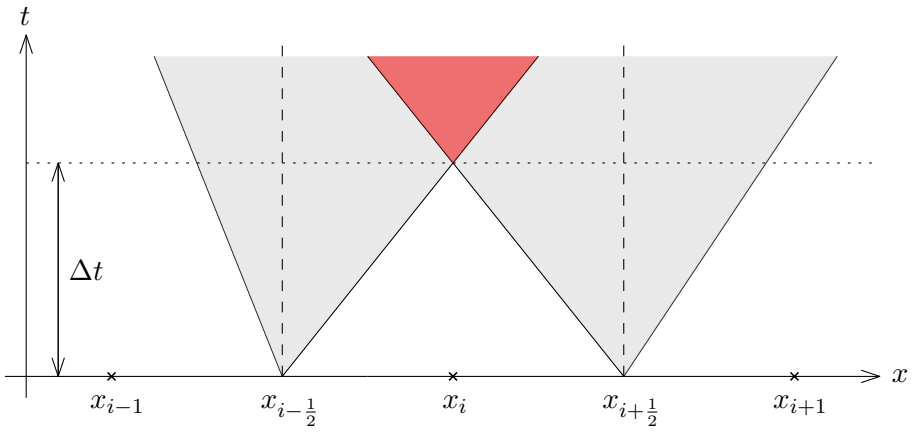

Figure 2.3 – Wave interaction to be prevented by the CFL condition (in red). The time step Δt is chosen so as to prevent the interaction.

From Figure 2.3, we deduce the following sufficient condition on Δt to prevent the interactions between waves:

$$\frac{\Delta t}{\Delta x} \max_{i\in\mathbb{Z}} \left(\left| \lambda^-_{i+\frac{1}{2}} \right|, \left| \lambda^+_{i+\frac{1}{2}} \right| \right) \leq \frac{1}{2}. \tag{2.11}$$

This condition is called the *Courant-Friedrichs-Lewy* (CFL) condition (see [55]). For all $i \in \mathbb{Z}$, it ensures that the waves from the Riemann problem located at $x_{i+\frac{1}{2}}$ do not penetrate within the cell (x_{i-1}, x_i) or the cell (x_{i+1}, x_{i+2}), thus preventing them from interacting with waves coming from neighboring Riemann problems.

The final ingredient we need to introduce Godunov's scheme is the following function W^Δ, which contains the juxtaposition of all the exact Riemann solutions:

$$\forall t \in (0, \Delta t]\,,\ \forall x \in [x_i, x_{i+1})\,,\ W^\Delta(t^n + t, x) = W_{\mathcal{R}}^{i+\frac{1}{2}}\left(\frac{x - x_{i+\frac{1}{2}}}{t} \right). \tag{2.12}$$

This function corresponds to the exact Riemann solution over the whole space domain, obtained from the initial condition $W^\Delta(t^n, x) = W_i^n \mathbb{1}_{c_i}(x)$. This juxtaposition function is displayed on Figure 2.4.

The main idea behind Godunov's scheme consists in noting that, for $t < \Delta t$, two consecutive Riemann solutions will not interact. Therefore, the exact solution of the Riemann problem (2.10) can be used in order to build a numerical flux (see [81]). The exact Riemann solution $W^\Delta(t^{n+1}, x)$ allows to define an updated approximate solution W_i^{n+1} within the cell c_i. However, in order to apply the same procedure at time t^{n+1}, the updated solution W_i^{n+1} must be constant in each cell c_i. Therefore, Godunov suggested to define W_i^{n+1} as the average of the juxtaposition $W^\Delta(t^{n+1}, x)$ of the exact Riemann problem solutions within the cell c_i, as

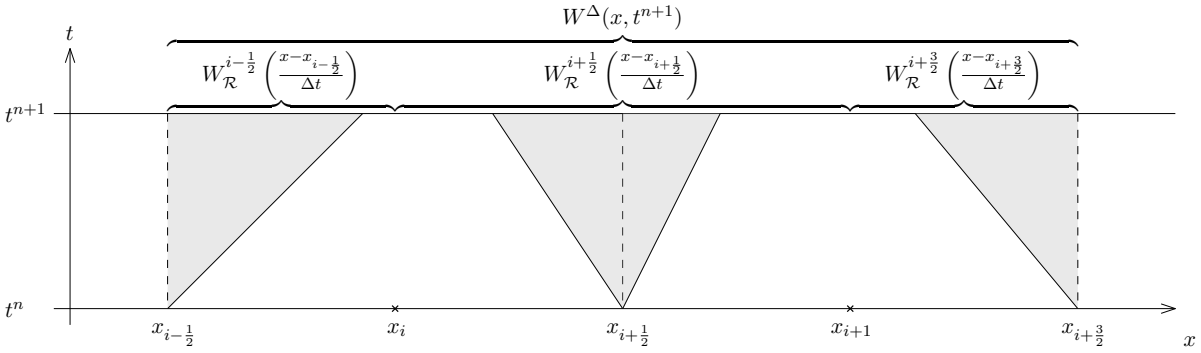

Figure 2.4 – Juxtaposition of exact Riemann solutions.

follows:

$$W_i^{n+1} = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W^{\Delta}(t^{n+1},x)\,dx. \tag{2.13}$$

Note that a wave with velocity λ travels a distance $\lambda\Delta t$ in a time Δt. Therefore, following Figure 2.4, the formula (2.13) for the updated approximated solution can be rewritten as follows:

$$\begin{aligned} W_i^{n+1} = &\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i-\frac{1}{2}}+\lambda^+_{i-\frac{1}{2}}\Delta t} W_{\mathcal{R}}^{i-\frac{1}{2}}\left(\frac{x-x_{i-\frac{1}{2}}}{t^{n+1}-t^n}\right)dx \\ &+\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}+\lambda^+_{i-\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}+\lambda^-_{i+\frac{1}{2}}\Delta t} W_i^n\,dx \\ &+\frac{1}{\Delta x}\int_{x_{i+\frac{1}{2}}+\lambda^-_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} W_{\mathcal{R}}^{i+\frac{1}{2}}\left(\frac{x-x_{i+\frac{1}{2}}}{t^{n+1}-t^n}\right)dx. \end{aligned} \tag{2.14}$$

In order to finalize this subsection devoted to Godunov's scheme, let us show that Godunov's scheme is conservative and consistent. To address this issue, we exhibit the numerical flux function associated to this scheme by computing the integrals of the exact Riemann solution present in (2.14). Arguing that the exact Riemann solution $W_{\mathcal{R}}^{i\pm\frac{1}{2}}$ is a solution of the conservation law (2.1) and integrating (2.1) over the rectangle $[t^n,t^{n+1}]\times[x_{i-\frac{1}{2}},x_{i-\frac{1}{2}}+\lambda^+_{i-\frac{1}{2}}\Delta t]$ yields:

$$\begin{aligned} &\int_{x_{i-\frac{1}{2}}}^{x_{i-\frac{1}{2}}+\lambda^+_{i-\frac{1}{2}}\Delta t}\left[W_{\mathcal{R}}^{i-\frac{1}{2}}\left(\frac{x-x_{i-\frac{1}{2}}}{t^{n+1}-t^n}\right)-W_i^n\right]dx \\ &+\int_{t^n}^{t^{n+1}}\left[F\left(W_{\mathcal{R}}^{i-\frac{1}{2}}\left(\frac{\lambda^+_{i-\frac{1}{2}}\Delta t}{t}\right)\right)-F\left(W_{\mathcal{R}}^{i-\frac{1}{2}}(0)\right)\right]dt=0 \end{aligned}$$

Note from Figure 2.3 that, for all $t\in[0,\Delta t]$ and for $x = x_{i-\frac{1}{2}}+\lambda^+_{i-\frac{1}{2}}\Delta t$, the exact solution of the Riemann problem is constant and equal to W_i^n. As a consequence, the first integral of (2.14) reads:

$$\begin{aligned} \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i-\frac{1}{2}}+\lambda^+_{i-\frac{1}{2}}\Delta t} W_{\mathcal{R}}^{i-\frac{1}{2}}\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta t}\right)dx = &\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i-\frac{1}{2}}+\lambda^+_{i-\frac{1}{2}}\Delta t} W_i^n\,dx \\ &-\frac{\Delta t}{\Delta x}\left(F(W_i^n))-F\left(W_{\mathcal{R}}^{i-\frac{1}{2}}(0)\right)\right). \end{aligned} \tag{2.15}$$

Similarly, the third integral of (2.13) is evaluated as follows:

$$\frac{1}{\Delta x}\int_{x_{i+\frac{1}{2}}+\lambda^-_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} W_{\mathcal{R}}^{i+\frac{1}{2}}\left(\frac{x-x_{i+\frac{1}{2}}}{\Delta t}\right)dx = \frac{1}{\Delta x}\int_{x_{i+\frac{1}{2}}+\lambda^-_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} W_i^n\,dx - \frac{\Delta t}{\Delta x}\left(F\left(W_{\mathcal{R}}^{i+\frac{1}{2}}(0)\right)-F(W_i^n)\right). \tag{2.16}$$

Plugging (2.15) and (2.16) into (2.14) yields:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left(F\left(W_{\mathcal{R}}^{i+\frac{1}{2}}(0)\right)-F\left(W_{\mathcal{R}}^{i-\frac{1}{2}}(0)\right)\right).$$

We have thus cast Godunov's scheme (2.14) into the conservative form (2.7), with the numerical flux function for Godunov's scheme being given by:

$$\mathcal{F}^n_{i+\frac{1}{2}} = F\left(W_{\mathcal{R}}^{i+\frac{1}{2}}(0)\right).$$

Therefore, Godunov's scheme is conservative, i.e. the property (2.8) is verified. Note that the scheme is also consistent, i.e. it satisfies (2.9). Indeed, we have $\mathcal{F}(W,W) = F(W_{\mathcal{R}}(0;W,W))$ for all $W \in \Omega$, with the quantity $W_{\mathcal{R}}(0;W,W)$ representing the exact Riemann solution with a uniform initial condition equal to W. This exact Riemann solution is thus nothing but the uniform initial condition W. Hence, we have $\mathcal{F}(W,W) = F(W)$, which proves the consistency property.

To summarize, Godunov's scheme can be written as a two-step procedure. The first step, the *evolution* step, consists in computing the exact solution of the Riemann problem at each interface. The second step, the *projection* step, consists in the averaging process (2.13) to define the updated numerical approximation.

We have thus completed the introduction of Godunov's scheme. It is the most natural conservative and consistent finite volume scheme to approximate solutions of the hyperbolic problem (2.1). As previously mentioned, the most important ingredient in the definition of Godunov's scheme is the knowledge of the exact solution of the Riemann problem (2.10). Unfortunately, computing this solution at each interface and for each time step is usually too costly, or even outright impossible since the exact Riemann solution is unknown for many systems. Furthermore, even if the exact solution is used, the projection step (2.13) only allows a first-order approximation of the solution. In light of such difficulties, a natural idea, introduced at the beginning of the 1980s by Roe in [135] and Harten, Lax and van Leer in [90], is to replace the exact solution of the Riemann problem with an approximate solution. Such an approach, leading to *Godunov-type schemes*, is described in the next subsection.

2.1.3 Godunov-type schemes

The main ingredient of Godunov-type schemes is the use of an approximate solution of the Riemann problem (2.10), instead of the exact one. Thus, Godunov-type schemes consist in replacing the *exact Riemann solver* with an *approximate Riemann solver*. We first discuss the construction of such a solver, and its associated Godunov-type scheme. Then, an example of

a Godunov-type scheme is presented.

The first issue in the derivation of the approximate Riemann solver is that the minimal and maximal exact wave speeds, $\lambda^-_{i+\frac{1}{2}}$ and $\lambda^+_{i+\frac{1}{2}}$, are not known anymore. Therefore, we use approximate wave speeds $\lambda^L_{i+\frac{1}{2}}$ and $\lambda^R_{i+\frac{1}{2}}$ as an approximation of the minimal and maximal exact wave speeds, respectively. They are chosen so as to ensure that no information is lost. Thus, the fan formed by the exact wave speeds must be included within the one formed by the approximate wave speeds, which yields the following required conditions on the approximate wave speeds, illustrated on Figure 2.5:

$$\lambda^L_{i+\frac{1}{2}} < \lambda^-_{i+\frac{1}{2}} \quad \text{and} \quad \lambda^R_{i+\frac{1}{2}} > \lambda^+_{i+\frac{1}{2}}.$$

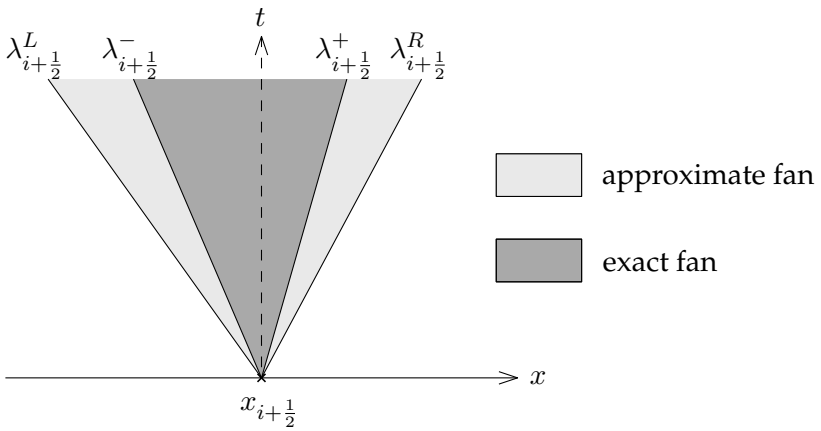

Figure 2.5 – Wave fans of the exact and approximate Riemann solvers.

We introduce the juxtaposition function for the approximate Riemann solver as follows:

$$\forall t \in (0, \Delta t]\,,\ \forall x \in [x_i, x_{i+1})\,,\ W^\Delta(t^n + t, x) = \widetilde{W}\left(\frac{x - x_{i+\frac{1}{2}}}{t}; W_i^n, W_{i+1}^n\right). \tag{2.17}$$

In (2.17), $\widetilde{W}$ represents the approximate Riemann solver. Note that, if the approximate Riemann solution is equal to the exact Riemann solution $W_{\mathcal{R}}$, then the juxtaposition function of Godunov's scheme (2.12) is recovered. Here, the approximate Riemann solver $\widetilde{W}$ provides an approximation of the exact solution $W_{\mathcal{R}}$ of the Riemann problem, and this juxtaposition function contains the approximate Riemann solution at each interface between cells.

We then define the approximate Riemann solver $\widetilde{W}$ as the following self-similar function:

$$\widetilde{W}\left(\frac{x}{t}; W_L, W_R\right) = \begin{cases} W_L & \text{if } x/t \le \lambda_L, \\ \widetilde{W}\left(\frac{x}{t}; W_L, W_R\right) & \text{if } \lambda_L < x/t < \lambda_R, \\ W_R & \text{if } x/t \ge \lambda_R. \end{cases} \tag{2.18}$$

Within the fan, i.e. for $\lambda_L < x/t < \lambda_R$, we take $\widetilde{W}$ made of $(n+1)$ intermediate states, separated by n discontinuities. We assume that all the intermediate states are constant; this choice is made in accordance with the approximate Riemann solvers suggested by Harten, Lax and van Leer in [90].

The goal is now to provide several properties that the approximate Riemann solver has to satisfy. To that end, we consider the following Riemann problem:

$$\begin{cases} \partial_t W + \partial_x F(W) = 0, \\ W(0,x) = \begin{cases} W_L & \text{if } x < 0, \\ W_R & \text{if } x > 0. \end{cases} \end{cases} \tag{2.19}$$

Note that this is the same Riemann problem as (2.10), rewritten with simpler notations and using the change of variables $x \mapsto x - x_{i+\frac{1}{2}}$. The states W_L and W_R are constant. We denote by $\lambda_L < \lambda_R$ the smallest and greatest approximate wave velocities, respectively. The approximate Riemann solution is made of at least two waves, the extremal waves λ_L and λ_R. This situation is illustrated by Figure 2.6.

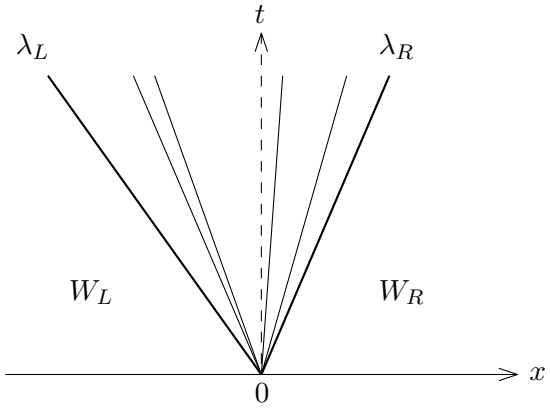

Figure 2.6 – Structure of the approximate solution of the Riemann problem (2.19). Specific case with six waves.

An approximate Riemann solver should satisfy two essential consistency properties. The first one states that $\widetilde{W}(x/t; W, W) = W$ for all $W \in \Omega$. This property is verified by the exact Riemann solver, and has to be also satisfied by the approximate Riemann solver.

In addition, in [89, 90], Harten and Lax introduced a property of *integral consistency* with the exact solution of the Riemann problem. This property reads as follows:

$$\frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_L, W_R\Big)\,dx = \frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} W_{\mathcal{R}}\Big(\frac{x}{\Delta t}; W_L, W_R\Big)\,dx. \tag{2.20}$$

We now consider an approximate Riemann solver satisfying this property. We can prove that the integral of the exact Riemann solution only depends on the left and right states. Indeed, arguing that the self-similar exact solution of (2.19) satisfies (2.1), we integrate (2.1) over the rectangle $[-\Delta x/2, \Delta x/2] \times [0, \Delta t]$, to get:

$$\begin{aligned} \int_{-\Delta x/2}^{\Delta x/2} W_{\mathcal{R}}\Big(\frac{x}{\Delta t}; W_L, W_R\Big)\,dx = &\int_{-\Delta x/2}^{\Delta x/2} W(0,x)\,dx \\ &- \int_0^{\Delta t} F\left(W_{\mathcal{R}}\left(\frac{\Delta x}{2t}; W_L, W_R\right)\right)dt \\ &+ \int_0^{\Delta t} F\left(W_{\mathcal{R}}\left(-\frac{\Delta x}{2t}; W_L, W_R\right)\right)dt. \end{aligned} \tag{2.21}$$

The initial condition of the Riemann problem (2.19) immediately yields

$$\int_{-\Delta x/2}^{\Delta x/2} W(0,x)\,dx = \frac{\Delta x}{2}(W_L + W_R). \tag{2.22}$$

In order to compute the flux integrals, we require the knowledge of the value of the exact Riemann solution along the $x = \pm\Delta x/2$ lines, for all $t \in [0, \Delta t]$. A sufficient CFL condition on Δt, which ensures that the exact Riemann solution is uniform along these lines, is exhibited on Figure 2.7.

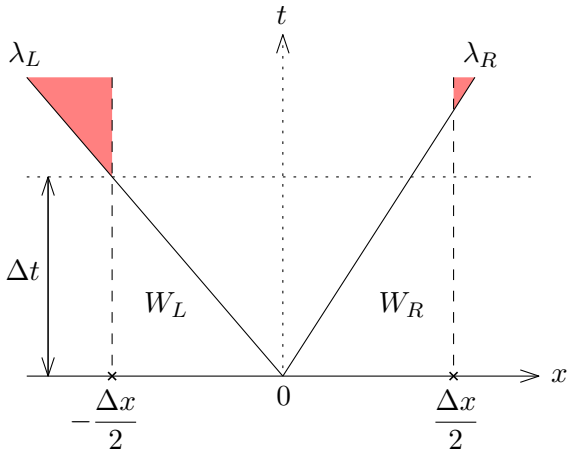

Figure 2.7 – CFL condition for Godunov-type schemes. The time step Δt is chosen to ensure that the exact Riemann solution is uniform along the $x = \pm\Delta x/2$ lines.

From Figure 2.7, the CFL condition on the time step Δt reads as follows:

$$\frac{\Delta t}{\Delta x}\max(|\lambda_L|, |\lambda_R|) \leq \frac{1}{2}. \tag{2.23}$$

Note that this condition is analogous to the CFL condition (2.11) exhibited for Godunov's scheme. This CFL condition has been chosen such that, for all $t \in (0, \Delta t]$,

$$W_{\mathcal{R}}\left(-\frac{\Delta x}{2t}; W_L, W_R\right) = W_L \quad \text{and} \quad W_{\mathcal{R}}\left(\frac{\Delta x}{2t}; W_L, W_R\right) = W_R. \tag{2.24}$$

Plugging (2.22) and (2.24) into (2.21) yields the following expression of the average of the solution to the Riemann problem (2.19):

$$\frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} W_{\mathcal{R}}\Big(\frac{x}{\Delta t}; W_L, W_R\Big)dx = \frac{W_L + W_R}{2} - \frac{\Delta t}{\Delta x}(F(W_R) - F(W_L)).$$

Therefore, the integral consistency condition (2.20) rewrites as follows:

$$\frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_L, W_R\Big)dx = \frac{W_L + W_R}{2} - \frac{\Delta t}{\Delta x}(F(W_R) - F(W_L)). \tag{2.25}$$

Equipped with a approximate Riemann solver $\widetilde{W}$ satisfying the integral consistency, we are able to define the Godunov-type scheme. Recall that, for a finite volume scheme, a Riemann problem occurs at each interface between cells. For a Godunov-type scheme, this Rie-

mann problem is approximately solved using the approximate Riemann solver $\widetilde{W}$. To that end, we use the juxtaposition function W^Δ introduced by (2.17). The Godunov-type scheme is then defined by averaging W^Δ over the cell c_i and at time $t^n + \Delta t$, as follows:

$$W_i^{n+1} = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W^\Delta(t^{n+1}, x)\, dx. \tag{2.26}$$

Note that this expression is similar to (2.13), but with the juxtaposition function W^Δ corresponding to the approximate Riemann solver $\widetilde{W}$, which is made of constant states separated by waves whose velocities are known. Therefore, the update formula (2.26) involves an integral that can always be computed explicitly, given an approximate Riemann solver.

We end this presentation of Godunov-type schemes by proving that the scheme (2.26) can be written under a conservative form. By definition (2.17) of the juxtaposition function W^Δ, the expression (2.26) rewrites as follows:

$$\begin{aligned} W_i^{n+1} = {} & \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_i} \widetilde{W}\Big(\frac{x - x_{i-\frac{1}{2}}}{\Delta t}; W_{i-1}^n, W_i^n\Big)\, dx \\ & + \frac{1}{\Delta x} \int_{x_i}^{x_{i+\frac{1}{2}}} \widetilde{W}\Big(\frac{x - x_{i+\frac{1}{2}}}{\Delta t}; W_i^n, W_{i+1}^n\Big)\, dx. \end{aligned} \tag{2.27}$$

After using the additivity property of integrals on the second integral and arguing the changes of variables $x \mapsto x - x_{i-\frac{1}{2}}$ and $x \mapsto x - x_{i+\frac{1}{2}}$, the expression (2.27) is rewritten as:

$$\begin{aligned} W_i^{n+1} = {} & \frac{1}{\Delta x} \int_0^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_{i-1}^n, W_i^n\Big)\, dx \\ & + \frac{1}{\Delta x} \int_{-\Delta x/2}^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_i^n, W_{i+1}^n\Big)\, dx \\ & - \frac{1}{\Delta x} \int_0^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_i^n, W_{i+1}^n\Big)\, dx. \end{aligned}$$

Recall that the approximate Riemann solver $\widetilde{W}$ satisfies the integral consistency condition (2.25). Therefore, the updated state W_i^{n+1} is given by:

$$\begin{aligned} W_i^{n+1} = {} & \frac{W_i^n + W_{i+1}^n}{2} - \frac{\Delta t}{\Delta x}\big(F(W_{i+1}^n) - F(W_i^n)\big) \\ & - \frac{1}{\Delta x} \int_0^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_{i-1}^n, W_i^n\Big)\, dx + \frac{1}{\Delta x} \int_0^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_i^n, W_{i+1}^n\Big)\, dx. \end{aligned} \tag{2.28}$$

Straightforward computations within (2.28) lead to the following conservative form of the updated approximation:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\big(\mathcal{F}\big(W_i^n, W_{i+1}^n\big) - \mathcal{F}\big(W_{i-1}^n, W_i^n\big)\big), \tag{2.29}$$

where the numerical flux function $\mathcal{F}$ is given by:

$$\mathcal{F}(W_L, W_R) = F(W_R) - \frac{\Delta x}{2\Delta t} W_R + \frac{1}{\Delta t} \int_0^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_L, W_R\Big)\, dx. \tag{2.30}$$

Note that, if we had transformed the first integral instead of the second one in (2.27), the computations would have yielded the following equivalent form of the numerical flux function:

$$\overline{\mathcal{F}}(W_L, W_R) = F(W_L) + \frac{\Delta x}{2\Delta t} W_L - \frac{1}{\Delta t} \int_{-\Delta x/2}^{0} \widetilde{W}\Big(\frac{x}{\Delta t}; W_L, W_R\Big)\, dx. \tag{2.31}$$

We remark that, as soon as the integral consistency condition is satisfied by the approximate Riemann solver $\widetilde{W}$, the equality $\mathcal{F}(W_L, W_R) = \overline{\mathcal{F}}(W_L, W_R)$ holds. Since $\widetilde{W}(x/t; W, W) = W$, both forms of the numerical flux function satisfy the consistency property, i.e. $\mathcal{F}(W, W) = F(W)$ and $\overline{\mathcal{F}}(W, W) = F(W)$. Therefore, the Godunov-type schemes are conservative and consistent provided the approximate Riemann solver they are based on satisfies both the integral consistency property and $\widetilde{W}(x/t; W, W) = W$.

To conclude this section on Godunov-type schemes, we derive the HLL scheme, based on an approximate Riemann solver with one intermediate state. This scheme has been suggested by Harten, Lax and van Leer in 1983 (see [90]). In the current framework of one constant intermediate state, the approximate Riemann solver (2.18) rewrites as follows:

$$\widetilde{W}\Big(\frac{x}{t}; W_L, W_R\Big) = \begin{cases} W_L & \text{if } x/t \le \lambda_L, \\ W_{HLL} & \text{if } \lambda_L < x/t < \lambda_R, \\ W_R & \text{if } x/t \ge \lambda_R, \end{cases} \tag{2.32}$$

where W_{HLL} denotes the value of the constant intermediate state. This approximate Riemann solver is displayed on Figure 2.8.

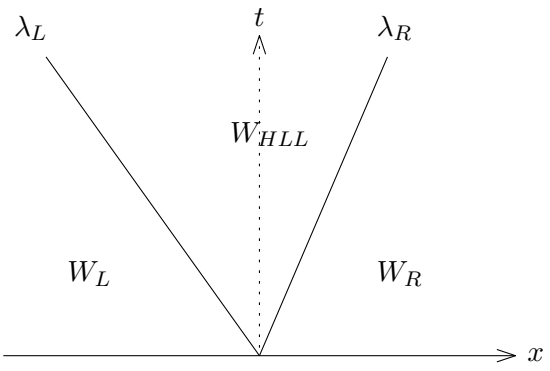

Figure 2.8 – Structure of the approximate Riemann solver (2.32).

The intermediate state W_{HLL} is determined in order for the approximate Riemann solver to satisfy the integral consistency property (2.25). We first consider the specific case where $\lambda_L < 0 < \lambda_R$. From (2.32), the average of $\widetilde{W}$ over $[-\Delta x/2, \Delta x/2]$ satisfies the following sequence of equalities:

$$\begin{aligned} \frac{1}{\Delta x} \int_{-\Delta x/2}^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_L, W_R\Big) dx &= \frac{1}{\Delta x} \left(\int_{-\Delta x/2}^{\lambda_L \Delta t} W_L\, dx + \int_{\lambda_L \Delta t}^{\lambda_R \Delta t} W_{HLL}\, dx + \int_{\lambda_R \Delta t}^{\Delta x/2} W_R\, dx \right) \\ &= \frac{W_L + W_R}{2} + \frac{\Delta t}{\Delta x} (\lambda_L W_L - \lambda_R W_R + (\lambda_R - \lambda_L) W_{HLL}). \end{aligned}$$

Using both the above equality and the integral consistency (2.25) immediately yields the fol-

lowing expression of W_{HLL}:

$$W_{HLL} = \frac{\lambda_R W_R - \lambda_L W_L}{\lambda_R - \lambda_L} - \frac{F(W_R) - F(W_L)}{\lambda_R - \lambda_L}. \tag{2.33}$$

Equipped with the expression of W_{HLL}, we can now compute the numerical flux function associated to the HLL scheme. Let us denote this function, which depends on W_L and W_R, by $\mathcal{F}_{HLL}$. Its expression is given by (2.30) (or, equivalently, (2.31)). We also denote $F(W_L)$ by W_L and $F(W_R)$ by W_R. Since the approximate Riemann solver is here defined by (2.32), the numerical flux $\mathcal{F}_{HLL}$ reads as follows:

$$\mathcal{F}_{HLL}(W_L, W_R) = F_R - \frac{\Delta x}{2\Delta t} W_R + \lambda_R W_{HLL} + \left(\frac{\Delta x}{2\Delta t} - \lambda_R\right) W_R. \tag{2.34}$$

Plugging the expression (2.33) of W_{HLL} into (2.34) yields, after straightforward computations:

$$\mathcal{F}_{HLL}(W_L, W_R) = \frac{\lambda_R F_L - \lambda_L F_R}{\lambda_R - \lambda_L} + \frac{\lambda_R \lambda_L (W_R - W_L)}{\lambda_R - \lambda_L}.$$

Recall that we have determined $\mathcal{F}_{HLL}$ in the specific case where $\lambda_L < 0 < \lambda_R$. On the one hand, in the case where $\lambda_L > 0$, since $\widetilde{W}(x/\Delta t; W_L, W_R) = W_L$ for $x/\Delta t < 0$, using the form (2.31) of the numerical flux immediately yields $\mathcal{F}_{HLL}(W_L, W_R) = F_L$. On the other hand, if $\lambda_R < 0$, using the form (2.30) yields $\mathcal{F}_{HLL}(W_L, W_R) = F_R$. Therefore, the numerical flux of the HLL scheme is given as follows:

$$\mathcal{F}_{HLL}(W_L, W_R) = \begin{cases} F_L & \text{if } \lambda_L \geq 0, \\ \dfrac{\lambda_R F_L - \lambda_L F_R}{\lambda_R - \lambda_L} + \dfrac{\lambda_R \lambda_L (W_R - W_L)}{\lambda_R - \lambda_L} & \text{if } \lambda_L < 0 < \lambda_R, \\ F_R & \text{if } \lambda_R \leq 0. \end{cases} \tag{2.35}$$

The goal of this manuscript is to derive a consistent, robust and well-balanced scheme for the shallow-water equations. Since the shallow-water system is hyperbolic, using a finite volume scheme based on an approximate Riemann solver is a suitable choice. However, the approximate Riemann solver we use cannot possess only one state. Indeed, for a one-state Riemann solver, merely arguing the consistency property yields the HLL scheme. As a consequence, we choose a two-state approximate Riemann solver to introduce more unknown intermediate states and recover the well-balance property. In addition, the two-state structure is in good agreement with the exact Riemann solution, discussed in Section 1.1.3, which possesses two intermediate states separated by a stationary contact discontinuity. This approximate Riemann solver will be derived in Chapter 3. However, other ingredients are required to enhance the scheme. These ingredients are discussed in the remainder of this chapter.

2.2 Second-order space accuracy in one dimension

After having derived first-order finite volume schemes in the previous section, we now turn to providing a second-order extension of these schemes. The purpose of such an exten-

sion is to improve the spatial order of accuracy of a scheme. The order of accuracy measures the rate at which the numerical approximation converges towards the exact solution as Δx diminishes. We define the average of the exact solution on a cell c_i as follows:

$$(W_{ex})_i^n := \frac{1}{\Delta x} \int_{c_i} W_{ex}(t^n, x) dx,$$

where $W_{ex}(t, x)$ is the exact solution of the initial value problem (2.1). The errors between the approximate solution $(W_i^n)_{i\in\mathbb{Z}}$ and the average of the exact solution $((W_{ex})_i^n)_{i\in\mathbb{Z}}$ are defined as follows:

$$L^1\text{-norm:} \quad e_1(\Delta x) = \sum_{i\in\mathbb{Z}} \Delta x \left|W_i^n - (W_{ex})_i^n\right|, \tag{2.36a}$$

$$L^2\text{-norm:} \quad e_2(\Delta x) = \left(\Delta x \sum_{i\in\mathbb{Z}} \left|W_i^n - (W_{ex})_i^n\right|^2\right)^{1/2}, \tag{2.36b}$$

$$L^\infty\text{-norm:} \quad e_\infty(\Delta x) = \max_{i\in\mathbb{Z}} \left|W_i^n - (W_{ex})_i^n\right|. \tag{2.36c}$$

Let $e \in \{e_1, e_2, e_\infty\}$ be the error in any norm. For a smooth exact solution, it is a well-known fact that, in any norm, the error between the approximate solution and the exact solution satisfies the following property when Δx tends to 0:

$$e(\Delta x) \underset{\Delta x\to 0}{=} \mathcal{O}(\Delta x^p),$$

where $e(\Delta x) > 0$ is the error for the considered space step Δx and p is the order of accuracy. We have presented first-order schemes (i.e. where $p = 1$) in the previous section. The goal is now to present second-order techniques, whose aim is ensuring that $p = 2$.

One class of second-order techniques (and high-order ones) is suited to the framework of Godunov-type schemes. They consist in replacing the piecewise constant approximation W_i^n in each cell with a piecewise linear approximation $\widehat{W}_i^n(x)$ (or piecewise polynomial in the case of higher-order schemes). This piecewise linear approximation $\widehat{W}_i^n(x)$ is called the *reconstruction*. An example of such a reconstruction, the MUSCL technique (see [112] for instance), is discussed later in this section.

We first introduce the following notations, that represent the values of the reconstruction at the inner interfaces of each cell:

$$W_i^- = \widehat{W}_i^n(x_{i-\frac{1}{2}}) \quad \text{and} \quad W_i^+ = \widehat{W}_i^n(x_{i+\frac{1}{2}}).$$

These values at the inner interfaces are then used in the numerical flux function $\mathcal{F}$ of the Godunov-type scheme. Instead of (2.29), the updated state W_i^{n+1} of the second-order scheme is given as follows:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left(\mathcal{F}\left(W_i^+, W_{i+1}^-\right) - \mathcal{F}\left(W_{i-1}^+, W_i^-\right)\right).$$

The reconstruction and the interface values are presented in Figure 2.9. In this figure, φ represents one component of the vector W, which thus lies in a subset of $\mathbb{R}$.

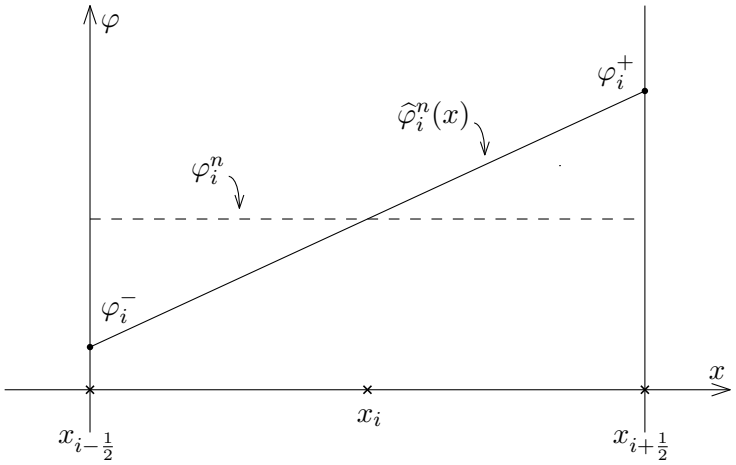

Figure 2.9 – Reconstruction within the cell c_i. The constant state φ_i^n (dashed line) is reconstructed as the linear function $\widehat{\varphi}_i^n(x)$ (solid line). The values of $\widehat{\varphi}_i^n(x)$ at the inner interfaces are denoted by φ_i^- and φ_i^+.

Many reconstruction procedures have been developed over the years. We mention the MUSCL (Monotonic Upstream-Centered Scheme for Conservation Laws) reconstruction, proposed by van Leer in [154] (see also [131, 132, 112] for instance). Another procedure, the MOOD reconstruction, provides a high-order polynomial approximation (see [46, 63, 65, 71]). We also mention the ENO (Essentially Non-Oscillatory) schemes (see [88]) and their extension, the WENO (Weighted ENO) schemes (see for instance [116, 99, 143]). Finally, the DG (Discontinuous Galerkin) method is mentioned as a high-order extension of finite volume schemes (see for instance [53, 52, 51]).

We conclude this section with a presentation of the MUSCL procedure. We assume known an approximation of the solution at time t^n, denoted by $(W_i^n)_{i\in\mathbb{Z}}$, constant in each cell c_i. The goal of the MUSCL reconstruction is to provide a linear reconstruction $\widehat{W}_i^n(x)$ of this piecewise constant approximation. For each cell c_i, this reconstruction is given by:

$$\widehat{W}_i^n(x) = W_i^n + \sigma_i^n(x - x_i),$$

where σ_i^n is the *slope* of the reconstruction. We immediately remark that $\widehat{W}_i^n(x_i) = W_i^n$, i.e. the piecewise constant approximation is recovered at the center of each cell. In addition, the values of the reconstruction at the inner interfaces satisfy:

$$W_i^{\pm} = W_i^n \pm \frac{\Delta x}{2}\sigma_i^n.$$

Now, the last ingredient we need to determine $\widehat{W}_i^n(x)$ is the slope σ_i^n. We choose σ_i^n under the following form:

$$\sigma_i^n := L\left(\frac{W_i^n - W_{i-1}^n}{\Delta x}, \frac{W_{i+1}^n - W_i^n}{\Delta x}\right),$$

where $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a function whose arguments are the slopes between the constant states on each side of both interfaces. Such a reconstruction is presented on Figure 2.10 for a component φ of the vector W.

In order to achieve the determination of the MUSCL reconstruction, we need to provide

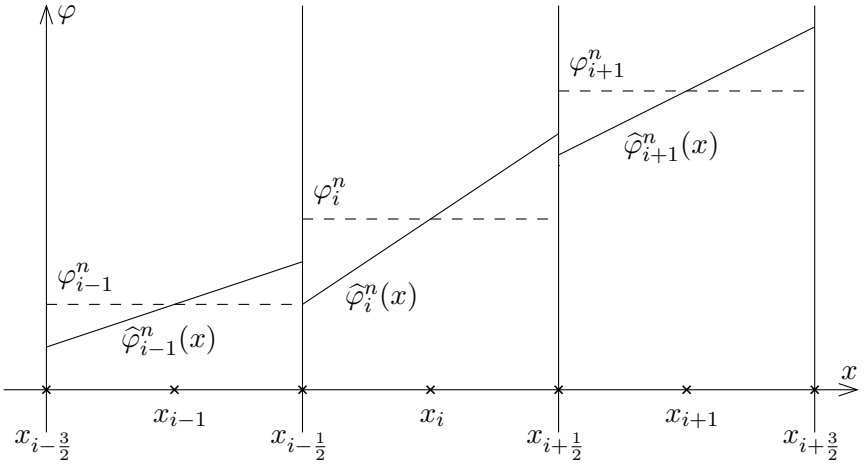

Figure 2.10 – The MUSCL reconstruction procedure. The constant states φ_i^n (dashed lines) are reconstructed to form the piecewise linear functions $\widehat{\varphi}_i^n(x)$ (solid lines).

an expression of L as a function of two slopes σ_L and σ_R. The most natural choice for L is the average of the slopes, as follows:

$$L(\sigma_L, \sigma_R) = \frac{\sigma_L + \sigma_R}{2}.$$

However, it is well-known that such a choice induces spurious oscillations. Therefore, the use of a *slope limiter* is usually suggested. The purpose of a slope limiter is to make sure that the slope σ_i^n is not too large, in order to reduce or nullify the amplitude of the oscillations. We here give a few examples of usual slope limiters:

- the *minmod limiter*: $L(\sigma_L, \sigma_R) = \text{minmod}(\sigma_L, \sigma_R)$, where

$$\text{minmod}(\sigma_L, \sigma_R) = \begin{cases} \min(\sigma_L, \sigma_R) & \text{if } \sigma_L > 0 \text{ and } \sigma_R > 0, \\ \max(\sigma_L, \sigma_R) & \text{if } \sigma_L < 0 \text{ and } \sigma_R < 0, \\ 0 & \text{otherwise;} \end{cases}$$

- the *superbee limiter*: $L(\sigma_L, \sigma_R) = \text{maxmod}(\text{minmod}(2\sigma_L, \sigma_R), \text{minmod}(\sigma_L, 2\sigma_R))$, where

$$\text{maxmod}(\sigma_L, \sigma_R) = \begin{cases} \max(\sigma_L, \sigma_R) & \text{if } \sigma_L > 0 \text{ and } \sigma_R > 0, \\ \min(\sigma_L, \sigma_R) & \text{if } \sigma_L < 0 \text{ and } \sigma_R < 0, \\ 0 & \text{otherwise;} \end{cases}$$

- the *Monotonized Central-Difference (MC) limiter*: $L(\sigma_L, \sigma_R) = \text{MC}(\sigma_L, \sigma_R)$, where

$$\text{MC}(\sigma_L, \sigma_R) = \begin{cases} \min\left(2\sigma_L, 2\sigma_R, \dfrac{\sigma_L + \sigma_R}{2}\right) & \text{if } \sigma_L > 0 \text{ and } \sigma_R > 0, \\ \max\left(2\sigma_L, 2\sigma_R, \dfrac{\sigma_L + \sigma_R}{2}\right) & \text{if } \sigma_L < 0 \text{ and } \sigma_R < 0, \\ 0 & \text{otherwise.} \end{cases}$$

2.3 Two-dimensional first-order finite volume schemes for hyperbolic problems

After having tackled the issues of first-order schemes for one-dimensional problems in Section 2.1 and second-order one-dimensional schemes in Section 2.2, we now turn to the approximation of two-dimensional hyperbolic systems of conservation laws. We consider the following initial value problem:

$$\begin{cases} \partial_t W + \boldsymbol{\nabla} \cdot \boldsymbol{F}(W) = 0, \\ W(0, \boldsymbol{x}) = W_0(\boldsymbol{x}). \end{cases} \tag{2.37}$$

Now, the space variable $\boldsymbol{x}$ lies within $\mathbb{R}^2$ instead of $\mathbb{R}$. Therefore, the physical flux $\boldsymbol{F}$ is now a function of W with values within $\mathcal{M}_{N,2}(\mathbb{R})$. The vector of conserved variables W lives in the admissible states space $\Omega \subset \mathbb{R}^N$, supposed to be convex and invariant. For more information on such systems and the approximation of their solutions, the reader is referred to [103, 112, 150] for instance.

Now, we focus on approximating solutions of such 2D systems of conservation laws. Once again, we elect to use finite volume schemes. Therefore, we start by presenting the finite volume discretization of the space domain and of the equations, to derive a 2D finite volume scheme. Then, we show that this scheme can actually be rewritten as a convex combination of 1D schemes, which ensures that some properties verified at the 1D level are still satisfied by the 2D scheme.

2.3.1 Finite volume discretization of the equations

In order to propose a numerical scheme for the 2D equations (2.37), we first need a discretization of the space domain $\mathbb{R}^2$. We elect to discretize this domain with polygonal cells c_i of center $\boldsymbol{x}_i$. Consider two neighboring cells c_i and c_j, i.e. two cells that possess a common edge. This edge is denoted by e_{ij} and the unit normal vector pointing from c_i to c_j is denoted by $\boldsymbol{n}_{ij}$. The area of the cell c_i will be denoted by $|c_i|$, and the length of the edge e_{ij} will be denoted by $|e_{ij}|$. The perimeter of the cell c_i is denoted by $|P_i|$. The notation ν_i represents the set of cells that share an edge with the cell c_i. These notations are illustrated on Figure 2.11 for a triangle mesh and on Figure 2.12 for a uniform Cartesian mesh.

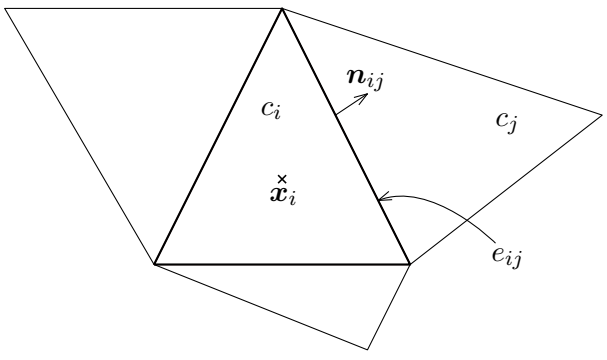

Figure 2.11 – 2D mesh made of triangles.

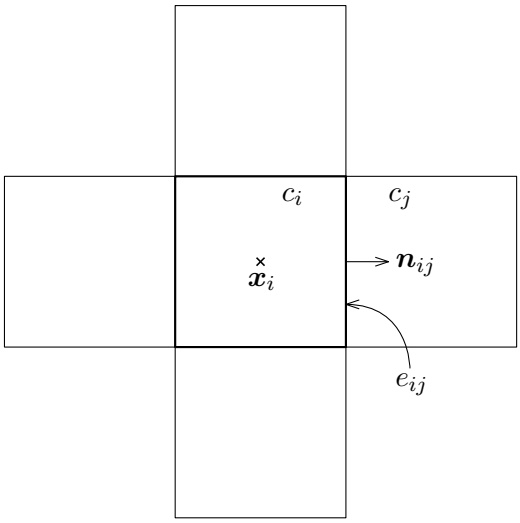

Figure 2.12 – Uniform 2D Cartesian mesh, made of squares.

Equipped with the mesh, we now derive a 2D finite volume scheme. First, we set the time step Δt. We assume that it is small; a more precise bound of the time step will be given later on. Then, the governing equations (2.37) are averaged over the cuboid $[t^n, t^{n+1}] \times c_i$, to get:

$$\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i}\int_{t^n}^{t^{n+1}} \partial_t W \, dt \, d\boldsymbol{x} + \frac{1}{|c_i|}\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\int_{c_i} \boldsymbol{\nabla}\cdot\boldsymbol{F}(W)\, d\boldsymbol{x}\, dt = 0. \tag{2.38}$$

The first integral of (2.38) satisfies the following equality:

$$\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i}\int_{t^n}^{t^{n+1}} \partial_t W \, dt \, d\boldsymbol{x} = \frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i} W(t^{n+1},\boldsymbol{x})\, d\boldsymbol{x} - \frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i} W(t^{n},\boldsymbol{x})\, d\boldsymbol{x}. \tag{2.39}$$

Now, we define the numerical approximation of the solution of (2.37) as piecewise constant on each cell. Within the cell c_i and at time t^n, this approximation is denoted by W_i^n, and it satisfies:

$$W_i^n \simeq \frac{1}{|c_i|}\int_{c_i} W(t^n,\boldsymbol{x})\, d\boldsymbol{x}.$$

As a consequence, (2.39) becomes:

$$\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i}\int_{t^n}^{t^{n+1}} \partial_t W \, dt \, d\boldsymbol{x} \simeq \frac{1}{\Delta t}\big(W_i^{n+1} - W_i^n\big). \tag{2.40}$$

We now turn to the second integral of (2.38). Arguing the divergence theorem, we have

$$\int_{c_i} \boldsymbol{\nabla}\cdot\boldsymbol{F}(W)\, d\boldsymbol{x} = \int_{\partial c_i} \boldsymbol{F}(W)\cdot\boldsymbol{n}\, d\boldsymbol{\sigma}, \tag{2.41}$$

where ∂c_i is the boundary of the cell c_i, $\boldsymbol{n}$ is the unit outer-pointing normal vector, and $d\boldsymbol{\sigma}$ is an element of length of ∂c_i. Note that the boundary of c_i satisfies the following relation:

$$\partial c_i = \bigcup_{j\in\nu_i} e_{ij}. \tag{2.42}$$

Using (2.42) within the integral in (2.41), we obtain:

$$\int_{c_i} \boldsymbol{\nabla} \cdot \boldsymbol{F}(W)\, d\boldsymbol{x} = \sum_{j \in \nu_i} \int_{e_{ij}} \boldsymbol{F}(W) \cdot \boldsymbol{n}_{ij}\, d\boldsymbol{\sigma}. \tag{2.43}$$

Substituting (2.43) into the second term of (2.38) yields:

$$\frac{1}{|c_i|} \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \int_{c_i} \boldsymbol{\nabla} \cdot \boldsymbol{F}(W)\, d\boldsymbol{x}\, dt = \frac{1}{|c_i|} \sum_{j \in \nu_i} \int_{e_{ij}} \left[\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \boldsymbol{F}(W) \cdot \boldsymbol{n}_{ij}\, dt \right] d\boldsymbol{\sigma}. \tag{2.44}$$

We remark that the integral within the brackets is nothing but the average of the physical flux function on the edge e_{ij} over time. Therefore, following the 1D case, we approximate this average with a numerical flux function $\mathcal{F}$, as follows:

$$\mathcal{F}_{ij}^n := \mathcal{F}(W_i^n, W_j^n; \boldsymbol{n}_{ij}) \simeq \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \boldsymbol{F}(W) \cdot \boldsymbol{n}_{ij}\, dt. \tag{2.45}$$

Note that the function $\mathcal{F}$ approximates the physical flux in the direction orthogonal to the edge e_{ij}. Thus, it can be viewed as a 1D numerical flux function, in the direction given by $\boldsymbol{n}_{ij}$. The numerical flux $\mathcal{F}_{ij}^n$ is then injected into (2.44), noting that $\mathcal{F}_{ij}^n$ does not depend on $\boldsymbol{\sigma}$, to get:

$$\frac{1}{|c_i|} \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \int_{c_i} \boldsymbol{\nabla} \cdot \boldsymbol{F}(W)\, d\boldsymbol{x}\, dt \simeq \sum_{j \in \nu_i} \frac{|e_{ij}|}{|c_i|} \mathcal{F}_{ij}^n. \tag{2.46}$$

Combining both equations (2.40) and (2.46) yields the following 2D first-order finite volume numerical scheme:

$$W_i^{n+1} = W_i^n - \Delta t \sum_{j \in \nu_i} \frac{|e_{ij}|}{|c_i|} \mathcal{F}_{ij}^n. \tag{2.47}$$

We now define the conservation, consistency and robustness properties. First, the robustness of the 2D scheme (2.47) is defined the same way as the robustness of the 1D scheme (2.7). Indeed, the 2D scheme is said to be robust if the following property holds:

$$\text{if, } \forall i \in \mathbb{Z},\ W_i^n \in \Omega, \text{ then } \forall i \in \mathbb{Z},\ W_i^{n+1} \in \Omega.$$

Second, the 2D numerical flux $\mathcal{F}$ is said to be consistent if it satisfies the following 2D analogue of (2.9):

$$\forall W \in \Omega,\ \forall \boldsymbol{n} \in \mathbb{R}^2,\ \mathcal{F}(W, W; \boldsymbol{n}) = \boldsymbol{F}(W) \cdot \boldsymbol{n}. \tag{2.48}$$

Third, the discrete conservation property reads:

$$\sum_{i \in \mathbb{Z}} |c_i| W_i^{n+1} = \sum_{i \in \mathbb{Z}} |c_i| W_i^n. \tag{2.49}$$

We now exhibit a sufficient condition on the numerical flux function $\mathcal{F}$ for this property to be satisfied. Plugging the value (2.47) of W_i^{n+1} within (2.49) immediately yields:

$$\sum_{i \in \mathbb{Z}} \sum_{j \in \nu_i} |e_{ij}| \mathcal{F}_{ij}^n = 0. \tag{2.50}$$

Let us perform a reindexation within the expression (2.50). Consider an interface k separating two cells c_{i_k} and c_{j_k}. We therefore have $|e_k| = |e_{i_k,j_k}| = |e_{j_k,i_k}|$. The expression (2.50) then rewrites as follows:

$$\sum_{k\in\mathbb{Z}} |e_k|\big(\mathcal{F}^n_{i_k,j_k} + \mathcal{F}^n_{j_k,i_k}\big) = 0. \tag{2.51}$$

A sufficient condition for (2.51) to be valid is that $\mathcal{F}^n_{i_k,j_k} = -\mathcal{F}^n_{j_k,i_k}$, for all $k \in \mathbb{Z}$. Performing the reverse reindexation, arguing the definition (2.45) of $\mathcal{F}$ and noting that $\boldsymbol{n}_{ij} = -\boldsymbol{n}_{ji}$ proves that this sufficient condition reads:

$$\forall i \in \mathbb{Z},\ \forall j \in \nu_i,\ \mathcal{F}\big(W^n_i, W^n_j; \boldsymbol{n}_{ij}\big) = -\mathcal{F}\big(W^n_j, W^n_i; -\boldsymbol{n}_{ij}\big).$$

This condition ensures that the numerical flux entering a cell through an edge is the opposite of the numerical flux leaving the cell though this edge. Therefore, it indeed corresponds to the conservation property.

2.3.2 2D schemes as convex combinations of 1D schemes

We have thus obtained the general form (2.47) of a finite volume scheme for a 2D conservation law. Following [132] (see also [19, 22, 17, 21]), we rewrite the 2D scheme (2.47) as a convex combination of 1D schemes. Such a process allows to easily check if properties that are valid in 1D are still satisfied in 2D. The following result states this convex combination.

Proposition 2.1. *Let $|P_i|$ be the perimeter of the cell $|c_i|$. Assume that the numerical flux of the 2D scheme (2.47) is consistent. Then, (2.47) rewrites under the following form:*

$$W^{n+1}_i = \sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|} \mathcal{W}^{n+1}_{ij}, \tag{2.52}$$

where $\mathcal{W}^{n+1}_{ij}$ is a 1D scheme in the direction given by $\boldsymbol{n}_{ij}$, given by:

$$\mathcal{W}^{n+1}_{ij} = W^n_i - \Delta t\, \frac{|P_i|}{|c_i|}\big[\mathcal{F}(W^n_i, W^n_j; \boldsymbol{n}_{ij}) - \mathcal{F}(W^n_i, W^n_i; \boldsymbol{n}_{ij})\big]. \tag{2.53}$$

Proving Proposition 2.1 means showing that the convex combination process (2.52) - (2.53) indeed yields the scheme (2.47). Let us start by noting that, by definition of the perimeter $|P_i|$, the following identity is satisfied:

$$|P_i| = \sum_{j\in\nu_i} |e_{ij}|. \tag{2.54}$$

Therefore, the combination (2.52) is indeed a convex combination, since all its coefficients are positive and their sum is equal to one. In addition, combining (2.52) and (2.53) yields:

$$W^{n+1}_i = \sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|}\left(W^n_i - \Delta t\, \frac{|P_i|}{|c_i|}\big[\mathcal{F}(W^n_i, W^n_j; \boldsymbol{n}_{ij}) - \mathcal{F}(W^n_i, W^n_i; \boldsymbol{n}_{ij})\big]\right). \tag{2.55}$$

The goal is now to prove that (2.55) holds. If that is the case, then Proposition 2.1 obviously also holds.

Proof of Proposition 2.1. The goal of this proof is to show (2.55). This equality rewrites as follows:

$$\begin{aligned} W_i^{n+1} &= \sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|} W_i^n \\ &- \Delta t \sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|}\frac{|P_i|}{|c_i|} \mathcal{F}(W_i^n, W_j^n; \boldsymbol{n}_{ij}) \\ &+ \Delta t \sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|}\frac{|P_i|}{|c_i|} \mathcal{F}(W_i^n, W_i^n; \boldsymbol{n}_{ij}). \end{aligned} \tag{2.56}$$

Using (2.54), the first sum in (2.56) rewrites:

$$\sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|} W_i^n = W_i^n \frac{\sum_{j\in\nu_i} |e_{ij}|}{|P_i|} = W_i^n. \tag{2.57}$$

Recall the definition (2.45) of the 2D numerical flux. Then, the second sum in (2.56) reads:

$$-\Delta t \sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|}\frac{|P_i|}{|c_i|} \mathcal{F}(W_i^n, W_j^n; \boldsymbol{n}_{ij}) = -\Delta t \sum_{j\in\nu_i} \frac{|e_{ij}|}{|c_i|} \mathcal{F}_{ij}^n. \tag{2.58}$$

Arguing the definition (2.45) of the 2D numerical flux and the consistency property (2.48), the third sum in (2.56) satisfies the following identity:

$$\Delta t \sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|}\frac{|P_i|}{|c_i|} \mathcal{F}(W_i^n, W_i^n; \boldsymbol{n}_{ij}) = \frac{\Delta t}{|c_i|} \sum_{j\in\nu_i} |e_{ij}| \boldsymbol{F}(W_i^n)\cdot \boldsymbol{n}_{ij}.$$

Now, we argue the divergence theorem. Since $F(W_i^n)$ is a constant, we have the following sequence of equalities:

$$\begin{aligned} \sum_{j\in\nu_i} |e_{ij}| \boldsymbol{F}(W_i^n)\cdot \boldsymbol{n}_{ij} &= \sum_{j\in\nu_i} \int_{e_{ij}} \boldsymbol{F}(W_i^n)\cdot \boldsymbol{n}_{ij}\, d\boldsymbol{\sigma} \\ &= \int_{\partial c_i} \boldsymbol{F}(W_i^n)\cdot \boldsymbol{n}\, d\boldsymbol{\sigma} \\ &= \int_{c_i} \boldsymbol{\nabla}\cdot \boldsymbol{F}(W_i^n)\, d\boldsymbol{x}. \end{aligned}$$

Since $\nabla \cdot \boldsymbol{F}(W_i^n) = 0$, the third sum in (2.56) vanishes, as follows:

$$\Delta t \sum_{j\in\nu_i} \frac{|e_{ij}|}{|c_i|} \mathcal{F}(W_i^n, W_i^n; \boldsymbol{n}_{ij}) = 0. \tag{2.59}$$

Combining the three evaluations (2.57) – (2.58) – (2.59) yields:

$$\sum_{j\in\nu_i} \frac{|e_{ij}|}{|P_i|} \left(W_i^n - \Delta t \frac{|P_i|}{|c_i|} [\mathcal{F}(W_i^n, W_j^n; \boldsymbol{n}_{ij}) - \mathcal{F}(W_i^n, W_i^n; \boldsymbol{n}_{ij})] \right) = W_i^n - \Delta t \sum_{j\in\nu_i} \frac{|e_{ij}|}{|c_i|} \mathcal{F}_{ij}^n.$$

Arguing (2.47), the right-hand side of the above expression is equal to W_i^{n+1}.

As a consequence, (2.55) holds and the convex combination (2.52) – (2.53) indeed allows the recovery of the 2D scheme (2.47). The proof is thus concluded. □

Corollary 2.2. *The convex combination process from Proposition 2.1 allows giving an upper bound for the time step Δt. It is constrained by the following CFL condition:*

$$\Delta t \max_{i\in\mathbb{Z}}\left(\frac{|P_i|}{|c_i|}\max_{j\in\nu_i}\left(|\lambda_{ij}^-|,|\lambda_{ij}^+|\right)\right)\leq\frac{1}{2}, \tag{2.60}$$

where λ_{ij}^- and λ_{ij}^+ respectively represent the minimum and maximum approximate wave speeds at the interface between the cells c_i and c_j. These two quantities depend on W_i^n and W_j^n.

Proof. The proof of this result relies on noticing that the time step of the 1D scheme (2.53) has to be constrained with a CFL-like condition. This CFL condition is nothing but the equation (2.60), which is the analogue of the 1D CFL condition (2.11). Thus, the proof is achieved. □

The main interest of such a convex combination process is to easily prove some properties on the 2D scheme with the mere knowledge of the 1D numerical flux function. Indeed, for instance, recall that the admissible states space Ω is assumed to be convex. Therefore, if the 1D scheme is robust, then the 2D scheme is also robust.

2.4 Two-dimensional high-order finite volume schemes

We now discuss the high-order extension of the first-order 2D scheme (2.47). Recall the definition (2.2) of the order of accuracy. Classical MUSCL techniques (see Section 2.2) may be applied to obtain a second-order space accuracy, i.e. $p = 2$. However, we focus here on high-order schemes, i.e. schemes with order $p \geq 3$. The MUSCL scheme used a piecewise linear reconstruction; high-order schemes require a piecewise polynomial reconstruction. Such schemes produce a better approximation of the exact solution, but also induce spurious oscillations, similarly to the 1D MUSCL case. As a consequence, specific techniques are required to prevent these oscillations. This section is dedicated to deriving a high-order finite volume scheme, and to presenting an oscillation prevention technique, the MOOD method (see [46, 63, 65]).

We begin by presenting the polynomial reconstruction procedure. Then, we derive a finite volume scheme that is high-order accurate in both space and time. Finally, we mention the MOOD method, which is a procedure to choose the optimal degree of the polynomial reconstruction, and ensure that some robustness properties are satisfied.

2.4.1 The polynomial reconstruction

In this subsection, we follow [46, 63, 65] to present a high-order polynomial reconstruction. Consider a component φ of the vector W. At this level, in each cell c_i of the mesh, we know constant values φ_i^n, which represent approximations of φ in the cell c_i and at time t^n. Within each cell c_i and at time t^n, we seek an expression $\hat{\varphi}_i^n(\boldsymbol{x};d)$ that is a polynomial of degree d, and that correctly approximates the solution of (2.37) within the cell c_i. A polynomial reconstruction of degree d will allow a spatial accuracy of order $d+1$ (recall the second-order MUSCL technique, where a linear reconstruction was applied).

We begin by requiring that the polynomial $\hat{\varphi}_i^n(\boldsymbol{x}; d)$ satisfies the following essential conservation property:

$$\frac{1}{|c_i|}\int_{c_i} \hat{\varphi}_i^n(\boldsymbol{x}; d)\, d\boldsymbol{x} = \varphi_i^n. \tag{2.61}$$

Therefore, we elect to use the following form for the polynomial reconstruction:

$$\hat{\varphi}_i^n(\boldsymbol{x}; d) = \varphi_i^n + \sum_{|\alpha|\in[\![1,d]\!]} R_i^\alpha((\boldsymbol{x}-\boldsymbol{x}_i)^\alpha - M_i^\alpha), \tag{2.62}$$

where we have defined:

- $\alpha \in \mathbb{N}^2$, a multi-index whose length is denoted by $|\alpha| = \alpha_1 + \alpha_2$;
- the usual notation $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_1}$;
- $R_i = (R_i^\alpha)_{|\alpha|\in[\![1,d]\!]}$, the unknown coefficients of the polynomial;
- M_i^α such that $M_i^\alpha = \dfrac{1}{|c_i|}\displaystyle\int_{c_i} (\boldsymbol{x}-\boldsymbol{x}_i)^\alpha d\boldsymbol{x}$.

Thanks to the presence of M_i^α, the expression (2.62) is immediately proven to satisfy the conservation property (2.61). In addition, it is worth noting that the definition (2.62) of $\hat{\varphi}_i^n(\boldsymbol{x}; d)$ does not ensure that $\hat{\varphi}_i^n(\boldsymbol{x}_i; d) = \varphi_i^n$, unlike in the 1D MUSCL case.

The last step in the full determination of the polynomial reconstruction is finding a value for the polynomial coefficients R_i. We begin by taking a stencil s_i^d, formed of N_d cells around each cell c_i, and which does not contain the cell c_i. The determination of the optimal stencil for a given mesh is still an open problem; therefore, for the moment, we do not explicitly give the stencil s_i^d of a given cell c_i. A lower bound of its size N_d will be provided shortly. The goal of this stencil is to provide a set in which the cells are considered close enough to the cell c_i to be used in the polynomial approximation of the solution in c_i. Following [46, 63, 65], we compute the polynomial coefficients R_i such that they minimize the least squares error between the reconstruction and the values φ_j^n of the piecewise constant approximation in the stencil cells $c_j \in s_i^d$. This condition is nothing but the minimization of the following functional:

$$E_i(R_i) = \frac{1}{2}\sum_{j\in s_i^d}\left[\frac{1}{|c_j|}\int_{c_j} \hat{\varphi}_i^n(\boldsymbol{x}; d)\, d\boldsymbol{x} - \varphi_j^n\right]^2.$$

Let us note that arguing the definition (2.62) of $\hat{\varphi}_i^n$ yields:

$$\frac{1}{|c_j|}\int_{c_j} \hat{\varphi}_i^n(\boldsymbol{x}; d)\, d\boldsymbol{x} = \varphi_i^n + \sum_{|\alpha|\in[\![1,d]\!]} R_i^\alpha\left(\frac{1}{|c_j|}\int_{c_j}(\boldsymbol{x}-\boldsymbol{x}_i)^\alpha\, d\boldsymbol{x} - M_i^\alpha\right).$$

Therefore, the functional E_i rewrites as follows:

$$E_i(R_i) = \frac{1}{2}\sum_{j\in s_i^d}\left[\sum_{|\alpha|\in[\![1,d]\!]} R_i^\alpha\left(\frac{1}{|c_j|}\int_{c_j}(\boldsymbol{x}-\boldsymbol{x}_i)^\alpha\, d\boldsymbol{x} - M_i^\alpha\right) + \varphi_i^n - \varphi_j^n\right]^2.$$

Minimizing E_i is therefore equivalent to minimizing the following L^2-norm:

$$E_i(R_i) = \frac{1}{2}\|X_i R_i - \Phi_i\|_2^2, \tag{2.63}$$

where we have set:

- X_i the matrix defined by $X_i = \left[\dfrac{1}{|c_j|}\displaystyle\int_{c_j} (\boldsymbol{x} - \boldsymbol{x}_i)^\alpha \, d\boldsymbol{x} - M_i^\alpha\right]_{j \in s_i^d,\, |\alpha| \in [\![1,d]\!]}$, and
- Φ_i the vector defined by $\Phi_i = \left(\varphi_j^n - \varphi_i^n\right)_{j \in s_i^d}$.

In order to ensure that there is at least one solution to the minimization problem (2.63), we exhibit a condition on the stencil size N_d. Indeed, we need more information from the stencil than we have polynomial coefficients. Therefore, we need $\#s_i^d > \#\{\alpha \in \mathbb{N}^2 \ ; \ |\alpha| \in [\![1, d]\!]\}$. After straightforward computations, we have the following lower bound on the size of the stencil:

$$N_d = \#s_i^d > \frac{(d+1)(d+2)}{2} - 1.$$

To solve the minimization problem (2.63), we use the normal equation approach. We know that R_i is a minimum of $E_i(R_i)$ if and only if R_i is a solution of the following equation, called the normal equation associated to the least squares problem (2.63):

$$X_i^T X_i R_i = X_i^T \Phi_i,$$

where X_i^T is the transpose of the matrix X_i. Now, assume that the matrix $X_i^T X_i$ is invertible. Since the matrix X_i only depends on the geometry, this invertibility property only depends on the stencil s_i^d. Therefore, an important ingredient in the choice of the stencil is to make sure that it leads to the matrix $X_i^T X_i$ being invertible. Equipped with this invertibility condition, the polynomial coefficients R_i satisfy:

$$R_i = (X_i^T X_i)^{-1} X_i^T \Phi_i. \tag{2.64}$$

The matrix $(X_i^T X_i)^{-1} X_i^T$ is called the Moore-Penrose pseudoinverse of X_i; more details can be found in [147]. The expression (2.64) makes the determination of the polynomial coefficients a lot easier. Indeed, since the matrix X_i only depends on the geometry of the mesh, which does not change over time, it is sufficient to compute the pseudoinverse $(X_i^T X_i)^{-1} X_i^T$ once for each cell c_i, at the very beginning of the time iterations of the scheme. Thus, solving the minimization problem only consists in performing the matrix-vector product (2.64) for each cell c_i and at each time step.

The procedure discussed above fully characterizes the polynomial coefficients R_i. However, the condition number of the matrix X_i may be very large, especially when dealing with a high polynomial degree. Therefore, after [1, 73], we suggest a rescaling of the matrix X_i to relax the dependence of the condition number on the geometry and the polynomial degree. The matrix X_i is rescaled as follows:

$$\widetilde{X}_i = \left[\frac{1}{|c_i|^{|\alpha|/2}}\left(\frac{1}{|c_j|}\int_{c_j} (\boldsymbol{x} - \boldsymbol{x}_i)^\alpha \, d\boldsymbol{x} - M_i^\alpha\right)\right]_{j \in s_i^d,\, |\alpha| \in [\![1,d]\!]}. \tag{2.65}$$

The equation (2.64) is then solved with the new matrix $\widetilde{X}_i$, to yield the rescaled polynomial coefficients $\widetilde{R}_i$, as follows:

$$\widetilde{R}_i = (\widetilde{X}_i^T \widetilde{X}_i)^{-1} \widetilde{X}_i^T \Phi_i.$$

Finally, the actual polynomial coefficients R_i to be used within (2.62) are obtained from the rescaled coefficients $\widetilde{R}_i$ by setting:

$$R_i = \left(\frac{\widetilde{R}_i^\alpha}{|c_i|^{|\alpha|/2}} \right)_{|\alpha| \in [\![1,d]\!]}. \tag{2.66}$$

2.4.2 Derivation of high-order two-dimensional schemes for balance laws

We now focus on approximating solutions of 2D systems of balance laws with a high-order accuracy. Such systems are governed by the following initial value problem:

$$\begin{cases} \partial_t W + \boldsymbol{\nabla} \cdot \boldsymbol{F}(W) = \mathfrak{S}(W), \\ W(0, \boldsymbol{x}) = W_0(\boldsymbol{x}). \end{cases} \tag{2.67}$$

In (2.67), as in (2.3), the quantity $\mathfrak{S}(W)$ represents a source term.

In order to provide a high-order approximation of solutions to the system (2.67), we first use the polynomial reconstruction (2.62) to derive a scheme that is high-order in space. Then, we use Runge-Kutta-type methods to provide a high-order time accuracy. This approach is detailed in the next two sections.

2.4.2.1 High-order space accuracy

This section is dedicated to proposing a high-order finite volume discretization of the 2D balance law (2.67). As usual, this discretization is obtained by averaging the balance law (2.67) on the cuboid $[t^n, t^{n+1}] \times c_i$. The main reason this high-order discretization is different from the first-order one presented in Section 2.3.1 is that the polynomial reconstruction (2.62) is used. Therefore, the approximate solution is no longer piecewise constant in each cell, but piecewise polynomial, as follows:

$$\forall i \in \mathbb{Z},\ \forall \boldsymbol{x} \in c_i,\ \widehat{W}_i^n(\boldsymbol{x}; d) \simeq W(t^n, \boldsymbol{x}), \tag{2.68}$$

where $\widehat{W}_i^n(\boldsymbol{x}; d)$ is the vector containing all the components $\hat{\varphi}_i^n(\boldsymbol{x}; d)$ given by (2.62).

Equipped with the polynomial approximation (2.68), we can proceed to determine a high-order finite volume discretization of the 2D balance law (2.67). To determine the high-order finite volume scheme, the system (2.67) is averaged over $[t^n, t^{n+1}] \times c_i$, as follows:

$$\begin{aligned} \frac{1}{\Delta t} \frac{1}{|c_i|} \int_{c_i} \int_{t^n}^{t^{n+1}} \partial_t W \, dt \, d\boldsymbol{x} + \frac{1}{\Delta t} \frac{1}{|c_i|} \int_{t^n}^{t^{n+1}} \int_{c_i} \boldsymbol{\nabla} \cdot \boldsymbol{F}(W) \, d\boldsymbol{x} \, dt \\ = \frac{1}{\Delta t} \frac{1}{|c_i|} \int_{c_i} \int_{t^n}^{t^{n+1}} \mathfrak{S}(W) \, dt \, d\boldsymbol{x}. \end{aligned} \tag{2.69}$$

The goal is now to provide an approximate value of the three integrals in (2.69), while keeping

the required order of accuracy.

