TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2018

46

Adrián Navas Montilla

Accurate simulation of shallow flows using arbitrary order ADER schemes and overcoming numerical shockwave anomalies

Departamento

Ciencia y Tecnología de Materiales y Fluidos

Director/es

Murillo Castarlenas, Javier

ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

ACCURATE SIMULATION OF SHALLOW FLOWS USING ARBITRARY ORDER ADER SCHEMES AND OVERCOMING NUMERICAL SHOCKWAVE ANOMALIES

Autor

Adrián Navas Montilla

Director/es

Murillo Castarlenas, Javier

UNIVERSIDAD DE ZARAGOZA

Ciencia y Tecnología de Materiales y Fluidos

2018

Universidad Zaragoza
1542

Tesis Doctoral

Accurate simulation of shallow flows Using arbitrary order ADER schemes and overcoming numerical shockwave anomalies

Autor

Adrián Navas Montilla

Director

Dr. Javier Murillo Castarlenas

Escuela de Ingeniería y Arquitectura

2018

ACCURATE SIMULATION OF **SHALLOW FLOWS**

Using arbitrary order ADER schemes and overcoming numerical shockwave anomalies

PhD thesis by

ADRIÁN NAVAS MONTILLA

Doctoral advisor · DR. JAVIER MURILLO CASTARLENAS

FEBRUARY 2018

PhD Programme in
Mechanical Engineering

Universidad Zaragoza
1542

Doctoral Thesis
Author · Adrián Navas Montilla
Supervisor · Dr. Javier Murillo Castarlenas

ACCURATE SIMULATION OF **SHALLOW FLOWS**

Using arbitrary order ADER schemes and overcoming numerical shockwave anomalies

ADRIÁN NAVAS MONTILLA

SUPERVISOR · DR. JAVIER MURILLO CASTARLENAS

AGRADECIMIENTOS

Quiero dedicar unas líneas para dotar de carácter humano este documento exclusivamente enfocado desde un punto de vista científico, ya que, en definitiva, es fruto de un trabajo personal en un entorno social particular. Creo necesario incluir un breve texto en el que mostrar mi gratitud a aquellas personas e instituciones que han contribuido de forma personal y profesional en el desarrollo de este trabajo.

Me gustaría comenzar reconociendo y agradeciendo profundamente la importante labor de Javier Murillo como Director de esta Tesis Doctoral. Queda ya muy atrás ese año en el que aparecí por su despacho preguntando por un Proyecto Final de Carrera pero su dedicación y seguimiento del trabajo siempre ha sido excelente durante todos estos años. Es ahora momento de expresar mi gratitud por el apoyo y motivación que me ha ofrecido, y más allá de todo ello, por su amistad.

Agradecer también a Pilar García por depositar su confianza en mí e introducirme profesionalmente en las actividades de investigación del Grupo de Hidráulica Computacional de la Universidad de Zaragoza. Asimismo, reflejar mi gratitud por ofrecerme la posibilidad de asistir a diferentes congresos y seminarios científicos así como a algunos cursos de formación en centros especializados.

Siguiendo en esta línea, reconocer el trabajo de algunas de las personas con las que coincidí en el Centro Europeo de Previsiones Meteorológicas a Plazo Medio (ECMWF) y en el Centro de Supercomputación de Julich (JSC) que me ofrecieron su ayuda de forma totalmente desinteresada.

También quiero expresar un especial agradecimiento a Carmelo Juez por diversos motivos. Por su ayuda, apoyo y consejo en incontables ocasiones, por mostrarme la "whole picture"de todo esto y por hacer posible mi estancia en el Laboratorio de Construcciones Hidráulicas (LCH) de la Escuela Politécnica Federal de Lausanne (EPFL). Relativo a esta estancia, agradecer igualmente a Mário Franca, Anton Schleiss y resto de miembros del LCH por acogerme en su laboratorio.

Me gustaría dedicar unas líneas para mostrar mi agradecimiento a mis compañeros del Área de Mecánica de Fluidos de la Universidad de Zaragoza. Dar las gracias a Diego por su desinteresada ayuda y atención, y en especial, por dedicar largas horas en ayudarme con los métodos Discontinuous Galerkin. También quiero agradecer a Mario y a Javi por su apoyo científico-técnico durante estos cuatro años y por nunca negar una contestación a mis preguntas al aire casi diarias. Dar las gracias a Isabel por ofrecerme siempre un aporte muy positivo, constructivo y entusiasta de las cosas, así como por el apoyo que me ha brindado en numerosas ocasiones. Asimismo, mostrar mi agradecimiento a aquellos que pasaron por aquí y de los que también aprendí, como Cifu, Asier y Daniel. Y agradecer también a los recién llegados, Geovanny y Sergio, por su buena actitud y por generar una muy buena atmósfera de trabajo. No me olvido de Víctor, que aunque todavía no le hemos convencido para compartir espacio físico con nosotros, siempre ha mostrado muy buena disposición para ayudarme con las múltiples consultas con las que le he abordado. Nombrar

también a Antonio Pascau y José Luis Gracia, quienes siempre han mostrado interés por mi trabajo y han sido de ayuda en diversas cuestiones.

Agradecer también a Valerio Caleffi, de la Universidad de Ferrara, y a Carmelo Juez, arriba citado, por su evaluación crítica y constructiva de este trabajo.

Por otro lado, me gustaría dar las gracias a Paula por su ayuda y asesoramiento técnico y también a Sara por sus valiosos consejos que, desde la experiencia, me han allanado el camino.

Finalmente, dar las gracias a mis padres y a Irene por su paciencia, apoyo y entendimiento durante estos años, lo que ha hecho posible la consecución de esta Tesis Doctoral.

RESUMEN

En la actualidad, gracias al desarrollo de algoritmos de simulación avanzados y de tecnologías computacionales eficientes que ha tenido lugar durante las últimas décadas, es posible simular problemas de elevada complejidad que hace unos años eran inalcanzables. Parte de estos problemas se modelan mediante ecuaciones en derivadas parciales de tipo hiperbólico. Este tipo de ecuaciones reproducen con fidelidad aquellos fenómenos que involucran la propagación de ondas. En situaciones realistas, es necesario tener en cuenta efectos dinámicos adicionales más allá de los fenómenos puramente convectivos. Dichos efectos se modelan matemáticamente mediante los llamados términos fuente, que dan lugar a sistemas de ecuaciones no homogéneos y suponen un desafío computacional importante en numerosas ocasiones. Sólo unas determinadas discretizaciones del término fuente garantizan la convergencia de la solución a una solución físicamente realista; cuando se utilizan métodos numéricos sofisticados, la complejidad en el tratamiento de los términos fuente aumenta de forma notable.

Esta tesis se centra en el desarrollo de esquemas numéricos de orden arbitrario para la resolución de sistemas hiperbólicos siguiendo la metodología ADER, que permite la extensión del esquema tradicional de Godunov a orden arbitrario. Los métodos que aquí se presentan están enfocados a la resolución de las ecuaciones de aguas poco profundas, pero se formulan de forma general para su posible aplicación a otros modelos matemáticos. La particularidad fundamental de los esquemas numéricos propuestos en esta tesis reside en la manera en la que se introducen los términos fuente en la formulación discreta. A diferencia de la mayoría de métodos comunmente utilizados, aquí se propone introducir los términos fuente en la formulación de los flujos numéricos, siguiendo una metodología de discretización *upwind*. Esto implica considerar los términos fuente en la formulación del problema de Riemann derivativo. De este modo, es posible garantizar un equilibrio perfecto entre flujos y términos fuente a nivel discreto y reproducir con precisión aquellas situaciones de equilibrio relevantes para los problemas estudiados. Para las ecuaciones de aguas poco profundas, aquellos esquemas que satisfacen esta propiedad se denominaron tradicionalmente *well-balanced*, aunque dicha atribución sólo hacía referencia a la preservación de estados de reposo estático.

Se muestra que sólo aquellos términos fuentes de tipo geométrico (por ejemplo, término de variación de fondo en las ecuaciones de aguas poco profundas) se deben incluir en la resolución del problema de Riemann derivativo. Otros términos fuente de distinta naturaleza se pueden integrar de forma tradicional utilizando reglas de cuadratura, o bien, se pueden reescribir como términos geométricos y pueden ser tratados del mismo modo. Siguiendo esta última aproximación, es posible garantizar la propiedad *well-balanced* sin perder el orden de convergencia arbitrario. Aquí se detalla la construcción de esquemas numéricos de orden arbitrario para las ecuaciones de aguas poco profundas con términos fuente de fondo, fricción y Coriolis, que satisfacen la propiedad *well-balanced*. Además, mediante consideraciones de conservación de energía a nivel discreto, dicha propiedad se extiende para situaciones de equilibrio unidimensionales que

involucran velocidades no nulas, desde una perspectiva de un esquema ADER.

Por último, en este trabajo también se estudian anomalías numéricas que pueden aparecer en la resolución de las ecuaciones de aguas poco profundas. Dichas anomalías son intrínsecas al método de volúmenes finitos y pueden dar lugar a oscilaciones severas de la solución numérica. Siguiendo estudios previos sobre anomalías numéricas en las ecuaciones de Euler, se formula un marco teórico para el estudio de dichas anomalías en las ecuaciones de aguas poco profundas. Se muestra que la presencia de resaltos hidráulicos genera oscilaciones numéricas en el caudal y se propone una corrección del flujo que lo solventa.

ABSTRACT

The combination of modern supercomputers with cutting-edge simulation tools has brought the society the capacity to solve complex evolution problems of technological and scientific interest. Some of those problems are modelled by hyperbolic systems of conservation laws, which arise in a broad variety of fields where wave propagation phenomena are dominant. When considering realistic scenarios, we usually have to account for additional physical processes, apart from the convective transport, by means of adding extra terms in the equations. Such terms are called source terms and their presence supposes a big challenge when considering the numerical resolution of the equations. The discretization of source terms is not a trivial task when seeking convergence to the physical solution and extra difficulties appear when designing sophisticated numerical methods.

This thesis focuses on the study and development of arbitrary order finite volume (FV) schemes for the resolution of hyperbolic conservation laws in 1 and 2 spatial dimensions, following the ADER approach. The numerical developments presented in this work are fundamentally devised for the application to the shallow water equations (SWE), but are presented in a general form for a further extension to other systems of equations. The proposed methods are designed to account for the contribution of source terms in the resolution of the Derivative Riemann problem (DRP) at cell interfaces, following the so-called augmented solver approach. This allows to ensure an exact discrete balance between sources and fluxes and eventually preserve the steady states of relevance for the problems of interest. In the framework of the SWE, the numerical schemes satisfying such property are called well-balanced schemes, concept introduced for those schemes which preserve the still water at rest.

It is shown that only those sources of geometric nature must be included in the resolution of the DRP. Other source terms of different nature can be either rewritten in geometric form and treated similarly or integrated conventionally using traditional quadrature rules. The latter approach is simpler in terms of implementation and computational cost but does not allow to construct a well-balanced scheme. In this work, arbitrary order well-balanced schemes are constructed for the SWE with bed elevation, friction and Coriolis. Furthermore, the well-balanced property is extended to satisfy the conservation of energy and preserve moving water equilibria in 1D.

The study of numerical shockwave anomalies in the framework of the SWE is also considered in this thesis. Such anomalies are intrinsic to FV schemes and may ruin the solution. Following previous developments for the Euler equations, the slowly-moving shock problem for the SWE is studied and circumvented.

CONTENTS

1 INTRODUCTION

1.1 Motivation

There is a broad variety of physical phenomena that can be explained from the perspective of fluid mechanics and play an important role in many natural processes. In addition to the traditional disciplines, such as environmental, aeronautical or astrophysical sciences, the rocketing technological growth of the manufacturing industry is demanding the study and resolution of new problems that can be addressed through the same approach. The aforementioned problems are mathematically represented by systems of integral or partial differential equations (PDEs), which formally describe the fundamental laws for the conservation of certain physical quantities inside a volume of reference. Such equations are difficult to solve and, actually, no general analytical solution has been found nor it is likely to be found in a close future.

For more than two centuries since Euler and Bernoulli first formulated some of those equations, the scientific community has made a great effort to obtain analytical solutions for certain simplified problems. However, the resolution of realistic problems in complex geometries was an impossible task until the first half of the 20^{th} century, when a new approach to such problems was introduced. Thanks to the discretization of the problem and the application of a numerical method, which could be eventually automated in a computer, the original differential equations could be transformed in a set of algebraic equations and the initial discrete data could be easily advanced in space and time. This supposed the beginning of a new discipline, the Computational Fluid Dynamics (CFD).

The pioneering work of Richardson [1], Courant, Friedrichs, and Lewy [2], Southwell [3], von Neumann [4], Richtmyer, Lax [5] and Godunov [6] established the foundations for the development of modern CFD. Most of their work focused on the correct computation of discontinuities in the flow field, such as shock waves. This is closely linked to the resolution of the Riemann problem (RP), presumably the most studied problem in CFD. In the following decades, stimulated by the growing needs of aerospace industry and thanks to the simultaneous evolution of computers, CFD experimented a quick development. This gave rise to improved simulation tools of application in a broad variety of fields, from aerodynamics and gas dynamics to weather prediction and environmental sciences.

Nowadays, the combination of modern supercomputers with cutting-edge simulation tools has brought society the capacity to solve very complex problems of technological and scientific interest and, what is more important, the capacity of being predictive. Numerical simulation of environmental events, such as atmospheric, oceanic or surface water flows, has come to a point where trustworthy predictions can be done at a reasonable cost, providing a quite a realistic picture of the potential threats intrinsically linked to those events. Some of the aforestated phenomena can be modeled under a shallow water type assumption, which

considers that the vertical scale is much smaller than the horizontal ones and hence vertical accelerations are neglected. In absence of dissipative processes, the resulting equations are of hyperbolic type and are called shallow water equations (SWE).

Hyperbolic PDEs arise in a broad variety of fields where wave propagation phenomena are dominant, for instance, gas dynamics, acoustics and geophysics, and there is a branch of CFD that is exclusively focused on such problems. Historically, many of the fundamental ideas were developed in the framework of the compressible Euler equations for gas dynamics by the aerospace community. At present, these methods have been extended to other applications such as free surface flows modeled by the SWE. As Euler equations, the SWE are nonlinear and their solutions suffer from loss of regularity: solutions that are initially smooth may eventually become discontinuous in the form of shock waves. The presence of shocks, of other nonlinear wave structures and their interaction lead to most of the computational challenges that stimulate the development of novel numerical methods nowadays [7].

The introduction of Godunov's scheme [6], a conservative finite volume (FV) type method, in combination with the development of the so-called Riemann solvers, which are algorithms that provide the solution of the RP, supposed a major step in the simulation of hyperbolic flows. Since then, FV schemes have been strongly improved. In the last decades, the focus has been put on the generation of more accurate schemes by means of advanced mesh adaption/refinement techniques, increasing the degree of the interpolating polynomials, overcoming numerical shockwave anomalies and seeking accurate discretizations of source terms that sometimes govern the dynamics of the problems (e.g the bed elevation and friction in the SWE). With all this, the resulting schemes would be able to provide more accurate solutions in a shorter computational time, becoming more efficient. See for instance Figure 1.1, where the solution provided by a novel high resolution method is compared with that of a traditional first order method

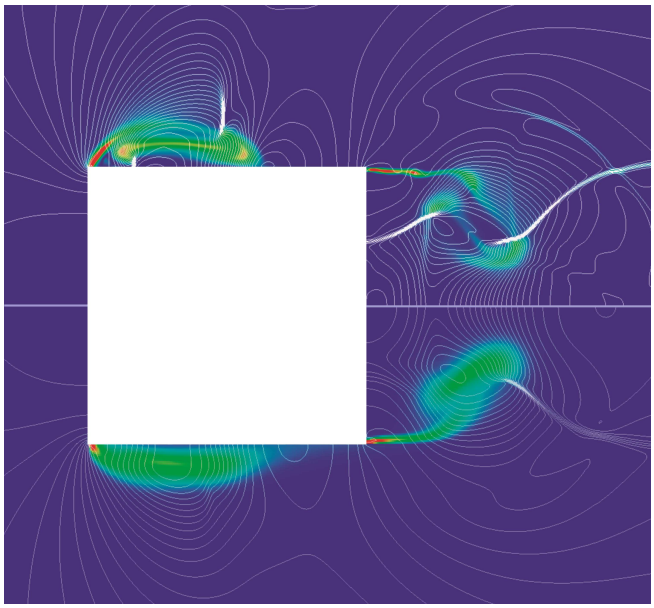


Figure 1.1: Computation of a subcritical water bore over varying bed hitting a square solid body. The solution is obtained using a high order scheme (upper half) and a first order scheme (lower half), in the same grid. Contour lines represent the water surface elevation and the color map the instantaneous vorticity.

Nowadays, the current trend is to design efficient schemes by means of a combination of: (a) very high order numerical schemes (e.g. WENO, ADER and DG schemes), which provide a higher resolution than lower order schemes using the same initial information and requiring shorter computational time and memory for the same accuracy, (b) adaptive mesh refinement techniques and other sophisticated meshes and (c) efficient parallel implementations for high performance computing, among others.

In the framework of the SWE, there is still room for improvement in the development of very high order

schemes that properly account for the contribution of source terms. The search of a suitable treatment of the source term in the numerical scheme is not a trivial task, but it is of utmost importance in order to preserve steady states of relevance. In the case of the SWE, such states are the still water at rest (quiescent equilibrium), the geostrophic equilibrium when considering rotation and the moving water steady state, when considering that dissipation is negligible (energy is conserved). Moreover, numerical shockwave anomalies in non-linear hyperbolic systems have been widely reported in the literature and circumvented for the Euler equations, but no cure has been presented for the moment for the SWE with source terms. In this thesis, the aforesaid issues are addressed and some novel methodologies are presented and evaluated.

1.2 State of the art

The state of the art herein presented has been divided in 4 subsections for the sake of clarity. The four topics are clearly interconnected but it is worth presenting them separately, in historical order. A more detailed state of the art is presented in some chapters when needed.

1.2.1 Early research on Godunov's schemes and Riemann solvers

Challenged by the complexity of the correct resolution of discontinuous flow fields, Courant, Friedrichs, and Lewy [2], Southwell [3] , von Neumann [4], Richtmyer, Lax [5] and Godunov [6] carried out a pioneering work on the development of accurate simulation techniques for such kind of problems. After some decades of intense research, conservative Godunov type FV schemes for the resolution of hyperbolic flows became more popular and experimented a fast development.

The importance of constructing conservative schemes was evidenced by Lax and Wendroff [8] and Hou [9], who showed that non-conservative schemes do not converge to the exact shock solution. In an attempt to obtain more accurate solutions, some authors focused their efforts on increasing the accuracy of Godunov's scheme to second order. Some of the first high-resolution schemes were introduced in the decade of the 70s and 80s and are still employed for many practical purposes. Such schemes are represented by the MUSCL schemes [10, 11], TVD schemes [12] and PPM schemes [13], which are second order accurate and resolve discontinuities in a sharper transition than first order schemes. They are often the best choice when seeking a good balance between computational cost and resolution for problems dominated by shocks with simple wave structures. Such schemes overcome the appearance of Gibbs oscillations [14] thanks to the utilization of different limiting techniques.

Simultaneously, an intense research on Riemann solvers was also taking place. Riemann solvers are the keystone of Godunov type methods, as they are used to provide accurate estimations of the numerical fluxes at cell interfaces. The presentation of the well-known Roe's approximate Riemann solver [15] supposed a major step in the construction of accurate and efficient Godunov-type schemes for hyperbolic problems with strong nonlinearities. It is based on upwinding the information using wave celerities. Other approximate solvers, such as the HLL solver [16] and the HLLC solver [17, 18], are also considered important contributions as they allow the construction of fast schemes for nonlinear systems of 2 and 3 equations respectively (the latter with at least one linearly degenerate field).

When dealing with geometric source terms, it was shown that the consideration of the source term in the resolution of the RP is more convenient in order to ensure certain properties of the numerical solution [19, 20]. Solvers accounting for the source term in the resolution of the RP are called augmented solvers [19] and were created to adequately characterize the influence of the source terms in the numerical solution [21, 22, 23].

1.2.2 Cutting-edge high order methods

In the last years, motivated by the recent technological development of computers, a genesis of new generations of high resolution schemes has occurred. In the framework of FV, the introduction of the ENO and WENO reconstruction techniques [24, 25, 26] supposed a major step when seeking arbitrary order of accuracy in space. On the other hand, the preservation of high order in time was generally done by means of a Runge-Kutta time integrators, which required to take multiple temporal sub-steps and sometimes proved to be inefficient due to Butcher's barrier [27]. This issue was addressed when using the ADER approach, pioneered by Toro et al. [28, 29], which provides a fully discrete scheme (one-step temporal update) of arbitrary order in space and time. ADER schemes consist of two steps: first, a high-order spatial reconstruction procedure and secondly, the resolution of a high order extension of the Riemann Problem (RP), called Derivative Riemann Problem (DRP) [30]. A broad variety of DRP solvers have been proposed up to date [31, 24, 32, 33, 34]. It is worth mentioning the work of Montecinos et al. [35], where a comparison of some of these solvers is presented.

Parallel to the development of WENO-ADER schemes, a rather different approach for the construction of arbitrary order schemes in space and time was also developed. This gave rise to a new family of schemes called Discontinuous Galerkin (DG) methods. Their origin is usually attributed to Reed and Hills in a paper published in 1973 on the numerical approximation of the neutron transport equation [36] . The fundamentals of such schemes can be found in a series of papers by Cockburn and Shu published in the 80s and 90s [37, 38]. After some years, the DG methods became more popular and were successfully applied to the resolution of evolution problems in continuum mechanics, among others. DG methods based on the ADER time integration challenge the performance of WENO-ADER schemes and combinations of both approaches are frequently used.

The ADER approach successfully allows the construction of arbitrary order schemes for systems of hyperbolic conservation laws [30, 39, 40, 27, 42, 43, 44, 45, 46] in Cartesian and triangular meshes. It is of particular interest in this thesis to consider the application of ADER schemes for the resolution of geophysical problems [47, 48, 49], specially for the resolution of the Shallow Water Equations (SWE) [31, 49, 50, 51]. Some alternatives for solving non-conservative complex physical models (mud flow, muptiphase flow), using a centered discretization, are also found [52, 53].

1.2.3 Application to the SWE

The application of all this technology to the resolution of surface flows, namely the SWE, has gained popularity over the last few decades [54, 55, 56]. Practical problems in shallow water flows include source terms that account for bottom variation, wind forces, Coriolis forces and many other physical effects. The numerical treatment of such source terms has been, presumably, the most studied issue in the framework of SWE. The idea of upwinding the source term was initially proposed by Roe [57] and consolidated by Bermudez and Vázquez-Cendón [58], allowing to obtain the correct effect of the source in the wave propagation. However, they, as well as Greenberg and Leroux [59], showed that upwinding the sources is not enough. A particular discretization of the source terms must be carried out in order to ensure an exact balance between fluxes and sources and converge to steady states or relevance.

For the SWE with bottom elevation, the very first property that a numerical scheme must satisfy is the preservation of the quiescent equilibrium. Such schemes are called well-balanced methods [58, 59]. There is a large variety of well-balanced methods based on Riemann solvers that ensure the preservation of the still water steady state [60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73]. See [74] for an exhaustive study on the treatment of bed steps, both in the framework of classical finite volume and path conservative approaches, when constructing well-balanced Discontinuous–Galerkin schemes.

The well-balanced property can still be enhanced. If neglecting friction in the SWE, mechanical energy is conserved under steady conditions in absence of hydraulic jumps. Such idea of energy conservation can be integrated in the numerical scheme, at the discrete level, allowing the extension of well-balanced

methods to exactly well-balanced methods, also called energy balanced (EB) methods [75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85]. For a complete description of 1D and 2D augmented solvers for the resolution of the SWE with bottom and friction, see [86].

As mentioned before, apart from bed elevation and friction, Coriolis forces are worth being included in the model when considering large scale flows. The SWE in the rotating frame represents a good model for large scale phenomena in geophysical flows, in which oceanic and atmospheric circulations are often perturbations of the so-called geostrophic equilibrium [87, 88]. In the last decade, a great effort has been put on the design of FV well-balanced numerical schemes capable to maintain the geostrophic equilibrium, some of high order of accuracy [89, 90, 91, 92, 93, 94, 87, 88] but not many of arbitrary order.

At present, the state-of-the-art challenges also include the derivation of more complex physical models that include morphodynamic effects [95, 96, 97] , steep slopes [98], multiphase and multilayer flows, granular and mud flows [52], among others.

1.2.4 Numerical shockwave anomalies

It has been widely reported in the literature that significant numerical anomalies arise in presence of shock waves. An example of such problems are the Carbuncle [99, 100], the slowly-moving shock [101, 102] and the wall-heating phenomenon [103], all of them leading to spurious numerical solutions, that are more visible when high order schemes are used.

Some of the problems related to numerical shockwave anomalies were first identified by Cameron and Emery [104, 105], who proposed some improvements based on the addition of artificial viscosity and modification of the grid. The slowly-moving shock problem was first investigated by Roberts in [101], who defined it as numerical noise generated in the discrete shock transition layer which is transported downstream. Such noise will be hereafter referred to as post-shock oscillations. In [101], the schemes of Godunov, Roe, and Osher were examined. Later on, Arora and Roe [102] carried out a thorough study on this problem and evidenced that it can be ruinous when, for instance, making calculations of shock-sound interaction. Far from being helpful, some authors showed that high order schemes accentuate this problem [106, 107, 108, 109, 110].

When designing numerical schemes for the computation of slowly-moving shocks, the addition of extra artificial viscosity seems to be the most preferred technique in the scientific community [104, 105, 101, 102, 107, 111, 112]. If we want to avoid extra diffusion, another possibility is the use of a flux interpolation method, which avoids using the evaluation of the physical fluxes in the untrustworthy intermediate cells corresponding to the shock discontinuity. This idea of flux interpolation was first presented by Zaide and Roe [113]. The authors claim that, by enforcing a linear shock structure and unambiguous sub-cell shock position, numerical shockwave anomalies are dramatically reduced.

Up to the present time, most studies have been carried out in the framework of Euler equations, but the growing needs for the computation of complex geophysical flows with a morphodynamical component motivate their application to the SWE.

1.3 Highlights and novel contributions

This thesis is devoted to the study and development of high resolution and efficient FV schemes with application to hyperbolic conservation laws with source terms, particularly the SWE. Numerical schemes are studied here with a focus on the resolution of the non-homogeneous (D)RP and the design of suitable Riemann solvers that properly account for the source term in the solution. This is important to ensure a suitable balance between sources and fluxes and converge to physically based solutions.

The work herein described is based on the so-called augmented solvers, that include the contribution of the source term as an extra wave of zero velocity. Numerical methods presented in this thesis will

be extensions/combinations of the ARoe and HLLS augmented solvers. To construct efficient high order schemes, the WENO-ADER approach is used, which allows to generate arbitrary order extensions of the aforementioned solvers. The novel solvers presented here are based on a new family of augmented DRP solvers, called AR(L) solvers. Unlike other solvers found in the literature, here we propose to solve the DRP by including the source term not only implicitly inside time derivatives via the CK procedure but also at cell interfaces, as the problem is non-homogeneous. Source terms can be of diverse nature. A study on the different discretization techniques for geometric and non-geometric source terms in the framework of WENO-ADER schemes is carried out.

The proposed schemes are extended to 2 space dimensions using a dimension-by-dimension approach, which is based on essentially 1D algorithms. Such approach is useful and efficient when using Cartesian grids. The WENO reconstruction is carried out by sweeping in the x and y directions sequentially and the numerical solvers are based on a 1D projection of the equations and variables onto the cell's normals. Inside cells, the integration of the source terms is carried out in the same way by means of a combination of Gaussian and Romberg integration. The latter allows to extend a particular second order quadrature rule to arbitrary order. Such quadrature rule can be designed to satisfy certain properties that ensure an exact balance between sources and fluxes and allows to preserve steady states of relevance.

The application to the SWE is the main focus of the work herein described. A broad variety of source terms can be considered, but we restrict to bed elevation, friction and Coriolis, which are of different nature. This set of sources is sufficient to exercise the different approaches proposed for the discretization of geometric and non-geometric sources. When considering 1D cases without dissipation, the discrete level of energy can be easily preserved. An EB WENO-ADER method using the AR(L) solver in combination with the ARoe and HLLS solvers is presented. On the other hand, when considering 2D cases, only the well-balanced property will be sought. A 2D WENO-ADER scheme based on the ARL solver is presented. This scheme ensures the well-balanced property for quiescent equilibrium as well as moving equilibrium under rotation.

The last part of the thesis is devoted to the study of numerical shockwave anomalies in the SWE, particularly the slowly-moving shock problem. Such problem is observed in presence of hydraulic jumps and produces a shedding of spurious oscillations that are only damped by the numerical diffusion of the scheme. A novel spike-reducing solver, based on previous work for Euler equations, is presented and applied to 1D and 2D cases. A theoretical study on the origin of the slowly-moving shock problem in the SWE is also included.

The application to other wave propagation problems, such as the linear scalar advection reaction equation and the linear acoustic equations is also considered in this thesis. The latter is used as a benchmark for comparing WENO-ADER schemes with DG-ADER schemes.

The thesis is divided in 3 main blocks. The first one is devoted to the development of Riemann solvers for the RP and DRP as well as to the construction of 1D schemes with application to the 1D SWE. The second block focuses on the extension of such methods to 2 space dimensions and their application to linear/nonlinear 2D problems. Finally, the third block is devoted to the study of numerical shockwave anomalies in the SWE. A more detailed structure of the thesis, including the relevant contributions, is displayed below:

- Block I: In Chapter 2 and 3, a brief introduction to hyperbolic conservation laws and FV schemes is presented. The mathematical tools for the analysis of such problems and the design of FV schemes are therein provided. The definition of the RP, the cornerstone of the methods herein presented, is also provided. In Chapter 4, traditional augmented Riemann solvers for scalar problems as well as systems of equations are presented. Such methods comprise the ARoe and HLLS solvers. The aim of Chapter 5 is to provide the fundamentals for the construction of WENO-ADER schemes, including the definition of the DRP. Chapter 6 is devoted to the development of Riemann solvers for the DRP. The (L)FS solver is proposed in this chapter and the resulting method when combining with the ARoe and HLLS solvers is presented. The main contributions are:

  - Study of 1D and 2D WENO reconstruction and sub-cell derivatives reconstruction techniques. Examination of different methods to avoid loss of accuracy around critical points.
  - Generation of 1D WENO-ADER schemes based on augmented Riemann solvers. A novel philosophy for the resolution of the DRP including the contribution of the source term is proposed: the (L)FS family of solvers. Two new solvers are proposed: the AR-(L)FS and HLLS-(L)FS solvers [83, 84].
  - Investigation on EB source term discretizations for the 1D SWE. Introduction of the SEBF source term discretization and generation of 1D EB ADER schemes for the SWE: the EB AR(L)-ADER and HLLS(L)-ADER methods [84, 85].

- Block II: In Chapter 7, some of the methods proposed in previous chapters are extended to 2 space dimensions. The DG-ADER scheme is also presented in this chapter. The next 4 chapters are devoted to the application of the schemes to different hyperbolic problems. In Chapter 8, we show the application of the WENO-ADER and DG-ADER scheme to linear problems, such as the advection-reaction equation and the acoustic system of equations. A comparison between methods is presented. In Chapters 9, 10 and 11, the WENO-ADER schemes using the novel solvers are applied to the resolution of the SWE. The main contributions are:
  - Generation of 2D WENO-ADER schemes (in Cartesian meshes) based on augmented Riemann solvers. A combination of Gaussian and Romberg integration is used for the integral of the source term inside cells.
  - Generation of a well-balanced WENO-ADER scheme for the 2D SWE with bed elevation, friction and Coriolis. Application to steady and transient flows.

- Block III: Chapter 12 is devoted to the study of the numerical shockwave anomalies in the SWE. Some numerical solvers that circumvent such anomalies are presented and applied to the 1D and 2D SWE. The main contributions are:
  - Investigation on numerical shockwave anomalies. A qualitative and quantitative study of the slowly-moving shock in the framework of the SWE is presented.
  - Generation of a spike-reducing solver in 1D and 2D (Cartesian mesh) for the SWE with bed elevation. Application to the resolution of steady and moving hydraulic jumps over complex bed geometries [85].

2 HYPERBOLIC CONSERVATION LAWS IN FLUID MECHANICS

In this chapter, we present the physical motivation for the mathematical problems of interest considered in this thesis. Conservation laws in the framework of fluid mechanics are presented both from the physical and mathematical point of view. The general formulation of hyperbolic systems of equations is derived from the concept of conservation law by means of the Reynolds Transport Theorem. Then, the mathematical tools for the analysis of such systems are presented. The wave nature of the solutions in hyperbolic conservation laws is detailed and the possible existence of discontinuous solutions is also explained. In the end, the concept of weak solution is introduced as a passage to Chapter 2 where the Finite Volume method is detailed.

2.1 The Reynolds Transport Theorem and conservation laws

A wide variety of physical events are described by systems of partial differential equations (PDEs) that correspond to conservation laws. In fluid mechanics, these conservation laws are commonly stated for extensive (integral) quantities such as mass, momentum and energy, among others. They result naturally from the application of the fundamental laws for the conservation of such quantities inside a *closed system*, hereafter referred to as *fluid volume* (V_f). This fluid volume is defined such that its boundaries move at the same velocity than the flow, therefore there is no relative velocity between them.

For instance, let us consider a fluid with density $\rho = \rho(x, y, z, t)$ inside a volume $V_f \subseteq \Omega \in \mathbb{R}^3$, then the equation for the conservation of mass inside V_f can be stated as follows

$$\left.\frac{dm}{dt}\right|_{V_f} = 0 \quad \Leftrightarrow \quad \frac{d}{dt}\iiint_{V_f} \rho dV = 0. \tag{2.1}$$

which evidences that the mass of the moving fluid parcel, that is, the fluid volume, V_f, is constant in time.

More generally, let us define $\mathbf{M}$ as a d-vector of any extensive property of the fluid (energy, mass, momentum...) and let $\mathbf{U} = d\mathbf{M}/dV$ be the intensive value of $\mathbf{M}$ per unit volume. Then, we can express the conservation of $\mathbf{M}$ in the fluid volume as

$$\left.\frac{d\mathbf{M}}{dt}\right|_{V_f} = 0 \quad \Leftrightarrow \quad \frac{d}{dt}\iiint_{V_f} \mathbf{U} dV = 0. \tag{2.2}$$

As mentioned above, fundamental physical laws have to be stated inside the fluid volume, that is, in the closed system, in order to ensure conservation. However, the integration of the variables and equations

in the fluid volume can become very difficult or even impossible. Then, the definition of a new integration volume is necessary. Such volume, hereafter referred to as control volume (CV), can be defined in a way that it fits the geometry of the problem. Whereas the integration in the CV is much simpler than in the fluid volume, it is not possible to state the conservation of extensive quantities in the CV as done in the fluid volume (2.2). This is because there is a relative velocity between the boundaries of the CV, hereafter referred to as control surface (CS), and the flow. Therefore, the flux across the CS must be accounted for in order to ensure conservation.

It seems necessary to find a way to relate variations inside the CV to variations inside the fluid volume. For this purpose, the Reynolds Transport Theorem, hereafter RTT, was introduced, allowing to express the variation of an extensive quantity inside the fluid volume as the variation of such quantity in a certain CV plus the flux of its associated intensive property across the CS. The utilization of this theorem supposes a great advantage since all calculations can be done over the CV, while ensuring the conservation of the physical quantity inside the fluid volume. The RTT is expressed as

$$\frac{d}{dt}\mathbf{M}_{V_f}(t) = \frac{d}{dt}\iiint_{CV}\mathbf{U}dV + \iint_{CS}\mathbf{U}(\mathbf{v}-\mathbf{v}_s)\cdot\hat{\mathbf{n}}dS \tag{2.3}$$

where the term on the left hand side of the equation stands for the total variation of quantity $\mathbf{M}$ inside the fluid volume, V_f, that must be either nil when the quantity is conserved or equal to a certain source. When existing, the sources will be considered acting on the CV.

It is remarkable to show that the RTT inside the fluid volume is given by Leibniz's rule for differentiation under the integral sign, that reads

$$\frac{d}{dt}\mathbf{M}_{V_f}(t) = \iiint_{V_f(t)}\frac{\partial\mathbf{U}}{\partial t}dV + \iint_{S_f(t)}\mathbf{U}(\mathbf{v}\cdot\hat{\mathbf{n}})dS. \tag{2.4}$$

2.2 Conservation laws: general formulation and hyperbolicity

The derivation of the differential formulation for a system of conservation laws is straightforward departing from the RTT by assuming an integration volume of infinitesimal size. Conservation laws described in the previous section can be expressed in their divergence form as

$$\frac{\partial\mathbf{U}}{\partial t} + \nabla\cdot\mathbf{E}(\mathbf{U}) = \mathbf{S}, \tag{2.5}$$

where $\mathbf{U} = \mathbf{U}(\mathbf{x},t) \in \mathscr{C} \subset \mathbb{R}^{N_\lambda}$ is the vector of conserved quantities that takes values on $\mathscr{C}$, the set of admissible states of $\mathbf{U}$, $\mathbf{E}(\mathbf{U}) : \mathscr{C} \longrightarrow \mathbb{R}^{N_\lambda \times d}$ is the matrix of fluxes, a nonlinear mapping of the conserved variables given by the physical flux and $\mathbf{S}$ is the vector of sources, yet to be defined. Normally, this vector of sources is of the form $\mathbf{S} = \mathbf{S}(\mathbf{U},\mathbf{x})$.

System in (2.5) can also be expressed as

$$\frac{\partial\mathbf{U}}{\partial t} + \sum_{j=1}^{d}\frac{\partial\mathbf{E}_j(\mathbf{U})}{\partial x_j} = \mathbf{S}, \tag{2.6}$$

where $\mathbf{E}_j(\mathbf{U})$ represents the flux in the i-th spatial direction. It is possible to apply the *chain rule* to derivatives in (2.6) yielding

$$\frac{\partial\mathbf{U}}{\partial t} + \sum_{j=1}^{d}\mathbf{J}_j(\mathbf{U})\frac{\partial\mathbf{U}}{\partial x_j} = \mathbf{S}, \tag{2.7}$$

with $\mathbf{J}_j(\mathbf{U})$ the Jacobian matrix of $\mathbf{E}_j(\mathbf{U})$, defined as

$$\mathbf{J}_j(\mathbf{U}) = \frac{\partial \mathbf{E}_j(\mathbf{U})}{\partial \mathbf{U}}. \tag{2.8}$$

Definition 1 *(Hyperbolic system). The system in (2.5) is said to be hyperbolic if the matrix* $\mathscr{J}(\mathbf{k}) \in \mathbb{R}^{N_\lambda \times N_\lambda}$ *defined as*

$$\mathscr{J}(\mathbf{k}) = \sum_{j=1}^{d} k_j \mathbf{J}_j(\mathbf{U}), \tag{2.9}$$

is diagonalizable with real eigenvalues for all $\mathbf{k} \in \mathbb{R}^d$ *and for all* $\mathbf{U} \in C$ *with* $C \subseteq \mathbb{R}^{N_\lambda}$ *the subset of physically relevant values of* $\mathbf{U}$*. If the* N_λ *eigenvalues are distinct, then the system is said to be strictly hyperbolic [114].*

Definition 2 *(Eliptic and parabolic systems). The system in (2.5) is said to be eliptic if none of the eigenvectors of* $\mathscr{J}(\mathbf{k}) \in \mathbb{R}^{N_\lambda \times N_\lambda}$ *is real. It is said to be parabolic if all eigenvectors are real and identical.*

In this work, problems in 2 space dimensions where $\mathbf{E}(\mathbf{U}) : \mathscr{C} \longrightarrow \mathbb{R}^{N_\lambda \times 2}$ are considered. The matrix of fluxes will be hereafter referred to as $\mathbf{E} = (\mathbf{F}, \mathbf{G})$, where $\mathbf{F} = \mathbf{F}(\mathbf{U}) : \mathscr{C} \longrightarrow \mathbb{R}^{N_\lambda}$ and $\mathbf{G} = \mathbf{G}(\mathbf{U}) : \mathscr{C} \longrightarrow \mathbb{R}^{N_\lambda}$ are the physical fluxes on the coordinate directions x and y. Note that $\mathbf{x} = (x, y)$.

It is possible to define two Jacobian matrices for the fluxes $\mathbf{F}(\mathbf{U})$ and $\mathbf{G}(\mathbf{U})$ as

$$\mathbf{A}(\mathbf{U}) = \frac{\partial \mathbf{F}(\mathbf{U})}{\partial \mathbf{U}}, \qquad \mathbf{B}(\mathbf{U}) = \frac{\partial \mathbf{G}(\mathbf{U})}{\partial \mathbf{U}}, \tag{2.10}$$

that provide sufficient information about the hyperbolicity of the system. According to Definition 1, the system is said to be *hyperbolic* if the matrix

$$\mathscr{J} = k_1 \mathbf{A}(\mathbf{U}) + k_2 \mathbf{B}(\mathbf{U}), \tag{2.11}$$

is diagonalizable with real eigenvalues $\forall \mathbf{k} \in \mathbb{R}^2$.

It is worth pointing out that the 2D system can be converted into a 1D system by projecting the flux in any direction of the space $\hat{\mathbf{n}} = (n_x, n_y)^T$ as

$$\mathscr{F}(\mathbf{U}) = n_x \mathbf{F}(\mathbf{U}) + n_y \mathbf{G}(\mathbf{U}), \tag{2.12}$$

leading to

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathscr{F}(\mathbf{U})}{\partial \check{x}} = \mathbf{S}, \tag{2.13}$$

where $\check{x}$ is the spatial coordinate in the direction of $\hat{\mathbf{n}}$. Analogously, for the Jacobian of $\mathscr{F}(\mathbf{U})$ we have that

$$\mathscr{J}(\mathbf{U}) = n_x \mathbf{A}(\mathbf{U}) + n_x \mathbf{B}(\mathbf{U}). \tag{2.14}$$

Comparing (2.11) and (2.14) is noticed that $\mathbf{k} = \hat{\mathbf{n}}$, hence we can state that if the 2D system of equations is hyperbolic, the projection of the system in any direction of the (x, y) plane will lead to a 1D hyperbolic system of equations.

2.2.1 Integral form of conservation laws

The integral form of (2.5) inside a discrete space-time domain is the keystone for the construction of FV schemes. For the derivation of the integral form of (2.5), it is sufficient to integrate the equation in the

domain $\mathscr{Q} = \Omega \times [0, \Delta t]$, with $\Omega \subseteq \mathbb{R}^d$ and $\mathbf{x} \in \Omega$, as

$$\int_0^{\Delta t} \int_\Omega \left(\frac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathbf{E}(\mathbf{U}) \right) dV dt = \int_0^{\Delta t} \int_\Omega \mathbf{S} dV dt \tag{2.15}$$

and applying Gauss-Ostrogradsky theorem, the following expression results

$$\int_\Omega \mathbf{U}(\mathbf{x}, \Delta t) dV = \int_\Omega \mathbf{U}(\mathbf{x}, 0) dV - \int_0^{\Delta t} \int_{\partial\Omega} \mathbf{E}(\mathbf{U}) \hat{\mathbf{n}} dS dt + \int_0^{\Delta t} \int_\Omega \mathbf{S}(\mathbf{U}, \mathbf{x}) dV dt \tag{2.16}$$

that represents that the integral of the conserved quantities at $t = \Delta t$ is equal to the integral of the conserved quantities at $t = 0$ minus the integral in time of the total leaving fluxes across the surface $\partial\Omega$, plus the contribution of the source terms.

2.3 Conservation laws in 1D

The methods used in this work sometimes involve the projection of the 2D original system of equations in certain directions of space, reducing it into a 1D system of equations. Hence, the analysis of 1D systems is of utmost importance and must be studied prior to the development of the numerical techniques. Nonlinear systems of conservation laws in 1D can be expressed as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{S}, \tag{2.17}$$

where $\mathbf{U} = \mathbf{U}(x, t) \in \mathscr{C} \subset \mathbb{R}^{N_\lambda}$ is the vector of conserved variables with $x \in \Omega \subseteq \mathbb{R}$, $\mathbf{F}(\mathbf{U}) : \mathscr{C} \longrightarrow \mathbb{R}^{N_\lambda}$ is the vector of fluxes and $\mathbf{S}$ the vector of sources.

It is possible to define a Jacobian matrix for the flux $\mathbf{F}(\mathbf{U})$ as

$$\mathbf{J}(\mathbf{U}) = \frac{\partial \mathbf{F}(\mathbf{U})}{\partial \mathbf{U}} \tag{2.18}$$

that provides sufficient information for the hyperbolicity of (2.17) according to Definition 1. Making use of the chain rule, system in (2.17) is rewritten as

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{J}(\mathbf{U}) \frac{\partial \mathbf{U}}{\partial x} = \mathbf{S}. \tag{2.19}$$

In the case when $\mathbf{F} = \mathbf{F}(\mathbf{U}, x)$, the previous approach must be rewritten as

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{J}(\mathbf{U}) \frac{\partial \mathbf{U}}{\partial x} + \frac{\delta \mathbf{F}(\mathbf{U}, x)}{\delta x} = \mathbf{S}. \tag{2.20}$$

Assuming that the system is hyperbolic with N_λ real eigenvalues

$$\lambda^1(\mathbf{U}) \leq \lambda^2(\mathbf{U}) \leq \ldots \leq \lambda^{N_\lambda}(\mathbf{U}) \tag{2.21}$$

and N_λ linearly independent right eigenvectors

$$\mathbf{e}^1(\mathbf{U}), \mathbf{e}^2(\mathbf{U}), \ldots, \mathbf{e}^{N_\lambda}(\mathbf{U}), \tag{2.22}$$

it is possible define two matrices $\mathbf{P}(\mathbf{U}) = (\mathbf{e}^1(\mathbf{U}), \mathbf{e}^2(\mathbf{U}), \ldots, \mathbf{e}^{N_\lambda}(\mathbf{U}))$ and $\mathbf{P}^{-1}(\mathbf{U})$ with the property that they diagonalize the Jacobian $\mathbf{J}$ as

$$\mathbf{J}(\mathbf{U}) = \mathbf{P}(\mathbf{U})\mathbf{\Lambda}(\mathbf{U})\mathbf{P}^{-1}(\mathbf{U}), \tag{2.23}$$

with $\mathbf{\Lambda}(\mathbf{U}) = \text{diag}\left(\lambda^1(\mathbf{U}), ..., \lambda^{N_\lambda}(\mathbf{U})\right)$ a diagonal matrix composed by the eigenvalues of the Jacobian.

2.3.1 Conservative vs non-conservative form

For the sake of simplicity, dependency of variables upon the conserved quantities is hereafter omitted. A generic conservative hyperbolic system is written as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} = 0, \tag{2.24}$$

where $\mathbf{U}$ is the vector of conserved quantities and $\mathbf{F}$ the vector of conservative fluxes. It can be expressed in its quasilinear form as

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{J}\frac{\partial \mathbf{U}}{\partial x} = 0, \tag{2.25}$$

where the Jacobian matrix $\mathbf{J} = d\mathbf{F}/d\mathbf{U}$ can be diagonalized with N_λ eigenvalues by means of N_λ linearly independent eigenvectors. The following relation is worth being shown

$$\mathbf{J}\cdot\mathbf{e}^m - \lambda^m\mathbf{e}^m = 0, \tag{2.26}$$

where λ^m and $\mathbf{e}^m$ are the eigenvalues and right eigenvectors of matrix $\mathbf{J}$.

Non-homogeneous hyperbolic conservation laws (2.17) cannot be expressed in an strict conservative form due to the presence of the source term. Instead, they can be expressed in non-conservative form as

$$\frac{\partial \hat{\mathbf{U}}}{\partial t} + \frac{\partial \hat{\mathbf{F}}(\hat{\mathbf{U}})}{\partial x} + \mathbf{H}\frac{\partial \hat{\mathbf{U}}}{\partial x} = 0, \tag{2.27}$$

where $\hat{\mathbf{U}} \in \mathscr{C} \subset \mathbb{R}^{N_\lambda+N_S}$ is the new vector of variables composed of the N_λ conserved variables in (2.17) plus additional N_S variables related to the source term, $\hat{\mathbf{F}}(\hat{\mathbf{U}}) : \mathscr{C} \longrightarrow \mathbb{R}^{N_\lambda+N_S}$ is the vector of conservative fluxes and $\mathbf{H}$ the matrix of non-conservative fluxes.

The non-conservative system in (2.27) can be more compactly expressed as

$$\frac{\partial \hat{\mathbf{U}}}{\partial t} + \mathscr{A}\frac{\partial \hat{\mathbf{U}}}{\partial x} = 0, \tag{2.28}$$

where $\mathscr{A} = \mathbf{J} + \mathbf{H}$ and with $\mathbf{J} = d\hat{\mathbf{F}}/d\hat{\mathbf{U}}$. Relation in (2.26) is now written as

$$\mathbf{J}\cdot\hat{\mathbf{e}}^m - \hat{\lambda}^m\hat{\mathbf{e}}^m = -\mathbf{H}\cdot\hat{\mathbf{e}}^m, \tag{2.29}$$

where $\hat{\lambda}^m$ and $\hat{\mathbf{e}}^m$ are the eigenvalues and right eigenvectors of matrix $\mathscr{A}$.

2.3.2 Linear conservation laws in 1D

When the Jacobian matrix in (2.18) does not depend either upon $\mathbf{U}$ or x, it will be constant and the system in (2.17) is said to be *linear*. In this case, the flux function can be expressed as

$$\mathbf{F}(\mathbf{U}) = \mathbf{J}\mathbf{U}, \tag{2.30}$$

leading to the following linear system of conservation laws

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{J}\frac{\partial \mathbf{U}}{\partial x} = \mathbf{S}, \tag{2.31}$$

where $\mathbf{J}$ is a matrix of constant coefficients. It is worth mentioning that the study of the properties of 1D linear systems is of utmost importance as the numerical solvers developed in this work for nonlinear systems of conservation laws are based in the linearization and projection of such systems in 1 space dimension.

Considering that the linear problem presented in (2.31) is hyperbolic, the diagonalization of the Jacobian matrix can be expressed as

$$\mathbf{J} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}, \tag{2.32}$$

where $\mathbf{P} = (\mathbf{e}^1, \mathbf{e}^2, ..., \mathbf{e}^{N_\lambda})$ and $\mathbf{\Lambda} = \text{diag}(\lambda^1, \lambda^2, ..., \lambda^{N_\lambda})$ are constant matrices composed of the eigenvectors of $\mathbf{J}$

$$\mathbf{e}^1,\ \mathbf{e}^2,\ \ldots,\ \mathbf{e}^{N_\lambda} \tag{2.33}$$

and the eigenvalues of $\mathbf{J}$

$$\lambda^1 \leq \lambda^2 \leq \ldots \leq \lambda^{N_\lambda} \tag{2.34}$$

respectively. In the case when the system in (2.31) is strictly hyperbolic, eigenvalues in (2.34) are all distinct.

Now, it is possible to define a new set of variables, denoted by $\mathbf{W} = (w^1, w^2, ..., w^{N_\lambda})$ and called *characteristic variables*, by means of the transformation

$$\mathbf{W} = \mathbf{P}^{-1}\mathbf{U} \tag{2.35}$$

that represent the projection of the conserved variables onto the Jacobian's eigenvectors basis. Considering that $\mathbf{P}$ is constant, the following relations are stated

$$\frac{\partial \mathbf{W}}{\partial t} = \mathbf{P}^{-1}\frac{\partial \mathbf{U}}{\partial t} \qquad \frac{\partial \mathbf{W}}{\partial x} = \mathbf{P}^{-1}\frac{\partial \mathbf{U}}{\partial x}. \tag{2.36}$$

Equivalently, a new set of variables, $\mathbf{B} = (\beta^1, \beta^2, ..., \beta^{N_\lambda})$, is defined for the source term as

$$\mathbf{B} = \mathbf{P}^{-1}\mathbf{S}, \tag{2.37}$$

ensuring the same relations presented for the derivatives of $\mathbf{W}$ and $\mathbf{U}$ in (2.36). From (2.32) and (2.36), it is possible to rewrite the initial system in (2.31) as a decoupled system of PDEs as

$$\frac{\partial \mathbf{W}}{\partial t} + \mathbf{\Lambda}\frac{\partial \mathbf{W}}{\partial x} = \mathbf{B} \tag{2.38}$$

that corresponds to the expression of the original system of PDEs on the Jacobian's eigenvectors basis.

System in (2.38) is composed of a set of independent linear scalar advection equations with source term, called *characteristic equations* and given by

$$\frac{\partial w^m}{\partial t} + \lambda^m \frac{\partial w^m}{\partial x} = \beta^m \quad \text{for } m = 1, ..., N_\lambda, \tag{2.39}$$

where λ^m is the eigenvalue associated to the m-th wave and represents its propagation velocity, called *characteristic speed*.

When the contribution of the source term is nil, the characteristic variables remain constant along the so-called characteristic lines, depicted in Figure 2.1 and defined as

Definition 3 *(Characteristic lines). Considering $w^m(x,t)$ a solution of (2.39), the curves $x = x(t)$ satisfying the initial value problem (IVP)*

$$\begin{cases} \dfrac{\partial x}{\partial t} = \lambda^m \\ \\ x(0) = x_0 \end{cases} \tag{2.40}$$

are called m-characteristic lines for the problem in (2.39).

When the source term is nil, we can rewrite (2.39) as

$$\frac{D}{Dt}(w^m) = 0 \quad \text{along } x = x_0 + \lambda^m t\,, \text{ for } m = 1, ..., N_\lambda, \tag{2.41}$$

where $\frac{D}{Dt}$ represents the material derivative operator, defined as

Definition 4 *(Material derivative operator). Operator*

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\,, \tag{2.42}$$

with ∇ the Del operator with respect to the spatial coordinates and $\mathbf{v}$ the velocity field, allows to calculate the total variation of a certain quantity as its variation in time plus its variation produced by its advection under the velocity field.

The solution for the original system in (2.31), $w^1(x,t), w^2(x,t), ..., w^m(x,t)$, can be obtained as a function of the solutions provided by the decoupled equations in (2.39). Regarding the previous results, the wave nature of the solution is noticed: the characteristic information will travel across the domain at different wave speeds given by λ^1, λ^2, ..., λ^m and the solution for the primitive variables will be obtained as a linear combination of the N_λ waves.

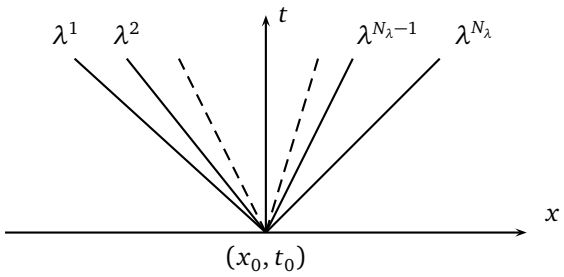


Figure 2.1: Characteristic lines passing through the point (x_0, t_0).

The initial condition for the decoupled system in Equation (2.39) is given by the projection of the initial condition $\mathring{\mathbf{U}} = \mathbf{U}(x,0)$ onto the Jacobian's eigenvectors basis, as

$$\mathring{\mathbf{W}} = \mathbf{P}^{-1}\mathring{\mathbf{U}}. \tag{2.43}$$

At a given point (x,t), it is possible to express the vector of primitive variables $\mathbf{U}(x,t)$ as a linear combination of the Jacobian's eigenvectors using the relation $\mathbf{U} = \mathbf{PW}$, as

$$\mathbf{U}(x,t) = \sum_{m=1}^{N_\lambda} w^m(x,t)\mathbf{e}^m\,, \tag{2.44}$$

where the scalar values $w^m(x,t)$ are the characteristic variables at the sought point and represent the strength of each wave.

When considering that $\mathbf{S} = 0$, characteristic equations in (2.39) are reduced to linear scalar transport equations. Therefore, the initial values for the characteristic variables $\mathring{w}^m(x,0)$ are simply advected at their corresponding wave speeds

$$w^m(x,t) = \mathring{w}^m(x-\lambda^m t) \quad \text{for } m = 1,...,N_\lambda\,, \tag{2.45}$$

with no change in shape. Then, the solution can be expressed as the superposition of the N_λ waves that have been advected independently, as

$$\mathbf{U}(x,t) = \sum_{m=1}^{N_\lambda} \mathring{w}^m(x-\lambda^m t)\mathbf{e}^m\,. \tag{2.46}$$

It is worth saying that numerical methods for the resolution of hyperbolic systems developed in this work are based on linear approximate solutions, being the previous results the foundations for such algorithms.

2.4 Integral curves and Riemann invariants

Let us consider the following hyperbolic system

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = 0 \tag{2.47}$$

where matrix $\mathbf{A} = d\mathbf{F}/d\mathbf{U}$ can be diagonalized with N_λ eigenvalues by means of N_λ linearly independent eigenvectors. Each eigenvalue $\lambda^m(\mathbf{U})$, or eigenvector $\mathbf{e}^m(\mathbf{U})$ equivalently, defines a *characteristic field* associated to it, for $m = 1,...,N_\lambda$. The properties of the characteristic fields will provide useful information about the solution.

Prior to the analysis of the characteristic fields, it is worth introducing the concept of state space. The state space, or phase plane, is the representation of a component of the state vector with respect to the other components. For instance, if considering a system of $N_\lambda = 2$ equations, with $\mathbf{U} = (u^1, u^2)$, the state space representation will be given by the representation of u^1, u^2 in a Cartesian coordinate system. In the state space representation, continuous elementary solutions for the system in (2.47) with piecewise constant initial data can be described by the so-called integral curves, which are next defined:

Definition 5 *(Integral curve). Let $\mathbf{U}(\xi)$ be a smooth curve through state space para- metrized by the scalar ξ. This curve is said to be an integral curve of the vector field $\mathbf{e}^m$ if at each point, the tangent vector to the curve, $d\mathbf{U}(\xi)/d\xi$ is an eigenvector of $\mathbf{J}(\mathbf{U}(\xi))$ corresponding to the eigenvalue $\lambda^m(\mathbf{U}(\xi))$. When considering a particular set of eigenvectors, the integral curve for $\mathbf{e}^m$ field is given by*

$$\frac{d\mathbf{U}(\xi)}{d\xi} = \nu(\xi)\cdot\mathbf{e}^m(\mathbf{U}(\xi))\,, \tag{2.48}$$

with $\nu(\xi)$ a constant parameter that depends on the normalization of the eigenvectors.

When analyzing the solution of hyperbolic systems of conservation laws, it is observed that the wave pattern present in the solution is related to the variation of the characteristic speed, $\lambda^m(\mathbf{U})$, along the

integral curve of the vector field $\mathbf{e}^m$. This variation can be expressed as the directional derivative of $\lambda^m(\mathbf{U})$ in the direction of the eigenvector

$$\frac{d}{d\xi}\lambda^m(\mathbf{U}(\xi)) = \nabla_u \lambda^m(\mathbf{U}(\xi)) \cdot \mathbf{e}^m(\mathbf{U}(\xi)). \tag{2.49}$$

When $\lambda^m(\mathbf{U})$ is constant along the integral curve, that is (2.49) is equal to zero, the characteristic field is said to be *linearly degenerate*. On the other hand, if $\lambda^m(\mathbf{U})$ varies along the integral curve, which means that the characteristic curves are compressing or expanding, the characteristic field is said to be *genuinely nonlinear*.

Definition 6 *(Linearly degenerate field). A λ^m-characteristic field is said to be linearly degenerate when*

$$\nabla_u \lambda^m(\mathbf{U}) \cdot \mathbf{e}^m(\mathbf{U}) = 0, \quad \forall \mathbf{U} \in \mathscr{C}, \tag{2.50}$$

with $\mathscr{C} \subseteq \mathbb{R}^{N_\lambda}$ and where ∇_u stands for the gradient with respect to the components of vector $\mathbf{U}$.

Definition 7 *(Genuinely nonlinear field). A λ^m-characteristic field is said to be genuinely nonlinear when*

$$\nabla_u \lambda^m(\mathbf{U}) \cdot \mathbf{e}^m(\mathbf{U}) \neq 0, \quad \forall \mathbf{U} \in \mathscr{C}, \tag{2.51}$$

with $\mathscr{C} \subseteq \mathbb{R}^{N_\lambda}$ and where ∇_u stands for the gradient with respect to the components of vector $\mathbf{U}$.

Definition 8 *(Riemann invariant). The scalar w^m is said to be a m-Riemann invariant when*

$$\nabla_u w^m(\mathbf{U}) \cdot \mathbf{e}^m(\mathbf{U}) = 0, \quad \forall \mathbf{U} \in \mathscr{C}, \tag{2.52}$$

with $\mathscr{C} \subseteq \mathbb{R}^{N_\lambda}$ and where ∇_u stands for the gradient with respect to the components of vector $\mathbf{U}$.

2.5 Loss of regularity and weak solutions

An important feature of nonlinear hyperbolic problems is the possible loss of regularity: solutions which are initially smooth may become discontinuous within a finite time. Therefore, it may not be possible to obtain a *classical solution* that satisfies the PDE (in the sense of an smooth and continuously differentiable solution) and only a *weak solution* may satisfy the equation. Weak solutions are not required to be continuous and differentiable, but they still satisfy the PDE in its integral form.

In order to illustrate the loss of regularity, let us consider the *Burgers' equation*, expressed as

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \tag{2.53}$$

where the flux is a convex function given by $f(u) = u^2/2$. Burgers' equation is a nonlinear equation that models a variety of processes, such as traffic flow or the dynamic of gases. It can be rewritten as

$$\frac{\partial u}{\partial t} + \lambda(u)\frac{\partial u}{\partial x} = 0, \tag{2.54}$$

where $\lambda(u) = u$ is the wave speed, defined as $\lambda(u) = \partial f(u)/\partial u$. It is noticed that (2.54) represents the total derivative of u, according to Definition 4. Hence, u is constant along the characteristic curves.

To find a classical solution for (2.54), it is sufficient to use the idea of the conservation of u along the characteristic lines. The information is translated along such lines in space and time and the solution is obtained. This technique is called the *method of characteristics*. For instance, let us consider Equation (2.54) and the following initial condition

$$u(x,0) = -0.8tanh(4x) + 1 \tag{2.55}$$

In Figure 2.2, the characteristic lines and solution for (2.54)–(2.55) are depicted. The characteristic lines are plotted every 0.05 spatial units. The plots on the left show a raw representation of the characteristic lines and the solution computed using the method of characteristics. It is observed that there is a region for $t > t^*$ where the characteristic lines cross. This means that there is a multi-valued solution within that region of space after t^*, as shown in Figure 2.2 (bottom left). As the solution cannot be multi-valued, a discontinuity forms due to the convergence of the characteristic lines and the solution becomes discontinuous after $t = t^*$. In Figure 2.2 (top right), the shock path is represented along the line where the characteristic lines cross. Under these conditions, classical solutions are not longer valid and only *weak solutions* may satisfy the equation for $t > t^*$. The weak solution that satisfies (2.54)–(2.55) is depicted in Figure 2.2 (bottom right), showing the discontinuous nature of such solution.

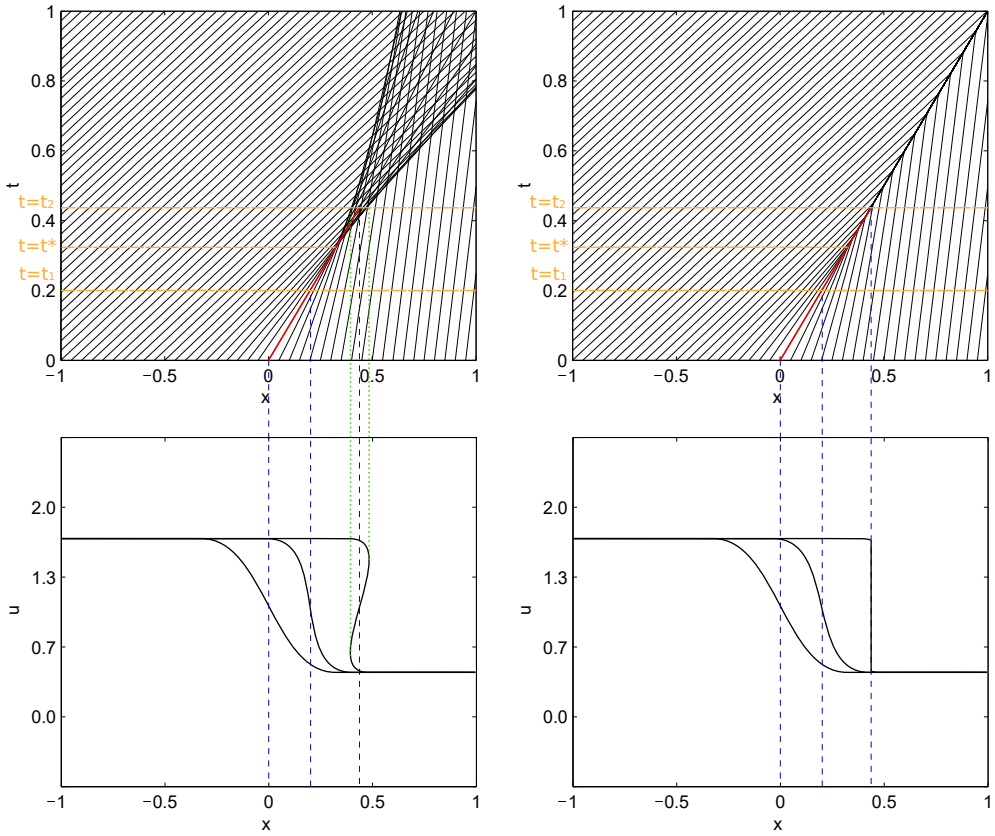


Figure 2.2: Characteristic lines (top) and solution (bottom) for the Burgers equation. The unphysical solution is depicted on the left and the physically feasible solution on the right.

Prior to the definition of the concept of weak solution, let us introduce some preliminary definitions. A subset of $\mathbb{R}^n$ is said to be *compact* if it is closed and bounded. A function $v : \mathbb{R}^n \longrightarrow \mathbb{R}$ is said to be function of compact support if it becomes zero outside of a compact set. Let us define the test function $v = v(\mathbf{x}, t) : \mathbb{R}^{d+1} \longrightarrow \mathbb{R}$ where d is the spatial dimension, $\mathbf{x} \in \mathbb{R}^d$ is the space vector and $t \in [0, \infty)$ is the time. Such function is smooth and $v \in \mathscr{C}_0^1$, where $\mathscr{C}_0^1 \equiv \mathscr{C}_0^1(\mathbb{R}^d \times [0, \infty))$ is the space of continuous functions with continuous first derivative that have compact support.

To introduce the concept of weak solution we will consider a homogeneous problem. Let us consider the left hand side of (2.5) and multiply it by the test function v as follows

$$\left[\frac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathbf{E}(\mathbf{U})\right] v = 0. \tag{2.56}$$

We integrate (2.56) in the domain $\mathscr{I} = \Omega \times [0, \infty)$ where $\Omega \subseteq \mathbb{R}^d$ is the spatial domain, yielding

$$\int_0^\infty \int_\Omega \left[\frac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathbf{E}(\mathbf{U})\right] v dV dt = 0 \tag{2.57}$$

and making use of the integration by parts, we obtain

$$-\int_0^\infty \int_\Omega \left[\mathbf{U}\frac{\partial v}{\partial t} + \mathbf{E}(\mathbf{U}) \cdot \nabla v\right] dV dt + \int_{\partial \mathscr{I}} v\left[\mathbf{U} n_t + \mathbf{E}(\mathbf{U}) \cdot \mathbf{n}\right] dS = 0\,, \tag{2.58}$$

where n_t is time component of the normal vector to $\partial \mathscr{I}$ and $\mathbf{n}$ is the space normal vector. For the sake of clarity, the complete normal vector to $\partial \mathscr{I}$ is written as $(n_t, \mathbf{n})^T$. $\partial \mathscr{I}$ is the boundary of $\mathscr{I}$.

As v is a function of compact support, we have that $v = 0$ on all the boundary hypersurface $\partial \mathscr{I}$ but at $t = 0$. Hence, Equation (2.58) becomes

$$\int_0^\infty \int_\Omega \left[\mathbf{U}\frac{\partial v}{\partial t} + \mathbf{E}(\mathbf{U}) \cdot \nabla v\right] dV dt + \int_\Omega v(\mathbf{x}, 0)\mathbf{U}(\mathbf{x}, 0) dV = 0 \tag{2.59}$$

Definition 9 *(Weak solution). The function* $\mathbf{U}(\mathbf{x}, 0)$ *is called a weak solution of*

$$\frac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathbf{E}(\mathbf{U}) = 0 \tag{2.60}$$

if the equality

$$\int_0^\infty \int_\Omega \left[\mathbf{U}\frac{\partial v}{\partial t} + \mathbf{E}(\mathbf{U}) \cdot \nabla v\right] dV dt + \int_\Omega v(\mathbf{x}, 0)\mathbf{U}(\mathbf{x}, 0) dV = 0 \tag{2.61}$$

holds for any $v \in \mathscr{C}^1$.

3 FINITE VOLUME NUMERICAL SCHEMES FOR HYPERBOLIC CONSERVATION LAWS

A common methodology for computing the solution of hyperbolic conservation laws is the discretization of the computational domain in volume cells where original PDEs can be integrated leading to an algebraic system of equations. Such equations represent the discrete balance of the quantities in the volume cells. Inside each cell, the conserved quantities are also integrated leading to a finite set of piecewise-constant (cell averaged) values that represent the approximate solution of the original system of PDEs. This approach is the so-called *Finite Volume (FV) method* [7, 115].

In this chapter, the general formulation of the FV method for the resolution of hyperbolic conservation laws is presented. Godunov's updating scheme is derived from the integral formulation of the original system of equations and the concept of numerical flux is introduced. Geometric source terms are presented here for the first time in the thesis and two possibilities for the numerical discretization of such terms in the updating scheme (centered and upwind integration) are also explained. In the end, the RP is presented both in strong and weak form, including the important integral relations, namely the Rankine-Hugoniot (RH) conditions, and the fundamental wave structures that can appear in the solution.

3.1 Introduction to Finite Volume schemes

Let us consider the system of conservation laws in Equation (2.5) for d spatial dimensions to compose the following Initial Boundary Value Problem (IVBP)

$$\begin{cases} \text{PDEs:} & \dfrac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathbf{E}(\mathbf{U}) = \mathbf{S} \\ \text{IC:} & \mathbf{U}(\mathbf{x},0) = \mathring{\mathbf{U}}(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega \\ \text{BC:} & \mathbf{U}(\mathbf{x},t) = \mathbf{U}_{\partial\Omega}(\mathbf{x},t) \quad \forall \mathbf{x} \in \partial\Omega \end{cases} \tag{3.1}$$

defined inside the domain $\Omega \times [0,T]$ with $\Omega \subseteq \mathbb{R}^d$ and $T \in \mathbb{R}^+$. As outlined before, the spatial domain is discretized in N volume cells, defined as $\Omega_i \subset \Omega$, such that $\Omega = \bigcup_{i=1}^{N} \Omega_i$. The volume contained in each of these cells is computed as

$$\vartheta_i = \int_{\Omega_i} d\Omega_i \qquad i = 1, ..., N \tag{3.2}$$

Inside each cell at time t^n, the conserved quantities are defined as cell averages as

$$\mathbf{U}_i^n = \frac{1}{\vartheta_i}\int_{\Omega_i} \mathbf{U}(\mathbf{x}, t^n) d\Omega_i \qquad i = 1, ..., N\,. \tag{3.3}$$

The conservation law in (3.1) is integrated inside each cell Ω_i following (2.16) and using definition in (3.3), leading to

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{1}{\vartheta_i}\left(\int_0^{\Delta t}\int_{\partial\Omega_i} \mathbf{E}(\mathbf{U})\hat{\mathbf{n}} dS dt + \int_0^{\Delta t}\int_{\Omega_i} \mathbf{S}(\mathbf{U}, \mathbf{x}) dV dt\right) \tag{3.4}$$

where $\mathbf{U}_i^{n+1}$ is the cell average at t^{n+1} and $\mathbf{U}_i^n$ is the cell average at t^n. $\mathbf{U}_i^{n+1}$ can be computed explicitly from $\mathbf{U}_i^n$ plus a suitable approximation of the integral of the fluxes over $\partial\Omega_i$ and the contribution of the source term inside Ω_i.

3.2 Godunov's method in 1D

When considering the particular case of one spatial dimension, the IVBP in (2.17) becomes

$$\begin{cases} \text{PDEs:} & \dfrac{\partial \mathbf{U}}{\partial t} + \dfrac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{S} \\ \text{IC:} & \mathbf{U}(x, 0) = \mathring{\mathbf{U}}(x) \\ \text{BC:} & \mathbf{U}(a, t) = \mathbf{U}_a(t) \quad \mathbf{U}(b, t) = \mathbf{U}_b(t) \end{cases} \tag{3.5}$$

defined inside the domain $[a, b] \times [0, T]$, with $\mathring{\mathbf{U}}(x)$ the initial condition and $\mathbf{U}_a(t)$ and $\mathbf{U}_b(t)$ the left and right boundary conditions. In this case, the computational grid is composed by N cells

$$a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < ... < x_{N-\frac{1}{2}} < x_{N+\frac{1}{2}} = b \tag{3.6}$$

as shown in Figure 3.1, with cells defined as

$$\Omega_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \qquad i = 1, ..., N \tag{3.7}$$

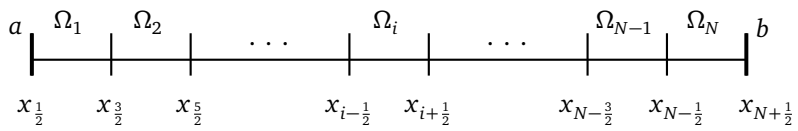


Figure 3.1: Mesh discretization

Cell sizes are derived from (3.2) and defined as

$$\Delta x_i = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} dx = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} \qquad i = 1, ..., N\,. \tag{3.8}$$

Inside each cell, the conserved quantities are defined as cell averages as

$$\mathbf{U}_i^n = \frac{1}{\Delta x_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{U}(x,t^n)dx \qquad i=1,...,N \tag{3.9}$$

at time t^n. The integration of (2.17) yields

$$\begin{aligned}\mathbf{U}_i^{n+1} = \quad & \mathbf{U}_i^n - \frac{1}{\Delta x_i}\left(\int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{U}(x_{i+\frac{1}{2}},t))dt - \int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{U}(x_{i-\frac{1}{2}},t))dt\right) \\ & + \int_{t^n}^{t^{n+1}} \frac{1}{\Delta x_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{S}(\mathbf{U}(x,t),x,t)dx\,dt\,,\end{aligned} \tag{3.10}$$

with $t^{n+1} = t^n + \Delta t$. If considering a first order explicit approximation of the integral in time of the physical fluxes at cell boundaries, it is possible to define the numerical fluxes and source term as

$$\mathbf{F}^-_{i+\frac{1}{2}} \approx \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{U}(x_{i+\frac{1}{2}},t))dt\,, \qquad \mathbf{F}^+_{i-\frac{1}{2}} \approx \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{U}(x_{i-\frac{1}{2}},t))dt \tag{3.11}$$

and

$$\bar{\mathbf{S}}_i \approx \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{S}(\mathbf{U}(x,t),x,t)dx\,dt \tag{3.12}$$

respectively.

When constructing a 1-st order scheme, all quantities are cell averaged and the numerical fluxes in (3.11) are constructed as a function of such discrete values at both sides of the cell interface as

$$\mathbf{F}^-_{i+\frac{1}{2}} = \mathbf{F}^-_{i+\frac{1}{2}}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n), \qquad \mathbf{F}^+_{i-\frac{1}{2}} = \mathbf{F}^+_{i-\frac{1}{2}}(\mathbf{U}_{i-1}^n, \mathbf{U}_i^n). \tag{3.13}$$

The numerical fluxes in (3.12) are computed by locally solving an initial value problem (IVP) composed of the system of PDEs and an initial condition given by the piecewise constant data at both sides of the interface. This is detailed in the next section.

Equivalently, the approximation of the integral of the source term in (3.12) can be expressed as $\bar{\mathbf{S}}_i = \bar{\mathbf{S}}_i(\mathbf{U}_i^n, x_i, t^n)$. Provided all those definitions, we can rewrite (3.10) as

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{\Delta x_i}\left(\mathbf{F}^-_{i+\frac{1}{2}} - \mathbf{F}^+_{i-\frac{1}{2}}\right) + \frac{\Delta t}{\Delta x_i}\bar{\mathbf{S}}_i \tag{3.14}$$

that represents the 1-st order Godunov's updating scheme [6].

Remark that depending on the nature of the source term, the centered integration of the source term used in (3.14) may prevent the numerical scheme from preserving the exact balance between fluxes and source term under steady state. This is the case of the so-called *geometric source terms*, which are of the form

$$\mathbf{S}(\mathbf{U},x) = \mathbf{S}_s(\mathbf{U})\frac{d}{dx}\phi(x), \tag{3.15}$$

with $\mathbf{S}_s(\mathbf{U})$ a function of the conserved quantities and $\phi(x)$ the geometric function that depends upon the position x and can be discontinuous. In this case, the so-called augmented solvers are of application, leading to the following updating formula

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{\Delta x_i}\left(\mathbf{F}^-_{i+\frac{1}{2}} - \mathbf{F}^+_{i-\frac{1}{2}}\right), \tag{3.16}$$

where $\mathbf{F}^-_{i+\frac{1}{2}} = \mathbf{F}^-_{i+\frac{1}{2}}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n, \bar{\mathbf{S}}_{i+1/2})$, $\mathbf{F}^+_{i-\frac{1}{2}} = \mathbf{F}^+_{i-\frac{1}{2}}(\mathbf{U}_{i-1}^n, \mathbf{U}_i^n, \bar{\mathbf{S}}_{i-1/2})$ are the numerical fluxes and $\bar{\mathbf{S}}_{i+1/2} = \bar{\mathbf{S}}_{i+1/2}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n, x_i, x_{i+1})$ is a suitable approximation of the integral of the source term across the cell edge.

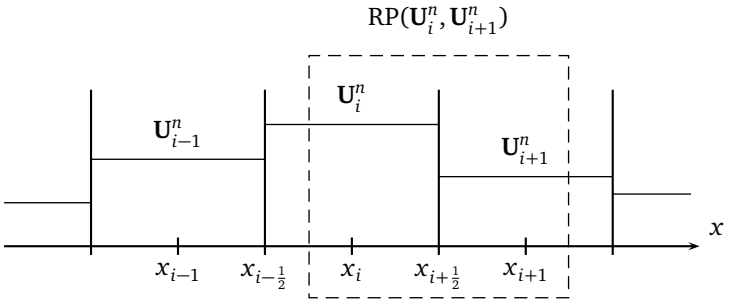


Figure 3.2: Neighbouring region of cell Ω_i and representation of piecewise defined data, showing RP at $x_{i+\frac{1}{2}}$ that will be referred to as RP($\mathbf{U}_i^n, \mathbf{U}_{i+1}^n$).

3.3 The Riemann Problem

At each interface, numerical fluxes in (3.11) can be computed by locally solving a initial value problem (IVP) composed of the system of PDEs and a initial condition given by piecewise constant data at both sides of the interface, as depicted in Figure 3.2. Such problem is defined at cell interface $x_{i+\frac{1}{2}}$ as

$$\begin{cases} \text{PDEs:} \quad \dfrac{\partial \mathbf{U}}{\partial t} + \dfrac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{S} \\ \\ \text{IC:} \quad \mathbf{U}(x, t^n) = \begin{cases} \mathbf{U}_i^n & x < x_{i+\frac{1}{2}} \\ \mathbf{U}_{i+1}^n & x > x_{i+\frac{1}{2}} \end{cases} \end{cases} \tag{3.17}$$

inside the domain $[x_{i+1/2} - \frac{\Delta x}{2}, x_{i+1/2} + \frac{\Delta x}{2}] \times [t^n, t^n + \Delta t]$. Problem in (3.17) is called Riemann Problem, hereafter RP. At interface $x_{i+\frac{1}{2}}$, it will be referred to as RP($\mathbf{U}_i^n, \mathbf{U}_{i+1}^n$). For the sake of clarity, spatial and temporal variables will be redefined setting the reference for the spatial coordinate at $x_{i+\frac{1}{2}}$ to $x = 0$ and for the time t^n to $t = 0$, leading to

$$\begin{cases} \dfrac{\partial \mathbf{U}}{\partial t} + \dfrac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{S} \\ \\ \mathbf{U}(x, 0) = \begin{cases} \mathbf{U}_i & x < 0 \\ \mathbf{U}_{i+1} & x > 0 \end{cases} \end{cases} \tag{3.18}$$

inside the domain $[-\frac{\Delta x}{2}, \frac{\Delta x}{2}] \times [0, \Delta t]$. The similarity solution is denoted by $\mathbf{U}(x/t)$ and composed of $N_\lambda + 1$ constant states separated by N_λ waves [115].

3.3.1 Integral relations in discontinuous solutions

It is of utmost importance to mention that there exists a certain relation between the wave speed and the jump of conserved quantities and fluxes across the discontinuities carried by the waves. This relation is called *Rankine-Hugoniot (RH) condition* or *jump condition*. When dealing with non-homogeneous systems

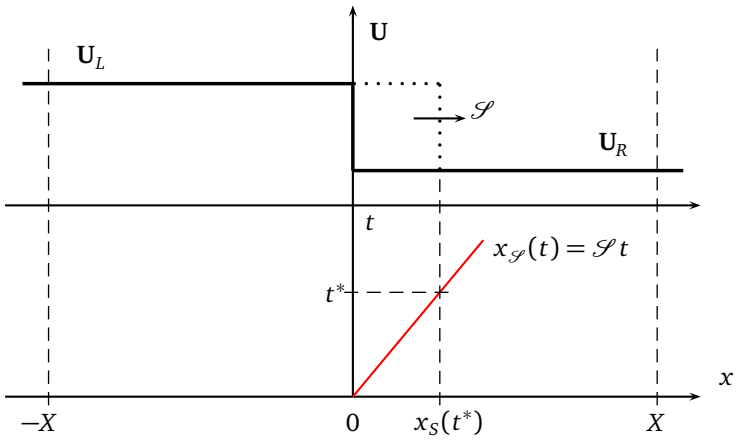


Figure 3.3: Discontinuity propagation in a non-linear system. The integration domain for the derivation of the Rankine-Hugoniot condition is depicted.

of equations, such condition must be extended to account for the contribution of the source term, leading to the *Generalized Rankine-Hugoniot (GRH) condition*.

Initial system in (3.17) is composed of N_λ waves, nevertheless, none of these waves are related to the source term and only conventional RH conditions could be defined across them. In order to study the more general case, where GRH can be defined, it is necessary to express the system in (3.17) in its non-conservative form according to Equation (2.27). In this way, the system is not only characterized by the N_λ eigenvalues associated to the conservative fluxes but also by other N_S eigenvalues, related to extra variables modelling the source term, as the dynamics of the source term is included, in some way, in the set of characteristic fields. For the sake of simplicity, N_S is hereafter set to 1.

The derivation of the GRH condition for the system in (3.18) with a geometric source term can be derived following two different approaches. The first one is to use equation (3.18) and consider the source term as a Dirac delta that moves with the wave [7]. The second option, the one we use here, is to derive the GRH condition by expressing the system in non-conservative form as (2.27). It is done by integrating (2.27) over an arbitrary domain $[-X,X]$ with X sufficiently large, as depicted in Figure 3.3. Notice that the displacement of the discontinuity represented in Figure 3.3 is done from $t = t_0$ to $t = t^* = t_0 + \delta t$, with δt of differential size. For each λ^m wave defining a characteristic field, the left and right states of the solution at each side of the discontinuity carried by wave λ^m are denoted by $\mathbf{U}_L$ and $\mathbf{U}_R$, and the speed of the discontinuity is denoted by $\mathscr{S}^m$. It is worth recalling that there are $N_\lambda + N_S$ characteristic fields. The integral of (2.27) over $[-X,X]$ reads

$$\int_{-X}^{X} \frac{\partial \hat{\mathbf{U}}}{\partial t} dx + \int_{-X}^{X} \frac{\partial \hat{\mathbf{F}}}{\partial x} dx + \int_{-X}^{X} \mathbf{H} \frac{\partial \hat{\mathbf{U}}}{\partial x} dx = 0\,. \tag{3.19}$$

Considering that the integration domain does not change in time, Equation (3.19) is rewritten as

$$\frac{d}{dt} \int_{-X}^{X} \hat{\mathbf{U}} dx + \left[\hat{\mathbf{F}}\right]_{-X}^{X} + \int_{-X}^{X} \mathbf{H} \frac{\partial \hat{\mathbf{U}}}{\partial x} dx = 0\,. \tag{3.20}$$

If separating the first term on the left hand side of Equation (3.20) as

$$\frac{d}{dt}\left(\int_{-X}^{x_S(t)} \hat{\mathbf{U}} dx + \int_{x_S(t)}^{X} \hat{\mathbf{U}} dx \right) = \frac{d}{dt}\left(\hat{\mathbf{U}}_L(X + \mathscr{S}^m t) + \hat{\mathbf{U}}_R(X - \mathscr{S}^m t) \right) \tag{3.21}$$

and taking the time derivative of the previous result, Equation (3.21) is rewritten as

$$\frac{d}{dt}\int_{-X}^{X}\hat{\mathbf{U}}dx = \mathscr{S}^m\left(\hat{\mathbf{U}}_L - \hat{\mathbf{U}}_R\right). \tag{3.22}$$

When combining the results obtained in (3.20) and (3.22), the following condition for the jump is obtained

$$\hat{\mathbf{F}}_R - \hat{\mathbf{F}}_L - \hat{\mathbf{D}} = \mathscr{S}^m\left(\hat{\mathbf{U}}_R - \hat{\mathbf{U}}_L\right), \tag{3.23}$$

where

$$\hat{\mathbf{D}} = -\int_{-X}^{X}\mathbf{H}\frac{\partial\hat{\mathbf{U}}}{\partial x}dx \tag{3.24}$$

is a suitable approximation of the integral of the source term. Notice that the case $\hat{\mathbf{D}} = 0$ corresponds to the traditional RH condition.

The condition for the application of GRH in (3.23) is that the source term is a Dirac delta moving at λ^m. When considering the particular case of a geometric source term as defined in (3.15), the geometric variable only depends on space $\phi = \phi(x)$. This way, it remains fixed in space in the Riemann solution and the jump will remain at $x = 0$. Hence, only when $\mathscr{S}^m = 0$ the GRH condition in (3.23) will be applied (at $x = 0$)

$$\hat{\mathbf{F}}_R - \hat{\mathbf{F}}_L = \hat{\mathbf{D}}. \tag{3.25}$$

Otherwise, when $\mathscr{S}^m \neq 0$ we have that $[\phi]_R = [\phi]_L$ and hence

$$\hat{\mathbf{D}} = 0, \tag{3.26}$$

recovering the traditional RH condition

$$\mathbf{F}_R - \mathbf{F}_L = \mathscr{S}^m(\mathbf{U}_R - \mathbf{U}_L) \tag{3.27}$$

for all $\mathscr{S}^m \neq 0$.

It is worth recalling that the set of right (left) states that can be connected to a given left (right) state by means of a discontinuous solution describe a curve in the phase space called Hugoniot Locus (HL), or Generalized Hugoniot Locus (GHL).

3.3.2 Fundamental wave structures in the Riemann solution

As the system in (3.18) is non-linear, the waves may lead to shocks, rarefaction waves or contact waves and the solution may become of high complexity. Definitions 6 and 7 can be used to determine the nature of each wave [115], making possible to define the following types of waves:

- **Shock wave:** If λ^m defines a *genuinely non-linear field* and the following conditions apply:
  - RH condition:

    $$\mathbf{F}(\mathbf{U}_L) - \mathbf{F}(\mathbf{U}_R) = \mathscr{S}^m(\mathbf{U}_L - \mathbf{U}_R) \tag{3.28}$$

    as any discontinuous solution must satisfy this integral relation.
  - Entropy conditions:

$$\lambda^m(\mathbf{U}_L) > \mathscr{S}^m > \lambda^m(\mathbf{U}_R) \tag{3.29}$$

as the shock arises from the convergence of the characteristic lines.

then left and right states $\mathbf{U}_L$ and $\mathbf{U}_R$ will be connected by a single jump discontinuity wave of speed $\mathscr{S}^m$ called shock wave.

- **Contact wave:** If λ^m defines a *linearly degenerate field* and the following conditions apply:
  - RH condition:
  $$\mathbf{F}(\mathbf{U}_L) - \mathbf{F}(\mathbf{U}_R) = \mathscr{S}^m(\mathbf{U}_L - \mathbf{U}_R) \tag{3.30}$$
  - Parallel characteristic condition:
  $$\lambda^m(\mathbf{U}_L) = \mathscr{S}^m = \lambda^m(\mathbf{U}_R) \tag{3.31}$$
  - Conservation of the Riemann Invariants across the wave if the system is homogeneous. The case of non-homogeneous systems is discussed next.

  then left and right states $\mathbf{U}_L$ and $\mathbf{U}_R$ will be connected by a single jump discontinuity wave of speed $\mathscr{S}^m$ called contact wave.

  Contact waves in non-conservative systems:

  The presence of contact discontinuities in RPs given by non-homogeneous systems of conservation laws has to be taken into account when constructing augmented solvers. It is of utmost importance to mention that given a initial left state, $\mathbf{U}_L$, the right state, hereafter denoted by $\mathbf{U}(\xi)$, does not necessarily lie on the integral curve, though it will always be related to the left state by means of the GRH condition [114]. Notice that $\mathbf{U}_L = \mathbf{U}(0)$.

  Let us consider the non-conservative system in (2.27) and assume that equality (2.50) holds for one of the characteristic fields, namely the m-th characteristic field, associated to eigenvalue λ^m and eigenvector $\mathbf{e}^m$. Then, the m-th field is linearly degenerate and the associated contact wave is given by

$$\mathbf{U}(x,t) = \begin{cases} \mathbf{U}_L & x < \mathscr{S}^m t \\ \mathbf{U}(\xi) & x > \mathscr{S}^m t \end{cases} \tag{3.32}$$

  with constant speed $\mathscr{S}^m = \lambda^m(\mathbf{U}(\xi)) = \lambda^m(\mathbf{U}_L)$. All possible $\mathbf{U}(\xi)$ states can be found by means of the generalized Hugoniot locus, which can be obtained from (3.23)

$$\mathbf{F}(\mathbf{U}(\xi)) - \mathbf{F}(\mathbf{U}_L) - \mathscr{S}^m(\mathbf{U}(\xi) - \mathbf{U}_L) = \mathbf{D}\,. \tag{3.33}$$

  In this way, $\mathbf{U}(\xi)$ will satisfy the GRH condition, however, the relevant m-Riemann invariants may not be conserved across the contact discontinuity, hence IC and GHL may not coincide. To find the condition so that such sets of states coincide, let us consider the differential form of (3.33)

$$\frac{d}{d\xi}\left[\mathbf{F}(\mathbf{U}(\xi)) - \mathscr{S}^m\mathbf{U}(\xi)\right] = \frac{d}{d\xi}\mathbf{D} \tag{3.34}$$

  that can be rewritten as

$$\frac{d\mathbf{F}}{d\mathbf{U}}\frac{d\mathbf{U}(\xi)}{d\xi} - \mathscr{S}^m\frac{d\mathbf{U}(\xi)}{d\xi} = \frac{d}{d\xi}\mathbf{D}\,. \tag{3.35}$$

To enforce the solution to lie on both the IC and the GHL, we set $\mathbf{U} = \mathbf{U}^m(\xi)$ to be the set of states lying on the IC according to Definition 5, yielding

$$\mathbf{J}\frac{d\mathbf{U}^m(\xi)}{d\xi} - \mathscr{S}^m\frac{d\mathbf{U}^m(\xi)}{d\xi} = \frac{d}{d\xi}\mathbf{D}, \tag{3.36}$$

where $d\mathbf{U}^m(\xi)/d\xi$ can be substituted by $\mathbf{e}^m$ as the solution follows the IC, and $\mathscr{S}^m$ by λ^m, leading to

$$\mathbf{J}\cdot\mathbf{e}^m - \lambda^m\cdot\mathbf{e}^m = \frac{d}{d\xi}\mathbf{D}, \tag{3.37}$$

that can be rewritten by means of (2.29) as

$$-\mathbf{H}\cdot\mathbf{e}^m = \frac{d}{d\xi}\mathbf{D}. \tag{3.38}$$

Only when relation in (3.38) is satisfied, the IC and GHL coincide and the Riemann invariants are conserved across the contact wave.

- **Rarefaction wave:** If λ^m defines a *genuinely non-linear field* and the following conditions apply:
  - Divergence of characteristic

$$\lambda^m(\mathbf{U}_L) < \mathscr{S}^m < \lambda^m(\mathbf{U}_R) \tag{3.39}$$

  - Conservation of the Riemann Invariants across the wave.

then left and right states $\mathbf{U}_L$ and $\mathbf{U}_R$ will be connected by a smooth transition called rarefaction wave.

3.3.3 Integral formulation of the RP

For this particular case of a RP, it can also be useful to derive its integral form. Integrating (3.18) over the control volume $[-\Delta x/2, \Delta x/2]\times[0,\Delta t]$

$$\int_{-\Delta x/2}^{\Delta x/2}\int_0^{\Delta t}\left(\frac{\partial\mathbf{U}}{\partial t} + \frac{\partial\mathbf{F}}{\partial x} - \mathbf{S}\right)dxdt = 0, \tag{3.40}$$

the following expression for the integral volume of $\mathbf{U}(x,\Delta t)$ is obtained

$$\int_{-\Delta x/2}^{\Delta x/2}\mathbf{U}(x,\Delta t)\,dx = x_{i+1}\mathbf{U}_{i+1} + x_i\mathbf{U}_i - (\delta\mathbf{F} - \bar{\mathbf{S}})_{i+\frac{1}{2}}\Delta t, \tag{3.41}$$

with $\delta(\cdot)_{i+\frac{1}{2}} = (\cdot)_{i+1} - (\cdot)_i$, $\mathbf{F}_{i+1} = \mathbf{F}(\mathbf{U}_{i+1})$ and $\mathbf{F}_i = \mathbf{F}(\mathbf{U}_i)$ and the source term integrated as

$$\int_{-\Delta x/2}^{\Delta x/2}\int_0^{\Delta t}\mathbf{S}(\mathbf{U}_i,\mathbf{U}_{i+1},t=0)\,dxdt = \Delta t\bar{\mathbf{S}}_{i+\frac{1}{2}}, \tag{3.42}$$

where $\bar{\mathbf{S}}_{i+\frac{1}{2}}$ is the approximation of the integral of the source term in (3.24), inside $[-\Delta x/2, \Delta x/2]$.

3.4 Concluding remarks

The highlights of this chapter are listed below:

- The first order FV formulation is derived from the integral formulation of the equations. When considering source terms, they can be included in the scheme either as a centered contribution in the cell or at cell interfaces in an upwind fashion. The latter choice is based on accounting for the source term in the definition of the RP, which helps to adequately represent the correct influence of the source in the solution and in the time step. This is the approach considered in this work for geometric source terms.
- When considering the source term in the definition of the RP, the traditional RH condition is extended to the GRH condition. Such condition is used in the generation of the solvers herein proposed.
- Across a contact wave originated by the presence of a geometric source term, the conservation of the Riemann invariants is not required and extra conditions must be taken into consideration. This idea will be considered again in the application to the SWE to show that either momentum or energy conservation can be imposed across such a wave.

4 FIRST ORDER APPROXIMATE RIEMANN SOLVERS

In the previous chapter, the RP was introduced as the fundamental component of FV schemes. In order for the numerical scheme to provide a good performance, the discretization of the fluxes at cell interfaces must be done in a particular way. It is in the resolution of the RP where the dynamics of the problem is represented, while Godunov's formula only provides the updating in time of the variables by means of a space-time integral balance. Such idea underscores the importance of a proper resolution of the RP in order to obtain an accurate computation of the numerical fluxes required in Godunov's scheme. Normally, the numerical fluxes are sought so that they are *conservative* (same flux is leaving and entering at the interface) and *consistent* (the numerical flux tends to the continuous flux as the mesh size vanishes) [7].

When dealing with linear problems, the solution of the RP can be analytically derived as done in Section 2.3.2. On the other hand, the resolution of non-linear RPs becomes a more challenging task and requires sophisticated algorithms called *Riemann solvers*. Such methods can be broadly divided in *exact* and *approximate* solvers. The former are computationally more expensive and also more accurate than the latter, whereas the latter are cheaper but with the disadvantage of sometimes providing gross estimations of the solution. This chapter is devoted to the study of first order approximate solvers for hyperbolic problems with source term. The methods presented here have the particularity of including the source term in the solution of the RP as an extra wave of velocity $\mathscr{S} = 0$.

In the first section, an augmented solver for non-linear scalar equations is presented. The methodology of this method is the basis for other Roe-type methods herein described. In the second section, two different methods for the resolution of non-linear systems of hyperbolic conservation laws, called Augmented Roe (ARoe) solver and HLLS solver, are presented. The former can be classified as a *complete solver*, as it includes the full wave structure of the system while the latter can be defined as an *incomplete solver*, as it only considers two waves representing the eigenstructure of the Jacobian plus the steady wave due to the source term. Only when the Jacobian matrix is of dimension 2, the HLLS solver will represent the full wave structure. With regards to the linearity of the solution, the ARoe solver is a *linear solver*, as it consider a linearization of the original system of equations, while the HLLS solver is a *nonlinear solver*, as it considers the original system of equations. It is worth noting that higher order methods described in the following chapters of this thesis are based on the use of the ARoe and HLLS solvers.

In what follows, $\delta(\cdot)_{i+1/2}$ operator will represent the difference between the right and left state of the RP centered in $i+1/2$ for a given variable, e.g. $\delta(\cdot)_{i+1/2} = (\cdot)_{(i+1)_L} - (\cdot)_{i_R}$ and $\delta(\cdot)_{i-1/2} = (\cdot)_{i_L} - (\cdot)_{(i-1)_R}$

4.1 First order augmented solver for scalar equations

Scalar version of RP in (3.18) reads

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial f(u)}{\partial x} = s \\ \\ u(x,0) = \begin{cases} u_i & x < 0 \\ u_{i+1} & x > 0 \end{cases} \end{cases} \tag{4.1}$$

where $u \in \mathbb{R}$ is the conserved variable, s the source term and $f(u)$ the physical flux, which is a nonlinear function of the conserved variable.

The integral form of (4.1) over the control volume $[0, \Delta t] \times [-x_L, x_R]$ is given by

$$\int_{-x_L}^{x_R} \int_0^{\Delta t} \left(\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} - s \right) dx dt = 0 \tag{4.2}$$

and the following expression for the integral volume of $u(x, \Delta t)$ inside $[-x_L, x_R]$ is obtained

$$\int_{-x_L}^{x_R} u(x, \Delta t)\, dx = x_R u_{i+1} + x_L u_i - (\delta f - \bar{s})_{i+\frac{1}{2}} \Delta t, \tag{4.3}$$

with $f_{i+1} = f(u_{i+1})$ and $f_i = f(u_i)$ and the source term integrated as

$$\int_{-x_L}^{x_R} \int_0^{\Delta t} s(u_i, u_{i+1}, t = 0)\, dx dt = \Delta t \bar{s}_{i+\frac{1}{2}}\,. \tag{4.4}$$

Problem in (4.1) can be approximated by the following constant coefficient linear RP

$$\begin{cases} \dfrac{\partial \hat{u}}{\partial t} + \widetilde{\lambda}_{i+\frac{1}{2}} \dfrac{\partial \hat{u}}{\partial x} = s \\ \\ \hat{u}(x,0) = \begin{cases} u_i & x < 0 \\ u_{i+1} & x > 0 \end{cases} \end{cases} \tag{4.5}$$

where $\hat{u}(x,t)$ is the approximate solution of (4.1) and $\widetilde{\lambda}_{i+\frac{1}{2}}$ is a constant wave velocity defined as a function of left and right states (u_i and u_{i+1}) that represents an approximation of the propagation velocity $\lambda(u) = \partial_u f(u)$ at $x_{i+\frac{1}{2}}$.

If expressing the integral form of (4.19) over the same control volume than in the previous case

$$\int_{-x_L}^{x_R} \hat{u}(x, \Delta t)\, dx = x_R u_{i+1} + x_L u_i - \left(\widetilde{\lambda} \delta u + \bar{s} \right)_{i+\frac{1}{2}} \Delta t \tag{4.6}$$

and imposing *consistency condition* between (4.3) and (4.6)

$$\int_{-x_L}^{x_R} \hat{u}(x, \Delta t)\, dx = \int_{-x_L}^{x_R} u(x, \Delta t)\, dx \tag{4.7}$$

the following constraint is noticed

$$\delta f_{i+\frac{1}{2}} = \widetilde{\lambda}_{i+\frac{1}{2}} \delta u_{i+\frac{1}{2}}\,, \tag{4.8}$$

that allows to compute the value of $\widetilde{\lambda}_{i+\frac{1}{2}}$.

The solution for $\hat{u}$ in (4.5) consists of three regions, as depicted in Figure 4.1 for the particular case when $\widetilde{\lambda}_{i+\frac{1}{2}} > 0$.

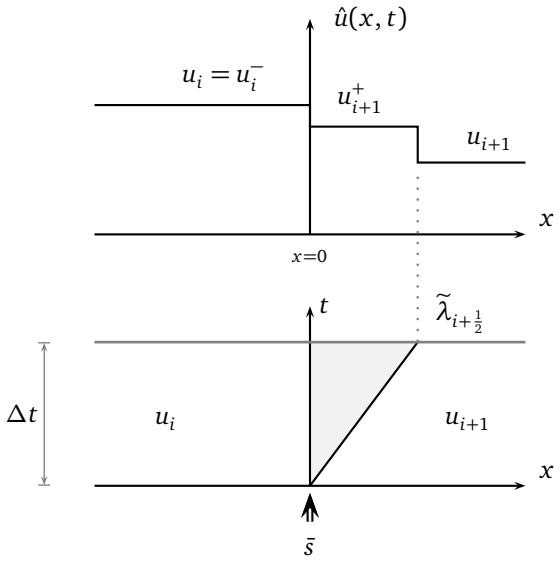


Figure 4.1: Values of the solution $\hat{u}(x,t)$ in each wedge of the (x,t) plane.

It is possible to define the solution on the left and right sides of the t axis, denoted by u_i^- and u_{i+1}^+ respectively as depicted in Figure 4.1. These values are defined as

$$u_i^- = \lim_{x\to 0^-} \hat{u}(x,t) \qquad u_{i+1}^+ = \lim_{x\to 0^+} \hat{u}(x,t). \tag{4.9}$$

In Figure 4.1, it is observed that the solution on the left hand side of the interface, u_i^-, is equal to the left state since the wave propagates to the right. However, a new state on the right hand side of the interface, u_{i+1}^+, appears. To find the value for u_{i+1}^+ the RH condition across the steady wave at the interface must be obtained first

$$f_{i+1}^+ - f_i^- - \bar{s}_{i+\frac{1}{2}} = 0. \tag{4.10}$$

On the other hand, if we assume that the difference of states and fluxes across the discontinuity are related using the approximate wave velocity in the following way

$$f_{i+1}^+ - f_i^- = \widetilde{\lambda}_{i+\frac{1}{2}}(u_{i+1}^+ - u_i^-), \tag{4.11}$$

then, the right state can be obtained by substitution of (4.11) in (4.10), yielding

$$u_{i+1}^+ = u_i^- + \left(\frac{\bar{s}}{\widetilde{\lambda}}\right)_{i+\frac{1}{2}} = u_i + \left(\frac{\bar{s}}{\widetilde{\lambda}}\right)_{i+\frac{1}{2}}. \tag{4.12}$$

It is also possible to apply the RH condition across the positive moving wave as

$$f_{i+1} - f_{i+1}^+ = \widetilde{\lambda}_{i+\frac{1}{2}}(u_{i+1} - u_{i+1}^+) \tag{4.13}$$

and substitution of (4.12) in (4.13) leads to the expression for the right state flux

$$f^+_{i+1} = f_i + \bar{s}_{i+\frac{1}{2}}. \tag{4.14}$$

The solution in the $x-t$ plane can be expressed as a piecewise constant function that depends upon x and t as

$$\hat{u}(x,t) = \begin{cases} u_i & \text{if} \quad x < \widetilde{\lambda}_{i+\frac{1}{2}} t \\ u_i + (\theta \delta u)_{i+\frac{1}{2}} & \text{if} \quad \widetilde{\lambda}_{i+\frac{1}{2}} t < x < 0 \\ u_{i+1} & \text{if} \quad 0 < x \end{cases} \tag{4.15}$$

when $\widetilde{\lambda}_{i+\frac{1}{2}} < 0$, and

$$\hat{u}(x,t) = \begin{cases} u_i & \text{if} \quad x < 0 \\ u_{i+1} - (\theta \delta u)_{i+\frac{1}{2}} & \text{if} \quad 0 < x < \widetilde{\lambda}_{i+\frac{1}{2}} t \\ u_{i+1} & \text{if} \quad \widetilde{\lambda}_{i+\frac{1}{2}} t < x \end{cases} \tag{4.16}$$

when $\widetilde{\lambda}_{i+\frac{1}{2}} > 0$, with

$$\theta_{i+\frac{1}{2}} = 1 - \left(\frac{\bar{s}}{\delta f} \right)_{i+\frac{1}{2}} \tag{4.17}$$

and from Equations (4.15) and (4.16) they yield

$$\begin{aligned} u^-_i &= \begin{cases} u_i & if \quad \widetilde{\lambda}_{i+\frac{1}{2}} > 0 \\ u_i + (\theta \delta u)_{i+\frac{1}{2}} & if \quad \widetilde{\lambda}_{i+\frac{1}{2}} < 0 \end{cases} \\ u^+_{i+1} &= \begin{cases} u_{i+1} - (\theta \delta u)_{i+\frac{1}{2}} & if \quad \widetilde{\lambda}_{i+\frac{1}{2}} > 0 \\ u_{i+1} & if \quad \widetilde{\lambda}_{i+\frac{1}{2}} < 0 \end{cases} \end{aligned} \tag{4.18}$$

4.2 First order augmented solver for systems of N_λ waves

4.2.1 Approximate solution using ARoe solver

RP in (3.18) can be approximated by exactly solving the following constant coefficient linear RP

$$\begin{cases} \dfrac{\partial \hat{\mathbf{U}}}{\partial t} + \widetilde{\mathbf{J}}_{i+\frac{1}{2}} \dfrac{\partial \hat{\mathbf{U}}}{\partial x} = \mathbf{S} \\ \\ \hat{\mathbf{U}}(x,0) = \begin{cases} \mathbf{U}_i & x < 0 \\ \mathbf{U}_{i+1} & x > 0 \end{cases} \end{cases} \tag{4.19}$$

where $\hat{\mathbf{U}}(x,t)$ is the approximate solution of (3.18) and $\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}}(\mathbf{U}_i, \mathbf{U}_{i+1})$ is a constant matrix defined as a function of left and right states ($\mathbf{U}_i$ and $\mathbf{U}_{i+1}$) that represents an approximation of the Jacobian at $x_{i+\frac{1}{2}}$.

If expressing the integral form of (4.19) over the same control volume than in the previous case

$$\int_{-x_L}^{x_R} \hat{\mathbf{U}}(x, \Delta t)\, dx = x_R \mathbf{U}_{i+1} + x_L \mathbf{U}_i - \left(\widetilde{\mathbf{J}} \delta \mathbf{U} + \bar{\mathbf{S}} \right)_{i+\frac{1}{2}} \Delta t \tag{4.20}$$

and imposing the *consistency condition*

$$\int_{-x_L}^{x_R} \hat{\mathbf{U}}(x,\Delta t)\,dx = \int_{-x_L}^{x_R} \mathbf{U}(x,\Delta t)\,dx, \tag{4.21}$$

the following constraint is noticed

$$\delta\mathbf{F}_{i+\frac{1}{2}} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}}\delta\mathbf{U}_{i+\frac{1}{2}}\,. \tag{4.22}$$

Matrix $\widetilde{\mathbf{J}}_{i+\frac{1}{2}}$ is considered to be diagonalizable with N_λ approximate real eigenvalues

$$\widetilde{\lambda}^1_{i+\frac{1}{2}} < \ldots < \widetilde{\lambda}^I_{i+\frac{1}{2}} < 0 < \widetilde{\lambda}^{I+1}_{i+\frac{1}{2}} < \ldots < \widetilde{\lambda}^{N_\lambda}_{i+\frac{1}{2}} \tag{4.23}$$

and N_λ eigenvectors $\widetilde{\mathbf{e}}^1,\ldots,\widetilde{\mathbf{e}}^{N_\lambda}$. With them, two approximate matrices, $\widetilde{\mathbf{P}}_{i+\frac{1}{2}} = (\widetilde{\mathbf{e}}^1,\ldots,\widetilde{\mathbf{e}}^{N_\lambda})_{i+\frac{1}{2}}$ and $\widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}$ are built with the following property

$$\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}}, \qquad \widetilde{\mathbf{\Lambda}}_{i+\frac{1}{2}} = \begin{pmatrix} \widetilde{\lambda}^1 & & 0 \\ & \ddots & \\ 0 & & \widetilde{\lambda}^{N_\lambda} \end{pmatrix}_{i+\frac{1}{2}}, \tag{4.24}$$

where $\widetilde{\mathbf{\Lambda}}_{i+\frac{1}{2}}$ is a diagonal matrix with approximate eigenvalues in the main diagonal. As done in Section 2.3.2, system in (4.19) can be transformed using $\widetilde{\mathbf{P}}^{-1}$ matrix as follows

$$\widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\left(\frac{\partial\hat{\mathbf{U}}}{\partial t} + \widetilde{\mathbf{J}}_{i+\frac{1}{2}}\frac{\partial\hat{\mathbf{U}}}{\partial x}\right) = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\mathbf{S}, \tag{4.25}$$

expressing (4.19) in terms of the characteristic variables $\hat{\mathbf{W}} = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\hat{\mathbf{U}}$, with $\hat{\mathbf{W}} = (\hat{w}^1,\ldots,\hat{w}^{N_\lambda})$. This transformation leads to a decoupled system that generates the following linear RP

$$\begin{cases} \dfrac{\partial\hat{\mathbf{W}}}{\partial t} + \widetilde{\mathbf{\Lambda}}_{i+\frac{1}{2}}\dfrac{\partial\hat{\mathbf{W}}}{\partial x} = \mathbf{B}_{i+\frac{1}{2}} \\ \hat{\mathbf{W}}(x,0) = \begin{cases} \mathbf{W}_i = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\mathbf{U}_i & \text{if} \quad x<0 \\ \mathbf{W}_{i+1} = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\mathbf{U}_{i+1} & \text{if} \quad x>0 \end{cases} \end{cases} \tag{4.26}$$

with $\mathbf{B}_{i+\frac{1}{2}} = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\mathbf{S} = (\beta^1,\ldots,\beta^{N_\lambda})_{i+\frac{1}{2}}$, where each equation

$$\frac{\partial\hat{w}^m}{\partial t} + \widetilde{\lambda}^m_{i+\frac{1}{2}}\frac{\partial\hat{w}^m}{\partial x} = \beta^m_{i+\frac{1}{2}}, \qquad m = 1,\ldots,N_\lambda \tag{4.27}$$

involves the variable $\hat{w}^m$ and the source term $\beta^m_{i+\frac{1}{2}}$. As equations in (4.27) are decoupled, RP in (4.26) can be decomposed in N_λ independent RPs

$$\begin{cases} \dfrac{\partial\hat{w}^m}{\partial t} + \widetilde{\lambda}^m_{i+\frac{1}{2}}\dfrac{\partial\hat{w}^m}{\partial x} = \beta^m_{i+\frac{1}{2}} \\ \hat{w}^m(x,0) = \begin{cases} w^m_i & \text{if} \quad x<0 \\ w^m_{i+1} & \text{if} \quad x>0 \end{cases} \end{cases} \tag{4.28}$$

The solution for each $\hat{w}^m$ characteristic variable is given by the solution of the scalar RP (4.28) [20] and consists of three regions as depicted in Figure 4.2.

