Analysis and development of finite volume methods for the new generation of cubed sphere dynamical cores for the atmosphere

Luan da Fonseca Santos

Report presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
for the Doctor of Science
qualifying examination

Program: Applied Mathematics
Advisor: Prof. Pedro da Silva Peixoto

During the development of this work the author was supported by CAPES and FAPESP (grant number 20/10280-4)

São Paulo
November, 2022

Analysis and development of finite volume methods for the new generation of cubed sphere dynamical cores for the atmosphere

Luan da Fonseca Santos

This is the original version of the
qualifying text prepared by candidate
Luan da Fonseca Santos, as submitted
to the Examining Committee.

Resumo

Luan da Fonseca Santos. **Análise e desenvolvimento de métodos de volumes finitos para modelos da nova geração da dinâmica atmosférica baseados na esfera cubada**. Exame de Qualificação (Doutorado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022.

O modelo atmosférico global FV3 do GFDL-NOAA-USA, inicialmente desenvolvido para malhas do tipo latitude-longitude, foi adaptado para a esfera cubada visando atingir melhor escalabilidade em super-computadores massivamente paralelos. Entretanto, neste tipo de malhas estamos mais sujeitos à problemas como o grid imprinting. Além disso, o modelo carece de algumas propriedades miméticas, que são altamente desejáveis. Este projeto de doutorado propõe-se a analisar as propriedades das discretizações de volumes finitos utilizadas no modelo FV3 na esfera cubada. Iremos investigar como propriedades das células da esfera cubada interferem na precisão dos esquemas numéricos. O estudo irá começar com a implementação de um código para gerar a esfera cubada e calcular os operados discretos do FV3. Então, iremos analisar como a malha interfere nos modelos de advecção e de águas rasas na esfera.

Palavras-chave: Núcleo dinâmico da atmosfera, esfera cubada, volumes finitos.

Abstract

Luan da Fonseca Santos. **Analysis and development of finite volume methods for the new generation of cubed sphere dynamical cores for the atmosphere**. Qualifying Exam (Doctorate). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2022.

The global atmospheric model FV3 from GFDL-NOAA-USA, which was originally designed for latitude-longitude grids, was adapted to the cubed sphere aiming to improve its scalability in massively parallel supercomputers. However, in this kind of grid, we are more likely to have grid imprinting problems. Besides that, the FV3 model lacks some highly desirable mimetic properties. This work aims to analyze the properties of the finite volume discretizations employed in the global atmospheric model FV3 on the cubed-sphere. We will investigate how the properties of the cells may impact on the accuracy of the numerical schemes. This study will firstly implement a cubed-sphere grid generator and the FV3 discrete operators on this grid. Then, we will and analyze how the cubed-sphere grid properties influence in the numerical schemes by assessing it using the advection and shallow-water equations on the sphere. We will study the numerical dispersion and conservations properties of the scheme aiming to propose modifications in the numerical schemes to develop a mimetic finite volume version of the model.

Keywords: Dynamical core, cubed-sphere, finite-volume.

Contents

Appendixes

Chapter 1

Introduction

1.1 Background

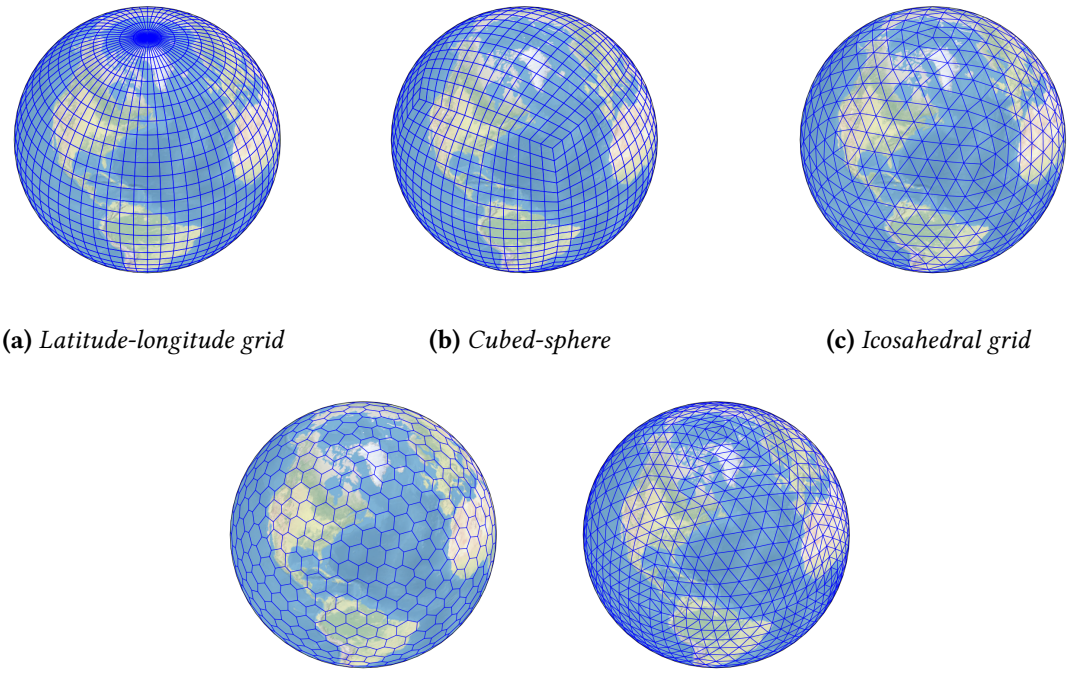
Weather and climate predictions are recognized as a good for mankind, due to the information they yield for diverse activities. For instance, short-range forecasts are useful for public use, while medium-range forecasts are helpful for industrial activities and agriculture. Seasonal forecasts (one up to three months) are important to energy planning and agriculture. At last, longer-range forecasts (one century, for instance) are useful for climate change projections that are important for government planning.

The first global Numerical Weather Prediction models emerged in the 1960s with applications to weather, seasonal and climate forecasts. All these applications are essentially based on the same set of Partial Differential Equations (PDEs) but with distinct time scales (D. L. Williamson, 2007). These PDEs are defined on the sphere and model the evolution of the atmospheric fluid given the initial conditions. One important component of global models is the dynamical core, which is responsible for solving the PDEs that governs the atmosphere dynamics on grid-scale. The development of numerical methods for dynamical cores has been an active research area since the 1960s.

Global models use the sphere as the computational domain and therefore they require a discretization of the sphere. The first global models used the latitude-longitude grid (Figure 1.1a), which is very suitable for finite-differences schemes due to its orthogonality. The major drawback of the latitude-longitude grid is the clustering of points at the poles, known as the "pole problem", which leads to extremely small time steps for explicit-in-time schemes due to the Courant-Friedrichs-Lewy (CFL) condition, making these schemes computationally very expensive.

The most successful method adopted in global atmospheric dynamical cores that overcomes the CFL restriction is the Semi-Implicit Semi-Lagrangian (SI-SL) scheme (Randall et al., 2018), which emerged in the 1980s and consists of the Lagrangian advection scheme applied at each time-step and the solution of fast gravity waves implicitly, allowing very large time steps despite the pole problem. The SI-SL approach combined with finite differences is still used nowadays, for instance in the UK Met Office global model ENDGame (Benacchio & Wood, 2016; Wood et al., 2014). The expensive part of the SI-SL approach is to

solve an elliptic equation at each time step, that comes from the semi-implicit discretization, which requires global data communication, being inefficient to run in massive parallel supercomputers. Besides that, Semi-Lagrangian schemes are inherently non-conservatives for mass, which is critical for climate forecasts (D. L. Williamson, 2007).



(a) *Latitude-longitude grid* **(b)** *Cubed-sphere* **(c)** *Icosahedral grid*

(d) *Pentagonal/Hexagonal grid* **(e)** *Octahedral grid*

Figure 1.1: *Examples of spherical grids: latitude-longitude grid (a) and grids based on Platonic solids (b)-(d).*

The emergence of the Fast Fourier Transform (FFT) in the 1960s with the work from Cooley and Tukey (1965) allowed the computation of discrete Fourier transforms with $N \log(N)$ complexity. The viability of the usage of FFTs for solving atmospheric flows was shown by Orszag (1970), using the barotropic vorticity equation on the sphere, and by Eliasen et al. (1970), using the primitive equations. The spectral transform method expresses latitude-longitude grid values, that represent some scalar field, using truncated spherical harmonics expansions, which consists of Fourier expansions in latitude circles and Legendre functions expansions in longitude circles. The coefficients in the spectral expansions are known as spectral coefficients and are usually thought to live in the so-called spectral space. Given the grid values, the spectral coefficients are obtained by performing a FFT followed by a Legendre Transform (LT). Conversely, given the spectral coefficients, the grid values are obtained by performing an inverse LT followed by an inverse FFT. The main idea of the spectral method is to apply the spectral transform, in order to go the spectral space, and evaluate spatial derivatives in the spectral space, which consists of multiplying the spectral coefficients by constants. Then, the method performs the inverse spectral transform in order to get back to grid space, and the nonlinear terms are treated on the grid space (Krishnamurti et al., 2006).

The spectral transform makes the use of SI-SL methods computationally cheap, since the solution to elliptic problems becomes easy, once the spherical harmonics are eigenfunctions of the Laplacian operator on the sphere. Therefore, the spectral transform method gets faster when combined with the SI-SL approach due to the larger times-steps allowed in this case. Due to these enhancements, the spectral transform dominated global atmospheric modeling (Randall et al., 2018) since the 1980s. Indeed, the spectral method is still used in many current operational Weather Forecasting models such as the Integrated Forecast System (IFS) from European Centre for Medium-Range Weather Forecasts (ECMWF), Global Forecast System (GFS) from National Centers for Environmental Prediction (NCEP) and the Brazilian Global Atmospheric Model (BAM) (Figueroa et al., 2016) from Center for Weather Forecasting and Climate Research [Centro de Previsão de Tempo e Estudos Climáticos (CPTEC)].

With the beginning of the multicore era in the 1990s, the global atmospheric models started to move towards parallel efficiency aiming to run at very high resolutions. Even though the spectral transform expansions have a global data dependency, some parallelization is feasible among all the computations of FFTs, LTs and their inverses (Barros et al., 1995). However, the parallelization of the spectral method requires data transpositions in order to compute FFTs and LTs in parallel. These transpositions demand a lot of global communication using, for instance, the Message Passing Interface (MPI) (Zheng & Marguinaud, 2018). Indeed, the spectral transform becomes the most expensive component of global spectral models when the resolution is increased due to the amount of MPI communications (Müller et al., 2019).

The adiabatic and frictionless continuous equations that govern the atmospheric flow have conserved quantities. Among them, some of the most important are mass, total energy, angular momentum and potential vorticity (Thuburn, 2011). Numerical schemes that are known for having discrete analogous of these conservative properties are known as mimetic schemes. As we pointed out, Semi-Lagrangian schemes lack mass conservation. Nevertheless, these schemes have been employed in dynamical cores for better computational performance. However, dynamical cores should have discrete analogous of the continuous conserved quantities, especially concerning for longer simulation runs.

Aiming for better performance in massively parallel computers and conservation properties, new dynamical cores have been developed since the beginning of the 2000s. Novel spherical grids have been proposed, in order to avoid the pole problem. A popular choice are grids based on Platonic solids (Staniforth & Thuburn, 2012). The construction of these grids relies on a Platonic circumscribed on the sphere and the projection of its faces onto the sphere, which leads to quasi-uniform and more isotropic spherical grids. Some examples of spherical grids based on Platonic solids employed in the new generation of dynamical cores are the cubed-sphere (Figure 1.1b), icosahedral grid (Figure 1.1c), the pentagonal/hexagonal or Voronoi grid (Figure 1.1d) and octahedral grid (Figure 1.1e), which are based on the cube, icosahedron, dodecahedron and octahedron, respectively (Ullrich et al., 2017).

1.2 Motivations

The cubed-sphere became a popular quasi-uniform grid for the new generation of dynamical cores. It was originally proposed by Sadourny (1972) and it was reviseted by Ronchi et al. (1996). Some of the cubed-sphere advantages are: uniformity; quadrilateral structure, making the grid indexing trivial; no overlappings; it is cheap to generate. However, the major drawbacks of the cubed-sphere are: non-orthogonal coordinate system, which leads to metric terms on the differetial operator; discontinuity of the coordinate system at the cube edges, which may generate numerical noise and demands special treatment of discrete operators at the cube edges.

Despite of its drawbacks, the cubed-sphere has been adpoted in some of the new generation dynamical cores. For instance, the cubed-sphere is used in the Community Atmosphere Model (CAM-SE) from the NCAR using spectral elements (Dennis et al., 2012) and in the Nonhydrostatic Unified Model of the Atmosphere (NUMA) from the US Navy using Discontinuous Galerkin methods (Giraldo et al., 2013). The cubed-sphere was also chosen to be used in the next UK Met Office global model using mixed finite elements (Kent et al., 2022). At last, the Finite Volume Cubed-Sphere dynamical core (FV3) from the Geophysical Fluid Dynamics Laboratory (GFDL) and the National Oceanic and Atmospheric Administration (NOAA) (L. M. Harris & Lin, 2013; Putman & Lin, 2007) is another example of new generation dynamical core based on the cubed-sphere.

The FV3 model is an extension of the Finite-Volume dynamical core (FVcore) from latitude-longitude grids to the cubed-sphere. The numerical methods from FVcore started to be developed with the transport scheme from the work Lin et al. (1994), which is based on the piecewise linear scheme from Van Leer (1977). This scheme was later improved, using the Piecewise Parabolic Method (PPM) (Carpenter et al., 1990; Colella & Woodward, 1984) using dimension splitting techniques that guarantee monotonicity and mass conservation, for the transport equation (Lin & Rood, 1996) and the shallow-water equations (Lin & Rood, 1997). An important feature is that the FVcore combines the Arakawa C- and D-grids (Arakawa & Lamb, 1977), where the C-grid values are computed in and intermediate time step. The full global model was then presented by Lin (2004).

The FVcore was adapted to the cubed-sphere grid (Putman, 2007; Putman & Lin, 2007), to reach better performance in parallel computers, leading to the FV3 model. Later, the FV3 also was improved to allow locally refinement grids through grid-nesting or grid-stretching (L. M. Harris & Lin, 2013). Currently, the FV3 model is capable of perfoming hydrostatic and non-hydrostatic atmospheric simulations and it was chosen as the new US global weather prediction model, indeed, it replaced the spectral transform Global Forecast System (GFS) in June, 2019 (Samenow, 2019).

However, a well-known problem that occurs on cubed-sphere models that use low-order numerical methods is the grid imprinting visible due to the coordinate system discontinuity, especially at larger scales, leading to the emergence of a wavenumber 4 pattern. This was reported in the paper of Rančić et al. (2017), where the authors employ a finite-difference numerical scheme on the Uniform Jacobian cubed-sphere using a Arakawa B-grid. The unpublished report from Whitaker (2015) shows grid imprinting in other models, including the FV3. Generally speaking, grid imprinting is the presence of artificial behaviors on

the numerical solution that is associated with the grid employed. It is important to stress out that other quasi-uniform grids may also suffer from grid imprinting. For instance, a popular mimetic method, known as TRiSK, was proposed in the literature by Thuburn et al. (2009) and Ringler et al. (2010) using finite difference and finite volume schemes. This scheme is designed for general orthogonal grids, such as the Voronoi and icosahedral grids, and ensures mass and total energy conservation. This method has been employed in the dynamical core of the Model for Prediction Across Scales (MPAS) from National Center for Atmospheric Research (NCAR) (Skamarock et al., 2012), which intended to work on general Voronoi grids, including locally refined Voronoi grids. However, the TRiSK scheme is a low-order scheme and also suffers from grid imprinting, *i.e.*, geometric properties of the grid, such as cell alignment, interfere with the method accuracy (Peixoto, 2016; Peixoto & Barros, 2013; Weller, 2012). Furthermore, in locally refined Voronoi grids, the scheme may become unstable due to ill-aligned cells and numerical dissipation is needed (Santos & Peixoto, 2021), breaking the total energy conservation of the method.

Despite being chosen as the new US global weather prediction model, there is a lack of numerical studies of the FV3 discretizations in the literature, especially regarding the grid imprinting problem and its mimetic properties. Numerical results for the advection equation on the cubed-sphere using the FV3 dynamical core was presented in Putman and Lin (2007) and some shallow-water simulations were presented in L. M. Harris and Lin (2013), considering cubed-spheres with local refinement through grid nesting. From the work L. M. Harris and Lin (2013) we can notice that the FV3 dynamical lack convergence on the maximum norm for the shallow-water model considering the classical balanced geostrophic flow test case from D. Williamson et al. (1992). The authors attribute these errors to the abrupt change in the grid resolution near the nested grid, but no quantitative results are shown considering the quasi-uniform grid. Many other papers available in the literature use the complete FV3 model in three-dimensional frameworks which make it harder to perform a numerical analysis study due not only to its computational cost but also due to the complexity of three-dimensional atmospheric models. There are no detailed works published in intermediate two-dimensional frameworks, using, for instance, the shallow-water equations on the sphere. Even though the advection equation on the sphere plays a key role in the dynamical core development, since it models the transport of scalar fields on the sphere, important features captured by the shallow-water equations on the sphere, such as the Coriolis effect, inertia-gravity waves, geostrophic adjustment, Rossby waves, among others, are not captured by a simple advection model. Hence, shallow-water equations provide an excellent benchmark to assess dynamical cores in general, since it is only two-dimensional but is a complex enough geophysical model for atmosphere dynamics.

1.3 Goals

The aim of this work is to fill the gap in the literature regarding numerical studies of the FV3 discrete operators that we pointed out before. More explicitly, the goals of this work are:

- Investigate the occurrence of grid imprinting on the cubed-sphere using the advection equations and the shallow-water equations on the sphere;

- Propose improvements on the FV3 discrete operators and modifications on the cubed-sphere that alleviate grid imprinting;
- Investigate how we can add more mimetic properties to the FV3 discretizations.

1.4 Outline

This report is outlined as follows. Chapter 2 is dedicated to review the Piecewise Parabolic Method (PPM) for the one-dimensional advection equation. Chapter 3 reviews the dimension splitting method, which allow us to use one-dimensional methods, such as the PPM, to solve the two-dimensional advection equation. Chapter 4 introduces the cubed-sphere grid and shows some of its geometric properties. Chapter 5 extends the ideas of Chapter 3 to the cubed-sphere grid. The dimension-splitting method on each cubed-sphere panel works as in the plane, with the addtion of metric terms, due to non-orthogonality of the grid, and interpolation between panels to obtain ghost cells values needed for stencil computations.

Chapter 2

One-dimensional finite-volume methods

The aim of this chapter is to give a detailed description of the celebrated Piecewise-Parabolic Method (PPM) proposed by Colella and Woodward (1984). As we shall see, the PPM is a one-dimensional finite-volume method for hyperbolic conservation laws that at each time step requires two tasks. The first task may be stated as: given the estimates of average values of the conservation laws solution, find a Piecewise-Parabolic function that approximates the function and preserves its local integral value (also referred as local mass). The second task is the following: given the Piecewise-Parabolic approximation (also known as reconstruction), solve the conservation law using the parabolas to obtain the solution at the next time-step. For instance, if the conservation law is the advection equation, the second step consists of advecting the parabolas. In the first step, we may also require some monotonization constraints on the parabolas, to ensure that no new extreme value is created in the Piecewise-Parabolic reconstruction, ensuring that the scheme is free of numerical oscillations. The steps required for PPM make it an REA (reconstruct, evolve, and average) algorithm, or also referred to as a Godonuv-type method, which was originally proposed by Godunov (1959).

The PPM approach has become popular in the literature for gas dynamics simulations, astrophysical phenomena modeling (Woodward, 1986) and later on atmospheric simulations (Carpenter et al., 1990). Indeed, the PPM has been implemented in the FV3 dynamical core on its latitude-longitude grid (Lin, 2004) and cubed-sphere (Putman & Lin, 2007) versions. We point out that the reconstruction function may be built using other basis functions rather than parabolas. In fact, PPM may be thought of as an extension of the Piecewise-Linear method from Van Leer (1977), which, on the other hand, was inspired by the Piecewise-Constant method attributed to Godunov (1959). Besides that, other schemes inspired by PPM were proposed in the literature using higher-order polynomials, such as quartic polynomials (White & Adcroft, 2008). For a review of general piecewise-polynomial reconstruction we refer to the technical report from Engwirda and Kelley (2016), Lauritzen et al. (2011) and the references therein. Even though many other shapes for the basis functions are available in the literature, as well higher order schemes, L. Harris et al. (2021) points out that the PPM scheme suits well the FV3 needs in the sense of being a flexible

method that can be modified to ensure low diffusivity or shape-preservation, for example. Besides that, a finite-volume numerical method usually requires monotonicity constraints, which by Godunov's theorem, limits the order of convergence to at most 1. Thus, a higher-order scheme needs to be well-balanced on the trade-off between computational cost increasing and potential benefits.

This chapter starts with a basic review of one-dimensional conservation laws in the integral form in Section 2.1, and in Section 2.2 we set the framework of general one-dimensional finite-volumes schemes, where we also introduce concepts such as consistency, convergence and stability. Section 2.3 describes the PPM method and its convergence order analysis of its reconstruction in given in Subsection 2.3.1. Subsection 2.3.2 is dedicated to introducing possible ways to monotonize the parabolas. Subsection 2.3.3 is dedicated to the description and investigation of the PPM flux computation considering the one-dimensional advection equation as the conservation law. Section 2.4 shows some numerical results using the PPM scheme for the advection equation. At last, Section 2.5 presents some conclusions. The usage of PPM to solve two-dimensional problems will be addressed in Chapter 3.

2.1 One-dimensional conservation law in integral form

In this section, we are going to present the derivation of one-dimensional conservation laws in the integral form. The derivation presented here follows LeVeque (1990) and LeVeque (2002) closely and will be useful to fix some notation. Let us assume that x and t represent the spatial and time coordinates, respectively. Given $[x_1, x_2] \subset \mathbb{R}$, $x_1 \leq x_2$, and a time interval $[t_1, t_2] \subset]0, +\infty[$, $t_1 \leq t_2$, we aim to describe how a state variable density given by a function $q : \mathbb{R} \times [0, +\infty[\to \mathbb{R}$ evolve within time in the considered time interval, assuming that we have neither sinks nor sources for the mass of the state variable and also assuming that the mass flow rate is known.

To set the problem in more mathematical terms, let us denote by $q : \mathbb{R} \times [0, +\infty[\to \mathbb{R}$, $q = q(x, t)$, the state variable. The mass of q in $[x_1, x_2]$ at time t is defined by:

$$M_{[x_1,x_2]}(t) := \int_{x_1}^{x_2} q(x, t)\, dx. \tag{2.1}$$

We are going to assume the following physical constraints concerning the total mass of the state variable:

1. No mass is created;

2. No mass is destroyed.

Also, let us assume that the mass flow rate in a point x and at a time $t > 0$ is given by $f(q(x, t))$, where $f : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable (C^1) function. This function f is known as flux function. With the physical constraints that we imposed, the following

equation must hold for the mass:

$$\frac{d}{dt}\left(\int_{x_1}^{x_2} q(x,t)\,dx \right) = f(q(x_1,t)) - f(q(x_2,t)). \tag{2.2}$$

Equation (2.2) is known as a conservation law written in integral form and tell us how the mass $M_{[x_1,x_2]}(t)$ varies with time. Another integral form of the conservation law may be obtained integrating Equation (2.2) with respect to time in $[t_1, t_2]$ leading to:

$$\int_{x_1}^{x_2} q(x,t_2)\,dx = \int_{x_1}^{x_2} q(x,t_1)\,dx + \int_{t_1}^{t_2} f(q(x_1,t))\,dt - \int_{t_1}^{t_2} f(q(x_2,t))\,dt. \tag{2.3}$$

Assuming that q is a C^1 function, we may write:

$$\int_{t_1}^{t_2} \frac{\partial q}{\partial t}(x,t)\,dt = q(x,t_2) - q(x,t_1), \tag{2.4}$$

and

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial x}(q(x,t))\,dx = f(q(x_2,t)) - f(q(x_1,t)). \tag{2.5}$$

Replacing Equations (2.4) and (2.5) in (2.3) we get the differential form of the conservation law:

$$\int_{t_1}^{t_2} \int_{x_1}^{x_2} \left(\frac{\partial q}{\partial t}(x,t) + \frac{\partial f}{\partial x}(q(x,t)) \right) dx\,dt = 0. \tag{2.6}$$

Since Equation (2.6) must hold for all x_1, x_2, t_1 and t_2 such that $[x_1, x_2] \times [t_1, t_2] \subset \mathbb{R}\times]0, +\infty[$, we obtain the differential form of the conservation law:

$$\frac{\partial q}{\partial t}(x,t) + \frac{\partial f}{\partial x}(q(x,t)) = 0, \quad \forall (x,t) \in \mathbb{R}\times]0, +\infty[. \tag{2.7}$$

Equation (2.7) is a hyperbolic partial differential equation (LeVeque, 1990). As we will specify latter, some initial conditions will also be supposed to be known as well.

Many physically relevant equations may be written as Equation (2.7). For instance, the Burgers equation, which is obtained when $m = 1$ and $f(q) = q^2$. The Burgers equation is well known for developing shocks, even for smooth initial conditions, and is a simple prototype to study shock formation. The linear advection equation is another interesting example, which is obtained when $m = 1$ and $f(q(x,t)) = u(x,t)q(x,t)$, where $u(x,t)$ is a given velocity. Strictly speaking, the linear advection is not in the form given by the Equation (2.7) since f depends on q but also on (x,t). But, one may check that Equation (2.7) is still hyperbolic in this case. The linear advection equation will play a key role in this work due to its importance to the development of atmospheric dynamical cores.

We say that q is a strong or classical solution to the conservation law (2.7) if it is C^1 and satisfies the Equation (2.7). Applying the steps from Equation (2.3) to Equation (2.7) in reverse order, one may check that if q is a strong solution, then it satisfies the integral form (2.3) for all x_1, x_2, t_1 and t_2 such that $[x_1, x_2] \times [t_1, t_2] \subset \mathbb{R}\times]0, +\infty[$. Therefore,

Equations (2.3) and (2.7) are equivalent when q is C^1. However, the problem (2.3) can be formulated to functions that are not C^1 and have discontinuities. More generally speaking, we say that q is a weak solution if it satisfies the Equation (2.3) for all x_1, x_2, t_1 and t_2 such that $[x_1, x_2] \times [t_1, t_2] \subset \mathbb{R}\times]0, +\infty[$. Of course q and u must be such that all the integrals in Equation (2.3) makes sense. Later (Problem 2.1), we specify the spaces where q and u belong. It can be shown that this notion of weak solution is equivalent to requiring that (LeVeque, 1990):

$$\int_{-\infty}^{+\infty} \int_{0}^{+\infty} \left(\frac{\partial \phi}{\partial t}(x,t) q(x,t) + \frac{\partial \phi}{\partial x}(x,t) f(q(x,t)) \right) dt\, dx = \int_{-\infty}^{+\infty} \phi(x,0) q(x,0)\, dx, \tag{2.8}$$

$\forall \phi \in C_0^1(\mathbb{R}\times[0, +\infty[)$ where $C_0^1(\mathbb{R}\times[0, +\infty[)$ denotes the set of all continuously differentiable functions with compact support in $\mathbb{R} \times [0, +\infty[$. This formulation of weak solution is more commonly employed in the construction of Discontinuous Galerkin methods (Nair et al., 2011).