The first integral of (2.69) satisfies:

$$\begin{aligned}\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i}\int_{t^n}^{t^{n+1}} \partial_t W\, dt\, d\boldsymbol{x} &= \frac{1}{\Delta t}\left[\frac{1}{|c_i|}\int_{c_i} W(t^{n+1},\boldsymbol{x})\, d\boldsymbol{x} - \frac{1}{|c_i|}\int_{c_i} W(t^n,\boldsymbol{x})\, d\boldsymbol{x}\right] \\ &\simeq \frac{1}{\Delta t}\left[\frac{1}{|c_i|}\int_{c_i} \widehat{W}_i^{n+1}(\boldsymbol{x};d)\, d\boldsymbol{x} - \frac{1}{|c_i|}\int_{c_i} \widehat{W}_i^{n}(\boldsymbol{x};d)\, d\boldsymbol{x}\right].\end{aligned}$$

Arguing the conservation property (2.61) of the polynomial reconstruction $\widehat{W}_i^n$, we have, for the first integral of (2.69):

$$\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i}\int_{t^n}^{t^{n+1}} \partial_t W\, dt\, d\boldsymbol{x} \simeq \frac{W_i^{n+1}-W_i^n}{\Delta t}. \tag{2.70}$$

The second integral of (2.69) concerns the physical flux. Therefore, its approximation will involve the numerical flux function $\mathcal{F}$. Arguing the divergence theorem yields, for this second integral, an expression similar to the one encountered in (2.44):

$$\begin{aligned}\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{t^n}^{t^{n+1}}\int_{c_i} \boldsymbol{\nabla}\cdot\boldsymbol{F}(W)\, d\boldsymbol{x}\, dt &= \frac{1}{|c_i|}\sum_{j\in\nu_i}\int_{e_{ij}}\left(\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \boldsymbol{F}(W(t,\boldsymbol{\sigma}))\cdot\boldsymbol{n}_{ij}\, dt\right) d\boldsymbol{\sigma} \\ &\simeq \frac{1}{|c_i|}\sum_{j\in\nu_i}\int_{e_{ij}} \mathcal{F}\Big(\widehat{W}_i^n(\boldsymbol{\sigma};d),\widehat{W}_j^n(\boldsymbol{\sigma};d);\boldsymbol{n}_{ij}\Big) d\boldsymbol{\sigma},\end{aligned}$$

where we have used the following approximation of the physical flux:

$$\mathcal{F}\Big(\widehat{W}_i^n(\boldsymbol{\sigma};d),\widehat{W}_j^n(\boldsymbol{\sigma};d);\boldsymbol{n}_{ij}\Big) \simeq \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \boldsymbol{F}(W(t,\boldsymbol{\sigma}))\cdot\boldsymbol{n}_{ij}\, dt.$$

We now introduce a quadrature formula on the edge e_{ij}. With $\boldsymbol{\sigma}\in e_{ij}$, the quadrature is given as follows, for any function $\phi : e_{ij}\to\mathbb{R}$:

$$\frac{1}{|e_{ij}|}\int_{e_{ij}} \phi(\boldsymbol{\sigma})\, d\boldsymbol{\sigma} \simeq \sum_{r=1}^{R}\xi_r\phi(\boldsymbol{\sigma}_r). \tag{2.71}$$

Appendix B (see also [2]) give the quadrature weights ξ_r and the quadrature points $\boldsymbol{\sigma}_r$, as well as their number R, so as to ensure a global accuracy of order $(d+1)$. Now, we approximate the integral of the numerical flux on the edge e_{ij} using the quadrature formula (2.71), as follows:

$$\int_{e_{ij}} \mathcal{F}\Big(\widehat{W}_i^n(\boldsymbol{\sigma};d),\widehat{W}_j^n(\boldsymbol{\sigma};d);\boldsymbol{n}_{ij}\Big)\, d\boldsymbol{\sigma} \simeq |e_{ij}|\sum_{r=1}^{R}\xi_r\,\mathcal{F}\Big(\widehat{W}_i^n(\boldsymbol{\sigma}_r;d),\widehat{W}_j^n(\boldsymbol{\sigma}_r;d);\boldsymbol{n}_{ij}\Big).$$

To shorten the notations, we set:

$$\mathcal{F}_{ij,r}^n := \mathcal{F}\Big(\widehat{W}_i^n(\boldsymbol{\sigma}_r;d),\widehat{W}_j^n(\boldsymbol{\sigma}_r;d);\boldsymbol{n}_{ij}\Big).$$

As a consequence, using the previous expressions, the second integral of (2.69) is approxi-

mated as follows:

$$\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{t^n}^{t^{n+1}}\int_{c_i}\boldsymbol{\nabla}\cdot\boldsymbol{F}(W)\,d\boldsymbol{x}\,dt \simeq \sum_{j\in\nu_i}\frac{|e_{ij}|}{|c_i|}\sum_{r=1}^{R}\xi_r\mathcal{F}_{ij,r}^n. \tag{2.72}$$

The third integral of (2.69) concerns the source term. We suggest the following approximation:

$$\begin{aligned}\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i}\int_{t^n}^{t^{n+1}}\mathfrak{S}(W)\,dt\,d\boldsymbol{x} &= \frac{1}{|c_i|}\int_{c_i}\left(\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\mathfrak{S}(W(t,\boldsymbol{x}))\,dt\right)d\boldsymbol{x}\\ &\simeq \frac{1}{|c_i|}\int_{c_i}\mathfrak{S}\Big(\widehat{W}_i^n(\boldsymbol{x};d)\Big)d\boldsymbol{x}.\end{aligned}$$

To deal with the integral in the above formula, we introduce a quadrature formula on the cell c_i. It is given as follows, for $\boldsymbol{x}\in c_i$ and $\psi : c_i\to\mathbb{R}$:

$$\frac{1}{|c_i|}\int_{c_i}\psi(\boldsymbol{x})\,d\boldsymbol{x}\simeq\sum_{q=1}^{Q}\chi_q\psi(\boldsymbol{X}_q), \tag{2.73}$$

where the weights χ_q and the points $\boldsymbol{X}_q$, as well as the number Q of quadrature points, are given by [2] (see also Appendix B). Using this quadrature formula (2.73) yields:

$$\frac{1}{|c_i|}\int_{c_i}\mathfrak{S}\Big(\widehat{W}_i^n(\boldsymbol{x};d)\Big)d\boldsymbol{x}\simeq\sum_{q=1}^{Q}\chi_q\,\mathfrak{S}\Big(\widehat{W}_i^n(\boldsymbol{X}_q;d)\Big).$$

We introduce the following shorter notation:

$$\mathcal{S}_{i,q}^n := \mathfrak{S}\Big(\widehat{W}_i^n(\boldsymbol{X}_q;d)\Big).$$

As a consequence, the approximation of the third integral of (2.69) reads:

$$\frac{1}{\Delta t}\frac{1}{|c_i|}\int_{c_i}\int_{t^n}^{t^{n+1}}\mathfrak{S}(W)\,dt\,d\boldsymbol{x}\simeq\sum_{q=1}^{Q}\chi_q\mathcal{S}_{i,q}^n. \tag{2.74}$$

We can finally derive the high-order finite volume numerical scheme for the 2D balance law (2.67). Combining the three approximations (2.70) – (2.72) – (2.74) and plugging them into the average (2.69) of the balance law (2.67) yields the following high-order numerical scheme:

$$W_i^{n+1} = W_i^n - \Delta t\sum_{j\in\nu_i}\frac{|e_{ij}|}{|c_i|}\sum_{r=1}^{R}\xi_r\mathcal{F}_{ij,r}^n + \Delta t\sum_{q=1}^{Q}\chi_q\mathcal{S}_{i,q}^n. \tag{2.75}$$

Concerning the initial condition, we average the function W_0 on each cell c_i, as follows:

$$W_i^0 = \frac{1}{|c_i|}\int_{c_i}W_0(\boldsymbol{x})\,d\boldsymbol{x}.$$

Using the quadrature formula on a cell (2.73) then yields the following initial condition:

$$W_i^0 = \sum_{q=1}^{Q} \chi_q W_0(\boldsymbol{X}_q).$$

Note that this expression ensures that the initial condition is approximated with the required order of accuracy.

2.4.2.2 High-order time accuracy

The scheme (2.75) is high-order accurate in space. However, it is only first-order accurate in time: therefore, it is globally first-order accurate. In this section, we suggest an extension of the scheme (2.75) to ensure a high-order time accuracy. There are several ways of providing a high-order time accuracy; one of them is the ADER approach (see [149, 150] for more information). However, here, we elect to use Strong Stability-Preserving Runge-Kutta (SSPRK) methods, as introduced in [84, 85]. The goal of such time integrators is to provide a high-order time accuracy while retaining some robustness property of the original scheme (2.75). In order to achieve such a time discretization, we use the second-order method SSPRK(2,2) when $d = 1$, the third-order method SSPRK(3,3) when $d = 2$, and the fourth-order method SSPRK(5,4) when $d \geq 3$. These techniques are described in [84, 137] (the reader is also referred to [144, 85, 138, 146, 139, 83, 102]). For a SSPRK(m,p) method, the number m represents the number of steps in the Runge-Kutta method, and the number p is the order of approximation of the time integrator. Note that the SSPRK(2,2) method is nothing but Heun's method.

We now briefly describe the high-order time integrators mentioned above. Let us rewrite the scheme (2.75) under the following condensed form:

$$W^{n+1} = \mathcal{H}(W^n),$$

where W^n is the vector containing all the constant values W_i^n, i.e. $W^n = (W_i^n)_{i\in\mathbb{Z}}$. After [144], the general form of Runge-Kutta methods reads as follows:

$$\begin{cases} W^{(0)} = W^n, \\ \forall l \in [\![1, m]\!],\ W^{(l)} = \displaystyle\sum_{k=0}^{l-1} \Big[(\alpha_{lk} - \beta_{lk}) W^{(k)} + \beta_{lk} \mathcal{H}(W^{(k)}) \Big], \\ W^{n+1} = W^{(m)}. \end{cases} \tag{2.76}$$

Note that, in [144], the scheme was written under the form $W^{n+1} = W^n + \Delta t \mathcal{L}(W^n)$, thus resulting in a slightly different expression of the Runge-Kutta method. The expression (2.76) is easily derived from the one present in [144]. In (2.76), the coefficients α_{lk} and β_{lk} depend on the required order of the time discretization. The values of α_{lk} and β_{lk} are given for each SSPRK method in Appendix C.

For the sake of completeness, we mention the formulas for the SSPRK(2,2) method and the SSPRK(3,3) method. They are obtained by evaluating the general Runge-Kutta expression (2.76) with the relevant values of α_{lk} and β_{lk} given in Appendix C. As a consequence, the

SSPRK(2,2) method (i.e. Heun's method) reads:

$$\begin{cases} W^{(0)} &= W^n, \\ W^{(1)} &= \mathcal{H}(W^{(0)}), \\ W^{(2)} &= \frac{1}{2}W^{(0)} + \frac{1}{2}\mathcal{H}(W^{(1)}), \\ W^{n+1} &= W^{(2)}, \end{cases}$$

while the SSPRK(3,3) method is given as follows:

$$\begin{cases} W^{(0)} &= W^n, \\ W^{(1)} &= \mathcal{H}(W^{(0)}), \\ W^{(2)} &= \frac{3}{4}W^{(0)} + \frac{1}{4}\mathcal{H}(W^{(1)}), \\ W^{(3)} &= \frac{1}{3}W^{(0)} + \frac{2}{3}\mathcal{H}(W^{(2)}), \\ W^{n+1} &= W^{(3)}. \end{cases}$$

We end this section by noting that the largest order of accuracy of the proposed Runge-Kutta schemes is 4 for the five-step SSPRK(5,4) scheme. Therefore, since the spatial order of accuracy is $p = d+1$, the global order of accuracy will be held back by the time order if $d \geq 4$. The time step Δt is still constrained with the classical CFL condition (2.60). In order to ensure an arbitrarily high order of time accuracy, the time step is modified as follows, with $\widetilde{\Delta t}$ to be used instead of Δt in the scheme:

$$\widetilde{\Delta t} \leq \Delta t^{\frac{\max(d,3)}{3}}. \tag{2.77}$$

2.4.3 The MOOD method

Thanks to the polynomial reconstruction, the integration of the 2D balance law and the relevant SSPRK time discretization, we have designed the scheme (2.75) – (2.76) to be high-order accurate in both space and time. However, this high-order accuracy comes with the loss of the robustness property, and the numerical solutions obtained with this scheme may present unwanted oscillations around the discontinuities (see [154, 111] for instance). Note that such issues were already present in the 1D second-order case. In the context of the 1D MUSCL reconstruction, slope limiters were used to prevent these non-physical oscillations and ensure the robustness of the scheme (see Section 2.2).

To address these issues in the context of a 2D high-order scheme, we use a MOOD technique. An overview of this method is presented in [46, 63, 65]. Several applications have also been suggested in recent years, for instance the recovery of the entropy preservation in [16], a coupling with the ADER technique in [117, 25], an application to the shallow-water equations in [71, 47, 50], and applications to some other systems in [49, 54, 64]. These applications are summarized in [48]. Following the MOOD paradigm, a subcell limiter technique for the discontinuous Galerkin method has been proposed in [68].

The goal of the MOOD procedure is to recover essential properties of a first-order scheme, for instance its robustness, by detecting whether these properties are verified by the high-

order approximation. This detection process is performed by several *detection criteria*, which check whether the properties are satisfied in each cell. If this verification fails in some cell, the degree of the approximation is lowered in this cell, until the properties are satisfied.

First, we present some detection criteria that are commonly used within the MOOD procedure. Their purpose is to preserve the robustness and control the spurious oscillations. For a more exhaustive description of these criteria, the reader is referred to [47, 71]. In this section, we use the notation $W^\star$ for the candidate solution, i.e. the solution obtained from W^n using the high-order scheme (2.75) – (2.76) presented in the previous subsection. This candidate solution is then tested against the detection criteria, to determine the cells where it is not acceptable. In such cells, computing a new candidate solution is required. Second, we state the detector chain, i.e. the order in which the detectors are used, as well as the MOOD loop.

The Physical Admissibility Detector (PAD)

The PAD determines whether the approximate solution lies within the admissible states space Ω. Thus, the PAD criterion fails within the cell c_i if $W_i^\star \notin \Omega$. Let us underline that, equipped with the PAD, the high-order scheme is robust.

The Discrete Maximum Principle detector (DMP)

Although the PAD ensures that the robustness is preserved, it does not prevent spurious oscillations from appearing in the vicinity of discontinuities. To address this issue, we use the DMP criterion to check for oscillations. The DMP criterion fails if, for some component φ of W, we have:

$$\min_{j\in\nu_i}(\varphi_j) - \varepsilon_M \leq (\varphi)_i^\star \leq \min_{j\in\nu_i}(\varphi_j) + \varepsilon_M, \tag{2.78}$$

where ε_M is a constant used to reduce the risk of falsely detecting an oscillation that could be due to a floating point error. In practice, we usually take $\varepsilon_M = \delta^3$, where

$$\delta = \frac{|c_i|}{|P_i|}.$$

Detecting physical oscillations: the u2 criterion

Unfortunately, the DMP criterion (2.78) is too restrictive. It will sometimes detect and eliminate physical oscillations, thus resulting in a false positive that reduces the accuracy of the scheme. Therefore, we need another criterion to detect whether an oscillation is physically admissible. To that end, we introduce the u2 criterion, which uses the constant second derivative of the second-degree polynomial reconstruction $\hat{\varphi}_i^n(\boldsymbol{x};2)$. With $\boldsymbol{x} = {}^t(x,y)$, we define the following curvatures on the cell c_i:

$$\mathcal{X}_i^{\min} = \min\Big(\partial_{xx}\hat{\varphi}_i^n, \min_{j\in\nu_i}\big(\partial_{xx}\hat{\varphi}_j^n\big)\Big), \qquad \mathcal{X}_i^{\max} = \max\Big(\partial_{xx}\hat{\varphi}_i^n, \max_{j\in\nu_i}\big(\partial_{xx}\hat{\varphi}_j^n\big)\Big),$$
$$\mathcal{Y}_i^{\min} = \min\Big(\partial_{yy}\hat{\varphi}_i^n, \min_{j\in\nu_i}\big(\partial_{yy}\hat{\varphi}_j^n\big)\Big), \qquad \mathcal{Y}_i^{\max} = \max\Big(\partial_{yy}\hat{\varphi}_i^n, \max_{j\in\nu_i}\big(\partial_{yy}\hat{\varphi}_j^n\big)\Big).$$

Equipped with the curvatures, we state three criteria, which will be combined to form the

u2 criterion (see [71, 47]). First, the plateau detector is defined as follows:

$$\max\left(\left|\mathcal{X}_i^{\min}\right|, |\mathcal{X}_i^{\max}|, \left|\mathcal{Y}_i^{\min}\right|, |\mathcal{Y}_i^{\max}|\right) \leq \delta. \tag{2.79}$$

This criterion detects whether the local curvatures are small enough to consider the approximation locally linear. In this case, the reconstruction should not be limited, and the plateau detector is hence activated. Next, the oscillation detector is given by:

$$\mathcal{X}_i^{\min}\mathcal{X}_i^{\max} \leq -\delta \quad \text{and} \quad \mathcal{Y}_i^{\min}\mathcal{Y}_i^{\max} \leq -\delta. \tag{2.80}$$

This oscillation detector is activated if the local curvatures undergo a change of sign in the vicinity of the cell. This behavior of the curvatures indicated that an oscillation is present, and therefore that the reconstruction should be limited in the cell. The third criterion involves a local smoothness detector, given as follows:

$$\frac{1}{2} \leq \frac{\min\left(\left|\mathcal{X}_i^{\min}\right|, |\mathcal{X}_i^{\max}|\right)}{\max\left(\left|\mathcal{X}_i^{\min}\right|, |\mathcal{X}_i^{\max}|\right)} \leq 1 \quad \text{and} \quad \frac{1}{2} \leq \frac{\min\left(\left|\mathcal{Y}_i^{\min}\right|, |\mathcal{Y}_i^{\max}|\right)}{\max\left(\left|\mathcal{Y}_i^{\min}\right|, |\mathcal{Y}_i^{\max}|\right)} \leq 1. \tag{2.81}$$

According to this detector, the solution is considered as locally smooth if the minimum and maximum curvatures are close enough. If the solution is determined to be locally smooth, the reconstruction should not be limited.

The u2 criterion is finally defined as a combination of these three detectors. Indeed, if a plateau is detected by (2.79) or if the solution is considered locally smooth by (2.81), then the DMP criterion becomes irrelevant and the u2 criterion succeeds, thus leading to a non-limited reconstruction. On the contrary, if a local oscillation is detected by (2.80), then the u2 criterion fails, and the polynomial degree is lowered in the cell.

The detector chain

Equipped with these detectors, we state the order in which they are checked. To address this issue, we introduce the *Cell Polynomial Degree* (CPD). The CPD is an integer, associated to a cell c_i, such that $\text{CPD}(i) \in [\![0, d]\!]$. If $\text{CPD}(i) = p$, then the polynomial reconstruction used in the cell c_i is of degree p. Figure 2.13 displays the detector chain for a cell where $\text{CPD}(i) = p > 0$, and the effect of each detector on the CPD.

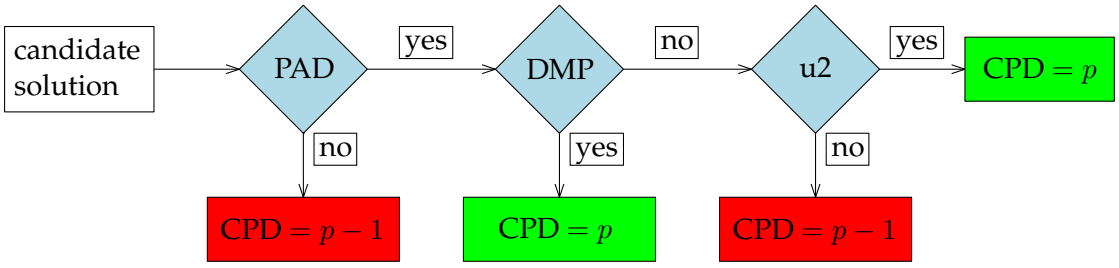

Figure 2.13 – The MOOD detector chain within a single cell. At the beginning of the chain, $\text{CPD}(i) = p$.

At the beginning of the chain, we consider a candidate solution W_i^* computed in the cell c_i with a polynomial reconstruction of degree p (i.e. $\text{CPD}(i) = p$). At the end of the chain, if

no criterion failed, then the candidate solution is declared suitable, and it is accepted as the updated approximate solution W^{n+1}. If one of the criteria did fail, then CPD(i) is set to $p-1$ and the candidate solution is not accepted. If that is the case, a new candidate solution is computed, using a polynomial reconstruction whose degree in the cell c_i is equal to CPD(i). We remark that, if a cell c_i *and its neighbors* are declared suitable, there is no need to compute a new candidate solution in this cell c_i. This remark help significantly reduce the computational cost the MOOD method by computing a new candidate solution only in the cells where it is required.

Note that this detector chain may be supplemented with additional detectors, to enforce other properties to be satisfied by the scheme. One such detector concerns the entropy preservation (see [16]). A MOOD-like method could also be used to recover the well-balance property of a scheme. Since the reconstruction procedure does not ensure that the well-balance property is satisfied by the high-order scheme, it is relevant to introduce a well-balance detection criteria to the detector chain. Such a criterion is suggested in Chapter 4.

The full MOOD loop

From the previous paragraphs, we know the MOOD detectors and the order in which they are applied. We now state the full MOOD loop for a desired reconstruction of degree d, i.e. a scheme of order $(d+1)$. For a single iteration in time of the SSPRK time discretization, the MOOD loop reads as follows.

1. In each cell c_i, initialize CPD$(i) = d$.
2. Compute the candidate solution $W^\star$ using the scheme (2.75) – (2.76) and the current CPD map.
3. Apply the detection process displayed Figure 2.13 to compute a potentially new CPD map and to decide whether to accept the candidate solution. If the candidate solution is rejected, go to step 2. Otherwise, go to step 4.
4. The candidate solution is accepted, and we set $W^{n+1} = W^\star$.

Note that this loop, at worst, makes the CPD of every cell equal to 1. This situation corresponds to using the first-order scheme (2.47), which is robust and non-oscillatory. Hence, it satisfies all the MOOD criteria. Therefore, the MOOD loop cannot be endless (see also [46] for a more formal proof). In practice, it is highly unlikely that such a situation happens.

3

A well-balanced scheme for the shallow-water equations

The shallow-water equations equipped with the topography and Manning friction source terms have been presented in Chapter 1. In addition, finite volume techniques have been discussed in Chapter 2. Equipped with these studies, the goal of this chapter is to derive a one-dimensional scheme that possesses the following properties:

- *consistency*: the scheme is consistent with the shallow-water equations with topography and Manning friction (1.31);
- *well-balance*: the scheme preserves the steady states for the shallow-water equations with topography and Manning friction, given in Section 1.2;
- *robustness*: the scheme ensures the non-negativity of the water height;
- capture of *wet/dry transitions*: the scheme is able to correctly model transitions between wet areas (where $h \neq 0$) and dry areas (where $h = 0$).

In order to obtain such properties, we elect to use a Godunov-type scheme (see Section 2.1.3 for more information). This scheme will be based on a two-state approximate Riemann solver. One of the most famous two-state approximate Riemann solvers is the HLLC Riemann solver, developed for the Euler system of fluid dynamics by Toro, Spruce and Speares in [151]. The HLLC (HLL – Contact) solver is based on the HLL solver. The goal of the HLLC scheme is to provide a good approximation of the contact discontinuity present in the Euler system of fluid dynamics. Compared to the HLL solver, it contains an additional wave, which corresponds to the contact wave in the Riemann problem. Note that adding a wave also adds unknowns to be determined. Additional relations may be imposed on these unknowns to satisfy several required properties. For instance, [76] deals with positive and entropy-satisfying approximate Riemann solvers applied to several systems. We also mention work on several other systems: a radiative transfer model in [14], a sediment transport model in [34], the Ripa model in [140], and the equations of chemotaxis in [15].

Several approximate Riemann solvers have also been developed in the framework of the shallow-water equations. For instance, we mention [75], where the author derives a general framework for positive and entropy-satisfying numerical approximate Riemann solvers. We also mention [7], where a two-state approximate Riemann solver is designed to be positive and to preserve the lake at rest steady state. Finally, in [12], the authors derive a positive and entropy-satisfying approximate Riemann solver that allows the preservation of all the steady state solutions of the shallow-water equations with just the topography. In both [7] and [12], the two states are separated by a wave whose velocity is zero. This choice is motivated by the presence of the stationary wave exhibited in Section 1.1.3 and created by the source terms.

This approach is used in this manuscript to derive an approximate Riemann solver that takes into account a generic source term on the discharge equation. The derivation of this scheme is presented in Section 3.1. Firstly, we derive a well-balanced approximate Riemann solver for a generic source term on the discharge equation, which may consist in the topography, the friction, or yet another source term. A correction is introduced to ensure the robustness of the scheme. Secondly, the scheme is applied to a specific class of source terms, to which the topography and the Manning friction source terms belong. For these two source terms, explicit expressions are given for the intermediate states of the approximate Riemann solver. A special treatment is made to consider vanishing water heights.

The scheme suggested in Section 3.1 is well-balanced and robust. However, the friction source term becomes stiff when a wet/dry transition is considered. Therefore, in order to correctly model the wet/dry transitions, we introduce in Section 3.2 a semi-implicitation of the scheme via a splitting technique. The scheme is first rewritten to exhibit the numerical flux function as well as the numerical source terms approximation. Then, a semi-implicitation technique is proposed for the Manning friction source term, in order to recover a good approximation of wet/dry fronts.

Finally, equipped with the well-balanced scheme, the last section of this chapter, Section 3.3, is dedicated to numerical experiments. First, the well-balance property is tested in the situations described in Section 1.2. The simulations of different lake at rest configurations, as well as several moving steady states for the source terms of topography and/or friction, are carried out. Second, validation experiments are performed. Namely, we present several dam-break experiments.

3.1 Well-balanced scheme for a generic source term on the discharge equation

In this section, we consider the shallow-water system endowed with a generic source term on the discharge equation. This system is governed by the following set of equations:

$$\begin{cases} \partial_t h + \partial_x q = 0, \\ \partial_t q + \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = S(W), \end{cases} \tag{3.1}$$

where $W = {}^t(h, q)$, and where $S(W)$ denotes a generic source term, which can be the topography, the friction, or another source term. Note that $S(W)$ may depend on other quantities than W, for example the topography function Z in the case of the topography source term. However, for the sake of simplicity in the notations, this dependence is not explicitly written. The equations (3.1) are rewritten under the following condensed form of a 1D balance law:

$$\partial_t W + \partial_x F(W) = \mathfrak{S}(W), \tag{3.2}$$

where

$$W = \begin{pmatrix} h \\ q \end{pmatrix} \quad ; \quad F(W) = \begin{pmatrix} q \\ \dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \end{pmatrix} \quad ; \quad \mathfrak{S}(W) = \begin{pmatrix} 0 \\ S(W) \end{pmatrix}. \tag{3.3}$$

In order to provide approximate solutions to this set of equations, we choose a Godunov-type scheme (see Section 2.1.3) equipped with an approximate Riemann solver made of two constant intermediate states. Recall from Section 2.1 that the finite volume schemes we use are based on a relevant discretization of the space domain $\mathbb{R}$. We briefly recall this procedure here. The space domain $\mathbb{R}$ is discretized in cells $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, of length Δx (see Figure 2.1). Then, the approximate solution W_i^n at time t^n is assumed to be piecewise constant in each cell $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$. The goal of a Godunov-type scheme is to provide an approximation W_i^{n+1} of the solution at time t^{n+1}, knowing the approximate solution within every cell at time t^n. This discretization leads to approximately solving the following Riemann problem at each interface between cells:

$$\begin{cases} \partial_t W + \partial_x F(W) = \mathfrak{S}(W), \\ W(0, x) = W_0(x) = \begin{cases} W_L \text{ if } x < 0, \\ W_R \text{ if } x > 0. \end{cases} \end{cases} \tag{3.4}$$

Here, the solution of this Riemann problem is approximated with the aforementioned two-state approximate Riemann solver. This approximate solver is defined as follows:

$$\widetilde{W}\Big(\frac{x}{t}; W_L, W_R\Big) = \begin{cases} W_L & \text{if } x/t \leq \lambda_L, \\ W_L^* & \text{if } \lambda_L < x/t < 0, \\ W_R^* & \text{if } 0 < x/t < \lambda_R, \\ W_R & \text{if } x/t \geq \lambda_R, \end{cases} \tag{3.5}$$

where W_L^* and W_R^* are the unknown intermediate states, to be determined in order to ensure that the required properties are satisfied. In addition, recall that the characteristic velocities for the Riemann problem (3.4) are given by (1.34). As a consequence, we define the approximate characteristic velocities λ_L and λ_R as follows:

$$\begin{aligned} \lambda_L &= \min(-|u_L| - c_L,\ -|u_R| - c_R,\ -\varepsilon_\lambda), \\ \lambda_R &= \max(|u_L| + c_L,\ |u_R| + c_R,\ \varepsilon_\lambda), \end{aligned} \tag{3.6}$$

where u is the velocity of the water, c is the sound speed, defined by (1.12), and $\varepsilon_\lambda > 0$ is a small constant to be fixed in the numerical applications. The constant ε_λ is introduced in order to add some numerical viscosity to the scheme. The definition (3.6) of the characteristic

velocities ensures the following crucial relation:

$$\lambda_L < 0 < \lambda_R. \tag{3.7}$$

Indeed, a stationary wave with velocity 0 is present in the approximate Riemann solver (3.5). The condition (3.7) ensures that the three waves with velocities λ_L, 0 and λ_R do not cross. The structure of the approximate Riemann solver (3.5) is displayed on Figure 3.1.

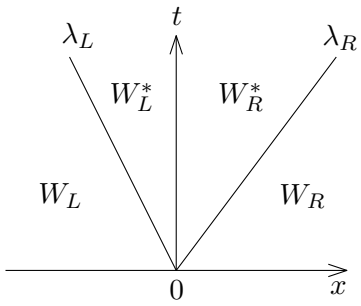

Figure 3.1 – Structure of the chosen approximate Riemann solver.

Equipped with $\widetilde{W}$, the goal is now to provide an expression for the updated state W_i^{n+1}. First, we take a time step constrained by the CFL condition (2.11), as follows:

$$\frac{\Delta t}{\Delta x} \max_{i\in\mathbb{Z}} \left(\left| \lambda^-_{i+\frac{1}{2}} \right|, \left| \lambda^+_{i+\frac{1}{2}} \right| \right) \leq \frac{1}{2},$$

where $\lambda^-_{i+\frac{1}{2}}$ and $\lambda^+_{i+\frac{1}{2}}$ are the approximate characteristic speeds for the Riemann problem located at the interface $x_{i+\frac{1}{2}}$. Then, we define the juxtaposition function W^Δ as follows:

$$\forall t \in (0, \Delta t]\,,\ \forall x \in [x_i, x_{i+1}),\ W^\Delta(t^n + t, x) = \widetilde{W}\left(\frac{x - x_{i+\frac{1}{2}}}{t}; W_i^n, W_{i+1}^n \right),$$

where $\widetilde{W}\Big((x - x_{i+\frac{1}{2}})/t; W_i^n, W_{i+1}^n \Big)$ is the approximate Riemann solver (3.5), given for $x \in [x_i, x_{i+1})$ and for $t \in (0, \Delta t]$ by

$$\widetilde{W}\left(\frac{x - x_{i+\frac{1}{2}}}{t}; W_i^n, W_{i+1}^n \right) = \begin{cases} W_i^n & \text{if } \dfrac{x - x_{i+\frac{1}{2}}}{t} \leq \lambda^L_{i+\frac{1}{2}}, \\ W^{L,*}_{i+\frac{1}{2}} & \text{if } \lambda^L_{i+\frac{1}{2}} < \dfrac{x - x_{i+\frac{1}{2}}}{t} < 0, \\ W^{R,*}_{i+\frac{1}{2}} & \text{if } 0 < \dfrac{x - x_{i+\frac{1}{2}}}{t} < \lambda^R_{i+\frac{1}{2}}, \\ W_{i+1}^n & \text{if } \dfrac{x - x_{i+\frac{1}{2}}}{t} \geq \lambda^R_{i+\frac{1}{2}}, \end{cases} \tag{3.8}$$

where $W^{L,*}_{i+\frac{1}{2}}$ and $W^{R,*}_{i+\frac{1}{2}}$ are the intermediate states of the approximate solution to the Riemann problem located at the interface $x_{i+\frac{1}{2}}$. This juxtaposition function, as well as the approximate Riemann solver, are displayed on Figure 3.2.

Finally, the updated solution W_i^{n+1} is obtained by integrating the juxtaposition function W^Δ on the cell $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$. In the current context of a two-state approximate Riemann solver,

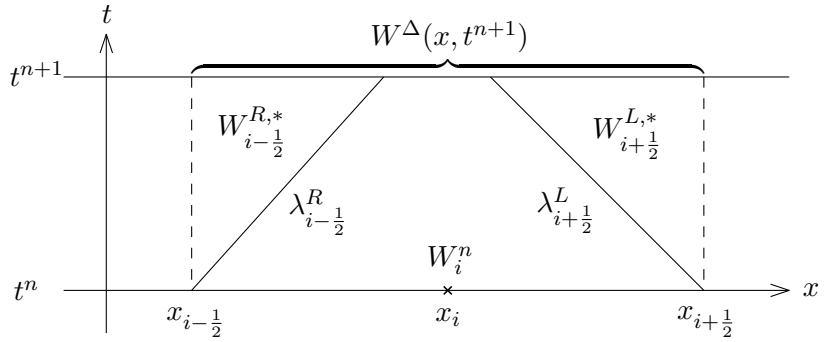

Figure 3.2 – The full Godunov-type scheme using the prescribed approximate Riemann solver.

the following sequence of equalities hold:

$$\begin{aligned} W_i^{n+1} &= \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W^{\Delta}(t^{n+1}, x)\, dx \\ &= \lambda^R_{i-\frac{1}{2}}\frac{\Delta t}{\Delta x} W^{R,*}_{i-\frac{1}{2}} + \left(\frac{x_{i+\frac{1}{2}}}{\Delta x} + \lambda^L_{i+\frac{1}{2}}\frac{\Delta t}{\Delta x} - \frac{x_{i-\frac{1}{2}}}{\Delta x} - \lambda^R_{i-\frac{1}{2}}\frac{\Delta t}{\Delta x}\right) W_i^n - \lambda^L_{i+\frac{1}{2}}\frac{\Delta t}{\Delta x} W^{L,*}_{i+\frac{1}{2}}. \end{aligned}$$

As a consequence, the updated state W_i^{n+1} is given by:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left[\lambda^L_{i+\frac{1}{2}}\left(W^{L,*}_{i+\frac{1}{2}} - W_i^n\right) - \lambda^R_{i-\frac{1}{2}}\left(W^{R,*}_{i-\frac{1}{2}} - W_i^n\right)\right]. \tag{3.9}$$

It is clear from (3.9) that the updated state W_i^{n+1} is fully determined as soon as an expression is given to the intermediate states $W^{L,*}_{i+\frac{1}{2}}$ and $W^{R,*}_{i+\frac{1}{2}}$, for all $i \in \mathbb{Z}$. Determining these intermediate states, first for a generic source term and then in the specific cases of the topography and the friction, is the focus of the remainder of this section.

3.1.1 Derivation of the intermediate states

Now, our goal is to propose a suitable approximation of the Riemann problem (3.4). To that end, we use the two-state approximate Riemann solver $\widetilde{W}$, defined by (3.5). It is made of four constant states separated by three discontinuities. Two of these states, W_L and W_R, are the known initial data of the Riemann problem. The other two, W_L^* and W_R^*, are unknown. The intermediate states W_L^* and W_R^* are each made of two unknowns, as follows:

$$W_L^* = \begin{pmatrix} h_L^* \\ q_L^* \end{pmatrix} \quad \text{and} \quad W_R^* = \begin{pmatrix} h_R^* \\ q_R^* \end{pmatrix}.$$

Note that, as soon as $W_L^* = W_L$ and $W_R^* = W_R$, the scheme (3.9) obviously becomes stationary, i.e. $W_i^{n+1} = W_i^n$. As a consequence, the intermediate states must satisfy $W_L^* = W_L$ and $W_R^* = W_R$ as soon as a steady state solution is considered. If this property is satisfied, then the scheme will be naturally well-balanced. Several other constraints have to be imposed on the intermediate states: namely, consistency and robustness. The consistency will be imposed by arguing the integral consistency property (2.20). Regarding the robustness, note that the

updated water height given by the scheme (3.9) rewrites as follows:

$$h_i^{n+1} = h_i^n \left(1 + \lambda_{i+\frac{1}{2}}^L \frac{\Delta t}{\Delta x} - \lambda_{i-\frac{1}{2}}^R \frac{\Delta t}{\Delta x}\right) + h_{i-\frac{1}{2}}^{R,*} \left(\lambda_{i-\frac{1}{2}}^R \frac{\Delta t}{\Delta x}\right) - h_{i+\frac{1}{2}}^{L,*} \left(\lambda_{i+\frac{1}{2}}^L \frac{\Delta t}{\Delta x}\right).$$

We assume that $h_i^n \geq 0$. Recall that $\lambda_{i+\frac{1}{2}}^L < 0$ and $\lambda_{i-\frac{1}{2}}^R > 0$ after (3.7), and recall the CFL condition (3.1), which constrains the time step Δt. As a consequence, we have

$$\Delta t \leq \frac{1}{2} \frac{\Delta x}{|\lambda_{i+\frac{1}{2}}^L|} \quad \text{and} \quad \Delta t \leq \frac{1}{2} \frac{\Delta x}{\lambda_{i-\frac{1}{2}}^R}.$$

Therefore, the following inequality holds:

$$1 + \lambda_{i+\frac{1}{2}}^L \frac{\Delta t}{\Delta x} - \lambda_{i-\frac{1}{2}}^R \frac{\Delta t}{\Delta x} \geq 0.$$

Furthermore, since $\lambda_{i+\frac{1}{2}}^L < 0$ and $\lambda_{i-\frac{1}{2}}^R > 0$, we get the following sufficient condition for h_i^{n+1} to be non-negative:

$$\text{if } h_{i-\frac{1}{2}}^{R,*} \geq 0 \text{ and } h_{i+\frac{1}{2}}^{L,*} \geq 0, \text{ then } h_i^{n+1} \geq 0.$$

Therefore, the scheme is robust as soon as the intermediate water heights are non-negative. The properties that the intermediate states W_L^* and W_R^* are required to satisfy are thus summarized as follows:

- integral consistency (2.20);
- robustness: $h_L^* \geq 0$ and $h_R^* \geq 0$;
- well-balance: $W_L^* = W_L$ and $W_R^* = W_R$ as soon as a steady state is reached, i.e. as soon as the steady relation $\partial_x F(W) = \mathfrak{S}(W)$ is satisfied in a discrete sense to be determined later.

In this section, we first briefly study the Riemann problem (3.4). Then, we determine the intermediate states W_L^* and W_R^* such that the required properties of consistency, well-balance and robustness are satisfied.

3.1.1.1 Properties of the Riemann problem

We now study the properties of the Riemann problem (3.4) for the shallow-water equations with a generic source term on the discharge equation. In Section 1.1.3, the shallow-water system has been studied in the case of the topography and the Manning friction source terms. The study with a generic source term is presented here in order to exhibit the wave structure and the Riemann invariants for the Riemann problem (3.4). These informations will be instrumental in the derivation of the intermediate states W_L^* and W_R^*.

The system under consideration reads as follows:

$$\begin{cases} \partial_t h + \partial_x q = 0, \\ \partial_t q + \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2\right) - S(W) \partial_x Y = 0, \\ \partial_t Y = 0, \end{cases}$$

where the quantity Y satisfies $Y(t,x) = x$ (and consequently $\partial_x Y = 1$). We now assume that $h \neq 0$. We also assume that h and q are smooth functions. Using the velocity $u = q/h$, the above system reads:

$$\begin{cases} \partial_t h + u\partial_x h + h\partial_x u = 0, \\ \partial_t u + g\partial_x h + u\partial_x u - Sh^{-1}\partial_x Y = 0, \\ \partial_t Y = 0, \end{cases}$$

where the dependence of S in W has been temporarily dropped for the sake of simplicity in the notations. Therefore, the shallow-water system with a generic source term can be cast under the following non-conservative form:

$$\partial_t U + A(U)\partial_x U = 0,$$

where the vector U and the matrix $A(U)$ are given by:

$$U = \begin{pmatrix} h \\ u \\ Y \end{pmatrix} \quad \text{and} \quad A(U) = \begin{pmatrix} u & h & 0 \\ g & u & -Sh^{-1} \\ 0 & 0 & 0 \end{pmatrix}.$$

The eigenvalues of this matrix are $\lambda_\pm(U) = u \pm \sqrt{gh}$ and $\lambda_0(U) = 0$. Concerning $\lambda_\pm(U)$, these characteristic velocities are both associated to GNL fields, through which the quantity Y is preserved (see Section 1.1.3 for the specific case where S is made of the topography and the Manning friction source terms). Regarding $\lambda_0(U)$, the eigenvector associated to this eigenvalue is given by:

$$R_0(U) = \begin{pmatrix} Sh \\ -Su \\ gh^2 - hu^2 \end{pmatrix}.$$

Since $\lambda_0(U) = 0$, the associated characteristic field is obviously linearly degenerate, and it will therefore produce a contact discontinuity. Recall that, across a contact discontinuity, the Riemann invariants are constant quantities. They are functions $\Phi(U)$ given by (1.13), as follows:

$$\nabla_U \Phi(U) \cdot R_0(U) = 0. \tag{3.10}$$

Note that, for $S = 0$ (i.e. a vanishing source term contribution), the quantities h and u are Riemann invariants. This behavior is to be expected since the stationary wave is created by the source term. Hence, without source term, there is no stationary wave. We now assume that $S \neq 0$. In this case, (3.10) rewrites:

$$\frac{dh}{Sh} = \frac{du}{-Su} = \frac{dY}{gh^2 - hu^2}. \tag{3.11}$$

The first equality of the relations (3.11) yields:

$$d(hu) = 0.$$

As a consequence, we recover, as expected, that the discharge $q = hu$ is constant across the

stationary wave. Equipped with this constant discharge, the second equality of the relations (3.11) yields:

$$\left(\frac{q^2}{h^2} - gh\right) dh + S dY = 0.$$

As a consequence, the Riemann invariants across the stationary wave are governed by the following two relations:

$$\begin{cases} dq = 0, \\ d\left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2\right) = S dY. \end{cases} \tag{3.12}$$

The equations (3.12) cannot be simplified further in the case of a generic source term S. In the specific cases of the topography and the Manning friction, the expressions from Section 1.1.3 are obtained.

3.1.1.2 Consistency

Using the algebraic properties of the shallow-water equations with a generic source term (3.1), we now derive suitable intermediate states for the approximate Riemann solver (3.5). We first determine a necessary condition on the intermediate states to ensure the consistency of the scheme. Recall from Section 2.1.3 that the following *integral consistency* condition (2.20) has to be prescribed on the intermediate states:

$$\frac{1}{\Delta x} \int_{-\Delta x/2}^{\Delta x/2} \widetilde{W}\Big(\frac{x}{\Delta t}; W_L, W_R\Big) dx = \frac{1}{\Delta x} \int_{-\Delta x/2}^{\Delta x/2} W_{\mathcal{R}}\Big(\frac{x}{\Delta t}; W_L, W_R\Big) dx. \tag{3.13}$$

In Section 2.1.3, the above integrals have been computed in the case of a hyperbolic conservation law, i.e. without source terms, to yield (2.25). We now perform these computations in the case of the shallow-water equations with a generic source term, given under the form (3.2), and for the approximate Riemann solver (3.5).

The average of the exact Riemann solution $W_{\mathcal{R}}$ rewrites as follows, by integrating (3.2) over the rectangle $[-\Delta x/2, \Delta x/2] \times [0, \Delta t]$, with the initial condition W_0 given by (3.4):

$$\begin{aligned} \int_{-\Delta x/2}^{\Delta x/2} W_{\mathcal{R}}\Big(\frac{x}{\Delta t}; W_L, W_R\Big) dx = & \int_{-\Delta x/2}^{\Delta x/2} W_0(x)\, dx \\ & - \int_0^{\Delta t} F\left(W_{\mathcal{R}}\left(-\frac{\Delta x}{2t}; W_L, W_R\right)\right) dt \\ & + \int_0^{\Delta t} F\left(W_{\mathcal{R}}\left(\frac{\Delta x}{2t}; W_L, W_R\right)\right) dt \\ & + \int_0^{\Delta t} \int_{-\Delta x/2}^{\Delta x/2} \mathfrak{S}\Big(W_{\mathcal{R}}\Big(\frac{x}{t}; W_L, W_R\Big)\Big)\, dx\, dt. \end{aligned} \tag{3.14}$$

Note that, due to the presence of the source term, uniform in space initial data is no longer solution to the balance law (3.2). Indeed, for a function $W(t)$ uniform in space, (3.2) rewrites as $\partial_t W = S(W)$, and W has to depend on the time to be a solution of this equation. However, in (3.14), we have made the approximation that the constant initial data is a solution. Therefore, the integral of the initial condition $W_0(x)$ does not depend on time. As a consequence,

performing straightforward computations and arguing the CFL condition (2.23) lead to the following expression of the average of the exact Riemann solution:

$$\begin{aligned}\frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} W_{\mathcal{R}}\Big(\frac{x}{\Delta t};W_L,W_R\Big)dx = \frac{W_L+W_R}{2} - \frac{\Delta t}{\Delta x}(F(W_R)-F(W_L))\\ + \frac{1}{\Delta x}\int_0^{\Delta t}\int_{-\Delta x/2}^{\Delta x/2}\mathfrak{S}\Big(W_{\mathcal{R}}\Big(\frac{x}{t};W_L,W_R\Big)\Big)\,dx\,dt.\end{aligned} \tag{3.15}$$

Note that this expression is very similar to (2.25), which had been obtained in the case without source terms. Indeed, only the average of the source terms contribution has been added to (2.25).

Now, using the expression (3.5) of $\widetilde{W}$, the integral of the approximate Riemann solver rewrites as follows:

$$\begin{aligned}\int_{-\Delta x/2}^{\Delta x/2}\widetilde{W}\Big(\frac{x}{\Delta t};W_L,W_R\Big)dx = \Big(\lambda_L\Delta t+\frac{\Delta x}{2}\Big)W_L + (0-\lambda_L\Delta t)W_L^*\\ + (\lambda_R\Delta t-0)W_R^* + \Big(\frac{\Delta x}{2}-\lambda_R\Delta t\Big)W_R.\end{aligned}$$

Therefore, the average of $\widetilde{W}$ is given by:

$$\frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2}\widetilde{W}\Big(\frac{x}{\Delta t};W_L,W_R\Big)dx = \frac{W_L+W_R}{2} - \lambda_R\frac{\Delta t}{\Delta x}(W_R-W_R^*) + \lambda_L\frac{\Delta t}{\Delta x}(W_L-W_L^*). \tag{3.16}$$

The integral consistency condition is then obtained by plugging (3.15) and (3.16) into (3.13), to get the following necessary condition on the intermediate states:

$$\begin{aligned}\lambda_R W_R^* - \lambda_L W_L^* = \lambda_R W_R - \lambda_L W_L - (F(W_R)-F(W_L))\\ + \frac{1}{\Delta t}\int_0^{\Delta t}\int_{-\Delta x/2}^{\Delta x/2}\mathfrak{S}\Big(W_{\mathcal{R}}\Big(\frac{x}{t};W_L,W_R\Big)\Big)\,dx\,dt.\end{aligned} \tag{3.17}$$

Now, recall that the sole intermediate state of the HLL approximate Riemann solver is given in [90] by (2.33), as follows:

$$W_{HLL} = \frac{\lambda_R W_R - \lambda_L W_L - (F_R - F_L)}{\lambda_R - \lambda_L}. \tag{3.18}$$

As a consequence, using (3.18), (3.17) reads:

$$\lambda_R W_R^* - \lambda_L W_L^* = (\lambda_R-\lambda_L)W_{HLL} + \frac{1}{\Delta t}\int_0^{\Delta t}\int_{-\Delta x/2}^{\Delta x/2}\mathfrak{S}\Big(W_{\mathcal{R}}\Big(\frac{x}{t};W_L,W_R\Big)\Big)\,dx\,dt.$$

In the context of the shallow-water equations with a source term on the discharge equation (3.2), we have $W = {}^t(h,q)$ and $\mathfrak{S}(W) = {}^t(0,S(W))$. Therefore, the above identity reads:

$$\lambda_R h_R^* - \lambda_L h_L^* = (\lambda_R-\lambda_L)h_{HLL}, \tag{3.19a}$$

$$\lambda_R q_R^* - \lambda_L q_L^* = (\lambda_R-\lambda_L)q_{HLL} + \frac{1}{\Delta t}\int_0^{\Delta t}\int_{-\Delta x/2}^{\Delta x/2} S\Big(W_{\mathcal{R}}\Big(\frac{x}{t};W_L,W_R\Big)\Big)\,dx\,dt, \tag{3.19b}$$

where h_{HLL} and q_{HLL} are given, after (3.18), by:

$$(\lambda_R - \lambda_L) h_{HLL} = \lambda_R h_R - \lambda_L h_L - [q], \tag{3.20a}$$

$$(\lambda_R - \lambda_L) q_{HLL} = \lambda_R q_R - \lambda_L q_L - \left[\frac{q^2}{h} + \frac{1}{2} g h^2\right]. \tag{3.20b}$$

With (3.19), we have obtained two equations linking the four unknowns h_L^*, h_R^*, q_L^* and q_R^*. We still need to exhibit two additional relations to uniquely determine these four unknowns. In addition, the average of the source term present in (3.19b) needs to be dealt with. Both these issues are addressed in the next section.

3.1.1.3 Well-balance parametrization

In order to deal with the source term average in (3.19b), we introduce a parameter $\overline{S}$ whose purpose is to approximate the source term average, as follows:

$$\overline{S} \simeq \frac{1}{\Delta t} \frac{1}{\Delta x} \int_0^{\Delta t} \int_{-\Delta x/2}^{\Delta x/2} S\Big(W_{\mathcal{R}}\Big(\frac{x}{t}; W_L, W_R\Big)\Big)\, dx\, dt. \tag{3.21}$$

The parameter $\overline{S}$ depends on W_L and W_R. It may also depend on other quantities, for instance the topography function Z in the case where S represents the topography source term. For the sake of simplicity, these dependencies are not explicitly written.

This parameter will be defined in the next section for the specific cases of the topography and the Manning friction. For the moment, we assume that such an approximation of the source term average is known, and that it is consistent. A more precise definition of the consistency of $\overline{S}$ will be given in the next section.

Equipped with $\overline{S}$, we impose that the intermediate states satisfy, instead of (3.19), the following equations, made by combining (3.19) with (3.21):

$$\lambda_R h_R^* - \lambda_L h_L^* = (\lambda_R - \lambda_L) h_{HLL}, \tag{3.22a}$$

$$\lambda_R q_R^* - \lambda_L q_L^* = (\lambda_R - \lambda_L) q_{HLL} + \overline{S} \Delta x. \tag{3.22b}$$

Now, let us introduce the steady state solutions of the balance law (3.2). The time derivative of such solutions vanishes. As a consequence, they are governed by the following equation:

$$\partial_x F(W) = \mathfrak{S}(W). \tag{3.23}$$

Arguing the definitions (3.3) of F and $\mathfrak{S}$ allows rewriting (3.23) as follows:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = S(W). \end{cases} \tag{3.24}$$

Therefore, the steady state solutions satisfy $q = \mathrm{cst}$. As usual, we denote this uniform value of the discharge by q_0.

We go back to the discrete level with the Riemann problem (3.4). We assume that $h_L > 0$

and $h_R > 0$. We elect to consider that the initial data of the Riemann problem (3.4) defines a steady state if the following equations hold:

$$\begin{cases} q_R - q_L = 0, \\ \left(\dfrac{q_R^2}{h_R} + \dfrac{1}{2} g h_R^2\right) - \left(\dfrac{q_L^2}{h_L} + \dfrac{1}{2} g h_L^2\right) = \overline{S} \Delta x. \end{cases} \tag{3.25}$$

The discrete steady relations (3.25) are nothing but a discrete version of the steady relations at the continuous level (3.24), with the source term $S(W)$ being approximated by $\overline{S}$. Using the usual jump notation $[X] = X_R - X_L$, the discrete steady state solutions are therefore defined as follows.

Definition 3.1. Two states $W_L = {}^t(h_L, q_L)$ and $W_R = {}^t(h_R, q_R)$, with $h_L > 0$ and $h_R > 0$, are said to *define a steady state* if the following relations hold:

$$\begin{cases} q_L = q_R = q_0, \\ q_0^2 \left[\dfrac{1}{h}\right] + \dfrac{g}{2} [h^2] = \overline{S} \Delta x. \end{cases} \tag{3.26}$$

Let us emphasize that $\overline{S}$ must be a consistent approximation of S, as evidenced by (3.21). As a consequence, the relations (3.26) impose that a suitable expression of $\overline{S}$ be derived. This expression must allow both consistency with S and recovery of the discrete steady state relations (3.26). This derivation is done in the next section, in the specific cases of the topography and the Manning friction source terms. At this level, we assume that such an expression is known.

Equipped with the discrete steady states (3.26), we can now propose a more precise definition of the well-balance property we seek. Indeed, we wish for intermediate states W_L^* and W_R^* which ensure that $W_L^* = W_L$ and $W_R^* = W_R$ as soon as a steady state is reached, i.e. as soon as W_L and W_R satisfy the discrete steady state relations (3.26). Note that the relations (3.24) coincide with the Riemann invariants (3.12) (since $Y(t, x) = x$). As a consequence, the definitions of the intermediate states are also based on the Riemann invariants.

Recall from Section 3.1.1.1 that the source term induces a stationary contact discontinuity, i.e. a contact discontinuity of velocity 0. Across this wave, the Riemann invariants (3.12) are constant. The approximate Riemann solver we are building involves three waves (see Figure 3.1), of respective velocities $\lambda_L < 0 < \lambda_R$. Hence, using the Riemann invariants for the stationary wave, as well as the source term approximation $\overline{S}$ given by (3.21), leads to imposing the following relations on the intermediate states W_L^* and W_R^*:

$$q_R^* - q_L^* = 0 \tag{3.27a}$$

$$\left(\frac{(q_R^*)^2}{h_R^*} + \frac{g}{2}(h_R^*)^2\right) - \left(\frac{(q_L^*)^2}{h_L^*} + \frac{g}{2}(h_L^*)^2\right) = \overline{S} \Delta x. \tag{3.27b}$$

As a consequence, (3.27a) imposes that q_L^* and q_R^* be taken equal. We take $q_L^* = q_R^*$, and we

denote this value by q^*. Equipped with q^*, the consistency relation (3.22b) rewrites as follows:

$$q^* = q_{HLL} + \frac{\overline{S}\Delta x}{\lambda_R - \lambda_L}. \tag{3.28}$$

Therefore, since $\overline{S}$ is assumed to be known, the above formula uniquely determines q^*.

Using that $q_L^* = q_R^* = q^*$, we compute a relation between h_L^* and h_R^*. The equation (3.27b) provides such a relation. However, this formula is nonlinear, and the formulas of h_L^* and h_R^* cannot be explicit. Note that (3.27b) rewrites as follows:

$$\left(-\frac{(q^*)^2}{h_L^* h_R^*} + \frac{g}{2}(h_L^* + h_R^*)\right)(h_R^* - h_L^*) = \overline{S}\Delta x. \tag{3.29}$$

In order to give explicit values to h_L^* and h_R^*, we consider the following linearization of (3.29):

$$\left(-\frac{(q^*)^2}{h_L h_R} + \frac{g}{2}(h_L + h_R)\right)(h_R^* - h_L^*) = \overline{S}\Delta x. \tag{3.30}$$

As a consequence, from the consistency relation (3.22a) and the linearized Riemann invariant (3.30), we obtain that h_L^* and h_R^* are solutions of the following linear system:

$$\begin{cases} \lambda_R h_R^* - \lambda_L h_L^* = (\lambda_R - \lambda_L) h_{HLL}, \\ \alpha(h_R^* - h_L^*) = \overline{S}\Delta x, \end{cases} \tag{3.31}$$

where the quantity α is defined by:

$$\alpha = -\frac{(q^*)^2}{h_L h_R} + \frac{g}{2}(h_L + h_R). \tag{3.32}$$

Solving (3.31) for h_L^* and h_R^*, we get:

$$h_L^* = h_{HLL} - \frac{\lambda_R \overline{S}\Delta x}{\alpha(\lambda_R - \lambda_L)}, \tag{3.33a}$$

$$h_R^* = h_{HLL} - \frac{\lambda_L \overline{S}\Delta x}{\alpha(\lambda_R - \lambda_L)}. \tag{3.33b}$$

Remark 3.2. Note that the expressions (3.28) of q^* and (3.33) of h_L^* and h_R^* ensure that we have $q^* = q_{HLL}$ and $h_L^* = h_R^* = h_{HLL}$ as soon as $\overline{S} = 0$. Therefore, if the approximate source term vanishes, then the suggested approximate Riemann solver degenerates into the HLL approximate Riemann solver, whose intermediate states are given by (3.18). Since the Godunov-type scheme associated to the HLL solver is entropy-satisfying, the stability of the current scheme is improved by having it degenerate to the HLL scheme in the absence of a source term.

The intermediate states $W_L^* = {}^t(h_L^*, q_L^*)$ and $W_R^* = {}^t(h_R^*, q_R^*)$ are thus completely and explicitly determined by $q_L^* = q_R^* = q^*$ and the relations (3.28) – (3.33). Note that these intermediate states are defined only for $h_L > 0$ and $h_R > 0$. Indeed, the quantity α is not defined as soon as $h_L = 0$ or $h_R = 0$, and the source term approximation $\overline{S}$ may also be

undefined when dealing with vanishing water heights.

Since $h_L > 0$ and $h_R > 0$ for now, we focus on a weaker notion of robustness, the *positivity preservation*. The scheme will be positivity-preserving if positive water heights at time t^n imply positive water heights at time t^{n+1}. From the expression (3.9) of the scheme, the positivity of h_L^* and h_R^* is a sufficient condition for the positivity of the scheme. However, we remark that the expressions (3.33) may lead to non-positive h_L^* or h_R^*, even if h_L and h_R are positive. Hence, the intermediate states (3.33) fail to ensure the positivity preservation of the scheme. A procedure to recover the positivity of the intermediate heights is presented in the next section.

3.1.1.4 Positivity

In this section, we suggest a modification of the intermediate water heights (3.33) to ensure the robustness of the scheme (3.9) while retaining the well-balance property. To address such an issue, we follow the procedure proposed in [7] (see also [15]). It consists in enforcing the positivity of h_L^* and h_R^*, while still ensuring that they satisfy the consistency relation (3.22a). Since h_L^* and h_R^* depend on h_{HLL}, we first state the following result, which concerns the sign of h_{HLL}.

Lemma 3.3. *With λ_L and λ_R defined by (3.6) and assuming that h_L and h_R are positive, the intermediate height of the HLL solver, defined by (3.20a) and labeled h_{HLL}, is necessarily positive.*

Proof. From (3.20a), we rewrite h_{HLL} as follows:

$$h_{HLL} = h_R \frac{\lambda_R - u_R}{\lambda_R - \lambda_L} + h_L \frac{u_L - \lambda_L}{\lambda_R - \lambda_L}. \tag{3.34}$$

Now, recall the definitions (3.6) of λ_L and λ_R. From these definitions, we immediately get:

$$\begin{aligned} \lambda_R &\geq |u_R| + c_R, \\ \lambda_L &\leq -|u_L| - c_L, \end{aligned}$$

where c_L and c_R are the left and right sound speeds, defined by $c_L = \sqrt{g h_L}$ and $c_R = \sqrt{g h_R}$. Therefore, (3.34) yields the following estimations of h_{HLL}:

$$\begin{aligned} h_{HLL} &\geq h_R \frac{|u_R| - u_R + c_R}{\lambda_R - \lambda_L} + h_L \frac{|u_L| + u_L + c_L}{\lambda_R - \lambda_L} \\ &\geq \frac{h_R c_R}{\lambda_R - \lambda_L} + \frac{h_L c_L}{\lambda_R - \lambda_L} \end{aligned}$$

Since $h_L > 0$ and $h_R > 0$, we have $c_L > 0$ and $c_R > 0$. As a consequence, we immediately get $h_{HLL} > 0$, which concludes the proof. □

In order to introduce the positivity preservation process, we define a small parameter ε. This parameter satisfies:

$$0 < \varepsilon \leq \min(h_L, h_R, h_{HLL}). \tag{3.35}$$

Equipped with the assumption that $h_L > 0$ and $h_R > 0$, as well as Lemma 3.3, the positivity

of ε is ensured, since it is lesser than the minimum of positive quantities. We now present the positivity preservation procedure.

(1) If $h_L^* < \varepsilon$, we take $h_L^* = \varepsilon$, and h_R^* is chosen according to (3.22a), to get:

$$\lambda_R h_R^* = \lambda_L \varepsilon + (\lambda_R - \lambda_L) h_{HLL},$$

which guarantees that $h_R^* > 0$ (see Figure 3.3).

(2) If $h_R^* < \varepsilon$, we take $h_R^* = \varepsilon$, and h_L^* is chosen according to (3.22a), to get:

$$\lambda_L h_L^* = \lambda_R \varepsilon - (\lambda_R - \lambda_L) h_{HLL},$$

which guarantees that $h_L^* > 0$, since $\lambda_L < 0$ (see Figure 3.3).

(3) Otherwise, we have $h_L^* \geq \varepsilon$ and $h_R^* \geq \varepsilon$: there is no need for the positivity procedure.

After the correction procedure, we have $h_L^* \geq \varepsilon$ and $h_R^* \geq \varepsilon$. As a consequence, we have recovered the positivity of the intermediate water heights.

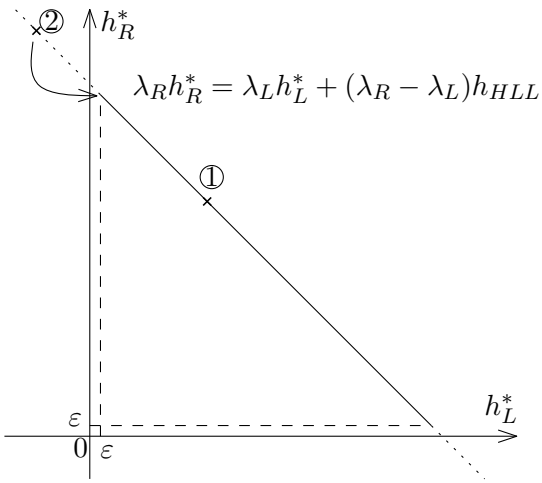

Figure 3.3 – Correction procedure to ensure positive and consistent intermediate water heights. The line represents the consistency equation (3.22a). If the point (h_L^*, h_R^*) belongs to the domain ①, then h_L^* and h_R^* are not modified. However, if (h_L^*, h_R^*) corresponds to a point within the domain ②, we replace (h_L^*, h_R^*) with $(\varepsilon, (1 - \frac{\lambda_L}{\lambda_R}) h_{HLL} + \frac{\lambda_L}{\lambda_R} \varepsilon)$, according to (3.22a).

We now combine the equations (3.28) – (3.33) with the positivity correction, to define the intermediate states W_L^* and W_R^* as follows:

$$W_L^* = \begin{pmatrix} h_L^* \\ q_L^* \end{pmatrix} \quad \text{and} \quad W_R^* = \begin{pmatrix} h_R^* \\ q_R^* \end{pmatrix}, \tag{3.36}$$

where the intermediate discharge and water heights are given by:

$$q_L^* = q_R^* = q^* = q_{HLL} + \frac{\overline{S}\Delta x}{\lambda_R - \lambda_L}, \tag{3.37a}$$

$$h_L^* = \min\left(\max\left(h_{HLL} - \frac{\lambda_R \overline{S}\Delta x}{\alpha(\lambda_R - \lambda_L)}, \varepsilon\right), \left(1 - \frac{\lambda_R}{\lambda_L}\right)h_{HLL} + \frac{\lambda_R}{\lambda_L}\varepsilon\right), \tag{3.37b}$$

$$h_R^* = \min\left(\max\left(h_{HLL} - \frac{\lambda_L \overline{S}\Delta x}{\alpha(\lambda_R - \lambda_L)}, \varepsilon\right), \left(1 - \frac{\lambda_L}{\lambda_R}\right)h_{HLL} + \frac{\lambda_L}{\lambda_R}\varepsilon\right), \tag{3.37c}$$

where α has been defined by (3.32), as follows:

$$\alpha = -\frac{(q^*)^2}{h_L h_R} + \frac{g}{2}(h_L + h_R), \tag{3.38}$$

and where the quantities h_{HLL} and q_{HLL} are defined by (3.20). The next section exhibits and proves the properties of the intermediate states we have derived.

3.1.1.5 Properties of the intermediate states

The following statement, regarding the properties of the intermediate states (3.36), holds.

Lemma 3.4. *Assume h_L and h_R to be positive. Then, the intermediate states W_L^* and W_R^* given by (3.36) satisfy the following properties:*

(i) consistency: the quantities h_L^*, h_R^*, q_L^* and q_R^* satisfy the equations (3.22);

(ii) positivity preservation: $h_L^* \geq \varepsilon$ and $h_R^* \geq \varepsilon$;

(iii) well-balance: if W_L and W_R define a steady state, i.e. if (3.26) holds, then $W_L^* = W_L$ and $W_R^* = W_R$.

Proof. Since ε is constrained by the estimations (3.35), we obviously get the required property (ii). Indeed, after (3.37b) and (3.37c), h_L^* and h_R^* stand for the minimum of quantities that are greater than or equal to ε. Hence, (ii) holds.

Next, let us set

$$\widetilde{h_L^*} = h_{HLL} - \frac{\lambda_R \overline{S}\Delta x}{\alpha(\lambda_R - \lambda_L)} \quad \text{and} \quad \widetilde{h_R^*} = h_{HLL} - \frac{\lambda_L \overline{S}\Delta x}{\alpha(\lambda_R - \lambda_L)}.$$

We immediately get the following identity:

$$\lambda_R \widetilde{h_R^*} - \lambda_L \widetilde{h_L^*} = (\lambda_R - \lambda_L)h_{HLL}, \tag{3.39}$$

which means that the heights $\widetilde{h_L^*}$ and $\widetilde{h_R^*}$ satisfy the consistency relation (3.22a). Since (3.22b) is obviously verified by q^*, the property (i) is established as soon as h_L^* and h_R^* are proven to satisfy (3.22a). Recall from Lemma 3.3 that $h_{HLL} > 0$. We have the following three configurations for the intermediate heights.

- If $\widetilde{h_L^*} \geq \varepsilon$ and $\widetilde{h_R^*} \geq \varepsilon$, then the relations (3.37) yield $h_L^* = \widetilde{h_L^*}$ and $h_R^* = \widetilde{h_R^*}$.
- If $\widetilde{h_L^*} < \varepsilon$, then from (3.37) we get $h_L^* = \varepsilon$ and $h_R^* = \left(1 - \frac{\lambda_L}{\lambda_R}\right)h_{HLL} + \frac{\lambda_L}{\lambda_R}\varepsilon$.

- Similarly, if $\widetilde{h_R^*} < \varepsilon$, then we have $h_R^* = \varepsilon$ and $h_L^* = \left(1 - \dfrac{\lambda_R}{\lambda_L}\right) h_{HLL} + \dfrac{\lambda_R}{\lambda_L}\varepsilon$.

We note that, in all three cases, the following identity systematically holds:

$$\lambda_R h_R^* - \lambda_L h_L^* = (\lambda_R - \lambda_L) h_{HLL}.$$

This identity turns out to be the consistency relation (3.22a). As a consequence, the property (i) is proven.

Finally, we have to check that the well-balance property is satisfied even in presence of the positivity correction. In order to prove the well-balance, we assume that W_L and W_R define a steady state, i.e. that (3.26) holds. Our goal is to show that, in this case, $W_L^* = W_L$ and $W_R^* = W_R$.

We begin by proving that $q_L^* = q_R^* = q_0$. From the definition (3.37a) of q^* and the steady relations (3.26) satisfied by W_L and W_R, we deduce that q^* satisfies the following sequence of equalities:

$$\begin{aligned} q^* &= \frac{\lambda_R q_0 - \lambda_L q_0}{\lambda_R - \lambda_L} - \frac{1}{\lambda_R - \lambda_L}\left[\frac{q_0^2}{h} + \frac{1}{2} g h^2\right] + \frac{\overline{S}\Delta x}{\lambda_R - \lambda_L} \\ &= q_0 - \frac{1}{\lambda_R - \lambda_L}\left(q_0^2\left[\frac{1}{h}\right] + \frac{g}{2}[h^2] - q_0^2\left[\frac{1}{h}\right] - \frac{g}{2}[h^2]\right) \\ &= q_0. \end{aligned}$$

As a consequence, we have $q_L^* = q_L = q_0$ and $q_R^* = q_R = q_0$.

We now prove that $h_L^* = h_L$ and $h_R^* = h_R$. First, let us compute $\overline{S}\Delta x/\alpha$ at the equilibrium using (3.26) and (3.38). We get the following equalities:

$$\frac{\overline{S}\Delta x}{\alpha} = \frac{q_0^2\left[\dfrac{1}{h}\right] + \dfrac{g}{2}[h^2]}{\dfrac{-q_0^2}{h_L h_R} + \dfrac{g}{2}(h_L + h_R)} = [h].$$

We then compute $\widetilde{h_L^*}$ at the equilibrium. According to (3.37b), we have:

$$\begin{aligned} \widetilde{h_L^*} &= \frac{\lambda_R h_R - \lambda_L h_L}{\lambda_R - \lambda_L} - \frac{[q]}{\lambda_R - \lambda_L} - \frac{\lambda_R \overline{S}\Delta x}{\alpha(\lambda_R - \lambda_L)} \\ &= \frac{\lambda_R h_R - \lambda_L h_L - \lambda_R h_R + \lambda_R h_L}{\lambda_R - \lambda_L} \\ &= h_L. \end{aligned}$$

Similarly, (3.37c) yields:

$$\begin{aligned} \widetilde{h_R^*} &= \frac{\lambda_R h_R - \lambda_L h_L}{\lambda_R - \lambda_L} - \frac{[q]}{\lambda_R - \lambda_L} - \frac{\lambda_L \overline{S}\Delta x}{\alpha(\lambda_R - \lambda_L)} \\ &= \frac{\lambda_R h_R - \lambda_L h_L - \lambda_L h_R + \lambda_L h_L}{\lambda_R - \lambda_L} \\ &= h_R. \end{aligned}$$

Moreover, from the definition (3.35) of ε, we have $h_L \geq \varepsilon$ and $h_R \geq \varepsilon$. Therefore $\widetilde{h_L^*} \geq \varepsilon$ and $\widetilde{h_R^*} \geq \varepsilon$. By construction of the positivity procedure, since $\widetilde{h_L^*}$ and $\widetilde{h_R^*}$ satisfy the consistency condition (3.39), we have $h_L^* = \widetilde{h_L^*}$ and $h_R^* = \widetilde{h_R^*}$ in this specific case of a discrete steady state governed by (3.26). As a consequence, $h_L^* = h_L$ and $h_R^* = h_R$.

Therefore, we have established that $W_L^* = W_L$ and $W_R^* = W_R$ as soon as W_L and W_R define a steady state. This concludes the proof of the well-balance property (iii), and Lemma 3.4 is thus proven. □

3.1.1.6 Properties of the scheme

Equipped with the intermediate states (3.36) and their properties given by Lemma 3.4, we can state the following result concerning the full scheme (3.9).

Theorem 3.5. *Consider $W_i^n \in \Omega^*$ for all $i \in \mathbb{Z}$, where Ω^* is the following restricted admissible states space:*

$$\Omega^* = \left\{ W = {}^t(h, q) \in \mathbb{R}^2 \, ; \, h > 0, \, q \in \mathbb{R} \right\}.$$

Assume that the intermediate states $W_{i+\frac{1}{2}}^{L,*}$ and $W_{i+\frac{1}{2}}^{R,*}$ are given, for all $i \in \mathbb{Z}$, by

$$W_{i+\frac{1}{2}}^{L,*} = \begin{pmatrix} h_L^*(W_i^n, W_{i+1}^n) \\ q_L^*(W_i^n, W_{i+1}^n) \end{pmatrix} \text{ and } \; W_{i+\frac{1}{2}}^{R,*} = \begin{pmatrix} h_R^*(W_i^n, W_{i+1}^n) \\ q_R^*(W_i^n, W_{i+1}^n) \end{pmatrix},$$

where q_L^* and q_R^* are defined by (3.37a), while h_L^* and h_R^* are respectively given by (3.37b) and (3.37c). Also, assume that the source term approximation $\overline{S}$ is consistent with the source term S according to (3.21). Finally, assume that, as soon as $(W_i^n)_{i\in\mathbb{Z}}$ defines a steady state, the approximation $\overline{S}$ verifies (3.26). Then, under the CFL restriction (3.1), the Godunov-type scheme (3.9) satisfies the following properties:

- *(i) consistency with the shallow-water system (3.1);*
- *(ii) positivity preservation: for all $i \in \mathbb{Z}$, $W_i^{n+1} \in \Omega^*$;*
- *(iii) well-balance: if $(W_i^n)_{i\in\mathbb{Z}}$ defines a steady state, i.e. if for all $i \in \mathbb{Z}$, W_i^n and W_{i+1}^n define a steady state, then for all $i \in \mathbb{Z}$, $W_i^{n+1} = W_i^n$.*

Proof. After [90], the consistency property (i) holds as soon as the approximate Riemann solver satisfies the integral consistency condition (3.13). After Lemma 3.4, the intermediate states (3.36) ensure that this integral consistency property is satisfied. As a consequence, (i) holds true.

We turn to proving the positivity preservation property (ii). By definition of Ω^*, this is equivalent to showing that, for all $i \in \mathbb{Z}$, $h_i^{n+1} > 0$ as soon as $h_i^n > 0$. We set $\varepsilon_i^n > 0$ constrained by (3.35), i.e. such that $\varepsilon_i^n \leq \min(h_i^n, \, h_{i+1}^n, \, h_{i+\frac{1}{2}}^{HLL})$, where $h_{i+\frac{1}{2}}^{HLL}$ is given by evaluating (3.20a) between the states W_i^n and W_{i+1}^n. The second item of Lemma 3.4 ensures that $h_{i+\frac{1}{2}}^{L,*} \geq \varepsilon_i^n$ and $h_{i+\frac{1}{2}}^{R,*} \geq \varepsilon_i^n$ as soon as $h_i^n > 0$ and $h_{i+1}^n > 0$. Since the scheme under consideration is given by (3.9), h_i^{n+1} turns out to be the sum of positive quantities, and (ii) is proven.

We finally need to prove the well-balance of the scheme (iii). Once again, this property comes from Lemma 3.4. Indeed, let us consider that $(W_i^n)_{i\in\mathbb{Z}}$ defines a steady state. Therefore,

for all $i \in \mathbb{Z}$, W_i^n and W_{i+1}^n define a steady state. As a consequence, from Lemma 3.4, we get for all $i \in \mathbb{Z}$ that $W_{i+\frac{1}{2}}^{L,*} = W_i^n$ and $W_{i+\frac{1}{2}}^{R,*} = W_{i+1}^n$. Hence, arguing the expression (3.9) of the scheme, we have $W_i^{n+1} = W_i^n$ for all $i \in \mathbb{Z}$, and the property (iii) holds. This concludes the proof of Theorem 3.5. □

Remark 3.6. Because of the arbitrary small parameter $\varepsilon > 0$, introduced in (3.36) to enforce the positivity of the intermediate water heights, the updated water height never vanishes. In the next section, we will present an extension of the scheme to deal with dry areas in the case where the expression of the source term is known and consists in the topography or the Manning friction. At this level, we reject vanishing water heights because of the unknown definitions of $\overline{S}$ and $\overline{S}/\alpha$, involved within the expressions (3.36) of the intermediate states. As soon as the full characterization of $\overline{S}$ is established, the scheme will be extended to allow $\varepsilon = 0$ in the definition (3.36).

3.1.2 Application to a specific class of source terms

With the intermediate states (3.36), Theorem 3.5 holds as soon as a suitable definition of the parameter $\overline{S}$ is provided. This section focuses on a specific class of source terms, to which the topography and the Manning friction source terms belong. We now assume that the generic source term S is given by:

$$S(W) = h^\beta f(q)\, \partial_x \sigma. \tag{3.40}$$

From now on, the topography source term will be labeled S^t, as follows:

$$S^t(W) = -gh\partial_x Z. \tag{3.41}$$

Note that S^t falls under the framework (3.40) if we set:

$$\beta = 1 \quad ; \quad f(q) = 1 \quad ; \quad \partial_x \sigma = -g\partial_x Z. \tag{3.42}$$

In addition, the Manning friction source term is now denoted by S^f, to get:

$$S^f(W) = -kq|q|h^{-\eta}. \tag{3.43}$$

The source term S^f can be written under the form (3.40) by taking:

$$\beta = -\eta \quad ; \quad f(q) = q|q| \quad ; \quad \partial_x \sigma = -k. \tag{3.44}$$

By adopting the source term given by (3.40), the smooth steady state solutions are governed by (3.24), as follows:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = h^\beta f(q)\, \partial_x \sigma. \end{cases} \tag{3.45}$$

The first equation of (3.45) obviously yields that the discharge $q = q_0$ is uniform. For smooth

steady states, the second equation reads as follows, after a division by h^β:

$$\left(-q_0^2 h^{-2-\beta} + g h^{1-\beta}\right)\partial_x h = f(q_0)\partial_x \sigma.$$

As a consequence, the following identity governs the steady state solutions:

$$\frac{q_0^2}{1+\beta}\partial_x\left(h^{-1-\beta}\right) + \frac{g}{2-\beta}\partial_x\left(h^{2-\beta}\right) - f(q_0)\partial_x\sigma = 0.$$

Hence, at the discrete level, the following algebraic relation describes the *smooth* steady state solutions for nonzero water heights:

$$\frac{q_0^2}{1+\beta}\left[h^{-1-\beta}\right] + \frac{g}{2-\beta}\left[h^{2-\beta}\right] - f(q_0)[\sigma] = 0. \tag{3.46}$$

Therefore, the discrete smooth steady states are governed by $q_L = q_R = q_0$, as well as the following relations, made by combining (3.26) and (3.46):

$$\begin{cases} q_0^2\left[\dfrac{1}{h}\right] + \dfrac{g}{2}[h^2] = \overline{S}\Delta x, \\ \dfrac{q_0^2}{1+\beta}\left[h^{-1-\beta}\right] + \dfrac{g}{2-\beta}\left[h^{2-\beta}\right] - f(q_0)[\sigma] = 0. \end{cases} \tag{3.47}$$

The system (3.47) is a nonlinear system of two equations, whose two unknowns are $\overline{S}$ and q_0. To solve this system, we note that the only unknown present in the second equation of (3.47) is q_0. As a consequence, this nonlinear equation may be solved to obtained the value of q_0. Then, the first equation yields an expression of $\overline{S}$ when W_L and W_R define a steady state. This expression depends only on W_L, W_R, σ_L and σ_R, and it can be used even when W_L and W_R do not define a steady state.