The solution can be expressed as a piecewise constant function that depends upon x and t as

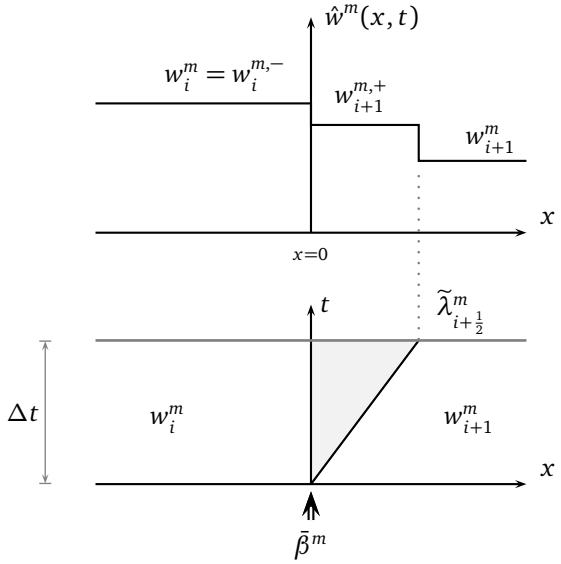


Figure 4.2: Values of the solution $\hat{w}(x,t)$ in the (x,t) plane.

$$\hat{w}^m(x,t) = \begin{cases} w_i^m & \text{if} \quad x < \widetilde{\lambda}_{i+\frac{1}{2}}^m t \\ w_i^m + (\theta\delta w)_{i+\frac{1}{2}}^m & \text{if} \quad \widetilde{\lambda}_{i+\frac{1}{2}}^m t < x < 0 \\ w_{i+1}^m & \text{if} \quad 0 < x \end{cases} \tag{4.29}$$

when $\widetilde{\lambda}_{i+\frac{1}{2}}^m < 0$, and

$$\hat{w}^m(x,t) = \begin{cases} w_i^m & \text{if} \quad x < 0 \\ w_{i+1}^m - (\theta\delta w)_{i+\frac{1}{2}}^m & \text{if} \quad 0 < x < \widetilde{\lambda}_{i+\frac{1}{2}}^m t \\ w_{i+1}^m & \text{if} \quad \widetilde{\lambda}_{i+\frac{1}{2}}^m t < x \end{cases} \tag{4.30}$$

when $\widetilde{\lambda}_{i+\frac{1}{2}}^m > 0$, with

$$\theta_{i+\frac{1}{2}}^m = 1 - \left(\frac{\bar{\beta}^m}{\widetilde{\lambda}^m \alpha^m}\right)_{i+\frac{1}{2}}, \tag{4.31}$$

where the set of wave strengths is defined as

$$\mathbf{A}_{i+\frac{1}{2}} = (\alpha^1, ..., \alpha^{N_\lambda})_{i+\frac{1}{2}}^T = \delta\mathbf{W}_{i+\frac{1}{2}} = (\widetilde{\mathbf{P}}^{-1}\delta\mathbf{U})_{i+\frac{1}{2}} \tag{4.32}$$

and the set of source strengths

$$\bar{\mathbf{B}}_{i+\frac{1}{2}} = (\bar{\beta}^1, ..., \bar{\beta}^{N_\lambda})_{i+\frac{1}{2}}^T = (\widetilde{\mathbf{P}}^{-1}\bar{\mathbf{S}})_{i+\frac{1}{2}}\,. \tag{4.33}$$

Analogously, it is possible to define the solution for each characteristic RP on the left and right sides of the t axis, denoted by $w_i^{m,-}$ and $w_{i+1}^{m,+}$ respectively as depicted in Figure 4.2. These values are defined as

$$w_i^{m,-} = \lim_{x\to 0^-} \hat{w}^m(x,t) \qquad w_{i+1}^{m,+} = \lim_{x\to 0^+} \hat{w}^m(x,t) \tag{4.34}$$

and from Equations (4.29) and (4.30) they yield

$$
\begin{aligned}
w_i^{m,-} &= \begin{cases} w_i^m & if \quad \widetilde{\lambda}_{i+\frac{1}{2}}^m > 0 \\ w_i^m + (\theta\delta w)_{i+\frac{1}{2}}^m & if \quad \widetilde{\lambda}_{i+\frac{1}{2}}^m < 0 \end{cases} \\
w_{i+1}^{m,+} &= \begin{cases} w_{i+1}^m - (\theta\delta w)_{i+\frac{1}{2}}^m & if \quad \widetilde{\lambda}_{i+\frac{1}{2}}^m > 0 \\ w_{i+1}^m & if \quad \widetilde{\lambda}_{i+\frac{1}{2}}^m < 0 \end{cases}
\end{aligned}
\tag{4.35}
$$

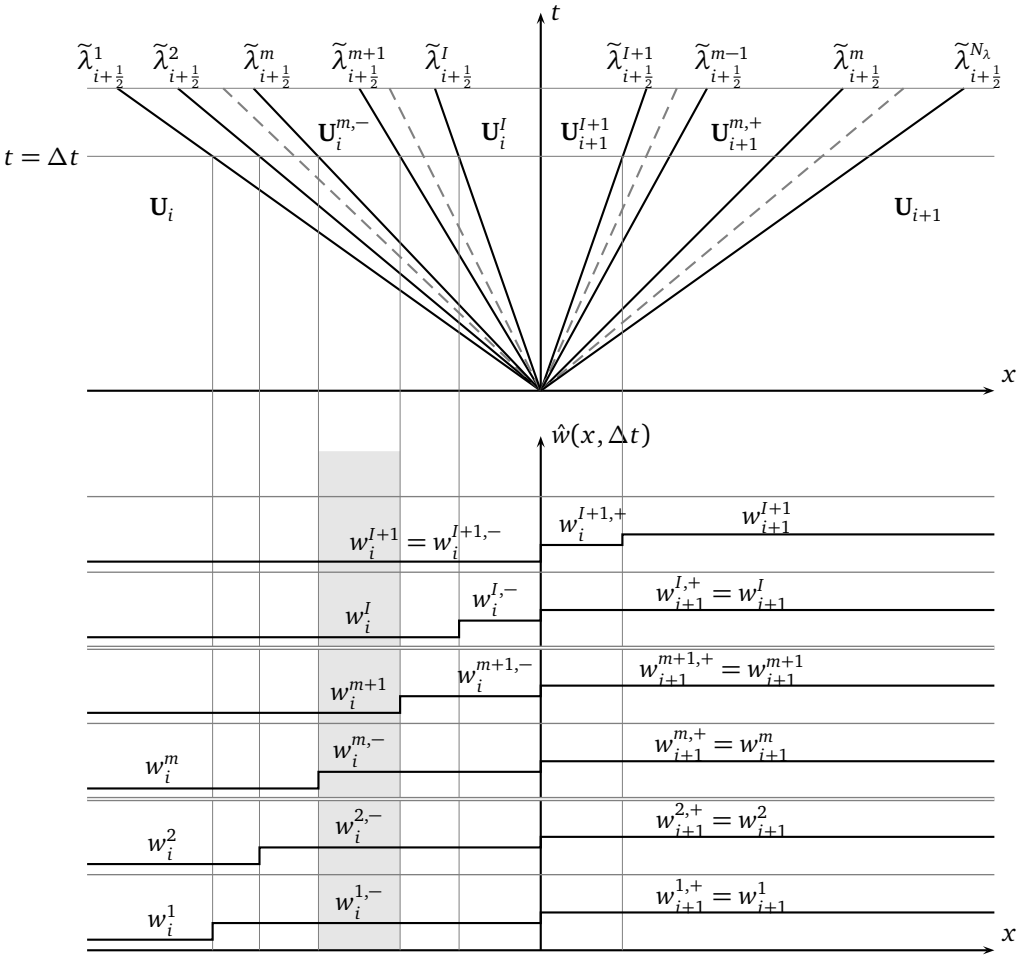


Figure 4.3: Upper: Approximate solution $\hat{\mathbf{U}}(x,t)$. The solution consist of N_λ inner constant states separated by a stationary contact discontinuity, with celerity $S = 0$ at $x = 0$. Lower: The solution for characteristic variables $\hat{w}^m(x,t)$ for $m = 1, ..., I+1$ is depicted at $t = \Delta t$.

The derivation of the general solution $\hat{\mathbf{U}}(x,t)$ for a linear system is based on the expansion of the solution as a linear combination of the vectors that compose the Jacobian's eigenvectors basis, using the relation $\mathbf{U} = \widetilde{\mathbf{P}}\mathbf{W}$, as follows

$$
\hat{\mathbf{U}}(x,t) = \sum_{m_1=1}^{N_\lambda} \hat{w}^{m_1}(x,t)\, \widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1}, \tag{4.36}
$$

where the scalar values $\hat{w}^{m_1}(x,t)$ are the characteristic approximate solutions at the sought point and represent the strength of each wave.

If focusing on a constant state on the left hand side of the t-axis, $\mathbf{U}^{m,-}$, defined between characteristic lines $\widetilde{\lambda}^{m}_{i+\frac{1}{2}}t$ and $\widetilde{\lambda}^{m+1}_{i+\frac{1}{2}}t$, the solution is given by the combination of the characteristic solutions in the spatial domain $[\widetilde{\lambda}^{m}_{i+\frac{1}{2}}t, \widetilde{\lambda}^{m+1}_{i+\frac{1}{2}}t]$. Following expansion in (4.36), $\mathbf{U}^{m,-}$ is given by

$$\mathbf{U}_i^{m,-} = \sum_{\widetilde{\lambda}^{m_1} \leq \widetilde{\lambda}^{m}} \hat{w}^{m_1}(x,t)\, \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} + \sum_{\widetilde{\lambda}^{m_1} \geq \widetilde{\lambda}^{m+1}} \hat{w}^{m_1}(x,t)\, \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} \,. \tag{4.37}$$

The solutions of the characteristic variables are given by (4.29) and (4.30) and are depicted in Figure 4.3 inside the interval $x = [\widetilde{\lambda}^{m}_{i+\frac{1}{2}}\Delta t, \widetilde{\lambda}^{m+1}_{i+\frac{1}{2}}\Delta t]$. The first term of the right hand side of equation (4.37) is then expressed as

$$\sum_{\widetilde{\lambda}^{m_1} \leq \widetilde{\lambda}^{m}} \hat{w}^{m_1}(x,t)\, \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} = \sum_{m_1=1}^{m} \left(w_i^{m_1} + (\theta \delta w)^{m_1}_{i+\frac{1}{2}} \right) \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} \tag{4.38}$$

and the second term becomes

$$\sum_{\widetilde{\lambda}^{m_1} \geq \widetilde{\lambda}^{m+1}} \hat{w}^{m_1}(x,t)\, \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} = \sum_{m_1=m+1}^{I} w_i^{m_1} \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} + \sum_{m_1=I+1}^{N_\lambda} w_i^{m_1} \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} \,. \tag{4.39}$$

Primitive vector solution in (4.37) can be expressed as

$$\mathbf{U}_i^{m,-} = \sum_{m_1=1}^{N_\lambda} w_i^{m_1} \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} + \sum_{m_1=1}^{m} (\theta \delta w \widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} \tag{4.40}$$

and considering that $\mathbf{U}_i = \sum_{m_1=1}^{N_\lambda} w_i^{m_1} \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}}$ and $\delta w^{m_1}_{i+\frac{1}{2}} = \alpha^{m_1}_{i+\frac{1}{2}}$, equation (4.40) can be rewritten as

$$\mathbf{U}_i^{m,-} = \mathbf{U}_i + \sum_{m_1=1}^{m} (\theta \alpha \widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} \,. \tag{4.41}$$

By separating the $\widetilde{\lambda}^{m}$-wave contribution from the summation,

$$\mathbf{U}_i^{m,-} = \mathbf{U}_i + \sum_{m_1=1}^{m-1} (\theta \alpha \widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} + (\theta \alpha \widetilde{\mathbf{e}})^{m}_{i+\frac{1}{2}} \tag{4.42}$$

it is noticed that $\mathbf{U}_i^{m,-}$ in (4.41) can be expressed in terms of its left adjacent state, $\mathbf{U}_i^{m-1,-}$, leading to the following jump between vector solutions

$$\mathbf{U}_i^{m,-} - \mathbf{U}_i^{m-1,-} = (\alpha \theta \widetilde{\mathbf{e}})^{m}_{i+\frac{1}{2}} \tag{4.43}$$

for $1 \leq m \leq I$. Remark that equation (4.42) can only provide solutions in the spatial domain $[\widetilde{\lambda}^{1}t, 0]$.

When seeking the primitive vector solution for a state defined on the right hand side of the t-axis, $\mathbf{U}_{i+1}^{m,+}$, it has to be defined between characteristic lines $\widetilde{\lambda}^{m-1}_{i+\frac{1}{2}}t$ and $\widetilde{\lambda}^{m}_{i+\frac{1}{2}}t$. Following expansion in (4.36), the combination of the characteristic solutions in the spatial domain $[\widetilde{\lambda}^{m-1}_{i+\frac{1}{2}}t, \widetilde{\lambda}^{m}_{i+\frac{1}{2}}t]$ provides

$$\mathbf{U}_{i+1}^{m,+} = \sum_{\widetilde{\lambda}^{m_1} \leq \widetilde{\lambda}^{m-1}} w^{m_1}(x,t) \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} + \sum_{\widetilde{\lambda}^{m_1} \geq \widetilde{\lambda}^{m}} w^{m_1}(x,t) \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} \tag{4.44}$$

and using characteristic solutions in (4.29) and (4.30), the first term of the primitive vector solution becomes

$$\sum_{\widetilde{\lambda}^{m_1}\leq\widetilde{\lambda}^{m-1}} w^{m_1}(x,t)\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} = \sum_{m_1=1}^{I} w^{m_1}_{i+1}\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} + \sum_{m_1=I+1}^{m-1} w^{m_1}_{i+1}\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} \tag{4.45}$$

and the second one

$$\sum_{\widetilde{\lambda}^{m_1}\geq\widetilde{\lambda}^{m}} w^{m_1}(x,t)\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} = \sum_{m_1=m}^{N_\lambda} \left(w^{m_1}_{i+1} - (\theta\delta w)^{m_1}_{i+\frac{1}{2}}\right)\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} \tag{4.46}$$

allowing to express $\mathbf{U}^{m,+}_{i+1}$ as follows

$$\mathbf{U}^{m,+}_{i+1} = \sum_{m_1=1}^{N_\lambda} w^{m_1}_{i+1}\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} - \sum_{m_1=m}^{N_\lambda} (\theta\delta w\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}}\,. \tag{4.47}$$

As done for (4.40), considering that $\mathbf{U}_{i+1} = \sum_{m_1=1}^{N_\lambda} w^{m_1}_{i+1}\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}}$, equation (4.47) can be rewritten as

$$\mathbf{U}^{m,+}_{i+1} = \mathbf{U}_{i+1} - \sum_{m_1=m}^{N_\lambda} (\theta\alpha\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} \tag{4.48}$$

and by separating the $\widetilde{\lambda}^m$-wave contribution from the summation,

$$\mathbf{U}^{m,+}_{i+1} = \mathbf{U}_{i+1} - \sum_{m_1=m+1}^{N_\lambda} (\theta\alpha\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} - (\theta\alpha\widetilde{\mathbf{e}})^{m}_{i+\frac{1}{2}} \tag{4.49}$$

it can be expressed in terms of its right adjacent state, $\mathbf{U}^{m+1,+}_{i+1}$, as follows

$$\mathbf{U}^{m+1,+}_{i+1} - \mathbf{U}^{m,+}_{i+1} = (\theta\alpha\widetilde{\mathbf{e}})^{m}_{i+\frac{1}{2}}\,, \tag{4.50}$$

leading to a jump between vector solutions. Now, equation (4.50) provides exclusively solutions in the spatial domain $[0, \widetilde{\lambda}^{N_\lambda}t]$.

In the vicinity of $x = 0$, left and right states denoted by $\mathbf{U}^-_i$ and $\mathbf{U}^+_{i+1}$ are defined inside spatial domains $[\widetilde{\lambda}_I t, 0]$ and $[0, \widetilde{\lambda}_{I+1}t]$ respectively, and also expressed as

$$\mathbf{U}^-_i = \lim_{x\to 0^-} \hat{\mathbf{U}}(x,t) \qquad \mathbf{U}^+_{i+1} = \lim_{x\to 0^+} \hat{\mathbf{U}}(x,t)\,. \tag{4.51}$$

Expressions for $\mathbf{U}^-_i$ and $\mathbf{U}^+_{i+1}$ can be derived from the previous results, setting $m = I$ in (4.41) and $m = I+1$ in (4.48) respectively, leading to

$$\begin{aligned} \mathbf{U}^-_i &= \mathbf{U}_i + \sum_{m_1=1}^{I} (\theta\alpha\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} \\ \mathbf{U}^+_{i+1} &= \mathbf{U}_{i+1} - \sum_{m_1=I+1}^{N_\lambda} (\theta\alpha\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} \end{aligned}\,. \tag{4.52}$$

The difference between left and right states across the interface can be expressed as

$$\mathbf{U}_{i+1}^{+} - \mathbf{U}_{i}^{-} = \mathbf{U}_{i+1} - \mathbf{U}_{i} - \sum_{m_1=1}^{N_\lambda} (\theta \alpha \tilde{\mathbf{e}})_{i+\frac{1}{2}}^{m_1} \tag{4.53}$$

where wave contributions can be written in their matrix form as

$$\sum_{m_1=1}^{N_\lambda} (\theta \alpha \tilde{\mathbf{e}})_{i+\frac{1}{2}}^{m_1} = \left(\tilde{\mathbf{P}} \Theta \mathbf{A}\right)_{i+\frac{1}{2}} \tag{4.54}$$

with $\Theta_{i+\frac{1}{2}} = diag(\theta_{i+\frac{1}{2}}^{1}, \theta_{i+\frac{1}{2}}^{2}, ..., \theta_{i+\frac{1}{2}}^{N_\lambda})$ a diagonal matrix that allows to rewrite $\tilde{\mathbf{P}}\Theta\mathbf{A} = \tilde{\mathbf{P}}\mathbf{A} - \tilde{\mathbf{P}}\tilde{\Lambda}^{-1}\tilde{\mathbf{B}}$. Substituting the previous results in (4.53) and noticing that $\tilde{\mathbf{P}}\mathbf{A}_{i+\frac{1}{2}} = \mathbf{U}_{i+1} - \mathbf{U}_{i}$, it becomes

$$\mathbf{U}_{i+1}^{+} - \mathbf{U}_{i}^{-} = \left(\tilde{\mathbf{P}}\tilde{\Lambda}^{-1}\tilde{\mathbf{B}}\right)_{i+\frac{1}{2}}, \tag{4.55}$$

from which it can be observed that the difference between left and right states is only due to the presence of the source term. Expressing $\tilde{\mathbf{B}}_{i+\frac{1}{2}} = \left(\tilde{\mathbf{P}}^{-1}\bar{\mathbf{S}}\right)_{i+\frac{1}{2}}$, the following relation is noticed

$$\bar{\mathbf{S}}_{i+\frac{1}{2}} = \left(\tilde{\mathbf{P}}\tilde{\Lambda}\tilde{\mathbf{P}}^{-1}\right)_{i+\frac{1}{2}} \left(\mathbf{U}_{i+1}^{+} - \mathbf{U}_{i}^{-}\right). \tag{4.56}$$

For RP in (4.26), an approximate flux function $\hat{\mathbf{F}}(x,t)$ can be also constructed from cell-averaged flux values $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$ with a similar structure than $\hat{\mathbf{U}}(x,t)$ as depicted in Figure 4.3. In this case, also intercell values for the fluxes can be defined at both sides of the t axis as

$$\mathbf{F}_{i}^{-} = \lim_{x \to 0^{-}} \hat{\mathbf{F}}(x,t) \qquad \mathbf{F}_{i+1}^{+} = \lim_{x \to 0^{+}} \hat{\mathbf{F}}(x,t). \tag{4.57}$$

The Rankine-Hugoniot condition across the stationary wave at $x = 0$ allows to relate approximate fluxes $\mathbf{F}_i^{-}$ and $\mathbf{F}_{i+1}^{+}$ with approximate solutions $\mathbf{U}_i^{-}$ and $\mathbf{U}_{i+1}^{+}$ as

$$\mathbf{F}_{i+1}^{+} - \mathbf{F}_{i}^{-} - \bar{\mathbf{S}}_{i+\frac{1}{2}} = \mathscr{S}(\mathbf{U}_{i+1}^{+} - \mathbf{U}_{i}^{-}) = 0 \tag{4.58}$$

that can be combined with Equation (4.56) to obtain the following relation among fluxes and conserved variables in the innermost regions

$$\mathbf{F}_{i+1}^{+} - \mathbf{F}_{i}^{-} = \tilde{\mathbf{J}}_{i+\frac{1}{2}} \left(\mathbf{U}_{i+1}^{+} - \mathbf{U}_{i}^{-}\right). \tag{4.59}$$

In order to provide a complete description of the approximate flux function $\hat{\mathbf{F}}(x,t)$, the inner constant fluxes on the left side of the (x,t) plane will be denoted by $\mathbf{F}_i^{m,-}$, where $1 \leq m \leq I$. On the right side of the (x,t) plane solution, inner constant states are denoted by $\mathbf{F}_{i+1}^{m,+}$, where $I+1 \leq m \leq N_\lambda$.

The approximate solution for the fluxes can be constructed defining appropriate RH condition across each moving wave, that will be given by

$$\mathbf{F}_{i}^{m,-} - \mathbf{F}_{i}^{m-1,-} = \tilde{\lambda}^{m}(\mathbf{U}_{i}^{m} - \mathbf{U}_{i}^{m-1}) = \left(\tilde{\lambda}\alpha\theta\tilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^{m} \tag{4.60}$$

for $1 \leq m \leq I$, where $\mathbf{F}_i^{I,-} = \mathbf{F}_i^{-}$, $\mathbf{F}_i^{0,-} = \mathbf{F}_i$, and

$$\mathbf{F}_{i+1}^{m+1,+} - \mathbf{F}_{i+1}^{m,+} = \tilde{\lambda}^{m}(\mathbf{U}_{i+1}^{m+1} - \mathbf{U}_{i+1}^{m}) = \left(\tilde{\lambda}\alpha\theta\tilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^{m} \tag{4.61}$$

for $I+1 \leq m \leq N_\lambda$, with $\mathbf{F}_{i+1}^{I+1,+} = \mathbf{F}_{i+1}^{+}$ and $\mathbf{F}_{i+1}^{N_\lambda+1,+} = \mathbf{F}_{i+1}$.

Approximate fluxes on the left and right side of the t axis, $\mathbf{F}_i^-$ and $\mathbf{F}_{i+1}^+$, can be derived using the telescopic property from results in (4.60) and (4.61) as

$$\begin{aligned}\mathbf{F}_i^- &= \mathbf{F}_i + \sum_{m_1=1}^{I} \left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^{m_1} \\ \mathbf{F}_{i+1}^+ &= \mathbf{F}_{i+1} - \sum_{m_1=I+1}^{N_\lambda} \left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^{m_1}.\end{aligned} \tag{4.62}$$

When considering an homogeneous RP, that is, the contribution of the source term is nil, RH condition across the interface yields $\mathbf{F}_i^- = \mathbf{F}_{i+1}^+$. Such fluxes are now a unique value and are denoted by $\mathbf{F}_{i+1/2}^*$, which can be expressed in terms of the left or right contributions according to (4.62) as follows

$$\begin{aligned}\mathbf{F}_{i+1/2}^* &= \mathbf{F}_i + \sum_{m_1=1}^{I} \left(\widetilde{\lambda}\alpha\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^{m_1} \\ \mathbf{F}_{i+1/2}^* &= \mathbf{F}_{i+1} - \sum_{m_1=I+1}^{N_\lambda} \left(\widetilde{\lambda}\alpha\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^{m_1}.\end{aligned} \tag{4.63}$$

Combination of the expressions in (4.63) leads to

$$\mathbf{F}_{i+1/2}^* = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2}\sum_{m_1=1}^{N_\lambda} \left(\left|\widetilde{\lambda}\right| \alpha\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^{m_1} \tag{4.64}$$

that can be rewritten in matrix form as

$$\mathbf{F}_{i+1/2}^* = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2}\left(\widetilde{\mathbf{P}} \mid \widetilde{\mathbf{\Lambda}} \mid \widetilde{\mathbf{A}}\right)_{i+\frac{1}{2}} \tag{4.65}$$

where

$$|\,\widetilde{\mathbf{\Lambda}}\,|_{i+\frac{1}{2}} = \begin{pmatrix} |\,\widetilde{\lambda}^1\,| & & 0 \\ & \ddots & \\ 0 & & |\,\widetilde{\lambda}^{N_\lambda}\,| \end{pmatrix}_{i+\frac{1}{2}}. \tag{4.66}$$

If defining $|\,\widetilde{\mathbf{J}}\,|_{i+\frac{1}{2}} = \left(\widetilde{\mathbf{P}}\left|\widetilde{\mathbf{\Lambda}}\right|\widetilde{\mathbf{P}}^{-1}\right)_{i+\frac{1}{2}}$, the last term in Equation (4.65) can be rewritten as

$$\left(\widetilde{\mathbf{P}} \mid \widetilde{\mathbf{\Lambda}} \mid \widetilde{\mathbf{A}}\right)_{i+\frac{1}{2}} = \left(\widetilde{\mathbf{P}} \mid \widetilde{\mathbf{\Lambda}} \mid \widetilde{\mathbf{P}}^{-1}\delta\mathbf{U}\right)_{i+\frac{1}{2}} = \left(|\,\widetilde{\mathbf{J}}\,| \,\delta\mathbf{U}\right)_{i+\frac{1}{2}} \tag{4.67}$$

leading to the following intercell homogeneous flux

$$\mathbf{F}_{i+1/2}^* = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2}\left(|\,\widetilde{\mathbf{J}}\,| \,\delta\mathbf{U}\right)_{i+\frac{1}{2}}. \tag{4.68}$$

Analogously, if defining $\delta\mathbf{F}_{i+1/2} = \widetilde{\mathbf{P}}_{i+1/2}\mathbf{\Gamma}_{i+1/2}$, it is straightforward to obtain the following relation

$$\mathbf{\Gamma}_{i+1/2} = \widetilde{\mathbf{\Lambda}}_{i+1/2}\widetilde{\mathbf{A}}_{i+1/2} \tag{4.69}$$

with $\mathbf{\Gamma}_{i+1/2} = (\gamma^1, ..., \gamma^{N_\lambda})_{i+1/2}$, that can be introduced in (4.65) to obtain

$$\mathbf{F}_{i+1/2}^* = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2}\mathrm{sgn}(\widetilde{\mathbf{J}}_{i+\frac{1}{2}})\delta\mathbf{F}_{i+1/2} \tag{4.70}$$

where $\text{sgn}(\widetilde{\mathbf{J}}_{i+\frac{1}{2}}) = \left(\widetilde{\mathbf{P}} \mid \widetilde{\mathbf{\Lambda}} \mid \widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{P}}^{-1}\right)_{i+\frac{1}{2}}$ is the upwinding matrix. The previous equation can be rewritten as follows

$$\mathbf{F}^*_{i+1/2} = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2}\sum_{m_1=1}^{N_\lambda} \left(\text{sgn}(\widetilde{\lambda})\gamma\widetilde{\mathbf{e}}\right)^{m_1}_{i+\frac{1}{2}} \tag{4.71}$$

or, analogously to equation (4.63)

$$\begin{aligned} \mathbf{F}^*_{i+1/2} &= \mathbf{F}_i + \sum_{m_1=1}^{I} (\gamma\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} \\ \mathbf{F}^*_{i+1/2} &= \mathbf{F}_{i+1} - \sum_{m_1=I+1}^{N_\lambda} (\gamma\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}}\,. \end{aligned} \tag{4.72}$$

If considering again the non-homogeneous flux, the corresponding intercell numerical fluxes for the approximate first order Godunov's method are given by

$$\mathbf{F}^-_{i+\frac{1}{2}} = \mathbf{F}^-_i \qquad \mathbf{F}^+_{i-\frac{1}{2}} = \mathbf{F}^+_i \tag{4.73}$$

and updating expression in (3.14) yields

$$\mathbf{U}^{n+1}_i = \mathbf{U}^n_i - (\mathbf{F}^-_i - \mathbf{F}^+_i)\frac{\Delta t}{\Delta x}\,. \tag{4.74}$$

Notice that source term is accounted for in the numerical fluxes and therefore no explicit contribution of the source appears in (4.74) as in (3.14).

4.2.2 Solution using HLLS solver for a 2-wave Riemann Problem

In the framework of first order Godunov's method, Harten, Lax and van Leer introduced a novel Riemann solver [16], called HLL solver. This solver was of application for homogeneous RPs of two waves, providing an estimation of the intercell numerical flux considering a single star region. Such flux is directly computed from the integral form of the governing equations. When dealing with non-homogeneous systems of PDEs, the HLL solver can not be used. A proper treatment of source terms in the framework of the HLL solver was proposed by Murillo in [22] with the generation of a new solver, called HLLS, that considers the presence of an additional stationary wave at $x = 0$. In this section, the HLLS solver presented in [22] is revisited.

Let us consider the original RP (3.18) with $\mathbf{U}(x,t) \in \mathscr{C} \subseteq \mathbb{R}^2$ and $\mathbf{F}(\mathbf{U}) : \mathscr{C} \to \mathbb{R}^2$, that is, (5.31) is a system of two equations characterized by two real eigenvalues $\lambda^1(\mathbf{U}) \leq \lambda^2(\mathbf{U})$ corresponding to the wave speeds, plus an extra wave of speed $\mathscr{S} = 0$ at $x = 0$ due to the presence of the source term. The complete wave structure of RP in (3.18) is depicted in Figure 4.4, noticing four constant states.

The integral form of (3.18) inside a control volume $[-x_L, x_R] \times [0, \Delta t]$ was detailed in (3.41). Recall that the expression for the integral volume of $\mathbf{U}(x, \Delta t)$ is expressed as

$$\int_{-x_L}^{x_R} \mathbf{U}(x,\Delta t)\,dx = x_R\mathbf{U}_{i+1} + x_L\mathbf{U}_i + (\mathbf{F}_i - \mathbf{F}_{i+1})\Delta t + \bar{\mathbf{S}}_{i+\frac{1}{2}}\Delta t \tag{4.75}$$

with $\mathbf{F}_{i+1} = \mathbf{F}(\mathbf{U}_{i+1})$ and $\mathbf{F}_i = \mathbf{F}(\mathbf{U}_i)$ and the source term integrated as in (3.42). The integral on the left hand side of (4.75) can be split considering a wave structure given by $\lambda^1 \leq 0 \leq \lambda^2$ as depicted in Figure 4.4 and with $-x_L < \lambda^1\Delta t$ and $x_R > \lambda^2\Delta t$

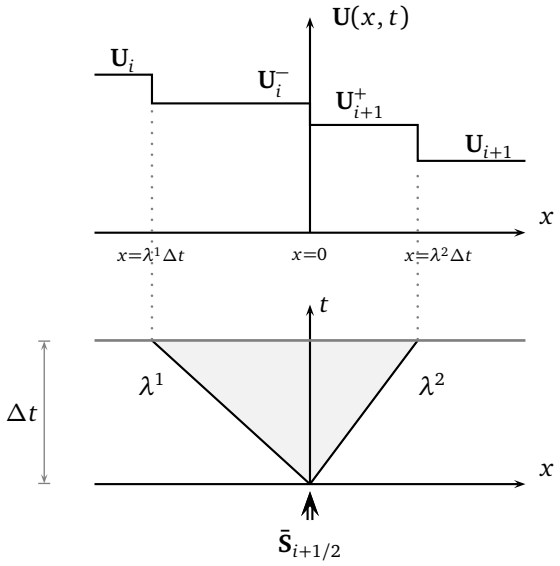


Figure 4.4: Values of the solution $\mathbf{U}(x,t)$ in each wedge of the (x,t) plane.

$$\begin{aligned}\int_{-x_L}^{x_R}\mathbf{U}(x,\Delta t)\,dx = &\int_{-x_L}^{\lambda^1\Delta t}\mathbf{U}(x,\Delta t)\,dx+\int_{\lambda^1\Delta t}^{0}\mathbf{U}(x,\Delta t)\,dx\\ &+\int_{0}^{\lambda^2\Delta t}\mathbf{U}(x,\Delta t)\,dx+\int_{\lambda^2\Delta t}^{x_R}\mathbf{U}(x,\Delta t)\,dx\end{aligned}\tag{4.76}$$

and considering the solution composed of four constant states as shown in Figure 4.4, it yields

$$\int_{-x_L}^{x_R}\mathbf{U}(x,\Delta t)\,dx=\mathbf{U}_i(\lambda^1\Delta t+x_L)+\mathbf{U}_{i+1}(x_L-\lambda^2\Delta t)+\mathbf{U}_i^-(-\lambda^1\Delta t)+\mathbf{U}_{i+1}^+(\lambda^2\Delta t).\tag{4.77}$$

Now, substitution of (4.77) in (4.75) leads to

$$(\mathbf{U}_i-\mathbf{U}_i^-)\lambda^1-(\mathbf{U}_{i+1}-\mathbf{U}_{i+1}^+)\lambda^2+\mathbf{F}_{i+1}-\mathbf{F}_i=\bar{\mathbf{S}}_{i+\frac{1}{2}}\tag{4.78}$$

where an extra condition is needed in order to obtain an expression for $\mathbf{U}_i^-$ and $\mathbf{U}_{i+1}^+$, due to the presence of the source term. For that purpose, let us define first an approximate flux function $\hat{\mathbf{F}}(x,t)$ with a similar structure than $\mathbf{U}(x,t)$ as depicted in Figure 4.4. In this case, also intercell values for the fluxes can be defined at both sides of the t axis as

$$\mathbf{F}_i^-=\lim_{x\to 0^-}\hat{\mathbf{F}}(x,t)\qquad\mathbf{F}_{i+1}^+=\lim_{x\to 0^+}\hat{\mathbf{F}}(x,t).\tag{4.79}$$

The following RH relations across waves between fluxes and conserved variables are stated

$$\mathbf{F}_i^--\mathbf{F}_i=\lambda^1(\mathbf{U}_i^--\mathbf{U}_i)\tag{4.80}$$

$$\mathbf{F}_{i+1}-\mathbf{F}_{i+1}^+=\lambda^2(\mathbf{U}_{i+1}-\mathbf{U}_{i+1}^+)\tag{4.81}$$

$$\mathbf{F}^+_{i+1} - \mathbf{F}^-_i - \bar{\mathbf{S}}_{i+\frac{1}{2}} = \mathscr{S}(\mathbf{U}^+_{i+1} - \mathbf{U}^-_i) = 0\,. \tag{4.82}$$

Moreover, using Roe's approach it is possible to define the following relation as done in Section 4.2.1

$$\mathbf{F}^+_{i+1} - \mathbf{F}^-_i = \widetilde{\mathbf{J}}_{i+\frac{1}{2}} \left(\mathbf{U}^+_{i+1} - \mathbf{U}^-_i\right). \tag{4.83}$$

where $\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}}(\mathbf{U}_i, \mathbf{U}_{i+1})$ is an approximation of the Jacobian matrix according to Equations (4.19) - (4.24). Combining (4.82) and (4.83), the following relation appears

$$\bar{\mathbf{S}}_{i+\frac{1}{2}} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}} \left(\mathbf{U}^+_{i+1} - \mathbf{U}^-_i\right) \tag{4.84}$$

and using (4.24), it is possible to write the jump on the conserved variables across the stationary wave at $x = 0$ as

$$\mathbf{U}^+_{i+1} - \mathbf{U}^-_i = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}} \bar{\mathbf{S}}_{i+\frac{1}{2}} = \bar{\mathbf{H}}_{i+\frac{1}{2}}\,. \tag{4.85}$$

Combination of (4.78) and (4.85) leads to the following values for the intermediate states

$$\mathbf{U}^-_i = \frac{\mathbf{F}_i - \mathbf{F}_{i+1} + \lambda^2 \mathbf{U}_{i+1} - \lambda^1 \mathbf{U}_i + \bar{\mathbf{S}}_{i+\frac{1}{2}} - \lambda^2 \bar{\mathbf{H}}_{i+\frac{1}{2}}}{\lambda^2 - \lambda^1}\,, \tag{4.86}$$

$$\mathbf{U}^+_{i+1} = \frac{\mathbf{F}_i - \mathbf{F}_{i+1} + \lambda^2 \mathbf{U}_{i+1} - \lambda^1 \mathbf{U}_i + \bar{\mathbf{S}}_{i+\frac{1}{2}} - \lambda^1 \bar{\mathbf{H}}_{i+\frac{1}{2}}}{\lambda^2 - \lambda^1}\,. \tag{4.87}$$

Remark that when the contribution of the source term is nil, relation in (4.85) shows that there exists only a unique intermediate state in the so-called star region. This state can be derived either from (4.86) or (4.87) indistinctly as

$$\mathbf{U}^* = \frac{\mathbf{F}_i - \mathbf{F}_{i+1} + \lambda^2 \mathbf{U}_{i+1} - \lambda^1 \mathbf{U}_i}{\lambda^2 - \lambda^1}\,. \tag{4.88}$$

Expressions for left and right intercell fluxes can be straightforward derived from (4.86) and (4.87) by applying the corresponding RH condition

$$\mathbf{F}^-_i = \frac{\lambda^2 \mathbf{F}_i - \lambda^1 \mathbf{F}_{i+1} + \lambda^1 \lambda^2 (\mathbf{U}_{i+1} - \mathbf{U}_i) + \lambda^1 \left(\bar{\mathbf{S}}_{i+\frac{1}{2}} - \lambda^2 \bar{\mathbf{H}}_{i+\frac{1}{2}}\right)}{\lambda^2 - \lambda^1}\,, \tag{4.89}$$

$$\mathbf{F}^+_{i+1} = \frac{\lambda^2 \mathbf{F}_i - \lambda^1 \mathbf{F}_{i+1} + \lambda^1 \lambda^2 (\mathbf{U}_{i+1} - \mathbf{U}_i) + \lambda^2 \left(\bar{\mathbf{S}}_{i+\frac{1}{2}} - \lambda^1 \bar{\mathbf{H}}_{i+\frac{1}{2}}\right)}{\lambda^2 - \lambda^1}\,. \tag{4.90}$$

Remark that the previous derivation was carried out under the assumption of $\lambda^1 \leq 0 \leq \lambda^2$, that is, a subcritical wave structure. When having supercritical cases, where $\lambda^1, \lambda^2 \geq 0$ or $\lambda^1, \lambda^2 \leq 0$, expressions in (4.89) and (4.90) can not be used. A general expression for the numerical fluxes can be provided

$$\mathbf{F}^-_{i+\frac{1}{2}} = \begin{cases} \mathbf{F}_i & \text{if} \quad \lambda^1 \geq 0 \\ \mathbf{F}^-_i & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \mathbf{F}_{i+1} - \bar{\mathbf{S}}_{i+\frac{1}{2}} & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{4.91}$$

$$\mathbf{F}^+_{i+\frac{1}{2}} = \begin{cases} \mathbf{F}_i \quad + \bar{\mathbf{S}}_{i+\frac{1}{2}} & \text{if} \quad \lambda^1 \geq 0 \\ \mathbf{F}^+_{i+1} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \mathbf{F}_{i+1} & \text{if} \quad \lambda^2 \leq 0 \end{cases}. \tag{4.92}$$

4.3 Concluding remarks

It is worth emphasizing the following issues addressed in this chapter:

- The augmented solver in Section 4.1, which correctly accounts for the presence of the source term at the interface, is revisited here. This solver provides the foundation for most numerical methods herein described.
- Two augmented solvers, the ARoe and the HLLS solvers, are revisited. Such solvers are later used to construct the novel Riemann solvers described in this thesis. Whereas the ARoe solver is a complete solver and contains the full wave structure, the HLLS solver is designed for systems of equations with 2 waves and geometric source term. The ARoe solver assumes a linearized approximation via discontinuous jumps, which encounters difficulties in the resolution of sonic or transonic rarefaction waves. Such difficulties appear in the form of unphysical, entropy violating shocks, which are far from the expected rarefaction wave. To address this problem, the so-called entropy correction is required. On the other hand, the HLLS solver provides a suitable representation of such kind of waves without extra corrections.
- Both, the ARoe and HLLS solvers are designed to satisfy the RH condition at the interface, leading to the fundamental relation

$$\mathbf{F}_{i+1}^{+} - \mathbf{F}_{i}^{-} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}} \left(\mathbf{U}_{i+1}^{+} - \mathbf{U}_{i}^{-} \right) . \tag{4.93}$$

- The mathematical expression of the approximate solution provided by the ARoe solver is formally derived here for the first time, although it was originally presented in [20].

5 INTRODUCTION TO ADER FINITE VOLUME SCHEMES

In this chapter, the fundamentals of ADER-FV numerical schemes are briefly recalled. In Section 5.1, ADER schemes are introduced and compared with the traditional first order Godunov's scheme. In the Section 5.2, the arbitrary order data reconstruction procedure used in this work, the WENO method, is detailed. The traditional WENO-JS procedure and other improved techniques, designed to circumvent some flaws of the original method, are briefly recalled. In Section 5.3, the mathematical formulation of ADER schemes is provided and the difference between the flux-expansion and state-expansion ADER methodology is explained. Finally, in Section 5.4, the DRP is introduced and the mathematical tools necessary for its resolution are provided. Among such tools, we put special emphasis on the derivation of the evolution equations for the derivatives of the problem and also on the procedure for the computation of time derivatives in terms of space derivatives, the CK procedure.

5.1 Introduction

ADER schemes were originally presented in [28, 29] and can be regarded as the natural extension to arbitrary order of the first order Godunov's scheme. They are explicit, fully-discrete and provide an arbitrary order of accuracy both in space and time. ADER schemes are composed of two steps: the first step is the piecewise reconstruction of the variables inside the cells using very high order procedures while the second step is the computation and time integration of the fluxes and sources, eventually updating the conserved variables. To compute the fluxes, a high order extension of the RP, called DRP, is required. ADER schemes successfully allow the construction of arbitrary order schemes for systems of hyperbolic conservation laws [30, 39, 40, 27, 42, 47, 48, 49, 50, 51, 53].

Unlike Godunov's first order scheme, where the information is represented by piecewise constant data inside cells, ADER schemes consider piecewise polynomial data of a degree matching the prescribed accuracy. Figure 5.1 shows the typical discretization of a 1D domain, including a representation of the cell averages and piecewise varying data. Subscripts L and R are defined with reference to the cell center. For the sake of clarity in further derivations, we introduce the notation

$$x_{i_R} = \lim_{x \to x^-_{i+1/2}} x\,, \qquad x_{(i+1)_L} = \lim_{x \to x^+_{i+1/2}} x\,. \tag{5.1}$$

For the spatial reconstruction, an arbitrary order reconstruction of the initial data which matches the order of the scheme is sought. One possibility is to reconstruct polynomials using a fixed stencil, which is called *linear reconstruction*. Such approach can make the solution become oscillatory, according to Go-

dunov's theorem. A radically different approach is to reconstruct polynomials on *variable* or *adaptive stencils*, rather than fixed stencils. Such reconstructions is called *non-linear reconstructions*. Among the most used reconstruction techniques, Total Variation Diminishing (TVD) schemes [12] and specially the ENO and WENO methods [24, 25] supposed a major step when seeking high order in space in the framework of FV schemes. On the other hand, the preservation of high order in time is based on the construction of Taylor power series expansions in time, where time derivatives are computed by means of the CK procedure.

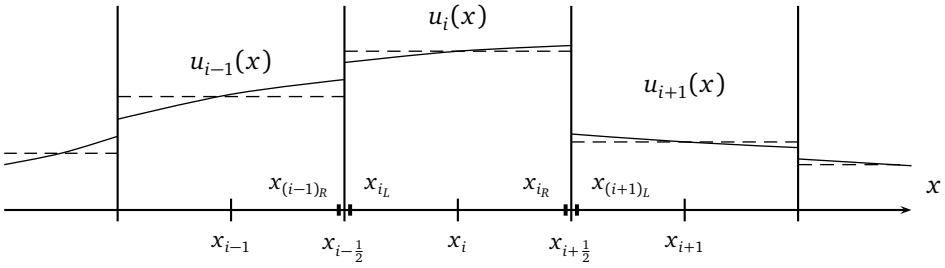


Figure 5.1: Mesh discretization, data reconstruction and notation in ADER schemes.

5.2 Non-oscillatory reconstruction procedures: the WENO method

In this work, we use the so-called weighted essentially non-oscillatory (WENO) reconstruction method [24, 25]. The WENO method uses a variable set of stencils where lower order polynomials are first constructed. Then, these lower order polynomials are combined either to create a higher order polynomial in smooth regions (optimal reconstruction) or an off-center reconstruction able to capture discontinuities in non-smooth regions. The definition of a smoothness indicator allows to distinguish between those two cases.

The traditional WENO reconstruction method, also referred to as WENO-JS, is briefly explained here. All variables are reduced to the scalar case for the sake of clarity. More details about the WENO reconstruction procedure, sub-cell WENO derivative reconstruction procedure and their extension to 2D can be found in Appendices A, B, C and D.

The departing data for the WENO reconstruction procedure are the cell averaged values of a function $u(x)$ defined in a computational grid composed of N cells, with cells and cell sizes defined by

$$\Omega_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \qquad \Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} \equiv \text{constant} \tag{5.2}$$

and cell averages of $u(x)$ are defined in the following way

$$u_i = \frac{1}{\Delta x_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(\xi)\, d\xi, \quad i = 1, 2, ..., N\,. \tag{5.3}$$

To construct a WENO reconstruction of degree $(2k-1)$ on the cell $\Omega_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ for the function $u(x)$, k different stencils linked to k cells are needed. These stencils are given by $S_r(i) = \{\Omega_{i-r}, ..., \Omega_{i+k-r-1}\}$ $(r = 0, ..., k-1)$, where r represents the number of cells on the left hand side of Ω_i. These stencils are used to generate a bigger stencil $\mathcal{T}(i) = \cup_{r=0}^{k-1} S_r(i) = \{\Omega_{i-k+1}, ..., \Omega_{i+k-1}\}$. The general procedure of the WENO reconstruction is summarized below:

a) *Definition of the optimal weights*

Following [26], there is a unique polynomial $p_r(x)$ defined in each stencil S_r, which is a k-th order approximation of the function $u(x)$ on the stencil $S_r(i)$ if this function is smooth inside it. The ex-

pression of $p_r(x)$ is expressed as a linear combination of the cell averages in the stencil. At x_{i_R} (see Figure 5.1), the approximation of $u(x_{i_R})$ is given by

$$u_{i_R}^{(r)} = p_r(x_{i+\frac{1}{2}}) = \sum_{j=0}^{k-1} c_{rj}^{(k)} u_{i-r+j} = u\left(x_{i_R}\right) + O\left(\Delta x^k\right), \tag{5.4}$$

where $c_{rj}^{(k)}$ are coefficients derived from the Lagrange interpolation formula. The same procedure can be used to obtain a polynomial $q(x)$, which is a $(2k-1)$-th order approximation of the function $u(x)$ on the big stencil $\mathscr{T}(i)$. At x_{i_R}, this approximation of $u(x_{i_R})$ is given by

$$u_{i_R} = q(x_{i+\frac{1}{2}}) = \sum_{j=1}^{2k-1} c_{k-1,j}^{(2k-1)} u_{i-k+j} = u\left(x_{i_R}\right) + O\left(\Delta x^{2k-1}\right). \tag{5.5}$$

Note that in (5.5), the value of r is fixed, with $r = k-1$, as the big stencil $\mathscr{T}(i)$ is symmetric. The $(2k-1)$-th order approximation in (5.5) can also be expressed as a linear convex combination of the k-th order reconstructions provided by (5.4) as

$$u_{i_R} = \sum_{r=0}^{k-1} \gamma_r u_{i_R}^{(r)} = u\left(x_{i_R}\right) + O\left(\Delta x^{2k-1}\right), \tag{5.6}$$

where γ_r are the optimal weights that can be easily computed relating $c_{k-1,j}^{(2k-1)}$ and $c_{rj}^{(k)}$ [26]. In the linear combination in (5.6) the optimal weights are calculated algebraically.

b) *Definition of the smoothness indicator: smoothness indicator for WENO-JS*

The so called smoothness indicator, β_r, which measure the smoothness of the initial data, is able to detect the presence of discontinuities. In the case of the traditional WENO reconstruction, called WENO-JS and proposed by Jiang and Shu [26], this indicator reads

$$\beta_r = \sum_{l=1}^{k-1} \int_{x_{i+\frac{1}{2}}}^{x_{i-\frac{1}{2}}} \Delta x^{2l-1} \left(\frac{\partial^l p_r(x)}{\partial x^l} \right)^2 dx, \qquad r = 0, \ldots, k-1 \tag{5.7}$$

and represents the variations in u inside the small stencils.

c) *Definition of the WENO-JS weights*

Departing from the optimal weights, it is possible to define the WENO-JS weights, denoted by ω_r^{JS}, that satisfy

$$\sum_{r=0}^{k-1} \omega_r^{JS} = 1, \quad \omega_r^{JS} \geq 0. \tag{5.8}$$

They generate a convex combination of the low order reconstructions to compute the final approximation. First the α_r^{JS} coefficients are formulated and then normalized leading to the WENO ω_r^{JS} weights

$$\alpha_r^{JS} = \frac{\gamma_r}{(\beta_r + \epsilon)^2} \qquad \omega_r^{JS} = \frac{\alpha_r}{\sum_{l=0}^{k-1} \alpha_l}, \qquad r = 0, \ldots, k-1 \tag{5.9}$$

with ϵ a properly defined small parameter. The final WENO approximation of $u(x)$ at $x_{i+\frac{1}{2}}$ is given by

$$u_{i_R} = \sum_{r=0}^{k-1} \omega_r^{JS} u_{i_R}^{(r)} = u\left(x_{i_R}\right) + O\left(\Delta x^{2k-1}\right). \tag{5.10}$$

When analyzing the performance of the WENO methods to preserve the order of accuracy in smooth regions, it is reported in the literature that the reconstruction becomes suboptimal in presence of critical points, where spatial derivatives vanish. This is due to a non-accurate recovery of the optimal weights around critical points. Some effort has been made to overcome the undesirable behavior of the loss of accuracy of the WENO-JS method around critical points [116, 117] in high order finite difference numerical schemes. As a result, some improvements in the calculation of the smoothness indicator in [118] and in the WENO-JS weights have been proposed.

In order to recover a formal 5-th order of convergence around critical points, in [116] a mapping function was presented leading to the WENO-5M scheme. In the same work, it was shown how the loss of formal convergence at the critical points, when using the WENO-JS method, was hidden due to the definition of tuning parameters to avoid division by zero. Improvements of the 5-th order WENO-JS and WENO-5M methods resulted in the WENO-Z scheme, proposed in [117]. WENO-Z scheme involves a global indicator based on a suitable combination of different evaluations of the smoothness indicator. The resulting scheme was extended to arbitrary order of accuracy in [119]. Although both the WENO-5M and the WENO-Z schemes ensure the formal order of convergence, the WENO-Z scheme provides more accurate results around shocks. A more recent approach introduced by S. Zhao et al. [120] proposes the combination of the WENO-5M and WENO-Z, leading to the so called WENO-MZ method. It is based on the mapping of the non-oscillatory weights provided by the WENO-Z method. In [120], the authors compare the utilization of a 5-th order WENO-MZ with a 5-th and 7-th order WENO-JS. Another novel technique, called the WENO-NS method, was recently introduced in [121]. This new approach provides a 5-th order of convergence in smooth regions, especially at critical points where only the first derivative vanishes. For one dimensional schemes, decisive advantages when compared to the WENO-5M and WENO-Z schemes were not found [120].

Some of the aforementioned approaches are detailed and exercised in Appendix A.3. The influence of the particular WENO reconstruction procedure on the reconstruction of spatial derivatives is shown in Appendix B. In Section 8, the WENO-JS and WENO-Z are compared for the resolution of the linear scalar advection equation in 1 and 2 space dimensions.

5.3 Fundamentals of ADER-type numerical schemes

Following the approach proposed by Godunov, a suitable arbitrary-order discretization of the system in (5.31) inside $\Omega_i \times [t^n, t^{n+1}]$ can be expressed as

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{\Delta x}[\mathbf{F}_{i+1/2}^- - \mathbf{F}_{i-1/2}^+] + \frac{\Delta t}{\Delta x}[\bar{\mathbf{S}}_{i_R,i_L}], \tag{5.11}$$

with the numerical fluxes $\mathbf{F}_{i+1/2}^-$ and $\mathbf{F}_{i-1/2}^+$ defined as time-integral averages of the time-dependent fluxes evolved in time at the interfaces

$$\mathbf{F}_{i+1/2}^- = \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}_{i_R}^-(\tau)\, d\tau\,, \qquad \mathbf{F}_{i-1/2}^+ = \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}_{i_L}^+(\tau)\, d\tau \tag{5.12}$$

and $\bar{\mathbf{S}}_{i_R,i_L}$ a suitable approximation of the integral of the source term inside the cell given by

$$\bar{\mathbf{S}}_{i_R,i_L} \approx \frac{1}{\Delta t}\int_0^{\Delta t}\int_{x_{i_L}}^{x_{i_R}} \mathbf{S}\; dx\; d\tau\,. \tag{5.13}$$

Analogously, Equation (5.11) can be rewritten in terms of fluctuations, generally denoted by $\delta\mathbf{M}$, leading to

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{\Delta x}[\delta\mathbf{M}^-_{i+1/2} + \delta\mathbf{M}_{i_R,i_L} + \delta\mathbf{M}^+_{i-1/2}], \tag{5.14}$$

where

$$\begin{aligned}
\delta\mathbf{M}^-_{i+1/2} &= \mathbf{F}^-_{i+1/2} - \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}_{i_R}(\tau)\,d\tau\,,\\
\delta\mathbf{M}_{i_R,i_L} &= \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}_{i_R}(\tau)\,d\tau - \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}_{i_L}(\tau)\,d\tau - \bar{\mathbf{S}}_{i_R,i_L}\,,\\
\delta\mathbf{M}^+_{i-1/2} &= \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}_{i_L}(\tau)\,d\tau - \mathbf{F}^+_{i-1/2}\,,
\end{aligned} \tag{5.15}$$

represent the contribution of the incoming waves to the right edge, the contribution due to the variation of the physical flux and source along the cell and the contribution of the incoming waves to the left edge, respectively. It is worth mentioning that fluctuation $\delta\mathbf{M}_{i_R,i_L}$ can be divided in as many terms as desired by making use of the telescopic property. For an arbitrary number of contributions, n_m, the centered fluctuation reads

$$\delta\mathbf{M}_{i_R,i_L} = \delta\mathbf{M}_{i_1,i_L} + \sum_{j=2}^{n_m-1} \delta\mathbf{M}_{i_j,i_{j-1}} + \delta\mathbf{M}_{i_R,i_{n_m-1}}\,, \tag{5.16}$$

requiring thus, extra inner cell information that may be provided by reconstruction procedures.

When the scheme in (5.14) is reduced to first order of accuracy, it is worth saying that the centered fluctuation, $\delta\mathbf{M}_{i_R,i_L}$, will be reduced to $\delta\mathbf{M}_{i_R,i_L} = \bar{\mathbf{S}}_{i_R,i_L}$ since the flux is constant inside the cell. In the case of using augmented Riemann solvers and a first order of accuracy, we must set $\delta\mathbf{M}_{i_R,i_L} = 0$ since the source term will be already accounted for in the weak solution of the RPs at the interfaces provided by the solver.

Physical fluxes at left and right cell edges, $\mathbf{F}_{i_L}(\tau)$ and $\mathbf{F}_{i_R}(\tau)$, can be approximated by a Taylor power series expansion in time as

$$\mathbf{F}_{i_L}(\tau) = \mathbf{F}^{(0)}_{i_L} + \sum_{k=1}^{K} \mathbf{R}^{(k)}_{i_L}\frac{\tau^k}{k!}\,, \qquad \mathbf{F}_{i_R}(\tau) = \mathbf{F}^{(0)}_{i_R} + \sum_{k=1}^{K} \mathbf{R}^{(k)}_{i_R}\frac{\tau^k}{k!}\,, \tag{5.17}$$

with $\mathbf{F}^{(0)}_{i_L}$ and $\mathbf{F}^{(0)}_{i_R}$ the so-called leading terms, computed as in (5.34) and with $\mathbf{R}^{(k)}_{i_L}$ and $\mathbf{R}^{(k)}_{i_R}$ the coefficients of the high order terms, computed as in (5.46).

As done for the fluxes, the source term inside cell Ω_i can also be approximated by a truncated Taylor power series expansion in time

$$\mathbf{S}_i(x,\tau) = \mathbf{S}_i(x,0) + \sum_{k=1}^{K}\left[\frac{\partial^k \mathbf{S}_i}{\partial t^k}\right]_{x,t=0}\frac{\tau^k}{k!}\,, \tag{5.18}$$

leading to the following expression for its integral inside $\Omega_i \times [0,\Delta t]$

$$\bar{\mathbf{S}}_{i_R,i_L} = \bar{\mathbf{S}}^{(0)}_{i_R,i_L} + \sum_{k=1}^{K} \bar{\mathbf{S}}^{(k)}_{i_R,i_L}\frac{\Delta t^k}{(k+1)!}\,, \tag{5.19}$$

with

$$\bar{\mathbf{S}}^{(0)}_{i_R,i_L} = \int_{x_{i_L}}^{x_{i_R}} \mathbf{S}_i(x,0)\, dx\,,$$
$$\bar{\mathbf{S}}^{(k)}_{i_R,i_L} = \int_{x_{i_L}}^{x_{i_R}} \mathbf{Q}^{(k)}(x,0)\, dx\,, \tag{5.20}$$

that will be integrated by means of approximated quadrature rules.

As outlined before, when dealing with geometric source terms of the type of (3.15), the contribution of the source is not only accounted for inside the cell but also at cell interfaces [20, 83] so that the scheme converges to the exact solution. In this case, the integral of the source and its derivatives has to be calculated at cell interfaces as

$$\bar{\mathbf{S}}_{i+1/2} = \bar{\mathbf{S}}^{(0)}_{i+1/2} + \sum_{k=1}^{K} \bar{\mathbf{S}}^{(k)}_{i+1/2} \frac{\Delta t^k}{(k+1)!}\,, \tag{5.21}$$

where

$$\bar{\mathbf{S}}^{(0)}_{i+1/2} = \int_{x^-_{i+1/2}}^{x^+_{i+1/2}} \mathbf{S}(x,0)\, dx\,,$$
$$\bar{\mathbf{S}}^{(k)}_{i+1/2} = \int_{x^-_{i+1/2}}^{x^+_{i+1/2}} \mathbf{Q}^{(k)}(x,0) dx\,, \tag{5.22}$$

that will be integrated using suitable approximations.

Two different approaches can be used to compute left and right intercell numerical fluxes $\mathbf{F}^-_{i_R}$ and $\mathbf{F}^+_{i_L}$ in (5.12). The first approach is called *state-expansion ADER* and proposes to obtain the solution for conserved variables at the interface by solving the DRP_K with a suitable solver and to evaluate the physical fluxes using this solution. The second option is to use the *flux-expansion ADER* approach, where fluxes $\mathbf{F}^-_{i_R}$ and $\mathbf{F}^+_{i_L}$ are constructed as a truncated power series expansion in time and the components of the expansion are functions of the approximate fluxes defined for each RP associated to the DRP_K.

a) **State-expansion ADER approach.** The numerical fluxes in (5.12) are evaluated using the time-dependent solutions of the DRP_K, $\mathbf{U}^-_{i_R}(\tau)$ and $\mathbf{U}^+_{(i+1)_L}(\tau)$, as

$$\mathbf{F}^-_{i+1/2} = \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}(\mathbf{U}^-_{i_R}(\tau))d\tau \qquad \mathbf{F}^+_{i+1/2} = \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{F}(\mathbf{U}^+_{(i+1)_L}(\tau))d\tau \tag{5.23}$$

with $\mathbf{U}^-_{i_R}(\tau)$ and $\mathbf{U}^+_{(i+1)_L}(\tau)$ computed using a suitable arbitrary-order Riemann solver.

b) **Flux-expansion ADER approach.** When adopting the flux-expansion ADER approach, we seek a truncated Taylor time expansion of the fluxes at the interfaces as

$$\mathbf{F}^-_{i_R}(\tau) = \mathbf{F}^{-,(0)}_{i_R} + \sum_{k=1}^{K} \mathbf{F}^{-,(k)}_{i_R} \frac{\tau^k}{k!} \qquad \mathbf{F}^+_{(i+1)_L}(\tau) = \mathbf{F}^{+,(0)}_{(i+1)_L} + \sum_{k=1}^{K} \mathbf{F}^{+,(k)}_{(i+1)_L} \frac{\tau^k}{k!} \tag{5.24}$$

that after integration, leads to the following expression of the numerical fluxes

$$\mathbf{F}^-_{i+1/2} = \mathbf{F}^{-,(0)}_{i_R} + \sum_{k=1}^{K} \mathbf{F}^{-,(k)}_{i_R} \frac{\Delta t^k}{(k+1)!}\,, \qquad \mathbf{F}^+_{i+1/2} = \mathbf{F}^{+,(0)}_{(i+1)_L} + \sum_{k=1}^{K} \mathbf{F}^{+,(k)}_{(i+1)_L} \frac{\Delta t^k}{(k+1)!}\,. \tag{5.25}$$

The coefficients of the expansion in (5.24)-(5.25) will be computed using suitable Riemann solvers. When the flux-expansion ADER approach is used, it is possible to rewrite fluctuations in (5.15) more compactly as

$$\delta \mathbf{M}_{i_R,i_L} = \delta \mathbf{M}^{(0)}_{i_R,i_L} + \sum_{k=1}^{K} \delta \mathbf{M}^{(k)}_{i_R,i_L} \tag{5.26}$$

for the centered contribution, with

$$\begin{aligned} \delta \mathbf{M}^{(0)}_{i_R,i_L} &= \mathbf{F}^{(0)}_{i_R} - \mathbf{F}^{(0)}_{i_L} - \bar{\mathbf{S}}^{(0)}_{i_R,i_L} \\ \delta \mathbf{M}^{(k)}_{i_R,i_L} &= \left(\mathbf{R}^{(k)}_{i_R} - \mathbf{R}^{(k)}_{i_L} - \bar{\mathbf{S}}^{(k)}_{i_R,i_L} \right) \frac{\Delta t^k}{(k+1)!} \end{aligned} \tag{5.27}$$

and

$$\begin{aligned} \delta \mathbf{M}^{-}_{i+1/2} &= \delta \mathbf{M}^{-,(0)}_{i+1/2} + \sum_{k=1}^{K} \delta \mathbf{M}^{-,(k)}_{i+1/2} \\ \delta \mathbf{M}^{+}_{i-1/2} &= \delta \mathbf{M}^{+,(0)}_{i-1/2} + \sum_{k=1}^{K} \delta \mathbf{M}^{+,(k)}_{i-1/2} \end{aligned} \tag{5.28}$$

for the contributions at the interfaces, with

$$\begin{aligned} \delta \mathbf{M}^{-,(0)}_{i+1/2} &= \mathbf{F}^{-,(0)}_{i_R} - \mathbf{F}^{(0)}_{i_R} \\ \delta \mathbf{M}^{-,(k)}_{i+1/2} &= \left(\mathbf{F}^{-,(k)}_{i_R} - \mathbf{R}^{(k)}_{i_R} \right) \frac{\Delta t^k}{(k+1)!} \end{aligned} \tag{5.29}$$

and

$$\begin{aligned} \delta \mathbf{M}^{+,(0)}_{i-1/2} &= \mathbf{F}^{+,(0)}_{i_L} - \mathbf{F}^{(0)}_{i_L} \\ \delta \mathbf{M}^{+,(k)}_{i-1/2} &= \left(\mathbf{F}^{+,(k)}_{i_L} - \mathbf{R}^{(k)}_{i_L} \right) \frac{\Delta t^k}{(k+1)!} \end{aligned} \tag{5.30}$$

5.4 The Derivative Riemann Problem

In Section 3.3 the Riemann Problem was described as an IVP whose initial condition is given by two piecewise constant states. This classic RP may be regarded as a first order approach to a general Cauchy problem with a discontinuity at $x = 0$. A second order approach (with piecewise linear data) to the Cauchy problem was introduced by Ben-Artzi and Falcovitz [122] and termed by them as Generalized Riemann Problem (GRP). More generally, an arbitrary order approach to the Cauchy problem is given by the Derivative Riemann Problem (DRP), originally studied by Toro and Titarev in [30]. In the DRP, the initial condition consists of piecewise polynomial data with K nontrivial derivatives, separated by a single discontinuity at $x = 0$ as

$$\begin{cases} \dfrac{\partial \mathbf{U}}{\partial t} + \dfrac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{S} \\ \mathbf{U}(x,0) = \begin{cases} \mathbf{U}_i(x) & x < 0 \\ \mathbf{U}_{i+1}(x) & x > 0 \end{cases} \end{cases} \tag{5.31}$$

where the initial states $\mathbf{U}_i(x)$ and $\mathbf{U}_{i+1}(x)$ are smooth functions of distance x that can be defined using suitable reconstruction procedures at the initial time. Recall that x stands for the local spatial coordinate, centered at $x_{i+1/2}$. Theoretical aspects regarding the DRP can be found in [123]. DRP in (5.31) is depicted in Figure 5.2 for the case when $N_\lambda = 2$.

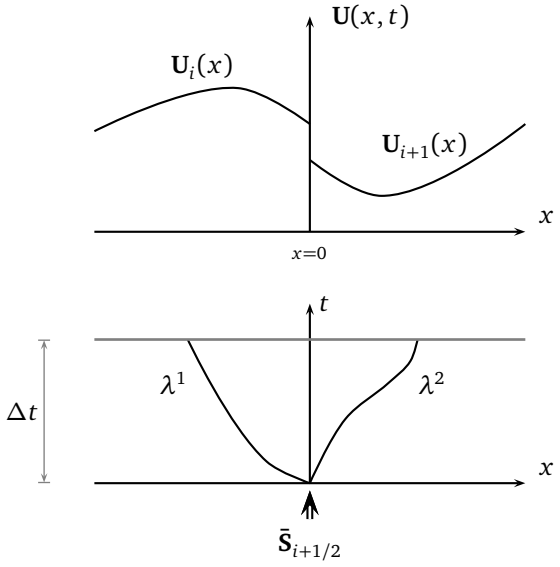


Figure 5.2: Graphical representation of the DRP_K showing the piecewise smooth states (upper figure) and wave velocities that depend upon time (lower figure).

For DRP in (5.31), it is possible to define the following values for vector $\mathbf{U}$ at the interface

$$\mathbf{U}_{i_R}^{(0)} = \lim_{x \to 0^-} \mathbf{U}_i(x), \qquad \mathbf{U}_{(i+1)_L}^{(0)} = \lim_{x \to 0^+} \mathbf{U}_{i+1}(x) \tag{5.32}$$

and for its derivatives

$$\mathbf{U}_{i_R}^{(k)} = \lim_{x \to 0^-} \frac{\partial^k}{\partial x^k} \mathbf{U}_i(x), \qquad \mathbf{U}_{(i+1)_L}^{(k)} = \lim_{x \to 0^+} \frac{\partial^k}{\partial x^k} \mathbf{U}_{i+1}(x), \tag{5.33}$$

at the initial time, with $k = 1, ..., K$.

Analogously, it is possible to define the following values for the physical fluxes $\mathbf{F}(\mathbf{U})$ at the interface

$$\mathbf{F}_{i_R}^{(0)} = \lim_{x \to 0^-} \mathbf{F}(\mathbf{U}_i(x)), \qquad \mathbf{F}_{(i+1)_L}^{(0)} = \lim_{x \to 0^+} \mathbf{F}(\mathbf{U}_{i+1}(x)) \tag{5.34}$$

and for their spatial derivatives

$$\mathbf{F}_{i_R}^{(k)} = \lim_{x \to 0^-} \frac{\partial^k}{\partial x^k} \mathbf{F}(\mathbf{U}_i(x)), \qquad \mathbf{F}_{(i+1)_L}^{(k)} = \lim_{x \to 0^+} \frac{\partial^k}{\partial x^k} \mathbf{F}(\mathbf{U}_{i+1}(x)), \tag{5.35}$$

at the initial time, with $k = 1, ..., K$.

The spatial reconstruction of the source term $\mathbf{S}(\mathbf{U},x,t)$ will be denoted in the same way

$$\mathbf{S}^{(0)}_{i_R} = \lim_{x\to 0^-} \mathbf{S}(\mathbf{U}_i(x),x,0), \qquad \mathbf{S}^{(0)}_{(i+1)_L} = \lim_{x\to 0^+} \mathbf{S}(\mathbf{U}_{i+1}(x),x,0) \tag{5.36}$$

and also its derivatives

$$\mathbf{S}^{(k)}_{i_R} = \lim_{x\to 0^-} \frac{\partial^k}{\partial x^k}\mathbf{S}(\mathbf{U}_i(x),x,0), \qquad \mathbf{S}^{(k)}_{(i+1)_L} = \lim_{x\to 0^+} \frac{\partial^k}{\partial x^k}\mathbf{S}(\mathbf{U}_{i+1}(x),x,0), \tag{5.37}$$

at the initial time, with $k = 1,...,K$.

High-order numerical methods of the ADER type require the solution at the interface position $x_{i+1/2}$ as a function of time t, allowing to compute the numerical fluxes and construct a numerical scheme of $K+1$-th order of accuracy in both space and time. Following [30, 39, 40] the solution will contain a leading term, provided by the DRP_0, equivalent to the classical piecewise constant data Riemann problem, associated with the first order Godunov scheme [6] and higher-order terms, associated with the K different RPs for the derivatives. It is worth saying that the Derivative Riemann Problem DRP_K can be decomposed in $K+1$ RPs where conventional Riemann Solvers are of application.

5.4.1 Evolution equation for derivatives

DRP in (5.31) provides the evolution equation for variable $\mathbf{U}$. Evolution equations for spatial or time derivatives of $\mathbf{U}$, denoted by $\partial_x^{(k)}\mathbf{U}$ and $\partial_t^{(k)}\mathbf{U}$ respectively, will be required for the resolution of the DRP. Such equations are straightforward obtained by taking succesive derivatives of (5.31), yielding

$$\frac{\partial}{\partial t}\left(\partial_x^{(k)}\mathbf{U}\right) + \frac{\partial}{\partial x}\left(\partial_x^{(k)}\mathbf{F}(\mathbf{U})\right) = \partial_x^{(k)}\mathbf{S} \qquad k = 1,...,K \tag{5.38}$$

and

$$\frac{\partial}{\partial t}\left(\partial_t^{(k)}\mathbf{U}\right) + \frac{\partial}{\partial x}\left(\partial_t^{(k)}\mathbf{F}(\mathbf{U})\right) = \partial_t^{(k)}\mathbf{S} \qquad k = 1,...,K \tag{5.39}$$

respectively. Algebraic manipulations of (5.38) yields

$$\frac{\partial}{\partial t}\left(\partial_x^{(k)}\mathbf{U}\right) + \mathbf{J}(\mathbf{U})\frac{\partial}{\partial x}\left(\partial_x^{(k)}\mathbf{U}\right) = \boldsymbol{\Upsilon}^{(k)} \tag{5.40}$$

where $\boldsymbol{\Upsilon}^{(k)} = \boldsymbol{\Upsilon}^{(k)}(\partial_x^{(k)}\mathbf{U}, \partial_x^{(k-1)}\mathbf{U}, ..., \mathbf{U}, \partial_x^{(k)}\mathbf{S}, \partial_x^{(k-1)}\mathbf{S}, ..., \mathbf{S})$ is a function of spatial derivatives of the fluxes and sources, that can be expressed as

$$\boldsymbol{\Upsilon}^{(k)} = -\frac{\partial^k}{\partial x^k}\left(\mathbf{J}(\mathbf{U})\frac{\partial \mathbf{U}}{\partial x}\right) + \mathbf{J}(\mathbf{U})\frac{\partial}{\partial x}\left(\frac{\partial^k \mathbf{U}}{\partial x^k}\right) + \frac{\partial^k \mathbf{S}}{\partial x^k} \tag{5.41}$$

Analogously, evolution equations for time derivatives of $\mathbf{U}$ are expressed as

$$\frac{\partial}{\partial t}\left(\partial_t^{(k)}\mathbf{U}\right) + \mathbf{J}(\mathbf{U})\frac{\partial}{\partial x}\left(\partial_t^{(k)}\mathbf{U}\right) = \boldsymbol{\Psi}^{(k)} \tag{5.42}$$

where $\boldsymbol{\Psi}^{(k)} = \boldsymbol{\Psi}^{(k)}(\partial_t^{(k)}\mathbf{U}, \partial_t^{(k-1)}\mathbf{U}, ..., \mathbf{U}, \partial_t^{(k)}\mathbf{S}, \partial_t^{(k-1)}\mathbf{S}, ..., \mathbf{S})$ is again a function of temporal derivatives of the fluxes and sources.

Different approaches can be done to find the solution for temporal derivatives in (5.40) or (5.42). In this work, the Jacobian will be considered as a constant coefficient matrix evaluated at $t = 0$, that means, spatial and temporal derivatives will be evolved using constant wave speeds corresponding to the eigenvalues of the Jacobian at the initial time. This leads to the following simplification of the evolution equations for

derivatives

$$\frac{\partial}{\partial t}\left(\partial_x^{(k)}\mathbf{U}\right)+\mathbf{J}(\mathbf{U}^{(0)})\frac{\partial}{\partial x}\left(\partial_x^{(k)}\mathbf{U}\right)=\partial_x^{(k)}\mathbf{S} \tag{5.43}$$

$$\frac{\partial}{\partial t}\left(\partial_t^{(k)}\mathbf{U}\right)+\mathbf{J}(\mathbf{U}^{(0)})\frac{\partial}{\partial x}\left(\partial_t^{(k)}\mathbf{U}\right)=\partial_t^{(k)}\mathbf{S} \tag{5.44}$$

noticing that derivatives of $\mathbf{U}$ and variable $\mathbf{U}$ are evolved using the same law.

5.4.2 Cauchy-Kowalevski (CK) Theorem

When dealing with EDPs of the type of (5.31), relations between temporal and spatial derivatives of $\mathbf{U}$ are provided by the CK theorem. Here, it is used to derive analytic expressions for time derivatives of $\mathbf{F}$ and $\mathbf{U}$ departing from the information provided by the spatial reconstruction method. It allows to express time derivatives of the physical fluxes at $t=0$ as functions $\mathbf{R}^{(k)}$ of spatial derivatives of $\mathbf{U}$ and $\mathbf{S}$

$$\partial_t^{(k)}\mathbf{F}=\mathbf{R}^{(k)}(\partial_x^{(k)}\mathbf{U},\partial_x^{(k-1)}\mathbf{U},...,\mathbf{U},\partial_x^{(k)}\mathbf{S},\partial_x^{(k-1)}\mathbf{S},...,\mathbf{S})\,. \tag{5.45}$$

Spatial derivatives defined at the cell interface $i+1/2$ in (5.33) are calculated using the sub-cell WENO reconstruction method [124]. They allow to compute the values of $\mathbf{R}^{(k)}$ at each side of the discontinuity in the DRP_K

$$\begin{aligned}\mathbf{R}_{i_R}^{(k)}&=\lim_{x\to 0^-}\mathbf{R}^{(k)}\approx\mathbf{R}^{(k)}(\mathbf{U}^{(k)},\mathbf{U}^{(k-1)},...,\mathbf{U}^0,\mathbf{S}^{(k)},\mathbf{S}^{(k-1)},...,\mathbf{S}^0)_{i_R}\\ \mathbf{R}_{(i+1)_L}^{(k)}&=\lim_{x\to 0^+}\mathbf{R}^{(k)}\approx\mathbf{R}^{(k)}(\mathbf{U}^{(k)},\mathbf{U}^{(k-1)},...,\mathbf{U}^0,\mathbf{S}^{(k)},\mathbf{S}^{(k-1)},...,\mathbf{S}^0)_{(i+1)_L}\,.\end{aligned} \tag{5.46}$$

Analogously, it is possible to construct temporal derivatives of $\mathbf{U}$ at $t=0$ as functions $\mathbf{D}^{(k)}$ of spatial derivatives of $\mathbf{U}$ and $\mathbf{S}$

$$\partial_t^{(k)}\mathbf{U}=\mathbf{D}^{(k)}(\partial_x^{(k)}\mathbf{U},\partial_x^{(k-1)}\mathbf{U},...,\mathbf{U},\partial_x^{(k)}\mathbf{S},\partial_x^{(k-1)}\mathbf{S},...,\mathbf{S}) \tag{5.47}$$

allowing to compute the values of $\mathbf{D}^{(k)}$ at each side of the discontinuity in the DRP_K

$$\begin{aligned}\mathbf{D}_{i_R}^{(k)}&=\lim_{x\to 0^-}\mathbf{D}^{(k)}\approx\mathbf{D}^{(k)}(\mathbf{U}^{(k)},\mathbf{U}^{(k-1)},...,\mathbf{U}^0,\mathbf{S}^{(k)},\mathbf{S}^{(k-1)},...,\mathbf{S}^0)_{i_R}\\ \mathbf{D}_{(i+1)_L}^{(k)}&=\lim_{x\to 0^+}\mathbf{D}^{(k)}\approx\mathbf{D}^{(k)}(\mathbf{U}^{(k)},\mathbf{U}^{(k-1)},...,\mathbf{U}^0,\mathbf{S}^{(k)},\mathbf{S}^{(k-1)},...,\mathbf{S}^0)_{(i+1)_L}\,.\end{aligned} \tag{5.48}$$

Temporal derivatives of the source term, $\mathbf{S}$, at $t=0$ can also be obtained using the CK procedure as functions $\mathbf{Q}^{(k)}$ of spatial derivatives of $\mathbf{U}$ and $\mathbf{S}$

$$\partial_t^{(k)}\mathbf{S}=\mathbf{Q}^{(k)}(\partial_x^{(k)}\mathbf{U},\partial_x^{(k-1)}\mathbf{U},...,\mathbf{U},\partial_x^{(k)}\mathbf{S},\partial_x^{(k-1)}\mathbf{S},...,\mathbf{S})\,. \tag{5.49}$$

5.5 Concluding remarks

The highlights of this chapter are listed below:

- The construction of ADER schemes has been considered. Two methodologies, the state-expansion and the flux-expansion approach, have been described. In this work, we adopt the flux-expansion approach, where the numerical flux is constructed as a power series expansion where the coefficients

are the solutions of the RPs composing the DRP. When solving linear problems, both approaches are the same.

- The traditional WENO-JS reconstruction procedure has been described here. As reported in the bibliography, the WENO-JS method has some flaws that can strongly reduce the order of convergence of the method. Moreover, when using the reconstruction provided by such method as the starting data to reconstruct spatial derivatives, those problems are more evident. The aforementioned problems are related to an inaccurate recovery of the optimal weights when the initial data has critical points (where derivatives vanish). This is due to an undesired behavior of the smoothness indicator in this particular case, hidden for some years by the tolerance used for the calculation of the corrected WENO weights. Different improved WENO procedures, found in the bibliography, are recalled. Some of them will be assessed numerically in following chapters, showing that convergence rates are improved when using them.

- The DRP has been described here. It is worth noting that the source term is included in the definition of the problem and will be herein considered in the resolution of the problem as well, following the *augmented solver* approach.

6 DRP SOLVERS FOR NONLINEAR CONSERVATION LAWS

When addressing the resolution of the DRP, the general fashion found in the literature is to neglect the presence of the source term and to solve a homogeneous DRP. Extra corrections are then required to ensure certain properties of the numerical scheme, such as equilibrium conditions.

An approach for the resolution of the DRP_K was first introduced in [30] and the proposed solver was called Toro–Titarev (TT) solver. This solver is based on the construction of a time–dependent solution at the interface as a power series expansion in time. It is worth emphasizing that the TT approach allows to reduce the DRP_K to a series of classical homogeneous Riemann problems where classical Riemann solvers are of application. The DRP_K consist of one RP for the leading term, referred to as DRP_0, plus K additional RPs for the derivatives. Up to date, a broad variety of Derivative Riemann solvers have been designed with the aim of providing accurate and fast solutions to different non-linear problems under a variety of conditions. Apart from the TT solver, the most common solvers we can find in the literature are the HEOC solver [31, 24] and the Castro–Toro (CT) solver [31], the latter constructed from the HEOC and the TT solvers. A semi–implicit version of those schemes was proposed by Montecinos [32], allowing to deal with more stiff source terms. It is worth pointing out that the traditional ADER approach using the CK procedure may become rather cumbersome when dealing with complex systems of equations and may not provide the expected performance when dealing with very stiff source terms. In such a case, a successful solution would be to replace the CK procedure by a local space–time Galerkin method, as done by the DET solver, proposed by Dumbser in [33].

In the TT and CT solvers, the solution is computed adopting the so-called state expansion approach. Such technique proposes the construction of the solution as a Taylor power series expansion in time. The leading term of the expansion is given by the solution of a classical Riemann problem for the evolution of the conserved quantities neglecting the presence of the source term. The calculation of the higher order terms varies from one solver to another. In the TT solver, derivatives are evolved by solving conventional homogeneous linearized RPs without source terms, defined for spatial derivatives of the conserved quantities. Then, the CK procedure is used to transform the evolved space derivatives to time derivatives. On the other hand, in the CT solver the higher order terms are directly given by the solution of conventional homogeneous linearized Riemann problems without source terms defined for time derivatives. In this case, the CK procedure is used to provide the initial data for such RPs. A radically different approach is used in the HEOC solver, which only considers the resolution of conventional homogeneous non-linear RPs whose initial data has already been evolved separately in time by means of Taylor power series expansions.

Alternatively to the state-expansion approach, as done in the TT-ADER and CT-ADER schemes, it is possible to directly compute the numerical flux as as a Taylor power series expansion in time at the interface. This is the so-called flux-expansion approach. The numerical schemes considered in this work are constructed

using the flux-expansion approach. It is worth pointing out that both methodologies are equivalent when solving linear problems.

The TT, CT and HEOC solvers assume that there is a single solution, referred to as the star solution [125], connecting the two initial states, with independence of the presence of source terms. However, when dealing with geometric source terms, it was shown that the consideration of the source term in the resolution of the DRP is more convenient in order to ensure certain properties of the numerical solution [19, 20]. In this chapter, a new family of DRP solvers for hyperbolic conservation laws with geometric source terms, constructed following the augmented-solver approach, is presented.

Two novel solvers are proposed here, the Flux-source (FS) solver and the Linearized flux-source (LFS) solver. They are based on the decomposition of the DRP into traditional RPs, one for the original evolution equations and K other for the k-th time derivative of the evolution equations, as done in [30, 31]. As mentioned before, unlike most DRP solvers found in the bibliography, the FS and LFS solvers consider the non-homogeneous evolution equations when geometric source terms are present. Other kind of source terms can be neglected when using such solvers.

Traditional Riemann solvers are required to compute the solution of each of the single RPs that compose the DRP. Augmented solvers are required to this end as we have to deal with non-homogeneous RPs. To this end, the ARoe solver and HLLS solvers are considered. It is worth pointing out that the FS and LFS solvers are based on the flux-expansion methodology. The numerical flux is finally constructed as a Taylor power series expansion in time whose coefficients are computed as the solution of the traditional RPs.

Whereas the FS solver considers nonlinear problems for both the leading term and the derivative terms, in the LFS solver, those RPs corresponding to the derivatives are linearized by means of a suitable approximation of the Jacobian matrix at the initial time. This leads to conventional linearized non-homogeneous RPs for the derivatives. Such strategy is done, for instance, in the TT and CT solvers [30, 31], where linearized (but homogeneous) equations are solved. When using the LFS solver, only time derivatives of the conserved quantities are required as initial data to obtain the solution of the RPs, avoiding the computation of time derivatives of the fluxes, required for the FS solver. However, it is done at the expense of a stronger restriction on the time step due to the linearization of the equations.

It is worth pointing out that when using the LFS solver, Riemann solvers are not required for the resolution of the linear RPs defined for the high order terms, as they have analytical solution. However, for the sake of simplicity, we propose to use the Riemann solver selected for the resolution of the leading term, which in this case will provide the exact solution (assuming that the source term can be integrated exactly).

In Section 6.1, the general formulation of the FS and LFS solvers is provided. Then, in Section 6.2 and 6.3, two particular solvers, termed by the author as AR-(L)FS and HLLS-(L)FS solvers, respectively, are presented. The former appears from the combination of the ARoe solver with the (L)FS solver whereas the latter appears from the combination of the HLLS solver with the (L)FS solver. Some conclusions are presented in Section 6.4.

6.1 The FS and LFS solvers

In this section, we provide a general definition of the Flux Source (FS) and Linear Flux Source (LFS) Derivative Riemann solver [83, 84]. Such solvers have been designed to overcome the difficulties arising in the resolution of the DRP in presence of geometric source terms.

When using the FS solver, the DRP_K in (5.31) is decomposed in $K+1$ conventional RPs, one for the evolution of the conserved quantities and K more problems for the evolution of the derivatives. The former is known as DRP_0 and corresponds to the following non-linear and non-homogeneous RP

$$\begin{cases} \dfrac{\partial \mathbf{U}}{\partial t} + \dfrac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{S} \\ \mathbf{U}(x,0) = \begin{cases} \mathbf{U}^{(0)}_{i_R} & \text{if} \quad x < 0 \\ \mathbf{U}^{(0)}_{(i+1)_L} & \text{if} \quad x > 0 \end{cases} \end{cases} \tag{6.1}$$

with the sought solution for the fluxes denoted as $\mathbf{F}^{-,(0)}_{i_R}$ and $\mathbf{F}^{+,(0)}_{(i+1)_L}$.

The K RPs associated to the high order terms of the DRP$_K$ are composed of the evolution equations for time derivatives of the conserved variables. The evolution equations for the derivatives are derived straightforward by differentiating the original system according to Equation (5.39), leading to the following RP

$$\begin{cases} \dfrac{\partial}{\partial t}\left(\partial_t^{(k)}\mathbf{U}\right) + \dfrac{\partial}{\partial x}\left(\partial_t^{(k)}\mathbf{F}(\mathbf{U})\right) = \partial_t^{(k)}\mathbf{S} \\ \partial_t^{(k)}\mathbf{U}(x,0) = \begin{cases} \mathbf{D}^{(k)}_{i_R} & \text{if} \quad x < 0 \\ \mathbf{D}^{(k)}_{(i+1)_L} & \text{if} \quad x > 0 \end{cases} \end{cases} \tag{6.2}$$

with the sought solution for the fluxes denoted as $\mathbf{F}^{-,(k)}_{i_R}$ and $\mathbf{F}^{+,(k)}_{(i+1)_L}$. It is worth saying that the reconstructed time derivatives of the fluxes at the interfaces, $\mathbf{R}^{(k)}_{i_R}$ and $\mathbf{R}^{(k)}_{(i+1)_L}$, are also required when solving (6.2).

We present now an alternative strategy to the FS solver which leads to the LFS solver. A similar strategy has been used by other authors [30, 31] and considers the resolution of linearized evolution equations for the derivatives. In [30, 31] linearized and homogeneous RPs for the derivatives are solved. The advantage of this approach is that only the reconstruction of time derivatives of the conserved quantities at the interfaces are required, unlike the FS solver that also requires time derivatives of the fluxes. In this case, we solve the following linearized RP for the evolution equations

$$\begin{cases} \dfrac{\partial}{\partial t}\left(\partial_t^{(k)}\mathbf{U}\right) + \widetilde{\mathbf{J}}_{i+1/2}\dfrac{\partial}{\partial x}\left(\partial_t^{(k)}\mathbf{U}\right) = \partial_t^{(k)}\mathbf{S} \\ \partial_t^{(k)}\mathbf{U}(x,0) = \begin{cases} \mathbf{D}^{(k)}_{i_R} & \text{if} \quad x < 0 \\ \mathbf{D}^{(k)}_{(i+1)_L} & \text{if} \quad x > 0 \end{cases} \end{cases} \tag{6.3}$$

where $\widetilde{\mathbf{J}}_{i+1/2} = \tilde{\mathbf{J}}_{i+1/2}(\mathbf{U}^{(0)}_{i_R}, \mathbf{U}^{(0)}_{(i+1)_L})$ is a constant matrix that allows to approximate time derivatives of the flux as

$$\partial_t^{(k)}\mathbf{F}(\mathbf{U}) = \widetilde{\mathbf{J}}_{i+1/2}\partial_t^{(k)}\mathbf{U}. \tag{6.4}$$

Notice that the source term is not neglected as in the TT, CT and HEOC solvers [30, 31, 24]. As the constructed ADER schemes are of the flux-ADER type, only the solution for the fluxes will be sought when solving the DRP$_K$ in (5.31).

When adopting the flux-expansion ADER approach we seek a truncated Taylor time expansion of the updated fluxes at the interfaces as (5.25) where $\mathbf{F}^{-,(0)}_{i_R}$ and $\mathbf{F}^{+,(0)}_{(i+1)_L}$ are computed from (6.1) and $\mathbf{F}^{-,(k)}_{i_R}$ and $\mathbf{F}^{+,(k)}_{(i+1)_L}$ are computed from (6.2) or (6.3).

6.2 The AR-(L)FS solver

A flux-ADER solver for the resolution of the DRP_K in (5.31) in presence of geometric source terms is presented. The solver is called Augmented Roe (Linearized) Flux Source (AR-(L)FS) and is presented for the first time in [83] and improved in [84]. The Augmented version of the Roe solver [15] (ARoe solver) in Section 4.2.1 [20, 22] is used here to solve the conventional RPs appearing when using the FS and LFS solvers. The ARoe solver takes into account the contribution of the source term in the solution, ensuring an exact balance between numerical fluxes and source terms.

Let us consider the hyperbolic nonlinear system of equations with source term in (5.31). Assuming that the convective part of (5.31) is strictly hyperbolic, with N_λ real eigenvalues $\lambda^1, ..., \lambda^{N_\lambda}$ and eigenvectors $\mathbf{e}^1, ..., \mathbf{e}^{N_\lambda}$, it is possible define two matrices $\mathbf{P} = (\mathbf{e}^1, ..., \mathbf{e}^{N_\lambda})$ and $\mathbf{P}^{-1}$ with the property that they diagonalize the Jacobian $\mathbf{J}$, as shown in (2.23).

The ARoe solver is based on the decomposition of the approximate Jacobian of the homogeneous part at the initial time $\tilde{\mathbf{J}}_{i+1/2}(\mathbf{U}^{(0)}_{i_R}, \mathbf{U}^{(0)}_{(i+1)_L})$

$$\delta\mathbf{F}^{(0)}_{i+1/2} = \widetilde{\mathbf{J}}_{i+1/2}\delta\mathbf{U}^{(0)}_{i+1/2}, \tag{6.5}$$

leading to a set of approximated eigenvalues $\tilde{\lambda}^m_{i+1/2}$ and eigenvectors $\widetilde{\mathbf{e}}^m_{i+1/2} = (e^m_1, ..., e^m_{N_\lambda})^T$. The approximate Jacobian $\widetilde{\mathbf{J}}_{i+1/2}$ can be expressed as

$$\widetilde{\mathbf{J}}_{i+1/2} = \widetilde{\mathbf{P}}_{i+1/2}\mathbf{\Lambda}_{i+1/2}\widetilde{\mathbf{P}}^{-1}_{i+1/2}, \tag{6.6}$$

with $\widetilde{\mathbf{P}}_{i+1/2} = \left(\widetilde{\mathbf{e}}^1, ..., \widetilde{\mathbf{e}}^{N_\lambda}\right)_{i+1/2}$ an invertible matrix composed by the eigenvectors of $\widetilde{\mathbf{J}}_{i+1/2}$ and $\mathbf{\Lambda}_{i+1/2}$ the diagonal matrix composed by the eigenvalues of $\widetilde{\mathbf{J}}_{i+1/2}$.

The leading terms of the expansion in (5.25) are computed by solving the DRP_0, given by Equation (6.1). When using the ARoe solver, the solution of the DRP_0 for the fluxes reads

$$\begin{aligned}
\mathbf{F}^{-,(0)}_{i_R} &= \mathbf{F}^{(0)}_{i_R} + \textstyle\sum_{m=1}^{N_\lambda}\left(\widetilde{\lambda}^- \alpha^{(0)} - \beta^{-,(0)}\right)^m_{i+1/2}\widetilde{\mathbf{e}}^m_{i+1/2}, \\
\mathbf{F}^{+,(0)}_{(i+1)_L} &= \mathbf{F}^{(0)}_{(i+1)_L} - \textstyle\sum_{m=1}^{N_\lambda}\left(\widetilde{\lambda}^+ \alpha^{(0)} - \beta^{+,(0)}\right)^m_{i+1/2}\widetilde{\mathbf{e}}^m_{i+1/2},
\end{aligned} \tag{6.7}$$

with $\mathbf{F}^{(0)}_{i_R}$ and $\mathbf{F}^{(0)}_{(i+1)_L}$ the physical fluxes defined in (5.34),

$$\left(\tilde{\lambda}^\pm\right)^m_{i+1/2} = \left(\frac{\widetilde{\lambda} \pm |\widetilde{\lambda}|}{2}\right)^m_{i+1/2}, \qquad \left(\beta^{\pm,(0)}\right)^m_{i+1/2} = \left(\frac{\widetilde{\lambda}^\pm}{\widetilde{\lambda}}\beta^{(0)}\right)^m_{i+1/2}, \tag{6.8}$$

$\alpha^{(0)}$ the wave strengths given by the projection of $\delta\mathbf{U}^{(0)}_{i+1/2}$ onto the Jacobian's eigenvectors basis as

$$\delta\mathbf{U}^{(0)}_{i+1/2} = \widetilde{\mathbf{P}}_{i+1/2}\mathbf{A}_{i+1/2}, \tag{6.9}$$

with $\mathbf{A}_{i+1/2} = \left(\alpha^{(0),1}, ..., \alpha^{(0),N_\lambda}\right)^T_{i+1/2}$ and $\beta^{(0)}$ the source strengths associated to each wave, given by

$$\bar{\mathbf{S}}^0_{i+1/2} = \widetilde{\mathbf{P}}_{i+1/2}\mathbf{B}^{(0)}_{i+1/2}, \tag{6.10}$$

with $\mathbf{B}^{(0)}_{i+1/2} = \left(\beta^{(0),1}, ..., \beta^{(0),N_\lambda}\right)^T_{i+1/2}$. As in the scalar case, a suitable approximation of the integral of the source term across the interface

$$\bar{\mathbf{S}}^{(0)}_{i+1/2} = \int_{x^-_{i+1/2}}^{x^+_{i+1/2}} \mathbf{S}(x,0)\, dx, \tag{6.11}$$

must be found when dealing with geometric source terms.

The same procedure is extended to derive the expression of the derivative terms of the numerical fluxes in (5.25). This is performed by solving the K RPs associated to the high order terms of the DRP_K, given by (6.2). When using the FS solver, the numerical solution for such RPs is computed using an extension of the ARoe solver that does not use derivatives of the conserved variables, $\mathbf{D}^{(k)}_{i_R}$ and $\mathbf{D}^{(k)}_{(i+1)_L}$, as initial condition but derivatives of the fluxes at the interfaces instead, namely $\mathbf{R}^{(k)}_{i_R}$ and $\mathbf{R}^{(k)}_{(i+1)_L}$. The proposed solution provides the following approximate fluxes at the interface

$$\begin{aligned}
\mathbf{F}^{-,(k)}_{i_R} &= \mathbf{R}^{(k)}_{i_R} + \textstyle\sum_{m=1}^{N_\lambda} \left(\gamma^{-,(k)} - \beta^{-,(k)}\right)^m_{i+1/2} \widetilde{\mathbf{e}}^m_{i+1/2}, \\
\mathbf{F}^{+,(k)}_{(i+1)_L} &= \mathbf{R}^{(k)}_{(i+1)_L} - \textstyle\sum_{m=1}^{N_\lambda} \left(\gamma^{+,(k)} - \beta^{+,(k)}\right)^m_{i+1/2} \widetilde{\mathbf{e}}^m_{i+1/2},
\end{aligned} \tag{6.12}$$

with

$$\left(\gamma^{\pm,(k)}\right)^m_{i+1/2} = \left(\frac{\widetilde{\lambda}^{\pm}}{\widetilde{\lambda}}\gamma^{(k)}\right)^m_{i+1/2}, \qquad \left(\beta^{\pm,(k)}\right)^m_{i+1/2} = \left(\frac{\widetilde{\lambda}^{\pm}}{\widetilde{\lambda}}\beta^{(k)}\right)^m_{i+1/2}. \tag{6.13}$$

The flux strengths, $\gamma^{(k)}$, are given in this case by the projection of the variation of $\mathbf{R}^{(k)}$ onto the Jacobian eigenvectors basis

$$\delta\mathbf{R}^{(k)}_{i+1/2} = \widetilde{\mathbf{P}}_{i+1/2}\boldsymbol{\Gamma}^{(k)}_{i+1/2}, \tag{6.14}$$

with $\boldsymbol{\Gamma}^{(k)}_{i+1/2} = \left(\gamma^{(k),1}, ..., \gamma^{(k),N_\lambda}\right)^T_{i+1/2}$, and the same for the source strengths, $\beta^{(k)}$, associated to each wave

$$\bar{\mathbf{S}}^{(k)}_{i+1/2} = \widetilde{\mathbf{P}}_{i+1/2}\mathbf{B}^{(k)}_{i+1/2}, \tag{6.15}$$

with $\mathbf{B}^{(k)}_{i+1/2} = \left(\beta^{(k),1}, ..., \beta^{(k),N_\lambda}\right)^T_{i+1/2}$. A suitable approximation of the integral of the source term across the interface

$$\bar{\mathbf{S}}^{(k)}_{i+1/2} = \int_{x^-_{i+1/2}}^{x^+_{i+1/2}} \mathbf{Q}^{(k)}\, dx, \tag{6.16}$$

must be found when dealing with geometric source terms.

As outlined in previous sections, $\delta\mathbf{M}^{+}_{i-1/2}$ and $\delta\mathbf{M}^{-}_{i+1/2}$ represent the contributions of the incoming waves at cell interfaces and are expressed as

$$\delta\mathbf{M}^{+,(0)}_{i-1/2} = \sum_{m=1}^{N_\lambda} \left[\left(\widetilde{\lambda}^{+}\alpha^{(0)} - \beta^{+,(0)}\right)\widetilde{\mathbf{e}}\right]^m_{i-1/2}, \tag{6.17}$$

$$\delta\mathbf{M}^{+,(k)}_{i-1/2} = \sum_{k=1}^{K}\sum_{m=1}^{N_\lambda} \left(\left(\gamma^{+,(k)} - \beta^{+,(k)}\right)\widetilde{\mathbf{e}}\right)^m_{i-1/2} \frac{\Delta t^k}{(k+1)!}, \tag{6.18}$$

$$\delta\mathbf{M}^{-,(0)}_{i+1/2} = \sum_{m=1}^{N_\lambda} \left[\left(\widetilde{\lambda}^{-}\alpha^{(0)} - \beta^{-,(0)}\right)\widetilde{\mathbf{e}}\right]^m_{i+1/2}, \tag{6.19}$$

$$\delta\mathbf{M}^{-,(k)}_{i+1/2} = \sum_{k=1}^{K}\sum_{m=1}^{N_\lambda}\left(\left(\gamma^{-,(k)} - \beta^{-,(k)}\right)\tilde{\mathbf{e}}\right)^m_{i+1/2}\frac{\Delta t^k}{(k+1)!} \tag{6.20}$$

6.2.1 Linear approach for the high order terms

When using the LFS Derivative Riemann solver, RPs for the derivatives are given by (6.3) instead of (6.2). The numerical solution for (6.3) is computed using an extension of the ARoe solver that, in this case, only uses derivatives of the conserved variables, $\mathbf{D}^{(k)}_{i_R}$ and $\mathbf{D}^{(k)}_{(i+1)_L}$, as initial condition. The proposed solution for the approximate derivative fluxes at the interfaces is

$$\begin{aligned}
\mathbf{F}^{-,(k)}_{i_R} &= \widetilde{\mathbf{J}}_{i+1/2}\mathbf{D}^{(k)}_{i_R} + \textstyle\sum_{m=1}^{N_\lambda}\left(\widetilde{\lambda}^{-}\alpha^{(k)} - \beta^{-,(k)}\right)^m_{i+1/2}\widetilde{\mathbf{e}}^m_{i+1/2}, \\
\mathbf{F}^{+,(k)}_{(i+1)_L} &= \widetilde{\mathbf{J}}_{i+1/2}\mathbf{D}^{(k)}_{(i+1)_L} - \textstyle\sum_{m=1}^{N_\lambda}\left(\widetilde{\lambda}^{+}\alpha^{(k)} - \beta^{+,(k)}\right)^m_{i+1/2}\widetilde{\mathbf{e}}^m_{i+1/2},
\end{aligned} \tag{6.21}$$

with

$$\left(\tilde{\lambda}^{\pm}\right)^m_{i+1/2} = \left(\frac{\widetilde{\lambda} \pm |\widetilde{\lambda}|}{2}\right)^m_{i+1/2}, \qquad \left(\beta^{\pm,(0)}\right)^m_{i+1/2} = \left(\frac{\widetilde{\lambda}^{\pm}}{\widetilde{\lambda}}\beta^{(0)}\right)^m_{i+1/2}. \tag{6.22}$$

The wave strengths, $\alpha^{(k)}$, are given in this case by the projection of the jump of $\mathbf{D}^{(k)}$ onto the Jacobian's eigenvectors basis

$$\delta\mathbf{D}^{(k)}_{i+1/2} = \widetilde{\mathbf{P}}_{i+1/2}\mathbf{A}^{(k)}_{i+1/2}, \tag{6.23}$$

with $\mathbf{A}^{(k)}_{i+1/2} = \left(\alpha^{(k),1},...,\alpha^{(k),N_\lambda}\right)^T_{i+1/2}$. The source strengths, $\beta^{(k)}$, associated to each wave are computed as proposed in the original approach, using (6.15).

Contributions of the incoming waves at cell interfaces, $\delta\mathbf{M}^{+}_{i-1/2}$ and $\delta\mathbf{M}^{-}_{i+1/2}$, will be now expressed as

$$\delta\mathbf{M}^{+,(0)}_{i-1/2} = \sum_{m=1}^{N_\lambda}\left[\left(\widetilde{\lambda}^{+}\alpha^{(0)} - \beta^{+,(0)}\right)\tilde{\mathbf{e}}\right]^m_{i-1/2}, \tag{6.24}$$

$$\delta\mathbf{M}^{+,(k)}_{i-1/2} = \sum_{k=1}^{K}\sum_{m=1}^{N_\lambda}\left(\left(\widetilde{\lambda}^{-}\alpha^{(k)} - \beta^{+,(k)}\right)\tilde{\mathbf{e}}\right)^m_{i-1/2}\frac{\Delta t^k}{(k+1)!}, \tag{6.25}$$

$$\delta\mathbf{M}^{-,(0)}_{i+1/2} = \sum_{m=1}^{N_\lambda}\left[\left(\widetilde{\lambda}^{-}\alpha^{(0)} - \beta^{-,(0)}\right)\tilde{\mathbf{e}}\right]^m_{i+1/2}, \tag{6.26}$$

$$\delta\mathbf{M}^{-,(k)}_{i+1/2} = \sum_{k=1}^{K}\sum_{m=1}^{N_\lambda}\left(\left(\widetilde{\lambda}^{+}\alpha^{(k)} - \beta^{-,(k)}\right)\tilde{\mathbf{e}}\right)^m_{i+1/2}\frac{\Delta t^k}{(k+1)!}, \tag{6.27}$$

6.3 The HLLS-(L)FS solver

In this section, the HLLS solver presented in Section 4.2.2 [22] is extended for the resolution of the DRP by means of the FS solver and LFS solver, allowing to construct an ADER type numerical scheme.

As done in the AR-ADER scheme, the resolution of the DRP is done by solving the $K+1$ RPs associated to the leading term and the derivatives. To this end, the HLLS solver [22] is used for the computation of the leading term, whereas two different approaches are proposed for the computation of the high order terms,

depending whether we use either the FS or LFS solver. If using the FS solver, the HLLS solver is applied for the resolution of the nonlinear evolution equations associated to the derivatives. On the other hand, if using the LFS solver, we linearize those equations and use the HLLS solver to obtain the solution. It is worth mentioning that the HLLS solver is a nonlinear Riemann solver unlike the ARoe solver, but it can also be used to provide the solution of linear RPs with source term.

We depart from the original DRP in (5.31) considering for this derivation that $\mathbf{U}(x,t) \in \mathscr{C} \subseteq \mathbb{R}^2$ and $\mathbf{F}(\mathbf{U}) : \mathscr{C} \to \mathbb{R}^2$, that is, (5.31) is a system of two equations characterized by two real eigenvalues $\lambda^1(\mathbf{U}) \leq \lambda^2(\mathbf{U})$ corresponding to the wave speeds plus an extra wave of speed $\mathscr{S} = 0$ at $x = 0$.

The leading terms, $\mathbf{F}^{-,0}_{i_R}$ and $\mathbf{F}^{+,0}_{(i+1)_L}$, are calculated by solving the so-called DRP$_0$, corresponding to RP in (6.1)and depicted in Figure 6.1.

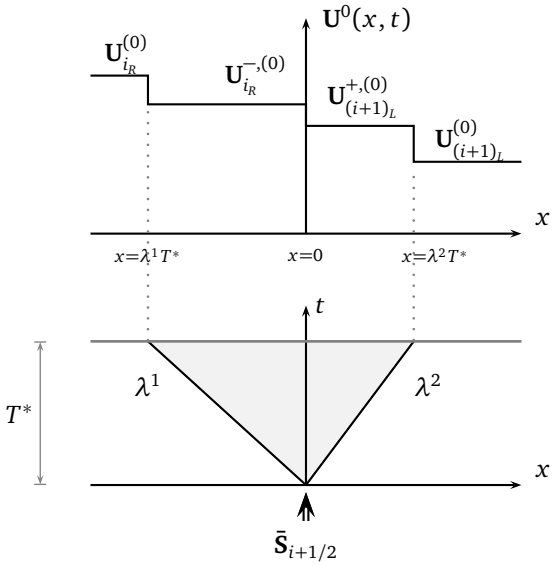


Figure 6.1: Values of the solution for the DRP$_0$, $\mathbf{U}^0(x,t)$, in each wedge of the (x,t) plane.