In order to develop finite-volume methods for a conservation law, it is useful to define the average value of the state variable q in the interval $[x_1, x_2]$ at a time t by:

$$Q(t) = \frac{1}{\Delta x} \int_{x_1}^{x_2} q(x,t)\, dx, \tag{2.9}$$

where $\Delta x = x_2 - x_1$. The Equation (2.2) may be rewritten in terms of Q as:

$$\frac{dQ}{dt}(t) = \frac{1}{\Delta x}(f(q(x_1,t)) - f(q(x_2,t))), \tag{2.10}$$

and so is Equation (2.3):

$$Q(t_2) = Q(t_1) + \frac{1}{\Delta x}\left(\int_{t_1}^{t_2} f(q(x_1,t))\, dt - \int_{t_1}^{t_2} f(q(x_2,t))\, dt \right). \tag{2.11}$$

To move towards finite volume schemes, we will restrict our attention to a conservation law in a bounded domain of the form $D = [a, b] \times [0, T]$, $a < b$, $T > 0$. However, we must impose some boundary conditions. One possible way that we will adopt in the text are the periodic boundary conditions:

$$q(a,t) = q(b,t), \quad \forall t \in [0, T]. \tag{2.12}$$

Also, we assume that an initial condition $q_0(x) = q(x, 0)$ is given. Thus, we have specified a Cauchy problem. We notice that Equations (2.10) and (2.11) hold for all x_1, x_2, t_1 and t_2 such that $[x_1, x_2]\times[t_1, t_2] \subset D$. So, let us discretize the domain D and write Equations (2.10) and (2.11) in terms of this discretization. Given a positive integer N_T, we define the time step $\Delta t = \frac{T}{N_T}$, $t^n = n\Delta t$, for $n = 0, 1, \cdots, N_T$. For the spatial discretization, we consider a uniformly spaced partition of $[a, b]$ given by:

$$[a,b] = \bigcup_{i=1}^{N} X_i, \text{ where } X_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \text{ and } a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N-\frac{1}{2}} < x_{N+\frac{1}{2}} = b. \tag{2.13}$$

Each interval X_i is referred to as the control volume. We shall use the notations $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and $x_i = \frac{1}{2}(x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}})$, $\forall i = 1, \cdots, N$, to define the control volume length and centroid, respectively. We also denote by $Q_i(t)$ as the average values of state variable at time t in the control volume X_i, $\forall i = 1, \cdots, N$. Replacing t_1, t_2, x_1 and x_2 by $t^n, t^{n+1}, x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$, respectively, in Equation (2.10), we get:

$$\frac{dQ_i}{dt}(t) = \frac{1}{\Delta x}(f(q(x_{i-\frac{1}{2}}, t)) - f(q(x_{i+\frac{1}{2}}, t))), \quad \forall i = 1, \cdots, N. \tag{2.14}$$

Similarly, Equation (2.11) becomes:

$$Q_i(t^{n+1}) = Q_i(t^n) + \frac{1}{\Delta x}\left(\int_{t^n}^{t^{n+1}} f(q(x_{i-\frac{1}{2}}, t))\, dt - \int_{t^n}^{t^{n+1}} f(q(x_{i+\frac{1}{2}}, t))\, dt \right), \quad \forall i = 1, \cdots, N, \quad \forall n = 0, \cdots, N_T - 1. \tag{2.15}$$

In order to use a more compact notation, it is helpful to use the following centered difference notation:

$$\delta_x g(x_i, t) = g(x_{i+\frac{1}{2}}, t) - g(x_{i-\frac{1}{2}}, t), \tag{2.16}$$

for an arbitrary vector valued function g. Using this notation, Equations (2.14) and (2.15) lead to:

$$\frac{dQ_i}{dt}(t) = -\frac{1}{\Delta x}\delta_x f(q(x_i, t)) \quad \forall i = 1, \cdots, N, \tag{2.17}$$

and

$$Q_i(t^{n+1}) = Q_i(t^n) - \frac{\Delta t}{\Delta x}\delta_x \left(\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(q(x_i, t))\, dt \right), \quad \forall i = 1, \cdots, N, \quad \forall n = 0, \cdots, N_T - 1, \tag{2.18}$$

respectively. It is worth pointing out that we have made no approximation in Equations (2.17) and (2.18). Indeed, if q satisfies Equation (2.2), $\forall [x_1, x_2] \subset [a, b]$ and $\forall t \in [0, T]$, then Equation (2.17) is just Equation (2.2) evaluated in the control volumes and written in terms of the average values Q. Similarly, if q satisfies Equation (2.3), $\forall [x_1, x_2] \times [t_1, t_2] \subset D$, then Equation (2.18) is just Equation (2.3) evaluated in the control volumes, at the time instants t^n, and written in terms of the average values Q.

Notice that in Equation (2.18) we divided and multiplied by Δt, so that we can interpret $\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f(q(x_i, t))\, dt$ as a time-averaged flux. This interpretation is very handy for the derivation of finite-volume schemes.

The formulations given by Equations (2.17) and (2.18) are the cornerstone of the development of finite volume methods for conservation laws. On the right-hand side of Equation (2.17), the flux function f may be discretized leading to an ordinary differential equation (ODE) that might be solved using classical ODE integrators. These methods are known as semi-discrete methods (LeVeque, 2002), since only the spatial coordinate is discretized. In this work, we shall restrict our attention to methods based on Equation (2.18), even though the PPM approach is applicable for semi-discrete methods (*e.g.* Suresh and Huynh (1997)).

2.2 The finite-volume approach

We summarize the problem of the system of conservation laws in the integral form discussed in Section 2.1 in Problem 2.1. For simplicity, hereafter we shall constrain our attention to the one-dimensional advection equation, that is, we are going to assume that the flux function has the form $f(q(x,t)) = u(x,t)q(x,t)$, where $u(x,t)$ is the velocity which is assumed to be given.

Since we are going to consider periodic boundary conditions, it is useful to introduce some spaces of periodic functions of period $b-a$. We are going to use the notation $\mathbb{S}^1 = [a,b]$ to represent the interval that we are interested. This notation is justified by thinking of periodic functions as functions defined on the circle of length $b-a$. Whenever we use the notation $\mathbb{S}^1$, a and b will be implicitly defined. With that said, we define:

$$\begin{aligned}
\mathcal{F}(\mathbb{S}^1) &= \{q : \mathbb{R} \to \mathbb{R}; \quad q(x+b-a) = q(x), \quad \forall x \in \mathbb{R}\},\\
\mathcal{F}(\mathbb{S}^1_T) &= \{q : \mathbb{R} \times [0,T] \to \mathbb{R}; \quad q(\cdot,t) \in \mathcal{F}(\mathbb{S}^1), \quad \forall t \in [0,T]\},\\
C^k(\mathbb{S}^1_T) &= \{q \in C^k(\mathbb{R} \times [0,T]) : q \in \mathcal{F}(\mathbb{S}^1_T)\}.
\end{aligned}$$

where we are using the notation $\mathbb{S}^1_T = \mathbb{S}^1 \times [0,T]$. We also introduce the following the locally integrable periodic functions:

$$\begin{aligned}
L^p_{\text{loc}}(\Omega) &= \{q : \Omega \to \mathbb{R}; \quad \int_K |q(x)|^p \, dx < +\infty, \quad \text{for all compact sets } K \subset \Omega\},\\
L^p_{\text{loc}}(\mathbb{S}^1) &= \{q \in \mathcal{F}(\mathbb{S}^1) : \quad q \in L^p_{\text{loc}}(\mathbb{R})\},\\
L^{p,x}_{\text{loc}}(\mathbb{S}^1_T) &= \{q \in \mathcal{F}(\mathbb{S}^1_T) : \forall t \in [0,T], \quad q(\cdot,t) \in L^p_{\text{loc}}(\mathbb{R})\},\\
L^{p,t}_{\text{loc}}(\mathbb{S}^1_T) &= \{q \in \mathcal{F}(\mathbb{S}^1_T) : \forall x \in \mathbb{R}, \quad q(x,\cdot) \in L^p_{\text{loc}}([0,T])\},\\
L^{p,x,t}_{\text{loc}}(\mathbb{S}^1_T) &= L^{p,x}_{\text{loc}}(\mathbb{S}^1_T) \cap L^{p,t}_{\text{loc}}(\mathbb{S}^1_T).
\end{aligned}$$

2.2.1 Discretization of the problem

Problem 2.1. *Given an initial condition $q_0 \in L^1_{loc}(\mathbb{S}^1) \cap L^2_{loc}(\mathbb{S}^1)$ and a velocity function $u \in L^{2,t}_{loc}(\mathbb{S}^1_T)$, we would like to find a weak solution $q \in L^{1,x,t}_{loc}(\mathbb{S}^1_T) \cap L^{2,x,t}_{loc}(\mathbb{S}^1_T)$ of the advection equation in the integral form:*

$$\int_{x_1}^{x_2} q(x,t_2)\,dx = \int_{x_1}^{x_2} q(x,t_1)\,dx + \int_{t_1}^{t_2} (uq)(x_1,t)\,dt - \int_{t_1}^{t_2} (uq)(x_2,t)\,dt,$$

$\forall [x_1,x_2] \times [t_1,t_2] \subset [a,b] \times [0,T]$, and $q(x,0) = q_0(x)$, $\forall x \in [a,b]$.

We point out that, for Problem 2.1, the total mass in $[a,b]$ satisfies:

$$M_{[a,b]}(t) = M_{[a,b]}(0), \quad \forall t \in [0,T]. \tag{2.19}$$

This is the conservation of total mass propriety and is highly desirable for any numerical scheme that intends to give a robust approximation of a conservation law solution. In Section 2.1 we introduced a version of Problem 2.1 considering a discretization of the domain $[a,b] \times [0,T]$. To define the discretization, we introduce the concepts of Δx-grid

and Δt-temporal grid and $(\Delta x, \Delta t, \lambda)$ discretization.

Definition 2.1 (Δx-grid). *For a given interval $[a,b]$ and a positive real number Δx such that $\Delta x = (b-a)/N$, for some positive integer N, we say that a N-tuple $\mathcal{X} = (X_i)_{i=1}^N$ is a Δx-grid for $[a,b]$ if $X_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, $a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N-\frac{1}{2}} < x_{N+\frac{1}{2}} = b$, $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$. Each X_i is called control volume or cell and $x_{i-\frac{1}{2}}$, $x_{i+\frac{1}{2}}$ are called edges of the control volume X_i.*

Remark 2.1. *We may define the cells X_i for i outside of the range $1, \cdots, N$ by $X_i = [a + (i-1)\Delta x, a + i\Delta x]$. These cells are called ghost cells.*

Definition 2.2 (Δt-temporal grid). *For a given interval $[0,T]$ and a positive real number Δt such that $\Delta t = T/N_T$, for some positive integer N_T, we say that a (N_T+1)-tuple $\mathcal{T} = (T_n)_{n=0}^{N_T}$ is a Δt-temporal grid for $[0,T]$ if $T_n = [t^n, t^{n+1}]$, $t^n = n\Delta t$, $\Delta t = \frac{T}{N_T}$, $\forall n = 0, \cdots, N_T$.*

Definition 2.3 ($(\Delta x, \Delta t, \lambda)$-discretization). *Given $[a,b] \times [0,T]$ and positive real numbers Δx and Δt, we say that $(\mathcal{X}, \mathcal{T})$ is a $(\Delta x, \Delta t, \lambda)-$discretization of $[a,b]\times[0,T]$ if $\mathcal{X}$ is a Δx-grid for $[a,b]$ and $\mathcal{T}$ is a Δt-temporal grid for $[0,T]$ and $\frac{\Delta t}{\Delta x} = \lambda$.*

Remark 2.2. *Whenever we mention a $\Delta x-$grid, or a Δt-temporal grid or a $(\Delta x, \Delta t, \lambda)$-discretization, then X_i, N, t^n and N_T shall be assumed implicitly defined.*

Now we define a discretized version of Problem 2.1 in Problem 2.2.

Problem 2.2. *Assume the framework of Problem 2.1 and that $(\mathcal{X}, \mathcal{T})$ is a $(\Delta x, \Delta t, \lambda)$-discretization of $[a,b] \times [0,T]$. Since we are in the framework of Problem 2.1, it follows that:*

$$Q_i(t^{n+1}) = Q_i(t^n) - \lambda \delta_x \left(\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} (uq)(x_i, t)\, dt \right), \quad \forall i = 1, \cdots, N, \quad \forall n = 0, \cdots, N_T - 1, \tag{2.20}$$

where $Q_i(t) = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x,t)\, dx$. Our problem now consists of finding the values $Q_i(t^n)$, $\forall i = 1, \cdots, N$, $\forall n = 0, \cdots, N_T - 1$, given the initial values $Q_i(0)$, $\forall i = 1, \cdots N$. In other words, we would like to find the average values of q in each control volume X_i at the considered time instants.

Next, we introduce the definitions of grid functions at cell centroids and edges.

Definition 2.4 (Δx- grid function). *For a Δx-grid $\mathcal{X}$, we say that Q is a Δx-grid function if $Q = (Q_1, \cdots, Q_N) \in \mathbb{R}^N$. We denote the space of Δx-grid functions by $\mathbb{R}^{\Delta x}$.*

Definition 2.5 (Δx-C-grid wind). *For a Δx-grid $\mathcal{X}$, we say that u is a Δx-C-grid function if $u = (u_{\frac{1}{2}}, \cdots, u_{N+\frac{1}{2}}) \in \mathbb{R}^{N+1}$. We denote the space of Δx-C-grid wind by $\mathbb{R}^{\Delta x+1}$.*

The definition of Δx-C-grid wind is based on the Arakawa grids (Arakawa & Lamb, 1977). Using a similar notation to Engwirda and Kelley (2016), we define the (r,s)-stencil and a grid function evaluated on a stencil as follows.

Definition 2.6 ((r,s)-stencil). *For a Δx-grid $\mathcal{X}$, for each $i = 0, \cdots, N$, we say define a (r,s)-stencil as $\mathcal{S}_{i+\frac{1}{2}} = \{i - r + 1, \cdots, i-1, i, i+1, \cdots, i+s\}$.*

Definition 2.7 (Grid function restricted to a stencil). *For a Δx-grid, a (r,s)-stencil $\mathcal{S}_{i+\frac{1}{2}}$ and Q a Δx-grid function, we define $Q(\mathcal{S}_{i+\frac{1}{2}}) = (Q_k)_{k \in \mathcal{S}_{i+\frac{1}{2}}}$.*

Remark 2.3. *When computing $Q(\mathcal{S}_{i+\frac{1}{2}})$, we may need values of Q_i that are out of the range $1, \cdots, N$. In this case, we may think of Q_i as a ghost cell value. Since we are under the assumption of periodic boundary conditions, this problem is overcome by assuming periodicity on the grid function Q. For instance, if for a $(3,3)-$stencil, we assume $Q_0 = Q_N$, $Q_{-1} = Q_{N-1}$, $Q_{-2} = Q_{N-2}$ and $Q_{N+1} = Q_1$, $Q_{N+2} = Q_2$, $Q_{N+2} = Q_3$. The same applies for Δx-C-grid winds.*

Remark 2.4. *For Problem 2.2, we define the Δx-grid functions q^n and $Q(t^n)$, where $q_i^n = q(x_i, t^n)$, $Q(t^n)_i = Q_i(t^n)$, for $n = 0, \cdots, N_T$. We also define the Δx-C-grid wind u^n where $u^n_{i+\frac{1}{2}} = u(x_{i+\frac{1}{2}}, t^n)$.*

Finally, we define the one-dimensional (1D) finite-volume (FV) scheme problem as follows in Problem 2.3.

Problem 2.3 (1D-FV scheme)**.** *Assume the framework defined in Problem 2.2. The finite-volume approach of Problem 2.2 consists of finding a scheme of the form:*

$$Q_i^{n+1} = Q_i^n - \lambda \delta_i F_i^n, \quad \forall i = 1, \cdots, N, \quad \forall n = 0, \cdots, N_T - 1, \tag{2.21}$$

where $\delta_i F_i^n = F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}$ and the $Q^n \in \mathbb{R}^{\Delta x}$ is intended to be an approximation of the other $Q(t^n) \in \mathbb{R}^{\Delta x}$ in some sense. We define $Q_i^0 = Q_i(0)$ or $Q_i^0 = q_i^0$. The terms $F^n_{i+\frac{1}{2}} = \mathcal{F}(Q^n(\mathcal{S}_{i+\frac{1}{2}}), \tilde{u}^n_{i+\frac{1}{2}})$ are known as numerical flux, where $\mathcal{S}_{i+\frac{1}{2}}$ is a (p,q)-stencil, $\mathcal{F} : \mathbb{R}^{p+q} \times \mathbb{R} \to \mathbb{R}$ is the numerical flux function, and it approximates $\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, t)\, dt$, $\forall i = 0, 1, \cdots, N$, or, in other words, it estimates the time-averaged fluxes at the control volume X_i boundaries. The value $\tilde{u}^n_{i+\frac{1}{2}}$ is related to the time-averaged velocity and depends on values of $u^n_{i+\frac{1}{2}}$.

Remark 2.5. *A scheme of the form from Equation (2.21) is referred to as a 1D-FV scheme and it is also known as a conservative scheme.*

For a 1D-FV the discrete total mass at the time-step n is given by

$$M^n = \Delta x \sum_{i=1}^{N} Q_i^n. \tag{2.22}$$

Therefore, the discrete total mass is constant for a 1D-FV scheme, which follows from a straightforward computation:

$$M^{n+1} = \Delta x \sum_{i=1}^{N} Q_i^{n+1} = M^n - \Delta t \sum_{i=1}^{N} (F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}) = M^n - \Delta t (F^n_{N+\frac{1}{2}} - F^n_{\frac{1}{2}}) = M^n,$$

where we are using that $F^n_{N+\frac{1}{2}} = F^n_{\frac{1}{2}}$, since we are assuming periodic boundary conditions. The numerical flux function for the PPM will be introduced in Section 2.3.3.

2.2.2 Consistency and convergence

Before moving to the definition of convergence, we point out an important relation between the average values of q and its value at the cell centroids. We mentioned in Problem 2.3 that the initial condition may be considered as q_i^0 instead of $Q_i(0)$. Furthermore, when analyzing the convergence of a 1D-FV scheme, we may want to compare Q_i^n with

q_i^n since $Q_i(t^n)$ requires the computation of an analytical integral, which may be too complicated to obtain in some cases. In the following Proposition 2.1, we give a simple proof of that q_i^n approximates $Q_i(t^n)$ with second order error when q is twice continuously differentiable.

Proposition 2.1. *If $q \in \mathcal{C}^2(\mathbb{S}_T^1)$, then $Q_i(t^n) - q_i^n = C_1 \Delta x^2$, where $C_1 = \frac{1}{24}\frac{\partial^2 q}{\partial x^2}(\eta, t^n)$, $\eta \in X_i$*

Proof. Just apply Theorem A.4 for the function $q(x, t^n)$. □

To move towards the convergence of 1D-FV schemes, for Problem 2.3 we introduce the local truncation error (LTE hereafter) τ_i^n following LeVeque (2002):

$$Q_i(t^{n+1}) = Q_i(t^n) - \lambda\left(\mathcal{F}(Q(t^n)(\mathcal{S}_{i+\frac{1}{2}}), \tilde{u}_{i+\frac{1}{2}}^n) - \mathcal{F}(Q(t^n)(\mathcal{S}_{i-\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n)\right) + \Delta t \tau_i^n. \tag{2.23}$$

Notice the LTE is obtained by replacing the exact solution in Equation (2.21). Since $Q_i(t^n)$ is the exact solution of Equation (2.20), the LTE may be rewritten as

$$\begin{aligned}\tau_i^n = \frac{1}{\Delta x}\Bigg[&\left(\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, t)\, dt - \mathcal{F}(Q(t^n)(\mathcal{S}_{i+\frac{1}{2}}), \tilde{u}_{i+\frac{1}{2}}^n)\right) + \\ &\left(\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, t)\, dt - \mathcal{F}(Q(t^n)(\mathcal{S}_{i-\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n)\right)\Bigg].\end{aligned} \tag{2.24}$$

The LTE gives a measure of how well the 1D-FV scheme approximates the integral form of the considered conservation law. Another interpretation of the LTE is that the LTE gives the error obtained after applying the scheme for a single time-step using the exact solution. The 1D-FV scheme is said to be consistent if the LTE converges to zero.

Given a $\Delta x-$grid and $Q \in \mathbb{R}^{\Delta x}$, we define the p-norm by

$$\|Q\|_{p,\Delta x} = \begin{cases} \left(\sum_{i=1}^N |Q_i|^p\right)^{\frac{1}{p}} & \text{if } 1 \le p < \infty, \\ \max_{i=1,\cdots,N} |Q_i| & \text{otherwise .} \end{cases} \tag{2.25}$$

We the define $\tau^n = (\tau_1^n, \cdots, \tau_N^n)$, which represent the LTEs at the time-step n. Now we can define consistency.

Definition 2.8 (Consistency)**.** *Let us consider the framework of Problem 2.3. A 1D-FV scheme is said to be consistency in the p-norm if for any sequence of $(\Delta x^{(k)}, \Delta t^{(k)}, \lambda)$-discretizations, $k \in \mathbb{N}$, with $\lim_{k\to\infty} \Delta x^{(k)} = \lim_{k\to\infty} \Delta t^{(k)} = 0$, we have:*

$$\lim_{k\to\infty}\left[\max_{1\le n\le N_T^{(k)}} \|\tau^n\|_{p,\Delta x^{(k)}}\right] = 0,$$

and it is said to be consistent with order P in the $p-$norm if

$$\max_{1\le n\le N_T^{(k)}} \|\tau^n\|_{p,\Delta x^{(k)}} = O(\Delta x^P).$$

From Equation (2.24), it follows that we basically need to ensure that the numerical flux function $\mathcal{F}$ converges to the time-averaged flux at edges when $\Delta x \to 0$ in order to guarantee consistency. In Section 2.3.3 we shall address how the numerical flux from PPM approximates the time-averaged flux at edges.

At last, we define the point-wise error at time-step n by:

$$E_i^n = Q_i(t^n) - Q_i^n, \quad i = 1, \cdots, N,$$

and we define the vector of errors by $E^n = (E_1^n, \cdots, E_N^n)$.

Definition 2.9 (Convergence). *Let us consider the framework of Problem 2.3. A 1D-FV scheme is said to be convergent in the p-norm if for any sequence of $(\Delta x^{(k)}, \Delta t^{(k)}, \lambda)$-discretizations, $k \in \mathbb{N}$, with $\lim_{k\to\infty} \Delta x^{(k)} = \lim_{k\to\infty} \Delta t^{(k)} = 0$, we have:*

$$\lim_{k\to\infty} \left[\max_{1\le n\le N_T^{(k)}} \|E^n\|_{p,\Delta x^{(k)}} \right] = 0,$$

and it is said to converge with order P in the $p-$norm if

$$\max_{1\le n\le N_T^{(k)}} \|E^n\|_{p,\Delta x^{(k)}} = O(\Delta x^P).$$

Subtracting Equation (2.21) from Equation (2.23) we get the following equation for the error:

$$\begin{aligned} E_i^{n+1} = E_i^n - \lambda \Bigg[& \left(\mathcal{F}(Q(t^n)(S_{i+\frac{1}{2}}), \tilde{u}_{i+\frac{1}{2}}^n) - \mathcal{F}(Q^n(S_{i+\frac{1}{2}}), \tilde{u}_{i+\frac{1}{2}}^n) \right) \\ & - \left(\mathcal{F}(Q(t^n)(S_{i-\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n) - \mathcal{F}(Q^n(S_{i-\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n) \right) \Bigg] + \tau_i^n \Delta t. \end{aligned} \tag{2.26}$$

Notice that if $q, u \in \mathcal{C}^3$, we can rewrite Equation 2.24 as:

$$\tau_i^n = \left[\frac{1}{\Delta x \Delta t} \int_{t^n}^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial (uq)}{\partial x}(x,t)\, dx\, dt - \left(\frac{\mathcal{F}(Q(t^n)(S_{i+\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n)) - \mathcal{F}(Q(t^n)(S_{i+\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n)}{\Delta x} \right) \right].$$

Using the midpoint rule for integration (Theorem A.4) and the mean value theorem for integrals (Theorem A.2), we have:

$$\begin{aligned} \tau_i^n &= \left[\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \left(\frac{\partial (uq)}{\partial x}(x_i,t) + \frac{\Delta x^2}{24} \frac{\partial (uq)}{\partial x}(\xi,t) \right) dt - \left(\frac{\mathcal{F}(Q(t^n)(S_{i+\frac{1}{2}}), \tilde{u}_{i+\frac{1}{2}}^n)) - \mathcal{F}(Q(t^n)(S_{i-\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n)}{\Delta x} \right) \right] \\ &= \left[\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \frac{\partial (uq)}{\partial x}(x_i,t)\, dt - \left(\frac{\mathcal{F}(Q(t^n)(S_{i+\frac{1}{2}}), \tilde{u}_{i+\frac{1}{2}}^n)) - \mathcal{F}(Q(t^n)(S_{i-\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n)}{\Delta x} \right) \right] + \frac{\Delta x^2}{24} \frac{\partial^3 (uq)}{\partial x^3}(\xi, \bar{t}), \end{aligned} \tag{2.27}$$

for $\xi \in X_i$ and $\bar{t} \in [t^n, t^{n+1}]$. Therefore, if $q, u \in \mathcal{C}^3$ the scheme is consistent, if and only if, $\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \frac{\partial (uq)}{\partial x}(x_i,t)\, dt$ is approximated by $\frac{\mathcal{F}(Q(t^n)(S_{i+\frac{1}{2}}), \tilde{u}_{i+\frac{1}{2}}^n)) - \mathcal{F}(Q(t^n)(S_{i-\frac{1}{2}}), \tilde{u}_{i-\frac{1}{2}}^n)}{\Delta x}$. This shall be very useful when we consider two-dimensional schemes, where we are going to use the discrete

operators to estimate the divergence of velocity fields.

2.2.3 Stability

In order to define the concept of stability, it is useful to introduce an operator representation of 1D-FV schemes. In the context of Problem 2.3, we define the operators $\mathcal{H}_{\Delta x,n} : \mathbb{R}^{\Delta x} \to \mathbb{R}^{\Delta x}$ whose i-th entry is given by:

$$[\mathcal{H}_{\Delta x,n}(Q)]_i = Q_i - \lambda\left(\mathcal{F}(Q(\mathcal{S}_{i+\frac{1}{2}}), \tilde{u}^n_{i+\frac{1}{2}}) - \mathcal{F}(Q(\mathcal{S}_{i-\frac{1}{2}}), \tilde{u}^n_{i-\frac{1}{2}})\right), \tag{2.28}$$

for $i = 1, \cdots, N$, $n = 0, \cdots, N_T - 1$. Notice that the dependence on n is due to the velocity that may be allowed to vary with time. As it is usual, we are assuming periodicity in the entries of Q when we apply the operator $\mathcal{H}_{\Delta x,n}$. Thus, Equation (2.21) may be rewritten in a vector form by

$$Q^{n+1} = \mathcal{H}_{\Delta x,n}(Q^n),$$

and Equation (2.23) in a vector form reads

$$Q(t^{n+1}) = \mathcal{H}_{\Delta x,n}(Q(t^n)) + \Delta t \tau^n,$$

and the error equation (2.26) is given by

$$E^{n+1} = \mathcal{H}_{\Delta x,n}(Q(t^n)) - \mathcal{H}_{\Delta x,n}(Q^n) + \Delta t \tau^n. \tag{2.29}$$

The stability theory focus on uniformly bounding the norm of $\mathcal{H}_{\Delta x,n}(Q(t^n)) - \mathcal{H}_{\Delta x,n}(Q^n)$ (LeVeque, 2002). We define stability as follows.

Definition 2.10 (Stability)**.** *In the context of Problem 2.3, a 1D-FV scheme is stable in the $p-$norm if for any $(\Delta x, \Delta t, \lambda)$-discretization of $[a, b] \times [0, T]$ we have:*

$$\|\mathcal{H}_{\Delta x,n}(Q) - \mathcal{H}_{\Delta x,n}(P)\|_{p,\Delta x} \leq (1 + \alpha\Delta t)\|Q - P\|_{p,\Delta x}, \tag{2.30}$$

for all $Q, P \in \mathbb{R}^{\Delta x}$ and α is a constant that does not depend neither on Δx nor on Δt.

Assuming that the scheme is stable in the $p-$norm, then it follows from Equation (2.29) that:

$$\begin{aligned}
\|E^{n+1}\|_{p,\Delta x} &\leq \|\mathcal{H}_{\Delta x,n}(Q(t^n)) - \mathcal{H}_{\Delta x,n}(Q^n)\|_{p,\Delta x} + \Delta t \max_{n=1,\cdots,N_T} \|\tau^n\|_{p,\Delta x} \\
&\leq (1 + \alpha\Delta t)\|E^n\|_{p,\Delta x} + \Delta t \max_{n=1,\cdots,N_T} \|\tau^n\|_{p,\Delta x} \\
&\leq (1 + \alpha\Delta t)^n\|E^0\|_{p,\Delta x} + \Delta t \max_{n=1,\cdots,N_T} \|\tau^n\|_{p,\Delta x} \sum_{k=0}^{n-1}(1 + \alpha\Delta t)^k \\
&\leq e^{\alpha T}(\|E^0\|_{p,\Delta x} + T \max_{n=1,\cdots,N_T} \|\tau^n\|_{p,\Delta x}),
\end{aligned} \tag{2.31}$$

where we used $n\Delta t \leq T$, $T = N\Delta t$ and the inequality $e^t > 1 + t$. When computing the initial average values using the value at the cell centroid, the initial error E^0 converges to zero provided q is twice continuously differentiable by Proposition 2.1. Therefore, it follows that if the scheme is stable and consistent then it is convergent. Furthermore, if

it is stable and consistent with order P, then the convergence order is at least equal to $\min\{P, 2\}$. In the case where both the conservation law and $\mathcal{H}_{\Delta x,n}$ are linear, this result is a particular case of the Lax-Ritchmyer stability and the convergence is guaranteed by the Lax equivalence theorem (LeVeque, 2002). In this Chapter, we are interested only in the linear advection equation. However, as we shall see in Section 2.3.3, the operator $\mathcal{H}_{\Delta x,n}$ may become non-linear when monotonicity constraints are activated.

Notice that, if $\mathcal{H}_{\Delta x,n}$ is linear, then stability is equivalent to require that

$$\|\mathcal{H}_{\Delta x,n}\|_{p,\Delta x} \leq 1 + \alpha \Delta t,$$

where

$$\|\mathcal{H}_{\Delta x,n}\|_{p,\Delta x} = \sup_{Q \in \mathbb{R}^{\Delta x}} \frac{\|\mathcal{H}_{\Delta x,n}(Q)\|_{p,\Delta x}}{\|Q\|_{p,\Delta x}},$$

is the operator p-norm.