However, solving the system (3.47) for $\overline{S}$ is not possible in the general case, where $f(q_0)$ is an unknown. In the next sections, we solve this system in the specific cases of the topography source term or the Manning friction source term. Comments are then given on combining both source terms and on vanishing water heights.

3.1.2.1 Approximate topography source term

In this section, we consider the topography source term S^t, given by (3.41). The goal of this section is to compute a suitable parameter $\overline{S}$ to approximate S^t. In this specific case of the topography source term, we shall denote this parameter by $\overline{S}^t$. We still, for the moment, assume that $h_L > 0$ and $h_R > 0$.

According to (3.42), the steady relations (3.47) rewrite as follows in the present case:

$$q_0^2\left[\frac{1}{h}\right] + \frac{g}{2}[h^2] = \overline{S}^t\Delta x, \tag{3.48a}$$

$$\frac{q_0^2}{2}\left[\frac{1}{h^2}\right] + g[h] + g[Z] = 0. \tag{3.48b}$$

We now exhibit the expression of $\overline{S}^t$ from the above identities. First, from (3.48b), we extract

the following expression of q_0^2:

$$q_0^2 = 2g[h+Z]\frac{h_L^2 h_R^2}{h_R^2 - h_L^2},$$

which is plugged into (3.48a) to get the following definition of $\overline{S}^t$:

$$\overline{S}^t \Delta x = \frac{g}{2}\left[h^2\right] - g[h+Z]\frac{2h_L h_R}{h_L + h_R}.$$

The above expression can be rewritten as follows, after straightforward computations:

$$\overline{S}^t \Delta x = -g[Z]\frac{2h_L h_R}{h_L + h_R} + \frac{g}{2}\frac{[h]^3}{h_L + h_R}. \tag{3.49}$$

Let us emphasize that such a definition of the approximate topography source term can be found in the literature. For instance, the reader is referred to [12, 13] (see also [128] for related expressions). When the water height and the topography are smooth functions, this expression of the approximate source term is consistent with the source term at the continuous level. This result is proven below.

Lemma 3.7. *If the water height is a smooth function, then the expression of $\overline{S}^t$ given by (3.49) is consistent with S^t.*

Proof. With a smooth water height and a smooth topography function, we take $h_L = h(x)$ and $h_R = h(x + \mathcal{O}(\Delta x))$, as well as $Z_L = Z(x)$ and $Z_R = Z(x + \mathcal{O}(\Delta x))$, in (3.49). Taylor's formula applied to the quantities h_R and Z_R yields:

$$h_R = h + \Delta x \partial_x h + \mathcal{O}(\Delta x^2) \quad \text{and} \quad Z_R = Z + \Delta x \partial_x Z + \mathcal{O}(\Delta x^2).$$

As a consequence, we get:

$$h_L + h_R = 2h + \mathcal{O}(\Delta x). \tag{3.50}$$

In addition, we immediately have $[h]^3 = \mathcal{O}(\Delta x^3)$, and we get, for the second term in the expression (3.49) of $\overline{S}^t$:

$$\frac{g}{2\Delta x}\frac{[h]_c^3}{h_L + h_R} = \frac{\mathcal{O}(\Delta x^2)}{2h + \mathcal{O}(\Delta x)} = \mathcal{O}(\Delta x^2). \tag{3.51}$$

Moreover, for the first term of $\overline{S}^t$, we have:

$$\frac{[Z]}{\Delta x} = \frac{Z + \Delta x \partial_x Z + \mathcal{O}(\Delta x^2) - Z}{\Delta x} = \partial_x Z + \mathcal{O}(\Delta x). \tag{3.52}$$

In addition, with (3.50), we get:

$$\frac{2h_L h_R}{h_L + h_R} = \frac{2h(h + \Delta x \partial_x Z + \mathcal{O}(\Delta x^2))}{2h + \mathcal{O}(\Delta x)} = h + \mathcal{O}(\Delta x). \tag{3.53}$$

Combining the equations (3.51), (3.52) and (3.53) finally yields $\overline{S}^t = -gh\partial_x Z + \mathcal{O}(\Delta x)$, which concludes the proof. □

An important ingredient in the consistency of the scheme is that the source term approximation $\overline{S}$ has to be consistent with the source term S. In the present case, $\overline{S}^t$ has to be consistent with the actual source term $-gh\partial_x Z$, assuming positive water heights. For instance, when the topography is flat, i.e. $[Z] = 0$, the actual topography source term vanishes. Therefore, in order for $\overline{S}^t$ to be consistent with the actual source term, we need $\overline{S}^t = \mathcal{O}(\Delta x)$ as soon as the topography is flat. However, as underlined in [12, 13, 123, 128], $\overline{S}^t$ is no longer consistent with zero when the topography is flat and the water height is not smooth. Indeed, in this case, we have

$$\overline{S}^t = \frac{g}{2(h_L + h_R)} \frac{[h]^3}{\Delta x} \neq \mathcal{O}(\Delta x).$$

In order to recover the required consistency, i.e. $\overline{S}^t = \mathcal{O}(\Delta x)$ for a flat topography, we adopt the strategy proposed in [12, 13, 123]. We modify $\overline{S}^t$ as follows:

$$\overline{S}^t \Delta x = -g[Z] \frac{2h_L h_R}{h_L + h_R} + \frac{g}{2} \frac{[h]_c^3}{h_L + h_R}. \tag{3.54}$$

In (3.54), $[h]_c$ is a cutoff of $[h] = h_R - h_L$, defined as follows:

$$[h]_c = \begin{cases} h_R - h_L & \text{if } |h_R - h_L| \leq C\Delta x, \\ \operatorname{sgn}(h_R - h_L)\, C\Delta x & \text{otherwise,} \end{cases} \tag{3.55}$$

with C a positive constant that does not depend on Δx. This new expression of $\overline{S}^t$ is consistent with the topography source term S^t. Indeed, for a flat topography, (3.54) becomes:

$$\overline{S}^t = \frac{g}{2(h_L + h_R)} \frac{[h]_c^3}{\Delta x}. \tag{3.56}$$

Note that the cutoff procedure enforces $|[h]_c| \leq C\Delta x$. Therefore, according to (3.56), we have $\overline{S}^t = \mathcal{O}(\Delta x^2)$ as soon as the topography is flat. However, the source term approximation $\overline{S}^t$ does not vanish when the topography is flat, and therefore the scheme does not reduce to a conservative scheme in that case.

Remark 3.8. For a smooth water height h, the relation $h_R - h_L = \mathcal{O}(\Delta x)$ obviously holds, and there exists $K \in \mathbb{R}_+^*$ such that $|h_R - h_L| \leq K\Delta x$. As a consequence, for a smooth water height, there exists C such that $[h]_c = [h]$, with $[h]_c$ given by (3.55). Indeed, taking $C < K$ suffices. In this case, $\overline{S}^t$ is given by (3.49). Thus, the relation (3.48a) holds by construction, and Theorem 3.5 ensures that the suggested intermediate states (3.37) are well-balanced. Hence, the cutoff procedure does not interfere with the well-balance property of the intermediate states.

Lemma 3.9. *The expression of $\overline{S}^t$ given by (3.54) is consistent with S^t.*

Proof. From Lemma 3.7, we know that the first term of $\overline{S}^t$ is consistent with $-gh\partial_x Z$. Now, note that, from the cutoff procedure (3.55), we have $|[h]_c| \leq C\Delta x$. Therefore, $[h]_c^3 = \mathcal{O}(\Delta x^3)$, and $\overline{S}^t$ is necessarily consistent with $-gh\partial_x Z$ for any h_L and h_R, which concludes the proof. □

3.1.2.2 Approximate friction source term

We now turn to the friction source term S^f, given by (3.43). In this section, we derive a suitable approximation $\overline{S}^f$ of S^f, to be plugged into the intermediate states (3.37). We still assume that $h_L > 0$ and $h_R > 0$. Since S^f is given by (3.44), the steady relations (3.47) now read:

$$q_0^2 \left[\frac{1}{h}\right] + \frac{g}{2}\left[h^2\right] = \overline{S}^f \Delta x, \tag{3.57a}$$

$$-\frac{q_0^2}{\eta - 1}\left[h^{\eta-1}\right] + \frac{g}{\eta+2}\left[h^{\eta+2}\right] + k q_0 |q_0| \Delta x = 0. \tag{3.57b}$$

From (3.57b), we get the following expression of q_0^2 for a steady state solution:

$$q_0^2 = \frac{g\dfrac{[h^{\eta+2}]}{\eta+2}}{\dfrac{[h^{\eta-1}]}{\eta-1} - k\mu_0\,\Delta x}, \tag{3.58}$$

where $\mu_0 = \mathrm{sgn}(q_0)$ denotes the direction of the steady water flow. Now, to obtain a suitable expression of $\overline{S}^f$, we take:

$$\overline{S}^f = -k\bar{q}|\bar{q}|\overline{h^{-\eta}}, \tag{3.59}$$

where the parameter $\bar{q}$ is consistent with q, and the parameter $\overline{h^{-\eta}}$ is consistent with $h^{-\eta}$. We emphasize that finding $\overline{S}^f$ now amounts to determining suitable parameters $\bar{q}$ and $\overline{h^{-\eta}}$. As a consequence, as soon as a steady state is considered, the quantity $\bar{q}$ has to be equal to q_0. Therefore, the steady relation (3.57a) becomes:

$$q_0^2 \left[\frac{1}{h}\right] + \frac{g}{2}\left[h^2\right] = -k q_0^2 \mu_0 \overline{h^{-\eta}} \Delta x.$$

Equipped with the formula (3.58), which gives the value of q_0^2 when a steady state is considered, the above equation yields the following formula for $\overline{h^{-\eta}}$:

$$\overline{h^{-\eta}} = \frac{[h^2]}{2}\frac{\eta+2}{[h^{\eta+2}]} - \frac{\mu_0}{k\,\Delta x}\left(\left[\frac{1}{h}\right] + \frac{[h^2]}{2}\frac{[h^{\eta-1}]}{\eta-1}\frac{\eta+2}{[h^{\eta+2}]}\right). \tag{3.60}$$

Concerning $\bar{q}$, we choose the following average:

$$\begin{cases} \bar{q} = \dfrac{2|q_L||q_R|}{|q_L| + |q_R|}\,\mathrm{sgn}(q_L + q_R) & \text{if } q_L \neq 0 \text{ and } q_R \neq 0; \\ \bar{q} = 0 & \text{if } q_L = 0, q_R = 0 \text{ or } k = 0. \end{cases} \tag{3.61}$$

This average indeed ensures that, if $q_L = q_R$, then $\bar{q} = q_L = q_R$. In particular, if a steady state is considered, we have $q_L = q_R = q_0$; hence, $\bar{q} = q_0$ in this case. In addition, $\bar{q}$ is consistent with q.

Now, note that the expression (3.60) of $\overline{h^{-\eta}}$ contains μ_0. This quantity depends on the steady state; it would have to be determined for non-steady states. To address such an issue,

we suggest the expression

$$\overline{h^{-\eta}} := \overline{h^{-\eta}}(h_L, h_R) = \frac{[h^2]}{2}\frac{\eta+2}{[h^{\eta+2}]} - \frac{\bar{\mu}}{k\,\Delta x}\left(\left[\frac{1}{h}\right] + \frac{[h^2]}{2}\frac{[h^{\eta-1}]}{\eta-1}\frac{\eta+2}{[h^{\eta+2}]}\right), \tag{3.62}$$

where $\bar{\mu}$ is the sign of the quantity $\bar{q}$ given by (3.61).

Lemma 3.10. *The expression of $\overline{h^{-\eta}}$ given by (3.62) is consistent with $h^{-\eta}$.*

Proof. With smooth water heights, we fix, in (3.62), $h_L = h(x)$ and $h_R = h(x + \mathcal{O}(\Delta x))$. Taylor's formula applied to h_R yields $h_R = h + \Delta x \partial_x h + \mathcal{O}(\Delta x^2)$. In order to evaluate the Taylor expansions of $[h^2]$, $[h^{\eta-1}]$, $[h^{\eta+2}]$ and $[h^{-1}]$, we now compute a Taylor expansion, for some $\gamma \in \mathbb{R}$, of the jump $[h^\gamma]$:

$$\begin{aligned}[h^\gamma] = h_R^\gamma - h_L^\gamma &= \big(h + \partial_x h \Delta x + \mathcal{O}(\Delta x^2)\big)^\gamma - h^\gamma \\ &= h^\gamma\big(1 + \gamma h^{-1}\partial_x h \Delta x + \mathcal{O}(\Delta x^2)\big) - h^\gamma \\ &= \gamma h^{\gamma-1}\partial_x h \Delta x + \mathcal{O}(\Delta x^2).\end{aligned}$$

Using the above evaluation, we have, for the first part of the expression of $\overline{h^{-\eta}}$:

$$\frac{[h^2]}{2}\frac{\eta+2}{[h^{\eta+2}]} = \frac{h\partial_x h \Delta x + \mathcal{O}(\Delta x^2)}{h^{\eta+1}\partial_x h \Delta x + \mathcal{O}(\Delta x^2)} = h^{-\eta} + \mathcal{O}(\Delta x). \tag{3.63}$$

Moreover, we have the following Taylor expansion:

$$\left[\frac{1}{h}\right] = -h^{-2}\partial_x h \Delta x + \mathcal{O}(\Delta x^2). \tag{3.64}$$

In addition, we get the following sequence of equalities:

$$\begin{aligned}\frac{[h^2]}{2}\frac{[h^{\eta-1}]}{\eta-1}\frac{\eta+2}{[h^{\eta+2}]} &= \frac{(h\partial_x h \Delta x + \mathcal{O}(\Delta x^2))(h^{\eta-2}\partial_x h \Delta x + \mathcal{O}(\Delta x^2))}{h^{\eta+1}\partial_x h \Delta x + \mathcal{O}(\Delta x^2)} \\ &= h^{-2}\partial_x h \Delta x + \mathcal{O}(\Delta x^2).\end{aligned} \tag{3.65}$$

Combining both equations (3.64) and (3.65) immediately yields the Taylor expansion of the second part of the expression (3.62) of $\overline{h^{-\eta}}$:

$$-\frac{\bar{\mu}}{k\,\Delta x}\left(\left[\frac{1}{h}\right] + \frac{[h^2]}{2}\frac{[h^{\eta-1}]}{\eta-1}\frac{\eta+2}{[h^{\eta+2}]}\right) = -\frac{\bar{\mu}}{k\,\Delta x}\mathcal{O}(\Delta x^2) = \mathcal{O}(\Delta x). \tag{3.66}$$

Since $\overline{h^{-\eta}}$ is given by (3.62), using both relations (3.63) and (3.66) gives $\overline{h^{-\eta}} = h^{-\eta} + \mathcal{O}(\Delta x)$, which concludes the proof. □

Equipped with the respective expressions (3.61) and (3.62) of $\bar{q}$ and $\overline{h^{-\eta}}$, we have fully determined the approximate friction source term $\overline{S}^f$, given by (3.59). After Lemma 3.10 and the expression of $\bar{q}$, this approximate source term is consistent with S^f. In addition, by construction, $\overline{S}^f$ satisfies the discrete steady state relation (3.57a) as soon as W_L and W_R define a steady state.

3.1.2.3 The case of both topography and friction source terms

Equipped with the approximations $\overline{S}^t$ of S^t and $\overline{S}^f$ of S^f, respectively given by (3.54) and (3.59), we now turn to the approximation of the source term made of both contributions of topography and friction. As a consequence, we consider the following source term in (3.1):

$$S(W) = S^t(W) + S^f(W) = -gh\partial_x Z - kq|q|h^{-\eta}.$$

The steady state solutions of the shallow-water system endowed with this source term are given by (3.24), as follows:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = S^t(W) + S^f(W). \end{cases} \tag{3.67}$$

Because of the presence of both source terms, the second equation of (3.67) cannot be put under an algebraic form similar to (3.46). Therefore, we cannot derive an approximation of the source term $S = S^t + S^f$ the same way we derived the approximations $\overline{S}^t$ and $\overline{S}^f$. Instead, we elect to discretize the second equation of (3.67) using $\overline{S}^t$ and $\overline{S}^f$, as follows:

$$\left[\frac{q^2}{h} + \frac{1}{2} g h^2 \right] = \overline{S}^t \Delta x + \overline{S}^f \Delta x. \tag{3.68}$$

This discretization has been obtained by taking $\overline{S} = \overline{S}^t + \overline{S}^f$ in (3.26). As a consequence, after (3.37), we define the following intermediate discharges and heights:

$$q_L^* = q_R^* = q^* = q_{HLL} + \frac{\overline{S}^t \Delta x}{\lambda_R - \lambda_L} + \frac{\overline{S}^f \Delta x}{\lambda_R - \lambda_L}, \tag{3.69a}$$

$$h_L^* = \min\left(\max\left(h_{HLL} - \frac{\lambda_R \overline{S}^t \Delta x}{\alpha(\lambda_R - \lambda_L)} - \frac{\lambda_R \overline{S}^f \Delta x}{\alpha(\lambda_R - \lambda_L)}, \varepsilon \right), \left(1 - \frac{\lambda_R}{\lambda_L} \right) h_{HLL} + \frac{\lambda_R}{\lambda_L} \varepsilon \right), \tag{3.69b}$$

$$h_R^* = \min\left(\max\left(h_{HLL} - \frac{\lambda_L \overline{S}^t \Delta x}{\alpha(\lambda_R - \lambda_L)} - \frac{\lambda_L \overline{S}^f \Delta x}{\alpha(\lambda_R - \lambda_L)}, \varepsilon \right), \left(1 - \frac{\lambda_L}{\lambda_R} \right) h_{HLL} + \frac{\lambda_L}{\lambda_R} \varepsilon \right), \tag{3.69c}$$

where the quantity α is given by (3.38).

3.1.2.4 Extension of the approximate source terms for vanishing water heights

The intermediate states (3.69) have been derived for nonzero water heights. We now suggest an extension of (3.69) to deal with vanishing water heights. To that end, we recall the assumption made earlier.

Assumption. When the water height vanishes, so does the velocity.

Now, we need to provide expressions of $\overline{S}^t$ and $\overline{S}^f$ when h_L or h_R vanishes, and when both h_L and h_R vanish. In addition, since α has been defined by (3.38) for positive water heights, the expressions $\overline{S}^t/\alpha$ and $\overline{S}^f/\alpha$ also need to be extended in order to take vanishing water heights into account.

Extension of $\overline{S}^t$ for vanishing water heights

We first determine a new expression of $\overline{S}^t$ for vanishing h_L and/or h_R. Since the expression (3.54) relied on the assumption that both h_L and h_R were positive, we cannot use this expression in the present case. We momentarily assume that the friction contribution vanishes, in order to derive an expression for the approximate topography source term $\overline{S}^t$.

In order to obtain a new formula for $\overline{S}^t$, we begin by assuming that W_L and W_R define a steady state with vanishing h_L or h_R, but not both h_L and h_R. From Proposition 1.14, we have $q_0 = 0$ as soon as W_L and W_R define a steady state. The steady state under consideration is therefore a lake at rest steady state governed by (1.43), that is to say a steady state with $[h + Z] = 0$. Note that (3.48a) can be rewritten as follows:

$$[hu^2] + \frac{g}{2}[h^2] = \overline{S}^t \Delta x \tag{3.70}$$

The above assumption ensures that $u_L = u_R = 0$. As a consequence, (3.70) reads:

$$g[h]\frac{h_R + h_L}{2} = \overline{S}^t \Delta x.$$

Now, plugging $[h] = -[Z]$ into this equality, we get the new expression of $\overline{S}^t \Delta x$, to be substituted to (3.54) as soon as either h_L or h_R vanishes:

$$\overline{S}^t \Delta x = -g(Z_R - Z_L)\frac{h_R + h_L}{2}. \tag{3.71}$$

The expression (3.71) of $\overline{S}^t$ is obviously consistent with the actual source term S^t given by (3.41).

Then, note that the lake at rest condition $[h + Z] = 0$, which comes from studying the smooth steady states, does not cover two cases of a physical lake at rest with a dry area and a discontinuous topography. Namely, the cases displayed on Figure 3.4 are physical steady states at rest which do not satisfy $[h + Z] = 0$.

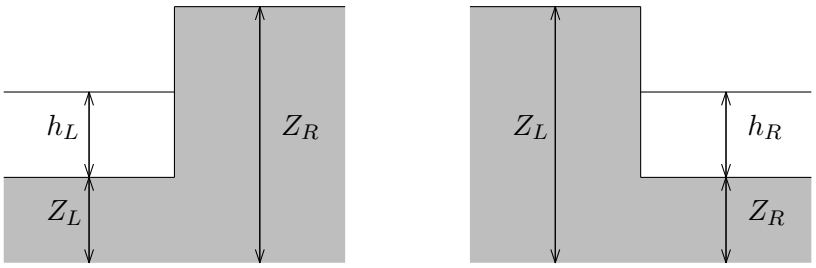

Figure 3.4 – Physical lake at rest configurations not governed by $[h + Z] = 0$. Left panel: lake at rest with $h_R = 0$ and $h_L + Z_L \leq Z_R$. Right panel: lake at rest with $h_L = 0$ and $h_R + Z_R \leq Z_L$.

Therefore, according to Figure 3.4, the states given by

$$\begin{cases} q_0 = 0, \\ h_R = 0, \\ h_L + Z_L \leq Z_R, \end{cases} \quad \text{or} \quad \begin{cases} q_0 = 0, \\ h_L = 0, \\ h_R + Z_R \leq Z_L, \end{cases} \tag{3.72}$$

do not satisfy $[h+Z]=0$ but are steady states at rest. Such steady state solutions are not included in (3.71), and considering only this expression will not allow their preservation. As a consequence, in the cases described by (3.72), we modify the expression of $\overline{S}^t$. Still with $u_L=u_R=0$, the equation (3.70) yields:

$$\overline{S}^t\Delta x=\begin{cases} g\dfrac{h_R^2}{2} & \text{if } q_L=q_R=0,\ h_L=0 \text{ and } h_R+Z_R\leq Z_L,\\ -g\dfrac{h_L^2}{2} & \text{if } q_L=q_R=0,\ h_R=0 \text{ and } h_L+Z_L\leq Z_R. \end{cases}$$

Note that $h_L=0$ implies $q_L=0$, and that $h_R=0$ implies $q_R=0$. Therefore, the above expressions read:

$$\overline{S}^t\Delta x=\begin{cases} g\dfrac{h_R^2}{2} & \text{if } q_R=0,\ h_L=0 \text{ and } h_R+Z_R\leq Z_L,\\ -g\dfrac{h_L^2}{2} & \text{if } q_L=0,\ h_R=0 \text{ and } h_L+Z_L\leq Z_R. \end{cases} \tag{3.73}$$

Finally, we handle the case where both h_L and h_R vanish (and thus $q_L=q_R=0$), i.e. the case where there is no water. Note that, in this case, we have $q_{HLL}=0$ after (3.20b). In order for the discharge to stay equal to zero, we have to make sure that $q^*=0$. To meet this requirement, we enforce

$$\overline{S}^t\Delta x=0. \tag{3.74}$$

as soon as both h_L and h_R are zero. This expression makes sense since the actual source term S^t also vanishes in the absence of water.

We now regroup the four cases (3.54), (3.71), (3.73) and (3.74), to get the following final expression of $\overline{S}^t$:

$$\begin{aligned}\overline{S}^t\Delta x &:= \overline{S}^t(h_L,h_R,q_L,q_R,Z_L,Z_R,\Delta x)\Delta x\\ &=\begin{cases} 0 & \text{if } h_L=0 \text{ and } h_R=0,\\ g\dfrac{h_R^2}{2} & \text{if } q_R=0,\ h_L=0 \text{ and } h_R+Z_R\leq Z_L,\\ -g\dfrac{h_L^2}{2} & \text{if } q_L=0,\ h_R=0 \text{ and } h_L+Z_L\leq Z_R,\\ -g[Z]\dfrac{h_R+h_L}{2} & \text{if } h_L=0 \text{ or } h_R=0,\\ -g[Z]\dfrac{2h_Lh_R}{h_L+h_R}+\dfrac{g}{2}\dfrac{[h]_c^3}{h_L+h_R} & \text{otherwise,}\end{cases}\end{aligned} \tag{3.75}$$

with $[h]_c$ given by (3.55).

Equipped with the expression (3.75) of $\overline{S}^t$, we now turn to providing a suitable expression of $\overline{S}^t/\alpha$. Recall that α has first been introduced to replace the leftmost term of (3.30). Note that this term is ill-defined for $h_L=0$ or $h_R=0$. In order to determine a suitable expression of α for $h_L=0$ or $h_R=0$, recall that, after Proposition 1.14, we have $q^*=q_0=0$ as soon as W_L and W_R define a steady state with a vanishing water height. Therefore, $u_L=u_R=0$ and,

as soon as $h_L = 0$ or $h_R = 0$, α is defined as follows:

$$\alpha = \frac{g}{2}(h_L + h_R). \tag{3.76}$$

Then, we consider $h_L = h_R = 0$. As a consequence of such a dry area, we have $q_L = q_R = 0$. In this case, after (3.20a), we have $h_{HLL} = 0$. For this area to stay dry, we need to ensure that $h_L^* = h_R^* = 0$. This requirement is met by enforcing, as soon as both h_L and h_R are zero, the following value of $\overline{S}^t \Delta x/\alpha$:

$$\frac{\overline{S}^t \Delta x}{\alpha} = 0. \tag{3.77}$$

Finally, we define $\overline{S}^t \Delta x/\alpha$ as follows, using (3.38), (3.71), (3.76) and (3.77):

$$\frac{\overline{S}^t \Delta x}{\alpha} = \begin{cases} 0 & \text{if } h_L = 0 \text{ and } h_R = 0, \\ h_R & \text{if } q_R = 0,\ h_L = 0 \text{ and } h_R + Z_R \leq Z_L, \\ -h_L & \text{if } q_L = 0,\ h_R = 0 \text{ and } h_L + Z_L \leq Z_R, \\ -[Z] & \text{if } h_L = 0 \text{ or } h_R = 0, \\ \dfrac{\overline{S}^t \Delta x}{-\frac{(q^*)^2}{h_L h_R} + \frac{g}{2}(h_L + h_R)} & \text{otherwise.} \end{cases} \tag{3.78}$$

Extension of $\overline{S}^f$ for vanishing water heights

Then, to extend the approximate friction source term $\overline{S}^f$, we recall the following assumption made on the friction source term in the presence of vanishing water heights.

Assumption. The friction source term vanishes as soon as the water height does.

In order for both quantities $\overline{S}^f$ and $\overline{S}^f/\alpha$ to satisfy this assumption, we have to impose that they vanish when h_L and/or h_R vanishes.

As a consequence, after (3.59) and the above assumption, $\overline{S}^f$ is given by:

$$\begin{aligned} \overline{S}^f \Delta x &:= \overline{S}^f(h_L, h_R, q_L, q_R, \Delta x)\Delta x \\ &= \begin{cases} 0 & \text{if } h_L = 0 \text{ and/or } h_R = 0, \\ -k\bar{q}|\bar{q}|\overline{h^{-\eta}} & \text{otherwise,} \end{cases} \end{aligned} \tag{3.79}$$

where $\bar{q}$ is given by (3.61) and $\overline{h^{-\eta}}$ is defined by (3.62). In addition, the quantity $\overline{S}^f/\alpha$ is given by:

$$\frac{\overline{S}^f \Delta x}{\alpha} = \begin{cases} 0 & \text{if } h_L = 0 \text{ and/or } h_R = 0, \\ \dfrac{-k\bar{q}|\bar{q}|\overline{h^{-\eta}}}{-\frac{(q^*)^2}{h_L h_R} + \frac{g}{2}(h_L + h_R)} & \text{otherwise.} \end{cases} \tag{3.80}$$

3.1.2.5 Properties of the scheme with both source terms

Equipped with the expressions (3.75), (3.78), (3.79) and (3.80) of the approximate topography and friction source terms, we can now extend the formulas (3.69) of the intermediate states in order to take vanishing water heights into account. Recall that the parameter $\varepsilon > 0$ prevented the intermediate heights from vanishing. To allow vanishing intermediate heights, we take $\varepsilon = 0$ in (3.69), to get the following intermediate states:

$$q_L^* = q_R^* = q^* = q_{HLL} + \frac{\overline{S}^t \Delta x}{\lambda_R - \lambda_L} + \frac{\overline{S}^f \Delta x}{\lambda_R - \lambda_L}, \tag{3.81a}$$

$$h_L^* = \min\left(\left(h_{HLL} - \frac{\lambda_R \overline{S}^t \Delta x}{\alpha(\lambda_R - \lambda_L)} - \frac{\lambda_R \overline{S}^f \Delta x}{\alpha(\lambda_R - \lambda_L)} \right)_+, \left(1 - \frac{\lambda_R}{\lambda_L}\right) h_{HLL} \right), \tag{3.81b}$$

$$h_R^* = \min\left(\left(h_{HLL} - \frac{\lambda_L \overline{S}^t \Delta x}{\alpha(\lambda_R - \lambda_L)} - \frac{\lambda_L \overline{S}^f \Delta x}{\alpha(\lambda_R - \lambda_L)} \right)_+, \left(1 - \frac{\lambda_L}{\lambda_R}\right) h_{HLL} \right), \tag{3.81c}$$

where $(X)_+ = \max(X, 0)$ denotes the positive part of a quantity X. The intermediate states (3.81) allow us to state the following extension of Lemma 3.4 for non-negative water heights.

Lemma 3.11. *Assume $h_L \geq 0$ and $h_R \geq 0$. Then, the intermediate states W_L^* and W_R^* given by (3.81) satisfy the following properties:*

(i) consistency: the quantities h_L^*, h_R^*, q_L^* and q_R^* satisfy the equations (3.22), where $\overline{S} = \overline{S}^t + \overline{S}^f$;

(ii) non-negativity preservation: $h_L^* \geq 0$ and $h_R^* \geq 0$;

(iii) well-balance: if W_L and W_R define a steady state, i.e. if (3.26) holds, then $W_L^* = W_L$ and $W_R^* = W_R$.

Proof. Concerning (i), q_L^* and q_R^* given by (3.81) are immediately shown to satisfy the consistency equations (3.22b) with $\overline{S} = \overline{S}^t + \overline{S}^f$. Let us introduce the following notations:

$$\widetilde{h_L^*} = h_{HLL} - \frac{\lambda_R \overline{S}^t \Delta x}{\alpha(\lambda_R - \lambda_L)} - \frac{\lambda_R \overline{S}^f \Delta x}{\alpha(\lambda_R - \lambda_L)} \quad \text{and} \quad \widetilde{h_R^*} = h_{HLL} - \frac{\lambda_L \overline{S}^t \Delta x}{\alpha(\lambda_R - \lambda_L)} - \frac{\lambda_L \overline{S}^f \Delta x}{\alpha(\lambda_R - \lambda_L)}.$$

The quantities $\widetilde{h_L^*}$ and $\widetilde{h_R^*}$ immediately satisfy the required consistency property (3.22a). Regarding h_L^* and h_R^*, the following three cases arise.

- If $\widetilde{h_L^*} \geq 0$ and $\widetilde{h_R^*} \geq 0$, then the relations (3.81) yield $h_L^* = \widetilde{h_L^*}$ and $h_R^* = \widetilde{h_R^*}$.
- If $\widetilde{h_L^*} < 0$, then from (3.37) we get $h_L^* = 0$ and $h_R^* = \left(1 - \dfrac{\lambda_L}{\lambda_R}\right) h_{HLL}$.
- Similarly, if $\widetilde{h_R^*} < 0$, then we have $h_R^* = 0$ and $h_L^* = \left(1 - \dfrac{\lambda_R}{\lambda_L}\right) h_{HLL}$.

In all three cases, the consistency relation (3.22a) holds. As a consequence, (3.22) is satisfied, and (i) holds.

The expressions (3.81b) and (3.81c) obviously yield that $h_L^* \geq 0$ and $h_R^* \geq 0$ Indeed, these intermediate heights are the minima of non-negative quantities, since $h_{HLL} > 0$ after Lemma 3.3. Therefore, (ii) is satisfied.

From Lemma 3.4, we know that the well-balance property (iii) is established as soon as the approximate source term $\overline{S}$ satisfies (3.26) when W_L and W_R define a steady state. After (3.48),

the approximate topography source term $\overline{S}^t$ satisfies this relation by construction. Similarly, the approximate friction source term $\overline{S}^f$ has been derived from (3.57), and thus the relation (3.26) holds. As a consequence, for the individual contributions of the topography and the friction, the property (iii) is verified. In addition, for both contributions, the steady relation we have elected to satisfy is (3.68). Therefore, (3.26) holds for $\overline{S} = \overline{S}^t + \overline{S}^f$. The proof of (iii) is thus achieved, which concludes the proof of Lemma 3.11. □

Remark 3.12. We note that using the definitions (3.81) and making the friction source term vanish allows the recovery of the intermediate states for topography only. Similarly, if the topography source term vanishes, we recover the intermediate states for friction only. As a consequence, (3.81) yields intermediate states that are well-balanced for the individual source terms of topography or friction. Let us recall that the steady states relation for the shallow-water system with both topography and friction source terms (3.67) cannot be written under the form of an algebraic relation for all Z. Therefore, we only manage to preserve the steady states up to the chosen discretization (3.68) of the steady relation (3.67) (for a similar approach, see [162] for the shallow-water equations with topography and [163, 120, 60, 101] for the Euler equations with gravity).

Lemma 3.11 allows us to state the following result, which is an extension of Theorem 3.5 to consider non-negative water heights.

Theorem 3.13. *Consider $W_i^n \in \Omega$ for all $i \in \mathbb{Z}$, where Ω is the admissible states space defined by (1.3). Assume that the intermediate states $W_{i+\frac{1}{2}}^{L,*}$ and $W_{i+\frac{1}{2}}^{R,*}$ are given, for all $i \in \mathbb{Z}$, by*

$$W_{i+\frac{1}{2}}^{L,*} = \begin{pmatrix} h_L^*(W_i^n, W_{i+1}^n) \\ q_L^*(W_i^n, W_{i+1}^n) \end{pmatrix} \text{ and } W_{i+\frac{1}{2}}^{R,*} = \begin{pmatrix} h_R^*(W_i^n, W_{i+1}^n) \\ q_R^*(W_i^n, W_{i+1}^n) \end{pmatrix}, \tag{3.82}$$

where q_L^* and q_R^* are defined by (3.81a), while h_L^* and h_R^* are respectively given by (3.81b) and (3.81c). Then, under the CFL restriction (3.1), the Godunov-type scheme (3.9) satisfies the following properties:

(i) consistency with the shallow-water system (1.1);

(ii) robustness: for all $i \in \mathbb{Z}$, $W_i^{n+1} \in \Omega$;

(iii) well-balance: if $(W_i^n)_{i\in\mathbb{Z}}$ defines a steady state, then for all $i \in \mathbb{Z}$, $W_i^{n+1} = W_i^n$.

Proof. After [90], if the approximate Riemann solver satisfies the integral consistency condition (3.13), then the consistency property (i) holds. This integral consistency property is satisfied: indeed, Lemma 3.11 ensures that the intermediate states (3.81) are consistent. Therefore, (i) holds.

By definition of Ω, proving the robustness property (ii) is equivalent to showing that for all $i \in \mathbb{Z}$, $h_i^{n+1} > 0$ as soon as $h_i^n > 0$. The second item of Lemma 3.4 ensures that $h_{i+\frac{1}{2}}^{L,*} \geq 0$ and $h_{i+\frac{1}{2}}^{R,*} \geq 0$ as soon as $h_i^n \geq 0$ and $h_{i+1}^n \geq 0$. As a consequence, after the expression (3.9) of the scheme, h_i^{n+1} is the sum of non-negative quantities, which proves (ii).

The well-balance property (iii) is then directly inferred from Lemma 3.11. Indeed, assume that $(W_i^n)_{i\in\mathbb{Z}}$ defines a steady state, i.e. that W_i^n and W_{i+1}^n define a steady state for all $i \in \mathbb{Z}$. Therefore, Lemma 3.11 yields that $W_{i+\frac{1}{2}}^{L,*} = W_i^n$ and $W_{i+\frac{1}{2}}^{R,*} = W_{i+1}^n$ for all $i \in \mathbb{Z}$.

Hence, $W_i^{n+1} = W_i^n$ for all $i \in \mathbb{Z}$, and the property (iii) holds, which concludes the proof of Theorem 3.13. □

3.2 Semi-implicitation of the scheme

The scheme (3.9) – (3.82) allows the simulation of wet/dry transitions. However, note that the friction source term becomes *stiff* (i.e. its value becomes arbitrarily large) in the vicinity of wet/dry transitions. As a consequence of this stiffness, spurious oscillations appear in the numerical approximation. In order to get rid of the oscillations, implicit schemes are usually needed to compute the numerical contribution of stiff source terms. The *splitting method* (see [26, 150] for instance) is used to avoid a fully implicit scheme, and rather suggests a semi-implicitation of the scheme, where only the stiff source terms are treated in an implicit way. The splitting method has been successfully applied to balance laws, especially in the presence of stiff source terms (see [113, 124] for instance).

To introduce such a semi-implicitation, we first rewrite the scheme (3.9) – (3.82) in order to exhibit the numerical flux function and the source terms contribution. Then, we adopt an explicit scheme for the flux and the topography, and an implicit scheme for the friction.

3.2.1 Rewriting the scheme

In this section, we exhibit the numerical flux function and the numerical source terms. The following result states a rewriting of the scheme (3.9) – (3.82) using these two functions (see for instance [90]).

Proposition 3.14. *The scheme (3.9) – (3.82) can be rewritten under the following form:*

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\Big(\mathcal{F}_{i+\frac{1}{2}}^n - \mathcal{F}_{i-\frac{1}{2}}^n\Big) + \frac{\Delta t}{2}\Big(\mathcal{S}_{i+\frac{1}{2}}^n + \mathcal{S}_{i-\frac{1}{2}}^n\Big), \tag{3.83}$$

where $\mathcal{F}_{i+\frac{1}{2}}^n = \mathcal{F}(W_i^n, W_{i+1}^n, Z_i^n, Z_{i+1}^n)$ is the numerical flux function evaluated at the interface $x_{i+\frac{1}{2}}$, and $\mathcal{S}_{i+\frac{1}{2}}^n = \mathcal{S}(W_i^n, W_{i+1}^n, Z_i^n, Z_{i+1}^n)$ is the numerical source term at the interface $x_{i+\frac{1}{2}}$. The numerical flux function is defined as follows:

$$\mathcal{F}_{i+\frac{1}{2}}^n = \frac{1}{2}\big(F(W_i^n) + F(W_{i+1}^n)\big) + \frac{\lambda_{i+\frac{1}{2}}^L}{2}\left(W_{i+\frac{1}{2}}^{L,*} - W_i^n\right) + \frac{\lambda_{i+\frac{1}{2}}^R}{2}\left(W_{i+\frac{1}{2}}^{R,*} - W_{i+1}^n\right), \tag{3.84}$$

while the numerical source term is given by:

$$\mathcal{S}_{i+\frac{1}{2}}^n = \begin{pmatrix} 0 \\ (S^t)_{i+\frac{1}{2}}^n + (S^f)_{i+\frac{1}{2}}^n \end{pmatrix}, \tag{3.85}$$

where the quantities $(S^t)_{i+\frac{1}{2}}^n$ and $(S^f)_{i+\frac{1}{2}}^n$ are approximations of the topography and the friction source terms, respectively. Adopting extended notations, they are given by:

$$(S^t)_{i+\frac{1}{2}}^n = \overline{S}^t\big(h_i^n, h_{i+1}^n, q_i^n, q_{i+1}^n, Z_i, Z_{i+1}, \Delta x\big), \tag{3.86a}$$

$$(S^f)_{i+\frac{1}{2}}^n = \overline{S}^f\big(h_i^n, h_{i+1}^n, q_i^n, q_{i+1}^n, \Delta x\big), \tag{3.86b}$$

where $\overline{S}^t$ and $\overline{S}^f$ are the approximate source terms already defined by (3.75) and (3.79).

Proof. After (3.84), the quantity $\mathcal{F}^n_{i+\frac{1}{2}} - \mathcal{F}^n_{i-\frac{1}{2}}$, present in (3.83), rewrites as follows:

$$\mathcal{F}^n_{i+\frac{1}{2}} - \mathcal{F}^n_{i-\frac{1}{2}} = \lambda^L_{i+\frac{1}{2}}\left(W^{L,*}_{i+\frac{1}{2}} - W^n_i\right) - \lambda^R_{i-\frac{1}{2}}\left(W^{R,*}_{i-\frac{1}{2}} - W^n_i\right) + \frac{1}{2}\left(\mathfrak{F}^n_{i+\frac{1}{2}} + \mathfrak{F}^n_{i-\frac{1}{2}}\right), \tag{3.87}$$

where $\mathfrak{F}^n_{i+\frac{1}{2}} = \mathfrak{F}(W^n_i, W^n_{i+1}, Z_i, Z_{i+1})$, with the function $\mathfrak{F}$ defined by:

$$\mathfrak{F}(W_L, W_R, Z_L, Z_R) = F(W_R) - F(W_L) + \lambda_R(W^*_R - W_R) - \lambda_L(W^*_L - W_L).$$

According to the above identity, the function $\mathfrak{F}$ satisfies the following sequence of equalities:

$$\begin{aligned}\mathfrak{F}(W_L, W_R, Z_L, Z_R) &= \lambda_R W^*_R - \lambda_L W^*_L - \left[\lambda_R W_R - \lambda_L W_L - \left(F(W_R) - F(W_L)\right)\right] \\ &= \lambda_R W^*_R - \lambda_L W^*_L - (\lambda_R - \lambda_L) W_{HLL},\end{aligned}$$

where W_{HLL} is the intermediate state of the HLL solver, defined by (3.18). Now, recall from Lemma 3.11 that the intermediate states (3.81) satisfy the integral consistency property, which is equivalent to the equations (3.22). Since $\overline{S} = \overline{S}^t + \overline{S}^f$ in the present context, arguing (3.22) yields:

$$\mathfrak{F}(W_L, W_R, Z_L, Z_R) = \begin{pmatrix} 0 \\ \overline{S}^t(h_L, h_R, q_L, q_R, Z_L, Z_R, \Delta x)\Delta x + \overline{S}^f(h_L, h_R, q_L, q_R, \Delta x)\Delta x \end{pmatrix}.$$

As a consequence, recalling the definition (3.85) – (3.86) of the numerical source term, we get:

$$-\frac{\Delta t}{\Delta x}\left(\mathcal{F}^n_{i+\frac{1}{2}} - \mathcal{F}^n_{i-\frac{1}{2}}\right) + \frac{\Delta t}{2}\left(\mathcal{S}^n_{i+\frac{1}{2}} + \mathcal{S}^n_{i-\frac{1}{2}}\right) = -\frac{\Delta t}{\Delta x}\left[\lambda^L_{i+\frac{1}{2}}\left(W^{L,*}_{i+\frac{1}{2}} - W^n_i\right) - \lambda^R_{i-\frac{1}{2}}\left(W^{R,*}_{i-\frac{1}{2}} - W^n_i\right)\right].$$

Arguing the definition (3.9) of the suggested numerical scheme, we get:

$$W^n_i - \frac{\Delta t}{\Delta x}\left(\mathcal{F}^n_{i+\frac{1}{2}} - \mathcal{F}^n_{i-\frac{1}{2}}\right) + \frac{\Delta t}{2}\left(\mathcal{S}^n_{i+\frac{1}{2}} + \mathcal{S}^n_{i-\frac{1}{2}}\right) = W^{n+1}_i,$$

which is nothing but the rewritten scheme (3.83). The proof is thus completed. □

3.2.2 Application to the topography and friction source terms

We now introduce a semi-implicit version of the Godunov-type scheme (3.83). The main idea of this section is to use a splitting method to reduce the impact of the aforementioned instabilities. The splitting strategy we use here is to first consider an explicit treatment of the flux and the topography source term, then an implicit treatment of the friction source term.

As a consequence, the first step, devoted to approximating solutions of the partial differential equation $\partial_t W + \partial_x F(W) = S^t(W)$ containing the flux and the topography source term, reads as follows:

$$\begin{pmatrix} h^{n+\frac{1}{2}}_i \\ q^{n+\frac{1}{2}}_i \end{pmatrix} = \begin{pmatrix} h^n_i \\ q^n_i \end{pmatrix} - \frac{\Delta t}{\Delta x}\left(\mathcal{F}^n_{i+\frac{1}{2}} - \mathcal{F}^n_{i-\frac{1}{2}}\right) + \frac{\Delta t}{2}\begin{pmatrix} 0 \\ (S^t)^n_{i+\frac{1}{2}} + (S^t)^n_{i-\frac{1}{2}} \end{pmatrix}. \tag{3.88}$$

The second and last step concerns the friction and consists in solving the following system of ordinary differential equations:

$$\begin{cases} \dfrac{dh}{dt} = 0, \\ \dfrac{dq}{dt} = -kq|q|h^{-\eta}, \end{cases} \quad \text{with initial data} \quad \begin{cases} h(0) = h_i^{n+\frac{1}{2}}, \\ q(0) = q_i^{n+\frac{1}{2}}. \end{cases}$$

This system can be solved to obtain an analytic expression of the solution. For $t \in [0, \Delta t]$, the exact solution of the above system reads as follows:

$$\begin{cases} h(t) = h(0), \\ q(t) = \dfrac{h(0)^\eta q(0)}{h(0)^\eta + k\, t\, |q(0)|}. \end{cases} \tag{3.89}$$

Note that the analytic expression (3.89) guarantees that, for all $t \in [0, \Delta t]$, the sign of $q(t)$ stays the same as the sign of $q(0)$, and that $|q(t)| < |q(0)|$. This was to be expected, since q is governed by the damping equation (3.2.2). This behavior is consistent with the fact that friction should only slow down the movement of the fluid, rather than changing its direction.

Then, evaluating (3.89) at $t = \Delta t$ and plugging the initial data yields the following updated state $W_i^{n+1} = {}^t(h_i^{n+1}, q_i^{n+1})$:

$$\begin{cases} h_i^{n+1} = h_i^{n+\frac{1}{2}}, & \text{(3.90a)} \\ q_i^{n+1} = \dfrac{\left(h_i^{n+1}\right)^\eta q_i^{n+\frac{1}{2}}}{\left(h_i^{n+1}\right)^\eta + k\,\Delta t\left|q_i^{n+\frac{1}{2}}\right|}. & \text{(3.90b)} \end{cases}$$

Let us note that the well-balance property is lost for the discharge. Indeed, if W_{i-1}^n, W_i^n and W_{i+1}^n define a steady state, we do not necessarily recover $q_i^{n+1} = q_i^n$, but we have $h_i^{n+1} = h_i^n$ since the semi-implicitation procedure does not change the evaluation of the water height. As a consequence, we decide to consider an approximation $(\overline{h^\eta})_i^{n+1}$ of $\left(h_i^{n+1}\right)^\eta$ in the discharge update equation (3.90b), thus replacing the update step (3.90) with the following expressions:

$$\begin{cases} h_i^{n+1} = h_i^{n+\frac{1}{2}}, & \text{(3.91a)} \\ q_i^{n+1} = \dfrac{(\overline{h^\eta})_i^{n+1} q_i^{n+\frac{1}{2}}}{(\overline{h^\eta})_i^{n+1} + k\,\Delta t\left|q_i^{n+\frac{1}{2}}\right|}. & \text{(3.91b)} \end{cases}$$

The approximation $(\overline{h^\eta})_i^{n+1}$ is now determined in order to ensure that the scheme satisfies the required well-balance property. In order to obtain such an expression of $(\overline{h^\eta})_i^{n+1}$, we momentarily suppose that W_{i-1}^n, W_i^n and W_{i+1}^n define a steady state. In this case, we need to ensure that $q_i^{n+1} = q_i^n$. Therefore, (3.91b) reads:

$$(\overline{h^\eta})_i^{n+1} q_i^{n+\frac{1}{2}} = (\overline{h^\eta})_i^{n+1} q_i^n + k\,\Delta t\left|q_i^{n+\frac{1}{2}}\right| q_i^n. \tag{3.92}$$

Now, note that the explicit scheme (3.83) can be rewritten as follows:

$$h_i^{n+1} = h_i^n - \frac{\Delta t}{\Delta x}\Big((\mathcal{F}^h)^n_{i+\frac{1}{2}} - (\mathcal{F}^h)^n_{i-\frac{1}{2}}\Big), \tag{3.93a}$$

$$q_i^{n+1} = q_i^n - \frac{\Delta t}{\Delta x}\Big((\mathcal{F}^q)^n_{i+\frac{1}{2}} - (\mathcal{F}^q)^n_{i-\frac{1}{2}}\Big) + \Delta t (S^t)^n_i + \Delta t (S^f)^n_i, \tag{3.93b}$$

where we have set:

$$\mathcal{F} = \begin{pmatrix} \mathcal{F}^h \\ \mathcal{F}^q \end{pmatrix} \quad ; \quad (S^t)^n_i = \frac{1}{2}\Big((S^t)^n_{i+\frac{1}{2}} + (S^t)^n_{i-\frac{1}{2}}\Big) \quad ; \quad (S^f)^n_i = \frac{1}{2}\Big((S^f)^n_{i+\frac{1}{2}} + (S^f)^n_{i-\frac{1}{2}}\Big). \tag{3.94}$$

Since the scheme (3.83) is well-balanced and we are considering steady states, the equation (3.93b) yields $q_i^{n+1} = q_i^n$, and it can be rewritten as:

$$(S^f)^n_i = \frac{1}{\Delta x}\Big((\mathcal{F}^q)^n_{i+\frac{1}{2}} - (\mathcal{F}^q)^n_{i-\frac{1}{2}}\Big) - (S^t)^n_i. \tag{3.95}$$

Since the evaluation of $q_i^{n+\frac{1}{2}}$ is obtained from (3.88), we get:

$$q_i^{n+\frac{1}{2}} = q_i^n - \frac{\Delta t}{\Delta x}\Big((\mathcal{F}^q)^n_{i+\frac{1}{2}} - (\mathcal{F}^q)^n_{i-\frac{1}{2}}\Big) + \Delta t (S^t)^n_i. \tag{3.96}$$

From (3.95) and (3.96), we immediately obtain

$$q_i^{n+\frac{1}{2}} = q_i^n - \Delta t (S^f)^n_i. \tag{3.97}$$

Thus, we are now able to determine the expression of $(\overline{h^\eta})^{n+1}_i$ that ensures the well-balance of the scheme. With $\mu_i^{n+\frac{1}{2}} = \operatorname{sgn} q_i^{n+\frac{1}{2}}$, plugging the expression (3.97) of $q_i^{n+\frac{1}{2}}$ into (3.92) yields:

$$(\overline{h^\eta})^{n+1}_i q_i^n - (\overline{h^\eta})^{n+1}_i \Delta t (S^f)^n_i = (\overline{h^\eta})^{n+1}_i q_i^n + k\,\Delta t\,\mu_i^{n+\frac{1}{2}}(q_i^n)^2 - k\,\Delta t^2\,\mu_i^{n+\frac{1}{2}} q_i^n (S^f)^n_i.$$

Hence, $(\overline{h^\eta})^{n+1}_i$ is immediately proven to satisfy:

$$(\overline{h^\eta})^{n+1}_i = \frac{-k(q_i^n)^2 \mu_i^{n+\frac{1}{2}}}{(S^f)^n_i} + k\,\Delta t\,\mu_i^{n+\frac{1}{2}} q_i^n. \tag{3.98}$$

Now, recall that the numerical source term $(\overline{S}^f)^n_i$ is defined by (3.94), and we get:

$$(S^f)^n_i = \frac{1}{2}\Big(-k\bar{q}^n_{i-\frac{1}{2}}|\bar{q}^n_{i-\frac{1}{2}}|(\overline{h^{-\eta}})^n_{i-\frac{1}{2}} - k\bar{q}^n_{i+\frac{1}{2}}|\bar{q}^n_{i+\frac{1}{2}}|(\overline{h^{-\eta}})^n_{i+\frac{1}{2}}\Big),$$

where the averages $\bar{q}^n_{i\pm\frac{1}{2}}$ and $(\overline{h^{-\eta}})^n_{i\pm\frac{1}{2}}$ are given with clear notations by (3.61) and (3.62), respectively. However, recall that the only requirement to choose the average $\bar{q}$ was that it be equal to q_0 as soon as a steady state was reached. In the current context, W^n_{i-1}, W^n_i and W^n_{i+1} define a steady state; hence, we have $q^n_{i-1} = q^n_i = q^n_{i+1} = q_0$. A relevant choice is therefore to take $\bar{q}^n_{i\pm\frac{1}{2}} = q^n_i$, which yields the following formula:

$$(S^f)^n_i = \frac{1}{2}\left(-kq_i^n|q_i^n|\left(\overline{h^{-\eta}}\right)^{n+1}_{i-\frac{1}{2}} - kq_i^n|q_i^n|\left(\overline{h^{-\eta}}\right)^{n+1}_{i+\frac{1}{2}}\right). \tag{3.99}$$

In (3.99), we have substituted $(\overline{h^{-\eta}})^n_{i\pm\frac{1}{2}}$ with $(\overline{h^{-\eta}})^{n+1}_{i\pm\frac{1}{2}}$. This substitution has no effect on the well-balance property, and it makes the scheme more implicit by considering the updated water height.

With this simplification in the source term approximation, we get the following expression for $(\overline{h^{\eta}})^{n+1}_i$, from (3.98) and (3.99):

$$(\overline{h^{\eta}})^{n+1}_i = \frac{2{\mu_i}^{n+\frac{1}{2}}\mu_i^n}{\left(\overline{h^{-\eta}}\right)^{n+1}_{i-\frac{1}{2}} + \left(\overline{h^{-\eta}}\right)^{n+1}_{i+\frac{1}{2}}} + k\,\Delta t\,{\mu_i}^{n+\frac{1}{2}} q_i^n. \tag{3.100}$$

Arguing the expressions of $(\overline{h^{-\eta}})^{n+1}_{i-\frac{1}{2}}$ and $(\overline{h^{-\eta}})^{n+1}_{i+\frac{1}{2}}$, the above equation can then be rewritten as

$$(\overline{h^{\eta}})^{n+1}_i = \frac{2k{\mu_i}^{n+\frac{1}{2}}\Delta x}{k\mu_i^n\Delta x\left(\beta^{n+1}_{i-\frac{1}{2}} + \beta^{n+1}_{i+\frac{1}{2}}\right) - \left(\gamma^{n+1}_{i-\frac{1}{2}} + \gamma^{n+1}_{i+\frac{1}{2}}\right)} + k\,\Delta t\,{\mu_i}^{n+\frac{1}{2}} q_i^n, \tag{3.101}$$

where we have set

$$\begin{aligned}
\beta^{n+1}_{i+\frac{1}{2}} &= \frac{\eta+2}{2}\frac{\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2}{\left(h^{n+1}_{i+1}\right)^{\eta+2} - \left(h^{n+1}_i\right)^{\eta+2}}\,, \text{ and} \\
\gamma^{n+1}_{i+\frac{1}{2}} &= \frac{1}{h^{n+1}_{i+1}} - \frac{1}{h^{n+1}_i} + \beta^{n+1}_{i+\frac{1}{2}}\frac{\left(h^{n+1}_{i+1}\right)^{\eta-1} - \left(h^{n+1}_i\right)^{\eta-1}}{\eta-1}.
\end{aligned} \tag{3.102}$$

Computations within the expression of $(\overline{h^{\eta}})^{n+1}_i$ show that it tends to 0 as soon as h^{n+1}_{i-1}, h^{n+1}_i or h^{n+1}_{i+1} tends to 0, which is a good behavior when dealing with wet/dry transitions. We have therefore devised a way to consider the friction contribution in an implicit way, while still retaining the well-balance property of the scheme. We can thus state the following result.

Theorem 3.15. *Assume that for all $i \in \mathbb{Z}$, $W_i^n \in \Omega$, with Ω the admissible states space defined by (1.3). The semi-implicit scheme (3.88) – (3.91) – (3.101) satisfies the following properties:*

(i) consistency with the shallow-water system (1.1);

(ii) robustness: for all $i \in \mathbb{Z}$, $W_i^{n+1} \in \Omega$;

(iii) well-balance: if $(W_i^n)_{i\in\mathbb{Z}}$ defines a steady state, then for all $i \in \mathbb{Z}$, $W_i^{n+1} = W_i^n$. Here, $(W_i^n)_{i\in\mathbb{Z}}$ is said to define a steady state if any of the three following cases arise:

- *topography steady state: for all $i \in \mathbb{Z}$, W_i^n and W_{i+1}^n satisfy (3.26), with $\overline{S} = \overline{S}^t$;*
- *friction steady state: for all $i \in \mathbb{Z}$, W_i^n and W_{i+1}^n satisfy (3.26), with $\overline{S} = \overline{S}^f$;*
- *topography and friction steady state: for all $i \in \mathbb{Z}$, W_i^n and W_{i+1}^n satisfy (3.26), with $\overline{S} = \overline{S}^t + \overline{S}^f$.*

Proof. We begin by proving the consistency property (i). Arguing Lemma 3.10, we get that $(\overline{h^{-\eta}})^{n+1}_{i-\frac{1}{2}}$ and $(\overline{h^{-\eta}})^{n+1}_{i+\frac{1}{2}}$ are consistent approximations of $h^{-\eta}$. Therefore, we clearly observe from (3.100) that $(\overline{h^{\eta}})^{n+1}_i$ is a consistent approximation of h^{η}. As a consequence, (3.91b) yields that q_i^{n+1} is indeed consistent with q. Finally, arguing Theorem 3.13 ensures that the expressions ${h_i}^{n+\frac{1}{2}}$ and ${q_i}^{n+\frac{1}{2}}$, given by (3.88), are respectively consistent with h and q. Therefore, the consistency property (i) holds.

Concerning the robustness, note that $h_i^{n+1} = h_i^{n+\frac{1}{2}}$ from (3.91a), with $h_i^{n+\frac{1}{2}}$ defined by (3.88). After Theorem 3.13, the scheme (3.88) is robust: therefore, the robustness property (ii) holds. We make the additional remark that, with $(\overline{h^\eta})_i^{n+1}$ given by (3.101), the expression (3.91b) of q_i^{n+1} ensures that this updated discharge vanishes as soon as a dry area is considered, which is a good behavior when considering transitions between wet and dry areas.

Now, to prove the well-balance, assume that W_{i-1}^n, W_i^n and W_{i+1}^n define a steady state, according to (3.26) with $\overline{S} = \overline{S}^t$, $\overline{S} = \overline{S}^t$ or $\overline{S} = \overline{S}^t + \overline{S}^f$. From Theorem 3.13, the scheme (3.88) is well-balanced. Therefore, since $h_i^{n+1} = h_i^{n+\frac{1}{2}}$, we immediately recover that $h_i^{n+1} = h_i^n$. To complete the proof, we now have to show that $q_i^{n+1} = q_i^n$. The updated discharge q_i^{n+1} is given by (3.90), with $q_i^{n+\frac{1}{2}}$ defined by (3.88). Since $(\overline{h^\eta})_i^{n+1}$ is given by (3.101) and has been chosen to ensure that $q_i^{n+1} = q_i^n$ as soon as a steady state is reached, the proof of the well-balance property (iii) is concluded. Hence, the proof of Theorem 3.15 is achieved. □

3.3 Numerical experiments

Numerical simulations are carried out to test the scheme derived in the previous sections. We start by recalling the two schemes we shall test:

- the *explicit* scheme (3.9) – (3.81);
- the *implicit* scheme (3.88) – (3.91) – (3.101).

In order to determine the properties of these schemes, we present two sets of numerical experiments.

The first set assesses the well-balance of the scheme, by considering steady states at rest and moving steady states with topography and/or friction. Namely, the steady state solutions exhibited in Section 1.2 are simulated. In addition, we consider the steady solutions reviewed by Goutal and Maurel in [86].

The second set consists in a numerical validation of the proposed explicit and implicit schemes. First, two experiments from [77] are presented, namely the drain on a non-flat bottom and a vacuum occurrence by double rarefaction. Then, the simulations of several dam-break situations are carried out. A wet dam-break and two dry dam-breaks are presented.

We also compare the proposed schemes with two classical schemes: the *HLL scheme* (2.35) (see [90]) and the *hydrostatic reconstruction* (*HR*) *scheme* (see [5]) applied to the HLL flux (2.35). Indeed, the HLL scheme is not well-balanced, while the HR scheme preserves the steady states at rest with a non-flat topography, but not the moving steady states.

Since the HLL scheme is designed for conservative systems, we take the topography and friction contributions into account by using a splitting method. The purpose of carrying out simulations with the HLL scheme is to highlight that the well-balance is an important property for a scheme to possess. In addition, the choice of the HLL scheme for comparisons is relevant since the construction of our scheme is based on a HLL-like construction.

Moreover, the friction is also introduced into the HR scheme through a splitting method. Since the expression used for the updated discharge in the splitting method is similar to (3.89), the friction contribution will be zero as soon as a solution at rest ($q = 0$) is considered. As a

consequence, even in the presence of the friction source term, the HR scheme still preserves the steady states at rest, for all k.

In order to assess the numerical accuracy of all the schemes, we compare the approximate solution with the exact solution. To that end, we compare the error estimates in L^1, L^2 and L^∞ norms and defined by (2.36).

Finally, we recall that the CFL condition (3.1) gives the time step Δt for each iteration, as follows:

$$\Delta t \leq \frac{\Delta x}{2\Lambda}, \quad \text{where } \Lambda = \max_{i\in\mathbb{Z}}\Big(-\lambda^L_{i+\frac{1}{2}}, \lambda^R_{i+\frac{1}{2}}\Big).$$

For the numerical experiments involving friction, a suitable value of the Manning coefficient k has to be chosen. For instance, the reader is referred to [45], where multiple values of k are given for different types of channel beds. Here, instead of following [45], we deliberately impose stronger Manning coefficients than in reality (up to 10 times). This choice is made to ensure that the friction source term is preponderant compared to the topography source term, in order to study the effects of the friction. The other constants are chosen as follows:

- in (1.1), $g = 9.81$ m.s^{-2};
- in (3.6), $\varepsilon_\lambda = 10^{-10}$ m.s^{-1}.

An important step in these numerical experiments in the choice of the parameter C, introduced in (3.55) to ensure the consistency of the approximate topography source term. In this manuscript, this parameter is chosen heuristically, and we give its value for each experiment. A better study of the stability of the scheme could provide several bounds for this parameter.

3.3.1 Well-balance assessment

In this first set of experiments, we assess the well-balance of the scheme, i.e. its ability to exactly preserve and capture steady state solutions. Recall that steady states are given by the equation (1.73), which prescribes a uniform discharge over the space domain, denoted by q_0.

First, we consider steady states at rest, i.e. $q_0 = 0$. Several different topography functions, continuous and discontinuous, are studied. In addition, the simulations of steady state solutions with dry areas are carried out. We also perform the simulation of a flow at rest with emerging bottom, proposed in [77].

Then, we consider moving steady states with a vanishing friction contribution, i.e. $k = 0$, and a non-flat topography. Such steady state solutions have been exhibited in Section 1.2.1, and examples of subcritical and supercritical steady states have been provided. The simulations of both kinds of steady states are therefore carried out.

Afterwards, steady state solutions for the friction source term only are studied, that is to say we impose $q_0 \neq 0$, $\partial_x Z = 0$ and $k \neq 0$. In Section 1.2.2, we have studied such steady states, and exhibited two specific examples, a subcritical solution and a supercritical solution. We perform the simulations of both these examples.

Subsequently, we consider steady states for both friction and topography, which are either analytic solutions in specific cases or steady states obtained by approximately solving (1.73). Section 1.2.3 provides two analytic steady state solutions, which we use to test the well-balance of the scheme. Afterwards, the equation (1.73) is approximately solved to exhibit a steady state solution, the simulation of which is carried out.

Finally, we perform the simulations of three well-known moving steady state solutions with a vanishing friction contribution and a non-flat topography, presented in [86]. Namely, the simulations of the *subcritical flow*, *transcritical flow* and *transcritical flow with shock* are carried out.

3.3.1.1 Steady states at rest

In this section, we consider steady state solutions at rest. We recall that these solutions all satisfy $q(t,x) = 0$ for all t and all x, i.e. $q_0 = 0$. After (1.43), the smooth steady state solutions are governed by $\partial_x(h + Z) = 0$, which corresponds to a lake at rest configuration. This condition can be extended to non-smooth (and even discontinuous) steady state solutions, to get $h + Z = \text{cst}$. As a consequence, at the discrete level, the approximate solution $(W_i^n)_{i\in\mathbb{Z}}$ is said to define a steady state at rest if the following relations hold:

$$\forall i \in \mathbb{Z}, \begin{cases} q_i^n = 0, \\ h_i^n + Z_i^n = h_{i+1}^n + Z_{i+1}^n. \end{cases}$$

We also recall that several physical steady state solutions at rest, given by (3.72), are not governed by $h + Z = \text{cst}$ (see Figure 3.4). We finally recall that, according to Proposition 1.14, a smooth steady state with a dry/wet transition is necessarily at rest. The goal of this section is therefore to perform simulations of all these cases.

Continuous topography

We begin with continuous steady states at rest, to assess the well-balance of the explicit and implicit schemes. In the two cases we consider, we have $q_0 = 0$ and $k = 10$. The two experiments are performed with 200 discretization cells, over the domain $[0,1]$ and until a final time $t_{end} = 1$s. The initial conditions are $q(0,x) = 0$ and $h(0,x) = (2 - Z_i(x))_+$, with topographies $(Z_i)_{i\in\{1,2\}}$ given by:

$$Z_1(x) = (1 - |4x - 2|)_+;$$
$$Z_2(x) = (4x - 1)_+.$$

These initial free surfaces are depicted on Figure 3.5. Note that the experiment where Z_2 is used involves a dry/wet transition. The exact solution at rest is imposed at the boundaries, in order to ensure that the boundary conditions do not interfere with the well-balance assessment. We used $C = +\infty$ in (3.55). Numerically, we set C as the upper bound of the double precision floating point numbers. The results of the simulations are presented in Table 3.1 and Table 3.2.

We observe on Table 3.1 and Table 3.2 that the HR, explicit and implicit schemes indeed preserve such lake at rest configurations, even in the case of a transition between a wet area and a dry area. In particular, we note, as expected, that the presence of the friction contribution in the HR scheme does not alter the preservation of this steady state at rest. We also remark that, for such steady states at rest, the implicit scheme degenerates into the explicit

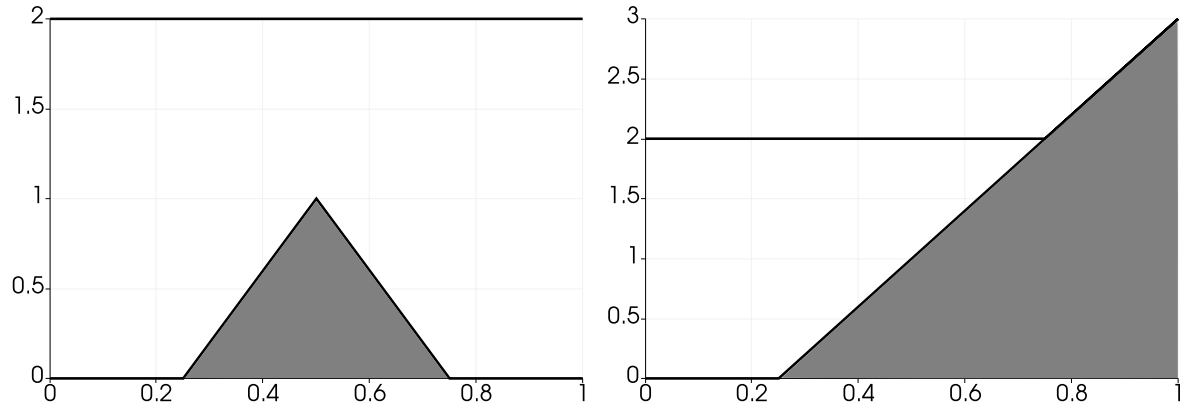

Figure 3.5 – From left to right: free surfaces for the lake at rest experiments with topographies given by Z_1 and Z_2.

	$h + Z_1$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.21e-04	8.78e-04	1.02e-02	1.90e-02	2.70e-02	4.41e-02
HR	0	0	0	3.37e-16	4.18e-16	1.58e-15
explicit	1.18e-15	1.30e-15	2.66e-15	1.76e-14	1.80e-14	2.36e-14
implicit	1.18e-15	1.30e-15	2.66e-15	1.76e-14	1.80e-14	2.36e-14

Table 3.1 – Free surface and discharge errors for the steady state at rest experiment with topography given by Z_1.

	$h + Z_2$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.43e-02	5.50e-02	3.20e-01	1.48e-02	2.21e-02	4.41e-02
HR	1.44e-17	9.55e-17	1.11e-15	2.72e-16	3.21e-16	1.58e-15
explicit	1.75e-16	3.05e-16	8.88e-16	8.27e-16	1.10e-15	3.65e-15
implicit	1.75e-16	3.05e-16	8.88e-16	8.27e-16	1.10e-15	3.65e-15

Table 3.2 – Free surface and discharge errors for the steady state at rest experiment with topography given by Z_2.

scheme. Thus, both schemes give the same results. However, concerning the HLL scheme, it provides an approximation of the steady state.

Discontinuous topography

We now turn to two experiments involving a discontinuous topography, and therefore a discontinuous water height, since the relation $h + Z = \mathrm{cst}$ holds. For these two experiments, we take $q_0 = 0$ and $k = 10$. The simulations are carried out over the domain $[0, 1]$, discretized with 200 cells. The final physical time is $t_{end} = 1$s. As initial conditions, we take the following steady state at rest: $q(0, x) = 0$ and $h(0, x) = (2 - Z_i(x))_+$, where the topography functions

$(Z_i)_{i\in\{3,4\}}$ are defined as follows:

$$Z_3(x) = \mathbb{1}_{[\frac{1}{2},1]}(x);$$
$$Z_4(x) = (4x-1)\mathbb{1}_{[\frac{1}{2},1]}(x).$$

These exact topography function and free surfaces are displayed on Figure 3.6. A transition between a wet area and a dry area is present when the topography is given by Z_4. At the boundaries, we choose to impose the exact solution. We also take $C = +\infty$ for these experiments. In Table 3.3 and Table 3.4, we present the results of the simulations.

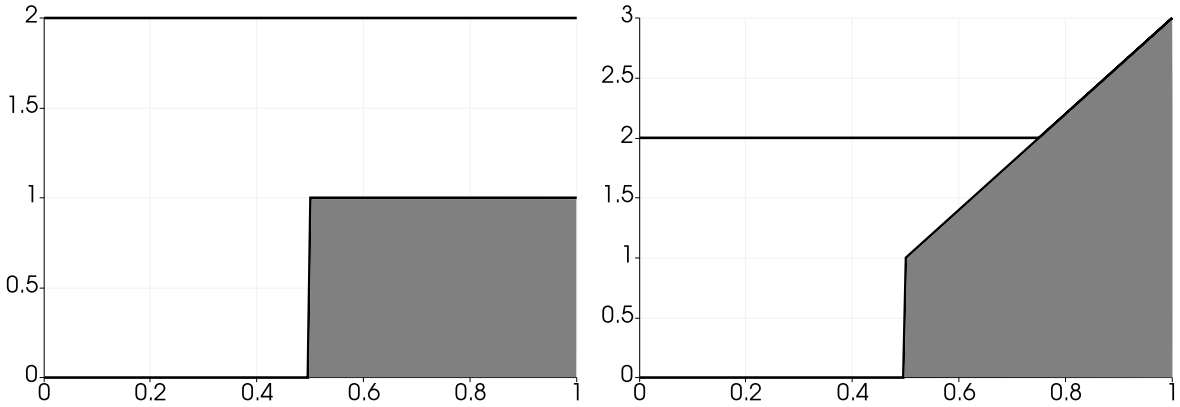

Figure 3.6 – From left to right: free surfaces for the lake at rest experiments with topographies given by Z_3 and Z_4.

	$h + Z_3$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	6.25e-03	2.63e-02	2.65e-01	2.18e-02	9.81e-02	1.04e+00
HR	0	0	0	0	0	0
explicit	0	0	0	0	0	0
implicit	0	0	0	0	0	0

Table 3.3 – Free surface and discharge errors for the steady state at rest experiment with topography given by Z_3.

	$h + Z_4$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.89e-02	6.09e-02	3.20e-01	1.57e-02	1.00e-01	1.05e+00
HR	1.44e-17	9.55e-17	1.11e-15	2.27e-16	2.28e-16	4.86e-16
explicit	5.55e-18	5.21e-17	6.66e-16	2.26e-16	2.28e-16	5.34e-16
implicit	5.55e-18	5.21e-17	6.66e-16	2.26e-16	2.28e-16	5.34e-16

Table 3.4 – Free surface and discharge errors for the steady state at rest experiment with topography given by Z_4.

Table 3.3 and Table 3.4 present the results of the four schemes at time $t_{end} = 1$s. The HR, explicit and implicit schemes exactly preserve the lake at rest steady state, even for such

discontinuous topography functions. On the contrary, the HLL scheme only provides an approximation of this lake at rest steady state.

Emerged discontinuous topography

This third set of experiments focuses on steady state solutions at rest governed by the equations (3.72) instead of $h + Z = \text{cst}$. As a consequence, we consider the following two topography functions on the space domain $[0, 1]$:

$$Z_5(x) = 3\mathbb{1}_{[\frac{1}{2},1]}(x);$$
$$Z_6(x) = 3\mathbb{1}_{[0,\frac{1}{2}]}(x).$$

Then, the initial condition for the experiment is given by $q(0, x) = 0$ and $h(0, x) = (2 - Z(x))_+$, as displayed on Figure 3.7. Since this experiment consists in a solution at rest, the height and discharge stay constant over time. Note that the topography source term discretization (3.75) has been derived in order to able to preserve such steady states.

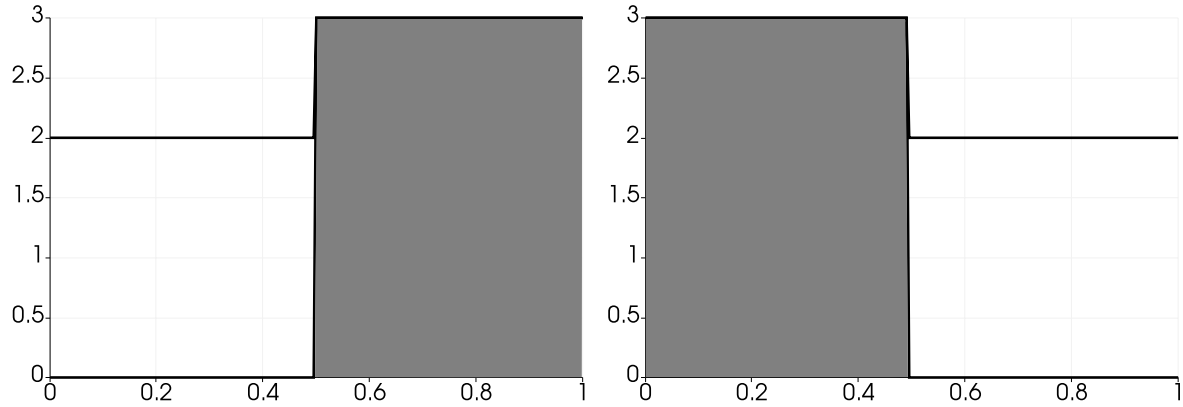

Figure 3.7 – From left to right: free surfaces for the lake at rest experiments with topographies given by Z_5 and Z_6.