The integral form of (6.1) inside a control volume $[-x_L, x_R] \times [0, T^*]$ provides the expression for the integral volume of $\mathbf{U}(x,T^*)$ as

$$\int_{-x_L}^{x_R} \mathbf{U}(x,T^*)\,dx = x_R \mathbf{U}^{(0)}_{(i+1)_L} + x_L \mathbf{U}^{(0)}_{i_R} + (\mathbf{F}^{(0)}_{i_R} - \mathbf{F}^{(0)}_{(i+1)_L})T^* + \bar{\mathbf{S}}^{(0)}_{i+\frac{1}{2}} T^*, \tag{6.28}$$

with conserved quantities, fluxes and source as defined in Section 2. The integral on the left hand side of (6.28) can be split considering a wave structure given by $\lambda^1 \leq 0 \leq \lambda^2$ as depicted in Figure 6.1 and with $-x_L < \lambda^1 T^*$ and $x_R > \lambda^2 T^*$

$$\begin{aligned}\int_{-x_L}^{x_R} \mathbf{U}(x,T^*)\,dx = &\int_{-x_L}^{\lambda^1 T^*} \mathbf{U}(x,T^*)\,dx + \int_{\lambda^1 T^*}^{0} \mathbf{U}(x,T^*)\,dx + \\ &\int_{0}^{\lambda^2 T^*} \mathbf{U}(x,T^*)\,dx + \int_{\lambda^2 T^*}^{x_R} \mathbf{U}(x,T^*)\,dx\end{aligned} \tag{6.29}$$

and considering the solution composed of four constant states as shown in Figure 6.1, it yields

$$\int_{-x_L}^{x_R} \mathbf{U}(x,T^*)\,dx = \mathbf{U}_{i_R}^{(0)}(\lambda^1 T^* + x_L) + \mathbf{U}_{(i+1)_L}^{(0)}(x_R - \lambda^2 T^*) + \mathbf{U}_{i_R}^{-,(0)}(-\lambda^1 T^*) + \mathbf{U}_{(i+1)_L}^{+,(0)}(\lambda^2 T^*)\,. \tag{6.30}$$

Now, substitution of (6.30) in (6.28) leads to

$$(\mathbf{U}_{i_R}^{(0)} - \mathbf{U}_{i_R}^{-,(0)})\lambda^1 - (\mathbf{U}_{(i+1)_L}^{(0)} - \mathbf{U}_{(i+1)_L}^{+,(0)})\lambda^2 + \mathbf{F}_{(i+1)_L}^{(0)} - \mathbf{F}_{i_R}^{(0)} = \bar{\mathbf{S}}_{i+\frac{1}{2}}^{(0)}\,, \tag{6.31}$$

where an extra condition is needed in order to obtain an expression for $\mathbf{U}_{i_R}^{-,(0)}$ and $\mathbf{U}_{(i+1)_L}^{+,(0)}$, due to the presence of the source term. For that purpose, let us define first an approximate flux function $\hat{\mathbf{F}}(x,t)$ with a similar structure than $\mathbf{U}(x,t)$ as depicted in Figure 6.1. In this case, also intercell values for the fluxes can be defined at both sides of the t axis as

$$\mathbf{F}_{i_R}^{-,(0)} = \lim_{x\to 0^-} \hat{\mathbf{F}}(x,t)\,, \qquad \mathbf{F}_{(i+1)_L}^{+,(0)} = \lim_{x\to 0^+} \hat{\mathbf{F}}(x,t)\,. \tag{6.32}$$

The following RH relations across waves between fluxes and conserved variables are stated

$$\mathbf{F}_{i_R}^{-,(0)} - \mathbf{F}_{i_R}^{(0)} = \lambda^1(\mathbf{U}_{i_R}^{-,(0)} - \mathbf{U}_{i_R}^{(0)})\,, \tag{6.33}$$

$$\mathbf{F}_{(i+1)_L}^{(0)} - \mathbf{F}_{(i+1)_L}^{+,(0)} = \lambda^2(\mathbf{U}_{(i+1)_L}^{(0)} - \mathbf{U}_{(i+1)_L}^{+,(0)})\,, \tag{6.34}$$

$$\mathbf{F}_{(i+1)_L}^{+,(0)} - \mathbf{F}_{i_R}^{-,(0)} - \bar{\mathbf{S}}_{i+\frac{1}{2}}^{(0)} = \mathscr{S}(\mathbf{U}_{(i+1)_L}^{+,(0)} - \mathbf{U}_{i_R}^{-,(0)}) = 0\,. \tag{6.35}$$

Moreover, using Roe's approach it is possible to define the following relation as done in the derivation of the AR-ADER scheme

$$\mathbf{F}_{(i+1)_L}^{+,(0)} - \mathbf{F}_{i_R}^{-,(0)} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}}\left(\mathbf{U}_{(i+1)_L}^{+,(0)} - \mathbf{U}_{i_R}^{-,(0)}\right), \tag{6.36}$$

where $\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}}(\mathbf{U}_{i_R}^{-,(0)}, \mathbf{U}_{(i+1)_L}^{+,(0)})$ is an approximation of the Jacobian matrix according to Equations (6.5) and (6.6). Combining (6.35) and (6.36), the following relation appears

$$\bar{\mathbf{S}}_{i+\frac{1}{2}}^{(0)} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}}\left(\mathbf{U}_{(i+1)_L}^{+,(0)} - \mathbf{U}_{i_R}^{-,(0)}\right), \tag{6.37}$$

being possible to write the jump on the conserved variables across the stationary wave at $x = 0$ as

$$\mathbf{U}_{(i+1)_L}^{+,(0)} - \mathbf{U}_{i_R}^{-,(0)} = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}}\bar{\mathbf{S}}_{i+\frac{1}{2}}^{(0)} = \bar{\mathbf{H}}_{i+\frac{1}{2}}^{(0)}\,. \tag{6.38}$$

Combination of (6.31) and (6.38) leads to the following values for the intermediate states

$$\mathbf{U}_{i_R}^{-,(0)} = \frac{\mathbf{F}_{i_R}^{(0)} - \mathbf{F}_{(i+1)_L}^{(0)} + \lambda^2\mathbf{U}_{(i+1)_L}^{(0)} - \lambda^1\mathbf{U}_{i_R}^{(0)} + \bar{\mathbf{S}}_{i+\frac{1}{2}}^{(0)} - \lambda^2\bar{\mathbf{H}}_{i+\frac{1}{2}}^{(0)}}{\lambda^2 - \lambda^1}\,, \tag{6.39}$$

$$\mathbf{U}_{(i+1)_L}^{+,(0)} = \frac{\mathbf{F}_{i_R}^{(0)} - \mathbf{F}_{(i+1)_L}^{(0)} + \lambda^2\mathbf{U}_{(i+1)_L}^{(0)} - \lambda^1\mathbf{U}_{i_R}^{(0)} + \bar{\mathbf{S}}_{i+\frac{1}{2}}^{(0)} - \lambda^1\bar{\mathbf{H}}_{i+\frac{1}{2}}^{(0)}}{\lambda^2 - \lambda^1}\,. \tag{6.40}$$

Application of the RH conditions in (6.33) - (6.35) allows to calculate the expression for the zero-th

order intercell fluxes. For this calculation, we will consider $\lambda^1 \leq 0 \leq \lambda^2$, that is, subcritical regime, and the approximate fluxes at both sides of the interface will be denoted by $\mathbf{F}^{-,(0),sub}_{i_R}$ and $\mathbf{F}^{+,(0),sub}_{(i+1)_L}$. Such fluxes read

$$\mathbf{F}^{-,(0),sub}_{i_R} = \frac{\lambda^2 \mathbf{F}^{(0)}_{i_R} - \lambda^1 \mathbf{F}^{(0)}_{(i+1)_L} + \lambda^1\lambda^2 \delta \mathbf{U}^{(0)}_{i+\frac{1}{2}} + \lambda^1 \left(\bar{\mathbf{S}}^{(0)}_{i+\frac{1}{2}} - \lambda^2 \bar{\mathbf{H}}^{(0)}_{i+\frac{1}{2}} \right)}{\lambda^2 - \lambda^1}, \tag{6.41}$$

$$\mathbf{F}^{+,(0),sub}_{(i+1)_L} = \frac{\lambda^2 \mathbf{F}^{(0)}_{i_R} - \lambda^1 \mathbf{F}^{(0)}_{(i+1)_L} + \lambda^1\lambda^2 \delta \mathbf{U}^{(0)}_{i+\frac{1}{2}} + \lambda^2 \left(\bar{\mathbf{S}}^{(0)}_{i+\frac{1}{2}} - \lambda^1 \bar{\mathbf{H}}^{(0)}_{i+\frac{1}{2}} \right)}{\lambda^2 - \lambda^1}. \tag{6.42}$$

with $\mathbf{F}^{(0)}_{i_R}$ and $\mathbf{F}^{(0)}_{(i+1)_L}$ the physical fluxes evaluated at both sides of the interface according to Equation (5.34), $\mathbf{U}^{(0)}_{i_R}$ and $\mathbf{U}^{(0)}_{(i+1)_L}$ the reconstructed variables at both sides of the interface according to (5.32) and $\bar{\mathbf{S}}^{(0)}_{i+\frac{1}{2}}$ a suitable approximation of the integral of the sources across the interface according to (5.22).

When considering all different combinations of wave speeds, that is, not only the subcritical case considered above but also supercritical regimes, the general expression for the zero-th order intercell numerical fluxes reads

$$\mathbf{F}^{-,(0)}_{i_R} = \begin{cases} \mathbf{F}^{(0)}_{i_R} & \text{if} \quad \lambda^1 \geq 0 \\ \mathbf{F}^{-,(0),sub}_{i_R} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \mathbf{F}^{(0)}_{(i+1)_L} - \bar{\mathbf{S}}^{(0)}_{i+\frac{1}{2}} & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.43}$$

$$\mathbf{F}^{+,(0)}_{(i+1)_L} = \begin{cases} \mathbf{F}^{(0)}_{i_R} + \bar{\mathbf{S}}^{(0)}_{i+\frac{1}{2}} & \text{if} \quad \lambda^1 \geq 0 \\ \mathbf{F}^{+,(0),sub}_{(i+1)_L} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \mathbf{F}^{(0)}_{(i+1)_L} & \text{if} \quad \lambda^2 \leq 0 \end{cases}. \tag{6.44}$$

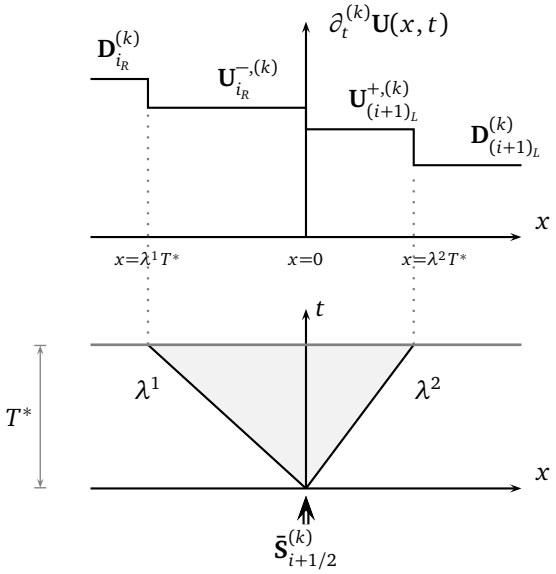


Figure 6.2: Values of the solution for the k-th order RP, $\partial_t \mathbf{U}^{(k)}(x,t)$, in each wedge of the (x,t) plane.

Terms associated to temporal derivatives of the fluxes in (5.25), $\mathbf{F}^{-,(k)}_{i_R}$ and $\mathbf{F}^{+,(k)}_{(i+1)_L}$, are calculated by solving the corresponding k-th order RPs in (6.2). The integral form of (6.2) inside a control volume $[-x_L, x_R] \times [0, T^*]$ is expressed as

$$\int_{-x_L}^{x_R} \partial_t^{(k)}\mathbf{U}(x,T^*)\,dx = x_R\mathbf{D}^{(k)}_{(i+1)_L} + x_L\mathbf{D}^{(k)}_{i_R} + (\mathbf{R}^{(k)}_{i_R} - \mathbf{R}^{(k)}_{(i+1)_L})T^* + \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}}T^*, \tag{6.45}$$

with $\mathbf{D}^{(k)}$ and $\mathbf{R}^{(k)}$ properly defined in (5.47) and (5.45) respectively and the source term integrated as described in (5.22), when necessary. The integral on the left hand side of (6.28) can be split considering a wave structure given by $\lambda^1 \leq 0 \leq \lambda^2$ and with $-x_L < \lambda^1 T^*$ and $x_R > \lambda^2 T^*$ and considering the solution composed of four constant states as shown in Figure 6.2, it yields

$$\begin{aligned}\int_{-x_L}^{x_R} \partial_t^{(k)}\mathbf{U}(x,T^*)\,dx = \quad & \mathbf{D}^{(k)}_{i_R}(\lambda^1 T^* + x_L) + \mathbf{D}^{(k)}_{(i+1)_L}(x_R - \lambda^2 T^*) + \\ & \mathbf{U}^{-,(k)}_{i_R}(-\lambda^1 T^*) + \mathbf{U}^{+,(k)}_{(i+1)_L}(\lambda^2 T^*),\end{aligned} \tag{6.46}$$

where

$$\mathbf{U}^{-,(k)}_{i_R} = \lim_{x\to 0^-}\left[\partial_t^{(k)}\mathbf{U}(x,t)\right]_{t=T^*} \qquad \mathbf{U}^{+,(k)}_{(i+1)_L} = \lim_{x\to 0^+}\left[\partial_t^{(k)}\mathbf{U}(x,t)\right]_{t=T^*} \tag{6.47}$$

are the values of the solution for temporal derivatives at each side of the interface, as depicted in Figure 6.2. Notice that wave speeds λ^1 and λ^1 are considered constant and are yet to be defined.

Now, substitution of (6.46) in (6.45) leads to

$$\left(\mathbf{D}^{(k)}_{i_R} - \mathbf{U}^{-,(k)}_{i_R}\right)\lambda^1 - \left(\mathbf{D}^{(k)}_{(i+1)_L} - \mathbf{U}^{+,(k)}_{(i+1)_L}\right)\lambda^2 + \mathbf{R}^{(k)}_{(i+1)_L} - \mathbf{R}^{(k)}_{i_R} = \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}}, \tag{6.48}$$

where an extra condition is needed in order to derive the expression for $\mathbf{U}^{-,(k)}_{i_R}$ and $\partial_t^{(k)}\mathbf{U}^{+}_{(i+1)_L}$, due to the presence of the source term. Before introducing this condition, let us define first an approximate function for the derivatives of the flux, denoted by $\partial_t^{(k)}\hat{\mathbf{F}}(x,t)$, with a similar structure than $\partial_t^{(k)}\mathbf{U}(x,t)$ as depicted in Figure 6.2. In this case, also intercell values for the fluxes can be defined at both sides of the t axis as

$$\mathbf{F}^{-,(k)}_{i_R} = \lim_{x\to 0^-}\left[\partial_t^{(k)}\hat{\mathbf{F}}(x,t)\right]_{t=T^*}, \qquad \mathbf{F}^{+,(k)}_{(i+1)_L} = \lim_{x\to 0^+}\left[\partial_t^{(k)}\hat{\mathbf{F}}(x,t)\right]_{t=T^*}. \tag{6.49}$$

For that purpose, the following RH relations across waves between derivatives of fluxes and conserved variables are stated

$$\mathbf{F}^{-,(k)}_{i_R} - \mathbf{R}^{(k)}_{i_R} = \lambda^1\left(\mathbf{U}^{-,(k)}_{i_R} - \mathbf{D}^{(k)}_{i_R}\right), \tag{6.50}$$

$$\mathbf{R}^{(k)}_{(i+1)_L} - \mathbf{F}^{+,(k)}_{(i+1)_L} = \lambda^2\left(\mathbf{D}^{(k)}_{(i+1)_L} - \mathbf{U}^{+,(k)}_{(i+1)_L}\right), \tag{6.51}$$

$$\mathbf{F}^{+,(k)}_{(i+1)_L} - \mathbf{F}^{-,(k)}_{i_R} - \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} = 0. \tag{6.52}$$

As mentioned before, velocities λ^1 and λ^2 were considered constant but have not been defined yet. The proposed approach is to assign the same velocities that those for the leading term, obtained from the approximate Jacobian evaluated at both sides of the discontinuity at $t = 0$, $\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = \widetilde{\mathbf{J}}(\mathbf{U}^{(0)}_{i_R}, \mathbf{U}^{(0)}_{(i+1)_L})$, using Roe's technique. Under this assumption, it is possible to derive the following relation from (6.36) as

$$\mathbf{F}^{+,(k)}_{(i+1)_L} - \mathbf{F}^{-,(k)}_{i_R} = \widetilde{\mathbf{J}}^{(0)}_{i+\frac{1}{2}}\left(\mathbf{U}^{+,(k)}_{(i+1)_L} - \mathbf{U}^{-,(k)}_{i_R}\right), \tag{6.53}$$

since derivatives of the Jacobian matrix are neglected.

Combining (6.52) and (6.53), the following relation appears

$$\bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} = \widetilde{\mathbf{J}}^{(0)}_{i+\frac{1}{2}} \left(\mathbf{U}^{+,(k)}_{(i+1)_L} - \mathbf{U}^{-,(k)}_{i_R} \right) \tag{6.54}$$

and using (6.6), it is possible to write the jump on the conserved variables across the stationary wave at $x = 0$ as

$$\mathbf{U}^{+,(k)}_{(i+1)_L} - \mathbf{U}^{-,(k)}_{i_R} = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}} \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} = \bar{\mathbf{H}}^{(k)}_{i+\frac{1}{2}}, \tag{6.55}$$

where $\widetilde{\mathbf{P}}$ is a matrix composed by the eigenvectors of the Jacobian that leads to the following diagonalization

$$\widetilde{\mathbf{J}}^{(0)}_{i+\frac{1}{2}} = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}}, \tag{6.56}$$

with $\widetilde{\mathbf{\Lambda}}$ a diagonal matrix composed by the eigenvalues of the Jacobian.

The expression for the intercell fluxes can be obtained by combining (6.55) with (6.48) and using the RH relations stated before

$$\mathbf{F}^{-,(k),sub}_{i_R} = \frac{\lambda^2 \mathbf{R}^{(k)}_{i_R} - \lambda^1 \mathbf{R}^{(k)}_{(i+1)_L} + \lambda^1\lambda^2 \delta\mathbf{D}^{(k)}_{i+\frac{1}{2}} + \lambda^1 \left(\bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} - \lambda^2 \bar{\mathbf{H}}^{(k)}_{i+\frac{1}{2}} \right)}{\lambda^2 - \lambda^1}, \tag{6.57}$$

$$\mathbf{F}^{+,(k),sub}_{(i+1)_L} = \frac{\lambda^2 \mathbf{R}^{(k)}_{i_R} - \lambda^1 \mathbf{R}^{(k)}_{(i+1)_L} + \lambda^1\lambda^2 \delta\mathbf{D}^{(k)}_{i+\frac{1}{2}} + \lambda^2 \left(\bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} - \lambda^1 \bar{\mathbf{H}}^{(k)}_{i+\frac{1}{2}} \right)}{\lambda^2 - \lambda^1}. \tag{6.58}$$

High order terms of the numerical fluxes in (5.25) are finally calculated using the following approximate flux that considers all possible wave structures

$$\mathbf{F}^{-,(k)}_{i_R} = \begin{cases} \mathbf{R}^{(k)}_{i_R} & \text{if} \quad \lambda^1 \geq 0 \\ \mathbf{F}^{-,(k),sub}_{i_R} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \mathbf{R}^{(k)}_{(i+1)_L} - \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.59}$$

$$\mathbf{F}^{+,(k)}_{(i+1)_L} = \begin{cases} \mathbf{R}^{(k)}_{i_R} + \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} & \text{if} \quad \lambda^1 \geq 0 \\ \mathbf{F}^{+,(k),sub}_{(i+1)_L} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \mathbf{R}^{(k)}_{(i+1)_L} & \text{if} \quad \lambda^2 \leq 0 \end{cases}. \tag{6.60}$$

Godunov's updating scheme can also be expressed in terms of contributions, namely those corresponding to RPs at the interfaces plus the contribution due to the variations of the variables inside the cell.

Fluctuations corresponding to the interfaces are next presented. After some algebraic manipulation of the equations, the following expressions for the contributions associated to the right interface are obtained for $\lambda^1 \leq 0 \leq \lambda^2$

$$\delta\mathbf{M}^{-,(0),sub}_{i+\frac{1}{2}} = \frac{-\lambda^1\delta\mathbf{F}^{(0)}_{i+\frac{1}{2}} + \lambda^1\lambda^2\delta\mathbf{U}^{(0)}_{i+\frac{1}{2}} + \lambda^1\left(\bar{\mathbf{S}}^{(0)}_{i+\frac{1}{2}} - \lambda^2\bar{\mathbf{H}}^{(0)}_{i+\frac{1}{2}}\right)}{\lambda^2-\lambda^1},$$
$$\delta\mathbf{M}^{-,(k),sub}_{i+\frac{1}{2}} = \left[\frac{-\lambda^1\delta\mathbf{R}^{(k)}_{i+\frac{1}{2}} + \lambda^1\lambda^2\delta\mathbf{D}^{(k)}_{i+\frac{1}{2}} + \lambda^1\left(\bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} - \lambda^2\bar{\mathbf{H}}^{(k)}_{i+\frac{1}{2}}\right)}{\lambda^2-\lambda^1}\right]\frac{\Delta t^k}{(k+1)!}. \tag{6.61}$$

Expressions for the contributions associated to the left interface are presented next, for $\lambda^1 \leq 0 \leq \lambda^2$

$$\delta\mathbf{M}^{+,(0),sub}_{i-\frac{1}{2}} = -\frac{-\lambda^2\delta\mathbf{F}^{(0)}_{i-\frac{1}{2}} + \lambda^1\lambda^2\delta\mathbf{U}^{(0)}_{i-\frac{1}{2}} + \lambda^2\left(\bar{\mathbf{S}}^{(0)}_{i-\frac{1}{2}} - \lambda^1\bar{\mathbf{H}}^{(0)}_{i-\frac{1}{2}}\right)}{\lambda^2-\lambda^1},$$
$$\delta\mathbf{M}^{+,(k),sub}_{i-\frac{1}{2}} = -\left[\frac{-\lambda^2\delta\mathbf{R}^{(k)}_{i-\frac{1}{2}} + \lambda^1\lambda^2\delta\mathbf{D}^{(k)}_{i-\frac{1}{2}} + \lambda^2\left(\bar{\mathbf{S}}^{(k)}_{i-\frac{1}{2}} - \lambda^1\bar{\mathbf{H}}^{(k)}_{i-\frac{1}{2}}\right)}{\lambda^2-\lambda^1}\right]\frac{\Delta t^k}{(k+1)!}. \tag{6.62}$$

As done for the fluxes, all possible combinations of wave speeds must be considered. The previous expressions provide the incoming wave contributions under subcritical flow regime but we also need to consider supercritical flow, that is $\lambda^1 > 0$ or $\lambda^2 < 0$. The complete expression for the fluctuations at the interfaces reads

$$\delta\mathbf{M}^{-,(0)}_{i+1/2} = \begin{cases} 0 & \text{if} \quad \lambda^1 \geq 0 \\ \delta\mathbf{M}^{-,(0),sub}_{i+1/2} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \delta\mathbf{F}^{(0)}_{i+1/2} - \bar{\mathbf{S}}^{(0)}_{i+1/2} & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.63}$$

$$\delta\mathbf{M}^{-,(k)}_{i+1/2} = \begin{cases} 0 & \text{if} \quad \lambda^1 \geq 0 \\ \delta\mathbf{M}^{-,(k),sub}_{i+1/2} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \left(\delta\mathbf{R}^{(k)}_{i+1/2} - \bar{\mathbf{S}}^{(k)}_{i+1/2}\right)\frac{\Delta t^k}{(k+1)!} & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.64}$$

$$\delta\mathbf{M}^{+,(0)}_{i-1/2} = \begin{cases} \delta\mathbf{F}^{(0)}_{i-1/2} - \bar{\mathbf{S}}^{(0)}_{i-1/2} & \text{if} \quad \lambda^1 \geq 0 \\ \delta\mathbf{M}^{+,(0),sub}_{i-1/2} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ 0 & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.65}$$

$$\delta\mathbf{M}^{+,(k)}_{i-1/2} = \begin{cases} \left(\delta\mathbf{R}^{(k)}_{i-1/2} - \bar{\mathbf{S}}^{(k)}_{i-1/2}\right)\frac{\Delta t^k}{(k+1)!} & \text{if} \quad \lambda^1 \geq 0 \\ \delta\mathbf{M}^{+,(k),sub}_{i-1/2} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ 0 & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.66}$$

Under steady conditions, all contributions must become nil in order to preserve the equilibrium of the numerical solution and achieve the steady regime. A proper factorization of the equations provided by the CK procedure allows to express temporal derivatives as a sum of factors multiplied by spatial derivatives of certain variables that are constant in space under steady regime. In this way, all temporal derivatives are enforced to be zero in steady state and therefore it is possible to affirm

$$\delta\mathbf{M}^{(k)}_{i_R,i_L} = \delta\mathbf{M}^{-,(k)}_{i+1/2} = \delta\mathbf{M}^{+,(k)}_{i-1/2} = 0\,. \tag{6.67}$$

On the other hand, contributions associated to the leading term, $\delta\mathbf{M}^{+,(0)}_{i-1/2}$ and $\delta\mathbf{M}^{-,(0)}_{i+1/2}$, will become zero if a proper equilibrium between sources and fluxes is guaranteed.

6.3.1 Linear approach for the high order terms

When using the LFS method, RPs corresponding to the linearized equations of evolution for the derivatives must be computed. The resolution of such problems using the HLLS solver is addressed in this section. The LFS Derivative Riemann solver is constructed now in combination with the HLLS-ADER solver. Recall that this strategy only requires the reconstruction of time derivatives of the conserved quantities at the interfaces, unlike the previous approach that also requires time derivatives of the fluxes.

The integral form of (6.3) inside a control volume $[-x_L, x_R] \times [0, T^*]$ is expressed as

$$\int_{-x_L}^{x_R} \partial_t^{(k)} \mathbf{U}(x, T^*)\, dx = x_R \mathbf{D}^{(k)}_{(i+1)_L} + x_L \mathbf{D}^{(k)}_{i_R} + \widetilde{\mathbf{J}}_{i+1/2}(\mathbf{D}^{(k)}_{i_R} - \mathbf{D}^{(k)}_{(i+1)_L}) T^* + \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} T^* \tag{6.68}$$

and the integral on the left hand side of (6.68) can be expressed as a sum of four constant states as done in (6.46). Substitution of (6.46) in (6.68) leads to an expression which is equivalent to (6.48) and reads

$$\left(\mathbf{D}^{(k)}_{i_R} - \mathbf{U}^{-,(k)}_{i_R}\right)\lambda^1 - \left(\mathbf{D}^{(k)}_{(i+1)_L} - \mathbf{U}^{+,(k)}_{(i+1)_L}\right)\lambda^2 + \widetilde{\mathbf{J}}_{i+1/2}\mathbf{D}^{(k)}_{(i+1)_L} - \widetilde{\mathbf{J}}_{i+1/2}\mathbf{D}^{(k)}_{i_R} = \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}}\,. \tag{6.69}$$

As in the original case, an extra condition is needed in order to derive the expression for $\mathbf{U}^{-,(k)}_{i_R}$ and $\mathbf{U}^{+,(k)}_{(i+1)_L}$, due to the presence of the source term. Such condition was derived before and presented in Equation (6.55).

The expression for the intercell fluxes, when $\lambda^1 \leq 0 \leq \lambda^2$, reads

$$\mathbf{F}^{-,(k),sub}_{i_R} = \frac{\widetilde{\mathbf{J}}_{i+1/2}\left(\lambda^2 \mathbf{D}^{(k)}_{i_R} - \lambda^1 \mathbf{D}^{(k)}_{(i+1)_L}\right) + \lambda^1\lambda^2\delta\mathbf{D}^{(k)}_{i+\frac{1}{2}} + \lambda^1\left(\bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} - \lambda^2 \bar{\mathbf{H}}^{(k)}_{i+\frac{1}{2}}\right)}{\lambda^2 - \lambda^1}, \tag{6.70}$$

$$\mathbf{F}^{+,(k),sub}_{(i+1)_L} = \frac{\widetilde{\mathbf{J}}_{i+1/2}\left(\lambda^2 \mathbf{D}^{(k)}_{i_R} - \lambda^1 \mathbf{D}^{(k)}_{(i+1)_L}\right) + \lambda^1\lambda^2\delta\mathbf{D}^{(k)}_{i+\frac{1}{2}} + \lambda^2\left(\bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} - \lambda^1 \bar{\mathbf{H}}^{(k)}_{i+\frac{1}{2}}\right)}{\lambda^2 - \lambda^1}. \tag{6.71}$$

The high order terms of the numerical fluxes in (5.25) are finally calculated using the following approximate flux that considers all possible wave structures

$$\mathbf{F}^{-,(k)}_{i_R} = \begin{cases} \widetilde{\mathbf{J}}_{i+1/2}\mathbf{D}^{(k)}_{i_R} & \text{if} \quad \lambda^1 \geq 0 \\ \mathbf{F}^{-,(k),sub}_{i_R} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \widetilde{\mathbf{J}}_{i+1/2}\mathbf{D}^{(k)}_{(i+1)_L} - \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.72}$$

$$\mathbf{F}^{+,(k)}_{(i+1)_L} = \begin{cases} \widetilde{\mathbf{J}}_{i+1/2}\mathbf{D}^{(k)}_{i_R} + \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} & \text{if} \quad \lambda^1 \geq 0 \\ \mathbf{F}^{+,(k),sub}_{(i+1)_L} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \widetilde{\mathbf{J}}_{i+1/2}\mathbf{D}^{(k)}_{(i+1)_L} & \text{if} \quad \lambda^2 \leq 0 \end{cases}. \tag{6.73}$$

The high order contributions of the incoming waves at the interfaces when considering subcritical flow regime are computed as

$$\delta \mathbf{M}^{-,(k),sub}_{i+1/2} =$$

$$\left[\frac{-\lambda^1 \widetilde{\mathbf{J}}_{i+1/2} \delta \mathbf{D}^{(k)}_{i+\frac{1}{2}} + \lambda^1 \lambda^2 \delta \mathbf{D}^{(k)}_{i+\frac{1}{2}} + \lambda^1 \left(\bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} - \lambda^2 \bar{\mathbf{H}}^{(k)}_{i+\frac{1}{2}} \right)}{\lambda^2 - \lambda^1} \right] \frac{\Delta t^k}{(k+1)!}, \tag{6.74}$$

$$\delta \mathbf{M}^{+,(k),sub}_{i-1/2} =$$

$$- \left[\frac{-\lambda^2 \widetilde{\mathbf{J}}_{i+1/2} \delta \mathbf{D}^{(k)}_{i-\frac{1}{2}} + \lambda^1 \lambda^2 \delta \mathbf{D}^{(k)}_{i-\frac{1}{2}} + \lambda^2 \left(\bar{\mathbf{S}}^{(k)}_{i-\frac{1}{2}} - \lambda^1 \bar{\mathbf{H}}^{(k)}_{i-\frac{1}{2}} \right)}{\lambda^2 - \lambda^1} \right] \frac{\Delta t^k}{(k+1)!} \tag{6.75}$$

and more generally, when considering all wave speed combinations, the resulting fluctuations read

$$\delta \mathbf{M}^{-,(k)}_{i+1/2} = \begin{cases} 0 & \text{if} \quad \lambda^1 \geq 0 \\ \delta \mathbf{M}^{-,(k),sub}_{i+1/2} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ \left(\widetilde{\mathbf{J}}_{i+1/2} \delta \mathbf{D}^{(k)}_{i+\frac{1}{2}} - \bar{\mathbf{S}}^{(k)}_{i+\frac{1}{2}} \right) \frac{\Delta t^k}{(k+1)!} & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.76}$$

$$\delta \mathbf{M}^{+,(k)}_{i-1/2} = \begin{cases} \left(\widetilde{\mathbf{J}}_{i+1/2} \delta \mathbf{D}^{(k)}_{i-\frac{1}{2}} - \bar{\mathbf{S}}^{(k)}_{i-\frac{1}{2}} \right) \frac{\Delta t^k}{(k+1)!} & \text{if} \quad \lambda^1 \geq 0 \\ \delta \mathbf{M}^{+,(k),sub}_{i-1/2} & \text{if} \quad \lambda^1 \leq 0 \leq \lambda^2 \\ 0 & \text{if} \quad \lambda^2 \leq 0 \end{cases}, \tag{6.77}$$

6.4 Concluding remarks

The highlights of this chapter are listed below:

- A novel methodology for the resolution of the DRP is presented. The proposed methods are called FS and LFS solvers. The highlights of such solvers are listed below:
  - The DRP is decomposed in traditional RPs, one for the evolution equations and others for the time derivatives of the evolution equations.
  - The IC for the RPs associated to the derivatives are time derivatives at $t = 0$ provided by the CK procedure, which include source terms.
  - The DRP is defined including the source term at the interface. Only geometric source terms must be taken into account.
  - Traditional Riemann solvers are required to solve the single RPs. When choosing the LFS solver, RPs defined for the high order terms are linear and Riemann solvers are not required for their resolution. In spite of this, we propose to use the same solver chosen for the DRP_0, that will provide the exact solution of the k-th RP and the same treatment of the source term as in the leading term.
  - The FS solver provides a higher accuracy when compared to the LFS solver and allows a more relaxed time step restriction. On the other hand, the computational cost is higher. This is evidenced using some numerical examples in the following chapters.
- An arbitrary order extension of the ARoe solver is proposed, following the FS and LFS methodology. The novel solver provides an exact equilibrium between sources and fluxes and approaches the numerical flux with arbitrary order of accuracy. The resulting solver is called AR-(L)FS solver and is a complete, linear solver.

- An arbitrary order extension of the HLLS solver is proposed, following the FS and LFS methodology. The novel solver provides an exact equilibrium between sources and fluxes and approaches the numerical flux with arbitrary order of accuracy. The resulting solver is called HLLS-(L)FS solver and is a 2-wave, nonlinear solver. If applied to 3-wave (or more) problems, such as the 2D SWE, it provides a poorer resolution of shear waves than the ARoe solver.

7 2D ADER SCHEMES FOR SYSTEMS OF CONSERVATION LAWS

This chapter is devoted to the numerical resolution of hyperbolic conservation laws in 2 space dimensions using the ADER approach. The basic formulation of the methods is outlined here and a more detailed insight into the numerical techniques used for each particular problem is provided in the following chapters.

In Section 7.1, the WENO-ADER scheme presented in Chapter 5 is extended to 2 dimensions. The scheme is derived for Cartesian grids and two strategies for the computation of the numerical fluxes are detailed: the resolution of the interface-normal DRP and the resolution of the x-split DRP using rotation of the variables.

In Section 7.2, a radically different numerical scheme, also using the ADER methodology, is detailed. Such scheme is called DG method and can be seen as a compromise between Finite Elements and FV. The reason for the study of the DG method in this thesis is to be compared to WENO-ADER schemes, as both families of methods offer competitive capabilities when solving hyperbolic problems.

7.1 WENO-ADER scheme in 2D Cartesian grid

Let us consider the system of conservation laws in (2.5) to compose the following Initial Boundary Value Problem (IVBP)

$$\begin{cases} \text{PDEs:} & \dfrac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathbf{E}(\mathbf{U}) = \mathbf{S} \\ \text{IC:} & \mathbf{U}(\mathbf{x},0) = \mathring{\mathbf{U}}(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega \\ \text{BC:} & \mathbf{U}(\mathbf{x},t) = \mathbf{U}_{\partial\Omega}(\mathbf{x},t) \quad \forall \mathbf{x} \in \partial\Omega \end{cases} \tag{7.1}$$

defined in the domain $\Omega \times [0,T]$, where $\Omega = [a,b] \times [c,d]$ is the spatial domain. Note that the initial condition is given by $\mathring{\mathbf{U}}(\mathbf{x})$ and the boundary condition by $\mathbf{U}_{\partial\Omega}(\mathbf{x},t)$. The spatial domain is discretized in $N_x \times N_y$ volume cells, defined as $\Omega_{ij} \subseteq \Omega$, such that $\Omega = \bigcup_{i,j=1}^{N} \Omega_{ij}$, with cell edges at

$$a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \ldots < x_{N_x-\frac{1}{2}} < x_{N_x+\frac{1}{2}} = b\,, \tag{7.2}$$

and

$$c = y_{\frac{1}{2}} < y_{\frac{3}{2}} < \ldots < y_{N_y-\frac{1}{2}} < y_{N_y+\frac{1}{2}} = d\,. \tag{7.3}$$

Cells and cell are sizes defined as

$$\Omega_{ij} = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \times \left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right], \qquad i = 1, ..., N_x,\ j = 1, ..., N_y \tag{7.4}$$

and

$$\vartheta_{ij} = (x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}) \cdot (y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}), \qquad i = 1, ..., N_x,\ j = 1, ..., N_y\,, \tag{7.5}$$

respectively, and in the case of regular grid we have $\vartheta_{ij} = \Delta x^2$.

Inside each cell at time $t^n = \sum_{l=1}^{n} \Delta t_l$, with Δt_l the time step dynamically calculated, the conserved quantities are defined as cell averages as

$$\mathbf{U}_{ij}^n = \frac{1}{\vartheta_{ij}} \int_{\Omega_{ij}} \mathbf{U}(\mathbf{x}, t^n) dV \qquad i = 1, ..., N_x,\ j = 1, ..., N_y\,. \tag{7.6}$$

Let us consider again the system in (2.5) and integrate it over the discrete domain $\Omega_{ij} \times \Delta t$, where $\Delta t = t^{n+1} - t^n$. Application of the Gauss-Ostrogradsky theorem yields

$$\mathbf{U}_{ij}^{n+1} = \mathbf{U}_{ij}^n - \frac{1}{\vartheta_{ij}} \int_0^{\Delta t} \int_{\partial\Omega_{ij}} \mathbf{E}\hat{\mathbf{n}} dS dt + \frac{1}{\vartheta_{ij}} \int_0^{\Delta t} \int_{\Omega_{ij}} \mathbf{S} dV dt\,. \tag{7.7}$$

If considering a regular Cartesian grid, all cells have a constant cell area Δx^2 and we obtain the following fully-discrete updating formula

$$\mathbf{U}_{ij}^{n+1} = \mathbf{U}_{ij}^n - \frac{\Delta t}{\Delta x^2} \left(\sum_{r=1}^{4} \mathscr{F}_r^- - \bar{\mathbf{S}}_{ij} \right), \tag{7.8}$$

where

$$\bar{\mathbf{S}}_{ij} \approx \frac{1}{\Delta t} \int_0^{\Delta t} \int_{x_{i+1/2}}^{x_{i-1/2}} \int_{y_{i+1/2}}^{y_{i-1/2}} \mathbf{S}\, dy dx\, d\tau\,, \tag{7.9}$$

is the approximation of the space-time integral of the source term inside the cell and $\mathscr{F}_r^-$ is the space-time integral of the numerical fluxes over the r-th cell edges. To construct a numerical scheme of order $K+1$-th, it is sufficient to approximate the integral of the flux, $\mathscr{F}_r^-$, using a $K+1$-th order Gaussian quadrature rule as

$$\mathscr{F}_r^- = \frac{\Delta x}{2} \sum_{q=1}^{k} w_q \mathscr{F}_{r,q}^-\,, \tag{7.10}$$

where w_l are the Gaussian weights inside the interval $[-1, 1]$ at the $q = 1, ..., k$ quadrature points along the cell edge and $\mathscr{F}_{r,q}^-$ the numerical fluxes at each of these points, computed by means of the resolution of a 1D approximation to the Cauchy problem with at least K non-trivial derivatives, to ensure high order not only in space but also in time. It is worth recalling that the use of k quadrature points for the Gaussian integration allows to construct a $2k-1$-th order approximation of the integral.

An arbitrary order approach to the Cauchy problem is given by the DRP_K, that is an IVP defined by a system of N_λ EDPs and an initial condition consisting of piecewise polynomial data with K nontrivial derivatives, separated by a single discontinuity at $x = 0$. The formulation and resolution of the DRP is purely one-dimensional, in the normal direction to cell interfaces. The formulation of the DRP at cell interfaces can be done by following two different approaches:

- Construct the normal DRP by projecting the fluxes, without rotation of the variables:

  To construct the normal DRP, we formulate the evolution equation for the conserved variables in the normal direction to the cell interface by carrying out a rotation of the axes $(x, y) \rightarrow (\check{x}, \check{y})$ as follows

$$\begin{pmatrix} \check{x} \\ \check{y} \end{pmatrix} = \begin{pmatrix} n_x & n_y \\ -n_y & n_x \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{7.11}$$

  yielding the following 1D DRP [84] (Figure 5.2)

$$\begin{cases} \dfrac{\partial \mathbf{U}}{\partial t} + \dfrac{\partial \mathscr{F}(\mathbf{U})}{\partial \check{x}} = \mathbf{S} \\ \\ \mathbf{U}(\check{x}, t=0) = \begin{cases} \mathbf{U}_{\xi}(\check{x}, \check{y}=0) & \check{x} < 0 \\ \mathbf{U}_{\xi+1}(\check{x}, \check{y}=0) & \check{x} > 0 \end{cases} \end{cases} \tag{7.12}$$

  where $\mathbf{U}_{\xi}(\check{x}, \check{y})$ and $\mathbf{U}_{\xi+1}(\check{x}, \check{y})$ are smooth functions of the position, defined using suitable reconstruction procedures, such as the WENO method, and $\mathscr{F} = \mathbf{E} \cdot \hat{\mathbf{n}}$ is the projection of the fluxes onto the normal direction to the edge, $\hat{\mathbf{n}}$. Such projection is computed as $\mathbf{E} \cdot \hat{\mathbf{n}} = \mathbf{F} \cdot n_x + \mathbf{G} \cdot n_y$. It is worth noting that the source term is included in the definition of the DRP, according to [84].

  Solution for the DRP in (7.12) can be constructed using the flux expansion approach as

$$\mathscr{F}^{-}_{r,q} = \mathscr{F}^{-,0}_{\xi_R} + \sum_{k=1}^{K} \mathscr{F}^{-,(k)}_{\xi_R} \frac{\Delta t^k}{(k+1)!}, \qquad \mathscr{F}^{+}_{r,q} = \mathscr{F}^{+,0}_{(\xi+1)_L} + \sum_{k=1}^{K} \mathscr{F}^{+,(k)}_{(\xi+1)_L} \frac{\Delta t^k}{(k+1)!}, \tag{7.13}$$

  where $\mathscr{F}^{-,(0)}_{\xi_R}$, $\mathscr{F}^{+,(0)}_{(\xi+1)_L}$, $\mathscr{F}^{-,(k)}_{\xi_R}$ and $\mathscr{F}^{+,(k)}_{(\xi+1)_L}$ are computed by solving the DRP_K.

- Construct the rotated DRP. We must formulate the evolution equation for the conserved variables in the x-direction and rotate the variables instead of projecting the fluxes. This approach is possible thanks to the rotating invariance property of the system, which reads

$$\mathbf{F} \cdot n_x + \mathbf{G} \cdot n_y = \mathbf{R}^{-1} \mathbf{F}(\mathbf{RU}) \tag{7.14}$$

  with

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & n_x & n_y \\ 0 & -n_y & n_x \end{pmatrix} \tag{7.15}$$

  the rotation matrix for a system with a scalar equation plus a vector equation in $\mathbb{R}^2$. If rotating the system of equations to the x-axis (clockwise) by doing

$$\frac{\partial \mathbf{RU}}{\partial t} + \frac{\partial \mathbf{R}(\mathbf{F} \cdot n_x + \mathbf{G} \cdot n_y)}{\partial x} = \mathbf{RS} \tag{7.16}$$

  and defining $\mathbf{V} = \mathbf{RU}$ and $\mathbf{T} = \mathbf{RS}$, we have the following DRP

$$\begin{cases} \dfrac{\partial \mathbf{V}}{\partial t} + \dfrac{\partial \mathbf{F}(\mathbf{V})}{\partial x} = \mathbf{T} \\ \\ \mathbf{V}(x, t=0) = \begin{cases} \mathbf{V}_{\xi}(x, y=0) & x < 0 \\ \mathbf{V}_{\xi+1}(x, y=0) & x > 0 \end{cases} \end{cases} \tag{7.17}$$

  where $\mathbf{V}_{\xi}(x, y)$ and $\mathbf{V}_{\xi+1}(x, y)$ are the reconstructions at each side of the edge.

The solution for the DRP in (7.17) can be constructed using the flux expansion approach as

$$\begin{aligned}\mathscr{F}^{-}_{r,q} &= \mathbf{R}^{-1}\mathbf{F}^{-,0}_{\xi_R} + \sum_{k=1}^{K}\mathbf{R}^{-1}\mathbf{F}^{-,(k)}_{\xi_R}\frac{\Delta t^k}{(k+1)!}\,,\\ \mathscr{F}^{+}_{r,q} &= \mathbf{R}^{-1}\mathbf{F}^{+,0}_{(\xi+1)_L} + \sum_{k=1}^{K}\mathbf{R}^{-1}\mathbf{F}^{+,(k)}_{(\xi+1)_L}\frac{\Delta t^k}{(k+1)!}\,,\end{aligned} \tag{7.18}$$

where $\mathbf{F}^{-,(0)}_{\xi_R}$, $\mathbf{F}^{+,(0)}_{(\xi+1)_L}$, $\mathbf{F}^{-,(k)}_{\xi_R}$ and $\mathbf{F}^{+,(k)}_{(\xi+1)_L}$ are computed by solving the DRP$_K$.

In this work, we only give the details for the resolution of the DRP$_K$ following the first approach. Te computation of the DRP in (7.17) is analogous.

The time step Δt is computed dynamically and is bounded by the maximum wave celerity (maximum eigenvalue of the Jacobian) and the cell size. In this way, the time step in a WENO-ADER method is computed as

$$\Delta t \leq \frac{\Delta x}{d\,|\lambda_{max}|}\,, \tag{7.19}$$

where d the number of spatial dimensions (in this case, 2) and λ_{max} the maximum eigenvalue. According to the traditional definition of CFL number

$$\Delta t = CFL\frac{\Delta x}{\lambda_{max}}\,, \tag{7.20}$$

we have that

$$0 < CFL \leq \frac{1}{2} \tag{7.21}$$

for a WENO-ADER scheme in a 2D Cartesian grid.

7.2 Discontinuous Galerkin ADER scheme

The DG scheme can be seen as a compromise between Finite Elements, when looking at the weak formulation and projection of the solution, and FV, as the basis functions are defined cell-wise and the concept of a numerical flux is required. When using the ADER approach, DRPs are defined and solved at cell interfaces.

In this section, the DG-ADER method is described, following [126]. In Subsection 1, the general formulation of the method is provided and in Subsection 2, the computation of the polynomial reconstruction basis is detailed.

7.2.1 General formulation of the one-step DG method

Let us consider the problem in (7.1), with $\mathbf{S} = 0$. The spatial domain, Ω, is discretized in N volume cells, denoted by Ω_i, such that $\Omega = \bigcup_{i=1}^{N}\Omega_i$. The DG method can be obtained by applying a traditional Galerkin projection method to each element. Inside each element, the solution is approximated by a linear combination of basis functions $\{\phi_l\}_{l=0,\dots,N_d}$ as follows

$$\mathbf{U}(\mathbf{x},t) \approx \mathbf{U}_i(\mathbf{x},t) = \sum_{l=0}^{N_d}\hat{\mathbf{U}}_{i,l}(t)\phi_l(\mathbf{x})\,, \qquad \forall\mathbf{x}\in\Omega_i\,, \tag{7.22}$$

where $\hat{\mathbf{U}}_{i,l}(t)$ is the l-th degree of freedom and N_d+1 is the number of degrees of freedom on each element.

It is worth noticing that the basis functions only depends upon space and the degrees of freedom are function of time only.

Projecting the governing equation onto each element of the basis set, the weak form of the problem can be written as

$$\int_{\Omega_i} \phi_k \frac{\partial \mathbf{U}}{\partial t} d\Omega + \int_{\Omega_i} \phi_k \nabla \cdot \mathbf{E} d\Omega = 0\,, \tag{7.23}$$

for $k = 0, ..., N_d$, and integrating by parts the second term of (7.23) using Green's first identity

$$\int_{\Omega_i} \phi_k \nabla \cdot \mathbf{E} d\Omega = \int_{\Gamma_i} \phi_k \mathbf{E} \cdot \hat{\mathbf{n}} d\Gamma - \int_{\Omega_i} \mathbf{E} \cdot \nabla \phi_k d\Omega\,, \tag{7.24}$$

with $\Gamma_i = \delta\Omega_i$ the cell boundary, we can rewrite Equation (7.23) as

$$\int_{\Omega_i} \phi_k \frac{\partial \mathbf{U}}{\partial t} d\Omega + \int_{\Gamma_i} \phi_k \mathbf{E} \cdot \hat{\mathbf{n}} d\Gamma - \int_{\Omega_i} \mathbf{E} \cdot \nabla \phi_k d\Omega = 0\,, \tag{7.25}$$

for $k = 0, ..., N_d$, where $\mathbf{E} \cdot \hat{\mathbf{n}}$ is the numerical flux at the interface, hereafter denoted by $\mathbf{F}^{-}(\mathbf{U}_L, \mathbf{U}_R)$ and computed from the solution of the DRP at the interface.

Substitution of $\mathbf{U} = \mathbf{U}_i(\mathbf{x}, t)$ in Equation (7.25) yields

$$\sum_{l=0}^{N_d} \left(\int_{\Omega_i} \phi_k \phi_l d\Omega \right) \frac{\partial \hat{\mathbf{U}}_{i,l}}{\partial t} + \int_{\Gamma_i} \phi_k \mathbf{F}^{-} d\Gamma - \int_{\Omega_i} \mathbf{E} \cdot \nabla \phi_k d\Omega = 0\,. \tag{7.26}$$

When choosing a orthogonal set of basis functions $\left\{\phi_0, \phi_1, ..., \phi_{N_d}\right\}$, the following property is satisfied

$$\int_{\Omega_i} \phi_k \phi_l d\Omega = a_l \delta_{kl} = \begin{cases} 0 & \text{if} \quad k \neq l \\ a_k & \text{if} \quad k = l \end{cases}\,, \tag{7.27}$$

for $k, l = 0, ..., N_d$, where δ_{kl} is the Kroenecker delta and $a_k \in \mathbb{R}$ is the value of the integral over Ω_i. Using the result in (7.27), (7.26) becomes a decoupled system of equations for the unknown degrees of freedom $\hat{\mathbf{U}}_{i,k}(t)$ that reads

$$a_k \frac{\partial \hat{\mathbf{U}}_{i,k}}{\partial t} + \int_{\Gamma_i} \phi_k \mathbf{F}^{-} d\Gamma_i - \int_{\Omega_i} \mathbf{E} \cdot \nabla \phi_k d\Omega = 0\,, \tag{7.28}$$

for $k = 0, ..., N_d$. In order to obtain a one-step fully discrete numerical scheme, Equation (7.28) is integrated over the time step $\Delta t = [t^n, t^{n+1}]$ as follows

$$a_k \int_{\Delta t} \frac{\partial \hat{\mathbf{U}}_{i,k}}{\partial t} dt + \int_{\Delta t} \int_{\Gamma_i} \phi_k \mathbf{F}^{-} d\Gamma dt - \int_{\Delta t} \int_{\Omega_i} \mathbf{E} \cdot \nabla \phi_k d\Omega dt = 0\,, \tag{7.29}$$

which allows to solve for the $N_d + 1$ degrees of freedom at t^{n+1}, denoted by $\hat{\mathbf{U}}_{i,k}^{n+1}$, as

$$\hat{\mathbf{U}}_{i,k}^{n+1} = \hat{\mathbf{U}}_{i,k}^{n} - \frac{\Delta t}{a_k} \left(\frac{1}{\Delta t} \int_{\Delta t} \int_{\Gamma_i} \phi_k \mathbf{F}^{-} d\Gamma dt - \frac{1}{\Delta t} \int_{\Delta t} \int_{\Omega_i} \mathbf{E} \cdot \nabla \phi_k d\Omega dt \right), \tag{7.30}$$

where the volume and boundary integrals are computed as:

- Integral over Ω_i: Departing from the volume integral

$$\frac{1}{\Delta t}\int_{\Delta t}\int_{\Omega_i}\mathbf{E}\cdot\nabla\phi_k d\Omega dt \tag{7.31}$$

and taking into account that at an arbitrary point $\mathscr{P}\in\Omega_i$ the physical flux $\mathbf{E}$ can be expressed as

$$\mathbf{E}(\mathbf{x}_{\mathscr{P}},t)=\mathbf{E}_{\mathscr{P}}(t)=\mathbf{E}_{\mathscr{P}}^{(0)}+\sum_{k=1}^{K}\mathbf{E}_{\mathscr{P}}^{(k)}\frac{\tau^k}{k!}, \tag{7.32}$$

where $\mathbf{E}_{\mathscr{P}}^{(0)}=\mathbf{E}(\mathbf{x}_{\mathscr{P}},0)$ and $\mathbf{E}_{\mathscr{P}}^{(k)}=\frac{\partial^k\mathbf{E}}{\partial t^k}(\mathbf{x}_{\mathscr{P}},0)$ is obtained using the CK procedure, the computation of the time integral is straightforward

$$\bar{\mathbf{E}}_{\mathscr{P}}=\frac{1}{\Delta t}\int_{\Delta t}\mathbf{E}_{\mathscr{P}}(t)dt=\mathbf{E}_{\mathscr{P}}^{(0)}+\sum_{k=1}^{K}\mathbf{E}_{\mathscr{P}}^{(k)}\frac{\Delta t^k}{(k+1)!}. \tag{7.33}$$

The notation $\bar{\mathbf{E}}_{\mathscr{P}}$ stands for $\bar{\mathbf{E}}(\mathbf{x}_{\mathscr{P}})$, where it can be noticed that the dependency upon time vanishes after integration in time. Now replacing (7.33) in (7.31) yields

$$\frac{1}{\Delta t}\int_{\Delta t}\int_{\Omega_i}\mathbf{E}\cdot\nabla\phi_k d\Omega dt=\int_{\Omega_i}\bar{\mathbf{E}}_{\mathscr{P}}\cdot\nabla\phi_k d\Omega=, \tag{7.34}$$

which can be integrated in space using Gaussian quadrature as follows

$$\frac{1}{\Delta t}\int_{\Delta t}\int_{\Omega_i}\mathbf{E}\cdot\nabla\phi_k d\Omega dt=\left|\mathbf{J}_{\Omega_i}\right|\sum_{q=1}^{k_1}\omega_q\left(\bar{\mathbf{E}}\cdot\nabla\phi_k\right)_{\mathscr{P}_q}, \tag{7.35}$$

where $\left|\mathbf{J}_{\Omega_i}\right|$ is the determinant of the Jacobian for the change of coordinates between the reference element and the true element, $\mathscr{P}_q$ is the q-th quadrature point and the vector $\nabla\phi_k$ is evaluated as $\nabla\phi_{k,\mathscr{P}_q}=\nabla\phi_k(\mathbf{x}_{\mathscr{P}_q})$.

- Integral over Γ_i: The space integral is split for each of the N_{edg} cell edges

$$\frac{1}{\Delta t}\int_{\Delta t}\int_{\Gamma_i}\phi_k\mathbf{F}^{-}d\Gamma dt=\sum_{r=1}^{N_{edg}}\frac{1}{\Delta t}\int_{\Delta t}\int_{\Gamma_r}\phi_k\mathbf{F}^{-}d\Gamma dt \tag{7.36}$$

where $\Gamma_r\in\Gamma_i$ is the r-th cell edge such that $\Gamma_i=\bigcup_{r=1}^{N_{edg}}\Gamma_r$. Considering that the numerical flux $\mathbf{F}^{-}$ is computed from the DRP at each point $\mathscr{P}\in\Gamma_i$ as a polynomial expansion in time

$$\mathbf{F}_{\mathscr{P}}^{-}(t)=\mathbf{F}_{\mathscr{P}}^{-,(0)}+\sum_{k=1}^{K}\mathbf{F}_{\mathscr{P}}^{-,(k)}\frac{t^k}{k!} \tag{7.37}$$

we can compute the time integral straightforward as

$$\bar{\mathbf{F}}_{\mathscr{P}}^{-}=\frac{1}{\Delta t}\int_{\Delta t}\mathbf{F}_{\mathscr{P}}^{-}(t)dt=\mathbf{F}_{\mathscr{P}}^{-,(0)}+\sum_{k=1}^{K}\mathbf{F}_{\mathscr{P}}^{-,(k)}\frac{\Delta t^k}{(k+1)!} \tag{7.38}$$

and substituting in (7.36) yields

$$\frac{1}{\Delta t}\int_{\Delta t}\int_{\Gamma_i}\phi_k\mathbf{F}^{-}d\Gamma dt=\sum_{r=1}^{N_{edg}}\int_{\Gamma_i}\phi_k\bar{\mathbf{F}}_{\mathscr{P}}^{-}d\Gamma. \tag{7.39}$$

The space integral can be computed using Gaussian quadrature, yielding

$$\frac{1}{\Delta t}\int_{\Delta t}\int_{\Gamma_i}\phi_k\mathbf{F}^{-}d\Gamma dt=\sum_{r=1}^{N_{edg}}\left|\mathbf{J}_{\Gamma_i}\right|\sum_{q=1}^{k_2}w_q\left(\phi_k\bar{\mathbf{F}}^{-}\right)_{\mathscr{P}_{r,q}}\,, \tag{7.40}$$

where $\mathscr{P}_{r,q}$ is a point on the r-th edge and q-th quadrature point and the basis function are evaluated at those points $\phi_{k,\mathscr{P}_{r,q}}=\phi_k(\mathbf{x}_{\mathscr{P}_{r,q}})$.

It is worth pointing out that the time step Δt is computed dynamically and may depend upon the maximum wave celerity (maximum eigenvalue of the Jacobian), the cell size and unlike FV-ADER schemes, also depends upon the degree of the polynomials of the basis. According to some authors, the time step in a DG-RK method is typically computed using [127]

$$\Delta t\leq\frac{l}{d(2k+1)\left|\lambda_{max}\right|}\,, \tag{7.41}$$

where l is the cell size, d the number of spatial dimensions and k the maximum degree of the polynomials. According to other authors, the time step should be calculated as [128]

$$\Delta t\leq\frac{l}{d(k+1)^2}\,. \tag{7.42}$$

As we can see, there is not a universal agreement on the calculation of the time step. In this work, we will consider the expression in (7.41). According to (7.41) and the traditional definition of CFL number

$$\Delta t=CFL\frac{l}{\lambda_{max}} \tag{7.43}$$

with $0<CFL\leq 1$, for the DG scheme we add an additional bound for the CFL as

$$0<CFL\leq\frac{1}{d(2k+1)}\,. \tag{7.44}$$

Equivalence between DG and FV Godunov method

It is worth showing that if choosing only one degree of freedom (first order in space), its corresponding basis function becomes $\phi_0=1$ and the solution is approximated by

$$\mathbf{U}(\mathbf{x},t)\approx\mathbf{U}_i(\mathbf{x},t)=\hat{\mathbf{U}}_{i,0}(t)\,,\qquad\forall\mathbf{x}\in\Omega_i\,, \tag{7.45}$$

Moreover, the set of equations in (7.28) becomes a single equation that reads

$$\mathbf{U}_i^{n+1}=\mathbf{U}_i^{n+1}-\frac{1}{a_k}\int_{\Delta t}\int_{\Gamma_i}\mathbf{F}^{-}d\Gamma dt\,, \tag{7.46}$$

because the second term on the right hand side of (7.28) is nil. If also considering first order in time and noticing that a_0 is the volume of the cell, that is $a_0=\vartheta_i$, it yields

$$\mathbf{U}_i^{n+1}=\mathbf{U}_i^{n+1}-\frac{\Delta t}{\vartheta_i}\int_{\Gamma_i}\mathbf{F}^{-}d\Gamma\,, \tag{7.47}$$

which corresponds to Godunov's first order updating formula.

7.2.2 The choice of a suitable basis: orthogonal basis

Orthogonal basis in 1D

When choosing a orthogonal set of basis functions $\{\phi_0, \phi_1, ..., \phi_{N_d}\}$, the following property is satisfied

$$\int_{\Omega_i} \phi_k \phi_l d\Omega = a_l \delta_{kl} = \begin{cases} 0 & \text{if} \quad k \neq l \\ a_k & \text{if} \quad k = l \end{cases}, \tag{7.48}$$

for $k, l = 0, ..., N_d$, where δ_{kl} is the Kroenecker delta and $a_k \in \mathbb{R}$ is the value of the integral. This property allows to obtain a diagonal mass matrix and decouple the system of equations, preventing from calculating the inverse of the mass matrix for solving the degrees of freedom.

Legendre polynomials are a suitable polynomial basis as they are orthogonal with respect to the L^2 norm, which satisfies (7.48). A recursive formula for Legendre polynomials, hereafter denoted by $P_k(x)$ where k stands for the degree, is given by

$$\begin{cases} P_0(x) = 1 \\ P_1(x) = x \\ P_k(x) = \left(\frac{2k-1}{k}\right) x P_{k-1}(x) - \left(\frac{k-1}{k}\right) P_{k-2}(x) \end{cases}, \tag{7.49}$$

Analogously, a recursive formula for the derivatives of the polynomials, $P_k^{(n)}(x)$, can be obtained

$$\begin{cases} P_0(x) = 1, P_0^{(1)}(x) = 0 \\ P_1(x) = x, P_0^{(1)}(x) = 1 \\ P_k^{(n)}(x) = \left(\frac{2k-1}{k}\right)\left(k P_{k-1}^{(n-1)}(x) x P_{k-1}^{(n)}(x)\right) - \left(\frac{k-1}{k}\right) P_{k-2}^{(n)}(x) \end{cases}, \tag{7.50}$$

Orthogonal basis in 2D for quadrilateral elements

The 2D extension of the 1D Legendre polynomials can be carried out using tensor product surfaces. This way, the construction of a 2D orthogonal basis is straightforward and can be regarded as a dimension-by-dimension reconstruction. The basis functions read

$$\phi_{kl}(x, y) = P_k(x) \cdot P_l(y) \tag{7.51}$$

and fulfill the orthogonality requirement as

$$\int_{\Omega_i} \phi_{kl}(x, y) \phi_{mn}(x, y) d\Omega = a_{kl} \delta_{km} \delta_{ln} = \begin{cases} 0 & \text{if} \quad k \neq m, l \neq n \\ a_{kl} & \text{if} \quad k = m \text{ and } l = n \end{cases}, \tag{7.52}$$

Cross derivatives of $\phi_{kl}(x, y)$ are easily computed as

$$\frac{\partial^n}{\partial x^p \partial x^{n-p}} \phi_{kl}(x, y) = \frac{d^p P_k(x)}{dx^p} \cdot \frac{d^{n-p} P_l(y)}{dy^{n-p}}. \tag{7.53}$$

For the sake of clarity, all combinations of subscripts k and l are remapped onto a new index k, allowing to write the basis functions as $\phi_k(x, y)$.

8 APPLICATION TO LINEAR PROBLEMS

This chapter is devoted to the application of the proposed schemes to different problems of linear nature. Such problems share a common feature: wave celerities do not depend upon the variables of the problem, they can be either constant or dependent on the spatial position. Hence, the phenomena of loss of regularity and shock formation can never occur. From the point of view of the numerical scheme, there is no need of using slope limiting techniques/non-oscillatory reconstruction when the initial data is smooth. Moreover, the Riemann solvers required for the resolution of the (D)RP will be reduced to the algorithm of the analytical solution of the problem. This evidences that linear problems are much less challenging and computationally demanding than nonlinear ones, hence they are a good starting point for testing the numerical schemes.

The linear scalar advection equation with a reaction source term is considered in Section 8.1. Both 1D and 2D problems, with constant and space-dependent advection speeds, are considered. Different WENO reconstruction techniques in combination with a traditional ADER scheme (TT-ADER scheme) are tested. It is worth recalling that the proposed ADER schemes as well as the TT and CT-ADER schemes reduce to the same algorithm when considering linear problems. In Section 8.2, the linear acoustic problem, which is a coupled system of PDEs and is derived from Euler equations, is solved in 2 space dimensions. This problem is used as a benchmark to compare WENO-ADER and DG-ADER schemes. Numerical errors and convergence rates are presented for all problems.

8.1 The linear scalar advection equation

In this section, the numerical resolution of the linear scalar equation in 1 and 2 space dimensions is considered. Such equation can be expressed as

$$\frac{\partial u}{\partial t} + \nabla \cdot (u\boldsymbol{\lambda}) = \zeta u\,, \tag{8.1}$$

where $u = u(x, y, t)$ is a scalar variable, $\nabla = (\partial_x, \partial_y)$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ is the vector of constant advection speeds in the x and y directions and ζ is a constant coefficient that represents the strength of the source term. For 1D cases we will neglect the y component in the previous definition.

8.1.1 1D linear advection-reaction equation

The following initial condition is imposed to Equation (8.1)

$$u(x,0) = (\sin \pi x)^4 \tag{8.2}$$

and it is numerically solved inside $[a,b] \times [0,t] = [0,2] \times [0,2]$, setting CFL=0.45. Cyclic boundary conditions are imposed in all cases. In this test λ is set equal to 1 and parameter ζ of the reactive term is set equal to 5. The transported function will suffer an exponential growth in time and classical first and second order numerical schemes do completely fail when simulating this test case. Very high order is mandatory if accurate solutions are sought.

The TT-ADER scheme is used here to compute the numerical solution and different WENO approaches will be compared. It is worth saying that the implementation of the AR-ADER scheme would lead to the TT-ADER algorithm, due to the linear nature of the problem and non-geometric nature of the source term.

Table 8.1 shows the numerical error and convergence rate measured with L_1 error norm for different grid refinements, for the 3-rd, 5-th, 7-th, 9-th and 11-th order TT-ADER schemes. For all norms explored the TT-ADER scheme in combination with the sub-cell derivative reconstruction proposed in this work, provides the expected rate of convergence. Numerical results for the 3-rd order TT-ADER scheme may seem to reproduce a suboptimal behavior, that can be easily overcome by setting further refinements.

When using the optimal reconstruction instead of a WENO method, as the function is smooth, the scheme achieves the prescribed accuracy for all orders up to 11-th order. It is expected that a powerful WENO reconstruction method must reproduce the same level of error and converge rate. When using WENO-JS method, the numerical results experience lack of precision due to the existence of 4 critical points in the initial condition (points where the derivatives vanish). When comparing the results provided by the WENO-JS and the WENO-Z methods, it can be observed that WENO-Z provides more accurate results for all error norms when using the 3-rd, 5-th, 7-th order TT-ADER schemes. When moving to higher orders the WENO-Z method provides worse results that the original WENO-JS method for all error norms.

8.1.2 1D linear advection of a discontinuous initial condition

For this test case, the reactive term of (8.1) is set to 0 leading to the scalar linear advection equation and $\lambda = 1$. The following discontinuous function composed of a square, triangular, Gaussian and sinusoidal wave is used as initial condition

$$f(x) = \sum_{l=1}^{4} f_l(x) \tag{8.3}$$

with

$$\begin{aligned}
f_1(x) &= H(x-20) - H(x-40) \\
f_2(x) &= 0.1\,[H(x-60) - H(x-70)](x-60) - 0.1\,[H(x-70) - H(x-80)](x-80) \\
f_3(x) &= \exp\left(-\tfrac{(x-120)^2}{150}\right) \\
f_4(x) &= [H(x-60) - H(x-70)] \sin\left(\tfrac{\pi}{20}(x-160)\right)
\end{aligned} \tag{8.4}$$

where $H(x-x_0)$ is the Heaviside function, with $H(x < x_0) = 0$ and $H(x \geq x_0) = 1$.

Figure 8.1 shows the numerical results provided by a 1-st, 3-rd, 5-th, 7-th and 9-th order TT-ADER scheme at t=2000, using the WENO-JS method and setting $\Delta x = 1$ and CFL= 0.45. For all cases, the essentially non-oscillatory property is retained and spurious oscillations do not appear. It is observed that numerical diffusion is dramatically reduced when increasing the order of the numerical scheme. When analyzing the result provided by the 1-st order scheme, it can be seen that the original shape of the function is not recovered. When moving to 3-rd order numerical scheme, the shape of the function is recovered but sharp discontinuities are not accurately captured. This issue is addressed when using the 5-th, 7-th, 9-th

Approach		Optimal rec.		WENO-JS		WENO-Z	
Scheme	N	L_1 error	L_1 order	L_1 error	L_1 order	L_1 error	L_1 order
ADER-3	8	13306.1047	-	14230.0031	-	13204.5891	-
	16	4236.8962	1.65	11305.9088	0.33	4227.95311	1.64
	20	3360.81026	1.04	9063.78775	0.99	3360.80944	1.03
	25	2503.53129	1.32	6964.92308	1.18	2503.53089	1.32
	32	1557.4721	1.92	4824.47974	1.49	1557.47206	1.92
ADER-5	8	8555.41998	-	13338.9284	-	11987.5178	-
	16	2975.91094	1.52	3432.32973	1.96	3226.90381	1.89
	20	1633.53804	2.69	2610.60903	1.23	1796.28241	2.63
	25	733.521245	3.59	1280.586	3.19	564.476899	5.19
	32	246.629784	4.42	500.379951	3.81	171.345515	4.83
ADER-7	8	4598.10939	-	10619.2407	-	10929.9357	-
	16	1927.94242	1.25	2601.30069	2.03	2290.76604	2.25
	20	607.305053	5.18	1172.9671	3.57	936.039569	4.01
	25	162.613454	5.90	368.269248	5.19	479.523701	3.00
	32	32.5639728	6.51	125.740234	4.35	97.9384238	6.43
ADER-9	8	2273.11647	-	8048.1982	-	8717.25759	-
	16	1051.29586	1.11	2104.74364	1.94	1957.18923	2.16
	20	202.364589	7.38	572.835132	5.83	514.762861	5.99
	25	34.6830785	7.90	89.7847532	8.30	253.23115	3.18
	32	4.31840113	8.44	31.5346655	4.24	115.92998	3.17
ADER-11	8	1091.2802	-	6327.61844	-	6504.90881	-
	16	528.751847	1.05	1972.29691	1.68	1998.0091	1.70
	20	65.701758	9.35	590.042993	5.41	824.56314	3.97
	25	7.42236012	9.77	99.2149885	7.99	131.965944	8.21
	32	0.57961132	10.33	11.488342	8.73	22.4695764	7.17

Table 8.1: Section 8.1.1. L_1 error norm and convergence rate at $t = 2$ using 3-rd, 5-th, 7-th, 9-th and 11-th order TT-ADER schemes comparing the utilization of optimal reconstruction, WENO-JS and WENO-Z ($p = k - 1$) approaches.

and 11-th order numerical schemes.

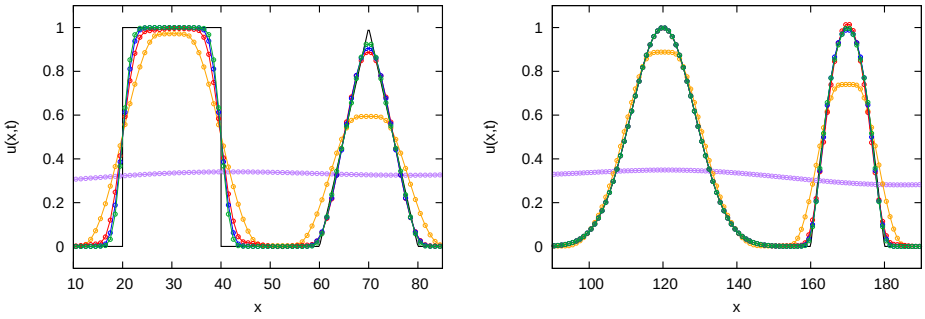


Figure 8.1: Section 8.1.2. Computational results for the advection equation with a discontinous initial condition using a 1-st (—●—), 3-rd (—●—), 5-th (—●—), 7-th (—●—) and 9-th (—●—) order TT-ADER numerical scheme and the WENO-JS method with $b = 20$. Results are compared with the exact solution (—), using a grid size $\Delta x = 1$.

8.1.3 2D linear advection of a Gaussian pulse

The following Gaussian function

$$u(x,y) = \exp\left(-\frac{(x-15)^2+(y-15)^2}{10}\right) \tag{8.5}$$

is used as initial condition for the linear scalar Equation in (8.1), setting $\lambda_1 = \lambda_2 = 1$. It is computed inside the spatial domain $\Omega = [0,30] \times [0,30]$, imposing cyclic boundary conditions.

Convergence rate tests for the solution at $t = 30$ are presented in Tables 8.2 and 8.3 where the optimal reconstruction and the traditional WENO-JS reconstruction are used respectively, in combination with the TT-ADER numerical scheme. Four refinement levels $\Delta x = \{15, 30, 60, 120\}$ have been used, setting $CFL = 0.45$. It is observed that, in general, best convergence results and lower numerical errors are achieved when using the optimal reconstruction since the initial condition is smooth and the problem does not lead to discontinuous solutions.

Scheme	N. of cells	L_1 error	L_1 order	L_∞ error	L_∞ order
ADER-3	15	2.12E-02	-	0.45553108	-
	30	6.36E-03	1.74	0.17686875	1.36
	60	1.06E-03	2.59	4.10E-02	2.11
	120	1.40E-04	2.92	6.00E-03	2.77
ADER-5	15	1.15E-02	-	0.24951337	-
	30	1.27E-03	3.18	3.77E-02	2.73
	60	5.19E-05	4.61	2.11E-03	4.16
	120	1.70E-06	4.93	7.29E-05	4.86
ADER-7	15	7.81E-03	-	0.15742517	-
	30	3.40E-04	4.52	9.97E-03	3.98
	60	3.75E-06	6.50	1.55E-04	6.00
	120	3.17E-08	6.89	1.38E-06	6.81
ADER-9	15	5.58E-03	-	0.10728565	-
	30	1.10E-04	5.67	3.13E-03	5.10
	60	3.52E-07	8.28	1.46E-05	7.75
	120	1.94E-09	7.51	5.38E-08	8.08

Table 8.2: L_1, L_2 and L_∞ error norms and corresponding convergence rates at $t = 30$ using a 3-rd, 5-th, 7-th and 9-th order ADER scheme in combination with the optimal reconstruction. CFL is set to 0.45. The number of cells appearing in the table corresponds to the number of cells in each direction when using a regular grid.

Figure 8.2 shows the numerical solution at $t = 60$ provided by a 1-st order Godunov scheme and by the 3-rd, 5-th and 7-th order 2D ADER numerical schemes in combination with the WENO-JS method, with a grid of size 30×30 cells and setting $CFL = 0.45$.

8.1.4 2D linear advection with space-dependent coefficients: Doswell frontogenesis

The kinematic approach to frontogenesis proposed by Doswell [129] provides a reliable benchmark for numerical models in meteorology. In [129], an idealized model of a vortex interacting with a initially straight frontal zone was developed. Local advection and frontogenesis were calculated analytically at the initial time and used to find the evolution of the system in time. In [130], an analytical solution for the advected scalar at a given time was obtained by solving the linear transport PDE for the scalar. Both publications explore the frontogenesis solution for a general nondivergent vortex flow.

Here, we use those results to reproduce numerically the advection of a scalar quantity under the effect of the frontogenesis using the 2D ADER scheme. Numerical results are compared with the exact solution derived in the aforementioned publications.

The kinematic model proposed in [129] consists of a hyperbolic vortex that represents a smooth ap-

Scheme	N. of cells	L_1 error	L_1 order	L_∞ error	L_∞ order
ADER-3	15	2.89E-02	-	0.65902957	-
	30	1.47E-02	0.98	0.4337996	0.60
	60	4.72E-03	1.64	0.24546794	0.82
	120	1.46E-03	1.69	0.11402813	1.11
ADER-5	15	1.49E-02	-	0.38028925	-
	30	2.15E-03	2.79	7.50E-02	2.34
	60	1.64E-04	3.71	5.74E-03	3.71
	120	6.66E-06	4.62	2.93E-04	4.29
ADER-7	15	9.25E-03	-	0.29682279	-
	30	6.92E-04	3.74	2.86E-02	3.37
	60	1.20E-05	5.85	9.59E-04	4.90
	120	1.70E-07	6.14	2.30E-05	5.38
ADER-9	15	7.76E-03	-	0.20997635	-
	30	3.16E-04	4.62	7.33E-03	4.84
	60	9.26E-07	8.41	2.64E-05	8.11
	120	2.49E-09	8.54	1.30E-07	7.67

Table 8.3: L_1, L_2 and L_∞ error norms and corresponding convergence rates at $t = 30$ using a 3-rd, 5-th, 7-th and 9-th order ADER scheme in combination with the WENO-JS reconstruction. CFL is set to 0.45. The number of cells appearing in the table corresponds to the number of cells in each direction when using a regular grid.

proximation to the Rankine combined vortex. It is worth mentioning that in many studies, the flow in real atmospheric vortices has been assumed to fit the Rankine Combined Vortex. The hyperbolic vortex in [129] is given by the following velocity profile in polar coordinates

$$\mathbf{v}(r,\theta) = \begin{pmatrix} 0 \\ V_T(r) \end{pmatrix} \tag{8.6}$$

where $V_T(r)$ represents a tangential wind given by

$$V_T(r) = V_{max}\,\mathrm{sech}^2(r)\tanh(r), \tag{8.7}$$

with $V_{max} = 2.5980762$ in order to normalize the maximum value of the wind profile. When expressing the velocity field on a Cartesian coordinate system, it reads

$$\mathbf{v}(x,y) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} = \begin{pmatrix} -V_T(r)\frac{y}{r} \\ V_T(r)\frac{x}{r} \end{pmatrix} \tag{8.8}$$

with $r = \sqrt{x^2+y^2}$. The kinematic properties of the vortex field can be analyzed by studying the linear representation of the velocity field, using the first order Taylor series expansion

$$\mathbf{v}(x_0+\delta x, y_0+\delta y) = \mathbf{v}(x_0,y_0) + \nabla(\mathbf{v})\cdot(\delta x,\delta y)^T, \tag{8.9}$$

where $\nabla(\mathbf{v})$ is the gradient of the velocity vector, with components $\frac{\partial u_i}{\partial x_j}$, that can be expressed as

$$\nabla(\mathbf{v}) = \mathbf{D} + \mathbf{R}, \tag{8.10}$$

where $\mathbf{D} \in \mathbb{R}^{2\times 2}$ is the deformation matrix with components $d_{i,j} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$ and $\mathbf{R} \in \mathbb{R}^{2\times 2}$ is the rotation matrix with components $r_{i,j} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}\right)$, with $i,j = 1,2$, $u_1 \equiv u$, $u_2 \equiv v$, $x_1 \equiv x$, $x_2 \equiv y$. From that, it is straightforward to notice [129] that for the hyperbolic vortex flow (8.8)

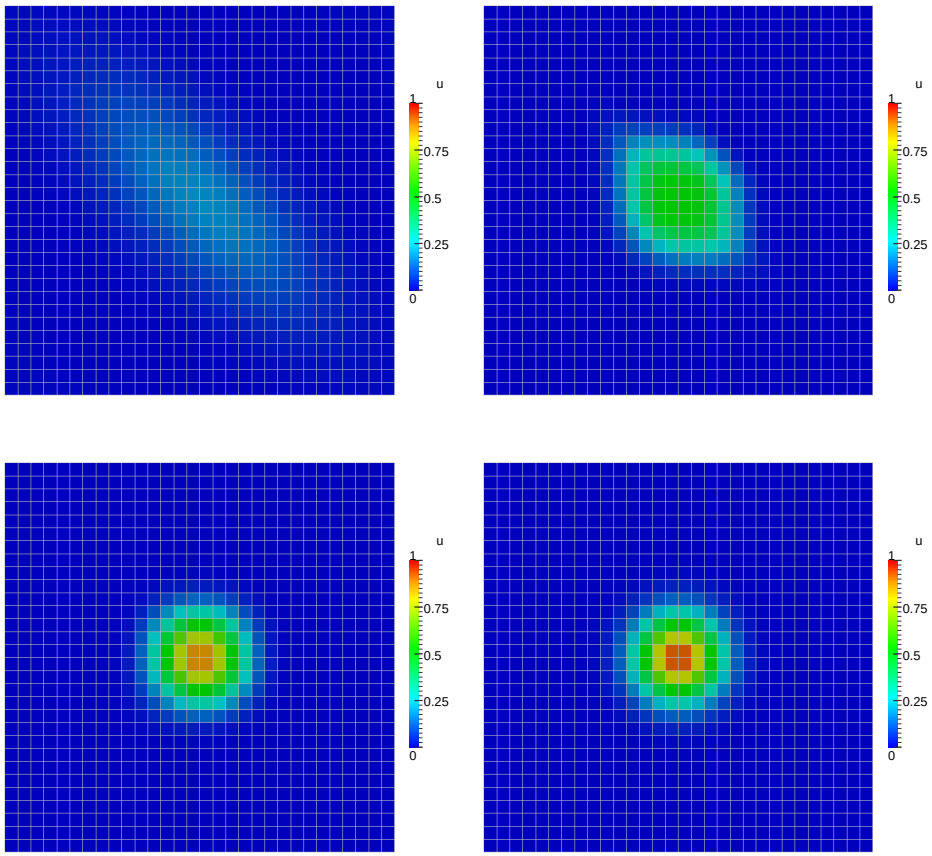


Figure 8.2: Numerical solution for the advection of the gaussian pulse at $t = 60$, using a 1-st order Godunov scheme and the 3-rd, 5-th and 7-th order 2D ADER numerical schemes. The computational grid is composed of 30×30 cells and CFL number is set to 0.45.

$$\mathbf{D} = \begin{pmatrix} \beta/2 & \alpha/2 \\ \alpha/2 & -\beta/2 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 0 & \gamma/2 \\ -\gamma/2 & 0 \end{pmatrix} \tag{8.11}$$

with

$$\alpha = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = V_{max}\,\mathrm{sech}^2(r)\left(\mathrm{sech}^2(r) - 2\tanh^2(r) - \frac{\tanh(r)}{r}\right)\cos(2\theta) \tag{8.12}$$

$$\beta = 2\frac{\partial u}{\partial x} = -2\frac{\partial v}{\partial y} = V_{max}\,\mathrm{sech}^2(r)\left(\frac{\tanh(r)}{r} - \frac{1}{2}\right)\sin(2\theta) \tag{8.13}$$

$$\gamma = \frac{\partial u}{\partial y} - \frac{\partial v}{\partial x} = -V_{max}\,\mathrm{sech}^2(r)\left(\mathrm{sech}^2(r) - 2\tanh^2(r) + \frac{\tanh(r)}{r}\right) \tag{8.14}$$

being $-\gamma$ the component of the vorticity vector $\nabla \times \mathbf{v}$ normal to the $x-y$ plane. It is worth mentioning the incompressible (non-divergent) characteristic of the flow, noticed as $\nabla \cdot \mathbf{v} = \mathrm{tr}(\mathbf{D}) = \beta/2 - \beta/2 = 0$.

Once the kinematic model has been studied, the initial condition for the scalar quantity $u = u(x, y, t)$ (e.g. temperature) must be provided. In [129], the following initial condition is proposed

$$u(x, y, 0) = \tanh\left(\frac{y}{\delta}\right), \tag{8.15}$$

modelling a straight front configuration. The evolution in time of the scalar field $u(x, y, t)$ with initial condition in (8.15) under the action of the vortex is given by the equation in (8.1) where $\lambda_1 = u(x, y)$ and $\lambda_2 = v(x, y)$ are the components of the velocity vector in (8.8).

Problem in (8.1), with $\zeta = 0$, is solved using the 2D ADER numerical scheme in combination with the WENO-JS reconstruction inside the spatial and temporal domains $\Omega = [0, 10] \times [0, 10]$ and $t \in [0, T]$ respectively, with the velocity field centered at $(x, y) = (5, 5)$. Parameter δ is set to 10^{-6}. Results of the computation of (8.1) using a 1-st order Godunov scheme and the 3-rd, 5-th and 7-th order 2D ADER numerical schemes at $t = 6$ are included in Figure 8.3, using $CFL = 0.45$ and a grid of 201 cells in each coordinate direction. It is observed that numerical diffusion is drastically reduced when moving from a 1-st order scheme to a 3-rd order ADER scheme. As the order of the numerical scheme is increased, the discontinuous solution of the frontogenesis is more accurately captured.

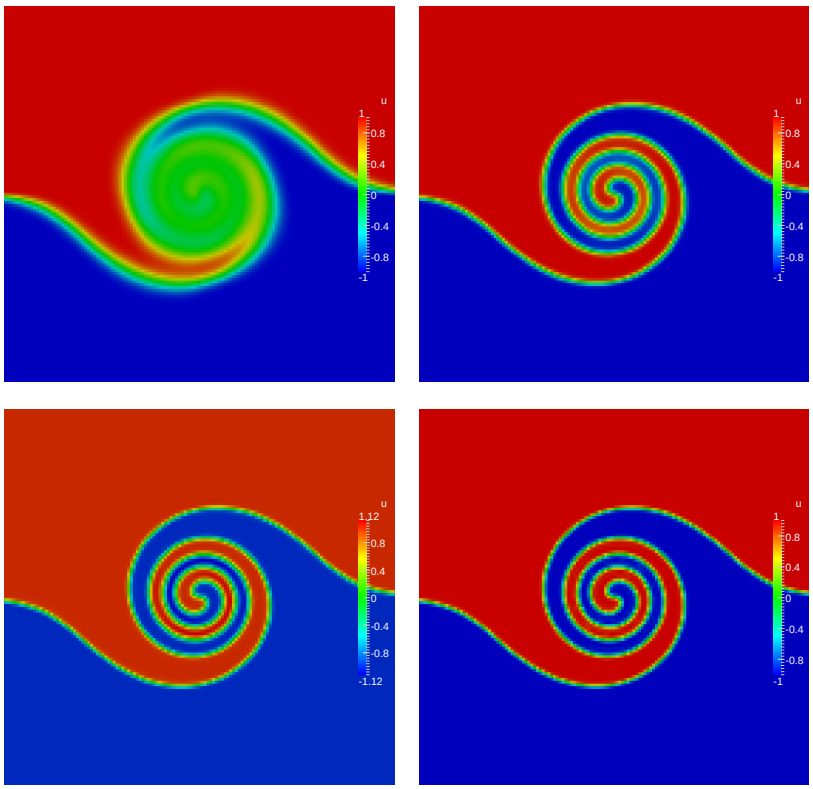


Figure 8.3: Numerical results for the Doswell frontogenesis test case in (8.1) at $t = 6$, using a 1-st order Godunov scheme and the 3-rd, 5-th and 7-th order 2D ADER numerical schemes. The computational grid is composed of 201×201 cells and CFL number is set to 0.45.

Longitudinal cuts in the y-direction at $x = 5$ of the solutions at $t = 4$ and $t = 6$ provided by a 1-st order Godunov scheme and the 3-rd, 5-th and 7-th order 2D ADER numerical schemes are presented in Figure 8.4, including the exact solution. It is observed that discontinuities are more accurately captured when increasing the order of the numerical scheme. Some oscillations are noticed when computing the solution using the 5-th order ADER scheme at $t = 6$, due to the fact that more than one discontinuity of the solution is included in the stencil of the non-oscillatory reconstruction, as reported in [131]. This issue appears to be masked when using the 7-th order ADER scheme, probably due to the even number of stencils.

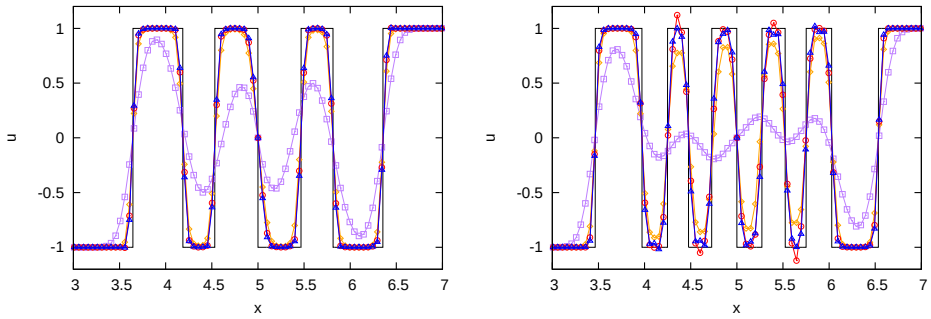


Figure 8.4: Numerical solution for the Doswell frontogenesis in (8.1) at $t = 4$ (left) and $t = 6$ (right) along the y-axis, using a 1-st order Godunov scheme and the 3-rd, 5-th and 7-th order 2D ADER numerical schemes. The computational grid is composed of 201 cells in the y-direction.

8.2 The acoustic problem: a linearization of Euler isentropic equations

The application of the Reynolds transport theorem for the mass and momentum of a gas inside a differential 1D control volume allows to obtain the following hyperbolic system of conservation laws

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial (\rho u)}{\partial x} = 0 \\ \dfrac{\partial \rho u}{\partial t} + \dfrac{\partial}{\partial x}\left(\rho u^2 + p\right) = 0\,, \end{cases} \tag{8.16}$$

where ρ is the density of the gas, u the velocity and p the pressure. Source terms for mass and momentum could be included on the right hand side of the equations, but we are not considering them for the moment. It is worth pointing out that the system in (8.16) is composed of two equations and three unknowns, therefore it could appear that an extra equation for p is required. However, p is not a conserved variable, therefore we need use an extra equation for other conserved quantity of the system: energy. Also, we will have to introduce an equation of state that relates p with the conserved variables: mass, momentum and energy.

The equation for the conservation of energy reads

$$\frac{\partial E}{\partial t} + \frac{\partial}{\partial x}\left((E + p)\,u\right) = 0\,, \tag{8.17}$$

where $E = \rho e + 1/2\rho u^2$ is the total energy, with e the specific internal energy. Such quantity can be described as a function of pressure and density alone, $e = e(p, \rho)$, which is the equation of state of the gas. Analogously, we can derive from fundamental principles that entropy is constant along particle paths in regions of smooth flows, leading to

$$\frac{\partial s}{\partial t} + u\frac{\partial s}{\partial x} = 0\,, \tag{8.18}$$

with s the entropy of the gas. In this way, we can express the full system of equations either using (8.16), (8.17) plus the equation of state or using (8.16), (8.18) plus the equation of state. The former option is a conservative description whereas the latter is non-conservative. Both systems are known as Euler equations.

Let us consider again what concerns the equation of state. When considering a polytropic gas, the following relation is satisfied

$$\frac{p}{\rho^n} = \mathscr{C} \tag{8.19}$$

with $\mathscr{C}$ a constant and $n \in \mathbb{R}$ the power exponent that defines the nature of the process. An special case is the so-called isentropic process, for which $n = \gamma$, with $\gamma = c_p/c_v$ the ratio of specific heats of the gas. In this kind of process, the flow is considered smooth and only small perturbations around a background state are possible, hence shocks do not appear. Under such conditions, entropy is simply advected with the flow and since it is initially uniform throughout the gas, it will remain constant in time. Therefore, we can drop Equation (8.18) (or (8.17) equivalently) from the description of the system and only solve (8.16) plus the equation of state.

For a isentropic flow, we can directly evaluate p as a function of ρ alone from (8.19) as

$$p(\rho) = \mathscr{C}\rho^{\gamma}\,. \tag{8.20}$$

The speed of sound for the Euler system can be derived by applying the conservation of mass and momentum across a discontinuity of differential size, which yields

$$c(\rho) = \sqrt{p'(\rho)}, \tag{8.21}$$

where $p'(\rho) = \frac{\partial p(\rho)}{\partial \rho}$. For the particular case of isentropic flow, the derivative of the pressure is computed from (8.20) as

$$p'(\rho) = \gamma\mathscr{C}\rho^{\gamma-1} \qquad \Longrightarrow \qquad p'(\rho) = \frac{\gamma p}{\rho}\,, \tag{8.22}$$

which leads to

$$c(\rho) = \sqrt{\frac{\gamma p(\rho)}{\rho}} \tag{8.23}$$

Equation (8.23) can be rewritten as

$$c(T) = \sqrt{\frac{\gamma RT}{M_m}} \tag{8.24}$$

by means of the relation $p/\rho = nRT$, where n is the number of moles, M_m is the molar mass of the gas and R is the universal gas constant. For dry air we have $R = 8.3145\ \mathrm{J}\cdot\mathrm{mol}^{-1}\cdot\mathrm{K}^{-1}$ and $M_m = 0.0289645\ \mathrm{kg/mol}$. At this point, it is also worth defining the bulk modulus, which measures the ratio of the infinitesimal pressure increase to the resulting relative decrease of the volume

$$K = -V\frac{\partial p}{\partial V} \tag{8.25}$$

and can be rewritten as

$$K = \rho p'(\rho) \qquad \Rightarrow \qquad K = \rho c^2\,. \tag{8.26}$$

Also, the definition of specific acoustic impedance (per unit area) is introduced

$$Z = \frac{p}{u} \qquad \Rightarrow \qquad Z = \rho c\,. \tag{8.27}$$

System in (8.16) can be expressed in general matrix form (2.17) with

$$\mathbf{U} = \begin{pmatrix} \rho \\ \rho u \end{pmatrix}, \mathbf{F} = \begin{pmatrix} \rho u \\ \rho u^2 + p(\rho) \end{pmatrix}, \tag{8.28}$$

We can always obtain a linear system from a non-linear problem by linearization about some state. This procedure is easily carried out from the quasilinear form by defining $\mathbf{A} = \mathbf{J}(\mathbf{U}_0)$ as a constant coefficient matrix, evaluated at the reference state $\mathbf{U}_0$. Linearized isentropic Euler equations represent a suitable model for the computation of acoustic waves, which are very small perturbations that propagate through the compressible gas and therefore are correctly represented by the linearized model. When evaluating the Jacobian at the reference state, the linear problem reads

$$\frac{\partial \hat{\mathbf{U}}}{\partial t} + \mathbf{A}\frac{\partial \hat{\mathbf{U}}}{\partial x} = 0 \tag{8.29}$$

where $\hat{\mathbf{U}} = \mathbf{U} - \mathbf{U}_0$ are the perturbations of pressure and velocity around the reference state. If writing out the system for $\hat{p}$ and $\hat{u}$ in general matrix form (8.29), it yields

$$\hat{\mathbf{U}} = \begin{pmatrix} \hat{p} \\ \hat{u} \end{pmatrix}, \qquad \mathbf{A} = \begin{pmatrix} u_0 & K_0 \\ 1/\rho_0 & u_0 \end{pmatrix} \tag{8.30}$$

The 2D extension of this problem can be expressed as

$$\frac{\partial \hat{\mathbf{U}}}{\partial t} + \mathbf{A}\frac{\partial \hat{\mathbf{U}}}{\partial x} + \mathbf{B}\frac{\partial \hat{\mathbf{U}}}{\partial y} = \mathbf{S} \tag{8.31}$$

with

$$\hat{\mathbf{U}} = \begin{pmatrix} \hat{p} \\ \hat{u} \\ \hat{v} \end{pmatrix}, \tag{8.32}$$

where $\hat{u}$ and $\hat{v}$ the velocities in x and y directions and

$$\mathbf{A} = \begin{pmatrix} u_0 & K_0 & 0 \\ 1/\rho_0 & u_0 & 0 \\ 0 & 0 & u_0 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} v_0 & 0 & K_0 \\ 0 & v_0 & 0 \\ 1/\rho_0 & 0 & v_0 \end{pmatrix} \tag{8.33}$$

the constant coefficient matrices. Hereafter in this work, we will consider perturbations around a reference state with zero velocity, that is $u_0 = v_0 = 0$.

8.2.1 The Riemann Problem for the acoustic equations

In this part, we consider the normal RP for the acoustic equations. To construct the normal RP, we formulate the evolution equation for the conserved variables in the cell-normal direction as

$$\begin{cases} \dfrac{\partial \mathbf{U}}{\partial t} + \mathbf{A}_r \dfrac{\partial \mathbf{U}}{\partial \breve{x}} = 0 \\ \\ \mathbf{U}(x, t=0) = \begin{cases} \mathbf{U}_L & \breve{x}_r < 0 \\ \mathbf{U}_R & \breve{x}_r > 0 \end{cases} \end{cases} \tag{8.34}$$

where $\mathbf{U} = (\hat{p}, \hat{u}, \hat{v})^T$ and $\mathbf{A}_r = \mathbf{A}n_x + \mathbf{B}n_y$ is the projection of the matrix $(\mathbf{A}, \mathbf{B})$ onto the r-face normal direction $\hat{\mathbf{n}} = (n_x, n_y)^T$. It is derived from from the projection of the flux $\mathscr{F} = \mathbf{E} \cdot \hat{\mathbf{n}}$, which yields

$$\mathscr{F} = \left(\mathbf{A}n_x + \mathbf{B}n_y\right)\mathbf{U}. \tag{8.35}$$

The dependence of the RP variables upon the cell number is omitted for the sake of clarity along this section. Matrix $\mathbf{A}_r$ reads

$$\mathbf{A}_r = \begin{pmatrix} 0 & K_0 n_x & K_0 n_y \\ 1/\rho_0 n_x & 0 & 0 \\ 1/\rho_0 n_y & 0 & 0 \end{pmatrix}. \tag{8.36}$$

We can define the eigenvalues and eigenvectors of $\mathbf{A}_r$ as

$$\begin{gathered} \lambda^1 = -c_0\,, \qquad \lambda^2 = 0\,, \qquad \lambda^3 = c_0\,, \\ \mathbf{e}^1 = \begin{pmatrix} -Z \\ n_x \\ n_y \end{pmatrix}, \qquad \mathbf{e}^2 = \begin{pmatrix} 0 \\ -n_y \\ n_x \end{pmatrix}, \qquad \mathbf{e}^3 = \begin{pmatrix} Z \\ n_x \\ n_y \end{pmatrix}. \end{gathered} \tag{8.37}$$

where

$$c_0 = \sqrt{\frac{K_0}{\rho_0}}, \qquad Z = \rho_0 c_0 \tag{8.38}$$

are the wave propagation speed and the impedance of the media, respectively.

The matrix that diagonalizes $\mathbf{A}_r$ is defined as $\mathbf{P} = (\mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3)$ so that $\mathbf{A_r} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ with $\mathbf{\Lambda} = \mathrm{diag}(\lambda^1, \lambda^2, \lambda^3)$. Using $\mathbf{P}$, it is possible to define the characteristic variables, denoted by $\mathbf{W} = (w^1, w^2, w^3)$, by means of the transformation $\mathbf{W} = \mathbf{P}^{-1}\mathbf{U}$. This equation can be used to construct the set of wave strengths as

$$(\alpha^1, \alpha^2, \alpha^3)^T = \delta\mathbf{W} = \mathbf{P}^{-1}\delta\mathbf{U}, \tag{8.39}$$

with $\delta(\cdot) = (\cdot)_R - (\cdot)_L$, which yields

$$\begin{aligned} \alpha^1 &= \frac{-\delta u^1 + Z\delta u^2 n_x + Z\delta u^3 n_y}{2Z}, \\ \alpha^2 &= -\delta u^2 n_y + \delta u^3 n_x\,, \\ \alpha^3 &= \frac{\delta u^1 + Z\delta u^2 n_x + Z\delta u^3 n_y}{2Z}. \end{aligned} \tag{8.40}$$

The calculation of the states at each side of the interface is straightforward

$$\mathbf{U}_L^- = \mathbf{U}_L + (\alpha\mathbf{e})^1\,, \qquad \mathbf{U}_R^+ = \mathbf{U}_R - (\alpha\mathbf{e})^3\,, \tag{8.41}$$

as well as the numerical fluxes

$$\mathbf{F}_L^- = \mathbf{F}_L + (\lambda\alpha\mathbf{e})^1\,, \qquad \mathbf{F}_R^+ = \mathbf{F}_R - (\lambda\alpha\mathbf{e})^3\,. \tag{8.42}$$

It is worth showing that

$$\mathbf{F}_R^+ - \mathbf{F}_L^- = (\lambda\alpha\mathbf{e})^2 = 0\,, \tag{8.43}$$

hence there is no flux variation across the interface and $\mathbf{F}_R^+ = \mathbf{F}_L^-$.

Let us study the internal structure of the solution for a particular case when the cell interface is perpen-

dicular to one of the coordinate directions on which the system and variables of the problem are expressed. If we consider for instance the x direction, then $\hat{\mathbf{n}} = (1,0)$, which yields the following fluctuations across each wave

$$(\alpha\mathbf{e})^1 = \begin{pmatrix} -Z\frac{-\delta\hat{p}+Z\delta\hat{u}}{2Z} \\ \frac{-\delta\hat{p}+Z\delta\hat{u}}{2Z} \\ 0 \end{pmatrix}, \qquad (\alpha\mathbf{e})^2 = \begin{pmatrix} 0 \\ 0 \\ \delta\hat{v} \end{pmatrix}, \qquad (\alpha\mathbf{e})^3 = \begin{pmatrix} Z\frac{\delta\hat{p}+Z\delta\hat{u}}{2Z} \\ \frac{\delta\hat{p}+Z\delta\hat{u}}{2Z} \\ 0 \end{pmatrix}, \tag{8.44}$$

allowing to derive the expression for the internal states

$$\mathbf{U}_L^- = \begin{pmatrix} \hat{p}_L - Z\frac{-\delta\hat{p}+Z\delta\hat{u}}{2Z} \\ \hat{u}_L + \frac{-\delta\hat{p}+Z\delta\hat{u}}{2Z} \\ \hat{v}_L \end{pmatrix}, \qquad \mathbf{U}_R^+ = \begin{pmatrix} \hat{p}_R - Z\frac{\delta\hat{p}+Z\delta\hat{u}}{2Z} \\ \hat{u}_R - \frac{\delta\hat{p}+Z\delta\hat{u}}{2Z} \\ \hat{v}_R \end{pmatrix}, \tag{8.45}$$

8.2.2 Convergence rate test using WENO-ADER schemes

In this test case, the real rate of convergence of the WENO-ADER scheme for the system in (8.31) is assessed. The solution is computed inside the domain $\Omega = [0,100]\times[0,100]$ at $t = 25$ s, setting CFL=0.4. Numerical errors are computed using the L_1 error norm and a reference solution for a very refined grid. The initial condition is given by

$$p(x,y) = \exp\left(-\frac{(x-50)^2+(y-50)^2}{10}\right), \qquad \forall (x,y)\in\Omega \tag{8.46}$$

and $u(x,y) = v(x,y) = 0\ \forall (x,y)\in\Omega$. Moreover, the properties of the media are defined as

$$K_0 = \rho_0 = 1\,, \qquad \forall (x,y)\in\Omega\,. \tag{8.47}$$

Numerical solutions for the pressure provided by a 1-st (top left), 3-rd (top right), 5-th (bottom left) and 7-th (bottom right) order WENO-ADER scheme in a 100×100 grid are presented in Figure 8.5. The WENO-JS reconstruction has been used. As expected, a higher numerical diffusion is observed for the schemes with lower accuracy. A convergence rate test using L_1 error norm is presented in Figure 8.6, where logarithmic plots of the L_1 against the number of cells (in one direction) and real computation time are presented. The simulations have been computed using a *2x Xeon E5-2697 v3* with 28 parallel threads. It is worth pointing out that errors for v are not presented because the problem considered here is symmetric, hence they are equal to those for u.

Numerical results in Figure 8.6 (top) evidence that the schemes achieve the prescribed rate of convergence, which has also been represented in the plots using a solid black line. A sligtly suboptimal convergence is noticed in the 3-rd order scheme due to the loss of accuracy at critical points experimented by the WENO-JS reconstruction. From Figure 8.6 (bottom) we can observe that the use of the WENO-ADER scheme enhances the efficiency of the simulation, that is to say, for a certain level of error, the higher the accuracy of the scheme, the faster the simulation is. For instance, let us consider the iso-error line at L_1 =1.00E-03, plotted in 8.6 (bottom left) and notice that the higher the order of the scheme, the further to the left is the intersection of the convergence line with the iso-error line.

8.2.3 Convergence rate test using DG-ADER schemes

The same test case defined in Section 8.2.2 is used here to assess the numerical performance of the DG-ADER scheme for the resolution of linear systems of equations. As in the previous case, the solution is

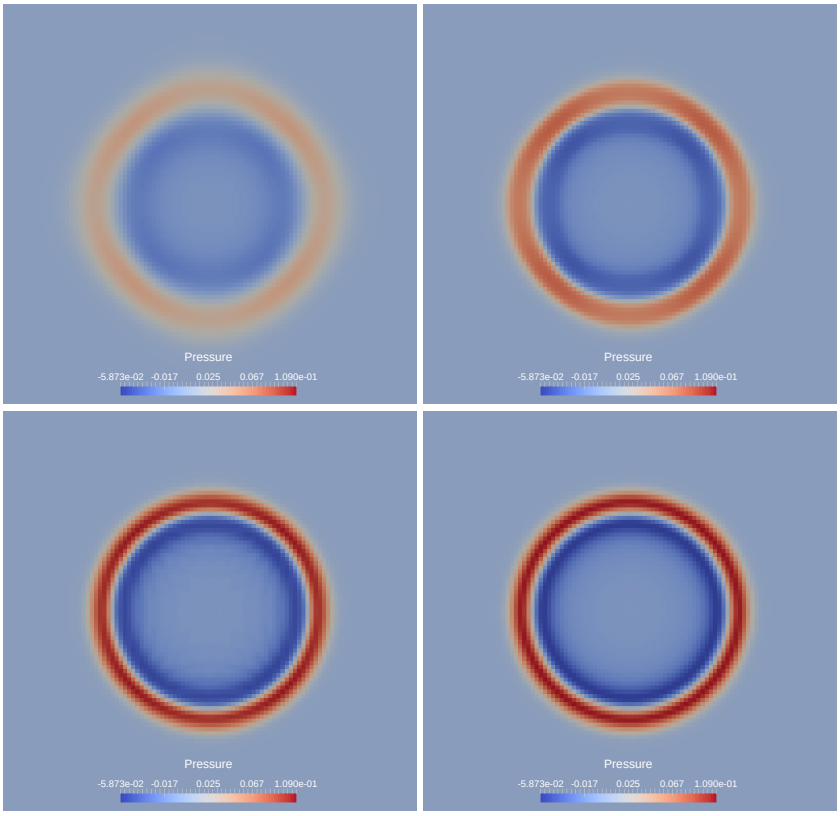


Figure 8.5: Numerical solution for p provided by a 1-st (top left), 3-rd (top right), 5-th (bottom left) and 7-th (bottom right) order WENO-ADER scheme at $t = 25$ s.

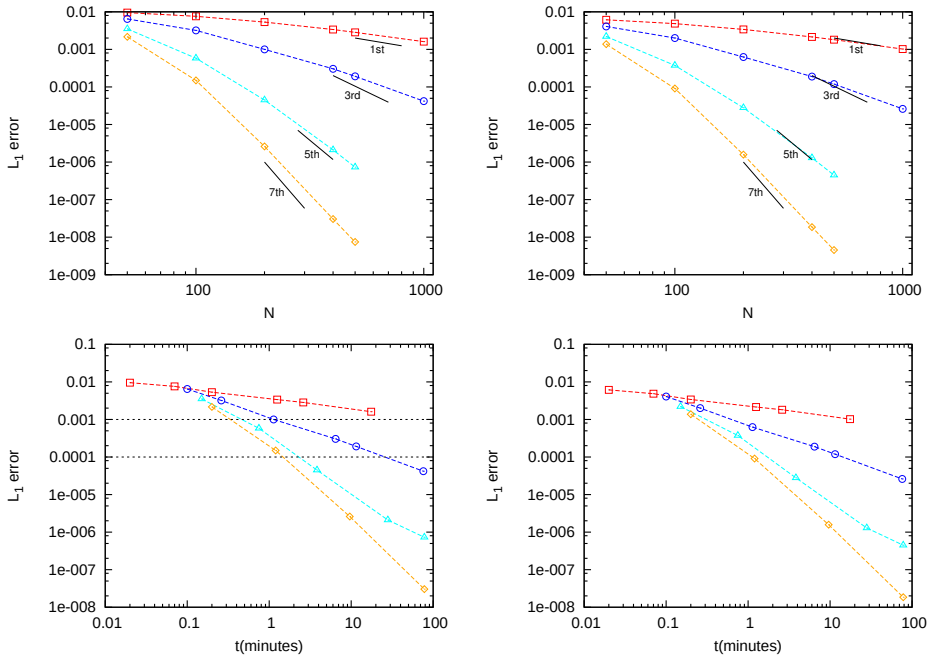


Figure 8.6: Convergence rate test: logarithmic plot of the L_1 error for p (left) and u (right) against the number of cells (top) and computation time (bottom) for the 1-st (red), 3-rd (blue), 5-th (cyan) and 7-th (orange) order WENO-ADER schemes.

computed inside the domain $\Omega = [0, 100] \times [0, 100]$ at $t = 25$ s, now setting CFL=0.04 which is sufficiently restrictive for all orders of the schemes.

Numerical solutions for the pressure provided by a 1-st (top left), 2-nd (top right), 3-rd (bottom left) and 4-th (bottom right) order WENO-ADER scheme in a 100×100 grid are presented in Figure 8.7. A convergence rate test using L_1 error norm is presented in Figure 8.8.

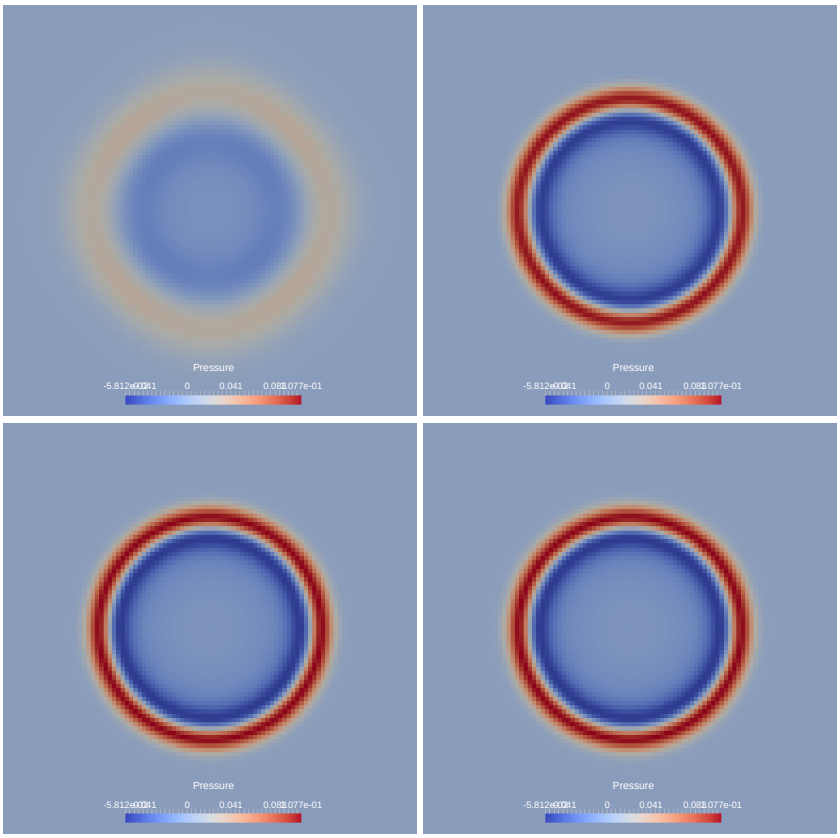


Figure 8.7: Numerical solution for p provided by a 1-st (top left), 2-nd (top right), 3-rd (bottom left) and 4-th (bottom right) order DG-ADER scheme at $t = 25$ s.