For linear operators, we may use the discrete Fourier transform (Trefethen, 2000) to estimate the 2-norm of $\mathcal{H}_{\Delta x,n}$. This approach is known as Von Neumann stability analysis. We define the nodes $\theta_i = i\frac{2\pi}{N}$, $i = 1, \cdots, N$, $\Delta\theta = \frac{2\pi}{N}$, $\theta = (\theta_1, \theta_2, \cdots, \theta_N)$. The imaginary unit is denoted by $\imath$. The Fourier modes are given by:

$$e^{\imath k\theta} = (e^{\imath k\theta_1}, e^{\imath k\theta_2}, \cdots, e^{\imath k\theta_N}) \in \mathbb{C}^N,$$

for $k = 1, \cdots, N$. Each k is referred to wavenumber and θ_k is called dimensionless wavenumber. The Fourier modes form an orthogonal basis of $\mathbb{C}^N$ with respect to the inner product

$$\langle Q, P \rangle = \frac{1}{N} \sum_{i=1}^{N} Q_i \overline{P_i},$$

for $P, Q \in \mathbb{C}$ and $\overline{z}$ denotes the complex conjugate of z. Given $Q \in \mathbb{R}^{\Delta x}$, we may may express it in terms of the Fourier modes

$$Q = \sum_{k=1}^{N} a_k \exp(\imath k\theta),$$

where $a_k \in \mathbb{C}$. The 2-norm of Q is then given by:

$$\|Q\|_{2,\Delta x} = \sqrt{N \sum_{k=1}^{N} |a_k|^2}.$$

The idea of Von Neumann stability analysis is to apply the operator $\mathcal{H}_{\Delta x,n}$ on each Fourier mode and analyze how it modifies its amplitude. For ease of analysis, we assume that the velocity is constant, which implies that the operator $\mathcal{H}_{\Delta x,n}$ has constant coefficients and does not depend on n. For the general case, where the velocity is not constant, the stability can be ensured using the frozen coefficients method (Strikwerda, 2004, p. 59). This method boils down to performing multiple times the stability analysis with a constant velocity being equal to each one of the possible values of the velocity on the grid. If the scheme is stable for all the possible constant velocities, then stability is ensured. Since the operator is

supposed to be linear with constant coefficients and we are assuming periodic boundaries conditions, we may write:

$$\mathcal{H}_{\Delta x,n}(e^{\imath k\theta}) = \rho(k)e^{\imath k\theta},$$

where the term $\rho(k)$ is called amplification factor and it is an eigenvalue of $\mathcal{H}_{\Delta x,n}$. The norm of $\mathcal{H}_{\Delta x,n}(Q)$ is bounded by:

$$\|\mathcal{H}_{\Delta x,n}(Q)\|^2_{2,\Delta x} = N\sum_{k=1}^{N}|a_k|^2|\rho(k)|^2 \leq \max_{k=1,\cdots,N}|\rho(k)|^2\|Q\|^2_{2,\Delta x}.$$

Therefore:

$$\|\mathcal{H}_{\Delta x,n}\|_{2,\Delta x} \leq \max_{k=1,\cdots,N}|\rho(k)|.$$

If we show that $\max_{k=1,\cdots,N}|\rho(k)| \leq 1+\alpha\Delta t$, with α independent of Δt, N and n, then we ensure the stability of $\mathcal{H}_{\Delta x,n}$. Generally speaking, the numerical flux can be written as a linear function

$$\mathcal{F}(Q(\mathcal{S}_{i+\frac{1}{2}}), \tilde{u}^n_{i+\frac{1}{2}}) = \sum_{l\in\mathcal{S}_{i+\frac{1}{2}}}\alpha_{l,i}Q_l,$$

when no monotonicity constraint is imposed, where the coefficients $\alpha_{l,i}$ depend on $\tilde{u}^n$, Δt and Δx. We can then express ρ in terms of $\alpha_{l,i}$. Indeed, when we apply the operator $\mathcal{H}_{\Delta x,n}$ in a Fourier mode, we get:

$$\begin{aligned}[\mathcal{H}_{\Delta x,n}(e^{\imath k\theta})]_i &= e^{\imath k\theta_i} - \lambda\left(\sum_{l\in\mathcal{S}_{i+\frac{1}{2}}}\alpha_{l,i}e^{\imath k\theta_l} - \sum_{l\in\mathcal{S}_{i-\frac{1}{2}}}\alpha_{l,i-1}e^{\imath k\theta_{l-1}}\right)\\ &= e^{\imath k\theta_i}\left(1-\lambda\left(\sum_{l\in\mathcal{S}_{i+\frac{1}{2}}}\alpha_{l,i}e^{\imath k\theta_{l-i}} - \sum_{l\in\mathcal{S}_{i-\frac{1}{2}}}\alpha_{l,i-1}e^{\imath k\theta_{l-1-i}}\right)\right).\end{aligned}$$

Hence, the amplification factor has the form

$$\rho(k) = 1-\lambda\left(\sum_{l\in\mathcal{S}_{i+\frac{1}{2}}}\alpha_{l,i}e^{\imath k\theta_{l-i}} - \sum_{l\in\mathcal{S}_{i-\frac{1}{2}}}\alpha_{l,i-1}e^{\imath k\theta_{l-1-i}}\right). \tag{2.32}$$

In Section 2.3.3 we shall analyse $|\rho(k)|$ in terms of the PPM coefficients.

2.3 The Piecewise-Parabolic Method

In this Section, we are going to review and analyze the Piecewise-Parabolic method (PPM). This method was proposed by Colella and Woodward (1984) for gas dynamic simulations and its viability for atmospheric simulations has been shown by Carpenter et al. (1990). This method is based on using parabolas to reconstruct the function from its average values, ensuring mass conservation and monotonicity. PPM is an extension of the Piecesiwe-Linear method from Van Leer (1977) and it is employed in the FV3 model using the dimension splitting method from Lin and Rood (1996). This section is organized as follows: in Subsection 2.3.1 we present and analyze the PPM reconstruction method and the monotonization and flux computation are presented and analyzed in Subsections 2.3.2

and 2.3.3, respectively.

2.3.1 Reconstruction

Let us consider a function $q \in L^1_{\text{loc}}(\mathbb{R})$, a Δx-grid $\mathcal{X}$ of $[a,b]$ and assume that we are given the average values $Q_i = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)\,dx$ on each control volume X_i, $\forall i = 1,\cdots,N$. In this context, it is useful to define the Δx-grid function $Q = (Q_1,\cdots,Q_N) \in \mathbb{R}^{\Delta x}$. We make use of the indicator function of each control volume X_i defined by:

$$\chi_i(x) = \begin{cases} 1 & \text{if } x \in X_i \\ 0 & \text{otherwise} \end{cases}$$

Using a notation similar to Stoer and Bulirsch (2002, Chapter 1), we assume that we have a family of functions $\Phi(\xi;\mu)$ defined for $\xi \in [0,1]$ depending on a parameter $\mu = (\mu_0,\mu_1,\cdots,\mu_d) \in \mathbb{R}^{d+1}$. The reconstruction problem consists of finding a piecewise function:

$$q_{Pd}(x;Q) = \sum_{i=1}^{N} \chi_i(x) q_i(x;Q), \tag{2.33}$$

where $q_i(x;Q) = \Phi\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta x};\alpha_i\right)$, $\alpha_i = (\alpha_{i0},\alpha_{i1},\cdots\alpha_{id}) \in \mathbb{R}^{d+1}$ is such that:

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q_{Pd}(x;Q)\,dx = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q_i(x;Q)\,dx = \int_0^1 \Phi(\xi;\alpha_i)\,d\xi = Q_i,$$

that is, $q_i(x;Q)$ preserves the mass on each control volume X_i.

Notice that since $q_i(x;Q) = \Phi\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta x};\alpha_i\right)$, it is reasonable that $\Phi(0;\alpha_i)$ should approximate $q_i(x_{i-\frac{1}{2}})$ and $\Phi(1;\alpha_i)$ should approximate $q(x_{i+\frac{1}{2}})$. Furthermore, if q and Φ are differentiable enough, then $\Phi^{(l)}(0;\alpha_i)$ should approximate $(\Delta x)^l q^{(l)}(x_{i-\frac{1}{2}})$ and $\Phi^{(l)}(1;\alpha_i)$ should approximate $(\Delta x)^l q^{(l)}(x_{i+\frac{1}{2}})$, whenever these derivatives exist. One way to estimate these values at the edges $x_{i+\frac{1}{2}}$ using the average values Q is to use the reconstruction method based on primitive functions (LeVeque, 2002, Chapter 17). Observe that if we define

$$Q(x) = \int_a^x q(\xi)\,d\xi, \tag{2.34}$$

we have $Q^{(l)}(x) = q^{(l-1)}(x)$. In particular, $Q^{(l)}(x_{i+\frac{1}{2}}) = q^{(l-1)}(x_{i+\frac{1}{2}})$ and $Q^{(l)}(x_{i+\frac{1}{2}}) = \Delta x\sum_{k=1}^{i} Q_k$, $\forall i = 0,\cdots,N$. Therefore we can use finite-difference schemes to estimate $q^{(l-1)}(x_{i+\frac{1}{2}})$ using the Δx-grid function Q, once this one is assumed to be given. Let us assume that the l–th derivative of Q at $x_{i+\frac{1}{2}}$ is approximated using a stencil $\mathcal{S}^{(l)}_{i+\frac{1}{2}}$ and weights $\beta^{(l)}_{k,i}$, $k \in \mathcal{S}^{l}_{i+\frac{1}{2}}$. If d is odd, then we may search for a parameter $\alpha_i \in \mathbb{R}^{d+1}$ that ensures mass conservation and approximation of q and its derivatives at edges by solving:

$$\begin{cases} \int_0^1 \Phi(\xi;\alpha_i)\,d\xi & = Q_i, \\ \Phi^{(l)}(0;\alpha_i) & = (\Delta x)^l \sum_{k\in\mathcal{S}^{(l)}_{i-\frac{1}{2}}} \beta^{(l)}_{k,i} Q_k, \quad \text{for } l = 0,\cdots,d-1. \end{cases} \tag{2.35}$$

If d is even, similarly we look for a parameter $\alpha_i \in \mathbb{R}^{d+1}$ that solves:

$$\begin{cases} \int_0^1 \Phi(\xi;\alpha_i)\,d\xi & = Q_i, \\ \Phi^{(l)}(0;\alpha_i) & = (\Delta x)^l \sum_{k\in S^{(l)}_{i-\frac{1}{2}}} \beta^{(l)}_{k,i} Q_k, \quad \text{for } l = 0,\cdots,\frac{d}{2}-1, \\ \Phi^{(l)}(1;\alpha_i) & = (\Delta x)^l \sum_{k\in S^{(l)}_{i+\frac{1}{2}}} \beta^{(l)}_{k,i} Q_k, \quad \text{for } l = 0,\cdots,\frac{d}{2}-1. \end{cases} \tag{2.36}$$

The reconstruction problem is linear if $\Phi(\xi;\mu)$ may be expressed as:

$$\Phi(\xi;\mu) = \sum_{k=0}^{d} \mu_k \Phi_k(\xi),$$

for functions Φ_k defined on $[0,1]$. In this case, Equation (2.35) and Equation (2.36) are $(d+1)\times(d+1)$ linear systems. It is usually common to assume that Φ_k's are linearly independent. Thus, we have described a method that allows us to reconstruct a function from its average values preserving its mass in each control value and approximates q at the edges that, in principle, works for functions Φ_k provided they are differentiable enough. For instance, we could choose $d = 0$, and $\Phi_0(\xi) = 1$. In this case, we have piecewise constant functions as used in Godunov (1959). If we choose $d = 1$, $\Phi_0(\xi) = 1$ and $\Phi_1(\xi) = \xi$, we have a piecewise linear reconstruction as in Van Leer (1977). For further polynomial reconstruction schemes, we refer to Engwirda and Kelley (2016) and the references therein.

Hereafter, we are going the focus on the piecewise parabolic method from Colella and Woodward (1984) that uses $d = 2$, $\Phi_0(\xi) = 1$, $\Phi_1(\xi) = \xi$, $\Phi_1(\xi) = (1-\xi)\xi$. In this case, we denote q_{Pd} by q_{PP}. In order to follow the notation from Colella and Woodward (1984), we write $\alpha_{0i} = q_{L,i}$, $\alpha_{1i} = \Delta q_i$ and $\alpha_{2i} = q_{6,i}$. Therefore, each q_i may be expressed as:

$$q_i(x;Q) = q_{L,i} + \Delta q_i z_i(x) + q_{6,i} z_i(x)(1-z_i(x)), \quad \text{where } z_i(x) = \frac{x - x_{i-\frac{1}{2}}}{\Delta x}, \quad x \in X_i, \tag{2.37}$$

where the values $q_{L,i}$, Δq_i and $q_{6,i}$ will be specified latter. Note that each z_i is just a normalization function that maps X_i onto $[0,1]$. It is easy to see that $\lim_{x\to x^+_{i-\frac{1}{2}}} q_i(x;Q) = q_{L,i}$. If we define $q_{R,i} = \lim_{x\to x^-_{i+\frac{1}{2}}} q_i(x;Q)$, then we have:

$$\Delta q_i = q_{R,i} - q_{L,i}. \tag{2.38}$$

The average value of q_i is given by:

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q_i(x;Q)\,dx = \frac{(q_{L,i}+q_{R,i})}{2} + \frac{q_{6,i}}{6}. \tag{2.39}$$

Under the hypothesis of mass conservation, we have:

$$q_{6,i} = 6\left(Q_i - \frac{(q_{L,i}+q_{R,i})}{2}\right). \tag{2.40}$$

Therefore, we have found the parameters Δq_i and $q_{6,i}$ as functions of the parameters $q_{L,i}$

and $q_{R,i}$, such that the parabola q_i from (2.33) guarantees mass conservation. To completely determine the parabola q_i, we need to set the values $q_{L,i}$ and $q_{R,i}$, which, as we have seen, represent the limits of q_i when x tends to the left and right boundaries of X_i, respectively. Hence, it is natural to seek for $q_{L,i}$ as an approximation of $q(x_{i-\frac{1}{2}})$ and $q_{R,i}$ as an approximation of $q(x_{i+\frac{1}{2}})$. As we mentioned before in after introducing Equation (2.34), this is achieved using finite-differences. An explicit expression for the approximation of $q(x_{i-\frac{1}{2}})$, denoted by $q_{i+\frac{1}{2}}$, is given by (Colella & Woodward, 1984):

$$q_{i+\frac{1}{2}} = \frac{1}{2}\Big(Q_{i+1} + Q_i\Big) - \frac{1}{6}\Big(\delta Q_{i+1} - \delta Q_i\Big), \tag{2.41}$$

where δQ_i is the average slope in the i-th control-volume:

$$\delta Q_i = \frac{1}{2}\Big(Q_{i+1} - Q_{i-1}\Big). \tag{2.42}$$

We notice that Formula (2.42) may be rewritten more explicitly as:

$$q_{i+\frac{1}{2}} = \frac{7}{12}\Big(Q_{i+1} + Q_i\Big) - \frac{1}{12}\Big(Q_{i+2} + Q_{i-1}\Big). \tag{2.43}$$

The Formula (2.43) is fourth-order accurate if q is at least C^4 (Colella & Woodward, 1984). Indeed, we prove this later in Proposition 2.2. An explicit expression for the values of $q_{R,i}$ and $q_{L,i}$ are given by:

$$q_{R,i} = q_{i+\frac{1}{2}} = \frac{7}{12}\Big(Q_{i+1} + Q_i\Big) - \frac{1}{12}\Big(Q_{i+2} + Q_{i-1}\Big), \tag{2.44}$$

$$q_{L,i} = q_{i-\frac{1}{2}} = \frac{7}{12}\Big(Q_i + Q_{i-1}\Big) - \frac{1}{12}\Big(Q_{i+1} + Q_{i-2}\Big). \tag{2.45}$$

We point out that a fifth-order accurate for the values of $q_{R,i}$ and $q_{L,i}$ is also possible, as it was developed by Putman and Lin (2007) based on the work Suresh and Huynh (1997). The fifth-order reconstruction formula reads:

$$q_{R,i} = \frac{1}{60}\Big(2Q_{i-2} - 13Q_{i-1} + 47Q_i + 27Q_{i+1} - 3Q_{i+2}\Big), \tag{2.46}$$

$$q_{L,i} = \frac{1}{60}\Big(-3Q_{i-2} + 27Q_{i-1} + 47Q_i - 13Q_{i+1} + 2Q_{i+2}\Big). \tag{2.47}$$

However, we notice that this reconstruction scheme allows discontinuity of the Piecewise-Parabolic function at the control volume edges since the stencil it is not symmetric.

PPM reconstruction order analysis

As we pointed out before, the approximation of q at the control volumes edges given by Equation (2.43) is fourth-order accurate when $q \in C^4(\mathbb{R})$. This is proved as a Corollary of the following Proposition 2.2.

Proposition 2.2. *Let $q \in C^4(\mathbb{R})$, $\overline{x} \in \mathbb{R}$ and $h > 0$. Then, the following identity holds:*

$$q(\overline{x}) = \frac{7}{12}\left(\frac{1}{h}\int_{\overline{x}}^{\overline{x}+h} q(x)\,dx + \frac{1}{h}\int_{\overline{x}-h}^{\overline{x}} q(x)\,dx\right) - \frac{1}{12}\left(\frac{1}{h}\int_{\overline{x}+h}^{\overline{x}+2h} q(x)\,dx + \frac{1}{h}\int_{\overline{x}-2h}^{\overline{x}-h} q(x)\,dx\right) + C_1 h^4, \tag{2.48}$$

where C_1 is a constant that depends on q and h.

Proof. We define $Q(x) = \int_a^x q(\xi)\,d\xi$ for fixed $a \in \mathbb{R}$ as in Equation (2.34). It follows that:

$$\int_{\overline{x}}^{\overline{x}+h} q(\xi)\,d\xi + \int_{\overline{x}-h}^{\overline{x}} q(\xi)\,d\xi = Q(\overline{x}+h) - Q(\overline{x}-h),$$
$$\int_{\overline{x}+h}^{\overline{x}+2h} q(\xi)\,d\xi + \int_{\overline{x}-2h}^{\overline{x}-h} q(\xi)\,d\xi = Q(\overline{x}+2h) - Q(\overline{x}-2h) - (Q(\overline{x}+h) - Q(\overline{x}-h)).$$

Using these identities, Equation (2.48) may be rewritten as:

$$q(\overline{x}) = \frac{4}{3}\left(\frac{Q(\overline{x}+h) - Q(\overline{x}-h)}{2h}\right) - \frac{1}{3}\left(\frac{Q(\overline{x}+2h) - Q(\overline{x}-2h)}{4h}\right) + C_1 h^4, \tag{2.49}$$

which consists of finite-difference approximations. Thus, Equation (2.48) follows from Lemma A.1 with:

$$C_1 = C_1(\mu_1, \mu_2) = \frac{1}{720}\left(6q^{(4)}(\mu 1) - 32q^{(4)}(\mu_2)\right), \tag{2.50}$$

where $\mu_1, \mu_2 \in [\overline{x} - 2h, \overline{x} + 2h]$, which concludes the proof. □

Corollary 2.1. *It follows from Proposition 2.2 with $\overline{x} = x_{i+\frac{1}{2}}$ and $h = \Delta x$ that $q_{i+\frac{1}{2}}$ given by Equation (2.43) satisfies:*

$$q(x_{i+\frac{1}{2}}) - q_{i+\frac{1}{2}} = C_1 \Delta x^4, \tag{2.51}$$

with C_1 given by Equation (2.50), whenever $q \in C^4(\mathbb{R})$.

Remark 2.6. *Similarly, one can show that the formulas are given by Equation (2.46) and Equation (2.46) are fifth-order accurate.*

The parabolic function from (2.37) given with coefficients specified before approximates q with order 3 when $q \in C^4(\mathbb{R})$. In order to check this, for $x \in X_i$ we rewrite Equation (2.37) as:

$$q_i(x; Q) = q_{L,i} + \frac{(\Delta q_i + q_{6,i})}{\Delta x}(x - x_{i-\frac{1}{2}}) - \frac{q_{6,i}}{\Delta x^2}(x - x_{i-\frac{1}{2}})^2 \tag{2.52}$$

and we write q using its Taylor expansion assuming $q \in C^4(\mathbb{R})$:

$$q(x) = q(x_{i-\frac{1}{2}}) + q'(x_{i-\frac{1}{2}})(x - x_{i-\frac{1}{2}}) + \frac{q''(x_{i-\frac{1}{2}})}{2}(x - x_{i-\frac{1}{2}})^2 + \frac{q^{(3)}(\theta_i)}{6}(x - x_{i-\frac{1}{2}})^3, \tag{2.53}$$

where $\theta_i \in X_i$. Comparing Equation (2.52) with Equation (2.53), it is reasonable to seek to

some bound to the expressions:

$$q'(x_{i-\frac{1}{2}}) - \frac{(\Delta q_i + q_{6,i})}{\Delta x}, \tag{2.54}$$

and:

$$\frac{q''(x_{i-\frac{1}{2}})}{2} - \left(- \frac{q_{6,i}}{\Delta x^2} \right). \tag{2.55}$$

We have seen that term $q_{L,i}$ gives a fourth-order approximation to $q(x_{i-\frac{1}{2}})$. The Corollary 2.2 shall prove that the term (2.54) has a bound proportional to Δx^2, and the Corollary 2.3 shall prove that the term (2.55) is bounded by a constant times Δx.

Before proving the desired bounds, it is useful to rewrite some terms explicitly as functions of the values of the Δx-grid function Q. Combining Equation (2.40) with Equations (2.44) and (2.45), we may write $q_{6,i}$ as:

$$q_{6,i} = \frac{1}{4}\left(Q_{i-2} - 6Q_{i-1} + 10Q_i - 6Q_{i+1} + Q_{i+2} \right). \tag{2.56}$$

Recalling the definition of Δq_i from Equation (2.38), and applying Equations (2.44) and (2.45), we may express Δq_i as:

$$\Delta q_i = \frac{1}{12}\left(Q_{i-2} - 8Q_{i-1} + 8Q_{i+1} - Q_{i+2} \right). \tag{2.57}$$

Finally, we combine Equations (2.56) and (2.57) and write their sum as:

$$\frac{(\Delta q_i + q_{6,i})}{\Delta x} = \frac{2Q_{i-2} - 13Q_{i-1} + 15Q_i - 5Q_{i+1} + Q_{i+2}}{6\Delta x}. \tag{2.58}$$

The next Proposition 2.3 proves that Equation (2.58) approximates $q'(x_{i-\frac{1}{2}})$ with order 2.

Proposition 2.3. *Let $q \in C^3(\mathbb{R})$, $\overline{x} \in \mathbb{R}$ and $h > 0$. Then, the following identity holds:*

$$\begin{aligned} q'(\overline{x}) = \frac{1}{6h}\Bigg(& \frac{2}{h}\int_{\overline{x}-2h}^{\overline{x}-h} q(x)\,dx - \frac{13}{h}\int_{\overline{x}-h}^{\overline{x}} q(x)\,dx + \frac{15}{h}\int_{\overline{x}}^{\overline{x}+h} q(x)\,dx \\ & - \frac{5}{h}\int_{\overline{x}+h}^{\overline{x}+2h} q(x)\,dx + \frac{1}{h}\int_{\overline{x}+2h}^{\overline{x}+3h} q(x)\,dx \Bigg) + C_2 h^2, \end{aligned} \tag{2.59}$$

where C_2 is a constant that depends on q and h.

Proof. We consider again $Q(x) = \int_a^x q(\xi)\,d\xi$ for $a \in \mathbb{R}$ fixed as in Equation (2.34). Like in

Proposition 2.3, we have:

$$\begin{aligned}\frac{1}{6h}\Bigg(\frac{2}{h}\int_{\overline{x}-2h}^{\overline{x}-h} q(x)\,dx - \frac{13}{h}\int_{\overline{x}-h}^{\overline{x}} q(x)\,dx + \frac{15}{h}\int_{\overline{x}}^{\overline{x}+h} q(x)\,dx - \frac{5}{h}\int_{\overline{x}+h}^{\overline{x}+2h} q(x)\,dx + \frac{1}{h}\int_{\overline{x}+2h}^{\overline{x}+3h} q(x)\,dx\Bigg)\\ = \frac{1}{6h}\Bigg(\frac{2}{h}\big(Q(\overline{x}-h) - Q(\overline{x}-2h)\big) - \frac{13}{h}\big(Q(\overline{x}) - Q(\overline{x}-h)\big) + \frac{15}{h}\big(Q(\overline{x}+h) - Q(\overline{x})\big)\\ -\frac{5}{h}\big(Q(\overline{x}+2h) - Q(\overline{x}+h)\big) + \frac{1}{h}\big(Q(\overline{x}+3h) - Q(\overline{x}+2h)\big)\Bigg)\\ = \frac{1}{6h^2}\Bigg(-2Q(\overline{x}-2h) + 15Q(\overline{x}-h) - 28Q(\overline{x}) + 20Q(\overline{x}+h) - 6Q(\overline{x}+2h) + Q(\overline{x}+3h)\Bigg),\end{aligned}$$

which consists of the finite-difference scheme from Lemma A.2. Therefore, Equation (2.59) follows from Lemma A.2 with:

$$C_2 = C_2(\mu_1, \mu_2) = \frac{1}{24}\Bigg(128q^{(3)}(\mu_1) - 116q^{(3)}(\mu_2)\Bigg), \tag{2.60}$$

where $\mu_1, \mu_2 \in [x_0 - 2h, x_0 + 3h]$, which concludes the proof. □

Corollary 2.2. *It follows from Proposition 2.3 with* $\overline{x} = x_{i-\frac{1}{2}}$ *and* $h = \Delta x$ *that* Δq_i *given by Equation (2.57) and* $q_{6,i}$ *given by Equation (2.56) satisfy:*

$$q'(x_{i-\frac{1}{2}}) - \frac{(\Delta q_i + q_{6,i})}{\Delta x} = C_2\Delta x^2, \tag{2.61}$$

with C_2 *given by Equation (2.60), whenever* $q \in C^3(\mathbb{R})$.

Now, we analyse the following expression:

$$-\frac{2q_{6,i}}{\Delta x^2} = -\frac{1}{2\Delta x^2}\Bigg(Q_{i-2} - 6Q_{i-1} + 10Q_i - 6Q_{i+1} + Q_{i+2}\Bigg). \tag{2.62}$$

deduced from Equation (2.56) and we prove in Proposition 2.4 that Equation (2.62) approximates $q''(x_{i-\frac{1}{2}})$ with order 1.

Proposition 2.4. *Let* $q \in C^3(\mathbb{R})$, $\overline{x} \in \mathbb{R}$ *and* $h > 0$. *Then, the following identity holds:*

$$\begin{aligned}q''(\overline{x}) = \frac{1}{2h^2}\Bigg(-\frac{1}{h}\int_{\overline{x}-2h}^{\overline{x}-h} q(x)\,dx + \frac{6}{h}\int_{\overline{x}-h}^{\overline{x}} q(x)\,dx - \frac{10}{h}\int_{\overline{x}}^{\overline{x}+h} q(x)\,dx\\ + \frac{6}{h}\int_{\overline{x}+h}^{\overline{x}+2h} q(x)\,dx - \frac{1}{h}\int_{\overline{x}+2h}^{\overline{x}+3h} q(x)\,dx\Bigg) + C_3h,\end{aligned} \tag{2.63}$$

where C_3 *is a constant that depends on* q *and* h.

Proof. Similarly to Proposition 2.3 using the same function Q, we have:

$$\begin{aligned}\frac{1}{2h^2}\Bigg(-\frac{1}{h}\int_{\overline{x}-2h}^{\overline{x}-h} q(x)\,dx + \frac{6}{h}\int_{\overline{x}-h}^{\overline{x}} q(x)\,dx - \frac{10}{h}\int_{\overline{x}}^{\overline{x}+h} q(x)\,dx + \frac{6}{h}\int_{\overline{x}+h}^{\overline{x}+2h} q(x)\,dx - \frac{1}{h}\int_{\overline{x}+2h}^{\overline{x}+3h} q(x)\,dx\Bigg)\\ = \frac{1}{2h^2}\Bigg(-\frac{1}{h}\big(Q(\overline{x}-h) - Q(\overline{x}-2h)\big) + \frac{6}{h}\big(Q(\overline{x}) - Q(\overline{x}-h)\big) - \frac{10}{h}\big(Q(\overline{x}+h) - Q(\overline{x})\big)\\ +\frac{6}{h}\big(Q(\overline{x}+2h) - Q(\overline{x}+h)\big) - \frac{1}{h}\big(Q(\overline{x}+3h) - Q(\overline{x}+2h)\big)\Bigg)\\ = \frac{1}{2h^3}\Big(Q(\overline{x}-2h) - 7Q(\overline{x}-h) + 16Q(\overline{x}) - 16Q(\overline{x}+h) + 7Q(\overline{x}+2h) - Q(\overline{x}+3h)\Big),\end{aligned}$$

which consists of the finite-difference scheme from Lemma A.3. Therefore, Equation (2.63) follows from Lemma A.3 with:

$$C_3 = C_3(\mu_1,\mu_2) = \frac{1}{48}\Big(104q^{(3)}(\mu_1) - 128q^{(3)}(\mu_2)\Big), \tag{2.64}$$

where $\mu_1, \mu_2 \in [x_0 - 2h, x_0 + 3h]$, which concludes the proof. □

Corollary 2.3. *It follows from Proposition 2.4 with $\overline{x} = x_{i-\frac{1}{2}}$ and $h = \Delta x$ that $q_{6,i}$ given by Equation (2.43) satisfies:*

$$q''(x_{i-\frac{1}{2}}) - \Big(-\frac{2q_{6,i}}{\Delta x^2}\Big) = C_3\Delta x, \tag{2.65}$$

with C_3 given by Equation (2.64), whenever $q \in C^3(\mathbb{R})$.

With the aid of Corollaries 2.1, 2.2, and 2.3, we are able to prove that the PPM reconstruction approximates q with order 3. Indeed, we prove this on the follow up Proposition 2.5.