For the numerical experiments, we take $k = 10$ and we discretize the domain $[0, 1]$ with 200 cells. The simulations are carried out until the final physical time $t_{end} = 1$s, and we set $C = +\infty$. The exact solution is prescribed as both initial and boundary conditions. The results are displayed in Table 3.5 and Table 3.6. Thanks to these tables, we observe that the HR scheme exactly preserves this steady state solution, which does not satisfy $h + Z = \text{cst}$. In addition, the specific cases introduced in the expression (3.75) of $\overline{S}^t$ allow the explicit and implicit schemes to preserve this steady state solution at rest. We also note that, as expected, the HLL scheme does not preserve this lake at rest.

Flow at rest with emerging bottom

This last experiment at rest involves an emerging bottom (see [77]). The space domain is $[0, 25]$, and the topography is given by $Z_7(x) = \left(0.2 - 0.05(x - 10)^2\right)_+$. We take $h(0, x) = (0.15 - Z_7(x))_+$ and $q(0, x) = 0$ as initial data. We present a graph of the free surface and the topography in Figure 3.8.

	$h + Z_5$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	6.28e-03	5.95e-02	8.06e-01	1.85e-02	1.89e-01	2.54e+00
HR	1.11e-16	1.57e-16	2.22e-16	2.22e-16	2.22e-16	2.22e-16
explicit	2.20e-16	2.42e-16	4.44e-16	7.33e-16	1.11e-15	3.61e-15
implicit	2.20e-16	2.42e-16	4.44e-16	7.33e-16	1.11e-15	3.61e-15

Table 3.5 – Free surface and discharge errors for the steady state at rest experiment with topography given by Z_5.

	$h + Z_6$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	6.28e-03	5.95e-02	8.06e-01	1.85e-02	1.89e-01	2.54e+00
HR	1.10e-16	1.56e-16	2.22e-16	2.22e-16	2.22e-16	2.22e-16
explicit	1.10e-16	1.56e-16	2.22e-16	1.22e-16	1.99e-16	1.55e-15
implicit	1.10e-16	1.56e-16	2.22e-16	1.22e-16	1.99e-16	1.55e-15

Table 3.6 – Free surface and discharge errors for the steady state at rest experiment with topography given by Z_6.

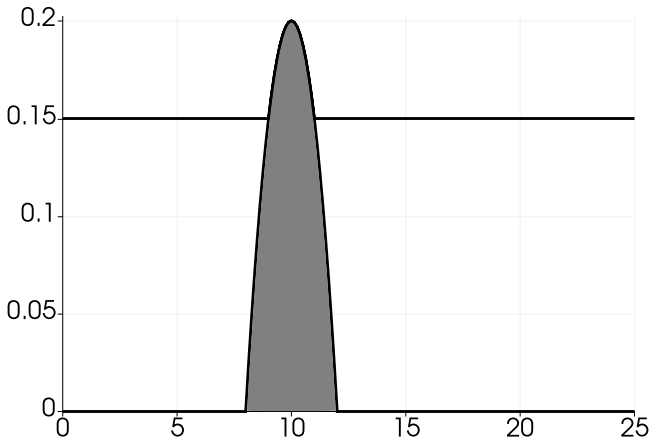

Figure 3.8 – Free surface and topography for the flow at rest with emerging bottom. The gray area represents the topography given by Z_7.

For this experiment, we set $C = +\infty$ and we use homogeneous Neumann boundary conditions. The simulation is carried out until the physical time $t_{end} = 100$s, using 200 discretization cells. In addition, we take a Manning coefficient $k = 10$. Such a nonzero friction is not present in the original experiment introduced in [77]. However, it does not change the steady state, since the friction contribution vanishes as soon as the discharge vanishes. The results of the four schemes are displayed in Table 3.7. This last experiment confirms once again the relevance of using a well-balanced scheme for the simulation of steady states at rest. Indeed, after Table 3.7, the HLL scheme only provides a first-order approximation of the steady state, while the HR, explicit and implicit schemes provide an exact preservation of this lake at rest.

	$h + Z_7$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.54e-02	2.89e-02	9.98e-02	5.12e-04	1.50e-03	7.69e-03
HR	2.83e-17	1.41e-16	1.42e-15	2.25e-16	2.26e-16	5.04e-16
explicit	4.64e-17	1.48e-16	1.08e-15	2.22e-16	2.22e-16	2.29e-16
implicit	4.64e-17	1.48e-16	1.08e-15	2.22e-16	2.22e-16	2.29e-16

Table 3.7 – Free surface and discharge errors for the flow at rest with emerging bottom.

3.3.1.2 Moving steady states for the topography source term

After steady states at rest, we now focus on the preservation of moving steady state solutions, with $q_0 \neq 0$. More precisely, we start with smooth moving steady states for the topography source term only, i.e. we take $k = 0$. Such solutions have been studied in Section 1.2.1; they are governed by (1.44). We here remark that, since the $k = 0$ and only the topography is considered, the implicit scheme degenerates into the explicit scheme.

In this section, we study the numerical preservation of the subcritical and supercritical steady solutions, exhibited as examples in Section 1.2.1. The topography function is defined as follows for this whole section:

$$Z(x) = \frac{1}{4} + \cos^2\Big(\pi(x - x_0) + \frac{\pi}{4}\Big).$$

For the numerical experiments, we consider the space domain $[0.75, 1.25]$ with 200 discretization cells. We consider an approximate solution of (1.45) on this space domain, obtained by using Newton's method with $x_0 = 0.75 - \Delta x$, $q_0 = \sqrt{g}$, and $h(x_0) = h_c = 1$. First, we focus on the subcritical solution, and then on the supercritical solution.

Subcritical topography steady state

We first consider the subcritical topography steady state. The water height for this steady state satisfies $h(t, x) = h(x) > h_c$. We take, as initial conditions, the subcritical solution $h_{sub}(x)$ obtained by using Newton's method to get an approximate solution of (1.45). As a consequence, we take $q(0, x) = q_0$ and $h(0, x) = h_{sub}(x)$. The boundary conditions are inhomogeneous Dirichlet boundary conditions, taken as the exact solution at points $0.75 - \Delta x$ and $1.25 + \Delta x$.

The first experiment consists in the preservation of the subcritical steady state. The simulations are carried out until the physical time $t_{end} = 1$, and $C = +\infty$, and their results are presented on Figure 3.9 and Table 3.8. On Figure 3.9, we note that the explicit and implicit schemes indeed preserve the subcritical topography steady state up to the machine precision. This observation is confirmed by Table 3.8, which highlights the fact that both the explicit and the implicit schemes exactly preserve this steady state, while the HLL and HR schemes only provide an approximation.

Now, we introduce a preservation of this subcritical steady state solution. We take the

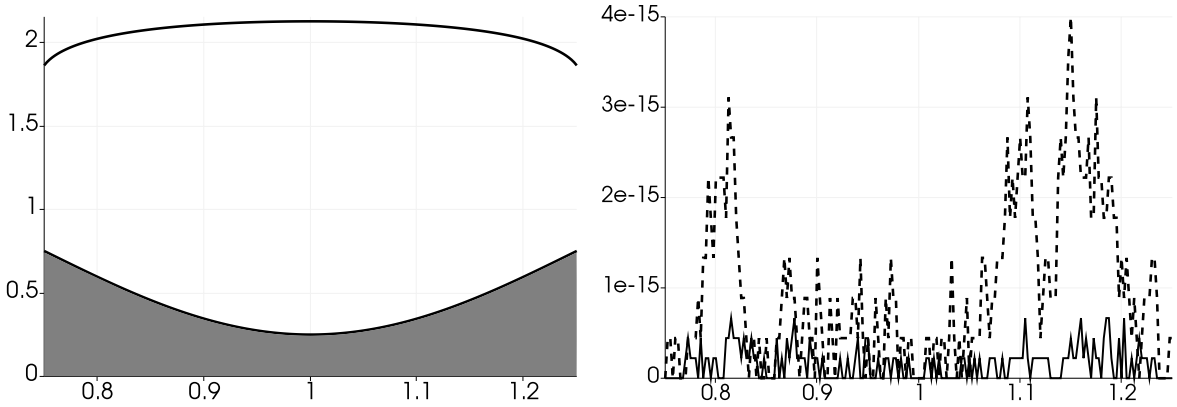

Figure 3.9 – Left panel: initial free surface for the subcritical topography steady state. Right panel: free surface (solid line) and discharge (dashed line) errors to the steady state after 1s, with the explicit scheme.

	$h+Z$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.17e-02	1.83e-02	1.44e-01	9.80e-03	1.09e-02	2.75e-02
HR	7.13e-03	1.53e-02	1.44e-01	5.37e-03	6.45e-03	2.40e-02
explicit	1.61e-16	2.40e-16	6.66e-16	9.79e-16	1.32e-15	4.00e-15
implicit	1.61e-16	2.40e-16	6.66e-16	9.79e-16	1.32e-15	4.00e-15

Table 3.8 – Free surface and discharge errors for the subcritical topography steady state.

following perturbed initial height:

$$h(0,x) = \begin{cases} h_{sub}(x) - 0.5 & \text{if } \dfrac{x-0.75}{1.25-0.75} \in \left[\dfrac{3}{7}, \dfrac{4}{7}\right]; \\ h_{sub}(x) & \text{otherwise.} \end{cases}$$

The initial discharge is still given by $q(0,x) = q_0$. The same boundary conditions as in the previous experiment are chosen.

After some time has elapsed, the approximate solution should converge to the original, unperturbed solution. Figure 3.10 shows the convergence to this solution, and Table 3.9 displays the error to the original solution after the time $t_{end} = 3$s. On Figure 3.10 and Table 3.9, we remark that the original steady state solution is indeed recovered. The convergence is obtained up to the machine precision for the explicit and implicit schemes, while the HLL and HR schemes only provide a first-order approximation of the steady state solution.

Supercritical topography steady state

We now consider the supercritical steady state solution $h_{sup}(x)$, which satisfies $h_{sup}(x) < h_c$. The initial conditions are $h(0,x) = h_{sup}(x)$ and $q(0,x) = q_0$. As boundary conditions, we take inhomogeneous Dirichlet boundary conditions, consisting in the steady state solution taken at $0.75 - \Delta x$ and $1.25 + \Delta x$. For the preservation of this steady solution, the final physical time is $t_{end} = 1$s, and we take $C = +\infty$. The results of the simulations are presented

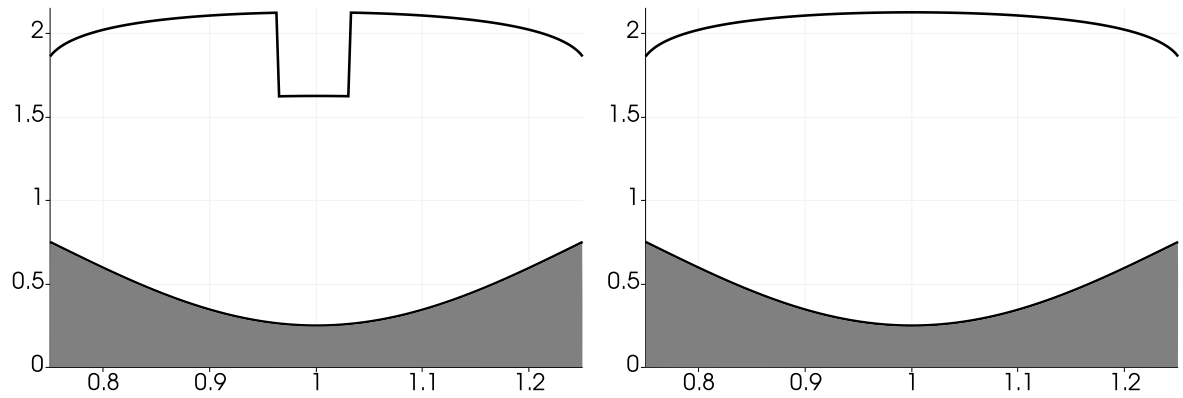

Figure 3.10 – Results of the explicit scheme for the perturbed subcritical topography steady state. Left panel: free surface at $t = 0$s. Right panel: free surface at $t = 3$s.

	$h+Z$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.17e-02	1.83e-02	1.44e-01	9.80e-03	1.09e-02	2.75e-02
HR	7.13e-03	1.53e-02	1.44e-01	5.37e-03	6.45e-03	2.40e-02
explicit	6.68e-15	8.12e-15	1.51e-14	1.23e-14	1.57e-14	2.93e-14
implicit	6.68e-15	8.12e-15	1.51e-14	1.23e-14	1.57e-14	2.93e-14

Table 3.9 – Free surface and discharge errors for the perturbed subcritical topography steady state.

on Figure 3.11 and in Table 3.10.

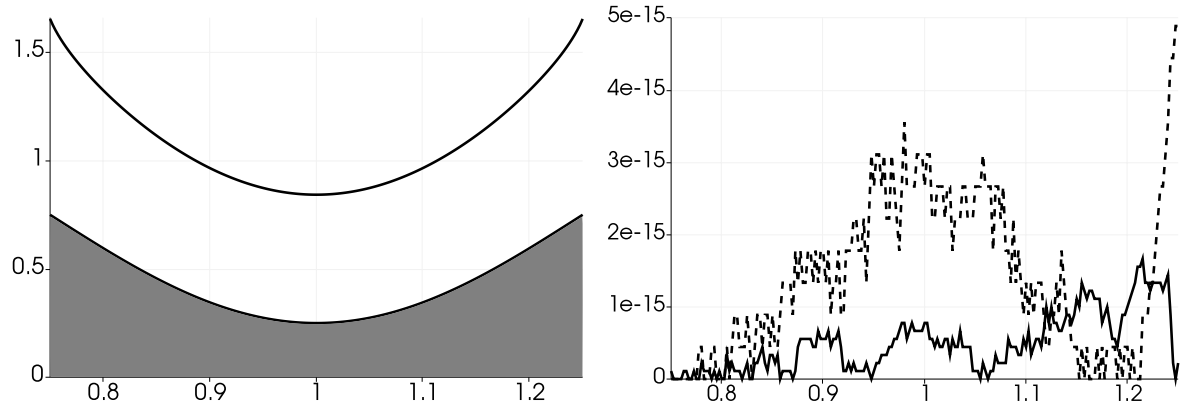

Figure 3.11 – Left panel: initial free surface for the supercritical topography steady state. Right panel: free surface (solid line) and discharge (dashed line) errors to the steady state after 1s, with the explicit scheme.

On the right panel of Figure 3.11, we check that the explicit scheme indeed preserves the steady state up to the machine precision. This observation is confirmed by Table 3.10, where the explicit and the implicit schemes are shown to exactly preserve the supercritical steady state. We also check that both HLL and HR schemes do not exactly preserve this moving steady state, as expected.

We know focus on a perturbation of this supercritical steady state solution. On the space

	$h+Z$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	9.21e-01	9.79e-01	1.25e+00	2.94e-02	3.36e-02	8.30e-02
HR	9.45e-01	9.93e-01	1.26e+00	3.07e-03	4.77e-03	3.71e-02
explicit	4.97e-16	6.39e-16	1.67e-15	1.41e-15	1.77e-15	4.88e-15
implicit	4.97e-16	6.39e-16	1.67e-15	1.41e-15	1.77e-15	4.88e-15

Table 3.10 – Free surface and discharge errors for the supercritical topography steady state.

domain $[0.75, 1.25]$, we take the following initial water height, which involves a perturbation:

$$h(0,x) = \begin{cases} h_{sup}(x) - 0.5 & \text{if } \dfrac{x-0.75}{1.25-0.75} \in \left[\dfrac{3}{7}, \dfrac{4}{7}\right]; \\ h_{sup}(x) & \text{otherwise.} \end{cases}$$

The discharge is still initialized to q_0, and we take the same boundary conditions as in the unperturbed case. For a large enough physical time, we should observe a convergence to the original, unperturbed supercritical steady state solution. To that end, we take $t_{end} = 3$s. The results of this simulation are displayed on Figure 3.12 and Table 3.11.

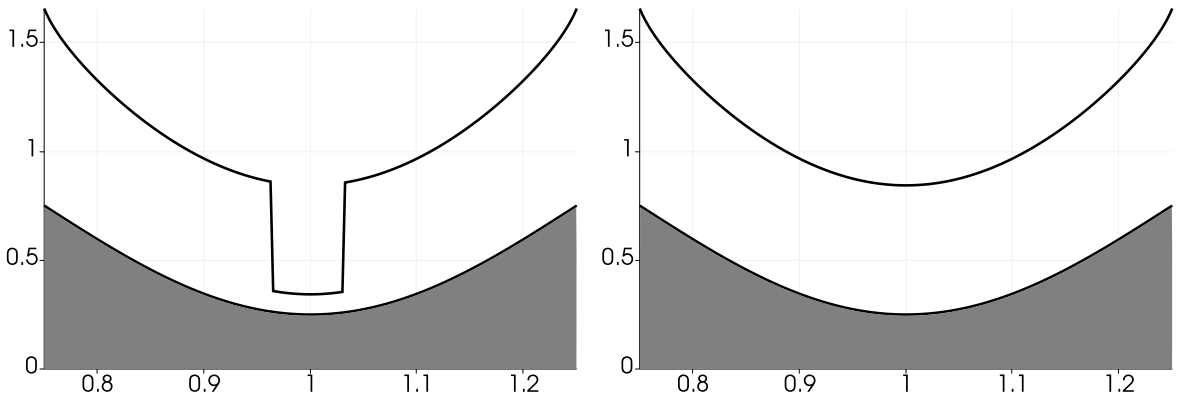

Figure 3.12 – Results of the explicit scheme. Left panel: free surface at $t = 0$s. Right panel: free surface at $t = 3$s.

	$h+Z$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	9.21e-01	9.78e-01	1.25e+00	3.03e-02	3.45e-02	9.23e-02
HR	9.44e-01	9.93e-01	1.26e+00	3.00e-03	4.71e-03	3.92e-02
explicit	8.93e-16	1.23e-15	3.44e-15	1.40e-15	1.88e-15	5.33e-15
implicit	8.93e-16	1.23e-15	3.44e-15	1.40e-15	1.88e-15	5.33e-15

Table 3.11 – Free surface and discharge errors for the perturbed supercritical topography steady state.

On Figure 3.12 and Table 3.11, we note that both the explicit and implicit schemes converge to the unperturbed supercritical steady state up to the machine precision. However, the HLL and HR schemes provide a first-order approximation of this supercritical steady state.

3.3.1.3 Moving steady states for the friction source term

We now focus on the preservation of the friction-only steady states, by assuming $q_0 \neq 0$ and a flat topography, i.e. $\partial_x Z = 0$. The smooth steady states are then given in Section 1.2.2.1, according to (1.53), or equivalently to (1.55).

In Section 1.2.2.1, the water height for a smooth friction steady state was obtained by considering a zero h of the nonlinear function χ defined by (1.57). Depending on the value of the difference between h and the critical height $h_c > 0$, defined by (1.49), we obtained either a subcritical solution (where $h > h_c$) or a supercritical solution (where $0 < h < h_c$). Examples of such solutions have been presented on Figure 1.21, and we once again consider these examples as the bases of the well-balance assessment of the proposed explicit and implicit schemes.

Subcritical friction steady state

For this experiment, the space domain is $[0.75, 0.9]$. We consider the subcritical solution of the steady state obtained by setting $q_0 = -\sqrt{g}/8$, $x_0 = 0.75 - \Delta x$ and $h_0 = h_c = 0.25$. The Manning coefficient k is chosen equal to 1.

The first experiment concerns the preservation of this steady state. We take $q(0, x) = q_0$ and the exact height $h_{sub}(x)$, obtained with Newton's method, as initial conditions. In addition, we impose the exact solution at the points $0.75 - \Delta x$ and $0.9 + \Delta x$ as inhomogeneous Dirichlet boundary conditions. Moreover, we set $C = 10^{-3}$, and we use a mesh made of 200 cells to compute the approximate solution until the final time $t_{end} = 1$s. The numerical results are presented on Figure 3.13, and the errors to the steady state are displayed in Table 3.12.

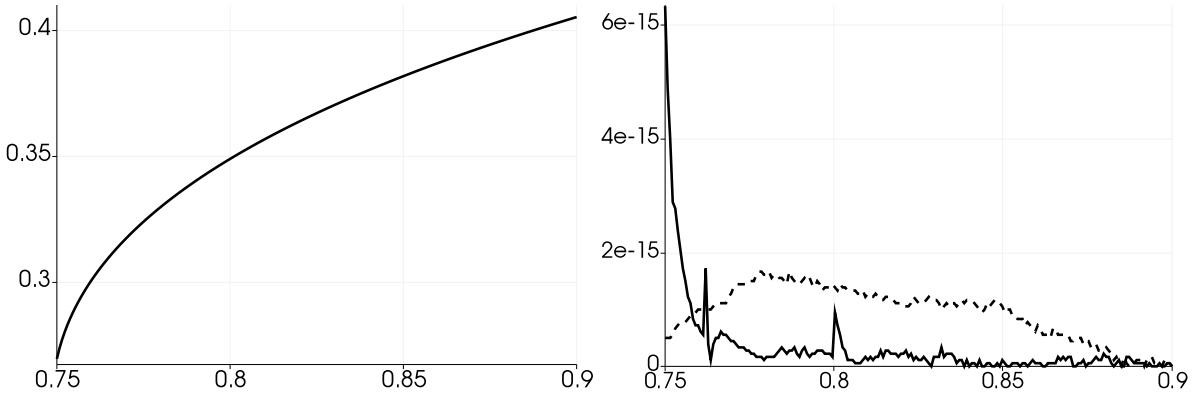

Figure 3.13 – Left panel: initial height for the subcritical friction steady state. Right panel: height (solid line) and discharge (dashed line) errors to the steady state after 1s, with the explicit scheme.

From Figure 3.13 and Table 3.12, we observe that this friction-only steady state is indeed preserved up to the machine precision by the explicit and implicit schemes, which now provide different numerical results because of the implicitation of the friction source term. However, the HLL and HR schemes do not preserve this steady state solution. It is worth noting that, here, the results from the HLL and the HR schemes are identical. Indeed, the topography is flat, the HR scheme is based on an HLL flux, and the treatment of the friction is identical

	h			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.21e-04	1.86e-04	5.51e-04	4.57e-05	7.27e-05	3.81e-04
HR	1.21e-04	1.86e-04	5.51e-04	4.57e-05	7.27e-05	3.81e-04
explicit	3.28e-16	8.00e-16	6.33e-15	9.47e-16	1.06e-15	1.67e-15
implicit	2.44e-16	7.33e-16	6.16e-15	3.72e-16	4.30e-16	7.77e-16

Table 3.12 – Height and discharge errors for the subcritical friction steady state.

for both schemes. As a consequence, the HR and the HLL schemes provide identical results in this case of a flat topography.

The second experiment introduces a perturbation of the subcritical steady state, as shown in Figure 3.14. With h_{sub} the exact height, this perturbation is defined by choosing the initial water height as follows:

$$h(0,x) = \begin{cases} h_{sub}(x) + 0.05 & \text{if } \dfrac{x-0.75}{0.9-0.75} \in \left[\dfrac{3}{7}, \dfrac{4}{7}\right]; \\ h_{sub}(x) & \text{otherwise.} \end{cases}$$

The initial discharge is unperturbed, and taken equal to $q_0 = -\sqrt{g}/8$ throughout the domain. The boundary conditions consist in the unperturbed exact solution. We use 100 discretization cells for the numerical simulation. Moreover, we take $C = 10^{-3}$. The computations are carried out until the final time $t_{end} = 5$s. Indeed, such a final time allows the perturbation to be dissipated and a steady state to be reached. In fact, this steady state turns our to be the original, unperturbed steady state. The results of the explicit scheme are presented in Figure 3.14 and an error comparison with the unperturbed steady state is provided in Table 3.13.

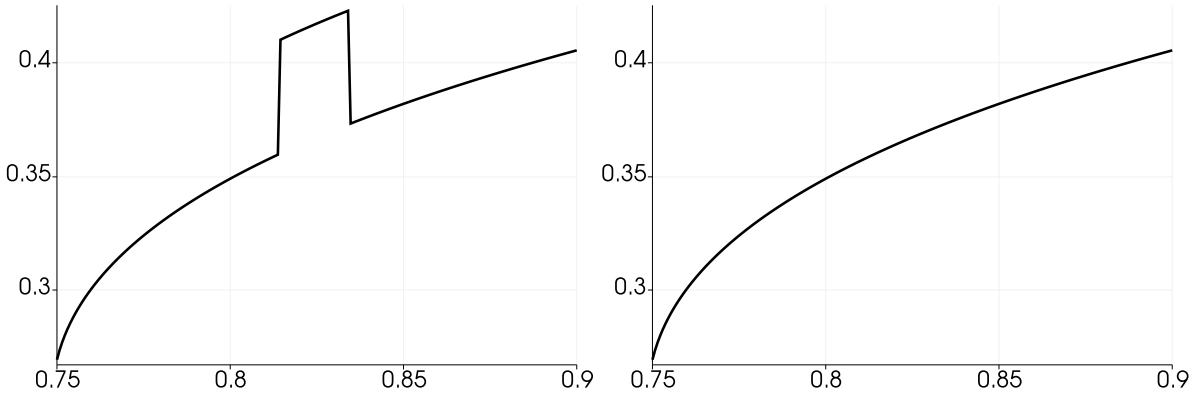

Figure 3.14 – Results of the explicit scheme for the perturbed subcritical friction steady state. Left panel: water height at $t = 0$s. Right panel: water height at $t = 5$s.

Figure 3.14 shows that the perturbation is eventually dissipated and that we recover the unperturbed steady state. This assertion is confirmed by the error analysis presented in Table 3.13, which shows that the explicit and implicit schemes recover the unperturbed steady state up to the machine precision. The HLL and HR schemes still provide an approximation of this steady state, and their results are identical.

	h			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	2.45e-04	2.72e-04	5.85e-04	8.86e-05	1.15e-04	4.70e-04
HR	2.45e-04	2.72e-04	5.85e-04	8.86e-05	1.15e-04	4.70e-04
explicit	1.87e-15	2.03e-15	7.33e-15	1.19e-15	1.33e-15	2.61e-15
implicit	4.24e-15	4.29e-15	8.27e-15	2.52e-15	2.90e-15	4.83e-15

Table 3.13 – Height and discharge errors for the perturbed subcritical friction steady state.

Supercritical friction steady state

We now consider the space domain $[0.75, 0.8]$, and we take $k = 1$. We focus on the supercritical branch of the previous steady state, obtained by assuming that $q_0 = -\sqrt{g}/8$, $x_0 = 0.75 - \Delta x$ and $h_0 = h_c = 0.25$.

The first experiment deals with the preservation of this supercritical steady state solution. The initial conditions of this experiment are $q(0, x) = q_0$ and $h(0, x) = h_{sup}(x)$, where $h_{sup}(x)$ is the supercritical water height obtained by approximately solving (1.53). The inhomogeneous Dirichlet boundary conditions consist in the initial condition at the points $0.75 - \Delta x$ and $0.8 + \Delta x$. The simulation is carried out until the final time $t_{end} = 1$s, on a mesh made of 200 discretization cells. In addition, we again take $C = 10^{-3}$. The results from the explicit scheme are depicted on Figure 3.15, and the errors to the steady state are presented in Table 3.14.

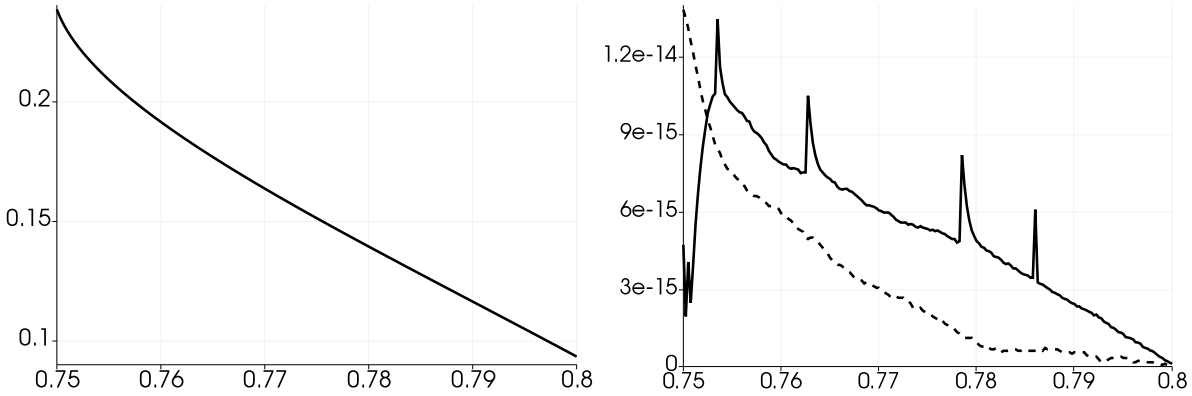

Figure 3.15 – Left panel: initial height for the supercritical friction steady state. Right panel: height (solid line) and discharge (dashed line) errors to the steady state after 1s, with the explicit scheme.

	h			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	3.62e-04	5.14e-04	2.45e-03	5.30e-08	7.94e-08	2.50e-07
HR	3.62e-04	5.14e-04	2.45e-03	5.30e-08	7.94e-08	2.50e-07
explicit	5.29e-15	6.03e-15	1.35e-14	3.15e-15	4.50e-15	1.38e-14
implicit	5.21e-15	5.91e-15	1.28e-14	3.11e-15	4.18e-15	1.23e-14

Table 3.14 – Height and discharge errors for the supercritical friction steady state.

We observe on Figure 3.15 and Table 3.14 that the explicit and implicit schemes exactly preserve this supercritical steady state solution. However, the HLL and HR provide a first-order approximation of this steady state.

We now study a perturbation of the aforementioned supercritical steady state, as shown in Figure 3.16. This perturbation is introduced by taking the following the initial water height:

$$h(0,x) = \begin{cases} h_{sup}(x) + 0.05 & \text{if } \dfrac{x-0.75}{0.9-0.75} \in \left[\dfrac{3}{7}, \dfrac{4}{7}\right]; \\ h_{sup}(x) & \text{otherwise,} \end{cases}$$

where h_{sup} is the supercritical water height.

We set the initial discharge as $q(0,x) = q_0$ on the whole space domain, and the boundary conditions still consist in the unperturbed supercritical solution. The numerical simulation uses 100 cells and we set $C = 10^{-3}$. The simulation is carried out until $t_{end} = 5$s, time at which the perturbed supercritical steady state has converged to the original supercritical steady state. Figure 3.16 depicts the results of the explicit scheme. Table 3.15 provides an error comparison between the HLL, HR, explicit and implicit schemes.

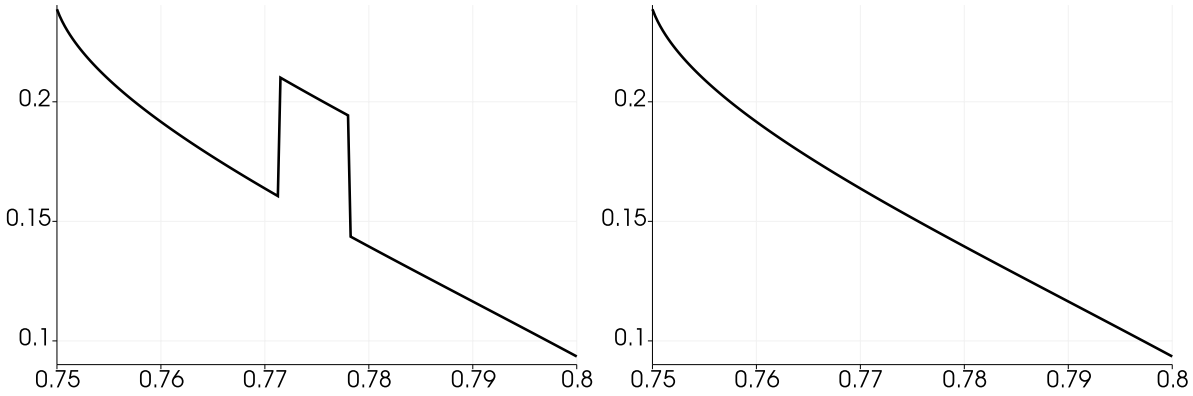

Figure 3.16 – Results of the explicit scheme for the perturbed supercritical friction steady state. Left panel: water height at $t = 0$s. Right panel: water height at $t = 5$s.

	h			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	3.62e-04	5.14e-04	2.45e-03	1.08e-07	1.62e-07	5.09e-07
HR	3.62e-04	5.14e-04	2.45e-03	1.08e-07	1.62e-07	5.09e-07
explicit	1.07e-14	1.40e-14	3.37e-14	2.11e-15	2.96e-15	6.44e-15
implicit	1.09e-14	1.43e-14	3.54e-14	2.07e-15	2.85e-15	5.88e-15

Table 3.15 – Height and discharge errors for the perturbed supercritical friction steady state.

From Figure 3.16, we note that the perturbed steady state indeed converges towards the unperturbed one. Table 3.15 shows that this convergence is valid up to the machine precision for the explicit and implicit schemes, and that the HLL and HR schemes provide an approximation of this steady state.

3.3.1.4 Moving steady states with both topography and friction source terms

We continue the verification of the well-balance property with numerical experiments consisting in the preservation of moving steady states involving both topography and friction. Thus, we set $k \neq 0$, $\partial_x Z \neq 0$, and $q_0 \neq 0$: the steady states are therefore given by the full equation (1.73). Recall that this equation cannot be rewritten under an algebraic form. Thus, to find a steady state solution, we have to either numerically solve the equation (1.73) in the general case, or exactly solve it in specific cases.

Uniform water height

We begin by considering the specific case where the height is uniform throughout the space domain. In that case, the derivative of the topography function is given in Section 1.2.3 by (1.74) (see also [42]). This specific case is tested numerically by taking $h_0 = q_0 = 1$, $k = 10$, $Z(0) = 0$ and the slope of Z given by (1.74). As a consequence, we have the following topography function:

$$Z(x) = -\frac{kx}{g}.$$

The space domain is $[0, 1]$ and is discretized using 100 cells. The initial and boundary conditions are the exact solution. The computations are carried out with all four schemes, and we take $t_{end} = 1$s as well as $C = +\infty$. The results are presented in Table 3.16.

	h			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	3.05e-01	3.69e-01	7.05e-01	8.64e-01	8.65e-01	9.41e-01
HR	3.07e-01	3.71e-01	7.08e-01	8.64e-01	8.65e-01	9.42e-01
explicit	1.24e-16	1.54e-16	2.22e-16	9.77e-17	1.59e-16	6.66e-16
implicit	2.22e-17	5.21e-17	2.22e-16	9.99e-17	1.84e-16	6.66e-16

Table 3.16 – Height and discharge errors for the topography and friction steady state with constant height.

Table 3.16 shows that this topography and friction steady state with constant height is indeed exactly preserved the explicit scheme and the implicit scheme. However, the HLL and HR only provide a first-order approximation of this steady state solution.

Uniform free surface

To build another exact solution of (1.73), we assume $h + Z = H_0$. We therefore have a constant free surface H_0 over the whole space domain $[0, 1]$. The exact height and topography functions for a steady state solution have been exhibited in Section 1.2.3. They are given by (1.77), under the existence condition (1.76). We choose the constants k, q_0 and h_0 such that h is positive over the whole domain $[0, 1]$, i.e. such that the condition (1.76) is satisfied. In the simulation, we set $x_0 = 0$ and $q_0 = h_0 = H_0 = 1$. As a consequence, the exact height and

topography functions are given by:

$$h(t,x) = (1+(\eta-1)kx)^{\frac{1}{\eta-1}},$$
$$Z(x) = 1-(1+(\eta-1)kx)^{\frac{1}{\eta-1}}.$$

We discretize the space domain with 100 cells and carry out the simulation until the time $t_{end} = 1$s. We take $C = +\infty$. An error comparison is displayed in Table 3.17.

	$h+Z$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.36e-03	1.37e-03	1.59e-03	8.76e-04	8.80e-04	9.59e-04
HR	4.00e-04	4.00e-04	4.66e-04	1.63e-03	1.63e-03	1.89e-03
explicit	1.66e-15	1.74e-15	2.66e-15	3.05e-14	3.06e-14	3.63e-14
implicit	5.00e-15	5.38e-15	6.88e-15	2.00e-14	2.01e-14	2.29e-14

Table 3.17 – Free surface and discharge errors for the topography and friction steady state with constant free surface.

The results presented in Table 3.17 show that the steady state is preserved up to the machine precision by the explicit and implicit schemes, while the HLL and HR approximate this steady state.

The general case

Finally, we derive a steady state for the shallow-water equations with topography and friction, without considering a constant height or free surface. Therefore, we approximately solve the discretization (3.68) of the full steady relation (1.73). First, we set $k = 0.01$ and we choose $[0,1]$ to be the space domain. We take the following topography function:

$$Z(x) = \frac{1}{2}\frac{e^{\cos(4\pi x)} - e^{-1}}{e^1 - e^{-1}}. \tag{3.103}$$

We set $q(0,x) = q_0 = 1$ throughout the domain. The equation (3.68) is then approximately solved using Newton's method, imposing $h(0,0) = 0.3$, to get the steady water height $h_{ex}(x)$. This procedure allows us to define the water height over the whole domain. This steady state ${}^t(h_{ex}, q_0)$ is then chosen as the initial and boundary conditions for this experiment. We take 100 discretization cells and $C = +\infty$; the numerical simulation runs until a final physical time $t_{end} = 1$s. The results of the explicit scheme are presented on Figure 3.17 and the errors to the steady state are displayed in Table 3.18.

Figure 3.17 shows that the explicit scheme exactly preserves this topography and friction steady state. Moreover, Table 3.18 shows the implicit scheme also exactly preserve such a steady state. On the contrary, the HLL and HR schemes produce an approximation of this solution.

Then, we focus on a perturbation of the above steady state. We denote by $h_{ex}(x)$ the water height of the previous topography and friction steady state. The initial water height of this

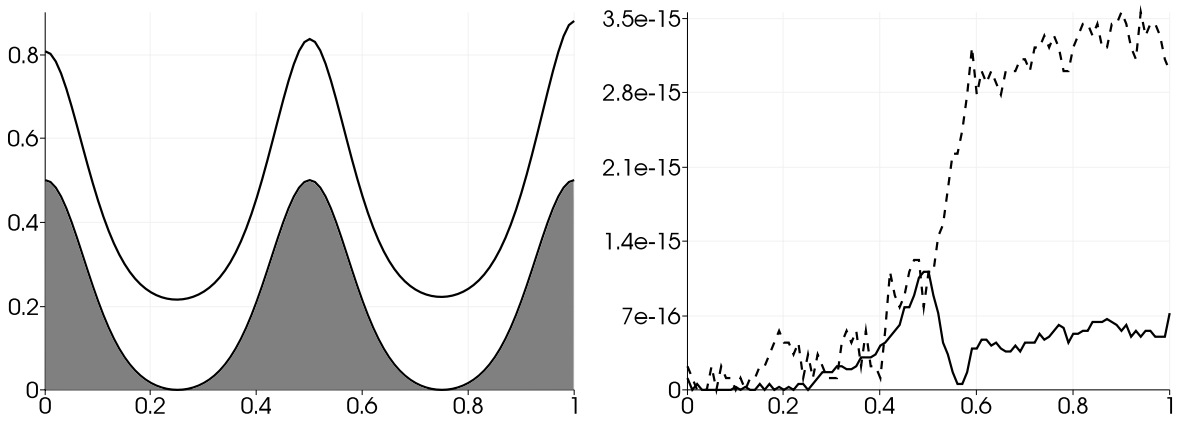

Figure 3.17 – Left panel: initial height for the topography and friction steady state. Right panel: height (solid line) and discharge (dashed line) errors to the steady state with the explicit scheme.

	$h+Z$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.84e-03	3.01e-03	1.20e-02	3.26e-09	3.64e-09	5.55e-09
HR	5.29e-02	1.43e-01	6.23e-01	4.63e-02	9.73e-02	2.94e-01
explicit	8.29e-16	1.36e-15	5.33e-15	1.07e-15	1.33e-15	3.33e-15
implicit	6.23e-16	9.68e-16	2.72e-15	2.45e-15	2.87e-15	5.11e-15

Table 3.18 – Free surface and discharge errors for the topography and friction steady state.

last experiment is defined by

$$h(0,x) = \begin{cases} h_{ex}(x) + 0.05 & \text{if } x \in \left[\frac{2}{7}, \frac{3}{7}\right] \bigcup \left[\frac{4}{7}, \frac{5}{7}\right], \\ h_{ex}(x) & \text{otherwise,} \end{cases}$$

and the initial discharge is defined by

$$q(0,x) = \begin{cases} q_0 + \frac{1}{2} & \text{if } x \in \left[\frac{2}{7}, \frac{3}{7}\right] \bigcup \left[\frac{4}{7}, \frac{5}{7}\right], \\ q_0 & \text{otherwise.} \end{cases}$$

The unperturbed steady state is prescribed as the boundary conditions. For this numerical experiment, the domain $[0,1]$ is discretized with 100 cells, and the simulation runs until the perturbation has been dissipated and the unperturbed steady state has been recovered. The final physical time we choose for these conditions to be met is $t_{end} = 2$s. In addition, we set $C = +\infty$. The evolution of the perturbation with the explicit scheme is depicted on Figure 3.18. Then, in Table 3.19, we present the errors to the original unperturbed steady state when the physical time has elapsed.

Figure 3.18 and Table 3.19 show that the schemes indeed allow to recover the original unperturbed steady state. This experiment emphasizes the ability of the explicit and implicit schemes to exactly capture a topography and friction steady state, even after a perturbation.

The HLL and HR schemes do not possess such an ability.

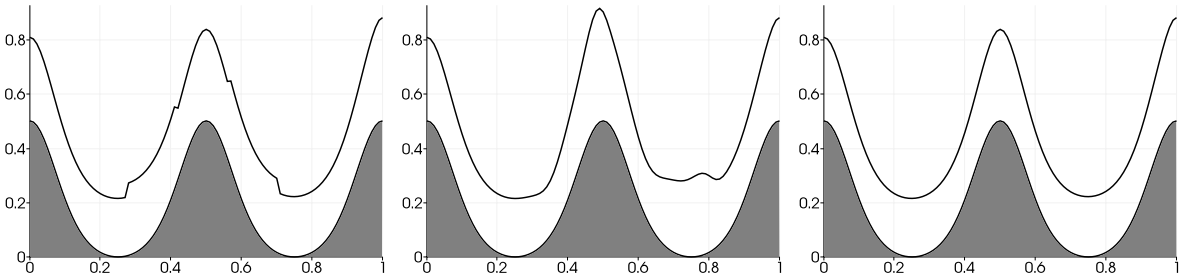

Figure 3.18 – Perturbed topography and friction steady state. From left to right: water height for $t = 0$s, $t = 0.015$s and $t = 2$s, with the explicit scheme.

	$h + Z$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	1.84e-03	3.01e-03	1.20e-02	9.20e-10	1.03e-09	1.56e-09
HR	3.18e-01	4.95e-01	9.22e-01	2.61e-02	3.36e-02	1.30e-01
explicit	5.71e-16	1.02e-15	4.16e-15	7.36e-16	1.08e-15	5.44e-15
implicit	1.47e-15	2.00e-15	5.72e-15	7.16e-16	9.17e-16	2.89e-15

Table 3.19 – Free surface and discharge errors for the perturbed topography and friction steady state.

3.3.1.5 Goutal and Maurel's steady flows

We finally carry out three experiments from Goutal and Maurel's test cases [86]. These benchmarks have been derived by considering the Bernoulli equation (1.44) that governs the steady state solutions of the shallow-water equations with non-flat topography and a vanishing friction contribution, i.e. $k = 0$. As a consequence, the implicit scheme and the explicit scheme will yield identical results.

Note that we have already assessed the preservation of moving steady state solutions in Section 3.3.1.2. However, the experiments presented in [86] are obtained after a transient state, i.e. from initial conditions which do not define a steady state. The goal of presenting these experiments is to assess the ability of the suggested scheme to capture steady state solutions, in addition to preserving them.

The experiments from [86] are called the *subcritical flow*, the *transcritical flow without shock* and the *transcritical flow with shock*. In this manuscript, they will be respectively labeled GM1, GM2 and GM3. The space domain is $0 < x < 25$ and the topography function is given by:

$$Z(x) = \left(0.2 - 0.05(x - 10)^2\right)_+.$$

The boundary conditions are given hereafter, in function of two quantities q_0 and h_0, whose values depend on the experiment studied:

- on the left boundary, the water height satisfies a homogeneous Neumann condition and the discharge is set to some q_0;

- on the right boundary, the water height is set to h_0 when the flow is subcritical (and a homogeneous Neumann boundary condition is prescribed otherwise), and the discharge follows a homogeneous Neumann boundary condition.

In addition, the initial conditions are $h(0,x) + Z(x) = h_0$ and $q(0,x) = 0$ throughout the domain. The values of q_0 and h_0 are:

- for GM1: $q_0 = 4.42\text{m}^3/\text{s}$ and $h_0 = 2\text{m}$;
- for GM2: $q_0 = 1.53\text{m}^3/\text{s}$ and $h_0 = 0.66\text{m}$;
- for GM3: $q_0 = 0.18\text{m}^3/\text{s}$ and $h_0 = 0.33\text{m}$.

Such initial and boundary conditions yield a transient state followed by a steady state, with uniform discharge q_0. For GM1 and GM2, this steady state is continuous, and it should thus be exactly obtained by the well-balanced scheme. However, the steady state in GM3 involves a stationary shock, which the well-balanced scheme is not able to capture exactly. After Section 1.2.1.3, this stationary shock is governed by the Rankine-Hugoniot relations and the discrete entropy inequality.

On the one hand, for the converged steady states associated to GM1 and GM2, note that $q = q_0$ and that the steady state equation (1.44) is verified. This equation is nothing but a statement of Bernoulli's principle, and it can be rewritten as follows:

$$\frac{q_0^2}{2h^2} + g(h + Z) = \mathcal{E},$$

where $\mathcal{E}$ is a uniform quantity, the total head (see [86] for instance). As a consequence, to evaluate the well-balance of the scheme on GM1 and GM2, we compute the error to the uniform discharge q_0 and the error to the uniform total head $\mathcal{E}$.

On the other hand, since GM3 presents a stationary shock, the discharge is constant but the total head is not. Indeed, it presents a discontinuity where the shock is located. Therefore, only the error to the uniform discharge q_0 is computed for this last experiment.

The final physical time t_{end} and the constant C are chosen as follows for each experiment:
– for GM1: $t_{end} = 500\text{s}$ and $C = +\infty$;
– for GM2: $t_{end} = 125\text{s}$ and $C = 2.5$;
– for GM3: $t_{end} = 1000\text{s}$ and $C = 1.1$.

Subcritical flow

We display on Figure 3.19 the results of the explicit scheme for the GM1 benchmark. Then, we compare in Table 3.20 the two well-balanced schemes with the HR scheme and the HLL scheme. These experiments are performed using a mesh of 200 cells. Table 3.20 shows that both HLL and HR schemes provide a first-order approximation of the moving steady state configuration GM1, while the proposed explicit and implicit schemes exactly preserve (i.e. up to the machine precision) such moving steady states. This result is also observed on Figure 3.19. Moreover, these schemes recover this steady state after a transient state, even though the steady state is not prescribed as initial condition.

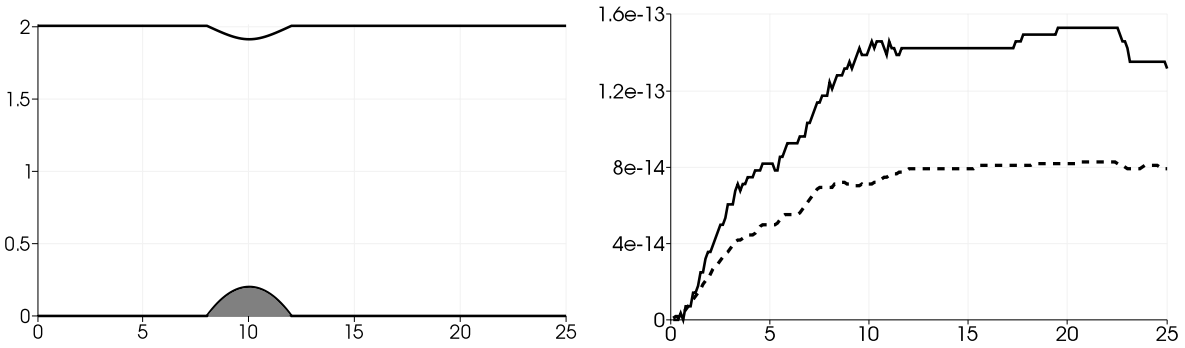

Figure 3.19 – Left panel: free surface and topography for the GM1 subcritical flow test case. Right panel: errors for the subcritical flow using the explicit scheme; the solid line is the total head error and the dashed line is the discharge error.

	$\mathcal{E}$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	8.24e-03	1.19e-02	7.41e-02	4.31e-03	1.22e-02	5.19e-02
HR	1.32e-02	1.97e-02	7.48e-02	2.37e-03	6.74e-03	2.74e-02
explicit	1.18e-13	1.25e-13	1.53e-13	6.65e-14	6.99e-14	8.26e-14
implicit	1.18e-13	1.25e-13	1.53e-13	6.65e-14	6.99e-14	8.26e-14

Table 3.20 – Total head and discharge errors for the GM1 subcritical flow experiment.

Transcritical flow

The results of the explicit scheme for the GM2 experiment are depicted on Figure 3.20. Then, we present in Table 3.21 the comparison between the well-balanced schemes and the non well-balanced schemes. We chose a mesh made of 200 discretization cells to carry out these experiments. From Figure 3.20 and Table 3.21, we get similar conclusions than the case of the GM1 experiment. Indeed, the HLL and HR approximate the steady state, while the suggested schemes exactly capture the steady state, even after a transient state.

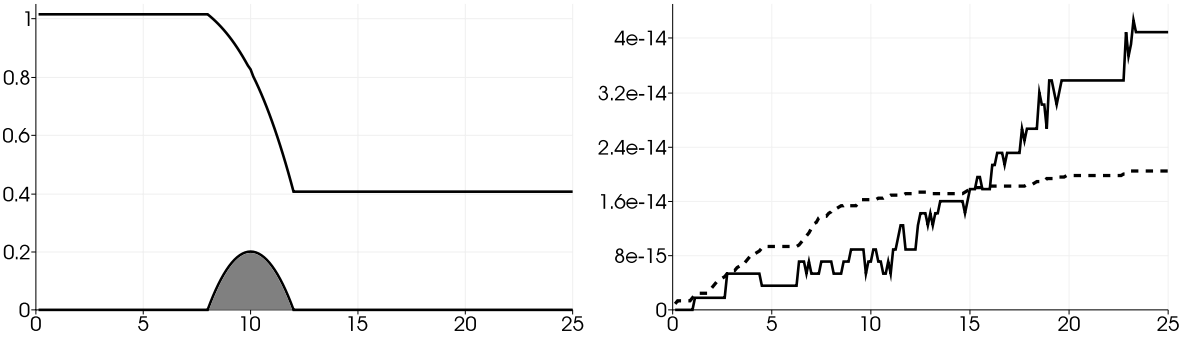

Figure 3.20 – Left panel: free surface and topography for the GM2 transcritical flow test case. Right panel: errors for the transcritical flow using the explicit scheme; the solid line is the total head error and the dashed line is the discharge error.

	$\mathcal{E}$			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	2.72e-02	3.50e-02	7.45e-02	1.54e-03	6.16e-03	3.70e-02
HR	4.79e-02	6.07e-02	8.12e-02	8.28e-04	3.30e-03	1.82e-02
explicit	1.67e-14	2.13e-14	4.26e-14	1.47e-14	1.58e-14	2.04e-14
implicit	1.67e-14	2.13e-14	4.26e-14	1.47e-14	1.58e-14	2.04e-14

Table 3.21 – Total head and discharge errors for the GM2 transcritical flow experiment.

Transcritical flow with shock

Finally, we turn to the GM3 test case. Since it contains a stationary shock, it is not exactly captured by the suggested explicit and implicit schemes, which are designed to capture smooth steady states. The results of the explicit scheme are displayed on Figure 3.21. Comparisons with respect to Δx and to the scheme used are presented on Figure 3.22, and comparisons with the HR and HLL schemes are presented in Table 3.22. The experiment is first carried out with 1000 discretization cells, and then with 4000 discretization cells.

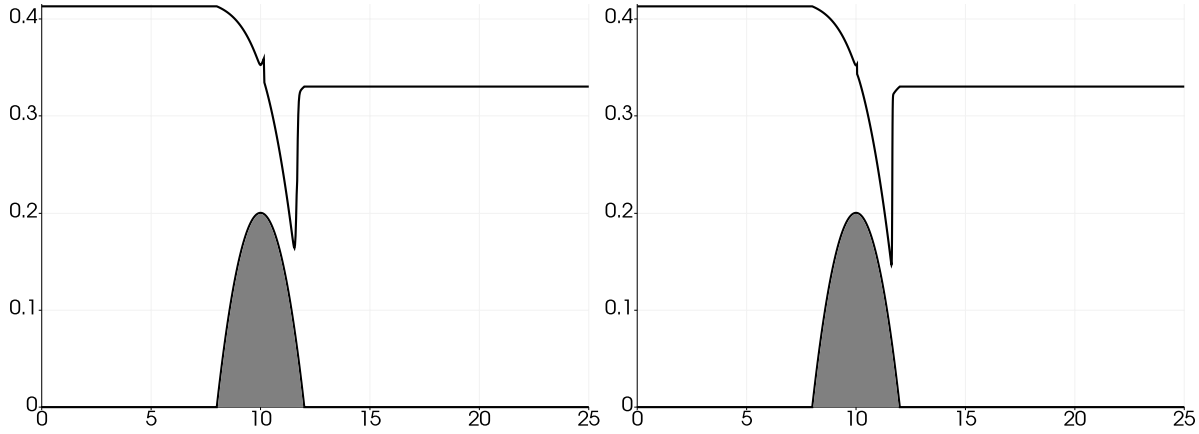

Figure 3.21 – Transcritical flow with shock experiment (GM3), with the explicit scheme. The topography is the gray area. Left panel: free surface and topography with 1000 discretization cells. Right panel: free surface and topography with 4000 discretization cells.

	q		
	L^1	L^2	L^∞
HLL	2.99e-04	1.84e-03	3.89e-02
HR	1.54e-04	1.53e-03	4.00e-02
explicit	2.54e-04	2.99e-03	5.01e-02
implicit	2.54e-04	2.99e-03	5.01e-02

Table 3.22 – Discharge errors for the experiment of the transcritical flow with shock (GM3) for 1000 discretization cells.

From Figure 3.21, we observe, as expected, that the GM3 experiment is not exactly captured by the explicit scheme. Note the presence of a small inconsistent discontinuity on the free surface in the vicinity of the top of the bump. The amplitude of this discontinuity de-

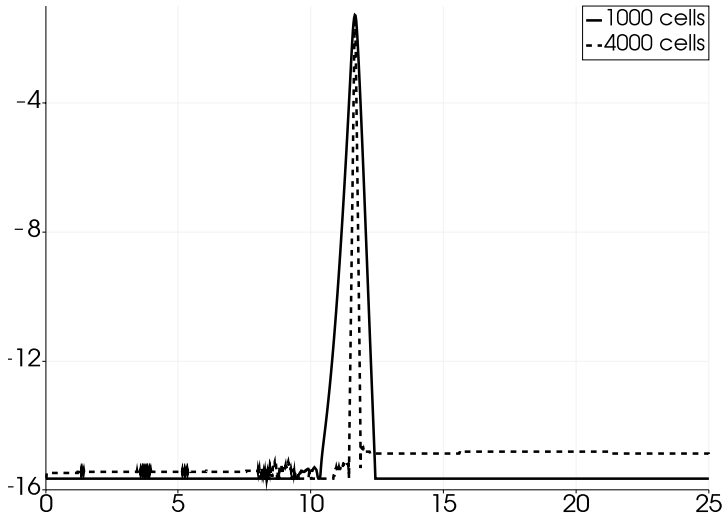

Figure 3.22 – Transcritical flow with shock (GM3) experiment: discharge error in logarithmic scale with the explicit scheme, with respect to the number of cells.

pends heavily on the value of the constant C. Moreover, this amplitude is reduced when Δx is reduced, which means that the explicit scheme indeed converges towards the required steady state when Δx tends to 0.

Table 3.22 gives the errors to the steady discharge q_0. We note that they are of the same order of magnitude with the four schemes under consideration.

On Figure 3.22, we observe the expected behavior of the discharge error. Although we do not exactly recover the exact solution, the shock becomes narrower as the number of cells increases.

Note that the proposed well-balanced schemes can also be compared to several other well-balanced schemes that preserve moving steady states. For instance, in [128, 157], error tables are provided, to show that the presented schemes indeed exactly preserve the studied moving steady states. However, it is worth noting that there is no evidence that these schemes are able to capture the steady states obtained after a transient state, contrary to the explicit and implicit schemes. We also mention the generalized hydrostatic reconstruction suggested in [33], which results in a scheme able to capture the moving steady states obtained after a transient state. However, this scheme is not robust when in the presence of large discontinuities in the topography function.

3.3.2 Validation experiments

In the previous section, we have assessed the well-balance of the suggested explicit and implicit schemes. We now turn to the other properties of the scheme: namely, the consistency and the robustness.

In this section, we perform several numerical experiments, whose goals are to show that the proposed schemes approximate the correct solutions when considering unsteady flows. We first present two experiments showing the influence of the parameter C present in the approximate topography source term $\overline{S}^t$ and used in (3.55). Then, we focus on the topography source term, and we take a vanishing friction contribution. To that end, we carry out

two experiments from [77]; both experiments present dry areas. Finally, several dam-break experiments are presented. The purpose of these experiments is to highlight the impact of the friction contribution on the solutions. The well-balance property is also shown to be important for these simulations.

3.3.2.1 Influence of the parameter C

Recall that the approximate source term $\overline{S}^t$, defined by (3.75), depends on C. Namely, this parameter is used to ensure the consistency of the approximate source term $\overline{S}^t$ with the actual source term S^t, by ensuring that the absolute value of the water height jump $[h]$ is no larger than $C\Delta x$. The purpose of the first set of experiments is to highlight the influence of this parameter C.

Shock waves over a flat topography

The first experiment we suggest concerns the propagation of shock waves over a flat topography. To that end, we consider the domain $[0, 1]$ with a flat topography (i.e. $Z \equiv 0$), homogeneous Neumann boundary conditions, and we take the following Riemann problem initial data:

$$h(0, x) = 1 \quad \text{and} \quad q(0, x) = \begin{cases} 7.5 & \text{if } x < 0.5, \\ -7 & \text{if } x \geq 0.5. \end{cases}$$

According to Section 1.1.2, such a Riemann problem corresponds to the two-shock case, and will thus produce two shock waves separating a constant intermediate state, the first one travelling towards the left of the initial discontinuity, and the second one towards its right. The exact solution of this Riemann problem can be computed using according to Section 1.1.2.

Recall that the cutoff involving C had been introduced to ensure the consistency of $\overline{S}^t$, especially for flat topographies. As a consequence, for this experiment, the value of C should be instrumental in getting the correct shocks waves and an accurate approximation of the intermediate state.

To carry out this experiment, we set a vanishing friction contribution (i.e. $k = 0$), we use 250 discretization cells, and we take the final time $t_{end} = 0.1$s. This experiment is performed with values of the parameter C ranging from 10 to 1000, i.e. for $C\Delta x$ ranging from 0.04 to 4.

The results of the explicit scheme are presented on Figure 3.23, where we display the exact solution as well as its numerical approximation obtained for $C = 10$ and $C = 1000$. The left panel shows that, for both values of C, the approximate shock waves are located at a consistent position and seem to have the correct amplitude. However, on the right panel, we note, on the one hand, that the intermediate state obtained with $C = 1000$ presents spurious oscillations, whose amplitude does not decrease when Δx decreases. On the other hand, with $C = 10$, the explicit scheme provides a good approximation of the intermediate state.

In order to quantify the loss of accuracy, we compute the errors, in L^1 and L^2 norms, between the approximate water height and the exact water height. These errors are presented with respect to the parameter C, for values of C ranging from 10 to 1000, on Figure 3.24. In both norms, the error increases as C increases. We also note that, for $C < 20$ in L^1-norm and for $C < 75$ in L^2-norm, the error stays the same. Similarly, for $C > 600$ in L^1-norm and

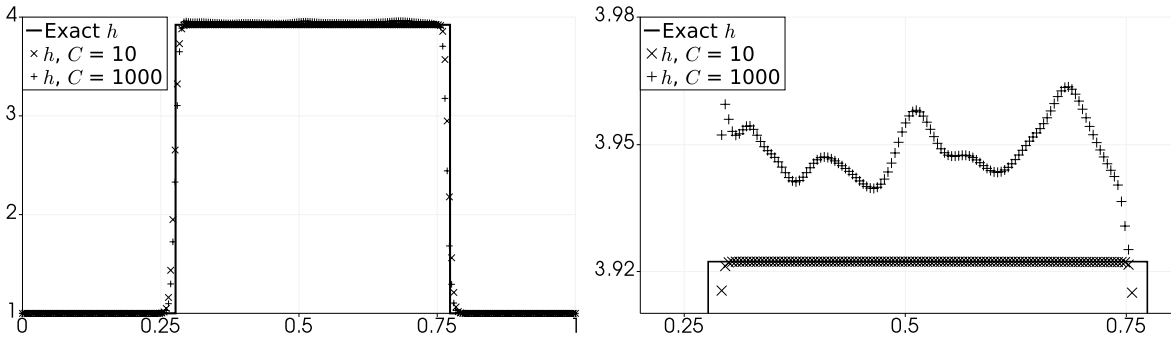

Figure 3.23 – Dam-break creating two shock waves over a flat bottom. Left panel: whole domain depicted at $t = 0.1$s. Right panel: zoom on the intermediate state of the dam-break problem.

for $C > 500$ in L^2-norm, the error is constant. As a consequence, this experiment is a good illustration of the fact that the cutoff (3.55) allows the recovery of a consistent scheme.

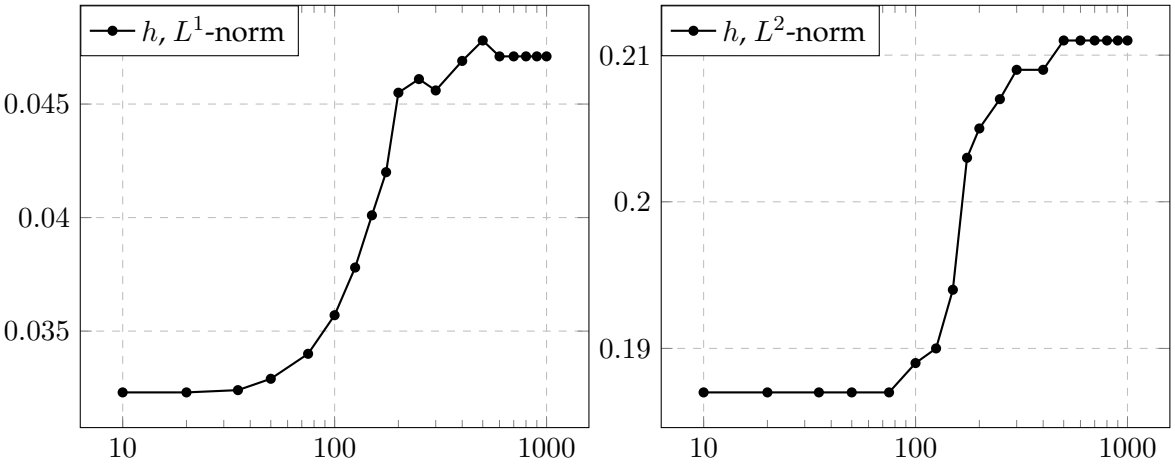

Figure 3.24 – Height error in L^1-norm (left panel) and L^2-norm (right panel) with respect to the parameter C.

Incident wave on an emerging bottom

The second experiment we perform in order to study the influence of the parameter C consists in an incident wave on an emerging bottom. To that end, we modify the flow at rest with emerging bottom experiment from [77], to add a wave perturbing the water at rest. On the space domain $[0, 15]$, we consider the topography function $Z(x) = \left(0.2 - 0.05(x - 10)^2\right)_+$ and the following initial conditions:

$$h(0,x) + Z(x) = \begin{cases} 0.2 & \text{if } x < 5, \\ 0.15 & \text{if } x \geq 5, \end{cases} \quad \text{and} \quad q(0,x) = 0.$$

Homogeneous Neumann boundary conditions are prescribed, and 4000 discretization cells are considered. We once again take a vanishing friction contribution, i.e. $k = 0$. The initial condition, as well as the reference solution, computed with the HR scheme at $t_{end} = 5$s, are displayed on Figure 3.25.

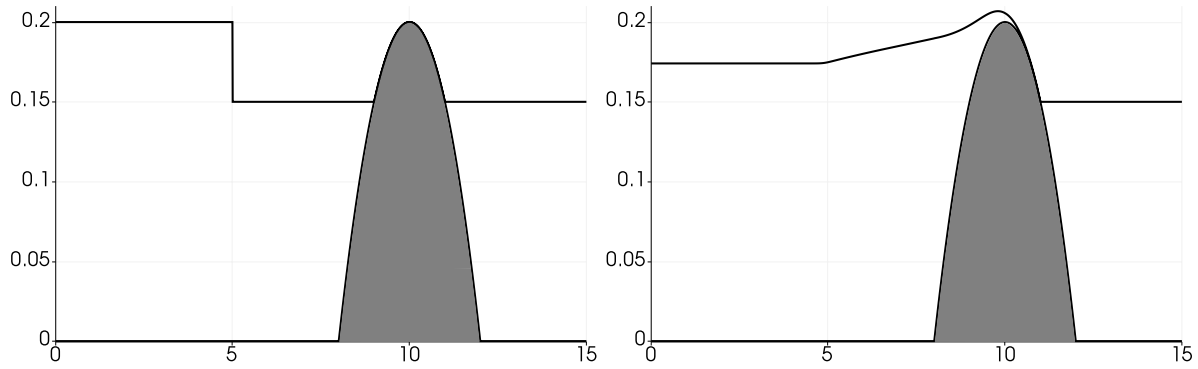

Figure 3.25 – Incident wave on an emerging bottom. Left panel: Initial free surface. Right panel: Reference free surface obtained with the HR scheme. On both panels, the gray area is the topography.

In order to study the influence of the parameter C, we consider the explicit scheme, and we carry out the simulation with $C = 1$ and $C = 10$. The results are displayed on Figure 3.26, where we once again observe that the consistency is ensured by the cutoff procedure. Indeed, with $C = 10$, large spurious oscillations appear, a wave is reflected to the left of the domain, and an inconsistent dry area appears on the bump. However, with $C = 1$, we only get small oscillations near $x = 9.75$; the amplitude of these oscillations decreases as Δx decreases.

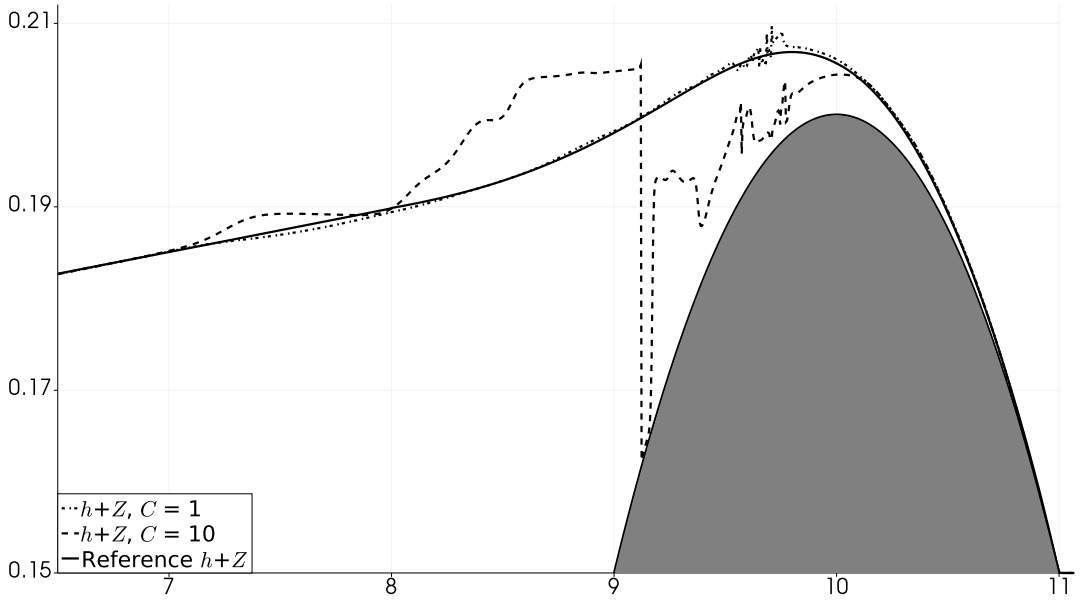

Figure 3.26 – Incident wave on an emerging bottom: zoomed comparison between the HR scheme, the explicit scheme with $C = 1$, and the explicit scheme with $C = 10$. The gray area is still the topography.

3.3.2.2 Drain on a non-flat bottom

The next validation experiment we propose is the drain on a non-flat bottom (see [77]). The topography is given on the space domain $[0, 25]$ by

$$Z(x) = \left(0.2 - 0.05(x - 10)^2\right)_+.$$

We take initial data at rest, as follows: $h(0, x) = 0.5 - Z(x)$ and $q(0, x) = 0$.

Concerning the boundary conditions, we assume that the left boundary is a solid wall and that the drain is done by the right boundary, where we impose an outlet condition on a dry bed (see [66, 30, 77] for more details on this boundary condition). These boundary conditions are given as follows. Let us denote by h_L and q_L the left boundary conditions, and by h_R and q_R the right boundary conditions. Let us assume that $(W_i^n)_{i\in[\![1,N]\!]}$ is the vector containing the approximate solution at time t^n. Then, the left boundary condition, which represents a solid wall, is taken as follows:

$$h_L = h_1^n \quad \text{and} \quad q_L = 0.$$

Concerning the right boundary condition, the process to obtain an outlet over a dry bed is detailed in [66, 30]. It consists in choosing the following values at the boundary:

$$h_R = \min\left(\frac{1}{9g}\left(u_N^n + 2\sqrt{gh_N^n}\right)^2, h_N^n\right) \quad \text{and} \quad q_R = \frac{h_R}{3}\left(u_N^n + 2\sqrt{gh_N^n}\right).$$

Note that the outlet on a dry bed boundary condition also requires that the numerical flux at the right boundary be the exact physical flux applied to ${}^t(h_R, q_R)$. This boundary condition enables the draining of the water through the right boundary.

The simulation is carried out with the implicit scheme, using a discretization of 200 cells, and until the final physical time $t_{end} = 1000$s. Note that, since the friction contribution is zero, the explicit scheme (3.9) – (3.81) and the implicit scheme (3.88) – (3.91) – (3.101) coincide. In addition, we take $C = 0.65$. The results are presented on Figure 3.27, where we observe that the implicit scheme provides results close to the ones from other schemes, given in [77, 20, 161, 18] for instance.

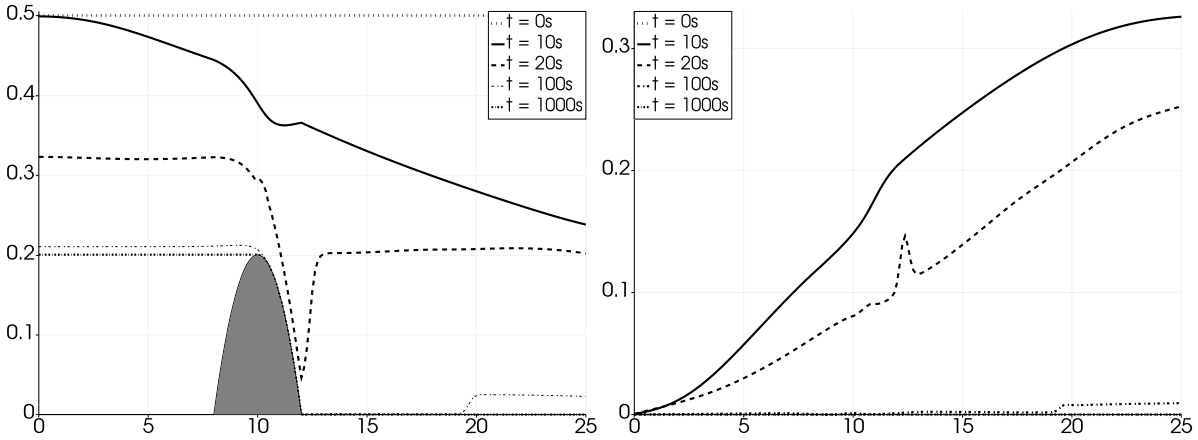

Figure 3.27 – Drain on a non-flat bottom. Left panel: free surface and topography (in gray). Right panel: discharge.

Note that this experiment converges to a steady state at rest (i.e. $q(t,x)=0$ over the whole domain), with the free surface equal to 0.2m at the left of the bump, and with a dry state at its right. Table 3.23 shows the convergence over time of the implicit scheme towards this steady state.

	h			q		
	L^1	L^2	L^∞	L^1	L^2	L^∞
150s	3.73e-03	4.27e-03	7.23e-03	6.79e-04	7.05e-04	3.50e-03
600s	4.13e-04	7.95e-04	1.55e-03	4.64e-05	4.94e-05	8.93e-05
2400s	9.59e-05	2.83e-04	3.57e-04	1.45e-06	2.26e-06	4.67e-06
19200s	2.11e-05	4.09e-05	8.50e-05	9.21e-08	3.47e-07	5.56e-07

Table 3.23 – Water height and discharge errors over time for the drain on a non-flat bottom.

3.3.2.3 Vacuum occurrence by a double rarefaction wave over a step

We then turn to another validation experiment, a vacuum occurrence deriving from a double rarefaction wave over a step, presented in [77]. We consider the space domain $[0,25]$, with a topography given by

$$Z(x)=\begin{cases}1 & \text{if } x\in\left(\frac{25}{3},\frac{25}{2}\right),\\ 0 & \text{if } x\in\left(0,\frac{25}{3}\right)\bigcup\left(\frac{25}{2},25\right).\end{cases}$$

The discontinuous initial data is given as follows:

$$h(0,x)=10 \quad \text{and} \quad q(0,x)=\begin{cases}-350 & \text{if } x<\frac{50}{3},\\ 350 & \text{otherwise.}\end{cases}$$

We prescribe homogeneous Neumann boundary conditions. The mesh consists in 200 discretization cells, and the simulation is carried out with the implicit scheme until a final physical time $t_{end}=0.65$s. We once again note that the explicit scheme and the implicit one coincide for a vanishing friction source term. We set $C=1$. The results are displayed on Figure 3.28, where we observe that the implicit scheme provides an approximation that is in good accordance with the ones obtained by several other schemes, given in [77, 20, 24, 158] for instance.