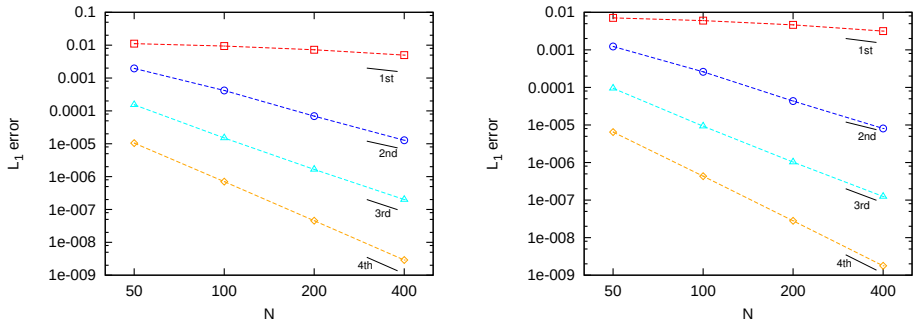


Figure 8.8: Convergence rate test: logarithmic plot of the L_1 error for p (left) and u (right) against the number of cells for the 1-st (red), 2-nd (blue), 3-rd (cyan) and 4-th (orange) order DG-ADER schemes.

The numerical implementation of the DG-ADER scheme has been carried out considering unstructured quadrilateral grids. A numerical example of the computation of the same test case in a half non-structured grid is presented in Figure 8.9. The grid has been created by randomly perturbing the right side of the structured 100×100 grid.

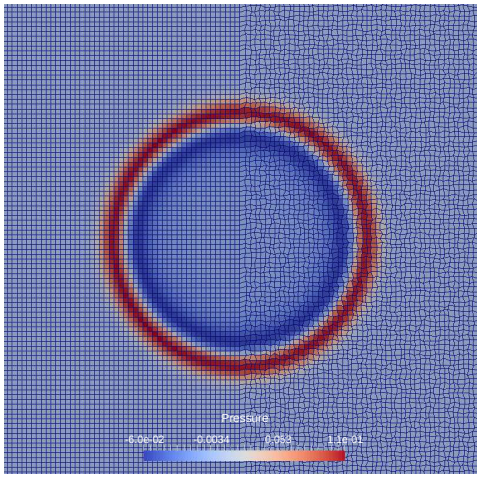


Figure 8.9: Numerical solution for test case 1.b provided by a 3-rd order DG-ADER scheme using a 100×100 grid, perturbed on the right half of the domain.

	WENO-opt ADER		WENO-JS ADER		DG-ADER	
N	L_1 error	Order	L_1 error	Order	L_1 error	Order
100	1.83E-03		4.35E-03		2.29E-05	
200	3.28E-04	2.48	1.46E-03	1.58	2.74E-06	3.06
400	4.45E-05	2.88	4.17E-04	1.81	3.38E-07	3.02
500	2.29E-05	2.98	2.54E-04	2.22	1.73E-07	3.00

Table 8.4: Numerical errors and convergence rates for p using L_1 error norm for the 3-rd order optimal WENO-ADER, WENO-JS ADER and DG-ADER schemes, using CFL=0.07.

8.2.4 Comparison of the numerical performance of the WENO-ADER and DG-ADER methods

In this case, we compare the numerical performance of the FV WENO-ADER and DG-ADER methods for the computation of linear acoustics. The test case proposed here is the same as before but considering a unique CFL number, 0.07, for both schemes, which is sufficiently restrictive for the DG-ADER scheme to be stable. The numerical solution is computed at $t = 25$ s inside the domain $\Omega = [0, 100] \times [0, 100]$. In Tables 8.4 and 8.5, numerical errors for p and wall-clock times are presented, respectively, for four different grids. Table 8.4 compares among the WENO-ADER scheme using the optimal reconstruction, the WENO-JS ADER and the DG-ADER scheme. It can be observed that to achieve the same error magnitude, the DG-ADER scheme requires fewer cells and less computational time than the WENO-ADER schemes. The observed results agree with those in [132], where WENO-FV and DG-RK methods were compared.

It is observed that the WENO-JS ADER scheme requires finer meshes to provide the prescribed order of accuracy, which is not completely reached in this test case, while the optimal WENO-ADER scheme is able to converge with the theoretical order. The explanation for this was commented in Section 8.1.1. On the other hand, the DG-ADER scheme converges at the prescribed rate, even for the coarsest grid, and provides numerical errors of lower magnitude.

The main difference between WENO-ADER and DG-ADER schemes is that in the former, the stencil for the spatial reconstruction grows with the order of the scheme, while the latter is based on a local sub-cell reconstruction. This explains why the WENO-ADER scheme allows a much more relaxed CFL condition while the DG-ADER scheme is very restrictive with the time step.

	WENO-ADER	DG-ADER
N	CPU time	CPU time
100	0.28	0.25
200	2.36	2.40
400	16.70	18.53
500	33.05	36.10

Table 8.5: Wall-clock times for the 3-rd order WENO-ADER and DG-ADER schemes, using CFL=0.07.

8.3 Concluding remarks

The highlights of this chapter are listed below:

- The WENO-ADER scheme has been used to compute linear scalar advection-reaction problems in 1D and 2D. The TT solver has been used, but would be equivalent to use any other DRP solver (CT, AR-(L)FS, etc.), as we are dealing with linear problems. Concerning the reaction source term, it would not be considered in the resolution of the DRP when using (L)FS solver types, as it is not of geometric nature.

- The performance of the WENO-JS, WENO-Z and optimal reconstruction has been assessed for those problems. It is observed that the optimal reconstruction always provides the prescribed convergence rates when the solution is smooth. On the other hand, the WENO-JS method is suboptimal in presence of critical points. This is circumvented when using the WENO-Z method. However, such method does not provide optimal results either when moving to very high order of accuracy.

- The linear acoustic equations have been considered and a linear wave propagation test has been computed using the WENO-ADER and DG-ADER schemes. Both schemes provide the expected accuracy. As reported in the literature, it is observed that the DG-ADER method is more accurate than the WENO-ADER scheme, for the same CFL number, but has a tighter time step restriction. This is because the DG approach is based on a local reconstruction, unlike the WENO reconstruction, which uses a stencil that grows with the order of the scheme. From the results, we can conclude that the DG-ADER method is more efficient than the WENO-ADER scheme, but further investigation must be carried out to overcome the strong restrictions in the time step.

9 THE SHALLOW WATER EQUATIONS

Many engineering and environmental problems involve the study of steady and transient free surface water flows where the vertical scale is much smaller than the horizontal ones. Such phenomena can be described by the SWE [133], a depth averaged model composed of the equations for the conservation of mass and momentum that considers a hydrostatic pressure distribution in the vertical direction. This assumption is feasible as the accelerations in the vertical dimension are considered to be negligible, hence the vertical pressure gradient is nearly hydrostatic. The SWE is a hyperbolic system of conservation laws which accurately describes non-linear wave propagation phenomena observed in the physical reality under the aforementioned hypothesis. Due to the non-linear hyperbolic nature of the equations, the numerical techniques described in this thesis are of application.

Most realistic problems involving shallow flows are dominated by source terms, such as bed variation and friction, which makes their numerical simulation quite a challenging task in what concerns the design of the numerical scheme. Suitable discretization techniques for the source terms are required in order for the schemes to provide physically based solutions, which is the only way to construct a trustworthy predictive tool. In previous chapters, particular discretization techniques for the source terms, specially for those of geometric nature, were presented and will now be applied to the SWE.

In this work, we consider the SWE with bed elevation, friction and rotation (Coriolis) source terms, which represents a good model for small and large scale geophysical flows. Among such terms, it is worth highlighting the particular nature of the bed elevation source term, which is geometric and has to be accounted for in the resolution of the (D)RP. The other aforementioned sources do not possess a geometric nature and do not necessarily have to be discretized in the same way. However, a geometric reinterpretation of those terms can be done in order to ensure the preservation of equilibrium states of relevance. Otherwise, extra corrections must be added to the scheme.

We will show that a geometric reinterpretation of the Coriolis term is necessary to ensure equilibrium in the rotating frame. On the other hand, we will also show that a centered discretization of the friction term is sufficient for the scheme to accurately represent the phenomena of interest.

In this thesis, the applications to the SWE have been divided in 3 chapters for the sake of clarity. The details concerning the mathematical model of the SWE and its analysis are presented in this chapter. On the other hand, the details for the construction of 1D and 2D schemes for the resolution of the SWE, based on the aforementioned ideas, are presented in Chapters 10 and 11 respectively. Numerical results obtained from the application of the proposed schemes to the resolution of a broad variety of test cases are also included.

In this Chapter, the general formulation of the SWE in 1 and 2 space dimensions is presented in Section

9.1. The analysis of such model, including the derivation of the characteristic fields and Riemann invariants as well as the study of the bed step contact discontinuity is presented in Section 9.2. We underscore the importance of the study of the bed step contact discontinuity for the derivation of a suitable discretization of the bed step at cell interfaces. To this end, the SWE are rewritten in non-conservative form, the resulting system is analyzed and the bed step discontinuity is studied from the point of view of energy/momentum conservation [134, 67, 135]. The conditions for the jump across the bed step contact wave are derived based on the conservation of specific mechanical energy, which seems to be the most reasonable choice and does not contradict the conservation of momentum [134]. Such conditions, which consist of the conservation of both the GRH and RI across the discontinuity, are used to derive a particular source term discretization that fulfills them and ensures the conservation of energy. Such discretization will be applied in Chapter 10 to design an EB numerical scheme.

9.1 The SWE model

9.1.1 2D formulation of the SWE

The SWE with bottom topography and friction in a rotating frame can be expressed in matrix form as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial y} = \mathbf{S}, \quad \mathbf{S} = \mathbf{S}_b + \mathbf{S}_f + \mathbf{S}_c, \tag{9.1}$$

where

$$\mathbf{U} = (h, hu, hv)^T, \quad \mathbf{F} = \left(hu, hu^2 + \frac{1}{2}gh^2, huv\right)^T, \quad \mathbf{G} = \left(hv, huv, hv^2 + \frac{1}{2}gh^2\right)^T, \tag{9.2}$$

are the vectors of conserved quantities and physical fluxes in the x and y directions and

$$\mathbf{S}_b = \left(0, -gh\frac{dz}{dx}, -gh\frac{dz}{dy}\right)^T, \quad \mathbf{S}_f = \begin{pmatrix} 0 \\ -c_f |\mathbf{v}| u \\ -c_f |\mathbf{v}| v \end{pmatrix}, \quad \mathbf{S}_c = \begin{pmatrix} 0 \\ fhv \\ -fhu \end{pmatrix}, \tag{9.3}$$

are the vectors of sources due to bed variation, friction and Coriolis force, respectively. Variable $\mathbf{v}$ is the velocity vector, z is the bed elevation, c_f is the friction coefficient and f the Coriolis coefficient. The friction coefficient c_f is computed as

$$c_f = \frac{gn^2}{h^{1/3}}, \tag{9.4}$$

with n the Manning coefficient, while the Coriolis coefficient will be computed using the β-plane approximation. Equations (9.1)–(9.3) represent a good model for both small and large scale phenomena in geophysical flows. Note that the bed source term is represented by a geometric source term.

The Jacobian matrices of the fluxes $\mathbf{F}(\mathbf{U})$ and $\mathbf{G}(\mathbf{U})$ read

$$\mathbf{A} = \frac{\partial \mathbf{F}(\mathbf{U})}{\partial \mathbf{U}} = \begin{pmatrix} 0 & 1 & 0 \\ c^2 - u^2 & 2u & 0 \\ -uv & v & u \end{pmatrix}, \qquad \mathbf{B} = \frac{\partial \mathbf{G}(\mathbf{U})}{\partial \mathbf{U}} = \begin{pmatrix} 0 & 0 & 1 \\ -uv & v & u \\ c^2 - u^2 & 0 & 2u \end{pmatrix} \tag{9.5}$$

respectively and can be projected onto the vector normal to the cell interface to obtain the normal Jacobian matrix

$$\mathbf{J} = \mathbf{A}n_x + \mathbf{B}n_y = \begin{pmatrix} 0 & n_x & n_y \\ c^2 n_x - u(\mathbf{v}\cdot\hat{\mathbf{n}}) & \mathbf{v}\cdot\hat{\mathbf{n}} + un_x & un_y \\ c^2 n_y - v(\mathbf{v}\cdot\hat{\mathbf{n}}) & vn_x & \mathbf{v}\cdot\hat{\mathbf{n}} + vn_y \end{pmatrix}. \tag{9.6}$$

9.1.2 x-split SWE and 1D model with bed elevation only

Generally, an alternative to the construction of the normal (D)RP is to rotate the variables and solve the x-split system of equations [86]. We consider here the x-split SWE with bed variation only and neglect the other sources, since this is the only source term of geometric nature and an special attention must be paid to it when defining the RP and DRP. The x-split SWE with bottom elevation reads

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{S}. \tag{9.7}$$

where

$$\mathbf{U} = \begin{pmatrix} h \\ hu \\ hv \end{pmatrix}, \ \mathbf{F} = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{pmatrix}, \ \mathbf{S} = \begin{pmatrix} 0 \\ -gh\frac{dz}{dx} \\ 0 \end{pmatrix}. \tag{9.8}$$

In the case of considering a pure 1D problem, the third equation in (9.8), corresponding to the passive advection of the shear velocity, is neglected.

9.2 Characteristic analysis

9.2.1 Characteristic analysis of the homogeneous part

The Jacobian matrix of the flux in (9.8) reads

$$\mathbf{J} = \begin{pmatrix} 0 & 1 & 0 \\ c^2 - u^2 & 2u & 0 \\ -uv & v & u \end{pmatrix}, \tag{9.9}$$

with $c = \sqrt{gh}$ the wave celerity. The eigenvalues and eigenvectors that diagonalize the Jacobian are given by

$$\lambda^1 = u - c, \qquad \lambda^2 = u, \qquad \lambda^3 = u + c \tag{9.10}$$

and

$$\mathbf{e}^1 = \begin{pmatrix} 1 \\ \lambda^1 \\ v \end{pmatrix}, \qquad \mathbf{e}^2 = \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}, \qquad \mathbf{e}^3 = \begin{pmatrix} 1 \\ \lambda^3 \\ v \end{pmatrix}. \tag{9.11}$$

Three characteristic fields are identified, associated to the eigenvectors $\mathbf{e}^1$, $\mathbf{e}^2$, $\mathbf{e}^3$. The nature of each characteristic field can be studied by analyzing the variation of the eigenvalue along the integral curve defined by the eigenvector associated to that field, as pointed out in section 2.3. For this particular case, such variations read

$$\begin{aligned}\nabla_u \lambda^1(\mathbf{U}) \cdot \mathbf{e}^1(\mathbf{U}) &= -\frac{\sqrt{g}}{2\sqrt{h}}, \\ \nabla_u \lambda^2(\mathbf{U}) \cdot \mathbf{e}^2(\mathbf{U}) &= 0, \\ \nabla_u \lambda^3(\mathbf{U}) \cdot \mathbf{e}^3(\mathbf{U}) &= \frac{\sqrt{g}}{2\sqrt{h}}\end{aligned} \tag{9.12}$$

and according to definitions 6 and 7, it can be concluded that characteristic fields associated to eigenvectors $\mathbf{e}^1$ and $\mathbf{e}^3$ are genuinely nonlinear fields while the characteristic field associated to eigenvector $\mathbf{e}^2$ is a linearly degenerated field. From the physical point of view, the nonlinear fields $\mathbf{e}^1$ and $\mathbf{e}^3$ are of this nature as they are associated to the nonlinear fluxes that govern the hydrodynamic of the problem. On the other hand, the transport of quantity hv is considered a passive advection, given by the velocity u, therefore the characteristic field $\mathbf{e}^2$ is linearly degenerate.

The characteristic fields associated to $\mathbf{e}^1$ and $\mathbf{e}^3$ are depicted in Figures 9.1 and 9.2 (left), including a contour plot of $\lambda^m(\mathbf{U})$ and the vector plot of $\nabla_u \lambda^m(\mathbf{U})$. It is possible to observe that $\nabla_u \lambda^m(\mathbf{U})$ is not orthogonal to the vector field for any of the cases, hence there is a variation of $\lambda^m(\mathbf{U})$ along the integral curves. In order to show this appreciation, a plot of the normalized scalar product

$$\zeta^m = \left| \frac{\nabla_u \lambda^m(\mathbf{U}) \cdot \mathbf{e}^m(\mathbf{U})}{|\nabla_u \lambda^m(\mathbf{U})| \cdot |\mathbf{e}^m(\mathbf{U})|} \right| \tag{9.13}$$

is presented in Figures 9.1 and 9.2 (right) for vector fields associated to $\mathbf{e}^1$ and $\mathbf{e}^3$. The quantity ζ accounts for the absolute value of the cosine of the angle between $\nabla_u \lambda^m(\mathbf{U})$ and $\mathbf{e}^m(\mathbf{U})$. Only when $\zeta^m = 0$, $\forall h, hu \in \mathscr{C}$, we say that the m-th field is linearly degenerate.

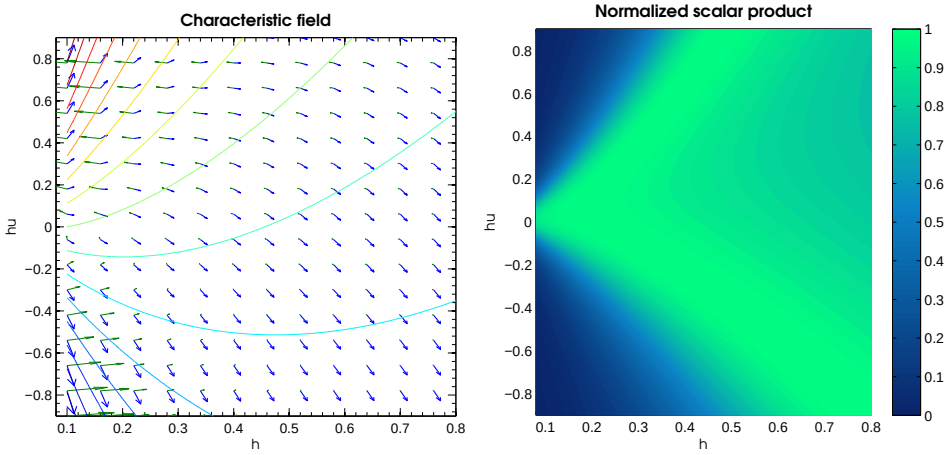


Figure 9.1: Left: Characteristic field associated to $\mathbf{e}^1$ (blue), including a contour plot of $\lambda^1(\mathbf{U})$ and the vector plot of $\nabla_u \lambda^1(\mathbf{U})$ (green). Right: Normalized scalar product ζ^1.

It is worth pointing out that the phase diagram for the 1-st and 3-rd characteristic fields in Figures 9.1 and 9.2 was constructed only considering variables h and hu for the sake of simplicity, though a 3D representation would be required to show all variables.

9.2.2 Characteristic analysis of the SWE system in its non-conservative form

For the sake of simplicity, the equation for the passive transport of hv in (9.8) will not be hereafter considered, as it does not play a role in the dynamics of the system. According to Equation (2.27), system in (9.8) can be expressed in its non-conservative form as

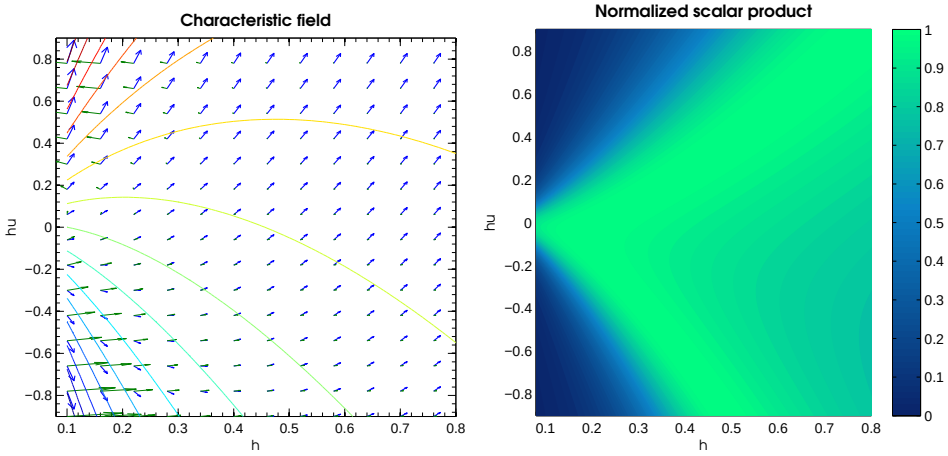


Figure 9.2: Left: Characteristic field associated to $\mathbf{e}^3$ (blue), including a contour plot of $\lambda^3(\mathbf{U})$ and the vector plot of $\nabla_u \lambda^3(\mathbf{U})$ (green). Right: Normalized scalar product ζ^2.

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} + \mathbf{H}(\mathbf{U})\frac{\partial \mathbf{U}}{\partial x} = 0, \tag{9.14}$$

where

$$\mathbf{U} = \begin{pmatrix} h \\ hu \\ z \end{pmatrix}, \ \mathbf{F} = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ 0 \end{pmatrix}, \ \mathbf{H} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & gh \\ 0 & 0 & 0 \end{pmatrix}. \tag{9.15}$$

The Jacobian matrix of the flux reads

$$\mathbf{J} = \begin{pmatrix} 0 & 1 & 0 \\ c^2 - u^2 & 2u & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{9.16}$$

and it can be used to construct the following matrix

$$\mathscr{A} = \mathbf{J} + \mathbf{H} = \begin{pmatrix} 0 & 1 & 0 \\ c^2 - u^2 & 2u & gh \\ 0 & 0 & 0 \end{pmatrix} \tag{9.17}$$

allowing to express the system in quasilinear form. The eigenvalues and eigenvectors that diagonalize $\mathscr{A}$ are given by

$$\lambda^1 = u - c, \qquad \lambda^2 = 0, \qquad \lambda^3 = u + c \tag{9.18}$$

and

$$\mathbf{e}^1 = \begin{pmatrix} 1 \\ \lambda^1 \\ 0 \end{pmatrix}, \qquad \mathbf{e}^2 = \begin{pmatrix} 1 \\ 0 \\ u^2/gh - 1 \end{pmatrix}, \qquad \mathbf{e}^3 = \begin{pmatrix} 1 \\ \lambda^3 \\ 0 \end{pmatrix}. \tag{9.19}$$

The nature of each characteristic field can be studied according to definitions 6 and 7, as pointed out in section 2.3. For this particular case, we have

$$\begin{aligned}
\nabla_u \lambda^1(\mathbf{U}) \cdot \mathbf{e}^1(\mathbf{U}) &= -\frac{\sqrt{g}}{2\sqrt{h}}, \\
\nabla_u \lambda^2(\mathbf{U}) \cdot \mathbf{e}^2(\mathbf{U}) &= 0, \\
\nabla_u \lambda^3(\mathbf{U}) \cdot \mathbf{e}^3(\mathbf{U}) &= \frac{\sqrt{g}}{2\sqrt{h}},
\end{aligned} \tag{9.20}$$

noticing that the 2-nd characteristic field associated to the bed step is linearly degenerate as eigenvalue λ^2 is zero $\forall \mathbf{U}$ (the step is regarded as a stationary discontinuity) while the two other characteristic fields are genuinely nonlinear.

The integral curves for each characteristic field are calculated next. For the sake of clarity and according to the notation used in definition 5, the integral curve associated to the m-th characteristic field will be referred to as $\mathbf{U}^m(\xi)$. Let us consider first the 1-st characteristic field, whose integral curve is defined as

$$\frac{d\mathbf{U}^1(\xi)}{d\xi} = \nu(\xi) \cdot \mathbf{e}^1(\mathbf{U}(\xi)), \tag{9.21}$$

which, after setting $\nu(\xi) = 1$, leads to the following equations

$$\begin{cases}
\dfrac{dh}{d\xi} = 1 \\
\dfrac{dhu}{d\xi} = \dfrac{hu}{h} - \sqrt{gh} \\
\dfrac{dz}{d\xi} = 0
\end{cases} \tag{9.22}$$

When choosing the parametrization by means of ξ, integration of the first equation in (9.22) yields

$$h(\xi) = C_1 + \xi, \tag{9.23}$$

with C_1 the integration constant. The equality $d\xi = dh$ noticed in the first equation allows to rewrite the second equation as

$$\frac{dhu}{dh} = \frac{hu}{h} - \sqrt{gh}, \tag{9.24}$$

that after integration yields

$$hu(h) = (C_2 - 2\sqrt{gh})h, \tag{9.25}$$

which is parametrized by h, with C_2 the integration constant. From (9.25), it can be noticed that the quantity

$$u + 2\sqrt{gh} \tag{9.26}$$

is constant along the curve. This quantity corresponds to the first Riemann invariant associated to the 1-characteristic field.

To find the expression of the integral curve, let us consider the curve starting at $(h, hu, z) = (h^*, (hu)^*, z^*)$, with $(hu)^* = h^* u^*$, and notice that $\xi = 0$ at that point. Using this information, Equation (9.23) become

$$h(\xi) = h^* + \xi, \tag{9.27}$$

and following the same procedure, C_2 in Equation (9.25) yields

$$(hu)^* = (C_2 - 2\sqrt{gh^*})h^* \quad \Rightarrow \quad C_2 = u^* + \sqrt{gh^*} \tag{9.28}$$

allowing to rewrite (9.25) as

$$hu(\xi) = (h^* + \xi)\left[u^* - 2(\sqrt{g(h^* + \xi)} - \sqrt{gh^*})\right]. \tag{9.29}$$

The integral curve associated to the 1-st field, parametrized by ξ and starting at $(h, hu, z) = (h^*, (hu)^*, z^*)$, reads

$$\mathbf{U}^1(\xi) = \begin{pmatrix} h(\xi) \\ hu(\xi) \\ z(\xi) \end{pmatrix} = \begin{pmatrix} h^* + \xi \\ (h^* + \xi)\left[u^* - 2(\sqrt{g(h^* + \xi)} - \sqrt{gh^*})\right] \\ z^* \end{pmatrix} \tag{9.30}$$

Let us consider now the 2-nd characteristic field, whose integral curve is defined as

$$\frac{d\mathbf{U}^2(\xi)}{d\xi} = \nu(\xi) \cdot \mathbf{e}^2(\mathbf{U}(\xi)), \tag{9.31}$$

which, after setting $\nu(\xi) = 1$, leads to the following equations

$$\begin{cases} \dfrac{dh}{d\xi} = 1 \\ \dfrac{dhu}{d\xi} = 0 \\ \dfrac{dz}{d\xi} = \dfrac{(hu)^2}{gh^3} - 1 \end{cases} \tag{9.32}$$

Integration of the first equation in (9.32) yields

$$h(\xi) = C_1 + \xi, \tag{9.33}$$

with C_1 the integration constant. Integration of the second equation

$$\frac{dhu}{\xi} = 0 \tag{9.34}$$

yields $hu = cte$, that is, constant discharge along the curve. It is worth pointing out that hu is the first Riemann invariant associated to the 2-characteristic field. The equality $d\xi = dh$ noticed in the first equation allows to rewrite the third equation in (9.32) as

$$\frac{dz}{dh} = \frac{(hu)^2}{gh^3} - 1, \tag{9.35}$$

that after integration yields

$$z(h) = C_3 - \frac{(hu)^2}{2gh^2} - h, \tag{9.36}$$

with C_3 the integration constant. From (9.36), it can be noticed that the quantity

$$\frac{u^2}{2g} + h + z \tag{9.37}$$

is constant along the curve. This quantity corresponds to the second Riemann invariant associated to the 2-characteristic field and represents the mechanical energy per unit mass of the system, denoted by E.

Considering again Equation (9.33), if setting the starting point of the curve at $(h, hu, z) = (h^*, (hu)^*, z^*)$, with $(hu)^* = h^* u^*$, and noticing that $\xi = 0$ at that point, such equation can be rewritten as

$$h(\xi) = h^* + \xi \,. \tag{9.38}$$

Using the same information, the discharge can be given the following constant value

$$hu = (hu)^* \tag{9.39}$$

in terms of the initial data. Analogously, C_3 in (9.36) can also be expressed in terms of such information as follows

$$C_3 = \frac{u^{*2}}{2g} + h^* + z^* \,. \tag{9.40}$$

Inserting (9.39) and (9.40) in (9.36), the latter yields

$$z(h) = \frac{u^{*2}}{2g} + h^* + z^* - \frac{(hu)^{*2}}{2gh^2} - h \,, \tag{9.41}$$

that can be parametrized in terms of ξ using Equation (9.38), leading to

$$z(\xi) = \frac{u^{*2}}{2g} + z^* - \frac{(hu)^{*2}}{2g(h^* + \xi)^2} - \xi \,. \tag{9.42}$$

The integral curve associated to the 2-nd characteristic field, parametrized by ξ and starting at $(h, hu, z) = (h^*, (hu)^*, z^*)$, reads

$$\mathbf{U}^2(\xi) = \begin{pmatrix} h(\xi) \\ hu(\xi) \\ z(\xi) \end{pmatrix} = \begin{pmatrix} h^* + \xi \\ (hu)^* \\ \frac{u^{*2}}{2g} + z^* - \frac{(hu)^{*2}}{2g(h^*+\xi)^2} - \xi \end{pmatrix} \tag{9.43}$$

Finally, let us consider the 3-rd characteristic field, whose integral curve is defined as

$$\frac{d\mathbf{U}^3(\xi)}{d\xi} = \nu(\xi) \cdot \mathbf{e}^3(\mathbf{U}(\xi)) \,, \tag{9.44}$$

with equations

$$\begin{cases} \dfrac{dh}{d\xi} = 1 \\ \dfrac{dhu}{d\xi} = \dfrac{hu}{h} + \sqrt{gh} \\ \dfrac{dz}{d\xi} = 0 \end{cases} \tag{9.45}$$

Following the same procedure than for the 1-st integral curve, the solution in this case reads

$$\mathbf{U}^3(\xi) = \begin{pmatrix} h(\xi) \\ hu(\xi) \\ z(\xi) \end{pmatrix} = \begin{pmatrix} h^* + \xi \\ (h^* + \xi)\left[u^* + 2(\sqrt{g(h^* + \xi)} - \sqrt{gh^*})\right] \\ z^* \end{pmatrix} \tag{9.46}$$

and the relevant Riemann invariant is given by

$$u - 2\sqrt{gh}. \tag{9.47}$$

Characteristic field	1-Riemann invariant	2-Riemann invariant
1	$u + 2\sqrt{gh}$	z
2	hu	$\frac{u^2}{2g} + h + z$
3	$u - 2\sqrt{gh}$	z

Table 9.1: Summary of Riemann invariants for the non-homogeneous SWE.

The resonance phenomenon

Prior to the explanation of the resonance phenomenon, it is worth recalling that the conserved variables, $\mathbf{U} = (h, hu, z)$, take values on $C \subseteq \mathbb{R}^3$, the set of physically admissible values. The resonance phenomenon happens under certain conditions, specifically when the flow speed coincides with the wave propagation celerity. Under these conditions, the conserved quantities lie on one of the following hypersurfaces

$$C^+ = \left\{(h, hu, z) | u = \sqrt{gh}\right\}, \tag{9.48}$$

$$C^- = \left\{(h, hu, z) | u = -\sqrt{gh}\right\}. \tag{9.49}$$

The Jacobian matrix of the non-conservative SWE, $\mathscr{A}(\mathbf{U})$, is strictly hyperbolic for $\mathbf{U} \in C \setminus (C^+ \cup C^-)$, as it has 3 linearly independent eigenvectors associated to 3 distinct eigenvalues. However, when $\mathbf{U} \in C^+ \cup C^-$, $\lambda^1 = \lambda^2 = 0$ or $\lambda^3 = \lambda^2 = 0$ and two of the eigenvectors are linearly dependent, hence $\mathscr{A}(\mathbf{U})$ is not diagonalizable. This is called in the literature the *resonance regime*. It occurs when passing through the sonic point, that is $u = \pm\sqrt{gh}$, which makes the 1-nonlinear field (when $\mathbf{U} \in C^+$) or the 3-nonlinear field (when $\mathbf{U} \in C^-$) superpose with the 2-linearly degenerate field of the contact discontinuity. Under those conditions, a number of waves greater than the number of characteristic fields may appear in the solution. In the resonance regime, the solution of the Riemann problem is no longer classical and the uniqueness is lost for some initial data [63].

9.2.3 Conservation of energy across the bed-step contact wave

As outlined in the previous section, the 2-characteristic field in the non-conservative SWE in (9.15) is a linearly degenerate field. This kind of field arises from the presence of the bed step and is characterized by a contact wave of zero celerity, $\lambda^2 = 0$, since the bed elevation does not vary in time.

Discontinuous solutions describing a contact wave are generally expressed by (3.32). For this particular case, the right state will be denoted by $\mathbf{U}_R$, hence (3.32) is rewritten as

$$\mathbf{U}(x,t) = \begin{cases} \mathbf{U}_L & x < 0 \\ \mathbf{U}_R & x > 0 \end{cases} \tag{9.50}$$

where $\mathbf{U}_L = (h_L, (hu)_L, z_L)^T$ and $\mathbf{U}_R = (h_R, (hu)_R, z_R)^T$ are the left and right states respectively. Notice that we may write $(hu)_L = h_L u_L$ for the sake of clarity and recall that this quantity represents the first Riemann invariant of the S-characteristic field, hence $h_L u_L = h_R u_R$. The second Riemann invariant is the specific mechanical energy, hence $u_L^2/2 + g(h+z)_L = u_R^2/2 + g(h+z)_R$.

Across the contact wave in (9.50), the GRH condition in (3.25) must hold for all variables. For this particular case, it reads

$$\begin{aligned} h_R u_R - h_L u_L &= 0\,, \\ \left(g\frac{h_R^2}{2} + h_R u_R^2\right) - \left(g\frac{h_L^2}{2} + h_L u_L^2\right) &= D\,, \end{aligned} \tag{9.51}$$

with D a suitable approximation of the integral of the source term across the bed step

$$D = -\int_{z_L}^{z_R} g h dz\,, \tag{9.52}$$

that can be rewritten as

$$D = -\int_{x_L}^{x_R} g h \frac{dz}{dx} dx\,. \tag{9.53}$$

As outlined before, GRH condition in (9.51) must be ensured so that (9.50) is a weak solution of the problem, hence the right state $(h_R, h_R u_R, z_R)$ must lie on the GHL for a given left state $(h_L, h_L u_L, z_L)$. However, this condition does not ensure the conservation of Riemann invariants across the contact wave. Only when condition in (3.38) holds, Riemann invariants are conserved and the IC coincide with the GHL. In other words, we can state that IC coincide with the GHL if (9.51) holds and the Riemann invariants of the 2-characteristic field in Table 9.1 are conserved.

It seems clear that the election of a suitable discretization of the integral of the source term in (9.53) is crucial. In [135], a particular source term discretization based on physical considerations that accounts for the dissipation of energy across the step was chosen. Under this assumption, they showed that equation (3.38) is not always satisfied and proved that the Riemann invariant associated to the specific mechanical energy was not anymore conserved across the step. In this way, they provided a coherent mathematical framework for the physically-based dissipative discretization of the bed step and they constructed a Riemann solver based on such ideas.

Unlike [135], in the present work the authors do not include any additional energy dissipation mechanism. Here, it is considered that energy must be conserved across the bed step, as claimed in [134, 67]. Hence, an energy-conservative source term discretization is sought and both the GRH condition and Equation (3.38) must hold, as Riemann invariants have to be conserved across the contact wave. Following [135], equation (3.38) is rewritten as

$$-\int_0^{\hat{\xi}} \mathbf{H} \cdot \mathbf{e}^2 d\xi = \mathbf{D}\,, \tag{9.54}$$

where $\hat{\xi} = h_R - h_L$ is the value of ξ on the right state. We define

$$h(\hat{\xi}) = h_R \qquad u(\hat{\xi}) = u_R \qquad z(\hat{\xi}) = z_R\,. \tag{9.55}$$

Our goal here is to find the expression for $\mathbf{D}$ satisfying (9.54) and to this end, we have to manipulate (9.54) using extra relations among left and right states. It is worth recalling that for the derivation of condition (9.54) (originally (3.38)), $\mathbf{U}(\xi)$ was imposed to lie on the IC, given by Equation (9.43). Here we will work under the same assumption, hence $\mathbf{U}(\hat{\xi}) = \mathbf{U}_R = (h_R, h_R u_R, z_R)$ lies on the IC provided the left state. Water depth along the IC can be expressed as

$$h(\hat{\xi}) = h_L + \hat{\xi} = h_R \tag{9.56}$$

and in the same way, the velocity along the IC is

$$u(\hat{\xi}) = \frac{h_L u_L}{h_L + \hat{\xi}} = \frac{h_R u_R}{h_R} = u_R \,, \tag{9.57}$$

with a constant discharge

$$hu(\hat{\xi}) = h_L u_L = h_R u_R \,, \tag{9.58}$$

and a variable bed elevation along the IC

$$z(\hat{\xi}) \equiv z_R = z_L + h_L - h_R + \frac{u_L^2}{2g} - \frac{u_R^2}{2g} \,. \tag{9.59}$$

In the following derivation, condition (9.54) will be combined with the relations between left and right states in (9.56)-(9.59), allowing to find the expression of **D** satisfying the RI and the GRH conditions. The product $\mathbf{H} \cdot \mathbf{e}^2$ reads

$$\mathbf{H} \cdot \mathbf{e}^2 = \begin{pmatrix} 0 \\ u^2(\xi) - gh(\xi) \\ 0 \end{pmatrix} \tag{9.60}$$

and using (9.57) in (9.60), the latter yields

$$-\int_0^{\hat{\xi}} \begin{pmatrix} 0 \\ \left(\frac{h_L u_L}{h_L + \xi}\right)^2 - g(h_L + \xi) \\ 0 \end{pmatrix} d\xi = \begin{pmatrix} 0 \\ D \\ 0 \end{pmatrix} . \tag{9.61}$$

From (9.61), only the second component will be considered

$$-\int_0^{\hat{\xi}} \left(\frac{h_L u_L}{h_L + \xi}\right)^2 d\xi + \int_0^{\hat{\xi}} g(h_L + \xi) d\xi = D \,, \tag{9.62}$$

that after integration and using the relation $h_L u_L = h_R u_R$ in (9.58) when required, yields

$$\left(g\frac{h_R^2}{2} + h_R u_R^2\right) - \left(g\frac{h_L^2}{2} + h_L u_L^2\right) = D \,, \tag{9.63}$$

with the right state laying on the IC in (9.43), provided the left state. It can be noticed that equation (9.63) coincides with the GRH condition for the conservation of momentum.

Now, combination of equation (9.63) with (9.59) allows to derive the particular source term discretization, D, that under the assumed hypotheses will ensure the conservation of the Riemann invariants and lead to an energy-conservative scheme. For the sake of clarity, equation (9.63) is rewritten as

$$\delta\left(g\frac{h^2}{2} + hu^2\right)_{L,R} = D \tag{9.64}$$

and so is (9.59), the equation for the conservation of energy

$$\delta\left(\frac{u^2}{2} + g(h+z)\right)_{L,R} = 0 \tag{9.65}$$

where $\delta(\cdot)_{L,R} = (\cdot)_R - (\cdot)_L$ is a difference operator. From (9.64), it is straightforward to obtain

$$\left(g\bar{h}\delta h + \bar{u}\delta(hu) + \overline{hu}\delta u\right)_{L,R} = D\,, \tag{9.66}$$

where

$$(\bar{\cdot})_{L,R} = \frac{(\cdot)_L + (\cdot)_R}{2} \tag{9.67}$$

is an average operator. For the sake of simplicity, subscript $(\cdot)_{L,R}$ is dropped in Equations (9.68)-(9.72) as they always refer to the left and right states of the contact wave in this derivation. Noticing that $\delta(hu)_{L,R} = h_R u_R - h_L u_L = 0$, Equation (9.66) yields

$$g\bar{h}\delta h + \overline{hu}\delta u = D\,. \tag{9.68}$$

The equation for the conservation of energy in (9.65) is multiplied by $\bar{h}$ and rewritten as

$$\bar{h}\bar{u}\delta u + g\bar{h}\delta h + g\bar{h}\delta z = 0\,, \tag{9.69}$$

from where the term $g\bar{h}\delta h$ can be expressed as

$$g\bar{h}\delta h = -\bar{h}u\delta u - g\bar{h}\delta z \tag{9.70}$$

and can be inserted in (9.68), leading to

$$D = -g\bar{h}\delta z + (\overline{hu} - \bar{h}\bar{u})\delta u\,. \tag{9.71}$$

It is straightforward to show that (9.71) can be rewritten as

$$D = -g\bar{h}\delta z + \delta(hu^2) - \bar{u}\delta(hu) - \bar{h}\delta\left(\frac{1}{2}u^2\right), \tag{9.72}$$

with $\delta(hu) = 0$ according to the GRH conditions, hence

$$D = -g\bar{h}\delta z + \delta(hu^2) - \bar{h}\delta\left(\frac{1}{2}u^2\right). \tag{9.73}$$

9.3 Concluding remarks

It is worth emphasizing the following issues addressed in this chapter:

- The SWE with bed elevation, friction and Coriolis have been considered. The 1D SWE are studied more in detail as the numerical schemes herein described are based on the resolution of 1D (D)RPs. Friction and Coriolis source terms have not been considered in the analysis as they do not have a geometric nature.
- The 1D SWE with bed elevation have been rewritten in non-conservative form to study the mathematical properties of the jump across the bed step discontinuity.
- When considering the non-conservative formulation of the SWE, the GRH condition is always ensured across the bed step contact discontinuity. On the other hand, RIs are only conserved when (3.38) is satisfied.
- The relevant RIs across the bed step discontinuity are the specific discharge and the specific mechanical energy. The former will always be conserved due to the GRH condition, whereas the latter will

only be conserved when (3.38) is fulfilled.

- In this work, the authors consider that energy must be conserved across the bed step, as claimed in [134, 67] and do not consider any additional dissipation mechanism, as done by Rosatti [135]. The conservation of energy across the bed step contact wave seems to be the only approach consistent with the classical SWE, which is compatible with the momentum balance across the step [134].
- To ensure the conservation of the energy across the bed step contact wave (the relevant RI), we have found a particular source term discretization, D, presented in (9.71), satisfying (3.38). Other choices for D will result in a gain or loss of energy across the bed step. If a singular energy loss is required at the bed step, additional terms as suggested by classical handbooks can be included [134].

10 ENERGY BALANCED SCHEMES FOR THE 1D SWE

In the framework of the SWE, the preservation of the still water at rest, also called quiescent equilibrium, is considered a fundamental feature the numerical scheme must posses. Numerical schemes able to preserve still water at rest are called well-balanced methods, concept introduced by Bermudez and Vázquez-Cendón [58] and Greenberg and Leroux [59]. There is a large variety of well-balanced methods based on Riemann solvers that ensure the preservation of the still water steady state [60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 136].

The well-balanced property can still be enhanced. If neglecting friction in the SWE, mechanical energy is conserved under steady conditions in absence of hydraulic jumps. Such idea of energy conservation can be integrated in the numerical scheme, at the discrete level, allowing the extension of well-balanced methods to exactly well-balanced methods, also called EB methods [75, 76, 77, 78, 79, 80]. In [81, 82], exactly well-balanced methods using the ARoe and HLLS solvers were presented, reproducing the exact solution for steady states with independence of the cell size. Such methods were based on a weighted EB formulation (WEBF) of the source term, which is recalled in this chapter. It is observed that the accuracy of the WEBF technique in the resolution of hydraulic jumps can still be enhanced [85]. A novel EB discretization technique, called selective EB formulation (SEBF) [85], is proposed here. Such technique is based on a shock detection algorithm and allows to accurately capture the position of the hydraulic jump in presence of bed variation.

The procedures for the extension of the aforementioned EB formulations to arbitrary order of accuracy [83, 84] are also described in this chapter. The approach in [75] is followed and the EB discretizations are extended to higher order of accuracy by extrapolation. The resulting approximation of the integral will now be of arbitrary order and ensure the conservation of energy.

The resulting schemes, the EB AR(L)-ADER and HLLS(L)-ADER methods, provide the exact solution for steady cases with moving water and irregular geometries, with independence of the grid. Moreover, such schemes are able to ensure convergence, at the prescribed rate, to the exact solution for transient problems including RPs that involve bed variations and resonant solutions [67]. It is worth mentioning that in order to ensure such capabilities when constructing the numerical schemes, specific data reconstruction techniques were required.

The Chapter is structured as follows. In Section 10.1, some preliminaries for the numerical resolution of the SWE including Roe's averages, are provided. In Section 10.2, different source term discretizations for the bed slope source term are studied. Such methods include the differential formulation (DF), integral formulation (IF), weighted EB formulation (WEBF) and selective EB formulation (SEBF), presented here for the first time. The aforementioned techniques are assessed using a steady transcritical test case that includes

the capture of a hydraulic jump. Finally, in Section 10.3, the EB source term discretization techniques are extended to arbitrary order of accuracy in the framework of WENO-ADER schemes and the EB AR(L)-ADER and HLLS(L)-ADER schemes are presented. The proposed methods are assessed using RPs and a smooth case where the convergence rates can be evaluated.

10.1 Numerical resolution of the 1D SWE

When applied to the SWE, the approximate Jacobian $\widetilde{\mathbf{J}}$ of the homogeneous part is given by [15]

$$\tilde{\mathbf{J}}_{i+1/2} = \begin{pmatrix} 0 & 1 \\ \widetilde{c}^2 - \widetilde{u}^2 & 2\widetilde{u} \end{pmatrix}_{i+1/2}, \tag{10.1}$$

where

$$\begin{aligned} \widetilde{\lambda}^1 &= \widetilde{u} - \widetilde{c}, \qquad \widetilde{\lambda}^2 = \widetilde{u} + \widetilde{c} \\ \widetilde{\mathbf{e}}^1 &= \begin{pmatrix} 1 \\ \widetilde{u} - \widetilde{c} \end{pmatrix}, \qquad \widetilde{\mathbf{e}}^2 = \begin{pmatrix} 1 \\ \widetilde{u} + \widetilde{c} \end{pmatrix} \end{aligned} \tag{10.2}$$

with

$$\widetilde{c} = \sqrt{g \frac{h^{(0)}_{i_R} + h^{(0)}_{(i+1)_L}}{2}}, \qquad \widetilde{u} = \frac{u^{(0)}_{(i+1)_L} \sqrt{h^{(0)}_{(i+1)_L}} + u^{(0)}_{i_R} \sqrt{h^{(0)}_{i_R}}}{\sqrt{h^{(0)}_{(i+1)_L}} + \sqrt{h^{(0)}_{i_R}}}, \tag{10.3}$$

with $h^{(0)}$ and $u^{(0)}$ the spatial reconstruction of the water depth and velocity, respectively.

The matrix and vectors in (10.1) and (10.2) are used to construct the HLLS and ARoe solver and their ADER versions. For such solvers, an appropiate numerical discretization of the bed step source term at the interface is required.

10.2 Numerical discretization of the source term at cell interfaces for augmented solvers

When using augmented solvers, such as the HLLS and ARoe solvers and their ADER versions, numerical approximations over the integral of the source term at cell interfaces are required. The approximation of the spatial integral of the source term at cell interface $i+1/2$, that is inside $[x_{i_R}, x_{(i+1)_L}]$, will be referred to as

$$\int_{x_{i_R}}^{x_{(i+1)_L}} -g\, h^{(0)} \frac{dz}{dx} dx \approx \bar{S}^{(0)}_{i+1/2}. \tag{10.4}$$

Analogously, the approximation of the spatial integral of time derivatives of the source term will be referred to as

$$\int_{x_{i_R}}^{x_{(i+1)_L}} -g\, h^{(k)} \frac{dz}{dx} dx \approx \bar{S}^{(k)}_{z_{i+1/2}}. \tag{10.5}$$

In this section we will only focus on the leading term of the source term, $\bar{S}^{(0)}_{i+1/2}$, and describe the different possible integration techniques. In the next section, we will also consider the integral in (10.5) and other

procedures required for the construction of arbitrary order schemes.

We can find in literature different approaches to compute the leading term of the source in Equation (10.4). One possibility is to compute it considering a smooth variation of the variables across the interface, as

$$\bar{S}^a_{i+1/2} = -g\bar{h}\delta z\,, \tag{10.6}$$

which will be referred to as differential formulation (DF) and where

$$\bar{h} = \frac{1}{2}(h^{(0)}_{(i+1)_L} + h^{(0)}_{i_R}), \qquad \delta z = z_{(i+1)_L} - z_{i_R} \tag{10.7}$$

The second possibility is the so-called integral formulation (IF), where $S^b_{i+1/2}$ is defined as

$$\bar{S}^b_{i+1/2} = -gh_j\delta z\,, \tag{10.8}$$

with

$$h_j = \begin{cases} h^{(0)}_{i_R} & \text{if} \quad \delta h^{(0)} > 0 \\ h^{(0)}_{(i+1)_L} & \text{if} \quad \delta h^{(0)} \leq 0 \end{cases} \tag{10.9}$$

where $\delta h^{(0)} = h^{(0)}_{(i+1)_L} - h^{(0)}_{i_R}$. In cases of still water with a continuous water level surface, both (10.6) and (10.8) do ensure quiescent equilibrium. In this particular case hydrostatic forces are exactly balanced.

In order to extend the well balanced property for static equilibrium to the exactly balanced property, that ensures exact equilibrium under for moving water steady state, it is necessary to impose extra conditions in the discretization of the source term. It is worth pointing out that those restrictions will only be included in the leading term of the source, as the higher order terms vanish under steady state.

Generally, under the assumption of conservation of energy across the bed step contact wave, the best choice for the discretization of the bed source term seems to be Equation (9.71). However, such a discretization does not allow to construct an explicit scheme as it depends upon the intermediate states at both sides of the bed step, $\mathbf{U}^{-,(0)}_{i_R}$ and $\mathbf{U}^{+,(0)}_{(i+1)_L}$.

Under steady conditions and considering no change in flow regime across the RP, it is straightforward to prove that $\mathbf{U}_{i_R} = \mathbf{U}^{-,(0)}_{i_R}$ and $\mathbf{U}_{(i+1)_L} = \mathbf{U}^{+,(0)}_{(i+1)_L}$, hence (9.71) can be rewritten in terms of the initial data as

$$\begin{aligned} D = \quad & -g\left(\frac{h^{(0)}_{(i+1)_L} + h^{(0)}_{i_R}}{2}\right)(z_{(i+1)_L} - z_{i_R}) + \\ & \left[\left(\frac{(hu)^{(0)}_{(i+1)_L} + (hu)^{(0)}_{i_R}}{2}\right) - \left(\frac{h^{(0)}_{(i+1)_L} + h^{(0)}_{i_R}}{2}\right)\left(\frac{u^{(0)}_{(i+1)_L} + u^{(0)}_{i_R}}{2}\right)\right](u^{(0)}_{(i+1)_L} - u^{(0)}_{i_R}). \end{aligned} \tag{10.10}$$

For the sake of clarity, notation for Equation (10.10) is simplified, considering variations and averages across the interface $i+1/2$, that is, the left and right states of the RP. By doing this, (10.10) is rewritten as

$$D = \left\{-g\bar{h}\delta z + (\overline{hu} - \bar{h}\bar{u})\delta u\right\}_{i+1/2}. \tag{10.11}$$

In shallow flows, there are physically feasible situations where energy is dissipated, such as hydraulic jumps. Ideally, such shock would be considered as a pure discontinuity where energy is suddenly dissipated, however, when using a finite volume formulation, the shock width is of the size of a cell, since the discretization considers constant values in each cell and the discontinuity cannot be kept anymore as a

discontinuity inside the cell. As a consequence, energy dissipation must be imposed at the interfaces of the cell containing the shock, as it is not possible to explicitly carry out the dissipation of energy inside the cell.

Murillo [81] proposed a novel approach for the discretization of the source term that allows to construct an exactly EB scheme. This approximation is based on the principle of conservation of mechanical energy and is only applied to the leading term, since higher order terms become nil in steady state when energy is conserved, as mentioned above.

Considering the IF and DF approaches for the discretization of the source term, it is possible to evaluate $\bar{S}^{(0)}_{i+1/2}$ as a combination of them as

$$\bar{S}^{(0)}_{i+1/2} = (1-\mathscr{A})S^{a}_{i+1/2} + \mathscr{A}S^{b}_{i+1/2}, \tag{10.12}$$

where $0 \leq \mathscr{A} \leq 1$. This formulation will be referred to as weighted energy balanced formulation (WEBF). In order to satisfy both energy and momentum conservation under steady conditions, a value $\mathscr{A}_E$ is defined as

$$\mathscr{A}_E = \frac{\delta(hu^2) - \bar{h}\delta\left(\frac{u^2}{2}\right)}{S^{b}_{i+1/2} - S^{a}_{i+1/2}}, \tag{10.13}$$

according to [81]. Coefficient $\mathscr{A}_E$ in (10.13) can be used in (10.12) to ensure the conservation of energy for smooth solutions. On the other hand, when considering transcritical jumps, energy must be dissipated, hence the value of weight coefficient $\mathscr{A}$ in (10.12) is set to 1. Considering these situations, the complete algorithm for the calculation of $\mathscr{A}$ reads [81]

$$\mathscr{A} = \begin{cases} \mathscr{A}_E & \text{if } u^{(0)}_{(i+1)_L} u^{(0)}_{i_R} > 0 \text{ and } \left|Fr_{(i+1)_L}\right| > 1 \text{ and } \left|Fr_{i_R}\right| > 1 \\ \mathscr{A}_E & \text{if } u^{(0)}_{(i+1)_L} u^{(0)}_{i_R} > 0 \text{ and } \left|Fr_{(i+1)_L}\right| < 1 \text{ and } \left|Fr_{i_R}\right| < 1 \\ 1 & \text{otherwise} \end{cases}, \tag{10.14}$$

where Fr_{i_R} and $Fr_{(i+1)_L}$ are the Froude numbers on the left and right sides of the interface. It is worth pointing out that $\mathscr{A}_E$ can be straightforward obtained from Equation (10.11).

Instead of imposing the exact amount of dissipation of energy across the shock by means of a tailored source term discretization at that point, here we propose to add an additional degree of freedom to the equations by means of using a traditional discretization of the source term at the interfaces surrounding the hydraulic jump while maintaining the energy conservative formulation in (10.11) for the rest. The differential discretization of the source term is chosen at those interfaces. This technique allows the numerical scheme to converge to the exact position of the shock while recovering the exact solution in both the subcritical and supercritical regions connected by the transcritical shock, with independence of the grid refinement.

The keystone of this novel technique is related to the way the information is propagated towards the jump. Let us consider an steady hydraulic jump that arises from the transition of supercritical flow on the left and subcritical flow on the right. On the left hand side of the jump, information is propagated only downstream whereas on the right hand side of the jump, information moves both upstream and downstream. By leaving free the condition of energy conservation at the left and right interfaces of the cell containing the hydraulic jump (using the differential discretization of the bed step) and imposing it at the rest of the interfaces, under steady state, the boundary conditions are enough to ensure the conservation of energy in both the subcritical and supercritical regions and the exact rate of dissipation at the shock.

The proposed approach is now explained in detail. We propose to use Roe celerities, $\tilde{\lambda}^m$ to locate the cell where the hydraulic jump is located, since it is known that both celerities at the left interface are positive (supercritical flow entering the cell) while celerities corresponding to subcritical conditions (one negative and the other one positive) are identified at the right interface. Let us consider the cells, Ω_i, as single items

contained in Ω such that $\Omega = \{\Omega_i \mid i \in [1,...,N]\}$. Considering the possibility of multiple hydraulic jumps within the domain, we denote the set of cells containing a positive-flow hydraulic jump as

$$\mathscr{D}^+ = \left\{\Omega_i \mid \Omega_i \in \Omega \wedge \tilde{\lambda}^1_{i-1/2} \cdot \tilde{\lambda}^1_{i+1/2} < 0 \wedge h_{i-1} < h_{i+1}\right\} \tag{10.15}$$

and the set of cells containing a negative-flow hydraulic jump as

$$\mathscr{D}^- = \left\{\Omega_i \mid \Omega_i \in \Omega \wedge \tilde{\lambda}^2_{i-1/2} \cdot \tilde{\lambda}^2_{i+1/2} < 0 \wedge h_{i-1} > h_{i+1}\right\} \tag{10.16}$$

and the set of Riemann Problems at their respective interfaces is

$$\mathscr{R}_1 = \left\{\mathrm{RP}_{i+1/2} \mid i \in \mathbb{N} \wedge \Omega_i \in \mathscr{D}^+ \cup \mathscr{D}^-\right\} \tag{10.17}$$

$$\mathscr{R}_2 = \left\{\mathrm{RP}_{i-1/2} \mid i \in \mathbb{N} \wedge \Omega_i \in \mathscr{D}^+ \cup \mathscr{D}^-\right\} \tag{10.18}$$

$$\mathscr{R} = \mathscr{R}_1 \cup \mathscr{R}_2\,, \tag{10.19}$$

where $\mathrm{RP}_{i+1/2}$ stands for the Riemann Problem at interface $i+1/2$. By using the previous definitions, the approximation of the integral of the source term at any interface is defined as follows

$$\tilde{S}^{(0)}_{i+1/2} = \begin{cases} -g\bar{h}\delta z + (\overline{hu} - \bar{h}\bar{u})\delta u & \text{if} \quad \mathrm{RP}_{i+1/2} \notin \mathscr{R} \\ -g\bar{h}\delta z & \text{if} \quad \mathrm{RP}_{i+1/2} \in \mathscr{R} \end{cases} \tag{10.20}$$

and the method will be hereafter referred to as selective energy balanced formulation (SEBF).

It is worth noting that the use of the DF leads to a well-balanced scheme. The proof will be shown in the section devoted to the applications to the 2D SWE.

10.2.1 Test case: 1-st order steady shock capturing for the SWE with bed topography

In this test case, steady solutions for the flow over the following bed elevation profile

$$z(x) = \begin{cases} 0 & \text{if} \quad x < 8 \\ 0.05(x-8) & \text{if} \quad 8 \le x \le 12 \\ 0.2 - 0.05\,(x-12)^2 & \text{if} \quad 12 \le x \le 14 \\ 0 & \text{if} \quad x > 12 \end{cases} \tag{10.21}$$

are computed using the 1-st order ARoe solver in combination with the different discretization techniques for the source term outlined before. The computational domain is $[0, 20]$ and the solution is computed for $t = 600$ s. CFL number is set to 0.45 for all cases. The discharge is imposed to 0.6 m^2/s upstream to obtain the critical point at the cell with maximum bed elevation, that is $z_{max} = 0.2$. Downstream, the water depth is also imposed to $h = 0.621$ m in order to generate a hydraulic jump downstream. Different computational grids, composed of 100, 200, 400, 800 and 1600 cells respectively, are used to compute the numerical solution.

Numerical solutions provided by the ARoe solver when using the different approximations of the source term presented before, namely the DF, the IF, the WEBF and the novel SEBF, are presented and compared with the exact solution in Figures 10.1, 10.2 and 10.3. In Figure 10.1, the numerical solutions for $h + z$ and q computed by the ARoe solver in combination with all the previous techniques on two grids of 100 and 400 cells are plotted together and compared with the exact solution. To study the effect of mesh refinement in the accuracy of the numerical solution and convergence to the exact position of the shock, a detailed plot of the solution provided by each one of the methods is presented in Figure 10.2 for three

different grids composed of 200, 400 and 800 cells respectively. Numerical results evidence that those approximations based on the IF of the source term, such as the WEBF approach from [81] and the IF itself, do not accurately capture the position of the shock, with independence of the grid. In any case, the former strategy provides much better results than the latter, as it is energy-conservative. On the other hand, it is evidenced that both the DF and the SEBF do accurately capture the shock position for any grid.

It is also noticed that a spurious spike in the numerical discharge appears for all methods and what is of utmost relevance, that the amplitude of this spike is not reduced with mesh refinement, as observed in Figure 10.1.

Analogously to Figure 10.2, similar plots of the same region but showing the numerical results computed in a very fine grid (1600 cells) as piecewise constant data are presented in Figure 10.3. Here, the left, right and averaged (Roe) characteristic speeds at cell interfaces, denoted by λ_L, λ_R and $\tilde{\lambda}$ respectively, as well as cell interfaces have also been represented with the aim of showing the transition of flow regime of the numerical solution. The exact position of the shock is around $x = 13.264$, hence it should be represented by the numerical scheme inside the cell with cell center at 13.275. It can be observed that only the numerical solution provided by the ARoe solver in combination with the DF and SEBF of the source term allows to capture the shock inside the aforementioned cell.

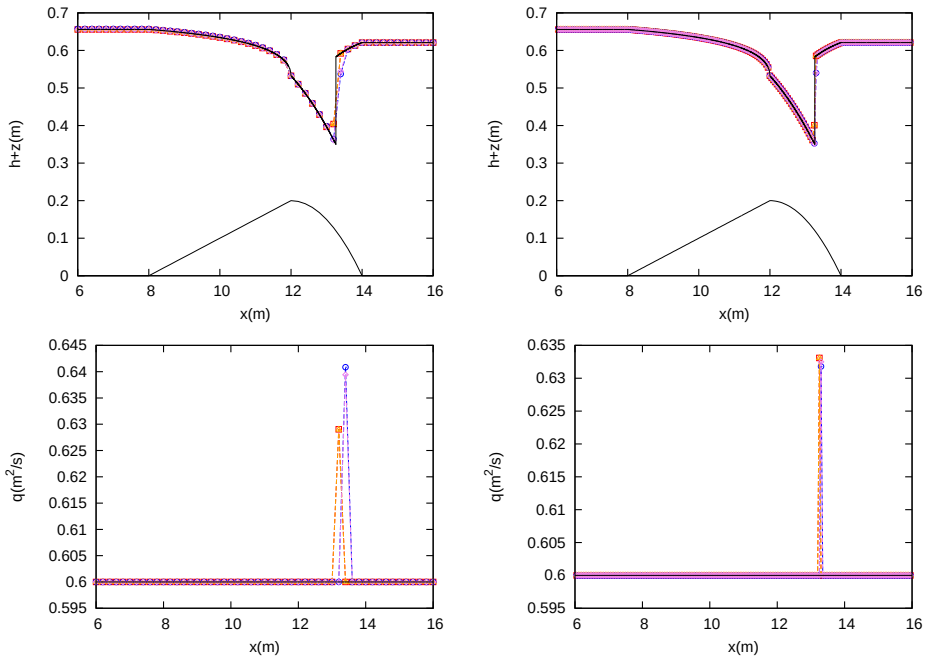


Figure 10.1: Section 10.2.1. Exact (—) and numerical solution for $h + z$ (top) and q (bottom) computed by the ARoe solver in combination with the DF (— △ —), IF (— ○ —), SEBF (—□—) and WEBF (— ◇ —), using 100 (left) and 400 cells (right).

The numerical solution for the specific mechanical energy, computed using the aforementioned techniques in the grids of 100 and 400 cells, is presented in Figure 10.4 left and right respectively. It is observed that only when using an EB source term discretization, such as the ARoe solver in combination with the SEBF or WEBF formulations, energy is conserved. On the other hand, when using the DF and IF formulations of the source term, energy is not conserved though it converges with mesh refinement. Among the assessed methods, the SEBF is the one providing the best performance, as it ensures the conservation of energy when required and accurately captures the position of the hydraulic jump. This method provides the exact solutions in all cells but the one containing the shock, with independence of the grid.

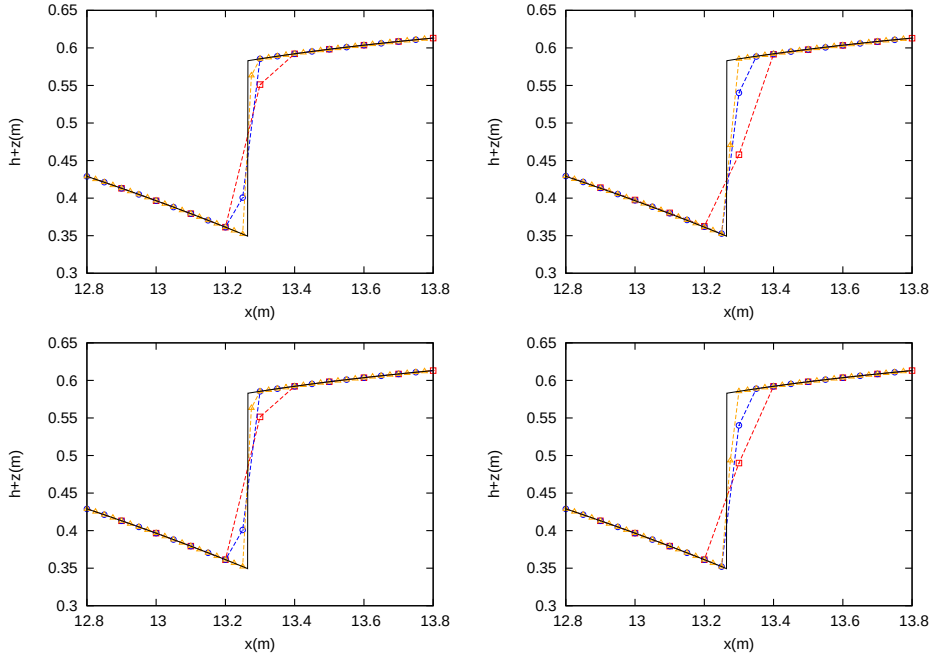


Figure 10.2: Section 10.2.1. Exact (—) and numerical solution for $h+z$ computed by the ARoe solver in combination with the DF (top left), IF (top right), SEBF (bottom left) and WEBF (bottom right) using 200 (—□—), 400 (— ∘ —) and 800 (— △ —) cells.

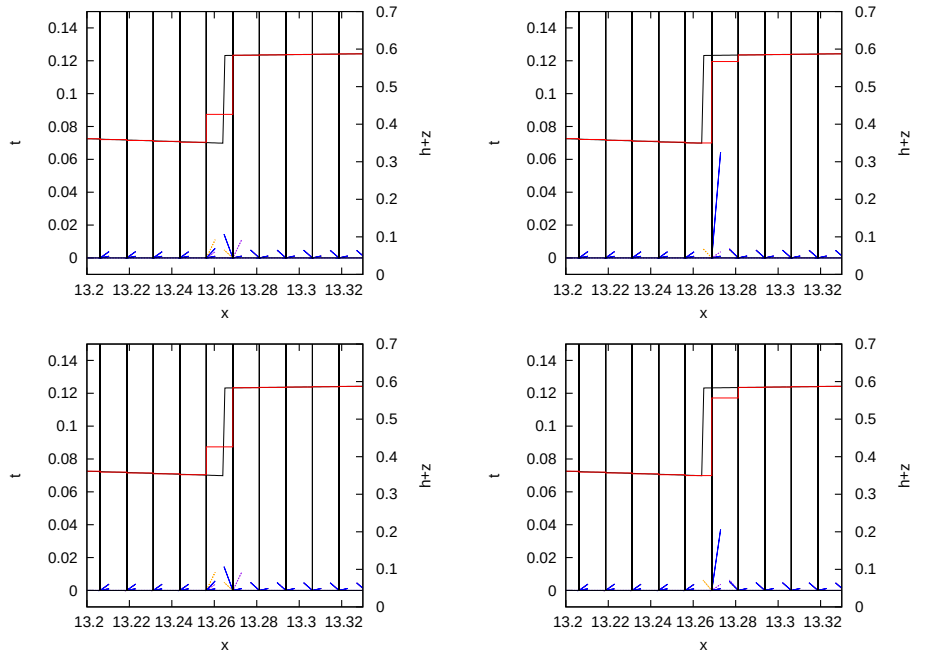


Figure 10.3: Section 10.2.1. Numerical solution for $h+z$ (—) computed by the ARoe solver in combination with the DF (top left), IF (top right), SEBF (bottom left) and WEBF (bottom right) using 1600 cells, including the representation of the exact solution (—) and wave speeds λ_L (—), λ_R (—) and $\tilde{\lambda}$ (—).

10.3 EB schemes with arbitrary order: the EB AR(L)-ADER HLLS(L)-ADER schemes

When constructing first order EB augmented schemes, the conservation of energy was only imposed in the discretization of the source term at cell interfaces [81] since variations of the variables along the cell length

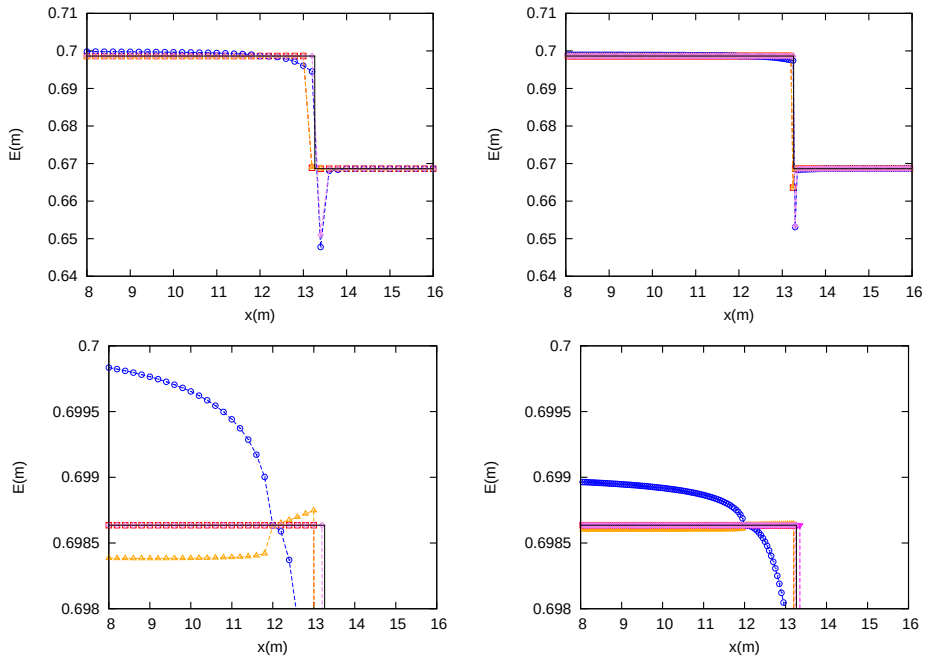


Figure 10.4: Section 10.2.1. Numerical solution for the specific mechanical energy computed by the ARoe solver in combination with the DF (— △ —), IF (— ○ —), SEBF (—□—) and WEBF (— ◇ —) (left) and detail of the solution (right).

were nil. When moving to high order schemes of the ADER type using (5.11), where spatial variations of the variables along the cell do exist, the conservation of energy must also be taken into account in the calculation of the integral of the source term inside the cell.

To construct an EB ADER scheme for the SWE, an energy-conservative discretization of the source term both at cell interfaces and inside cell is required. Such EB discretization is only mandatory for the leading terms of (5.21) and (5.19), that is, for the 0-th order coefficient of the integral of the source term at the interface, $\bar{S}^{(0)}_{i+1/2}$, and for the 0-th order coefficient of the centered integral of the source term, $\bar{S}^{(0)}_{i_L,i_R}$, respectively. On the other hand, the coefficients of higher order terms of the Taylor series expansion are time derivatives of the source term, which vanish under steady state. Hence, no particular EB discretization is required for such terms, for instance we can use

$$\bar{S}^{(k)}_{i+1/2} = -g\left(\frac{h^{(k)}_{(i+1)_L} + h^{(k)}_{i_R}}{2}\right)\delta z \tag{10.22}$$

to approach (10.5), with $h^{(k)}$ the k-th time derivative of h.

It is worth pointing out that the centered integral of the source term, $\bar{S}^{(k)}_{i_L,i_R}$, (both for the leading term and higher order terms) has to be computed using a quadrature rule that ensures the sought order of convergence. To this end, the Romberg integration method will be used following [75].

As mentioned in previous chapters, when constructing ADER schemes, an spatial reconstruction procedure is required. In this work, we use the WENO method. To construct an EB scheme, the WENO reconstruction must be carried out in a particular way so that the energy is conserved at the discrete level.

In order to preserve the EB property, the spatial reconstruction must be carried out for the specific mechanical energy and the unitary discharge

$$E = \frac{u^2}{2g} + h + z\,, \qquad q = hu\,. \tag{10.23}$$

Then, the water depth h is computed from Equation (10.23) where all the remaining terms are given by the spatial reconstructions of E and q.

Time derivatives of the fluxes, conserved variables and sources are computed by means of spatial derivatives of the conserved variables using the CK procedure. When carrying out this procedure, time derivatives should be expressed as a sum of terms where spatial derivatives of E and q are common factors. In this way, functions $\mathbf{R}^{(k)}$, $\mathbf{D}^{(k)}$ and $\mathbf{Q}^{(k)}$ become nil as there is no spatial variation of energy or discharge, ensuring the discrete equilibrium.

A summary of the different reconstruction possibilities and the properties expected for the scheme is listed below:

a) Reconstruct 3 variables:

- $p_h \approx h$
- $p_q \approx q$
- $p_z \approx z$

Not well-balanced.

b) Reconstruct 3 variables:

- $p_{h+z} \approx h + z \Rightarrow$ *well-balanced equilibrium variable*
- $p_q \approx q$
- $p_z \approx z$

and calculate $h = p_{h+z} - p_z$.

Well-balanced, not constant discharge.

c) Reconstruct 3 variables:

- $p_E \approx \frac{1}{2g}u^2 + h + z \Rightarrow$ *energy balanced equilibrium variable*
- $p_q \approx q$
- $p_z \approx z$

and calculate h from energy equation (10.23) as $h = h(p_E, p_q, p_z)$ using Cardano's algorithm.

Well-balanced (if using DF), EB (if using SEBF or WEBF).

For the sake of clarity we denote by p_π the spatial reconstruction of quantity π provided by the WENO reconstruction.

10.3.1 Integral of the source term inside the cell

The leading term of the centered approximation of the integral of the source term, $\bar{S}^{(0)}_{i_L,i_R}$, is computed using the same approach than in (10.20). When the cell does not contain an hydraulic jump, the integral of the source term is approached as

$$\bar{S}^{(0)}_{i_L,i_R} = -g\bar{h}\delta z + (\overline{hu} - \bar{h}\bar{u})\delta u\,, \tag{10.24}$$

where the second term on the right hand side can be expressed as

$$\breve{s} = (\overline{hu} - \bar{h}\bar{u})\delta u\,. \tag{10.25}$$

Equation (10.25) represents the correction of the source term required to exactly balance the difference of fluxes. It is evidenced that the reason why the correction must be made is that $\widetilde{hu} \neq \widetilde{h}\widetilde{u}$. It is worth showing that $\check{s}$ can also be written as

$$\check{s} = \delta(hu^2) - \bar{h}\delta\left(\frac{1}{2}u^2\right) = \mathscr{A}(S_z^b - S_z^a), \tag{10.26}$$

using the definition of $\mathscr{A}$ in [81]. In this way, the correction in (10.25) or (10.26) can be given a physical meaning, that is the difference between the differential discretization, S_z^a, and the integral discretization, S_z^b. In [75], they showed that $\check{s}$ can be rewritten as a third order difference

$$\check{s} = \frac{1}{4}\delta h\,(\delta u)^2\ . \tag{10.27}$$

As outlined before, the correcting term is a third-order difference and therefore (10.27) is second-order accurate as a quadrature for the source term [75]. It is evidenced that difficulties may arise when constructing high order EB numerical schemes, as it may become a hard task to preserve the discrete energy balance property while ensuring the theoretical order of accuracy. To address this issue, in [75] a fifth order EB scheme was constructed by approximating the mentioned integral by a fourth order extrapolation of a second order EB discretization of the source term.

In the present work, the approach in [75] is followed and the second order approximation of the integral of the source term provided by (10.12) is extended to higher order of accuracy by extrapolation. The resulting approximation of the integral will now be of arbitrary order while still ensuring the conservation of energy, however, the computational cost will be higher since a recursive procedure is required to carry out the extrapolation.

To obtain a $K+1$-th order scheme, it is neccessary to extend this integration technique to arbitrary order in space. To this end, we can use Romberg integration, which is a result that can be obtained from Richardson's extrapolation. This technique can be used for the DF, WEBF and SEBF approaches. Here, the derivation of a 4-th order integration rule for the x-direction integral is shown. Let us define $I(S)$ as the exact integral

$$I(S) = \int_{x_{i_L}}^{x_{i_R}} S\ dx\ , \tag{10.28}$$

inside the integration domain $x \in \Upsilon = [x_{i_L}, x_{i_R}]$. On the one hand, we can approximate the previous integral using the trapezoid rule (taking the two extrema of the domain) as

$$I(S) = \left\{\bar{S}_i^{(0)}\right\}_0^0 + K_1\Delta x^2 + \mathscr{O}(\Delta x^4), \tag{10.29}$$

where $\left\{\bar{S}_i^{(0)}\right\}_0^0$ as defined by the DF, WEBF or SEBF, for instance (10.24) for the latter. On the other hand, we can approximate $I(S)$ by dividing Υ into two intervals and using the trapezoid rule inside each of them

$$I(S) = \left\{\bar{S}_i^{(0)}\right\}_1^0 + K_2\frac{\Delta x}{2}^2 + \mathscr{O}(\Delta x^4), \tag{10.30}$$

where $\left\{\bar{S}_i^{(0)}\right\}_1^0 = \bar{S}_{i_R,i}^{(0)} + \bar{S}_{i,i_L}^{(0)}$.

Combination of (10.29) and (10.30) allows to find a 4-th order approximation as

$$I(S) = \frac{4\left\{\bar{S}_i^{(0)}\right\}_1^0 - \left\{\bar{S}_i^{(0)}\right\}_0^0}{3} + \mathscr{O}(\Delta x^4), \tag{10.31}$$

which will be referred to as

$$\left\{\bar{S}_i^{(0)}\right\}_2^1 = \frac{4\left\{\bar{S}_i^{(0)}\right\}_2^0 - \left\{\bar{S}_i^{(0)}\right\}_1^0}{3}. \tag{10.32}$$

It is worth pointing out that an arbitrary order integral is denoted as $\left\{\bar{S}_i^{(0)}\right\}_m^n$, where m is the number of subdivisions of the initial interval Υ, with step size

$$h_m = \frac{\Delta x}{2^m} \tag{10.33}$$

and n the number of Romberg iterations, with a magnitude of the residual of $\mathcal{O}(\Delta x^{2(n+1)})$. To construct a $K+1$-th order ADER scheme (with K non-trivial derivatives), both n and m will take values up to $\left\lceil \frac{K-1}{2} \right\rceil$ and the order of accuracy of the method will be $\mathcal{O}(\Delta x^{K+2})$ if K is even or $\mathcal{O}(\Delta x^{K+1})$ if K is odd. The expression for $\left\{\bar{S}_i^{(0)}\right\}_m^n$ is computed recursively departing from the trapezoid integrals, that is $n=0$ and $m=0,...,\left\lceil \frac{K-1}{2} \right\rceil$, and computing the following levels $n=1,...,\left\lceil \frac{K-1}{2} \right\rceil$ as

$$\left\{\bar{S}_i^{(0)}\right\}_{m+1}^{n+1} = \frac{4^n\left\{\bar{S}_i^{(0)}\right\}_{m+1}^n - \left\{\bar{S}_i^{(0)}\right\}_m^n}{4^n - 1}. \tag{10.34}$$

Concerning the derivative terms, there is no need of a particular discretization technique of the source term to ensure the well-balanced property as time derivatives vanish under steady state. Here, we use a 2D Gaussian integration

$$\bar{S}_{i_L,i_R}^{(k)} = \sum_{\alpha=1}^{k} w_\alpha \left(-g h^{(k)} \partial_x z\right)_\alpha, \tag{10.35}$$

where $h_\alpha^{(k)}$ is the k-th time derivative of h at the quadrature point.

The expression for the integral of the source term inside the cell (5.19) reads

$$\bar{\mathbf{S}}_{i_R,i_L} = \left\{\bar{\mathbf{S}}_i^{(0)}\right\}_m^n + \sum_{k=1}^{K} \bar{\mathbf{S}}_{i_R,i_L}^{(k)} \frac{\Delta t^k}{(k+1)!}. \tag{10.36}$$

It is worth pointing out that under steady conditions $\left\{\bar{S}_i^{(0)}\right\}_m^n = \left\{\bar{S}_i^{(0)}\right\}_0^0$, for any m and n, therefore the EB condition is always satisfied.

When using the proposed EB arbitrary order integration of the source term in (10.36) in combination with the AR-(L)FS or HLLS-(L)FS solvers, the resulting schemes are termed AR(L)-ADER and HLLS(L)-ADER schemes.

10.3.2 Test case: steady subcritical flow over a hump

In this test case, a subcritical flow over a hump is computed using the AR-ADER scheme. Two different source term discretizations, the DF and the SEBF, are used, leading to a well-balanced and a EB scheme, respectively. Moreover, two different choices for the spatial reconstruction of the variables are assessed. The first one is to reconstruct over $h+z$ and the second choice is to reconstruct over E.

The test case is set up as follows. The bed elevation is given by the following function

$$z(x) = \begin{cases} 0 & \text{if} \quad x < 8 \\ 0.2 - 0.05\,(x-12)^2 & \text{if} \quad 10 \le x \le 14 \\ 0 & \text{if} \quad x > 12 \end{cases} \tag{10.37}$$

inside the domain $[0, 25]$ and the boundary conditions are set to $q = 4.42$ m^2/s at the inlet and $h = 2$ m at the outlet. CFL is set to 0.4 and the acceleration of gravity to 9.8 m/s^2.

The numerical solution is computed with the 3-rd order AR-ADER scheme and the three following combinations of data reconstruction set and source term discretization:

a) Reconstruction over $h+z$ and DF source term integration.

b) Reconstruction over E and DF source term integration.

c) Reconstruction over E and SEBF source term integration.

Image 10.5 shows the numerical solution for $h+z$, q and E computed using the 3 previously mentioned approaches. It is observed that the approach a), which is the basic well-balanced formulation, does not ensure neither the conservation of energy nor a constant discharge. On the other hand, when using approaches b) and c), a constant discharge is ensured but only approach c) allows to preserve the discrete energy level.

It is worth mentioning that the DF formulation does ensure a constant discharge when using a 1-st order augmented scheme, as reported in Section 10.2.1.

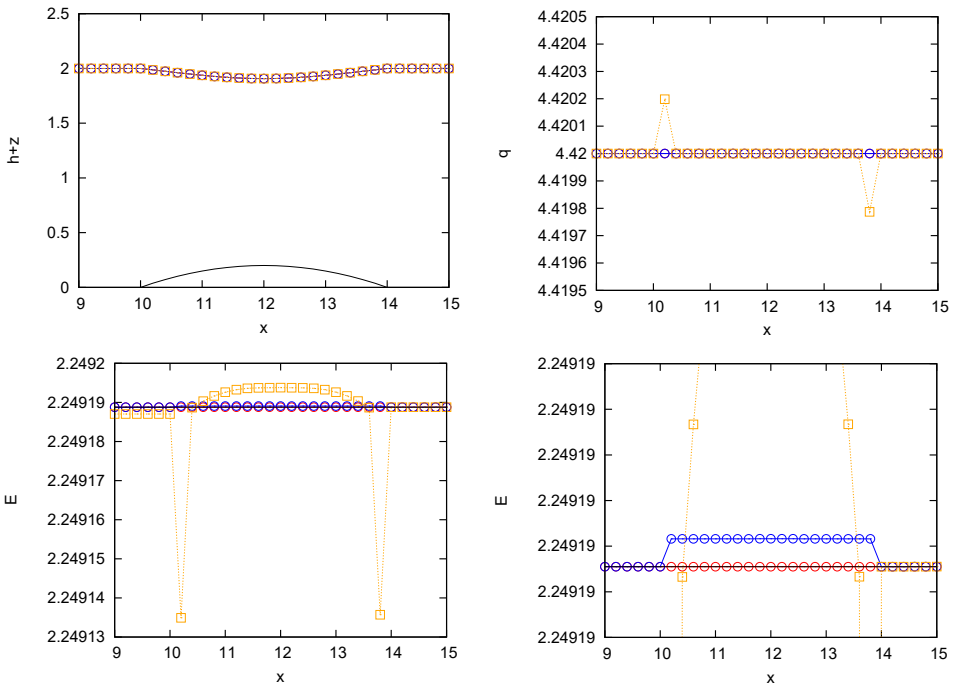


Figure 10.5: Exact (—) and numerical solution for $h+z$ (top left) and q (top right) computed by the 3-rd order AR-ADER scheme using approach a) (— △ —), b) (— ∘ —) and c) (—□—) using $\Delta x = 0.2$.

10.3.3 Test case: numerical performance in RPs

The numerical resolution of RPs using the 1-st, 3-rd and 5-th order AR-ADER, ARL-ADER, HLLS-ADER and HLLSL-ADER methods is considered in this test case. 4 Different RPs, involving different combinations of wave patterns in presence of bed discontinuities, are proposed. The ICs for such RPs are summarized in Table 10.1. The problems 2 and 3 are included in a list of RPs defined by LeFloch and Duc-Thanh [67]. The spatial domain is given by $[-0.5, 0.5]$, the bottom step is positioned at $x = 0$ and the acceleration of gravity is set to g=9.8 m/s^2. The domain is divided in 500 cells and 1000 cells to show the convergence

RP	h_L	h_R	u_L	u_R	z_L	z_R
1	4.0	0.69196567	0.0	0.0	0.0	1.5
2	0.3	0.39680194	2.0	2.2	1.0	1.0
3	1.0	1.2	3.0	0.1	1.1	1.0
4	1.0	2.0	2.0	4.0	1.1	1.0

Table 10.1: Summary of test cases.

of the schemes when the grid is refined. Numerical solutions are plotted for $h+z$ and q at time $t = 0.01$ s. For all the problems, we set CFL=0.2.

The RP 1 is a dam-break type problem with a initial condition consisting of two columns of water of different height and zero velocity, with a discontinuity in bed elevation. The solution contains a left-moving rarefaction wave, a stationary discontinuity at the step and a right-moving shock wave. Numerical solutions are compared with the exact solution in Figures 10.6, 10.7, 10.8, 10.9. All the schemes converge to the exact solution when the grid is refined and are able to accurately capture the stationary discontinuity generated by the source term at $x = 0$. When comparing the numerical solutions provided by the different schemes, it is possible to notice that the HLLS-ADER scheme provide a less diffusive numerical solution than the others. There are not noticeable differences between the AR-ADER, the ARL-ADER and the HLLSL-ADER schemes.

Supercritical motion from left to right is considered in RP 2. Numerical solutions are compared with the exact solution in Figures 10.10, 10.11, 10.12, 10.13. The four numerical schemes converge to the exact solution when the grid is refined. It is observed that the presence of the discontinuity in bed elevation at $x = 0$ does not introduce any disturbance in the wave patterns on the right side. When seeking differences among the numerical solutions provided by the different schemes, it is possible to observe that the numerical solution provided by the HLLS-ADER scheme is slightly less diffusive than those solutions provided by the other methods, among which no relevant differences can be found.

RP 3 is a resonant problem that admits only one solution given by a sequence of shocks. Numerical solutions are compared with the exact solution in Figures 10.14, 10.15, 10.16, 10.17. It can be observed that all the schemes converge to the exact solution when the grid is refined. Here, numerical results evidence again that the HLLS-ADER scheme provides a more accurate solution around discontinuities.

As a general observation on the previous test cases, we can say that all the numerical schemes provide accurate results for the water level surface and unitary discharge at the bed discontinuity and convergence is ensured with mesh refinement or when numerical order is increased.

RP 4 also is a resonant problem that admits only one solution. The solution begins with a rarefaction, followed by a stationary contact, continued by a shock wave and finally ends in a rarefaction. The numerical solutions are compared with the exact solution in Figures 10.18, 10.19, 10.20, 10.21. When using ARoe type schemes, namely the AR-ADER and the ARL-ADER schemes, convergence to the exact solution is ensured. On the other hand, when using HLLS-type schemes for RP 4, namely the HLLS-ADER and HLLSL-ADER schemes, such methods are unable to converge to the exact solution, as shown in Figures 10.20 and 10.21. This can be caused due to an improper choice of the wave celerities.

10.3.4 Convergence rate test

The following test case corresponds to the computation of the evolution of a smooth initial condition over a smooth bed elevation, given by

$$z(x) = \begin{cases} 0 & \text{if} \quad x < 2 \\ 0.01\sin(\pi x)^4 & \text{if} \quad 2 \le x \le 3 \\ 0 & \text{if} \quad x > 3 \end{cases} \tag{10.38}$$

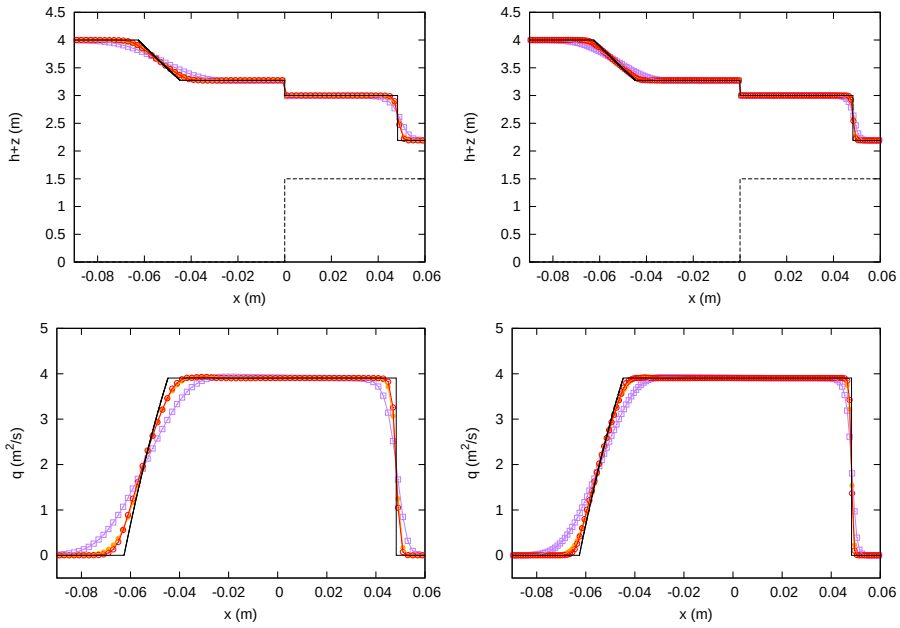


Figure 10.6: Section 10.3.3. RP 1. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order AR-ADER method using (left) 500 and (right) 1000 cells.

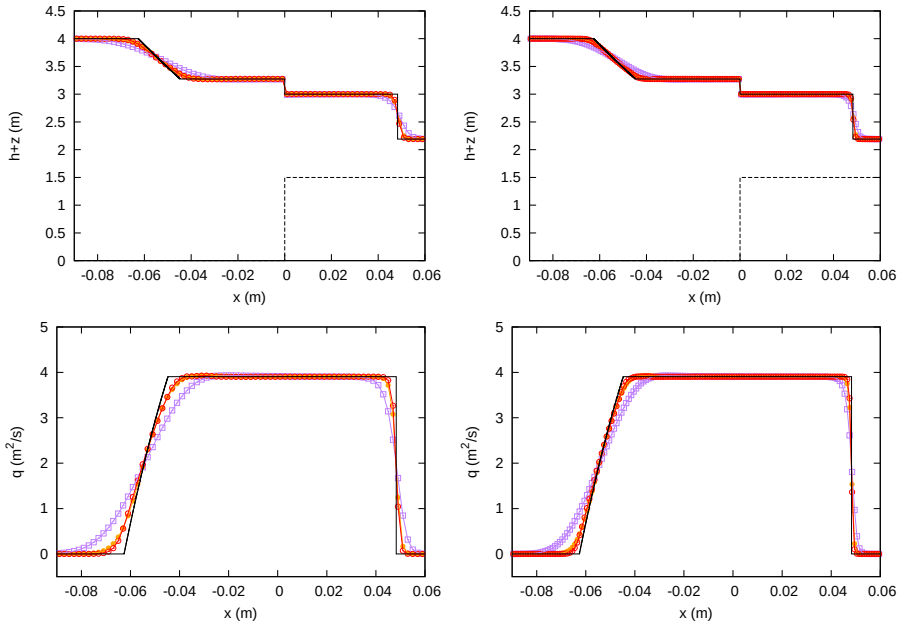


Figure 10.7: Section 10.3.3. RP 1. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order ARL-ADER method using (left) 500 and (right) 1000 cells.

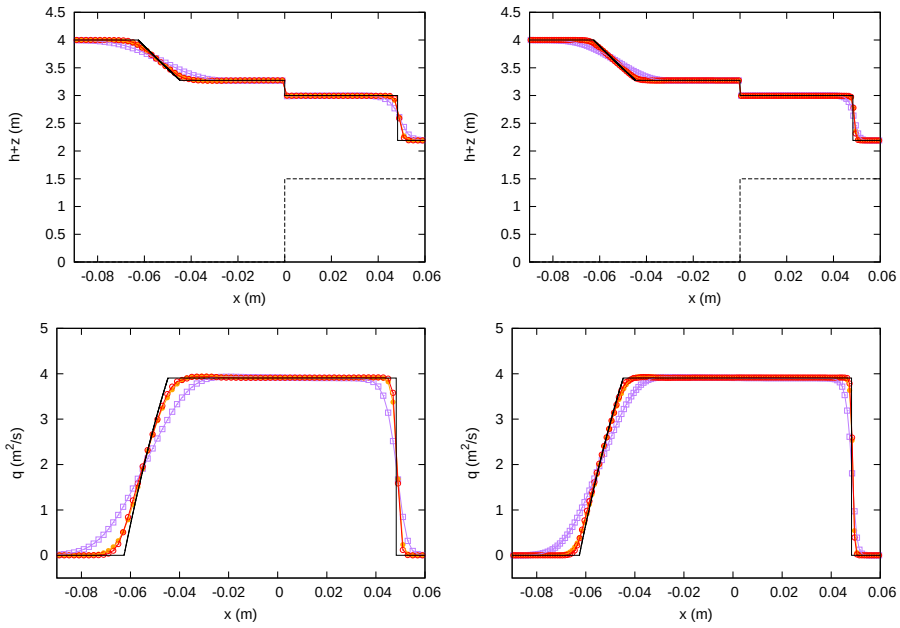


Figure 10.8: Section 10.3.3. RP 1. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order HLLS-ADER method using (left) 500 and (right) 1000 cells.

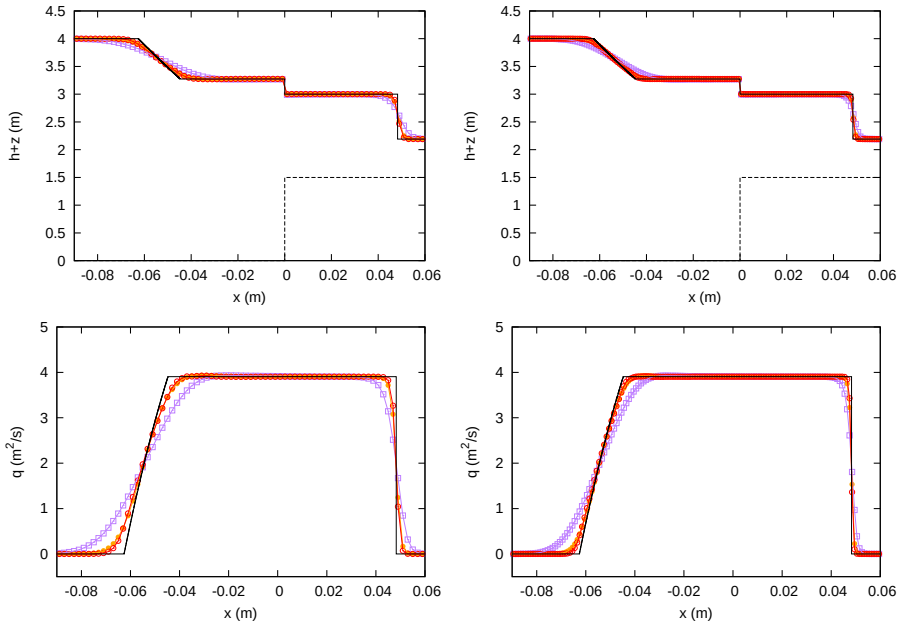


Figure 10.9: Section 10.3.3. RP 1. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order HLLSL-ADER method using (left) 500 and (right) 1000 cells.

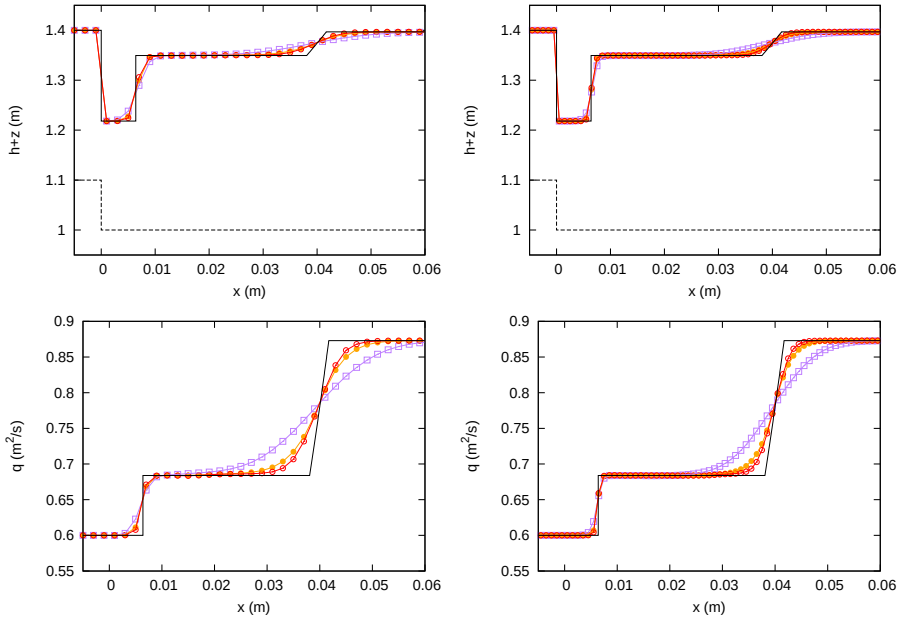


Figure 10.10: Section 10.3.3. RP 2. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order AR-ADER method using (left) 500 and (right) 1000 cells.

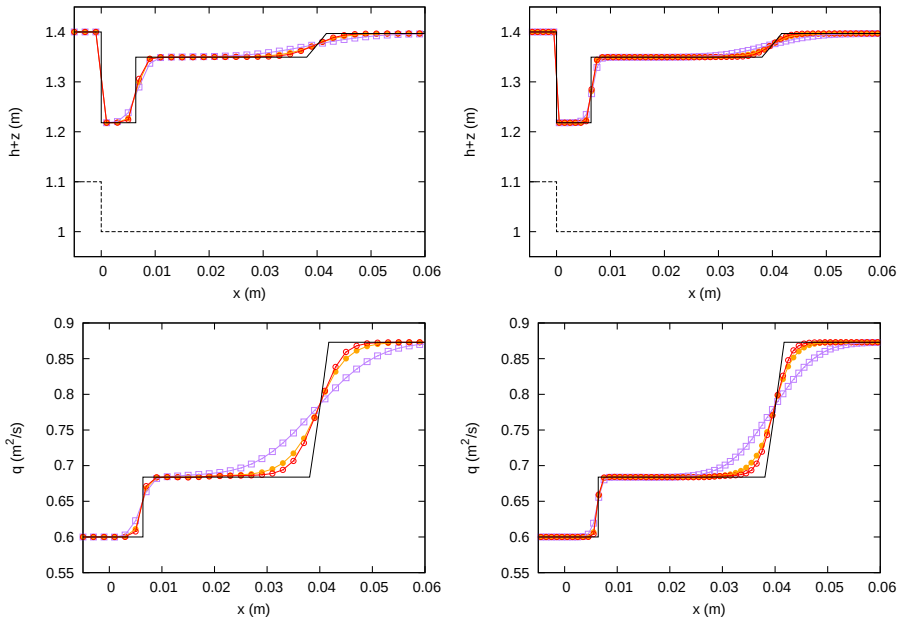


Figure 10.11: Section 10.3.3. RP 2. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order ARL-ADER method using (left) 500 and (right) 1000 cells.

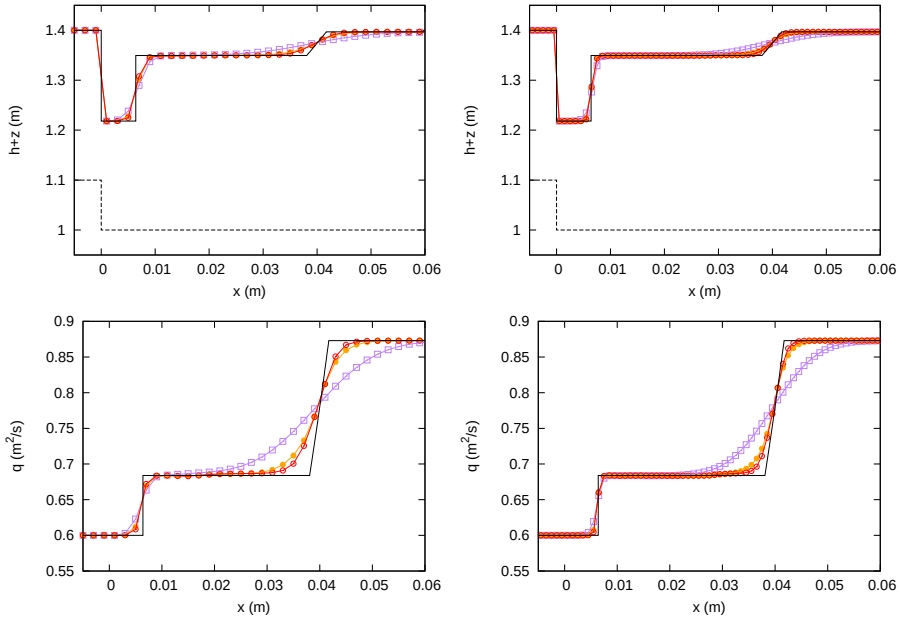


Figure 10.12: Section 10.3.3. RP 2. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order HLLS-ADER method using (left) 500 and (right) 1000 cells.

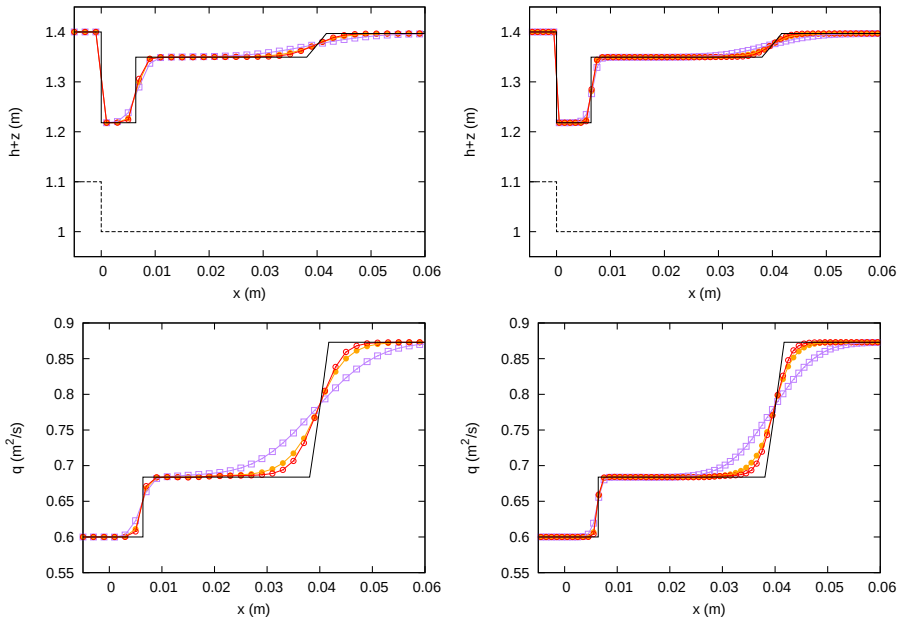


Figure 10.13: Section 10.3.3. RP 2. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order HLLSL-ADER method using (left) 500 and (right) 1000 cells.

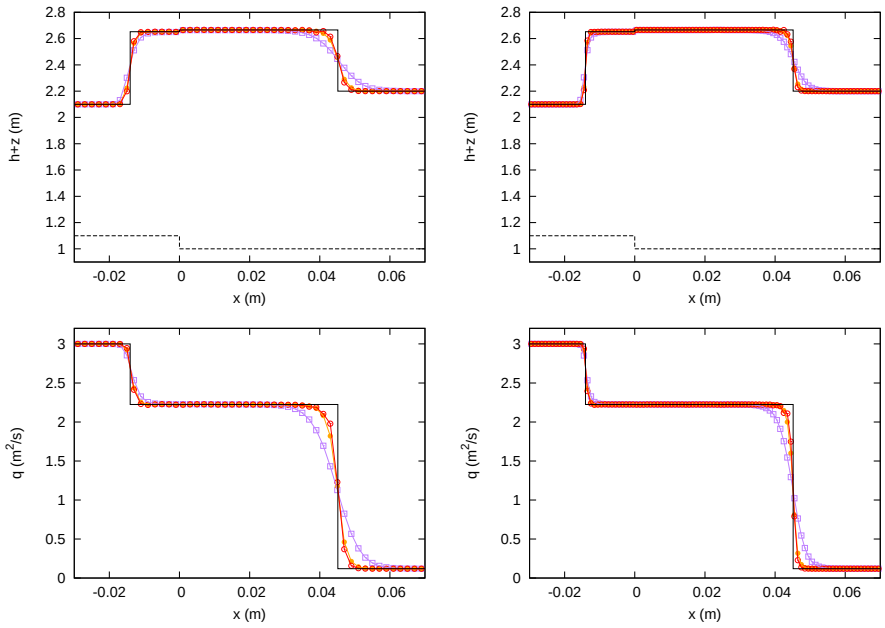


Figure 10.14: Section 10.3.3. RP 3. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order AR-ADER method using (left) 500 and (right) 1000 cells.

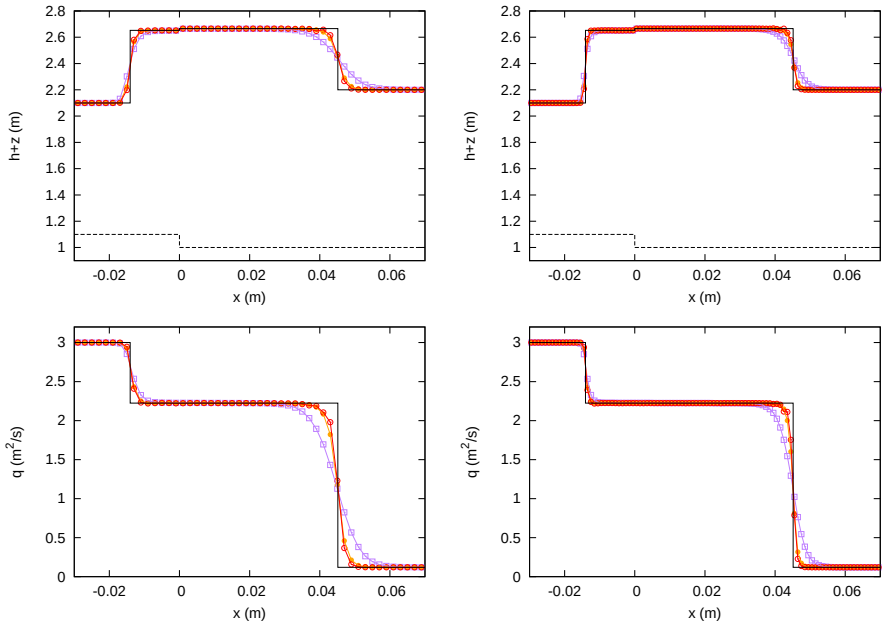


Figure 10.15: Section 10.3.3. RP 3. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order ARL-ADER method using (left) 500 and (right) 1000 cells.