Proposition 2.5. *Let $q \in C^4([a,b])$. Then, the Piecewise-Parabolic function given by Equation (2.37) with the parameters $q_{R,i}$ and $q_{L,i}$ obeying Equations (2.44) and (2.45) gives a third-order approximation to q on the control volume X_i. Namely, there exist constants M_1 and M_2 such that*

$$|q(x) - q_i(x;Q)| \le M_1\Delta x^4 + M_2\Delta x^3, \quad \forall x \in X_i.$$

Proof. For $x \in X_i$, from Equations (2.53) and (2.52), we have:

$$\begin{aligned}q(x) - q_i(x;Q) = (q'(x_{i-\frac{i}{2}}) - q_{L,i}) + \Big(q'(x_{i-\frac{1}{2}}) - \frac{(\Delta q_i + q_{6,i})}{\Delta x}\Big)(x - x_{i-\frac{1}{2}})\\ +\Big(\frac{q''(x_{i-\frac{1}{2}})}{2} + \frac{q_{6,i}}{\Delta x^2}\Big)(x - x_{i-\frac{1}{2}})^2 + \frac{q^{(3)}(\theta_i)}{6}(x - x_{i-\frac{1}{2}})^3.\end{aligned}$$

Using this fact with Corollaries 2.1, 2.2, and 2.3, we have:

$$q(x) - q_i(x;Q) = C_1\Delta x^4 + C_2\Delta x^2(x - x_{i-\frac{1}{2}}) + \frac{C_3}{2}\Delta x(x - x_{i-\frac{1}{2}})^2 + C_4(x - x_{i-\frac{1}{2}})^3,$$

where C_1, C_2 and C_3 are given by Equations (2.50), (2.60) and (2.64), respectively, and

$$C_4 = C_4(\theta_i) = \frac{q^{(3)}(\theta_i)}{6}. \tag{2.66}$$

For $x \in X_i$, we have $|x - x_{i-\frac{i}{2}}| \leq \Delta x$, thus:

$$|q(x) - q_i(x; Q)| \leq M_1 \Delta x^4 + M_2 \Delta x^3,$$

where

$$M_1 = \frac{38}{720} \sup_{\xi \in [a,b]} |q^{(4)}(\xi)|,$$

$$M_2 = \left(\frac{244}{24} + \frac{232}{96} + \frac{1}{6} \right) \sup_{\xi \in [a,b]} |q^{(3)}(\xi)| = \frac{143}{12} \sup_{\xi \in [a,b]} |q^{(3)}(\xi)|,$$

which concludes the proof. □

Remark 2.7. *Replacing the formulas for $q_{R,i}$ and $q_{L,i}$ given by Equations (2.44)and (2.45) by the formulas given by Equations (2.46) and (2.47), does not change the order of convergence of the parabolic approximation.*

2.3.2 Monotonization

This section is dedicated to presenting possible ways of ensuring the creation of new extrema values in the PPM reconstruction. We are going to present the original monotonic scheme from Colella and Woodward (1984) and an alternative scheme from Lin (2004), which was an attempt to reduce the diffusion of the original scheme Colella and Woodward (1984) and is currently employed in the FV3 dynamical core (L. Harris et al., 2021).

Limiter from Colella and Woodward (1984)

To avoid numerical oscillations in the parabolas, especially when discontinuities are present, Colella and Woodward (1984) ensures that the reconstructed value at cell edges (namely, $q_{i+\frac{1}{2}}$) does not stay outside of the range of its neighbors average values (Q_i and Q_{i+1}). This can be achieved by replacing the term δQ_i in Equation (2.41) by the values $\delta_m Q_i$ given by:

$$\delta_m Q_i = \begin{cases} \max(|\delta Q_i|, 2|Q_{i+1} - Q_i|, 2|Q_i - Q_{i-1}|) \cdot \operatorname{sgn}(\delta Q_i) & \text{if } (Q_{i+1} - Q_i)(Q_i - Q_{i-1}) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2.67}$$

where sgn denotes the sign function. To ensure, monotonicity we also must ensure that the parabola has values between $q_{R,i}$ and $q_{L,i}$. This step will introduce a discontinuity on the edges of the PPM approximation. If Q_i is the local maximum/minimum, then we make the parabola constant. This is expressed as:

$$q_{L,i} \leftarrow Q_i, \quad q_{R,i} \leftarrow Q_i, \quad \text{if} \quad (Q_{R,i} - Q_i)(Q_i - Q_{L,i}) \geq 0 \tag{2.68}$$

This step eliminates the introduction of new extremes when we already have an extremum. The other case where we need to modify the values $q_{L,i}$ and $q_{R,i}$ is when the extrema of the parabola falls in $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. It is easy to see from Equation (2.52) that, the extrema of the parabola falling in $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ is equivalent to $|\Delta q_i| \leq |q_{6,i}|$. In this case, the values are updated as follows:

$$\begin{cases} q_{L,i} \leftarrow 3Q_i - 2q_{R,i} & \text{if } \Delta q_i \cdot q_{6,i} > (\Delta q_i)^2, \\ q_{R,i} \leftarrow 3Q_i - 2q_{L,i} & \text{if } -(\Delta q_i)^2 > \Delta q_i \cdot q_{6,i} \end{cases} \tag{2.69}$$

In this step, we are changing the value at the edge where the extreme is closer and ensuring again that no new extreme is created.

Limiter from Lin (2004)

Similarly to Colella and Woodward (1984), Lin (2004) reduces numerical oscilations in the parabolas replacing the term δQ_i in Equation (2.41) by the values $\delta_m Q_i$ given by:

$$\delta_m Q_i = \max(|\delta Q_i|, 2\delta Q_{\min,i}, 2\delta Q_{\max,i}) \cdot \operatorname{sgn}(\delta Q_i), \tag{2.70}$$

where $\delta Q_{\min,i} = Q_i - \min(Q_{i+1}, Q_i, Q_{i-1})$ and $\delta Q_{\max,i} = \max(Q_{i+1}, Q_i, Q_{i-1}) - Q_i$. The monotonicity is achieved by the following scheme:

$$q_{L,i} \leftarrow Q_i - \max(|\delta_m Q_i|, |q_{L,i} - Q_i|) \cdot \operatorname{sgn}(\delta_m Q_i), \tag{2.71}$$

$$q_{R,i} \leftarrow Q_i - \max(|\delta_m Q_i|, |q_{R,i} - Q_i|) \cdot \operatorname{sgn}(\delta_m Q_i). \tag{2.72}$$

This scheme may be further improved to reduce the diffusion even more as described by Lin (2004), but we are not going to assess this approach here.

2.3.3 Flux

Time-averaged flux

As we pointed out in Problem 2.3, we must approximate the time-averaged flux $\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, t)\, dt$ at the cell edges $x_{i+\frac{1}{2}}$ to deduce a finite-volume scheme. One way to do that is to first write this temporal integral as a spatial integral at time t^n. After that, we can estimate this spatial integral using the reconstruction function at time t^n. As we shall see soon, this spatial integral domain is related with the definition of departure point.

To introduce the definition of departure point, for each $s \in [t^n, t^{n+1}]$, we consider the following Cauchy problem backward in time:

$$\begin{cases} \frac{\partial X}{\partial t}(t, s; \alpha) = u(X(t, s; \alpha), t), \quad t \in [t^n, s] \\ X(s, s; \alpha) = \alpha. \end{cases} \tag{2.73}$$

The point $X(t^n, s; \alpha)$ is called departure point at time t^n of the point α at time s. Integrating

Equation (2.73) over the interval $[t, s]$, we get:

$$X(t, s; \alpha) = \alpha - \int_t^s u(X(\theta, s; \alpha), \theta)\, d\theta. \tag{2.74}$$

In the next proposition, we show how the time-averaged integral is related to a spatial integral over the departure points.

Proposition 2.6. *Assume the framework of Problem 2.2. If q and u are C^1 functions, then:*

$$\int_{t^n}^{t^{n+1}} (uq)(\alpha, s)\, ds = \int_{X(t^n, t^{n+1}; \alpha)}^{\alpha} q(x, t^n)\, dx \tag{2.75}$$

Proof. Using the Leibniz rule for integration in Equation (2.74), it follows that:

$$\begin{aligned} \frac{\partial X}{\partial s}(t, s; \alpha) &= -\left(u(\alpha, s) + \int_t^s \frac{du}{ds}(X(\theta, s; \alpha), \theta)\, d\theta \right) \\ &= -u(\alpha, s) - \int_t^s \frac{\partial u}{\partial x}(X(\theta, s; \alpha), \theta) \frac{\partial X}{\partial s}(\theta, s; \alpha)\, d\theta. \end{aligned} \tag{2.76}$$

Taking the derivative with respect to t of Equation (2.76), we have:

$$\frac{\partial}{\partial t}\left(\frac{\partial X}{\partial s} \right)(t, s; \alpha) = \frac{\partial u}{\partial x}(X(t, s; \alpha), t) \frac{\partial X}{\partial s}(t, s; \alpha). \tag{2.77}$$

Using standard ODE techniques, we get that X that solves Equations (2.76) and (2.77) is given by:

$$\frac{\partial X}{\partial s}(t, s; \alpha) = -\exp\left(\int_t^s \frac{\partial u}{\partial x}(X(\theta, s; \alpha), \theta)\, d\theta \right) u(\alpha, s). \tag{2.78}$$

Computing q on the trajectory give by $X(t, s; \alpha)$ and taking its time derivative, we obtain:

$$\begin{aligned} \frac{dq}{dt}(X(t, s; \alpha), t) &= \frac{\partial q}{\partial t}(X(t, s; \alpha), t) + u(X(t, s; x_{i+\frac{1}{2}}), t) \frac{\partial q}{\partial x}(X(t, s; \alpha), t) \\ &= -\frac{\partial u}{\partial x}(X(t, s; \alpha), t) q(X(t, s; \alpha), t), \end{aligned} \tag{2.79}$$

where we used that q satisfies the linear advection equation on its differential form and that $X(t, s; x_{i+\frac{1}{2}})$ solves Equation (2.73). Using again standard ODE techniques, we get that q that solves Equation (2.79) is given by:

$$q(X(t, s; \alpha), t) = \exp\left(-\int_t^s \frac{\partial u}{\partial x}(X(\theta, s; \alpha), \theta)\, d\theta \right) q(\alpha, s). \tag{2.80}$$

Notice that if u does not depend on x, then q is constant along the trajectory $X(t, s; \alpha)$.

Let us consider the mapping $s \in [t^n, t^{n+1}] \to X(t^n, s, \alpha)$. Integrating q over all departure

points at time t^n from α at time s, we have

$$\int_{X(t^n,t^{n+1};\alpha)}^{X(t^n,t^{n+1};\alpha)} q(x,t^n)\,dx = \int_{t^n}^{t^{n+1}} q(X(t^n,s;\alpha),t^n)\frac{\partial X}{\partial s}(t^n,s;\alpha)\,ds, \tag{2.81}$$

where we are just using the variable change integration formula. Then, it follows from Equations (2.78) and (2.80) with $t = t^n$ that:

$$\int_{\alpha}^{X(t^n,t^{n+1};\alpha)} q(x,t^n)\,dx = -\int_{t^n}^{t^{n+1}} (uq)(\alpha,s)\,ds,$$

which is the desired formula. □

Therefore, we can conclude from Proposition 2.6 that the flux computation at an edge $\alpha = x_{i+\frac{1}{2}}$ requires three step: reconstruction of the solution, computation of the departure points from $x_{i+\frac{1}{2}}$ and computation of the spatial integral. Since the reconstruction problem has been already addressed in Section 2.3.1, we now focus on the departure point computation.

Departure point computation

It follows from Proposition 2.6 that the evaluation of the time-averaged flux may be replaced by the departure point calculation. We recall the definition of the Courant-Friedrichs-Lewy (CFL) condition.

Definition 2.11. *For Problem 2.3, we say that the CFL condition holds if* $\frac{\Delta t}{\Delta x}|u^n_{i+\frac{1}{2}}| \leq 1$, $\forall i = 0,\cdots,N, n = 0,\cdots,N_T$. *The CFL number is defined by* $\frac{\Delta t}{\Delta x}\max\{|u^n_{i+\frac{1}{2}}| : i = 0,\cdots,N, n = 0,\cdots,N_T\}$.

Before addressing the question of how compute the departure point, we point out that Equation (2.74) is useful to show that the trajectory $X(t,s;x_{i+\frac{1}{2}})$ remains on a single control volume, provided some basic assumptions, as we show on the next proposition.

Proposition 2.7. *If* $u \in C^1$, *the CFL condition is satisfied and if* Δt *and* Δx *are small enough, then for any* $s \in [t^n, t^{n+1}]$, $t \in [t^n, s]$, *we have that* $X(t,s;x_{i+\frac{1}{2}}) \in X_i$ *if* $u^n_{i+\frac{1}{2}} > 0$ *and* $X(t,s;x_{i+\frac{1}{2}}) \in X_{i+1}$ *if* $u^n_{i+\frac{1}{2}} < 0$.

Proof. Let us assume $u^n_{i+\frac{1}{2}} > 0$. Since u is continuous and the CFL condition is satisfied, there exist Δt and Δx such that $|u(X(t,s;x_{i+\frac{1}{2}}),t)|\frac{\Delta t}{\Delta x} \leq 1$ and $u(X(t,s;x_{i+\frac{1}{2}}),t) > 0$, $\forall s \in [t^n, t^{n+1}]$, $t \in [t^n, s]$. Hence, it follows from Equation (2.74) and the mean value theorem for integrals (see Theorem A.2) that:

$$X(t,s;x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - (t-s)u(X(\theta_1,s;x_{i+\frac{1}{2}}),\theta_1),$$

for some $\theta_1 \in [t,s]$. Therefore:

$$X(t,s;x_{i+\frac{1}{2}}) \geq x_{i+\frac{1}{2}} - \Delta t u(X(\theta_1,s;x_{i+\frac{1}{2}}),\theta_1) \geq x_{i+\frac{1}{2}} - \Delta x = x_{i-\frac{1}{2}},$$

from which the claim follows. The case $u^n_{i+\frac{1}{2}} < 0$ is very similar to the case $u^n_{i+\frac{1}{2}} > 0$. □

Equation (2.74) allow us to compute or estimate the departure point. For instance, if u is constant, then the departure point at time t^n of the point $x_{i+\frac{1}{2}}$ at time t^{n+1} is given by:

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - u\Delta t. \tag{2.82}$$

For the general case where u may depend on x and t, we follow the approach of the original PPM scheme from Colella and Woodward (1984) using:

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t + C\Delta t^2, \tag{2.83}$$

with $\tilde{u}^n_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}}$. This scheme shall be referred to as **RK1**. Our aim now is to exhibit the constant C on the next proposition. It is useful to before introduce the material derivative

$$\frac{Dh}{Dt} = \frac{\partial h}{\partial t} + u\frac{\partial h}{\partial x},$$

for a function $h \in C^1$.

Proposition 2.8. *If $u \in C^1$, then X that solves Equation (2.73) satisfies:*

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - u^n_{i+\frac{1}{2}}\Delta t + C\Delta t^2, \tag{2.84}$$

for a constant C depending on u.

Proof. Using Corollary A.1 for the function $f(t) = u(X(t, t^{n+1}; x_{i+\frac{1}{2}}), t)$ in Equation (2.74), we get:

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - u(X(t^n, t^{n+1}; x_{i+\frac{1}{2}}), t^n)\Delta t + \frac{1}{2}\frac{Du}{Dt}(\tilde{x}_1, \tilde{t}_1)\Delta t^2, \tag{2.85}$$

for $\tilde{x}_1 \in X_i \cup X_{i+1}$, $\tilde{t}_1 \in [t^n, t^{n+1}]$, for small Δx and Δt as in Proposition 2.7. Using the Taylor's expansion of $u(X(t, t^{n+1}; x_{i+\frac{1}{2}}), t^n)$ we have:

$$u(X(t^n, t^{n+1}; x_{i+\frac{1}{2}}), t^n) = u^n_{i+\frac{1}{2}} - \left(u\frac{\partial u}{\partial x}\right)(\tilde{x}_2, \tilde{t}_2)\Delta t, \tag{2.86}$$

for $\tilde{t}_2 \in [t^n, t^{n+1}]$. Hence, replacing Equation (2.86) in Equation (2.85) we obtain the desired constant C given by:

$$C = C(\tilde{x}_1, \tilde{x}_2, \tilde{t}_1, \tilde{t}_2) = \left(\frac{1}{2}\frac{Du}{Dt}(\tilde{x}_1, \tilde{t}_1) - \left(u\frac{\partial u}{\partial x}\right)(\tilde{x}_2, \tilde{t}_2)\right)\Delta t. \tag{2.87}$$

□

The problem of estimating the departure point is very common in Semi-Lagrangian scheme, which are quite popular in atmospheric modeling. For a review on departure point calculation methods, we refer to Tumolo (2011, Chapter 3) and the references therein. There are different approaches to compute the departure point, such as integrating the ODE from Equation 2.6 using different time integrators (D. Durran, 2011) backward in time. The Runge-Kutta methods are a possible choice to compute the departure point (*cf. e.g.* Guo et al. (2014), Lu et al. (2022)). In this work, we shall consider a second-order

Runge-Kutta to compute the departure point, which we express in terms of $\tilde{u}^n_{i+\frac{1}{2}}$ using the following equations (D. R. Durran, 2010)

$$\begin{aligned} x^d_{i+\frac{1}{2}} &= x_{i+\frac{1}{2}} - u^n_{i+\frac{1}{2}} \frac{\Delta t}{2}, \\ \tilde{u}^n_{i+\frac{1}{2}} &= u\left(x^d_{i+\frac{1}{2}}, t^n + \frac{\Delta t}{2} \right). \end{aligned} \tag{2.88}$$

Notice that this scheme requires values of u at points that are not grid points, both in time and space. This problem is addressed firstly using a second-order extrapolation in time

$$u^{n+\frac{1}{2}}_{i+\frac{1}{2}} = \frac{3}{2} u^n_{i+\frac{1}{2}} - \frac{1}{2} u^{n-1}_{i+\frac{1}{2}}. \tag{2.89}$$

and then we use a linear interpolation in space,

$$\tilde{u}^n_{i+\frac{1}{2}} = \begin{cases} \left(1 - \frac{x_{i+\frac{1}{2}} - x^d_{i+\frac{1}{2}}}{\Delta x} \right) u^{n+\frac{1}{2}}_{i+\frac{1}{2}} + \left(\frac{x_{i+\frac{1}{2}} - x^d_{i+\frac{1}{2}}}{\Delta x} \right) u^{n+\frac{1}{2}}_{i-\frac{1}{2}} & \text{if } u^n_{i+\frac{1}{2}} \geq 0, \\ \left(\frac{x_{i+\frac{1}{2}} - x^d_{i+\frac{1}{2}}}{\Delta x} \right) u^{n+\frac{1}{2}}_{i+\frac{3}{2}} + \left(1 - \frac{x_{i+\frac{1}{2}} - x^d_{i+\frac{1}{2}}}{\Delta x} \right) u^{n+\frac{1}{2}}_{i+\frac{1}{2}} & \text{if } u^n_{i+\frac{1}{2}} < 0. \end{cases} \tag{2.90}$$

This scheme leads to a 3-th order error in departure point (see *e.g.* D. R. Durran (2010, Section 7.1.2)):

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}} \Delta t + C \Delta t^3, \tag{2.91}$$

where C is a constant depending on u and its derivatives. This scheme shall be referred to as **RK2**. Notice that for this scheme, we need ghost values for the velocity, namely $u^n_{-\frac{1}{2}}$ and $u^n_{N+\frac{3}{2}}$. The linear interpolation described makes sense when we assume the CFL condition. When we have a larger CFL number, it is straightforward to modify the linear interpolation to compute the velocity needed.

Numerical flux

Let us assume the framework from Problem 2.3. Supposing that $Q^n \in \mathbb{R}^{\Delta x}$ is known, we would like to compute the values Q^{n+1}. This is achieved using a scheme of the type given in Problem 2.3. Therefore, we need to estimate the time-averaged flux. From Proposition 2.6, we conclude that the time-averaged flux may be compute as a spatial integral, which requires an estimate of the departure point using, for instance, a time-average velocity as in Equations (2.83) and (2.88). For each control volume edge $i = 0, \cdots, N$ and $y > 0$ we define the following average of the Piecewise-Parabolic approximation defined in Equation (2.33) for Q^n (Colella & Woodward, 1984):

$$F_{L,i+\frac{1}{2}}(y) = \frac{1}{y} \int_{x_{i+\frac{1}{2}} - y}^{x_{i+\frac{1}{2}}} q_{PP}(\xi; Q^n)\, d\xi, \tag{2.92}$$

and

$$F_{R,i+\frac{1}{2}}(y) = \frac{1}{y} \int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}} + y} q_{PP}(\xi; Q^n)\, d\xi, \tag{2.93}$$

If $y \leq \Delta x$, then both of the above integral domains are constrained to a single control volume. Thus, it follows from a straightforward computation using Equation (2.37) that:

$$F_{L,i+\frac{1}{2}}(y) = \frac{1}{y}\int_{x_{i+\frac{1}{2}}-y}^{x_{i+\frac{1}{2}}} q_i(\xi;Q^n)\,d\xi = q_{R,i} + \frac{(q_{6,i}-\Delta q_i)}{2\Delta x}y - \frac{q_{6,i}}{3\Delta x^2}y^2, \tag{2.94}$$

and

$$F_{R,i+\frac{1}{2}}(y) = \frac{1}{y}\int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}}+y} q_{i+1}(\xi;Q^n)\,d\xi = q_{L,i+1} + \frac{(q_{6,i+1}+\Delta q_{i+1})}{2\Delta x}y - \frac{q_{6,i+1}}{3\Delta x^2}y^2. \tag{2.95}$$

The numerical flux function is then defined by:

$$\mathcal{F}(Q^n(\mathcal{S}_{i+\frac{1}{2}}),\tilde{u}^n_{i+\frac{1}{2}}) = F^n_{i+\frac{1}{2}} = \begin{cases} \tilde{u}^n_{i+\frac{1}{2}}F_{L,i+\frac{1}{2}}(\tilde{u}^n_{i+\frac{1}{2}}\Delta t) & \text{if } \tilde{u}^n_{i+\frac{1}{2}} \geq 0, \\ \tilde{u}^n_{i+\frac{1}{2}}F_{R,i+\frac{1}{2}}(-\tilde{u}^n_{i+\frac{1}{2}}\Delta t) & \text{if } \tilde{u}^n_{i+\frac{1}{2}} < 0, \end{cases} \tag{2.96}$$

where $\tilde{u}^n_{i+\frac{1}{2}}$ is the velocity used in the departure point estimation. The numerical flux function may be also expressed as:

$$\mathcal{F}(Q^n(\mathcal{S}_{i+\frac{1}{2}}),\tilde{u}^n_{i+\frac{1}{2}}) = \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-\tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q_{PP}(x;Q^n)\,dx. \tag{2.97}$$

From this Equation (2.97), observe that we could compute the fluxes using reconstruction functions of the form (2.33) for general Φ rather than just parabolas. Notice that if we define:

$$c^n_{i+\frac{1}{2}} = \tilde{u}^n_{i+\frac{1}{2}}\frac{\Delta t}{\Delta x},$$

the requirement $y \leq \Delta x$ for Equation (2.96) is equivalent to require that $|c^n_{i+\frac{1}{2}}| \leq 1$ for all i, which is the CFL condition. This requirement is needed to ensure that the integration domain is contained in a single control volume, hence the integral may be evaluated using only $q_i(x;Q^n)$. However, if Δt is such that the integration domain contains more than one control volume, due to the local mass preservation of the reconstruction, we may evaluate the integral using the accumulated mass from the point $x_{i+\frac{1}{2}}$ to the control-volume that contains the departure point and using the local reconstruction function on the control-volume that contains the departure point. This approach allows larger-time steps (Y. Chen et al., 2017).

If the monotonic scheme from Lin (2004) is employed, then $\mathcal{S}_{i+\frac{1}{2}} = \{i-3, i-2, i-1, i, i+1, i+2, i+3\}$. Otherwise the stencil is given by $\mathcal{S}_{i+\frac{1}{2}} = \{i-2, i-1, i, i+1, i+2\}$ for all the other schemes that we presented. In the absence of monotonization, it follows from Equations (2.44), (2.45), (2.56) and (2.57) that the numerical flux may be expressed as the following sum:

$$\mathcal{F}(Q(\mathcal{S}_{i+\frac{1}{2}}),\tilde{u}^n_{i+\frac{1}{2}}) = \tilde{u}^n_{i+\frac{1}{2}}\sum_{k=-2}^{3}\alpha_{i,k}Q^n_{i+k},$$

where the coefficients are satisfies:

$$12\alpha_{i,-2} = \begin{cases} c_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ 0 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$12\alpha_{i,-1} = \begin{cases} -1 - 5c_{i+\frac{1}{2}} + 6c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ -1 + 2c_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$12\alpha_{i,0} = \begin{cases} 7 + 15c_{i+\frac{1}{2}} - 10c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ 7 - 13c_{i+\frac{1}{2}} + 6c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$12\alpha_{i,1} = \begin{cases} 7 - 13c_{i+\frac{1}{2}} + 6c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ 7 + 15c_{i+\frac{1}{2}} - 10c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$12\alpha_{i,2} = \begin{cases} -1 + 2c_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ -1 - 5c_{i+\frac{1}{2}} + 6c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$12\alpha_{i,3} = \begin{cases} 0 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ c_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

If the reconstruction at the edges is calculated using Equation (2.46) and (2.47), which is leads to a scheme called hybrid PPM (Putman & Lin, 2007), the flux stencil coefficients may be written as:

$$60\alpha_{i,-2} = \begin{cases} 2 - c_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ 0 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$60\alpha_{i,-1} = \begin{cases} -13 - c_{i+\frac{1}{2}} + 14c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ -3 + 4c_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$60\alpha_{i,0} = \begin{cases} 47 + 39c_{i+\frac{1}{2}} - 26c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ 27 - 41c_{i+\frac{1}{2}} + 14c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$60\alpha_{i,1} = \begin{cases} 27 - 41c_{i+\frac{1}{2}} + 14c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ 47 + 39c_{i+\frac{1}{2}} - 26c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$60\alpha_{i,2} = \begin{cases} -3 + 4c_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ -13 - c_{i+\frac{1}{2}} + 14c_{i+\frac{1}{2}}^2 & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases}$$

$$60\alpha_{i,3} = \begin{cases} 0 & \text{if } \tilde{u}^n_{i+\frac{1}{2}} \geq 0, \\ 2 - c_{i+\frac{1}{2}} - c^2_{i+\frac{1}{2}} & \text{if } \tilde{u}^n_{i+\frac{1}{2}} < 0, \end{cases}$$

Flux numerical analysis

With the stencil coefficients, we can compute the amplification factor (Equation (2.32)) for the PPM and the hybrid PPM schemes, both without monotonization. We assume a constant velocity equal to one and $N = 100$ (number of control volumes). In Figure 2.1 we show the amplification factor for both PPM and hybrid PPM schemes considering different CFL numbers. We can observe that both schemes damp most of the Fourier modes for larger k, regardless of the CFL number. Besides that, the hybrid scheme is more effective when reducing the Fourier modes amplitude. We point out that both schemes are exact when the CFL number is equal to 1. From this analysis, we can conclude that the PPM and hybrid PPM schemes satisfy the Von Neumann stability criteria when the CFL restriction is respected. In order to investigate the consistency of the PPM scheme, we notice that when

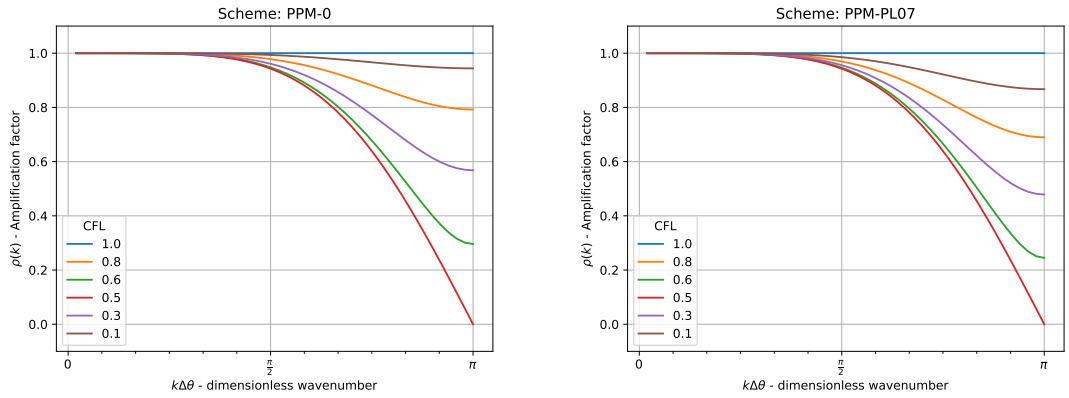


Figure 2.1: *Amplification factor for the PPM (left) and hybrid PPM (right) schemes for different CFL numbers.*

we are deducing the time average flux, we are making approximations of the form:

$$\int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, t)\, dt \approx \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q_{PP}(x; Q^n)\, dx, \tag{2.98}$$

as we can see from Equations (2.92) and (2.93), which basically replace q by q_{PP} and $X(t^n, t^{n+1}, x_{i+\frac{1}{2}})$ by $x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t$ on the right-hand side of Equation (2.75). As we shall see, the approximation (2.98) has two sources of error: one related to the departure point estimation and another due to the parabolic approximation. The next Proposition (2.9) investigates how the departure point error impacts on the approximation (2.98) replacing q_{PP} by q.