Note that, if the physical time t is large enough, the space domain should be completely drained of water, i.e. $h\equiv 0$ and $q\equiv 0$. We now analyze whether the four schemes at our disposal tend to that limit behavior. We introduce the time t_∞, such that for all $x\in[0,25]$, $q(t_\infty,x)<\varepsilon_{machine}$, where $\varepsilon_{machine}\simeq 2.22\times 10^{-16}$ is the lower bound of the double precision floating point numbers. The results are presented in Table 3.24, where we note that the HLL scheme, the explicit scheme and the implicit scheme all allow the space domain to be completely devoid of water for $t\geq t_\infty$. However, with the HR scheme, even in presence of a very

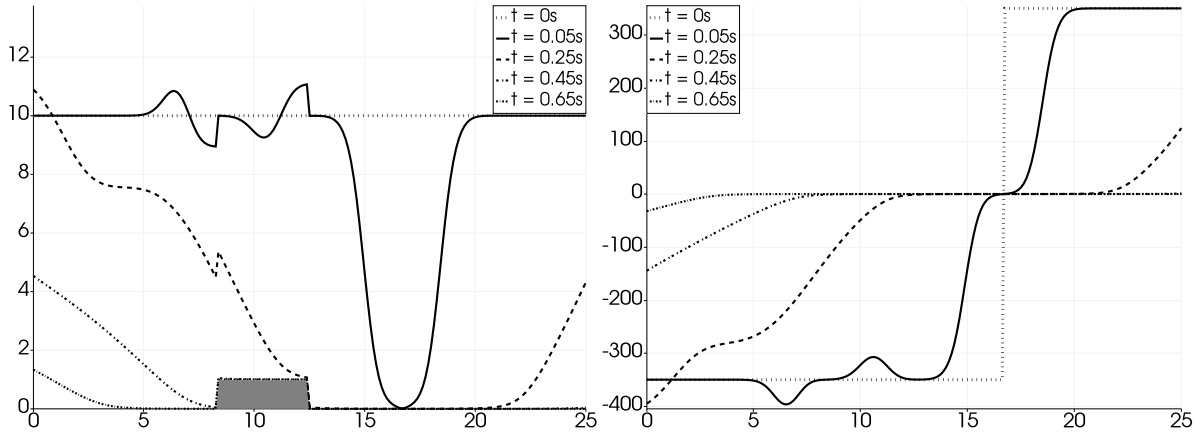

Figure 3.28 – Vacuum occurrence by a double rarefaction wave over a step. The gray area represents the topography. Left panel: free surface and topography. Right panel: discharge.

large physical time, there is always some water at rest remaining to the right of the bump, even if the discharge has reached $\varepsilon_{machine}$.

		h			q		
	t_∞	L^1	L^2	L^∞	L^1	L^2	L^∞
HLL	2.45s	4.05e-16	2.39e-15	3.37e-14	3.00e-16	3.20e-16	5.60e-16
HR	$> 10^4$s	7.02e-06	9.93e-06	1.40e-05	2.22e-16	2.22e-16	2.22e-16
explicit	1.78s	2.28e-16	2.29e-16	3.78e-16	2.40e-16	2.50e-16	6.33e-16
implicit	1.78s	2.28e-16	2.29e-16	3.78e-16	2.40e-16	2.50e-16	6.33e-16

Table 3.24 – Vacuum occurrence by a double rarefaction wave over a step experiment. Time t_∞ at which the water has come to a stop.

3.3.2.4 Wet dam-break

We now turn to dam-break experiments. The first dam-break experiments under consideration are wet dam-breaks. We introduce the following topography function:

$$Z_{dam}(x) = \frac{1}{2}\cos^2(\pi x). \tag{3.104}$$

We study the following four wet dam-break cases to highlight the behavior of both source terms:

DAM1: $Z(x) = 0$ and $k = 0$;
DAM2: $Z(x) = 0$ and $k = 5$;
DAM3: $Z(x) = Z_{dam}(x)$ and $k = 0$;
DAM4: $Z(x) = Z_{dam}(x)$ and $k = 5$.

The space domain is $[-1, 1]$, and we choose the same initial data for the four experiments, as follows:

$$h(0,x) + Z(x) = \begin{cases} 5 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0, \end{cases} \quad \text{and} \quad q(0,x) = 0.$$

With this initial data, note that the first wet dam-break experiment corresponds to the Riemann problem studied in Section 1.1.2. In particular, we know the exact solution for this experiment.

For the numerical simulations, we prescribe homogeneous Neumann boundary conditions at both boundaries. The results are presented at the final time $t_{end} = 0.1$s. We use 200 discretization cells for the implicit scheme, and present a reference solution computed using the HLL scheme with 2000 cells. Finally, we set $C = 10$.

Flat topography

The results of the wet dam-break experiments with flat topography are displayed on Figure 3.29. The implicit scheme yields a correct approximation of the reference solution. The action of the friction is clearly visible.

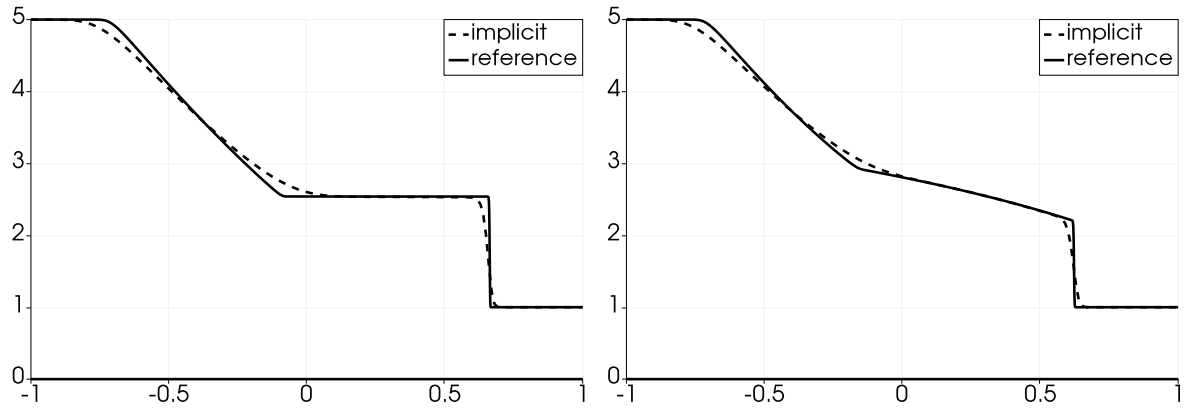

Figure 3.29 – Wet dam-break on a flat topography: free surface observed at the final physical time with the implicit scheme. Left panel: $k = 0$; right panel: $k = 5$.

Non-flat topography

Figure 3.30 depicts the results of the wet dam-break experiments with the topography function (3.104). Once again, we note that the approximate solution provided by the implicit scheme is in good agreement with the reference solution.

Note that there is a lake at rest configuration in the regions untouched by the waves. Indeed, the free surface untouched by the rarefaction wave or the shock wave should remain unperturbed. This means that $h(t,x) + Z(x) = 2$ and $q(t,x) = 0$ for all x inferior to the position of the head of the rarefaction wave, and $h(t,x) + Z(x) = 1$ and $q(t,x) = 0$ for all x superior to the position of the shock wave. This lake at rest behavior is exactly preserved by the explicit and implicit schemes. Therefore, this experiment highlights the interest of a using a well-balanced scheme for such simulations, even if the whole domain does not involve a steady state solution.

3.3.2.5 Dry dam-break

We then focus on dry dam-break experiments. We study four experiments, with the same topography functions and values of the Manning coefficient as in the wet dam-breaks DAM1,

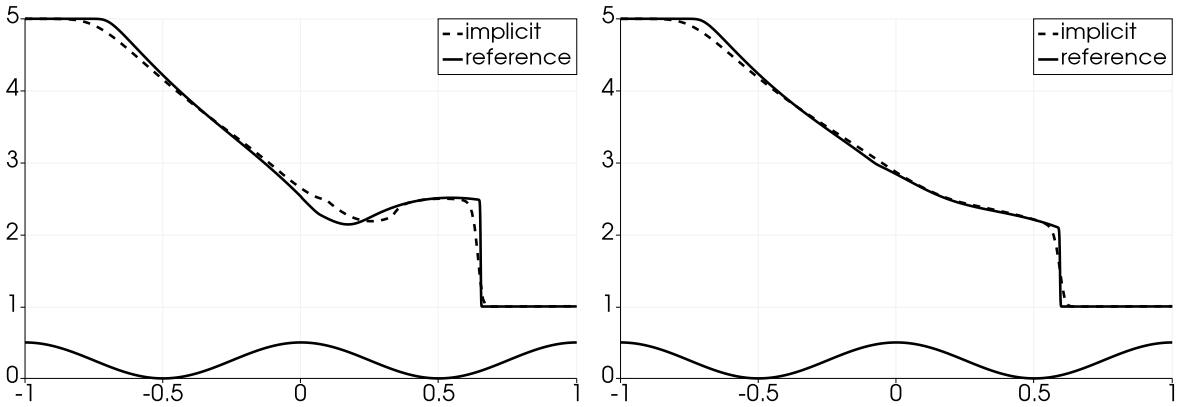

Figure 3.30 – Wet dam-break on a non-flat topography: free surface observed at the final physical time with the implicit scheme. Left panel: $k = 0$; right panel: $k = 5$.

DAM2, DAM3 and DAM4, but whose initial data is now given on $[-1, 1]$ by:

$$h(0,x) = \begin{cases} 1.5 - Z(x) & \text{if } x < 0, \\ 0 & \text{if } x \geq 0, \end{cases} \quad \text{and} \quad q(0,x) = 0.$$

Note that the initial water height vanishes for $x \geq 0$. As a consequence, a dry area is present to the right of the dam. We remark that, in the first case, where $Z(x) = 0$ and $k = 0$, we recover the Riemann problem with a dry area presented in Section 1.1.2.

In the numerical simulations, both boundaries are endowed with homogeneous Neumann boundary conditions. The final physical time is $t_{end} = 0.1$s. The simulation is carried out with 200 cells for the implicit scheme, and with 2000 cells for the reference HLL solution. We still set $C = 10$.

Flat topography

Figure 3.31 displays the results of the dry dam-break simulations with a flat topography. We note that the implicit scheme provides an approximation that is in good agreement with the reference solution. In addition, we here remark the effects of the friction, especially in the shape of the wet/dry front. This font has also been slowed down by the friction.

Non-flat topography

On Figure 3.32, the results of the dry dam-break experiments with a non-flat topography function are depicted. As in the case of a flat topography, the implicit scheme shows good agreement with the reference solution, and the slowing effects of the friction are noted.

3.3.2.6 Dry dam-break with two bumps

This next dry dam-break experiment presents a more complicated topography, which consists in two bumps. The space domain is $[0, 5]$ and we choose to use 10^4 discretization cells with the implicit scheme to have a relevant simulation. The two boundaries at $x = 0$m and

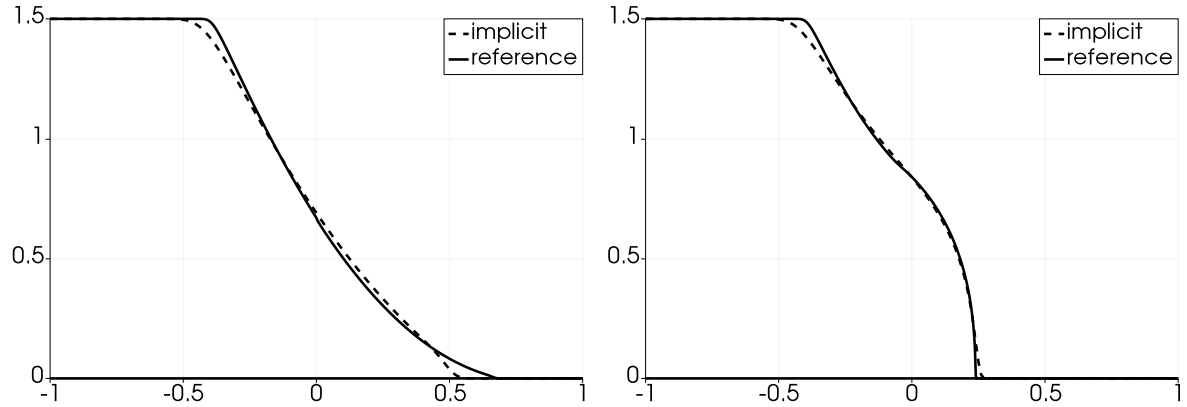

Figure 3.31 – Dry dam-break on a flat topography: free surface observed at the final physical time with the implicit scheme. Left panel: $k = 0$; right panel: $k = 5$.

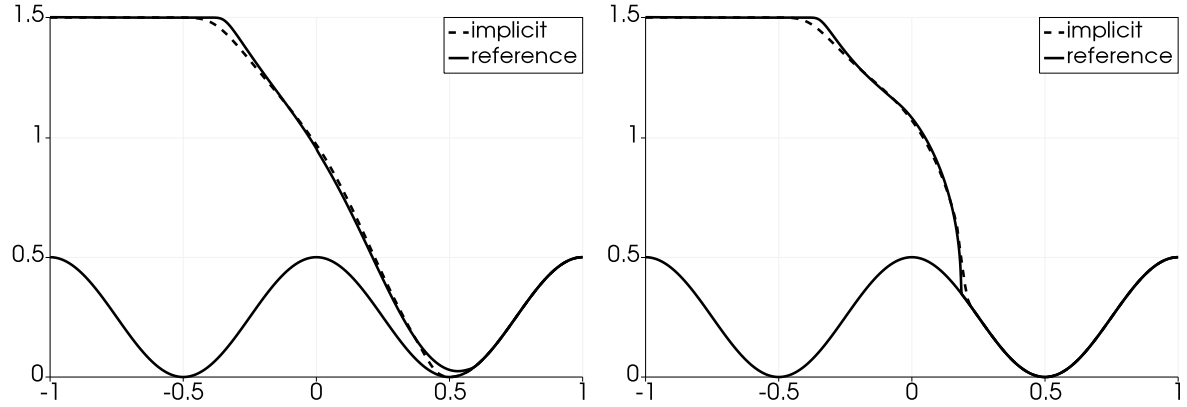

Figure 3.32 – Dry dam-break on a non-flat topography: free surface observed at the final physical time with the implicit scheme. Left panel: $k = 0$; right panel: $k = 5$.

$x = 5$m are solid walls. The topography is defined by

$$Z(x) = \frac{1}{2}\left(1 - \frac{(x-5/2)^2}{1/25}\right)_+ + 2\left(1 - \frac{(x-4)^2}{1/25}\right)_+,$$

and indeed consists in two quadratic bumps, a smaller one followed by a larger one. The dam is located at $x_D = 0.7$m, breaks at $t = 0$s, and contains an initial water height $h_L = 6$m. The domain $x > x_D$ contains no water, i.e. $h_R = 0$. We choose a Manning coefficient k equal to 1. The left panel of Figure 3.33 shows that the initial water height is significantly larger than the bumps. Indeed, we elected to have a larger mass of water whose energy is important enough not to be completely dissipated by the bottom friction. We choose $C = 0.1$, and we depict the results of the implicit scheme on Figure 3.33 and Figure 3.34.

On Figure 3.33, we observe several waves and reflections created by the two bumps. In particular, on the right panel, we remark the characteristic profile of the dry dam-break solution with friction. In addition, the first bump has created a reflection, which is seen on the left panel of Figure 3.34 to be propagating to the left of the domain. Note, on the left panel of Figure 3.34, that a wave reflected from the right bump has appeared and is traveling to the left, while some water has also leaked above this right bump. Finally, on the right panel

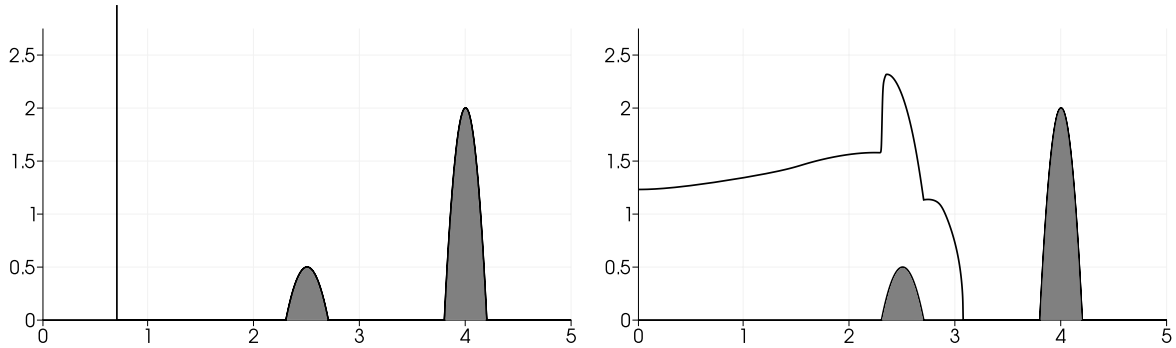

Figure 3.33 – Free surface for the double bump test case at different times. The gray area is the topography. Left panel: solution at $t = 0$s; right panel: solution at $t = 0.38$s.

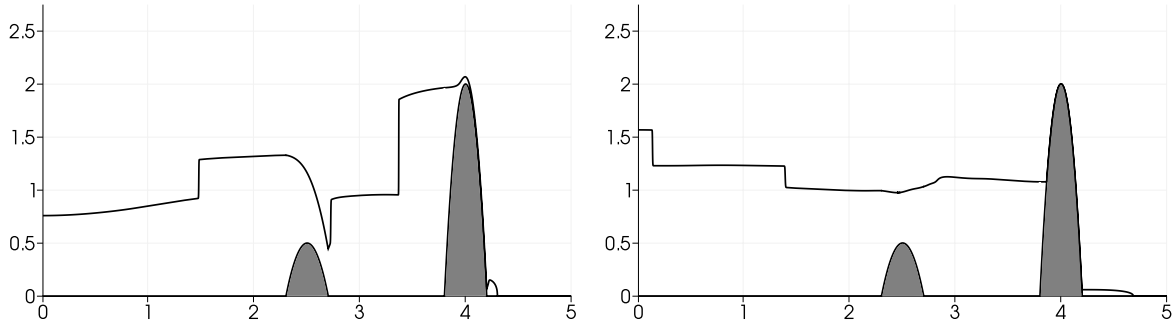

Figure 3.34 – Free surface for the double bump test case at different times. The gray area is the topography. Left panel: solution at $t = 0.74$s; right panel: solution at $t = 1.70$s.

of Figure 3.34, the water level is starting to converge towards a steady state at rest to the left of the right bump. Several waves are still present, interacting among themselves, with the bumps and with the solid walls. The small quantity of water that had leaked is still traveling to the right of the domain.

3.3.2.7 Dry dam-break on a sloping channel: asymptotic solution

The last dam-break experiment we consider comes from [96, 97] and consists in a dry dam-break on a sloping channel. In [96, 97], the author suggests an asymptotic approximation of the water height, valid far enough away from the dam. The goal of this numerical experiment is to compare the implicit scheme with Hunt's asymptotic approximation and with experimental data.

Let us first mention that, in [96, 97], the author does not use Manning's friction law (3.43). Instead, the Darcy-Weisbach friction law is used (see for instance [145, 57] and references therein), and the friction source term $S^f(W)$ is given by:

$$S^f(W) = -\frac{f}{8}\frac{q|q|}{h^2},$$

where f is the Darcy-Weisbach friction coefficient. Note that this friction source term can be rewritten under the more general form $S^f(W) = -kq|q|h^{-\eta}$, with $k = f/8$ and $\eta = 2$. As a consequence, we are able to adapt the implicit scheme to this new source term simply by

taking Manning's coefficient k equal to $f/8$ and by setting $\eta = 2$ instead of $\eta = 7/3$.

The initial data of the experiment is presented on Figure 3.35. On an infinitely long sloping channel, a reservoir of length L holds water, up to a dam of height H, which breaks at time $t = 0$. Note that the slope of the channel is H/L: therefore, the topography function is given by

$$Z(x) = -\frac{H}{L}x.$$

As a consequence, the initial data is given as follows:

$$h(0,x) = \begin{cases} \max(0, -Z(x)) & \text{if } x < L, \\ 0 & \text{if } x \geq L, \end{cases} \quad \text{and} \quad q(0,x) = 0,$$

where the x-axis, as displayed on Figure 3.35, follows the slope of the channel.

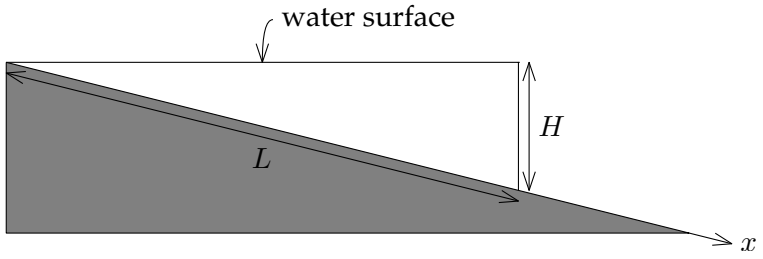

Figure 3.35 – Initial water height for the dry dam-break on a sloping channel.

Then, following Hunt [96, 97], we introduce the uniform flow velocity at a normal depth of H, denoted by U. After having introduced the dimensionless variables $u_* = u/U$ and $h_* = h/H$, Hunt suggests using the *kinematic-wave approximation*, i.e. $u_*^2 = h_*$, where a sharp shock is observed at the wet/dry front. Under this approximation, Hunt derives the following two quantities:

- the shock amplitude $h_s(t) = H\left(\dfrac{L}{Ut}\right)^{2/3}$;
- the shock position $x_s(t) = \dfrac{3L}{2}\left(\dfrac{Ut}{L}\right)^{2/3}$.

In addition, under this kinematic-wave approximation, the shallow-water system can be solved, to yield the following *outer solution*:

$$h_o(t,x) = H\left(\frac{2}{3}\frac{x}{Ut}\right)^2.$$

This formula is valid far from the dam: Hunt suggests using it for $x/L > 5$.

Afterwards, Hunt derives the *inner solution*, valid near the wet/dry front. To that end, he introduces the actual position of the wet/dry front, given by:

$$x_0(t) = L\left(\frac{3}{2}\left(\frac{Ut}{L}\right)^{2/3} + \frac{1}{2}\left(\frac{L}{Ut}\right)^{2/3}\right).$$

Then, the inner solution h_i is given as the unique solution within $[0, h_s(t))$ of the following

nonlinear equation:

$$\frac{h_i(t,x)}{h_s(t)} + \ln\left(1 - \frac{h_i(t,x)}{h_s(t)}\right) + \frac{1}{2} + \frac{x_s(t) - x}{L}\frac{H}{h_s(t)} = 0.$$

Finally, Hunt combines the outer solution and the inner solution to get the *composite solution*, as follows:

$$h_c(t,x) = \begin{cases} h_o(t,x) + h_i(t,x) - h_s(t) & \text{if } 0 \leq x \leq x_s(t), \\ h_i(t,x) & \text{if } x_s(t) \leq x \leq x_0(t), \\ 0 & \text{if } x > x_0(t). \end{cases}$$

This composite solution, according to Hunt, is valid for $x/L > 5$.

Our goal is now to compare the results of the implicit scheme with Hunt's composite solution. To that end, we propose two checks: the water height with respect to the time at a fixed position, far enough away from the dam, and the water height with respect to the position at a fixed time. In both cases, we take homogeneous Neumann boundary conditions and we use 400 discretization cells for the implicit scheme. In addition, the constants are chosen according to Table 3.25, where we note that the slope of the channel H/L is very mild (about $2.5°$). For both schemes, the cutoff constant C is taken equal to 1.

H	A	U	f
0.04 m	0.932 m	1.195 m.s^{-1}	0.0932

Table 3.25 – Values of the constants for the dry dam-break on a sloping channel.

Water height with respect to the time at a fixed position

We first follow Hunt [96, 97] and we set $x/L = 5.7$ to observe the time evolution of the solution over the time domain $[0, 14]$. Experimental values for the water height are given in [96]. Note that the position $x/L = 5.7$ is well within the domain of validity of Hunt's asymptotic approximation.

The result of the implicit scheme, as well as the experimental points and Hunt's solution, are displayed on Figure 3.36, where we note that both the implicit scheme and Hunt's approximation have a correct shape compared to the experimental points. In addition, for both approximate solutions, the water arrives at the position $x/L = 5.7$ at roughly the same time, which corresponds to the experimental result. We also note that the implicit scheme provides a correct approximation of the maximum experimental water height, while Hunt's solution presents an overshoot. This overshoot behavior was already documented in [97]

Water height with respect to the position at a fixed time

We then compare Hunt's composite solution with the implicit scheme at the fixed time $t = 6$ over the space domain $[0, 7]$. Hunt's solution should be valid for $x/L > 5$, i.e. $x > 4.66$m.

The water height produced by the implicit scheme and Hunt's composite solution are presented on Figure 3.37. On this figure, we first note that the wet/dry front is located at roughly the same position for both approximations. This position is given by $x_0(6) \simeq 5.57$. We remark that $x_0(6) > 4.66$, and therefore that the wet/dry front is located within the domain of validity of Hunt's approximation. We also note that the maximum of Hunt's composite solution is once again higher than the maximum of the water height approximated by the implicit scheme.

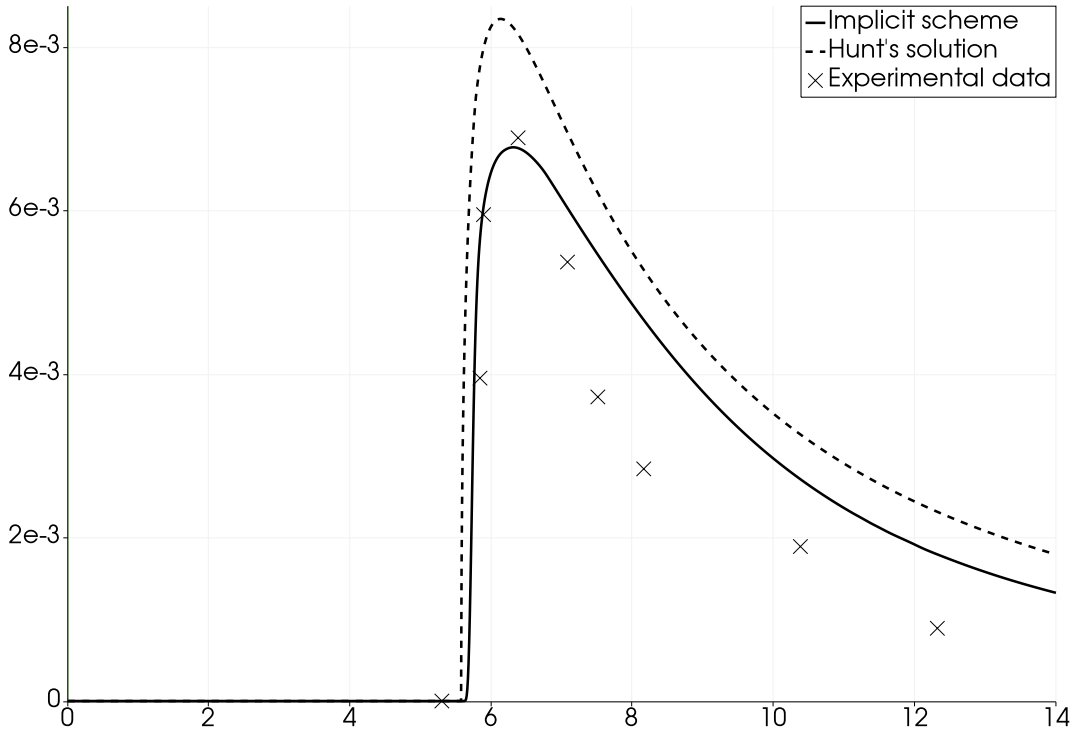

Figure 3.36 – Water height with respect to the time at the position $x/L = 5.7$ for the dry dam-break on a sloping channel. Comparison between the experimental data (crosses), Hunt's composite solution (dashed line), and the result of the implicit scheme (solid line).

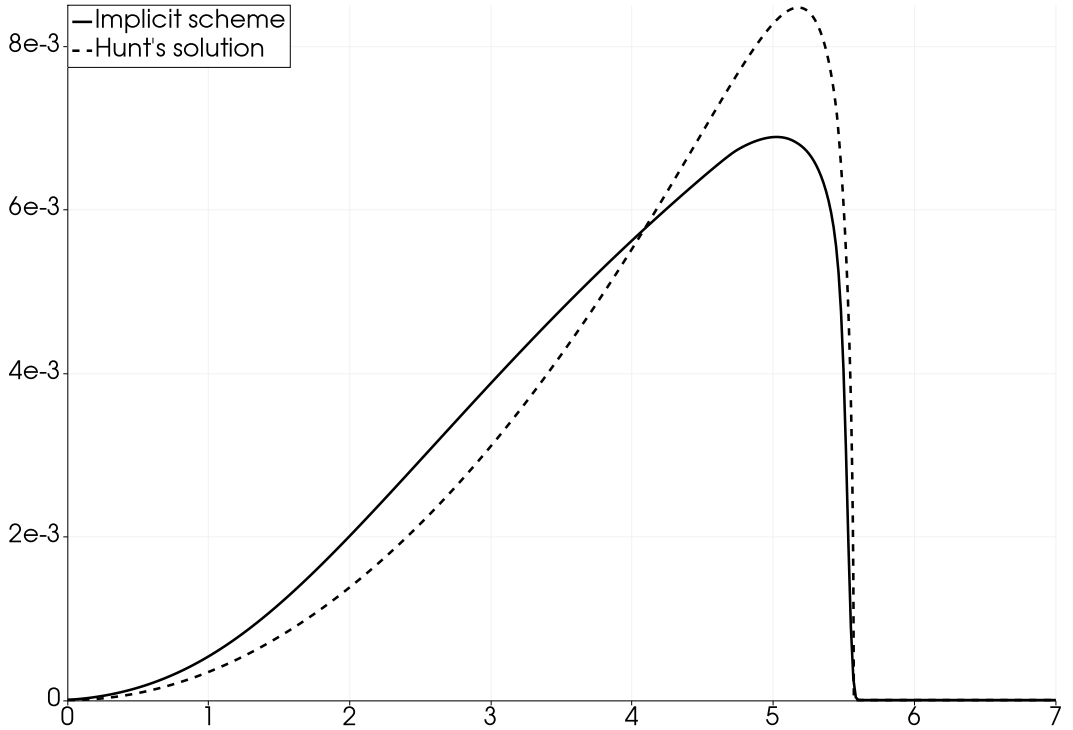

Figure 3.37 – Water height with respect to the position at the time $t = 6$s for the dry dam-break on a sloping channel. Comparison between Hunt's composite solution (dashed line) and the result of the implicit scheme (solid line).

4

Two-dimensional and high-order extensions

In the previous chapter, we have suggested a well-balanced scheme for the shallow-water equations with topography and Manning friction (1.1) in one space dimension. This scheme is based on a suitable approximate Riemann solver and a semi-implicitation of the friction contribution. In addition, it has been built to be consistent, non-negativity preserving, and able to deal with interfaces between wet and dry areas. The goal of this chapter is to propose two extensions of this scheme, a first one to consider two-dimensional geometries, and a second one to provide a high-order accuracy. A requirement for these extensions is to recover as many properties of the one-dimensional first-order scheme as possible.

We begin with a brief presentation of the shallow-water equations in two space dimensions. The homogeneous 2D shallow-water equations have been extensively studied (see [3, 4, 156, 112, 36, 58] for instance). With the source terms, the balance law is governed by the following equations:

$$\begin{cases} \partial_t h + \boldsymbol{\nabla} \cdot \boldsymbol{q} = 0, \\ \partial_t \boldsymbol{q} + \boldsymbol{\nabla} \cdot \left(\dfrac{\boldsymbol{q} \otimes \boldsymbol{q}}{h} + \dfrac{1}{2} g h^2 \mathbb{I}_2 \right) = -gh \boldsymbol{\nabla} Z - k \boldsymbol{q} \|\boldsymbol{q}\| h^{-\eta}, \end{cases} \tag{4.1}$$

where $\mathbb{I}_2$ represents the identity matrix of $\mathcal{M}_{2,2}(\mathbb{R})$ and the notation $\|.\|$ stands for the euclidean norm of a vector, defined for $\boldsymbol{X} = {}^t(X_1, X_2) \in \mathbb{R}^2$ by $\|\boldsymbol{X}\| = \sqrt{X_1^2 + X_2^2}$. Lastly, the symbol $\otimes$ represents the tensor product of two vectors, which is a matrix defined as follows:

$$\forall\, \boldsymbol{X} = {}^t(X_1, X_2) \in \mathbb{R}^2, \forall\, \boldsymbol{Y} = {}^t(Y_1, Y_2) \in \mathbb{R}^2, \ \boldsymbol{X} \otimes \boldsymbol{Y} = \begin{pmatrix} X_1 Y_1 & X_1 Y_2 \\ X_2 Y_1 & X_2 Y_2 \end{pmatrix}.$$

The equations (4.1) can be rewritten under the following condensed form (2.67) of a 2D bal-

ance law:

$$\partial_t W + \boldsymbol{\nabla} \cdot \boldsymbol{F}(W) = \mathfrak{S}(W),$$

where we have set $\boldsymbol{q} = {}^t(q_x, q_y)$, and:

$$W = \begin{pmatrix} h \\ q_x \\ q_y \end{pmatrix} \quad ; \quad \boldsymbol{F}(W) = \begin{pmatrix} q_x & q_y \\ \dfrac{q_x^2}{h} + \dfrac{1}{2}gh^2 & \dfrac{q_x q_y}{h} \\ \dfrac{q_x q_y}{h} & \dfrac{q_y^2}{h} + \dfrac{1}{2}gh^2 \end{pmatrix} \quad ; \quad \mathfrak{S}(W) = \begin{pmatrix} 0 \\ -gh\partial_x Z - \dfrac{kq_x\|\boldsymbol{q}\|}{h^\eta} \\ -gh\partial_y Z - \dfrac{kq_y\|\boldsymbol{q}\|}{h^\eta} \end{pmatrix}.$$

The vector W must be taken in the following 2D admissible states space:

$$\Omega = \left\{ W = {}^t(h, q_x, q_y) \in \mathbb{R}^3 \, ; \, h \geq 0, \, q_x \in \mathbb{R}, q_y \in \mathbb{R} \right\}.$$

Finally, we turn to describing some steady state solutions for the 2D shallow-water model with topography and/or friction. The steady state solutions are as usual defined by making the time derivatives vanish, as follows:

$$\begin{cases} \boldsymbol{\nabla} \cdot \boldsymbol{q} = 0, \\ \boldsymbol{\nabla} \cdot \left(\dfrac{\boldsymbol{q} \otimes \boldsymbol{q}}{h} + \dfrac{1}{2}gh^2 \mathbb{I}_2 \right) + gh\boldsymbol{\nabla} Z + k\boldsymbol{q}\|\boldsymbol{q}\|h^{-\eta} = \boldsymbol{0}. \end{cases}$$

We therefore no longer have a uniform discharge as soon as a steady state is involved. Instead, the discharge field is divergence free, i.e. the divergence of the discharge vanishes, as follows:

$$\boldsymbol{\nabla} \cdot \boldsymbol{q} = 0. \tag{4.2}$$

Hence, studying the steady state solutions of the shallow-water equations in two dimensions is much harder than in one dimension.

However, several specific steady states can still be recovered. For instance, if we take a smooth steady state at rest (i.e. $\boldsymbol{q} = 0$), then we once again get the lake at rest steady state, defined as usual by $h + Z = \text{cst}$. In addition, the 1D moving steady states can obviously be recovered, by taking, with $q_0 \neq 0$, $\boldsymbol{q} = {}^t(q_0, 0)$ or $\boldsymbol{q} = {}^t(0, q_0)$, as well as a one-dimensional water flow. As a consequence, our goal is not to preserve all the steady state solutions of the 2D shallow-water equations, but rather to preserve the pseudo-1D steady states, i.e. those along the x-axis and the y-axis, as well as the lake at rest in every direction. This restricted well-balance property is called the *well-balance by direction*; a formal definition will be given later. The two-dimensional extension of the 1D scheme developed in Chapter 3 is therefore the focus of Section 4.1. A Cartesian mesh is considered in order to allow the scheme to be well-balanced by direction. This 2D extension is performed by suggesting a convex combination involving the 1D scheme derived in the previous chapter. We then state several properties satisfied by the 2D scheme.

This 2D scheme is then supplemented by a high-order extension in Section 4.2. This high-order extension consists in providing a polynomial reconstruction of the variables in each cell.

Afterwards, the high-order strategy from Section 2.4 is used to derive a finite volume scheme for a 2D balance law with a high order of spatial accuracy. The time accuracy of the scheme is improved by Strong Stability-Preserving Runge-Kutta (SSPRK) methods. Then, we apply the MOOD method to the current case, in order to recover the robustness of the scheme and to eliminate non-physical oscillations. This MOOD method is supplemented by a procedure to recover the well-balance of the scheme.

The Fortran implementation of this scheme is then discussed in Section 4.3. Namely, speedup and efficiency graphs for the OpenMP parallelization are provided.

Finally, we propose in Section 4.4 several numerical experiments to assess the properties of the scheme. First, the well-balance property is tested with the 2D first-order and high-order schemes. Then, we check the order of accuracy of the scheme. Finally, several validation experiments are suggested, and real-world simulations are presented.

4.1 Two-dimensional extension on a Cartesian grid

The goal of this section is to derive a two-dimensional scheme for the shallow-water equations on a Cartesian grid. This 2D scheme is based on the 1D scheme developed in the previous chapter. The choice of a Cartesian grid is motivated by the fact that we want the scheme to be able to preserve steady state solutions along the x-axis and the y-axis.

We now introduce the discretization of the space domain $\mathbb{R}^2$, which consists in a Cartesian mesh of uniform cells $c_{i,j}$, defined by:

$$c_{i,j} = \left(x_{i,j} - \frac{\Delta x}{2}, x_{i,j} + \frac{\Delta x}{2} \right) \times \left(y_{i,j} - \frac{\Delta y}{2}, y_{i,j} + \frac{\Delta y}{2} \right),$$

where $(x_{i,j}, y_{i,j})$ is the center of the cell $c_{i,j}$. Thus, all cells are rectangles of length Δx and width Δy. From now on, we denote by $|c_{i,j}| = \Delta x \Delta y$ the area of the cell $c_{i,j}$. The piecewise constant approximate solution, within the cell $c_{i,j}$ and at time t^n, will henceforth be denoted by $W_{i,j}^n$.

In order to propose a way to update the piecewise constant approximate solution in time, we suggest a two-step scheme. In Section 4.1.1, we introduce the proposed discretization of the equations, which consists in a two-step semi-implicit scheme. Its first step is devoted to the flux and the topography source term, and its second step uses a splitting method to take the friction contribution into account. Then, in Section 4.1.2, we state the properties of this 2D scheme. To that end, we rewrite the 2D scheme as a convex combination of 1D schemes in the spirit of Section 2.3.2.

4.1.1 Derivation of a two-dimensional scheme

In order to derive a numerical scheme for the two-dimensional shallow-water equations (4.1), we first cast this system into the following form:

$$\partial_t W + \partial_x F(W) + \partial_y G(W) = S^t(W) + S^f(W), \tag{4.3}$$

where the fluxes F in the x-direction and G in the y-direction are defined by

$$F(W) = \begin{pmatrix} q_x \\ \dfrac{q_x^2}{h} + \dfrac{1}{2}gh^2 \\ \dfrac{q_x q_y}{h} \end{pmatrix} \quad \text{and} \quad G(W) = \begin{pmatrix} q_y \\ \dfrac{q_x q_y}{h} \\ \dfrac{q_y^2}{h} + \dfrac{1}{2}gh^2 \end{pmatrix},$$

while the topography and friction source terms S^t and S^f are given by

$$S^t(W) = \begin{pmatrix} 0 \\ -gh\partial_x Z \\ -gh\partial_y Z \end{pmatrix} \quad \text{and} \quad S^f(W) = \begin{pmatrix} 0 \\ -kq_x\|\boldsymbol{q}\|h^{-\eta} \\ -kq_y\|\boldsymbol{q}\|h^{-\eta} \end{pmatrix}. \tag{4.4}$$

Following the 1D case, we suggest a two-step semi-implicit scheme to approximate solutions of (4.3). The first step of the scheme is devoted to the flux and the topography, while the second step concerns the implicit treatment of the friction source term.

The first step requires an approximation of the following system:

$$\partial_t W + \partial_x F(W) + \partial_y G(W) = S^t(W). \tag{4.5}$$

Following the 2D scheme (2.47), we introduce two numerical flux functions. The numerical flux function in the x-direction, denoted by $\mathcal{F}$, has already been introduced in the previous chapter. It is given by (3.84) in one dimension. In the two-dimensional case, we set

$$\mathcal{F} = \begin{pmatrix} \mathcal{F}^h \\ \mathcal{F}^{q_x} \\ \mathcal{F}^h v_y^t \end{pmatrix},$$

where the functions $\mathcal{F}^h$ and $\mathcal{F}^{q_x}$ are defined as follows:

$$\mathcal{F}^h(W_L, W_R, Z_L, Z_R, \Delta x) = \frac{1}{2}\big(F^h(W_L) + F^h(W_R)\big) + \frac{\lambda_L^x}{2}(h_L^* - h_L) + \frac{\lambda_R^x}{2}(h_R^* - h_R),$$
$$\mathcal{F}^{q_x}(W_L, W_R, Z_L, Z_R, \Delta x) = \frac{1}{2}\big(F^{q_x}(W_L) + F^{q_x}(W_R)\big) + \frac{\lambda_L^x}{2}(q_x^* - (q_x)_L) + \frac{\lambda_R^x}{2}(q_x^* - (q_x)_R),$$

where we have set $F = {}^t(F^h, F^{q_x}, F^{q_y})$, and where the intermediate states are given by:

$$\begin{aligned} h_L^* &= h_L^*(h_L, h_R, (q_x)_L, (q_x)_R, Z_L, Z_R, \Delta x), \\ h_R^* &= h_R^*(h_L, h_R, (q_x)_L, (q_x)_R, Z_L, Z_R, \Delta x), \\ q_x^* &= q^*(h_L, h_R, (q_x)_L, (q_x)_R, Z_L, Z_R, \Delta x), \end{aligned} \tag{4.6}$$

with h_L^*, h_R^* and q^* given by (3.81). Moreover, the characteristic velocities are defined by (3.6), as follows:

$$\lambda_L^x = \min(-|(v_x^n)_L| - c_L,\ -|(v_x^n)_R| - c_R,\ -\varepsilon_\lambda),$$
$$\lambda_R^x = \max(|(v_x^n)_L| + c_L,\ |(v_x^n)_R| + c_R,\ \varepsilon_\lambda),$$

where $(v_x^n)_L$ and $(v_x^n)_R$ are the normal velocities, defined by $v_x^n = q_x/h$. In addition, the tangential velocity v_y^t satisfies

$$v_y^t = \begin{cases} \dfrac{(q_y)_L}{h_L} & \text{if } \mathcal{F}^h > 0, \\ \dfrac{(q_y)_R}{h_R} & \text{if } \mathcal{F}^h < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the numerical flux $\mathcal{F}$ is extended to the y-direction, to get a new numerical flux function, denoted by $\mathcal{G}$. This function is obtained by arguing rotational invariance properties (see [80]). We set $\mathcal{G} = {}^t(\mathcal{G}^h, \mathcal{G}^h v_x^t, \mathcal{G}^{q_y})$, where $\mathcal{G}^h$ and $\mathcal{G}^{q_y}$ are defined by:

$$\begin{aligned} \mathcal{G}^h(W_L, W_R, Z_L, Z_R, \Delta y) &= \frac{1}{2}\big(G^h(W_L) + G^h(W_R)\big) + \frac{\lambda_L^y}{2}(h_L^* - h_L) + \frac{\lambda_R^y}{2}(h_R^* - h_R), \\ \mathcal{G}^{q_y}(W_L, W_R, Z_L, Z_R, \Delta y) &= \frac{1}{2}\big(G^{q_y}(W_L) + G^{q_y}(W_R)\big) + \frac{\lambda_L^y}{2}\big(q_y^* - (q_y)_L\big) + \frac{\lambda_R^y}{2}\big(q_y^* - (q_y)_R\big), \end{aligned}$$

where $G = {}^t(G^h, G^{q_x}, G^{q_y})$, and where we define the intermediate states as follows:

$$\begin{aligned} h_L^* &= h_L^*(h_L, h_R, (q_y)_L, (q_y)_R, Z_L, Z_R, \Delta y), \\ h_R^* &= h_R^*(h_L, h_R, (q_y)_L, (q_y)_R, Z_L, Z_R, \Delta y), \\ q_y^* &= q^*(h_L, h_R, (q_y)_L, (q_y)_R, Z_L, Z_R, \Delta y). \end{aligned} \tag{4.7}$$

In addition, the characteristic velocities are defined by:

$$\begin{aligned} \lambda_L^y &= \min\big(-|(v_y^n)_L| - c_L,\ -|(v_y^n)_R| - c_R,\ -\varepsilon_\lambda\big), \\ \lambda_R^y &= \max\big(|(v_y^n)_L| + c_L,\ |(v_y^n)_R| + c_R,\ \varepsilon_\lambda\big), \end{aligned}$$

where the normal velocity v_y^n is given by $v_y^n = q_y/h$. Finally, we define the tangential velocity v_x^t as follows:

$$v_x^t = \begin{cases} \dfrac{(q_x)_L}{h_L} & \text{if } \mathcal{G}^h > 0, \\ \dfrac{(q_x)_R}{h_R} & \text{if } \mathcal{G}^h < 0, \\ 0 & \text{otherwise.} \end{cases}$$

As a consequence, the first step of the 2D scheme, devoted to the approximation of (4.5), reads as follows:

$$W_{i,j}^{n+\frac{1}{2}} = W_{i,j}^n - \frac{\Delta t}{\Delta x}\Big(\mathcal{F}_{i+\frac{1}{2},j}^n - \mathcal{F}_{i-\frac{1}{2},j}^n\Big) - \frac{\Delta t}{\Delta y}\Big(\mathcal{G}_{i,j+\frac{1}{2}}^n - \mathcal{G}_{i,j-\frac{1}{2}}^n\Big) + \Delta t \begin{pmatrix} 0 \\ (\boldsymbol{S}_{\text{WB}}^t)_{i,j}^n \end{pmatrix}, \tag{4.8}$$

where we have set the following shorter notations for the numerical fluxes:

$$\mathcal{F}_{i+\frac{1}{2},j}^n = \mathcal{F}(W_{i,j}^n, W_{i+1,j}^n, Z_{i,j}, Z_{i+1,j}, \Delta x) \quad \text{and} \quad \mathcal{G}_{i,j+\frac{1}{2}}^n = \mathcal{G}(W_{i,j}^n, W_{i,j+1}^n, Z_{i,j}, Z_{i,j+1}, \Delta y).$$

In addition, the numerical source term $(\boldsymbol{S}^t_{\text{WB}})^n_{i,j}$ is defined by

$$(\boldsymbol{S}^t_{\text{WB}})^n_{i,j} = \frac{1}{2}\begin{pmatrix} (\overline{S}^t_x)^n_{i-\frac{1}{2},j} + (\overline{S}^t_x)^n_{i+\frac{1}{2},j} \\ (\overline{S}^t_y)^n_{i,j-\frac{1}{2}} + (\overline{S}^t_y)^n_{i,j+\frac{1}{2}} \end{pmatrix}, \tag{4.9}$$

where the subscript helps identify the well-balanced scheme. In (4.9), the approximate topography source terms $\overline{S}^t_x$ and $\overline{S}^t_y$ are given as follows:

$$\begin{aligned} (\overline{S}^t_x)^n_{i+\frac{1}{2},j} &= \overline{S}^t\Big(h^n_{i,j}, h^n_{i+1,j}, (q_x)^n_{i,j}, (q_x)^n_{i+1,j}, Z_{i,j}, Z_{i+1,j}, \Delta x\Big), \\ (\overline{S}^t_y)^n_{i,j+\frac{1}{2}} &= \overline{S}^t\Big(h^n_{i,j}, h^n_{i,j+1}, (q_y)^n_{i,j}, (q_y)^n_{i,j+1}, Z_{i,j}, Z_{i,j+1}, \Delta y\Big), \end{aligned}$$

where $\overline{S}^t$ is the approximate topography source term, given by (3.75).

Now, the second step of the two-step scheme is devoted to the contribution of the friction source term. Hence, the following system of ordinary differential equations is considered:

$$\partial_t W = S^f(W).$$

As a consequence, the second step consists in solving the following initial value problem:

$$\begin{cases} \dfrac{dh}{dt} = 0, \\ \dfrac{d\boldsymbol{q}}{dt} = -k\,\boldsymbol{q}\|\boldsymbol{q}\|h^{-\eta}, \end{cases} \text{ with initial data } \begin{cases} h(0) = h^{n+\frac{1}{2}}_{i,j}, \\ \boldsymbol{q}(0) = \boldsymbol{q}^{n+\frac{1}{2}}_{i,j}. \end{cases} \tag{4.10}$$

Straightforward computations show that this initial value problem admits a unique analytical solution. This solution is given for all $t \in [0, \Delta t]$ by:

$$\begin{cases} h(t) = h(0), \\ \boldsymbol{q}(t) = \dfrac{h(0)^\eta\,\boldsymbol{q}(0)}{h(0)^\eta + k\,t\,\|\boldsymbol{q}(0)\|}. \end{cases} \tag{4.11}$$

Therefore, the solution to the initial value problem (4.10) reads as follows:

$$\begin{cases} h^{n+1}_{i,j} = h^{n+\frac{1}{2}}_{i,j}, \\ \boldsymbol{q}^{n+1}_{i,j} = \dfrac{\left(h^{n+1}_{i,j}\right)^\eta \boldsymbol{q}^{n+\frac{1}{2}}_{i,j}}{\left(h^{n+1}_{i,j}\right)^\eta + k\,\Delta t\left\|\boldsymbol{q}^{n+\frac{1}{2}}_{i,j}\right\|}. \end{cases} \tag{4.12}$$

However, merely plugging this analytical solution as the second step of the scheme does not allow the recovery of the well-balance property. According to the 1D case, we replace the expression $(h^{n+1}_{i,j})^\eta$ with an average in the formula (4.12) that yields $\boldsymbol{q}^{n+1}_{i,j}$. For the x-discharge, we take

$$(q_x)^{n+1}_{i,j} = \frac{(\overline{h^\eta_x})^{n+1}_{i,j}\,(q_x)^{n+\frac{1}{2}}_{i,j}}{(\overline{h^\eta_x})^{n+1}_{i,j} + k\,\Delta t\left\|\boldsymbol{q}^{n+\frac{1}{2}}_{i,j}\right\|}, \tag{4.13a}$$

while we set, for the y-discharge:

$$(q_y)_{i,j}^{n+1} = \frac{(\overline{h_y^\eta})_{i,j}^{n+1}\,(q_y)_{i,j}^{n+\frac{1}{2}}}{(\overline{h_y^\eta})_{i,j}^{n+1} + k\,\Delta t\left\|\boldsymbol{q}_{i,j}^{n+\frac{1}{2}}\right\|}. \tag{4.13b}$$

In (4.13), we have introduced two averages of $(h_{i,j}^{n+1})^\eta$. The average in the x-direction is denoted by $(\overline{h_x^\eta})_{i,j}^{n+1}$, while the average in the y-direction is denoted by $(\overline{h_y^\eta})_{i,j}^{n+1}$. These averages are given by the 1D formula (3.101) evaluated in both space directions, as follows:

$$(\overline{h_x^\eta})_{i,j}^{n+1} = \frac{2k(\mu_x)_{i,j}^{n+\frac{1}{2}}\Delta x}{k(\mu_x)_{i,j}^n\Delta x\left(\beta_{i-\frac{1}{2},j}^{n+1}+\beta_{i+\frac{1}{2},j}^{n+1}\right)-\left(\gamma_{i-\frac{1}{2},j}^{n+1}+\gamma_{i+\frac{1}{2},j}^{n+1}\right)} + k\,\Delta t\,(\mu_x)_{i,j}^{n+\frac{1}{2}}(q_x)_{i,j}^n,$$

$$(\overline{h_y^\eta})_{i,j}^{n+1} = \frac{2k(\mu_y)_{i,j}^{n+\frac{1}{2}}\Delta y}{k(\mu_y)_{i,j}^n\Delta x\left(\beta_{i,j-\frac{1}{2}}^{n+1}+\beta_{i,j+\frac{1}{2}}^{n+1}\right)-\left(\gamma_{i,j-\frac{1}{2}}^{n+1}+\gamma_{i,j+\frac{1}{2}}^{n+1}\right)} + k\,\Delta t\,(\mu_y)_{i,j}^{n+\frac{1}{2}}(q_y)_{i,j}^n,$$

where ${}^t(\mu_x,\mu_y) = {}^t(\mathrm{sgn}(q_x),\mathrm{sgn}(q_y))$, and where $\beta_{i+\frac{1}{2},j}^{n+1}$, $\gamma_{i+\frac{1}{2},j}^{n+1}$, $\beta_{i,j+\frac{1}{2}}^{n+1}$ and $\gamma_{i,j+\frac{1}{2}}^{n+1}$ are given with clear notations by (3.102).

The full 2D scheme has thus been derived. It is a two-step scheme, given by (4.8) and (4.13). The next section is devoted to exhibiting the properties satisfied by this scheme.

4.1.2 Properties verified by the scheme

In order to highlight the properties satisfied by the 2D scheme (4.8) – (4.13), we first rewrite its first step (4.8) under the form of a convex combination of one-dimensional schemes.

Proposition 4.1. *The first step (4.8) of the two-step scheme can be rewritten as follows:*

$$W_{i,j}^{n+\frac{1}{2}} = \frac{1}{4}\left(\mathcal{W}_{i+\frac{1}{2},j}^{n+\frac{1}{2}} + \mathcal{W}_{i-\frac{1}{2},j}^{n+\frac{1}{2}} + \mathcal{W}_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} + \mathcal{W}_{i,j-\frac{1}{2}}^{n+\frac{1}{2}}\right), \tag{4.14}$$

where we have set

$$\begin{aligned}
\mathcal{W}_{i-\frac{1}{2},j}^{n+\frac{1}{2}} &= W_{i,j}^n - \frac{4\Delta t}{\Delta x}\Big(\mathcal{F}\big(W_{i,j}^n, W_{i,j}^n\big) - \mathcal{F}\big(W_{i,j}^n, W_{i-1,j}^n\big)\Big) + 2\Delta t\begin{pmatrix}0\\(\overline{S}_x^t)_{i-\frac{1}{2},j}^n + (\overline{S}_x^t)_{i,j}^n\\0\end{pmatrix},\\
\mathcal{W}_{i+\frac{1}{2},j}^{n+\frac{1}{2}} &= W_{i,j}^n - \frac{4\Delta t}{\Delta x}\Big(\mathcal{F}\big(W_{i,j}^n, W_{i+1,j}^n\big) - \mathcal{F}\big(W_{i,j}^n, W_{i,j}^n\big)\Big) + 2\Delta t\begin{pmatrix}0\\(\overline{S}_x^t)_{i,j}^n + (\overline{S}_x^t)_{i+\frac{1}{2},j}^n\\0\end{pmatrix},\\
\mathcal{W}_{i,j-\frac{1}{2}}^{n+\frac{1}{2}} &= W_{i,j}^n - \frac{4\Delta t}{\Delta y}\Big(\mathcal{G}\big(W_{i,j}^n, W_{i,j}^n\big) - \mathcal{G}\big(W_{i,j}^n, W_{i,j-1}^n\big)\Big) + 2\Delta t\begin{pmatrix}0\\0\\(\overline{S}_y^t)_{i,j-\frac{1}{2}}^n + (\overline{S}_y^t)_{i,j}^n\end{pmatrix},\\
\mathcal{W}_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} &= W_{i,j}^n - \frac{4\Delta t}{\Delta y}\Big(\mathcal{G}\big(W_{i,j}^n, W_{i,j+1}^n\big) - \mathcal{G}\big(W_{i,j}^n, W_{i,j}^n\big)\Big) + 2\Delta t\begin{pmatrix}0\\0\\(\overline{S}_y^t)_{i,j}^n + (\overline{S}_y^t)_{i,j+\frac{1}{2}}^n\end{pmatrix},
\end{aligned} \tag{4.15}$$

where the quantities $(\overline{S}_x^t)_{i,j}^n$ *and* $(\overline{S}_y^t)_{i,j}^n$ *are defined by:*

$$(\overline{S}_x^t)_{i,j}^n = \overline{S}^t\Big(h_{i,j}^n, h_{i,j}^n, (q_x)_{i,j}^n, (q_x)_{i,j}^n, Z_{i,j}, Z_{i,j}, \Delta x\Big),$$
$$(\overline{S}_y^t)_{i,j}^n = \overline{S}^t\Big(h_{i,j}^n, h_{i,j}^n, (q_y)_{i,j}^n, (q_y)_{i,j}^n, Z_{i,j}, Z_{i,j}, \Delta y\Big).$$

Proof. For the numerical fluxes, the proof of is contained within the proof of Proposition 2.1. Concerning the approximate source terms, the expression (3.75) of $\overline{S}^t$ implies that $(\overline{S}_x^t)_{i,j}^n$ and $(\overline{S}_y^t)_{i,j}^n$ vanish. As a consequence, (4.14) immediately yields the expression (4.8) of the first step. The proof is thus achieved. □

Note that (4.15) represents a collection of four one-dimensional schemes, and that (4.14) is nothing but a convex combination of these schemes. These 1D schemes are written under the form (3.83). As a consequence, each of these schemes enjoy the same properties as the actual 1D scheme (3.83), stated in Theorem 3.13. However, the scheme will not be able to preserve all the 2D steady state solutions. In order to state the weaker well-balance property satisfied by the 2D scheme, we introduce the property of *well-balance by direction* in the following definition.

Definition 4.2. The vector $(W_{i,j}^n)_{(i,j)\in\mathbb{Z}^2}$ is said to define a steady state in the x-direction if:

- $\forall (i,j) \in \mathbb{Z}^2$, $W_{i,j+1}^n = W_{i,j}^n$;
- $\forall (i,j) \in \mathbb{Z}^2$, $(q_y)_{i,j}^n = 0$;
- $\forall (i,j) \in \mathbb{Z}^2$, the pairs ${}^t(h_{i,j}^n, (q_x)_{i,j}^n)$ and ${}^t(h_{i+1,j}^n, (q_x)_{i+1,j}^n)$ satisfy (3.68).

Similarly, $(W_{i,j}^n)_{(i,j)\in\mathbb{Z}^2}$ is said to define a steady state in the y-direction if:

- $\forall (i,j) \in \mathbb{Z}^2$, $W_{i+1,j}^n = W_{i,j}^n$;
- $\forall (i,j) \in \mathbb{Z}^2$, $(q_x)_{i,j}^n = 0$;
- $\forall (i,j) \in \mathbb{Z}^2$, the pairs ${}^t(h_{i,j}^n, (q_y)_{i,j}^n)$ and ${}^t(h_{i,j+1}^n, (q_y)_{i,j+1}^n)$ satisfy (3.68).

Equipped with Proposition 4.1 and Definition 4.2, we can state the properties of the 2D two-step scheme (4.8) – (4.13).

Theorem 4.3. *Under the CFL condition (2.60), the following properties are satisfied by the two-dimensional two-step scheme (4.8) – (4.13).*

(i) Robustness: if $W_{i,j}^n \in \Omega$ *for all* $(i,j) \in \mathbb{Z}^2$, *then* $W_{i,j}^{n+1} \in \Omega$ *for all* $(i,j) \in \mathbb{Z}^2$.

(ii) Well-balance by direction: if $(W_{i,j}^n)_{(i,j)\in\mathbb{Z}^2}$ *defines a steady state in the* x*-direction or in the* y*-direction, then for all* $(i,j) \in \mathbb{Z}^2$, $W_{i,j}^{n+1} = W_{i,j}^n$.

(iii) Preservation of steady states at rest: if, for all $(i,j) \in \mathbb{Z}^2$, $(q_x)_{i,j}^n = (q_y)_{i,j}^n = 0$ *and* $h_{i,j}^n + Z_{i,j}^n = \mathrm{cst}$, *then for all* $(i,j) \in \mathbb{Z}^2$, $W_{i,j}^{n+1} = W_{i,j}^n$.

Proof. We start by proving (i). Note that the second step (4.13) of the scheme does not introduce a modification of the updated height. As a consequence, the robustness of the scheme relies only on the first step of the scheme. Recall Proposition 4.1, which states that the first step (4.8) of the 2D scheme can be written as a convex combination of 1D schemes. Therefore, the 2D scheme is robust if and only if the 1D schemes are robust. Each 1D scheme defined by

(4.15) enjoys the same properties as the first step (3.88) of the truly 1D scheme. From Theorem 3.15, the 1D schemes are robust. Hence, the robustness of the two-step scheme is proven, and (i) holds.

In order to establish the well-balance property, assume that $(W_{i,j}^n)_{(i,j)\in\mathbb{Z}^2}$ defines a steady state in the x-direction. Therefore, the sum of the vertical fluxes in (4.8) vanishes, as does the y contribution of the topography. Thus, the first step (4.8) of the scheme becomes the exact same as in the 1D case, and it is given by (3.88). Then, the y contribution of the friction source term vanishes, leaving only the x contribution to the second step (4.13). In the present context, this second step is the same as in the 1D case; it is therefore given by (3.91). Therefore, Theorem 3.15 applies, and $W_{i,j}^{n+1} = W_{i,j}^n$ for all $(i,j) \in \mathbb{Z}^2$. A similar chain of arguments can be applied to prove the preservation of the steady states in the y-direction, which completes the proof of (ii).

Now, to prove (iii), we consider a steady state at rest. According to Definition 4.2, the relations defining a steady state at rest define both a steady state in the x-direction and in the y-direction. As a consequence, such data is exactly preserved by the scheme after (ii), and (iii) holds as a specific case of (ii). The proof of Theorem 4.3 is thus achieved. □

4.2 High-order extension

In the previous section, we have derived a 2D extension of the 1D well-balanced scheme. The properties possessed by this 2D scheme have been stated in Theorem 4.3. We now focus on a high-order extension of this 2D scheme, in order to improve the space and time accuracy. The general idea of the high-order extension we suggest has been introduced in Section 2.4. This section shows the application of such techniques to the present case, and the recovery of several essential properties such as the robustness and the well-balance.

First, we apply in Section 4.2.1 the results from Section 2.4 to derive a high-order scheme. Then, noting that the reconstruction procedure alters the steady state solutions, we remark that the well-balance property is not satisfied by the high-order scheme. In order to recover this property, we introduce a convex combination between the first-order scheme and the high-order scheme in Section 4.2.2. This convex combination procedure favors the well-balanced scheme when a steady solution is present, i.e. where this scheme is exact. On the contrary, for an unsteady solution, the high-order scheme is favored by the convex combination. In addition, as explained in Section 2.4, the derived scheme is not naturally robust, and some additional treatment has to be applied. In Section 4.2.3, following Section 2.4.3, we elect to apply a MOOD procedure to recover the robustness of the scheme, and eliminate spurious oscillations. Finally, in Section 4.2.4, we discuss the full MOOD loop, including the standard MOOD procedure and the convex combination with the well-balanced scheme.

4.2.1 Application of the high-order strategy to a Cartesian mesh

Recall from Section 2.4 that two ingredients are necessary to achieve high-order accuracy:

- a polynomial reconstruction, see Section 2.4.1;
- a scheme that possesses a high order of accuracy in space and time, see Section 2.4.2.

The goal of the present section is to apply these ingredients to the current case of a Cartesian mesh.

The polynomial reconstruction procedure consists in replacing, in each cell, the constant quantity $W_{i,j}^n$ with a polynomial $\widehat{W}_{i,j}^n(\boldsymbol{x};d)$ of degree d, defined by (2.61). In the case of the shallow-water equations, we elect to provide a reconstruction of the following variables:

$$h \quad ; \quad q_x \quad ; \quad q_y \quad ; \quad h+Z.$$

The coefficients of this polynomial are given by (2.66). Note that computing these coefficients requires knowing the stencil s_i^d, which, in the general case, depends on the cell and the polynomial degree. However, for the particular case of a Cartesian mesh, we can choose the same stencil s^d for each cell. This stencil is taken as the smallest stencil leading to an invertible matrix $\widetilde{X}_i^T\widetilde{X}_i$, where $\widetilde{X}_i$ is given by (2.65) and is used to compute the coefficients (2.66) of the polynomial. With respect to the polynomial degree, the stencil is chosen according to Figure 4.1.

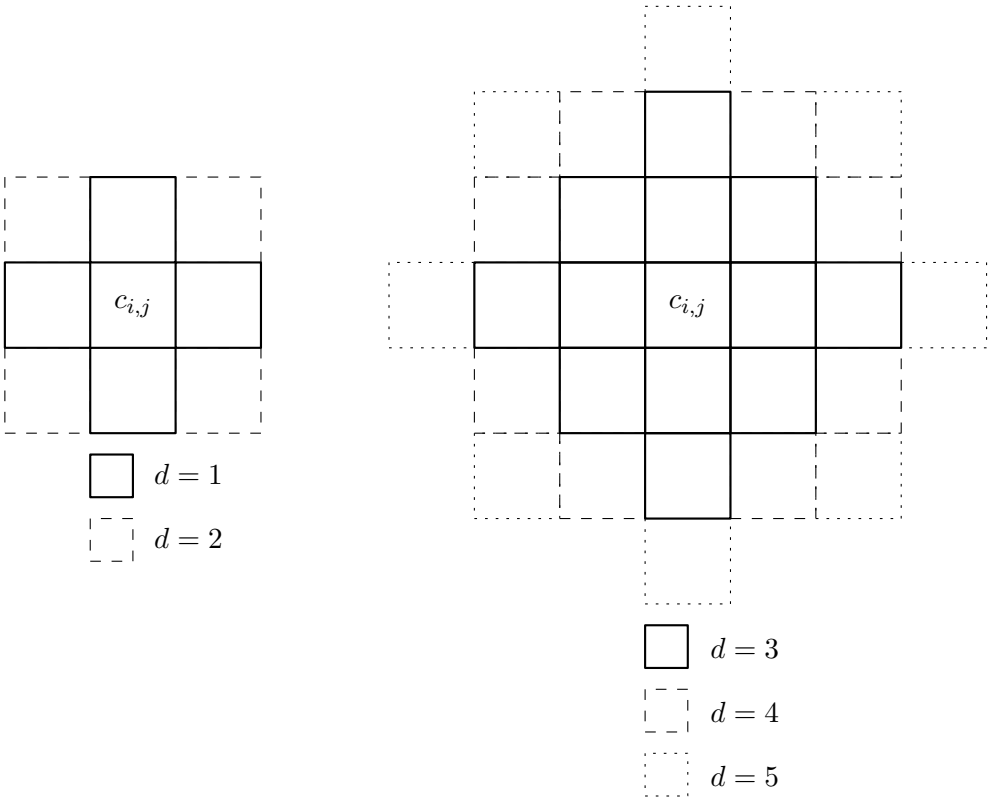

Figure 4.1 – Representation of the stencil for $d \in [\![1,5]\!]$. The lower order stencils are always included in the higher order ones. For the sake of simplicity, we have taken $\Delta x = \Delta y$ in this figure.

Equipped with the polynomial reconstruction, we are able to introduce the high-order scheme. For the spatial high-order, the scheme has been derived in Section 2.4.2.1, to get the expression (2.75). Since this expression has been obtained for a general mesh and for any balance law, we here apply it to the case of the shallow-water equations with topography and

Manning friction. The high-order scheme is therefore given by:

$$W_{i,j}^{n+1} = W_{i,j}^n - \sum_{r=1}^{N_G} \xi_r \left[\frac{\Delta t}{\Delta x} \left(\mathcal{F}_{i+\frac{1}{2},j,r}^n - \mathcal{F}_{i-\frac{1}{2},j,r}^n \right) \right] - \sum_{r=1}^{N_G} \xi_r \left[\frac{\Delta t}{\Delta y} \left(\mathcal{G}_{i,j+\frac{1}{2},r}^n - \mathcal{G}_{i,j-\frac{1}{2},r}^n \right) \right] + \Delta t \sum_{q=1}^{Q} \chi_q \Big((\mathcal{S}^t)_{i,j,q}^n + (\mathcal{S}^f)_{i,j,q}^n \Big). \tag{4.16}$$

In Chapter 2, we introduced the quadrature formulas (2.71) and (2.73), respectively concerning the approximation of the integral over an edge and a cell. These quadrature formulas define the coefficients ξ_r and χ_q, as well as points $\boldsymbol{\sigma}_r$ and $\boldsymbol{X}_q$ within edges and cells. For a cell $c_{i,j}$, the quadrature points on the inner edges are denoted with clear notations by $\boldsymbol{\sigma}_{i\pm\frac{1}{2},j}^r$ and $\boldsymbol{\sigma}_{i,j\pm\frac{1}{2}}^r$. Similarly, the quadrature points within the cell are denoted by $\boldsymbol{X}_{i,j}^q$. Equipped with these quadrature formulas, the high-order numerical fluxes are defined as follows:

$$\begin{aligned} \mathcal{F}_{i+\frac{1}{2},j,r}^n &= \mathcal{F}\Big(\widehat{W}_{i,j}^n(\boldsymbol{\sigma}_{i+\frac{1}{2},j}^r; d),\ \widehat{W}_{i+1,j}^n(\boldsymbol{\sigma}_{i+\frac{1}{2},j}^r; d) \Big), \\ \mathcal{G}_{i,j+\frac{1}{2},r}^n &= \mathcal{G}\Big(\widehat{W}_{i,j}^n(\boldsymbol{\sigma}_{i,j+\frac{1}{2}}^r; d),\ \widehat{W}_{i,j+1}^n(\boldsymbol{\sigma}_{i,j+\frac{1}{2}}^r; d) \Big), \end{aligned}$$

where $\widehat{W}_{i,j}^n(\boldsymbol{x}; d)$ represents the polynomial reconstruction within the cell $c_{i,j}$.

After (4.6) – (4.7), the intermediate states used in the numerical fluxes $\mathcal{F}$ and $\mathcal{G}$ involve the approximate friction source term $\overline{S}^f$, since they are given by (3.81). However, in order to obtain a high-order scheme, the definition (3.79) of $\overline{S}^f$ has to be replaced. Within $\mathcal{F}$, i.e. in the x-direction, we suggest the following expression:

$$\begin{aligned} \overline{S}^f \Delta x &:= \overline{S}^f(h_L, h_R, q_L, q_R, \Delta x) \Delta x^{d+1} \\ &= \begin{cases} 0 & \text{if } h_L = 0 \text{ and/or } h_R = 0, \\ -k\bar{q}|\bar{q}|\overline{h^{-\eta}} \Delta x^{d+1} & \text{otherwise,} \end{cases} \end{aligned} \tag{4.17}$$

where $\bar{q}$ is defined by (3.61) and $\overline{h^{-\eta}}$ is given, instead of (3.62), by:

$$\overline{h^{-\eta}} = \frac{[h^2]}{2} \frac{\eta + 2}{[h^{\eta+2}]} - \frac{\bar{\mu}}{k\, \Delta x^{d+1}} \left(\left[\frac{1}{h} \right] + \frac{[h^2]}{2} \frac{[h^{\eta-1}]}{\eta - 1} \frac{\eta + 2}{[h^{\eta+2}]} \right). \tag{4.18}$$

In the above expressions, the quantity Δx present in the first-order case has been raised to the power of $(d+1)$. As a consequence, the expression of $\overline{S}^f/\alpha$ is no longer given by (4.19), but rather by:

$$\frac{\overline{S}^f \Delta x}{\alpha} = \begin{cases} 0 & \text{if } h_L = 0 \text{ and/or } h_R = 0, \\ \dfrac{-k\bar{q}|\bar{q}|\overline{h^{-\eta}} \Delta x^{d+1}}{-\frac{(q^*)^2}{h_L h_R} + \frac{g}{2}(h_L + h_R)} & \text{otherwise.} \end{cases} \tag{4.19}$$

Then, within the numerical flux $\mathcal{G}$ in the y-direction, similar expressions are used with Δy instead of Δx. Note that, if $d = 0$, the expressions (4.17), (4.18) and (4.19) coincide with (3.79), (3.62) and (3.80). Thus, if $d = 0$, the numerical fluxes are not modified.

Then, the high-order numerical source terms $(\mathcal{S}^t)_{i,j,q}^n$ and $(\mathcal{S}^f)_{i,j,q}^n$ are defined by evaluat-

ing the source terms (4.4) at the quadrature points, as follows:

$$(\mathcal{S}^t)^n_{i,j,q} = S^t\Big(\widehat{W}^n_{i,j}(\boldsymbol{X}^q_{i,j};d)\Big) \quad \text{and} \quad (\mathcal{S}^f)^n_{i,j,q} = S^f\Big(\widehat{W}^n_{i,j}(\boldsymbol{X}^q_{i,j};d)\Big).$$

Hence, the numerical source terms are given by:

$$(\mathcal{S}^t)^n_{i,j,q} = -g\hat{h}^n_{i,j}(\boldsymbol{X}^q_{i,j};d)\begin{pmatrix}0\\(\boldsymbol{\nabla}\widehat{Z})^n_{i,j}(\boldsymbol{X}^q_{i,j};d)\end{pmatrix}, \tag{4.20a}$$

$$(\mathcal{S}^f)^n_{i,j,q} = -k\|\hat{\boldsymbol{q}}^n_{i,j}(\boldsymbol{X}^q_{i,j};d)\|(\hat{h}^n_{i,j}(\boldsymbol{X}^q_{i,j};d))^{-\eta}\begin{pmatrix}0\\\hat{\boldsymbol{q}}^n_{i,j}(\boldsymbol{X}^q_{i,j};d)\end{pmatrix}. \tag{4.20b}$$

Since the first components of these high-order numerical source terms are zero, we also introduce their nonzero components $(\boldsymbol{\mathcal{S}}^t_{\text{HO}})^n_{i,j}$ and $(\boldsymbol{\mathcal{S}}^f_{\text{HO}})^n_{i,j}$, defined as follows:

$$\sum_{q=1}^{Q}\chi_q(\mathcal{S}^t)^n_{i,j,q} = \begin{pmatrix}0\\(\boldsymbol{\mathcal{S}}^t_{\text{HO}})^n_{i,j}\end{pmatrix} \quad \text{and} \quad \sum_{q=1}^{Q}\chi_q(\mathcal{S}^f)^n_{i,j,q} = \begin{pmatrix}0\\(\boldsymbol{\mathcal{S}}^f_{\text{HO}})^n_{i,j}\end{pmatrix}. \tag{4.21}$$

Let us make the important remark that, from Chapter 2, the derivation of the high-order accurate scheme involves the approximation of the integral of the source terms over a cell. In order to preserve the high-order accuracy, these integrals need to be approximated with a quadrature formula, and the high-order numerical source terms cannot involve the averages $\overline{S}^t$ and $\overline{S}^f$, defined by (3.75) and (3.79) to ensure the well-balance of the scheme. As a consequence, there is no way for the high-order scheme (4.16) to be well-balanced without an additional treatment.

The scheme (4.16) has a high-order spatial accuracy. However, its time accuracy is still of order one. Hence, we use the SSPRK (Strong Stability-Preserving Runge-Kutta) methods described in Section 2.4.2.2 to improve the time accuracy of the scheme. In order to set up such techniques, we first rewrite the scheme (4.16) as $W^{n+1} = \mathcal{H}(W^n)$, where $W^n = (W^n_i)_{i\in\mathbb{Z}}$. The generic Runge-Kutta technique is then given by (2.76), where the coefficients α_{lk} and β_{lk} for the SSPRK method depend on its order, chosen by following Table C.1. These methods require the use of the modification (2.77) of the time step (2.60). Enhanced with a SSPRK time integrator, the scheme (4.16) has a high order of space and time accuracy.

4.2.2 Recovery of the well-balance property

The high-order scheme (4.16) derived in the previous section is not well-balanced by direction. Indeed, the reconstruction procedure modifies the approximate solution at the interfaces, and the approximate source terms $\overline{S}^t$ and $\overline{S}^f$ are no longer present to allow an exact preservation of the steady state solutions: as mentioned in Section 4.2.1, the scheme uses quadrature formulas instead of $\overline{S}^t$ and $\overline{S}^f$ to approximate the contributions of the source terms.