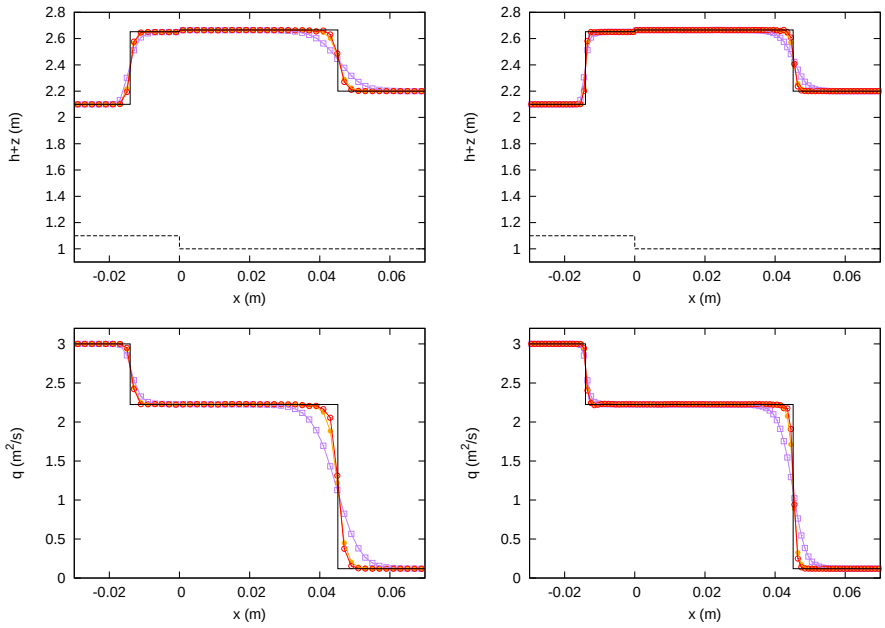


Figure 10.16: Section 10.3.3. RP 3. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order HLLS-ADER method using (left) 500 and (right) 1000 cells.

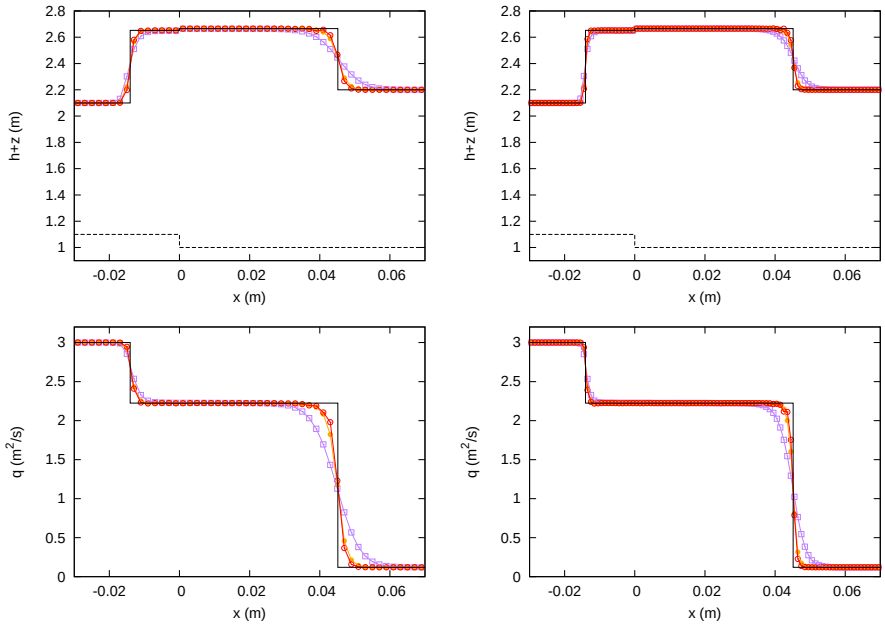


Figure 10.17: Section 10.3.3. RP 3. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order HLLSL-ADER method using (left) 500 and (right) 1000 cells.

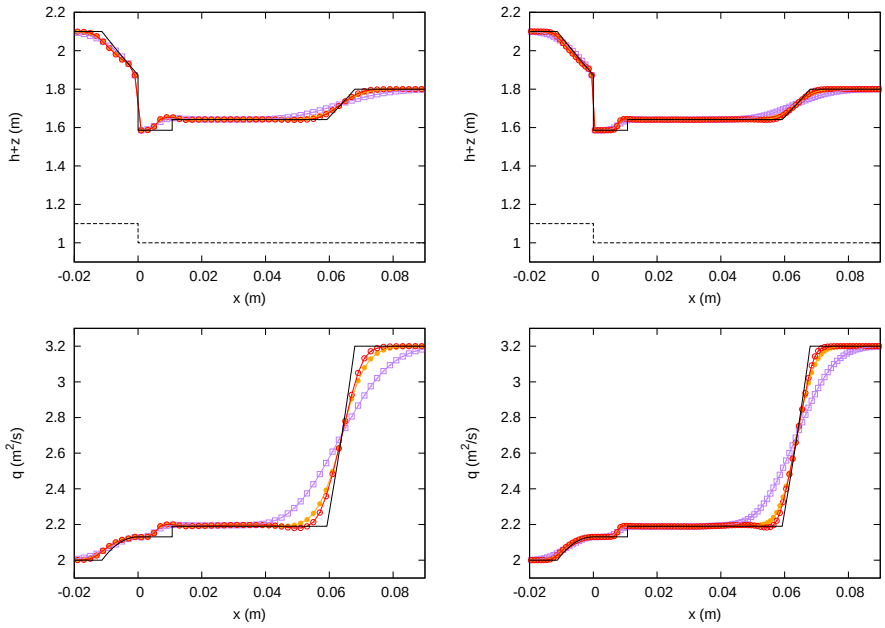


Figure 10.18: Section 10.3.3. RP 4. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order AR-ADER method using (left) 500 and (right) 1000 cells.

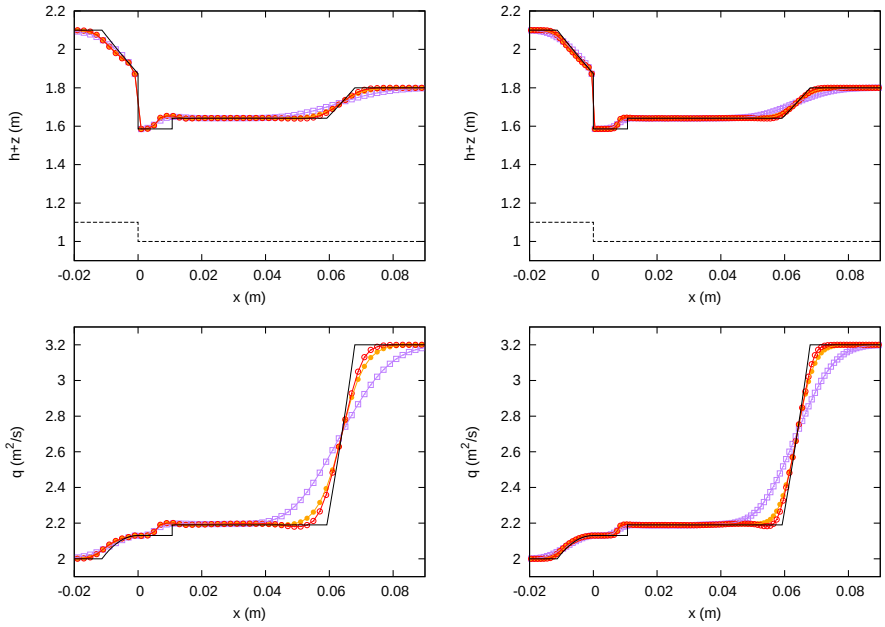


Figure 10.19: Section 10.3.3. RP 4. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order ARL-ADER method using (left) 500 and (right) 1000 cells.

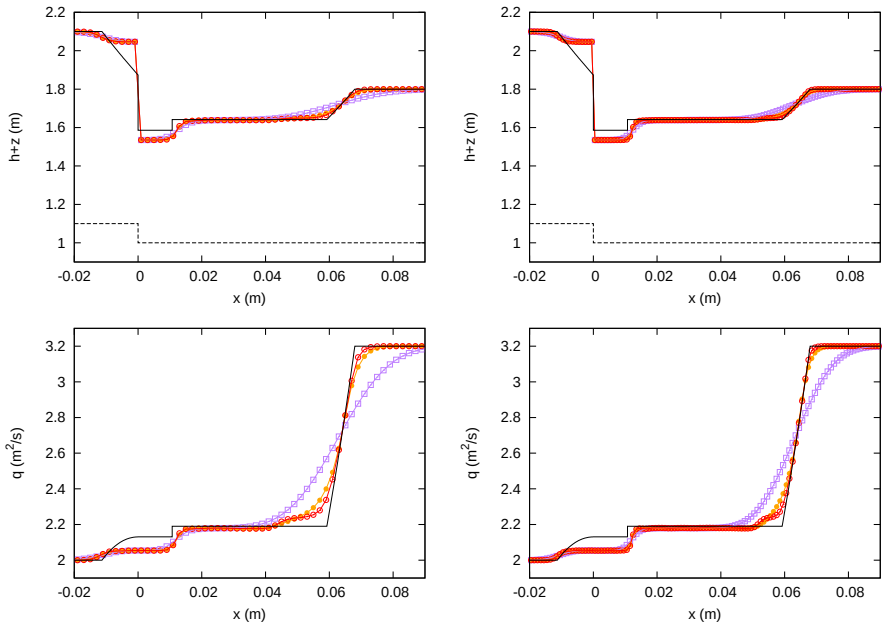


Figure 10.20: Section 10.3.3. RP 4. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order HLLS-ADER method using (left) 500 and (right) 1000 cells.

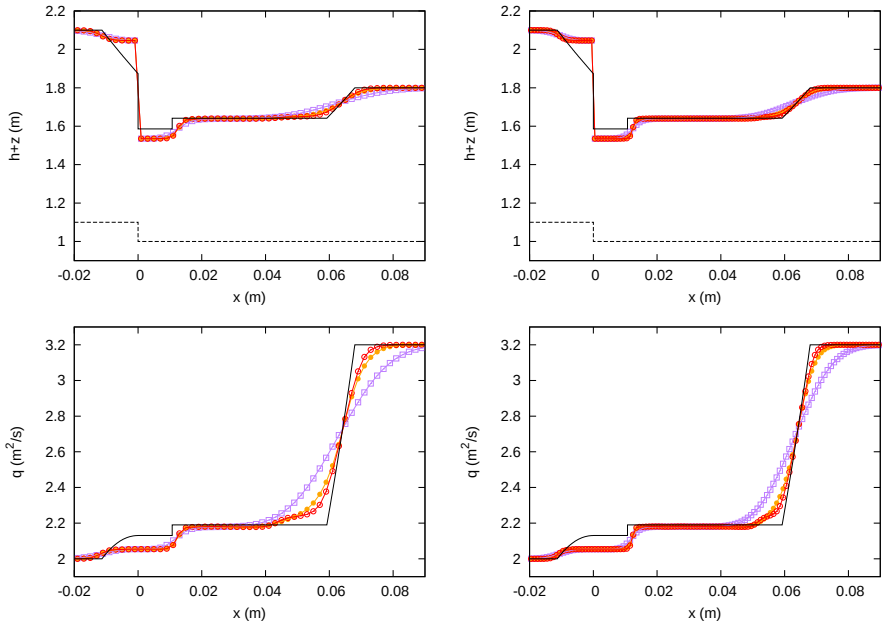


Figure 10.21: Section 10.3.3. RP 4. Exact solution (—) and numerical solutions using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order HLLSL-ADER method using (left) 500 and (right) 1000 cells.

		AR-ADER				HLLS-ADER			
		h		q		h		q	
Scheme	N	L_1 error	Order	L_1 error	Order	L_1 error	Order	L_1 error	Order
3-rd	40	1.02E-03		2.74E-03		1.10E-03		2.91E-03	
	60	4.78E-04	1.87	1.12E-03	2.19	5.09E-04	1.90	1.22E-03	2.14
	80	2.30E-04	2.54	5.27E-04	2.64	2.48E-04	2.50	5.76E-04	2.61
	100	1.27E-04	2.65	2.89E-04	2.69	1.39E-04	2.60	3.19E-04	2.65
	125	6.77E-05	2.83	1.51E-04	2.91	7.40E-05	2.82	1.67E-04	2.91
5-th	40	5.58E-04		1.13E-03		5.68E-04		1.21E-03	
	60	1.28E-04	3.63	2.39E-04	3.84	1.40E-04	3.46	2.66E-04	3.74
	80	3.63E-05	4.38	6.76E-05	4.39	3.98E-05	4.36	7.53E-05	4.39
	100	1.32E-05	4.55	2.49E-05	4.48	1.45E-05	4.52	2.77E-05	4.48
	125	4.86E-06	4.47	9.00E-06	4.55	5.37E-06	4.46	1.01E-05	4.54

Table 10.2: Section 10.3.4. Convergence rate test for h and q using the L_1 error norm for the 3-rd and 5-th order EB AR-ADER and HLLS-ADER scheme. CFL=0.3.

		ARL-ADER				HLLSL-ADER			
		h		q		h		q	
Scheme	N	L_1 error	Order	L_1 error	Order	L_1 error	Order	L_1 error	Order
3-rd	40	1.21E-03		3.23E-03		1.21E-03		3.23E-03	
	60	5.75E-04	1.84	1.31E-03	2.22	5.75E-04	1.84	1.31E-03	2.22
	80	2.89E-04	2.39	6.46E-04	2.47	2.89E-04	2.39	6.46E-04	2.47
	100	1.60E-04	2.64	3.54E-04	2.7	1.60E-04	2.64	3.54E-04	2.7
	125	8.54E-05	2.82	1.87E-04	2.85	8.54E-05	2.82	1.87E-04	2.85
5-th	40	7.12E-04		1.33E-03		7.12E-04		1.33E-03	
	60	1.50E-04	3.84	2.68E-04	3.95	1.50E-04	3.84	2.68E-04	3.95
	80	4.45E-05	4.22	8.10E-05	4.16	4.45E-05	4.22	8.10E-05	4.16
	100	1.58E-05	4.64	2.89E-05	4.61	1.58E-05	4.64	2.89E-05	4.61
	125	5.87E-06	4.44	1.04E-05	4.6	5.87E-06	4.44	1.04E-05	4.6

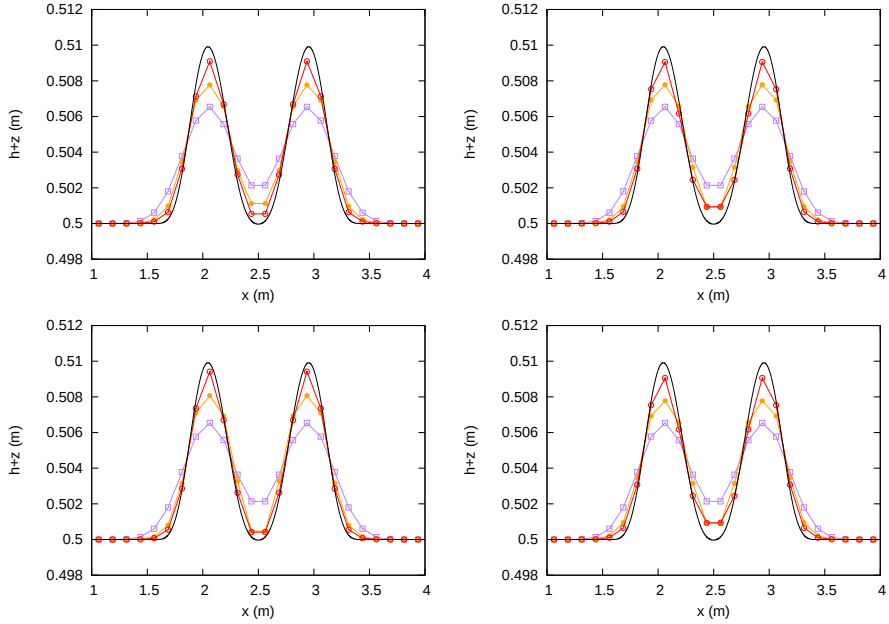
Table 10.3: Section 10.3.4. Convergence rate test for h and q using the L_1 error norm for the 3-rd and 5-th order EB ARL-ADER and HLLSL-ADER scheme. CFL=0.05.

$$h(x,0) = 0.5 + z(x), \qquad q(x,0) = 0 \tag{10.39}$$

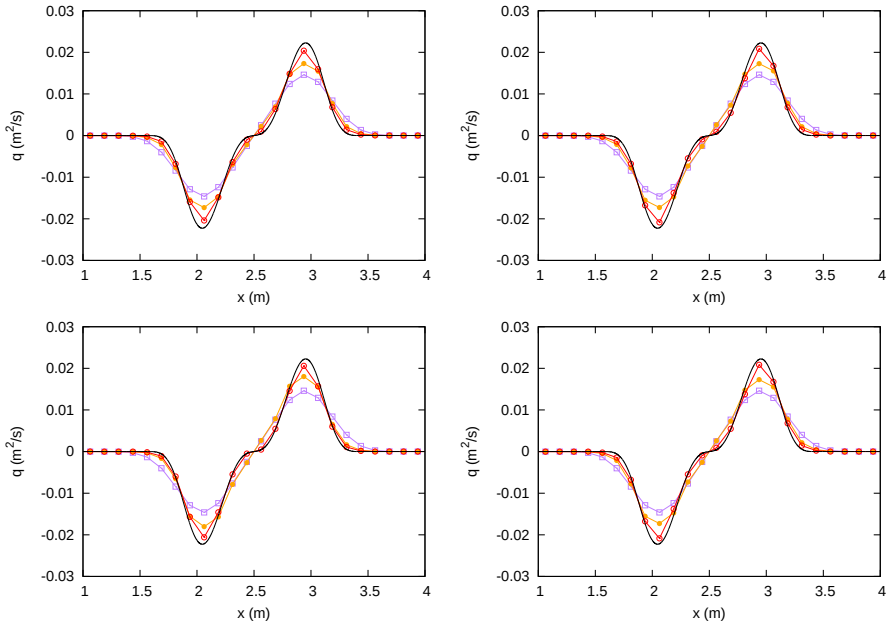
inside the spatial domain $\Omega = [0,5]$ with $x \in \Omega$. The initial condition for the water depth in (10.39), setting the unitary discharge to zero at the initial time, leads to two symmetric waves that move in opposite directions, as shown in Figures 10.22a and 10.22b. Numerical results are computed using the first order ARoe and HLLS schemes, as well as their higher order ADER versions presented in this work, setting CFL to 0.3. Such solutions are compared with a reference solution, computed with a 5-th order AR-ADER scheme using 8000 cells. Numerical solutions for water surface elevation and unitary discharge provided by all the numerical schemes as well as the reference solution for such quantities are plotted at time $t = 0.2$ s in Figures 10.22a and 10.22b respectively.

In Tables 10.2 and 10.3 numerical errors provided by the 3-rd and 5-th order AR-ADER, ARL-ADER, HLLS-ADER and HLLSL-ADER EB schemes are presented. Convergence rate tests have been carried out setting CFL=0.3 for the AR-ADER and HLLS-ADER schemes and CFL=0.05 for their linearized version, the ARL-ADER and HLLSL-ADER schemes. Numerical results in Tables 10.2 and 10.3 evidence that all the numerical schemes converge to the reference solution at the prescribed rate.

It has been observed that those numerical schemes using the LFS solver, namely the ARL-ADER and HLLSL-ADER schemes, may become suboptimal for high CFL numbers. To study this particular behavior, we have repeated the convergence rate tests using different CFL numbers for the 3-rd and 5-th order HLLS-ADER and HLLSL-ADER schemes. Numerical results are presented in Figure 10.23 for CFL= 0.6, CFL= 0.3, CFL= 0.15 and CFL= 0.08. It is observed that the 3-rd order version of the HLLS-ADER and HLLSL-ADER scheme do converge at the expected rate. However, when considering the 5-th order version of such schemes, the HLLS-ADER scheme does converge at the prescribed rate but the HLLSL-ADER scheme appears



(a) Refined solution and numerical solutions for the water surface elevation $h+z$.



(b) Refined solution and numerical solutions for the discharge q.

Figure 10.22: Section 10.3.4. Refined solution (—) and numerical solutions for the water surface elevation $h+z$ and discharge q using the 1-st (—□—), 3-rd (—●—) and 5-th (—○—) order AR-ADER scheme (upper left), ARL-ADER scheme (upper right), HLLS-ADER scheme (lower left) and HLLSL-ADER scheme (lower right), using 40 cells.

to be suboptimal for the higher CFL numbers. This behavior is due to the linearization that has been carried out to construct the HLLSL-ADER scheme by means of the LFS Derivative Riemann solver. Similar results have been reported for the AR-ADER and ARL-ADER schemes.

Therefore, the choice of such a small CFL number when using the linearized LFS solver is related to the recovery of the optimal accuracy of the numerical scheme rather than to any issue related to stability. When setting the CFL number to 0.3 or even higher, LFS-based schemes, namely the ARL-ADER and HLLSL-

ADER, are stable and still converge to the reference solution, however, their convergence rate is suboptimal as Figure 10.23 shows.

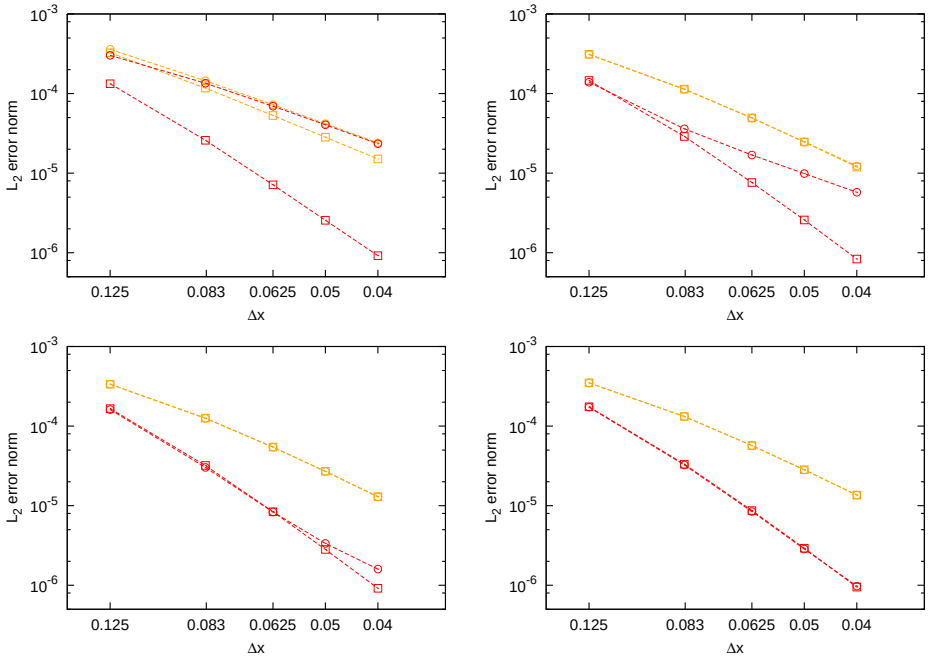


Figure 10.23: Section 10.3.4. L_1 error norm for the water depth h using the 3-rd order HLLS-ADER scheme (—□—), the 3-rd order HLLSL-ADER scheme (— ○ —) , the 5-th order HLLS-ADER sheme (—□—) and the 5-th order HLLSL-ADER sheme (— ○ —). Results computed setting CFL= 0.6 (upper left), CFL= 0.3 (upper right), CFL= 0.15 (lower left) and CFL= 0.08 (lower right).

CPU times for the present test case when using the 3-rd and 5-th order EB HLLSL-ADER, HLLS-ADER, ARL-ADER and AR-ADER schemes are presented in Table 10.4. The simulation time is set to $t = 3$ s and the CFL number to 0.3. Results are presented for two different grids composed of 80 and 160 cells respectively. Speed-ups of the linearized schemes, namely HLLSL-ADER and ARL-ADER, with respect to their nonlinear versions are also shown as a percentage. It is observed that the linearized LFS solver offers increasingly higher speed-ups with respect to the FS solver as the order of the numerical scheme is increased, for a fixed CFL number. While the 3-rd order HLLSL-ADER and ARL-ADER schemes only save around a 10% of the computational time required for the HLLS-ADER and AR-ADER schemes, the 5-th order schemes offer up to a 40% of computational time saving.

It is worth mentioning that the LFS solver avoids the computation of the CK procedure for the fluxes, which is a very tedious and expensive process regarding CPU time. The expression for the time derivatives of the fluxes, obtained by means of the CK procedure, become exponentially larger as the required order of accuracy is increased. That is why much larger CPU time saving is shown in Table 10.4 for the 5-th order LFS-based schemes than for the 3-rd order schemes. When using the LFS solver, CPU time saving is expected to increase as the order of accuracy of the scheme is increased. Therefore, when moving to very high order of accuracy, it is much more efficient to use a LFS-based scheme with a low CFL number than a FS-based scheme with higher CFL number. This issue is to be explored in the future.

10.4 Concluding remarks

The highlights of this chapter are listed below:

- The SEBF for the discretization of the source term, which allows to construct a EB numerical scheme

		HLLSL-ADER		HLLS-ADER	ARL-ADER		AR-ADER
Cells	Order	Time (s)	Speed-up	Time (s)	Time (s)	Speed-up	Time (s)
160	3	2.00	11%	2.25	1.98	11%	2.22
	5	54.92	39%	90.62	54.55	40%	90.26
80	3	0.50	11%	0.57	0.50	15%	0.58
	5	13.54	40%	22.52	13.56	41%	22.85

Table 10.4: CPU times for test case in Section 10.3.4 at $t = 3$ s, setting CFL=0.3. Times are shown for the 3-rd and 5-th order EB HLLSL-ADER, HLLS-ADER, ARL-ADER and AR-ADER schemes. Speed-ups of the HLLSL-ADER and ARL-ADER schemes with respect to their nonlinear version are shown as a percentage.

for the SWE, has been presented. The DF, IF and WEBF have also been recalled. The performance of the aforementioned formulations in the resolution of a steady hydraulic jump over a varying bed has been compared. Numerical results evidence that the SEBF outperforms the other techniques.

- The first order EB method has been extended to arbitrary order of accuracy. To this end, the bed slope source term has to be discretized using an EB formulation not only at cell interfaces, but also inside cells. The integral of the source term and flux fluctuations must be exactly balanced inside cells.

- A procedure for an arbitrary order EB discretization of the source term inside cells has been presented. It is based on a high order extrapolation of the first order EB formulation by means of Romber integration, as done in [75].

- The resulting schemes, the EB AR(L)-ADER and HLLS(L)-ADER methods, provide the exact solution for steady cases with moving water and irregular geometries, with independence of the grid. They ensure convergence to the exact solution at the prescribed rate for transient problems including complex RPs that involve bed variations and resonant solutions [67]. The numerical results evidence that the 4 schemes achieve the expected accuracy up to 5-th order. Higher orders of convergence have not been tested as the CK procedure becomes very tedious.

- It has been observed that the linearized version of the schemes, based on the LFS solver, are around a 10% (for 3-rd order) and a 40% (for 5-th order) faster then their corresponding non-linear versions for test case in Section 10.3.4. However, they require a tighter CFL restriction.

11 WELL-BALANCED SCHEMES FOR THE 2D SWE

This chapter is devoted to the resolution of the 2D SWE with source terms of different nature, such as bed variation, friction and Coriolis. Among the different schemes proposed in previous chapters, we choose the LFS solver in combination with the ARoe solver and extend it to 2 space dimensions using Cartesian meshes. The motivation for this choice is that the numerical method resulting from such combination, the ARL-ADER scheme [84], provides a good balance between accuracy and computational cost. It is worth recalling that this scheme is designed using augmented solvers and therefore is suitable for the numerical treatment of source terms of different nature, such as those considered for the SWE, and allows to preserve the discrete equilibrium when required.

As outlined in previous chapters, the preservation of equilibrium states is a matter of great importance when designing a numerical scheme. Most flows are often perturbations of certain equilibrium states. The very first property that the numerical scheme must satisfy is the preservation of the quiescent equilibrium at the discrete level, called well-balanced property [58, 59]. As shown before, the well-balanced property can still be enhanced by considering energy conservation criteria in the numerical scheme. This allowed the extension of well-balanced methods to exactly well-balanced methods, also called EB methods [75, 76, 77, 78, 79, 80]. When considering rotation, equilibrium states become more complex as they now include the circulation of the flow in particular directions. For the SWE with Coriolis, the most relevant equilibrium solution to be considered in the design of the numerical scheme is the so-called geostrophic equilibrium. It arises from the balance of the Coriolis force with the hydrostatic pressure change due to the surface elevation gradient. This steady state is often referred to as *jet in the rotating frame* [88].

The SWE in the rotating frame represents a good model for large scale phenomena in geophysical flows, in which oceanic and atmospheric circulations are often perturbations of the so-called geostrophic equilibrium [87, 88]. When addressing the numerical resolution of such model, the well-balanced property that allowed to correctly simulate the still water at rest must be extended to preserve the geostrophic equilibrium. This question is more delicate than the quiescent equilibrium for two reasons: it is an essentially 2D problem and it involves a non-zero velocity field [87]. In the last decade, a great effort has been put on the design of FV well-balanced numerical schemes capable to maintain the geostrophic equilibrium by following different approaches, some of high order of accuracy [89, 90, 91, 92, 93, 94, 87, 88] but not many of arbitrary order.

In this work, we propose a novel approach for the preservation of equilibrium states for the SWE with geometric source terms by using high order augmented solvers. We first develop the 2D extension of the 1D ARL-ADER scheme in [84] for N-dimensional hyperbolic systems of equations with geometric source terms that allows equilibrium scenarios in a Cartesian grid. This is achieved by means of a particular procedure for the integration of the source term that ensures an exact balance between flux fluctuations and source terms

at cell interfaces and inside cells. This procedure does not only preserve equilibrium states with machine precision but also guarantees an arbitrary order of accuracy thanks to the use of Romberg's integration method [75], provided an optimal derivation of time derivatives. Note that the design of the procedure is based on the use of the fluctuation form of the updating scheme [83, 84].

For the particular case of the SWE with bed elevation and Coriolis, we define two directional primitive variables for the x and y contributions of the Coriolis force, as done in [93, 87, 88], which allows to express again the Coriolis source term as a geometric source term. By doing this, both source terms can be merged into a single geometric source where the scalar variable can be regarded as an apparent topography [93, 88]. The numerical techniques designed here for the discretization of the bed slope source term can be extended to the Coriolis source term while retaining the well-balanced property, even in cases with discontinuous bed topography. It is worth pointing out that the numerical scheme is designed to preserve exactly the still water at rest and the geostrophic equilibrium in the Cartesian directions. Contrary to 1D geostrophic equilibrium problems, when considering the 2D geostrophic equilibrium, the exact steady solution cannot be preserved [87]. However, the high order of accuracy provided by the scheme will help to accurately converge to such solution. It is worth pointing out that the preservation of the geostrophic equilibrium is done at the cost of loosing the optimal accuracy of the integration in time, as the CK procedure has to be carried out in a particular way that combines the use of conserved and primitive source term variables.

In this chapter, we also address the discretization of the friction source term. Unlike the bed elevation and Coriolis terms, the friction term is not discretized as a geometric source term. Hence, it is not included in the definition of the DRP and is only accounted for in the centered integral of the source inside the cell. As a consequence, the numerical scheme will not provide an exact equilibrium between friction slope and bed slope, but will ensure convergence to such equilibrium with mesh refinement. In spite of using a centered discretization of the friction source term, the well-balanced property is still satisfied, even when considering bed variation, friction and Coriolis at the same time.

The chapter is structured as follows. In Section 11.1, the extension of the ARL-ADER scheme to 2 space dimensions is detailed. In Section 11.2, the design of a well-balanced scheme for the SWE with bed elevation is described, including the details for the integration of the source term inside cells, based on the combination of Gaussian and Romberg integration methods. In Section 11.3, the same is done for the SWE with bed elevation and Coriolis, as the latter is treated as a geometric source term in order to preserve the geostrophic equilibrium. The numerical discretization of the friction source term is briefly explained in Section 11.4. In Section 11.5, we present a broad variety of test cases where the performance of the scheme is assessed. Such cases comprise theoretical problems such as 2D dam breaks, steady and transient flows passing over obstacles, shock-vortex interaction problems, quiescent and geostrophic equilibrium cases and other more realistic test cases. Among the latter, we include the propagation of Rossby and Kelvin waves in the equatorial line as well as the anticyclonic propagation of an eddy in the northern hemisphere and the simulation of the seiche phenomenon in a channel with lateral cavities, where the numerical results are compared with experimental measurements. Convergence rates are also examined using cases that consider the propagation of smooth perturbations.

11.1 The ARL scheme for the 2D SWE

The approximate Jacobian $\widetilde{\mathbf{J}}$ for the DRP_K in (6.5), defined at the interface $x_{\xi+1/2}$ reads

$$\widetilde{\mathbf{J}}_{\xi+1/2} = \begin{pmatrix} 0 & n_x & n_y \\ \widetilde{c}^2 n_x - \widetilde{u}(\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}}) & \widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}} + \widetilde{u}n_x & \widetilde{u}n_y \\ \widetilde{c}^2 n_y - \widetilde{v}(\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}}) & \widetilde{v}n_x & \widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}} + \widetilde{v}n_y \end{pmatrix}, \tag{11.1}$$

and is constructed using the Roe averages [15].

$$\widetilde{c}_{\xi+1/2} = \sqrt{g\widetilde{h}}, \tag{11.2}$$

$$\widetilde{u}_{\xi+1/2} = \frac{u^{(0)}_{(\xi+1)_L}\sqrt{h^{(0)}_{(\xi+1)_L}} + u^{(0)}_{\xi_R}\sqrt{h^{(0)}_{\xi_R}}}{\sqrt{h^{(0)}_{(\xi+1)_L}} + \sqrt{h^{(0)}_{\xi_R}}}, \quad \widetilde{v}_{\xi+1/2} = \frac{v^{(0)}_{(\xi+1)_L}\sqrt{h^{(0)}_{(\xi+1)_L}} + v^{(0)}_{\xi_R}\sqrt{h^{(0)}_{\xi_R}}}{\sqrt{h^{(0)}_{(\xi+1)_L}} + \sqrt{h^{(0)}_{\xi_R}}} \tag{11.3}$$

with

$$\widetilde{h}_{\xi+1/2} = \frac{h^{(0)}_{\xi_R} + h^{(0)}_{(\xi+1)_L}}{2}, \tag{11.4}$$

and $h^{(0)}_{(\cdot)}$ and $u^{(0)}_{(\cdot)}$ the spatial reconstruction of the water depth and velocity, respectively. The Jacobian in $\widetilde{\mathbf{J}}$ is diagonalized by the following eigenvectors

$$\widetilde{\mathbf{e}}^1_{\xi+1/2} = \begin{pmatrix} 1 \\ \widetilde{u} - \widetilde{c}n_x \\ \widetilde{v} - \widetilde{c}n_y \end{pmatrix}, \quad \widetilde{\mathbf{e}}^2_{\xi+1/2} = \begin{pmatrix} 1 \\ -\widetilde{c}n_y \\ \widetilde{c}n_x \end{pmatrix}, \quad \widetilde{\mathbf{e}}^3_{\xi+1/2} = \begin{pmatrix} 1 \\ \widetilde{u} + \widetilde{c}n_x \\ \widetilde{v} + \widetilde{c}n_y \end{pmatrix}, \tag{11.5}$$

leading to the diagonal matrix $\widetilde{\mathbf{\Lambda}}_{\xi+1/2} = diag(\widetilde{\lambda}^1_{\xi+1/2}, \widetilde{\lambda}^2_{\xi+1/2}, \widetilde{\lambda}^3_{\xi+1/2})$, where

$$\widetilde{\lambda}^1_{\xi+1/2} = (\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}} - \widetilde{c})_{\xi+1/2}, \quad \widetilde{\lambda}^2_{\xi+1/2} = (\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}})_{\xi+1/2}, \quad \widetilde{\lambda}^3_{\xi+1/2} = (\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}} + \widetilde{c})_{\xi+1/2}. \tag{11.6}$$

The wave strengths for the leading term yield

$$\begin{aligned}
\alpha^{(0),1}_{\xi+1/2} &= \left(\frac{\delta h^{(0)}}{2} + \frac{1}{2\widetilde{c}}\left(\delta h^{(0)}(\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}}) - \delta(h\mathbf{v})^{(0)}\cdot\hat{\mathbf{n}}\right)\right)_{\xi+1/2}, \\
\alpha^{(0),2}_{\xi+1/2} &= \frac{1}{\widetilde{c}}\left(\delta h^{(0)}(\widetilde{u}n_y - \widetilde{v}n_x) - \delta(hu)^{(0)}n_y + \delta(hv)^{(0)}n_x\right)_{\xi+1/2}, \\
\alpha^{(0),3}_{\xi+1/2} &= \left(\frac{\delta h^{(0)}}{2} - \frac{1}{2\widetilde{c}}\left(\delta h^{(0)}(\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}}) - \delta(h\mathbf{v})^{(0)}\cdot\hat{\mathbf{n}}\right)\right)_{\xi+1/2},
\end{aligned} \tag{11.7}$$

where $\mathbf{U}^{(0)}_{(\cdot)} = (h^{(0)}, hu^{(0)}, hv^{(0)})^T$ is the vector of reconstructed quantities, provided by the WENO scheme. Identically, we can derive the expression for the wave strengths for the k-th terms

$$\begin{aligned}
\alpha^{(k),1}_{\xi+1/2} &= \left(\frac{\delta h^{(k)}}{2} + \frac{1}{2\widetilde{c}}\left(\delta h^{(k)}(\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}}) - \delta(h\mathbf{v})^{(k)}\cdot\hat{\mathbf{n}}\right)\right)_{\xi+1/2}, \\
\alpha^{(k),2}_{\xi+1/2} &= \frac{1}{\widetilde{c}}\left(\delta h^{(k)}(\widetilde{u}n_y - \widetilde{v}n_x) - \delta(hu)^{(k)}n_y + \delta(hv)^{(k)}n_x\right)_{\xi+1/2}, \\
\alpha^{(k),3}_{\xi+1/2} &= \left(\frac{\delta h^{(k)}}{2} - \frac{1}{2\widetilde{c}}\left(\delta h^{(k)}(\widetilde{\mathbf{v}}\cdot\hat{\mathbf{n}}) - \delta(h\mathbf{v})^{(k)}\cdot\hat{\mathbf{n}}\right)\right)_{\xi+1/2},
\end{aligned} \tag{11.8}$$

where $\mathbf{D}^{(k)}_{(\cdot)} = (h^{(k)}, hu^{(k)}, hv^{(k)})^T$ is the vector of reconstructed time derivatives of the conserved quantities, provided by the CK procedure.

Regarding the source strengths for the leading term, it is worth recalling that only when considering geometric source terms, the source strengths are not nil. Otherwise, the source term is not included in the definition of the DRP. According to (9.1)–(9.3), source terms are only acting on the momentum equations, hence we only consider non-zero source components such equations as follows

$$\bar{\mathbf{S}}^{(0)}_{\xi+1/2} = \begin{pmatrix} 0 \\ \bar{S}^{x,(0)}_{\xi+1/2} \\ \bar{S}^{y,(0)}_{\xi+1/2} \end{pmatrix}, \tag{11.9}$$

with $\bar{\mathbf{S}}^{(0)}_{\xi+1/2}$ defined in (6.11). Then, we can define the following vector $\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2} \in \mathbb{R}^2$, where ^{M} stands for *momentum*, as

$$\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2} = \begin{pmatrix} \bar{S}^{x,(0)}_{\xi+1/2} \\ \bar{S}^{y,(0)}_{\xi+1/2} \end{pmatrix}, \tag{11.10}$$

and use it for the definition of the source strengths as

$$\beta^{(0),1}_{\xi+1/2} = -\frac{1}{2\widetilde{c}}\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}}, \quad \beta^{(0),2}_{\xi+1/2} = -\frac{1}{\widetilde{c}}\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}}_\perp, \quad \beta^{(0),3}_{\xi+1/2} = -\beta^{(0),1}_{\xi+1/2} \tag{11.11}$$

where $\hat{\mathbf{n}}_\perp = \mathbf{R}_{\pi/2}\hat{\mathbf{n}}$ is the unitary vector parallel to the cell interface and $\mathbf{R}_{\pi/2}$ a $\pi/2$ rad rotation matrix. The expression of the source strengths for the higher order terms can be derived analogously.

If considering the geometric source term in (3.15), the projection $\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}}$ must be an approximation of

$$\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}} \approx \frac{1}{\Delta t}\int_0^{\Delta t}\int_{\check{x}^-_{i+1/2}}^{\check{x}^+_{i+1/2}} S_s(\mathbf{U})\nabla\phi\cdot\hat{\mathbf{n}}d\check{x}\,, \tag{11.12}$$

where, for instance, $S_s(\mathbf{U}) = -gh$ and $\phi = z$ in the case of only considering the bed slope source term as geometric source term. Using the relations $dx = d\check{x}n_x$ and $dy = d\check{x}n_y$, (11.12) can be approached by

$$\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}} = S_s(\bar{\mathbf{U}}^{(0)}_{\xi+1/2})\int_{\mathbf{x}^-_{i+1/2}}^{\mathbf{x}^+_{\xi+1/2}}\nabla\phi\, d\mathbf{x} = S_s(\bar{\mathbf{U}}^{(0)}_{\xi+1/2})\delta(\phi)^{(0)}_{\xi+1/2}\,, \tag{11.13}$$

On the other hand, the projection $\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}}_\perp$ must be an approximation of

$$\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}}_\perp \approx \frac{1}{\Delta t}\int_0^{\Delta t}\int_{\check{x}^-_{\xi+1/2}}^{\check{x}^+_{\xi+1/2}} S_s(\mathbf{U})\nabla\phi\cdot\hat{\mathbf{n}}_\perp d\check{x}\,, \tag{11.14}$$

where $\nabla\phi\cdot\hat{\mathbf{n}}_\perp$ is the directional derivative of ϕ in the direction parallel to the cell interface. According to the definition of the DRP$_K$, we only consider variations of the variables in the normal direction to the cell interface, hence $\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}}_\perp = 0$ and $\beta^{(0),2}_{\xi+1/2} = 0$.

In the following subsections, a detailed derivation of the approximation of the source term for the SWE with bottom elevation and Coriolis, to construct a well-balanced scheme in Cartesian grids, is presented.

11.2 Resolution of the SWE with bed elevation

For now, we will consider the SWE in (9.1)–(9.3) with $\mathbf{S}_c = \mathbf{S}_f = 0$, that is, with no friction and in a fixed frame. The well-balanced formulation can be regarded as a weaker exact conservation property than the EB formulation. The latter ensures the exact conservation property for both quiescent and moving equilibrium cases by considering the mechanical energy, E, whereas the former only satisfies the aforementioned property when $\mathbf{v} = 0$, that is, still water at rest.

Under steady conditions, when the velocity vanishes, $\mathbf{v} = 0$, the equation for the conservation of energy

and momentum yield the same result

$$\nabla (h+z) = 0\,, \tag{11.15}$$

which is known in the literature as *lake at rest* condition. At the discrete level and considering a Cartesian grid, between two points $\mathbf{x}_1$ and $\mathbf{x}_2$, Equation (11.15) can be decomposed into the Cartesian directions as

$$\delta (h+z)_{x_2,x_1} = 0\,, \qquad \delta (h+z)_{y_2,y_1} = 0\,. \tag{11.16}$$

To construct a well-balanced scheme, the previous discrete conditions in (11.16) must be satisfied. This can only be achieved if the WENO reconstruction method is applied to $\eta = h+z$ and z first, and h is computed from the difference of these reconstructions as

$$h^{(0)}_{(\cdot)} = \eta^{(0)}_{(\cdot)} - z^{(0)}_{(\cdot)}\,, \tag{11.17}$$

where $\eta^{(0)}_{(\cdot)}$ and $z^{(0)}_{(\cdot)}$ are the reconstructed water surface elevation and bottom elevation and $h^{(0)}_{(\cdot)}$ the computed water depth. Otherwise, the exact conservation cannot be ensured. A complete relation of the reconstructed variables and reconstruction procedures is presented in Table 11.1. It is worth pointing out that when using this reconstruction procedure, $\eta^{(0)}_{(\cdot)} = constant$ for still water at rest.

Variable	Reconstruction method	Departing data	Notation
$h+z$	WENO rec.	Cell averages	$\eta^{(0)}_{(\cdot)}$
z	WENO rec.	Cell averages	$z^{(0)}_{(\cdot)}$
hu	WENO rec.	Cell averages	$hu^{(0)}_{(\cdot)}$
hv	WENO rec.	Cell averages	$hu^{(0)}_{(\cdot)}$
h	$\eta^{(0)}_{(\cdot)} - z^{(0)}_{(\cdot)}$	Point data	$h^{(0)}_{(\cdot)}$

Table 11.1: Data reconstruction technique to construct a well-balanced 2D scheme.

The design of a well-balanced scheme is next detailed. Let us consider again the updating scheme in (7.8). To design a well-balanced scheme, the fluctuation form of the scheme in (7.8) is more suitable, as the fluctuation terms stand for the discrete variations of sources and fluxes between different positions along the grid. This formulation reads

$$\mathbf{U}^{n+1}_{ij} = \mathbf{U}^{n}_{ij} - \frac{\Delta t}{\Delta x^2}\left(\sum_{r=1}^{4} \delta \mathbf{M}^{-}_{r} + \delta \mathbf{M}_{ij}\right), \tag{11.18}$$

where $\delta \mathbf{M}^{-}_{r}$ are the contribution of the incoming waves at cell interfaces and $\delta \mathbf{M}_{ij}$ is the centered fluctuation, that accounts for the variation of physical fluxes and source terms inside the cell. To construct a well-balanced scheme, it is required that all fluctuations become nil under quiescent conditions, that is $\delta \mathbf{M}^{-}_{r} = \delta \mathbf{M}_{ij} = 0$.

Centered fluctuations read

$$\delta \mathbf{M}_{ij} = \sum_{r=1}^{4} \frac{\Delta x}{2} \sum_{q=1}^{k} w_q \mathscr{F}_{r,q} - \bar{\mathbf{S}}_{ij}\,, \tag{11.19}$$

where

$$\bar{\mathbf{S}}_{ij} \approx \begin{pmatrix} 0 \\ \dfrac{1}{\Delta t}\displaystyle\int_0^{\Delta t}\int_{x_{i+1/2}}^{x_{i-1/2}}\int_{y_{i+1/2}}^{y_{i-1/2}} S^x \, dydx \, d\tau \\ \dfrac{1}{\Delta t}\displaystyle\int_0^{\Delta t}\int_{x_{i+1/2}}^{x_{i-1/2}}\int_{y_{i+1/2}}^{y_{i-1/2}} S^y \, dydx \, d\tau \end{pmatrix} \tag{11.20}$$

is the approximation of the integral of the source term inside the cell. In this case, a suitable approximation of the integral in (11.20) is required to ensure $\delta\mathbf{M}_{ij} = 0$.

On the other hand, upwind fluctuations are given by

$$\delta\mathbf{M}_r^- = \sum_{r=1}^{4} \frac{\Delta x}{2} \sum_{q=1}^{k} w_q \left(\mathbf{F}_{r,q}^- - \mathscr{F}_{r,q} \right), \tag{11.21}$$

where $\mathscr{F}_{r,q} = \mathscr{F}(\mathbf{U}_{r,q}) \cdot \hat{\mathbf{n}}$. For (11.21), the equilibrium condition is $\mathbf{F}_{r,q}^- = \mathscr{F}_{r,q}$. Note that q stands for the quadrature point, which will be hereafter denoted by α in the y direction and by β in the x direction.

Let us consider first the equilibrium condition for centered fluctuations, $\delta\mathbf{M}_{ij} = 0$. First of all, we propose to decompose (11.20) in each coordinate direction by performing a Gaussian integration along the transverse direction with respect to the direction of the variation of the geometric term while seeking a suitable discretization of the source term for the integral along the former direction. This means

$$\bar{\mathbf{S}}_{ij} = \begin{pmatrix} 0 \\ \dfrac{\Delta x}{2} \displaystyle\sum_{\alpha=1}^{k} w_\alpha \bar{S}^x_{ij,\alpha} \, dx \\ \dfrac{\Delta x}{2} \displaystyle\sum_{\beta=1}^{k} w_\beta \bar{S}^y_{ij,\beta} \, dy \end{pmatrix}, \tag{11.22}$$

where

$$\begin{aligned} \bar{S}^x_{ij,\alpha} &\approx \frac{1}{\Delta t}\int_0^{\Delta t}\int_{x_{i-1/2}}^{x_{i+1/2}} S^x(x, y_\alpha, \tau) \, dx \, d\tau \\ \bar{S}^y_{ij,\beta} &\approx \frac{1}{\Delta t}\int_0^{\Delta t}\int_{y_{i-1/2}}^{y_{i+1/2}} S^y(x_\beta, y, \tau) \, dy \, d\tau \end{aligned} \tag{11.23}$$

are the sought approximations. Equation (11.19) can also be decomposed in each of the Cartesian directions as follows

$$\begin{aligned} \delta\mathbf{M}_{ij} = \; & \tfrac{\Delta x}{2} \textstyle\sum_{\alpha=1}^{k} w_\alpha \left(\mathscr{F}_{2,\alpha} + \mathscr{F}_{4,\alpha} - \begin{pmatrix} 0 \\ \bar{S}^x_{ij,\alpha} \\ 0 \end{pmatrix} \right) + \\ & \tfrac{\Delta x}{2} \textstyle\sum_{\beta=1}^{k} w_\beta \left(\mathscr{F}_{3,\beta} + \mathscr{F}_{1,\beta} - \begin{pmatrix} 0 \\ 0 \\ \bar{S}^y_{ij,\beta} \end{pmatrix} \right) \end{aligned} \tag{11.24}$$

where the source term has been decomposed in each of the coordinate directions. It is worth recalling that the physical flux and source term are constructed as a power series expansion in time as follows

$$\mathscr{F}_{ij}(x,y,\tau) = \mathscr{F}_{ij}(x,y,0) + \sum_{k=1}^{K} \left[\frac{\partial^k \mathscr{F}_{ij}}{\partial t^k} \right]_{x,y,t=0} \frac{\tau^k}{k!}, \tag{11.25}$$

$$\mathbf{S}_{ij}(x,y,\tau) = \mathbf{S}_{ij}(x,y,0) + \sum_{k=1}^{K} \left[\frac{\partial^k \mathbf{S}_{ij}}{\partial t^k} \right]_{x,y,t=0} \frac{\tau^k}{k!}, \tag{11.26}$$

and we can rewrite the 1D discretizations of the source term as

$$\begin{aligned} \bar{S}^x_{ij,\alpha} &\approx \int_{x_{i-1/2}}^{x_{i+1/2}} \left(S^x(x,y_\alpha,\tau) + \sum_{k=1}^{K} \left[\frac{\partial^k S^x}{\partial t^k} \right]_{x,y_\alpha,t=0} \frac{\Delta t^k}{(k+1)!} \right) dx \ , \\ \bar{S}^y_{ij,\beta} &\approx \int_{y_{i-1/2}}^{y_{i+1/2}} \left(S^y(x_\beta,y,\tau) + \sum_{k=1}^{K} \left[\frac{\partial^k S^y}{\partial t^k} \right]_{x_\beta,y,t=0} \frac{\Delta t^k}{(k+1)!} \right) dy \ , \end{aligned} \tag{11.27}$$

or in its compact form

$$\begin{aligned} \bar{S}^x_{ij,\alpha} &= \bar{S}^{x,(0)}_{ij,\alpha} + \sum_{k=1}^{K} \bar{S}^{x,(k)}_{ij,\alpha} \frac{\Delta t^k}{(k+1)!} \, dx \ , \\ \bar{S}^y_{ij,\beta} &= \bar{S}^{y,(0)}_{ij,\beta} + \sum_{k=1}^{K} \bar{S}^{y,(k)}_{ij,\beta} \frac{\Delta t^k}{(k+1)!} \, dy \ , \end{aligned} \tag{11.28}$$

To derive the well-balanced formulation, steady conditions are considered. Hence, for the derivation of the suitable approximation of the 1D integrals of the source term in (11.28), we only consider the leading terms of the equations, as time derivatives vanish in the steady state. Equation (11.24) is rewritten for the leading term only

$$\begin{aligned} \delta \mathbf{M}^{(0)}_{ij} = \ & \tfrac{\Delta x}{2} \textstyle\sum_{\alpha=1}^{k} w_\alpha \left(\mathscr{F}^{(0)}_{2,\alpha} + \mathscr{F}^{(0)}_{4,\alpha} - \begin{pmatrix} 0 \\ \bar{S}^{x,(0)}_{ij,\alpha} \\ 0 \end{pmatrix} \right) + \\ & \tfrac{\Delta x}{2} \textstyle\sum_{\beta=1}^{k} w_\beta \left(\mathscr{F}^{(0)}_{3,\beta} + \mathscr{F}^{(0)}_{1,\beta} - \begin{pmatrix} 0 \\ 0 \\ \bar{S}^{y,(0)}_{ij,\beta} \end{pmatrix} \right) \end{aligned} \tag{11.29}$$

from which we notice the one-dimensional conditions

$$\mathbf{F}^{(0)}_{i_R,\alpha} - \mathbf{F}^{(0)}_{i_L,\alpha} - \begin{pmatrix} 0 \\ \bar{S}^{x,(0)}_{ij,\alpha} \\ 0 \end{pmatrix} = 0 \,, \qquad \forall \alpha = 1,...,k \,, \tag{11.30}$$

$$\mathbf{G}^{(0)}_{j_R,\beta} - \mathbf{G}^{(0)}_{j_L,\beta} - \begin{pmatrix} 0 \\ 0 \\ \bar{S}^{y,(0)}_{ij,\beta} \end{pmatrix} = 0 \,, \qquad \forall \beta = 1,...,k \,. \tag{11.31}$$

It is worth mentioning that in Equations (11.30) and (11.30) and in what follows, the 1D notation described in Figure 11.1 is used.

If we consider the condition in the x-direction (11.30) (the condition in y is the same due to the rotational invariance), we can rewrite it as

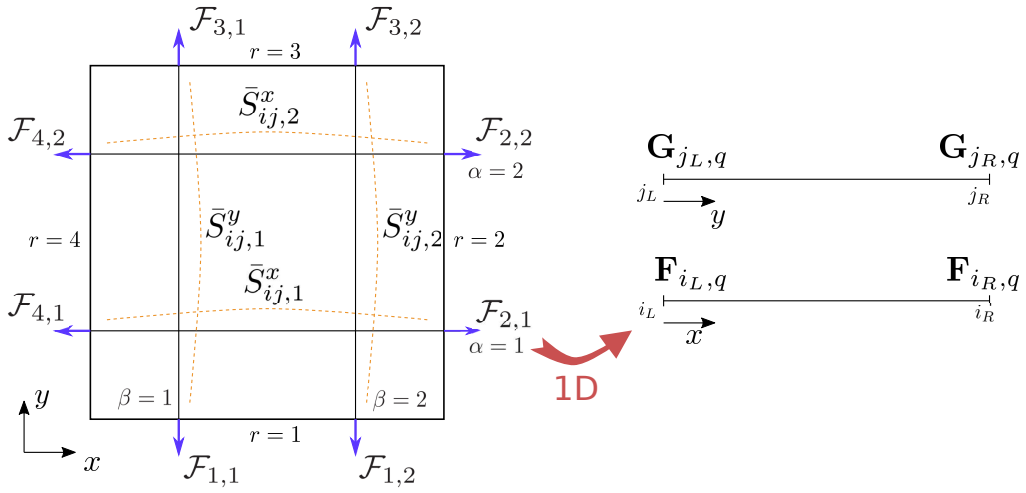


Figure 11.1: Sub-cell integration lines in the Cartesian directions and 1D analogy.

$$\mathbf{F}^{(0)}_{i_R,\alpha} - \mathbf{F}^{(0)}_{i_L,\alpha} - \bar{\mathbf{S}}^{x,(0)}_{ij,\alpha} = 0 \qquad \forall \alpha = 1,...,k \tag{11.32}$$

and under quiescent steady conditions, we have that $\mathbf{F}(\mathbf{U}) = \frac{1}{2}gh^2$ yielding

$$(g\bar{h}\delta h)^{(0)}_{i_R,i_L} - \bar{S}^{x,(0)}_{ij,\alpha} = 0 \qquad \forall \alpha = 1,...,k\,, \tag{11.33}$$

where $\bar{S}^{x,(0)}_{ij,\alpha}$ is the approximation of

$$\bar{S}^{x,(0)}_{ij,\alpha} \approx -\int_{x_{i_L}}^{x_{i_R}} gh\frac{\partial z}{\partial x}dx\,. \tag{11.34}$$

If approaching the previous integral as

$$\bar{S}^{x,(0)}_{ij,\alpha} = -g\bar{h}\delta z^{(0)}_{i_L,i_R}\,, \tag{11.35}$$

under steady state we have $g\bar{h}\delta(h+z)^{(0)}_{i_L,i_R} = 0$, where one notices that $(h+z)^{(0)}_{(\cdot)} = \eta^{(0)}_{(\cdot)}$ is the equilibrium reconstruction variable and therefore $g\bar{h}\delta\eta^{(0)}_{i_L,i_R} = 0$. Note again that it is a must to reconstruct over the quantity $h+z$ and z and then compute h from the reconstructions.

The discretization of the source term used above is only 2-nd order accurate in space. To obtain a $K+1$-th order scheme, it is neccessary to extend this integration technique to arbitrary order in space. To this end, we can use Romberg integration, which is a result that can be obtained from Richardson's extrapolation. An arbitrary order integral of the source term is denoted as $\left\{\bar{S}^{x,(0)}_{ij,\alpha}\right\}^n_m$, where m is the number of subdivisions of the initial interval $\Upsilon = [x_{i-1/2}, x_{i+1/2}]$, with step size $h_m = \frac{\Delta x}{2^m}$ and n the number of Romberg iterations, having a magnitude of the residual of $\mathcal{O}(\Delta x^{2(n+1)})$. To construct a $K+1$-th order ADER scheme, both n and m will take values up to $\left\lceil \frac{K-1}{2} \right\rceil$ and the order of accuracy of the method will be either $\mathcal{O}(\Delta x^{K+2})$ if K is even or $\mathcal{O}(\Delta x^{K+1})$ if K is odd. The expression for $\left\{\bar{S}^{x,(0)}_{ij}\right\}^n_m$ is computed recursively departing from the trapezoid integrals, that is $n=0$ and $m=0,...,\left\lceil \frac{K-1}{2} \right\rceil$, and computing the following levels $n=1,...,\left\lceil \frac{K-1}{2} \right\rceil$ as

$$\left\{\bar{S}^{x,(0)}_{ij,\alpha}\right\}^{n+1}_{m+1} = \frac{4^n\left\{\bar{S}^{x,(0)}_{ij,\alpha}\right\}^n_{m+1} - \left\{\bar{S}^{x,(0)}_{ij,\alpha}\right\}^n_m}{4^n - 1}\,. \tag{11.36}$$

For $n=0$, the integrals are given by $\left\{\bar{S}^{x,(0)}_{ij,\alpha}\right\}^0_0 = (-g\bar{h}\delta z)^{(0)}_{i_R,i_L}$ and $\left\{\bar{S}^{x,(0)}_{ij,\alpha}\right\}^0_1 = (-g\bar{h}\delta z)^{(0)}_{i_R,i} + (-g\bar{h}\delta z)^{(0)}_{i,i_L}$, for $m=0$ and $m=1$ respectively.

Concerning the derivative terms, there is no need of a particular discretization technique of the source term to ensure the well-balanced property as time derivatives vanish under steady state. Here, we use a 2D Gaussian integration

$$\bar{S}^{x,(k)}_{ij} = \sum_{\alpha=1}^{k} w_\alpha \sum_{\beta=1}^{k} w_\beta \left(-gh^{(k)}\partial_x z\right)_{\alpha,\beta}\,, \tag{11.37}$$

$$\bar{S}^{y,(k)}_{ij} = \sum_{\alpha=1}^{k} w_\alpha \sum_{\beta=1}^{k} w_\beta \left(-gh^{(k)}\partial_y z\right)_{\alpha,\beta}\,, \tag{11.38}$$

where $h^{(k)}_{\alpha,\beta}$ is the k-th time derivative of h at the quadrature point. The complete integral of the source term will be computed as

$$\bar{\mathbf{S}}_{ij} = \frac{\Delta x}{2}\begin{pmatrix} 0 \\ \displaystyle\sum_{\alpha=1}^{k} w_\alpha \left\{\bar{S}^{x,(0)}_{ij,\alpha}\right\}^n_m + \sum_{k=1}^{K} \bar{S}^{x,(k)}_{ij}\frac{\Delta t^k}{(k+1)!} \\ \displaystyle\sum_{\beta=1}^{k} w_\beta \left\{\bar{S}^{y,(0)}_{ij,\beta}\right\}^n_m + \sum_{k=1}^{K} \bar{S}^{y,(k)}_{ij}\frac{\Delta t^k}{(k+1)!} \end{pmatrix}. \tag{11.39}$$

It is worth pointing out that under steady conditions, $h+z$ is constant and it is straightforward to derive that $\left\{\bar{S}^{x,(0)}_{ij}\right\}^n_m = \left\{\bar{S}^{x,(0)}_{ij}\right\}^0_0$, for any m and n, therefore the well-balanced condition is always satisfied.

On the other hand, when considering the upwind fluctuations in (11.21), it is straightforward to prove that the following approximation for the leading term

$$\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}} = \left(-g\bar{h}\delta z\right)^{(0)}_{\xi+1/2}\,, \tag{11.40}$$

satisfies the steady state equilibrium condition $\mathbf{F}^-_{r,q} = \mathscr{F}_{r,q}$, as

$$\left(\widetilde{\lambda}^- \alpha^{(0)} - \beta^{-,(0)}\right)^m_{\xi+1/2} = 0\,, \tag{11.41}$$

in Equation (6.7). Higher order terms are computed as

$$\bar{\mathbf{S}}^{M,(k)}_{\xi+1/2}\cdot\hat{\mathbf{n}} = \left(-g\bar{h}^{(k)}\delta z^{(0)}\right)_{\xi+1/2}\,. \tag{11.42}$$

As mentioned above, according to the definition of the DRP_K, gradients in the direction parallel to cell interfaces are considered nil, hence

$$\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2}\cdot\hat{\mathbf{n}}_\perp = \bar{\mathbf{S}}^{M,(k)}_{\xi+1/2}\cdot\hat{\mathbf{n}}_\perp = 0\,. \tag{11.43}$$

11.3 Resolution of the SWE with bed elevation in the rotating frame

We now consider the SWE with bed elevation in a rotating frame by means of including both the bed slope and the Coriolis source terms, $\mathbf{S}_b$ and $\mathbf{S}_c$ respectively. Equations (9.1)–(9.3) represent a good model for large scale phenomena in geophysical flows, in which oceanic and atmospheric circulations are often perturbations of the so-called geostrophic equilibrium [88].

When designing a numerical scheme for the resolution of a particular system of equations, it is of

importance to design the scheme in such a way that allows the preservation of the steady-state equilibrium solutions, since many phenomena of interest are often perturbations of those equilibrium states. It is worth recalling that when the Coriolis effect is neglected, the SWE in (9.1)–(9.3) satisfies the quiescent equilibrium steady state in (11.15). Numerical schemes satisfying (11.15) in the discrete form are called well-balanced schemes.

When considering the Coriolis source term, equilibrium states become more complex as they now include the circulation of the flow in particular directions. For the system in (9.1)–(9.3), the most relevant equilibrium solution to be considered in the design of the numerical scheme is the so-called geostrophic equilibrium state, which arises from the balance of the Coriolis force with the hydrostatic pressure change due to the surface elevation gradient. This steady state is often referred to as *jet in the rotating frame* [88]. According to [88], the geostrophic equilibrium satisfies

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0\,,\quad g\frac{\partial}{\partial x}(h+z)=fv\,,\quad g\frac{\partial}{\partial y}(h+z)=-fu\,, \tag{11.44}$$

which can be rewritten as

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0\,,\quad g\frac{\partial}{\partial x}(h+z-V)=0\,,\quad g\frac{\partial}{\partial y}(h+z+U)=0\,, \tag{11.45}$$

with

$$\frac{\partial V}{\partial x}=\frac{fv}{g}\,,\quad \frac{\partial U}{\partial y}=\frac{fu}{g}\,, \tag{11.46}$$

the primitive functions of the Coriolis force, also called *apparent topography* [90]. For the sake of simplicity, we will define the potentials

$$L=h+z-V\,,\quad K=h+z+U\,, \tag{11.47}$$

which are functionals that are conserved in the geostrophic equilibrium. We can identify two particular jets in the rotating frame [88], which satisfy

$$u=0\quad \frac{\partial v}{\partial y}=0\,,\quad \frac{\partial h}{\partial y}=0\,,\quad \frac{\partial z}{\partial y}=0\,,\quad L\equiv constant\,, \tag{11.48}$$

and

$$v=0\quad \frac{\partial u}{\partial x}=0\,,\quad \frac{\partial h}{\partial x}=0\,,\quad \frac{\partial z}{\partial x}=0\,,\quad K\equiv constant\,. \tag{11.49}$$

In this section, a well-balanced WENO-ADER scheme using the Augmented Roe solver, which preserves the jets in (11.48) and (11.49), is proposed. The keystone of this scheme is the treatment of the Coriolis source terms as geometric sources in order to discretize them in the same way than the bed elevation source term. As done for the SWE with variable bed elevation and no Corolis terms, it is necessary to identify first which quantities the reconstruction procedure will be applied to and which other quantities will be computed from the reconstructed data in order to satisfy the discrete equilibrium. When Coriolis was not considered, the reconstruction technique was applied to hu, hv, z and $h+z$ (the equilibrium variable) and then, the water depth h was computed as detailed in Table 11.1. When Coriolis forces are present, the equilibrium variable is not anymore $h+z$ and instead, the exact conservation must be ensured for both K and L. In order to satisfy this, the WENO reconstruction will be carried out for hu, hv, h, z, L and K first, and then, V and U will be computed from the reconstructed data as

$$V_{(\cdot)}^{(0)}=h_{(\cdot)}^{(0)}+z_{(\cdot)}^{(0)}-L_{(\cdot)}^{(0)}\,,\qquad U_{(\cdot)}^{(0)}=K_{(\cdot)}^{(0)}-h_{(\cdot)}^{(0)}-z_{(\cdot)}^{(0)}\,, \tag{11.50}$$

where $h^{(0)}_{(\cdot)}$, $z^{(0)}_{(\cdot)}$, $L^{(0)}_{(\cdot)}$ and $K^{(0)}_{(\cdot)}$ are the reconstructed water depth, bottom elevation, L potential and K potential, and $V^{(0)}_{(\cdot)}$ and $U^{(0)}_{(\cdot)}$ the computed Coriolis primitive variables.

When using WENO-ADER schemes, the problem data is discretized in the form of cell averages, which are required in the first step of the scheme to carry out the WENO reconstructions. The discretization of h, z, hu and hv as cell averages is straightforward, however, for V and U is not that simple and has to be done in a fancier way. According to the definitions of V and U, we can express them as the following integrals

$$V(x,y) = \int_0^x \frac{fv(\chi,y)}{g} d\chi + V(0), \quad U(x,y) = \int_0^y \frac{fu(x,\chi)}{g} d\chi + U(0), \tag{11.51}$$

where $V(0) = U(0) = 0$. When considering piecewise constant data in a Cartesian grid with cell size Δx, we can compute V and U at cell interfaces as

$$V_{i+1/2,j} = \sum_{t=0}^{i} \left(\frac{fv}{g}\right)_{t,j} \Delta x, \quad U_{i,j+1/2} = \sum_{t=0}^{j} \left(\frac{fu}{g}\right)_{i,t} \Delta x \tag{11.52}$$

and then calculate the cell averages as

$$V_{ij} = \frac{1}{2}\left(V_{i+1/2,j} + V_{i-1/2,j}\right), \quad U_{ij} = \frac{1}{2}\left(U_{i,j+1/2} + U_{i,j-1/2}\right). \tag{11.53}$$

After computing the cell averages for V and U, those for K and L can be computed and the reconstruction procedure for all variables can be carried out.

Concerning the integration of the source terms, $\mathbf{S}_b$ and $\mathbf{S}_c$, to satisfy the well-balance property, we must follow the same approach than in the previous section for the SWE with variable bed. The numerical scheme must be written in fluctuation form (11.18) and a particular discretization of the source term must be sought in such a way that all fluctuations, those at cell interfaces and those inside the cell, must become nil under geostrophic conditions, satisfying (11.48) and (11.49). As done in the previous section, the scheme is designed to satisfy the discrete equilibrium in the Cartesian directions and therefore the source term integration is reduced to a one-dimensional formulation. For the x-geostrophic balance in (11.48), the 0-th discretization of the source term reads

$$\bar{S}^{x,(0)}_{ij,\alpha} = -(g\bar{h}\delta z)^{(0)}_{i_L,i_R} + (g\bar{h}\delta V)^{(0)}_{i_L,i_R}, \tag{11.54}$$

yielding

$$g\bar{h}\delta(h+z-V)^{(0)}_{i_L,i_R} = 0, \tag{11.55}$$

where one notices that $(h+z-V)^{(0)}_{(\cdot)} = L^{(0)}_{(\cdot)}$ is the equilibrium reconstruction variable and therefore

$$g\bar{h}\delta L^{(0)}_{i_L,i_R} = 0 \tag{11.56}$$

is always satisfied. To extend the 2-nd order integral in (11.54) to higher order of accuracy, we use the Romberg integration method detailed in the previous section. Concerning the derivative terms, there is no need of a particular discretization technique of the source term to ensure the well-balanced property, as outlined in the previous section. Here, we use a 2D Gaussian integration

$$\bar{S}^{x,(k)}_{ij} = \sum_{\alpha=1}^{k} w_\alpha \sum_{\beta=1}^{k} w_\beta \left(-gh^{(k)}\partial_x z + f(hv)^{(k)}\right)_{\alpha,\beta}, \tag{11.57}$$

$$\bar{S}_{ij}^{y,(k)} = \sum_{\alpha=1}^{k} w_\alpha \sum_{\beta=1}^{k} w_\beta \left(-gh^{(k)}\partial_y z - f(hu)^{(k)}\right)_{\alpha,\beta}, \tag{11.58}$$

where $h_{\alpha,\beta}^{(k)}$ and $hu_{\alpha,\beta}^{(k)}$ are the k-th time derivative of h and hu at the quadrature point.

This is because all derivatives vanish under steady regime, such as the geostrophic equilibrium or the lake-at-rest equilibrium. However, this is only achieved when using a particular expression of the CK procedure for the computation of time derivatives in terms of space derivatives. The keystone to ensure that all time derivatives are nil in the geostrophic equilibrium is to use spatial derivatives of V and U when possible in the Coriolis source term, instead of directly computing the derivatives of fhv and fhu.

By means of the CK procedure, we can express the first derivatives of the conserved quantities as

$$\partial_t h = -\left(\partial_x hu + \partial_y hv\right), \tag{11.59}$$

$$\partial_t (hu) = -\partial_x \left(hu^2 + \frac{1}{2}gh^2\right) - \partial_y (huv) - gh\partial_x z + gh\partial_x V, \tag{11.60}$$

$$\partial_t (hv) = -\partial_y \left(hv^2 + \frac{1}{2}gh^2\right) - \partial_x (huv) - gh\partial_y z - gh\partial_y U, \tag{11.61}$$

and it is straightforward to notice that under the x-geostrophic equilibrium in (11.48), they yield

$$\partial_t h = 0, \tag{11.62}$$

$$\partial_t (hu) = -\partial_x \left(\frac{1}{2}gh^2\right) - gh\partial_x z + gh\partial_x V, \tag{11.63}$$

$$\partial_t (hv) = 0. \tag{11.64}$$

If substituting $L = h + z - V$ in Equation (11.63), it yields

$$\partial_t (hu) = -gh\partial_x L = 0, \tag{11.65}$$

where

$$\partial_x L = \partial_x h + \partial_x z - \partial_x V = 0, \tag{11.66}$$

hence the equilibrium is satisfied for all first derivatives. The same can be done recursively for higher order derivatives up to the desired accuracy.

Let us consider now the upwind fluctuations in (11.21). For the particular case of Cartesian grid, we can express the approximation of the projection of the integral of the source term onto the cell normal as follows

$$\bar{\mathbf{S}}_{\xi+1/2}^{M,(0)} \cdot \hat{\mathbf{n}} = \left(-g\bar{h}\delta z\right)_{\xi+1/2}^{(0)} + \alpha(g\bar{h}\delta V)_{i_L,i_R}^{(0)} - \beta(g\bar{h}\delta V)_{i_L,i_R}^{(0)}, \tag{11.67}$$

where $\alpha = abs(n_x)$ and $\beta = abs(n_y)$, satisfying the steady state equilibrium condition. Higher order terms are computed as

$$\bar{\mathbf{S}}_{\xi+1/2}^{M,(k)} \cdot \hat{\mathbf{n}} = \left(-g\bar{h}^{(k)}\delta z^{(0)}\right)_{\xi+1/2}, \tag{11.68}$$

where the contribution of the derivatives of the Coriolis term have been neglected, due to its non-geometric nature. According to (11.43), $\bar{\mathbf{S}}^{M,(0)}_{\xi+1/2} \cdot \hat{\mathbf{n}}_{\perp} = 0$.

11.4 Resolution of the SWE with friction

As outlined in the introduction of the chapter, the friction term is discretized here as a centered source term, which means that it is not accounted for in the definition of the DRP. The approach taken here does not ensure an exact equilibrium between bed slope and friction slope but ensures convergence with arbitrary order to this equilibrium state. The following 2D Gaussian quadrature is proposed to approximate the integral of the leading term inside the cell

$$\bar{S}^{x,(0)}_{ij} = \sum_{\alpha=1}^{k} w_\alpha \sum_{\beta=1}^{k} w_\beta \left(-c_f \left|\mathbf{v}^{(0)}\right| u^{(0)}\right)_{\alpha,\beta} , \tag{11.69}$$

$$\bar{S}^{y,(0)}_{ij} = \sum_{\alpha=1}^{k} w_\alpha \sum_{\beta=1}^{k} w_\beta \left(-c_f \left|\mathbf{v}^{(0)}\right| v^{(0)}\right)_{\alpha,\beta} , \tag{11.70}$$

with $c_f = c_f(\mathbf{U}^{(0)}, n)$ and n the Manning coefficient. To construct a Gaussian quadrature for the derivative terms of the source term, the CK procedure must be used first to provide an approximation of the time derivatives of the source at the quadrature points

$$S^{x,(k)}_{\alpha,\beta} = \frac{\partial^k}{\partial t^k}\left(-c_f \left|\mathbf{v}\right| u\right)_{\alpha,\beta} \tag{11.71}$$

$$S^{y,(k)}_{\alpha,\beta} = \frac{\partial^k}{\partial t^k}\left(-c_f \left|\mathbf{v}\right| v\right)_{\alpha,\beta} \tag{11.72}$$

Then, we can construct the 2D Gaussian quadrature as follows

$$\bar{S}^{x,(k)}_{ij} = \sum_{\alpha=1}^{k} w_\alpha \sum_{\beta=1}^{k} w_\beta S^{x,(k)}_{\alpha,\beta} , \tag{11.73}$$

$$\bar{S}^{y,(k)}_{ij} = \sum_{\alpha=1}^{k} w_\alpha \sum_{\beta=1}^{k} w_\beta S^{x,(k)}_{\alpha,\beta} . \tag{11.74}$$

Bed elevation, friction and Coriolis source terms can be combined and considered at the same time. However, for each combination of them, a particular CK procedure must be derived as the equations are changed.

11.5 Numerical results

11.5.1 Well-balanced property assessment

Quiescent equilibrium with bed variation

This test case consist of a 2D water surface at rest over a irregular bed. The computational domain is $\Omega = [0, 100] \times [0, 100]$ and the initial condition is given by

$$h(x, y) = 2 , \quad u(x, y) = v(x, y) = 0 , \tag{11.75}$$

	1-st order	3-rd order
t(s)	L_∞ error	L_∞ error
5	2.22044E-016	6.66133E-016
10	2.22044E-016	8.88178E-016
500	2.22044E-016	3.77475E-015

Table 11.2: Section 11.5.1. Numerical errors for h provided by the 3-rd order ARL-ADER scheme, measured with L_∞ error norm at $t = 5$ and $t = 10$ s. Double precision is used.

	1-st order	3-rd order
t(s)	L_∞ error	L_∞ error
5	4.44088E-016	4.44089E-016
10	4.44088E-016	4.44089E-016
500	4.44088E-016	4.44089E-016

Table 11.3: Section 11.5.1. Numerical errors for K provided by the 1-st and 3-rd order ARL-ADER scheme, measured with L_∞ error norm at $t = 5$ and $t = 10$ s. Double precision is used.

with a bottom elevation

$$z(x,y) = \exp\left(-\frac{(x-50)^2+(y-50)^2}{80}\right), \qquad \forall (x,y) \in \Omega \tag{11.76}$$

The solution is computed at $t = 5$ and $t = 10$ s using a 1-st and 3-rd ARL-ADER scheme. The numerical error for h is measured using L_∞ error norm and presented in Table 11.2. It is observed that the scheme preserves the discrete quiescent equilibrium with machine precision (double precision is used).

Geostrophic equilibrium with bed variation

This test case consist of a flow in the y-direction that is initially at geostrophic equilibrium [88]. The computational domain is $y \in [-5,5]$ and the initial condition is given by

$$K(x,y) = 2\,, \quad v(x,y) = \frac{2g}{f} y \exp(-y^2)\,, \quad u(x,y) = 0\,, \tag{11.77}$$

where $f = 1$ s^{-1} and $g = 9.8$ ms^{-2} and the bottom elevation is given by

$$z(x,y) = 0.1 \sin(0.2\pi y)\,. \tag{11.78}$$

The numerical solution for $h+z$, hu, hv and K at time $t = 10$ s provided by the 3-rd order ARL-ADER scheme is presented in Figure 11.2 and is compared with the exact solution. The proposed scheme preserves the geostrophic equilibrium up to machine precision, with round-off errors for K of magnitude 10^{-16}, as shown in Table 11.3. Hence, the scheme is perfectly balanced and satisfies the well-balanced property with machine precision.

11.5.2 Convergence test for the SWE with bed elevation

In this section, a convergence rate test for the ARL-ADER well-balanced scheme is presented. The following initial condition is imposed

$$z(x,y) = 0.1 \exp\left(-\frac{(x-50)^2+(y-50)^2}{80}\right), \qquad \forall (x,y) \in \Omega \tag{11.79}$$

$h(x,y,0) = 1$ and $hu(x,y,0) = hv(x,y,0) = 0$ $\forall (x,y) \in \Omega$. The computational domain is $\Omega = [0,100] \times$

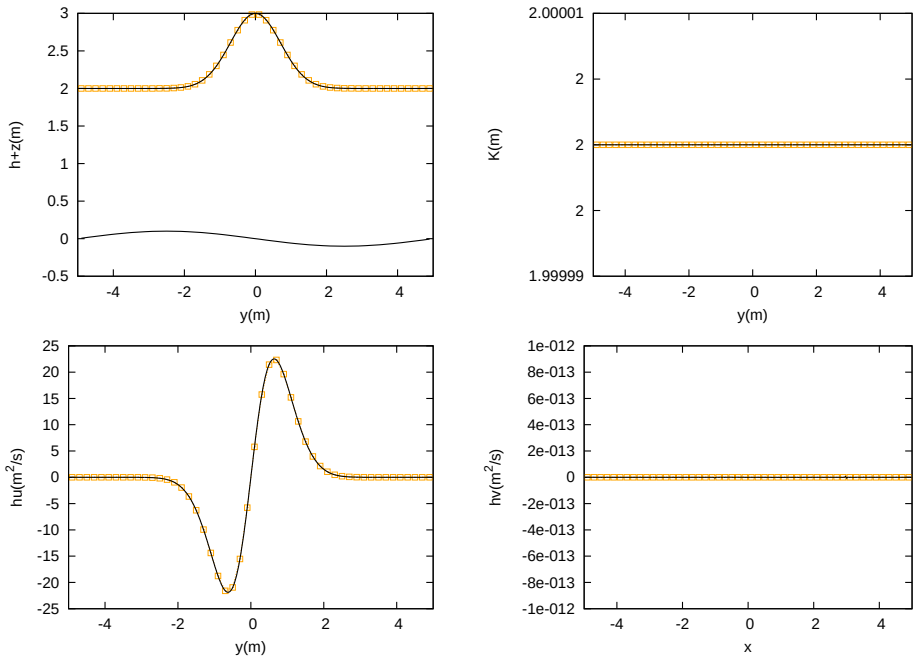


Figure 11.2: Section 11.5.1. Numerical solution for $h+z$, K, hu and hv at $t = 10$ s provided by the 3-rd order ARL-ADER scheme.

Scheme	N	L_1 error	Order	L_2 error	Order	L_∞ error	Order
1-st	50	1.13E-03		2.16E-05		8.65E-03	
	100	6.43E-04	0.81	1.25E-05	0.79	4.65E-03	0.89
	200	3.46E-04	0.90	6.79E-06	0.88	2.44E-03	0.93
	400	1.80E-04	0.94	3.55E-06	0.94	1.26E-03	0.96
3-rd	50	6.45E-05		1.24E-06		4.09E-04	
	100	8.69E-06	2.89	1.71E-07	2.86	5.83E-05	2.81
	200	1.10E-06	2.99	2.17E-08	2.98	7.39E-06	2.98
	400	1.36E-07	3.01	2.71E-09	3.00	9.21E-07	3.00

Table 11.4: Section 11.5.2. Convergence rate test for h using L_1 and L_2 and L_∞ error norms for the 1-st and 3-rd order ARL-ADER schemes. CFL=0.2.

$[0, 100]$ and the solution is computed at $t = 5$ s setting $CFL = 0.2$ using the 1-st and 3-rd order scheme.

Numerical errors and convergence rates for h and hu computed in fours different grids composed of 50×50, 100×100, 200×200 and 400×400 cells are presented in Tables 11.4 and 11.5, respectively. Numerical errors have been computed using a reference solution computed by the 3-rd order scheme in a 2000×2000 grid and are measured using the L_1, L_2 and L_∞ error norms. In Figure 11.3, a logarithmic plot of the numerical errors for the water depth and discharge provided by the 1-st and 3-rd order schemes are represented against the computational time. It can be observed that the theoretical convergence rates are achieved.

11.5.3 Shock reflection and diffraction against a solid body

Reflection and diffraction of shock waves at walls and corners are common features that appear in many problems of interest in the framework of shallow water flows. In this section, numerical results for the simulation of a shock reflection and diffraction against a solid body are presented. The computational domain is defined as $\Omega = [0, 100] \times [0, 100]$ with a solid body define by the points (40,40), (40,60), (60,60) and (60,40). The incoming shock is configured as $Fr = 1.559$ and $h + z = 6$ m and located at $x = 10$ m at the initial time. In the unperturbed region, we set $hu = hv = 0$ m^2/s and $h + z = 1$ m. For the

Scheme	N	L_1 error	Order	L_2 error	Order	L_∞ error	Order
1-st	50	2.17E-03		4.52E-05		1.86E-02	
	100	1.25E-03	0.80	2.67E-05	0.76	1.12E-02	0.73
	200	6.74E-04	0.89	1.47E-05	0.87	6.22E-03	0.85
	400	3.51E-04	0.94	7.70E-06	0.93	3.29E-03	0.92
3-rd	50	1.37E-04		3.02E-06		1.30E-03	
	100	1.82E-05	2.91	4.12E-07	2.87	1.80E-04	2.85
	200	2.30E-06	2.98	5.25E-08	2.97	2.35E-05	2.94
	400	2.86E-07	3.01	6.56E-09	3.00	3.00E-06	2.97

Table 11.5: Section 11.5.2. Convergence rate test for hu using L_1 and L_2 and L_∞ error norms for the 1-st and 3-rd order ARL-ADER schemes. CFL=0.2.

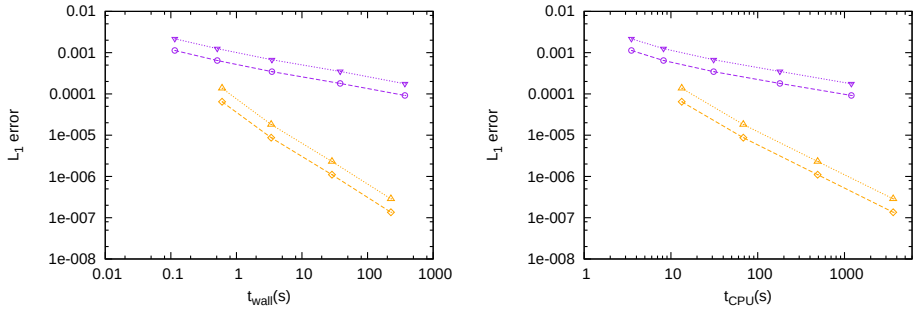


Figure 11.3: Section 11.5.2. Convergence rate test: logarithmic plot of the L_1 error against the wall-clock time (bottom-left) and CPU time (bottom-right). Solution computed using a 1-st (purple) and 3-rd (orange) order schemes.

calculation of the RH condition we have considered $g = 9.8$ m/s^2. The bed elevation is given by

$$z(x,y) = \begin{cases} 0 & \text{if} \quad x < 15 \\ 0.01(x-15) & \text{if} \quad x \geq 15 \end{cases} \tag{11.80}$$

The numerical solution is depicted in Figure 11.4 (top) at time $t = 5.5$ s showing the water surface elevation gradient and the velocity vector field. Numerical results for the second case are depicted in Figure 11.4 at $t = 5$ s. For both cases, the solution is computed using a 1-st and 3-rd order ARL-ADER numerical scheme, setting CFL=0.4.

The reflection of the incident shock is accurately captured by the schemes as well as the diffraction along the right-side corners of the square. The numerical results evidence that only when using the 3-rd order scheme, the small-scale features of the flow are captured.

11.5.4 Shock-vortex interaction

This test case is devoted to the resolution of the interaction between shock waves and vortices. Such wave pattern is generated by the diffraction of a single shock wave behind a square solid body. The computational domain is given by $\Omega = [0,100] \times [0,100]$ and the vertices of the solid body by the coordinates (40,40), (40,60), (60,60) and (60,40). The incoming shock is configured as $Fr = 0.50016$ and $h + z = 6$ m and located at $x = 10$ m at the initial time. In the unperturbed region, we set $hu = hv = 0$ m^2/s and $h + z = 1$ m. For the calculation of the RH condition we have considered $g = 9.8$ m/s^2. The bed elevation is given by (11.80). The solution is computed at $t = 20$ s using the ARL-ADER scheme in a 800×800 grid.

In Figure 11.5, the vorticity magnitude and water surface elevation provided by the 3-rd order ARL-ADER scheme is presented at different times. The scheme is able to accurately capture the reflection and diffraction of the initial shock wave, as well as the generation and transport of vorticity at the trailing and

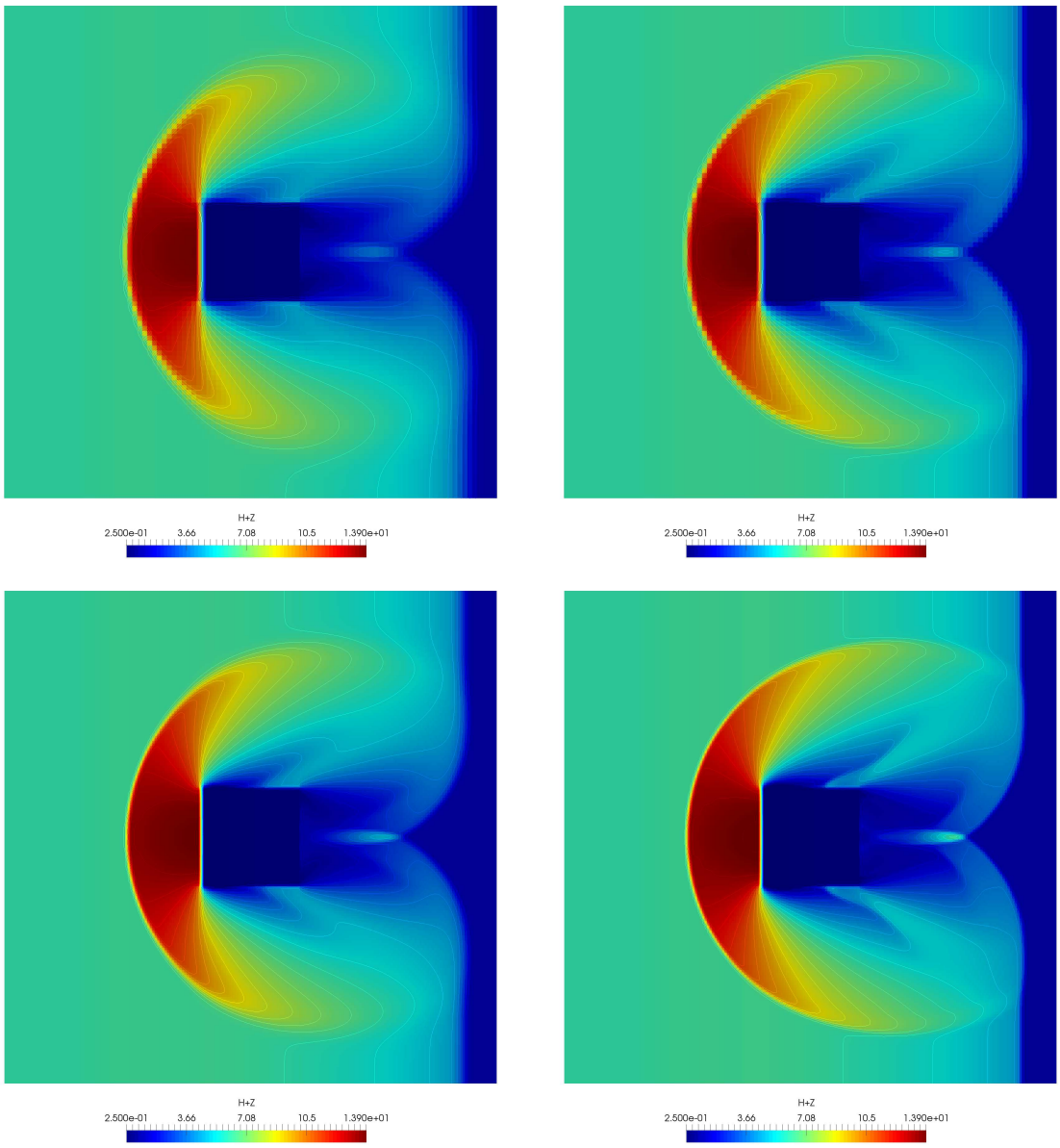


Figure 11.4: Water surface elevation computed by a 1-st (left) and 3-rd order (right) scheme using $\Delta x = 1$ (top) and $\Delta x = 0.5$ (bottom). 25 contour lines are plotted, equally spaced in the interval of the color scale.

leading corner edges. In the beginning, the diffracted shocks interact, collide (second image) and propagate in opposite directions. Such shocks eventually enter in the vortices and generate a complex shock-vortex interaction (fourth plot) which is accurately captured by the scheme.

In Figure 11.6, the solutions provided by the 1-st and 3-rd order ARL-ADER schemes at $t = 20$ s are compared. The numerical results evidence that the 1-st order scheme is unable to capture neither the vortical structures with detail nor the complex shock wave pattern appearing from the interactions.

11.5.5 Steady supercritical flow against a solid wedge: Mach reflection

The accurate capture of 2D shocks has been a challenging task in the framework of FV schemes for hyperbolic conservation laws. The resolution of complex wave reflection patterns in presence of solid bodies

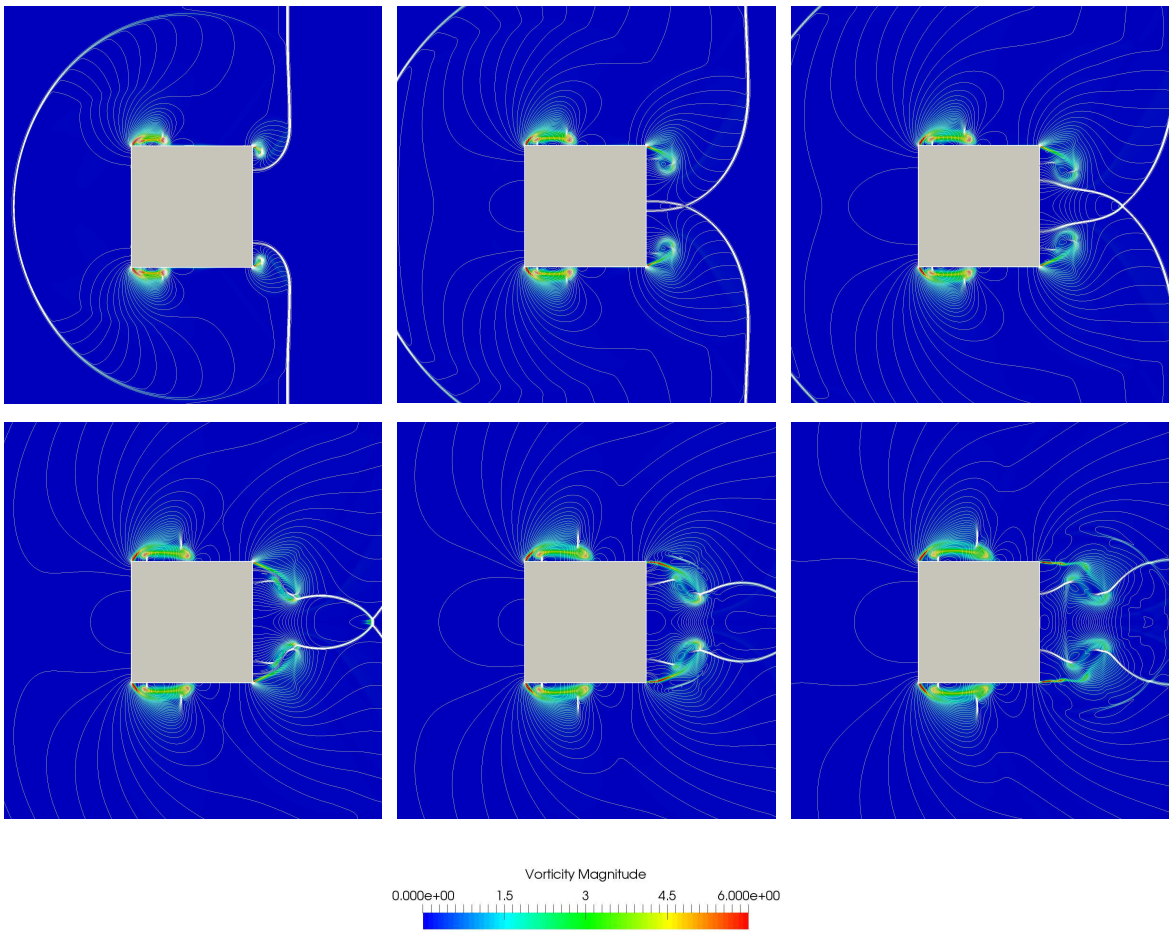


Figure 11.5: Vorticity magnitude and water surface elevation provided by the 3-rd order ARL-ADER scheme.

is not a trivial task, even when considering a flat bed. In this case, we consider the resolution of a Mach reflection (MR) pattern that arises in the reflection of an oblique shock against a solid wall.

Oblique shocks normally appear when a supercritical straight flow encounters a wedge that deflects it. The presence of the wedge involves a change in the flow field and aligns the flow in its direction, θ. The region of influence of the wedge corresponds to the region downstream the resulting oblique shock. The discontinuity between both regions is a shock wave with an angle β. When the incident oblique shock, hereafter denoted by I, encounters a solid wall, the MR may appear as depicted in Figure 11.7. This type of reflection leads to the so-called 3-shock solution.

The 3-shock solution for a given incoming Froude number and deflection angle can be obtained using the so-called shock polar diagram. Such diagram is a representation of $h/h_0 = h/h_0(\theta)$, with θ in the x-axis and h/h_0 in the y-axis. The solution for the MR will be located on the intersection between the curves $h_1/h_0 = h_1/h_0(\theta)$ and $h_{2,3}/h_0 = h_{2,3}/h_0(\theta)$, where $h_{2,3}$ is the water depth in regions (2) and (3). Note that $h_{2,3}/h_0$ can be easily computed as $h_{2,3}/h_0 = h_1/h_0(\theta_1) \cdot h_{2,3}/h_1(\theta^{'} - \theta_1)$, where $\theta^{'}$ is the deflection angle with respect to θ_1.

In this test case, we consider a supercritical flow aligned to the x-axis and confined in a straight channel with solid walls. The flow is defined by $Fr_0 = 4$ and $h_0 = 1$ m, and is deflected by a wedge of $\theta_1 = 23.3048^o$, generating an incident attached shock, I, which is eventually reflected by the top wall. The computational domain is given by $\Omega = [0, 100] \times [0, 55]$ and the solid domain is defined by the points $(15, 0)$, $(80, 28)$ and $(80, 0)$. Solid BCs are considered on the lateral walls, while a supercritical BC is considered at the inlet

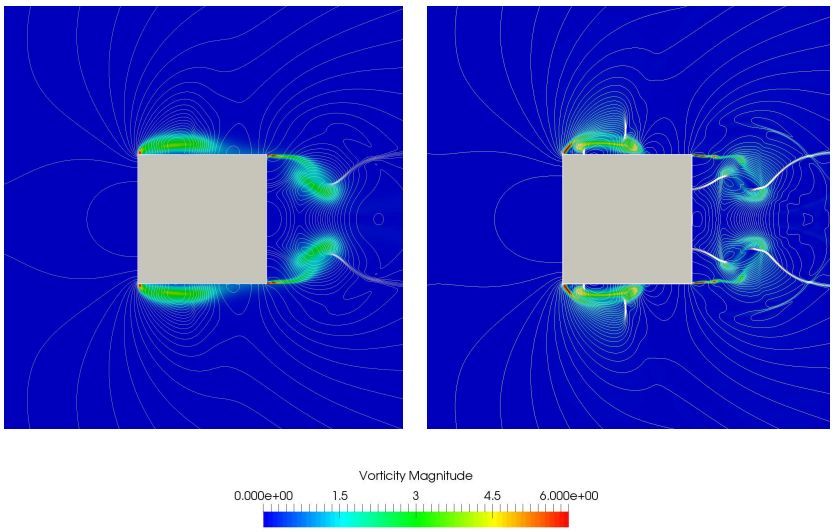


Figure 11.6: Vorticity magnitude and water surface elevation provided by the 1-st (left) and 3-rd order ARL-ADER scheme (right) at $t = 20$ s.

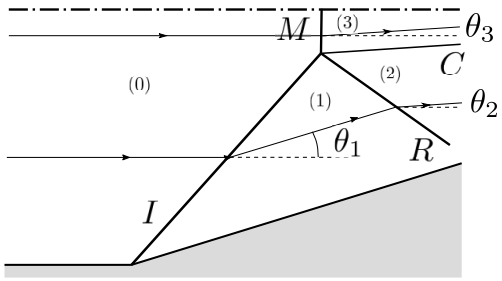


Figure 11.7: MR wave pattern, including relevant angles and states.

and a transmissive BC at the outlet. The solution is computed at $t = 200$ s using the 1-st and 3-rd order ARL-ADER scheme in a 200×110 grid (square) and 800×440 grid, and the traditional Roe method in a triangular grid with 285363 triangles.

The analytical solution is depicted in Figure 11.8, where the Mach polar is represented. It is observed that the intersection between such curves is located at the M branch (strong shock), hence a MR occurs. From this intersection, we see that in regions (2) and (3) the water depth is $h_{2,3} = 5.1754$ m and the angle of deflection of the slip line is $\theta_{2,3} = 7.92^o$. The numerical solutions are also shown in such plot. In Table 11.6, the numerical solution for h, $\theta_{2,3}$ and β_R is presented.

It is observed that all the schemes provide an accurate estimation of the I, M and R angles, but when using Cartesian grids, only the 3-rd order scheme accurately captures the direction of the slip line. Additionally, the presence of a spurious boundary layer in the region near the wedge is observed when using Cartesian grids. Such layer is due to the representation of the solid body by means of squares, in a stair-like shape. This introduces reflections in the x-direction and eventually produces the aforementioned spurious behavior. On the other hand, the use of triangular cells allows to construct boundary conforming grids. In this case, the flow is tangential to the real direction of the wall.

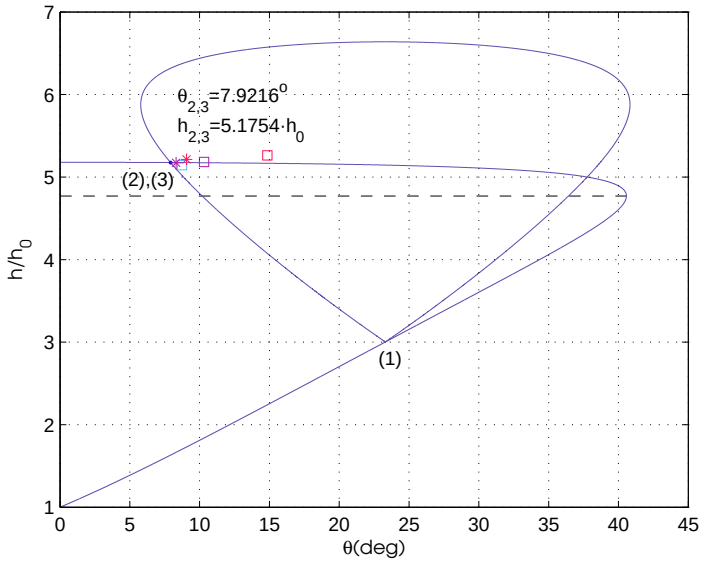


Figure 11.8: Mach polar diagram for an incident flow with $Fr_0 = 4$ and $h_0 = 1$ m, deflected by a wedge of $\theta_1 = 23.3048^o$. The numerical solution is depicted for the 1-st order scheme (square) and 3-rd order scheme (asterisk) in a 800×440 grid (purple), 200×110 (red) and triangular grid (blue)

Scheme	Grid	h (m)	$\theta_{2,3}$ (deg)	β_R (deg)
1-st	Cartesian 200×110	5.263	14.854	36.920
	Cartesian 800×440	5.182	10.316	32.622
	Triangular 285363 cells	5.144	8.739	31.699
3-rd	Cartesian 200×110	5.213	9.063	34.084
	Cartesian 800×440	5.173	8.310	31.257

Table 11.6: Numerical solution for h, $\theta_{2,3}$ and β_R provided by the 1-st and 3-rd order schemes in Cartesian and triangular grids.

11.5.6 Steady supercritical flow against forward-facing step

In early stages of the development of FV schemes, the *Mach 3 wind tunnel with a step* test problem was introduced [105] in the framework of Euler equations. This problem has proven to be a useful test for a large number of methods and a large number of years. An exhaustive comparison of the numerical performance of different methods when solving this problem can be found in [137].

Here, we propose the analogous problem for the shallow water equations with variable bed elevation and use it to test the numerical performance of the ADER schemes when solving complex 2D shock structures. The configuration of the case is detailed below.

We consider a computational domain given by $\Omega = [0,70] \times [0,25] \setminus \Pi_s$ where $\Pi_s = [15,70] \times [0,5]$ is a solid obstacle which mimics the step and is not included in the computational domain. The boundary condition at the inlet ($x = 0$) is defined as supercritical flow setting $h = 1$ m and $hu = 10.5$ m^2/s, while at the outlet ($x = 75$), free flow condition is imposed. On the remaining boundaries, we impose solid wall conditions. The initial condition is given by $h(x,y,0) = 1$ m, $hu(x,y,0) = 10.5$ m^2/s and $hv(x,y,0) = 0$ m^2/s $\forall (x,y) \in \Omega$.

Bed elevation is not constant in this case and is given by the following piecewise function

$$z(x,y) = \begin{cases} 0 & \text{if} \quad x \leq 5 \\ 0.3\,x^{0.3} - 0.486 & \text{if} \quad x > 5 \end{cases} \tag{11.81}$$

Two different grids are used in this test: a coarse grid, composed of 100×280 cells and a fine grid,

composed of 400×1120 cells. The numerical solution for the water surface elevation at $t = 100$ computed by the 1-st and 3-rd order well-balanced ARL-ADER scheme in each grid is presented in Figure 11.9. We have used $CFL = 0.3$.

In order to give a closer insight to the structure and features of the solution, a representation of the numerical solution of $|\nabla(h+z)|$, Froude number and velocity magnitude is presented in Figures 11.10, 11.11 and 11.12 respectively. The solution is represented at the time when the Kelvin-Helmholtz instability is better observed. From such figures we can see that the initial effect of the presence of the obstacle in the supercritical flow is the formation of a bow shock (hydraulic jump) that is reflected on the top solid wall and eventually forms a Mach stem that is joined to the incident wave (bow shock) and reflected wave at the so-called triple point. The main reflected wave is again reflected on the bottom solid wall and generates a secondary, but smaller and weaker, Mach stem and a secondary reflected wave. It is also worth mentioning that at the triple point, a shear layer appears and a Kelvin-Helmholtz instability develops, triggered by the numerical shockwave instabilities at the Mach stem and amplified by the physical instability at the slip line.

In Figure 11.9, it is observed that the position of the triple point does depend on the grid and the accuracy of the numerical scheme. The coarser the grid and the lower the accuracy is, the higher the triple point is located. Moreover, we observe that the solution provided by the 3-rd order scheme in the coarse grid is of about the same accuracy than the solution provided by the 1-st order scheme in the fine grid, which has 16 times more cells. The position of the triple point/upper Mach stem proved to be a good estimation for the accuracy of the scheme and can be used to assess the convergence [137]. In Figure 11.13, the x position of the Mach stem at $y = 24.5$ computed by the 1-st and 3-rd order schemes is plotted against the number of cells in the x-direction, denoted by N_x, chosen in the discretization. It is observed that for the 3-rd order scheme, when discretizing the domain with more than 1000 cells in the x-direction, the variation in the position of the Mach stem is of the order of 10^{-2}, while it is more than an order of magnitude higher for the 1-st order scheme.

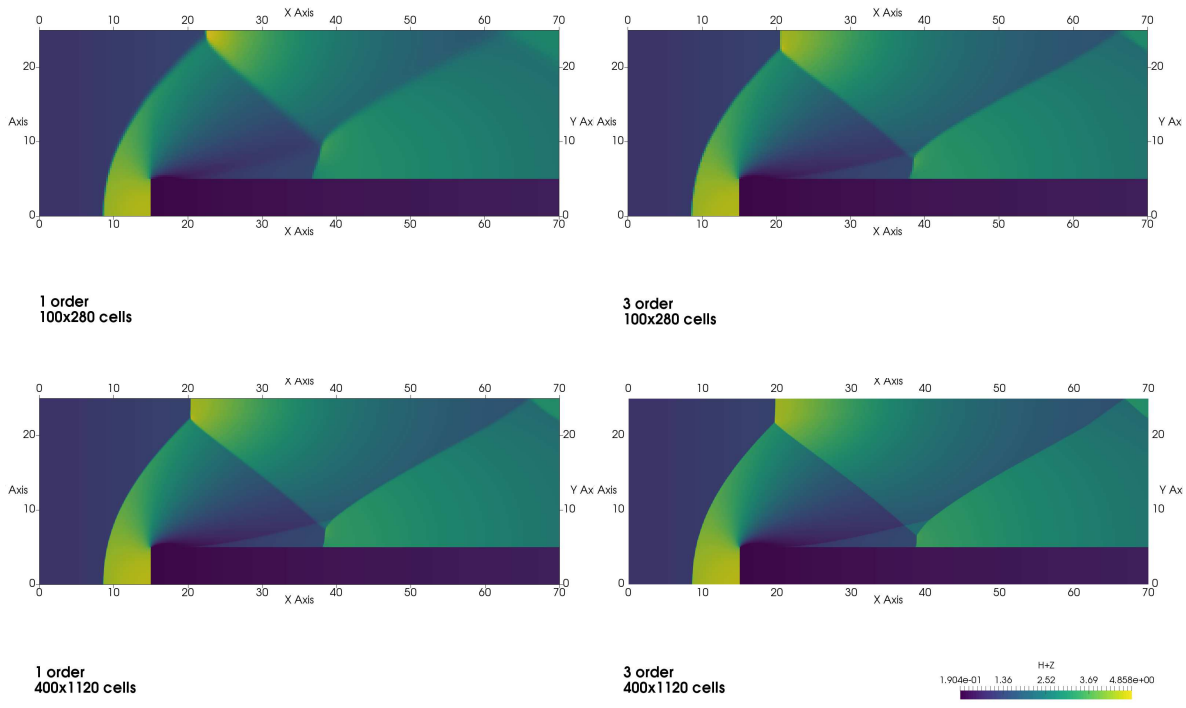


Figure 11.9: Section 11.5.6. Numerical solution for $h+z$ at $t = 100$ provided by the 1-st order (left) and 3-rd order scheme (right) using the 100×280 grid (top) and the 400×1120 grid (bottom).

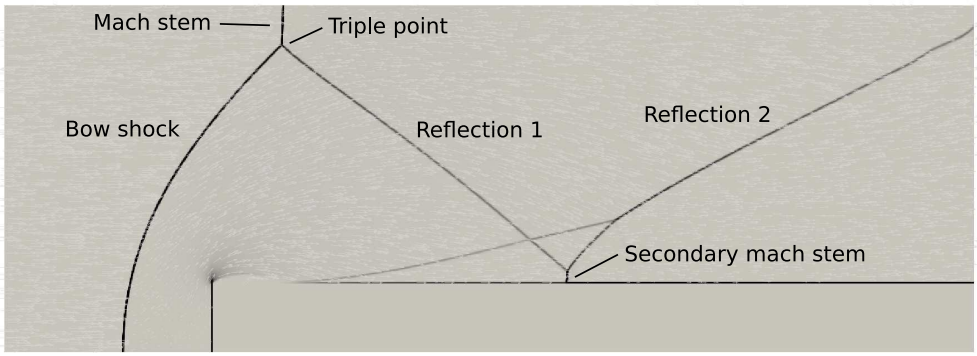


Figure 11.10: Section 11.5.6. Numerical solution for the water surface elevation gradient, including the relevant features of this particular flow.