Proposition 2.9. *Assume the framework of Problem 2.2 with $q \in C^1$ and $u \in C^P$, for some $P \geq 2$. Furthermore, assume the CFL condition and that Δx and Δt are small enough as in Proposition 2.7.*

Assume also that for some time-average velocity $\tilde{u}^n_{i+\frac{1}{2}}$, *we have* $X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t + C_{i+\frac{1}{2}}\Delta t^P$, *where the constants* $C_{i+\frac{1}{2}}$ *can be written as* $C_{i+\frac{1}{2}} = F(\tilde{x}_{i+\frac{1}{2}}, \tilde{t}_{i+\frac{1}{2}})$ *for a* C^1 *function* $F : [a,b]^d \times [0,T]^d \to \mathbb{R}$. *Besides that, assume* $\tilde{x}_{i+\frac{1}{2}} \in [x_{i+\frac{1}{2}} - k\Delta x, x_{i+\frac{1}{2}} + k\Delta x]^d$, *where* k *do not depend on* Δx *and* $\tilde{t}_{i+\frac{1}{2}} \in [t^n, t^{n+1}]^d$. *Under all these assumptions, we have:*

$$\left| \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s)\, ds - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q(x, t^n)\, dx - \left(\int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, s)\, ds - \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}}\Delta t}^{x_{i-\frac{1}{2}}} q(x, t^n)\, dx \right) \right|$$
$$\leq K_1 \Delta t^{P+1},$$

where K_1 *depends on* q *and* u.

Proof. Using Equation (2.75) and the mean value theorem for integrals, we get:

$$\int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s)\, ds - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q(x, t^n)\, dx = \int_{X(t^n, t^{n+1}; x_{i+\frac{1}{2}})}^{x_{i+\frac{1}{2}}} q(x, t^n)\, dx - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q(x, t^n)\, dx$$
$$= \int_{X(t^n, t^{n+1}; x_{i+\frac{1}{2}})}^{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t} q(x, t^n)\, dx = \left(X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) - x_{i+\frac{1}{2}} + \tilde{u}^n_{i+\frac{1}{2}}\Delta t\right) q(\mu_i, t^n) = F(\tilde{x}_{i+\frac{1}{2}}, \tilde{t}_{i+\frac{1}{2}})\Delta t^P q(\mu_{i+\frac{1}{2}}, t^n),$$

for some $\mu_{i+\frac{1}{2}} \in X_i \cup X_{i+1}$. Similarly, we have:

$$\int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, s)\, ds - \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}}\Delta t}^{x_{i-\frac{1}{2}}} q(x, t^n)\, dx = F(\tilde{x}_{i-\frac{1}{2}}, \tilde{t}_{i-\frac{1}{2}})\Delta t^P q(\mu_{i-\frac{1}{2}}, t^n),$$

and again, $\mu_{i-\frac{1}{2}} \in X_{i-1} \cup X_i$ We introduce the following auxiliary C^1 function:

$$G(v) = F(v_1, v_2) q(v_3, t^n).$$

where $v = (v_1, v_2, v_3)$, $v_1 \in [a,b]^d$, $v_2 \in [0,T]^d$, $v_3 \in [a,b]$. Introducing $v_{i+\frac{1}{2}} = (\tilde{x}_{i+\frac{1}{2}}, \tilde{t}_{i+\frac{1}{2}}, \mu_{i+\frac{1}{2}})$, $v_{i-\frac{1}{2}} = (\tilde{x}_{i-\frac{1}{2}}, \tilde{t}_{i-\frac{1}{2}}, \mu_{i-\frac{1}{2}})$ and using the mean value theorem, we have:

$$|G(v_{i+\frac{1}{2}}) - G(v_{i-\frac{1}{2}})| \leq \left(\sup_{v \in [a,b]^d \times [0,T]^d \times [a,b]} \|\nabla G(v)\|_{2d+1} \right) \|v_{i+\frac{1}{2}} - v_{i-\frac{1}{2}}\|_{2d+1}$$
$$\leq \left(\sqrt{(d(2k+1)^2 + 9)\lambda^2 + d} \sup_{v \in [a,b]^d \times [0,T]^d \times [a,b]} \|\nabla G(v)\|_{2d+1} \right) \Delta t,$$

where $\|\cdot\|_D$ is the 2-norm of $\mathbb{R}^D$ and we used that $\|\tilde{x}_{i+\frac{1}{2}} - \tilde{x}_{i-\frac{1}{2}}\|_d^2 \leq d(2k+1)^2\Delta x^2$, $|\tilde{\mu}_{i+\frac{1}{2}} - \tilde{\mu}_{i-\frac{1}{2}}| \leq 3\Delta x$, and $\|\tilde{t}_{i+\frac{1}{2}} - \tilde{t}_{i-\frac{1}{2}}\|_d^2 \leq d\Delta t^2$. Finally, we have the desired bound:

$$\left| \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s)\, ds - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q(x, t^n)\, dx - \left(\int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, s)\, ds - \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}}\Delta t}^{x_{i-\frac{1}{2}}} q(x, t^n)\, dx \right) \right|$$
$$= |(G(v_{i+\frac{1}{2}}) - G(v_{i-\frac{1}{2}}))|\Delta t^P \leq K_1 \Delta t^{P+1},$$

where $K_1 = \sqrt{(d(2k+1)^2 + 9)\lambda^2 + d}\, \sup_{v \in [a,b]^d \times [0,T]^d \times [a,b]} \|\nabla G(v)\|_{2d+1}$. □

Remark 2.8. *If the departure point is computed using Equation (2.83), it follows from Proposition 2.8 that $P = 2$, $k = 3$, $d = 2$. The function F is defined by the right-hand side of Equation (2.87).*

The next proposition gives a measure of the impact of the Piecewise-Parabolic approximation on the time average flux, considering this last computed using an estimated departure point.

Proposition 2.10. *Assume the framework of Problem 2.2. If $q \in \mathcal{C}^5$ and $u \in \mathcal{C}^1$, then:*

$$\left|\frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-\tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q(x,t^n)\,dx - \mathcal{F}(Q(t_n)(S_{i+\frac{1}{2}}),\tilde{u}^n_{i+\frac{1}{2}}) - \left(\frac{1}{\Delta t}\int_{x_{i-\frac{1}{2}}-\tilde{u}^n_{i-\frac{1}{2}}\Delta t}^{x_{i-\frac{1}{2}}} q(x,t^n)\,dx - \mathcal{F}(Q(t_n)S_{i-\frac{1}{2}},\tilde{u}^n_{i-\frac{1}{2}})\right)\right| \leq K_2\Delta x^4,$$

where K_2 depends on q and u.

Proof. We denote by q_{PP} the piecewise-parabolic approximation of $Q(t^n)$. Then:

$$\frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-\tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q(x,t^n)\,dx - \mathcal{F}(Q(t_n)(S_{i+\frac{1}{2}}),\tilde{u}^n_{i+\frac{1}{2}}) = \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-\tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q(x,t^n)\,dx - \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-\tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} q_{PP}(x;Q(t^n))\,dx$$

and

$$\frac{1}{\Delta t}\int_{x_{i-\frac{1}{2}}-\tilde{u}^n_{i-\frac{1}{2}}\Delta t}^{x_{i-\frac{1}{2}}} q(x,t^n)\,dx - \mathcal{F}(Q(t_n)(S_{i-\frac{1}{2}}),\tilde{u}^n_{i-\frac{1}{2}}) = \frac{1}{\Delta t}\int_{x_{i-\frac{1}{2}}-\tilde{u}^n_{i-\frac{1}{2}}\Delta t}^{x_{i-\frac{1}{2}}} q(x,t^n)\,dx - \frac{1}{\Delta t}\int_{x_{i-\frac{1}{2}}-\tilde{u}^n_{i-\frac{1}{2}}\Delta t}^{x_{i-\frac{1}{2}}} q_{PP}(x;Q(t^n))\,dx.$$

Similarly to Proposition 2.5, we can write:

$$\begin{aligned} q(x,t^n) - q_{PP}(x;Q(t^n)) =& C_1(\mu_1,\mu_2)\Delta x^4 + C_2(\mu_3,\mu_4)\Delta x^2(x-x_L) + \frac{C_3}{2}(\mu_5,\mu_6)\Delta x(x-x_L)^2 \\ &+ C_4(\mu_7)(x-x_L)^3, \end{aligned}$$

where C_1, C_2, C_3 and C_4 are given by Equations (2.50), (2.60), (2.64) and (2.66) respectively, x_L is the left boundary of the control volume that contains x (X_i or X_{i+1}) and $\mu_k \in [x_{i+\frac{1}{2}} - 3\Delta x, x_{i+\frac{1}{2}} + 3\Delta x]$, $\forall k = 1,\cdots,7$. Similarly to Proposition 2.9, using the mean value theorem for integrals, one can write:

$$\int_{x_{i+\frac{1}{2}}-\tilde{u}^n_{i+\frac{1}{2}}\Delta t}^{x_{i+\frac{1}{2}}} (q(x,t^n) - q_{PP}(x;Q(t^n)))\,dx = F(\mu^1)\Delta x^4, \tag{2.99}$$

and

$$\int_{x_{i-\frac{1}{2}}-\tilde{u}^n_{i-\frac{1}{2}}\Delta t}^{x_{i-\frac{1}{2}}} (q(x,t^n) - q_{PP}(x;Q(t^n)))\,dx = F(\mu^2)\Delta x^4, \tag{2.100}$$

for an auxiliary function $F : [a,b]^8 \to \mathbb{R}, F \in \mathcal{C}^1$, where F depends on q, u, and C_1, C_2, C_3 and C_4. Subtracting Equation (2.100) from Equation (2.99) and using the mean value theorem, we get the desired inequality. □

Now we are able to tackle the consistency problem on the next proposition.

Proposition 2.11. *Assume the same hypothesis of Proposition 2.9 and Proposition 2.10. Denote by q_{PP} the Piecewise-Parabolic approximation of $q(x,t^n)$. Then, the LTE given by Equation (2.24) satisfies:*

$$|\tau_i^n| \leq M_1 \Delta t^{P-1} + M_2 \Delta x^3, \tag{2.101}$$

where M_1 and M_2 are constants depending only on q and u.

Proof. We have:

$$\Delta x \tau_i^n = \left| \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s)\, ds - \mathcal{F}(Q(t_n)(S_{i+\frac{1}{2}}), \tilde{u}^n_{i+\frac{1}{2}}) - \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, s)\, ds + \mathcal{F}(Q(t_n)(S_{i-\frac{1}{2}}), \tilde{u}^n_{i-\frac{1}{2}}) \right|$$

$$\frac{1}{\Delta t} \left| \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s)\, ds - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}} \Delta t}^{x_{i+\frac{1}{2}}} q_{PP}(x)\, dx - \left(\int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, s)\, ds - \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}} \Delta t}^{x_{i-\frac{1}{2}}} q_{PP}(x)\, dx \right) \right| =$$

$$\frac{1}{\Delta t} \left| \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s)\, ds - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}} \Delta t}^{x_{i+\frac{1}{2}}} q(x,t^n)\, dx + \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}} \Delta t}^{x_{i+\frac{1}{2}}} q(x,t^n)\, dx - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}} \Delta t}^{x_{i+\frac{1}{2}}} q_{PP}(x)\, dx \right.$$

$$\left. - \left(\int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, s)\, ds - \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}} \Delta t}^{x_{i-\frac{1}{2}}} q(x,t^n)\, dx + \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}} \Delta t}^{x_{i-\frac{1}{2}}} q(x,t^n)\, dx - \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}} \Delta t}^{x_{i-\frac{1}{2}}} q_{PP}(x)\, dx \right) \right| \leq$$

$$\frac{1}{\Delta t} \left| \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s)\, ds - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}} \Delta t}^{x_{i+\frac{1}{2}}} q(x,t^n)\, dx - \left(\int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, s)\, ds - \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}} \Delta t}^{x_{i-\frac{1}{2}}} q(x,t^n)\, dx \right) \right| +$$

$$\frac{1}{\Delta t} \left| \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}} \Delta t}^{x_{i+\frac{1}{2}}} q(x,t^n)\, dx - \int_{x_{i+\frac{1}{2}} - \tilde{u}^n_{i+\frac{1}{2}} \Delta t}^{x_{i+\frac{1}{2}}} q_{PP}(x)\, dx - \left(\int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}} \Delta t}^{x_{i-\frac{1}{2}}} q(x,t^n)\, dx - \int_{x_{i-\frac{1}{2}} - \tilde{u}^n_{i-\frac{1}{2}} \Delta t}^{x_{i-\frac{1}{2}}} q_{PP}(x)\, dx \right) \right|$$

Therefore, it follows from Propositions 2.9 and 2.10 that

$$|\tau_i^n| \leq \frac{1}{\Delta x \Delta t} K_1 \Delta t^{P+1} + \frac{1}{\Delta x} K_2 \Delta x^4 = K_1 \lambda \Delta t^{P-1} + K_2 \Delta x^3,$$

from which the proposition follows. □

Thus, it follows from Proposition 2.10 that the PPM scheme is consistent in the ∞-norm and has two sources of error: one related to the departure point calculation and another due to the parabolic approximation. In particular, the PPM flux using the departure point from Equation (2.83), we have a first-order error related to the departure point computation. We point out that, if the velocity is constant, then no error is obtained using the PPM flux, except for the approximation q_{PP} to q.

2.4 Numerical experiments

This Section is dedicated to presenting the numerical results of the PPM and its variations discussed here. For non-monotonic schemes, we are going to consider the original PPM from Colella and Woodward (1984)(hereafter will be referred as **PPM-0**) and the hybrid PPM from Putman and Lin (2007) (hereafter **PPM-PL07**). For monotonic schemes, we are going to consider the monotonization schemes from Colella and Woodward (1984) (hereafter **PPM-CW84**) and Lin (2004) (hereafter **PPM-L04**), which are referred to as CW84 monotonization and L04 monotonization hereafter. In Subsection 2.4.2 we present results using the linear advection equation with constant velocity and in Subsection 2.4.3 the results are based on the linear advection equation with variable velocity. The code used in this Section may be found in Appendix C.

2.4.1 Reconstruction at edges accuracy

Before moving toward advection simulations, we are going to assess the accuracy the order of the reconstruction schemes PPM-0, PPM-CW84, PPM-PL07 and PPM-L04 at the edges of the control volumes. To this purpose we consider the function to be reconstructed given by:

$$q(x) = \sin(2\pi k x) + 1, \quad x \in [0, 1]. \tag{2.102}$$

Here k denotes the wavenumber and we adopt $k = 5$. Inspired by Trefethen (2000), we also consider the following periodic Gaussian profile:

$$q(x) = \exp(-10\cos^2(2\pi x)), \quad x \in [0, 1]. \tag{2.103}$$

Both functions from Equations (2.103) and (2.103) are smooth. To check the convergence of the error at the edges, we consider $\Delta x^{(k)}-$grids with $\Delta x^{(k)} = 1/2^k$ for $k = 4, \cdots, 10$ and then we compute the normalized maximum error:

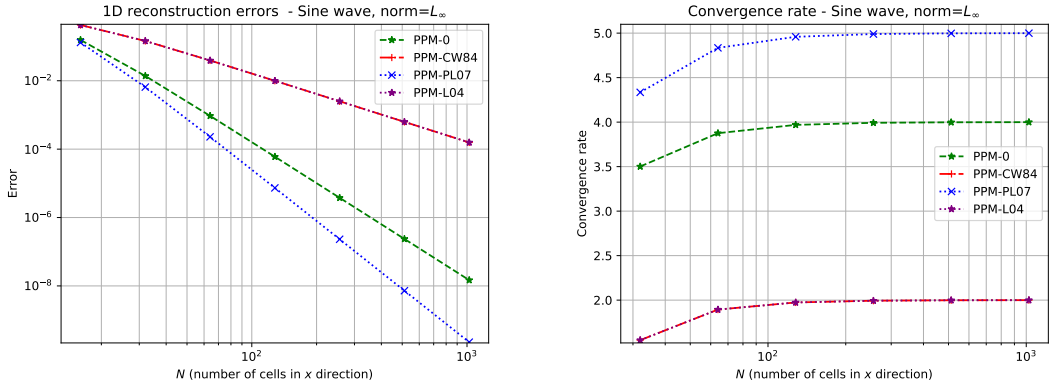
$$E_k = \frac{\max_{i=1,\cdots 2^k}\{|q_{L,i} - q_0(x_{i-\frac{1}{2}})|, |q_{R,i} - q_0(x_{i+\frac{1}{2}})|\}}{\max_{i=0,\cdots 2^k}\{|q_0(x_{i+\frac{1}{2}})(x)|\}},$$

and the values $q_{L,i}$ and $q_{R,i}$ where explained in Subsection 2.3.1 (Equation (2.37)). The convergence rate is defined by

$$CR_k = \frac{\ln\left(\frac{E_k}{E_{k-1}}\right)}{\ln 2}, \quad \text{for} \quad k = 5, \cdots 10.$$

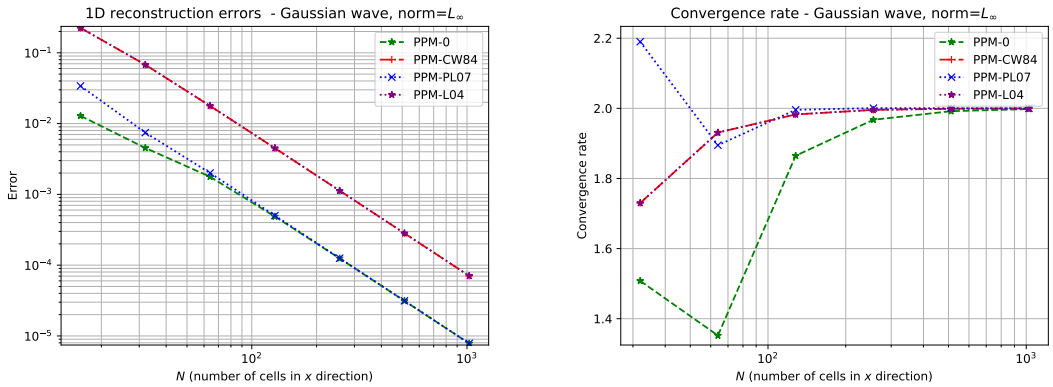
The major difference between the functions from Equations (2.102) and (2.103) is the fact of the function from (2.102) having an explicit formula for its primitive, allowing us to compute its average values exactly, while the function from (2.102) does not have an explicit formula of its primitive. As discussed in Section 2.3.1 (Proposition 2.2), we may approximate the average value using the value at the cell centroid with second-order accuracy for functions differentiable enough. Therefore, we expect that for the function from Equation (2.102), the error should converge with fourth and fifth order for the schemes PPM-0 and PPM-L07, respectively. For the function from (2.103) we expect at

least a second-order error for both schemes due to the approximation of the average value.



(a) *Error.* **(b)** *Convergence rate.*

Figure 2.2: *Convergence of the relative error at edges (a) and convergence rate (b) in the maximum norm for the reconstruction schemes PPM-0, PPM-CW84, and PPM-PL07 with PPM-L04 applied to the function given by Equation* (2.102).



(a) *Error.* **(b)** *Convergence rate.*

Figure 2.3: *Similar to Figure 2.2 the using function given by Equation* (2.102)

We can observe from Figure 2.2 that the reconstruction error for the function agrees with the theory: PPM-0 has fourth-order accuracy while PPM-PL07 has fifth-order accuracy. Besides that, both monotonic schemes PPM-CW84 and PPM-L04 converges with second-order having almost the same error. However, when we consider the Gaussian wave, due to the average value approximation by the centroid value, we observe that all schemes converge with second-order, also as expected, but the non-monotonic schemes have an error almost tenfold smaller than the monotonic schemes.

2.4.2 Linear advection equation with constant velocity simulations

For the linear advection equation with constant velocity we shall adopt the $u = 0.2$ and a CFL number equal to 0.8. The spatial domain will be given by $[0, 1]$ and the time integration interval will be $[0, 5]$. We again use $\Delta x^{(k)}$−grids with $\Delta x^{(k)} = 1/2^k$ for $k = 4, \cdots, 10$. Since we are going to assume periodic boundary conditions, the period is equal to 5. Hence, the simulations presented here shall advect an initial profile for one time period. The departure schemes RK1 and RK2 introduced in Section 2.3.3 compute the departure point exactly in this case, therefore we are going only to use the RK1 scheme. This shall be the general setup for all simulations presented in this subsection. What will distinguish the simulations is the initial condition. We consider q_0 given by Equations (2.102) and (2.103) as in Subsection 2.4.2. We also consider a discontinuous initial condition given by:

$$q_0(x) = \begin{cases} 1 & \text{if } x \in [0.4, 0.6], \\ 0 & \text{otherwise.} \end{cases} \tag{2.104}$$

It is easy to check that the exact solution of Problem 2.1 is given by $q_0(x - ut)$ for all q_0 presented here. We are going to consider the relative error in the maximum norm:

$$E_k = \max_{n=0,\cdots,N_T} \frac{\|Q^n - Q(t^n)\|_{\infty,\Delta x}}{\|Q(t^n)\|_{\infty,\Delta x}}.$$

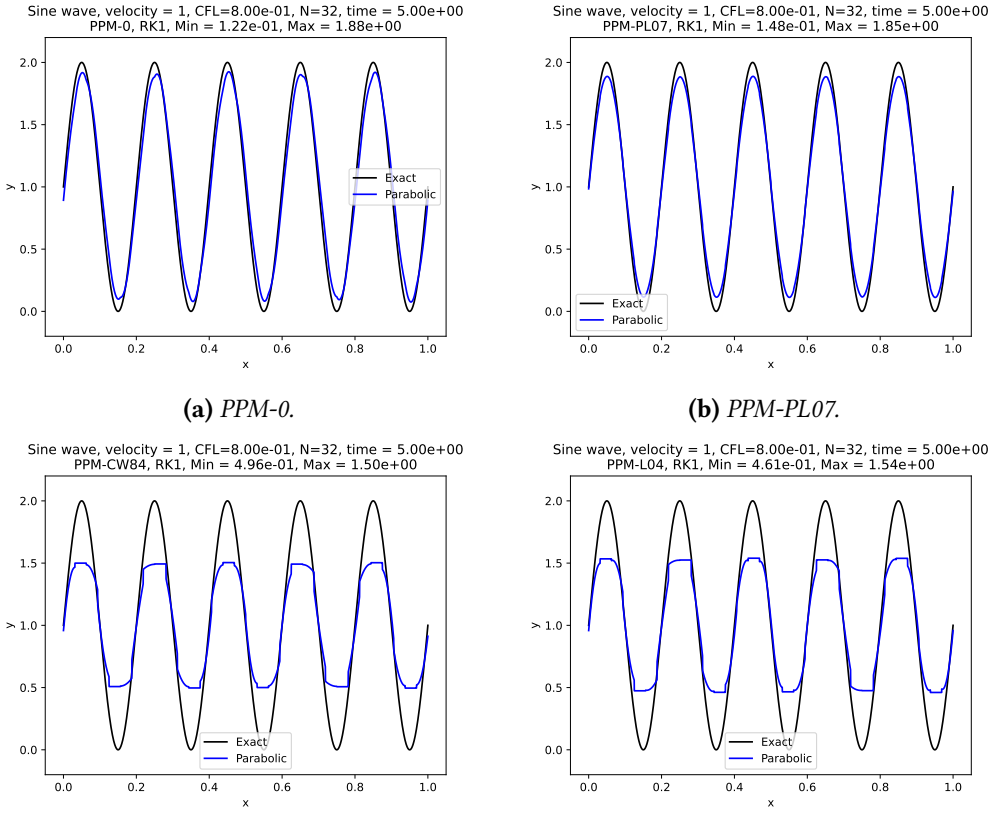
The convergence rate is defined by

$$CR_k = \frac{\ln\left(\frac{E_k}{E_{k-1}}\right)}{\ln 2}, \quad \text{for} \quad k = 5, \cdots 10.$$

As pointed in Subsection 2.4.1, when q_0 is given by Equation (2.103), we are going to compute the initial average values $Q_i(0)$ using the initial values of q_i^0 at the control volume centroids, which is second-order accurate by Proposition 2.1. In the error calculation, only when q_0 is given by Equation (2.103), we replace $Q_i(t^n)$ by its centroid value $q_i(t^n)$, which again gives a second-order approximation by Proposition 2.1. Therefore, since q_0 is smooth, we expect that the error convergence shall be at least second-order accurate. The relative change at time-step n in the mass is computed as

$$\frac{|M^n - M^0|}{|M_0|},$$

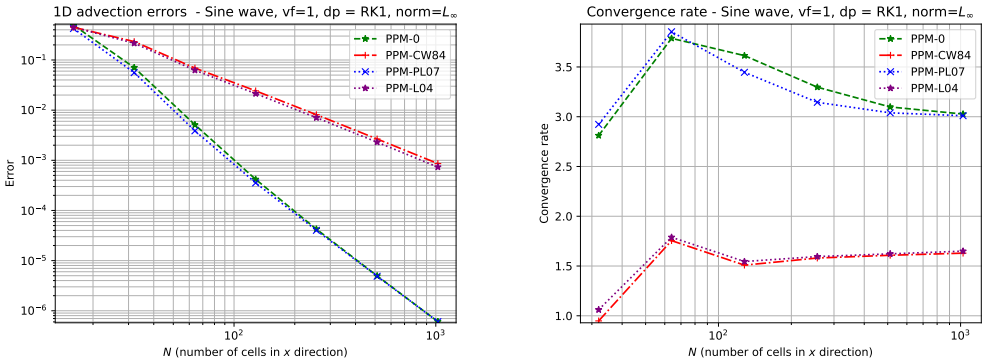
where M^n is given by Equation (2.22). For all the simulations, the mass is preserved with machine precision.



(a) *PPM-0.* **(b)** *PPM-PL07.*

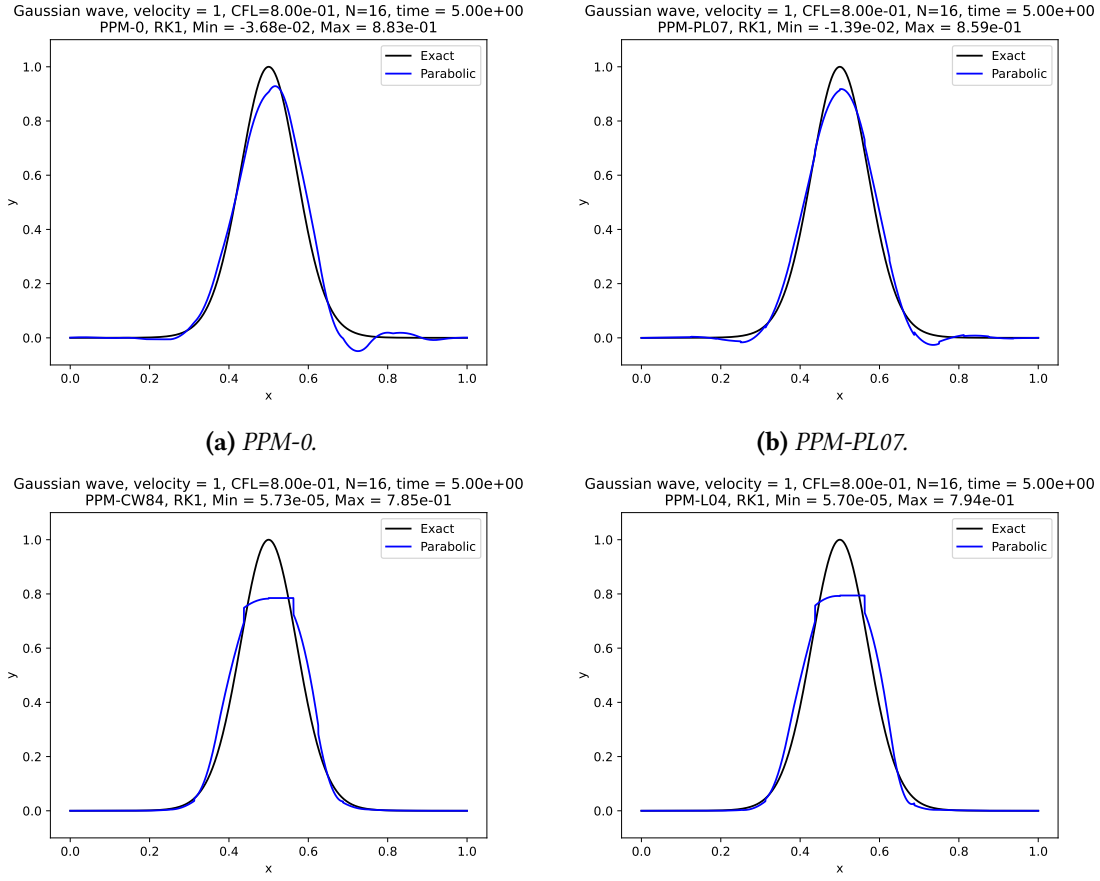
(c) *PPM-CW84.* **(d)** *PPM-L04.*

Figure 2.4: *Linear advection experiment using a constant velocity equal to* 0.1*, a CFL number equal to* 0.8*,* $N = 32$ *cells, and the initial condition is given by Equation* (2.102)*. These figures show the advected profile after 5 time units (one time period). Reconstruction schemes employed: PPM-0 (a), PPM-PL07 (b), PPM-CW84 (c) and PPM-L04 (d).*