We now propose a way to restore this essential property, by introducing a convex combination between the high-order scheme and the first-order well-balanced scheme. A similar technique has been used in [94, 164, 114], where the authors introduce a convex combination

between two schemes to obtain a high-order positivity-preserving scheme. The goal of the present convex combination is to use the high-order scheme when the solution is unsteady and the first-order well-balanced scheme when the solution is steady. The first-order scheme is exact, i.e. its order is infinite, for steady state solutions. As a consequence, such a convex combination would be carried out between a high-order accurate scheme and an exact scheme. The resulting scheme should thus still be at least high-order accurate.

The convex combination is based on a steady state detector, which we first introduce. Then, this detector is used to derive a convex combination favoring either the well-balanced scheme or the high-order scheme.

4.2.2.1 A one-dimensional steady state detector

Since the two-step 2D scheme (4.8) – (4.13) is well-balanced by direction (see Theorem 4.3), it makes sense to define a steady state detector in each space dimension. Hence, we momentarily consider the 1D shallow-water equations (1.1) and a 1D space discretization.

Recall that steady state solutions in one space dimension are given by (1.40), as follows:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -gh\partial_x Z - kq|q|h^{-\eta}. \end{cases} \tag{4.22}$$

From (3.68), two states W_L and W_R are said to define a steady state if the following discretization of (4.22) is satisfied:

$$\begin{cases} [q] = 0, \\ \left[\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right] = \overline{S}^t \Delta x + \overline{S}^f \Delta x, \end{cases} \tag{4.23}$$

with $\overline{S}^t$ given by (3.75) and $\overline{S}^f$ given by (3.79). As a consequence, it makes sense to consider the following steady state detector:

$$\varphi(W_L, W_R, Z_L, Z_R, \Delta x) = \left\| \begin{pmatrix} q_R - q_L \\ \dfrac{q_R^2}{h_R} - \dfrac{q_L^2}{h_L} + \dfrac{g}{2}(h_R^2 - h_L^2) - (\overline{S}^t)\Delta x - (\overline{S}^f)\Delta x \end{pmatrix} \right\|_2. \tag{4.24}$$

We immediately note that, if W_L and W_R define a steady state according to (4.23), then the quantity $\varphi(W_L, W_R, Z_L, Z_R, \Delta x)$ vanishes. Therefore, φ detects whether a steady state is defined by two pairs (W_L, Z_L) and (W_R, Z_R).

Now, consider two 1D cells c_i and c_{i+1}, where approximate solutions W_i^n and W_{i+1}^n are defined. In order to detect whether W_i^n and W_{i+1}^n define a steady state, we consider the steady state detector evaluated at the interface between the cells c_i and c_{i+1}, i.e. the following quantity:

$$\begin{aligned} \varphi_{i+\frac{1}{2}}^n &:= \varphi(W_i^n, W_{i+1}^n, Z_i, Z_{i+1}, \Delta x) \\ &= \left\| \begin{pmatrix} q_{i+1}^n - q_i^n \\ \dfrac{q_{i+1}^2}{h_{i+1}} - \dfrac{q_i^2}{h_i} + \dfrac{g}{2}(h_{i+1}^2 - h_i^2) - (\overline{S}^t)_{i+\frac{1}{2}}^n \Delta x - (\overline{S}^f)_{i+\frac{1}{2}}^n \Delta x \end{pmatrix} \right\|_2, \end{aligned}$$

where $(\overline{S}^t)^n_{i+\frac{1}{2}} = \overline{S}^t(h^n_i, h^n_{i+1}, q^n_i, q^n_{i+1}, Z_i, Z_{i+1}, \Delta x)$ and $(\overline{S}^f)^n_{i+\frac{1}{2}} = \overline{S}^f(h^n_i, h^n_{i+1}, q^n_i, q^n_{i+1}, \Delta x)$. We also define the following steady state detector, which determines whether a steady state is present between the three states W^n_{i-1}, W^n_i and W^n_{i+1}, i.e. if both pairs (W^n_{i-1}, W^n_i) and (W^n_i, W^n_{i+1}) define a steady state:

$$\varphi^n_i = \varphi^n_{i-\frac{1}{2}} + \varphi^n_{i+\frac{1}{2}}. \tag{4.25}$$

It is clear that W^n_{i-1}, W^n_i and W^n_{i+1} define a steady state as soon as φ^n_i vanishes.

From the steady state detector (4.25), we can derive a suitable convex combination process. This convex combination shall rely on a parameter favoring the well-balanced scheme when a steady state is reached. Indeed, the well-balanced scheme is exact for steady states, and it should be used whenever the solution is close to defining a steady state. As a consequence, we define a parameter θ^n_i, which lives in $[0, 1]$. We wish θ^n_i to vanish if the equilibrium error φ^n_i is small enough, i.e. if W^n_{i-1}, W^n_i and W^n_{i+1} are close to defining a steady state. Moreover, we want θ^n_i to be equal to 1 if these vectors are far from defining a steady state, i.e. if φ^n_i is large enough. Therefore, we elect to define θ^n_i as follows:

$$\theta^n_i = \begin{cases} 0 & \text{if } \varphi^n_i < m\,\Delta x, \\ \dfrac{\varphi^n_i - m\Delta x}{M\Delta x - m\Delta x} & \text{if } m\,\Delta x \leq \varphi^n_i \leq M\,\Delta x, \\ 1 & \text{if } \varphi^n_i > M\,\Delta x, \end{cases} \tag{4.26}$$

with $M \geq m \geq 0$. If $M = 0$, then $\theta^n_i = 1$ and the high-order scheme is used. If $M = m$, then either $\theta^n_i = 0$ or $\theta^n_i = 1$. The process we used to define θ^n_i is highlighted on Figure 4.2.

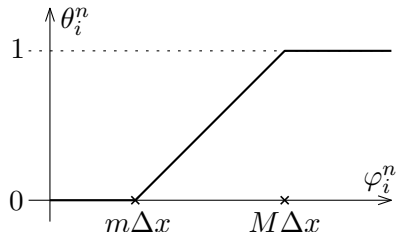

Figure 4.2 – Graph of θ^n_i with respect to φ^n_i, according to (4.26).

4.2.2.2 The two-dimensional convex combination

Equipped with the parameter θ^n_i given by (4.26), we now use it to define a suitable convex combination process ensuring that the high-order scheme (4.16) is well-balanced. This convex combination is done between the high-order scheme (4.16) and the two-step well-balanced scheme (4.8) – (4.13).

Since the parameter θ^n_i has been defined for the 1D shallow-water equations, we have to extend it to two space dimensions. We consider $\boldsymbol{\theta}^n_{i,j} = {}^t((\theta_x)^n_{i,j}, (\theta_y)^n_{i,j})$, where $(\theta_x)^n_{i,j}$ is obtained by evaluating (4.26) in the x-direction, and $(\theta_y)^n_{i,j}$ results from the evaluation of (4.26) in the y-direction. Therefore, after Definition 4.2, the parameter $(\theta_x)^n_{i,j}$ detects steady states in the x-direction, while $(\theta_y)^n_{i,j}$ detects steady states in the y-direction. To define $(\theta_x)^n_{i,j}$ and $(\theta_y)^n_{i,j}$, we have to introduce four additional constants, m_x, M_x, m_y and M_y, to mimic the

role of m and M in (4.26). As a consequence, $(\theta_x)_{i,j}^n$ is defined as follows:

$$(\theta_x)_{i,j}^n = \begin{cases} 0 & \text{if } (\varphi_x)_{i,j}^n < m_x\,\Delta x, \\ \dfrac{(\varphi_x)_{i,j}^n - m_x\Delta x}{M_x\Delta x - m_x\Delta x} & \text{if } m_x\,\Delta x \leq (\varphi_x)_{i,j}^n \leq M_x\,\Delta x, \\ 1 & \text{if } (\varphi_x)_{i,j}^n > M_x\,\Delta x, \end{cases} \tag{4.27}$$

while $(\theta_y)_{i,j}^n$ is given by:

$$(\theta_y)_{i,j}^n = \begin{cases} 0 & \text{if } (\varphi_y)_{i,j}^n < m_y\,\Delta y, \\ \dfrac{(\varphi_y)_{i,j}^n - m_y\Delta y}{M_y\Delta y - m_y\Delta y} & \text{if } m_y\,\Delta y \leq (\varphi_y)_{i,j}^n \leq M_y\,\Delta y, \\ 1 & \text{if } (\varphi_y)_{i,j}^n > M_y\,\Delta y. \end{cases} \tag{4.28}$$

In (4.27), we have defined the steady state detector in the x-direction $(\varphi_x)_{i,j}^n$ as follows:

$$\begin{aligned} (\varphi_x)_{i,j}^n &= \varphi_{i-\frac{1}{2},j}^n + \varphi_{i+\frac{1}{2},j}^n \\ &= \varphi(W_{i-1,j}^n, W_{i,j}^n, Z_{i-1,j}, Z_{i,j}, \Delta x) + \varphi(W_{i,j}^n, W_{i+1,j}^n, Z_{i,j}, Z_{i+1,j}, \Delta x). \end{aligned}$$

Similarly, the steady state detector in the y-direction $(\varphi_y)_{i,j}^n$, used in (4.28), is given by:

$$\begin{aligned} (\varphi_y)_{i,j}^n &= \varphi_{i,j-\frac{1}{2}}^n + \varphi_{i,j+\frac{1}{2}}^n \\ &= \varphi(W_{i,j-1}^n, W_{i,j}^n, Z_{i,j-1}, Z_{i,j}, \Delta y) + \varphi(W_{i,j}^n, W_{i,j+1}^n, Z_{i,j}, Z_{i,j+1}, \Delta y). \end{aligned}$$

We are now able to state the convex combination. The first step, devoted to the flux and the topography source term, reads:

$$\begin{aligned} W_{i,j}^{n+\frac{1}{2}} = W_{i,j}^n &- (\theta_x)_{i,j}^n \frac{\Delta t}{\Delta x} \sum_{r=1}^{N_G} \xi_r \Big(\mathcal{F}_{i+\frac{1}{2},j,r}^n - \mathcal{F}_{i-\frac{1}{2},j,r}^n \Big) \\ &- \Big(1 - (\theta_x)_{i,j}^n\Big) \frac{\Delta t}{\Delta x} \big(\mathcal{F}\big(W_{i,j}^n, W_{i+1,j}^n\big) - \mathcal{F}\big(W_{i-1,j}^n, W_{i,j}^n\big) \big) \\ &- (\theta_y)_{i,j}^n \frac{\Delta t}{\Delta y} \sum_{r=1}^{N_G} \xi_r \Big(\mathcal{G}_{i,j+\frac{1}{2},r}^n - \mathcal{G}_{i,j-\frac{1}{2},r}^n \Big) \\ &- \Big(1 - (\theta_y)_{i,j}^n\Big) \frac{\Delta t}{\Delta y} \big(\mathcal{G}\big(W_{i,j}^n, W_{i,j+1}^n\big) - \mathcal{G}\big(W_{i,j-1}^n, W_{i,j}^n\big) \big) \\ &+ \Delta t \begin{pmatrix} 0 \\ \boldsymbol{\theta}_{i,j}^n \cdot (\boldsymbol{\mathcal{S}}_{\mathrm{HO}}^t)_{i,j}^n + \Big(1 - \boldsymbol{\theta}_{i,j}^n\Big) \cdot (\boldsymbol{S}_{\mathrm{WB}}^t)_{i,j}^n \end{pmatrix}, \end{aligned} \tag{4.29}$$

where the high-order numerical topography source term $(\boldsymbol{\mathcal{S}}_{\mathrm{HO}}^t)_{i,j}^n$ is given by (4.21), and where the well-balanced numerical topography source term $(\boldsymbol{S}_{\mathrm{WB}}^t)_{i,j}^n$ is defined by (4.9). The expression (4.29) of the scheme is nothing but a convex combination by direction of the first step (4.8) of the well-balanced scheme and the flux and topography contributions of the high-order scheme (4.16).

Concerning the updated water heights, we take $h_{i,j}^{n+1} = h_{i,j}^{n+\frac{1}{2}}$, since the second step is

devoted to the friction source term and therefore does not impact the water height. Following (4.13), let $(\boldsymbol{q}_{\text{WB}})_{i,j}^{n+1}$ be the vector containing the discharge obtained after the second step of the first-order scheme: we set

$$(\boldsymbol{q}_{\text{WB}})_{i,j}^{n+1} = \begin{pmatrix} (q_x)_{i,j}^{n+1} \\ (q_y)_{i,j}^{n+1} \end{pmatrix}. \tag{4.30}$$

We also define $(\boldsymbol{q}_{\text{HO}})_{i,j}^{n+1}$ as the discharge obtained using the high-order friction source term, as follows:

$$(\boldsymbol{q}_{\text{HO}})_{i,j}^{n+1} = \boldsymbol{q}_{i,j}^{n+\frac{1}{2}} + \Delta t (\boldsymbol{\mathcal{S}}_{\text{HO}}^{f})_{i,j}^{n}. \tag{4.31}$$

The second step of the high-order well-balanced scheme consists in the following convex combination of the discharge (4.30) of the well-balanced scheme with the discharge (4.31) of the high-order scheme:

$$\boldsymbol{q}_{i,j}^{n+1} = \boldsymbol{\theta}_{i,j}^{n} \cdot (\boldsymbol{q}_{\text{HO}})_{i,j}^{n+1} + \left(1 - \boldsymbol{\theta}_{i,j}^{n}\right) \cdot (\boldsymbol{q}_{\text{WB}})_{i,j}^{n+1}. \tag{4.32}$$

The two-step scheme (4.29) – (4.32) allows to use the high-order scheme or the first-order well-balanced one, or even a scheme that is a combination of these two schemes. Indeed, if $\boldsymbol{\theta}_{i,j}^{n} = 1$, then the high-order scheme is used, while the first-order well-balanced scheme is used if $\boldsymbol{\theta}_{i,j}^{n} = 0$. The well-balance property satisfied by this scheme is summarized in the following result.

Lemma 4.4. *The high-order two-step scheme (4.29) – (4.32) is well-balanced by direction.*

Proof. The goal of this proof is to show that, if a steady state in the x-direction or the y-direction is considered, then the scheme (4.29) – (4.32) yields $W_{i,j}^{n+1} = W_{i,j}^{n}$ for all $(i,j) \in \mathbb{Z}$.

Assume that $(W_{i,j}^{n})_{(i,j)\in\mathbb{Z}^2}$ defines a steady state in the x-direction, as prescribed by Definition 4.2. In this case, by construction, we have $(\theta_x)_{i,j}^{n} = 0$ and $(\theta_y)_{i,j}^{n} = 0$. As a consequence, each of the steps (4.29) – (4.32) of the two-step high-order scheme degenerates into the steps (4.8) – (4.13) of the well-balanced scheme. Recall from Theorem 4.3 that this scheme is well-balanced by direction. Therefore, if $(W_{i,j}^{n})_{(i,j)\in\mathbb{Z}^2}$ defines a steady state in the x-direction, then $W_{i,j}^{n+1} = W_{i,j}^{n}$ for all $(i,j) \in \mathbb{Z}$.

The same chain of arguments can be applied to show that, if $(W_{i,j}^{n})_{(i,j)\in\mathbb{Z}^2}$ defines a steady state in the y-direction, then $W_{i,j}^{n+1} = W_{i,j}^{n}$ for all $(i,j) \in \mathbb{Z}$. The proof of Lemma 4.4 is thus completed. □

4.2.3 The MOOD method

The high-order procedure described in Section 4.2.1, in addition to inducing a loss of the well-balance property, produces spurious oscillations and a loss of robustness. In addition, since the high-order scheme is no longer semi-implicit, the friction contribution is treated explicitly, and the stiffness of the friction source term near wet/dry interfaces will also cause spurious oscillations if the time step is not modified. The well-balance has been recovered in Section 4.2.2. To address the issue of the oscillations, we elect to use a MOOD method, presented in the general case in Section 2.4.3. It consists in lowering the degree of the polynomial reconstruction in specific cells if the approximate solution does not satisfy certain criteria. The

approximate solution, called the candidate solution and denoted by $W^\star_{i,j}$, is tested against several detection criteria. These criteria have been introduced in Section 2.4.3, and we state them below in the context of the shallow-water equations.

The Physical Admissibility Detector (PAD)

The PAD determines whether the approximate solution is out of the admissible states space Ω. In the case of the shallow-water equations, the PAD checks whether the water height is non-negative. Thus, the PAD criterion fails within the cell $c_{i,j}$ if

$$h^\star_{i,j} < 0. \tag{4.33}$$

We emphasize that the PAD ensures that the high-order scheme is non-negativity preserving, since this property is satisfied by the first-order scheme.

The Discrete Maximum Principle detector (DMP)

The DMP criterion (2.78) checks for oscillations. Let $\nu_{i,j}$ be the set of cells connected to $c_{i,j}$ with an edge or a vertex. The DMP criterion fails if one of the following three checks fail:

$$\begin{aligned}
\min_{l\in\nu_{i,j}}(h_l + Z_l) - \varepsilon_M &\leq h^\star_{i,j} + Z^\star_{i,j} \leq \min_{l\in\nu_{i,j}}(h_l + Z_l) + \varepsilon_M,\\
\min_{l\in\nu_{i,j}}((q_x)_l) - \varepsilon_M &\leq (q_x)^\star_{i,j} \leq \min_{l\in\nu_{i,j}}((q_x)_l) + \varepsilon_M,\\
\min_{l\in\nu_{i,j}}((q_y)_l) - \varepsilon_M &\leq (q_y)^\star_{i,j} \leq \min_{l\in\nu_{i,j}}((q_y)_l) + \varepsilon_M,
\end{aligned}$$

where $\varepsilon_M = \min(\Delta x, \Delta y)^3$.

The u2 criterion

The goal of the u2 criterion is to ensure that the DMP does not eliminate physical oscillations. This criterion is made of three detectors, already defined in Section 2.4.3:

- the plateau detector (2.79);
- the oscillation detector (2.80);
- the smoothness detector (2.81).

If the plateau or the smoothness detectors are activated, then the DMP criterion was a false positive, and the u2 criterion succeeds. However, if the oscillation detector is activated, then the u2 criterion fails.

The detector loop

Equipped with the three detectors, the loop is similar to the one defined in Section 2.4.3. The only difference is that, in the present case, the Cell Polynomial Degree (CPD) goes from d to 0 in the first iteration of the MOOD method, instead of gradually decreasing from d to 0. This behavior has been chosen to be consistent with the well-balance recovery presented in Section 4.2.2. Indeed, this procedure involves a convex combination between the high-order

scheme and the first-order well-balanced scheme. As a consequence, here, it does not make sense to gradually decrease the degree of the polynomial reconstruction when applying the MOOD technique; rather, we brutally decrease the degree from d to 0. In this context, the cascade of detectors is displayed on Figure 4.3.

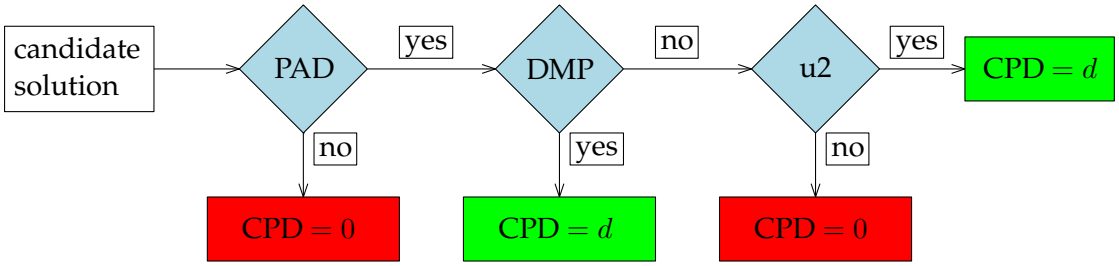

Figure 4.3 – The MOOD detector chain.

4.2.4 Combining the well-balance recovery with MOOD

In the two previous sections, we have suggested two methods to restore, for the high-order scheme, the properties satisfied by the first-order scheme. First, the well-balance property is recovered using the convex combination technique presented in Section 4.2.2. Then, the oscillations induced by the high-order scheme are eliminated thanks to the MOOD process introduced in Section 4.2.3. The goal is now to combine these two procedures and to add the MOOD technique to the high-order well-balanced scheme (4.29) - (4.32).

Since the well-balance correction is an *a priori* procedure, it makes sense to check *a priori* for the physical admissibility of the reconstruction, in addition to using the PAD detector. The reconstruction will be considered physically admissible in a cell $c_{i,j}$ if we have $\hat{h}^n_{i,j} > 0$ at all edge and cell Gauss points $\boldsymbol{\sigma}_r$ and $\boldsymbol{X}_q$. The admissibility of the reconstruction is checked twice, once when computing the reconstructed heights at the edge Gauss points $\boldsymbol{\sigma}_r$ to determine the high-order fluxes, and once when computing the numerical approximation of the mean of the friction source term, using the evaluation of $\hat{h}^n_{i,j}$ at the cell Gauss points $\boldsymbol{X}_q$.

Algorithm 4.5. *For a single iteration in time of the SSPRK time integrator, the MOOD loop reads as follows.*

(1) For each cell $c_{i,j}$, initialize CPD$(i,j) = d$.

(2) For each cell $c_{i,j}$, compute the pair of correction parameters $\boldsymbol{\theta}^n_{i,j}$. If $(\theta_x)^n_{i,j} = (\theta_y)^n_{i,j} = 0$, i.e if $\boldsymbol{\theta}^n_{i,j} = \mathbf{0}$, then a steady state is considered within the cell $c_{i,j}$. In this case, set CPD$(i,j) = 0$ to ensure that the well-balanced scheme is used to exactly preserve this steady state solution.

(3) Compute the reconstruction within each cell $c_{i,j}$, to be evaluated at the edge and cell Gauss points $\boldsymbol{\sigma}_r$ and $\boldsymbol{X}_q$, and test its physical admissibility, as follows:

(3a) if $\hat{h}^n_{i,j}(\boldsymbol{\sigma}_r) < 0$ for some edge Gauss point $\boldsymbol{\sigma}_r$, then the reconstruction is rejected in $c_{i,j}$, and CPD(i,j) is set to zero;

(3b) if $\hat{h}^n_{i,j}(\boldsymbol{X}_q) < 0$ for some cell Gauss point $\boldsymbol{X}_q$, then the reconstruction is rejected in $c_{i,j}$, and CPD(i,j) is set to zero.

(4) Equipped with the new CPD map, compute the candidate solution $W^\star$, using the high-order well-balanced scheme (4.29) – (4.32).

(5) Apply the detection process displayed on Figure 4.3 to compute a potentially new CPD map and to decide whether to accept the candidate solution. If the candidate solution is rejected, go to step (4). Otherwise, go to step (6).

(6) The candidate solution is accepted: set $W^{n+1} = W^\star$.

Equipped with Algorithm 4.5, the following result holds.

Theorem 4.6. *Algorithm 4.5 yields a scheme that is robust and well-balanced by direction.*

Proof. The MOOD procedure includes the PAD detection criterion (4.33), which ensures that the updated water height is non-negative. Indeed, at worst, it is computed with the first-order scheme, which is positivity-preserving. In addition, the well-balance property is ensured by Lemma 4.4. Therefore, the scheme defined by Algorithm 4.5 is robust and well-balanced by direction, which concludes the proof of Theorem 4.6. □

4.3 Implementation in Fortran

The scheme proposed in Algorithm 4.5 was implemented in Fortran from scratch. It was also equipped with an OpenMP parallelization (see [38, 40] for instance). This section describes this process.

First, we implemented both 1D schemes, the explicit scheme (3.9) – (3.81) and the semi-implicit one (3.88) – (3.91) – (3.101). This implementation was straightforward, and no difficulties were encountered. Since the proposed 1D numerical experiments did not take a long CPU time, the code was not parallelized at this stage. Thanks to preprocessor directives, both the explicit and the semi-implicit scheme were implemented in the same code; when compiling the code, the user chooses either the explicit scheme or the semi-implicit scheme.

Then, regarding the 2D scheme, we first had to create a mesh. Since we focused on a Cartesian mesh, this step did not require the use of additional software. The mesh was created within the Fortran code using several customized types.

Afterwards, we computed the matrix $\widetilde{X}_i$, given by (2.65) and used in the polynomial reconstruction. Noting that all the cells are rectangles of length Δx and width Δy, we remarked that this matrix does not actually depend on the cell c_i, and that a single matrix $\widetilde{X}$ had to be computed. Equipped with $\widetilde{X}$, the next step was the computation of the pseudoinverse $(\widetilde{X}_i^T \widetilde{X}_i)^{-1} \widetilde{X}_i^T$. To that end, we elected to use two LAPACK routines, `DGETRF` to compute the LU factorization of $\widetilde{X}_i^T \widetilde{X}_i$ and `DGETRI` to actually compute the inverse matrix from the LU factorization. Thanks to the rescaling (2.65) which lowers its condition number, inverting the matrix $\widetilde{X}_i^T \widetilde{X}_i$ was not problematic. Without the rescaling, the condition number became very high, especially when dealing with large polynomial degrees, which introduced potentially damaging errors.

Then, we followed Algorithm 4.5 to implement the high-order well-balanced scheme. We were able to provide a straightforward OpenMP parallelization of all the loops. For instance, in order to compute the maximum of the characteristic velocities, we added a `REDUCTION`

clause to the OpenMP `DO` loop. The steps from Algorithm 4.5 were finally supplemented with the SSPRK technique, in order to ensure the high order accuracy in time as well as in space.

In order to test the parallelization, we used a machine equipped with two Intel Xeon X5650 processors, each with 12 cores at 2.66 GHz, 6 physical and 6 logical. The code was compiled with GNU Fortran version 4.9.2, using the `-O3` optimization flag. In order to handle the dependencies between the numerous modules, the program `makedepf90` was used. The speedup and the efficiency of the parallelization were tested. The speedup is defined as the time gained by using the parallelization. With t_N the time taken using N cores, the speedup $\mathscr{S}$ is defined by:

$$\mathscr{S} = \frac{t_1}{t_N}.$$

The optimal speedup is equal to the number of cores N. The efficiency $\mathscr{E}$ of the parallelization is closely related to the speedup: indeed, it is a percentage defined as the speedup divided by the number of cores, as follows:

$$\mathscr{E} = 100\frac{\mathscr{S}}{N}.$$

Since the optimal speedup is equal to N, the optimal efficiency is 100%. The results of the test are displayed on Figure 4.4; they show a good speedup and efficiency of the parallelization.

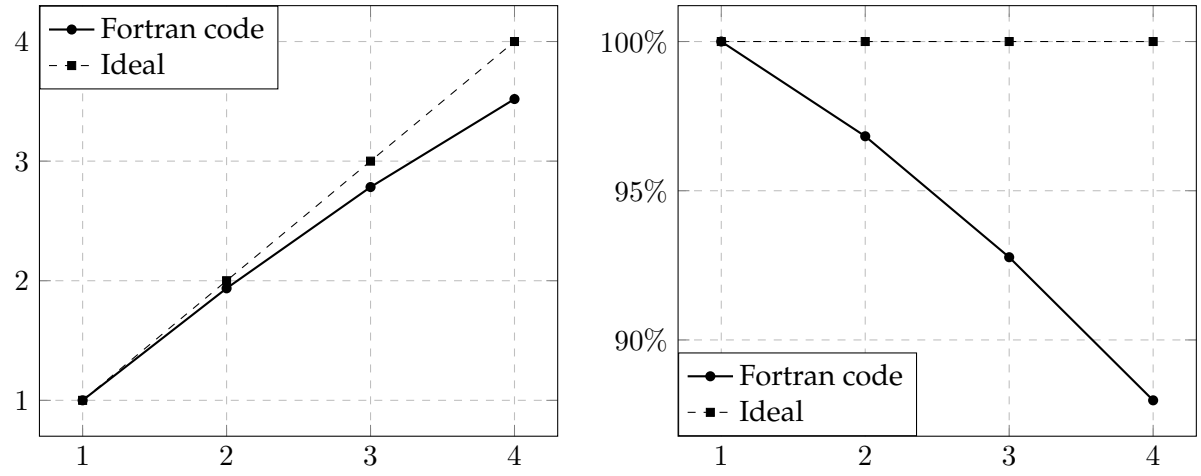

Figure 4.4 – Speedup $\mathscr{S}$ and efficiency $\mathscr{E}$ for the OpenMP parallelization. Left panel: speedup; right panel: efficiency.

Finally, the output of the code consists either in ASCII `.vtk` files or in plain text `.csv` files. These files are to be read by Paraview. In addition, all the 2D figures and some 1D figures were made with Paraview version 5.0.1; the rest of the 1D figures were made with PGFPLOTS version 1.13. The few 3D figures are also made with Paraview from `.csv` files, using the "Table to Points" and the "Delaunay 2D" filters.

4.4 Numerical experiments

This last section is devoted to numerical experiments, designed to highlight some essential properties of the scheme. We first introduce the following notations, used to concisely

represent the schemes that will be tested.

- The scheme that uses a polynomial reconstruction of degree d, i.e. whose order of accuracy is $(d+1)$, is denoted by $\mathbb{P}_d$. This notation includes the first-order well-balanced scheme, which is thus denoted by $\mathbb{P}_0$. Note that, since the well-balance correction is not active for the $\mathbb{P}_d$ scheme, we have $M_x = M_y = 0$.
- For $d \geq 1$, the $\mathbb{P}_d$ scheme equipped with the well-balance correction, i.e. with $M_x > 0$ and/or $M_y > 0$, will be denoted by $\mathbb{P}_d^{\text{WB}}$.

For the $\mathbb{P}_d$ scheme, Algorithm 4.5 is applied without the well-balanced correction, while the $\mathbb{P}_d^{\text{WB}}$ uses the full loop present in Algorithm 4.5.

In addition, in order to assess the well-balance and the high-order accuracy of the scheme, we need to compute errors between the exact solution $W^{ex}(t,x,y)$ and the approximate solution. Consider a uniform Cartesian mesh made of $N = N_x \times N_y$ cells. We denote by $W_{i,j}^{ex}$ the average of the exact solution over the cell $c_{i,j}$ at time t, as follows:

$$W_{i,j}^{ex}(t) = \frac{1}{|c_{i,j}|} \int_{c_{i,j}} W^{ex}(t,x,y)\, dx\, dy. \tag{4.34}$$

Equipped with this notation, we compute the errors in L^1, L^2 and L^∞ norms following (2.36), with $W_{i,j}^n$ the approximate solution at time t^n:

$$L^1 \text{ error: } \frac{1}{N} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left| W_{i,j}^n - W_{i,j}^{ex}(t^n) \right|, \tag{4.35a}$$

$$L^2 \text{ error: } \sqrt{\frac{1}{N} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left(W_{i,j}^n - W_{i,j}^{ex}(t^n) \right)^2}, \tag{4.35b}$$

$$L^\infty \text{ error: } \max_{\substack{1 \leq i \leq N_x \\ 1 \leq j \leq N_y}} \left| W_{i,j}^n - W_{i,j}^{ex}(t^n) \right|. \tag{4.35c}$$

In order to compute the errors, we have to compute $W_{i,j}^{ex}(t)$, given by the integral from (4.34), for all cells $c_{i,j}$. Such a computation is achieved by using a relevant quadrature rule, of the same order as the scheme, for instance the quadrature rule on a cell given by (2.73) and described in Appendix B. To assess the well-balance and the accuracy of the scheme, we will usually evaluate these errors at the final physical time t_{end}.

Let us recall here that, given Δx and Δy, the time step Δt is given by the CFL-like condition (2.77), as follows:

$$\Delta t \leq \frac{\delta^{\frac{\max(d,3)}{3}}}{2\Lambda},$$

where $\delta = \min(\Delta x, \Delta y)$ and where Λ is the maximum of the absolute values of all the characteristic velocities at each interface between cells.

In this section, we determine whether the two-dimensional scheme satisfies the required properties. Namely, we start by proving that the scheme is well-balanced by direction. First, a 2D steady state at rest with a wet/dry transition is considered. Such a steady state should be preserved by the scheme, since it falls within the scope of Definition 4.2. Second, we focus on a perturbed moving steady state with topography and Manning friction in the x-direction.

This experiment is similar to the one presented for the 1D scheme in Section 3.3.1.4. The 2D scheme should also exactly capture such a steady state since it satisfies Definition 4.2.

Then, we focus on the order of accuracy of the scheme. To that end, we consider two specific exact solutions of the 2D shallow-water model with topography and/or friction. Both of these exact solutions are 2D steady state solutions which do not fall within the framework of Definition 4.2, and therefore are not exactly preserved by the 2D scheme. As a consequence, they are good candidates to test the order of accuracy of the 2D scheme. The first exact solution we consider is a steady state obtained with just the topography source term, and the second one is obtained with both the topography and the friction.

Afterwards, validation experiments are performed. First, we consider a 1D dry dam-break, to evaluate the impact of the convex combination process present in the 2D scheme. In this experiment, several regions coexist: a steady state at rest, an unsteady flow, and a dry area. The goal of this experiment is to study the behavior of the convex combination parameter in such a situation. Second, an experiment analogous to the one presented in Section 3.3.2.6 is considered. The topography for this experiment is a truly 2D function, which also possesses two bumps. Last, we consider a partial dam-break experiment, whose main goal is to study the role of the friction source term.

Finally, we carry out two real-world simulations. The first one is the simulation of the 2011 Japan tsunami. The simulated data is compared to real data, captured by several buoys equipped with tide sensors. The second one concerns a tsunami on an urban topography. It depicts the buildings within a city begin flooded by a tsunami wave, and how the water behaves around the buildings.

4.4.1 Well-balance assessment

In this section, we perform numerical experiments to assess the well-balance of the $\mathbb{P}_5^{\text{WB}}$ scheme. Note that, if the $\mathbb{P}_5^{\text{WB}}$ scheme is well-balanced, then all $\mathbb{P}_d^{\text{WB}}$ schemes with $d < 5$ are also well-balanced. The first experiment concerns the preservation of a lake at rest steady state with a dry area, while the second one focuses on capturing a one-dimensional moving steady state with friction and topography that has been perturbed. Both of these experiments feature steady state solutions by direction; after Theorem 4.6, these steady states should be exactly captured by the $\mathbb{P}_5^{\text{WB}}$ scheme.

4.4.1.1 Steady state at rest

We begin the well-balance numerical experiments with the preservation of a lake at rest steady state. This experiment involves a nonzero Manning coefficient $k = 10$, a non-flat topography and a dry area. On the space domain $[0, 1] \times [0, 1]$, the topography is given by:

$$Z(x, y) = \sqrt{x^2 + y^2}.$$

To ensure that h stays non-negative, the water height and the discharge of this steady state at rest are defined as follows:

$$h(t, x, y) = (1 - Z(x, y))_+ \qquad \text{and} \qquad \boldsymbol{q}(t, x, y) = \mathbf{0}.$$

Both initial and boundary conditions consist in the exact solution. A three-dimensional view of the exact height and the topography is displayed on Figure 4.5. Note that this steady state at rest involves a dry area.

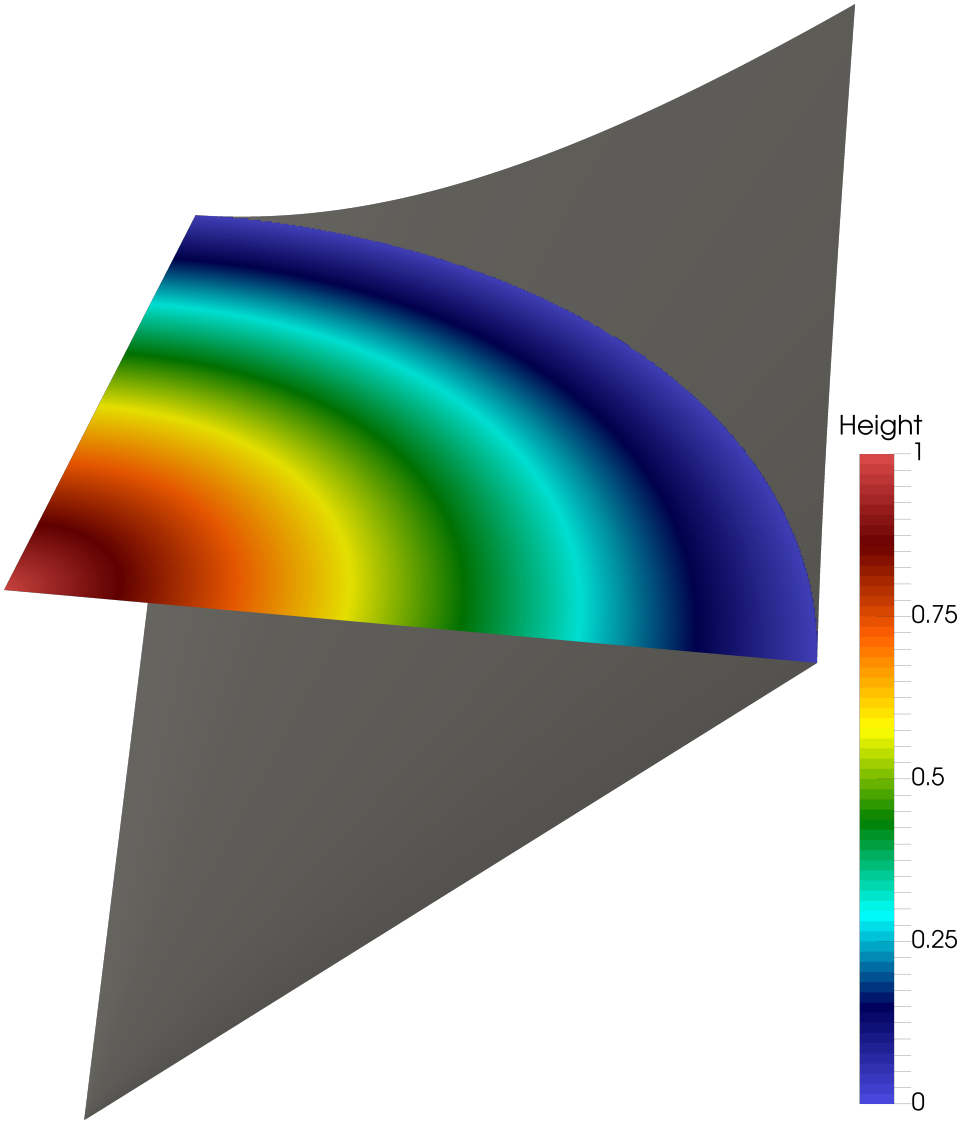

Figure 4.5 – Topography and exact water height for the lake at rest experiment with 250000 (500×500) cells. The gray surface represents the topography.

In order to highlight the relevance of the well-balance correction, the simulation is carried out using the first-order $\mathbb{P}_0$ scheme and the sixth-order $\mathbb{P}_5^{\text{WB}}$ and $\mathbb{P}_5$ schemes, with and without correction. The results of the experiment are presented in Table 4.1, for 10000 (100×100) cells and at time $t_{end} = 0.1$s. For this simulation, we set $C = +\infty$. Moreover, for the $\mathbb{P}_5^{\text{WB}}$ scheme, we set $m_x = m_y = 10^{-12}$, and $M_x = M_y = 10^{-11}$.

On Table 4.1, we observe that the first-order well-balanced scheme, labeled $\mathbb{P}_0$, indeed preserves the lake at rest. However, the sixth-order $\mathbb{P}_5$ scheme, as expected, does not preserve the lake at rest but instead approximates this steady state. The relevance of the proposed well-balance correction is thus highlighted here. Indeed, the sixth-order scheme equipped with the correction, labeled $\mathbb{P}_5^{\text{WB}}$, preserves the lake at rest up to the machine precision.

	$h+Z$			$\|\boldsymbol{q}\|$		
	L^1	L^2	L^∞	L^1	L^2	L^∞
$\mathbb{P}_0$	5.50e-18	2.51e-17	2.22e-16	6.90e-17	1.24e-16	7.68e-16
$\mathbb{P}_5$	1.70e-05	5.81e-05	8.41e-04	2.11e-05	8.86e-05	3.95e-03
$\mathbb{P}_5^{\text{WB}}$	1.81e-17	5.58e-17	7.77e-16	3.16e-16	4.57e-16	3.06e-15

Table 4.1 – Free surface and discharge norm errors for the lake at rest experiment.

4.4.1.2 Moving steady state with topography and friction

We propose another experiment, whose 1D counterpart was presented in Section 3.3.1.4, to assess the well-balance of the scheme. For this experiment, we consider a moving flow of water (i.e. $q_0 \neq 0$), involving both source terms of topography and friction. This moving flow defines a one-dimensional steady state, as per (3.67). The current experiment is set up similarly to the one presented in Section 3.3.1.4. Indeed, we introduce a perturbation of the moving steady state, we take this perturbation as the initial solution, and we carry out the simulation of the dissipation of the perturbation. The schemes converge to the original unperturbed steady state, which should be exactly captured by the well-balanced schemes $\mathbb{P}_0$ and $\mathbb{P}_5^{\text{WB}}$, and approximated by the high-order $\mathbb{P}_5$ scheme without well-balance correction.

This experiment is intended to assess the relevance of the convex combination technique in order to recover the well-balance by direction of the first-order scheme. The experiment presented in the previous section proved that steady states at rest were indeed preserved by the $\mathbb{P}_5^{\text{WB}}$ scheme, and the purpose of the current experiment is to tackle the case of a moving steady state for the topography and friction source terms. We here present the experiment in the x-direction. The same conclusions can be drawn from the experiment in the y-direction, and we do not present this second experiment here.

To set up this experiment, we follow Section 3.3.1.4. First, we look for an approximate solution of the equation (3.67) on the domain $[0,1]$. To address this issue, we set $k = 0.01$ and we take the 1D topography function given by (3.103), as follows:

$$Z(x,y) = \frac{1}{2}\frac{e^{\cos(4\pi x)} - e^{-1}}{e^1 - e^{-1}}.$$

The exact discharge is given by $q_x(t,x,y) = q_0 = 1$ and $q_y(t,x,y) = 0$. Then, to determine the corresponding steady state, we approximately solve (3.67) using Newton's method, in order to find the water height $h^{\text{steady}}(x)$ of the steady state. Here, the water height h^{steady} depends on x, but not on t and y, since we seek a steady state solution in the x-direction. The exact solution ${}^t(h^{\text{steady}}, q_x, q_y) = {}^t(h^{\text{steady}}, q_0, 0)$ indeed defines a steady state in the x-direction after Definition 4.2.

Equipped with the steady state height h^{steady} and discharge q_0, we now introduce a perturbation on the domain

$$\mathcal{P} = \left[\frac{2}{7}, \frac{3}{7}\right] \bigcup \left[\frac{4}{7}, \frac{5}{7}\right].$$

The perturbed initial water height is defined as follows:

$$h(0,x,y) = \begin{cases} h^{\text{steady}}(x) + 0.05 & \text{if } x \in \mathcal{P}, \\ h^{\text{steady}}(x) & \text{otherwise,} \end{cases}$$

while the perturbed initial discharge in the x-direction is given by:

$$q_x(0,x,y) = \begin{cases} q_0 + 0.5 & \text{if } x \in \mathcal{P}, \\ q_0 & \text{otherwise.} \end{cases}$$

The discharge in the y-direction, q_y, is left unperturbed and equal to zero. The initial free surface is displayed on Figure 4.6.

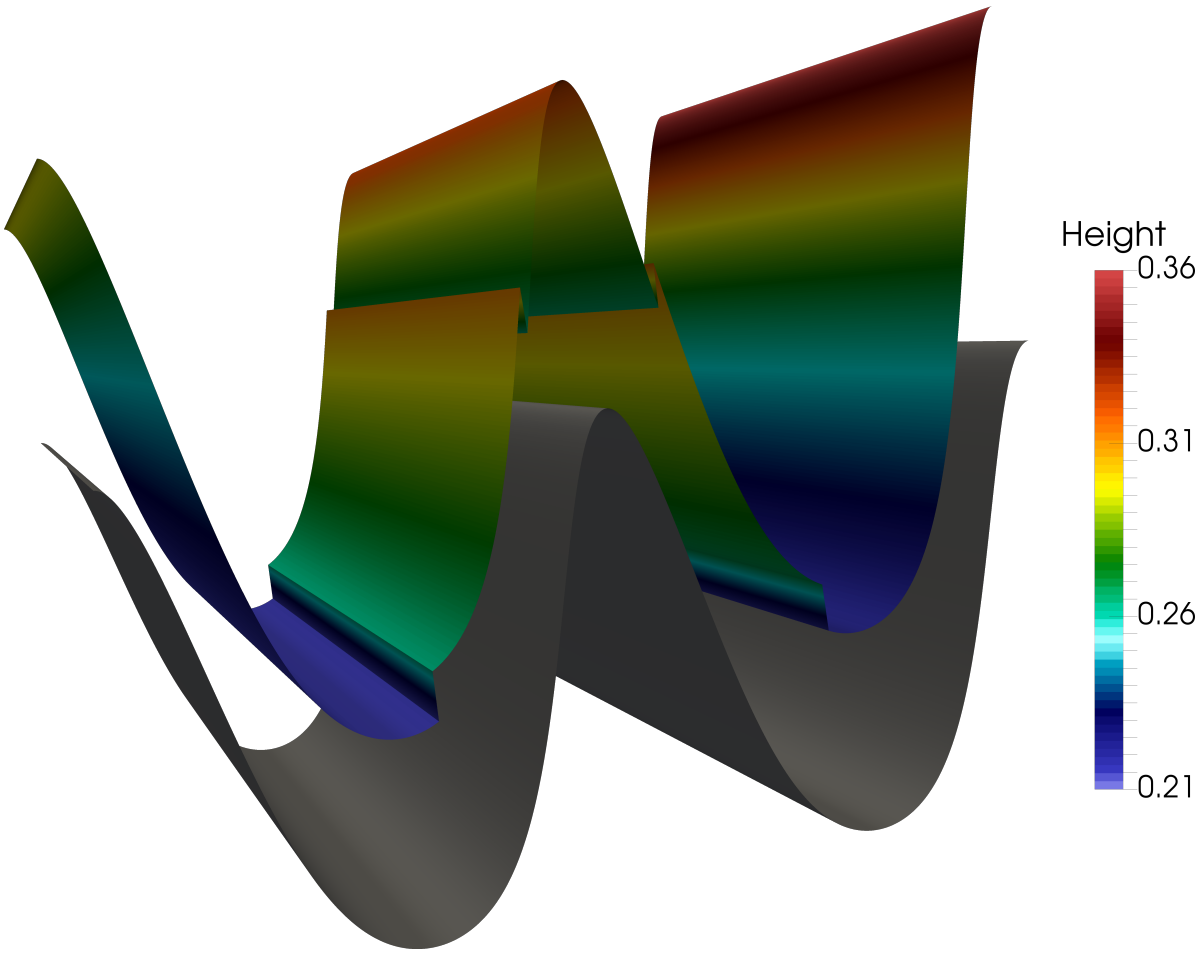

Figure 4.6 – Three-dimensional view of the initial condition for the topography and friction steady state, with $100000 = 1000 \times 100$ cells. The gray surface is the topography. The perturbations are clearly visible on the free surface.

In order to set up the experiment, we set the exact unperturbed solution as the initial and boundary conditions. We take a uniform Cartesian mesh, composed of 300 (100×3) cells, of the domain $[0,1]^2$. The simulation is once again carried out with the $\mathbb{P}_0$, $\mathbb{P}_5$ and $\mathbb{P}_5^{\text{WB}}$ schemes. In addition, we choose $C = +\infty$ for all schemes, and we take $m_x = 0.01$, $M_x = 1$, and $m_y = M_y = 0$ for the $\mathbb{P}_5^{\text{WB}}$ scheme. The results of this simulation are presented at time $t_{end} = 2$s, once the perturbation is fully dissipated, on Table 4.2. Note that, if plotted for a

fixed y coordinate, the free surface over time would be given by Figure 3.18, which has been obtained in the 1D case.

	h			$\|\boldsymbol{q}\|$		
	L^1	L^2	L^∞	L^1	L^2	L^∞
$\mathbb{P}_0$	1.22e-15	1.71e-15	6.27e-15	2.34e-15	3.02e-15	9.10e-15
$\mathbb{P}_5$	5.01e-05	1.47e-04	1.16e-03	2.32e-04	2.63e-04	1.18e-03
$\mathbb{P}_5^{\text{WB}}$	8.50e-14	1.05e-13	3.35e-13	2.82e-13	3.37e-13	6.76e-13

Table 4.2 – Height and discharge norm errors for the topography and friction steady state along the x axis.

Once again, this experiment emphasizes the relevance of the well-balance correction. Indeed, the $\mathbb{P}_5$ scheme does not exactly capture the unperturbed steady state, while the $\mathbb{P}_5^{\text{WB}}$ scheme captures it exactly, like the $\mathbb{P}_0$ scheme.

We also make the important remark that this well-balance correction also reduces the CPU time, which decreases from 1463s with the $\mathbb{P}_5$ scheme to 707s with the $\mathbb{P}_5^{WB}$ scheme. Indeed, by downgrading to the first-order scheme when the approximate solution is close to a steady state, the $\mathbb{P}_5^{WB}$ scheme manages to be both faster and more accurate than the uncorrected $\mathbb{P}_5$ scheme.

4.4.2 Order of accuracy assessment

We now turn to verifying the order of accuracy of the high-order scheme. As previously mentioned, this check is done using truly 2D steady state solutions, which are not steady states by direction and thus do not fall within the scope of Definition 4.2. This choice is made to ensure that an exact solution is known. Indeed, we can derive truly 2D steady state solutions by choosing a discharge field that satisfies (4.2), i.e. whose divergence vanishes.

In order to compute the order of accuracy, we consider the results from two simulations, one carried out on a mesh composed of N discretization cells, and the other one with $N' > N$ cells. The errors are then computed according to (4.35). Let e_N be the value of error, in any of the three norms, for a mesh with N cells. The order of accuracy p is then defined as follows:

$$p = -\frac{\ln(e_N) - \ln(e_{N'})}{\ln\sqrt{N} - \ln\sqrt{N'}}. \tag{4.36}$$

In order to have a relevant computation of the order of accuracy, we take $N' = 4N$ in (4.36). Thus, the definition of the order of accuracy used in this section is the following:

$$p = \frac{\ln(e_N) - \ln(e_{4N})}{\ln 2}. \tag{4.37}$$

In this section, we suggest two different 2D steady state solutions. The first one is obtained by assuming a vanishing friction contribution, while the second one is computed with both source terms. For both of these solutions, we compute the order of accuracy of the schemes according to (4.37).

4.4.2.1 Steady vortex experiment

The first experiment we consider involves a steady state solution with a vanishing friction contribution (i.e. $k = 0$), the *steady vortex* (see [47]). We set $\boldsymbol{r} = {}^t(x, y)$ and we take a radial topography, given by $Z(x, y) = 0.2e^{0.5(1-\|\boldsymbol{r}\|^2)}$. Then, the water height is defined as follows:

$$h(t, x, y) = 1 - \frac{1}{4g} e^{2(1-\|\boldsymbol{r}\|^2)} - Z(x, y),$$

and the x- and y-velocities are given by:

$$u(t, x, y) = y\, e^{1-\|\boldsymbol{r}\|^2} \quad \text{and} \quad v(t, x, y) = -x\, e^{1-\|\boldsymbol{r}\|^2}.$$

For such initial data, the discharge $\boldsymbol{q} = {}^t(hu, hv)$ indeed satisfies (4.2), but it is not a constant. This steady state is depicted on Figure 4.7 on the space domain $[-3, 3]^2$.

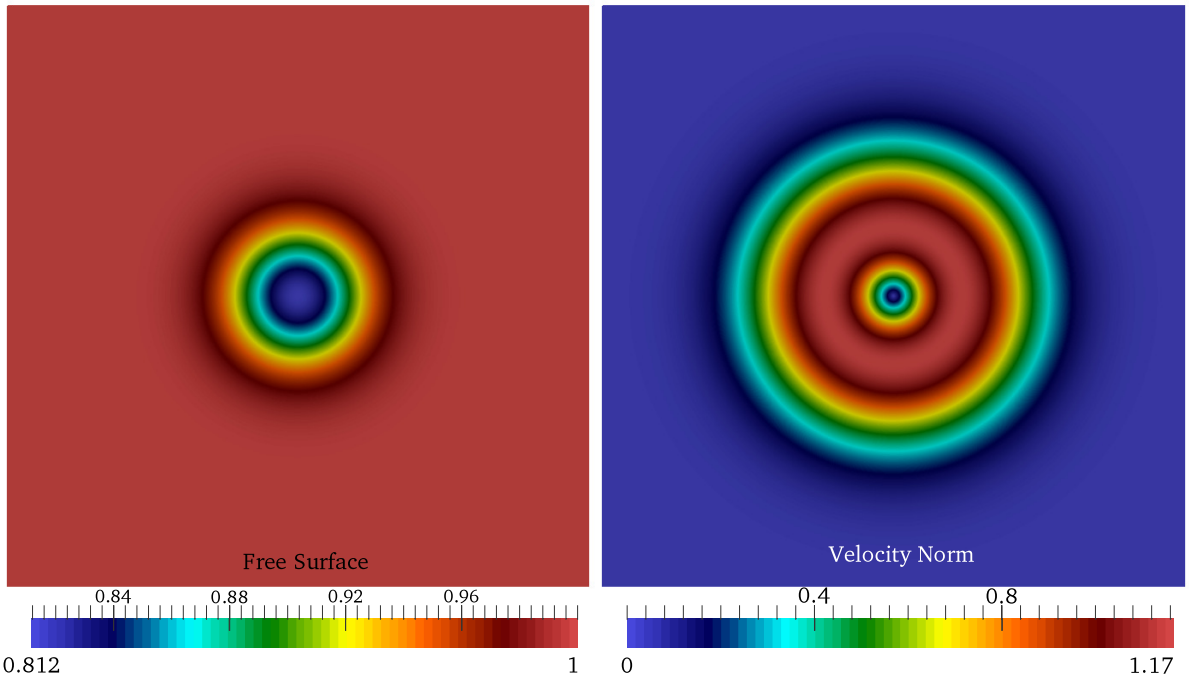

Figure 4.7 – Steady vortex. Left panel: free surface. Right panel: velocity norm (the vortex flows clockwise).

The simulations are carried out with the $\mathbb{P}_3^{\text{WB}}$, $\mathbb{P}_4^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes, until a final physical time $t_{end} = 1$s. In addition, we take $C = +\infty$ and $m_x = M_x = m_y = M_y = +\infty$ for the three schemes. The results of the simulations are presented in:

- Table 4.3 and Table 4.4 for $d = 3$;
- Table 4.5 and Table 4.6 for $d = 4$;
- Table 4.7 and Table 4.8 for $d = 5$.

These results show good agreement with the theory. Indeed, in all cases, the order of accuracy is roughly equal to $d + 1$, as expected. This order of accuracy is maintained thanks to the u2 detection criteria. Indeed, on such smooth solutions, the DMP criterion (2.78) is not sufficient, for it would wrongly lower the CPD in some cells. Here, the smoothness detector (2.81) is used to prevent over-detection from the DMP criterion. The reader is referred to [47]

N	h, L^1		h, L^2		h, L^∞	
900	5.90e-05	—	1.06e-04	—	5.94e-04	—
3600	3.32e-06	4.15	5.12e-06	4.37	2.92e-05	4.35
14400	1.90e-07	4.12	2.80e-07	4.19	1.44e-06	4.34
57600	1.16e-08	4.04	1.75e-08	4.00	1.35e-07	3.41

Table 4.3 – Height error for the steady vortex experiment using the $\mathbb{P}_3^{\text{WB}}$ scheme.

N	$\|\boldsymbol{q}\|$, L^1		$\|\boldsymbol{q}\|$, L^2		$\|\boldsymbol{q}\|$, L^∞	
900	3.34e-04	—	6.25e-04	—	3.18e-03	—
3600	2.08e-05	4.01	3.68e-05	4.08	2.15e-04	3.89
14400	1.21e-06	4.10	2.01e-06	4.20	8.82e-06	4.61
57600	7.38e-08	4.04	1.18e-07	4.08	5.07e-07	4.12

Table 4.4 – Discharge norm error for the steady vortex experiment using the $\mathbb{P}_3^{\text{WB}}$ scheme.

N	h, L^1		h, L^2		h, L^∞	
900	8.87e-05	—	1.85e-04	—	1.65e-03	—
3600	3.96e-06	4.49	8.99e-06	4.37	9.32e-05	4.14
14400	1.44e-07	4.78	3.20e-07	4.81	2.94e-06	4.98
57600	5.62e-09	4.68	1.16e-08	4.78	7.97e-08	5.21

Table 4.5 – Height error for the steady vortex experiment using the $\mathbb{P}_4^{\text{WB}}$ scheme.

N	$\|\boldsymbol{q}\|$, L^1		$\|\boldsymbol{q}\|$, L^2		$\|\boldsymbol{q}\|$, L^∞	
900	5.94e-04	—	1.12e-03	—	5.83e-03	—
3600	2.52e-05	4.56	5.41e-05	4.37	4.03e-04	3.86
14400	8.54e-07	4.88	1.86e-06	4.86	1.48e-05	4.76
57600	2.87e-08	4.89	5.99e-08	4.95	4.98e-07	4.90

Table 4.6 – Discharge norm error for the steady vortex experiment using the $\mathbb{P}_4^{\text{WB}}$ scheme.

N	h, L^1		h, L^2		h, L^∞	
900	2.04e-05	—	5.22e-05	—	7.84e-04	—
3600	3.07e-07	6.05	6.88e-07	6.25	9.94e-06	6.30
14400	3.93e-09	6.29	5.82e-09	6.88	5.53e-08	7.49
57600	5.74e-11	6.10	7.27e-11	6.32	3.30e-10	7.39

Table 4.7 – Height error for the steady vortex experiment using the $\mathbb{P}_5^{\text{WB}}$ scheme.

N	$\|\boldsymbol{q}\|$, L^1		$\|\boldsymbol{q}\|$, L^2		$\|\boldsymbol{q}\|$, L^∞	
900	1.37e-04	—	3.46e-04	—	2.90e-03	—
3600	1.90e-06	6.17	5.27e-06	6.04	5.10e-05	5.83
14400	2.33e-08	6.35	5.33e-08	6.63	4.98e-07	6.68
57600	3.08e-10	6.24	5.76e-10	6.53	4.42e-09	6.82

Table 4.8 – Discharge norm error for the steady vortex experiment using the $\mathbb{P}_5^{\text{WB}}$ scheme.

for a comparison of the order with and without the u2 criterion. In [47], the authors show that it is necessary to use the u2 criterion for this experiment in order to recover the expected order of accuracy.

Error graphs in L^2-norm are provided on Figure 4.8 for the $\mathbb{P}_3^{\text{WB}}$ and the $\mathbb{P}_5^{\text{WB}}$ schemes. We can clearly observe on this figure that the $\mathbb{P}_3^{\text{WB}}$ scheme is roughly of order 4 and the $\mathbb{P}_5^{\text{WB}}$ scheme is roughly of order 6. Indeed, in logarithmic scale, the slope of the error with respect to the number of discretization cells corresponds to the order of the scheme.

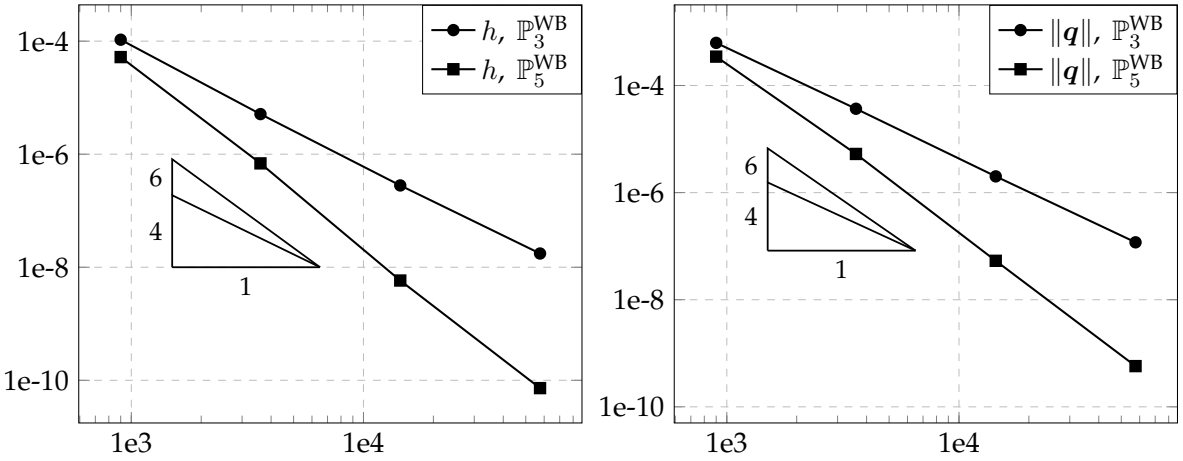

Figure 4.8 – Error plots for the steady vortex experiment, in L^2-norm, for the $\mathbb{P}_3^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes. Left panel: water height errors; right panel: discharge errors.

4.4.2.2 2D steady state with topography and friction

We now turn to another 2D steady state solution. This new steady state is obtained by considering both contributions of topography and friction (i.e. $k \neq 0$). For this solution, we assume that $\|\boldsymbol{r}\| \neq 0$, and we take the following topography function:

$$Z(x,y) = \frac{2k\|\boldsymbol{r}\| - 1}{2g\|\boldsymbol{r}\|^2}.$$

In addition, the exact water height and discharge are given by:

$$h(t,x,y) = 1 \qquad \text{and} \qquad \boldsymbol{q}(t,x,y) = \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|^2}.$$

Note that such a definition of the discharge ensures that (4.2) is satisfied, i.e. that $\boldsymbol{\nabla}\cdot\boldsymbol{q} = 0$. On the space domain $[-0.3, 0.3] \times [0.4, 1]$ and for $k = 10$, the topography is depicted on Figure 4.9, while Figure 4.10 shows the discharge field in both directions.

In order to highlight the high-order accuracy of the schemes, this experiment is carried out with the $\mathbb{P}_3^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes. The final physical time is $t_{end} = 0.1$s, and we take once again $C = +\infty$ and $m_x = M_x = m_y = M_y = +\infty$. We again take the exact solution as initial and boundary conditions. The results of the simulations are presented in:

- Table 4.9, Table 4.10, and Table 4.11 for $d = 3$;
- Table 4.12, Table 4.13, and Table 4.14 for $d = 5$.

N	h		q_x		q_y	
900	5.00e-07	—	1.95e-06	—	2.00e-06	—
3600	3.12e-08	4.00	1.22e-07	4.00	1.24e-07	4.00
14400	1.91e-09	4.03	7.54e-09	4.02	7.64e-09	4.02
57600	1.17e-10	4.03	4.65e-10	4.02	4.70e-10	4.02

Table 4.9 – L^1 errors for the friction and topography 2D steady state using the $\mathbb{P}_3^{\text{WB}}$ scheme.

N	h		q_x		q_y	
900	7.59e-07	—	2.94e-06	—	2.61e-06	—
3600	4.54e-08	4.06	1.75e-07	4.07	1.61e-07	4.02
14400	2.60e-09	4.13	1.05e-08	4.06	9.82e-09	4.04
57600	1.48e-10	4.13	6.37e-10	4.04	6.01e-10	4.03

Table 4.10 – L^2 errors for the friction and topography 2D steady state using the $\mathbb{P}_3^{\text{WB}}$ scheme.

N	h		q_x		q_y	
900	5.63e-06	—	1.06e-05	—	1.07e-05	—
3600	4.42e-07	3.67	6.39e-07	4.05	7.44e-07	3.85
14400	3.15e-08	3.81	3.68e-08	4.12	4.45e-08	4.06
57600	2.12e-09	3.89	2.14e-09	4.10	2.64e-09	4.07

Table 4.11 – L^∞ errors for the friction and topography 2D steady state using the $\mathbb{P}_3^{\text{WB}}$ scheme.

N	h		q_x		q_y	
900	2.37e-08	—	8.00e-08	—	1.12e-07	—
3600	3.77e-10	5.98	1.28e-09	5.96	1.82e-09	5.94
14400	5.89e-12	6.00	1.99e-11	6.01	2.91e-11	5.96
57600	1.24e-14	8.89	2.06e-13	6.60	1.20e-13	7.92

Table 4.12 – L^1 errors for the friction and topography 2D steady state using the $\mathbb{P}_5^{\text{WB}}$ scheme.

N	h		q_x		q_y	
900	3.20e-08	—	1.30e-07	—	1.48e-07	—
3600	5.07e-10	5.98	2.05e-09	5.98	2.40e-09	5.94
14400	7.98e-12	5.99	3.17e-11	6.02	3.84e-11	5.97
57600	5.31e-14	7.23	5.08e-13	5.96	3.67e-13	6.71

Table 4.13 – L^2 errors for the friction and topography 2D steady state using the $\mathbb{P}_5^{\text{WB}}$ scheme.

N	h		q_x		q_y	
900	1.04e-07	—	5.20e-07	—	5.57e-07	—
3600	1.80e-09	5.86	8.15e-09	6.00	1.02e-08	5.77
14400	3.38e-11	5.73	1.25e-10	6.02	1.71e-10	5.89
57600	8.33e-13	5.34	2.26e-12	5.79	2.59e-12	6.05

Table 4.14 – L^∞ errors for the friction and topography 2D steady state using the $\mathbb{P}_5^{\text{WB}}$ scheme.

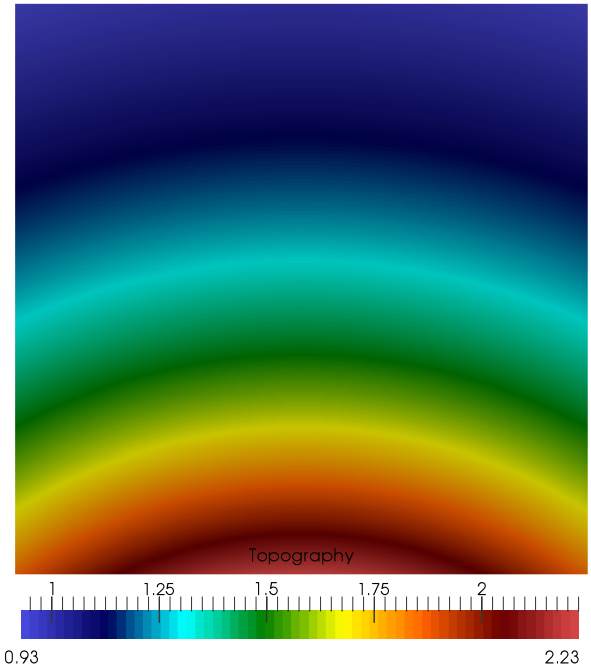

Figure 4.9 – Topography for the 2D steady state with topography and friction.

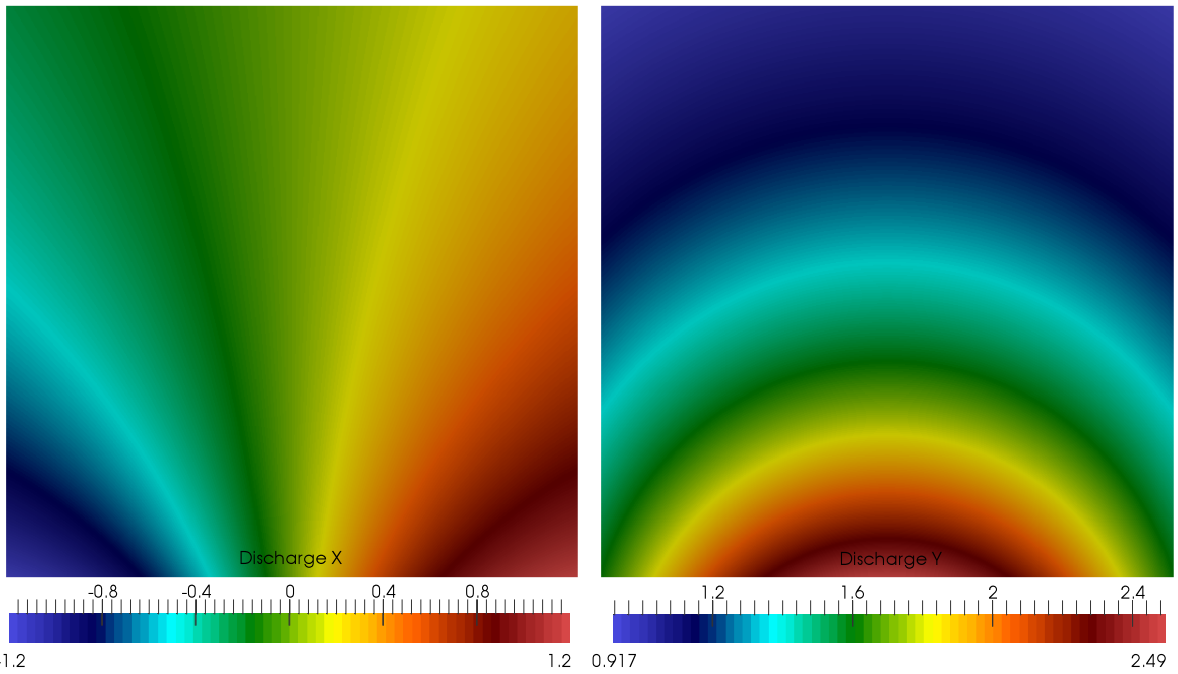

Figure 4.10 – Discharge for the 2D steady state with topography and friction. Left panel: discharge in the x-direction. Right panel: discharge in the y-direction.

Once again, we recover the expected order of accuracy, roughly equal to $d+1$. Similarly to the previous experiment, this order of accuracy is recovered only thanks to the use of the u2 criterion in addition to the DMP criterion. We also present the following error graphs:

- the error for the water height in all norms is depicted on Figure 4.11;
- the error for the discharge in both directions, in the L^2-norm, is displayed on Figure 4.12.

On these figures, the orders of accuracy of the schemes are clearly visible. For h, q_x and q_y, the $\mathbb{P}_3^{\text{WB}}$ scheme is of order 4, and the $\mathbb{P}_5^{\text{WB}}$ scheme is of order 6.

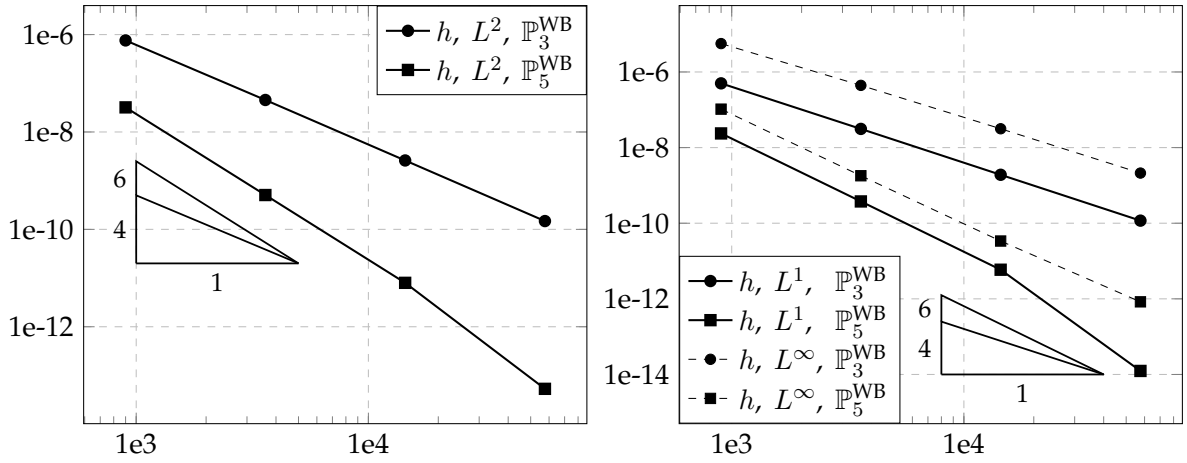

Figure 4.11 – Water height error plots for the steady vortex experiment, for the $\mathbb{P}_3^{\mathrm{WB}}$ and $\mathbb{P}_5^{\mathrm{WB}}$ schemes. Left panel: errors in the L^2-norm; right panel: errors in the L^1- and L^∞-norms.

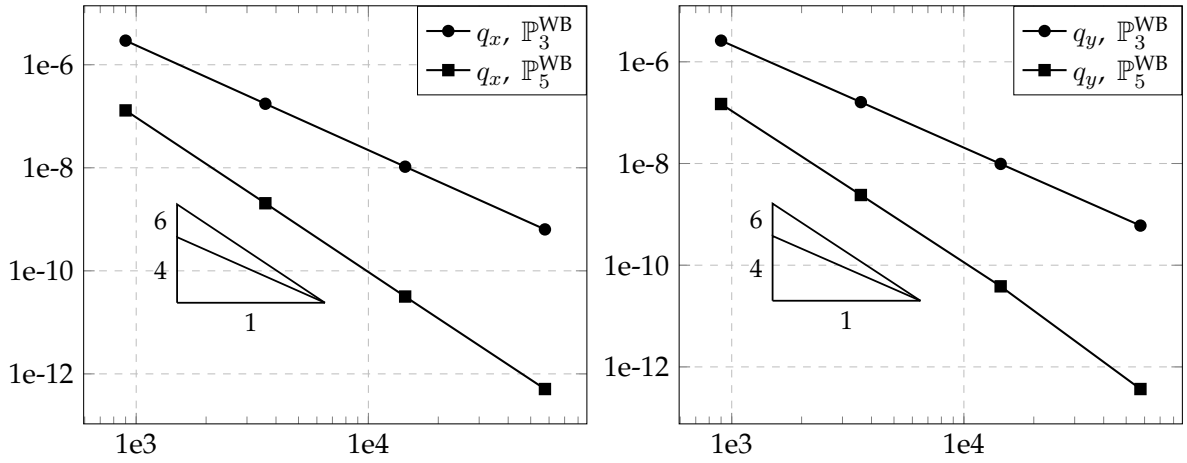

Figure 4.12 – Steady vortex experiment: error plots, in L^2-norm, for the x-discharge and for the y-discharge. Left panel: errors for the x-discharge; right panel: errors for the y-discharge.

4.4.3 Validation experiments

After having presented the well-balance property and the high-order accuracy satisfied by the suggested scheme, we now turn to its numerical validation, by focusing on more complex experiments. First, we present the simulation of a dam-break over a dry bottom in one space direction. This simulation also highlights the relevance of the well-balance correction and of the MOOD procedure. Next, we present a two-dimensional dam-break simulation, on a topography involving two bumps. Afterwards, we present a two-dimensional partial dam-break.

4.4.3.1 Dry dam-break

This subsection focuses on a double dam-break over a dry sinusoidal bottom. The space domain is $[0, 1] \times [0, 0.1]$, and the topography is chosen as follows:

$$Z(x, y) = \frac{1}{2}\cos^2(2\pi x).$$

The initial free surface consists in a double dam-break; it is given by:

$$h(0,x,y) + Z(x,y) = \begin{cases} 2 & \text{if } x \in \mathcal{D}, \\ Z(x,y) & \text{otherwise,} \end{cases}$$

where the domain $\mathcal{D}$ is defined by:

$$\mathcal{D} = \left[0, \frac{1}{3}\right] \bigcup \left[\frac{2}{3}, 1\right].$$

In addition, the initial discharge is zero: $\boldsymbol{q}(0,x,y) = \mathbf{0}$. Note that such a free surface corresponds to a steady state at rest on $\mathcal{D} \times [0, 0.1]$. In $(\{1/3\} \times [0, 0.1]) \cup (\{2/3\} \times [0, 0.1])$, the free surface is discontinuous, thus producing the initial dam-break conditions. The domain $([0,1] \setminus \mathcal{D}) \times [0, 0.1]$ is dry. As a consequence, this experiment will highlight three crucial parts of the scheme: the well-balance property, the ability to handle dry/wet and wet/dry transitions, and the consistency with the 2D shallow-water equations. This experiment is presented in the x-direction for simplicity, but can be carried out in the y-direction or even in a transverse direction, yielding the same conclusions.