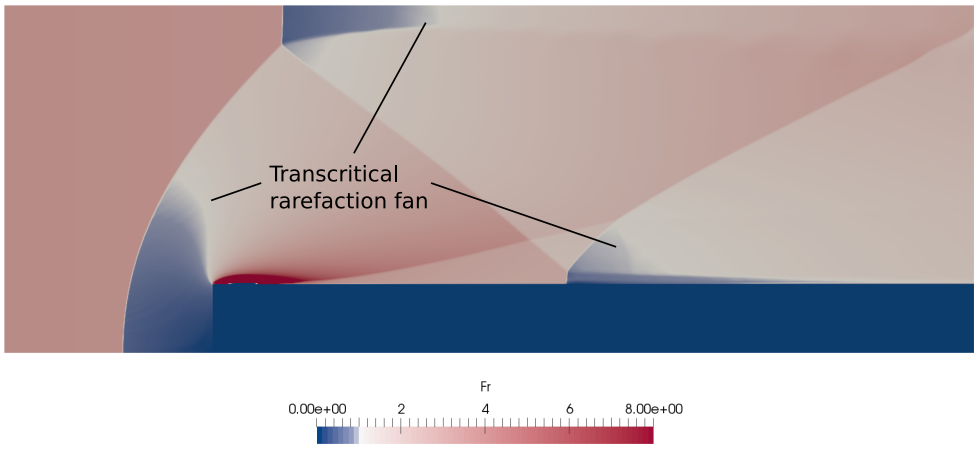


Figure 11.11: Section 11.5.6. Numerical solution for the Froude number.

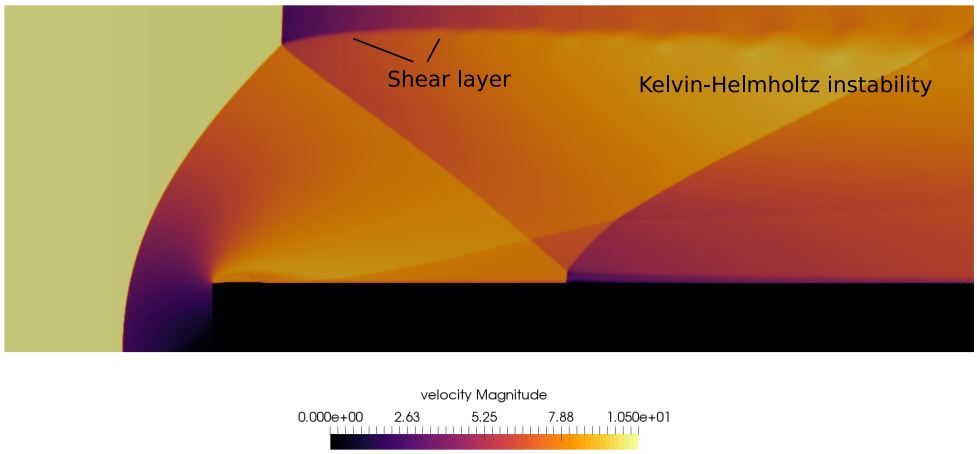


Figure 11.12: Section 11.5.6. Numerical solution for the velocity magnitude.

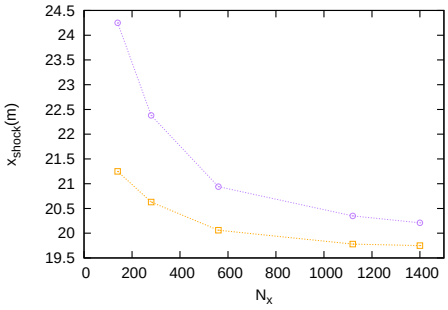


Figure 11.13: Section 11.5.6. Representation of the x position of the Mach stem at $y = 24.5$ computed by the 1-st (purple) and 3-rd order (orange) ARL-ADER schemes against the number of cells in the x-direction

11.5.7 2D Riemann problem

Compared with the relatively simple 1D RPs, the solution of 2D RPs include complex geometric wave patterns that pose a computational challenge for high-resolution numerical schemes. In this section, we solve a 2D RP whose initial condition is given by piecewise constant data in each of the four quadrants of the Cartesian plane. Hereafter, such quadrants will be denoted by Q_1, Q_2, Q_3, Q_4. The problem is solved using the 1-st and 3-rd order ADER scheme in a grid composed of 800×800 cells using $CFL = 0.4$ in the domain $\Omega = [0, 100] \times [0, 100]$, at a time $t = 3$.

The initial condition is given by

$$h(x,y) = \begin{cases} 10 & \text{if} \quad (x,y) \in Q_1 \\ 3 & \text{if} \quad (x,y) \in Q_2 \\ 0.2 & \text{if} \quad (x,y) \in Q_3 \\ 3 & \text{if} \quad (x,y) \in Q_4 \end{cases}, z(x,y) = \begin{cases} 1 & \text{if} \quad (x,y) \in Q_1 \\ 0 & \text{if} \quad (x,y) \in Q_2 \\ -0.5 & \text{if} \quad (x,y) \in Q_3 \\ 0 & \text{if} \quad (x,y) \in Q_4 \end{cases} \tag{11.82}$$

$$u(x,y) = \begin{cases} 0 & \text{if} \quad (x,y) \in Q_1 \\ 3 & \text{if} \quad (x,y) \in Q_2 \\ 3 & \text{if} \quad (x,y) \in Q_3 \\ 0 & \text{if} \quad (x,y) \in Q_4 \end{cases}, v(x,y) = \begin{cases} 0 & \text{if} \quad (x,y) \in Q_1 \\ 0 & \text{if} \quad (x,y) \in Q_2 \\ 3 & \text{if} \quad (x,y) \in Q_3 \\ 3 & \text{if} \quad (x,y) \in Q_4 \end{cases} \tag{11.83}$$

The structure of the solution is composed by a complex variety of waves: at the four edges of the quadrants limiting with the axis, there are four contact discontinuities due to the bed steps. In the first quadrant, there are two interacting rarefaction waves moving in the x and y directions. In the second and fourth quadrants, there are two rarefaction waves moving in the y and x direction respectively, which interact with a transverse shock wave. Finally, the most complex wave pattern can be found in the third quadrant. Here, there are two pairs of shocks in each coordinate direction, moving perpendicularly with respect to each other, and from their interaction a jet stream pointing to the bottom-left corner of the domain is generated. Such jet is bounded by multiple shocks and a strong recirculation is observed at both sides of the jet. It is worth pointing out that this RP is a resonant problem as the number of waves is greater than the number of eigenvalues.

A shadowgraph for the numerical $h + z$ and the velocity vector field has been represented in Figure 11.14. A cross-sectional representation of the numerical $h + z$ and q computed using a 1-st and 3-rd order scheme in a 200×200 and 800×800 grid is also presented in Figure 11.15 and compared with a reference solution that has been computed using the 3-rd order scheme in a very fine mesh.

It is observed that only when using the 3-rd order scheme, the finest details such as the normal shock produced by the deceleration of the reverse flow in the jet in the third quadrant, can be captured. Moreover, the velocity peak in the jet is underestimated when using the 1-st order scheme. Important differences can also be found in the first quadrant, where the solution provided by the 1-st order scheme shows larger

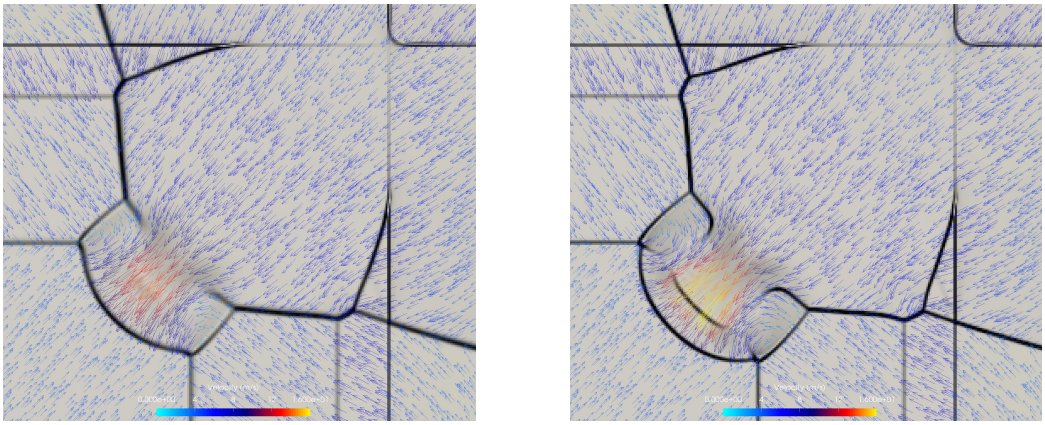


Figure 11.14: Section 11.5.7. Water surface elevation gradient and velocity field provided by the 1-st order scheme (left) and 3-rd order scheme (right) in a 200×200 grid.

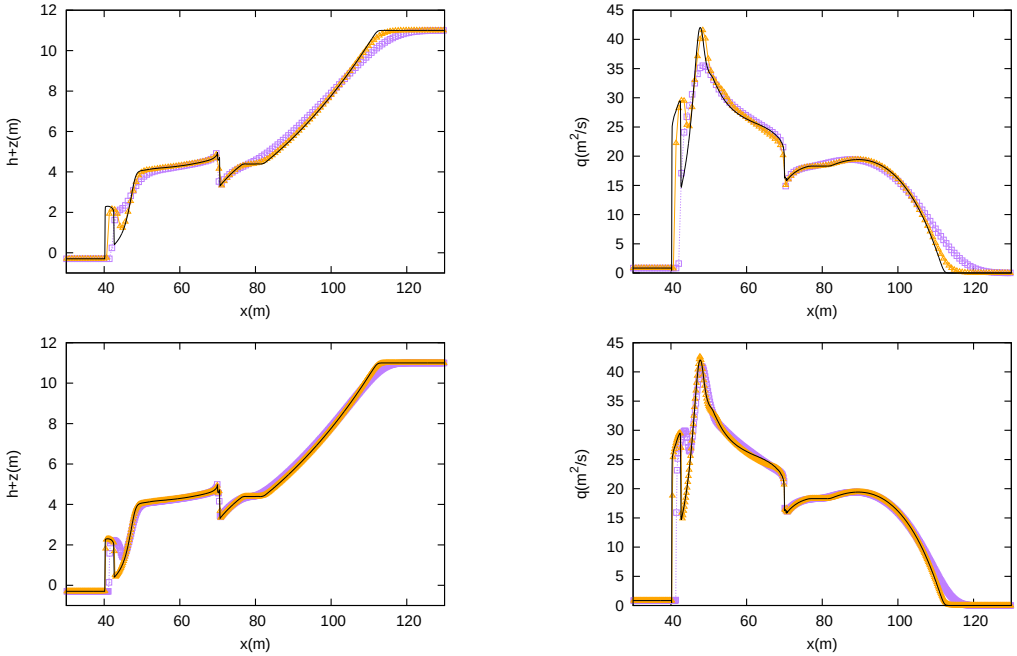


Figure 11.15: Section 11.5.7. Cross-sectional representation of the numerical $h+z$ (left) and q (right) computed using a 1-st and 3-rd order scheme in a 200×200 (top) and 800×800 (bottom) grid.

diffusion and does not capture the correct shape of the interaction of the rarefaction fans.

11.5.8 Convergence test for the SWE with bed elevation and Coriolis

A convergence rate test for the ARL-ADER well-balanced scheme with bed elevation and Coriolis is presented. The following initial condition is imposed

$$z(x,y) = 0.1 \exp\left(-\frac{(x-50)^2+(y-50)^2}{80}\right), \qquad \forall (x,y) \in \Omega \tag{11.84}$$

$h(x,y,0) = 0.5$ and $hu(x,y,0) = hv(x,y,0) = 0$ $\forall (x,y) \in \Omega$ and $f = 2$ s^{-1}. The computational domain is $\Omega = [0,100] \times [0,100]$ and the solution is computed at $t = 3$ s setting $CFL = 0.2$ using the 3-rd order scheme.

Variable	N	L_1 error	Order	L_2 error	Order	L_∞ error	Order
h	50	1.81E-06		8.30E-08		1.04E-04	
	100	3.72E-07	2.28	1.63E-08	2.35	2.00E-05	2.37
	200	6.62E-08	2.49	2.78E-09	2.55	3.39E-06	2.56
	400	1.24E-08	2.41	4.70E-10	2.57	5.48E-07	2.63
hu	50	2.19E-05		7.84E-07		5.40E-04	
	100	3.24E-06	2.76	1.16E-07	2.75	8.27E-05	2.71
	200	5.05E-07	2.68	1.85E-08	2.65	1.49E-05	2.48
	400	8.87E-08	2.51	3.44E-09	2.43	3.14E-06	2.24

Table 11.7: Section 11.5.8. Convergence rate test for h and hu using L_1 and L_2 and L_∞ error norms for the 3-rd order ARL-ADER scheme. CFL=0.2.

Numerical errors and convergence rates for h and hu computed in fours different grids composed of 50×50, 100×100, 200×200 and 400×400 cells are presented in Table 11.7. Numerical errors have been computed for h and hu using a reference solution computed by the 3-rd order scheme in a 2000×2000 grid and are measured using the L_1, L_2 and L_∞ error norms. Numerical errors for hv are not presented due to the symmetry of the case.

It is observed that the theoretical convergence rate is not fully achieved. This may be due to a sub-optimal time integration method. Such integral is computed by integrating the Taylor polynomial, which is constructed using time derivatives obtained with the CK procedure. First order time derivatives were derived expressing the source term in terms of the primitive variables. This is required to ensure the well–balanced property. By contrast, second and higher order time derivatives were derived considering the original source terms fhu and fhv. It seems that the use of this combined approach for the derivation of time derivatives makes the evolution in time of suboptimal accuracy.

11.5.9 2D geostrophic adjustment

Here we consider the test case proposed in [89] (see also [88, 92]), which consists of a initial asymmetrical column of water that falls under a strong rotation that leads to a 2D geostrophic adjustment the numerical scheme must be able to reproduce. The computational domain is $\Omega = [-10,10]\times[-10,10]$, the bottom topography is flat and the initial condition is given by

$$h(x,y) = 1 + 0.5A_0\left(1-\tanh\left(\frac{\sqrt{(\sqrt{\lambda}x)^2+(y/\sqrt{\lambda})^2}-R_i}{R_E}\right)\right), \tag{11.85}$$

$$u(x,y) = v(x,y) = 0\,, \tag{11.86}$$

where $A_0 = 0.5$, $\lambda = 2.5$, $R_E = 0.1$, $R_i = 1$ and the gravity and Coriolis parameters are set to $g = 1$ m/s^2 and $f = 1$ s^{-1}. The numerical solution is computed using the 3-rd order ARL-ADER scheme using a grid of 400×400 cells and setting CFL=0.4. Initially, the elliptical column of water is not at equilibrium and evolves in an nonaxisymetric way due to the rotation effect. Two successive shock (gravity) waves are generated and leave behind a small smooth hump that is slowly spinning clockwise. The numerical result for the water surface elevation at times $t = 0$, $t = 4$, $t = 8$, $t = 12$, $t = 16$ and $t = 20$ s are presented in Figure 11.16. It is observed that the expected behavior of the evolution of the solution is reproduced by the numerical scheme and that the numerical results are identical to those presented in [88, 92].

A cross sectional representation of the solution for $h+z$ and L at $y = 10$ and $t = 4$ s, provided by a 1-st and 3-rd order ARL-ADER scheme in two grids composed of 101×101 and 401×401 cells is depicted in Figure 11.17.

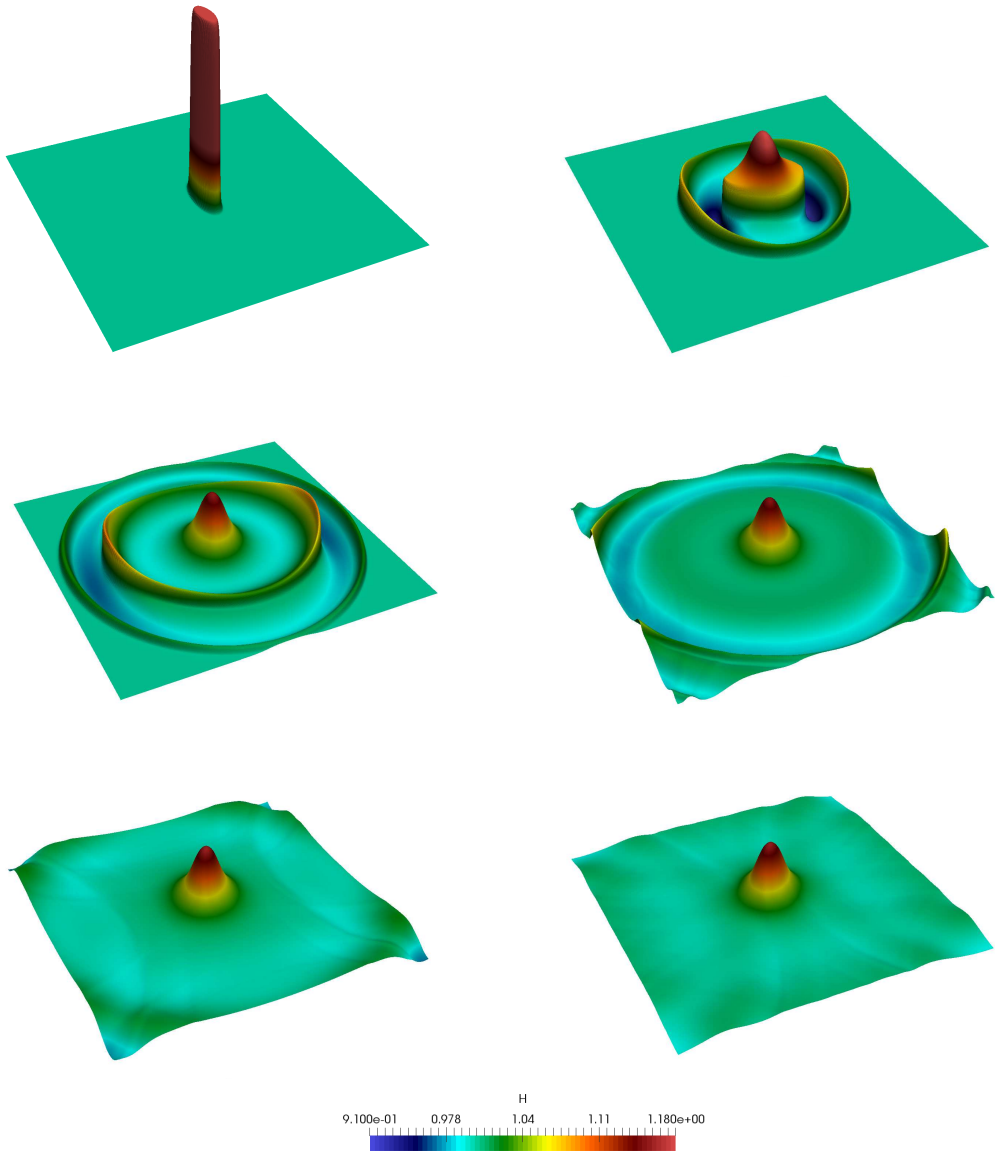


Figure 11.16: Section 11.5.9. Numerical $h+z$ at $t=0$ (top-left), $t=4$ (top-right), $t=8$ s (middle-left), $t=12$ (middle-right), $t=16$ (bottom-left) and $t=20$ s (bottom-right) provided by the 3-rd order ARL-ADER scheme in a 400×400 grid.

11.5.10 2D propagation of Rossby waves on the equatorial β-plane

This test case considers the propagation of a Rossby soliton on the equatorial beta-plane, for which an asymptotic solution exists to the inviscid SWE. Theoretically, the soliton should propagate to the west at fixed phase speed, without change of shape. Since the uniform propagation and shape preservation of the soliton are achieved through a fine balance between linear wave dynamics and nonlinearity, this is a good context in which to look for erroneous wave dispersion and/or numerical damping and has proven to be a good benchmark for atmosphere and ocean models (http://marine.rutgers.edu/po /index.php?model=test-problems) [138]. The interest in this test problem is to assess the spurious dispersion and dissipation effects of the numerical scheme, and how they relate to the choice of grid resolution and the accuracy of the scheme.

Long, weakly nonlinear, equatorial Rossby waves are governed by either Korteweg–de Vries (KDV) or

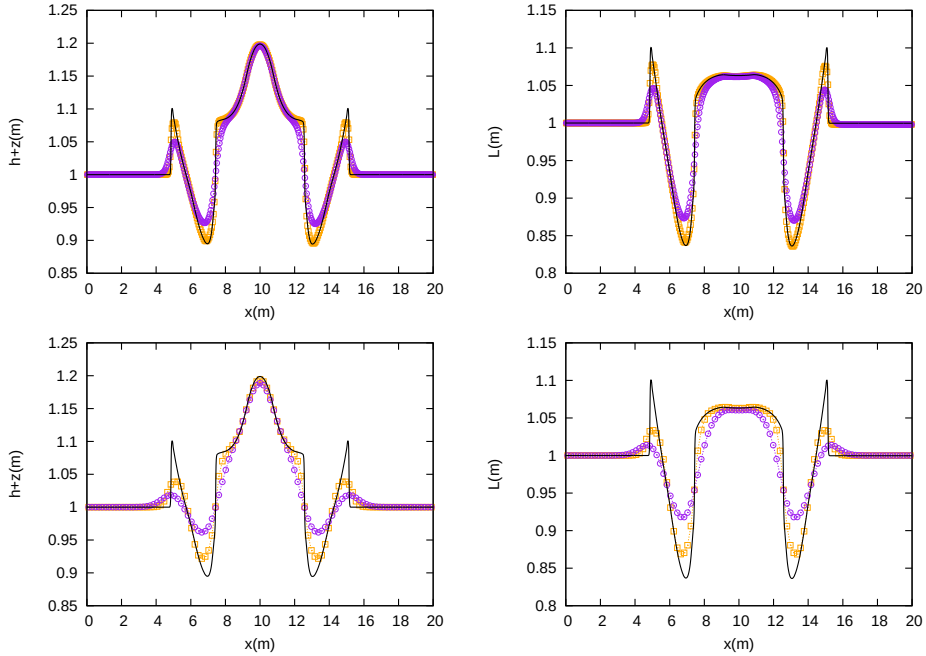


Figure 11.17: Section 11.5.9. Cross sectional representation of the solution for $h+z$ and L at $y=10$ and $t=4$ provided by a 1-st and 3-rd order ARL-ADER scheme in a 101×101 and 401×401 grid.

the modified Korteweg–de Vries (MKDV) equation [138, 139]. Here, a $zero$-th order asymptotic solution to the SWE is used [139]. The initial condition for a dipole at $(x,y)=(0,0)$ can be found in [138, 139]. Here, the dipole is translated to $(x,y)=(72,6)$ by means of evaluating the initial condition at $x'=x-72$ and $y'=y-6$. The gravity constant is set to $g=1$ m/s^2 and the Coriolis parameter is calculated using the β-plane approximation $f(y)=f_0+\beta y$ with $f_0=0$ s^{-1} and $\beta=1$ m^{-1}s^{-1}.

The case considered here consists of the $zero$-th order soliton described above over a flat bed, that is $z(x,y)=0$, computed inside the domain $\Omega=[0,96]\times[0,12]$ at time $t=120$ s. The numerical solution provided by the 1-st order (11.18) and 3-rd order ARL-ADER scheme (11.19) at times $t=0$, $t=30$, $t=60$, $t=90$ and $t=120$ s using two grid sizes of $\Delta x=0.2$ and $\Delta x=0.1$ m, are presented and compared with the exact solution at $t=120$ s. CFL is set to 0.4. It is observed that the 1-st order scheme in the coarsest grid does not perform well as it generates spurious waves. When moving to the finest grid size, the performance of the scheme is improved though it is still very diffusive and dispersive. The 3-rd order scheme does provide an accurate solution with both grids and ensures a much lower dispersion and diffusion of the soliton.

To assess the performance of the numerical schemes, we have used the following metrics: the damping factor, ν, which accounts for the numerical damping of the solution and the relative speed, c_r, which accounts for the numerical dispersion of the solution, defined in [138].

It is worth noting that all maximum and minimum water depth values are cell-averaged values and no interpolation is used. Numerical values for the metrics described above and other related data is presented in Table 11.8 using the results provided by the 1-st and 3-rd AR-ADER scheme in two grids of $\Delta x=0.2$ and $\Delta x=0.1$ m. It is evidenced that the 3-rd order scheme outperforms the 1-st order scheme in terms of numerical dispersion and damping, as it was expected. If comparing with the numerical results in [138] it is observed that the measures for dispersion and damping are of the same magnitude.

	3-rd order		1-st order	
	$\Delta x = 0.2$	$\Delta x = 0.1$	$\Delta x = 0.2$	$\Delta x = 0.1$
$h_{max}(t=0)$	1.16948	1.17032	1.16948	1.17032
$h_{min}(t=0)$	1.00000	1.00000	1.00000	1.00000
$h_{max}(t=120)$	1.14019	1.15332	1.05989	1.08135
$h_{min}(t=120)$	0.99509	0.99263	≈ 0	0.99835
ν	0.975	0.985	0.906	0.924
x_{end}	24.9	24.84	26.9	26.2
c_r	0.981	0.983	0.940	0.954
Time (CPU)	212.8	1777.5	7.2	39.5
Time (wall-clock)	12.55	100.75	0.95	6.7

Table 11.8: Numerical values of the relevant metrics for the assessment of numerical dispersion and damping of the scheme.

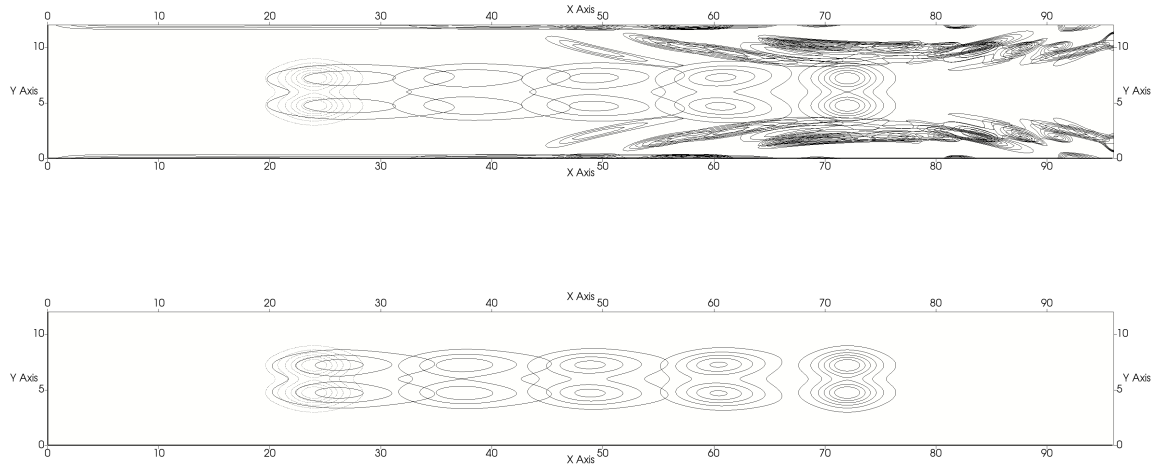


Figure 11.18: Section 11.5.10. Numerical solution provided by the 1-st order scheme at times $t = 0$, $t = 30$, $t = 60$, $t = 90$ and $t = 120$ s, using two different grids with $\Delta x = 0.2$ (top) and $\Delta x = 0.1$ m (bottom). The contour plot has been generated using 6 intervals from 1.02 to 1.14.

11.5.11 Kelvin front generation on the equatorial β-plane

In this section, the generation of nonlinear planetary (Rossby) and Kelvin waves at Earth's equatorial line is reproduced by the ARL-ADER numerical scheme. When the equatorial area is perturbed (by changing winds, for instance), its adjustment to the new equilibrium state is done by means of wave propagation. Such perturbations are usually of a very low frequency and therefore gravity waves are not excited, instead, only certain type of waves such as Kelvin waves, mixed waves and planetary waves (Rossby waves) appear. The short wavelength Kelvin waves carry energy eastward direction, whereas the long wavelength planetary waves carry energy to westward direction.

An additional phenomena is considered in this test case. It has been recognized for some time that nonlinear equatorial Kelvin waves can steepen and break, forming a broken wave, or front, propagating eastward [140]. This leads to the generation of equatorially trapped inertial-gravity (or Poincare) waves, which is an analogous mechanism than for nonlinear coastal Kelvin waves.

In this test case, we aim to show that the proposed numerical scheme is able to reproduce the formation and propagation of both Rossby and Kelvin waves over a non-flat bed elevation and eventually the generation of the Kelvin front and resonant formation of Poincare waves. The computational domain for this test case is $\Omega = [0, 70] \times [0, 12]$. The initial perturbation is given by a Gaussian water surface elevation

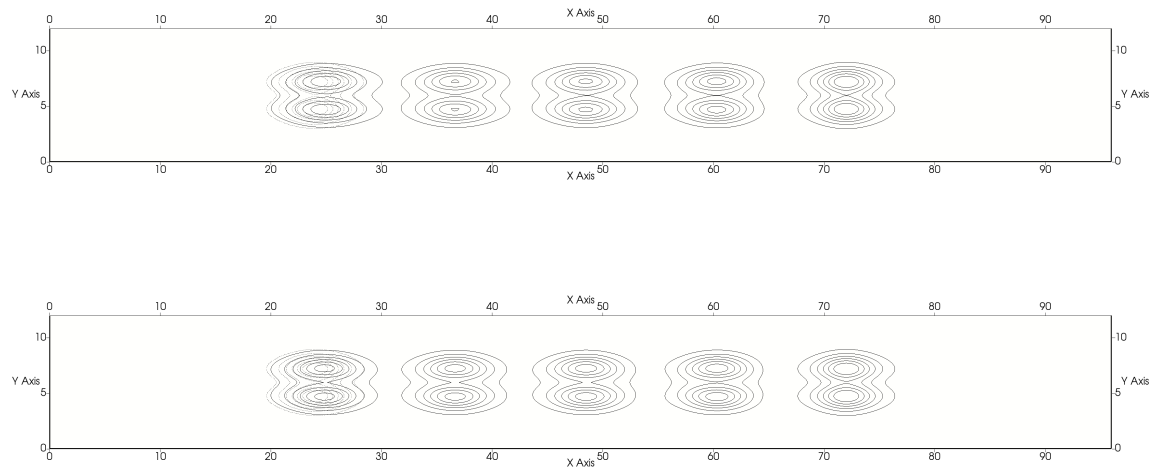


Figure 11.19: Section 11.5.10. Numerical solution provided by the 3-rd order scheme at times $t = 0$, $t = 30$, $t = 60$, $t = 90$ and $t = 120$ s, using two different grids with $\Delta x = 0.2$ (top) and $\Delta x = 0.1$ m (bottom). The contour plot has been generated using 6 intervals from 1.02 to 1.14.

anomaly, which reads

$$h(x,y) = h_0 + 0.8\exp\left(-\frac{(x-30)^2+(y-6)^2}{3}\right) \tag{11.87}$$

where $h_0 = 2$ m. The bed elevation is given by

$$z(x,y) = \begin{cases} 0 & \text{if } x \geq 40 \\ 0.025x - 1 & \text{if } x > 40 \end{cases} \tag{11.88}$$

The numerical solution is computed using the 1-st and 3-rd order ARL-ADER scheme in a 700×120 and 1400×240 grids, using $CFL = 0.4$. The solution for $h + z$ at $t = 40$ s is depicted in Figures 11.20 and 11.21 using a contour plot with 20 intervals from 1.94 to 2.36 m. It is observed that only when using the 3-rd order scheme, an accurate resolution of the Kelvin front formation is possible and Poincare waves are captured. Regarding the planetary waves moving westward, it is worth mentioning that both schemes are able to reproduce the expected physical behavior, though the first order scheme is more diffusive and dispersive.

11.5.12 Anticyclonic propagation in the β-plane

The proposed scheme is applied here to a more realistic case from [141] that consists of a initially symmetric vortex propagating westward due to the effect of the variation of the Coriolis coefficient in the y-direction. The domain extent is an idealized 2000 km $\times$ 1200 km rectangular basin and the initial condition is given by a Gaussian distribution of the free surface centered at the origin of the domain, prescribed together with a velocity field which is in geostrophic balance. The water depth at the initial time is given by

$$h(x,y) = h_0 + \zeta(x,y), \tag{11.89}$$

where $h_0 = 1.631$ m is the water depth reference level and $\zeta(x,y)$ is the surface height anomaly given by

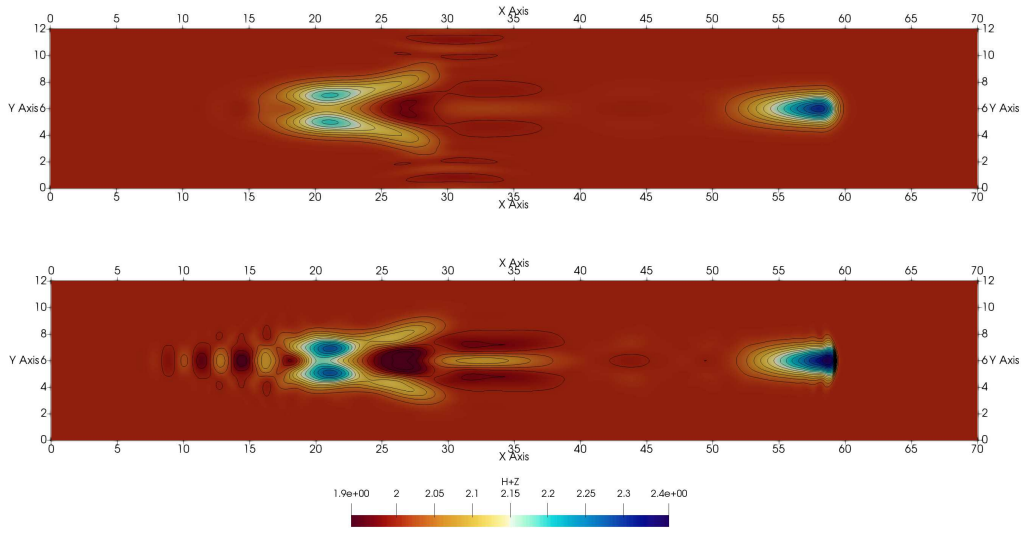


Figure 11.20: Section 11.5.11. Numerical solution for $h+z$ provided by the 1-st (top) and 3-rd order (bottom) ARL-ADER scheme in the coarse mesh at $t = 40$ s using $CFL = 0.4$. The contour plot has been generated using 20 intervals from 1.94 to 2.36.

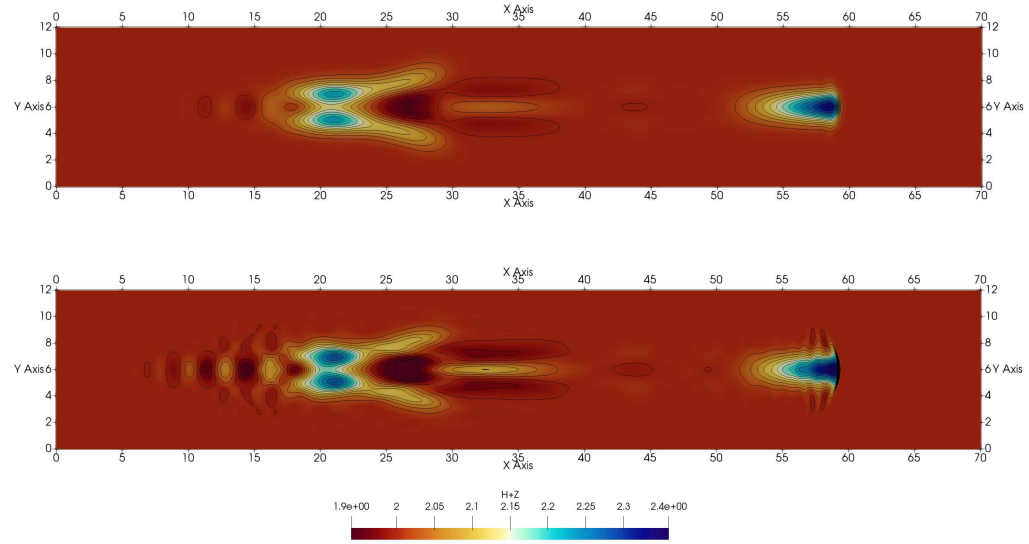


Figure 11.21: Section 11.5.11. Numerical solution for $h+z$ provided by the 1-st (top) and 3-rd order (bottom) ARL-ADER scheme in the fine mesh at $t = 40$ s using $CFL = 0.4$. The contour plot has been generated using 20 intervals from 1.94 to 2.36.

$$\zeta(x,y) = Ae^{-(x^2+y^2)/B^2} \tag{11.90}$$

with $A = 0.95$ m and $B = 130$ km. The initial velocity field is given by

$$u_1(x,y) = 2A\frac{g}{f(y)}\frac{y}{B^2}e^{-(x^2+y^2)/B^2}, \quad u_2(x,y) = -2A\frac{g}{f(y)}\frac{x}{B^2}e^{-(x^2+y^2)/B^2}, \tag{11.91}$$

with $f(y)$ the Coriolis parameter, evaluated at 25^o N using the β-plane approximation with $f_0 = 6.1635 \times 10^{-5}$ s^{-1} and $\beta = 2.0746 \times 10^{-11}$ m^{-1}s^{-1}. The gravity constant is set to $g = 9.81$ m/s^2.

The solution is computed using the 1-st and 3-rd order ARL-ADER schemes at $t = 8$ weeks. Figure 11.22 shows the numerical solution for the 1-st (left) and 3-rd order scheme (right) using a 100×60 grid (top) and a 200×120 grid. As reported in [92], the first order scheme is not able to reproduce the physical behavior of the solution and is much more mesh dependent than the 3-rd order scheme, which provides a rather accurate solution even for the coarsest grid.

In Figure 11.23, the trajectory of the center of the moving vortex computed by the 1-st and 3-rd order schemes is plotted in the $x-y$ plane. The trajectories are computed using three different grids composed of 100×60 cells, 200×120 cells and 400×240 cells. It is observed that the 3-rd order scheme provides an accurate prediction of the trajectory of the westward moving eddy, even for the coarsest grid, while the 1-st order scheme has a much lower convergence rate and requires more than 400×240 cells to predict the trajectory within an acceptable level of accuracy.

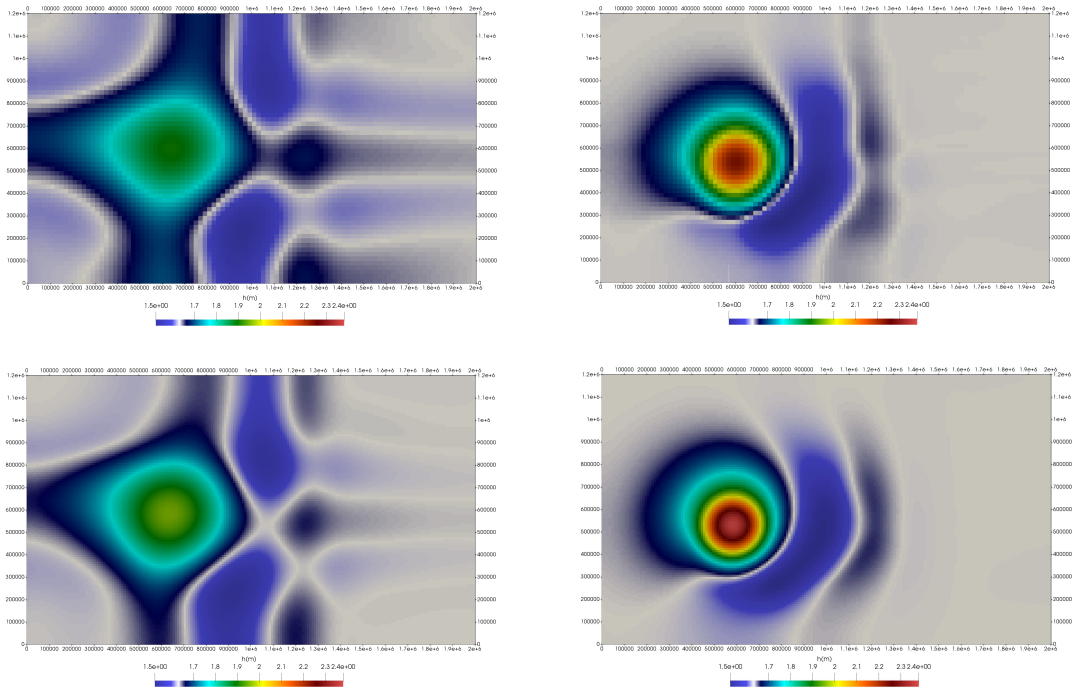


Figure 11.22: Section 11.5.12. Numerical solution computed by the 1-st (left) and 3-rd order scheme (right) using a 100×60 grid (top) and a 200×120 grid.

11.5.13 Convergence test for the SWE with bed elevation and friction

A convergence rate test for the ARL-ADER well-balanced scheme with bed elevation and friction is presented. The following initial condition is imposed

$$z(x,y) = 0.2 \exp\left(-\frac{(x-50)^2 + (y-50)^2}{80}\right), \qquad \forall (x,y) \in \Omega \tag{11.92}$$

$h(x,y,0) = 1.0$ and $hu(x,y,0) = hv(x,y,0) = 0$ $\forall (x,y) \in \Omega$, $g = 9.81$ m/s^2 and $n = 0.05$ s/m$^{1/3}$. The computational domain is $\Omega = [0,100] \times [0,100]$ and the solution is computed at $t = 5$ s using the 3-rd order scheme, setting $CFL = 0.1$.

Numerical errors and convergence rates for h and hu computed in fours different grids composed of

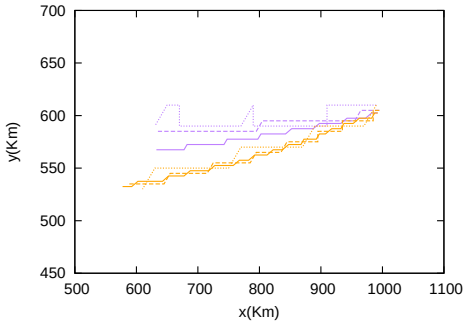


Figure 11.23: Section 11.5.12. Trajectory of the center of the moving vortex in the $x-y$ plane computed by the 1-st (purple) and 3-rd order scheme (orange) in a 100×60 cell grid (dotted line), 200×120 cell grid (dashed line) and 400×240 cell grid (solid line).

Variable	N	L_1 error	Order	L_2 error	Order	L_∞ error	Order
h	50	1.30E+00		2.47E-02		8.02E-04	
	100	1.72E-01	2.92	3.37E-03	2.87	1.25E-04	2.68
	200	2.05E-02	3.07	4.07E-04	3.05	1.67E-05	2.91
	400	2.26E-03	3.17	4.54E-05	3.16	2.07E-06	3.01
hu	50	2.82E+00		6.22E-02		2.75E-03	
	100	3.75E-01	2.91	8.46E-03	2.88	4.01E-04	2.78
	200	4.52E-02	3.05	1.03E-03	3.04	5.42E-05	2.89
	400	5.16E-03	3.13	1.17E-04	3.14	7.06E-06	2.94

Table 11.9: Section 11.5.13. Convergence rate test for h and hu using L_1 and L_2 and L_∞ error norms for the 3-rd order ARL-ADER scheme. CFL=0.1.

50×50, 100×100, 200×200 and 400×400 cells are presented in Table 11.9. Numerical errors have been computed for h and hu using a reference solution computed by the 3-rd order scheme in a 2000×2000 grid and are measured using the L_1, L_2 and L_∞ error norms. The prescribed convergence rate is achieved.

11.5.14 Simulation of the seiche phenomenon in channels with lateral cavities

Generally, a seiche is a standing gravity wave in a partially enclosed body of water. In the particular case of a channel with lateral cavities, it is produced by the coupling between the instability of the separated turbulent layer along the opening of the cavities and a gravity standing wave within the cavities. Such coupling is associated with large-scale coherent vortical structures in the unstable shear layer and periodic oscillations of the free surface within the cavity [142]. Figure 11.24 depicts a typical configuration of a channel with lateral cavities and includes the geometrical parameters of relevance, namely the channel width, b, the cavity width, W and the cavity length, L. The region where the vortex shedding takes place and the direction of the induced standing gravity wave are also depicted in the figure.

The presence of standing gravity waves in the cavities can be analytically examined. If considering a linearization of the system (9.2), neglecting the source terms, we obtain the traditional wave equation, which in 1D reads

$$\frac{\partial \hat{h}^2}{\partial t^2} - c^2 \frac{\partial \hat{h}^2}{\partial y^2} = 0, \tag{11.93}$$

where $c = \sqrt{gh}$ is the wave celerity, $\hat{h}$ is a perturbation around the reference depth, h, and y is the spatial coordinate in the transverse direction to the channel. Equation (11.93) models the propagation of linear gravity waves and admits periodic solutions as follows

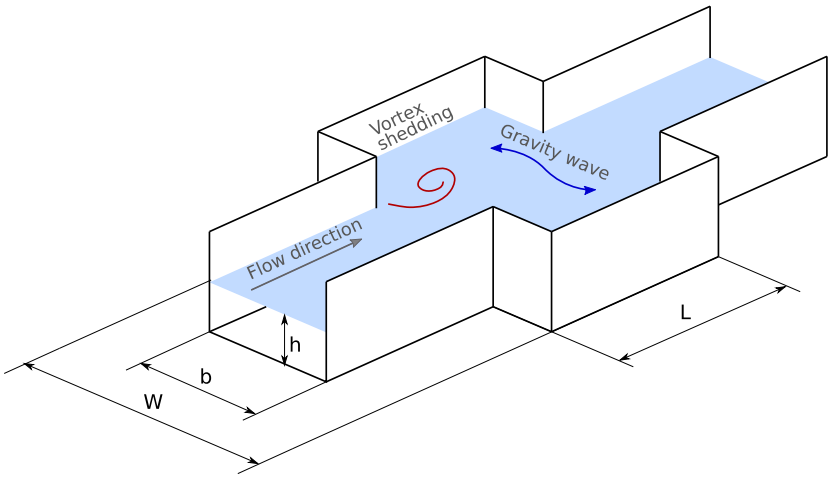


Figure 11.24: Representation of a sector of the channel with lateral cavities including the relevant geometric dimensions and flow features.

$$\hat{h}(y,t) = \hat{h}_0 \sin(\omega t - ky)\,, \tag{11.94}$$

where $\omega = 2\pi/T$ is the angular frequency, $k = 2\pi/\lambda$ is the wave number, λ is the wave length and T is the period. The following relation is satisfied

$$\omega = kc\,. \tag{11.95}$$

This work focuses on the formation of standing gravity waves, which are the origin of the seiche phenomenon. Standing waves can be constructed as the superposition of two travelling waves moving in opposite directions at the same celerity. Such waves are given by

$$\hat{h}_1(y,t) = \hat{h}_0 \sin(\omega t - ky)\,, \qquad \hat{h}_2(y,t) = \hat{h}_0 \sin(\omega t + ky) \tag{11.96}$$

and the resulting wave yields

$$\hat{h}_T(y,t) = \hat{h}_1(y,t) + \hat{h}_2(y,t) = -2\hat{h}_0 \sin(\omega t)\cos(ky)\,. \tag{11.97}$$

As the oscillation is free at the extrema (open tube analogy), an integer number of half-wavelengths must fit in the transverse length, W, hence

$$\lambda = \frac{2W}{n} \tag{11.98}$$

with n the harmonic number (1 for the fundamental). Combination of (11.95) and (11.98) yields the frequency of oscillation

$$f = \frac{nc}{2W}\,. \tag{11.99}$$

Channel configuration

The channel configuration considered in this thesis consists of a channel with lateral embayments where the seiche phenomenon appears. This configuration was studied by Juez et. al [143] and the experimental results obtained by the authors are used here for the evaluation of the mathematical model and numerical

scheme herein proposed.

The geometry of the channel including the relevant dimensions and the location of the 15 probes measuring the water depth are depicted in Figure 11.26. The cavity where experimental PIV measurements were carried out is highlighted in blue. The channel width is $W = 1$ m and $b = 0.5$ m. The slope of the channel is 0.1%. In the experiments, the flow was uniform, with $h = 0.05$ m and $Q = 8.5$ l/s.

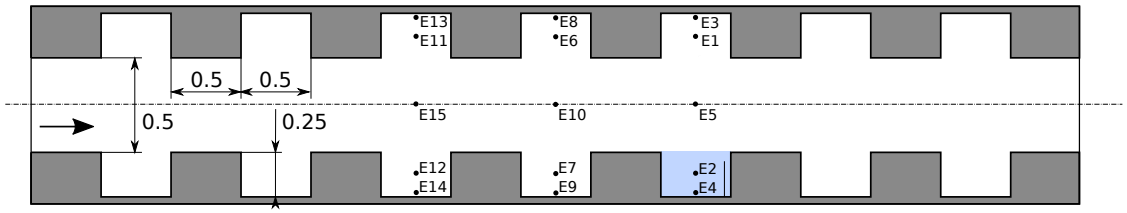


Figure 11.25: Top view representation of the channel configuration 2, including the relevant geometric dimensions and the location of the probes. The cavity used in the experimental measurements is highlighted in blue.

Assessment of the coupling frequency and identification of the seiche

The eigenfrequency (1-st mode) of the system is given by (11.99), yielding $f = 0.35$ Hz. In Figure 11.26, an schematic representation of a pair of symmetric cavities (top view on the left and cross sectional cut on the right) is presented. The location of the probes measuring the water depth, $E3$ and $E5$, is shown and the standing wave producing the seiche phenomenon is represented. Theoretically, the node of the wave should be located at the position of the probe $E5$ while the maximum amplitude of the wave should be found at $E3$. The peak-to-peak value for the water depth will be hereafter referred to as h_{pp}.

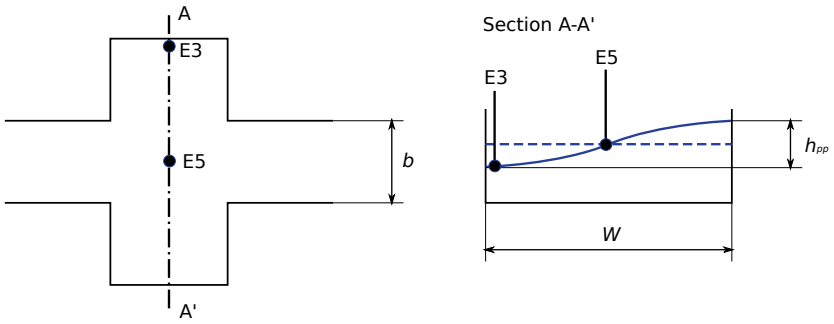


Figure 11.26: Schematic representation of a pair of symmetric cavities (top view on the left and cross sectional cut on the right).

In what follows, an accurate description of the seiche phenomenon is necessary. To identify the presence of the seiche, we propose to use the following statistical estimator

$$\gamma = \frac{h_{pp}(E3)}{h_{pp}(E5)} \tag{11.100}$$

where $h_{pp}(E3)$ and $h_{pp}(E5)$ are the estimated peak-to-peak amplitude of the water depth oscillation at probes E3 and E5, computed as

$$h_{pp}(E3) = 2\sqrt{2}\sigma(E3) \qquad h_{pp}(E5) = 2\sqrt{2}\sigma(E5) \tag{11.101}$$

with $\sigma(E3)$ and $\sigma(E5)$ the standard deviation of the data collected by probes E3 and E5 respectively. It is worth pointing out that the assumption of a perfect sinusoidal wave is used to estimate h_{pp}.

The greater γ is, the more likely a seiche oscillation is present. On the other hand, if $\gamma < 1$ the oscillation at the central node is of higher amplitude than at the extrema and a pure seiche oscillation cannot be present. However, this could be the case when the seiche phenomenon coexists with longitudinal travelling waves that make the amplitude at the central node increase. To unequivocally identify the presence of the seiche, we also know that (considering the 1-st mode of oscillation) there is a phase lag of half of a period (or π rad) between the amplitude at the extrema of two symmetric cavities when the seiche is present. The aforementioned criteria are herein used to assess the presence of the seiche.

Experimental results

The measured water depth at E3 and E5 for the *channel configuration 3.1* is depicted in Figure 11.27. The analysis of the data is complemented by computing the power spectrum of the signals using the Fast Fourier Transform (FFT). The FFT is applied to the signals collected by probes E3 and E5 and the results are plotted in Figure 11.28. It is observed that the frequency of the seiche is $f = 0.35$ Hz, which is in agreement with the theory.

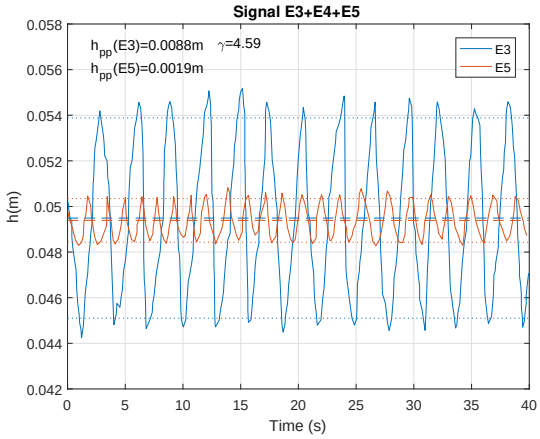


Figure 11.27: Measured water depth at E3 and E5.

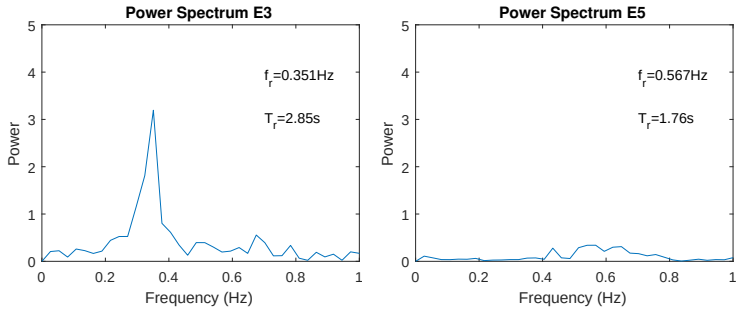


Figure 11.28: Power density spectrum for the measured signals at E3 and E5.

Computational results

Numerical results are computed using a 3-rd order WENO-ADER scheme in a Cartesian grid with cell size $\Delta x = 0.00625$ m and setting $CFL = 0.4$. The Manning coefficients for the wall and bed are $n = 0.03$ and $n = 0.01$ respectively.

In Figure 11.29, four snapshots showing the instantaneous velocity (left), vorticity (middle) and cross-sectional water depth (right), taken at four equally spaced phase positions $\theta = 0, \pi/2, \pi, 3\pi/2$ rad, are

presented. Two cross-sectional profiles at $x/L = 0.5$ and $x/L = 0.8$ are provided, namely $A-A'$ (red line) and $B-B'$ (blue line) respectively. The aim here is to illustrate the relationship between the transverse standing wave and the periodic shedding of vortices at the opening of the cavities. A coupling between the standing gravity wave within the cavities and the shedding of vortices at the opening of the cavities is observed.

On the first row of Figure 11.29, the solution in the beginning of the period ($\theta = 0$) is depicted. At this time, the water depth is maximum in the lower cavity and minimum in the upper cavity. The transverse flow is nil and the vortices are symmetrically located around the center of the cavities. There is a gradient of pressure in the transverse direction, which will trigger a transverse flow. When looking at phase $\theta = \pi/2$ (second row), it is observed that the water depth profile is virtually flat (with some local fluctuations due to the presence of the vortices). A transverse flow has been triggered by the gradient of pressure in the transverse direction previously observed. At this time, a new vortex is shed along the opening of the upper cavity. This makes the streamlines bend upwards and enhances a transverse flow in the direction of the upper cavity. At the same time, the vortex in the lower cavity, which is already developed, is close to the impingement corner and prevents the flow from entering in the cavity. The streamlines along the shear layer of the lower cavity are also bent upwards. After a half of a period (third row, $\theta = \pi$), the situation with no transverse flow is recovered. In this case, the maximum water depth is found in the upper cavity while the minimum water depth is found in the lower cavity. Again, there is a gradient of pressure in the transverse direction, which will again produce a transverse flow. After three quarters of a period (fourth row, $\theta = 3\pi/2$), the water surface is again flat and the transverse flow is now in the direction of the lower cavity.

A grid refinement analysis is carried out. The 3-rd order WENO-ADER scheme is used to compute the solution in 2 different grids with cell size $\Delta x = 0.008$ and $\Delta x = 0.00625$. The numerical solution for the water depth collected by probes E3, E4 and E5 is presented in Figure 11.30. The power density spectrum at E3 and E5 is presented in Figure 11.31.

The results evidence that when using a coarse grid, the numerical solution reproduces the experimental observations. When the grid is refined, the numerical solution becomes more disorganized, showing a higher noise in the probe E5. This is due to the low dissipation provided by the scheme when refining the grid. The mathematical model in (9.2) does not include any physical dissipation term and when using very fine grids, the numerical diffusion is virtually nil. In order to reproduce a physically feasible solution in a very fine grid, extra viscosity terms have to be added in the equations. This would allow to account for the physical dissipation produced by the low-scale vortical structures. The use of a turbulence model will be studied in a future work.

11.6 Concluding remarks

The highlights of this chapter are listed below:

- The ARL-ADER scheme, constructed as the combination between the LFS solver and the ARoe solver, has been extended to 2 space dimensions using Cartesian meshes.
- The proposed method has been applied to the 2D SWE with bed elevation, friction and Coriolis. For such model, the relevant equilibrium states considered in the design of the scheme are the quiescent equilibrium (lake at rest) and the geostrophic equilibrium (jet in the rotating frame).
- To ensure the geostrophic equilibrium, the Coriolis source term has been rewritten as a geometric source term, which allows its combination with the bed slope source term. This leads to a single geometric source where the scalar variable can be regarded as an apparent topography. The friction source term is integrated by means of a traditional Gaussian quadrature rule.
- Generally, we would proceed as follows for the integration of the source term:

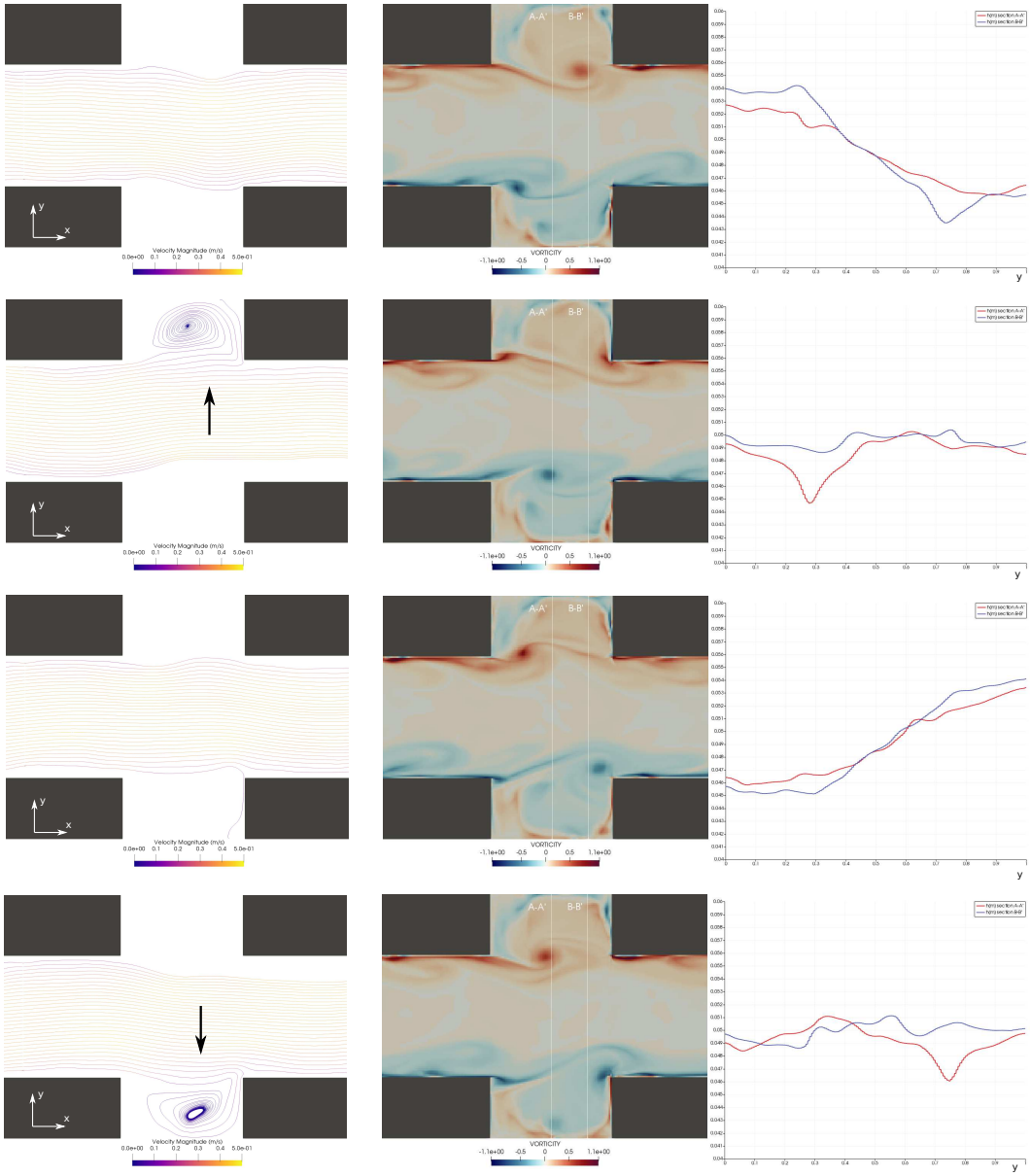


Figure 11.29: Numerically computed instantaneous velocity (left), vorticity (middle) and cross-sectional water depth (right), taken at four equally spaced phase positions $\theta = 0$ (first row), $\theta = \pi/2$ (second row) ,$\theta = \pi$ (third row) and $\theta = 3\pi/2$ rad (fourth row).

  - If the source term is geometric: must be included in the definition of the DRP and integrated inside cells using a particular well-balanced discretization that can be extended to high order via Romberg integration. Example: bed elevation.
  - If the source term is non-geometric but we want to enforce a certain equilibrium state: must be included in the definition of the DRP and integrated inside cells using a particular well-balanced discretization that can be extended to high order via Romberg integration. Example: Coriolis (if requiring geostrophic equilibrium), friction (if requiring bed slope equal to friction slope).
  - If the source term is non-geometric: must not be included in the definition of the DRP and can be integrated inside cells using traditional Gaussian quadrature. Example: Coriolis, friction, wind forces, etc.

- The numerical discretization of geometric source terms (natural or enforced) at cell interfaces has

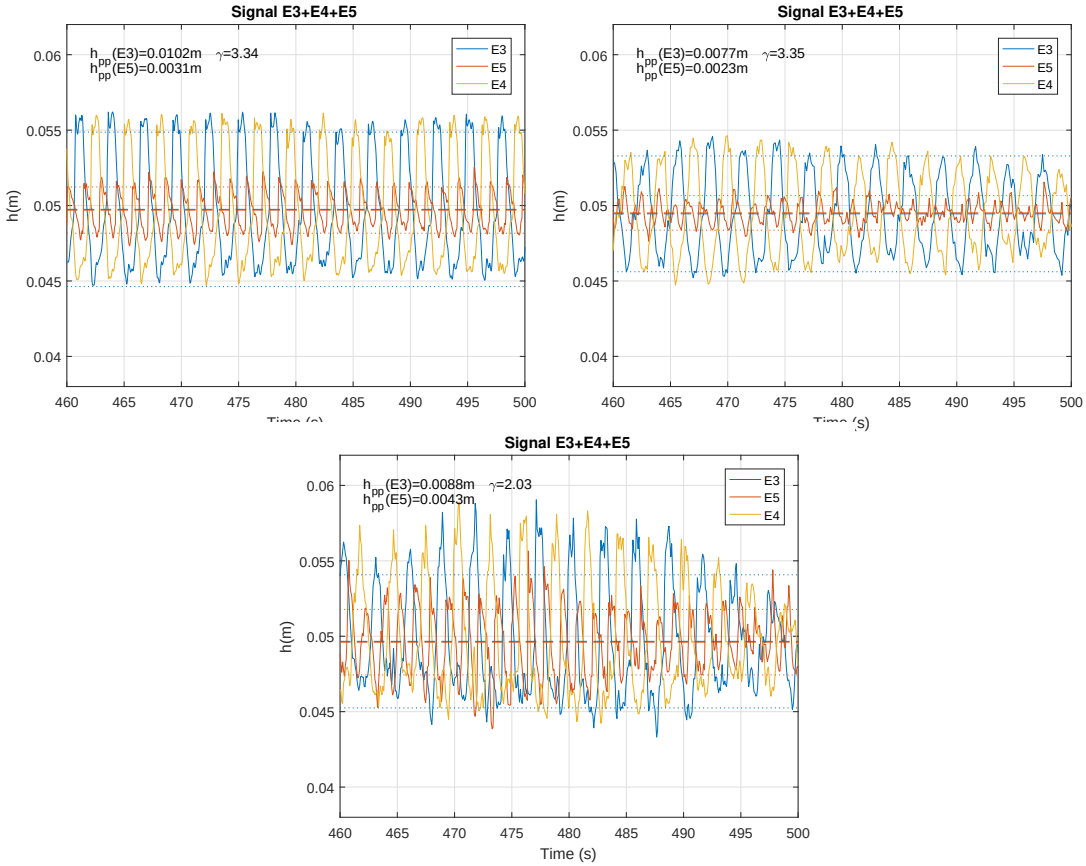


Figure 11.30: Numerical solution for the water depth at E3, E4 and E5 computed in grids with cell size $\Delta x = 0.008$ (top-left), $\Delta x = 0.00625$ (top-right) and $\Delta x = 0.005$ (bottom).

been done as proposed in the previous chapters (the DF has been used). Inside cells, geometric source terms also have to be exactly balanced with flux fluctuations. To ensure such balance in 2D, the problem has been reduced to two 1D problems in the Cartesian directions. This allows to derive a 2D arbitrary order approximation of the integral of the source term by means of the combination of 1D arbitrary order integrals using Gaussian quadrature and Romberg integration. The keystone for the preservation of equilibrium with very high order is the use of the Romberg integration method, which allows to extend the well–balanced discretization of the source term to arbitrary order. As a result, the proposed scheme is able to preserve the steady states of relevance while retaining a high order of accuracy for the resolution of transient wave propagation.

- The convergence rate of the solution has been experimentally assessed. It is observed that the scheme achieves the expected accuracy when considering bed variation and/or friction. On the other hand, when including the Coriolis source term, a suboptimal behavior is noticed. The reason for this could be that the integration in time when considering Coriolis is not optimally done. This underscores the importance of an optimal derivation of time derivatives and shows that when considering complex source terms, such as the geometric reinterpretation of the Coriolis source term, the CK procedure may become rather cumbersome and other techniques may be worth being used [33].

- The numerical results evidence that high order schemes are not only recommended, due to the improved efficiency of the methods, but sometimes necessary when the first order schemes are not able to reproduce the physical solution (Section 11.5.12). The simulation tools proposed in this thesis provide an appropriate balance between computational efficiency and complexity of implementation. They are able to reproduce a broad variety of flows dominated by source terms by using a

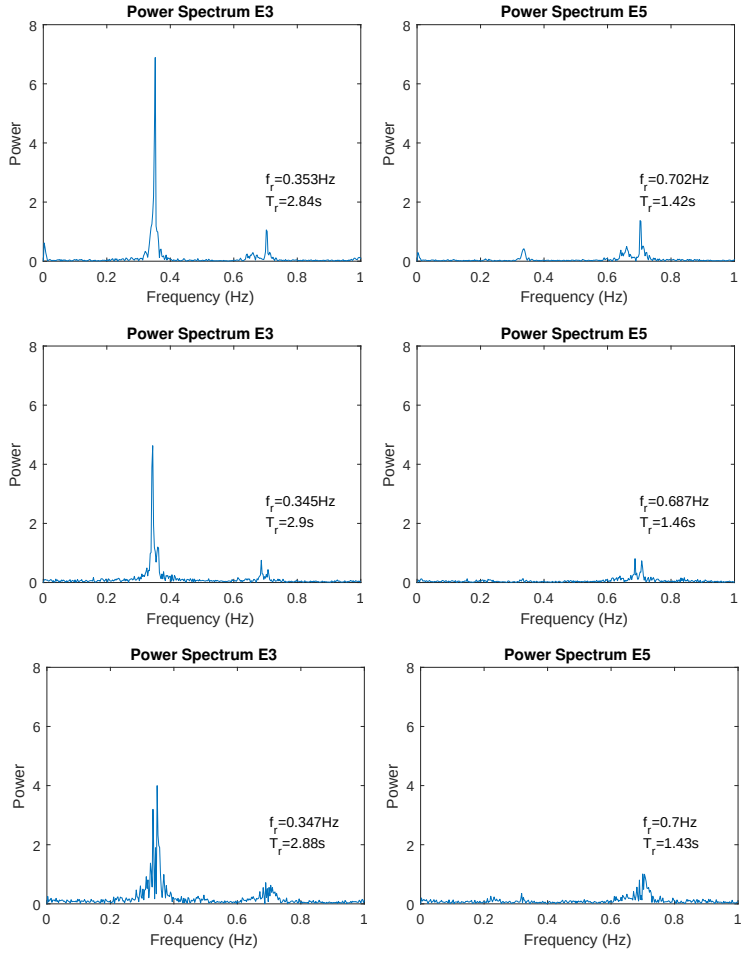


Figure 11.31: Power density spectrum for grids with cell size $\Delta x = 0.008$ (first row), $\Delta x = 0.00625$ (second row) and $\Delta x = 0.005$ (bottom).

unified strategy based on the consideration of geometric source terms that allow to satisfy certain equilibrium conditions. The schemes use Cartesian mesh, but no mesh–dependency of the solution is observed due to the high order of accuracy. The scheme can be easily extended to other systems of equations with source terms of a broad variety.

- The mathematical model of the SWE used in this thesis may not be sufficiently representative of the physical phenomena governing certain flows. This does not depend on the choice of numerical scheme used for the integration of the equations. This is the case of the free surface steady flow in the channel with lateral cavities in Section 11.5.14, where a turbulence model (diffusion term) should be added to the SWE in order to converge to the experimental observations.

12 NUMERICAL SHOCKWAVE ANOMALIES

In this thesis, a special emphasis has been put on high order schemes, which provide a good balance between accuracy and computational cost when compared to lower order schemes. However, a high order of accuracy also preserves the effect of undesirable numerical anomalies. The utilization of such schemes in presence of spurious oscillations prevents from the dissipation of such oscillations, as high order schemes have a very reduced numerical diffusion. It has been widely reported in the literature that significant numerical anomalies arise in presence of shock waves. An example of such problems are the Carbuncle [99, 100], the slowly-moving shock [101, 102] and the wall-heating phenomenon [103], all of them leading to spurious numerical solutions. In this chapter, we focus on the slowly-moving shock problem.

Shockwaves are typical solutions for nonlinear hyperbolic systems of conservation laws and their numerical treatment is of utmost importance to provide accurate solutions. As mentioned by Zaide and Roe [113], physical shockwaves have a finite width which is determined by the physical dissipation processes, however, when considering numerical shockwaves, a numerical width, usually much greater than the physical width, is enforced. This leads to the appearance of intermediate states which cannot be given a direct physical interpretation. Such states cannot be removed even when refining the grid, therefore a special emphasis must be put on this issue when designing a numerical scheme. Up to the present time, most studies have been carried out in the framework of Euler equations. In this work, we will extend those techniques to the SWE.

Some of the problems related to numerical shockwave anomalies were first identified by Cameron and Emery [104, 105], who proposed some improvements based on the addition of artificial viscosity and modification of the grid. Here, we focus on the slowly-moving shock problem, which is associated to hydraulic jumps in the SWE. The slowly-moving shock problem was first investigated by Roberts in [101], who defined it as numerical noise generated in the discrete shock transition layer which is transported downstream. Such noise will be hereafter referred to as post-shock oscillations. In [101], the schemes of Godunov, Roe, and Osher were examined and the source of this error as also provided by using the Hugoniot locus. It was also observed that the slowly-moving shock problem only appears for systems of equations and not for scalar equations, where such schemes perform correctly. It is worth pointing out that even for non-linear systems, the slowly-moving shock problem does not appear if the Hugoniot curves are linear [106], as happens in the system in [144]. Later on, Arora and Roe [102] carried out a thorough study on this problem and evidenced that it can be ruinous when, for instance, making calculations of shock-sound interaction.

One could think that by increasing the order of accuracy of the scheme, the slowly-moving shock problem could be circumvented. However, as mentioned by other authors [106, 107, 108], the slowly-moving shock problem will only be accentuated when increasing the accuracy of the scheme. Such an increase of accuracy will be translated into a longer preservation of post-shock oscillations as they provide a better resolution of

the spurious physics. When using a high order scheme, the order is reduced to first order in the vicinity of the shock and the numerical solution within this region will behave according to what is expected from a first order scheme [109, 110]. Away from the shock, the order of accuracy is higher and therefore the spurious oscillations will be better resolved, preventing them from vanishing as one would desire. It must be borne in mind that even when using high order interpolations with limiting techniques, such as Total Variation Diminishing (TVD) interpolations and Essentially Non-oscillatory (ENO) schemes, the slowly-moving shock problem is accentuated [108].

The slowly-moving shock problem has been deeply studied for homogeneous systems of equations (e.g. the Euler equations) but scarcely studied for systems dominated by source terms. In [108], numerical results for the computation of a 1D compressible flow through a divergent nozzle by means of different first and high order schemes were presented, showing the inability of all schemes to converge to the exact solution in presence of shocks. This is the slowly-moving shock problem in the limit when shock speed is nil. The SWE are analogous to the 1D compressible flow with varying area and the slowly-moving shock problem also appears, as broadly reported in the literature.

Here we focus on the slowly-moving shock problem in the SWE. To this end, we identify the conditions for the aforementioned problem to appear by studying the Hugoniot locus of the SWE and by seeking slowly-moving shock waves. We notice that they are only produced when dealing with a kind of transcritical shocks called hydraulic jumps, characterized by a change of sign of the relevant eigenvalue across them. A complete description of such kind of waves is provided and a thorough study on the shock structure, comparing exact and Godunov type solutions, is carried out in phase space. The slowly-moving shock problem in the SWE is a well-known problem in the scientific community, characterized by a spike in the discharge at the cell where the hydraulic jump is contained. In fact, it seems that this problem is more a feature than a problem when considering steady solutions of the SWE containing hydraulic jumps. The presence of the spurious spike in the discharge has been taken for granted as it does not perturb the rest of the solution. However, when considering transient cases, it produces a very undesirable shedding of oscillations downstream that should be avoided.

When designing numerical schemes for the computation of slowly-moving shocks, the addition of extra artificial viscosity seems to be the most preferred technique in the scientific community [104, 105, 101, 102, 107, 111, 112]. If we want to avoid extra diffusion, another possibility is the use of a flux interpolation method, which avoids using the evaluation of the physical fluxes in the untrustworthy intermediate cells corresponding to the shock discontinuity. This idea of flux interpolation was first presented by Zaide and Roe [113], who proposed to find the fluxes in the intermediate cells by extrapolation from trustworthy neighbors. The authors claim that, by enforcing a linear shock structure and unambiguous sub-cell shock position, numerical shockwave anomalies are dramatically reduced. It could be said that their method is also based on the addition of artificial viscosity, as the flux interpolation functions can be regarded as the traditional Roe flux plus a viscosity term. The difference is that such flux functions use dissipation to control the shock structure rather than to approach the true viscous solution and therefore they do not expand the shock profile [106].

In this chapter, we use the approach in [113] to propose a novel spike-reducing flux function for the SWE with varying bed. Such function is designed to satisfy certain properties of convergence and consistence. Prior to the presentation of the proposed technique, the flux functions in [106] are applied to the SWE with flat bed, showing their spike-reducing nature. The proposed technique is assessed in a variety of situations, including steady and transient cases, with continuous and discontinuous bed profiles, proving the expected spike-reducing behavior. The analogous SWE problem of the 1D nozzle problem in [108], which is the steady flow over a hump, is reproduced in this work, showing that the proposed scheme leads to a convergent solution, even when measured with L_∞ error norm. The proposed techniques are also extended to 2 space dimensions.

The chapter is structured as follows. In Section 12.1, a theoretical study on the hydraulic jump and the presence of the spike of discharge is presented. In Section 12.2, the flux fixes proposed in [106] are recalled and a novel flux function that accounts for the presence of the source term is proposed and assessed in a

variety of cases involving the resolution of steady and transient jumps. In Section 12.3, such techniques are extended to 2 space dimensions and are validated with two test cases that involve the resolution of hydraulic jumps (bow shocks) around 2D obstacles.

12.1 Numerical shockwave anomalies in the SWE: the hydraulic jump

It has been widely reported in the literature that significant numerical anomalies arise in presence of shock waves. An example of such problems are the Carbuncle, the slowly-moving shock and the wall-heating phenomenon, all of them leading to spurious numerical solutions. The aforementioned problems have been deeply studied in the framework of Euler equations and some authors have proposed different numerical techniques to address them. Here, we will focus on the numerical anomalies present when computing steady and moving hydraulic jumps, which are a particular type of shock waves in the framework of the Shallow Water Equations (SWE). Specifically, our interest lies in the reduction of the spike in the discharge, reported in the previous section.

The hydraulic jump occurs when a supercritical flow suddenly changes to subcritical conditions, generating a steep free surface elevation where intense mixing takes place and a large amount of mechanical energy is dissipated. Mathematically, hydraulic jumps are modelled by a discontinuity corresponding to a shock wave and the relation between the states at each side of the discontinuity is provided by the RH conditions.

12.1.1 Hugoniot locus of the hydraulic jump

To understand the mathematical treatment of the hydraulic jump and the numerical anomalies arising from such a wave, it is worth studying first the analytical solution of this type of wave under the simplest conditions, that is over flat bed. From Rankine-Hugoniot (RH) conditions, all possible values connecting the left and right states can be determined and represented in phase space as $(h(\xi), hu(\xi))$ by means of the so-called Hugoniot locus

$$\mathbf{U}(\xi) = \begin{pmatrix} h(\xi) \\ hu(\xi) \end{pmatrix} = \begin{pmatrix} h_L + \xi \\ (hu)_L + \xi\left(u_L \pm \sqrt{g h_L + \frac{1}{2} g \xi \left(3 + \frac{\xi}{h_L}\right)}\right) \end{pmatrix}, \tag{12.1}$$

where $\xi = h - h_L$, with h the independent variable used for the parametrization. From (12.1), we notice that two families of curves are possible, denoted by Ψ^1 and Ψ^2, which are associated to the 1-wave and 2-wave respectively. Such curves are defined by

$$\Psi^1(\xi) = \begin{pmatrix} \psi_1^1(\xi) \\ \psi_2^1(\xi) \end{pmatrix} = \begin{pmatrix} h_L + \xi \\ (hu)_L + \xi\left(u_L - \sqrt{g h_L + \frac{1}{2} g \xi \left(3 + \frac{\xi}{h_L}\right)}\right) \end{pmatrix}, \tag{12.2}$$

$$\Psi^2(\xi) = \begin{pmatrix} \psi_1^2(\xi) \\ \psi_2^2(\xi) \end{pmatrix} = \begin{pmatrix} h_L + \xi \\ (hu)_L + \xi\left(u_L + \sqrt{g h_L + \frac{1}{2} g \xi \left(3 + \frac{\xi}{h_L}\right)}\right) \end{pmatrix}. \tag{12.3}$$

It must be borne in mind that not every choice of subcritical state that is connected to a given supercritical state represents a hydraulic jump. For instance, let us consider a left supercritical state given by $h_L = 0.85$ and $hu_L = 3.411764705882353$ and let us find two possible right states connected to it, each of them laying on each branch of the Hugoniot locus. This is depicted in Figure 12.1, where the original left state is denoted by F, the right state lying on the 1-curve, Ψ^1, is denoted by G and the right state lying on the 2-curve, Ψ^2, is denoted by J. The entropically inadmissible region of the curves has been represented by dashed line. It is observed that both G and J lie on the subcritical region of the phase plane and they

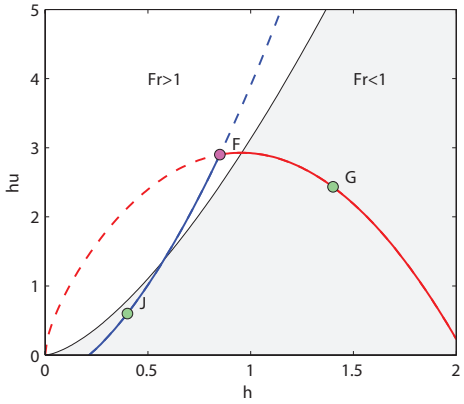


Figure 12.1: Phase space $(h, hu) \in \mathbb{R}^+ \times \mathbb{R}^+$ with the subcritical region depicted in green background and the supercritical region in white background, showing the Hugoniot locus Ψ^1 in red and Ψ^2 in blue.

are both entropically admissible, however, only the combination of states F–G leads to a hydraulic jump, because G, unlike J, has a higher water depth than F and, what is decisive in this case, wave celerities of F and G have opposite sign. More generally, we can define an hydraulic jump as:

Definition 10 *(Hydraulic jump). Let the following discontinuous solution*

$$\mathbf{U}(x,t) = \begin{cases} (h,hu)_L & x < 0 \\ (h,hu)_R & x > 0 \end{cases} \tag{12.4}$$

be a weak solution of the SWE system, where $(h,hu)_L$ and $(h,hu)_R$ are two different states laying on Ψ^m and satisfying the entropy condition $\lambda^m(\mathbf{U}_L) > \mathscr{S}^m > \lambda^m(\mathbf{U}_R)$, with $\mathscr{S}^m$ the speed of the jump, that undergoes a flow transition as $Fr_L < 1 < Fr_R$ or $Fr_R < 1 < Fr_L$. Solution in (12.4) is termed as hydraulic jump if and only if $\lambda^m(\mathbf{U}_L) > 0 > \lambda^m(\mathbf{U}_R)$.

Notice that, according to the previous definition, hydraulic jumps admit that $\mathscr{S}^m$ be nil, hence they are the only shock-type solution for the SWE that can be stationary at a fixed position.

In Figure 12.2 the Hugoniot locus Ψ^1 in red and Ψ^2 in blue for the left state $(h,hu)_L = (0.5, 3)$ is depicted, showing three possible solutions in the form of a hydraulic jump: a steady jump (top-right), a right-moving jump (bottom-left) and a left-moving jump (bottom-right). The speed of the jump, $\mathscr{S}$, is the slope of the straight line depicted in magenta.

12.1.2 Analytical study and comparison of the exact solution for 2 and 3-states hydraulic jumps.

Prior to analyzing the numerical solutions of Godunov's scheme to the hydraulic jump, it is worth studying the analytical solutions to this problem, which will help to understand the nature and characteristics of the numerical (discrete) solution to it. It is well known that an intermediate state appears in the numerical solution provided by Godunov's scheme, with independence of the solver [113]. The presence of this intermediate state, hereafter denoted by $\mathbf{U}_M$, is not of any physical relevance as it provides an unrealistic estimation of the average discharge in the intermediate cell (spike) which does not match the constant value of discharge. However, when using conservative schemes the intermediate value may be useful to compute a rough estimate of the shock position. The position of the shock inside the cell can be computed imposing conservation of mass as

$$x_S = \frac{h_M - h_R}{h_L - h_R}, \tag{12.5}$$

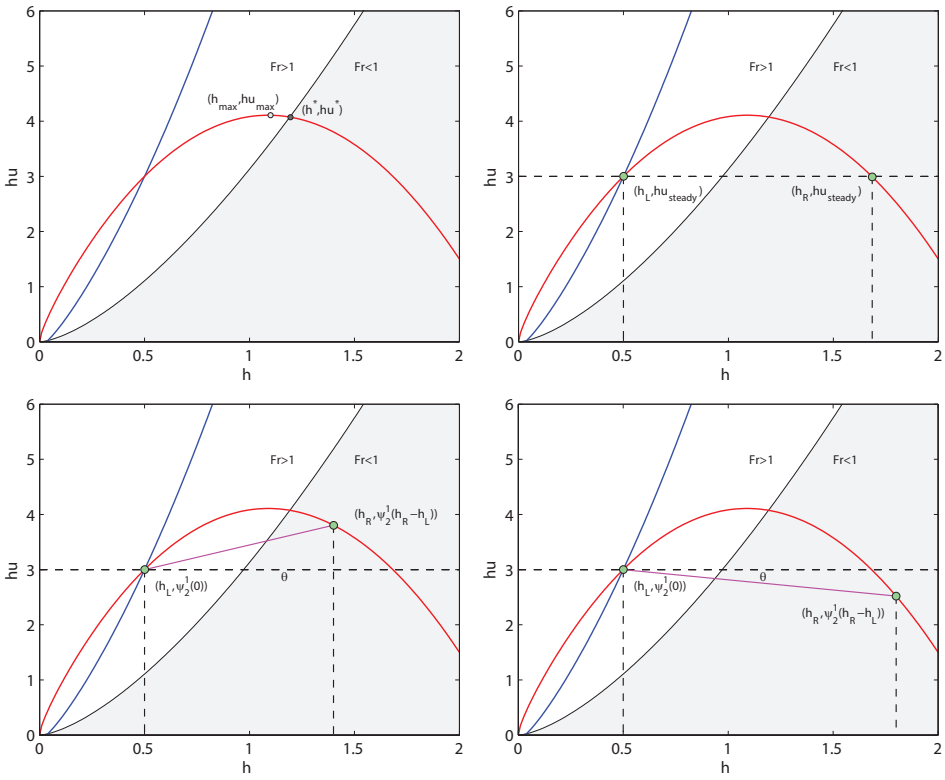


Figure 12.2: Hugoniot locus Ψ^1 in red and Ψ^2 in blue for the left state $(h, hu) = (0.5, 3)$, showing three possible solutions in the form of a hydraulic jump: a steady jump (top-right), a right-moving jump (bottom-left) and a left-moving jump (bottom-right).

where $x_S \in [0,1]$ represents the normalized position of the shock (where 0≡left interface, 0.5≡middle position and 1≡right interface) [113].

As a first approach and before getting into the numerical issues concerning hydraulic jumps, let us compare analytically the solution for the ideal steady hydraulic jump (pure discontinuity) with another solution for the steady hydraulic jump that includes an intermediate state, which resembles the discrete solution provided by Godunov's scheme. Both solutions are weak solutions of the equations and they are both valid. Whereas the former is characterized by two states, namely $\mathbf{U}_L$ and $\mathbf{U}_R$, the latter is given by $\mathbf{U}_L$, $\mathbf{U}_M$ and $\mathbf{U}_R$. Moreover, the latter does not experience a sudden transition of flow regime, hence it cannot be considered a pure, or ideal, hydraulic jump.

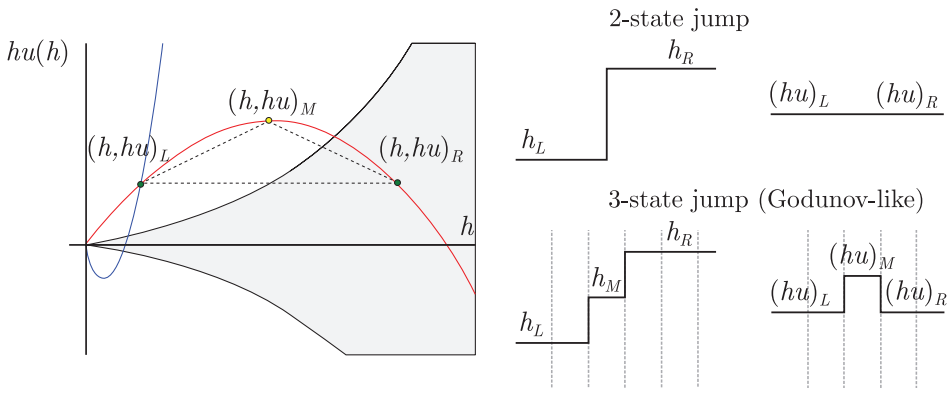


Figure 12.3: Hugoniot Locus and sketch of the analytical solutions for a 2-state and 3-state hydraulic jumps.

Let us consider first the ideal hydraulic jump composed of two states. This solution consists of a supercritical right-moving steady flow that suddenly decelerates through a pure discontinuity to subcritical conditions, as depicted schematically in Figure 12.3 (top-right). The Hugoniot locus that connects the left and right states of the jump, Ψ^1, is depicted in Figure 12.3 (left), showing that such states are located at the intersection of the Hugoniot Locus with the curve of constant discharge $(hu)_L = (hu)_R$, ensuring the steady regime.

On the other hand, when seeking a weak solution of the equations that includes an intermediate state, $\mathbf{U}_M$, as depicted in Figure 12.3 (bottom-right), we need to look for this additional state on the Hugoniot curve. According to Figure 12.3 (left), the intermediate state $(h_M, (hu)_M)$ (yellow point) will lie on Hugoniot Locus and is connected to the left and right states (green points) through this curve. From the previous observations, we realize that only a linear Hugoniot Locus would ensure a constant discharge in the intermediate state [113].

If a curve of the family of

$$\breve{\Psi}(\xi) = \begin{pmatrix} h(\xi) \\ (hu)_{steady} \end{pmatrix} \tag{12.6}$$

was considered in state space, with $(hu)_{steady} \in \mathbb{R}^+$ for a right-moving flow, a constant discharge for the intermediate state would be possible. Only if Ψ^1 was of the type of $\breve{\Psi}$, constant discharge would be ensured across the intermediate cell. This means that we would have a linear Hugoniot [113]. This concept can be extended to moving hydraulic jumps by examination of Figure 12.2 (bottom left). Let us redefine the states denoted in the plot by (h', hu') and (h'', hu'') as left state (h_L, hu_L) and right state (h_R, hu_R), respectively. The linear Hugoniot must lie on the line depicted in magenta, with slope $\theta = (h_R - h_L)/(hu_R - hu_R)$ and can be parametrized in terms of x_S in (12.5). Hence, it can be expressed as

$$\breve{\Psi}(x_S) = \begin{pmatrix} h(x_S) \\ hu(x_S) \end{pmatrix}, \tag{12.7}$$

where $h(x_S) = x_S(h_R - h_L) + h_L$,

$$hu(x_S) = hu_L + \theta h(x_S) \tag{12.8}$$

and $x_S \in [0, 1]$. Note that parametrization $\breve{\Psi}(\xi)$ is straighfoward as $\xi = (h_R - h_L)x_S$.

Considering again the steady case described above and depicted in Figure 12.3, we can observe that the exact Hugoniot is neither linear nor monotone and ψ_2^1 has a global maxima hu_{max} at $h_{max} \in [h_L, h_R] \subset \mathbb{R}^+$ therefore, for any $h_M \in [h_L, h_R] \subset \mathbb{R}^+$, we have that $(hu)_M \geq (hu)_L = (hu)_R \equiv (hu)_{steady}$. This can be observed in Figure 12.3 (bottom-right), where a spike in the discharge appears.

12.1.3 Properties of the intermediate state in discrete Godunov-type solutions

Up to this point throughout this section, we have only considered exact solutions to the hydraulic jump. Theoretically, when considering the exact solution, the presence of an intermediate constant state $\mathbf{U}_M = (h_M, (hu)_M)$ is not stable, that is, it cannot be kept under steady conditions. The reason for this is that both jumps (left to middle and middle to right) have non-zero wave velocities of opposite sign, hence both jumps would converge to form a unique jump. This behavior, shown in Figure 12.4, is only present in the exact solution. On the other hand, when considering a discrete solution in a computational grid, both waves could be kept at a stationary position (at the cell interfaces of the intermediate cell) and the intermediate cell could keep the intermediate value in the steady regime. The reason for this is that the overall flux fluctuation inside the cell is nil and the numerical fluxes are equal to the left and right fluxes

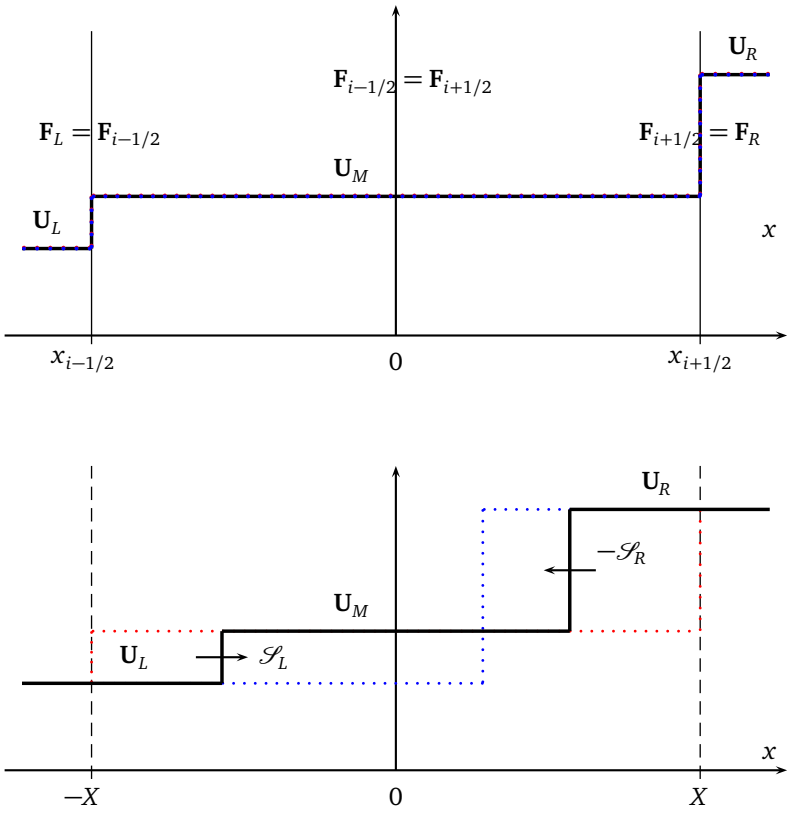


Figure 12.4: Initial condition considering an intermediate state (red), transient evolution of the discontinuities $\mathbf{U}_L$-$\mathbf{U}_M$ and $\mathbf{U}_M$-$\mathbf{U}_R$ (black) and final steady solution (blue).

$$\delta\mathbf{M}^{-}_{i+1/2} + \delta\mathbf{M}^{+}_{i-1/2} = 0\,, \quad \mathbf{F}^{-}_{i+1/2} = \mathbf{F}^{+}_{i-1/2} = \mathbf{F}_R = \mathbf{F}_L \tag{12.9}$$

when considering the numerical resolution of the problem by means of FV Godunov's scheme in (5.11). On the other hand, the physical flux evaluated at the intermediate state, $\mathbf{F}_M$, hereafter referred to as equilibrium flux, does not match the left and right fluxes

$$\mathbf{F}_M = \mathbf{F}(\mathbf{U}_M)\,, \quad \mathbf{F}_M \neq \mathbf{F}_L = \mathbf{F}_R\,. \tag{12.10}$$

Figure 12.4 depicts the contrasting behavior of the 3-state hydraulic jump when considering the discrete (top) and exact (bottom) solution. The initial condition is represented by red dotted line, the final solution (when steadiness is achieved) is represented by blue dotted line and the solution at an arbitrary time before reaching the steady state is represented by black solid line. It can be observed that the initial condition is maintained in the discrete solution, where the intermediate state, $\mathbf{U}_M$, has been defined inside the cell $[x_{i-1/2}, x_{i+1/2}]$.

We are going to show in detail how the discrete equilibrium is achieved in the solution provided by the Riemann solver. Source terms are considered nil, hence we work with homogeneous fluxes

$$\mathbf{F}^{*}_{i+1/2} \equiv \mathbf{F}^{+}_{i+1/2} = \mathbf{F}^{-}_{i+1/2} \qquad \mathbf{F}^{*}_{i-1/2} \equiv \mathbf{F}^{+}_{i-1/2} = \mathbf{F}^{-}_{i-1/2}\,. \tag{12.11}$$

When considering RPs between states $\mathbf{U}_L$, $\mathbf{U}_M$ and $\mathbf{U}_M$, $\mathbf{U}_R$, defined at the interfaces of the intermediate cell, it is noticed that the left-middle RP, hereafter RP_{LM}, connects two supercritical states while the middle-

right RP, RP_{MR}, is subcritical. When using the HLL solver under the aforementioned hypotheses, we know that $\mathbf{F}^*_{i_M-1/2} = \mathbf{F}_L$ as the flow is supercritical. Hence, we only require that

$$\mathbf{F}^*_{i_M+1/2} = \mathbf{F}_R \tag{12.12}$$

to ensure a steady regime, where i_M is the cell index for the intermediate cell. This can be proven by noticing that (12.9) is satisfied. To impose (12.12), the RH condition across the subcritical (left-moving) λ_1 wave at $x_{i_M+1/2}$ is considered

$$\mathbf{F}_M = \mathbf{F}_R - \lambda^1 (\mathbf{U}_R - \mathbf{U}_M) \,. \tag{12.13}$$

When applied to the SWE, we choose the Roe wave velocities $\lambda^1 = \widetilde{u} - \widetilde{c}$ and $\lambda^2 = \widetilde{u} + \widetilde{c}$ and (12.13) becomes

$$\begin{cases} (hu)_M = & (hu)_R - \lambda^1 (h_R - h_M) \\ \left(\frac{1}{2} g h^2 + hu^2\right)_M = & \left(\frac{1}{2} g h^2 + hu^2\right)_R - \lambda^1 ((hu)_R - (hu)_M) \end{cases} \tag{12.14}$$

which provides the equilibrium intermediate state that satisfies (12.9) when using the Roe and HLL solver (with Roe celerities).

It is possible to find a curve in the phase space for all pairs h_M, $(hu)_M$ satisfying (12.14), for a given right state. Such a curve is analogous to the analytical Hugoniot locus previously presented, Ψ^1, with the difference that now the curve does not satisfy the exact relations but those particular relations given by the approximate solver. Only when using an exact solver both curves coincide. Figure 12.5 shows the exact Hugoniot locus in red and the approximate locus for the approximate Riemann solver in purple, when considering Roe celerities. It is observed that both curves have the same tendency and share two points (apart from the origin) in the phase space: the left and the right states. This implies that only when the intermediate state coincides with the left or right states, the approximate solver would provide the exact solution. Hence, only when the shock position is located at the interface, the approximate solver provides the exact solution [145, 146]. A exhaustive comparison of the numerical performance in shock-capturing of different flux functions in the framework of Euler equations can be found in [147].

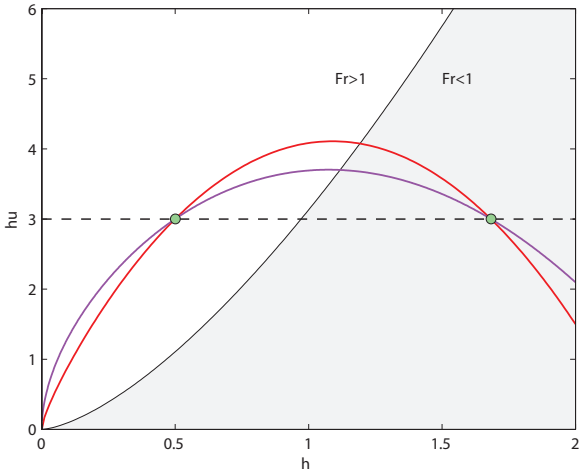


Figure 12.5: Exact Hugoniot locus (red) and approximate locus for the Riemann solver (purple) that connect the left and right states.

A numerical experiment is proposed to show the aforementioned affirmations. We have chosen a computational domain $[0, 450]$, discretized in 900 computational cells.

Test case	h_L	$(hu)_L$	h_M	$(hu)_M$	h_R	$(hu)_R$
A	0.5	3	1.6828655896	3	1.6828655896	3
B	0.5	3	0.8793959880	3.6256825146	1.6828655896	3
C	0.5	3	0.8793959880	3.9682712891	1.6828655896	3

Table 12.1: Initial intermediate state values for the three test cases proposed in this section.

The initial conditions for the three numerical tests are listed in Table 12.1. For the sake of clarity, the intermediate state $\mathbf{U}_M$ is represented in the phase space in Figure 12.6 for case A (green), B (yellow) and C (blue). Initial conditions for h and hu are constructed as follows

$$\mathbf{U}(x,0)\begin{cases}\mathbf{U}_L & \text{if} \quad 0 \leq x < x_{i_M-1/2} \\ \mathbf{U}_M & \text{if} \quad x_{i_M-1/2} \leq x \leq x_{i_M+1/2} \\ \mathbf{U}_R & \text{if} \quad x_{i_M-1/2} < x \leq 450\end{cases} \tag{12.15}$$

where i_M is the intermediate cell and is set to 451, hence $x_{i_M-1/2} = 225$.

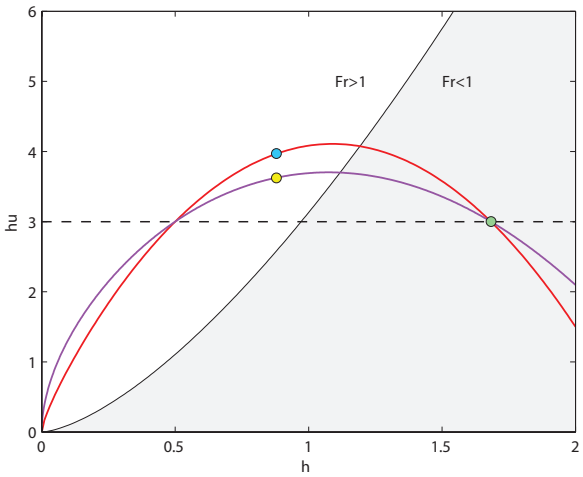


Figure 12.6: Intermediate state $\mathbf{U}_M$ depicted for case A (green), B (yellow) and C (blue). The Hugoniot locus is represented in red and the locus for the HLL solver in purple.

Figure 12.7 depicts the numerical solution provided by the HLL scheme (Roe celerities) for the cell averages (dashed line). The internal structure of the solution provided by the solver is also depicted in the figure with continuous line. The width of the internal states, that is to say, of the star region, is considered constant. Numerical results for test cases A, B and C are depicted in Figure 12.7 top, middle and bottom, respectively. It is observed that only for test cases A and B, the internal structure of the solution inside the intermediate cell matches the solution on the left and right cells, hence the solution keeps the steady state. However, when considering the test case C, it is observed that the star solution on the right subcritical RP does not match the right value, hence equilibrium is not maintained.

12.2 Flux fixes for the computation of the hydraulic jump

In this section, some spike-reduction numerical techniques based on flux interpolation are recalled and applied to the Shallow Water Equations (SWE). This idea of flux interpolation was first presented by Zaide and Roe [113], who proposed to find the fluxes in the untrustworthy intermediate cells by extrapolation from trustworthy neighbors and presented two new flux functions. The first one, named by the authors

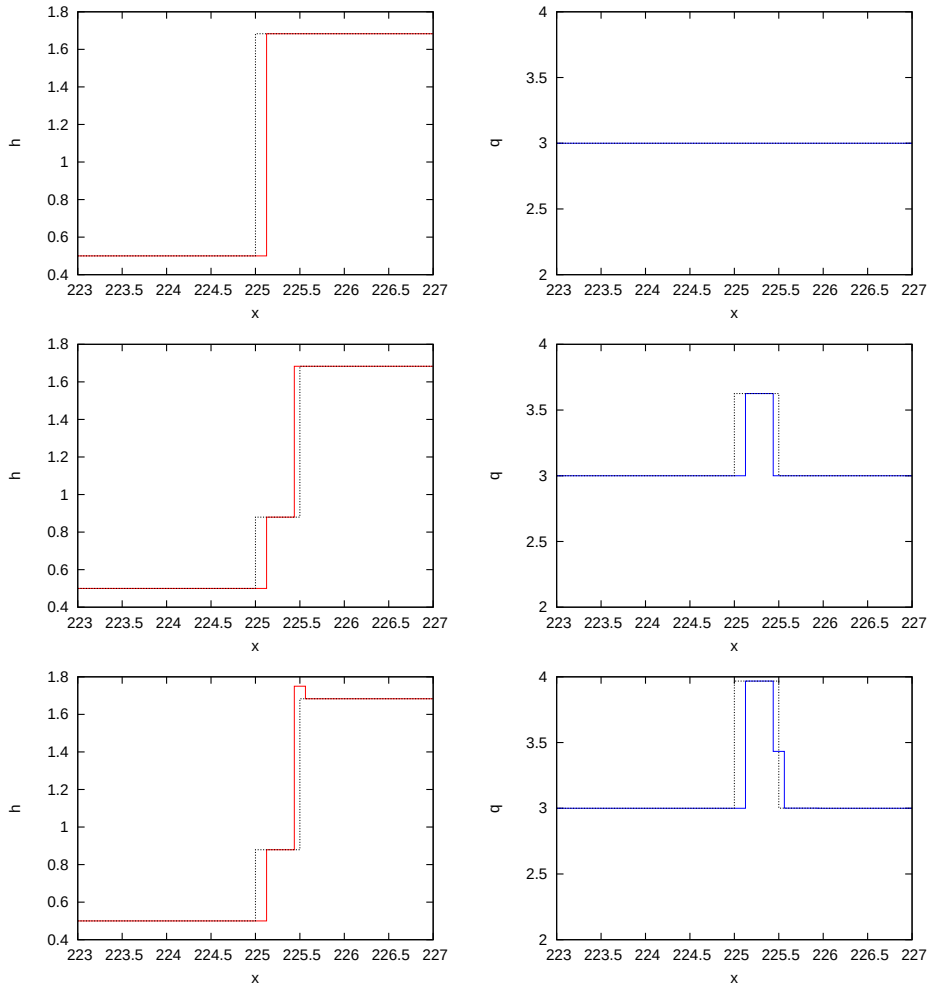


Figure 12.7: Numerical solution depicted as cell averages (dashed line) and showing the internal structure (continuous line) for h (left) and hu (right) provided by the HLL solver after one time step for cases A (top row), B (middle row) and C (bottom row).

flux function A, was constructed based on the flux-wave approach, by computing the fluctuations in the interpolated fluxes across each wave. The second one, called flux function B, is based on the classical Roe solver and relies on conserved variables to determine the jumps across each wave and the contribution of each wave to the numerical flux. The authors claim that, by enforcing a linear shock structure and unambiguous sub-cell shock position, numerical shockwave anomalies are dramatically reduced.

Zaide and Roe [113] proposed to compute the fluxes in the intermediate cells by extrapolation from neighboring cells, hence a more general idea of a homogeneous flux function of the type $\mathbf{F}^{\star}_{i+1/2} = \mathbf{F}^{\star}_{i+1/2}(\mathbf{U}_{i-m}, ..., \mathbf{U}_{i-n})$ was introduced, rather than a Riemann solver that computes the numerical flux as $\mathbf{F}^{\star}_{i+1/2} = \mathbf{F}^{\star}_{i+1/2}(\mathbf{U}_i, \mathbf{U}_{i+1})$, with m and n two integer numbers. The authors in [113] outline that the conserved variables must be trusted since this is the only way to ensure conservation, however, the flux values should not be trusted.

Prior to the construction of the novel numerical fluxes $\mathbf{F}^{\star}_{i+1/2}$, physical fluxes (which are the cell centered fluxes, $\mathbf{F}_i$) are used to construct a novel approximation of the fluxes in every cell. Cell-centered fluxes, $\mathbf{F}_i$, are re-computed by means of extrapolation from neighboring cells. At every cell, the new flux is calculated as

$$\check{\mathbf{F}}_i = \frac{1}{2}(\mathbf{F}_{i+1} + \mathbf{F}_{i-1}) - \frac{1}{2}\tilde{\mathbf{J}}_{i-1,i+1}(\mathbf{U}_{i+1} - 2\mathbf{U}_i + \mathbf{U}_{i-1}), \tag{12.16}$$

with $\widetilde{\mathbf{J}}_{i-1,i+1} = \widetilde{\mathbf{J}}_{i-1,i+1}(\mathbf{U}_{i+1}, \mathbf{U}_{i-1})$ a Jacobian Roe's matrix,

$$\mathbf{F}_{i+1} - \mathbf{F}_{i-1} = \widetilde{\mathbf{J}}_{i-1,i+1}(\mathbf{U}_{i+1} - \mathbf{U}_{i-1}). \tag{12.17}$$

To construct those more general numerical fluxes, two alternatives, named flux function A and flux function B, are proposed in [113]. Such alternatives, as well as the traditional Roe flux, are detailed below:

- Traditional Roe homogeneous flux:

  The traditional Roe homogeneous flux (4.70) in 4.72 is used. It is constructed using Roe's matrix $\widetilde{\mathbf{J}}_{i+\frac{1}{2}}$,

  $$\mathbf{F}^{\star,Roe}_{i+1/2} = \frac{1}{2}(\mathbf{F}_i + \mathbf{F}_{i+1}) - \frac{1}{2} \mid \widetilde{\mathbf{J}}_{i+1/2} \mid \delta\mathbf{U}_{i+1/2}, \tag{12.18}$$

  evaluated conventionally as $\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = \widetilde{\mathbf{J}}_{i+\frac{1}{2}}(\mathbf{U}_i, \mathbf{U}_{i+1})$.

- Flux function A:

  The extrapolated fluxes, $\check{\mathbf{F}}_i$, computed by (12.16), can be directly projected onto the Jacobian's eigenvectors basis and upwinded according to the propagation velocities of the Jacobian. The resulting numerical flux is constructed using (4.70), yielding [113]

  $$\mathbf{F}^{\star,A}_{i+1/2} = \frac{1}{2}\left(\check{\mathbf{F}}_i + \check{\mathbf{F}}_{i+1}\right) - \frac{1}{2}\mathrm{sgn}\left(\widetilde{\mathbf{J}}_{i+\frac{1}{2}}\right)\delta\check{\mathbf{F}}_{i+1/2}. \tag{12.19}$$

- Flux function B:

  This new flux function is computed by means of a novel Roe's matrix that spans a wider set of cells, instead of just the two cells at each side of the discontinuity. It reads [113]

  $$\mathbf{F}^{\star,B}_{i+1/2} = \frac{1}{2}\left(\check{\mathbf{F}}_i + \check{\mathbf{F}}_{i+1}\right) - \frac{1}{2} \mid \bar{\mathbf{J}}_{i+1/2} \mid \delta\mathbf{U}_{i+1/2}, \tag{12.20}$$

  with $\bar{\mathbf{J}}_{i+1/2} = \bar{\mathbf{J}}_{i+1/2}(\mathbf{U}_{i-1}, \mathbf{U}_{i+2})$ Roe's matrix computed with cells $i-1$ and $i+2$.