(a) *Error.* **(b)** *Convergence rate.*

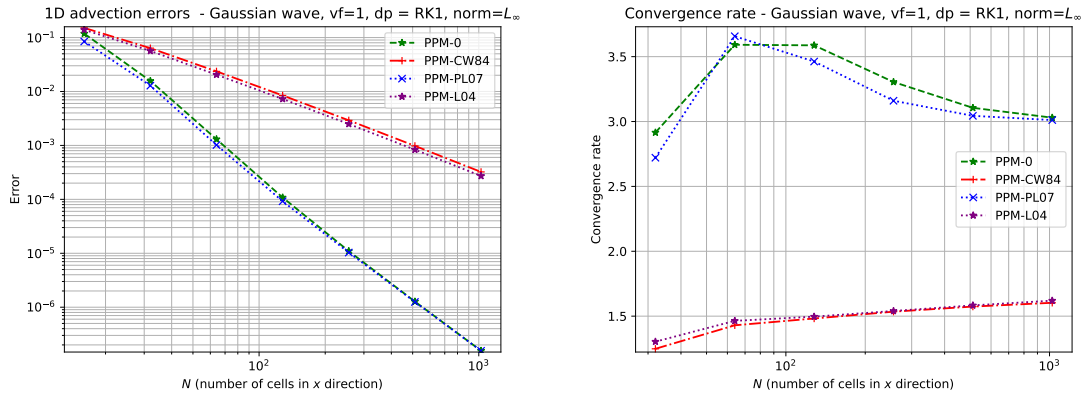
Figure 2.5: *Convergence of the error (a) and convergence rate (b) for the schemes PPM, hybrid PPM, PPM with the CW84 monotonization and PPM with L04 monotonization applied to the linear advection problem using a constant velocity equal to* 0.1*, a CFL number equal to* 0.8*, a final time of integration equal to 5 time units and the initial condition given by Equation* (2.102)*.*



(a) *PPM-0.* **(b)** *PPM-PL07.*

(c) *PPM-CW84.* **(d)** *PPM-L04.*

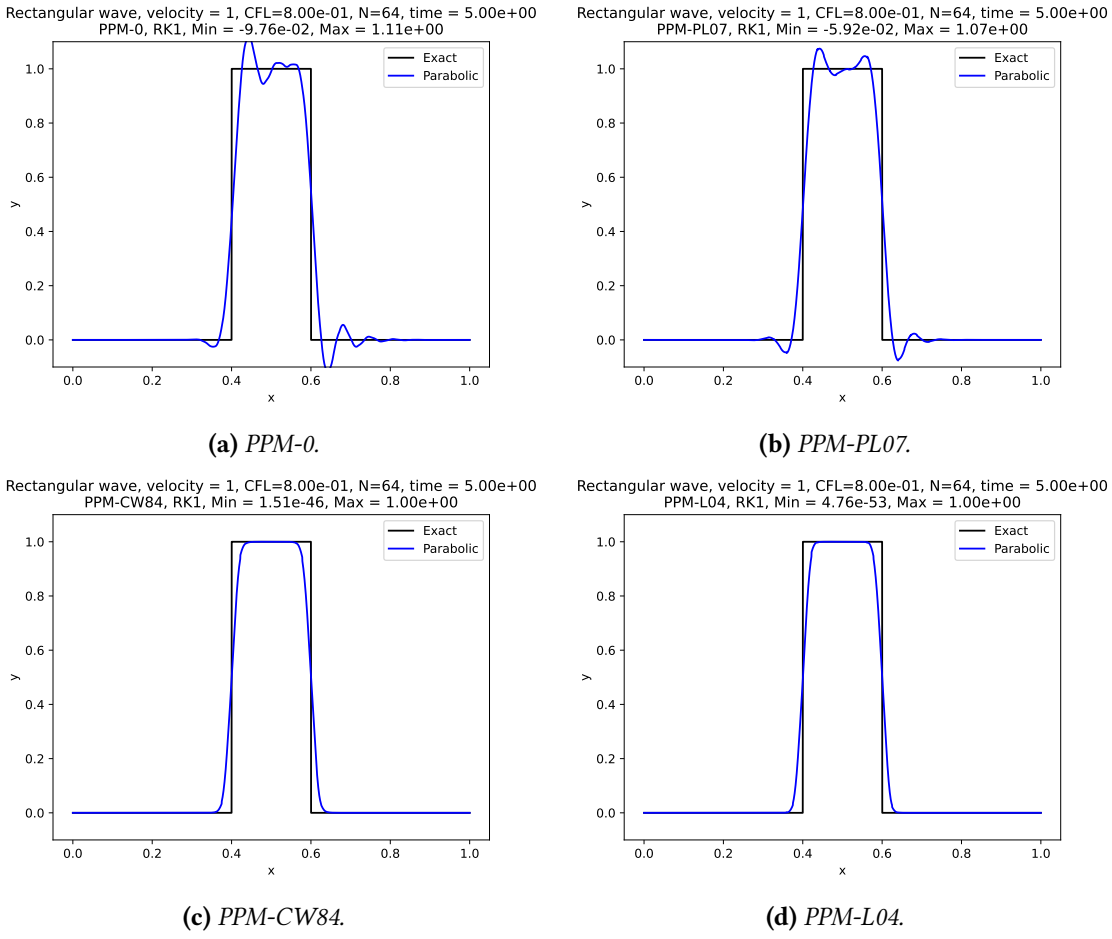
Figure 2.6: *Similar to Figure 2.4 but using* $N = 16$ *and the initial condition given by Equation* (2.103).



(a) *Error.* **(b)** *Convergence rate.*

Figure 2.7: *Similar to Figure 2.5 but using the initial condition given by Equation* (2.103).

When q_0 is given by Equation (2.104), it follows from Figure 2.4 that monotonic scheme PPM-L04 is less dissipative than PPM-CW84, and PPM-L07 is less dissipative than PPM-0, which agrees with the amplification factors from Figure 2.1. This may be noticed when we compare the extreme values. Furthermore, both monotonic schemes avoid negative values. For the Gaussian wave, we make similar conclusions from Figure 2.6. The order of convergence of the PPM-0 and PPM-PL07 are equal to three as expected (Figure 2.5) for q_0 from Equation (2.104). Notice that, even though the initial average of the Gaussian profile is computed with a second-order approximation, the final error is third-order accurate as shown in Figure 2.7. In all cases, the convergence order of the monotonic schemes is approximately 1.5 and this order reduction is expected by Godunov's theorem. The major advantage of the monotonic schemes is observed in Figure 2.8, where we observe that the monotonic schemes prevent all the strong oscillations presented in the non-monotonic schemes, and they avoid generation of new extrema as well.



(a) *PPM-0.* **(b)** *PPM-PL07.*

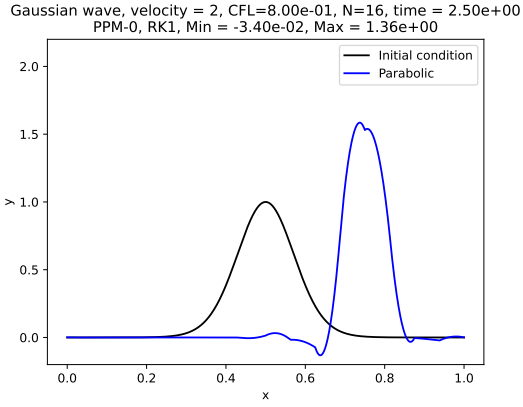
(c) *PPM-CW84.* **(d)** *PPM-L04.*

Figure 2.8: *Similar to Figure 2.4 but using* $N = 64$ *and the initial condition given by Equation* (2.104).

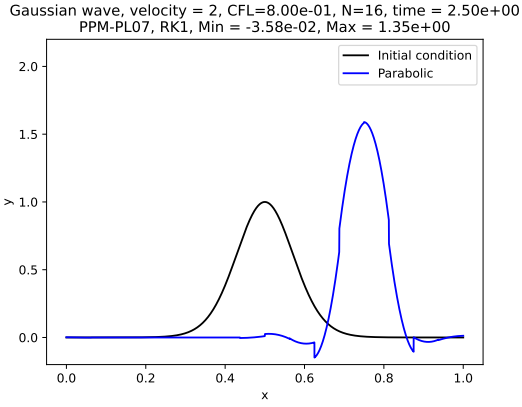
2.4.3 Linear advection equation with variable velocity simulations

In this Subsection, we shall investigate the how the PPM schemes behave when the velocity is variable. The initial condition is given by Equation (2.103). The relative errors

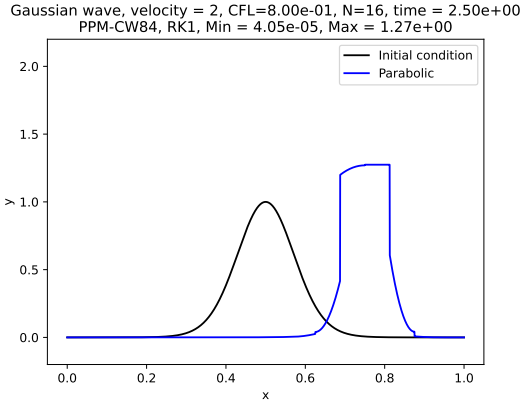
are computed using the centroid values of q as described in Section 2.4.3.



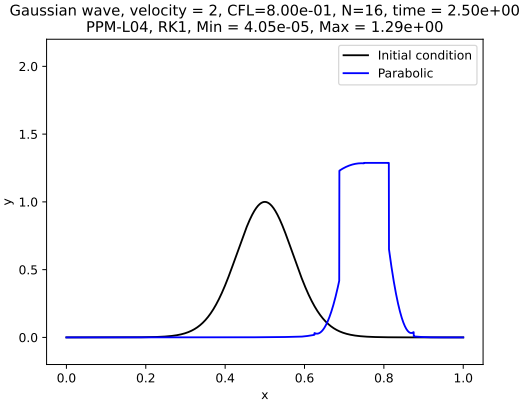
(a) *PPM-0.*



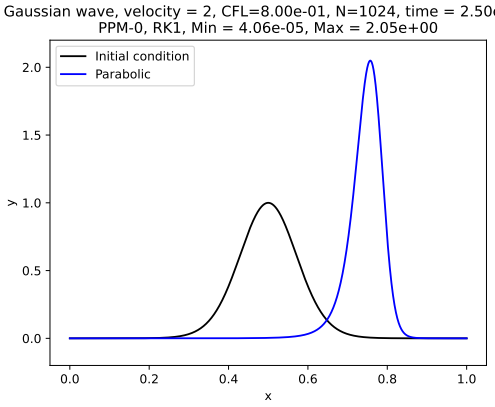
(b) *PPM-PL07.*



(c) *PPM-CW84.*



(d) *PPM-L04.*



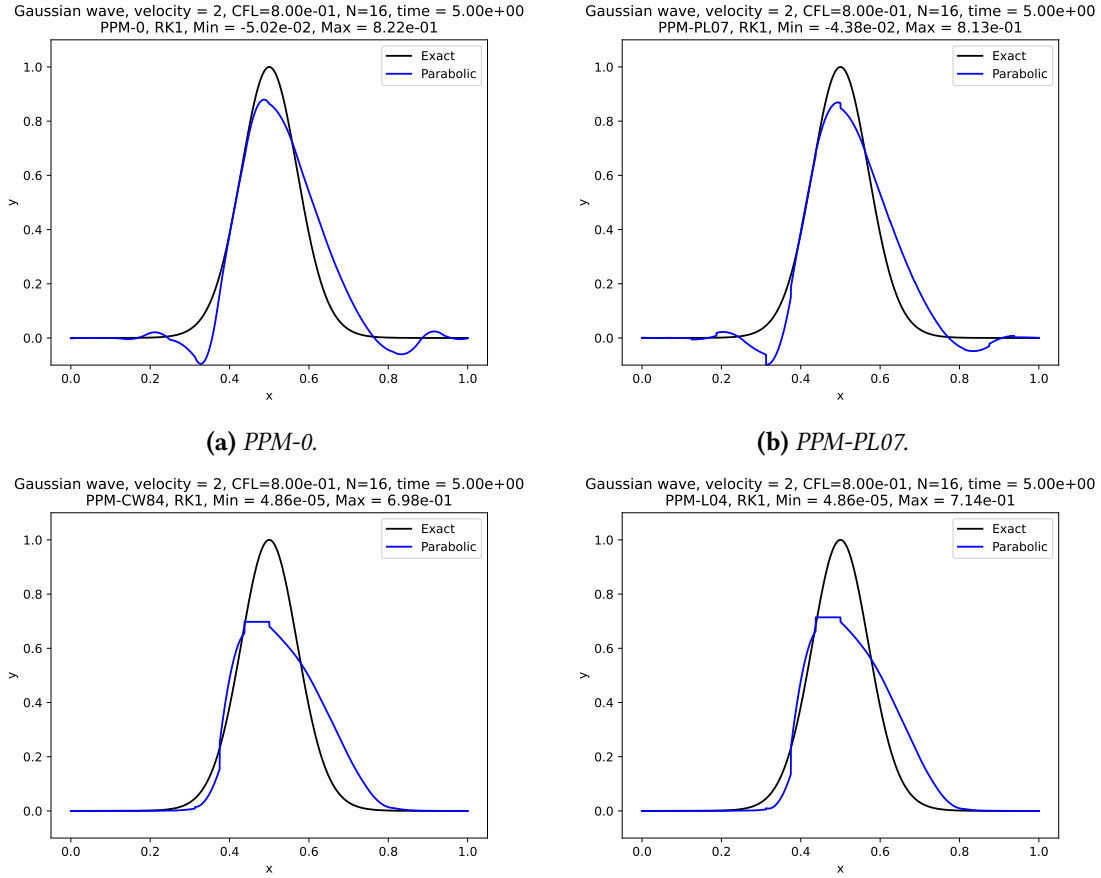
(e) *Reference solution.*

Figure 2.9: *Similar to Figure 2.4 but using $N = 16$, the initial condition given by Equation (2.103), the variable velocity given by Equation (2.105) and the final time is 2.5 (half a period). In (e) we show a reference solution, using the PPM-0 scheme with 1024 cells.*

We are going to consider the velocity

$$u(x,t) = u_0 \cos\left(\frac{\pi t}{T}\right) \sin^2(\pi x). \tag{2.105}$$

We adopt the parameters $u_0 = 0.2$ and $T = 5$. In this case, the solution has a period equal to 5, then the profile returns to its initial shape and position after 5 time units we can compute the error. We point out that the velocity from Equation (2.105) is based on the deformational flow test case from on Nair and Lauritzen (2010). Since the velocity is variable, we are going to use the departure point schemes RK1 and RK2 described in Section 2.3.3.



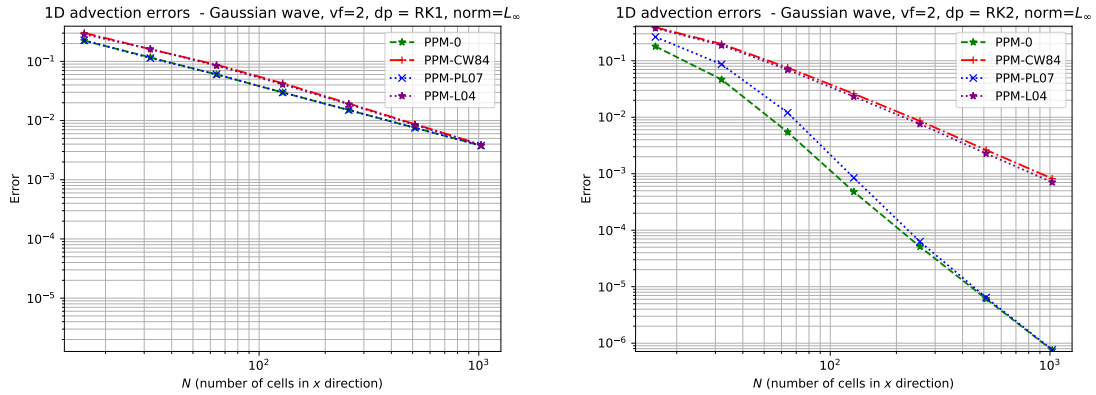
(a) *PPM-0.* **(b)** *PPM-PL07.*

(c) *PPM-CW84.* **(d)** *PPM-L04.*

Figure 2.10: *Similar to Figure 2.4 but using $N = 16$, the initial condition given by Equation (2.103) and the variable velocity given by Equation (2.105) .*

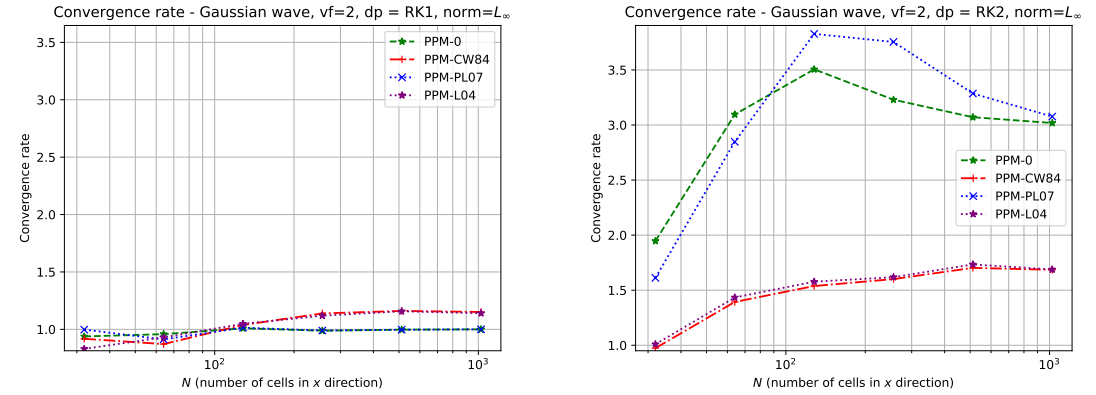
In Figure 2.9 we show the numerical after half period using the RK1 scheme and $N = 16$. We also depict a reference solution in Figure 2.9e at a high resolution ($N = 1024$). In Figure 2.10 we show the profile obtained for each PPM scheme after one period. From Figures 2.9 and 2.10 we get a similar conclusion about the dissipation of each scheme as in Section 2.4.3. The difference between the schemes RK1 and RK2 is clear when we observe the relative error in Figure 2.11 and the convergence rate in Figure 2.12. The RK1 scheme leads to a first-order in the departure point that dominates the total error for all PPM schemes,

in agreement with Proposition 2.11. When we employ the RK2 scheme, we can achieve third-order for the schemes PPM-0 and PPM-L07 which is better than the expected from Proposition 2.11. This experiment illustrates the impact of the error in the departure point calculation in the total error.



(a) *RK1.* **(b)** *RK2.*

Figure 2.11: *Relative error for different PPM schemes using the RK1 (left) and RK2 (right) departure point scheme for the initial condition given by Equation* (2.103) *and the variable velocity given by Equation* (2.105).



(a) *RK1.* **(b)** *RK2.*

Figure 2.12: *Convergence rate for different PPM schemes using the RK1 (left) and RK2 (right) departure point scheme for the initial condition given by Equation* (2.103) *and the variable velocity given by Equation* (2.105).

2.5 Concluding remarks

In this Chapter, we gave a general overview of 1D finite-volume schemes for the advection equation. We saw that 1D-FV schemes require two tasks. The first task is to reconstruct a function from its average values. The second task is to compute the departure point of the control volume edges. Each task introduces an error in the consistency error, impacting the final error, as we have shown theoretically. The first task was performed

using the PPM from Colella and Woodward (1984) and some of its variants. Without monotonicity constraints, we were able to achieve third in the reconstruction process. The second task was performed using the first-order departure point calculation from Colella and Woodward (1984) (RK1). We also explored a third-order approach using a three stages Runge-Kutta scheme (RK2) to integrate the departure point ODE.

From the numerical experiments, we observe that the PPM-L07 (Putman & Lin, 2007), which uses a fifth-order reconstruction at the edges, leads to a third-order but more accurate than the also third-order scheme PPM-0 (Colella & Woodward, 1984), which uses a fourth-order reconstruction at the edges. For the monotonic schemes, we observe that both of them were able the avoid overshoots, with the PPM-L04 (Lin, 2004) being more accurate than (Colella & Woodward, 1984).

The difference between the departure point schemes was observed when we performed a test with variable velocity, where the simulation performed with RK1 scheme led to a final first-order error, despite of third-order accuracy in the space, while the RK2 scheme preserves to third-order accuracy of the scheme, despite the RK2 scheme is only second-order accurate. We expect that, in general, the PPM combined with the RK2 scheme must be at least second-order accurate. Clearly, the RK2 leads to a more expensive scheme, since we need to compute a time extrapolation and a linear interpolation of the velocity field. A possible way to reduce its cost would be to use a Semi-Lagrangian version of the PPM when computing the numerical flux (Y. Chen et al., 2017), which allows large time steps, as we shortly explained in Section 2.3.3.

Chapter 3

Two-dimensional finite-volume methods

In Chapter 2, we tackled the problem of solving the one-dimensional linear advection equation using the finite-volume method based on PPM. In this Chapter, we are interested in solving the two-dimensional linear advection equation using the finite-volume method. This step plays a key role in this work, since as we shall see in Chapter 5, the solution of the linear advection equation on the cubed-sphere boils down to solving one two-dimensional linear advection equations at each cube face with interpolation between the adjacent panels.

A natural way to derive a finite-volume method for the two-dimensional linear advection equation would be to extend the PPM for two dimensions. Indeed, a piecewise bi-parabolic extension of PPM was proposed by Rančić (1992) using a semi-lagrangian temporal discretization. However, the major problem of this method is its quite expensive computational cost. A popular alternative, and usually computationally cheaper, is to use dimension-splitting methods. These schemes replace two-dimensional problem with a sequence of one-dimensional problems. For instance, we can solve the two-dimensional linear advection equation by solving a sequence one-dimensional linear advection equations using the PPM from Chapter 2. Further, we could, in principle, use any numerical method that solves the one-dimensional linear advection. A comparison between two-dimension and dimension splitting semi-lagrangian schemes on the plane has been investigated by Y. Chen et al. (2017) using the PPM as the one-dimensional solver and distorted two-dimensional grids. Their major conclusion is that the dimension splitting schemes are more sensitive to grid distortions but it is computationally cheaper and more accurate than their two-dimensional methods, specially considering large CFL numbers.

The main aim of this chapter is to give a detailed description of the dimension splitting method proposed by Lin and Rood (1996), which is currently employed in the FV3 dynamical core, applied to the two-dimensional linear advection equation using the one-dimensional finite-volume schemes from Chapter 2. Similarly to Chapter 2, we start this Chapter start with a review of the two-dimensional advection equation in the integral form in Section 3.1, and in Section 3.2 we set the framework of general two-dimensional finite-volumes schemes. Section 3.3 presents the dimension splitting method and numerical

experiments are shown in Section 3.4.

3.1 Two-dimensional advection equation in integral form

Let us consider a C^1 velocity field given by $\boldsymbol{u} : \mathbb{R}^2 \to \mathbb{R}^2$, $\boldsymbol{u} = (u, v)$, where u is the velocity in x-direction and v is the velocity in x and y direction. The two-dimensional advection equation in the differential form in a domain $\Omega = [a, b] \times [c, d] \subset \mathbb{R}^2$ associated to the velocity field $\boldsymbol{u}$ is given by:

$$\frac{\partial q}{\partial t}(x, y, t) + \nabla \cdot (q\boldsymbol{u})(x, y, t) = 0, \quad \forall (x, y, t) \in \Omega^{o} \times]0, +\infty[, \tag{3.1}$$ [1]

where $\nabla \cdot (q\boldsymbol{u})$ is the spatial divergence

$$\nabla \cdot (q\boldsymbol{u})(x, y, t) = \frac{\partial (uq)}{\partial x}(x, y, t) + \frac{\partial (vq)}{\partial y}(x, y, t). \tag{3.2}$$

We recall that we say the u is **non-divergent** if $\nabla \cdot \boldsymbol{u} = 0$. A classical or strong solution to the two-dimensional advection equation is a C^1 function q satisfying Equation (3.1). As we did in Section 2.1, our goal is to deduce an integral form of Equation (3.1). Thus, let us consider $[x_1, x_2] \times [y_1, y_2] \subset \Omega^{o}$ and $[t_1, t_2] \subset [0, +\infty[$. Integrating Equation (3.1) over $[x_1, x_2] \times [y_1, y_2]$ yields:

$$\begin{aligned} \frac{d}{dt}\left(\int_{x_1}^{x_2} \int_{y_1}^{y_2} q(x, y, t)\, dx\, dy \right) = &- \int_{y_1}^{y_2} \Big((uq)(x_2, y, t) - (uq)(x_1, y, t) \Big)\, dy \\ &- \int_{x_1}^{x_2} \Big((vq)(x, y_2, t) - (vq)(x, y_1, t) \Big)\, dx. \end{aligned} \tag{3.3}$$

Integrating Equation (3.3) over the time interval $[t_1, t_2]$, we have:

$$\begin{aligned} \int_{x_1}^{x_2} \int_{y_1}^{y_2} q(x, y, t_{n+1})\, dx\, dy = &\int_{x_1}^{x_2} \int_{y_1}^{y_2} q(x, y, t_n)\, dx\, dy \\ &- \int_{t_1}^{t_2} \int_{y_1}^{y_2} \Big((uq)(x_2, y, t) - (uq)(x_1, y, t) \Big)\, dy\, dt \\ &- \int_{t_1}^{t_2} \int_{x_1}^{x_2} \Big((vq)(x, y_2, t) - (vq)(x, y_1, t) \Big)\, dx\, dt. \end{aligned} \tag{3.4}$$

Equation (3.4) is the integral form of Equation (3.1). We say that q is a weak solution to the advection equation (3.1) if q satisfies the integral form (3.4), $\forall [x_1, x_2] \times [y_1, y_2] \subset \Omega^{o}$ and $\forall [t_1, t_2] \subset [0, +\infty[$. Similarly to Section 2.1, these problems are equivalent when q is a C^1 function. We consider an initial condition q_0, $q(x, y, 0) = q_0(x, y)$, $\forall (x, y) \in \Omega$. Boundary conditions will be assumed bi-periodic. Therefore, we are again dealing with a Cauchy problem.

[1] Ω^{o} denotes the interior of Ω. Namely, $\Omega^{o} =]a, b[\times]c, d[$.

To move in the direction of a discrete version of Equation (3.4), let us discretize the domain $D = \Omega \times [0,T]$ following the notations of Section 2.1. Given a positive integer N_T, we define the time step $\Delta t = \frac{T}{N_T}$, $t_n = n\Delta t$, for $n = 0, 1, \cdots, N_T$. The spatial discretization is constructed through an uniformly spaced partition of Ω given by:

$$[a,b] = \bigcup_{i=1}^{N} X_i, \text{ where } X_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \text{ and } a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N-\frac{1}{2}} < x_{N+\frac{1}{2}} = b, \tag{3.5}$$

$$[c,d] = \bigcup_{j=1}^{M} Y_j, \text{ where } Y_j = [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}] \text{ and } c = y_{\frac{1}{2}} < y_{\frac{3}{2}} < \cdots < y_{M-\frac{1}{2}} < y_{M+\frac{1}{2}} = d, \tag{3.6}$$

$$\Omega = \bigcup_{i=1}^{N}\bigcup_{j=1}^{M} \Omega_{ij}, \text{ where } \Omega_{ij} = X_i \times Y_j. \tag{3.7}$$

The regions Ω_{ij} are known as control volumes or cells. Similarly to Chapter 2 we employ the notations $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, $\Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$ and $x_i = \frac{1}{2}(x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}})$ $y_j = \frac{1}{2}(y_{j+\frac{1}{2}} + y_{j-\frac{1}{2}})$, $\forall i = 1, \cdots, N, \forall j = 1, \cdots, M$, to define the control volume lengths and centroids, respectively. Finally, we denote by $Q_{ij}(t)$ as the average values of state variable at time t in the control volume Ω_{ij}, that is:

$$Q_{ij}(t) = \frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j-\frac{1}{2}}} q(x,y,t)\, dx. \tag{3.8}$$

Substituting t_1, t_2, x_1, x_2, y_1 and y_2 by $t_n, t_{n+1}, x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}, y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}$, respectively, in Equation (3.4), we obtain:

$$\begin{aligned} Q_{ij}(t_{n+1}) = Q_{ij}(t_n) &- \frac{\Delta t}{\Delta x \Delta y}\delta_x \left(\frac{1}{\Delta t} \int_{t_1}^{t_2} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (uq)(x_i, y, t)\, dy\, dt \right) \\ &- \frac{\Delta t}{\Delta x \Delta y}\delta_y \left(\frac{1}{\Delta t} \int_{t_1}^{t_2} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (vq)(x, y_j, t)\, dx\, dt \right), \end{aligned} \tag{3.9}$$

where we are using the centered finite-difference notation:

$$\delta_x h(x_i, y, t) = h(x_{i+\frac{1}{2}}, y, t) - h(x_{i-\frac{1}{2}}, y, t), \tag{3.10}$$

$$\delta_y h(x, y_j, t) = h(x, y_{j+\frac{1}{2}}, t) - h(x, y_{j-\frac{1}{2}}, t), \tag{3.11}$$

for any function h. The Equation (3.9) is useful to motivate two-dimensional finite-volume schemes, as we shall see in the next section.