For this experiment, the Manning coefficient k is set to 10 and the boundaries are assumed to be solid walls, i.e. we set $q_x(t,0,y) = q_x(t,1,y) = 0$ and $q_y(t,x,0) = q_y(t,x,0.1) = 0$ for all t, x and y. The experiment is carried out with the $\mathbb{P}_0$ and $\mathbb{P}_5^{\text{WB}}$ schemes, to compare the first-order scheme with the high-order well-balanced scheme. The final physical time is $t_{end} = 0.03$s, and we set $C = 7.5$, $m_x = m_y = 10^{-10}$, and $M_x = M_y = 0.5$. The results are presented on Figure 4.13 and on Figure 4.14.

- Figure 4.13 displays a comparison between the results obtained with the $\mathbb{P}_0$ scheme and those obtained with the $\mathbb{P}_5^{\text{WB}}$ scheme, with $200 = 100 \times 2$ cells in each case. We also display a reference solution, obtained using the $\mathbb{P}_0$ scheme with $8000 = 4000 \times 2$ discretization cells.
- On Figure 4.14, CPD(i,j) and $(\theta_x)_{i,j}^n$ are depicted, as well as the free surface and the topography, for $t = t_{end}/10$ and $t = t_{end}$, with the $\mathbb{P}_5^{\text{WB}}$ scheme.

Figure 4.13 highlights the relevance of using a high-order well-balanced scheme for such an experiment. First, the results from the $\mathbb{P}_5^{\text{WB}}$ scheme are visibly closer to the reference solution than those of the $\mathbb{P}_0$ scheme. Moreover, the approximations of the interfaces between dry and wet areas are in good agreement with the reference solution. In addition, note that the free surface should be unperturbed close to the edges of the domain. Indeed, the waves from the dam-break have not yet reached the edges of the domain at $t = t_{end}$, and the water close enough to the edges is in a lake at rest configuration. This essential property exactly holds for the $\mathbb{P}_0$ scheme. It also holds for the $\mathbb{P}_5^{\text{WB}}$ scheme thanks to the well-balance correction, which forces the well-balanced scheme to be used in lake at rest-type situations. Figure 4.14 displays more details on the role played by the well-balance convex combination.

On the left panel of Figure 4.14, we observe that $(\theta_x)_{i,j}^n$ is zero in areas that have not yet been impacted by the waves, i.e. in the areas where a lake at rest configuration is found. As a consequence, in these areas (namely the center of the domain and close to its edges),

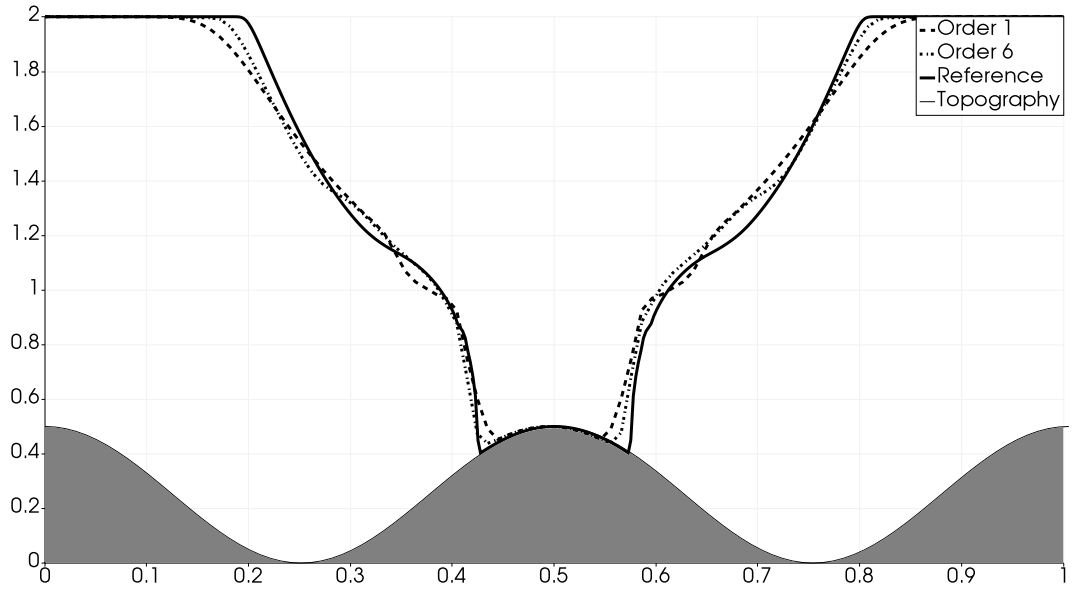

Figure 4.13 – Free surface for the dam-break over a dry sinusoidal bottom: reference solution and results of the $\mathbb{P}_0$ and $\mathbb{P}_5^{\text{WB}}$ schemes. The gray area represents the topography.

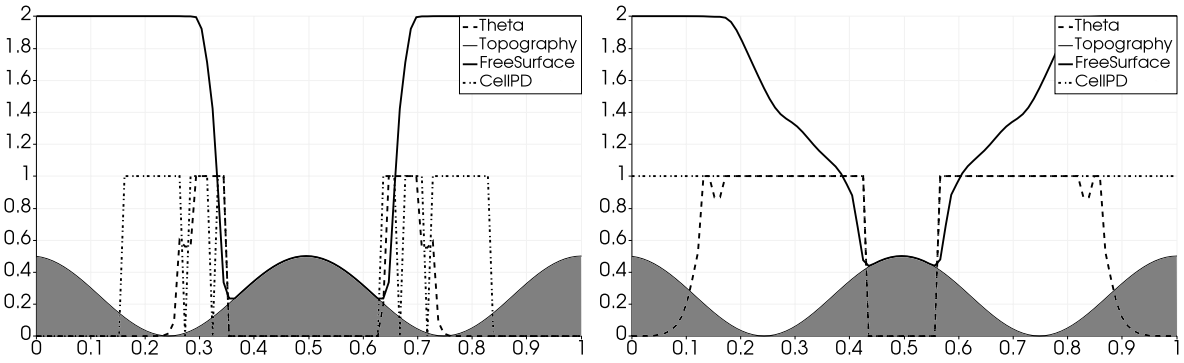

Figure 4.14 – Free surface, CPD map and convex combination coefficient θ_x for the the dam-break over a dry sinusoidal bottom with the $\mathbb{P}_5^{\text{WB}}$ scheme. The gray area represents the topography. Left panel: $t = 3.10^{-3}$s. Right panel: $t = 3.10^{-2}$s.

the CPD is equal to 0 and the well-balanced scheme is used. We also notice that the CPD is zero in two cells within each wave. Indeed, in those cells, the DMP detector (2.78) has been triggered. Similar conclusions can be drawn from the right panel of Figure 4.14. The center of the domain is still considered to be at rest, and the convex combination parameter is very close to zero on the edges of the domain, where the water is almost at rest. We do not have $(\theta_x)_{i,j}^n = 0$ in those cells because the numerical diffusion created by the waves has been introducing small perturbations that travel faster than the actual waves. The amplitude of these perturbations is high enough to cause the steady state error (4.27) to be greater than 0. However, since $(\theta_x)_{i,j}^n < 1$, the steady state error is still lower than M_x near the edges of the domain.

4.4.3.2 2D dam-break with two bumps

The second 2D experiment is a dry dam-break with a topography presenting two bumps. It is heavily inspired from an experiment presented in [19], which did not include the friction source term. The Manning coefficient is $k = 0.1$, and the topography function is given by

$$Z(x,y) = \frac{1}{2}\left(1 - 25\left(\left(x - \frac{5}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2\right)\right)_+ + 2\left(1 - 25\left((x-4)^2 + \left(y - \frac{1}{2}\right)^2\right)\right)_+.$$

The space domain is $[0,5] \times [0,1]$. The initial discharge is zero in both directions, i.e. we take $\boldsymbol{q}(0,x,y) = \mathbf{0}$, and the initial water height is given by

$$h(0,x,y) = \begin{cases} 6 & \text{if } x < 0.7, \\ 0 & \text{otherwise.} \end{cases}$$

The simulation runs until a physical time $t_{end} = 1.35$s with the $\mathbb{P}_1^{\text{WB}}$ scheme, using $C = 1$, $m_x = m_y = 10^{-5}$ and $M_x = M_y = 25$. We take 288000 discretization cells (1200 in the x direction and 240 in the y direction). In addition, we prescribe wall boundary conditions. The results are presented on Figure 4.15, Figure 4.16, Figure 4.17, and Figure 4.18.

This experiment has been carried out to make sure that the numerical scheme still behaves correctly in a truly 2D setting and in the presence of dry/ wet transition. We recover a numerical solution involving the friction source term, which can be compared to the numerical solution without friction presented in [19]. In addition, this 2D experiment is similar to the 1D double bump experiment we presented in Section 3.3.2.6. Indeed, the behavior of the water before it comes into contact with the first bump is the same in both experiments.

4.4.3.3 Partial dam-break

The last dam-break experiment is a two-dimensional partial dam-break (see [126, 47]). An extensive study of this experiment, focused on the differences between various reconstruction degrees and MOOD criteria, has been presented in [47]. However, in [47], the friction source term was not present, and the authors studied the effects of the topography only. Thus, our study is mainly focused on the effects of the friction source term. To that end, we carry out the simulation with three different Manning coefficients.

For this experiment, the space domain is $[-100,100] \times [-100,100]$, and we take the following topography function:

$$Z(x,y) = \begin{cases} 1 & \text{if } x \leq -5, \\ 0 & \text{if } x \geq 5, \\ 0.1(5-x) & \text{if } -5 < x < 5 \text{ and } -40 < y < 40, \\ 12 & \text{if } -5 < x < 5 \text{ and } y \in [-100,-40] \cup [40,100]. \end{cases}$$

Hence, it represents a 12 meters high, 10 meters wide broken dam. Initially, the reservoir (to

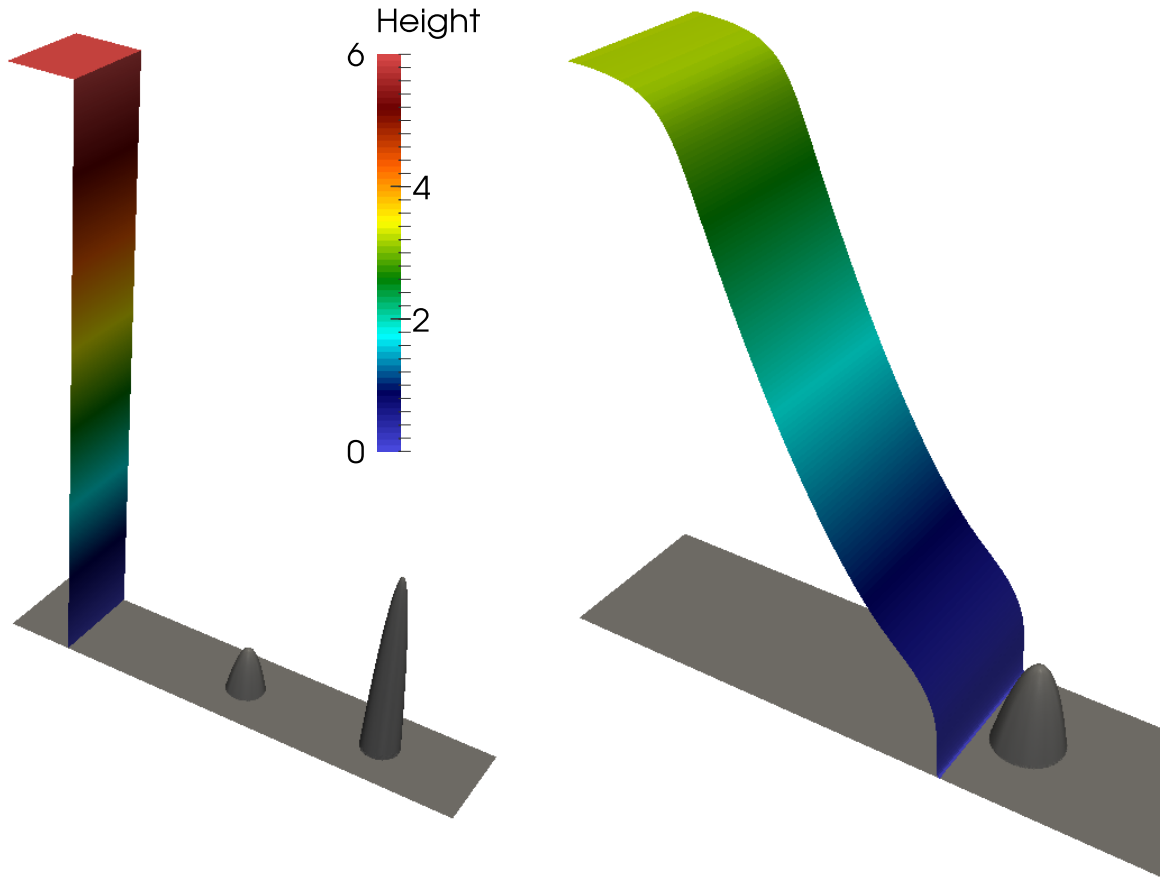

Figure 4.15 – Left panel: initial condition of the 2D dam-break over a double bump experiment. Note that the same color scale for the water height is used for Figure 4.15, Figure 4.16, Figure 4.17, and Figure 4.18, and that the solid gray color represents the topography. Right panel: approximate solution at $t = 0.15$s, just before the water hits the first bump. Note the shape of the front of the water, due to the nonzero bottom friction.

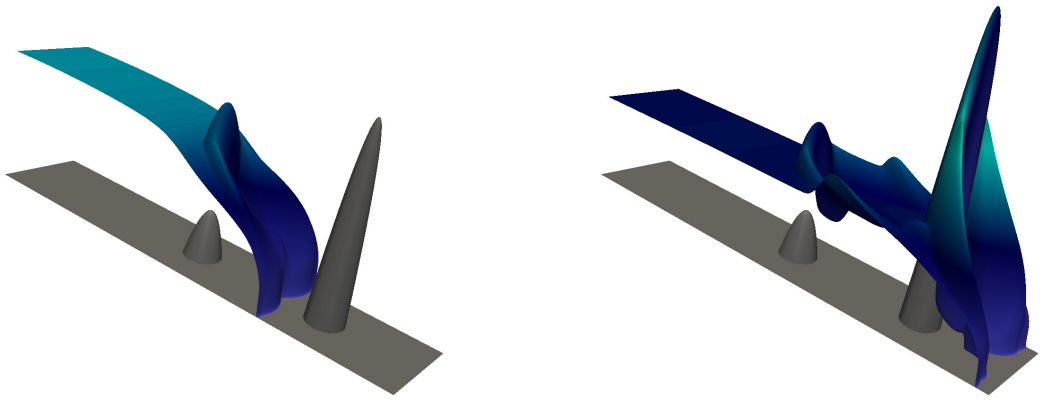

Figure 4.16 – Approximate solution of the 2D dam-break over a double bump experiment, displayed at times $t = 0.3$s (left panel) and $t = 0.45$s (right panel).

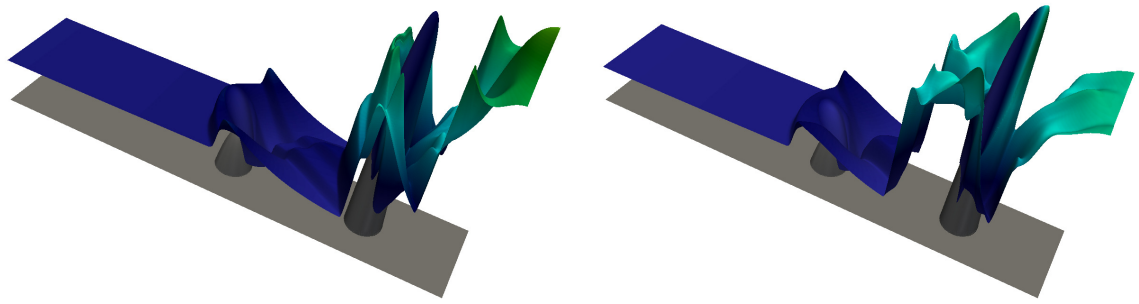

Figure 4.17 – Approximate solution of the 2D dam-break over a double bump experiment, displayed at times $t = 0.75$s (left panel) and $t = 0.9$s (right panel).

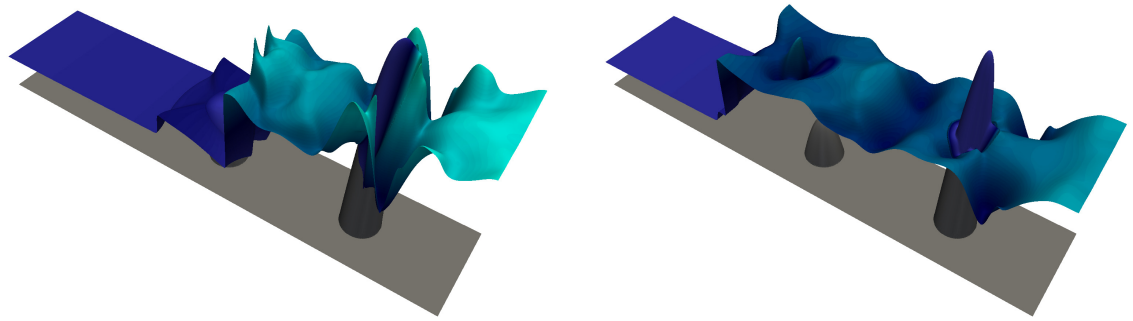

Figure 4.18 – Approximate solution of the 2D dam-break over a double bump experiment, displayed at times $t = 1.05$s (left panel) and $t = 1.35$s (right panel).

the left) is filled, as follows:

$$h(0,x,y) = \begin{cases} 10 - Z(x,y) & \text{if } x \leq -5, \\ 5 - Z(x,y) & \text{if } x \geq 5, \\ 5 - Z(x,y) & \text{if } -5 < x < 5 \text{ and } -40 < y < 40, \\ 0 & \text{if } -5 < x < 5 \text{ and } y \in [-100,-40] \cup [40,100]. \end{cases}$$

The water is initially at rest, i.e. $\boldsymbol{q}(0,x,y) = \mathbf{0}$. In addition, we use homogeneous Neumann boundary conditions, and we take the final physical time $t_{end} = 7$s

In order for the simulation to be relevant, we elected to set $\text{CPD}(i,j) = 0$ for cells where the topography gradient is too large. In order to conserve the high-order behavior of the scheme, we only set $\text{CPD}(i,j) = 0$ for cells possessing at least one vertex that belongs to the dam, where $Z(x,y) = 12$. Indeed, in such cells, the high-order approximation of the topography source term (4.20a) becomes too large, leading to spurious oscillations in their vicinity.

Influence of the friction coefficient

We now compare the results from the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes with various Manning coefficients, namely $k = 0$, $k = 0.25$ and $k = 2$. All of these comparisons have been carried

out using $40000 = 200 \times 200$ discretization cells. In addition, we set $C = 0.5$, $m_x = m_y = 10^{-10}$, and $M_x = M_y = 0.5$. The results of the simulations are displayed on the following figures, with the same color scale:

- Figure 4.19 for the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes with $k = 0$;
- Figure 4.20 for the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes with $k = 0.25$;
- Figure 4.21 for the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes with $k = 2$.

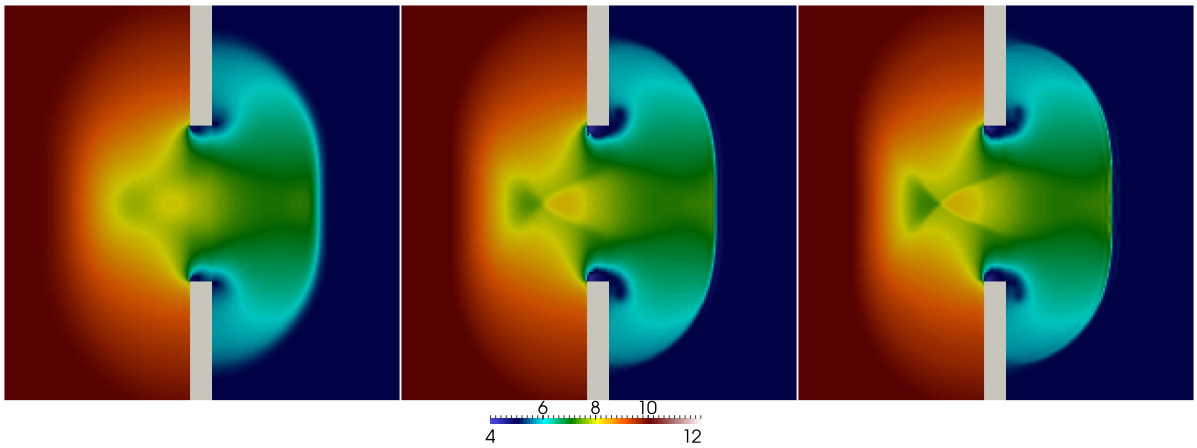

Figure 4.19 – Free surface for the partial dam-break simulation with $k = 0$. From left to right: results of the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes.

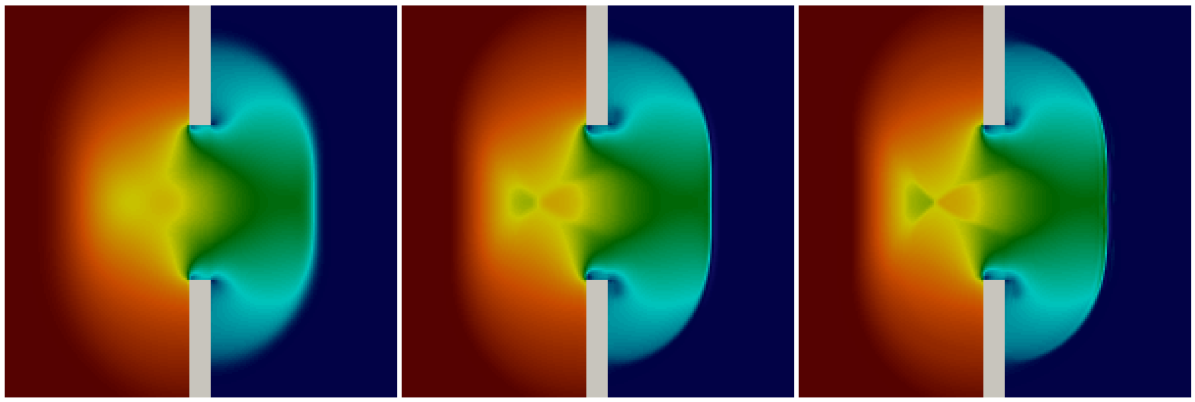

Figure 4.20 – Free surface for the partial dam-break simulation with $k = 0.25$. From left to right: results of the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes.

On all three figures, we observe important differences between the results of the three schemes. Indeed, for instance on Figure 4.19 and Figure 4.20, we observe that the shock wave to the right of the dam and the rarefaction wave to the left of the dam are visibly more smeared when using the $\mathbb{P}_0$ scheme instead of the $\mathbb{P}_1^{\text{WB}}$ or the $\mathbb{P}_5^{\text{WB}}$ scheme. In addition, the structure at the center of the water flow is not visible with the $\mathbb{P}_0$ scheme. It becomes visible, although smeared, with the $\mathbb{P}_1^{\text{WB}}$ scheme, and it is very well-defined with the $\mathbb{P}_5^{\text{WB}}$ scheme. For Figure 4.21, the conclusions are similar. The smearing of the shock wave and the rarefaction wave is still present unless a high-order scheme is used, but the presence of an important friction has caused the central structure to nearly disappear.

An important remark we make here concerns the vortices present at the edges of the dam in Figure 4.19 and Figure 4.20. First, note that the presence of the friction source term dampens

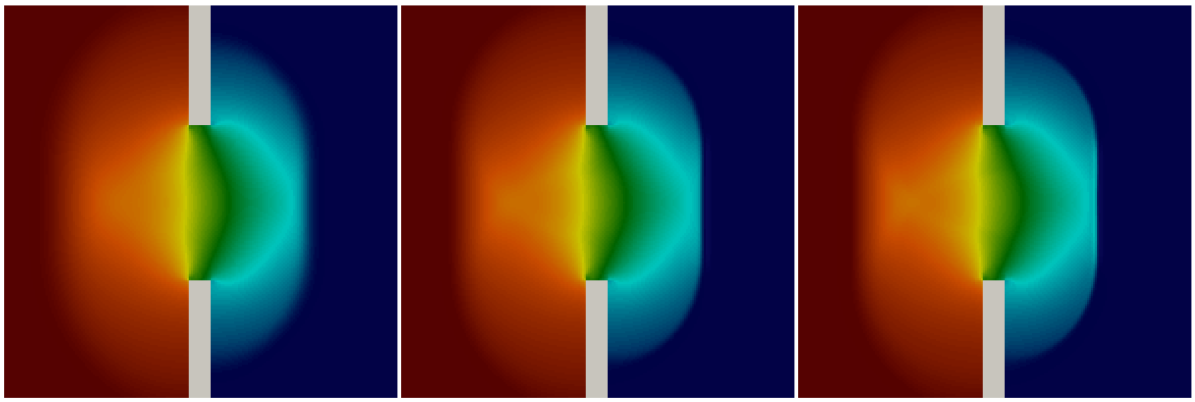

Figure 4.21 – Free surface for the partial dam-break simulation with $k = 2$. From left to right: results of the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_5^{\text{WB}}$ schemes.

the depth, as well as the size, of these vortices. This behavior is highlighted in Table 4.15, in which the approximate size and the depth of a vortex are collected. For this table, we focused on the top vortex, whose characteristics are similar to the bottom one since the experiment is symmetric with respect to the $y = 0$ line.

Manning coefficient	Vortex size	Water depth
$k = 0$	84m^2	4.28m
$k = 0.25$	17m^2	5.45m
$k = 2$	0m^2	7.23m

Table 4.15 – Water and approximate size for the deepest vortex, for the $\mathbb{P}_5^{\text{WB}}$ scheme. For the case where $k = 2$, there is no vortex, and the table displays the free surface at the point where the vortex would be located if the Manning coefficient were lower.

The last part of the study of this experiment is the analysis, over the $y = 0$ line, of the left rarefaction wave and the right shock wave. Some relevant quantities are the position of the head of the rarefaction wave, the width of its fan, and its amplitude. Those quantities are displayed in Table 4.16, where we chose to compute the amplitude of the rarefaction wave by subtracting the water height at the tail from the water height at the head. Concerning the shock wave, we are interested in its position and its amplitude, which are displayed in Table 4.17. Similarly, the amplitude of the shock wave is obtained by computing the difference between the water height to the left of the wave and the water height to its right. Note that, since those computations are performed on the numerical results of the $\mathbb{P}_5^{\text{WB}}$ scheme, the shock wave takes only a couple of cells, and the evaluation of its position is fairly accurate. We observe that the amplitude of the shock presented for $k = 0$ in Table 4.17 is very similar to the results obtained in [47], although the authors do not use the same scheme.

From Table 4.16 and Table 4.17, we observe that the friction produces the expected effects. Indeed, it dampens the amplitude of both the rarefaction wave and the shock wave. Moreover, an increase in friction is accompanied by a diminution of the size of the rarefaction wave, and a decrease in the distance traveled by the shock wave. This behavior is expected, as an increase in friction leads to a decrease in discharge, as evidenced by the expressions (3.89) in

Manning coefficient	Fan width	Amplitude	Head position
$k = 0$	39m	2.68m	$x = -74$m
$k = 0.25$	38m	2.28m	$x = -74$m
$k = 2$	31m	1.29m	$x = -74$m

Table 4.16 – Left rarefaction wave: approximate width of the fan, water height amplitude and position of the head, with respect to the Manning coefficient.

Manning coefficient	Shock position	Amplitude
$k = 0$	$x = 60$m	2.28m
$k = 0.25$	$x = 58$m	1.96m
$k = 2$	$x = 53$m	0.98m

Table 4.17 – Right shock wave: approximate position and water height amplitude, with respect to the Manning coefficient.

1D and (4.11) in 2D. Thus, this decrease in discharge leads to a slower travel time of the shock wave, which directly means that the wave will travel less distance.

Finally, we observe from Table 4.16 that the friction does not change the position of the head of the rarefaction wave. This behavior is also expected. Indeed, recall the expression of the friction source term given by (1.1) in 1D and (4.1) in 2D. Near the head of the rarefaction wave, the water is almost at rest, since no wave has yet perturbed the initial condition at rest. Hence, since only wet areas are considered, the impact of the friction source term is negligible, and the head of the rarefaction wave travels at nearly the same speed for $k = 0$, $k = 0.25$ or $k = 2$. Therefore, the value of the Manning coefficient does not alter the position of the head of the rarefaction wave.

High-resolution simulations

We also include simulations performed using the $\mathbb{P}_5^{\text{WB}}$ scheme, with $k = 0$, $k = 0.25$ and $k = 2$, on a much finer discretization grid, made of one million (1000×1000) cells. In addition, the same values were taken for the parameters C, m_x, m_y, M_x and M_y. Depending on the value of the friction coefficient, the simulations took between 5 hours for $k = 2$ and 6 hours for $k = 0$, with an OpenMP parallelization on 24 cores (12 physical and 12 logical).

The results of this simulation are displayed on Figure 4.22 (free surface) and on Figure 4.23 (discharge). Note that the color scales are different for each figure, in order to display all the details of each experiment. As expected, we once again remark that the vortices are deeper and that the shock wave travels further with less friction. In addition, on Figure 4.23, we again note that the maximum discharge is considerably lower with a higher friction coefficient.

4.4.4 Simulations on a real-world topography

We conclude the numerical experiments of the 2D high-order well-balanced scheme by presenting two real-world simulations. The first one concerns the 2011 Great East Japan

tsunami, in Tōhoku, Japan. The second one consists in an urban topography being flooded by a tsunami.

4.4.4.1 Simulation of the 2011 Tōhoku tsunami

This experiment concerns the simulation of the Great East Japan tsunami. This catastrophic event occurred on the 11th of March, 2011. The numerical simulation of such destructive phenomena is of prime importance for risk assessment and prevention. To address this issue, we consider a Cartesian mesh of the topography of the area, made of around 13 million cells. The emerged topography is displayed on Figure 4.24, and the submerged topography (i.e. the bathymetry) is depicted on Figure 4.25. The water height data related to the initial shape and the position of the tsunami is displayed on Figure 4.26. We also know the water height data from three mareographs, i.e. buoys equipped with tide sensors. The goal of this simulation is to compare the water height from the numerical scheme with the experimental data.

In order to carry out this experiment, we use the $\mathbb{P}_0$ scheme. We take homogeneous Neumann boundary conditions, and we set $C = 100$. The Manning coefficient is chosen according to [45] (page 109); we take $k = 0.025$. The experimental data from the mareographs is available for one hour, and therefore the final time is $t_{end} = 3600$s. The results of the simulation are presented on Figure 4.27 and Figure 4.28. The simulated water height is close to the experimental one, even using the $\mathbb{P}_0$ scheme.

4.4.4.2 Urban topography

The last experiment is a simulation of a city being hit by a wave. We consider the space domain $[0, 1000] \times [0, 1000]$. The topography consists in a gentle upwards slope leading to a flat surface, upon which buildings are placed. Disregarding the buildings, the bottom has the following topography:

$$Z(x, y) = \begin{cases} x/50 & \text{if } x < 500, \\ 10 & \text{otherwise.} \end{cases}$$

The 100 meters high buildings occupy the flat part of the topography, i.e. buildings are only present for $x > 500$. Figure 4.29 displays the shapes and the positions of the buildings, for a uniform Cartesian mesh of 10^6 cells (1000 in each direction).

The initial conditions are $W(0, x, y) = 0$ for all x and y in the space domain. Indeed, the boundary conditions help create the flood and the wave that hits the city. We prescribe homogeneous Neumann boundary conditions for each boundary of the domain, except the left boundary, where a time-dependent boundary condition the x-discharge q_x is applied, as follows:

$$\begin{cases} q_x(t, 0, y) = 15 & \text{if } t < 350, \\ \partial_x q_x(t, 0, y) = 0 & \text{otherwise.} \end{cases} \tag{4.38}$$

Such a boundary condition creates water that fills the sloping part of the topography and creates a wave that hits the city. At time $t = 300$s, some time before the water stops being injected, the free surface is displayed on Figure 4.29.

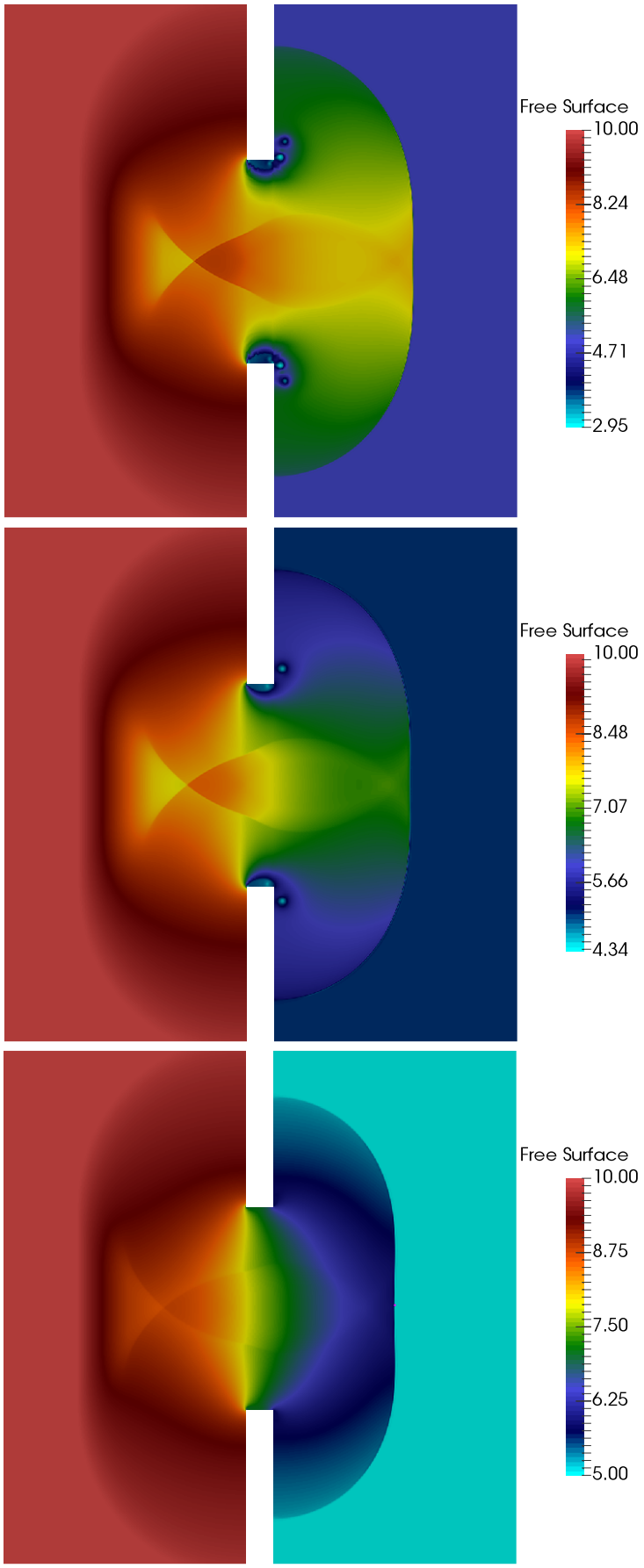

Figure 4.22 – Partial dam-break: free surface using the $\mathbb{P}_5^{\mathrm{WB}}$ and 10^6 cells. From top to bottom: $k = 0$, $k = 0.25$ and $k = 2$. Note that the color scale is different on each figure.

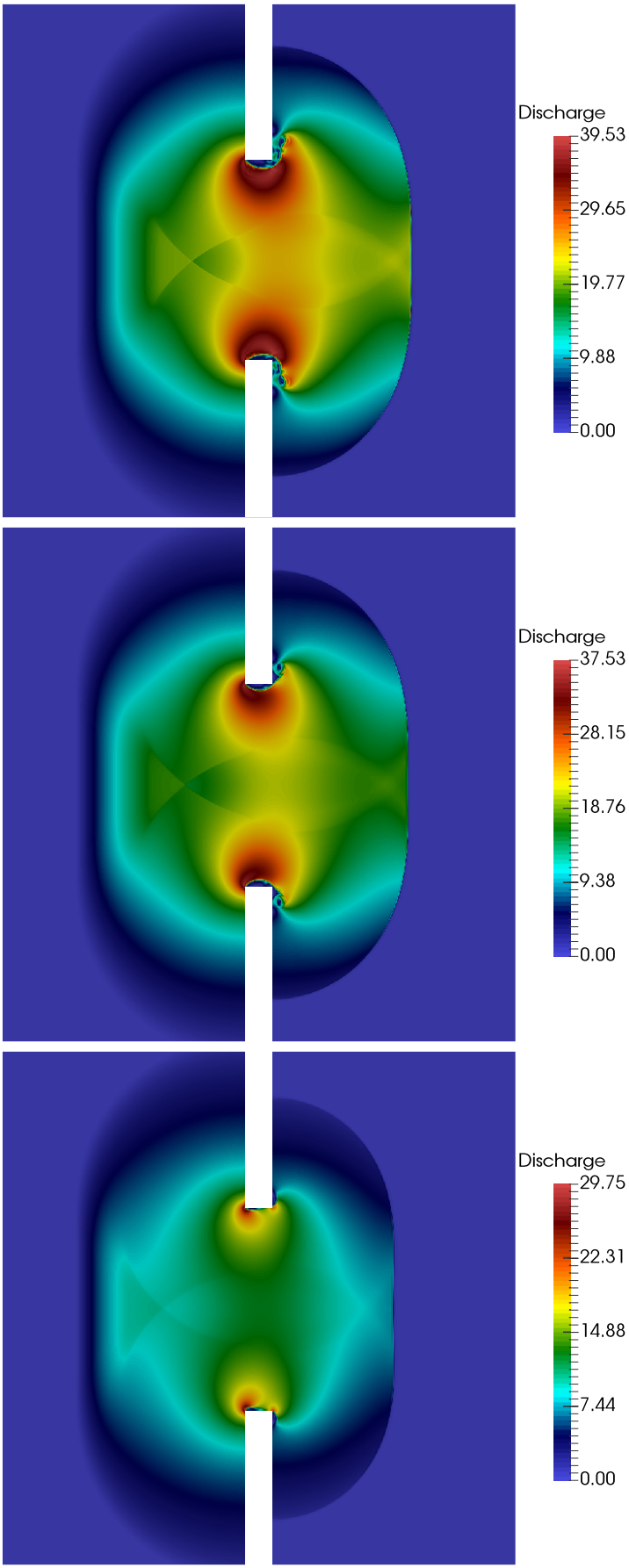

Figure 4.23 – Partial dam-break: discharge using the $\mathbb{P}_5^{\mathrm{WB}}$ and 10^6 cells. From top to bottom: $k = 0$, $k = 0.25$ and $k = 2$. Note that the color scale is different on each figure.

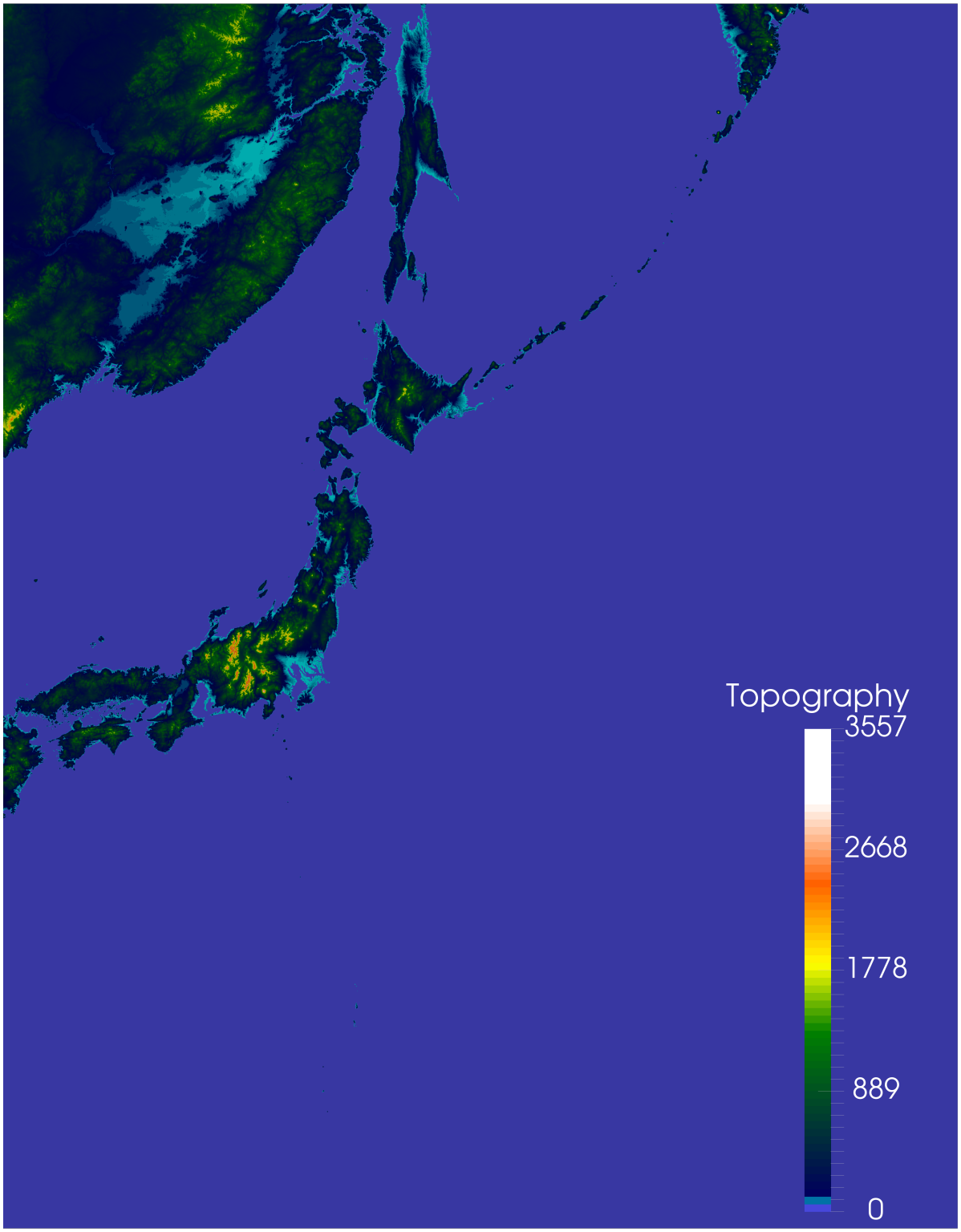

Figure 4.24 – Emerged topography for the Tōhoku tsunami simulation.

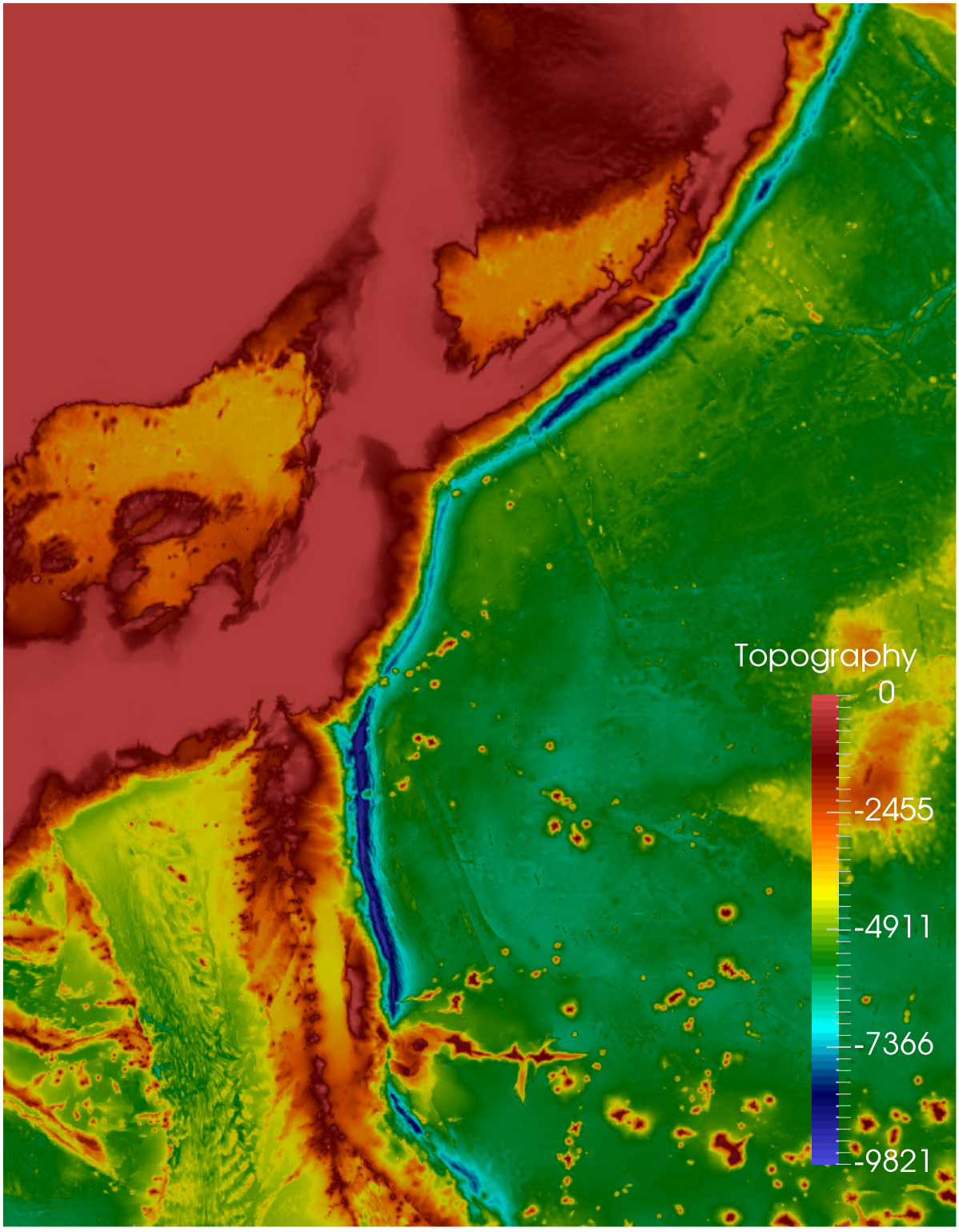

Figure 4.25 – Submerged topography (bathymetry) for the Tōhoku tsunami simulation.

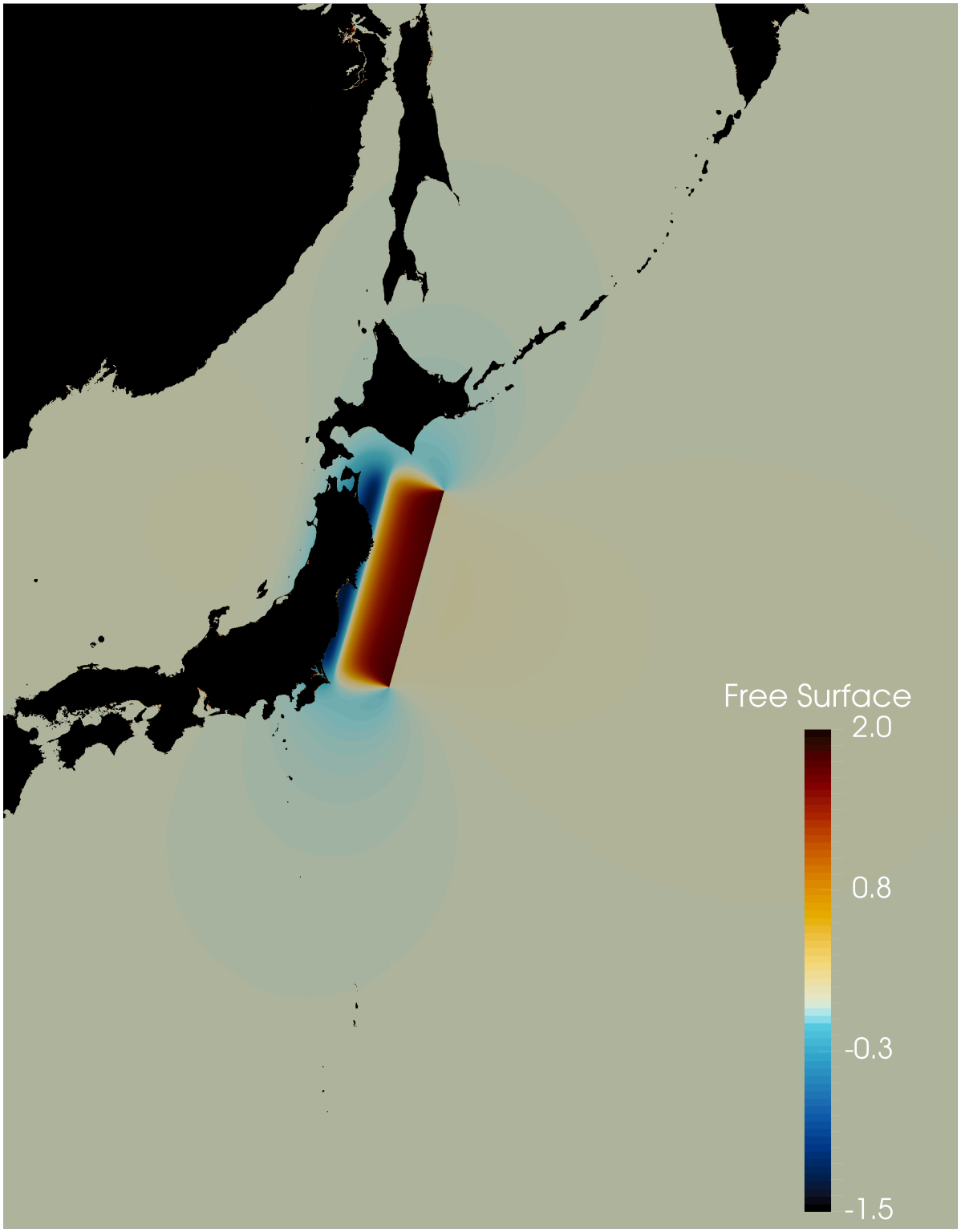

Figure 4.26 – Initial free surface for the Tōhoku tsunami simulation.

We consider a nonzero Manning coefficient $k = 1$ and a final time $t_{end} = 850$s. The simulation is carried out using the $\mathbb{P}_1^{\text{WB}}$ scheme, and we take $C = 10^{-2}$, $m_x = m_y = 10^{-5}$ and $M_x = M_y = 1$. It took around 7.5 hours to complete with 24 cores (12 physical and 12 logical).

We display on Figure 4.30 the free surface and the discharge along the line $x = 225$m, located between the second and third row of square buildings. On the left panel, at $t = 300$s, the Dirichlet boundary condition is active, and the discharge is equal to $15\text{m}^2.\text{s}^{-1}$ on the left boundary. On the right panel, at $t = 355$s, the boundary condition has become a homogeneous Neumann one, and the discharge has started diminishing near the boundary. On both graphs, note that the water front has the shape expected when dealing with the Manning friction source term.

The results of the numerical simulation are displayed on Figure 4.31, Figure 4.32, Figure 4.33, Figure 4.34, and Figure 4.35.

The left panel of Figure 4.29 shows the wave created by the Dirichlet boundary condition arriving on the city. Because of the friction, this wave presents a rather steep front. On the left panel of Figure 4.31, the wave has hit the first buildings located at the south of the city. Note that the space between the first two columns of buildings is still dry. Also note that, as per (4.38), the boundary condition imposed on the x-discharge q_x is now a homogeneous Neumann boundary condition, and no more water is injected into the domain. The right panel of Figure 4.31 displays the wave about to hit the square building located at the middle of the city. As expected, between the southern buildings, the wave is slowed down.

On the left panel of Figure 4.32, the wave has reflected on the southwestern side of the square building, and it has thus moved faster towards the buildings to the south. On the right panel of Figure 4.32, the waves reflected from the square building are moving south and north. Moreover, the back of the "S"-shaped building will soon be flooded.

The flooding of the back of the "S"-shaped building is happening on Figure 4.33, with only a small area still dry on the right panel. In addition, on the right panel of Figure 4.33, the wave has almost hit the small square building on the bottom right of the city.

Figure 4.34 and Figure 4.35 display the final phases of the flooding of the city. Note that the southern buildings are mostly uniformly flooded and that the inner courtyard of the square building is still dry. Moreover, the water at the back of the "S"-shaped building is less deep than at other points of the same vertical line.

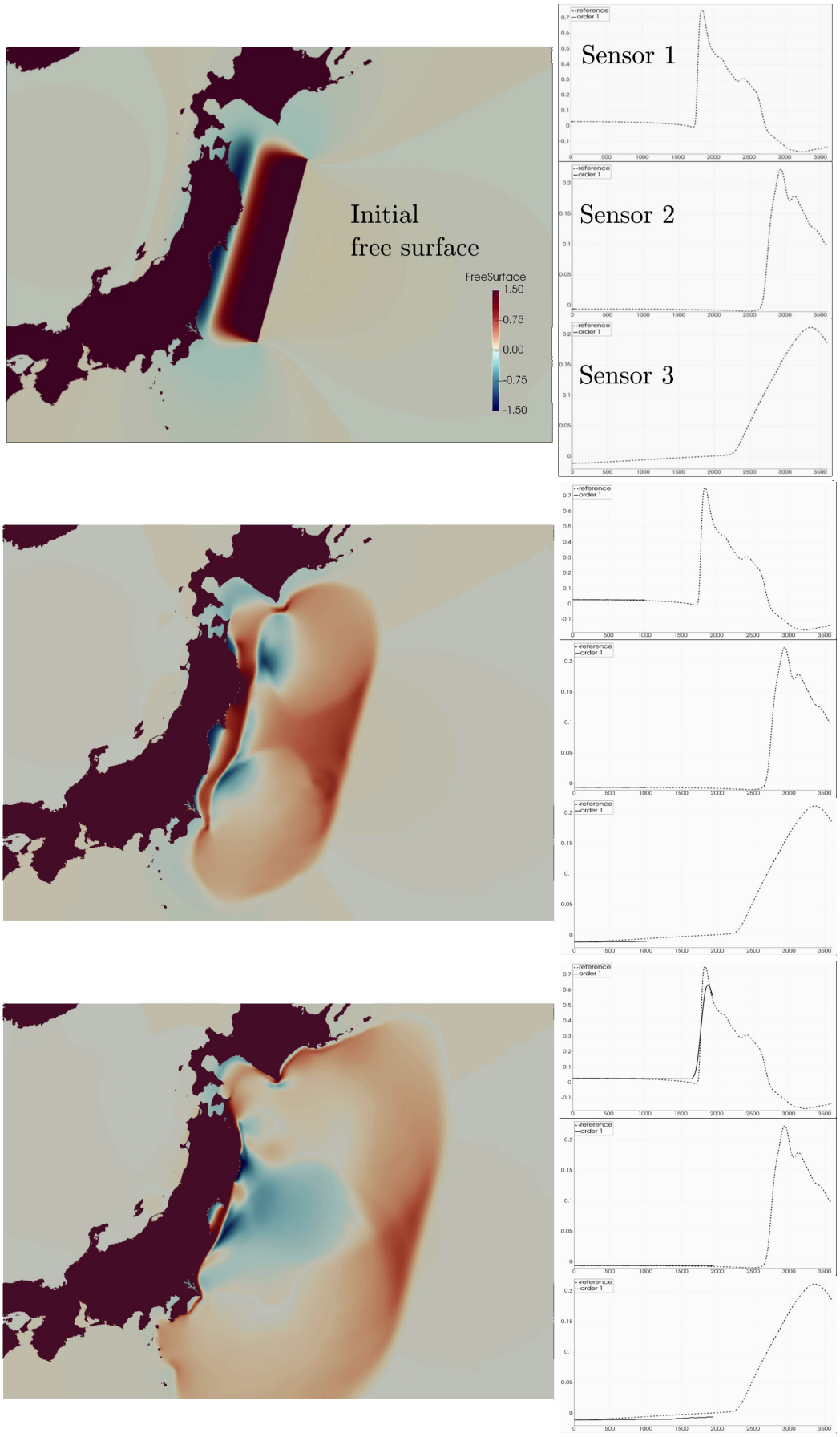

Figure 4.27 – Tōhoku tsunami simulation. From top to bottom: free surface at $t = 0$s, $t = 1000$s and $t = 1900$s. The sensor data is displayed on the right.

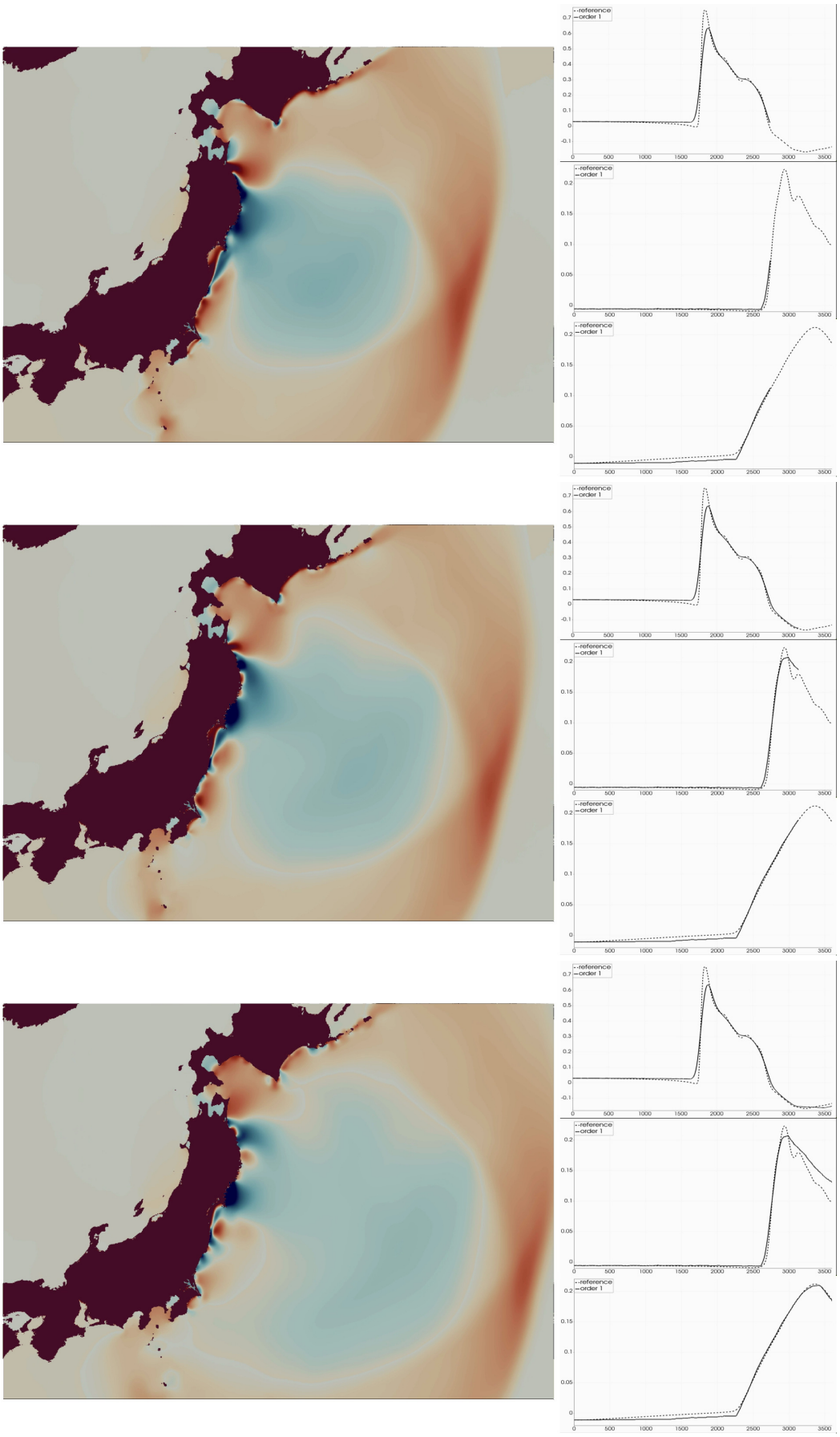

Figure 4.28 – Tōhoku tsunami simulation. From top to bottom: free surface at $t = 2750$s, $t = 3200$s and $t = 3600$s. The sensor data is displayed on the right.

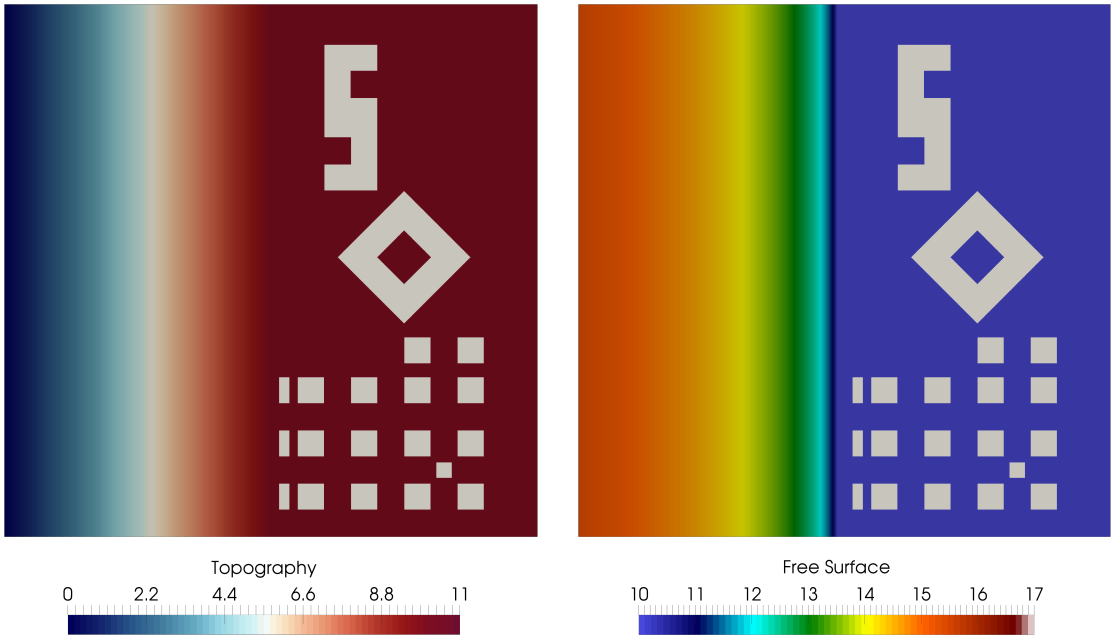

Figure 4.29 – Wave on an urban topography simulation. Left panel: topography of the city. The buildings are actually 100 meters high, and are represented in white in this figure. One can see the upwards slope on the left, leading to the city itself. Right panel: free surface at $t = 300$s. The wave is present to the left of the figure. Note that the same free surface color scale will be used in the next figures.

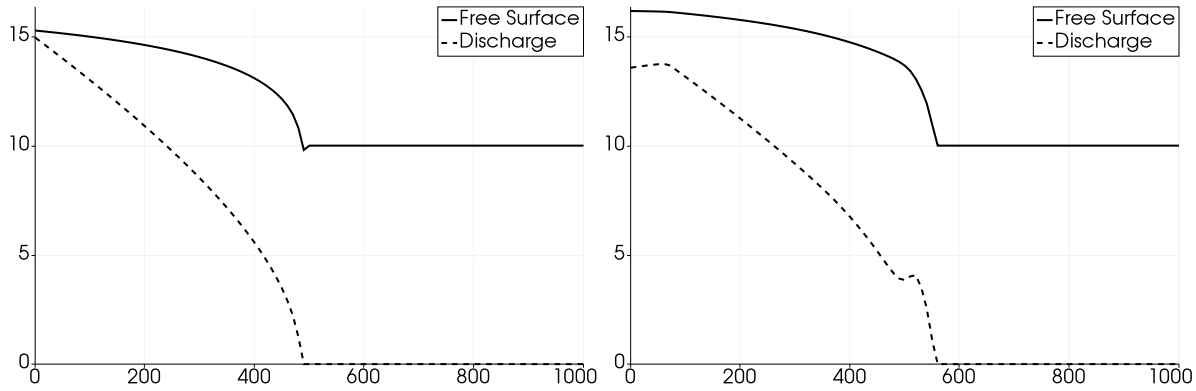

Figure 4.30 – Free surface and discharge along the line $x = 225$m for the urban topography simulation, at $t = 300$s (left panel) and $t = 355$s (right panel).

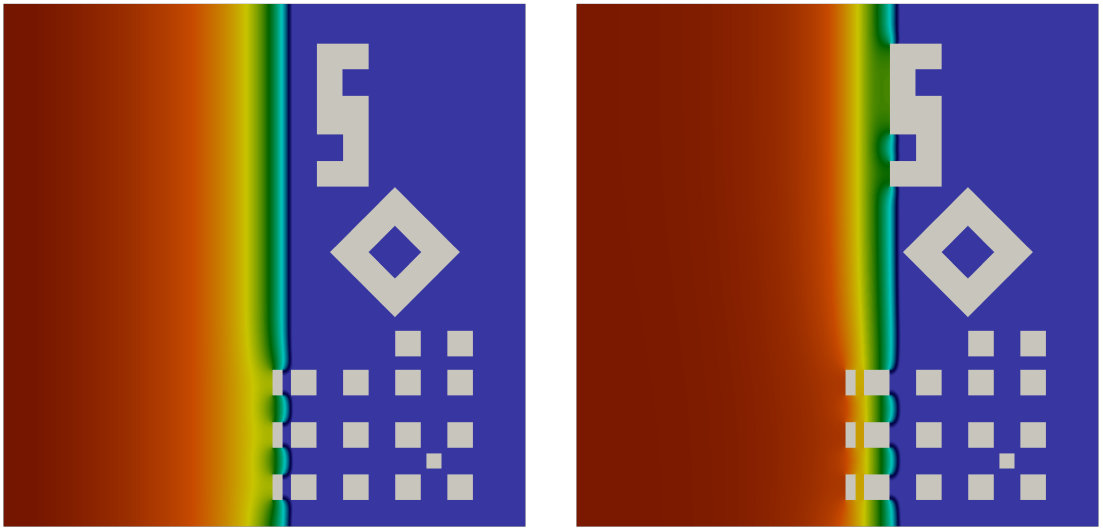

Figure 4.31 – Free surface for the urban topography simulation at $t = 355$s (left panel) and $t = 410$s (right panel).

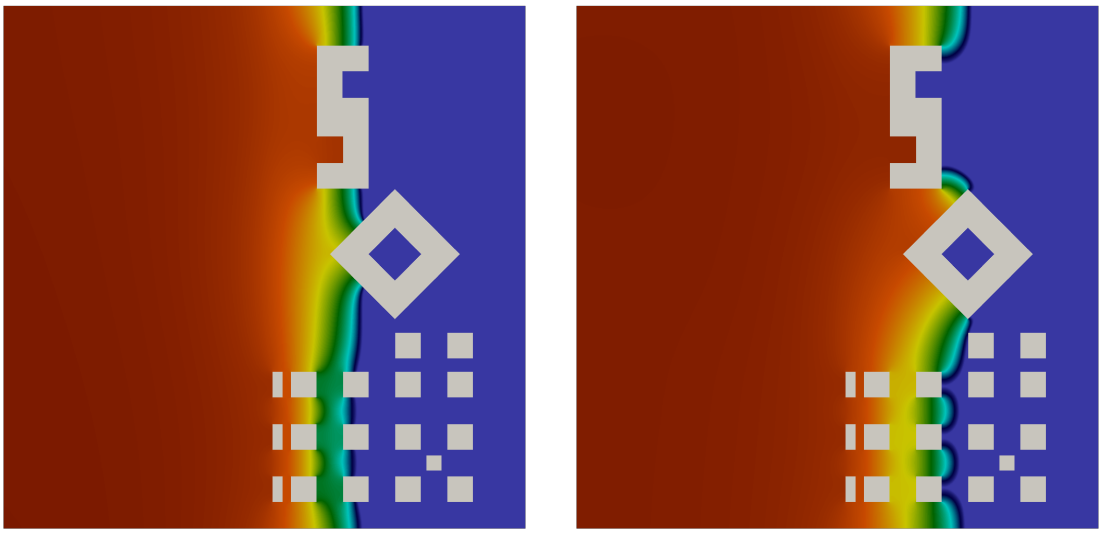

Figure 4.32 – Free surface for the urban topography simulation at $t = 465$s (left panel) and $t = 520$s (right panel).

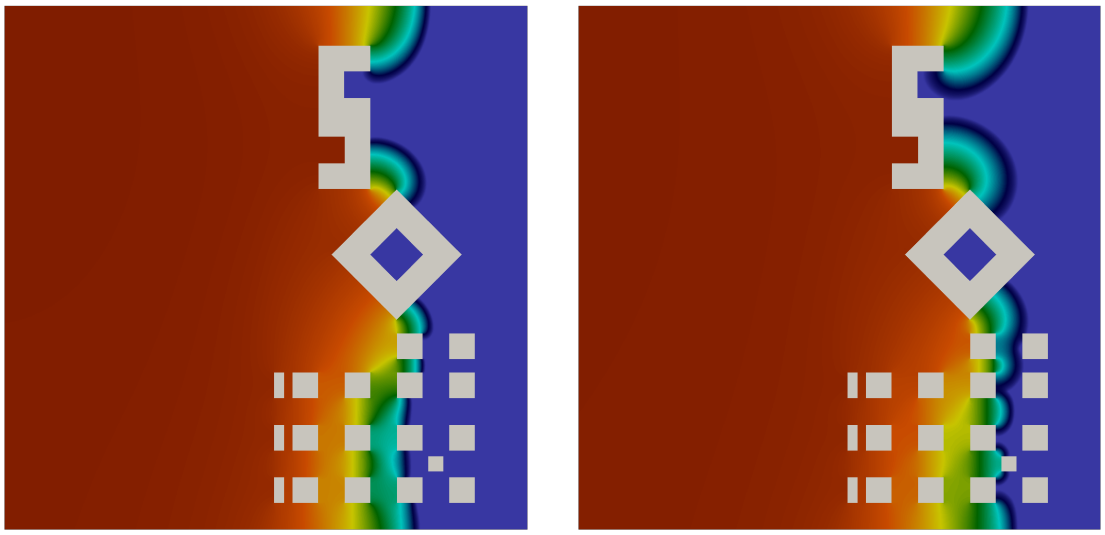

Figure 4.33 – Free surface for the urban topography simulation at $t = 575$s (left panel) and $t = 630$s (right panel).

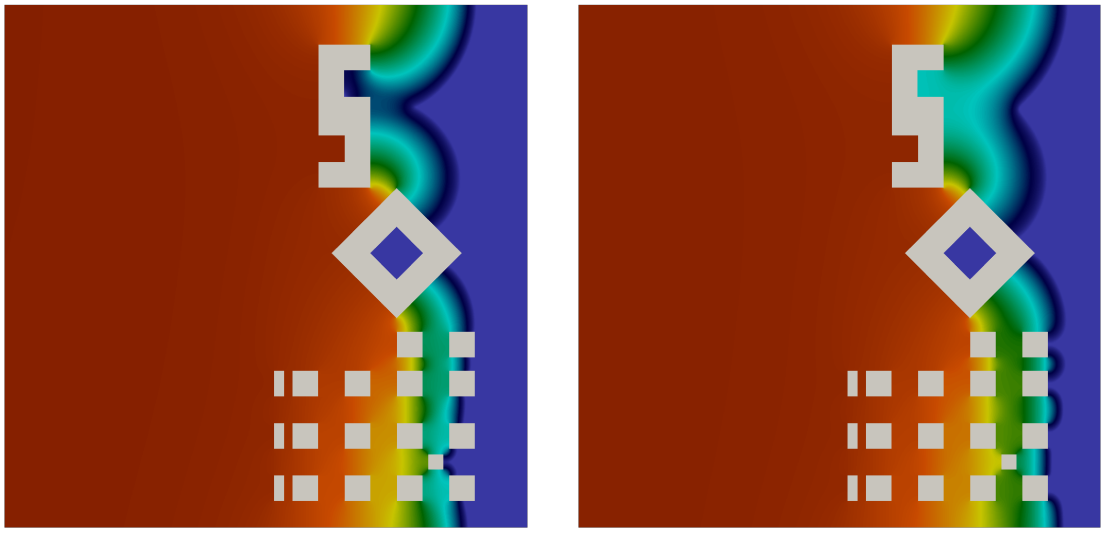

Figure 4.34 – Free surface for the urban topography simulation at $t = 685$s (left panel) and $t = 740$s (right panel).

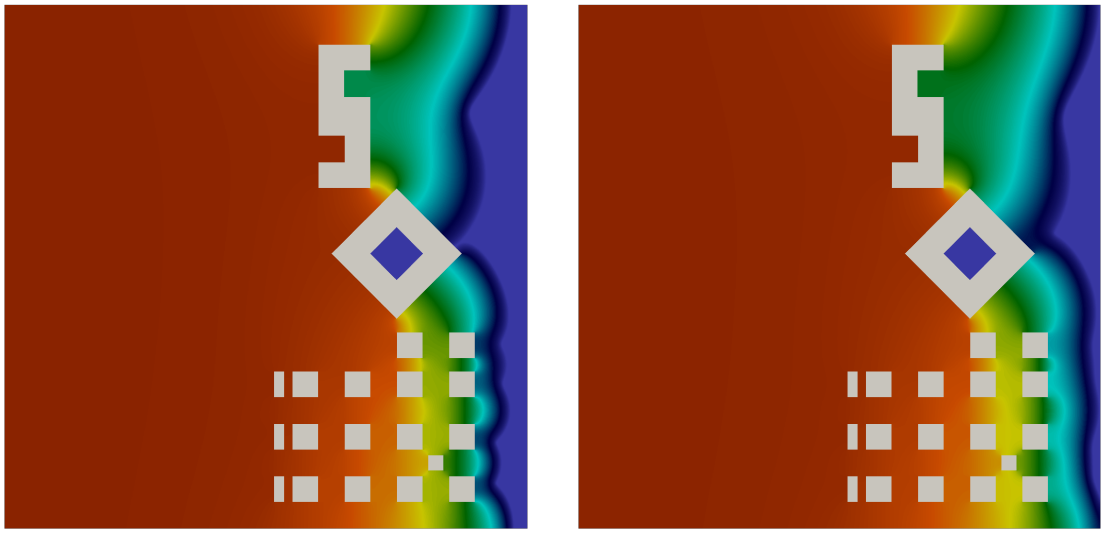

Figure 4.35 – Free surface for the urban topography simulation at $t = 795$s (left panel) and $t = 850$s (right panel).

Conclusion & perspectives

Version française

Conclusion

Dans ce manuscrit, nous avons étudié, analytiquement et numériquement, le système de Saint-Venant muni des termes source de topographie et de friction de Manning.

Dans un premier temps, nous avons étudié les effets du terme source de friction sur les propriétés algébriques du système ainsi que sur ses solutions stationnaires. Nous avons montré que, à l'instar du terme source de topographie, le terme source de friction ajoutait un champ caractéristique stationnaire au système. Par conséquent, lorsque les deux termes source sont présents, ce champ caractéristique est associé à la valeur propre double 0. De plus, dans un souci de complétude, les solutions stationnaires ont été rappelées pour le terme de topographie, tandis qu'une étude complète des solutions stationnaires a été effectuée pour le terme de friction. En particulier, pour les deux termes source, chercher les solutions stationnaires revient à chercher les zéros d'une fonction non-linéaire. Nous avons montré que, si ce problème possède une solution, alors soit elle est unique, soit il y en a exactement deux. Si la solution stationnaire est unique, sa hauteur d'eau, égale à une hauteur critique, est la même pour les termes de topographie et de friction. Si deux solutions cohabitent, l'une d'entre elles est subcritique (hauteur supérieure à la hauteur critique), tandis que l'autre est supercritique (hauteur inférieure à la hauteur critique).