12.2.1 Test case: assessment of flux functions A and B for the SWE

In order to test flux functions A and B in the framework of the SWE and compare their performance with the traditional homogeneous Roe flux, the following numerical experiment is proposed. It consists of a RP with initial data $h_L = 0.5$, $(hu)_L = 3$, $h_R = 1.6$ and $(hu)_R = 3.28787832816$, that generates a moving shock wave with speed $\mathscr{S} = 0.26171$. The computational domain is set to $[0, 450]$, with the discontinuity located at $x = 225$. Regarding the numerical discretization, the computational domain is divided in 900 cells of size $\Delta x = 0.5$ and the CFL number is set to 0.8. The simulation time is 25 s.

This test case is computed using the traditional Roe flux in (12.18) as well as the flux functions A and B in (12.19) and (12.20) respectively. The numerical solution for the discharge provided by such methods is plotted in space and time in Figure 12.8.

In Figure 12.8, it is clearly evidenced that whereas the traditional Roe solver leads to a high spike in the discharge, which generates a shedding of spurious waves, when using the novel flux functions the spike is dramatically reduced and hence the shedding of such waves. A closer examination of the numerical results evidences that flux function A provides the best performance concerning the reduction of the spike, on the other hand, flux function B does also reduce this anomalous behavior at the cell where the shock is contained but still leaves a small spike behind it. This particularity of flux function B noticed, as the spikes appear to be shifted to the left, which means that it occurs on the right side of the wavefront. This is observed in Figure 12.8 (bottom).

In Figure 12.9 (left), the numerical solutions provided by the traditional Roe solver, the solver using flux function A and the solver using flux function B is depicted at $t = 25$ s in purple, green and magenta, respectively. It is observed that both the Roe flux and the flux A capture the exact position of the shock whereas the flux B underestimates the shock speed, hence providing a slightly shifted, though convergent, shock position.

The analysis of the properties of the novel flux functions from [113] can be completed by plotting the numerical results in the phase space. Figure 12.9 (right) shows the exact and approximate Hugoniot locus for the intermediate states between the left and right states of the RP. The exact Hugoniot locus is represented by a red continuous line, the approximate locus for the traditional Roe flux by purple dots, the approximate locus for flux function A by green dots and that for flux function B by magenta dots. As outlined in [113], the optimal locus that prevents the numerical solution from exhibiting any spike and spurious waves is the straight line between the left and right state. It can be observed in Figure 12.9 (right) that only flux function A achieves this requirement and therefore it is the preferred technique for the reduction of the spike in the SWE.

12.2.2 Extension of the flux function A to the SWE with source term

It is evidenced that flux function A is a better choice than B for the resolution of moving hydraulic jumps as it provides a better estimate of the shock speed. Previous numerical experiments do not include the presence of source terms, but most realistic cases are dominated by the action of those sources. In this section, the extension of flux function A to non-homogeneous equations is carried out by means of a suitable correction of the interpolation technique that ensures a virtually exact equilibrium between fluxes and source term. In addition to this, the numerical fluxes at the interfaces must be rewritten to account for the source term.

First, it is time to find out which is the most suitable correction of the flux extrapolation to reduce the spike of discharge in both transient and steady cases. Following a similar procedure than in [113], the idea is to find an approximation of such fluxes that ensures the exact equilibrium between fluxes and source term across cell interfaces under steady conditions, while keeping the idea of having an interpolated flux in the cell contanining the shock in order to prevent the scheme from using the equilibrium flux, which leads to the spike. To this end, it is first required to find the cell where the shock is contained. We propose to use Roe celerities, $\tilde{\lambda}^m$ to unequivocally locate such a cell, since it is known that both celerities at the left interface are positive (supercritical flow entering the cell) while a combination of celerities corresponding

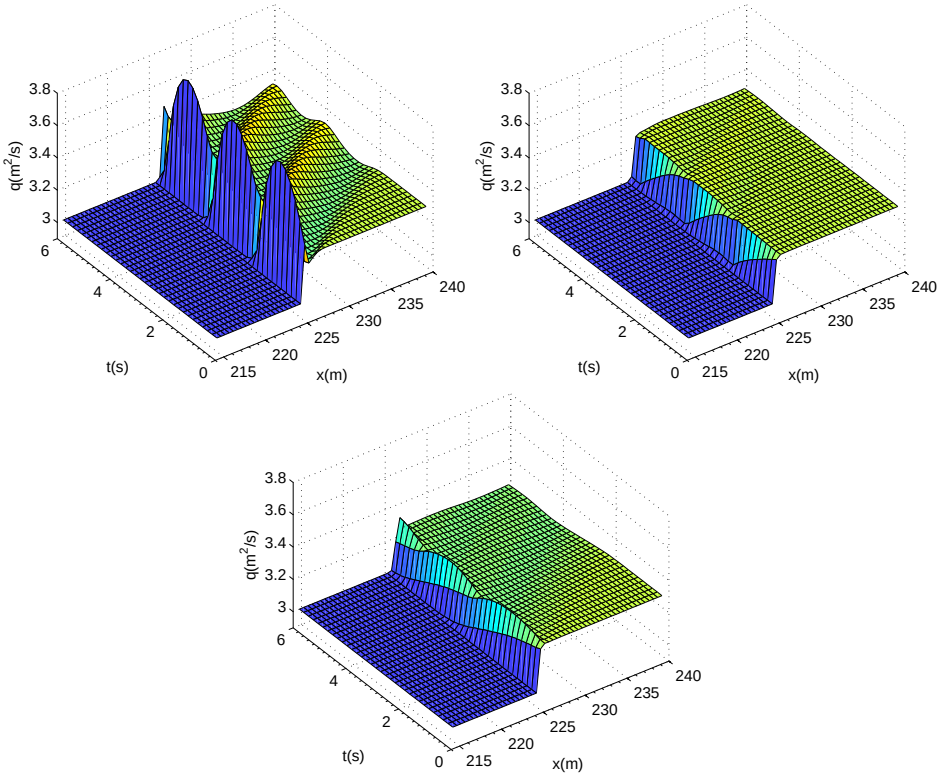


Figure 12.8: Section 12.2.1. Numerical solution provided by the traditional Roe solver (top-left) as well as the flux functions A (top-right) and B (bottom) proposed in [113] within the time interval $[0, 6]$ s.

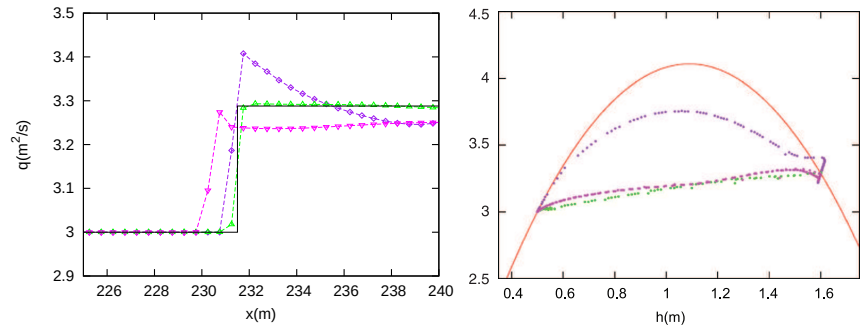


Figure 12.9: Section 12.2.1. Left: numerical solution using the Roe flux (—◇—), flux function A (—△—) and flux function B (—▽—) at $t = 25$ s. Right: exact Hugoniot locus and approximate locus for the Roe flux, flux function A and flux function B.

to subcritical conditions (one negative and the other one positive) is identified at the right interface.

Let us consider the cells, Ω_i, as single items contained in the domain Ω such that $\Omega = \{\Omega_i \mid i \in [1,...,N]\}$. Considering the possibility of multiple hydraulic jumps within the domain, we denote the set of cells containing a positive-flow hydraulic jump as

$$\mathscr{D}^+ = \left\{\Omega_i \mid \Omega_i \in \Omega \wedge \tilde{\lambda}^1_{i-1/2} \cdot \tilde{\lambda}^1_{i+1/2} < 0 \wedge h_{i-1} < h_{i+1}\right\} \tag{12.21}$$

and the set of cells containing a negative-flow hydraulic jump as

$$\mathscr{D}^- = \left\{\Omega_i \mid \Omega_i \in \Omega \wedge \tilde{\lambda}^2_{i-1/2} \cdot \tilde{\lambda}^2_{i+1/2} < 0 \wedge h_{i-1} > h_{i+1}\right\}. \tag{12.22}$$

In those cells containing a shock, which are in the set $\Omega_i \in \mathscr{D}^+ \cup \mathscr{D}^-$, the flux extrapolation, $\check{\mathbf{F}}_i$, in (12.16) is used to construct the following flux function

$$\hat{\mathbf{F}}_i = \check{\mathbf{F}}_i + \varphi\,, \tag{12.23}$$

where φ is a correction term, due to the presence of the source term, which has to satisfy the following properties:

1. Left-convergence: Under steady state, $\hat{\mathbf{F}}_i$ must hold, or at least approximate with first order of accuracy, the GRH condition at $x_{i-1/2}$, given by $\hat{\mathbf{F}}_i - \mathbf{F}_{i-1} = \bar{\mathbf{S}}_{i-1/2}$. We require at least $\hat{\mathbf{F}}_i = \mathbf{F}_i^{exact} + \mathscr{O}(\Delta x)$, with $\mathbf{F}_i^{exact}$ the exact intermediate flux that holds the GRH condition.
2. Right-convergence: Under steady state, $\hat{\mathbf{F}}_i$ must hold, or at least approximate with first order of accuracy, the GRH condition at $x_{i+1/2}$, given by $\mathbf{F}_{i+1} + \hat{\mathbf{F}}_i = \bar{\mathbf{S}}_{i+1/2}$. We require at least $\hat{\mathbf{F}}_i = \mathbf{F}_i^{exact} + \mathscr{O}(\Delta x)$.
3. Consistence: When data is smooth, the novel flux function $\hat{\mathbf{F}}_i$ converges to $\mathbf{F}_i$ with at least first order of accuracy, $\hat{\mathbf{F}}_i = \mathbf{F}_i + \mathscr{O}(\Delta x)$

Note that the two first conditions lead to obtain a flux $\hat{\mathbf{F}}_i$ that is in equilibrium (under steady conditions) with the left and right fluxes through the GRH condition. This new flux is different from the equilibrium flux, which preserved the steady state through the equilibrium of RH conditions inside the Riemann solution. Now, the new flux exactly balances the cell-values through the GRH conditions at the interfaces.

The correction term φ has to be derived to satisfy the sought properties. To this end, let us consider the expression for $\check{\mathbf{F}}_i$ in (12.16) and suppose that the intermediate state $\mathbf{U}_i$ can be expressed as a linear combination of the left and right states (linear Hugoniot)

$$\mathbf{U}_i = x_{\mathscr{S},i}\mathbf{U}_{i-1} + (1 - x_{\mathscr{S},i})\mathbf{U}_{i+1}\,, \tag{12.24}$$

where $\mathbf{U}_{i-1}$, $\mathbf{U}_i$ and $\mathbf{U}_{i+1}$ are any arbitrary left, middle and right states defining a hydraulic jump as depicted in Figure 12.4. Parameter $x_{\mathscr{S},i}$ accounts for the normalized position of the shock inside the cell, here approximated by

$$x_{\mathscr{S},i} = \frac{h_i - h_{i+1}}{h_{i-1} - h_{i+1}}\,. \tag{12.25}$$

If inserting (12.24) in (12.16), we obtain

$$\check{\mathbf{F}}_i = \frac{1}{2}(\mathbf{F}_{i+1} + \mathbf{F}_{i-1}) - \left(\frac{1}{2} - x_{\mathscr{S},i}\right)\tilde{\mathbf{J}}_{i-1,i+1}(\mathbf{U}_{i+1} - \mathbf{U}_{i-1})\,, \tag{12.26}$$

Considering now steady state conditions, we can substitute $\mathbf{F}_{i+1} = \mathbf{F}_{i-1} + \bar{\mathbf{S}}_{i-1,i+1}$ and $\tilde{\mathbf{J}}_{i-1,i+1}(\mathbf{U}_{i+1} - \mathbf{U}_{i-1}) = \bar{\mathbf{S}}_{i-1,i+1}$ in (12.26), yielding

$$\check{\mathbf{F}}_i = \mathbf{F}_{i-1} + \left(1 - x_{\mathscr{S},i}\right)\bar{\mathbf{S}}_{i-1,i+1}\,, \tag{12.27}$$

In order to satisfy the GRH condition at $x_{i-1/2}$, $\hat{\mathbf{F}}_i - \mathbf{F}_{i-1} = \bar{\mathbf{S}}_{i-1/2}$, the following equality must hold

$$\left(\mathbf{F}_{i-1} + \left(1 - x_{\mathscr{S},i}\right)\bar{\mathbf{S}}_{i-1,i+1} + \varphi\right) - \mathbf{F}_{i-1} = \bar{\mathbf{S}}_{i-1/2}\,, \tag{12.28}$$

hence, φ reads

$$\varphi = \bar{\mathbf{S}}_{i-1/2} - \left(1 - x_{\mathscr{S},i}\right)\bar{\mathbf{S}}_{i-1,i+1}\,. \tag{12.29}$$

If considering the GRH condition at $x_{i+1/2}$ and carry out an analogous derivation of φ, we obtain

$$\varphi = x_{\mathscr{S},i}\bar{\mathbf{S}}_{i-1,i+1} - \bar{\mathbf{S}}_{i+1/2}\,. \tag{12.30}$$

From the equality of Equations (12.29) and (12.30), we obtain the following condition

$$\bar{\mathbf{S}}_{i-1,i+1} = \bar{\mathbf{S}}_{i-1/2} + \bar{\mathbf{S}}_{i+1/2}\,. \tag{12.31}$$

that is to say, if the integrals at cell interfaces are computed using the trapezoidal rule, the centered integral should be computed using a composite trapezoidal rule. For instance

$$\bar{\mathbf{S}}_{i-1,i+1} = \begin{pmatrix} 0 \\ -g\frac{h_{i-1}+h_i}{2}(z_i - z_{i-1}) - g\frac{h_i+h_{i+1}}{2}(z_{i+1} - z_i) \end{pmatrix}, \tag{12.32}$$

and

$$\bar{\mathbf{S}}_{i-1/2} = \begin{pmatrix} 0 \\ -g\frac{h_{i-1}+h_i}{2}(z_i - z_{i-1}) \end{pmatrix}. \tag{12.33}$$

Proof of properties 1 and 2:

It is worth pointing out that the corrected flux in (12.41) provides an approximation of the cell-centered flux in the shock cell that converges to the exact steady flux, $\mathbf{F}_i^{exact}$, unlike traditional methods, that only converge to an equilibrium flux (different to the exact flux) that allows the steadiness of the solution. The reason why the proposed technique does not always ensure the exact flux with independence of the grid is due to the assumption we make for the definition of (12.41): the intermediate state (at cell Ω_i where the shock is located) lies on a linear Hugoniot between the left and right states, according to (12.24), which is not completely true under the presence of a bed step source term. The exact linear Hugoniot would be expressed instead as

$$\mathbf{U}_i^{exact} = x_{\mathscr{S},i}\mathbf{U}_i^- + (1 - x_{\mathscr{S},i})\mathbf{U}_i^+\,, \tag{12.34}$$

where $\mathbf{U}_i^-$ and $\mathbf{U}_i^+$ are the left and right intermediate states at the interfaces of cell Ω_i. In spite of this, the approximation in (12.24) provides a trustworthy approximation of the shock position when solving for $x_{\mathscr{S},i}$ and what is of most importance, it converges to the exact position as the grid is refined, when dealing with a smooth bed topography. This can be proved by considering Equation (4.56) under steady conditions

$$\mathbf{U}_i^- = \mathbf{U}_{i+1} - \widetilde{\mathbf{J}}_{i+\frac{1}{2}}^{-1}\bar{\mathbf{S}}_{i+\frac{1}{2}}\,, \tag{12.35}$$

which yields

$$\mathbf{U}_i^- = \mathbf{U}_{i+1} - \widetilde{\mathbf{J}}_{i+\frac{1}{2}}^{-1} \left(0, -g\bar{h}_{i+1/2}\partial z \Delta x, g\bar{h}_{i+1/2}\partial z \Delta x\right)^T , \tag{12.36}$$

allowing to show the first order convergence of $\mathbf{U}_i^-$ to $\mathbf{U}_{i+1}$ under steady conditions

$$\mathbf{U}_i^- = \mathbf{U}_{i+1} + \mathscr{O}(\Delta x). \tag{12.37}$$

If doing the same for $\mathbf{U}_i^+$, it can be proved that the approximation of $\mathbf{U}_i$ in (12.24) is

$$\mathbf{U}_i = \mathbf{U}_i^{exact} + \mathscr{O}(\Delta x), \tag{12.38}$$

hence, convergence of the solution in the cell containing the shock is ensured. Hence, the numerical error measured with L_∞ error norm will exhibit a first order convergence rate and $\hat{\mathbf{F}}_i = \mathbf{F}_i^{exact} + \mathscr{O}(\Delta x)$.

Proof of property 3:

The third property for $\hat{\mathbf{F}}_i$ is the convergence to the original flux when the solution is smooth. This property is now proven, although the proposed flux function will only be used inside the cell containing a shock, in most of the 1D cases. However, it is necessary to ensure convergence because if extending this method to 2D cases, shock profiles may span more than a cell and the flux correction must be applied in cells with smoother data around the shock.

Following [113], we can take a Taylor expansion of (12.23) to obtain

$$\hat{\mathbf{F}}_i = \mathbf{F}_i + \left(\partial_{xx}\mathbf{F} - \tilde{\mathbf{J}}_{i-1,i+1}\partial_{xx}\mathbf{U}\right)\frac{\Delta x^2}{2} - \left(x_{\mathscr{S},i}\bar{h}_{i-1/2}\partial_x z - \left(1 - x_{\mathscr{S},i}\right)\bar{h}_{i+1/2}\partial_x z\right) g \Delta x , \tag{12.39}$$

which can be written as

$$\hat{\mathbf{F}}_i = \mathbf{F}_i + \mathscr{O}_1(\Delta x) + \mathscr{O}_2(\Delta x^2), \tag{12.40}$$

proving first order convergence.

Having demonstrated that the novel flux function satisfies all the properties, we propose to use it in 1D problems as follows

$$\hat{\mathbf{F}}_i = \begin{cases} \mathbf{F}_i & \text{if} \quad \Omega_i \notin \mathscr{D}^+ \cup \mathscr{D}^- \\ \check{\mathbf{F}}_i - (1 - x_{\mathscr{S},i})\bar{\mathbf{S}}_{i-1,i+1} + \bar{\mathbf{S}}_{i-1/2} & \text{if} \quad \Omega_i \in \mathscr{D}^+ \cup \mathscr{D}^- \end{cases} \tag{12.41}$$

that is, only on those cells containing a hydraulic jump.

Finally, the expression for the numerical fluxes at cell interfaces is presented. Using definitions in Section 4.2.1, we can write the non-homogeneous version of the numerical flux in (12.19) to account for the contribution of the source term as

$$\begin{aligned} \mathbf{F}_{i+1/2}^- &= \hat{\mathbf{F}}_i + \sum_{m=1}^{I} [(\hat{\gamma} - \beta)\widetilde{\mathbf{e}}]_{i+\frac{1}{2}}^m , \\ \mathbf{F}_{i+1/2}^+ &= \hat{\mathbf{F}}_{i+1} - \sum_{m=I+1}^{N_\lambda} [(\hat{\gamma} - \beta)\widetilde{\mathbf{e}}]_{i+\frac{1}{2}}^m , \end{aligned} \tag{12.42}$$

where $\hat{\gamma}$ are the components of $\hat{\mathbf{\Gamma}}_{i+1/2} = \widetilde{\mathbf{P}}_{i+1/2}^{-1}\delta\hat{\mathbf{F}}_{i+1/2}$, the projection of the jump in the extrapolated fluxes across cell interfaces, $\delta\hat{\mathbf{F}}_{i+1/2} = \hat{\mathbf{F}}_{i+1} - \hat{\mathbf{F}}_i$.

12.2.3 Test case: steady jump over smoothly varying bed profile

In this test case, steady solutions for the flow over the following bed elevation profile

$$z(x) = \begin{cases} 0 & \text{if} \quad x < 8 \\ 0.05(x-8) & \text{if} \quad 8 \leq x \leq 12 \\ 0.2 - 0.05\,(x-12)^2 & \text{if} \quad 12 \leq x \leq 14 \\ 0 & \text{if} \quad x > 12 \end{cases} \tag{12.43}$$

are computed using the proposed technique. The computational domain is $[0, 20]$ and the solution is computed for $t = 400$ s. CFL number is set to 0.45 for all cases and the computational domain is discretized in 100 cells. The discharge is imposed to 0.6 m^2/s upstream to obtain the sonic point at the cell with the maximum bed elevation, that is $z_{max} = 0.2$. Downstream, the water depth is also imposed in order to generate the hydraulic jump. Different values for h downstream, are chosen to generate the jump at different locations and assess the performance of the proposed scheme. The complete configuration of boundary conditions is presented in Table 12.2.

Numerical results provided by the novel scheme are presented for test case 1.A in Figure 12.10 (top) and compared with the results provided by the traditional Roe solver, depicted in Figure 12.10 (bottom). No differences can be noticed when considering the solution for the water surface elevation, but it is clearly evidenced that the spike in the solution for the discharge at the cell where the shock is located is strongly reduced when using the novel numerical technique.

Case	$q_{BC:left}$ (m^2/s)	$h_{BC:right}$ (m)	Shock position (m)	$x_{\mathscr{S}}$
1.A	0.6	0.6185	13.298	0.01
1.B	0.6	0.6200	13.278	0.11
1.C	0.6	0.6220	13.252	0.24
1.D	0.6	0.6256	13.201	0.495
1.E	0.6	0.6280	13.166	0.67
1.F	0.6	0.6300	13.135	0.825
1.G	0.6	0.6320	13.102	0.99

Table 12.2: Different boundary condition configurations.

To study the behavior of this spike, the solution for the discharge in the shock cell is depicted for tests cases 1.A-1.G in Figure (12.11) (left). In this plot, the value of discharge against the normalized shock position has been depicted for the results provided by the traditional Roe solver as well as the modified solver using flux interpolation in [113] and the proposed technique. It can be observed that the method in [113] already helps decreasing the spike of discharge but only when including the correction term, as done in the novel method, the spike is virtually reduced to zero.

As outlined before, the proposed scheme does not always provide the exact discharge in the shock cell, however, the numerical estimate of the discharge in this cell converges to the exact value as the grid is refined. This property is of utmost importance, as the novel scheme can be considered L_1, L_2 and L_∞ convergent, while previous schemes were not able to converge when regarding L_∞ error norm. Convergence rate results for L_∞ error norm are presented in Figure 12.11 (right) for the traditional Roe solver and for the proposed scheme. The convergence rate test has been carried out for case 1.D using four different grids, composed of 100, 200, 400 and 800 cells. It is worth mentioning that the grid is shifted in order to keep a constant distance between the exact position of the jump and the right cell interface. It is clearly evidenced that the proposed technique allows the scheme to converge to the exact solution as the grid is refined, unlike the traditional Roe solver that does not exhibit any convergence with grid refinement because the equilibrium discharge at the shock cell is always different than the exact discharge when the shock is not located at cell interfaces.

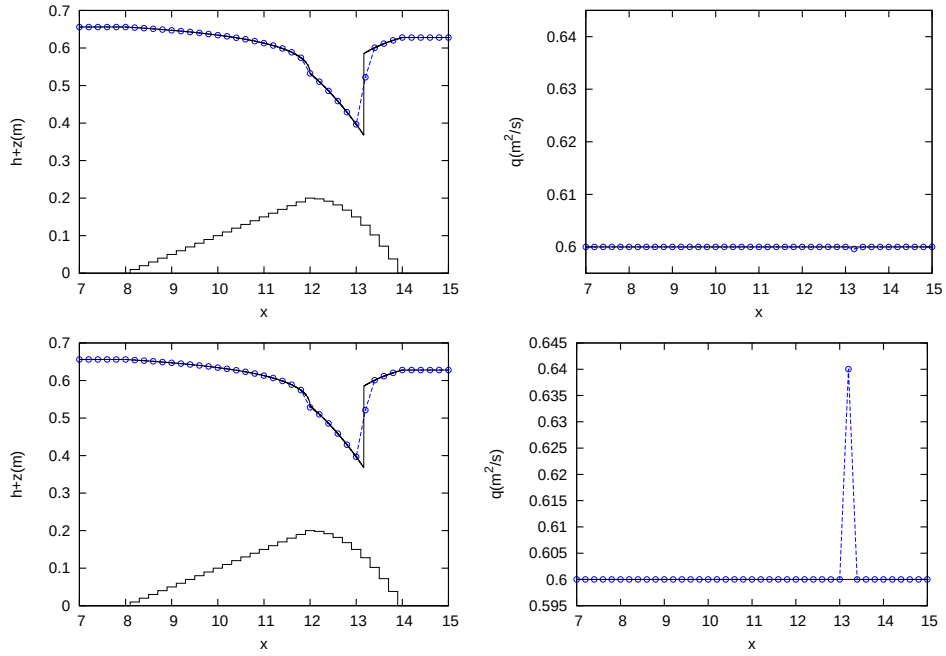


Figure 12.10: Section 12.2.3. Numerical results for $h+z$ (left) and q (right) provided by the proposed spike-reducing method (top) and by the traditional Roe solver (bottom), compared to the exact solution, using 100 cells and CFL=0.45.

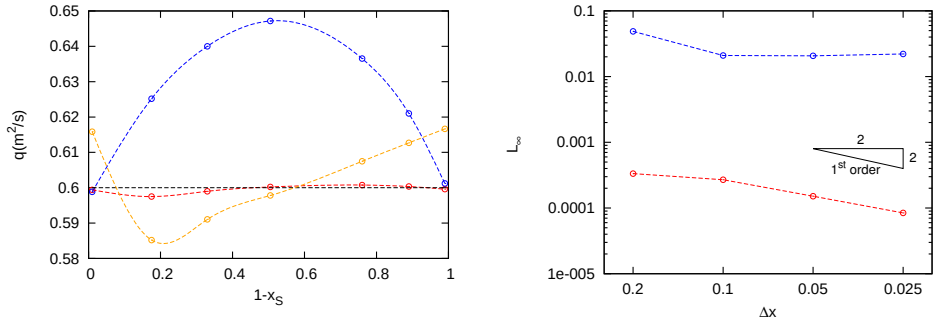


Figure 12.11: Section 12.2.3. Left: representation of the spike of discharge against the position of the shock within the cell for the traditional Roe flux (—○—), for the method using the interpolated flux in [113] (—○—) and for the proposed spike-reducing method (—○—), using 100 cells and CFL=0.45. Right: convergence rate test for the traditional Roe method (—○—) and for the proposed method (—○—), using CFL=0.45.

12.2.4 Test case: traveling jump over different bed profiles

In this test case, traveling shock waves over different bed elevation profiles $z(x)$ are computed. For all bed profiles, the maximum bed elevation is $z_{max} = 0.2$ m and the bed elevation at the boundaries is zero. To construct a solution consisting of a single jump traveling across the domain, we first compute a steady transcritical solution over the bed profile by imposing a constant discharge upstream of $q = 0.6$ m^2/s. When the steady regime is reached, the boundary condition upstream is redefined, imposing now $q = 0.556749458405104$ m^2/s and $h = 0.12$ m, which generates a supercritical state that is connected with the original subcritical state by means of a traveling hydraulic jump, according to the Hugoniot locus. The computational domain is $[0, 560]$ and the solution is computed at $t = 610$ s. The CFL number is set to 0.45 and the domain is discretized in 140 computational cells.

The bed profile will be constructed as

$$z(x) = \begin{cases} \frac{0.2}{276}(x-4) + g(x) & \text{if} \quad 4 \leq x < 280 \\ 0.2 - \frac{0.2}{276}(x-280) & \text{if} \quad 280 \leq x \leq 556 \\ 0 & \text{otherwise} \end{cases} \tag{12.44}$$

where $g(x)$ is an additional geometric function that allows to make variations in the basic constant slope profile (when $g(x) = 0$). Three different bed slopes are defined:

- *Constant slope:* The first test is carried out over a constant slope profile, setting $g(x) = 0$ in (12.44).
- *Sinusoidal variations in a constant slope:* Now, a sinusoidal variation is added to (12.44) by means of

$$g(x) = \begin{cases} 0.02\sin(0.04\pi(x-12)) & \text{if} \quad 12 \leq x < 212 \\ 0 & \text{otherwise} \end{cases} \tag{12.45}$$

- *Discontinuities in the constant slope:* Here, some discontinuities are added to (12.44) by means of

$$g(x) = \begin{cases} 0.02 & \text{if} \quad 12 \leq x < 32 \\ -0.02 & \text{if} \quad 32 \leq x < 52 \\ 0.04 & \text{if} \quad 52 \leq x < 72 \\ -0.04 & \text{if} \quad 72 \leq x < 92 \\ 0 & \text{otherwise} \end{cases} \tag{12.46}$$

Numerical results for such tests are presented in Figures 12.12, 12.13 and 12.14. Figure 12.12 shows the numerical solution at $t = 610$ s for the water surface elevation and discharge provided by the ARoe scheme and by the proposed spike-reducing method in Section 12.2.2. For all the tests, the SEBF discretization of the source term is chosen. In the figures mentioned above, major differences are observed in the solution of the discharge, which is much more oscillatory when computed by the ARoe method. On the other hand, differences on the water surface elevation are less sensitive to the spike. A space-time representation of the numerical discharge is presented in Figure 12.13, where the elimination of post-shock oscillations can be observed. In order to carry out an exhaustive analysis on the spike reducing effect of the proposed method, the evolution in time of the numerical solution for the discharge in cells 2 to 11, computed by means of the aforementioned schemes, is plotted in Figure 12.14. It is evidenced that the numerical solution provided by the proposed scheme completely reduces the spike and only leaves very small peaks that are virtually bounded by the values of the discharge at each side of the shock, hence they are not of any relevance.

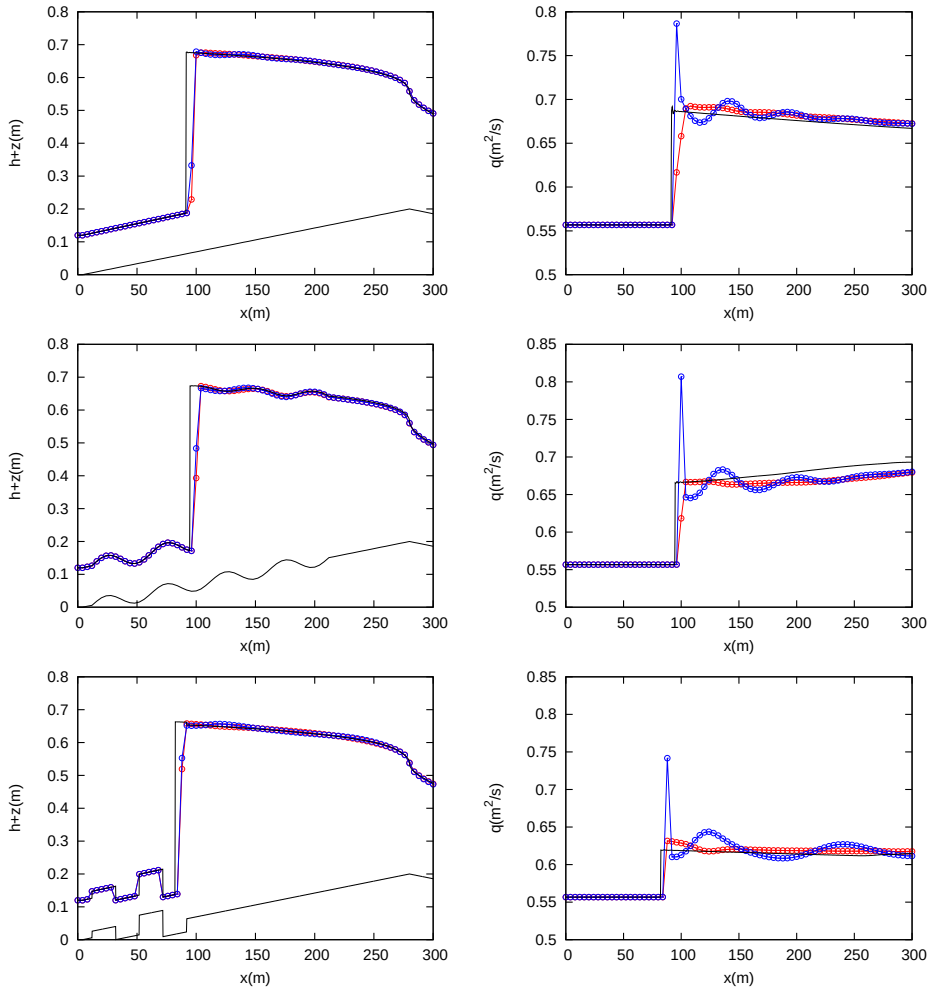


Figure 12.12: Section 12.2.4. Numerical solution at $t = 610$ s for the water surface elevation (left) and discharge (right) provided by the traditional Roe flux (—∘—) and by the proposed spike-reducing method (—∘—), using 140 cells and CFL=0.45.

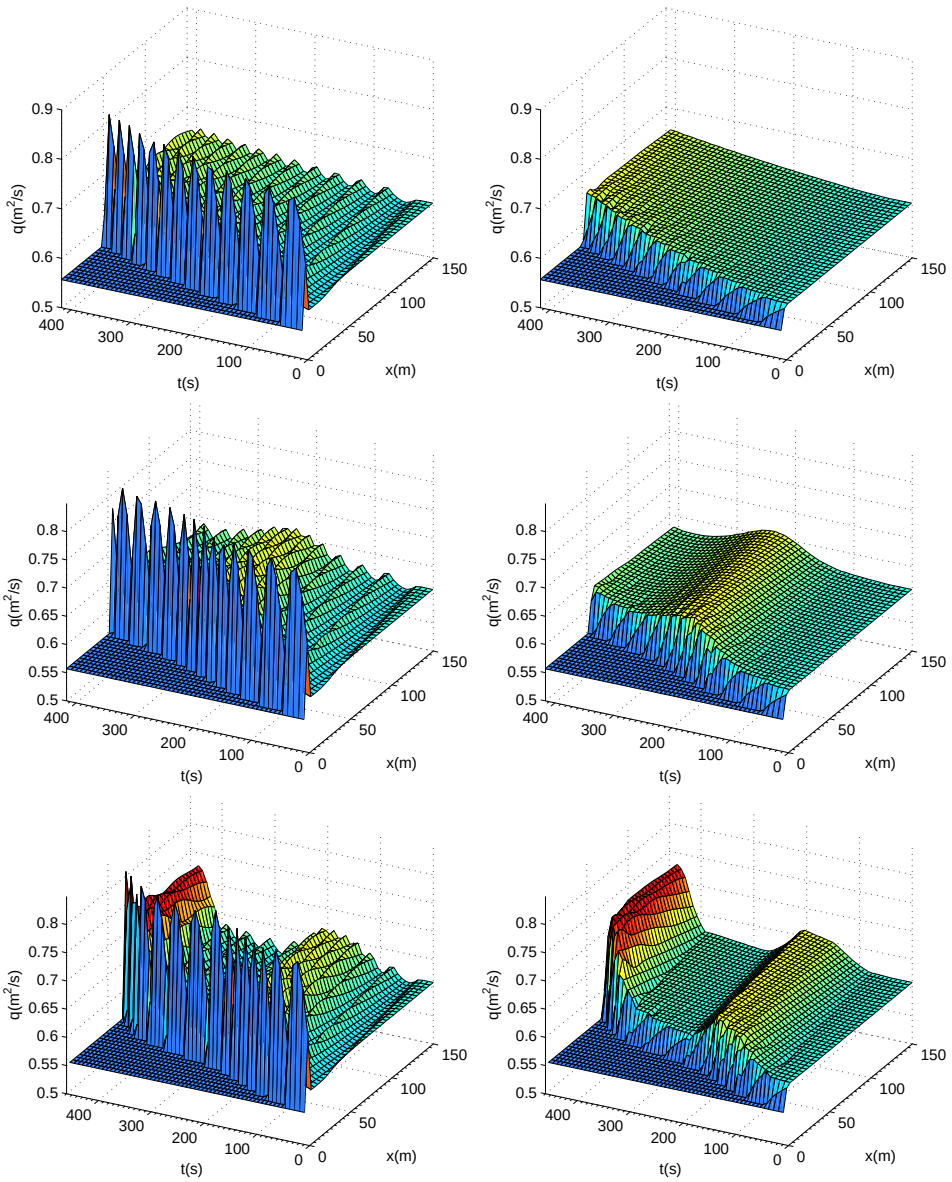


Figure 12.13: Section 12.2.4. Space-time representation of the numerical discharge provided by the traditional Roe flux (left) and by the proposed spike-reducing method (right), using 140 cells and CFL=0.45.

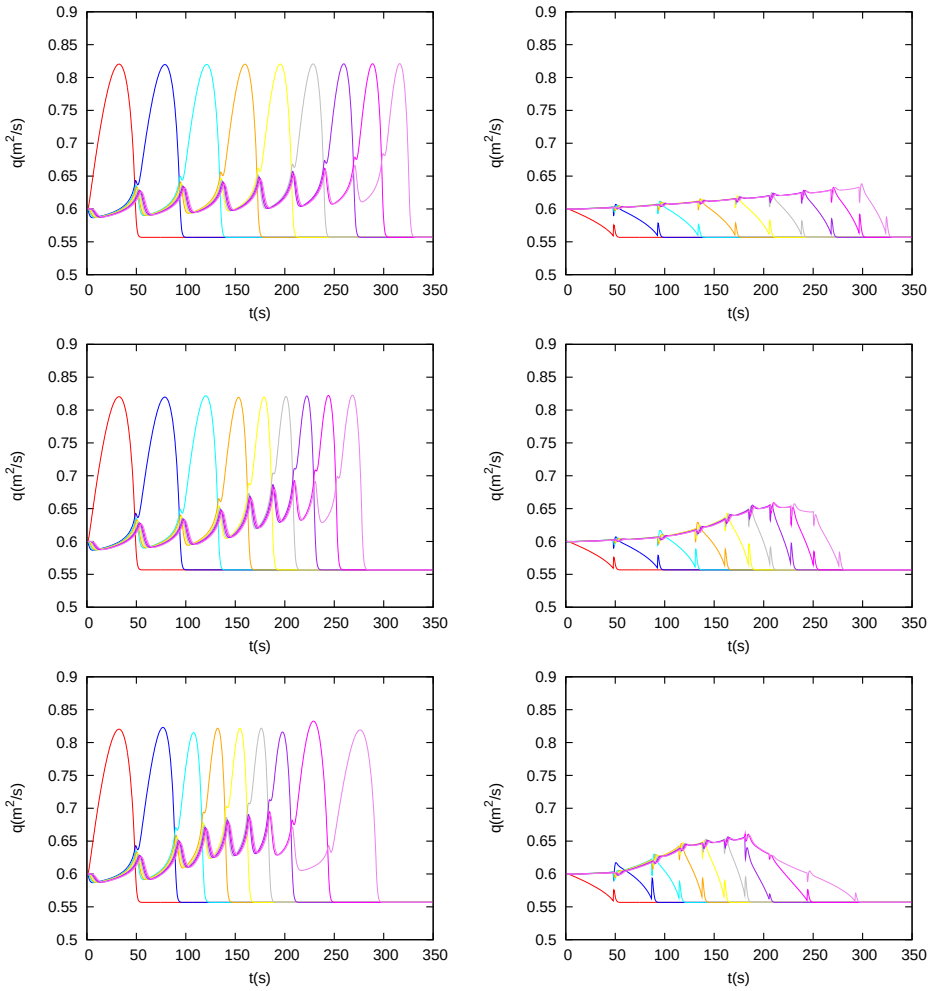


Figure 12.14: Section 12.2.4. Evolution in time of the numerical solution for the discharge inside cells 2 to 11 provided by the traditional Roe flux (left plot) and by the proposed spike-reducing method (right plot), using 140 cells and CFL=0.45.

12.3 Extension to 2 dimensions

The 2D extension of the spike-reducing solver presented in the previous section is developed by applying the 1D methodology to each direction independently. We now consider 2D problems as (2.6), with $\mathbf{E} = (\mathbf{F}, \mathbf{G})$ the fluxes in x and y directions. When considering a Cartesian mesh, it is possible to define the interpolated fluxes as done in the 1D case

$$\check{\mathbf{F}}_{i,j} = \frac{1}{2}(\mathbf{F}_{i+1,j} + \mathbf{F}_{i-1,j}) - \frac{1}{2}\tilde{\mathbf{J}}_{(i-1,i+1),j}(\mathbf{U}_{i+1,j} - 2\mathbf{U}_{i,j} + \mathbf{U}_{i-1,j}), \tag{12.47}$$

$$\check{\mathbf{G}}_{i,j} = \frac{1}{2}(\mathbf{G}_{i,j+1} + \mathbf{G}_{i,j-1}) - \frac{1}{2}\tilde{\mathbf{J}}_{i,(j-1,j+1)}(\mathbf{U}_{i,j+1} - 2\mathbf{U}_{i,j} + \mathbf{U}_{i,j-1}). \tag{12.48}$$

In [106], the author outlines that in the stationary case, each intermediate shock state is adjacent to at least two end states of the shock, but not necessarily aligned in the x or y direction. Therefore, the interpolated flux in the x-direction should lie on a straight line in flux space with the interpolated flux in the y-direction. This requires a genuinely two-dimensional method, using interpolated fluxes computed from information in both directions, however, the dimension-by-dimension method proposed here is powerful enough to provide the sought results.

As in the 1D case, a shock-detection algorithm is required. A dimension-by-dimension shock detection procedure is proposed. When dealing with oblique shocks in Cartesian meshes, the dimension-by-dimension detection of shocks may provide discrete shock profiles given by cells $\Omega_{i,j} \in \mathscr{D}$, with $\mathscr{D}$ the set of cells containing the shock, such that two consecutive cells only share one vertex and no edges. This means a 2D discontinuity of the shock profile, which can reduce the robustness of the spike-reducing method.

Unlike in the 1D case, the shock profile detected by the algorithm will now span a width of 3 cells in order to avoid instabilities. The set of cells definfing a positive-flow hydraulic jump in the x and y directions are given by

$$\mathscr{D}^{+,x} = \left\{\Omega_{i-1,j} \cup \Omega_{i,j} \cup \Omega_{i+1,j} \mid \tilde{\lambda}^1_{i-1/2,j} \cdot \tilde{\lambda}^1_{i+1/2,j} < 0 \wedge h_{i-1,j} < h_{i+1,j}\right\} \tag{12.49}$$

and

$$\mathscr{D}^{+,y} = \left\{\Omega_{i,j-1} \cup \Omega_{i,j} \cup \Omega_{i,j+1} \mid \tilde{\lambda}^1_{i,j-1/2} \cdot \tilde{\lambda}^1_{i,j+1/2} < 0 \wedge h_{i,j-1} < h_{i,j+1}\right\} \tag{12.50}$$

respectively. The set of cells defining a negative jump is defined equivalently. The sets of cells containing x and y shocks are finally constructed as

$$\mathscr{D}^x = \mathscr{D}^{+,x} \cup \mathscr{D}^{-,x}, \quad \mathscr{D}^y = \mathscr{D}^{+,y} \cup \mathscr{D}^{-,y} \tag{12.51}$$

and is used to define the novel fluxes, $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$, as follows

$$\hat{\mathbf{F}}_{i,j} = \begin{cases} \mathbf{F}_{i,j} & \text{if} \quad \Omega_{i,j} \notin \mathscr{D}^x \\ \check{\mathbf{F}}_{i,j} - (1 - x_{\mathscr{S},i,j})\bar{\mathbf{S}}_{(i-1,i+1),j} + \bar{\mathbf{S}}_{i-1/2,j} & \text{if} \quad \Omega_{i,j} \in \mathscr{D}^x \end{cases} \tag{12.52}$$

$$\hat{\mathbf{G}}_{i,j} = \begin{cases} \mathbf{G}_{i,j} & \text{if} \quad \Omega_{i,j} \notin \mathscr{D}^y \\ \check{\mathbf{G}}_{i,j} - (1 - y_{\mathscr{S},i,j})\bar{\mathbf{S}}_{i,(j-1,j+1)} + \bar{\mathbf{S}}_{i,j-1/2} & \text{if} \quad \Omega_{i,j} \in \mathscr{D}^y \end{cases} \tag{12.53}$$

The fluxes in (12.52) and (12.53) are used to construct the normal flux and to define RPs at cell interfaces, as done in the 1D case.

12.3.1 Test case: 2D shock wave over an inclined plane

This numerical experiment consist of a supercritical flow that hits a circular obstacle and generates an upstream-propagating shock wave. The computational domain is $\Omega = [0, 80] \times [0, 100]$. The water depth and discharge at the inlet are set as $h_L = 1$ m and $hu_L = 9$ m^2/s respectively, while at the other boundaries, transmissive BCs are imposed. The solid body is given by the set of points

$$\mathscr{W} = \left\{ \mathbf{x} \,|\, (x-80)^2 + (y-50)^2 \leq 400 \,, \mathbf{x} \in \Omega \right\} \tag{12.54}$$

and the bed elevation is given by

$$z(x,y) = \begin{cases} 0 & \text{if} \quad x < 5 \\ 0.01(x-5.0) & \text{if} \quad x \geq 5 \end{cases} \tag{12.55}$$

The solution is computed at $t = 150$ s, using $\Delta x = 1$, and is presented in Figures 12.15 and 12.16. In Figure 12.15, a 3D representation of the numerical $h + z$ and z is presented, computed using the spike-reducing solver. In Figure 12.16, the numerical discharges computed by the traditional ARoe solver and the spike-reducing solver are presented. The reduction of the spike along the shock profile, provided by the spike-reducing solver, is remarkable.

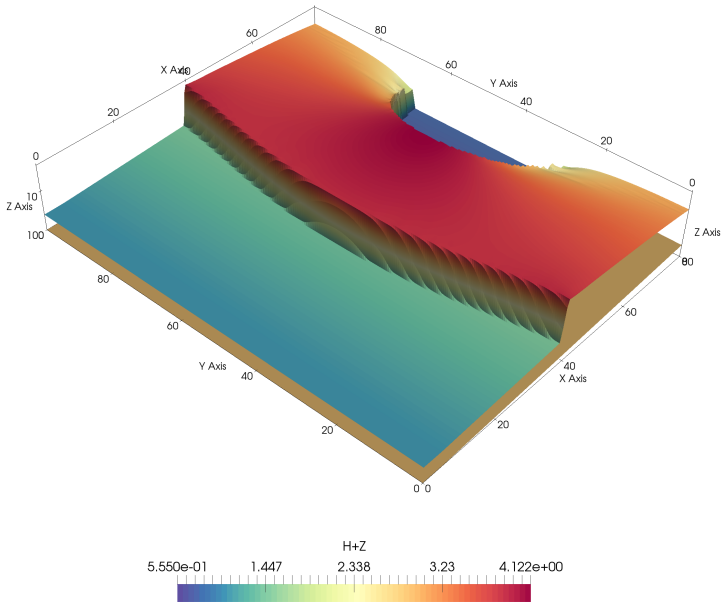


Figure 12.15: Section 12.3.1. Computed water surface elevation and bottom surface at $t = 150$ s.

12.3.2 Test case: 2D shock wave over a sinusoidal inclined plane

This test case is analogous to the previous one, but a different bed topography and inlet BC are considered. In this case, the shock wave remains stationary, generating a bow shock around the solid body. A strong influence of the topography is observed in the shape of the shock. The water depth and discharge at the inlet are now set as $h_L = 0.8$ m and $hu_L = 9$ m^2/s respectively. At the other boundaries, transmissive BCs

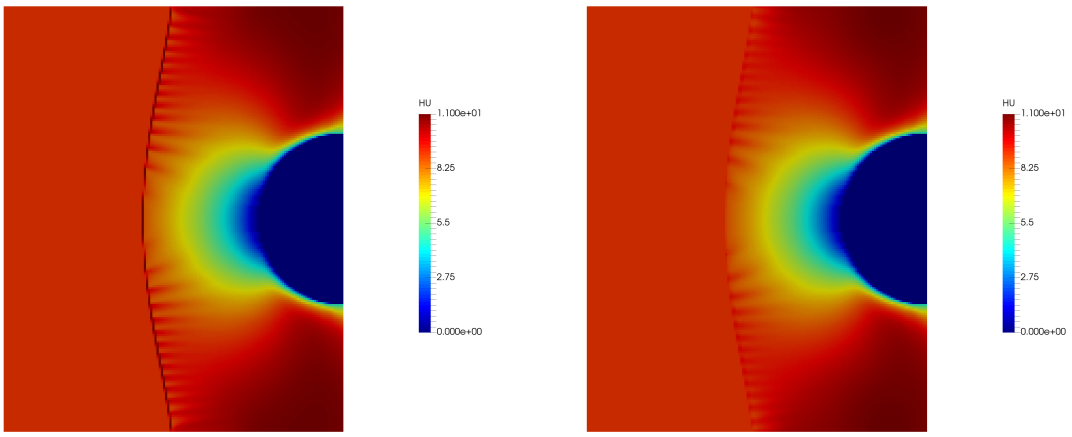


Figure 12.16: Section 12.3.1. Computed discharge at $t = 150$ s using the traditional ARoe scheme (left) and the spike-reducing solver (right).

are imposed. The solid body is defined in (12.54) and the bed elevation is given by

$$z(x,y) = \begin{cases} 0 & \text{if} \quad x < 5 \\ 0.01(x-5.0) + 0.3\sin(0.1\pi(x-5))\cos(0.1\pi y) & \text{if} \quad 5 \leq x \leq 65 \\ 0.01(x-5.0) & \text{if} \quad x > 65 \end{cases} \tag{12.56}$$

The solution is computed at $t = 150$ s and is presented in Figures 12.15 and 12.16. In Figure 12.15, a 3D representation of the numerical $h+z$ and z is presented, computed using the spike-reducing solver. In Figure 12.16, the numerical discharges computed by the traditional ARoe solver and the spike-reducing solver, using two different grids with using $\Delta x = 1$ and $\Delta x = 0.5$, are presented. The reduction of the spike along the shock profile, provided by the spike-reducing solver, is again remarkable.

12.4 Concluding remarks

The highlights of this chapter are listed below:

- The slowly-moving shock anomaly in the framework of the SWE has been theoretically studied. It has been shown that the presence of the spike in discharge when using Godunov's scheme can be explained by the non-linearity of the Hugoniot locus. At the intermediate cell, it exists an equilibrium flux so that the RH conditions at the cell interfaces are satisfied.

- Different spike-reducing techniques found in the literature and have been tested for the resolution of the homogeneous SWE. The flux function A from [106] outperforms the other candidate functions.

- A novel flux function for the SWE with bed elevation, based on the flux function A, has been proposed. Such function is constructed by means of a flux interpolation plus a correction term that accounts for the presence of the source term. Then, the resulting flux is upwinded using the information contained in the Jacobian matrix. The novel flux function ensures two properties:

  - Convergence to the exact flux that holds the GRH condition at cell interfaces.
  - Consistency, as it yields the physical flux when refining the grid.

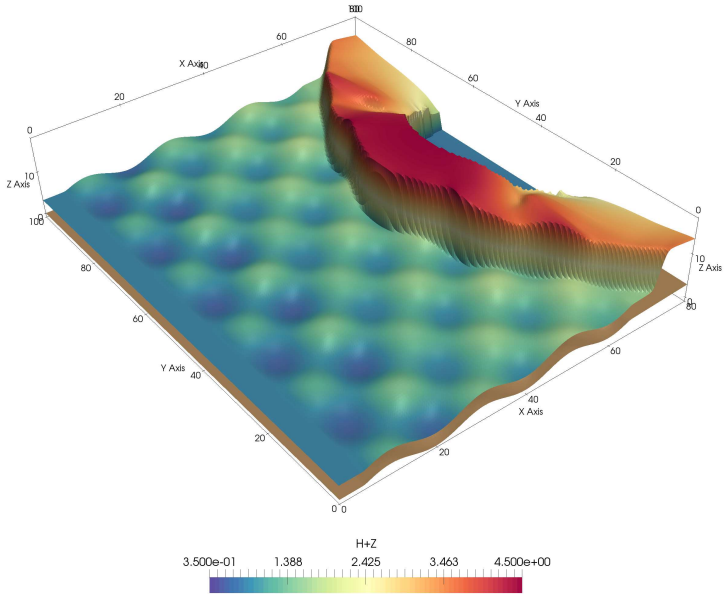


Figure 12.17: Section 12.3.2. Computed water surface elevation and bottom surface at $t = 150$ s.

- The novel flux function has been assessed using a variety of cases involving the computation of steady and transient hydraulic jumps. Numerical results evidence that the aforementioned properties are satisfied.
- To the knowledge of the author, convergence to the exact solution when solving steady hydraulic jumps (measured with L_∞ error norm) has been ensured for the first time. Traditional schemes always produce the spike in discharge, even when refining the grid, and convergence of the solution inside the cell containing the shock cannot be guaranteed.
- The methods have been successfully extended to 2 space dimensions using a dimension-by-dimension approach, though the problem is essentially bidimensional. Such approach provides acceptable results as shown in the 2D numerical experiments. The method reduces the spurious shock line for the discharge, that appears where the front of the hydraulic jump is located.

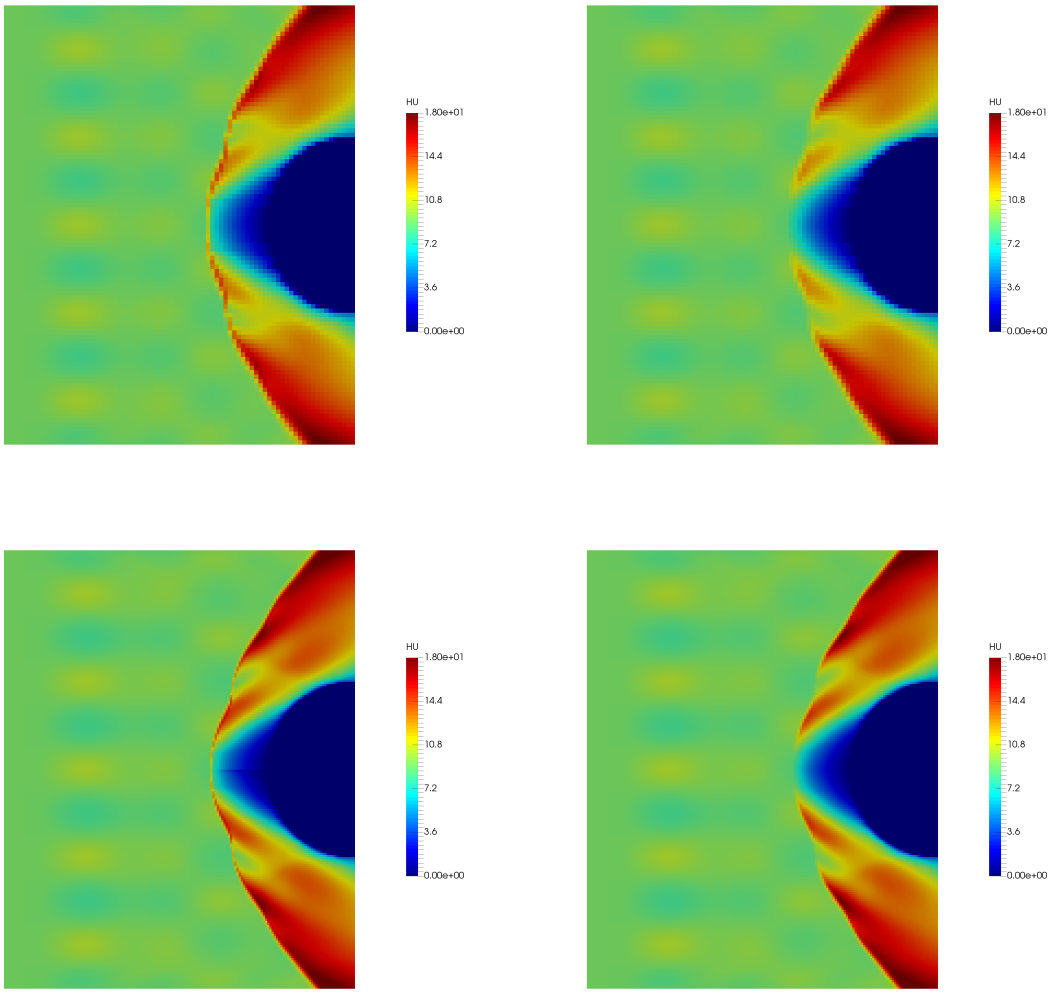


Figure 12.18: Section 12.3.2. Computed discharge at $t = 150$ s using the traditional ARoe scheme (left) and the spike-reducing solver (right) using a $\Delta x = 1$ grid (top) and $\Delta x = 0.5$ grid (bottom).

13 CONCLUDING REMARKS AND FUTURE WORK

In this thesis, new methodologies for the construction of high resolution FV schemes have been presented. The WENO-ADER approach has been followed, putting an especial emphasis on the numerical treatment of source terms. The proposed schemes have been applied to the resolution of linear advection-reaction equations, the acoustic equations and the SWE. For the latter, numerical shockwave anomalies, inherent to the use of FV schemes, have been studied and circumvented.

Unlike previous work found in the literature, the arbitrary order WENO-ADER schemes presented in this work are based on augmented solvers. Such solvers include the contribution of the source term at cell interfaces (in the resolution of the DRP). This allows to correctly upwind the source term while satisfying GRH conditions at cell interfaces, which ensures an exact balance between fluxes and sources. This approach shows a good performance when solving RPs with large discontinuities in the source terms at cell interfaces. A new family of solvers for the DRP, called (L)FS solvers, has been presented.

The application to the SWE is the main focus of this thesis. Along with the use of augmented solvers, the mathematical and numerical treatment of the bed step discontinuity is an issue of paramount importance. This has been deeply studied here, following [134, 67, 135]. The conditions for the jump across the bed step contact wave have been derived based on the conservation of specific mechanical energy, which seems to be the most reasonable choice and does not contradict the conservation of momentum [134]. Such conditions, which consist of the conservation of both the GRH and RI across the discontinuity, have been used to derive a particular energy conservative source term discretization, called SEBF. Additionally, it has been observed that the discretization of the bed slope in presence of hydraulic jumps is not a trivial task. The DF and SEBF ensure convergence to the exact position of the jump under steady state while other techniques based on the IF do not. On the other hand, when computing moving jumps, the exact positioning of the jump is not guaranteed with any technique. Among the source term formulations studied in the thesis, we choose the SEBF, but further investigation should be done in the resolution of travelling jumps over varying bed.

Using energy-conservation arguments (the SEBF for instance), EB augmented WENO-ADER schemes based on the ARoe and HLLS solvers have been presented. Following [75], Romberg's method has been used to approximate the integral of the source term inside cells. Such method allows to extend the particular SEBF of the source term, which is second order accurate, to arbitrary order. The resulting schemes outperform their respective well-balanced versions as they provide the exact solution not only under quiescent equilibrium but also under moving equilibrium. It has also been evidenced that there is a requirement of reconstructing (using the WENO method) the equilibrium quantity: for the well-balanced scheme, $h+z$ must be reconstructed, while for the EB scheme, we use $1/2u^2 + g(h+z)$. Then, hu is also reconstructed and h is computed from the reconstructions.

The proposed methods are one-step methods, where time integration is done by means of exact integration of polynomial expansions in time of the variables. Such expansions require the reconstruction of time derivatives using the CK procedure. It is worth pointing out that, in order to obtain a well-balanced or EB scheme, the CK procedure also has to be derived in terms of the equilibrium variables so that derivatives vanish under steady state. For most of the applications considered in this thesis, the CK procedure is well suited. When dealing with very stiff source terms, it would be recommended to explore new methodologies that provide more accurate results regarding the evolution in time (see for instance [34, 33]). In a future work, the derivation of a more accurate estimation of time derivatives for the SWE with Coriolis will be studied.

The methods have been extended from 1 to 2 spatial dimensions in Cartesian meshes using a dimension-by-dimension approach. Such methodology allows an straightforward application of 1D algorithms, namely the data reconstruction, Riemann solver and source term integration, to the 2D case and leads to the construction of efficient schemes of simple implementation.

The main flaw when using Cartesian meshes is the potential mesh dependency of the solution. For the schemes considered in this thesis, mesh dependency has only been observed for low orders of accuracy (e.g. 1-st order Godunov's scheme), while high order schemes do not exhibit significant mesh dependency. Another additional drawback is that they provide a poorer resolution of the flow near solid boundaries with complex geometries. This could be fixed by means of a Cartesian cut-cell approach, which allows to retain a boundary conforming grid. On the other hand, Cartesian meshes offer very favorable features as they are beneficial for memory saving purposes as well as for ensuring a good homogeneity among cells while allowing very simple implementations. They are also appropriate for the utilization of adaptive mesh refinement algorithms, which will be object of future work.

For the 2D WENO-ADER schemes proposed in this work, the integration of the source terms inside cells is carried out by means of a combination of Gaussian and Romberg integration, following again a dimension-by-dimension approach. Romberg's method allows to extend a particular second order quadrature rule to arbitrary order. Such quadrature rule can be designed to satisfy certain properties that ensure an exact balance between sources and fluxes and allows to preserve steady states of relevance. For the 2D SWE, such states are the quiescent equilibrium and the geostrophic equilibrium. The extension of the EB property to 2D is not trivial and will be considered in the future.

It has been shown that all kind of sources can be included in the definition of the RP when constructing 1-st order solvers, however, only geometric source terms must be accounted for when constructing higher order schemes. This is because the integral of the source term is carried out over $[-\Delta x/2, \Delta x/2]$ when constructing a 1-st order scheme whereas it only accounts for the jump at the interface, $[0^-, 0^+]$, when considering higher order schemes. Hence, only geometric source terms offer a jump-like contribution. Nevertheless, non-geometric source terms can be rewritten in geometric form in order to satisfy certain equilibrium properties. This is the case of the Coriolis source term, which is discretized here like the bed slope source term. Using simple mathematical transformations, we can define the so-called apparent topography and satisfy the quiescent and geostrophic equilibrium.

Broadly speaking, the proposed methods are much more accurate and efficient than their 1-st order versions (see Figure 11.3), however, there is still room for improvement. In the future, the implementation of h (cell size) and p (polynomial degree) adaptivity should be considered in order to increase the computational efficiency of the methods. Moreover, the substitution of the CK procedure by more advanced techniques will also be considered to avoid suboptimal convergence rates when dealing with stiff sources. Regarding the discretization of source terms, the assumption of pure 1D spatial gradients of the source term at cell interfaces in Equation 11.14 should be reconsidered and, in a broader sense, multidimensional Riemann solvers ought to be explored. The utilization of 1D Riemann solvers for 2D problems leads to a poor representation of discontinuities lying oblique to the grid, especially shear and contact discontinuities, and an excess of dissipation at low speeds [148].

On the other hand, the mathematical model of the SWE, which is sufficient to represent most advective

nonlinear phenomena considered in this work, exhibits some shortages when considering highly turbulent flows in presence of unstable shear layers where intense mixing is taking place. The addition of turbulence models in combination with high order solvers must be explored. Obtaining the adequate balance between numerical and physical diffusion is not a trivial task, but the use of higher order schemes may help. Furthermore, the system of equations can be enhanced by accounting for the projections of the gravity force in order to deal with large slopes [98] as well as including bed load transport [95, 96, 97] and other physical phenomena.

The final part of the thesis is devoted to the generation of numerical fixes that overcome numerical shockwave anomalies in the SWE, such as the slowly-moving shock that appears in presence of hydraulic jumps. Such anomaly has been taken for granted for a long time, but it may eventually ruin the solution under certain particular conditions. A theoretical framework of study for this anomaly has been provided and a 1D/2D spike reducing solver has been presented, based on a previous work by Zaide and Roe [106]. The resulting scheme has been exercised in a variety of scenarios and outperforms, by far, the traditional ARoe scheme. Moreover, convergence to the exact solution in presence of hydraulic jumps (measured with the L_∞ error norm), is achieved for the first time to the knowledge of the author. There is still room for improvement when considering spike reduction techniques in presence of Carbuncle-like instabilities. More advanced solvers able to suppress the spurious spikes while eliminating the unphysical Carbuncles have not been presented yet.

CONCLUSIONES

En esta tesis se presenta una nueva estrategia para la construcción de esquemas numéricos de orden arbitrario utilizando volúmenes finitos. Los métodos propuestos se construyen utilizando la metodología WENO-ADER y poniendo un énfasis especial en el tratamiento numérico de los términos fuente. Se considera la aplicación de los mismos a problemas de transporte lineal con y sin reacción, al problema acústico lineal y a las ecuaciones de aguas poco profundas. Para estas últimas, se realiza un estudio detallado de las anomalías numéricas inherentes a los métodos de volúmenes finitos y se proponen soluciones.

A diferencia de la mayoría de métodos propuestos hasta la fecha, los esquemas WENO-ADER que se proponen en este trabajo se basan en *Riemann solvers* aumentados, que consideran la contribución del término fuente en la resolución del problema de Riemnann derivativo. Esta aproximación garantiza un equilibrio perfecto entre flujos y términos fuente en situaciones estacionarias y una convergencia con orden arbitrario en casos transitorios. Además, proporciona excelentes resultados en la resolución de problemas de Riemann con grandes discontinuidades. Se propone una nueva familia de algoritmos para la resolución del problema de Riemnann derivativo, denominados *(L)FS solvers*. Se observa que sólo aquellos términos fuente de naturaleza geométrica deben ser considerados en el problema de Riemnann derivativo.

La resolución de las ecuaciones de aguas poco profundas con términos fuente es el enfoque principal de este trabajo. Paralelamente al desarrollo de esquemas numéricos aumentados para la resolución de dichas ecuaciones, es necesario encontrar una correcta discretización de los términos fuente que aproxime la física del problema con fidelidad. Por su naturaleza geométrica, este trabajo profundiza en el tratamiento numérico del término fuente de variación de fondo [134, 67, 135], especialmente cuando existen discontinuidades en el mismo. Se comprueba que la hipótesis más razonable para su discretización es la conservación de la energía mecánica específica, $1/2u^2 + g(h + z)$, lo cual no contradice la conservación del momento lineal [134] y además es fiel al modelo matemático no disipativo que se resuelve. Bajo esta hipótesis, se comprueba que en una discontinuidad del fondo se debe cumplir la condición de Rankine-Hugoniot generalizada así como conservar los invariantes de Riemann. Siguiendo estas restricciones, se deriva una formulación para el término fuente de fondo, denominada SEBF, que conserva la energía. Además, se muestra que la elección de una formulación única para el término fuente de fondo no es un asunto trivial en presencia de resaltos hidráulicos. Se concluye que formulaciones de tipo IF no son capaces de reproducir la posición de resalto hidráulico en casos estacionarios mientras que las de tipo DF, no lo hacen en casos transitorios, aunque sí que proporcionan una convergencia con el refinamiento de malla.

Utilizando una discretización del término fuente que garantice la conservación de la energía, se construyen esquemas numéricos de tipo EB basados en los métodos ARoe y HLLS y siguiendo la metodología WENO-ADER. Siguiendo el trabajo de [75], se propone una extensión de la formulación SEBF del término fuente a orden arbitrario usando el método de integración de Romberg. Los esquemas numéricos

resultantes proporcionan resultados claramente superiores a esquemas puramente *well-balanced* y a otros esquemas EB de primer orden. La característica fundamental es que son capaces de reproducir con precisión de máquina situaciones estacionarias que involucren velocidades no nulas mientras proporcionan convergencia con orden arbitrario en casos transitorios. Además, se ha estudiado la dependencia de las propiedades de conservación del esquema en función de las variables reconstruidas con el método WENO. Se concluye que para construir un método *well-balanced*, es necesario reconstruir $h+z$ mientras que para un método EB, hay que reconstruir $1/2u^2+g(h+z)$.

Los métodos aquí propuestos son esquemas de un único paso de actualización, en los que la integración numérica se realiza mediante una integración exacta de una expansión polinómica en el tiempo de las variables. Para construir dicha expansión, se requiere el cálculo de derivadas temporales utilizando el procedimiento de Cauchy-Kovalevskaya que permite expresar las mismas en función de derivadas espaciales previamente reconstruidas. La derivación de este procedimiento no es trivial cuando se busca un esquema de tipo *well-balanced* o EB. Para la mayoría de las aplicaciones consideradas en esta tesis este método proporciona buenos resultados, pero en el caso de términos fuente de mayor intensidad es necesario explorar otras alternativas [34, 33].

Los métodos se proponen inicialmente en 1 dimensión espacial y luego se extienden a 2 dimensiones en mallas Cartesianas. Algunos algoritmos en 2D se construyen como combinación de sus versiones 1D en x e y. La principal desventaja del uso de mallas Cartesianas es la potencial dependencia de la solución con la malla, debido a la existencia de dos direcciones preferenciales de propagación de la información. Para los esquemas propuestos en esta tesis, esta dependencia sólo se observa cuando se trabaja con primer orden de convergencia, mientras que con mayores órdenes se consigue la eliminación de las direcciones preferenciales. También cabe destacar que el uso de mallas Cartesianas no permite obtener una buena resolución del flujo en fronteras sólidas con geometrías complicadas. Esto se puede solventar fácilmente utilizando una aproximación de tipo *Cartesian cut-cell*, que será objeto de trabajo futuro. Por otro lado, el uso de este tipo de mallas proporciona unas características muy ventajosas en términos de uso de memoria, homogeneidad entre celdas y coste de implementación de los métodos. También permiten una implementación sencilla de algoritmos de refinamiento adaptativo de malla.

Para la construcción de métodos WENO-ADER en 2 dimensiones espaciales que garanticen la propiedad *well-balanced* es necesario utilizar procedimientos particulares de integración de los términos fuente que proporcionen un equilibrio exacto entre flujos y los mismos. Para ello, se propone un método de integración 2D basado en una combinación de fórmulas de cuadratura Gausiana e integración de Romberg, que se aplican recursivamente para obtener una integral bidimensional. Con este método se construyen métodos *well-balanced* para las ecuaciones de aguas poco profundas con variación de fondo y Coriolis. La extensión de métodos EB a 2D será objeto de estudio en el futuro.

En trabajos anteriores se menciona que la utilización de algoritmos aumentados para la resolución de problemas de Riemann con términos fuente de distinta naturaleza es adecuada para la construcción de esquemas de primer orden. Sin embargo, cuando se construyen esquemas de mayor orden, sólo aquellos términos fuente de tipo geométrico deberán incluirse en la definición del problema de Riemann derivativo. La razón es que la integral del término fuente se realiza en el dominio $[-\Delta x/2, \Delta x/2]$ cuando se trata de un esquema de orden 1 mientras que para un esquema de mayor orden, el dominio se reduce a $[0^-, 0^+]$. Por lo tanto, en este último caso sólo es necesario contabilizar el salto de la variable geométrica en $x=0$, ya que el resto de términos fuente no pueden ser integrados en la discontinuidad. Sin embargo, aquellos términos fuente de naturaleza no geométrica pueden ser escritos de manera geométrica con la finalidad de satisfacer determinadas condiciones de equilibrio y la propiedad *well-balanced*. En este trabajo, el término fuente de Coriolis se reescribe en forma geométrica, lo que da lugar a la definición de una topografía aparente que permite satisfacer el equilibrio geostrófico con precisión de máquina.

En términos generales, los métodos propuestos son mucho más precisos eficientes que sus versiones de primer orden, aunque todavía queda mucho por mejorar. Como trabajo futuro se propone la implementación de algoritmos adaptativos *hp* con la finalidad de incrementar notablemente la eficiencia computacional. Además, se debería considerar la sustitución del procedimiento de Cauchy-Kovalevskaya por

técnicas más novedosas, para evitar tasas de convergencia subóptimas. En lo que respecta a la discretización de términos fuente, las hipótesis de variaciones puramente unidimensionales que se asumen en (11.14) se deberían reconsiderar. En un sentido más amplio, sería interesante explorar el uso de *Riemann solvers* multidimiensionales [148].

Respecto a la utilización del modelo de aguas poco profundas, cabe destacar que dicho modelo es de utilidad para la representación de fenómenos de propagación de onda no lineales. Sin embargo, muestra importantes deficiencias en presencia de flujos muy turbulentos producidos por esfuerzos cortantes elevados entre capas. Para mejorar el comportamiento predictivo del modelo en estos casos, es necesario añadir modelos de turbulencia que añadan disipación viscosa a las ecuaciones. En el futuro se explorará el uso de modelos complejos de turbulencia en combinación con esquemas numéricos de muy alto orden. Otros aspectos a mejorar del modelo involucran una mejor representación de las fuerzas gravitacionales para poder resolver flujos en sobre grandes pendientes [98]. Por otro lado, también sería interesante introducir otros fenómenos físicos relevantes como transporte de sedimento y fondo [95, 96, 97].

La parte final de la tesis centra en el estudio de anomalías numéricas inherentes a los métodos de volúmenes finitos en la resolución de las ecuaciones de aguas poco profundas. En particular, el estudio se reduce al problema del pico de caudal que aparece en presencia de resaltos hidráulicos. Esta anomalía ha sido considerada durante años una particularidad de la solución en vez de un problema real, sin embargo, aquí se muestra que en determinadas circunstancias puede destruir la convergencia de la solución por completo. Siguiendo estudios previos sobre anomalías numéricas en las ecuaciones de Euler [106], se formula un marco teórico para el estudio de dichas anomalías en las ecuaciones de aguas poco profundas. Por otro lado, se proponen algoritmos que solventan este problema para las ecuaciones de aguas poco profundas con término fuente de fondo basándose en métodos formulados para sistemas hiperbólicos homogéneos [106]. Los esquemas resultantes proporcionan resultados notablemente superiores al esquema ARoe tradicional. Además, se consigue por primera vez garantizar la convergencia de la solución a la solución exacta en presencia de resaltos hidráulicos. El autor subraya la importancia de la reducción de las anomalías numéricas, especialmente si se trabaja con esquemas numéricos de muy alto orden, ya que capturan y transportan con precisión cualquier oscilación, sea física o numérica.

BIBLIOGRAPHY

[1] L.F. Richardson, The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam, Phil. Trans. R. Soc. London, Series A, 210 (1910) 307–57.

[2] R. Courant, K.O. Friedrichs, H. Lewy, Uber die partiellen Differenzen-gleichungen der Math. Physik, Math. Ann., 100 (1928) 32–74.

[3] R.V. Southwell, Relaxation methods in engineering science. London, UK, Oxford University Press (1940).

[4] J. von Neumann, R.D. Richtmeyer, A method for the numerical calculation on the hydrodynamic shocks, J. Appl. Phys., 21 (1950) 232–7.

[5] P.D. Lax, Weak solution of nonlinear hyperbolic equations and their numerical computation, Commun. Pure Appl. Math., 7 (1954) 159–63

[6] S.K. Godunov, A finite difference method for the computation of discontinuous solutions of the equations of fluid dynamics, Mat. Sb. 47 (1959) 357–393

[7] Leveque, R. Finite Volume Methods for Hyperbolic Problem. Cambridge University Press, New York, 2002.

[8] P.D. Lax, B. Wendroff, Systems of Conservation Laws, Comm. Pure Appl. Math. 13 (1960) 217–237.

[9] T.Y. Hou, P. LeFloch, Why Non-Conservative Schemes Converge to the Wrong Solutions:Error Analysis, Math. Comput. 62 (1994) 497–530.

[10] B. van Leer, Towards the Ultimate Conservative Difference Scheme I. The Quest for Monotonicity, Lecture Notes in Physics 18 (1973) 163–168.

[11] B. van Leer, Towards the ultimate conservative difference scheme V. A second order sequel to Godunov's method, J. Comput. Phys. 32 (1979) 101–136.

[12] A. Harten, High resolution schemes for hyperbolic conservation laws, J. Comput. Phys. 49 (1983) 357–393.

[13] P. Colella and P.R. Woodward, The piecewise parabolic method (PPM) for gas dynamical simulations, J. Comput. Phys. 54 (1984) 174–201.

[14] J. W. Gibbs, Fourier's Series, Nature 59 (1898).

[15] P.L. Roe, Approximate Riemann solvers, parameter vectors, and difference schemes, J. Comput. Phys. 43 (1981) 357–372.

[16] A. Harten, P. Lax and B. van Leer, On upstream differencing and Godunov type methods for hyperbolic conservation laws, SIAM review. 25 (1983) 35–61.

[17] E.F. Toro, M. Spruce, W. Speares, Restoration of the Contact Surface in the HLL Riemann Solver, Technical Report COA-9204, College of Aeronautics, Cranfield Institute of Technology, UK, 114 (1992).

[18] E.F. Toro, M. Spruce, W. Speares, Restoration of the Contact Surface in the HLL-Riemann Solver, Shock Waves 4 (1994) 25–34.

[19] D.L. George. Augmented Riemann solvers for the shallow water equations over variable topography with steady states and inundation. J. Comput. Phys. 227 (2008) 3089–3113.

[20] J. Murillo, P. García-Navarro, Weak solutions for partial differential equations with source terms: application to the shallow water equations, J. Comput. Phys. 229 (2010) 4327–4368.

[21] J. Murillo, J. Burguete, P. Brufau, and P. García-Navarro. The influence of source terms on stability, accuracy and conservation in two-dimensional shallow flow simulation using triangular finite volumes, Int. J. Numer. Meth. Fluids (2007) 54 543–590.

[22] J. Murillo, P. García-Navarro, Augmented versions of the HLL and HLLC Riemann Solvers including source terms in one and two dimensions for shallow flow applications, J. Comput. Phys. 231 (2012) 6861–6906.

[23] J. Murillo, P. García-Navarro, Augmented Roe's approaches for Riemann problems including source terms: definition of stability region with application to the shallow water equations with rigid and deformable bed. In M. E. Vázquez-Cendón and A. Hidalgo and P. García-Navarro and L. Cea, eds., Numerical Methods for Hyperbolic Equations. Theory and Applications, pages 149–154. Taylor-Francis Group, 2013.

[24] A. Harten, B. Engquist, S. Osher, S.R. Chakravarthy, Uniformly high order accuracy essentially non-oscillatory schemes III, J. Comput. Phys. 71 (1987) 231–303.

[25] X.-D. Liu, S. Osher, T. Chan, Weighted essentially non-oscillatory schemes, J. Comput Phys. 115 (1994) 200–212.

[26] C.-W. Shu, Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws, in Advanced Numerical Aproximation of Nonlinear Hyperbolic Equations, edited by B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor, Lect. Notes in Math. Springer-Verlag, Berlin/New York, 1998 Vol. 1697.

[27] T. Schwartzkopff, M. Dumbser, C.-D. Munz, Fast high order ADER schemes for linear hyperbolic equations, J. Comput. Phys. 197 (2004) 532–539.

[28] E.F. Toro, R.C. Millington, and L.A.M. Nejad. Primitive upwind methods for hyperbolic partial differential equations. In C. H. Bruneau, editor, Sixteenth International Conference on Numerical Methods for Fluid Dynamics. Lecture Notes in Physics, pages 421–426. Springer-Verlag, 1998.

[29] E.F. Toro, R.C. Millington, and L.A.M. Nejad. Towards very high order Godunov schemes. In E. F. Toro, editor, Godunov Methods. Theory and Applications, pages 907–940. Kluwer/Plenum Academic Publishers, 2001.

[30] E.F. Toro, V.A. Titarev, Solution of the generalised Riemann problem for advection-reaction equations, Proc. Roy. Soc. London A 458 (2002) 271–281.

[31] C.E. Castro, E.F. Toro, Solvers for the high-order Riemann problem for hyperbolic balance laws, J. Comput. Phys. 227 (2008) 2481–2513.

[32] E.F. Toro, G. Montecinos, Implicit, semi-analytical solution of the generalized Riemann problem for stiff hyperbolic balance laws, J. Comput. Phys. 303 (2015) 146–172.

[33] M. Dumbser, C. Enaux, E.F. Toro, Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws, J. Comput. Phys., 227 (2008) 3971–4001.

[34] C. R. Goetz, M. Dumbser, A Novel Solver for the Generalized Riemann Problem Based on a Simplified LeFloch–Raviart Expansion and a Local Space–Time Discontinuous Galerkin Formulation. Journal of Scientific Computing, 69 (2016) 805–840.

[35] G. Montecinos, C. E. Castro, M. Dumbser, E. F. Toro, Comparison of solvers for the generalized Riemann problem for hyperbolic systems with source terms, J. Comput. Phys. 231 (2012) 6472–6494.

[36] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation, National topical meeting on mathematical models and computational techniques for analysis of nuclear systems, (1973).

[37] B. Cockburn and Chi-Wang Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework, Math. Comput., 52 (1989) 411–435.

[38] B. Cockburn and Chi-Wang Shu. TVB runge-kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case, Math. Comput., 54 (1990) 545–581.

[39] E.F. Toro, V.A. Titarev, ADER schemes for scalar hyperbolic conservation laws with source terms in three space dimensions, J. Comput. Phys. 202 (1) (2005) 196–215.

[40] E.F. Toro, V.A. Titarev, Derivative Riemann solvers for systems of conservation laws and ADER methods, J. Comput Phys. 212 (1) (2006) 150–165.

[41] E.F. Toro, V.A. Titarev, ADER schemes for three-dimensional non-linear hyperbolic systems, J. Comput Phys. 204 (2) (2005) 715–736.

[42] O. Zanotti, F. Fambri, M. Dumbser, A. Hidalgo, Space–time adaptive ADER discontinuous Galerkin finite element schemes with sub-cell finite volume limiting, Comput. Fluids. 118 (2015) 204–224.

[43] M. Dumbser, Arbitrary High Order Schemes for the Solution of Hyperbolic Conservation Laws in Complex Domains, PhD thesis, Institut fur Aero und Gasdynamik, Universitat Stuttgart, Germany (2005)

[44] M. Dumbser, C.D. Munz, ADER Discontinuous Galerkin Schemes for Aeroacoustics, Comptes Rendus Mécanique, 333 (2005) 683–687 .

[45] M. Käser, A. Iske, Adaptive ADER Schemes for the Solution of Scalar Non-Linear Hyperbolic Problems, J. Comput. Phys. 205 (2005) 486–508.

[46] M. Dumbser, I. Peshkov, E. Romenski, O. Zanotti, High order ADER schemes for a unified first order hyperbolic formulation of continuum mechanics: viscous heat-conducting fluids and elastic solids, J. Comput. Phys. 314 (2016) 824–862.

[47] M. Dumbser, M. Käser and E.F. Toro, An arbitrary high-order Discontinuous Galerkin method for elastic waves on unstructured meshes-V. Local time stepping and p-adaptivity, Geophys. J. Int. 171 (2007) 695–717.

[48] M. Dumbser, M. Castro, C. Parés, E. F. Toro, ADER schemes on unstructured meshes for nonconservative hyperbolic systems: Applications to geophysical flows, Comput. Fluids. 38 (2009) 1731–1748.

[49] C.E. Castro, E.F. Toro and M. Käser, ADER scheme on unstructured meshes for shallow water: simulation of tsunami waves, Geophys. J. Int. 189 (2021) 1505–1520.

[50] G. Vignoli, V.A. Titarev, E.F. Toro, ADER schemes for the shallow water equations in channel with irregular bottom elevation, J. Comput. Phys. 227 (2008) 2463–2480.

[51] G. Montecinos, C.E. Castro, M. Dumbser and E.F. Toro, Comparison of solvers for the generalized Riemann problem for hyperbolic systems with source terms, J. Comput. Phys. 231 (2012) 6472 – 6494.

[52] E. F. Toro, A. Siviglia, PRICE: primitive centred schemes for hyperbolic systems, International journal for numerical methods in fluids, 42 (2003) 1263–1291.

[53] A. Canestrelli, A. Siviglia, M. Dumbser, E. F. Toro, Well-balanced high-order centered schemes for non-conservative hyperbolic systems. Applications to shallow water equations with fixed and mobile bed, Adv. Water Resour. 32 (2009) 834–844.

[54] P. Garcia-Navarro, F. Alcrudo, 1D Open Channel Flow Simulation using TVD Mccormack Scheme, J. Hydraul. Engrg., ASCE 118 (1992) 1359–1373.

[55] L. Fraccarollo, E.F. Toro, Experimental and Numerical Assessment of the Shallow Water Model for Two-Dimensional Dam-Break Type Problems, J. Hydraul. Res., 33 (1995) 843–864.

[56] E.F. Toro, Shock-Capturing Methods for Free-Surface Shallow Flows, Wiley and Sons Ltd (2001).

[57] P.L. Roe, Characteristic Based Schemes for the Euler Equations, Ann. Rev. Fluid Mech., Annual Reviews, (1986) 337–365

[58] A. Bermudez and M.E. Vázquez-Cendón, Upwind methods for hyperbolic conservation laws with source terms, Comput. Fluids. 23 (1994) 1049–1071.

[59] J.M. Greenberg, A.Y. Leroux, A well-balanced scheme for the numerical processing of source terms in hyperbolic equations, SIAM J. Numer. Anal. 33 (1996) 1–16.

[60] P. García-Navarro, M.E. Vázquez-Cendón. On numerical treatment of the source terms in the shallow water equations, Comput. and Fluids. 29 (2000) 951–979.

[61] J. Burguete and P. García-Navarro. Efficient construction of high-resolution TVD conservative schemes for equations with source terms: application to shallow water flows, Int. J. Numer. Meth. Fluids 37 (2001) 209–248.

[62] J. Burguete and P. García-Navarro. Implicit schemes with large time step for non-linear equations: application to river flow hydraulics, Int. J. Numer. Meth. Fluids 46 (2004) 607–636.

[63] A. Chinnayya, A. Y. LeRoux, N. Seguin, A well-balanced numerical scheme for the approximation of the shallow-water equations with topography: the resonance phenomenon, Int. J. Finite Volumes 1 (2004) 1–33.

[64] R. Bernetti, V.A. Titarev, E.F. Toro, Exact solution of the Riemann problem for the shallow water equations with discontinuous bottom geometry, J. Comput. Phys. 227 (2008) 3212–3243.

[65] G. Rosatti, L. Begnudelli, The Riemann Problem for the one-dimensional, free-surface Shallow Water Equations with a bed step: theoretical analysis and numerical simulations, J. Comput. Phys. 229 (2010) 760-787.

[66] P.G. LeFloch, M.D. Thanh, The Riemann problem for shallow water equations with discontinuous topography, Commun. Math. Sci. 5 (2007) 865–885.

[67] P.G. LeFloch, M.D. Thanh, A Godunov-type method for the shallow water equations with discontinuous topography in the resonant regime, J. Comput. Phys. 230 (2011) 7631–7660.

[68] X. Yulong and S. Chi-Wang , High order finite difference WENO schemes with the exact conservation property for the shallow water equations, J. Comput. Phys. 208 (2005) 206–227.

[69] F. Alcrudo, F. Benkhaldoun, Exact solutions to the Riemann problem of the shallow water equations with a bottom step, Comput. Fluids 30 (2001) 643–671.

[70] B.F. Sanders, D.A. Jaffe and A.K. Chu. Discretization of integral equations describing flow in nonprismatic channels with uneven beds, J. Hydraul. Eng–ASCE 129(3) (2003) 235–244.

[71] G. Kesserwani, R. Ghostine, J. Vazquez, A. Ghenaim and R. Mosé. Application of a second-order Runge-Kutta discontinuous Galerkin scheme for the shallow water equations with source terms, Int. J. Numer. Meth. Fluids 56 (2008) 805–821.

[72] M. Catella, E. Paris and L. Solari. Conservative scheme for numerical modeling of flow in natural geometry, J. Hydraul. Eng–ASCE 134(6) (2008) 736–748.

[73] S.H. Lee and N.G. Wright. Simple and efficient solution of the shallow water equations with source terms, Int. J. Numer. Meth. Fluids 63 (2010) 313–340.

[74] V. Caleffi, A. Valiani, G. Li, A comparison between bottom-discontinuity numerical treatments in the DG framework, Appl. Math. Model. (2015).

[75] Noelle, S., Xing, Y. and Shu, C., High-order well-balanced finite volume WENO schemes for shallow water equation with moving water, J. Comput. Phys. 226 (2007) 29–58.

[76] U.S. Fjordholm, S. Mishra, E. Tadmor,Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography, J. Comput. Phys. 230 (2011) 5587–5609.

[77] M.J. Castro Díaz , J.A. López-García, Carlos Parés, High order exactly well-balanced numerical methods for shallow water systems, J. Comput. Phys. 246 (2013) 242–264.

[78] Y. Xing,Exactly well-balanced discontinuous Galerkin methods for the shallow water equations with moving water equilibrium, J. Comput. Phys. 257 (2014) 536–553.

[79] V. Caleffi, A. Valiani, Well balancing of the SWE schemes for moving-water steady flows, J. Comput. Phys., 342 (2017) 85–116.

[80] Y. Cheng, A. Chertock, M. Herty, A. Kurganov, A New Approach for Designing Moving-Water Equilibria Preserving Schemes for the Shallow Water Equations. Submitted preprint.

[81] J. Murillo, P. García-Navarro, Energy balance numerical schemes for shallow water equations with discontinuous topography , J. Comput. Phys. 236 (2012) 119–142.

[82] J. Murillo, P. García-Navarro, Accurate numerical modeling of 1D flow in channels with arbitrary shape. Application of the energy balanced property, J. Comput. Phys. 260 (2014) 222–248.

[83] A. Navas-Montilla, J. Murillo, Energy balanced numerical schemes with very high order. The Augmented Roe Flux ADER scheme. Application to the shallow water equations, J. Comput. Phys. 290 (2015) 188–218

[84] A. Navas-Montilla, J. Murillo, Asymptotically and exactly energy balanced augmented flux-ADER schemes with application to hyperbolic conservation laws with geometric source terms, J. Comput. Phys. 317 (2016) 108–147.

[85] A. Navas-Montilla, J. Murillo, *Overcoming numerical shockwave anomalies using energy balanced numerical schemes. Application to the Shallow Water Equations with discontinuous topography*, J. Comput. Phys. 340 (2017) 575–616.

[86] J. Murillo and A. Navas-Montilla, A comprehensive explanation and exercise of the source terms in hyperbolic systems using Roe type solutions. Application to the 1D-2D shallow water equations, Adv. Water Resour. 98 (2016) 70–96.

[87] E. Audusse, R. Klein, D. D. Nguyen, S. Vater, Preservation of the discrete geostrophic equilibrium in shallow water flows. Finite Volumes for Complex Applications VI Problems and Perspectives, 4 (2011) 59–67.

[88] A. Chertock, M. Dudzinski, A. Kurganov and M. Lukacova-Medvidova, Well-balanced schemes for the shallow water equations with Coriolis forces. Submitted to Numer. Math., (2014).

[89] A. C. Kuo, L. M. Polvani, Nonlinear geostrophic adjustment, cyclone/anticyclone asymmetry, and potential vorticiy rearrangement. Phys. of Fluids, 12 (2000) 1087–1100.

[90] F. Bouchut, J. Le Sommer, V. Zeitlin, Frontal geostrophic adjustment and nonlinear wave phenomena in one-dimensional rotating shallow water. Part 2. High-resolution numerical simulations. J. Fluid Mech. 514 (2004) 35–63.

[91] N. Pankratz, J. R. Natvig, B. Gjevik, S. Noelle, High-order well-balanced finite-volume schemes for barotropic flows: Development and numerical comparisons, Ocean Model., 18 (2007) 53–79.

[92] M. J. Castro, J. A. Lopez and C. Pares, Finite volume simulation of the geostrophic adjustment in a rotating shallow-water system. SIAM J. Sci. Comput., 31 (2008) 444–477.

[93] E. Audusse, R. Klein, A. Owinoh, Conservative discretization of Coriolis force in a finite volume framework, J. Comp. Phys. 228 (2009) 2934–2950.

[94] F. Bouchut, V. Zeitlin, A robust well-balanced scheme for multi-layer shallow water equations. Discrete Continuous Dyn. Syst. Ser. B 13 (2010) 739–758.

[95] A. Siviglia, G. Stecca, D. Vanzo, G. Zolezzi, E.F. Toro, M. Tubino, Numerical modelling of two-dimensional morphodynamics with applications to river bars and bifurcations, Adv. Water Resour., 52 (2013) 243–260.

[96] C. Juez, J. Murillo, and P. García-Navarro, A 2D weakly-coupled and efficient numerical model for transient shallow flow and movable bed, Adv. Water Resour., 71 (2014) 93–109.

[97] V. Caleffi, A. Valiani, A. Bernini, High-order balanced CWENO scheme for movable bed shallow water equations, Adv. Water Resour., 30 (2007) 730–741.

[98] C. Juez, J. Murillo, and P. García-Navarro, 2D simulation of granular flow over irregular steep slopes using global and local coordinates, J. Comput. Phys., 255 (2013) 166–204.

[99] K.M. Peery and S.T. Imlay, Blunt-body flow simulations, AIAA paper, (1988) 88–2924.

[100] K. Kitamura, E. Shima, and P.L. Roe, Three-dimensional carbuncles and euler fluxes, Proceedings of the 48th AIAA Aerospace Sciences Meeting (2010).

[101] T.W. Roberts, The behavior of flux difference splitting schemes near slowly moving shock waves, J. Comput. Phys., 90 (1990) 141–160.

[102] M. Arora, P.L. Roe, On postshock oscillations due to shock capturing schemes in unsteady flows, J. Comput. Phys., 130 (1997) 25–40.

[103] W.F. Noh, Errors for calculations of strong shocks using an artificial viscosity and an artificial heat flux, J. Comput. Phys., 72 (1987) 78–120.

[104] G. Cameron, An analysis of the errors caused by using artificial viscosity terms to represent steady-state shock waves, J. Comput. Phys. 1 (1966) 1–20.

[105] A. Emery, An evaluation of several differencing methods for inviscid fluid flow problems, J. Comput. Phys., 2 (1968) 306–331.

[106] D. W. Zaide, Numerical Shockwave Anomalies, PhD thesis, Aerospace Engineering and Scientific Computing, University of Michigan, 2012.

[107] S. Karni, S. Canic, Computations of slowly moving shocks, J. Comput. Phys., 136 (1997) 132–139.

[108] S. Jin, J. G. Liu, The Effects of Numerical Viscosities, J. Comput. Phys., 126 (1996) 373–389.

[109] M. H. Carpenter, J. H. Casper, Accuracy of Shock Capturing in Two Spatial Dimensions, AIAA Journal, 37 (1999) 1072–1079.

[110] N. K. Yamaleev, M. H. Carpenter, On accuracy of adaptive grid methods for captured shocks, J. Comput. Phys., 181 (2002) 280–316.

[111] Y. Stiriba, R. Donat, A numerical study of postshock oscillations in slowly moving shock waves, Comput. Math. with Appl., 46 (2003) 719–739.

[112] E. Johnsen, S. K. Lele, Numerical errors generated in simulations of slowly moving shocks, Center for Turbulence Research, Annual Research Briefs, (2008) 1–12.

[113] D. W. Zaide, P. L. Roe, Flux functions for reducing numerical shockwave anomalies. ICCFD7, Big Island, Hawaii, (2012) 9–13.

[114] E. Godlewski, P.-A. Raviart Numerical Approximation of Hyperbolic Systems of Conservation Laws. Springer Science and Business Media, Berlin, 2013.

[115] E.F. Toro. Riemann Solvers and Numerical Methods for Fluid Dynamics. Springer-Verlag, Berlin, 1999.

[116] A.K. Henrick, T.D. Aslam, J.M. Powers, Mapped weighted essentially non-oscillatory schemes: achieving optimal order near critical points, J. Comput Phys. 207 (2005) 542–567.

[117] R. Borges, C. Carmona, B. Costa, W.S. Don, An improved weighted essentially non-oscillatory scheme for hyperbolic conservation laws, J. Comput Phys. 227 (6) (2008) 3101–3211.

[118] G.S. Jiang, C.W. Shu, Efficient implementation of weighted ENO schemes, J. Comput Phys. 126 (1996) 202–228.

[119] M. Castro, B. Costa, W.S. Don, High order weighted essentially non-oscillatory WENO-Z schemes for hyperbolic conservation laws, J. Comput Phys. 230 (6) (2011) 1766–1792.

[120] S. Zhao, N. Lardjane, I. Fedioun, Comparison of improved finite-difference WENO schemes for the implicit large eddy simulation of turbulent non-reacting and reacting high-speed shear flows, Comput. Fluids, 95 (2014) 74-87.

[121] Y. Ha, C.H. Kim, Y. J. Lee, J. Yoon, An improved weighted essentially non-oscillatory scheme with a new smoothness indicator, J. Comput Phys. 232 (2013) 68–86.

[122] M. Ben-Artzi, J. Falcovitz, A second order Godunov-type scheme for compressible fluid dynamics, J. Comput. Phys. 55 (1984) 1–32.

[123] P. Le Floch, P. A. Raviart, An Asymptotic Expansion for the Solution of the Generalized Riemann Problem. Part 1: General Theory. Ann. Inst. Henri Poincare. Analyse non Lineare, 5 (1988) 179—207.

[124] J.B. Cheng, E. F. Toro, S. Jiang, W. Tang, A sub-cell WENO reconstruction method for spatial derivatives in the ADER scheme , J. Comput Phys. 251 (2013) 53–80.

[125] E.F. Toro, Riemann solvers and numerical methods for fluid dynamics: a practical introduction, third ed., Springer-Verlag, Berlin, Heidelberg, 2009.

[126] Dumbser, M., Munz, C. D, Building blocks for arbitrary high order discontinuous Galerkin schemes. J. of Sci. Comp., 27 (2006) 215–230.

[127] L. Krivodonova and R. Qin, An analysis of the spectrum of the discontinuous galerkin method, Appl. Num. Math., 64 (2013) 1–18.

[128] D. Gottlieb and E. Tadmor, The CFL condition for spectral approximations to hyperbolic initial-boundary value problems, Math. Comp., 56 (1994) 565–588.

[129] Charles A. Doswell III, 1984: A Kinematic Analysis of Frontogenesis Associated with a Nondivergent Vortex, J. Atmospheric Sci. 41 (1984) 1242–1248.

[130] R. Davies-Jones, Comments on " Kinematic analysis of frontogenesis associated with a nondivergent vortex", J. Atmospheric Sci. 42 (1985) 2073–2075.

[131] V. A. Titarev. Derivative Riemann Problem and ADER Schemes. PhD thesis, Department of Mathematics, University of Trento, Italy, 2005.

[132] T. Zhou, Y. Li and C. W. Shu, Numerical comparison of WENO finite volume and Runge–Kutta discontinuous Galerkin methods, J. Sci. Comp., 16 (2001) 145–171.

[133] Cunge, J.A., Holly, F.M. and Verwey, A. Practical Aspects of Computational River Hydraulics, Pitman,London, U.K, (1980)

[134] Valiani, A., Caleffi, V, Momentum balance in the shallow water equations on bottom discontinuities, Adv. Water Resour., 100 (2017) 1–13.

[135] G. Rosatti, L. Begnudelli, The Riemann Problem for the one-dimensional, free-surface Shallow Water Equations with a bed step: theoretical analysis and numerical simulations, J. Comput. Phys., 229 (2010) 760-787.

[136] V. Caleffi, A. Valiani, A. Bernini, Fourth-order balanced source term treatment in central WENO schemes for shallow water equations, J. Comput. Phys., 218 (2006) 228–245.

[137] P. Woodward, P. Colella, The numerical simulation of two-dimensional fluid flow with strong shocks, J. Comput. Phys., 54 (1984) 115–173.

[138] W.Y. Sun, O.M. Sun, Numerical simulation of Rossby wave in shallow water, Comput. Fluids, 76 (2013) 116–127.

[139] J.P. Boyd, Equatorial solitary waves. Part-1: Rossby solitons, J. Phys. Oceanogr. 10 (1980) 699–717.

[140] A.V. Fedorov and W.K. Melville, Kelvin fronts on the equatorial thermocline, J. Phys. Oceanogr., 30 (2000) 1692–1705.

[141] D. Y. Le Roux, V. Rostand, and B. Pouliot, Analysis of Numerically Induced Oscillations in 2D Finite Element Shallow Water Models Part I: Inertia Gravity Waves, SIAM J. Sci. Comp. 29 (2007) 331–360.

[142] B.A. Tuna, E. Tinar, Rockwell, Shallow flow past a cavity: globally coupled oscillations as a function of depth, D. Exp Fluids (2013) 1586.

[143] C. Juez, I. Buhlmann, G. Maechler, A.J. Schleiss, M.J. Franca, Transport of suspended sediments under the influence of bank macro–roughness, Earth Surface Processes and Landforms, 43 (2018) 271–284.

[144] R. S. Myong, P.L. Roe, Shock waves and rarefaction waves in magnetohydrodynamics. part 2. the mhd system. J. Plasma Ph., 58 (1997) 21–552.

[145] T.J. Barth, Some Notes on Shock-Resolving Flux Functions Part 1: Stationary Characteristics, NASA TM-101087 (1989).

[146] P.L. Roe, Fluctuations and Signals - A Framework for Numerical Evolution Problems, Numerical Methods for Fluid Dynamics, edited by K. W. Morton, and M. J. Baines, Academic Press, New York, (1982) 219–257.

[147] K. Kitamura, P.L. Roe, F. Ismail, Evaluation of Euler fluxes for hypersonic flow computations, AIAA Journal, 47 (2009) 44–53

[148] P. Roe, Is Discontinuous Reconstruction Really a Good Idea? J. Sci. Comp., 73 (2017) 1094–1114.

[149] A. Harten, High resolution schemes for hyperbolic conservation laws, J. Comput Phys. 49 (1983) 357–393.

[150] A. Harten, B. Enquist, S. Osher, S. Chakravarthy , Uniformly high order accurate essentially non-oscillatory schemes, J. Comput Phys. 131 (1997) 3–7

[151] G.S. Jiang, C.-W. Shu, Efficient implementation of weighted ENO schemes, J. Comput Phys. 126 (1996) 202-–228.

[152] Y. Liu., C-W. Shu, M. Zhang, On the positivity of linear weights in WENO approximations. Springer-Verlag, 25 (2009) 503–538.

[153] E. Carlini, R. Ferretti and G. Russo, A weighted essentially non-oscillatory, large time-step scheme for Hamilton-Jacobi equations, SIAM Journal of Scientific Computing, 2005.

LIST OF FIGURES

LIST OF TABLES

A WENO RECONSTRUCTION PROCEDURES

The preservation of high accuracy in both space and time for system of conservation laws with source terms has been and is a major step in the resolution of complex flows. If a reconstruction procedure is performed to provide a high order approximation of the conserved variables, fluxes and source terms, it must be considered that discontinuous solutions may be present. Discontinuities may introduce spurious oscillations in the numerical solution and the choice of a proper reconstruction technique is decisive for their rejection.

In this chapter, the WENO method is introduced. The acronym of WENO stands for *Weighted Essentially Non-Oscillatory*. Its name arises from the way data is reconstructed and how the solution behaves around discontinuities: any possible oscillatory behavior is eliminated, leading to a very stable non-oscillatory reconstruction.

Before the appearance of the WENO method, many other approaches addressed the issue of the generation of spurious oscillations in finite differences schemes, leading to the family of total-variation diminishing (TVD) schemes [149]. Later on, in the search of appropriate reconstruction techniques, the essentially non-oscillatory (ENO) method was proposed by Harten et al. [150]. Based on the definition of an smoothness indicator, the ENO method selects among different candidate stencils. The stencil in which the solution is smoothest is selected, avoiding oscillatory effects produced by the discontinuities. Founded in the ENO approach, the WENO method was then developed by Liu et al. in [25], allowing a r-th order ENO reconstruction be transformed into an $(r+1)$-th order WENO reconstruction.

The WENO reconstruction procedure uses a dynamic set of stencils where lower order polynomials are constructed first. These lower order polynomials are combined either to create a higher order polynomial in smooth regions (optimal reconstruction) or an off-center reconstruction able to capture discontinuities in non-smooth regions. The definition of a smoothness indicator allows to distinguish between both cases. Also, it is desirable that the selected indicator preserves the desired order of accuracy in smooth regions while retaining the essentially non-oscillatory property. Focusing in the preservation of the order of accuracy, Jiang and Shu [151] proposed a smoothness indicator linked to each small stencil, leading to an improved 5-th order WENO method. This indicator was established as the basis of an arbitrary order WENO method, referred here to as WENO-JS.

We will first review simple data reconstruction in 1D, focusing on the WENO method afterward. The two first sections of this chapter are based on the work of C.W. Shu presented in [26].

A.1 Interpolation and reconstruction in 1D

In this section, the problem of data reconstruction at an arbitrary point inside a cell by means of polynomial interpolation when departing from cell averages is considered.

The function $u(x)$ will be defined departing from the starting data, that will be considered as the average value of this function in each cell. The definition of $u(x)$ is useful for the derivation of the reconstruction procedure but its analytical expression will be unknown in most cases. The computational grid, shown in Figure A.1, is composed by N cells as

$$a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < ... < x_{N-\frac{1}{2}} < x_{N+\frac{1}{2}} = b \tag{A.1}$$

with cells and cell sizes defined by

$$I_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \tag{A.2}$$

$$\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} \equiv \text{constant} \tag{A.3}$$

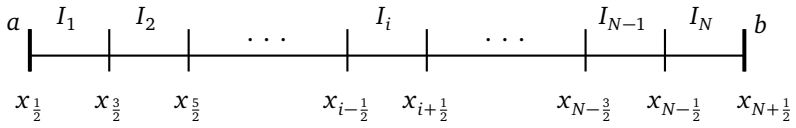


Figure A.1: Mesh discretization

With the previous definitions, the starting data set is now defined as the the average value of the function $u(x)$ in each cell

$$\bar{u}_i = \frac{1}{\Delta x_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(\xi)\, d\xi, \quad i = 1, 2, ..., N \tag{A.4}$$

The problem we face is to find a polynomial $p_r(x)$ of **degree at most** $k-1$ for each cell I_i, such that it is a k**-th order accurate** approximation of the function $u(x)$ inside I_i

$$p_r(x) = u(x) + O\left(\Delta x^k\right), \quad x \in I_i, \quad i = 1, 2, ..., N \tag{A.5}$$

The polynomial in (A.5) provides an approximation to the values of the function at the boundaries of cell I_i when evaluating $p_r(x)$ at $x_{i+\frac{1}{2}}$ and $x_{i-\frac{1}{2}}$, as follows

$$u^+_{i-\frac{1}{2}} = p_r\left(x_{i-\frac{1}{2}}\right), \quad u^-_{i+\frac{1}{2}} = p_r\left(x_{i+\frac{1}{2}}\right), \quad i = 1, 2, ..., N \tag{A.6}$$

Due to the properties of this form of interpolation, the resulting values for the approximation at the cell boundaries (A.6) will be defined as a linear combination of cell averages [26]. This linear combination is given by a set of constants c_{rj} which depend on the polynomial degree and the grid geometry, but not on the function $u(x)$. The expression for the approximation to the values of the function at the cell boundaries is written as

$$u^{-}_{i+\frac{1}{2}} = \sum_{j=0}^{k-1} c_{rj}\bar{u}_{i-r+j}, \qquad u^{+}_{i-\frac{1}{2}} = \sum_{j=0}^{k-1} \tilde{c}_{rj}\bar{u}_{i-r+j} \tag{A.7}$$

with $\tilde{c}_{rj} = c_{r-1,j}$.

The reconstructed values at the cell boundaries are a k-th approximation to those of the function $u(x)$ at these points

$$u^{-}_{i+\frac{1}{2}} = u\left(x_{i+\frac{1}{2}}\right) + O\left(\Delta x^k\right), \qquad u^{+}_{i-\frac{1}{2}} = u\left(x_{i-\frac{1}{2}}\right) + O\left(\Delta x^k\right) \tag{A.8}$$

The derivation of (A.7) is detailed next. First of all, it is necessary to clarify some notions. First, the concept of *stencil* is introduced. A stencil is defined as a group of connected cells. In this section, the reconstruction will be performed using information contained in only one stencil. Therefore, each cell, I_i, will be linked to a stencil $S_r(i)$ composed by cell I_i plus r cells to the left and s cells to the right. Hence, **the number of cells in the stencil** will be $r+s+1$ which **agrees with the order of accuracy of the polynomial for that stencil**, $k = r+s+1$. For all cases, the condition $r,s \geq 0$ must be satisfied. The stencil will be denoted by:

$$S_r(i) = \{I_{i-r}, ..., I_i, ..., I_{i+s}\} \tag{A.9}$$

It is worth mentioning that polynomial $p_r(x)$ refers to stencil $S_r(i)$. Therefore, it is possible to define k $p_r(x)$ independent polynomials of k-th order, with r variable, that will be used to provide information inside cell I_i. This will be used for the WENO reconstruction procedure in the next chapter.

The steps required to generate the reconstructing polynomial departing from cell averages are listed below:

a) Stencil selection.

   Given the cell I_i and the order of accuracy required k, we must first choose a stencil $S_r(i)$ with $k = r+s+1$ cells.