3.2 The finite-volume approach

This Section is basically an extension to two dimensions of the concepts presented in Section 2.2. We introduce the following spaces of bi-periodic functions:

$$\mathcal{F}(\mathbb{T}^2) = \{q : \mathbb{R}^2 \to \mathbb{R}; \quad q(x+b-a, y) = q(x,y), \quad q(x, y+d-c) = q(x,y), \quad \forall (x,y) \in \mathbb{R}^2\},$$
$$\mathcal{F}(\mathbb{T}_T^2) = \{q : \mathbb{R}^2 \times [0,T] \to \mathbb{R}; \quad q(\cdot,\cdot,t) \in \mathcal{F}(\mathbb{T}^2), \quad \forall t \in [0,T]\},$$
$$C^k(\mathbb{T}_T^2) = \{q \in C^k(\mathbb{R}^2 \times [0,T]) : q \in \mathcal{F}(\mathbb{T}_T^2)\}.$$

where we are using the notation $\mathbb{T}^2_T = \mathbb{T}^2 \times [0,T]$ and $\mathbb{T}^2 = [a,b] \times [c,d]$ denotes the torus of lengths $b-a$ and $d-c$. We are using the notation $\mathbb{T}^2$ since we may think of bi-periodic functions as functions defined on the torus of lengths $b-a$ and $d-c$. Whenever we use the notation $\mathbb{T}^2$, a, b, c and d will be implicitly defined. We also introduce the following the locally integrable periodic functions:

$$
\begin{aligned}
L^p_{\text{loc}}(\Omega) &= \{q : \Omega \to \mathbb{R}; \quad \int_K |q(x,y)|^p \, dx\, dy < +\infty, \quad \text{for all compact sets } K \subset \Omega\},\\
L^p_{\text{loc}}(\mathbb{T}^2) &= \{q \in \mathcal{F}(\mathbb{T}^2) : \quad q \in L^p_{\text{loc}}(\mathbb{R}^2)\},\\
L^{p,x,y}_{\text{loc}}(\mathbb{T}^2_T) &= \{q \in \mathcal{F}(\mathbb{T}^2_T) : \forall t \in [0,T], \quad q(\cdot,\cdot,t) \in L^p_{\text{loc}}(\mathbb{R}^2)\},\\
L^{p,y,t}_{\text{loc}}(\mathbb{T}^2_T) &= \{q \in \mathcal{F}(\mathbb{T}^2_T) : \forall x \in \mathbb{R}, \quad q(x,\cdot,\cdot) \in L^p_{\text{loc}}(\mathbb{R} \times [0,T])\},\\
L^{p,x,t}_{\text{loc}}(\mathbb{T}^2_T) &= \{q \in \mathcal{F}(\mathbb{T}^2_T) : \forall y \in \mathbb{R}, \quad q(\cdot,y,\cdot) \in L^p_{\text{loc}}(\mathbb{R} \times [0,T])\}.
\end{aligned}
$$

3.2.1 Discretization of the problem

The problem of two-dimensional advection equation in the integral form presented Section 3.1 is written in a concise way in Problem 3.1.

Problem 3.1. *Given an initial condition $q_0 \in L^1_{loc}(\mathbb{T}^2) \cap L^2_{loc}(\mathbb{T}^2)$ and a velocity function $u \in L^{2,y,t}_{loc}(\mathbb{T}^2_T)$ in the x-direction, a velocity function $v \in L^{2,x,t}_{loc}(\mathbb{T}^2_T)$ in the y-direction, we would like to find a weak solution $q \in L^{1,x,y}_{loc}(\mathbb{T}^1_T) \cap L^{2,x,t}_{loc}(\mathbb{T}^2_T) \cap L^{2,y,t}_{loc}(\mathbb{T}^2_T)$ of the two-dimensional advection equation in the integral form:*

$$
\begin{aligned}
\int_{x_1}^{x_2}\int_{y_1}^{y_2} q(x,y,t)\,dx\,dy = &\int_{x_1}^{x_2}\int_{y_1}^{y_2} q(x,y,t)\,dx\,dy\\
&- \int_{t_1}^{t_2}\int_{y_1}^{y_2} \Big((uq)(x_2,y,t) - (uq)(x_1,y,t) \Big)\,dy\,dt\\
&- \int_{t_1}^{t_2}\int_{x_1}^{x_2} \Big((vq)(x,y_2,t) - (vq)(x,y_1,t) \Big)\,dx\,dt.
\end{aligned}
$$

$\forall [x_1,x_2] \times [y_1,y_2] \times [t_1,t_2] \subset [a,b] \times [c,d] \times [0,T]$, and $q(x,y,0) = q_0(x,y)$, $\forall (x,y) \in [a,b] \times [c,d]$.

For Problem 3.1, the total mass in Ω is defined by:

$$
M_\Omega(t) = \int_\Omega q(x,y,t)\,dx\,dy, \quad \forall t \in [0,T], \tag{3.12}
$$

and is conserved within time:

$$
M_\Omega(t) = M_\Omega(0), \quad \forall t \in [0,T]. \tag{3.13}
$$

considering a discretization of the domain $[a,b] \times [0,T]$. Similar to Definitions 3.1 and 2.3, we introduce the concepts of $(\Delta x, \Delta y)$-grid and $(\Delta x, \Delta y, \Delta t, \lambda_x, \lambda_y)$ discretization.

Definition 3.1 ($(\Delta x, \Delta y)$-grid)**.** *Given $[a,b] \times [c,d]$ and positive real numbers Δx and Δy such that $\Delta x = (b-a)/N$, $\Delta y = (d-c)/M$, for positive integers N and M, we say that a*

(N, M)-tuple $\mathcal{D} = (\Omega_{ij})_{i=1,\cdots,N, j=1,\cdots,M}$ is a $(\Delta x, \Delta y)$-grid for $[a,b]\times[c,d]$ if $\Omega_{ij} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]\times[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, $a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N-\frac{1}{2}} < x_{N+\frac{1}{2}} = b$, $c = y_{\frac{1}{2}} < y_{\frac{3}{2}} < \cdots < y_{M-\frac{1}{2}} < y_{M+\frac{1}{2}} = d$, $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, $\Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$. Each Ω_{ij} is called control volume or cell.

Remark 3.1. *We may define the cells Ω_{ij} for i outside of the range $1,\cdots,N$ or j outside of the range $1,\cdots,M$ by $\Omega_{ij} = [a+(i-1)\Delta x, a+i\Delta x]\times[c+(j-1)\Delta x, c+j\Delta x]$. These cells are called ghost cells.*

Definition 3.2 ($(\Delta x, \Delta y, \Delta t, \lambda_x, \lambda_y)$-discretization)**.** *Given $[a,b]\times[c,d]\times[0,T]$ and positive real numbers Δx Δy and Δt, we say that $(\mathcal{D}, \mathcal{T})$ is a $(\Delta x, \Delta y, \Delta t, \lambda_x, \lambda_y)-$discretization of $[a,b]\times[c,d]\times[0,T]$ if Ω is a $(\Delta x, \Delta y)$-uniform grid for $[a,b]\times[c,d]$ and $\mathcal{T}$ is a Δt-temporal grid for $[0,T]$, $\frac{\Delta t}{\Delta x} = \lambda_x$ and $\frac{\Delta t}{\Delta y} = \lambda_y$.*

Remark 3.2. *Whenever we mention a $(\Delta x, \Delta y)-$grid, or a $(\Delta x, \Delta y, \Delta t, \lambda_x, \lambda_y)$-discretization, then Ω_{ij}, N and M are implicitly defined.*

Section 3.1 introduced a version of Problem 3.1 considering a discretization of the domain $[a,b]\times[c,d]\times[0,T]$. This version is also summarized in Problem 3.2.

Problem 3.2. *Assume the framework of Problem 3.1 and that $(\mathcal{D}, \mathcal{T})$ is a $(\Delta x, \Delta y, \Delta t, \lambda_x, \lambda_y)$-discretization of $[a,b]\times[c,d]\times[0,T]$. Since we are in the framework of Problem 3.1, it follows that:*

$$Q_{ij}(t_{n+1}) = Q_{ij}(t_n) - \lambda_x \delta_x \left(\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (uq)(x_i, y, t)\, dy\, dt \right) - \lambda_y \delta_y \left(\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (vq)(x, y_j, t)\, dx\, dt \right),$$

where $Q_{ij}(t) = \frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q(x,y,t)\, dx\, dy$.

Our problem now consists of finding the values $Q_{ij}(t_n)$, $\forall i = 1,\cdots,N$, $\forall j = 1,\cdots,M$, $\forall n = 1,\cdots,N_T$, given the initial values $Q_{ij}(0)$, $\forall i = 1,\cdots N$, $\forall j = 1,\cdots,M$. In other words, we would like to find the average values of q in each control volume Ω_{ij} at the considered time instants.

Next, we introduce the definitions of grid functions at cell centroids and C-grid functions.

Definition 3.3 ($(\Delta x, \Delta y)$-grid function)**.** *For a $(\Delta x, \Delta y)$-grid $\mathcal{D}$, we say that $Q \in \mathbb{R}^{N\times M}$ is a $(\Delta x, \Delta y)$-grid function, where we assume that Q is ordered by the lines in the x direction of the grid, i.e.,*

$$Q = (Q_{11}, Q_{21}\cdots, Q_{M1}, Q_{12}, Q_{22}\cdots, Q_{M2}, \cdots, Q_{1N}, Q_{2N}\cdots, Q_{NM}).$$

We denote the space of $(\Delta x, \Delta y)$-grid functions by $\mathbb{R}^{\Delta x\times\Delta y}$.

Definition 3.4 ($(\Delta x, \Delta y)$-C grid wind)**.** *For a $(\Delta x, \Delta y)$-grid $\mathcal{D}$, we say that (u, v) is a $(\Delta x, \Delta y)$-C grid wind if $u \in \mathbb{R}^{(N+1)\times M}$, $v \in \mathbb{R}^{N\times(M+1)}$, where we assume that u is ordered by the*

lines in the x direction of the grid, i.e.,

$$u = (u_{\frac{1}{2},1}, u_{\frac{3}{2},1}, \cdots, u_{N+\frac{1}{2},1}, u_{\frac{1}{2},2}, u_{\frac{3}{2},2}, \cdots, u_{N+\frac{1}{2},2}, \cdots, u_{\frac{1}{2},M}, u_{\frac{3}{2},M}, \cdots, u_{N+\frac{1}{2},M}),$$

v is ordered by the lines in the y direction of the grid, i.e.,

$$v = (v_{1,\frac{1}{2}}, v_{1,\frac{3}{2}}, \cdots, v_{1,M+\frac{1}{2}}, v_{2,\frac{1}{2}}, v_{2,\frac{3}{2}}, \cdots, v_{2,M+\frac{1}{2}}, \cdots, v_{N,\frac{1}{2}}, v_{N,\frac{3}{2}}, \cdots, v_{N,M+\frac{1}{2}}).$$

Remark 3.3. *When computing stencils, we may need values of Q_{ij} such that the indexes i and j are out of the range $1, \cdots, N$, $1, \cdots, M$. These values are called ghost cell values. Since we are under the assumption of periodic boundary conditions, this problem is overcome by assuming periodicity on the grid function Q. The same applies for $(\Delta x, \Delta y)$-C grid functions.*

Finally, we define the two-dimensional (2D) finite-volume (FV) scheme problem as follows in Problem 3.3.

Remark 3.4. *For Problem 2.2, we define the $(\Delta x, \Delta y)$-grid functions q^n and $Q(t^n)$, where $q^n_{ij} = q(x_i, y_j, t^n)$, $Q(t^n)_{ij} = Q_{ij}(t^n)$, for $n = 0, \cdots, N_T$. We also define the $(\Delta x, \Delta y)$-C grid wind (u^n, v^n) where $u^n_{i+\frac{1}{2},j} = u(x_{i+\frac{1}{2}}, y_j, t^n)$ and $v^n_{i,j+\frac{1}{2}} = v(x_i, y_{j+\frac{1}{2}}, t^n)$.*

Problem 3.3 (2D-FV scheme)**.** *Assume the framework defined in Problem 3.2. The finite-volume approach of Problem 3.1 consists of a finding a scheme of the form:*

$$Q^{n+1}_{ij} = Q^n_{ij} - \lambda_x \delta_i F^n_{i,j} - \lambda_y \delta_j G^n_{i,j}, \tag{3.14}$$
$$\forall i = 1, \cdots, N, \quad \forall j = 1, \cdots, M, \quad \forall n = 0, \cdots, N_T - 1,$$

where $\delta_i F^n_{ij} = F^n_{i+\frac{1}{2},j} - F^n_{i-\frac{1}{2},j}$, $\delta_j G^n_{ij} = G^n_{i,j+\frac{1}{2}} - G^n_{i,j-\frac{1}{2}}$ and Q^n is intended to be an approximation of $Q(t_n)$ in some sense. We define $Q^0_{ij} = Q_{ij}(0)$ or $Q^0_{ij} = q^0_{i,j}$.

The term $F^n_{i+\frac{1}{2},j} = \mathbb{F}(Q^n, \tilde{u}^n, \tilde{v}^n; i, j)$ is known as numerical flux in the x direction, where $\mathbb{F}$ is the numerical flux function, and it approximates $\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (uq)(x_{i+\frac{1}{2}}, y, t)\, dy\, dt$, $\forall i = 0, 1, \cdots, N$, and $G^n_{i,j+\frac{1}{2}} = \mathbb{G}(Q^n, \tilde{u}^n, \tilde{v}^n; i, j)$ is known as numerical flux in the y direction, where $\mathbb{G}$ is the numerical flux function, and it approximates $\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (vq)(x, y_{j+\frac{1}{2}}, t)\, dx\, dt$, $\forall j = 0, 1, \cdots, M$, or, in other words, they estimate the time-averaged fluxes at the control volume Ω_{ij} boundaries. The values $\tilde{u}^n_{i+\frac{1}{2},j}$ and $\tilde{v}^n_{i,j+\frac{1}{2}}$ are related to the time-averaged velocities and depend on values of $u^n_{i+\frac{1}{2},j}$ and $v^n_{i,j+\frac{1}{2}}$.

Remark 3.5. *A scheme of the form from Equation (3.14) is referred to as a 2D-FV scheme and it is also known as a conservative scheme.*

Remark 3.6. *For Problem 3.3, we define the CFL number in the x and y direction by $\max\{|u^n_{i+\frac{1}{2},j}|\}$ and $\max\{|v^n_{i,j+\frac{1}{2}}|\}$, respectively. The CFL number is maximum between these numbers and we say that the CFL conditions is satisfied if the CFL number is less than one.*

Definition 3.5 (Discrete divergence)**.** *For Problem 3.3, we define the discrete divergence as*

a grid function $\mathbb{D}^n = \mathbb{D}^n(Q, \tilde{u}^n, \tilde{v}^n)$ *given by*

$$\mathbb{D}^n_{ij} = \left(\frac{\mathbb{F}(Q, \tilde{u}^n, \tilde{v}^n; i, j) - \mathbb{F}(Q, \tilde{u}^n, \tilde{v}^n; i-1, j)}{\Delta x} \right) + \left(\frac{\mathbb{G}(Q, \tilde{u}^n, \tilde{v}^n; i, j) - \mathbb{G}(Q, \tilde{u}^n, \tilde{v}^n; i, j-1)}{\Delta y} \right). \tag{3.15}$$

For a 2D-FV the discrete total mass at the time-step n is given by

$$M^n = \Delta x \Delta y \sum_{i=1}^{N} \sum_{j=1}^{M} Q^n_{ij}.$$

Therefore, the discrete total mass is constant for a 2D-FV scheme, which follows from a straightforward computation:

$$\begin{aligned} M^{n+1} &= \Delta x \sum_{i=1}^{N} \sum_{j=1}^{M} Q^{n+1}_{ij} = M^n - \Delta t \sum_{i=1}^{N} \sum_{j=1}^{M} (F^n_{i+\frac{1}{2},j} - F^n_{i-\frac{1}{2},j}) - \Delta t \sum_{i=1}^{N} \sum_{j=1}^{M} (G^n_{i,j+\frac{1}{2}} - G^n_{i,j-\frac{1}{2}}) \\ &= M^n - \Delta t \sum_{j=1}^{M} (F^n_{N+\frac{1}{2},j} - F^n_{\frac{1}{2},j}) - \Delta t \sum_{i=1}^{N} (G^n_{i,M+\frac{1}{2}} - G^n_{i,\frac{1}{2}}) = M^n, \end{aligned}$$

where we are using that $F^n_{N+\frac{1}{2},j} = F^n_{\frac{1}{2},j}$, $G^n_{i,M+\frac{1}{2}} = G^n_{i,\frac{1}{2}}$ since we are assuming bi-periodic boundary conditions.

3.2.2 Convergence, consistency and stability

As we mentioned in Problem 3.3, the initial condition may be assumed as q^0_{ij} or $Q_{ij}(0)$. For two-dimensional simulations, we are going to assume q^0_{ij} as initial data to avoid the computation of integrals. Furthermore, the errors will be calculated using the values q^n_{ij} instead of $Q_{ij}(t_n)$. Similarly to Proposition 2.1, we have that the centroid value approximates the average value with second order, as Proposition 3.1 shows.

Proposition 3.1. *If* $q \in C^2$, *then* $|Q_{ij}(t^n) - q^n_{ij}| \leq C_1 \Delta x^2 + C_2 \Delta x \Delta y + C_3 \Delta y^2$, *where* C_1, C_2 *and* C_3 *are constants.*

Proof. Just apply Theorem A.5 for the function $q(x, y, t^n)$. □

The notions of convergence, consistency and stability for a 2D-FV schemes are straightforward from these notions for 1D-FV schemes (see Subsections 2.2.2 and 2.2.3). Indeed, in the context of Problem 3.3, we define the operators $\mathcal{H}_{\Delta x, \Delta y, n} : \mathbb{R}^{\Delta x \times \Delta y} \to \mathbb{R}^{\Delta x \times \Delta y}$ whose (i, j) entry is given by:

$$\begin{aligned} [\mathcal{H}_{\Delta x, \Delta y, n}(Q)]_{ij} = Q_{ij} &- \lambda_x \left(\mathbb{F}(Q, \tilde{u}^n, \tilde{v}^n; i, j) - \mathbb{F}(Q, \tilde{u}^n, \tilde{v}^n; i-1, j) \right) \\ &- \lambda_y \left(\mathbb{G}(Q, \tilde{u}^n, \tilde{v}^n; i, j) - \mathbb{G}(Q, \tilde{u}^n, \tilde{v}^n; i, j-1) \right) \end{aligned}$$

for $i = 1, \cdots, N$, $j = 1, \cdots, M$, $n = 0, \cdots, N_T - 1$. The 2D-FV is then expressed as

$$Q^{n+1} = \mathcal{H}_{\Delta x, \Delta y, n}(Q^n).$$

The local error truncation $\tau^n \in \mathbb{R}^{\Delta x \times \Delta y}$ is given by

$$Q(t^{n+1}) = \mathcal{H}_{\Delta x, \Delta y, n}(Q(t^n)) + \Delta t \tau^n.$$

The error equation is given by

$$E^{n+1} = \mathcal{H}_{\Delta x, \Delta y, n}(Q(t^n)) - \mathcal{H}_{\Delta x, \Delta y, n}(Q^n) + \Delta t \tau^n. \tag{3.16}$$

Given $r = (r_{ij})_{i=1,\cdots,N, j=1,\cdots,M} \in \mathbb{R}^{\Delta x \times \Delta y}$, we define the p-norm by

$$\|r\|_{p,\Delta x \times \Delta y} = \begin{cases} \left(\sum_{i=1}^{N} \sum_{j=1}^{M} |r_{ij}|^p \right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty, \\ \max_{i=1,\cdots,N, j=1,\cdots,M} |r_{ij}| & \text{otherwise .} \end{cases} \tag{3.17}$$

The stability in the p-norm is defined as in the 1D case.

Definition 3.6. *A 2D-FV scheme is stable in the $p-$norm if*

$$\|\mathcal{H}_{\Delta x, \Delta y, n}(Q) - \mathcal{H}_{\Delta x, \Delta y, n}(P)\|_{p,\Delta x \times \Delta y} \leq (1 + \alpha \Delta t) \|Q - P\|_{p,\Delta x \times \Delta y}, \tag{3.18}$$

for all $Q, P \in \mathbb{R}^{\Delta x \times \Delta y}$ and α is a constant that does not depend neither on Δx, Δy, Δt nor on n.

If a 2D-FV scheme is stable in the $p-$norm, similarly to Equation (2.31) we have:

$$\|E^{n+1}\|_{p,\Delta x \times \Delta y} \leq e^{\alpha T} (\|E^0\|_{p,\Delta x \times \Delta y} + T \max_{n=1,\cdots,N_T} \|\tau^n\|_{p,\Delta x \times \Delta y}).$$

Again, we point out that from Proposition 3.1, we have that the initial error E^0 shall be second-order accurate. Consistency is defined as in Definition 2.8 and convergence is defined as in Definition 2.9.

The Von Neumann analysis can be applied when $\mathcal{H}_{\Delta x, \Delta y, n}$ is linear, since we are considering periodic boundary conditions. The idea is the same as in the one-dimensional case, we just apply the operator $\mathcal{H}_{\Delta x, \Delta y, n}$ on the Fourier modes to obtain the amplification factor. We introduce the nodes $\theta_i = i\frac{2\pi}{N}$, $i = 1, \cdots, N$, $\Delta\theta = \frac{2\pi}{N}$, $\theta_i = (\theta_1, \theta_2, \cdots, \theta_N)$, $\phi_j = j\frac{2\pi}{M}$, $j = 1, \cdots, M$, $\Delta\phi = \frac{2\pi}{M}$, $\phi = (\phi_1, \phi_2, \cdots, \phi_M)$. For $k_1 = 1, \cdots, N$, $k_2 = 1, \cdots, M$, the two-dimensional Fourier mode $\boldsymbol{k} = (k_1, k_2)$ from $\mathbb{C}^{N \times M}$ has its (i, j) entry given by $[e^{\imath \boldsymbol{k} \theta}]_{ij} = e^{\imath k_1 \theta_i} e^{\imath k_2 \phi_j}$.

Notice that if $q, u, v \in C^3$, we can rewrite Equation (2.24) as:

$$\tau_{ij}^n = \left[\frac{1}{\Delta x \Delta y \Delta t} \int_{t^n}^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \nabla \cdot (\boldsymbol{u} q)(x, y, t) \, dy \, dx \, dt - \left(\frac{\mathbb{F}(Q, \tilde{u}^n, \tilde{v}^n; i, j) - \mathbb{F}(Q, \tilde{u}^n, \tilde{v}^n; i-1, j)}{\Delta x} \right) \right.$$
$$\left. - \left(\frac{\mathbb{G}(Q, \tilde{u}^n, \tilde{v}^n; i, j) - \mathbb{G}(Q, \tilde{u}^n, \tilde{v}^n; i, j-1)}{\Delta y} \right) \right].$$

Using the midpoint rule for integration (Theorem A.5), the mean value theorem for integrals (Theorem A.2) and recalling the discrete divergence (Definition 3.5), we have:

$$\tau_{ij}^n = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \nabla\cdot(\boldsymbol{u}q)(x_i, y_j, t)\,dt - \mathbb{D}_{ij}^n + O(\Delta x^2) + O(\Delta y^2). \tag{3.19}$$

Therefore, in order to investigate the consistency, we may compare how well the discrete divergence approximates the divergence.

3.3 Dimension splitting

Before introducing the dimension splitting scheme from Lin and Rood (1996), it is useful to investigate a little bit of general operator splitting schemes, since the dimension splitting technique is a particular case of operator splitting methods. For a time interval $[0, T]$, we consider a $\Delta t-$temporal grid. We consider the abstract Cauchy problems

$$\begin{cases} \frac{dq}{dt}(t) &= Aq(t), \quad t \in [t^n, t^{n+1}], \\ q(t^n) &= q_n, \end{cases}$$

for $n = 0, \cdots, N_T - 1$, where $q(t) \in \mathcal{B}$ for some Banach space $\mathcal{B}$, and $A : \mathcal{B} \to \mathcal{B}$ is a linear operator following the framework of Richtmyer and Morton (1968, Chapter 3). We are interested in finding $q(t^{n+1})$ given q_n. Assuming that $A = A_1 + A_2$ for two linear operators $A_1, A_2 : \mathcal{B} \to \mathcal{B}$, we consider the following abstract Cauchy sub-problems:

$$\begin{cases} \frac{dq^1}{dt}(t) &= A_1 q(t), \quad t \in [t^n, t^{n+1}], \\ q^1(t^n) &= q_n, \end{cases}$$

and

$$\begin{cases} \frac{dq^{21}}{dt}(t) &= A_2 q(t), \quad t \in [t^n, t^{n+1}], \\ q^{21}(t^n) &= q^1(t^{n+1}). \end{cases}$$

Then we can approximate $q(t_0 + \Delta t)$ by $q^{21}(t^n + \Delta t)$ with error $O(\Delta t)$, if A_1 and A_2 do not commute; otherwise, this method is exact. This approach is known as Lie-Trotter splitting. Observe the Lie-Trotter splitting may be performed in a reverse order solving the sub-problems

$$\begin{cases} \frac{dq^2}{dt}(t) &= A_2 q(t), \quad t \in [t^n, t^{n+1}], \\ q^2(t^n) &= q_n, \end{cases}$$

and

$$\begin{cases} \frac{dq^{21}}{dt}(t) &= A_1 q(t), \quad t \in [t^n, t^{n+1}], \\ q^{12}(t^n) &= q^1(t^{n+1}), \end{cases}$$

and again we estimate $q(t^{n+1})$ by $q^{12}(t^{n+1})$ with error $O(\Delta t)$. As observed by Strang (1968), we can consider

$$q^*(t^{n+1}) = \frac{q^{21}(t^{n+1}) + q^{12}(t^{n+1})}{2}, \tag{3.20}$$

which is a second-order ($O(\Delta t^2)$) symmetric scheme to approximate $q(t^{n+1})$. We shall refer to this scheme as average Lie-Trotter splitting. This process of averaging two Lie-Trotter splitting is a particular case of methods known in the literature as weighted sequential splitting methods. Furthermore, this process of averaging schemes may be extended to achieve higher-order schemes (c.f., e.g. Jia and Li (2011)). For an accuracy analysis of weighted sequential splitting methods we refer to Csomós et al. (2005).

We point out that one of the most widely used in the literature second-order splitting schemes is the Strang splitting (Strang, 1968). This scheme requires the solution of 3 sub-problems per time-step, with one of them at time $t_n + \frac{\Delta t}{2}$, while the average Lie-Trotter splitting requires the solution of 4 sub-problems per time-step. Hence, the Strang splitting is computationally cheaper. However, as we shall see in this Chapter, the average Lie-Trotter splitting applied for the linear advection equation allows a modification that eliminates a splitting error that appears when we consider a constant scalar field and non-divergent velocity as observed by Lin and Rood (1996).