Nous avons ensuite dérivé un schéma équilibre robuste pour le système de Saint-Venant avec topographie et friction de Manning. Ce schéma vérifie les propriétés suivantes :

(i) consistance avec les équations de Saint-Venant avec topographie et friction de Manning ;
(ii) préservation et capture exactes de toutes les solutions stationnaires (celles au repos et les solutions à vitesse non nulle) des équations de Saint-Venant avec topographie ou friction ;
(iii) robustesse, c'est-à-dire préservation de la positivité de la hauteur d'eau.

Ensuite, nous avons rendu ce schéma semi-implicite, afin de s'assurer que les transitions entre zones mouillées et zones sèches n'induisaient pas d'oscillations dues à la raideur du terme de friction quand la hauteur d'eau devient petite. Cette semi-implicitation consiste en un traitement explicite du flux et de la topographie, et en un traitement implicite de la friction. Des cas-tests numériques visant à vérifier la propriété de préservation des états stationnaires et à valider numériquement le schéma ont ensuite été présentés.

Nous avons enfin étendu ce schéma pour prendre en compte des géométries bidimensionnelles cartésiennes, et nous avons réalisé une montée en ordre. Pour l'extension à deux

dimensions d'espace, nous avons effectué une combinaison convexe par direction du schéma 1D afin d'en préserver les propriétés. Comme nous nous y attendions, il n'a pas été possible d'étendre complètement la propriété de préservation des états stationnaires. En effet, le schéma 2D préserve les états stationnaires dans les directions des axes x et y, ainsi que tous les états stationnaires au repos. Ensuite, nous avons réalisé une montée en ordre utilisant une reconstruction polynomiale. Cependant, après cette procédure de reconstruction, le schéma produisait des oscillations et ne préservait plus les états stationnaires. Afin de remédier à ces problèmes, nous avons suggéré l'utilisation d'une méthode de type MOOD. En particulier, la propriété de préservation des états stationnaires a été recouvrée en utilisant une combinaison convexe entre le schéma d'ordre élevé et le schéma d'ordre un. Cette combinaison convexe favorise le schéma d'ordre un lorsque la solution approchée est proche d'une solution stationnaire (lorsque le schéma d'ordre un est exact), ce qui résulte en un schéma au moins d'ordre élevé. Nous avons ensuite réalisé un couplage entre cette combinaison convexe et une méthode MOOD plus classique, dans le but d'obtenir un schéma d'ordre élevé, préservant les solutions stationnaires, et ne produisant pas d'oscillations. Finalement, nous avons déterminé les propriétés de ce schéma en codant un programme Fortran parallèle. Nous avons fourni des précisions sur l'implémentation en Fortran ainsi que sur les bibliothèques externes utilisées dans le programme. Nous avons ensuite proposé des cas-tests pour vérifier la préservation des états stationnaires et l'ordre élevé du schéma, avant d'effectuer plusieurs cas-tests de ruptures de barrage afin de valider le schéma. Enfin, nous avons proposé deux simulations d'événements réels : celle du tsunami ayant frappé le Japon en 2011 et celle d'une vague inondant une topographie urbaine.

Perspectives

Nous pouvons envisager plusieurs perspectives aux travaux contenus dans ce manuscrit. Naturellement, nous pouvons penser à appliquer la méthode générique développée dans le troisième chapitre à d'autres termes source. De plus, en notant que les états intermédiaires du solveur de Riemann approché sont assez proches des états intermédiaires du solveur HLL, nous pourrions étudier l'entropie du schéma proposé, par exemple en utilisant des résultats connus sur le schéma HLL. Enfin, nous pourrions travailler un peu plus sur l'extension à l'ordre élevé, notamment en essayant de formuler une preuve rigoureuse de l'ordre du schéma.

Application à d'autres termes source

Dans ce manuscrit, nous avons proposé un schéma numérique pour les équations de Saint-Venant avec un terme source générique (3.1). Nous avons montré que ce schéma permettait de préserver les solutions stationnaires dès qu'une moyenne pertinente $\overline{S}$ du terme source était donnée. Si le terme source est donné sous la forme (3.40), les relations gouvernant les états stationnaires incluent une relation algébrique. Dans ce cas, les relations gouvernant les états stationnaires discrets, données par (3.47), permettent d'obtenir $\overline{S}$. À présent, nous nous intéressons plus particulièrement à deux exemples de termes source : la force de Coriolis et un terme source qui permet de prendre en compte la largeur du canal.

Le terme source représentant la force de Coriolis est utilisé en océanographie, son utilité première étant de simuler la préservation d'un équilibre (voir [121]). Par conséquent, dans ce cas, il serait pertinent d'utiliser un schéma préservant les états stationnaires. Les équations de Saint-Venant munies de ce terme source ont déjà été étudiées, analytiquement (voir [148, 165, 59, 118]) et numériquement (voir [27, 8, 91, 43]). Les auteurs de ces articles étudient le modèle en 2D, où la rotation induite par la force de Coriolis est facilement modélisée par un terme source. Cependant, il est possible de dériver un modèle unidimensionnel, en étudiant les équations sous la forme suivante :

$$\begin{cases} \partial_t h + \partial_x q = 0, \\ \partial_t q + \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -fq, \end{cases}$$

où $f \in \mathbb{R}$ représente le coefficient de la force de Coriolis. Nous pouvons donc écrire le terme source de Coriolis $S^c(W) = -fq$ sous la forme (3.40), en posant :

$$\beta = 0 ; \quad f(q) = q ; \quad \partial_x \sigma = -f.$$

Par conséquent, les états stationnaires discrets sont gouvernés par (3.47), comme suit :

$$\begin{cases} q_0^2 \left[\dfrac{1}{h} \right] + \dfrac{g}{2} \left[h^2 \right] = \overline{S}^c \Delta x, \\ q_0^2 \left[\dfrac{1}{h} \right] + \dfrac{g}{2} \left[h^2 \right] + q_0 f \Delta x = 0, \end{cases} \tag{F3}$$

où $\overline{S}^c$ est une moyenne de S^c. Ces relations sont valides pour h_L et h_R distincts et strictement positifs. La seconde équation de (F3) permet d'obtenir une expression de q_0 pour les états stationnaires, qui fait intervenir, entre autres, la direction de l'écoulement. Cette expression doit ensuite être injectée dans la première équation afin d'obtenir une formule donnant $\overline{S}^c$.

Après avoir introduit le terme source représentant la force de Coriolis, nous nous tournons vers la prise en compte de la largeur du canal. Ce deuxième terme source fut introduit dans [155] (voir aussi [95, 78]), pour donner le système suivant :

$$\begin{cases} \partial_t h + \partial_x q = -q \dfrac{\partial_x B}{B}, \\ \partial_t q + \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -\dfrac{q^2}{h} \dfrac{\partial_x B}{B}, \end{cases} \tag{F4}$$

où la fonction $B(x)$ représente la largeur du canal. Nous remarquons qu'un terme source est présent sur l'équation de conservation de la hauteur d'eau ; ce cas n'est pas pris en compte dans le schéma développé dans ce manuscrit. Afin de remédier à ce problème, nous introduisons un nouveau jeu de variables (voir [130] par exemple) : $H = hB$ et $Q = qB$. En utilisant ces nouvelles variables et en supposant une solution régulière dont la hauteur est strictement

positive, le système (F4) se réécrit comme suit :

$$\begin{cases} \partial_t H + \partial_x Q = 0, \\ \partial_t Q + \partial_x \left(\dfrac{Q^2}{H} + \dfrac{g}{2}\dfrac{H^2}{B} \right) = \dfrac{g}{2}\dfrac{H^2}{B^2}\partial_x B, \end{cases}$$

où la première équation ne contient plus de terme source. En revanche, le terme de pression dans la dérivée spatiale du flux a été modifié. Le terme source représentant la largeur du canal est donc défini par :

$$S^b = \frac{g}{2}\frac{H^2}{B^2}\partial_x B.$$

De plus, nous pouvons montrer que les états stationnaires associés à cette équation sont gouvernés par :

$$\begin{cases} Q = \mathrm{cst} = Q_0, \\ \partial_x \left(\dfrac{H}{B} + \dfrac{Q_0^2}{2gH^2} \right) = 0. \end{cases}$$

Par conséquent, en notant $\overline{S}^b$ une moyenne pertinente du terme source S^b, les relations définissant un état stationnaire discret sont les suivantes :

$$\begin{cases} Q_0^2 \left[\dfrac{1}{H} \right] + \dfrac{g}{2} \left[\dfrac{H^2}{B} \right] = \overline{S}^b \Delta x, \\ \left[\dfrac{H}{B} \right] + \dfrac{Q_0^2}{2g} \left[\dfrac{1}{H^2} \right] = 0. \end{cases} \tag{F5}$$

Comme précédemment, la seconde équation de (F5) fournit une expression de Q_0^2 valide pour un état stationnaire, qui est ensuite injectée dans la première équation afin d'obtenir une formule pour $\overline{S}^b$. Notons que cette formule dépendra de la variable $H = hB$. Des manipulations algébriques seront donc requises pour concilier ce terme source approché avec ceux obtenus pour la topographie, la friction ou la force de Coriolis.

Stabilité du schéma

Dans ce manuscrit, nous ne nous sommes pas posé la question de la stabilité du schéma. L'expression (3.6) des vitesses caractéristiques a été choisie pour s'assurer que $\lambda_L < 0 < \lambda_R$, ce qui augmente la diffusion numérique du schéma, et qui entraîne donc une augmentation de sa stabilité. Ce choix de vitesses caractéristiques nous a donc permis de retarder l'étude de la stabilité.

Nous pouvons aussi nous intéresser à l'entropie associée au schéma dérivé dans le troisième chapitre. En effet, les états intermédiaires utilisés dans le solveur de Riemann approché de ce schéma donc donnés par (3.81). Nous remarquons que ces états intermédiaires sont en fait les états intermédiaires du solveur HLL, définis par (3.20), auxquels un terme supplémentaire, dépendant linéairement du pas d'espace, a été rajouté. Ces états intermédiaires peuvent donc être vus comme une perturbation des états intermédiaires du schéma HLL, qui est entropique (voir [90]). Par conséquent, une étude précise de cette perturbation pourrait permettre d'obtenir une inégalité d'entropie pour le schéma équilibre, et ce dans le cas des termes source

individuels de topographie et de friction, ou même en présence des deux termes source. Les vitesses caractéristiques λ_L et λ_R joueraient certainement un rôle dans cette inégalité. De plus, la constante C, introduite dans (3.54) afin de faire en sorte que le terme source approché de topographie soit consistant, pourrait aussi jouer un rôle dans cette inégalité, dans le cas de la topographie.

En supposant qu'une telle inégalité puisse être déterminée, un nouveau critère de détection pourrait s'ajouter aux critères déjà présents dans la méthode MOOD utilisée dans le schéma d'ordre élevé. En effet, dans l'esprit de [16], nous pourrions utiliser l'inégalité d'entropie afin de diminuer le degré de la reconstruction polynomiale jusqu'à ce que cette inégalité soit vérifiée.

Ordre élevé : résultats et améliorations

Le schéma proposé dans le quatrième chapitre est d'ordre élevé et permet de préserver les solutions stationnaires, comme nous l'avons illustré grâce aux cas-tests présentés à la fin de ce chapitre. Cependant, à cause de la procédure de combinaison convexe entre les schémas d'ordre élevé et d'ordre un, ce dernier est utilisé lorsque la solution est proche d'une solution stationnaire. Le schéma est donc, au final, au moins d'ordre élevé, puisque le schéma d'ordre un n'est utilisé que lorsqu'il est exact (d'ordre infini). Une preuve rigoureuse de cet ordre élevé pourrait cependant être étudiée.

De plus, dans [33, 31, 35], les auteurs ont proposé une reconstruction polynomiale basée sur les états stationnaires à vitesse non nulle. Cette procédure permet d'obtenir un schéma d'ordre élevé préservant naturellement les solutions stationnaires, sans avoir besoin d'introduire de combinaison convexe. Par conséquent, la reconstruction polynomiale de [33, 31, 35] pourrait être une extension intéressante du schéma proposé dans le quatrième chapitre. Cependant, afin d'utiliser cette reconstruction, il faut résoudre approximativement les équations non-linéaires gouvernant les solutions stationnaires à tout moment où la reconstruction doit être calculée ; un des avantages du schéma présenté dans ce manuscrit est le fait que de telles résolutions approchées d'équations non linéaires n'intervenaient pas dans sa dérivation.

English version

Conclusion

In this manuscript, we have studied, both numerically and analytically, the shallow-water system equipped with the source terms of topography and Manning friction, governed by (1.1).

First, we studied the effects of the Manning friction source term on the shallow-water system, regarding either its algebraic properties or its steady state solutions. This friction source term, as expected, was proven to add a stationary characteristic field, as does the topography source term. As a consequence, when both source terms are present, the stationary characteristic field is associated to the double eigenvalue 0. In addition, the steady state solutions have been recalled, for the sake of completeness, in the case of the topography source term. In the case of the friction source term, they have been studied in detail. In particular, in both cases, we have shown that finding a steady state was equivalent to finding the zeros of a nonlinear equation; this problem has zero, one or two solutions. When the solution is unique, the water height of this steady state is equal to a critical water height, which has the same value for both source terms. When two solutions exist, one is subcritical (strictly superior to the critical height) and the other one is supercritical (strictly inferior to the critical height).

We then derived a suitable numerical scheme for the shallow-water equations with both topography and friction. We derived a well-balanced and robust scheme, i.e. a scheme that:

(i) is consistent with the shallow-water equations with topography and friction;
(ii) exactly preserves and captures all the steady state solutions (the steady states at rest and the moving steady states) of the shallow-water equations with topography or friction;
(iii) preserves the non-negativity of the water height.

In addition, a semi-implicit extension of the scheme was introduced to ensure that the transitions between wet and dry areas did not induce oscillations in the water height. This semi-implicitation consists in an explicit discretization of the flux and the topography and in an implicit treatment of the friction, in order to account for the stiffness of the friction source term near dry areas. Afterwards, numerical experiments were carried out in order to check all the properties of the scheme. Namely, the well-balance of the scheme was assessed, and several validation experiments were performed.

We finally focused on an extension of the scheme to two-dimensional Cartesian geometries, as well as the derivation of a high-order accurate scheme from the 2D first-order one. First, the 2D extension was obtained using a convex combination technique in order to preserve the properties satisfied by the 1D scheme. As expected, the resulting 2D scheme turned out to be well-balanced by direction, i.e. only the steady state solutions along the x-axis and the y-axis were exactly preserved, in addition to all the steady states at rest. Then, a high-order extension of this 2D scheme was proposed. However, due to the reconstruction procedure, the scheme lost its well-balance property and produced oscillating solutions. A MOOD-like method was suggested to deal with these shortcomings. Namely, the well-balance property was recovered thanks to a convex combination between the first-order well-balanced scheme and the high-order scheme. This convex combination favors the well-balanced scheme close to steady state solutions, i.e. in areas where this scheme is exact, thus resulting in a scheme

that is at least high-order accurate. The convex combination was then coupled to a more classical MOOD method to yield a non-oscillatory and well-balanced high-order 2D scheme. To assess the properties of this scheme, several benchmark simulations were then carried out using a parallel Fortran code, made from scratch. An explanation of the Fortran implementation and the external libraries used was provided. Regarding the numerical experiments, the well-balance property and the high-order accuracy were first checked. Then, several validation dam-breaks experiments were performed. Finally, two real-world simulations were proposed: the 2011 Tōhoku tsunami, and a wave impacting an urban topography.

Perspectives

Several perspectives of this work can be envisioned. Namely, the generic approach developed in Chapter 3 could be extended to take other source terms into account. Also, noting that the expressions of the intermediate states of the suggested 1D scheme are quite close to those of the intermediate states of the HLL scheme, the entropy stability of the proposed scheme could be studied, by using existing results on the entropy stability of the HLL scheme. Other perspectives concern the high-order well-balanced scheme: for instance, a rigorous proof of the high-order accuracy could be studied, or an alternate polynomial reconstruction could be considered.

Application to other source terms

Let us recall that the 1D well-balanced scheme for the shallow-water equations with a generic source term (3.1) involves the intermediate states (3.37). Thanks to Theorem 3.5, we know the approximate Riemann solver obtained with these intermediate states is well-balanced as soon as a relevant average $\overline{S}$ of the source term is provided. With a source term given by (3.40), the steady state relations involve an algebraic equation. As a consequence, the discrete steady state relations, needed to get a suitable average $\overline{S}$, are given by (3.47). We now provide two examples of source terms the generic strategy could be applied to: a source term representing the Coriolis force and another one taking into account the variations of the channel breadth.

We first consider the Coriolis force source term; it is widely used in oceanography, mostly to simulate the perturbation of an equilibrium (see [121]). Therefore, using a well-balanced scheme would be particularly relevant in this context. The shallow-water equations equipped with this source term have already been studied, both analytically (see for instance [148, 165, 59, 118]) and numerically (see for instance [27, 8, 91, 43]). In these articles, the model is studied in two space dimensions, where the rotation induced by the Coriolis force makes more sense. However, it is possible to derive a one-dimensional model of the Coriolis force. Equipped with just the Coriolis force source term, the shallow-water equations are given as follows:

$$\begin{cases} \partial_t h + \partial_x q = 0, \\ \partial_t q + \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -f q, \end{cases}$$

where $f \in \mathbb{R}$ is the coefficient of the Coriolis force. We can therefore write the Coriolis force

source term $S^c(W) = -fq$ under the form (3.40), by setting:

$$\beta = 0 \quad ; \quad f(q) = q \quad ; \quad \partial_x \sigma = -f.$$

As a consequence, (3.47) yields the following discrete steady relations, with $\overline{S}^c$ a suitable average of S^c:

$$\begin{cases} q_0^2 \left[\frac{1}{h}\right] + \frac{g}{2}\left[h^2\right] = \overline{S}^c \Delta x, \\ q_0^2 \left[\frac{1}{h}\right] + \frac{g}{2}\left[h^2\right] + q_0 f \Delta x = 0. \end{cases} \tag{E3}$$

These relations are valid for $h_L > 0$ and $h_R > 0$ such that $h_L \neq h_R$. The second equation of (E3) provides a value of q_0 depending on h_L and h_R. The sign of q_0, i.e. the direction of the steady water flow has to be taken into account to solve this equation. Then, the expression of q_0 is plugged into the first equation, to finally get the expression of $\overline{S}^c$.

Having introduced the Coriolis force, we now turn to the breadth variation source term. This source term has been introduced in [155] (see also [95, 78] for instance) to deal with the variations of the channel breadth, to get the following shallow-water model:

$$\begin{cases} \partial_t h + \partial_x q = -q \dfrac{\partial_x B}{B}, \\ \partial_t q + \partial_x \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2 \right) = -\dfrac{q^2}{h} \dfrac{\partial_x B}{B}, \end{cases} \tag{E4}$$

where the function $B(x) > 0$ represents the breadth of the channel. Note that a source term is present on the height equation. This case is not taken into account by the scheme suggested in Chapter 3. To address such an issue, we introduce the new set of variables $H = hB$ and $Q = qB$ (see [130] for instance). Using H and Q and assuming a smooth solution with $h > 0$, the shallow-water model (E4) rewrites as follows:

$$\begin{cases} \partial_t H + \partial_x Q = 0, \\ \partial_t Q + \partial_x \left(\dfrac{Q^2}{H} + \dfrac{g}{2} \dfrac{H^2}{B} \right) = \dfrac{g}{2} \dfrac{H^2}{B^2} \partial_x B, \end{cases}$$

where we have successfully eliminated the source term on the height equation. However, the pressure term in the spatial derivative now contains the breadth function B. According to the above system, the breadth source term is thus defined by:

$$S^b = \frac{g}{2} \frac{H^2}{B^2} \partial_x B.$$

In addition, seeking smooth steady state solutions for this system leads, after straightforward computations, to the following ordinary differential equation:

$$\begin{cases} Q = \text{cst} = Q_0, \\ \partial_x \left(\dfrac{H}{B} + \dfrac{Q_0^2}{2gH^2} \right) = 0. \end{cases}$$

Therefore, with $\overline{S}^b$ a suitable average of the breadth source term S^b, the discrete steady relations read, in this case:

$$\begin{cases} Q_0^2 \left[\frac{1}{H}\right] + \frac{g}{2}\left[\frac{H^2}{B}\right] = \overline{S}^b \Delta x, \\ \frac{Q_0^2}{2}\left[\frac{1}{H^2}\right] + g\left[\frac{H}{B}\right] = 0. \end{cases} \tag{E5}$$

As usual, the second equation of (E5) yields an expression of Q_0^2, to be plugged into the first equation in order to get an formula for $\overline{S}^b$. Note that this expression will depend on the variable $H = hB$. Therefore, several algebraic manipulations will have to be made on the equations in order to combine the approximation of this source term with the approximations of the topography, friction and Coriolis force source terms.

Stability of the scheme

The question of the stability of the scheme is not raised in the present manuscript. The choice (3.6) of the characteristic velocities ensures that $\lambda_L < 0 < \lambda_R$, which increases the numerical diffusion of the scheme. As a consequence of this increased diffusion, the scheme is more stable. Therefore, the choice of the characteristic velocities allowed us to postpone a more precise study of the stability.

The question of the entropy preservation of the well-balanced scheme derived in Chapter 3 could also be raised. Indeed, for any source term, the intermediate states (3.81) of this well-balanced scheme are written as the intermediate states of the HLL scheme (3.20) with an additional term. This term depends linearly on the space step. Thus, the well-balanced intermediate states can be viewed as perturbations of the intermediate states of the HLL scheme. Moreover, the HLL scheme is known to be entropy-preserving (see [90]). Therefore, quantifying the extent of that perturbation could provide an adequate entropy inequality for the well-balanced scheme, in order to determine whether this scheme is entropy-preserving for the topography source term, the Manning friction source term, or even both source terms. The characteristic velocities λ_L and λ_R would certainly play a role in this entropy inequality. In addition, the value of the cutoff constant C present in (3.54), introduced in order to make the approximate topography source term $\overline{S}^t$ consistent, could also be relevant to uncover the entropy inequality.

Assuming the scheme was indeed entropy-preserving, another detector could then be added to the MOOD method, following [16]. In [16], the authors suggest using the entropy inequality to introduce a new MOOD criterion. If the entropy inequality is not satisfied because of the reconstruction procedure, then the degree of the reconstruction is lowered until the inequality is verified. Such a criterion would supplement the PAD, DMP and u2 detection criteria already present in the MOOD loop.

High-order accuracy: results and improvements

The scheme suggested in Chapter 4 is well-balanced and high-order accurate, as shown in the numerical experiments proposed in Section 4.4.1 and Section 4.4.2. However, because of the convex combination procedure introduced in Section 4.2.2, the first-order scheme is used

when the solution is close to a steady state, and the high-order scheme is used otherwise. This results in a formally high-order scheme, since it is either high-order for unsteady states, or exact (i.e. of infinite order) for steady states. However, a rigorous proof of this high-order accuracy could be studied.

In addition, in [33, 31, 35], the authors proposed a reconstruction procedure based on the moving steady state solutions. Such a procedure ensures that the suggested scheme is truly high-order accurate, without the need for a convex combination, and that it exactly preserves the moving steady states. As a consequence, adapting the procedure from [33, 31, 35] to the scheme developed in this manuscript could be an interesting addition. However, this reconstruction requires approximately solving the steady state relations, which leads to solving a nonlinear equation every time the reconstruction is computed; one of the advantages of the scheme proposed in this manuscript was to ensure that no such nonlinear solvers were required in its derivation.

Appendices

A The Rankine-Hugoniot relations for a balance law

The purpose of this appendix is to derive the Rankine-Hugoniot relations in the case of a balance law. We consider the following Cauchy problem for a 1D balance law:

$$\begin{cases} \partial_t W + \partial_x F(W) = S(W), \\ W(0,x) = W_0(x). \end{cases} \tag{A.1}$$

We begin by stating the definition of a weak solution of (A.1).

Definition A.1. A *weak solution* of the Cauchy problem (A.1) is a function $W \in L^\infty_{loc}(\mathbb{R}_+ \times \mathbb{R})$ such that for all $\varphi \in \mathcal{C}^1_0(\mathbb{R}_+ \times \mathbb{R})$, the following identity holds:

$$\int_{\mathbb{R}_+ \times \mathbb{R}} (W \partial_t \varphi + F(W) \partial_x \varphi + S(W)\varphi)\, dt\, dx = -\int_{\mathbb{R}} W_0(x)\varphi(0,x) dx. \tag{A.2}$$

Equipped with this definition, we can state the Rankine-Hugoniot relations.

Proposition A.2. Consider $W \in L^\infty_{loc}(\Omega)$, with $\Omega \subset \mathbb{R}_+ \times \mathbb{R}$. Assume that W is a weak solution of the Cauchy problem (A.1). In addition, assume that a discontinuity curve Γ, parameterizable in t, separates the domain Ω in two sub-domains Ω_L and Ω_R, as displayed on Figure A.1. We also assume that W is a classical solution of (A.1) on Ω_L and Ω_R. We denote by W_L and W_R the limits of W to the left and the right of the discontinuity curve. Then, the following *Rankine-Hugoniot relations* hold:

$$\sigma(W_R - W_L) = F(W_R) - F(W_L),$$

where σ is the velocity of the discontinuity.

Proof. Let W be a weak solution of (A.1) on Ω. Let $\varphi \in \mathcal{T}$ be a test function, with

$$\mathcal{T} = \{\varphi \in \mathcal{C}^1_0(\mathbb{R}_+ \times \mathbb{R}) \mid \varphi(0,x) = 0\}.$$

Therefore, (A.2) yields:

$$\int_\Omega (W \partial_t \varphi + F(W) \partial_x \varphi)\, dt\, dx = \int_\Omega S(W)\varphi\, dt\, dx.$$

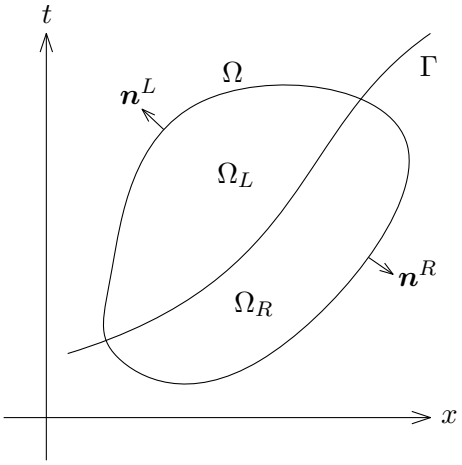

Figure A.1 – The set Ω and the discontinuity curve Γ in the (x,t)-plane.

Since $\Omega = \Omega_L \cup \Omega_R$, we have:

$$\begin{aligned}\int_{\Omega_L} (W\partial_t\varphi + F(W)\partial_x\varphi)\, dt\, dx + \int_{\Omega_R} (W\partial_t\varphi + F(W)\partial_x\varphi)\, dt\, dx \\ + \int_{\Omega_L} S(W)\varphi\, dt\, dx + \int_{\Omega_R} S(W)\varphi\, dt\, dx = 0.\end{aligned} \tag{A.3}$$

Let $\boldsymbol{n}^L = {}^t(n_x^L, n_t^L)$ and $\boldsymbol{n}^R = {}^t(n_x^R, n_t^R)$ be the respective unit outer-pointing normal vectors to the subsets Ω_L and Ω_R. These vectors are depicted on Figure A.1.

Arguing Green's theorem, the first integral of (A.3) rewrites:

$$\begin{aligned}\int_{\Omega_L} (W\partial_t\varphi + F(W)\partial_x\varphi)\, dt\, dx = -\int_{\Omega_L} \varphi(\partial_t W + \partial_x F(W))\, dt\, dx \\ + \int_{\partial\Omega_L} \big(W\varphi n_t^L + F(W)\varphi n_x^L\big) ds,\end{aligned} \tag{A.4}$$

where ds is an element of length of $\partial\Omega_L$. Since φ is compactly supported, we have:

$$\int_{\partial\Omega_L} \big(W\varphi n_t^L + F(W)\varphi n_x^L\big) ds = \int_{\Omega\cap\Gamma} \big(W_L\varphi n_t^\Gamma + F(W_L)\varphi n_x^\Gamma\big) ds, \tag{A.5}$$

where $\boldsymbol{n}^\Gamma = {}^t(n_x^\Gamma, n_t^\Gamma)$ is the unit normal vector to Γ, pointing from Ω_L to Ω_R. Combining equations (A.4) and (A.5) yield the following evaluation of the first integral of (A.3):

$$\begin{aligned}\int_{\Omega_L} (W\partial_t\varphi + F(W)\partial_x\varphi)\, dt\, dx = -\int_{\Omega_L} \varphi(\partial_t W + \partial_x F(W))\, dt\, dx \\ + \int_{\Omega\cap\Gamma} \big(W_L\varphi n_t^\Gamma + F(W_L)\varphi n_x^\Gamma\big) ds.\end{aligned} \tag{A.6}$$

Performing similar computations within the second integral of (A.3), we get:

$$\begin{aligned}\int_{\Omega_R} (W\partial_t\varphi + F(W)\partial_x\varphi)\, dt\, dx = -\int_{\Omega_R} \varphi(\partial_t W + \partial_x F(W))\, dt\, dx \\ + \int_{\Omega\cap\Gamma} \big(W_R\varphi(-n_t^\Gamma) + F(W_R)\varphi(-n_x^\Gamma)\big) ds.\end{aligned} \tag{A.7}$$

Finally, plugging (A.6) and (A.7) into (A.3) yields:

$$\begin{aligned}\int_{\Omega_L} \varphi(\partial_t W + \partial_x F(W) - S(W))\, dt\, dx + \int_{\Omega_R} \varphi(\partial_t W + \partial_x F(W) - S(W))\, dt\, dx \\ = \int_{\Omega\cap\Gamma} \varphi\Big[(W_L - W_R)\, n_t^\Gamma + (F(W_L) - F(W_R))\, n_x^\Gamma\Big] ds.\end{aligned} \tag{A.8}$$

Recall that W is a classical solution of the Cauchy problem (A.1) on both Ω_L and Ω_R. Therefore, on these sets, the identity $\partial_t W + \partial_x F(W) = S(W)$ is satisfied. As a consequence, the left-hand side of (A.8) vanishes, and we get:

$$\int_{\Omega\cap\Gamma} \varphi\Big[(W_L - W_R)\, n_t^\Gamma + (F(W_L) - F(W_R))\, n_x^\Gamma\Big] ds = 0. \tag{A.9}$$

Recall that Γ is parameterizable in t. We denote by $\gamma(t)$ its parameterization. This property ensures that the first component n_x^Γ of the normal vector to the curve Γ never vanishes. Therefore, the following sequence of equalities holds:

$$\frac{n_t^\Gamma}{n_x^\Gamma} = \frac{d\gamma}{dt} = \sigma, \tag{A.10}$$

where σ is the velocity of the discontinuity. Arguing (A.10), (A.9) rewrites as follows:

$$\int_{\Omega\cap\Gamma} \varphi\Big[(W_L - W_R)\, \sigma + (F(W_L) - F(W_R))\Big] n_x^\Gamma\, ds = 0.$$

Since the above identity is valid for all $\varphi \in \mathcal{T}$, the following equation necessarily holds true:

$$\sigma(W_R - W_L) = F(W_R) - F(W_L). \tag{A.11}$$

The relations (A.11) turn out to be the Rankine-Hugoniot relations we sought. The proof is thus achieved. □

B High-order quadrature rules

In this appendix, we give the coefficients of the quadrature rules (2.71) and (2.73). These formulas are given in [2], but we recall them here for the sake of completeness.

We first focus on the quadrature formula (2.71), used to approximate the integral of a function φ on an edge $\mathfrak{e}$. The quadrature formula reads:

$$\frac{1}{|\mathfrak{e}|}\int_{\mathfrak{e}} \varphi(\boldsymbol{\sigma}) d\boldsymbol{\sigma} \simeq \sum_{r=1}^{R} \xi_r \varphi(\boldsymbol{\sigma}_r), \tag{B.1}$$

where R denotes the number of quadrature points, $(\xi_r)_{r\in[\![1,R]\!]}$ are the quadrature weights, and $(\boldsymbol{\sigma}_r)_{r\in[\![1,R]\!]}$ are the quadrature points. Let $\boldsymbol{a} \in \mathbb{R}^2$ and $\boldsymbol{b} \in \mathbb{R}^2$ be the endpoints of the edge $\mathfrak{e}$. Then, the quadrature points $\boldsymbol{\sigma}_r$ are such that

$$\boldsymbol{\sigma}_r = \frac{1}{2}[\boldsymbol{a}(1-\alpha_r) + \boldsymbol{b}(1+\alpha_r)]. \tag{B.2}$$

The coefficients α_r, as well as the weights ξ_r from (B.1), are taken according to Table B.1.

degree	R	α_r	ξ_r
1	1	0	1
2, 3	2	$-\sqrt{\frac{1}{3}}; \sqrt{\frac{1}{3}}$	$\frac{1}{2}; \frac{1}{2}$
4, 5	3	$-\sqrt{\frac{3}{5}}; 0; \sqrt{\frac{3}{5}}$	$\frac{5}{18}; \frac{4}{9}; \frac{5}{18}$

Table B.1 – High-order quadrature rule on an edge.

Note that, in the case of a Cartesian mesh discussed in Chapter 4, we only have vertical or horizontal edges. Let us consider the cell defined as follows:

$$\mathfrak{c} = \left(-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right) \times \left(-\frac{\Delta y}{2}, \frac{\Delta y}{2}\right). \tag{B.3}$$

In this case, the horizontal and vertical edges are respectively given by:

$$\mathfrak{e}_H = \left(-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right) \times \left\{\pm\frac{\Delta y}{2}\right\} \quad \text{and} \quad \mathfrak{e}_V = \left\{\pm\frac{\Delta x}{2}\right\} \times \left(-\frac{\Delta y}{2}, \frac{\Delta y}{2}\right).$$

Therefore, following (B.2), the quadrature points on the horizontal edges are respectively defined by:

$$\boldsymbol{\sigma}_r^H = \left(\alpha_r \frac{\Delta x}{2}; \pm\frac{\Delta y}{2}\right) \quad \text{and} \quad \boldsymbol{\sigma}_r^V = \left(\pm\frac{\Delta x}{2}; \alpha_r \frac{\Delta y}{2}\right). \tag{B.4}$$

We finally focus on the quadrature formula (2.73) applied to the Cartesian cell $\mathbf{c}$ defined by (B.3). For a function ψ defined on the cell, we get:

$$\frac{1}{|\mathbf{c}|}\int_{\mathbf{c}} \psi(\boldsymbol{x})d\boldsymbol{x} \simeq \sum_{q=1}^{Q} \eta_q \psi(\boldsymbol{X}_q),$$

where Q denotes the number of quadrature points, $(\eta_q)_{q\in[\![1,Q]\!]}$ are the quadrature weights, and $(\boldsymbol{X}_q)_{q\in[\![1,Q]\!]}$ are the quadrature points. For the square cell $\mathbf{c}$, we heavily rely on the edge quadrature to build the cell quadrature. Indeed, we take $Q = R^2$, and we set the following values of the quadrature weights and points:

$$(\eta_q)_{1\le q\le R^2} = (\xi_i\xi_j)_{1\le i,j\le R} \quad \text{and} \quad (\boldsymbol{X}_q)_{1\le q\le R^2} = \left(\alpha_i \frac{\Delta x}{2}; \alpha_j \frac{\Delta y}{2}\right)_{1\le i,j\le R}.$$

Note that the above quadrature is obtained by combining the horizontal and vertical edge quadrature points given by (B.4) and multiplying the relevant edge quadrature weights.

C Coefficients of the SSPRK methods

The SSPRK (Strong Stability-Preserving Runge-Kutta) methods consist in giving adequate values to the coefficients α_{lk} and β_{lk} of the general Runge-Kutta method (2.76). We denote by SSPRK(m,p) the SSPRK method of order p and made of m steps; for α_{lk} and β_{lk}, we have $1 \leq l \leq m$ and $0 \leq k \leq l-1$.

First, we display on Table C.1 the choice of the time integrator with respect to the degree d of the reconstruction. For $d > 3$, recall that the time step has to be modified according to (2.77).

$d = 1$	$d = 2$	$d \geq 3$
SSPRK(2,2)	SSPRK(3,3)	SSPRK(5,4)

Table C.1 – Choice of the SSPRK method with respect to the degree d of the reconstruction.

Then, after [84, 137], we give in the following tables the values of the coefficients α_{lk} and β_{lk} for the above methods:

- in Table C.2: values of α_{lk} and β_{lk} for the SSPRK(2,2) integrator;
- in Table C.3: values of α_{lk} and β_{lk} for the SSPRK(3,3) integrator;
- in Table C.4: values of α_{lk} for the SSPRK(5,4) integrator;
- in Table C.5: values of β_{lk} for the SSPRK(5,4) integrator.

1	
$1/2$	$1/2$

1	
0	$1/2$

Table C.2 – Coefficients α_{lk} (left table) and β_{lk} (right table) for the SSPRK(2,2) method. Rows: $1 \leq l \leq 2$; columns: $0 \leq k \leq 1$.

1		
$3/4$	$1/4$	
$1/3$	0	$2/3$

1		
0	$1/4$	
0	0	$2/3$

Table C.3 – Coefficients α_{lk} (left table) and β_{lk} (right table) for the SSPRK(3,3) method. Rows: $1 \leq l \leq 3$; columns: $0 \leq k \leq 2$.

1				
0.444370493651235	0.555629506348765			
0.620101851488403	0	0.379898148511597		
0.178079954393132	0	0	0.821920045606868	
0	0	0.517231671970585	0.096059710526147	0.386708617503269

Table C.4 – Coefficients α_{lk} for the SSPRK(5,4) method. Rows: $1 \leq l \leq 5$; columns: $0 \leq k \leq 4$.

0.391752226571890				
0	0.368410593050371			
0	0	0.251891774271694		
0	0	0	0.544974750228521	
0	0	0	0.063692468666290	0.226007483236906

Table C.5 – Coefficients β_{lk} for the SSPRK(5,4) method. Rows: $1 \leq l \leq 5$; columns: $0 \leq k \leq 4$.

Bibliography

[1] R. Abgrall. On essentially non-oscillatory schemes on unstructured meshes: Analysis and implementation. *J. Comput. Phys.*, 114(1):45–58, 1994. Cited page 93.

[2] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York, 1992. Reprint of the 1972 edition. Cited pages 95, 96, and 239.

[3] F. Alcrudo and P. Garcia-Navarro. A high-resolution Godunov-type scheme in finite volumes for the 2D shallow-water equations. *Internat. J. Numer. Methods Fluids*, 16(6):489–505, 1993. Cited page 175.

[4] K. Anastasiou and C. T. Chan. Solution of the 2D shallow water equations using the finite volume method on unstructured triangular meshes. *Internat. J. Numer. Methods Fluids*, 24(11):1225–1245, 1997. Cited page 175.

[5] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput.*, 25(6):2050–2065, 2004. Cited pages 12, 19, 26, 27, 53, 69, and 137.

[6] E. Audusse and M.-O. Bristeau. A well-balanced positivity preserving "second-order" scheme for shallow water flows on unstructured meshes. *J. Comput. Phys.*, 206(1):311–333, 2005. Cited page 69.

[7] E. Audusse, C. Chalons, and P. Ung. A simple well-balanced and positive numerical scheme for the shallow-water system. *Commun. Math. Sci.*, 13(5):1317–1332, 2015. Cited pages 15, 22, 47, 69, 104, and 115.

[8] E. Audusse, R. Klein, and A. Owinoh. Conservative discretization of Coriolis force in a finite volume framework. *J. Comput. Phys.*, 228(8):2934–2950, 2009. Cited pages 229 and 233.

[9] M. A. Baptista, J. M. Miranda, R. Omira, and C. Antunes. Potential inundation of Lisbon downtown by a 1755-like tsunami. *Nat. Hazards Earth Syst. Sci.*, 11(12):3319–3326, dec 2011. Cited pages 11 and 18.

[10] A. J.-C. Barré de Saint-Venant. Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit. *Extrait des Comptes rendus de l'Académie des Sciences, tome LXXIII, séances des 17 et 24 juillet 1871*, 1871. Cited page 25.

[11] A. Bermudez and M. E. Vazquez. Upwind methods for hyperbolic conservation laws with source terms. *Comput. & Fluids*, 23(8):1049–1071, 1994. Cited pages 12, 19, and 68.

[12] C. Berthon and C. Chalons. A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations. *Math. Comp.*, 85(299):1281–1307, 2016. Cited pages 27, 41, 53, 69, 104, 122, and 123.

[13] C. Berthon, C. Chalons, S. Cornet, and G. Sperone. Fully well-balanced, positive and simple approximate Riemann solver for shallow water equations. *Bull. Braz. Math. Soc. (N.S.)*, 47(1):117–130, 2016. Cited pages 69, 122, and 123.

[14] C. Berthon, P. Charrier, and B. Dubroca. An HLLC scheme to solve the M_1 model of radiative transfer in two space dimensions. *J. Sci. Comput.*, 31(3):347–389, 2007. Cited page 103.

[15] C. Berthon, A. Crestetto, and F. Foucher. A Well-Balanced Finite Volume Scheme for a Mixed Hyperbolic/Parabolic System to Model Chemotaxis. *J. Sci. Comput.*, 67(2):618–643, 2016. Cited pages 69, 103, and 115.

[16] C. Berthon and V. Desveaux. An entropy preserving MOOD scheme for the Euler equations. *Int. J. Finite Vol.*, 11, 2014. Cited pages 98, 101, 231, and 235.

[17] C. Berthon, B. Dubroca, and A. Sangam. An entropy preserving relaxation scheme for ten-moments equations with source terms. *Commun. Math. Sci.*, 13(8):2119–2154, 2015. Cited page 89.

[18] C. Berthon and F. Foucher. Hydrostatic upwind schemes for shallow-water equations. In *Finite volumes for complex applications. VI. Problems & perspectives. Volume 1, 2*, volume 4 of *Springer Proc. Math.*, pages 97–105. Springer, Heidelberg, 2011. Cited page 163.

[19] C. Berthon and F. Foucher. Efficient well-balanced hydrostatic upwind schemes for shallow-water equations. *J. Comput. Phys.*, 231(15):4993–5015, 2012. Cited pages 47, 69, 89, and 209.

[20] C. Berthon and F. Marche. A positive preserving high order VFRoe scheme for shallow-water equations: a class of relaxation schemes. *SIAM J. Sci. Comput.*, 30(5):2587–2612, 2008. Cited pages 163 and 164.

[21] C. Berthon, G. Moebs, C. Sarazin-Desbois, and R. Turpault. An asymptotic-preserving scheme for systems of conservation laws with source terms on 2D unstructured meshes. *Commun. Appl. Math. Comput. Sci.*, 11(1):55–77, 2016. Cited page 89.

[22] C. Berthon, G. Moebs, and R. Turpault. An asymptotic-preserving scheme for systems of conservation laws with source terms on 2D unstructured meshes. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, volume 77 of *Springer Proc. Math. Stat.*, pages 107–115. Springer, Cham, 2014. Cited page 89.

[23] A. Bollermann, G. Chen, A. Kurganov, and S. Noelle. A well-balanced reconstruction of wet/dry fronts for the shallow water equations. *J. Sci. Comput.*, 56(2):267–290, 2013. Cited page 69.

[24] A. Bollermann, S. Noelle, and M. Lukáčová-Medviďová. Finite volume evolution Galerkin methods for the shallow water equations with dry beds. *Commun. Comput. Phys.*, 10(2):371–404, 2011. Cited page 164.

[25] W. Boscheri, R. Loubère, and M. Dumbser. Direct arbitrary-Lagrangian-Eulerian ADER-MOOD finite volume schemes for multidimensional hyperbolic conservation laws. *J. Comput. Phys.*, 292:56–87, 2015. Cited page 98.

[26] F. Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2004. Cited pages 47, 69, and 132.

[27] F. Bouchut, J. Le Sommer, and V. Zeitlin. Frontal geostrophic adjustment and nonlinear wave phenomena in one-dimensional rotating shallow water. II. High-resolution numerical simulations. *J. Fluid Mech.*, 514:35–63, 2004. Cited pages 229 and 233.

[28] F. Bouchut and T. Morales de Luna. A subsonic-well-balanced reconstruction scheme for shallow water flows. *SIAM J. Numer. Anal.*, 48(5):1733–1758, 2010. Cited page 69.

[29] S. Bryson, Y. Epshteyn, A. Kurganov, and G. Petrova. Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system. *ESAIM Math. Model. Numer. Anal.*, 45(3):423–446, 2011. Cited page 69.

[30] T. Buffard, T. Gallouët, and J.-M. Hérard. A sequel to a rough Godunov scheme: application to real gases. *Comput. & Fluids*, 29(7):813–847, 2000. Cited page 163.

[31] M. Castro, J. M. Gallardo, J. A. López-García, and C. Parés. Well-balanced high order extensions of Godunov's method for semilinear balance laws. *SIAM J. Numer. Anal.*, 46(2):1012–1039, 2008. Cited pages 231 and 236.

[32] M. Castro, J. M. Gallardo, and C. Parés. High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. Applications to shallow-water systems. *Math. Comp.*, 75(255):1103–1134, 2006. Cited page 69.

[33] M. J. Castro, A. Pardo Milanés, and C. Parés. Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique. *Math. Models Methods Appl. Sci.*, 17(12):2055–2113, 2007. Cited pages 69, 159, 231, and 236.

[34] M. J. Castro Díaz, E. D. Fernández-Nieto, T. Morales de Luna, G. Narbona-Reina, and C. Parés. A HLLC scheme for nonconservative hyperbolic problems. Application to turbidity currents with sediment transport. *ESAIM Math. Model. Numer. Anal.*, 47(1):1–32, 2013. Cited page 103.

[35] M. J. Castro Díaz, J. A. López-García, and C. Parés. High order exactly well-balanced numerical methods for shallow water systems. *J. Comput. Phys.*, 246:242–264, 2013. Cited pages 69, 231, and 236.

[36] T. Chacón Rebollo, A. Domínguez Delgado, and E. D. Fernández Nieto. Numerical schemes for 2D shallow water equations with variable depth and friction effects. In *Mathematical and numerical aspects of wave propagation—WAVES 2003*, pages 506–511. Springer, Berlin, 2003. Cited page 175.

[37] C. Chalons, F. Coquel, E. Godlewski, P.-A. Raviart, and N. Seguin. Godunov-type schemes for hyperbolic systems with parameter-dependent source. The case of Euler system with friction. *Math. Models Methods Appl. Sci.*, 20(11):2109–2166, 2010. Cited page 69.

[38] R. Chandra, L. Dagum, D. Kohr, D. Maydan, J. McDonald, and R. Menon. *Parallel Programming in OpenMP*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001. Cited page 193.

[39] P. Chandrashekar and C. Klingenberg. A second order well-balanced finite volume scheme for Euler equations with gravity. *SIAM J. Sci. Comput.*, 37(3):B382–B402, 2015. Cited page 69.

[40] B. Chapman, G. Jost, and R. van der Pas. *Using OpenMP: Portable Shared Memory Parallel Programming (Scientific and Engineering Computation)*. The MIT Press, 2007. Cited page 193.

[41] A. Chertock, S. Cui, A. Kurganov, Ş. N. Özcan, and E. Tadmor. Well-Balanced Central-Upwind Schemes for the Euler Equations with Gravitation. working paper or preprint, 2015. Cited page 69.

[42] A. Chertock, S. Cui, A. Kurganov, and T. Wu. Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms. *Internat. J. Numer. Methods Fluids*, 78(6):355–383, 2015. Cited pages 64, 69, and 152.

[43] A. Chertock, M. Dudzinski, A. Kurganov, and M. Lukáčová-Medviďová. Well-Balanced Schemes for the Shallow Water Equations with Coriolis Forces. working paper or preprint, 2014. Cited pages 69, 229, and 233.

[44] A. Chinnayya, A.-Y. LeRoux, and N. Seguin. A well-balanced numerical scheme for the approximation of the shallow-water equations with topography: the resonance phenomenon. *Int. J. Finite Vol.*, 1(1):33, 2004. Cited pages 13, 16, 20, 22, 46, 47, and 69.

[45] V. T. Chow. *Open-channel hydraulics*. McGraw-Hill civil engineering series. McGraw-Hill, 1959. Cited pages 138 and 215.

[46] S. Clain, S. Diot, and R. Loubère. A high-order finite volume method for systems of conservation laws—Multi-dimensional Optimal Order Detection (MOOD). *J. Comput. Phys.*, 230(10):4028–4050, 2011. Cited pages 15, 21, 84, 91, 92, 98, and 101.

[47] S. Clain and J. Figueiredo. The MOOD method for the non-conservative shallow-water system. working paper or preprint, oct 2014. Cited pages 69, 98, 99, 100, 201, 203, 209, and 213.

[48] S. Clain, J. Figueiredo, R. Louère, and S. Diot. An overview of the multidimensional optimal order detection method. In *SYMCOMP 2015, Faro, March 26-27, 2015, Portugal*, pages 69–87. ECCOMAS, 2015. Cited page 98.

[49] S. Clain, G. J. Machado, J. M. Nóbrega, and R. M. S. Pereira. A sixth-order finite volume method for multidomain convection-diffusion problem with discontinuous coefficients. *Comput. Methods Appl. Mech. Eng.*, 267:43–64, 2013. Cited page 98.

[50] S. Clain, C. Reis, R. Costa, J. Figueiredo, M. A. Baptista, and J. M. Miranda. The Tagus 1969 tsunami simulation with a finite volume solver and the hydrostatic reconstruction technique. working paper or preprint, December 2015. Cited pages 11, 18, and 98.

[51] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, editors. *Discontinuous Galerkin methods*, volume 11 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, 2000. Theory, computation and applications, Papers from the 1st International Symposium held in Newport, RI, May 24–26, 1999. Cited page 84.

[52] B. Cockburn, S. Y. Lin, and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. III. One-dimensional systems. *J. Comput. Phys.*, 84(1):90–113, 1989. Cited page 84.

[53] B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. *Math. Comp.*, 52(186):411–435, 1989. Cited page 84.

[54] R. Costa, G. J. Machado, and S. Clain. A sixth-order finite volume method for the 1D biharmonic operator: application to intramedullary nail simulation. *Int. J. Appl. Math. Comput. Sci.*, 25(3):529–537, 2015. Cited page 98.

[55] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100:32–74, 1928. Cited page 74.

[56] O. Delestre and P.-Y. Lagrée. A well-balanced finite volume scheme for 1D hemodynamic simulations. *ESAIM, Proc.*, 35:222–227, 2012. Cited page 69.

[57] O. Delestre, C. Lucas, P.-A. Ksinant, F. Darboux, C. Laguerre, T.-N.-T. Vo, F. James, and S. Cordier. SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies. *Internat. J. Numer. Methods Fluids*, 72(3):269–300, 2013. Cited pages 11, 18, 39, and 169.

[58] A. I. Delis and T. Katsaounis. Numerical solution of the two-dimensional shallow water equations by the application of relaxation methods. *Appl. Math. Model.*, 29(8):754 – 783, 2005. Cited page 175.

[59] P. J. Dellar and R. Salmon. Shallow water equations with a complete Coriolis force and topography. *Phys. Fluids*, 17(10):19, 2005. Cited pages 229 and 233.

[60] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. A well-balanced scheme for the Euler equation with a gravitational potential. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, volume 77 of *Springer Proc. Math. Stat.*, pages 217–226. Springer, Cham, 2014. Cited pages 69 and 131.

[61] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. A well-balanced scheme to capture non-explicit steady states in the Euler equations with gravity. *Internat. J. Numer. Methods Fluids*, 81(2):104–127, 2016. Cited page 69.

[62] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. Well-balanced schemes to capture non-explicit steady states: Ripa model. *Math. Comp.*, 85(300):1571–1602, 2016. Cited page 69.

[63] S. Diot, S. Clain, and R. Loubère. Improved detection criteria for the multi-dimensional optimal order detection (MOOD) on unstructured meshes with very high-order polynomials. *Comput. & Fluids*, 64:43–63, 2012. Cited pages 15, 21, 84, 91, 92, and 98.

[64] S. Diot, M. M. François, and E. D. Dendy. A higher-order unsplit 2D direct Eulerian finite volume method for two-material compressible flows based on the MOOD paradigms. *Internat. J. Numer. Methods Fluids*, 76(12):1064–1087, 2014. Cited page 98.

[65] S. Diot, R. Loubère, and S. Clain. The multidimensional optimal order detection method in the three-dimensional case: very high-order finite volume method for hyperbolic systems. *Internat. J. Numer. Methods Fluids*, 73(4):362–392, 2013. Cited pages 15, 21, 84, 91, 92, and 98.

[66] F. Dubois and P. G. LeFloch. Boundary conditions for nonlinear hyperbolic systems of conservation laws. *J. Differ. Equations*, 71(1):93–122, 1988. Cited page 163.

[67] M. Dudzinski and M. Lukáčová-Medviďová. Well-balanced bicharacteristic-based scheme for multilayer shallow water flows including wet/dry fronts. *J. Comput. Phys.*, 235:82–113, 2013. Cited page 69.

[68] M. Dumbser, O. Zanotti, R. Loubère, and S. Diot. *A posteriori* subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *J. Comput. Phys.*, 278:47–75, 2014. Cited page 98.

[69] A. Duran, Q. Liang, and F. Marche. On the well-balanced numerical discretization of shallow water equations on unstructured meshes. *J. Comput. Phys.*, 235:565–586, 2013. Cited page 69.

[70] E. D. Fernández-Nieto, D. Bresch, and J. Monnier. A consistent intermediate wave speed for a well-balanced HLLC solver. *C. R. Math. Acad. Sci. Paris*, 346(13-14):795–800, 2008. Cited page 69.

[71] J. Figueiredo and S. Clain. Second-order finite volume MOOD method for the shallow water with dry/wet interface. In *SYMCOMP 2015, Faro, March 26-27, 2015, Portugal*, pages 191–205. ECCOMAS, 2015. Cited pages 84, 98, 99, and 100.

[72] U. S. Fjordholm, S. Mishra, and E. Tadmor. Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography. *J. Comput. Phys.*, 230(14):5587–5609, 2011. Cited page 69.

[73] O. Friedrich. Weighted Essentially Non-Oscillatory Schemes for the Interpolation of Mean Values on Unstructured Grids. *J. Comput. Phys.*, 144(1):194–212, 1998. Cited page 93.

[74] J. M. Gallardo, M. Castro, C. Parés, and J. M. González-Vida. On a well-balanced high-order finite volume scheme for the shallow water equations with bottom topography and dry areas. In *Hyperbolic problems: theory, numerics, applications*, pages 259–270. Springer, Berlin, 2008. Cited pages 47 and 69.

[75] G. Gallice. Solveurs simples positifs et entropiques pour les systèmes hyperboliques avec terme source. *C. R. Math. Acad. Sci. Paris*, 334(8):713–716, 2002. Cited page 104.

[76] G. Gallice. Positive and entropy stable Godunov-type schemes for gas dynamics and MHD equations in Lagrangian or Eulerian coordinates. *Numer. Math.*, 94(4):673–713, 2003. Cited page 103.

[77] T. Gallouët, J.-M. Hérard, and N. Seguin. Some approximate Godunov schemes to compute shallow-water equations with topography. *Comput. & Fluids*, 32(4):479–513, 2003. Cited pages 137, 138, 142, 143, 160, 161, 163, and 164.

[78] P. Garcia-Navarro and M. E. Vazquez-Cendon. On numerical treatment of the source terms in the shallow water equations. *Comput. & Fluids*, 29(8):951–979, 2000. Cited pages 229 and 234.

[79] E. Godlewski and P.-A. Raviart. *Hyperbolic systems of conservation laws*, volume 3/4 of *Mathématiques & Applications (Paris) [Mathematics and Applications]*. Ellipses, Paris, 1991. Cited pages 31, 32, 67, and 68.

[80] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1996. Cited pages 13, 19, 26, 68, and 179.

[81] S. K. Godunov. A Difference Scheme for Numerical Solution of Discontinuous Solution of Hydrodynamic Equations. *Mat. Sb., Nov. Ser.*, 47:271–306, 1959. Cited pages 14, 21, 73, and 74.

[82] L. Gosse. A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. *Comput. Math. Appl.*, 39(9-10):135–159, 2000. Cited pages 12, 19, and 68.

[83] S. Gottlieb. On high order strong stability preserving Runge-Kutta and multi step time discretizations. *J. Sci. Comput.*, 25(1-2):105–128, 2005. Cited page 97.

[84] S. Gottlieb and C.-W. Shu. Total variation diminishing Runge-Kutta schemes. *Math. Comp.*, 67(221):73–85, 1998. Cited pages 15, 21, 97, and 241.

[85] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112, 2001. Cited page 97.

[86] N. Goutal and F. Maurel. Proceedings of the 2nd Workshop on Dam-Break Wave Simulation. Technical report, Groupe Hydraulique Fluviale, Département Laboratoire National d'Hydraulique, Electricité de France, 1997. Cited pages 16, 22, 47, 51, 53, 68, 137, 139, 155, and 156.

[87] J. M. Greenberg and A.-Y. LeRoux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.*, 33(1):1–16, 1996. Cited pages 12, 19, 47, and 68.

[88] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes, {III}. *J. Comput. Phys.*, 71(2):231–303, 1987. Cited page 84.

[89] A. Harten and P. D. Lax. A random choice finite difference scheme for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 18:289–315, 1981. Cited page 78.

[90] A. Harten, P. D. Lax, and B. van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Rev.*, 25(1):35–61, 1983. Cited pages 14, 21, 76, 77, 78, 81, 111, 119, 131, 132, 137, 230, and 235.

[91] H. S. Hassan, K. T. Ramadan, and S. N. Hanna. Numerical Solution of the Rotating Shallow Water Flows with Topography Using the Fractional Steps Method. *AM*, 01(02):104–117, 2010. Cited pages 229 and 233.

[92] C. Hirsch. *Numerical computation of internal and external flows. Volume 1: Fundamentals of numerical discretization.* Chichester (UK) etc.: Wiley, 1988. Cited page 71.

[93] C. Hirsch. *Numerical computation of internal and external flows. Volume 2: Computational methods for inviscid and viscous flows.* Chichester etc.: John Wiley & Sons, 1990. Cited page 71.

[94] X. Y. Hu, N. A. Adams, and C.-W. Shu. Positivity-preserving method for high-order conservative schemes solving compressible Euler equations. *J. Comput. Phys.*, 242:169–180, 2013. Cited page 186.

[95] M. E. Hubbard and P. Garcia-Navarro. Flux difference splitting and the balancing of source terms and flux gradients. *J. Comput. Phys.*, 165(1):89–125, 2000. Cited pages 229 and 234.

[96] B. Hunt. Asymptotic solution for dam-break problem. *J. Hydr. Eng. Div.*, 108(1):115–126, 1982. Cited pages 169, 170, and 171.

[97] B. Hunt. Perturbation solution for dam-break floods. *J. Hydraul. Eng.*, 110(8):1058–1071, 1984. Cited pages 169, 170, and 171.

[98] M.J. Ivings, D.M. Causon, and E.F. Toro. On Riemann solvers for compressible liquids. *Int. J. Numer. Methods Fluids*, 28(3):395–418, 1998. Cited page 28.

[99] G.-S. Jiang and C.-W. Shu. Efficient implementation of weighted ENO schemes. *J. Comput. Phys.*, 126(1):202–228, 1996. art. no. 0130. Cited page 84.

[100] S. Jin. A steady-state capturing method for hyperbolic systems with geometrical source terms. *M2AN Math. Model. Numer. Anal.*, 35(4):631–645, 2001. Cited page 69.

[101] R. Käppeli and S. Mishra. Well-balanced schemes for the Euler equations with gravitation. *J. Comput. Phys.*, 259:199–219, 2014. Cited pages 69 and 131.

[102] D. I. Ketcheson and A. C. Robinson. On the practical importance of the SSP property for Runge-Kutta time integrators for some common Godunov-type schemes. *Int. J. Numer. Methods Fluids*, 48(3):271–303, 2005. Cited page 97.

[103] D. Kröner. Finite volume schemes in multidimensions. In *Numerical analysis 1997. Proceedings of the 17th Dundee biennial conference, University of Dundee, GB, June 24–27, 1997*, pages 179–192. Harlow: Longman, 1998. Cited page 86.

[104] C. Y. Kuo, Y. C. Tai, F. Bouchut, A. Mangeney, M. Pelanti, R. F. Chen, and K. J. Chang. Simulation of Tsaoling landslide, Taiwan, based on Saint Venant equations over general topography. *Engineering Geology*, 104(3–4):181 – 189, 2009. Cited pages 11 and 18.

[105] A. Kurganov and G. Petrova. A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system. *Commun. Math. Sci.*, 5(1):133–160, 2007. Cited page 69.

[106] A. Kurganov and G. Petrova. Central-upwind schemes for two-layer shallow water equations. *SIAM J. Sci. Comput.*, 31(3):1742–1773, 2009. Cited page 69.

[107] P. D. Lax. Hyperbolic systems of conservation laws. II. *Commun. Pure Appl. Math.*, 10:537–566, 1957. Cited pages 26 and 32.

[108] P. G. LeFloch. *Hyperbolic systems of conservation laws. The theory of classical and nonclassical shock waves.* Basel: Birkhäuser, 2002. Cited page 26.

[109] P. G. LeFloch and M. D. Thanh. The Riemann problem for the shallow water equations with discontinuous topography. *Commun. Math. Sci.*, 5(4):865–885, 2007. Cited page 41.

[110] P. G. LeFloch and M. D. Thanh. A Godunov-type method for the shallow water equations with discontinuous topography in the resonant regime. *J. Comput. Phys.*, 230(20):7631–7660, 2011. Cited page 41.

[111] R. J. LeVeque. *Numerical methods for conservation laws*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 1992. Cited pages 71 and 98.

[112] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002. Cited pages 13, 19, 26, 28, 31, 32, 83, 84, 86, and 175.

[113] R. J. LeVeque and H. C. Yee. A study of numerical methods for hyperbolic conservation laws with stiff source terms. *J. Comput. Phys.*, 86(1):187–210, 1990. Cited page 132.

[114] C. Liang and Z. Xu. Parametrized maximum principle preserving flux limiters for high order schemes solving multi-dimensional scalar hyperbolic conservation laws. *J. Sci. Comput.*, 58(1):41–60, 2014. Cited page 186.

[115] Q. Liang and F. Marche. Numerical resolution of well-balanced shallow water equations with complex source terms. *Adv. Water Resour.*, 32(6):873–884, 2009. Cited page 69.

[116] X.-D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *J. Comput. Phys.*, 115(1):200–212, 1994. Cited page 84.

[117] R. Loubère, M. Dumbser, and S. Diot. A new family of high order unstructured MOOD and ADER finite volume schemes for multidimensional systems of hyperbolic conservation laws. *Commun. Comput. Phys.*, 16(3):718–763, 2014. Cited page 98.

[118] C. Lucas. Cosine effect on shallow water equations and mathematical properties. *Quart. Appl. Math.*, 67(2):283–310, 2009. Cited pages 229 and 233.

[119] M. Lukáčová-Medviďová, S. Noelle, and M. Kraft. Well-balanced finite volume evolution Galerkin methods for the shallow water equations. *J. Comput. Phys.*, 221(1):122–147, 2007. Cited page 69.

[120] J. Luo, K. Xu, and N. Liu. A well-balanced symplecticity-preserving gas-kinetic scheme for hydrodynamic equations under gravitational field. *SIAM J. Sci. Comput.*, 33(5):2356–2381, 2011. Cited page 131.

[121] A. Majda. *Introduction to PDEs and waves for the atmosphere and ocean*, volume 9 of *Courant Lecture Notes in Mathematics*. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2003. Cited pages 229 and 233.

[122] R. Manning. On the flow of water in open channels and pipes. *Transactions of the Institution of Civil Engineers of Ireland*, 20:161–207, 1890. Cited pages 11, 18, and 25.

[123] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography. *Comput. Math. Appl.*, 72(3):568–593, 2016. Cited pages 47 and 123.

[124] L. A. Monthé. A study of splitting scheme for hyperbolic conservation laws with source terms. *J. Comput. Appl. Math.*, 137(1):1–12, 2001. Cited page 132.

[125] R. Natalini, M. Ribot, and M. Twarogowska. A well-balanced numerical scheme for a one dimensional quasilinear hyperbolic model of chemotaxis. *Commun. Math. Sci.*, 12(1):13–39, 2014. Cited page 69.

[126] I. K. Nikolos and A. I. Delis. An unstructured node-centered finite volume scheme for shallow water flows with wet-dry fronts over complex topography. *Comput. Methods Appl. Mech. Engrg.*, 198(47-48):3723–3750, 2009. Cited page 209.

[127] S. Noelle, N. Pankratz, G. Puppo, and J. R. Natvig. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *J. Comput. Phys.*, 213(2):474–499, 2006. Cited page 69.

[128] S. Noelle, Y. Xing, and C.-W. Shu. High-order well-balanced finite volume WENO schemes for shallow water equation with moving water. *J. Comput. Phys.*, 226(1):29–58, 2007. Cited pages 69, 122, 123, and 159.

[129] R. Omira, M. A. Baptista, J. M. Miranda, E. Toto, C. Catita, and J. Catalão. Tsunami vulnerability assessment of Casablanca-Morocco using numerical modelling and GIS tools. *Nat Hazards*, 54(1):75–95, sep 2009. Cited pages 11 and 18.

[130] C. Parés and M. Castro. On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems. *M2AN Math. Model. Numer. Anal.*, 38(5):821–852, 2004. Cited pages 229 and 234.

[131] B. Perthame and Y. Qiu. A variant of Van Leer's method for multidimensional systems of conservation laws. *J. Comput. Phys.*, 112(2):370–381, 1994. Cited page 84.

[132] B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for Euler equations. *Numer. Math.*, 73(1):119–130, 1996. Cited pages 84 and 89.

[133] B. Perthame and C. Simeoni. A kinetic scheme for the Saint-Venant system with a source term. *Calcolo*, 38(4):201–231, 2001. Cited page 69.

[134] A. Ritter. Die Fortpflanzung der Wasserwellen. *Z. Ver. Dtsch. Ing.*, 36(33):947–954, 1892. Cited page 39.

[135] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43:357–372, 1981. Cited pages 14, 21, and 76.

[136] G. Russo and A. Khe. High order well balanced schemes for systems of balance laws. In *Hyperbolic problems: theory, numerics and applications*, volume 67 of *Proc. Sympos. Appl. Math.*, pages 919–928. Amer. Math. Soc., Providence, RI, 2009. Cited page 69.

[137] S. J. Ruuth. Global optimization of explicit strong-stability-preserving Runge-Kutta methods. *Math. Comp.*, 75(253):183–207, 2006. Cited pages 97 and 241.

[138] S. J. Ruuth and R. J. Spiteri. Two barriers on strong-stability-preserving time discretization methods. *J. Sci. Comput.*, 17(1-4):211–220, 2002. Cited page 97.

[139] S. J. Ruuth and R. J. Spiteri. High-order strong-stability-preserving Runge-Kutta methods with downwind-biased spatial discretizations. *SIAM J. Numer. Anal.*, 42(3):974–996, 2004. Cited page 97.

[140] C. Sánchez-Linares, T. Morales de Luna, and M. J. Castro Díaz. A HLLC scheme for Ripa model. *Appl. Math. Comput.*, 272(part 2):369–384, 2016. Cited page 103.

[141] D. Serre. *Systèmes de lois de conservation. I*. Fondations. [Foundations]. Diderot Editeur, Paris, 1996. Hyperbolicité, entropies, ondes de choc. [Hyperbolicity, entropies, shock waves]. Cited page 67.

[142] D. Serre. *Systems of conservation laws. 1*. Cambridge University Press, Cambridge, 1999. Hyperbolicity, entropies, shock waves, Translated from the 1996 French original by I. N. Sneddon. Cited page 67.

[143] C.-W. Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In *Advanced numerical approximation of nonlinear hyperbolic equations. Lectures given at the 2nd session of the Centro Internazionale Matematico Estivo (C. I. M. E.) held in Cetraro, Italy, June 23–28, 1997*, pages 325–432. Berlin: Springer, 1998. Cited page 84.

[144] C.-W. Shu and S. Osher. Efficient implementation of essentially nonoscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439–471, 1988. Cited page 97.

[145] M. W. Smith, N. J. Cox, and L. J. Bracken. Applying flow resistance equations to overland flows. *Prog. Phys. Geog.*, 31(4):363–387, 2007. Cited page 169.

[146] R. J. Spiteri and S. J. Ruuth. Non-linear evolution using optimal fourth-order strong-stability-preserving Runge-Kutta methods. *Math. Comput. Simul.*, 62(1-2):125–135, 2003. Cited page 97.

[147] G. W. Stewart. *Matrix algorithms. Vol. I.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1998. Basic decompositions. Cited page 93.

[148] W. C. Thacker. Some exact solutions to the nonlinear shallow-water wave equations. *J. Fluid Mech.*, 107:499–508, 1981. Cited pages 229 and 233.

[149] V. A. Titarev and E. F. Toro. ADER: arbitrary high order Godunov approach. In *Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala)*, volume 17, pages 609–618, 2002. Cited page 97.

[150] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics. A practical introduction*. Springer-Verlag, Berlin, third edition, 2009. Cited pages 28, 31, 71, 73, 86, 97, and 132.

[151] E. F. Toro, M. Spruce, and W. Speares. Restoration of the contact surface in the HLL-Riemann solver. *Shock Waves*, 4(1):25–34, 1994. Cited page 103.

[152] R. Touma and C. Klingenberg. Well-balanced central finite volume methods for the Ripa system. *Appl. Numer. Math.*, 97:42–68, 2015. Cited page 69.

[153] A. Valiani, V. Caleffi, and A. Zanni. Case Study: Malpasset Dam-Break Simulation Using a Two-Dimensional Finite Volume Method. *Journal of Hydraulic Engineering*, 128(5):460–472, 2002. Cited pages 11 and 18.

[154] B. van Leer. Towards the Ultimate Conservative Difference Scheme, V. A Second Order Sequel to Godunov's Method. *J. Com. Phys.*, 32:101–136, 1979. Cited pages 14, 21, 84, and 98.

[155] M. E. Vázquez-Cendón. Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. *J. Comput. Phys.*, 148(2):497–526, 1999. Cited pages 229 and 234.

[156] J.-W. Wang and R.-X. Liu. A comparative study of finite volume methods on unstructured meshes for simulation of 2D shallow water wave problems. *Math. Comput. Simulation*, 53(3):171–184, 2000. Cited page 175.

[157] Y. Xing. Exactly well-balanced discontinuous Galerkin methods for the shallow water equations with moving water equilibrium. *J. Comput. Phys.*, 257(part A):536–553, 2014. Cited pages 69 and 159.

[158] Y. Xing and C.-W. Shu. High-order finite volume WENO schemes for the shallow water equations with dry states. *Adv. Water Resour.*, 34(8):1026–1038, 2011. Cited pages 69 and 164.

[159] Y. Xing and C.-W. Shu. A survey of high order schemes for the shallow water equations. *J. Math. Study*, 47(3):221–249, 2014. Cited page 69.

[160] Y. Xing, C.-W. Shu, and S. Noelle. On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations. *J. Sci. Comput.*, 48(1-3):339–349, 2011. Cited page 69.

[161] Y. Xing, X. Zhang, and C.-W. Shu. Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations. *Adv. Water Resour.*, 33(12):1476–1493, 2010. Cited pages 69 and 163.

[162] K. Xu. A well-balanced gas-kinetic scheme for the shallow-water equations with source terms. *J. Comput. Phys.*, 178(2):533–562, 2002. Cited page 131.

[163] K. Xu, J. Luo, and S. Chen. A well-balanced kinetic scheme for gas dynamic equations under gravitational field. *Adv. Appl. Math. Mech.*, 2(2):200–210, 2010. Cited pages 69 and 131.

[164] Z. Xu. Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: one-dimensional scalar problem. *Math. Comp.*, 83(289):2213–2238, 2014. Cited page 186.

[165] V. Zeitlin, S. B. Medvedev, and R. Plougonven. Frontal geostrophic adjustment, slow manifold and nonlinear wave phenomena in one-dimensional rotating shallow water. I. Theory. *J. Fluid Mech.*, 481:269–290, 2003. Cited pages 229 and 233.

[166] F. Zhou, G. Chen, S. Noelle, and H. Guo. A well-balanced stable generalized Riemann problem scheme for shallow water equations using adaptive moving unstructured triangular meshes. *Internat. J. Numer. Methods Fluids*, 73(3):266–283, 2013. Cited page 69.

List of Tables

List of Figures

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Victor MICHEL-DANSAC

Développement de schémas équilibre d'ordre élevé pour des écoulements géophysiques

Development of high-order well-balanced schemes for geophysical flows

Résumé

L'objectif de ce travail est de proposer un schéma numérique pertinent pour les équations de Saint-Venant avec termes source de topographie et de friction de Manning.
Le premier chapitre est dédié à l'étude du système de Saint-Venant muni des termes source. Dans un premier temps, les propriétés algébriques de ce système sont obtenues. Dans un second temps, nous nous intéressons à ses états stationnaires, qui sont étudiés pour les termes source individuels de topographie et de friction.
Le deuxième chapitre permet de rappeler des notions concernant la méthode des volumes finis. Nous évoquons des schémas aux volumes finis pour des systèmes de lois de conservation unidimensionnels et bidimensionnels, et nous en proposons une extension permettant d'assurer un ordre élevé de précision.
Le troisième chapitre concerne la dérivation d'un schéma numérique pour les équations de Saint-Venant avec topographie et friction. Ce schéma permet :

- de préserver tous les états stationnaires ;
- de préserver la positivité de la hauteur d'eau ;
- d'approcher les transitions entre zones mouillées et zones sèches, et ce même en présence de friction.

Des cas-tests mettant en lumière les propriétés du schéma sont présentés.
Le quatrième chapitre permet d'étendre le schéma proposé précédemment, pour prendre en compte des géométries bidimensionnelles et pour assurer un ordre élevé de précision. Des cas-tests numériques sont aussi présentés, y compris des simulations de phénomènes réels.

Mots-clés

équations de Saint-Venant, friction de Manning, schémas de type Godunov, schémas équilibre, états stationnaires en mouvement

Abstract

This manuscript is devoted to a relevant numerical approximation of the shallow-water equations with the source terms of topography and Manning friction.
The first chapter concerns the study of the shallow-water equations, equipped with the aforementioned source terms. Algebraic properties of this system are first obtained. Then, we focus on its steady state solutions for the individual source terms of topography and friction.
The second chapter introduces the finite volume method, which is used throughout the manuscript. One-dimensional and two-dimensional systems of conservation laws are studied, and a high-order strategy is presented.
The third chapter deals with the numerical approximation of the shallow-water equations with topography and friction. We derive a scheme that:

- preserves all the steady states;
- preserves the non-negativity of the water height;
- is able to deal with transitions between wet and dry areas.

Relevant numerical experiments are presented to exhibit these properties.
The fourth chapter is dedicated to extensions of the scheme developed in the third chapter. Namely, the scheme is extended to two space dimensions, and we suggest a high-order extension. Numerical experiments are once again provided, including real-world simulations.

Keywords

shallow-water equations, Manning friction, Godunov-type schemes, well-balanced schemes, moving steady states