   There is a unique polynomial $p_r(x)$ of degree at most $k-1$ whose cell average value for each cell in the stencil agrees with that of the function $u(x)$ [26]

$$\frac{1}{\Delta x_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p_r(\xi)\,d\xi = \bar{u}_j, \qquad j = i-r, ..., i+s \tag{A.10}$$

b) Definition of the primitive function.

   In order to find the interpolating polynomial $p_r(x)$ of degree $k-1$ and k-th order of accuracy, a new function is introduced. This new function is the primitive function of $u(x)$, denoted by $U(x)$, which is defined as the cumulative integral of $u(x)$ from $-\infty$ to x.

$$U(x) = \int_{-\infty}^{x} u(\xi)\,d\xi \tag{A.11}$$

   For a random location in the grid, i, the value of this cumulative integral at the right boundary of the cell I_i can be computed by the summation of the average values of each cell multiplied by the cell size, from $-\infty$ to the cell I_i, as follows:

$$U\left(x_{i+\frac{1}{2}}\right) = \int_{-\infty}^{x_{i+\frac{1}{2}}} u(\xi)\,d\xi = \sum_{j=-\infty}^{i}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(\xi)\,d\xi = \sum_{j=-\infty}^{i}\bar{u}_j\Delta x_j \tag{A.12}$$

Also, a polynomial $P_r(x)$ is defined as the **unique polynomial of degree at most k which interpolates $U(x)$ with $k+1$-th order of accuracy in $k+1$ nodes** (which are all the cell boundaries in the stencil) and we denote its derivative by $p_r(x)$:

$$p_r(x) = P_r'(x) \tag{A.13}$$

Note that $p_r(x)$ is a polynomial of degree $k-1$ and k-th order, defined by k cells. Polynomial $P_r(x)$ is one order greater, and as the number of cells does not change, $k+1$ interpolation points are necessary. It is worth noticing that this new $k+1$ points are defined in the nodes, even the value of $u(x)$ is not a priori defined at these locations.

Using $P_r'(x)$ it is possible to prove the equality in (A.10)

$$\begin{aligned}
&\frac{1}{\Delta x_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p_r(\xi)\,d\xi = \frac{1}{\Delta x_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} P_r'(\xi)\,d\xi = \frac{1}{\Delta x_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} dP_r(\xi) = \\
&= \frac{1}{\Delta x_j}\left(P_r\left(x_{j+\frac{1}{2}}\right) - P_r\left(x_{j-\frac{1}{2}}\right)\right) \approx \frac{1}{\Delta x_j}\left(U\left(x_{j+\frac{1}{2}}\right) - U\left(x_{j-\frac{1}{2}}\right)\right) = \\
&= \frac{1}{\Delta x_j}\left(\int_{-\infty}^{x_{j+\frac{1}{2}}} u(\xi)\,d\xi - \int_{-\infty}^{x_{j-\frac{1}{2}}} u(\xi)\,d\xi\right) = \frac{1}{\Delta x_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(\xi)\,d\xi = \bar{u}_j
\end{aligned} \tag{A.14}$$

for any $j = i-r, \ldots, i+s$, being j the subscript that indicates the cell of the stencil we are dealing with. The approximation symbol stands for the approximation of $U(x)$ by the interpolating polynomial $P_r(x)$. This interpolation is a $k+1$-th order approximation

$$P_r(x) = U(x) + O\left(\Delta x^{k+1}\right), \qquad x \in I_i \tag{A.15}$$

and that of its derivative, a k-th order approximation

$$P_r'(x) = U'(x) + O\left(\Delta x^{k}\right), \qquad x \in I_i \tag{A.16}$$

Therefore it can be concluded that we must first get $P_r(x)$ by interpolating the primitive function $U(x)$ and then we must take the derivative of $P_r(x)$ to find $p_r(x)$.

c) Lagrange interpolation

In [26], the use of the Lagrange form of the interpolating polynomial is proposed to achieve what it is conveyed in the previous lines. This kind of interpolation is said to be nodal since each weight takes the value of 1 in its node and 0 in the rest of the nodes. Therefore, the result of the interpolation at each node is the value of the function at that node, since the other terms of the summation will be zero and have no contribution. The generic expression for the Lagrange interpolating polynomial, at $b+1$ nodes $(x_0, y(x_0))\ldots(x_b, y(x_b))$, for a function $y(x)$, is as follows:

$$L(x) = \sum_{i=0}^{b} y(x_i) l_i(x) \tag{A.17}$$

with the weighting functions

$$l_i(x) = \prod_{\substack{l=0 \\ l\neq i}}^{b} \frac{(x - x_l)}{(x_i - x_l)} \tag{A.18}$$

The plots of the weighting functions $l_i(x)$ is shown in Figure A.2. In this case, the number of nodes for the interpolation is 4 reaching a 4-th order of accuracy. The cell size Δx is constant. As we can see in Figure A.2, $l_1(x)$ is equal to 1 at $x = 0$, $l_2(x)$ is 1 at $x = 1$ and so on.

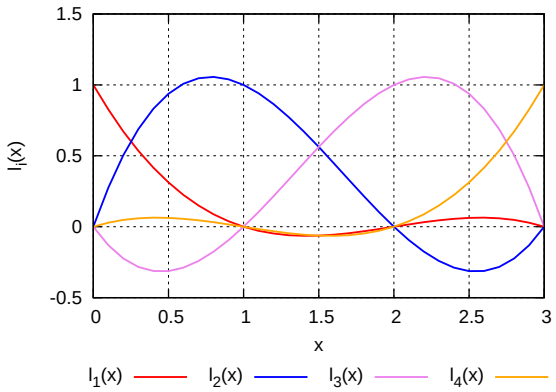


Figure A.2: Weighting functions for $k = 3$, with nodes $x_i = \{0, 1, 2, 3\}$

Now, we write the expression for $P_r(x)$ following the Lagrange interpolating polynomial in (A.17), by imposing the values of function $U(x)$ at the $k+1$ nodes of the stencil $S(i)$:

$$P_r(x) = \sum_{m=0}^{k} U(x_{i-r+m-\frac{1}{2}}) \prod_{\substack{l=0 \\ l\neq m}}^{k} \frac{(x - x_{i-r+l-\frac{1}{2}})}{(x_{i-r+m-\frac{1}{2}} - x_{i-r+l-\frac{1}{2}})} \tag{A.19}$$

For an easier manipulation, the constant value $U(x_{i-r-\frac{1}{2}})$ is going to be subtracted from the previous expression (A.19). This way, the starting point for the calculation of the integral will shift from $-\infty$ to the first wall of the stencil.

$$P_r(x) - U(x_{i-r-\frac{1}{2}}) = \sum_{m=0}^{k} \left(U(x_{i-r+m-\frac{1}{2}}) - U(x_{i-r-\frac{1}{2}}) \right) \prod_{\substack{l=0 \\ l\neq m}}^{k} \frac{(x - x_{i-r+l-\frac{1}{2}})}{(x_{i-r+m-\frac{1}{2}} - x_{i-r+l-\frac{1}{2}})} \tag{A.20}$$

The difference between the primitive function evaluated in any wall of the stencil, $U(x_{i-r+m-\frac{1}{2}})$, and the same function evaluated in the first wall of the stencil, $U(x_{i-r-\frac{1}{2}})$, will be a measure of the cumulative integral from the beginning of the stencil to that wall at $x_{i-r+m-\frac{1}{2}}$. This can be clearly seen in the following expression

$$U(x_{i-r+m-\frac{1}{2}}) - U(x_{i-r-\frac{1}{2}}) = \sum_{j=0}^{m-1} \bar{u}_{i-r+j} \Delta x_{i-r+j} \tag{A.21}$$

Taking the derivative on both terms of (A.20) and noticing the previous equality, we get the expression for the polynomial $p_r(x)$, which performs the reconstruction using the average values of $u(x)$ in the cells, unlike $P_r(x)$, which used boundary values

$$p_r(x) = \sum_{m=0}^{k} \sum_{j=0}^{m-1} \bar{u}_{i-r+j} \Delta x_{i-r+j} \left(\frac{\sum_{\substack{l=0 \\ l\neq m}}^{k} \prod_{\substack{q=0 \\ q\neq m,l}}^{k} \left(x - x_{i-r+q-\frac{1}{2}} \right)}{\prod_{\substack{l=0 \\ l\neq m}}^{k} \left(x_{i-r+m-\frac{1}{2}} - x_{i-r+l-\frac{1}{2}} \right)} \right) \tag{A.22}$$

A simpler expression for $p_r(x)$ can be derived from equation (A.22) taking the cell averages as common factors. The resulting expression represents the reconstructing polynomial as a linear combination of the cell averages as

$$p_r(x) = \sum_{j=0}^{k-1} \left(\sum_{m=j+1}^{k} \frac{\sum_{l=0, l\neq m}^{k} \prod_{q=0, q\neq m,l}^{k} \left(x - x_{i-r+q-\frac{1}{2}} \right)}{\prod_{l=0, l\neq m}^{k} \left(x_{i-r+m-\frac{1}{2}} - x_{i-r+l-\frac{1}{2}} \right)} \right) \bar{u}_{i-r+j} \Delta x_{i-r+j}, \qquad r = 0, ..., k-1 \tag{A.23}$$

If defining

$$C_{rj}^{(k)}(x) = \left(\sum_{m=j+1}^{k} \frac{\sum_{l=0, l\neq m}^{k} \prod_{q=0, q\neq m,l}^{k} \left(x - x_{i-r+q-\frac{1}{2}} \right)}{\prod_{l=0, l\neq m}^{k} \left(x_{i-r+m-\frac{1}{2}} - x_{i-r+l-\frac{1}{2}} \right)} \right) \Delta x_{i-r+j} \tag{A.24}$$

it is possible to express Equation (A.23) as

$$p_r(x) = \sum_{j=0}^{k-1} C_{rj}^{(k)}(x) \bar{u}_{i-r+j}, \qquad r = 0, ..., k-1 \tag{A.25}$$

Where $C_{rj}^{(k)}(x)$ are constants at a given x and provide the weights for the linear combination of cell averages. The superscript k of the linear coefficient $C_{rj}^{(k)}(x)$ stands for the dimension of the stencil where $p_r(x)$ is defined, it is useful to not mix up these coefficients when different stencils are used at the same time.

For the sake of clarity, the evaluation of $C_{rj}^{(k)}(x)$ at $x_{i+\frac{1}{2}}$ or $x_{i-\frac{1}{2}}$ will be denoted as $c_{rj}^{(k)}$ and $\tilde{c}_{rj}^{(k)}$, respectively, as follows

$$C_{rj}^{(k)}(x = x_{i+\frac{1}{2}}) = c_{rj}^{(k)}, \qquad C_{rj}^{(k)}(x = x_{i-\frac{1}{2}}) = \tilde{c}_{rj}^{(k)} \tag{A.26}$$

Expression in (A.25) can expressed in a more compact form as

$$p(r, k, v, \bar{\mathbf{v}}) = p_r(v) = \sum_{j=0}^{k-1} C_{rj}^{(k)}(v) \bar{v}_j, \tag{A.27}$$

where r and k describe the stencil and the position of the reconstruction cell, v stands for the spatial variable and $\bar{\mathbf{v}}$ for the vector of cell averages in the stencil, with components $\bar{v}_j = \bar{u}_{i-r+j}$ for $j = 0, ..., k-1$.

d) Computation of the linear $c_{rj}^{(k)}$ coefficients.

The evaluation of Equation (A.23) at $x_{i+\frac{1}{2}}$ (right boundary of the cell I_i) provides the approximation to the value $u(x_{i+\frac{1}{2}})$, denoted by $u_{i+\frac{1}{2}}$

$$u_{i+\frac{1}{2}} = p_i(x_{i+\frac{1}{2}}) = \sum_{j=0}^{k-1} \left(\sum_{m=j+1}^{k} \frac{\sum_{l=0, l\neq m}^{k} \prod_{q=0, q\neq m,l}^{k} \left(x_{i+\frac{1}{2}} - x_{i-r+q-\frac{1}{2}} \right)}{\prod_{l=0, l\neq m}^{k} \left(x_{i-r+m-\frac{1}{2}} - x_{i-r+l-\frac{1}{2}} \right)} \right) \bar{u}_{i-r+j} \Delta x_{i-r+j} \tag{A.28}$$

As outlined before, this expression may be seen as a summation of constants, denoted by $c_{rj}^{(k)}$, multiplied by the cell averages $\bar{u}_{i-r+j}$, following the equation (A.7). The expression for these constants $c_{rj}^{(k)}$ results from the evaluation of (A.24) at $x_{i+\frac{1}{2}}$

$$c_{rj}^{(k)} = \left(\sum_{m=j+1}^{k} \frac{\sum_{l=0, l\neq m}^{k} \prod_{q=0, q\neq m,l}^{k} \left(x_{i+\frac{1}{2}} - x_{i-r+q-\frac{1}{2}} \right)}{\prod_{l=0, l\neq m}^{k} \left(x_{i-r+m-\frac{1}{2}} - x_{i-r+l-\frac{1}{2}} \right)} \right) \Delta x_{i-r+j} \tag{A.29}$$

If we now rewrite this expression for the particular case of uniform grid (constant Δx), we finally obtain

$$c_{rj}^{(k)} = \sum_{m=j+1}^{k} \frac{\sum_{\substack{l=0 \\ l\neq m}}^{k} \prod_{\substack{q=0 \\ q\neq m,l}}^{k} (r-q+1)}{\prod_{\substack{l=0 \\ l\neq m}}^{k} (m-l)} \tag{A.30}$$

A.2 Weighted Essentially Non-Oscillatory (WENO) reconstruction

In this section, the procedure to construct a WENO reconstruction are provided. Before starting, the reader must notice that there are different sorts of problems for which WENO procedures are designed such as WENO interpolation, WENO integration, WENO approximation to the first derivative and WENO reconstruction [152]. WENO interpolation departs from pointwise information instead of cell-averaged values and it is used in finite difference methods. WENO integration provides an approximation to the integral of a function, given its values at grid points. WENO approximation to the first derivative and WENO reconstruction are equivalent and depart from the cell averages of a function.

The case analyzed in this text is the WENO reconstruction: from cell averages, we have to compute the value of the function at the cell boundaries. This procedure is widely used in the numerical solution of conservation laws.

In the previous section it was detailed how to perform a simple data reconstruction using linear interpolation: from cell averages in the stencil $S_r(i)$, an approximation to the cell boundary values of cell I_i was computed. Now, the procedure goes further and the reconstruction will depend on the shape of the function (on its smoothness), preventing the solution from being oscillatory. Moreover, the starting data set for the interpolation will be broader than in the previous case. Instead of computing the approximation of the function inside one cell with the data stored in only one stencil of k cells, a combination of k different stencils composed of k cells each one will be used. This leads to a reconstruction of $2k-1$-th order of accuracy.

The smoothness of the function inside each stencil is measured by a suitable smoothness indicator. The final reconstruction combines the k different stencils, where the weight associated to each of them is determined by this indicator.

The reconstruction is computed in two steps. The first one is related to the calculation of the coefficients that ensure the equality between the polynomial high order approximation in the big stencil and the linear combination of polynomial lower order approximations in the smaller stencils. These coefficients will be referred to as *optimal weights*. The second step focuses on the calculation of the non-oscillatory weights, modifying the optimal weights by means of the smoothness indicators.

A.2.1 First part: Computation of the optimal weights

Before starting, a reconstruction domain must be chosen. In the previous section, the reconstruction procedure used data from only one stencil (composed of k cells). It was shown that depending on the selection of r, and keeping in mind that $k = r+s+1$, k different $p_r(x)$ polynomials associated to k different stencils could be found to approximate the value of $u(x)$ inside a cell. The keystone of the WENO method is to combine these k different $p_r(x)$ polynomials to generate a $(2k-1)$-th order reconstruction.

To construct a **WENO reconstruction of $(2k-1)$-th order** on the cell $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ for the function $u(x)$, we need the k **different stencils** for $r = 0, ..., k-1$, denoted by $S_r(i)$ and defined as

$$S_r(i) = \{I_{i-r}, ..., I_{i+k-r-1}\}, \qquad r = 0, ..., k-1 \tag{A.31}$$

where $s = k - r - 1$.

Stencils $S_r(i)$ are overlapped on the interval $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, as follows

$$\cap_{r=0}^{k-1} S_r(i) = I_i \tag{A.32}$$

Then, they are used to generate a bigger stencil that will contain all the cells from the smaller stencils, denoted by

$$\mathscr{T}(i) = \cup_{r=0}^{k-1} S_r(i) = \{I_{i-k+1}, ..., I_{i+k-1}\} \tag{A.33}$$

As it was defined in equation (A.10), there is a **unique polynomial $p_r(x)$ associated to each stencil S_r**, which is a **k-th order approximation** to the function $u(x)$ on the stencil $S_r(i)$ if this function is smooth inside it, as follows

$$p_r(x) = u(x) + O\left(\Delta x^k\right), \qquad x \in S_r,\ r = 0, ..., k-1 \tag{A.34}$$

The expression for $p_r(x)$ was derived in equation (A.23). If we evaluate it at $x_{i+\frac{1}{2}}$ or $x_{i-\frac{1}{2}}$, it provides approximations to the cell boundary values

$$u_{i+\frac{1}{2}}^{(r)} = p_r(x_{i+\frac{1}{2}}) = \sum_{j=0}^{k-1} c_{rj}^{(k)} \bar{u}_{i-r+j}, \qquad u_{i-\frac{1}{2}}^{(r)} = p_r(x_{i-\frac{1}{2}}) = \sum_{j=0}^{k-1} \tilde{c}_{rj}^{(k)} \bar{u}_{i-r+j} \tag{A.35}$$

The procedure in (A.23) can be extended to obtain a **polynomial $q(x)$, which is a $2k-1$-th order accurate approximation of the function $u(x)$ on the big stencil $\mathscr{T}(i)$**, denoted by

$$q(x) = \sum_{j=1}^{2k-1} \left(\sum_{m=j+1}^{2k} \frac{\sum_{\substack{l=1 \\ l\neq m}}^{2k} \prod_{\substack{q=1 \\ q\neq m,l}}^{2k} \left(x - x_{i-k+q-\frac{1}{2}}\right)}{\prod_{\substack{l=1 \\ l\neq m}}^{2k} \left(x_{i-k+m-\frac{1}{2}} - x_{i-k+l-\frac{1}{2}}\right)} \right) \bar{u}_{i-k+j} \Delta x_{i-k+j} \tag{A.36}$$

that can be written as

$$q(x) = \sum_{j=1}^{2k-1} C_{rj}^{(2k-1)}(x) \bar{u}_{i-k+j} \tag{A.37}$$

where superscript $2k-1$ only refers to the order of the approximation. The approximation at the right boundary of I_i is denoted as

$$u_{i+\frac{1}{2}} = q(x = x_{i+\frac{1}{2}}) = \sum_{j=1}^{2k-1} c_{k-1,j}^{(2k-1)} \bar{u}_{i-k+j} \tag{A.38}$$

Note that in (A.38), the value of r is fixed, $r = k-1$, as the big stencil $\mathscr{T}(i)$ is always symmetric. Now, the goal is to express the coefficients $c_{k-1,j}^{(2k-1)}$ of the big stencil as a linear combination of the previously computed coefficients $c_{rj}^{(k)}$ obtained for the small stencils. By doing this, it will be possible to express polynomial $q(x)$ in terms of the k $p_r(x)$ polynomials. At a certain point x, the evaluation of $q(x)$ will be expressed as a linear combination of the evaluations provided by $p_r(x)$ and the coefficients that provide this linear combination are the so called optimal weights.

Computation of optimal weights for a 5-th Order WENO reconstruction

For the sake of clarity, a simple example of the procedure for the computation of the optimal weights is given in this text. The details concerning the calculation of the *optimal weights* for a 5-th order WENO

reconstruction, based on 3-cell stencil reconstruction with 3-th order polynomials ($k = 3$) are given below. A uniform grid will be assumed.

The three different stencils are given by

$$\begin{aligned} S_0(i) &= \{I_i, I_{i+1}, I_{i+2}\} \\ S_1(i) &= \{I_{i-1}, I_i, I_{i+1}\} \\ S_2(i) &= \{I_{i-2}, I_{i-1}, I_i\} \end{aligned} \tag{A.39}$$

and the stencil $\mathcal{T}(i)$ can be constructed

$$\mathcal{T}(i) = S_0 \cup S_1 \cup S_2 = \{I_{i-2}, I_{i-1}, I_i, I_{i+1}, I_{i+2}\} \tag{A.40}$$

Stencils in (A.39) and (A.40) are depicted in Figure A.3 for a random cell I_i.

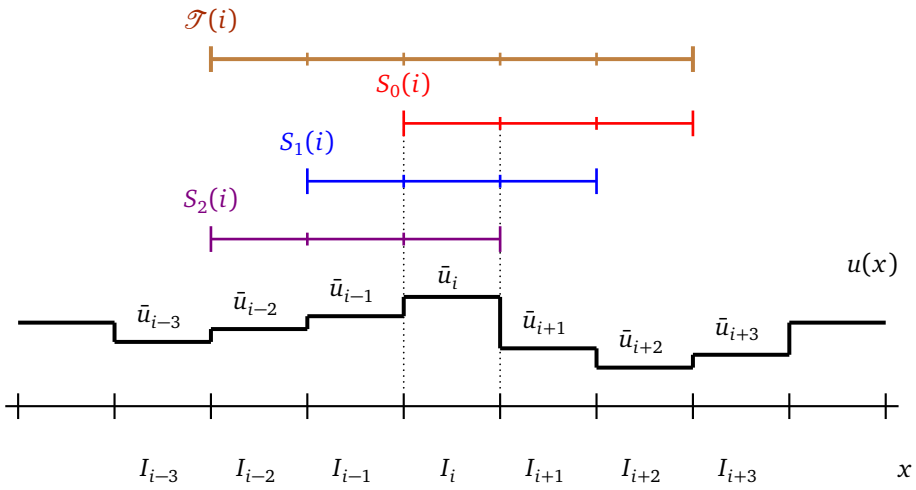


Figure A.3: Stencil combination for a 5-th order WENO reconstruction

For each stencil, boundary values of $u(x)$ in I_i are obtained using (A.35). At the right boundary, the 3 polyomials $p_r(x)$ provide the following approximations

$$u^{(0)}_{i+\frac{1}{2}} = \sum_{j=0}^{2} c^{(3)}_{0j} \bar{u}_{i+j} = \frac{1}{3}\bar{u}_i + \frac{5}{6}\bar{u}_{i+1} - \frac{1}{6}\bar{u}_{i+2} \tag{A.41}$$

$$u^{(1)}_{i+\frac{1}{2}} = \sum_{j=0}^{2} c^{(3)}_{1j} \bar{u}_{i-1+j} = -\frac{1}{6}\bar{u}_{i-1} + \frac{5}{6}\bar{u}_i + \frac{1}{3}\bar{u}_{i+1} \tag{A.42}$$

$$u^{(2)}_{i+\frac{1}{2}} = \sum_{j=0}^{2} c^{(3)}_{2j} \bar{u}_{i-2+j} = \frac{1}{3}\bar{u}_{i-2} - \frac{7}{6}\bar{u}_{i-1} - \frac{11}{6}\bar{u}_i \tag{A.43}$$

which are a 3-th order approximation to the value of the function $u(x)$ at $x_{i+\frac{1}{2}}$ if the function is smooth inside each stencil [152]

$$u^{(0)}_{i+\frac{1}{2}} = u\left(x_{i+\frac{1}{2}}\right) + O\left(\Delta x^3\right), \qquad \text{if } u(x) \text{ is smooth inside } S_0(i) \tag{A.44}$$

$$u^{(1)}_{i+\frac{1}{2}} = u\left(x_{i+\frac{1}{2}}\right) + O\left(\Delta x^3\right), \qquad \text{if } u(x) \text{ is smooth inside } S_1(i) \tag{A.45}$$

$$u^{(2)}_{i+\frac{1}{2}} = u\left(x_{i+\frac{1}{2}}\right) + O\left(\Delta x^3\right), \qquad \text{if } u(x) \text{ is smooth inside } S_2(i) \tag{A.46}$$

When the 5-th order centered reconstruction generated using $q(x)$, Equation (A.35) leads to

$$u_{i+\frac{1}{2}} = \sum_{j=1}^{5} c^{(5)}_{2j} \bar{u}_{i-3+j} = \frac{1}{30}\bar{u}_{i-2} - \frac{13}{60}\bar{u}_{i-1} + \frac{47}{60}\bar{u}_i + \frac{9}{20}\bar{u}_{i+1} - \frac{1}{20}\bar{u}_{i+2} \tag{A.47}$$

which is a 5-th order approximation to the value of the function $u(x)$ at the boundary $x_{i+\frac{1}{2}}$ if the function is smooth inside $\mathscr{T}(i)$

$$u_{i+\frac{1}{2}} = u\left(x_{i+\frac{1}{2}}\right) + O\left(\Delta x^5\right) \tag{A.48}$$

The 5-th order reconstruction in (A.47) may also be expressed as a convex combination of the 3-th order approximations, $u^{(r)}_{i+\frac{1}{2}}$ in (A.41-A.43). The coefficients that determine this combination will be unique and denoted by γ_0, γ_1, γ_2, giving

$$u_{i+\frac{1}{2}} = \gamma_0 u^{(0)}_{i+\frac{1}{2}} + \gamma_1 u^{(1)}_{i+\frac{1}{2}} + \gamma_2 u^{(2)}_{i+\frac{1}{2}} \tag{A.49}$$

These coefficients are called *optimal weights*. They can be easily computed by imposing the equality between (A.47) and (A.49) as

$$\begin{aligned} \frac{1}{30}\bar{u}_{i-2} - \frac{13}{60}\bar{u}_{i-1} + \frac{47}{60}\bar{u}_i + \frac{9}{20}\bar{u}_{i+1} - \frac{1}{20}\bar{u}_{i+2} = \gamma_0\left(\frac{1}{3}\bar{u}_i + \frac{5}{6}\bar{u}_{i+1} - \frac{1}{6}\bar{u}_{i+2}\right) + \\ +\gamma 1\left(-\frac{1}{6}\bar{u}_{i-1} + \frac{5}{6}\bar{u}_i + \frac{1}{3}\bar{u}_{i+1}\right) + \gamma_2\left(\frac{1}{3}\bar{u}_{i-2} - \frac{7}{6}\bar{u}_{i-1} - \frac{11}{6}\bar{u}_i\right) \end{aligned} \tag{A.50}$$

From equation (A.50) we can obtain the 3 coefficients in (A.49) formulating 2 different equations to satisfy the equality of weights for 2 cell averages in this particular case, and one more equation to satisfy the unit sum of the weights $\sum_{r=0}^{k-1} \gamma_r = 1$. The following equations, starting from $\bar{u}_{i-2}$, appear

a) The equality of weights for $\bar{u}_{i-2}$ leads to

$$\begin{aligned} \gamma_2 \frac{1}{3} &= \frac{1}{30} \\ \gamma_2 &= \frac{1}{10} \end{aligned} \tag{A.51}$$

b) And the same for $\bar{u}_{i-1}$ gives

$$\begin{aligned} -\gamma_2 \frac{7}{6} - \gamma_1 \frac{1}{6} &= -\frac{13}{60} \\ \gamma_1 &= \frac{6}{10} \end{aligned} \tag{A.52}$$

c) Finally, an additional equation to fulfill the unit sum allows to compute γ_0

$$\begin{aligned} \gamma_2 + \gamma_1 + \gamma_0 &= 1 \\ \gamma_0 &= \frac{3}{10} \end{aligned} \tag{A.53}$$

The same could be done to compute the weights at the left boundary, denoted by $\tilde{\gamma}_r$.

Generalization of the procedure

The generalization of the procedure and the steps for the computation of the optimal weights for a $(2k-1)$-th order accurate approximation are presented next. First, the $(2k-1)$-th order polynomial, $q(x)$, is expressed in terms of the k-th order polynomials, $p_r(x)$, as

$$q(x) = \sum_{r=0}^{k-1} \Gamma_r(x)\, p_r(x) \tag{A.54}$$

where $\Gamma_r(x)$ are the weights, which are rational functions [152]. Inserting in this equation the expressions for the polynomials $p_r(x)$ in (A.25) and $q(x)$ in (A.37), it yields

$$\sum_{j=1}^{2k-1} C_{k-1,j}^{(2k-1)}(x)\bar{u}_{i-k+j} = \sum_{r=0}^{k-1} \Gamma_r(x) \sum_{j=0}^{k-1} C_{rj}^{(k)}(x)\bar{u}_{i-r+j} \tag{A.55}$$

For the sake of clarity, Equation (A.55) is evaluated at $x = x_{i+\frac{1}{2}}$ though any other point could be used for this derivation. The sought optimal weights are only valid at the point where the evaluation is carried out, in this case at $x = x_{i+\frac{1}{2}}$. As a result of this evaluation, Equation (A.55) becomes

$$\sum_{j=1}^{2k-1} c_{k-1,j}^{(2k-1)}\bar{u}_{i-k+j} = \sum_{r=0}^{k-1} \gamma_r \sum_{j=0}^{k-1} c_{rj}^{(k)}\bar{u}_{i-r+j} \equiv u_{i+\frac{1}{2}} \tag{A.56}$$

where the reconstruction coefficients $c_{k-1,j}^{(2k-1)}$ and $c_{rj}^{(k)}$ are now constant according to (A.26) and with $\gamma_r = \Gamma_r\left(x = x_{i+\frac{1}{2}}\right)$ the sought weights.

From (A.56) and taking into account the condition $\sum_{r=0}^{k-1}\gamma_r = 1$, the following system of equations for the optimal weights γ_r is formulated

$$\mathbf{M}\cdot\gamma = \mathbf{c} \tag{A.57}$$

where $\mathbf{M} \in \mathbb{R}^{k\times k}$, $\gamma \in \mathbb{R}^k$ and $\mathbf{c} \in \mathbb{R}^k$

$$\begin{pmatrix} c_{k-1,0}^{(k)} & 0 & 0 & 0 & \cdots & 0 \\ c_{k-1,1}^{(k)} & c_{k-2,0}^{(k)} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ c_{k-1,k-2}^{(k)} & c_{k-2,k-3}^{(k)} & c_{k-3,k-4}^{(k)} & \cdots & c_{1,0}^{(k)} & 0 \\ 1 & 1 & 1 & \cdots & \cdots & 1 \end{pmatrix} \begin{pmatrix} \gamma_{k-1} \\ \gamma_{k-2} \\ \vdots \\ \gamma_0 \end{pmatrix} = \begin{pmatrix} c_{k-1,0}^{(2k-1)} \\ c_{k-1,1}^{(2k-1)} \\ \vdots \\ c_{k-1,k-1}^{(2k-1)} \\ 1 \end{pmatrix}$$

The components of the matrix $\mathbf{M}$ at a location (α, β) are given by

k	γ_0	γ_1	γ_2	γ_3	γ_4
5	$\frac{5}{126}$	$\frac{20}{63}$	$\frac{10}{21}$	$\frac{10}{63}$	$\frac{1}{126}$
4	$\frac{4}{35}$	$\frac{18}{35}$	$\frac{12}{35}$	$\frac{1}{35}$	
3	$\frac{3}{10}$	$\frac{6}{10}$	$\frac{1}{10}$		
2	$\frac{2}{3}$	$\frac{1}{3}$			
1	1				

Table A.1: Linear coefficients γ_r for $k = 1, 2, 3, 4, 5$

$$M_{\alpha,\beta} = \begin{cases} c^{(k)}_{k-\beta,\alpha-\beta} & \text{if } \alpha \geq \beta, \alpha \neq k \\ 1 & \text{if } \alpha = k \\ 0 & \text{if } \alpha < \beta \end{cases}$$

for $1 \leq \alpha \leq k$ and $1 \leq \beta \leq k$, where α stands for the row and β stands for the column inside the matrix.

Once the coefficients γ_r have been computed, the expression for the approximation to the value of $u(x)$ at the right boundary of I_i can be computed as follows

$$u_{i+\frac{1}{2}} = \sum_{r=0}^{k-1} \gamma_r u^{(r)}_{i+\frac{1}{2}} \tag{A.58}$$

where $u_{i+\frac{1}{2}}$ is $(2k-1)$-th order accurate as long as the function $u(x)$ is smooth inside the stencil $\mathscr{T}(i)$. Table A.1 shows the coefficients γ_r computed using (A.57) for five different values of k, from 1 to 5. They can be used to construct up to a 9-th order reconstruction at $x_{i+1/2}$.

Recall that different optimal weights are obtained depending on the point at which the reconstruction has to be computed. Therefore, the same procedure should be repeated at each point where the reconstruction has to be calculated. For instance, to compute a $(2k-1)$-th order reconstruction at the left boundary of I_i we use

$$u_{i-\frac{1}{2}} = \sum_{r=0}^{k-1} \tilde{\gamma}_r u^{(r)}_{i-\frac{1}{2}} \tag{A.59}$$

and for the particular case of a uniform grid, $\tilde{\gamma}_r = \gamma_{k-1-r}$, for $r = 0, ..., k-1$, due to the symmetry of stencil $\mathscr{T}(i)$.

A.2.2 Second part: Calculation of the non-oscillatory weights

As outlined before, the approximation in (A.58) of the value of the function at the cell boundaries will be $(2k-1)$-th accurate as long as the function is smooth inside the big stencil $\mathscr{T}(i)$. If the function is non-smooth or discontinuous, a lower order of accuracy is reached. Moreover, oscillations will appear due to the presence of discontinuities. This fact motivates the idea of using a modified set of coefficients instead of the optimal weights in order to reduce the weight of the contributions associated to those stencils including discontinuities.

Thus, instead of computing the $(2k-1)$-th order approximation using the optimal weights, γ_r, as in (A.58), non-oscillatory WENO weights, denoted by ω_r, will be used. The non-oscillatory reconstruction of $u(x)$ at the right cell boundary will be computed now as

$$u_{i+\frac{1}{2}} = \sum_{r=0}^{k-1} \omega_r u_{i+\frac{1}{2}}^{(r)} \tag{A.60}$$

where the set of non-oscillatory weights ω_r is sought to provide a linear convex combination using the k different low order reconstructions. Therefore we require

$$\sum_{r=0}^{k-1} \omega_r = 1 \qquad \text{and} \qquad \omega_r \geq 0 \tag{A.61}$$

In the case when $u(x)$ is smooth, both expressions (A.58) and (A.60) are equivalent and provide a $(2k-1)$-th order approximation. According to the properties of the WENO reconstruction [151], WENO weights ω_r are a $k-1$-th order approximation to the optimal weights γ_r,

$$\omega_r = \gamma_r + O(\Delta x^{k-1}) \tag{A.62}$$

in smooth monotone regions. If there is a discontinuity in the stencil, then

$$\omega_r = \gamma_r + O(\Delta x) \tag{A.63}$$

In order to compute the WENO nonlinear weights ω_r, the nonlinear coefficients α_r are formulated first

$$\alpha_r = \frac{\gamma_r}{(\beta_r + \epsilon)^2}, \qquad r = 0, ..., k-1 \tag{A.64}$$

with ϵ a properly defined small parameter (see [153], p.4). The smoothness indicator, β_r, can be computed following [151],

$$\beta_r = \sum_{l=1}^{k-1} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \Delta x^{2l-1} \left(\frac{\partial^l p_r(x)}{\partial x^l} \right)^2 dx, \qquad r = 0, ..., k-1 \tag{A.65}$$

defined as the sum of the L^2 norms of all derivatives of the interpolating polynomial $p_r(x)$ over the interval $\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$. The term Δx^{2l-1} is included to remove Δx-dependent factors in the derivatives of the polynomials.

Once computed, the α_r coefficients are normalized so that their sum is equal to the unity, leading to the desired non-oscillatory weights

$$\omega_r = \frac{\alpha_r}{\sum_{l=0}^{k-1} \alpha_l}, \qquad r = 0, ..., k-1 \tag{A.66}$$

Repeating the procedure for $\tilde{\omega}_r$, which are the non-oscillatory weights associated to the linear weights at the left interface, $\tilde{\gamma}_r$, the WENO reconstruction of $u(x)$ at the cell boundaries will be computed as

$$u_{i+\frac{1}{2}} = \sum_{r=0}^{k-1} \omega_r u_{i+\frac{1}{2}}^{(r)}, \qquad u_{i-\frac{1}{2}} = \sum_{r=0}^{k-1} \tilde{\omega}_r u_{i-\frac{1}{2}}^{(r)} \tag{A.67}$$

Considering again the general case, it was shown that $q(x)$ can be constructed in two different ways: in Equation (A.54) it was defined as a linear combination of lower order polynomials $p_r(x)$ while in Equation (A.37) it was defined by directly constructing a $(2k-1)$-th order polynomial using the general polynomial reconstruction procedure. Following the first approach, it is possible to construct a polynomial $q(x)$ reducing the contribution of certain $p_r(x)$ polynomials which may generate oscillations. This new polynomial is defined, $q_{\text{WENO}}(x)$

$$q_{\text{WENO}}(x) = \sum_{r=0}^{k-1} \Omega_r(x)\, p_r(x) \tag{A.68}$$

where $\Omega_r(x)$ is the general expression for the WENO non-oscillatory weights, for instance

$$\Omega_r\left(x_{i+\frac{1}{2}}\right) = \omega_r \qquad \Omega_r\left(x_{i-\frac{1}{2}}\right) = \tilde{\omega}_r \tag{A.69}$$

For the sake of clarity when moving to 2D WENO reconstruction procedures, expression in (A.68) can be expressed in a more compact form as

$$q(k, v, \mathbf{p}) = q_{\text{WENO}}(v) = \sum_{r=0}^{k-1} \Omega_r(v)\, p_r(v) \tag{A.70}$$

where k is the size of the small stencils, v stands for the spatial variable ($v \equiv x$ in this case) and $\mathbf{p} = \{p_r(v)\}_{r=0,\ldots,k-1}$ for the vector of low order polynomials used to generate the high order reconstruction.

Computation of β_r

In order to compute the smoothness indicator, β_r, using (A.65), a general expression for the the n-th derivative of $p_r(x)$ must be obtained departing from formulation in (A.25), as

$$\frac{\partial^n (p_r(x))}{\partial x^n} = \frac{\partial^n}{\partial x^n}\left(\sum_{j=0}^{k-1} C_{rj}^{(k)}(x)\bar{u}_{i-r+j}\right) = \sum_{j=0}^{k-1} \frac{\partial^n}{\partial x^n}\left(C_{rj}^{(k)}(x)\right)\bar{u}_{i-r+j} \tag{A.71}$$

with $C_{rj}^{(k)}(x)$ defined in (A.24). The term $\frac{\partial^n}{\partial x^n}\left(C_{rj}^{(k)}(x)\right)$ is the n-th derivative of the expression in (A.24), which is expressed as

$$\frac{\partial^n}{\partial x^n}\left(C_{rj}^{(k)}(x)\right) = \left(\sum_{m=j+1}^{k} \frac{\sum_{l_1=0,\, l_1\neq m}^{k} \sum_{l_2=0,\, l_2\neq m,l_1}^{k} \cdots \sum_{l_{n+1}=0,\, l_{n+1}\neq m,l_1,\ldots,l_n}^{k} \prod_{q=0,\, q\neq m,l_1,\ldots,l_{n+1}}^{k} \left(x - x_{i-r+q-\frac{1}{2}}\right)}{\prod_{l=0,\, l\neq m}^{k}\left(x_{i-r+m-\frac{1}{2}} - x_{i-r+l-\frac{1}{2}}\right)}\right)\Delta x_{i-r+j} \tag{A.72}$$

with $n = 1, \ldots, k-2$.

To compute β_r using (A.65), numerical integration must be carried out. A suitable quadrature formula must be used for the integration of the $k-2$ first terms of the summation. For the last term ($l = k-1$), numerical integration is not needed since the derivative is a constant value.

Using (A.71) it is possible to compute the value of the n-th derivative of $p_r(x)$ at a certain point. When having uniform grid, derivatives of the coefficients for (A.71) at the right boundary are given by

$$\frac{\partial^n}{\partial x^n}\left(c_{rj}^{(k)}\right) = \left(\sum_{m=j+1}^{k} \frac{\sum_{l_1=0,\, l_1\neq m}^{k} \sum_{l_2=0,\, l_2\neq m,l_1}^{k} \cdots \sum_{l_{n+1}=0,\, l_{n+1}\neq m,l_1,\ldots,l_n}^{k} \prod_{q=0,\, q\neq m,l_1,\ldots,l_{n+1}}^{k} (r-q+1)}{\prod_{l=0,\, l\neq m}^{k}(m-l)}\right)\Delta x^{-n} \tag{A.73}$$

where $c_{rj}^{(k)} = C_{rj}^{(k)}(x = x_{i+\frac{1}{2}})$ and $n = 1, \ldots, k-2$. Derivatives of degree up to $k-2$ at the cell center are given by

$$\frac{\partial^n}{\partial x^n}\left(\breve{c}_{rj}^{(k)}\right)=\left(\sum_{m=j+1}^{k}\frac{\sum_{l_1=0}^{k}\sum_{l_2=0}^{k}\cdots\sum_{l_{n+1}=0}^{k}\prod_{q=0}^{k}(r-q+1/2)}{\prod_{\substack{l=0\\l\neq m}}^{k}(m-l)}\right)\Delta x^{-n} \tag{A.74}$$

with $\breve{c}_{rj}^{(k)}=C_{rj}^{(k)}(x_i)$. Derivatives of degree up to $k-2$ at the left boundary can be obtained using

$$\frac{\partial^n}{\partial x^n}\left(\tilde{c}_{rj}^{(k)}\right)=\frac{\partial^n}{\partial x^n}\left(c_{r-1,j}^{(k)}\right) \tag{A.75}$$

with $\tilde{c}_{rj}^{(k)}=C_{rj}^{(k)}(x=x_{i-\frac{1}{2}})$

Equation (A.72) is only valid for the derivatives of order $n=1,\ldots,k-2$, since the product term cannot be computed for the case when $n=k-1$. To calculate the $(k-1)$-th derivative, which is constant inside the cell, the following equation is used

$$\frac{\partial^{k-1}}{\partial x^{k-1}}\left(C_{rj}^{(k)}(x)\right)=\left(\sum_{m=j+1}^{k}\frac{k!}{\prod_{\substack{l=0\\l\neq m}}^{k}(m-l)}\right)\Delta x^{-(k-1)} \tag{A.76}$$

Numerical results of the computation of derivatives for Gaussian type function

The following Gaussian function is considered:

$$f(x)=1+e^{-\frac{(x+100)^2}{150}} \tag{A.77}$$

The first four derivatives of (A.77) are computed in the domain $x=[0,200]$ using $k=5$, $\Delta x=2$ and $N=100$. Numerical results at cell boundaries and cell center are plotted in Figure A.4 and compared with the exact solution. Notice that the fourth derivative shown in Figure A.4 is constant inside each cell since reconstructing polynomials are of 4-th degree when setting $k=5$.

A.3 Improved WENO procedures

A.3.1 WENO-5M

The mapped WENO approach was first introduced in [116] as a fix for the convergence issues that appeared at critical points when using the WENO-JS finite differences scheme. The resulting numerical scheme was only designed to reach fifth order of accuracy and was called WENO-5M [116].

In the WENO-JS scheme conditions in (A.61) and (A.62) were required to ensure the formal order of accuracy of the WENO reconstruction. But the achievement of formal order of accuracy require that WENO weights become a 3-rd order approximation of the optimal weights γ_r at critical points [116]. This constrain can be enforced through a mapping procedure. In [116] the following mapping function was proposed

$$g_r(\omega)=\frac{\omega(\gamma_r+\gamma_r^2-3\gamma_r\omega+\omega^2)}{\gamma_r^2+\omega(1-2\gamma_r)}\qquad r=0,1,2 \tag{A.78}$$

and combines the WENO-JS weights and the optimal weights. The WENO-5M weights, ω_r^M, are computed as follows

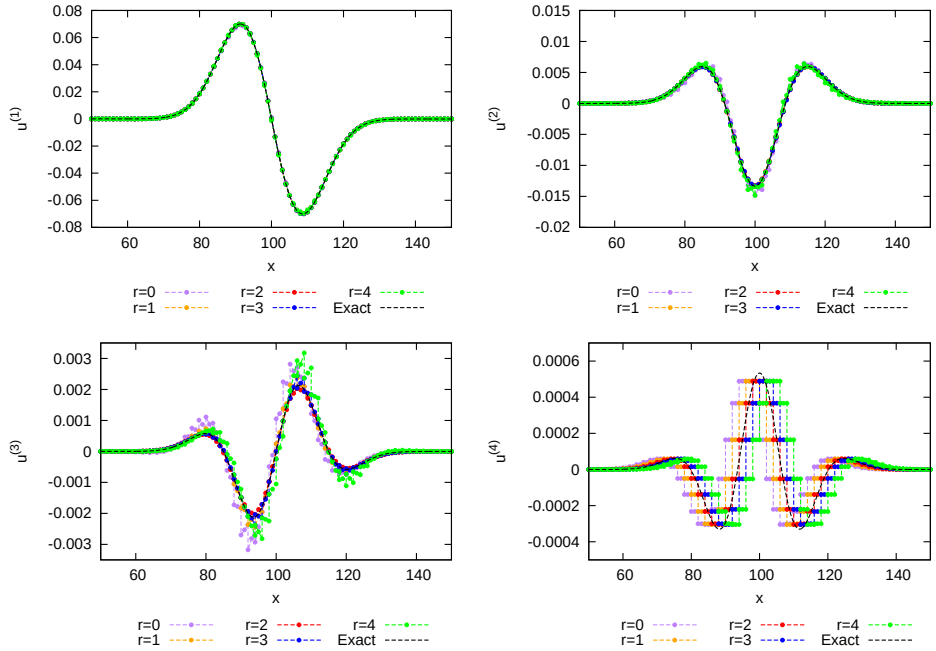


Figure A.4: Numerical results of the computation of first four derivatives of function in (A.77) using $k = 5$, $\Delta x = 2$ and $N = 100$.

$$\alpha_r^M = g_r(\omega_r^{JS}) \qquad \omega_r^M = \frac{\alpha_r^M}{\sum_{l=0}^{k-1} \alpha_l^M}, \qquad r = 0, 1, 2 \tag{A.79}$$

Function $g_r(\omega)$ in (A.78) becomes flat in the neighborhood of the r-th optimal weight γ_r, and in smooth regions where the deviation of the original WENO-JS weights ω_r^{JS} from the optimal weights γ_r is relatively small, function $g_r(\omega)$ maps those weights providing more accurate values, closer to γ_r. In non-smooth regions the original weights ω_r^{JS} may get extreme values (close to 0 or 1), and $g_r(\omega)$ provides a mapping close to the identity mapping ensuring $g_r(0) = 0$ and $g_r(1) = 1$. A drawback of this new technique is the extra computational cost needed for the rendering of the new weights [117].

A.3.2 WENO-Z

The WENO-Z weights in [117, 119] provide an alternative to the smoothness indicator β_r in (A.65), defining a more sophisticated indicator β_r^Z

$$\beta_r^Z = \frac{\beta_r + \epsilon}{\beta_r + \tau_{2k-1} + \epsilon} \tag{A.80}$$

where τ_{2k-1} is the global smoothness indicator, derived from the examination of the Taylor expansions of the Lagrange polynomials that provide the $c_{rj}^{(k)}$ coefficients in (A.35). This indicator will be either computed as $\tau_{2k-1} = |\beta_0 - \beta_{k-1}|$ when k is odd or computed as $\tau_{2k-1} = |\beta_0 - \beta_1 + \beta_{k-2} - \beta_{k-1}|$ when k is even. The general expression for the α_r^Z and ω_r^Z weights is given by

$$\alpha_r^Z = \frac{\gamma_r}{\beta_r^Z} = \gamma_r \left(1 + \left(\frac{\tau_{2r-1}}{\beta_r + \epsilon}\right)^{p_Z}\right) \qquad \omega_r^Z = \frac{\alpha_r^Z}{\sum_{l=0}^{k-1} \alpha_l^Z}, \qquad r = 0, ..., k-1 \tag{A.81}$$

with $p_Z = k - 1$, ensuring the necessary order of accuracy of the non-oscillatory weights at critical points. Even both the WENO-5M and the WENO-Z schemes ensure all conditions to achieve the formal order

of convergence, the WENO-Z scheme provides more accurate results around shocks avoiding the extra computational cost of a mapping procedure [117, 119].

A.3.3 The WENO-MZ method

In [120], an improved technique based on the combination of the WENO-M and WENO-Z methods was proposed. First, the non-oscillatory weights are calculated using the WENO-Z approach, following the procedure in Section A.3.2. Then, the ω_r^Z weights are mapped into new weights that should be closer to the optimal weights in smooth regions. These new weights will be denoted by ω_r^{MZ} weights and are computed following the procedure in Section A.3.1 as

$$\alpha_r^{MZ} = g_r(\omega_r^Z) \qquad \omega_r^{MZ} = \frac{\alpha_r^{MZ}}{\sum_{l=0}^{k-1} \alpha_l^{MZ}}, \qquad r = 0, \ldots, k-1 \tag{A.82}$$

where $g_r(\omega)$ is the mapping function in (A.78).

B SUB-CELL WENO RECONSTRUCTION OF DERIVATIVES

WENO sub-cell derivative reconstruction procedures in [124, 30] provide suitable approximations of the derivatives of the function in ADER schemes. Reconstruction procedure of spatial derivatives in [124] is more efficient and leads to a better solution of the ADER scheme. The reconstruction of the $2k-2$ derivatives for a $(2k-1)$-th order ADER scheme is performed by means of a $(2k-1)$-th order WENO reconstruction using k stencils in [124], while in [30] $(2k-1)$ stencils are needed.

This method is based on the construction of a polynomial $\phi_i(x)$ inside each cell I_i. As $2k-2$ points inside I_i are defined, $2k-2$ WENO reconstructions of $(2k-1)$-th order are required. Derivatives of polynomial $\phi_i(x)$ are an approximation of the exact derivatives of function $u(x)$. The procedure for the estimation of the $2k-2$ derivatives is summarized in the next subsection.

B.1 Procedure for the reconstruction of the derivatives

The procedure for the WENO sub-cell derivative reconstruction procedure in [124] is composed of the following four steps:

a) *Define sub-cell points*

Sub-cell points for the cell I_i are denoted by $x_i^{(b)}$ for $b = 1, ..., 2k-2$. In [124], uniformly distributed points inside the cell are proposed, but this leads to negative optimal weights, γ_r, in the WENO reconstruction. Thus the WENO procedure needs to be modified in order to give a good treatment to the negative weights.

All sub-cell points are chosen to be inside the positive interval of the optimal weights [152]. Note that the positive intervals of the optimal weights widely cover the cell boundaries but the center, as depicted in Figure B.1, therefore sub-cell points are taken close to cell interfaces. Moreover, it is worth mentioning that there points inside the cell where optimal weights do not exist. The asymptotic behavior of the weights around these points can be observed in Figure B.1 that plots the minimum optimal weight against x inside a normalized cell with $\Delta x = 1$. Notice that the number of singular points is equal to $k-1$ and also that the center of the cell is a singular point when k is even.

The following formula is proposed, with a geometrical refinement of $1/2$ between consecutive points of the same cell side:

$$x_i^{(b)} = \begin{cases} x_{i-\frac{1}{2}} & \text{if } b = 1 \\ x_{i-\frac{1}{2}} + \frac{\Delta x}{2 \cdot 2^{k-b}} & \text{if } 2 \le b \le k-1 \\ x_{i+\frac{1}{2}} - \frac{\Delta x}{2 \cdot 2^{b-k+1}} & \text{if } k \le b \le 2k-3 \\ x_{i+\frac{1}{2}} & \text{if } b = 2k-2 \end{cases}$$

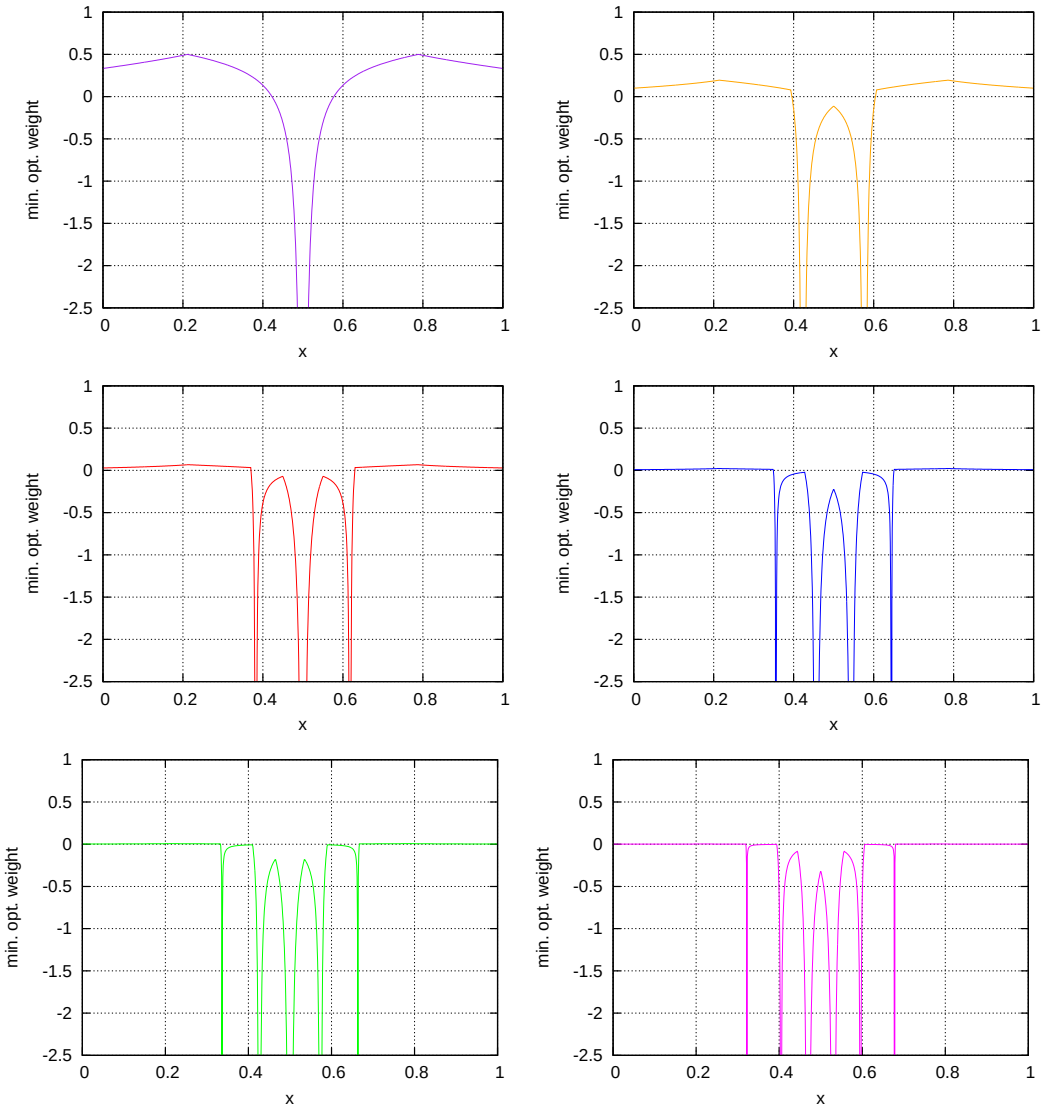


Figure B.1: Minimum optimal weight value inside a cell with cell size $\Delta x = 1$ for a 3-rd, 5-th, 7-th, 9-th, 11-th and 13-th polynomial reconstruction procedure.

b) *Reconstruct* $(2k-2)$ *point-wise values of* $u(x)$

A $(2k-1)$-th order WENO approximation in (A.68) has to be used to get $u_i^{(b)}$ with $b = 1, 2, ..., 2k-2$, the reconstructed point-wise values of u at the sub-cell points $x_i^{(b)}$ inside I_i. Different WENO weights, denoted by $\omega_r^{(b)} = \Omega_r(x_i^{(b)})$, will appear at each sub-cell point. Expression in (A.64) is used to compute $\alpha_r^{(b)}$ weights

$$\alpha_r^{(b)} = \frac{\gamma_r^{(b)}}{(\beta_r + \epsilon_i)^2}, \qquad r = 0, ..., k-1, \quad b = 1, ..., 2k-2 \tag{B.1}$$

where $\gamma_r^{(b)} = \Gamma_r(x_i^{(b)})$. Then, expression in (A.66) is used to compute the non-oscillatory weights as

$$\omega_r^{(b)} = \frac{\alpha_r^{(b)}}{\sum_{l=0}^{k-1} \alpha_l^{(b)}}, \qquad r = 0, ..., k-1, \quad b = 1, ..., 2k-2 \tag{B.2}$$

The $2k-1$-th WENO reconstruction at $x_i^{(b)}$ is given by (A.60), yielding

$$u_i^{(b)} = \sum_{r=0}^{k-1} \omega_r^{(b)} u_i^{(b)(r)} \tag{B.3}$$

c) *Construct* $\phi_i(x)$

Following [124], the expression for the polynomial that approximates $u(x)$ in I_i is

$$\phi_i(x) = \sum_{l=0}^{2k-2} a_l \left(\frac{x - x_{i-\frac{1}{2}}}{\Delta x} \right)^l \tag{B.4}$$

where the coefficients a_l $(l = 0, ..., 2k-2)$ have to be determined with the following $2k-1$ equations for each cell I_i:

- From the $2k-2$ reconstructed values of $u(x)$ at the $2k-2$ sub-cell points, the following equations are formulated

$$\phi_i\left(x_i^{(b)}\right) = u_i^{(b)}, \qquad b = 1, ..., 2k-2 \tag{B.5}$$

- From the cell average, $\bar{u}_i$, the following equation is formulated

$$\frac{1}{\Delta x} \int_{I_r} \phi_i(x)\, dx = \sum_{l=0}^{2k-2} \frac{a_l}{l+1} \left[\left(\frac{x - x_{i-\frac{1}{2}}}{\Delta x} \right)^{l+1} \right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} = \bar{u}_i \tag{B.6}$$

These $2k-2$ equations in (B.5) and the equation in (B.6) are used to formulate a linear system with the coefficients a_l $(l = 0, ..., 2k-2)$ in the vector of unknowns.

d) *Evaluate derivatives at the cell boundaries*

We get the approximation to the m-th derivative of $u(x)$ $(m = 1, ..., 2k-2)$ at any desired point taking the m-th derivative of $\phi_i(x)$

$$\frac{d^m \phi_i(x)}{dx^m} = \frac{d^m}{dx^m} \left[\sum_{l=0}^{2k-2} a_l \left(\frac{x - x_{i-\frac{1}{2}}}{\Delta x} \right)^l \right] = \sum_{l=m}^{2k-2} \frac{l!}{(l-m)!} \frac{a_l}{\Delta x^l} \left(x - x_{i-\frac{1}{2}} \right)^{(l-m)} \tag{B.7}$$

From (B.7), it becomes clear that the quality of the WENO the reconstruction at the $2k-2$ points will determine the accuracy in the computation of the $2k-2$ derivatives or the equivalent a_l coefficients in (B.7), and the actual convergence to an ADER scheme of $(2k-1)$-th order of accuracy.

B.2 Results

B.2.1 Reconstruction of a smooth function and its derivatives

The performance of the different WENO approaches when applied to the WENO sub-cell derivative reconstruction method is first analyzed here. A polynomial type function is chosen. The quality of the reconstruction of the a_l coefficients in (B.4) is of utmost importance, as they will provide the direct estimation of the $2k-2$ non trivial derivatives required to generate a $2k-1$-th ADER numerical scheme.

In this test case, a fifth order polynomial of the form

$$u(x) = \prod_{r=1}^{4}(x - x_r) \tag{B.8}$$

with roots $x_r = \{-621.9210, 43.5780, 103.2702, 175.0728\}$, is reconstructed in the domain $x = [0, 200]$. As outlined in the previous section, when using the WENO sub-cell derivative reconstruction procedure a $\phi_i(x)$ function is defined inside each cell following (B.4). Also, for any cell I_i, it is possible to exactly express polynomial in (B.8) as

$$u_i(x) = \sum_{l=0}^{2k-2} a_l^*(x - x_{i-1/2})^l \tag{B.9}$$

following the same form used to define $\phi_i(x)$ in (B.4), where the a_l^* coefficients depend on the selected cell. In case that the reconstruction procedure provides the exact solution, the equality

$$a_l^* = \frac{a_l}{\Delta x^l} \qquad l = 0, ..., 2k-2 \tag{B.10}$$

will be satisfied and therefore $u_i(x) = \phi_i(x)$. It is worth recalling that the order of the reconstructing polynomial must higher or equal to that in $u_i(x)$.

By setting $\Delta x = 1$ in a domain $x = [0, 200]$, the recovery of the a_l^* coefficients is analyzed in two different cells: I_{71} with $x_{i-1/2} = 70$ and I_{101} with $x_0 = 100$. At these cells, functions $\phi_{71}(x)$ and $\phi_{101}(x)$ are defined respectively. In I_{71} there is a first order critical point where the first derivative vanishes, while in I_{101} all derivatives are different from zero. To measure the error in the computation of a_l^*, we introduce the following indicator

$$\Theta = \|\boldsymbol{\theta}\|_\infty \tag{B.11}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_{2k-1})$ is a vector. The components of $\boldsymbol{\theta}$ are given by

$$\theta_{l+1} = \frac{a_l^* - \frac{a_l}{\Delta x^l}}{a_l^*}, \qquad l = 0, ..., 2k-2 \tag{B.12}$$

and account for the relative error in the computation of each coefficient. This indicator provides the maximum relative error value among all the coefficients.

Table B.1 shows the exact polynomial coefficients a_l^* in cell I_{71} and the estimation of the same coefficients provided by the 5-th order WENO-JS and WENO-Z methods. Due to the presence of a critical point inside cell I_{71}, when using the WENO-JS method, coefficient a_5^* is recovered with a huge relative error. The WENO-Z method provides better results than those given by the WENO-JS scheme, reducing the error in 2 orders of magnitude.

Table B.2 shows numerical errors in cell I_{101}. The best reconstruction is provided by the WENO-Z, though there are not big differences when compared to the WENO-JS. Since there are not any critical points in I_{101} and its neighboring cells, the error $\boldsymbol{\theta}$ for the WENO-JS is already small enough as the optimal weights are adequately recovered.

l	a_l^*	$(a_l/\Delta x^l)_{\omega_{JS}}$	$(a_l/\Delta x^l)_{\omega_{Z},p=2}$
0	6.391	6.39099972	6.3909999
1	-1.8E-03	-1.80068E-03	-1.800003E-03
2	-7.76E-03	-7.75322E-03	-7.759969E-03
3	5.8E-05	4.75146E-05	5.795266E-05
4	1.0E-07	5.06439E-06	1.224175E-07
Error (Θ):		49.6439	0.2241

Table B.1: Section B.2.1. Coefficients of exact and reconstructed polynomial ϕ_{71} with the sub-cell reconstruction procedure, $\Delta x = 1$.

l	a_l^*	$(a_l/\Delta x^l)_{\omega_{JS}}$	$(a_l/\Delta x^l)_{\omega_{Z},p=2}$
0	1	0.999999941	1.0
1	-0.3	-0.299999943	-0.3
2	-2.0E-03	-1.9998E-03	-2.0E-03
3	7.0E-05	6.9870E-05	7.0E-05
4	1.0E-07	1.0601E-07	1.0E-07
Error (Θ):		6.0085E-02	1.7124E-07

Table B.2: Section B.2.1. Coefficients of exact and reconstructed polynomial ϕ_{101} with the sub-cell reconstruction procedure, $\Delta x = 1$.

C 2D EXTENSION OF THE WENO RECONSTRUCTION METHOD

C.1 Interpolation and reconstruction in 2D

In this section, the problem of data reconstruction in 2D at an arbitrary point inside a cell by means of polynomial interpolation when departing from cell averages is considered.

The function $u(x,y)$ will be defined departing from the starting data, that will be considered as the average value of this function in each cell. The definition of $u(x,y)$ is useful for the derivation of the reconstruction procedure but its analytical expression will be unknown in most cases. The computational grid, shown in Figure C.1, is composed by $N_x \times N_y$ cells as

$$\Omega = [a,b] \times [c,d] \tag{C.1}$$

with

$$\begin{aligned} a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \ldots < x_{N_x-\frac{1}{2}} < x_{N_x+\frac{1}{2}} = b \\ c = y_{\frac{1}{2}} < y_{\frac{3}{2}} < \ldots < y_{N_y-\frac{1}{2}} < y_{N_y+\frac{1}{2}} = d \end{aligned} \tag{C.2}$$

with cells and cell sizes defined by

$$I_{i,j} = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \times \left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right] \tag{C.3}$$

$$\begin{aligned} \Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} \equiv \text{constant} \\ \Delta y_j = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}} \equiv \text{constant} \end{aligned} \tag{C.4}$$

With the previous definitions, the starting data set is now defined as the the average value of the function $u(x,y)$ in each cell

$$\bar{u}_i = \frac{1}{\Delta x_i \Delta y_j} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u(\xi,\eta)\, d\eta d\xi, \quad i = 1,2,\ldots,N_x \quad j = 1,2,\ldots,N_y \tag{C.5}$$

The problem we face is to find a polynomial function $p_{r_1,r_2}(x,y)$ of **degree at most** $k-1$ for each cell $I_{i,j}$, such that it is a k**-th order accurate** approximation of the function $u(x,y)$ inside $I_{i,j}$

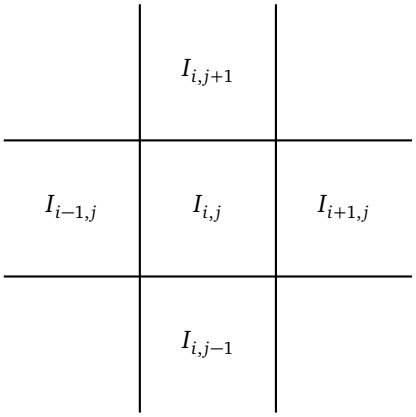


Figure C.1: Mesh discretization

$$p_r(x,y) = u(x,y) + O\left(\Delta x^k\right), \qquad x \in I_{i,j}, \quad i = 1,2,...,N_x \quad j = 1,2,...,N_y \tag{C.6}$$

In the two-dimensional case, the concept of *stencil* is generalized to a group of surface connected cells. For each cell, $I_{i,j}$, it is possible to define a stencil $S_{r_1,r_2}(i,j)$ composed by cell $I_{i,j}$ plus r_1 cells to the left, s_1 cells to the right, r_2 cells to the top and s_2 cells to the bottom. If considering $S_{r_1,r_2}(i,j)$ with the same number of cells k, in both directions, we can affirm $k = r_1 + s_1 + 1 = r_2 + s_2 + 1$. For all cases, the condition $r, s \geq 0$ must be satisfied. The stencil can be expressed as

$$S_{r_1,r_2}(i,j) = \bigcup_{l,m \in [0,...,k-1]} I_{i-r_1+l,j-r_2+m} \tag{C.7}$$

The steps required to generate the reconstructing polynomial departing from cell averages are listed below:

a) Stencil selection.

Given the cell $I_{i,j}$ and the order of accuracy required k, we must first choose a stencil $S_{r_1,r_2}(i,j)$ with $k = r_1 + s_1 + 1 = r_2 + s_2 + 1$ cells.

There is a unique polynomial $p_{r_1,r_2}(x,y)$ of degree at most $k-1$ whose cell average value for each cell in the stencil agrees with that of the function $u(x,y)$ [26]

$$\frac{1}{\Delta x_m \Delta y_l} \int_{x_{m-\frac{1}{2}}}^{x_{m+\frac{1}{2}}} \int_{y_{l-\frac{1}{2}}}^{y_{l+\frac{1}{2}}} p_{r_1,r_2}(\xi,\eta)\, d\eta d\xi = \bar{u}_{m,l} \tag{C.8}$$

with $m = i - r_1, ..., i + s_1$ and $l = j - r_2, ..., i + s_2$.

b) Definition of the primitive function.

In order to find the interpolating polynomial $p_{r_1,r_2}(x,y)$ of degree $k-1$ and k-th order of accuracy, a new function is introduced. This new function is the primitive function of $u(x,y)$, denoted by $U(x,y)$, which is defined as the cumulative integral of $u(x,y)$ from $-\infty$ to x and y

$$U(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} u(\xi,\eta)\, d\eta d\xi \tag{C.9}$$

For a random location in the grid, i, j, the value of this cumulative integral at the right boundary of the cell $I_{i,j}$ can be computed by the summation of the average values of each cell multiplied by the

cell size, from $-\infty$ to the cell $I_{i,j}$, as follows:

$$U\left(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}\right) = \int_{-\infty}^{x_{i+\frac{1}{2}}} \int_{-\infty}^{y_{j+\frac{1}{2}}} u(\xi,\eta)\, d\eta d\xi = \sum_{m=-\infty}^{i} \sum_{l=-\infty}^{j} \bar{u}_{m,l} \Delta x_m \Delta y_l \tag{C.10}$$

Also, a polynomial $P_{r_1,r_2}(x,y)$ is defined as the **unique polynomial function of degree at most** k **which interpolates** $U(x,y)$ **with** $k+1$**-th order of accuracy in** $k+1$ **nodes** (which are all the cell boundaries in the stencil) and we denote its derivative by $p_{r_1,r_2}(x,y)$:

$$p_{r_1,r_2}(x,y) = \frac{\partial^2}{\partial x \partial y} P_{r_1,r_2}(x,y) \tag{C.11}$$

Note that $p_{r_1,r_2}(x,y)$ is a polynomial of degree $k-1$ and k-th order, defined by k^2 cells. Polynomial $P_r(x)$ is one order greater, and as the number of cells does not change, $k+1$ interpolation points are necessary. It is worth noticing that this new $k+1$ points are defined at the nodes, even though the value of $u(x,y)$ is not initially defined at these locations.

Using (C.11) it is possible to prove the equality in (C.8)

$$\begin{aligned}
&\frac{1}{\Delta x_m \Delta y_l} \int_{x_{m-\frac{1}{2}}}^{x_{m+\frac{1}{2}}} \int_{y_{l-\frac{1}{2}}}^{y_{l+\frac{1}{2}}} p_{r_1,r_2}(\xi,\eta)\, d\eta d\xi = \frac{1}{\Delta x_m \Delta y_l} \int_{x_{m-\frac{1}{2}}}^{x_{m+\frac{1}{2}}} \int_{y_{l-\frac{1}{2}}}^{y_{l+\frac{1}{2}}} \frac{\partial^2}{\partial \xi \partial \eta} P_{r_1,r_2}(\xi,\eta)\, d\eta d\xi = \\
&\frac{1}{\Delta x_m \Delta y_l} \int_{x_{m-\frac{1}{2}}}^{x_{m+\frac{1}{2}}} \partial \int_{y_{l-\frac{1}{2}}}^{y_{l+\frac{1}{2}}} \partial P_{r_1,r_2}(\xi,\eta) = \frac{1}{\Delta x_m \Delta y_l} \int_{x_{m-\frac{1}{2}}}^{x_{m+\frac{1}{2}}} \partial \left[P_{r_1,r_2}\left(\xi, y_{l+\frac{1}{2}}\right) - P_{r_1,r_2}\left(\xi, y_{l-\frac{1}{2}}\right) \right] = \\
&\frac{1}{\Delta x_m \Delta y_l} \left[P_{r_1,r_2}\left(x_{m+\frac{1}{2}}, y_{l+\frac{1}{2}}\right) - P_{r_1,r_2}\left(x_{m+\frac{1}{2}}, y_{l-\frac{1}{2}}\right) - P_{r_1,r_2}\left(x_{m-\frac{1}{2}}, y_{l+\frac{1}{2}}\right) + P_{r_1,r_2}\left(x_{m-\frac{1}{2}}, y_{l-\frac{1}{2}}\right) \right] \\
&\quad \approx \frac{1}{\Delta x_m \Delta y_l} \left[U\left(x_{m+\frac{1}{2}}, y_{l+\frac{1}{2}}\right) - U\left(x_{m+\frac{1}{2}}, y_{l-\frac{1}{2}}\right) - U\left(x_{m-\frac{1}{2}}, y_{l+\frac{1}{2}}\right) + U\left(x_{m-\frac{1}{2}}, y_{l-\frac{1}{2}}\right) \right] = \\
&= \frac{1}{\Delta x_m \Delta y_l} \int_{x_{m-\frac{1}{2}}}^{x_{m+\frac{1}{2}}} \int_{y_{l-\frac{1}{2}}}^{y_{l+\frac{1}{2}}} \frac{\partial^2}{\partial \xi \partial \eta} U(\xi,\eta)\, d\eta d\xi = \frac{1}{\Delta x_m \Delta y_l} \int_{x_{m-\frac{1}{2}}}^{x_{m+\frac{1}{2}}} \int_{y_{l-\frac{1}{2}}}^{y_{l+\frac{1}{2}}} u(\xi,\eta)\, d\eta d\xi = \bar{u}_{m,l}
\end{aligned} \tag{C.12}$$

for any $m = i - r_1, ..., i + s_1$ and $l = j - r_2, ..., i + s_2$. The approximation symbol stands for the approximation of $U(x,y)$ by the interpolating polynomial $P_{r_1,r_2}(x,y)$. This interpolation is a $k+1$-th order approximation

$$P_{r_1,r_2}(x,y) = U(x,y) + O\left(\Delta x^{k+1}\right), \qquad x,y \in I_{i,j} \tag{C.13}$$

and that of its derivative, a k-th order approximation

$$\frac{\partial^2}{\partial x \partial y} P_{r_1,r_2}(x,y) = \frac{\partial^2}{\partial x \partial y} U(x,y) + O\left(\Delta x^k\right), \qquad x,y \in I_{i,j}\,. \tag{C.14}$$

c) Lagrange interpolation

In [26], the use of the Lagrange form of the interpolating polynomial is proposed. This kind of interpolation is said to be nodal since each weight takes the value of 1 at the corresponding node and 0 at the rest of the nodes. The expression for the 2D Lagrange interpolating polynomial for structured meshes of $n_x \cdot n_y$ points is given by

$$L(x) = \sum_{i=0}^{n_x} \sum_{j=0}^{n_y} u(x_i, y_j) L_{i,j}(x,y) \tag{C.15}$$

where

$$L_{i,j}(x,y) = l_i(x) \cdot l_j(y) \tag{C.16}$$

considering the 2D mesh as the intersection of two 1D meshes defined by the points $\{x_1, x_2, ..., x_{n_x}\}$ and $\{y_1, y_2, ..., y_{n_y}\}$ respectively.

The weighting functions are defined as in the 1D case as

$$\begin{aligned} l_i(x) &= \prod_{\substack{d=0 \\ d\neq i}}^{n_x} \frac{(x - x_d)}{(x_i - x_d)} \\ l_j(y) &= \prod_{\substack{d=0 \\ d\neq j}}^{n_y} \frac{(y - y_d)}{(y_j - y_d)} \end{aligned} \tag{C.17}$$

Making use of the Lagrange formula in (C.15), it is possible to write the expression for the interpolating polynomial $P_{r_1,r_2}(x,y)$ at all nodes of the stencil $S(i,j)$ where the values of function $U(x,y)$ are known, yielding

$$P_{r_1,r_2}(x,y) = \sum_{m=0}^{k}\sum_{l=0}^{k} \tilde{U}_{m,l} \prod_{\substack{d=0 \\ d\neq m}}^{k} \frac{(x - x_{i-r_1+d-\frac{1}{2}})}{(x_{i-r_1+m-\frac{1}{2}} - x_{i-r_1+d-\frac{1}{2}})} \prod_{\substack{d=0 \\ d\neq l}}^{k} \frac{(y - y_{j-r_2+d-\frac{1}{2}})}{(y_{j-r_2+l-\frac{1}{2}} - y_{j-r_2+d-\frac{1}{2}})} \tag{C.18}$$

where $\tilde{U}_{m,l}$ is a redefined primitive function with origin at $(x_{i-\frac{1}{2}-r_1}, y_{j-\frac{1}{2}-r_2})$ and given by

$$\tilde{U}_{m,l} = \int_{x_{i-\frac{1}{2}-r_1}}^{x_{i-\frac{1}{2}-r_1+m}} \int_{y_{j-\frac{1}{2}-r_2}}^{y_{j-\frac{1}{2}-r_2+l}} u(\xi,\eta)\, d\eta d\xi \tag{C.19}$$

The use of $\tilde{U}_{m,l}$ allows to express (C.18) in terms of exclusively cell averages inside the stencil, since

$$\tilde{U}_{m,l} = \sum_{e=0}^{m-1}\sum_{d=0}^{l-1} \bar{u}_{i-r_1+e,j-r_2+d} \Delta x_{i-r_1+e} \Delta y_{j-r_2+d} \tag{C.20}$$

Taking the cross derivative of (C.18) with respect to x and y and inserting the previous result, a expression for polynomial $p_{r_1,r_2}(x,y)$ is obtained

$$p_{r_1,r_2}(x,y) = \sum_{m=0}^{k}\sum_{l=0}^{k} \left(\mathscr{L}_m \mathscr{L}_l \sum_{e=0}^{m-1}\sum_{d=0}^{l-1} \bar{u}_{i-r_1+e,j-r_2+d} \Delta x_{i-r_1+e} \Delta y_{j-r_2+d} \right) \tag{C.21}$$

with

$$\mathscr{L}_m = \left(\frac{\sum_{t=0,\, t\neq m}^{k} \prod_{n=0,\, n\neq m,t}^{k} \left(x - x_{i-r_1+n-\frac{1}{2}}\right)}{\prod_{\substack{n=0 \\ n\neq m}}^{k} \left(x_{i-r_1+m-\frac{1}{2}} - x_{i-r_1+n-\frac{1}{2}}\right)} \right) \tag{C.22}$$

$$\mathscr{L}_l = \left(\frac{\sum_{t=0,\, t\neq l}^{k} \prod_{n=0,\, n\neq l,t}^{k} \left(y - y_{j-r_2+n-\frac{1}{2}}\right)}{\prod_{\substack{n=0 \\ n\neq l}}^{k} \left(y_{j-r_2+l-\frac{1}{2}} - y_{j-r_2+n-\frac{1}{2}}\right)} \right) \tag{C.23}$$

A simpler expression for $p_{r_1,r_2}(x,y)$ can be derived from equation (C.21) taking the cell averages as common factors. The resulting expression represents the reconstructing polynomial function as a linear combination of the cell averages inside the stencil as

$$p_{r_1,r_2}(x,y) = \sum_{e=0}^{k-1}\sum_{d=0}^{k-1}\left(\sum_{m=e+1}^{k}\sum_{l=d+1}^{k}\mathscr{L}_m\,\mathscr{L}_l\right)\bar{u}_{i-r_1+e,j-r_2+d}\,\Delta x_{i-r_1+e}\,\Delta y_{j-r_2+d} \tag{C.24}$$

that can be rewritten as

$$p_{r_1,r_2}(x,y) = \sum_{e=0}^{k-1}\sum_{d=0}^{k-1}\left(\sum_{m=e+1}^{k}\mathscr{L}_m\sum_{l=d+1}^{k}\mathscr{L}_l\right)\bar{u}_{i-r_1+e,j-r_2+d}\,\Delta x_{i-r_1+e}\,\Delta y_{j-r_2+d} \tag{C.25}$$

If defining

$$C^{(k)}_{r_1,e}(x) = \left(\sum_{m=e+1}^{k}\mathscr{L}_m\right)\Delta x_{i-r_1+e} \tag{C.26}$$

$$C^{(k)}_{r_2,d}(y) = \left(\sum_{l=d+1}^{k}\mathscr{L}_l\right)\Delta y_{j-r_2+d} \tag{C.27}$$

it is possible to express Equation (C.25) as

$$p_{r_1,r_2}(x,y) = \sum_{e=0}^{k-1}\sum_{d=0}^{k-1} C^{(k)}_{r_1,e}(x)\,C^{(k)}_{r_2,d}(y)\;\bar{u}_{i-r_1+e,j-r_2+d} \tag{C.28}$$

Where $C^{(k)}_{r_1,e}(x)$ and $C^{(k)}_{r_2,d}(y)$ are constants at a given x and provide the weights for the linear combination of cell averages. The superscript k of these coefficients stands for the dimension of the stencil in each direction. Remark that coefficients $C^{(k)}_{r_1,e}(x)$ and $C^{(k)}_{r_2,d}(y)$ are equivalent to those obtained for the 1D case in (A.24).

d) Calculation of $C^{(k)}_{r_1,e}(x)$ and $C^{(k)}_{r_2,d}(y)$ coefficients at the sought point and computation of the reconstruction using (C.28).

C.2 Dimension-by-dimension 2D reconstruction

In the previous part, the procedure for the generation of a 2D reconstruction departing from cell averages was shown. The expression for the reconstructing polynomial function was obtained in (C.28). Considering this result, it is straightforward to compute the reconstruction at a certain point by calculating first the coefficients $C^{(k)}_{r_1,e}(x)$ and $C^{(k)}_{r_2,d}(y)$ at the desired point and substituting then in (C.28), getting the sought value.

Another possibility would be to obtain the 2D reconstruction by carrying out two 1D reconstructions recursively, for each of the variables, x and y, each time. For instance, let us consider the first 1D reconstruction for variable y. First, for a fixed x value, the function $u(x,y)$ can be reconstructed along the y coordinate using Lagrange interpolation as in the 1D case. Then, the resulting set of reconstructed values at each x position, which depends upon y, can be used to generate another Lagrange interpolation polynomial that depends upon x and y and corresponds to the sought reconstructing function.

For instance, let us consider the first polynomial reconstruction for variable y. The interpolation polynomial will be denoted by $P^m_{r_1,r_2}(y)$ and constructed using the Lagrange basis, leading to the following expression

$$P^m_{r_1,r_2}(y) = \sum_{l=0}^{k-1} \tilde{U}_{m,l} \prod_{\substack{d=0 \\ d\neq l}}^{k} \frac{(y - y_{j-r_2+d-\frac{1}{2}})}{(y_{j-r_2+l-\frac{1}{2}} - y_{j-r_2+d-\frac{1}{2}})} \tag{C.29}$$

where m stands for the x position. Substitution of $\tilde{U}_{m,l}$ by the expression provided in (C.20) leads to

$$P^m_{r_1,r_2}(y) = \sum_{l=0}^{k-1} \left(\sum_{e=0}^{m-1} \sum_{d=0}^{l-1} \bar{u}_{i-r_1+e,j-r_2+d} \Delta x_{i-r_1+e} \Delta y_{j-r_2+d} \right) \prod_{\substack{d=0 \\ d\neq l}}^{k} \frac{(y - y_{j-r_2+d-\frac{1}{2}})}{(y_{j-r_2+l-\frac{1}{2}} - y_{j-r_2+d-\frac{1}{2}})} \tag{C.30}$$

Taking the derivative of (C.30) with respect to y, it yields

$$\frac{\partial P^m_{r_1,r_2}(y)}{\partial y} = \sum_{l=0}^{k-1} \left[\mathscr{L}_l \sum_{d=0}^{l-1} \left(\sum_{e=0}^{m-1} \bar{u}_{i-r_1+e,j-r_2+d} \Delta x_{i-r_1+e} \right) \Delta y_{j-r_2+d} \right] \tag{C.31}$$

with $\mathscr{L}_l$ defined in (C.23). This expression can be rewritten as

$$\frac{\partial P^m_{r_1,r_2}(y)}{\partial y} = \sum_{d=0}^{k-1} \sum_{l=d+1}^{k} \mathscr{L}_l \Delta y_{j-r_2+d} \sum_{e=0}^{m-1} \bar{u}_{i-r_1+e,j-r_2+d} \Delta x_{i-r_1+e} \tag{C.32}$$

and making use of the coefficient $C^{(k)}_{r_2,d}(y)$ defined in (C.27), Equation (C.32) can be expressed as

$$\frac{\partial P^m_{r_1,r_2}(y)}{\partial y} = \sum_{d=0}^{k-1} C^{(k)}_{r_2,d}(y) \sum_{e=0}^{m-1} \bar{u}_{i-r_1+e,j-r_2+d} \Delta x_{i-r_1+e} \,. \tag{C.33}$$

Factorization of the previous expression for each value of e yields

$$\frac{\partial P^m_{r_1,r_2}(y)}{\partial y} = \sum_{e=0}^{m-1} \left(\sum_{d=0}^{k-1} C^{(k)}_{r_2,d}(y) \bar{u}_{i-r_1+e,j-r_2+d} \right) \Delta x_{i-r_1+e} \,. \tag{C.34}$$

noticing that the term inside brackets corresponds to a 1D polynomial reconstruction for a given value of e. The general expression for a 1D reconstruction is provided in (A.27) and referred to as $p(r, k, \nu, \bar{\mathbf{v}})$, where in this case $r = r_2$, $\nu = y$ and the vector of cell averages will include the superscript e, r_1 and r_2 to denote dependency upon the x position and the stencils respectively, becoming $\bar{\mathbf{v}}^e_{r_1,r_2}$ and with components $\left\{\bar{v}^e_{r_1,r_2}\right\}_d = \bar{u}_{i-r_1+e,j-r_2+d}$, for $d = 0, ..., k-1$. In this case, Equation (A.27) becomes

$$p(r_2, k, y, \bar{\mathbf{v}}^e_{r_1,r_2}) = \sum_{d=0}^{k-1} C^{(k)}_{r_2,d}(y) \bar{u}_{i-r_1+e,j-r_2+d} \,. \tag{C.35}$$

Making use of (C.35), the derivative $\partial P^m_{r_1,r_2}(y)/\partial y$ in Equation (C.34) is finally expressed as

$$\frac{\partial P^m_{r_1,r_2}(y)}{\partial y} = \sum_{e=0}^{m-1} p(r_2, k, y, \bar{\mathbf{v}}^e_{r_1,r_2}) \Delta x_{i-r_1+e} \,. \tag{C.36}$$

On the other hand, an analogous polynomial interpolation can be carried out in the x direction by means of the Lagrange formula, given by the following expression

$$P_{r_1,r_2}(x, y) = \sum_{m=0}^{k-1} P^m_{r_1,r_2}(y) \prod_{\substack{d=0 \\ d\neq m}}^{k} \frac{(x - x_{i-r_1+d-\frac{1}{2}})}{(x_{i-r_1+m-\frac{1}{2}} - x_{i-r_1+d-\frac{1}{2}})} \tag{C.37}$$

where $P^m_{r_1,r_2}(y)$ are the values of the 1D interpolating polynomial in (C.30) along y at the k different x positions according to the selected stencil.

In order to obtain the sought 2D reconstructing function, $p_{r_1,r_2}(x,y)$, according to definition in (C.11), we take the second order cross derivative of (C.37) and obtain

$$p_{r_1,r_2}(x,y) = \sum_{m=0}^{k-1} \frac{\partial}{\partial y}\left(P^m_{r_1,r_2}(y)\right) \frac{\partial}{\partial x}\left(\prod_{\substack{d=0\\d\neq m}}^{k} \frac{(x - x_{i-r_1+d-\frac{1}{2}})}{(x_{i-r_1+m-\frac{1}{2}} - x_{i-r_1+d-\frac{1}{2}})}\right) \tag{C.38}$$

Inserting (C.36) in (C.38), the latter yields

$$p_{r_1,r_2}(x,y) = \sum_{m=0}^{k-1}\left(\sum_{e=0}^{m-1} p(r_2,k,y,\bar{\mathbf{v}}^e_{r_1,r_2})\Delta x_{i-r_1+e}\right) \frac{\partial}{\partial x}\left(\prod_{\substack{d=0\\d\neq m}}^{k} \frac{(x - x_{i-r_1+d-\frac{1}{2}})}{(x_{i-r_1+m-\frac{1}{2}} - x_{i-r_1+d-\frac{1}{2}})}\right) \tag{C.39}$$

and noticing that the derivative of the product with respect to x is equal to $\mathscr{L}_m$ in (C.22), Equation (C.39) can be expressed more compactly as

$$p_{r_1,r_2}(x,y) = \sum_{m=0}^{k-1}\left(\sum_{e=0}^{m-1} p(r_2,k,y,\bar{\mathbf{v}}^e_{r_1,r_2})\Delta x_{i-r_1+e}\right) \mathscr{L}_m \tag{C.40}$$

that can be rewritten taking cell averages as common factors, leading to

$$p_{r_1,r_2}(x,y) = \sum_{e=0}^{k-1}\sum_{m=e+1}^{k} \mathscr{L}_m \Delta x_{i-r_1+e}\, p(r_2,k,y,\bar{\mathbf{v}}^e_{r_1,r_2}) \tag{C.41}$$

Making use of the definition of $C^{(k)}_{r_1,e}(x)$ in (C.26), Equation (C.41) can be expressed in a more compact form as

$$p_{r_1,r_2}(x,y) = \sum_{e=0}^{k-1} C^{(k)}_{r_1,e}(x)\, p(r_2,k,y,\bar{\mathbf{v}}^e_{r_1,r_2}) \tag{C.42}$$

that corresponds to a 1D reconstruction along x with departing data provided by the 1D reconstruction $p(r_2,k,y,\bar{\mathbf{v}}^e_{r_1,r_2})$. Equation (C.42) can be expressed in its compact form using (A.27), as

$$p_{r_1,r_2}(x,y) = p\left(r_1,k,x,\left\{p(r_2,k,y,\bar{\mathbf{v}}^e_{r_1,r_2})\right\}_{e=0,\ldots,k-1}\right) \tag{C.43}$$

Notice that substitution of (C.35) in (C.42) leads to the general expression for the reconstruction in 2D presented in (C.28), that can be rewritten in recursive (dimension-by-dimension) form as

$$p_{r_1,r_2}(x,y) = \sum_{e=0}^{k-1} C^{(k)}_{r_1,e}(x)\left(\sum_{d=0}^{k-1} C^{(k)}_{r_2,d}(y)\, \bar{u}_{i-r_1+e,j-r_2+d}\right) \tag{C.44}$$

A simpler and analogous derivation of the dimension-by-dimension approach for 2D polynomial reconstruction can be carried out by introducing the variable $\bar{\zeta}_i(y)$ as

$$\bar{\zeta}_i(y) = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} u(x,y)\,dx \tag{C.45}$$

that stands for the x-line averages along y, inside cells $I_{i,j}$ at column i. The use of this variable makes possible to rewrite cell averages as

$$\bar{u}_{i,j} = \frac{1}{\Delta x \Delta y} \int \int_{x,y \in I_{i,j}} u(x,y)\, dx\, dy = \frac{1}{\Delta y} \int_{y \in I_{i,j}} \bar{\zeta}_i(y)\, dy \tag{C.46}$$

and to notice that the application of the dimension-by-dimension reconstruction is straightforward. First, line averages $\bar{\zeta}_i(y)$ are reconstructed for a certain y value by means of Equation (C.35) and departing from cell averages $\bar{u}_{i,j}$. This reconstruction is carried out for all columns composing the stencil, given by parameter e, and provides the new 1D averages used as departing data in the second reconstruction. Finally, this second reconstruction is carried out using polynomial in (C.42) for a certain x value, leading to Equation (C.44).

C.3 Dimension-by-dimension 2D WENO reconstruction

As outlined in the previous section, it is possible to construct a conventional 2D reconstruction by means of two nested 1D reconstructions in each of the coordinate directions. In the same way, it will be possible to generate a 2D WENO reconstruction by carrying out two successive 1D WENO reconstructions.

For the generation of a $2k-1$-th order 2D WENO reconstruction inside the cell $I_{i,j}$, k^2 different stencils will be needed. The candidate stencils are given by

$$S_{r_1,r_2}(i,j) \qquad \forall r_1, r_2 = 0, ..., k-1 \tag{C.47}$$

with $S_{r_1,r_2}(i,j)$ defined in (C.7). Moreover, a bigger stencil is defined as the union of the smaller stencils $S_{r_1,r_2}(i,j)$

$$\mathscr{T}(i,j) = \bigcup_{r_1,r_2 \in [0,...,k-1]} S_{r_1,r_2}(i,j) \tag{C.48}$$

noticing the following property

$$\bigcap_{r_1,r_2 \in [0,...,k-1]} S_{r_1,r_2}(i,j) = I_{i,j} \tag{C.49}$$

To compute a 2D WENO reconstruction inside cell $I_{i,j}$, the first step is to obtain $2k-1$ 1D WENO reconstructions along y, referred to as $\tilde{q}^e_{r_1}(y)$, for each x column of $\mathscr{T}(i,j)$ departing from cell averages grouped in k 1D stencils according to parameter r_2. These reconstructions will provide new one-dimensional average-like values, grouped in k 1D stencils according to parameter r_1, to generate another 1D WENO reconstruction along x, referred to as $\tilde{\tilde{q}}(x,y)$. The procedure to compute a 5-th order 2D WENO reconstruction inside cell $I_{i,j}$ is entirely depicted in Figure C.2.

As pointed out in the previous paragraph, the 1D reconstructions along y are carried out for each of the $2k-1$ columns of the big stencil $\mathscr{T}(i,j)$. For each column, k different stencils are taken in the y direction according to parameter r_2. Moreover, for each stencil characterized by r_2, k-th order 1D reconstructing polynomials $\tilde{p}^e_{r_2}(y)$ are calculated using Equation (C.35) as

$$\tilde{p}^e_{r_2}(y) = p\left(r_2, k, y, \left\{\bar{v}^e_{r_1,r_2}\right\}_{d=0,...,k-1}\right) \tag{C.50}$$

with $r_1 = k-1$, $e = 0, ..., 2k-2$ and $r_2 = 0, ..., k-1$.

The $2k-1$-th order WENO reconstruction $\tilde{q}^e_{r_1}(y)$ based on polynomials $\tilde{p}^e_{r_2}(y)$ is expressed according to (A.70) as

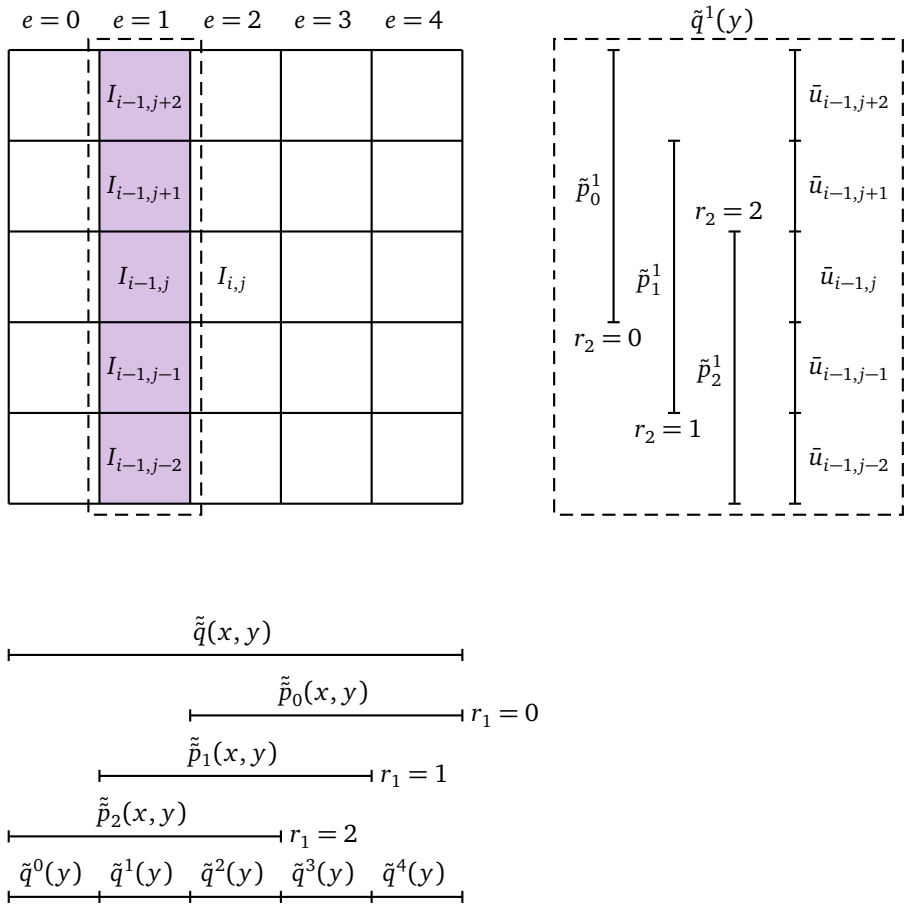


Figure C.2: 5-th order ($k=3$) 2D WENO reconstruction for cell $I_{i,j}$ inside stencil $\mathscr{T}(i,j)$ using two 1D sweeps. The first 1D sweep, along y direction, is depicted for $e=1$.

$$\tilde{q}^e(y) = q\left(k, y, \left\{\tilde{p}^e_{r_2}(y)\right\}_{r_2=0,\ldots,k-1}\right) \tag{C.51}$$

for each $e=0,\ldots,2k-2$.

WENO reconstruction in (C.51) provides new one-dimensional average-like values along the x direction. Therefore, it is possible to repeat the previous procedure but in this case departing from $\tilde{q}^e(y)$ instead of cell averages of u. Now, as in the previous case k different stencils are taken according to parameter r_1 with corresponds to the x direction. Inside each stencil, the the following polynomials are constructed

$$\tilde{\tilde{p}}_{r_1}(x,y) = p\left(r_1, k, x, \left\{\tilde{q}^e(y)\right\}_{e=i-r_1,\ldots,i-r_1+k-1}\right) \tag{C.52}$$

Finally, a $2k-1$-th order WENO reconstruction is carried out using polynomials in (C.52) as

$$\tilde{\tilde{q}}(x,y) = q\left(k, x, \left\{\tilde{\tilde{p}}_{r_1}(x,y)\right\}_{r_1=0,\ldots,k-1}\right) \tag{C.53}$$

Up to this point, the reconstruction procedures have been only considered inside a cell $I_{i,j}$ and therefore subscripts denoting for row and column position, i and j respectively, were dropped from polynomials. In what follows, the 1D WENO reconstruction in y direction will be denoted by $\tilde{q}_{i,j}(y)$, which is equivalent to $\tilde{q}^{k-1}(y)$ according to Equation (C.51). Similarly, the 2D WENO reconstruction will be denoted by $\tilde{\tilde{q}}_{i,j}(x,y)$.

D 2D SUB-CELL WENO RECONSTRUCTION OF DERIVATIVES

The previously presented 1D sub-cell WENO reconstruction of derivatives is now extended to 2 spatial dimensions, again by means of a dimension-by-dimension reconstruction approach.

D.1 Derivation and description of the procedure

The procedure for the WENO sub-cell derivative reconstruction procedure is composed of the following steps:

a) *Define sub-cell points*

Sub-cell points for the cell $I_{i,j}$ are denoted by $[x_i^{(b)}, y_j^{(v)}]$ for $b, v = 1, ..., 2k-2$. In Section 1, it was shown that choosing uniformly distributed points inside the cell leads to negative optimal weights, γ_r, in the WENO reconstruction procedure. Therefore, the WENO procedure would require some modifications in order to give a good treatment to the negative weights.

As done in the 1D case, all sub-cell points are chosen to be inside the positive interval of the optimal weights, according to Figure B.1. The same formula is used for the generation of the subcell grid, for each of the coordinate directions:

$$x_i^{(b)} = \begin{cases} x_{i-\frac{1}{2}} & \text{if } b = 1 \\ x_{i-\frac{1}{2}} + \frac{\Delta x}{2 \cdot 2^{k-b}} & \text{if } 2 \leq b \leq k-1 \\ x_{i+\frac{1}{2}} - \frac{\Delta x}{2 \cdot 2^{b-k+1}} & \text{if } k \leq b \leq 2k-3 \\ x_{i+\frac{1}{2}} & \text{if } b = 2k-2 \end{cases}$$

$$y_j^{(v)} = \begin{cases} x_{j-\frac{1}{2}} & \text{if } v = 1 \\ x_{j-\frac{1}{2}} + \frac{\Delta x}{2 \cdot 2^{k-v}} & \text{if } 2 \leq v \leq k-1 \\ x_{j+\frac{1}{2}} - \frac{\Delta x}{2 \cdot 2^{v-k+1}} & \text{if } k \leq v \leq 2k-3 \\ x_{j+\frac{1}{2}} & \text{if } v = 2k-2 \end{cases}$$

b) *Reconstruct an interpolating polynomial that approximates line averages of $u(x, y)$ and its derivatives, in y direction, inside each cell:*

Line averages of $u(x, y)$ in y direction are given by $\bar{\zeta}_i(y)$, defined in (C.45). The 1D WENO sub-cell derivative reconstruction procedure in Section .. is applied to obtain an approximation of $\bar{\zeta}_i(y)$

and its spatial derivatives departing from cell averages $\bar{u}_{i,j}$. The polynomial that approximates $\bar{\zeta}_i(y)$ inside cell $I_{i,j}$ is denoted by $\bar{\phi}_{i,j}(y)$ and defined as

$$\bar{\phi}_{i,j}(y) = \sum_{l=0}^{2k-2} a_l \left(\frac{y - y_{j-\frac{1}{2}}}{\Delta y} \right)^l \tag{D.1}$$

where the coefficients a_l $(l = 0, ..., 2k-2)$ have to be determined with the following $2k-1$ equations for each cell $I_{i,j}$:

- Using the $2k-2$ reconstructed values of $\bar{\zeta}_i(y)$ at the $2k-2$ sub-cell points obtained with the WENO procedure, given by

$$\bar{\zeta}_{i,j}^{(v)} \approx \bar{\zeta}_i(y_j^{(v)}) \tag{D.2}$$

  we set

$$\bar{\phi}_{i,j}\left(y_j^{(v)}\right) = \bar{\zeta}_{i,j}^{(v)}, \qquad v = 1, ..., 2k-2 \tag{D.3}$$

- Using the cell average, $\bar{u}_i$, the following equation is formulated

$$\frac{1}{\Delta y} \int_{I_{i,j}} \bar{\phi}_{i,j}(y)\, dy = \sum_{l=0}^{2k-2} \frac{a_l}{l+1} \left[\left(\frac{y - y_{j-\frac{1}{2}}}{\Delta y} \right)^{l+1} \right]_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} = \bar{u}_{i,j} \tag{D.4}$$

Polynomial $\bar{\phi}_{i,j}(y)$ is determined, providing the following approximations

$$\begin{aligned} \bar{\phi}_{i,j}(y) &\approx \bar{\zeta}_i(y) \\ \frac{\partial^m}{\partial y^m} \bar{\phi}_{i,j}(y) &\approx \frac{\partial^m}{\partial y^m} \bar{\zeta}_i(y) \end{aligned} \tag{D.5}$$

c) *Reconstruct an interpolating polynomial that approximates $u(x, y)$ and its derivatives in x, inside each cell and at each quadrature point:*

The 1D WENO sub-cell derivative reconstruction procedure in Section .. is applied now in the x direction to obtain an approximation of $u(x, y)$ and its spatial derivatives in x, at each quadrature y-point $\mathscr{G}_{y_j}^{(e_2)}$, departing from line averages $\bar{\zeta}_i(\mathscr{G}_{y_j}^{(e_2)})$ obtained in (D.5). The polynomial that approximates $u(x, y)$ inside cell $I_{i,j}$, along x and at $\mathscr{G}_{y_j}^{(e_2)}$, is denoted by $\phi_{i,j}^{(0),(e_2)}(x)$ and defined as

$$\phi_{i,j}^{(0),(e_2)}(x) = \sum_{l=0}^{2k-2} b_l^{(0)} \left(\frac{x - x_{i-\frac{1}{2}}}{\Delta x} \right)^l \tag{D.6}$$

where the coefficients $b_l^{(0)}$ $(l = 0, ..., 2k-2)$ have to be determined with the following $2k-1$ equations for each cell $I_{i,j}$ and for each quadrature point:

- Using the $2k-2$ reconstructed values of $u(x, \mathscr{G}_{y_j}^{(e_2)})$ at the $2k-2$ x-sub-cell points obtained with the WENO procedure, denoted by

$$u_{i,j}^{(b,e_2)} \approx u(x_i^{(b)}, \mathscr{G}_{y_j}^{(e_2)}) \tag{D.7}$$

  we set

$$\phi_{i,j}^{(0),(e_2)}\left(x_i^{(b)}\right) = u_{i,j}^{(b,e_2)}, \qquad b = 1, ..., 2k-2 \tag{D.8}$$

- Using the cell average, $\bar{\zeta}_i(\mathscr{G}_{y_j}^{(e_2)})$, the following equation is formulated

$$\frac{1}{\Delta x}\int_{I_{i,j}} \phi_{i,j}^{(0),(e_2)}(x)\,dx = \sum_{l=0}^{2k-2} \frac{b_l^{(0)}}{l+1}\left[\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta x}\right)^{l+1}\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} = \bar{\zeta}_i(\mathscr{G}_{y_j}^{(e_2)}) \tag{D.9}$$

Polynomials $\phi_{i,j}^{(0),(e_2)}(x)$ are determined at each $\mathscr{G}_{y_j}^{(e_2)}$, providing the following approximations

$$\begin{aligned} \phi_{i,j}^{(0),(e_2)}(x) &\approx u(x,\mathscr{G}_{y_j}^{(e_2)}) \\ \frac{\partial^n}{\partial x^n}\phi_{i,j}^{(0),(e_2)}(x) &\approx \frac{\partial^n}{\partial x^n}u(x,\mathscr{G}_{y_j}^{(e_2)}) \end{aligned} \tag{D.10}$$

and allow to compute the approximation of $u(x,y)$ and its derivatives at the quadrature points $(x,y)=(\mathscr{G}_{x_i}^{(e_1)},\mathscr{G}_{y_j}^{(e_2)})$ by evaluating (D.10) at $x=\mathscr{G}_{x_i}^{(e_1)}$.

d) *Reconstruct interpolating polynomials for cross derivatives and y derivatives of $u(x,y)$, along x, inside each cell and for each quadrature point:*

The calculation of derivatives of $u(x,y)$ in the x direction is straightforward when departing from information related to x-averaged values, as done in the previous step. However, the computation of derivatives in the y direction as well as cross derivatives require an additional step since the former were already calculated in the second step as averages in x.

Now, we seek derivatives of the type

$$\frac{\partial^{n+m}}{\partial x^n \partial y^m}u(x,y) \tag{D.11}$$

with $m+n \leq 2k-2$ and $m>0$, since the case when $m=0$ corresponds to the previous step. The departing data will be x-averages of (D.11) with $n=0$ and $m=1,...,2k-2$, that can be expressed as derivatives of line averages $\bar{\zeta}_i(y)$

$$\frac{\partial^m}{\partial y^m}\bar{\zeta}_i(y) = \int_{x_{i-1/2}}^{x_{i+1/2}} \frac{\partial^m}{\partial y^m}u(x,y)dx \tag{D.12}$$

computed straightforward from derivatives of $\bar{\phi}_{i,j}(y)$ in (D.5).

For each value of m, the 1D WENO sub-cell derivative reconstruction procedure is applied in the x direction to obtain an approximation of $\frac{\partial^m}{\partial y^m}u(x,y)$ and its spatial derivatives in x (cross derivatives). This procedure will be carried out at each quadrature y-point $\mathscr{G}_{y_j}^{(e_2)}$.

The polynomial that approximates $\frac{\partial^m}{\partial y^m}u(x,y)$ inside cell $I_{i,j}$, along x and at $\mathscr{G}_{y_j}^{(e_2)}$, is denoted by $\phi_{i,j}^{(m),(e_2)}(x)$ and defined as

$$\phi_{i,j}^{(m),(e_2)}(x) = \sum_{l=0}^{2k-2} b_l^{(m)}\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta x}\right)^l \tag{D.13}$$

where the coefficients $b_l^{(m)}$ $(l=0,...,2k-2)$ have to be determined with the following $2k-1$ equations for each cell $I_{i,j}$ and for each quadrature point:

- Using the $2k-2$ reconstructed values of $\frac{\partial^m}{\partial y^m}u(x,y)\Big|_{y=\mathscr{G}_{y_j}^{(e_2)}}$ at the $2k-2$ x-sub-cell points obtained with the WENO procedure, we set

$$\phi_{i,j}^{(m),(e_2)}\left(x_i^{(b)}\right) = \frac{\partial^m}{\partial y^m} u(x,y)\Big|_{\substack{y=\mathscr{G}_{y_j}^{(e_2)} \\ x=x_i^{(b)}}}, \qquad b = 1,\dots,2k-2 \tag{D.14}$$

- Using the cell average of the m-th derivative in y direction, $\frac{\partial^m}{\partial y^m}\bar{\zeta}_i(y)\Big|_{y=\mathscr{G}_{y_j}^{(e_2)}}$, the following equation is formulated

$$\frac{1}{\Delta x}\int_{I_{i,j}} \phi_{i,j}^{(m),(e_2)}(x)\,dx = \sum_{l=0}^{2k-2} \frac{b_l^{(m)}}{l+1}\left[\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta x}\right)^{l+1}\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} = \frac{\partial^m \bar{\zeta}_i(y)}{\partial y^m}\Big|_{y=\mathscr{G}_{y_j}^{(e_2)}} \tag{D.15}$$

Polynomials $\phi_{i,j}^{(m),(e_2)}(x)$ are determined at each quadrature point $\mathscr{G}_{y_j}^{(e_2)}$, providing the following approximations

$$\begin{aligned} \phi_{i,j}^{(m),(e_2)}(x) &\approx \frac{\partial^m}{\partial y^m} u(x, \mathscr{G}_{y_j}^{(e_2)}) \\ \frac{\partial^n}{\partial x^n}\phi_{i,j}^{(m),(e_2)}(x) &\approx \frac{\partial^n}{\partial x^n}\left(\frac{\partial^m}{\partial y^m} u(x, \mathscr{G}_{y_j}^{(e_2)})\right) \end{aligned} \tag{D.16}$$

with $n \leq 2k-2-m$. They allow to compute the approximation of $u(x,y)$ and its derivatives at the quadrature points $(x,y) = (\mathscr{G}_{x_i}^{(e_1)}, \mathscr{G}_{y_j}^{(e_2)})$ by evaluating (D.16) at $x = \mathscr{G}_{x_i}^{(e_1)}$.