To move towards to the scheme from Lin and Rood (1996), let us consider Problem 3.1 in its differential form:

$$\begin{cases} \frac{\partial q}{\partial t}(x,y,t) + \frac{\partial (uq)}{\partial x}(x,y,t) + \frac{\partial (vq)}{\partial y}(x,y,t) = 0, \quad \forall (x,y,t) \in \mathbb{T}_T^2 \\ q(x,y,0) = q_0(x,y), \end{cases}$$

where $q, u, v \in C^1(\mathbb{T}_T^2)$. We are going to consider the one-dimensional advection equation in the x-direction

$$\frac{\partial q^x}{\partial t}(x,y,t) + \frac{\partial (uq^x)}{\partial x}(x,y,t) = 0,$$

for each $y = y_j$, and the one-dimensional advection equation in the y-direction

$$\frac{\partial q^y}{\partial t}(x,y,t) + \frac{\partial (vq^y)}{\partial y}(x,y,t) = 0,$$

for each $x = x_i$. We shall assume that these problems are solved using a 1D-FV scheme as in Problem 2.3 with numerical flux functions $\mathcal{F}$ and $\mathcal{G}$, respectively. We introduce the auxiliary functions

$$\mathbf{F}(Q_{ij}^n) = -\lambda_x \left(\mathcal{F}(Q_{ij}^n(\mathcal{S}_{i+\frac{1}{2}}); \tilde{u}_{i+\frac{1}{2},j}^n) - \mathcal{F}(Q_{ij}^n(\mathcal{S}_{i-\frac{1}{2}}); \tilde{u}_{i-\frac{1}{2},j}^n) \right),$$

and

$$\mathbf{G}(Q_{ij}^n) = -\lambda_y \left(\mathcal{G}(Q_{ij}^n(\mathcal{S}_{j+\frac{1}{2}}); \tilde{v}_{i,j+\frac{1}{2}}^n) - \mathcal{G}(Q_{ij}^n(\mathcal{S}_{j-\frac{1}{2}}); \tilde{v}_{i,j-\frac{1}{2}}^n) \right),$$

which are the numerical flux update of the 1D-FV schemes in the x and y direction,

respectively, that is

$$\mathcal{F}(Q_{ij}^n(\mathcal{S}_{i+\frac{1}{2}});\tilde{u}_{i+\frac{1}{2},j}^n) = \frac{1}{\Delta t}\int_{t_n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, y_j, t)\,dt + O(\Delta t^P),$$
$$\mathcal{G}(Q_{ij}^n(\mathcal{S}_{j+\frac{1}{2}});\tilde{v}_{i,j+\frac{1}{2}}^n) = \frac{1}{\Delta t}\int_{t_n}^{t^{n+1}} (vq)(x_i, y_{j+\frac{1}{2}}, t)\,dt + O(\Delta t^P),$$

where P is the order of accuracy of the flux and $\tilde{u}^n$ and $\tilde{v}^n$ are the time-averaged velocities. The Lie-Trotter splitting is obtained by solving the advection in the x direction

$$Q_{ij}^{x,n+1} = Q_{ij}^n + \mathbf{F}(Q_{ij}^n),$$

for $j = 1, \cdots, M$, and then we advect in the y direction with initial data $Q_{ij}^{x,n+1}$

$$Q_{ij}^{yx,n+1} = Q_{ij}^{x,n+1} + \mathbf{G}(Q_{ij}^{x,n+1}),$$

for $i = 1, \cdots, N$. To get the average Lie-Trotter splitting we repeat the process in the reverse order by solving the advection equation in the y direction

$$Q_{ij}^{y,n+1} = Q_{ij}^n + \mathbf{G}(Q_{ij}^n),$$

for $i = 1, \cdots, N$, and then we advect in the x-direction with initial data $Q_{ij}^{y,n+1}$

$$Q_{ij}^{xy,n+1} = Q_{ij}^{y,n+1} + \mathbf{F}(Q_{ij}^{y,n+1}),$$

for $j = 1, \cdots, M$, and thus we have the average Lie-Trotter solution:

$$Q_{ij}^{n+1} = \frac{(Q^{xy,n+1} + Q^{yx,n+1})}{2} = Q_{ij}^n + \frac{1}{2}\mathbf{F}(Q_{ij}^n) + \frac{1}{2}\mathbf{G}(Q_{ij}^n) + \frac{1}{2}\mathbf{F}\left(Q_{ij}^n + \frac{1}{2}\mathbf{G}(Q_{ij}^n)\right) + \frac{1}{2}\mathbf{G}\left(Q_{ij}^n + \frac{1}{2}\mathbf{F}(Q_{ij}^n)\right),$$

assuming that the numerical flux functions are linear in the input Q, we may rewrite a computationally cheaper version of the average Lie-Trotter splitting as (Lin & Rood, 1996):

$$Q_{ij}^{n+1} = \frac{(Q^{xy,n+1} + Q^{yx,n+1})}{2} = Q_{ij}^n + \mathbf{F}\left(Q_{ij}^n + \frac{1}{2}\mathbf{G}(Q_{ij}^n)\right) + \mathbf{G}\left(Q_{ij}^n + \frac{1}{2}\mathbf{F}(Q_{ij}^n)\right). \tag{3.21}$$

The numerical flux functions defined in Chapter 2 are indeed linear the input Q if there are monotonic constrain, as we mention in Chapter 2, but we are going to consider this scheme even when there are monotonic constraints since it requires fewer operations. For a while let us assume that the numerical flux functions are exactly the time-averaged fluxes, which they must approximate, as pointed out in Problem 2.3. Under this assumption, we have:

$$\mathbf{F}(Q_{ij}^n) = -\lambda_x \delta_x \left(\frac{1}{\Delta t}\int_{t_n}^{t^{n+1}} (uq)(x_i, y_j, t)\,dt\right),$$
$$\mathbf{G}(Q_{ij}^n) = -\lambda_y \delta_y \left(\frac{1}{\Delta t}\int_{t_n}^{t^{n+1}} (vq)(x_i, y_j, t)\,dt\right).$$

Further, if we if assume that $q = \overline{q}$ is constant and $\nabla \cdot \boldsymbol{u} = 0$ then the solution remains constant and then, assuming also that $\boldsymbol{u}$ does not depend on t, then $\mathbf{F}$ and $\mathbf{G}$ are given by

$$\mathbf{F}(Q_{ij}^n) = -\overline{q}\lambda_x \delta_x u(x_i, y_j),$$
$$\mathbf{G}(Q_{ij}^n) = -\overline{q}\lambda_y \delta_y v(x_i, y_j).$$

However, if we compute the updated solution using Equation (3.21), we have that the error is given by

$$Q_{ij}^{n+1} - \overline{q} = -\Delta t\left(\frac{\delta_x u(x_i, y_j)}{\Delta x} + \frac{\delta_y v(x_i, y_j)}{\Delta y}\right) - \Delta t^2 \overline{q}\left(\frac{\delta_y v \delta_x u(x_i, y_j) + \delta_x u \delta_y v(x_i, y_j)}{2\Delta x \Delta y}\right) \tag{3.22}$$

$$= \Delta t(O(\Delta x^2) + O(\Delta y^2)) - \Delta t^2 \overline{q}\left(\frac{\delta_y v \delta_x u(x_i, y_j) + \delta_x u \delta_y v(x_i, y_j)}{2\Delta x \Delta y}\right). \tag{3.23}$$

Thus, the terms in the equation above multiplied by Δt^2 are related to an splitting error, even if we consider the exact fluxes. Aiming to eliminate the error from, Lin and Rood (1996) proposes to consider a modification of the average Lie-Trotter splitting as

$$Q_{ij}^{n+1} = Q_{ij}^n + \mathbf{F}\left(Q_{ij}^n + \frac{1}{2}\mathbf{g}(Q_{ij}^n)\right) + \mathbf{G}\left(Q_{ij}^n + \frac{1}{2}\mathbf{f}(Q_{ij}^n)\right), \tag{3.24}$$

where $\mathbf{f}$ and $\mathbf{g}$ are called inner advective operators and approximate $-\Delta t u \frac{\partial q}{\partial x}$ and $-\Delta t v \frac{\partial q}{\partial y}$.

In this work, we shall consider the following inner advective operator proposed by Lin (2004) (hereafter, **L04**) and the one proposed by Putman and Lin (2007) (hereafter, **PL07**). The PL07 scheme is currently used in the FV3 dynamical core. We also shall consider the average Lie-Trotter splitting (hereafter, **AVLT**). All the expressions of each inner advective operator mentioned are shown in Table 3.1. It is easy to see that both operators L04 and PL07 eliminate the term multiplied by Δt^2 that appeared in Equation (3.22) when we apply these operators for a constant grid function Q^n and a non-divergent velocity field in Equation (3.24). Therefore, these inner advective operators eliminate the splitting error for a constant field and a non-divergent velocity field.

Scheme	$\mathbf{f}(Q_{ij}^n)$	$\mathbf{g}(Q_{ij}^n)$
AVLT	$\mathbf{F}(Q_{ij}^n)$	$\mathbf{G}(Q_{ij}^n)$
L04	$\mathbf{F}(Q_{ij}^n) + Q_{ij}^n \frac{\Delta t}{\Delta x}(\tilde{u}_{i+\frac{1}{2},j}^n - \tilde{u}_{i-\frac{1}{2},j}^n)$	$\mathbf{G}(Q_{ij}^n) + Q_{ij}^n \frac{\Delta t}{\Delta y}(\tilde{v}_{i,j+\frac{1}{2}}^n - \tilde{v}_{i,j-\frac{1}{2}}^n)$
PL07	$\frac{1}{2}\left(-Q_{ij}^n + \frac{Q_{ij}^n + \mathbf{F}(Q_{ij}^n)}{1 - \frac{\Delta t}{\Delta x}\left(\tilde{u}_{i+\frac{1}{2},j}^n - \tilde{u}_{i-\frac{1}{2},j}^n\right)}\right)$	$\frac{1}{2}\left(-Q_{ij}^n + \frac{Q_{ij}^n + \mathbf{G}(Q_{ij}^n)}{1 - \frac{\Delta t}{\Delta y}\left(\tilde{v}_{i,j+\frac{1}{2}}^n - \tilde{v}_{i,j-\frac{1}{2}}^n\right)}\right)$

Table 3.1: *Expression of the inner advective operators considered in this work. AVLT stands for the average Lie-Trotter scheme, while L04 and PL07 stands for the inner advective operators from Lin (2004) and from Putman and Lin (2007), respectively.*

Recalling the definition of discrete divergence (Definition 3.5) we have:

$$\mathbb{D}_{ij}^{n} = -\frac{1}{\Delta t}\left[\mathbf{F}\left(Q_{ij}^{n} + \frac{1}{2}\mathbf{g}(Q_{ij}^{n})\right) + \mathbf{G}\left(Q_{ij}^{n} + \frac{1}{2}\mathbf{f}(Q_{ij}^{n})\right)\right], \tag{3.25}$$

and as pointed out in Section 3.2.2, we may use the discrete divergence to check the scheme consistency.

3.4 Numerical experiments

To assess the dimension splitting schemes introduce previously, we are going to consider the linear advection equation on $[0,1]\times[0,1]$ with bi-biperiodic boundary conditions. For the 1D schemes, we consider the reconstruction schemes PPM-0, PPM-PL07, PPM-CW84, PPM-L04 and the departure point schemes as in Section 2.4. For all simulations, the CFL number equal to 0.8 and the time integration interval is $[0,5]$, which represents the period of the exact solution for all tests. We employ $(\Delta x^{(k)}, \Delta y^{(k)})-$grids with $\Delta x^{(k)} = \Delta y^{(k)} = 1/2^k$ for $k = 4,\cdots,10$. We introduce the relative error in the maximum norm:

$$E_k = \max_{n=0,\cdots,N_T} \frac{\|Q^n - q(t^n)\|_{\infty,\Delta x\times\Delta y}}{\|q(t^n)\|_{\infty,\Delta x\times\Delta y}},$$

and the convergence rate is defined as in Section 2.4.2, as well the total mass variation, which in all experiments presented here is preserved with machine precision. Notice that in the error computation we are going to use the centroid values instead of the exact average values to avoid computations of analytical integrals. As we pointed in Proposition 3.1, this approximation introduces a second-order error.

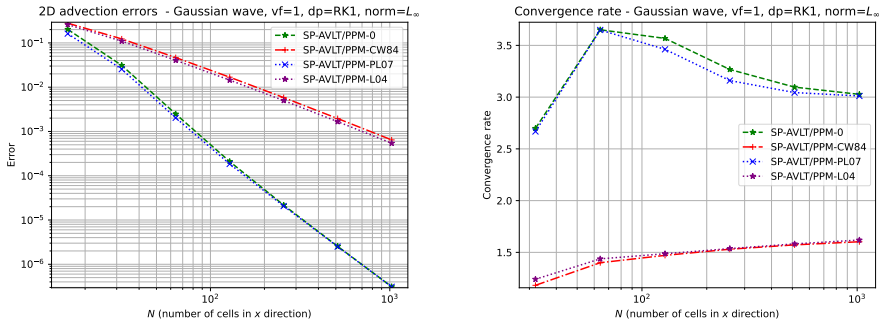
3.4.1 Linear advection equation with constant velocity simulations

For a constant velocity we shall adopt the $\boldsymbol{u} = (0.2, -0.2)$. We are going to consider as initial condition a Gaussian profile given by:

$$q_0(x,y) = \exp(-10\cos^2(2\pi x))\exp(-10\cos^2(2\pi y)), \quad (x,y)\in[0,1]\times[0,1], \tag{3.26}$$

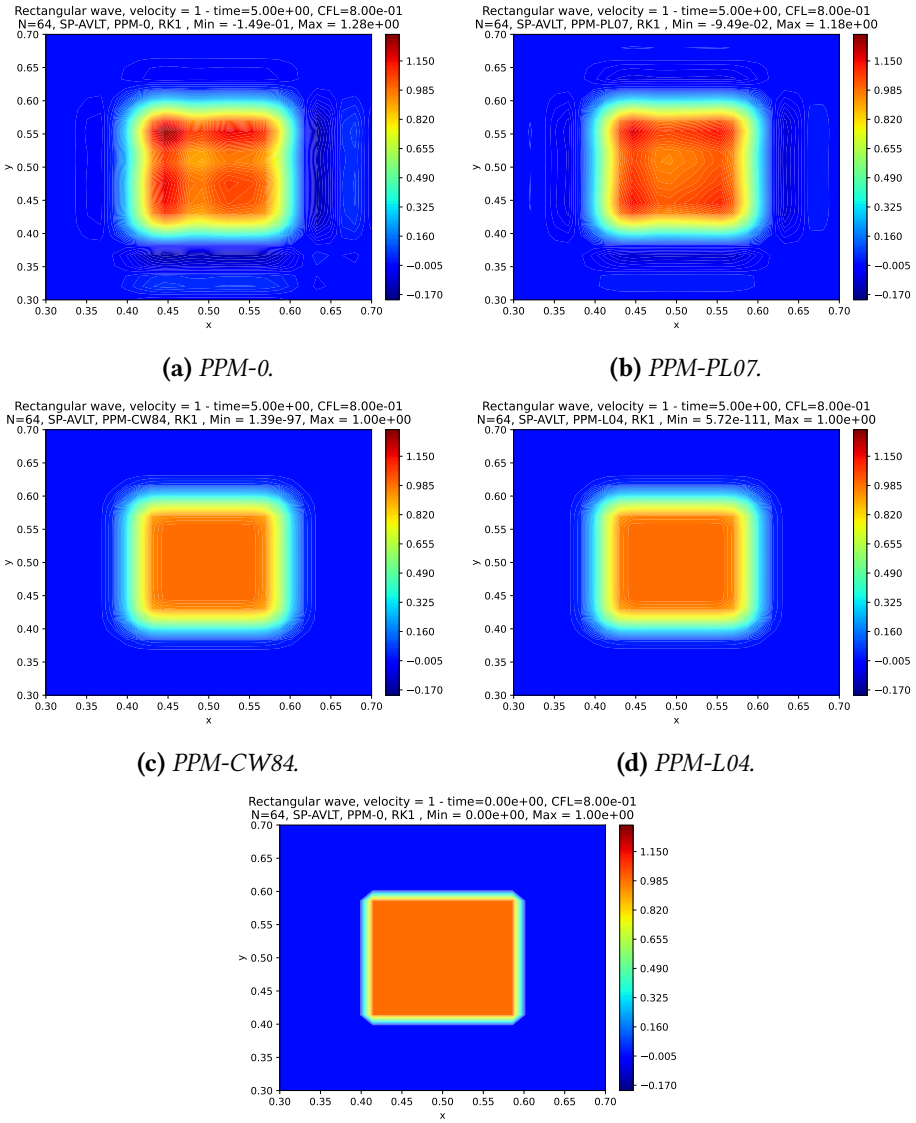
and a rectangular profile given by:

$$q_0(x,y) = \begin{cases} 1 & \text{if } (x,y)\in[0.4,0.6]\times[0.4,0.6], \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$



(a) *Error.* **(b)** *Convergence rate.*

Figure 3.1: *Relative error convergence (a) and convergence rate (b) for PPM schemes with AVLT splitting applied to the advection equation using a constant velocity* $\boldsymbol{u} = (0.2, -0.2)$*, a CFL number equal to* 0.8*, a final time equal to 5 time units and the initial condition given by Equation* (3.26).



(a) *PPM-0.*

(b) *PPM-PL07.*

(c) *PPM-CW84.*

(d) *PPM-L04.*

(e) *Initial condition.*

Figure 3.2: *Linear advection experiment using a constant velocity equal to* $\boldsymbol{u} = (0.2, -0.2)$*, a CFL number equal to* 0.8*,* $N = M = 64$*, and the initial condition is given by Equation* (3.27)*. We use the schemes PPM-0, PPM-PL07, PPM-CW84 and PPM-L04 with AVLT splitting. These figures show the advected profile after 5 time units (one time period). The initial condition is shown in (e).*

The exact solution of Problem 3.1 in this case is $q_0(x - ut)$ for both q_0. Since the velocity field is constant, all the splitting schemes introduce in Section 3.3 are the same, hence we are just going to consider the AVLT splitting. Besides that, it is easy to see that the Lie-Trotter splitting is exact in this case (cf. eg. LeVeque, 1990, p. 202-203). For 1D schemes, we use the RK1 to compute the departure point, since this scheme is exact if the velocity is constant.

The conclusions for the constant velocity tests are very similar to the 1D tests from Section 2.4.2. In fact, the error and convergence rate (Figure 3.1) for the Gaussian profile is very similar to the one-dimensional case (Figure 2.7). This behavior is due to the fact

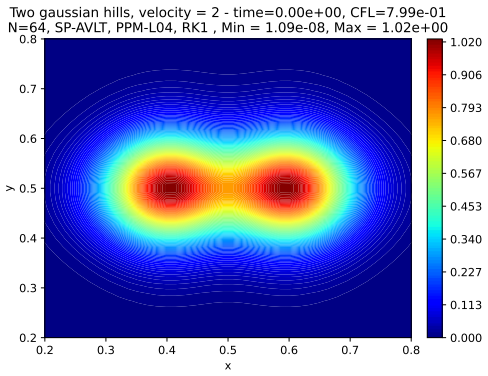
of no splitting error is introduced when the velocity is constant. From Figure 3.2, we can see that the AVLT splitting preserves the monotonicity when we use the monotonic 1D schemes PPM-CW84 and PPM-L04. For the non-monotonic schemes, we observe as in Figure 2.8 that PPM-PL07 produces less numerical dispersion than PPM-0.

3.4.2 Linear advection equation with variable velocity simulations

For variable velocity testing, we consider two Gaussian hills given by:

$$\begin{aligned} q_0(x,y) = &\exp(-10\cos^2(2\pi(x-0.1))\exp(-10\cos^2(2\pi y))+ \\ &\exp(-10\cos^2(2\pi(x+0.1)))\exp(-10\cos^2(2\pi y)), \end{aligned} \tag{3.28}$$

defined in $[0,1]\times[0,1]$, whose graph is shown in Figure 3.3a.



(a) *Initial condition from Equation* (3.28).

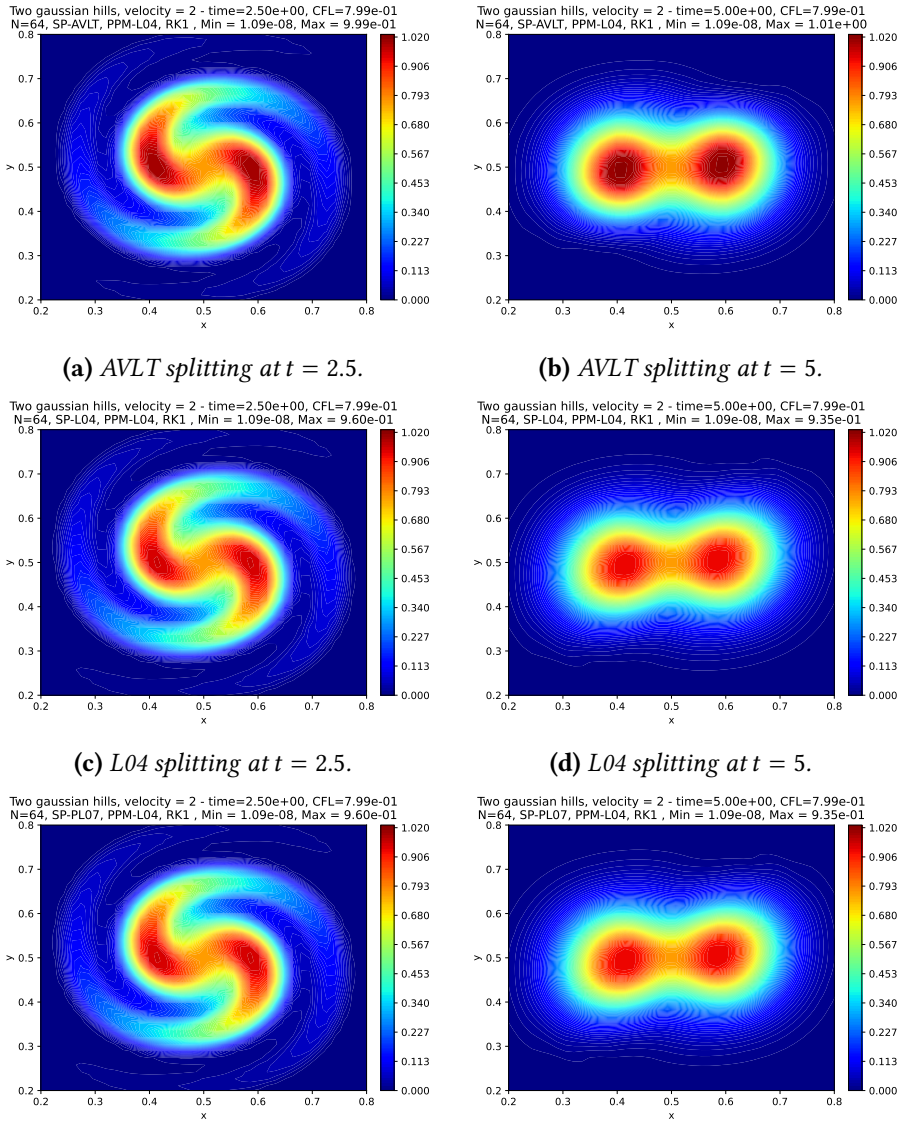
We consider the velocity proposed by Nair and Lauritzen (2010):

$$\begin{cases} u(x,y,t) &= \sin(\pi x)^2\sin(2\pi y)\cos\left(\frac{\pi t}{T}\right), \\ v(x,y,t) &= -\sin(\pi y)^2\sin(2\pi x)\cos\left(\frac{\pi t}{T}\right), \end{cases} \tag{3.29}$$

where $T=5$. We also consider the Cartesian version of the deformational flow test case on the sphere from Nair and Lauritzen (2010) proposed by Y. Chen et al. (2017). The velocity is given by:

$$\begin{cases} u(x,y,t) &= c\frac{\pi}{L_y}\sin^2(\alpha_1)(2\cos(\alpha_2)\sin(\alpha_2))(\cos(\alpha_3)) - \frac{L_x}{T}, \\ v(x,y,t) &= \frac{-c}{\pi}\frac{2\pi}{L_x}(2\sin(\alpha_1)\cos(\alpha_1)\cos^2(\alpha_2))\cos(\alpha_3), \end{cases} \tag{3.30}$$

where $L_x = 2\pi, L_y = \pi, T = 5, c = \frac{10}{T}(\frac{L_x}{2\pi})^2, \alpha_1 = 2\pi\left(\frac{X}{L_x} - \frac{t}{T}\right), \alpha_2 = \frac{\pi Y}{L_y}, \alpha_3 = \frac{\pi t}{T}, X = -\pi + 2\pi x$, $Y = -\frac{\pi}{2} + \pi y$. Y. Chen et al. (2017) uses periodic boundary conditions in the $x-$direction and zero-gradient in the $y-$direction. We shall employ biperiodic boundary conditions to simplify the problem. Both velocity fields are divergence-free and they deform the initial condition after 5 time units the scalar field returns to its initial position and shape, therefore we can compute the error.



(a) *AVLT splitting at* $t = 2.5$.

(b) *AVLT splitting at* $t = 5$.

(c) *L04 splitting at* $t = 2.5$.

(d) *L04 splitting at* $t = 5$.

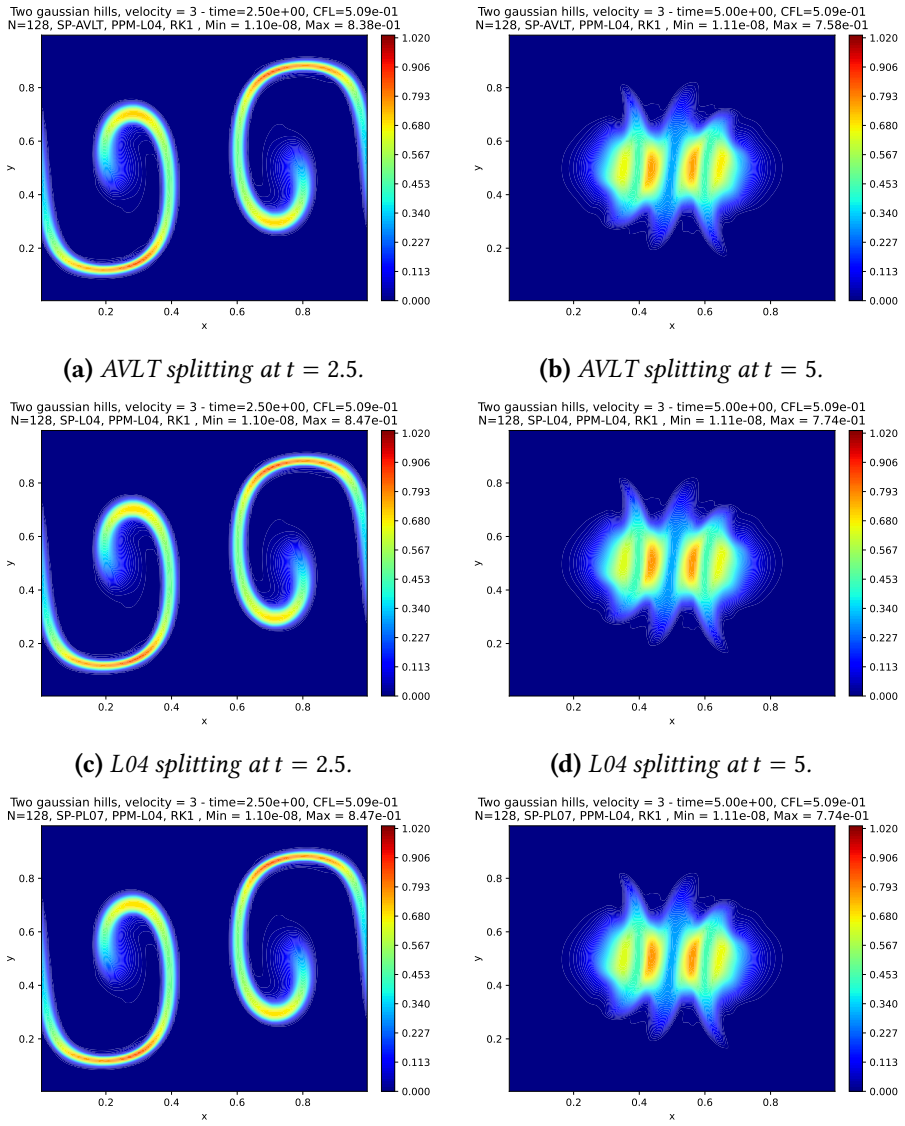
(e) *PL07 splitting at* $t = 2.5$.

(f) *PL07 splitting at* $t = 5$.

Figure 3.4: *Linear advection experiment using the velocity from Equation* (3.29)*, a CFL number equal to* 0.8*,* $N = M = 64$*, and the initial condition is given by Equation* (3.28)*. We use the scheme PPM-L04 with AVLT (a and b), L04 (c and d) and PL07 (e and f) splitting and RK1 departure point scheme. These figures show the advected profile after 2.5 (left) and 5 (right) time units (one time period).*

In Figure 3.4 we depict the results obtained using the two Gaussian hills and the velocity field from Equation (3.30) using the PPM-L04 with AVLT, L04, and PL07 splitting and RK1 as the departure point scheme. We can observe how the scalar field has deformed and returns to its initial position. However, we can notice that the PL07 and L04 splitting are more diffusive than the AVLT splitting when we look at the solution after 5 time units in Figure 3.4. Employing the RK1 scheme and also considering the PPM-PL07 scheme, from Figure 3.6 and Figure 3.7 we observe that the AVLT has the larger error considering and all schemes have first-order of accuracy (Figure 3.7a). However, when we use the RK3 scheme, the most accurate schemes are obtained with AVLT with PPM-L07, which reaches third-order (Figure 3.7b), despite AVLT being second-order, followed by AVLT with

PPM-L04 (Figure 3.6b). The other schemes seem only to not improve but get a larger error using RK3. Therefore, we conclude that the L04 and PL07 splitting introduce a first-order error. Furthermore, these schemes do not seem to differ from each other.



(a) *AVLT splitting at* $t = 2.5$.

(b) *AVLT splitting at* $t = 5$.

(c) *L04 splitting at* $t = 2.5$.

(d) *L04 splitting at* $t = 5$.

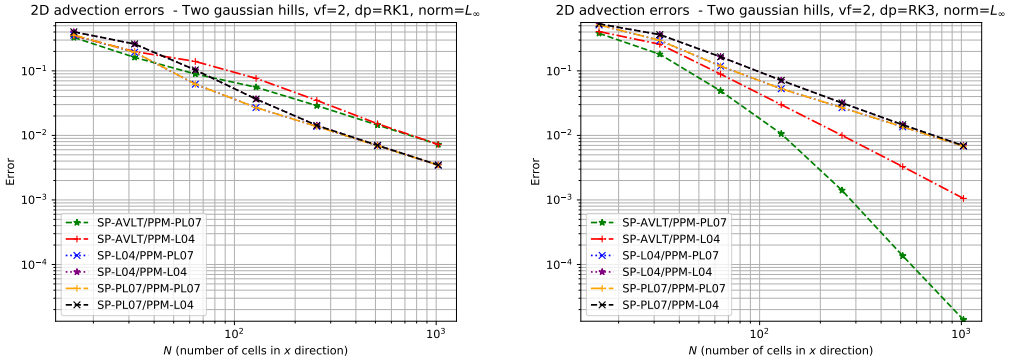
(e) *PL07 splitting at* $t = 2.5$.

(f) *PL07 splitting at* $t = 5$.

Figure 3.5: *Similar to Figure 3.5 but using the velocity from Equation* (3.30) *and* $N = 128$.

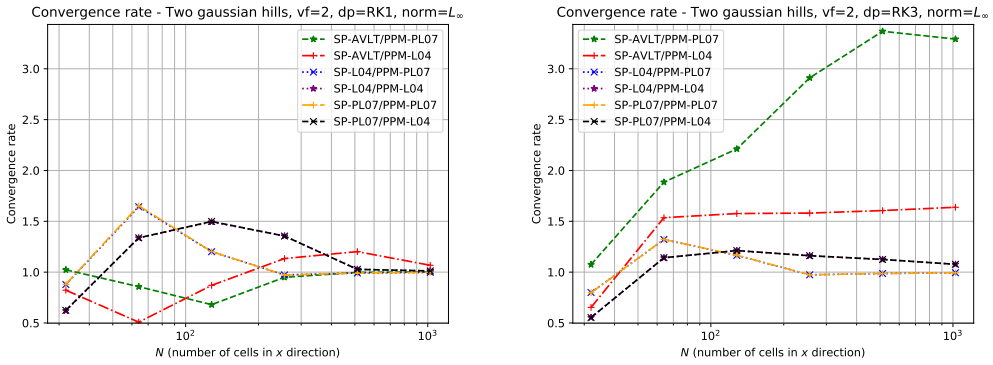
In Figure 3.5 we depict the results using the same setup and schemes as before but using the velocity from Equation (3.30). As in the previous test, we see the velocity deformarting the Gaussian hills. Again the PL07 and L04 schemes produce almost the same results, however, in contrast with previous test, they are less diffusive than the AVLT. From Figure 3.8a, notice that, when using the RK1 schemes, the PL07 and L04 schemes are slightly more accurate than AVLT for the 1D PPM-PL07 schemes. All these schemes reaches a converge-order superior to two (Figure 3.9a). When we use the 1D scheme PPM-L04, all the splitting method are very similar, regardless the departure point scheme. However, when use the RK3 schemes, the AVLT is the most accurate and reaches third-order (Figure

3.9b).



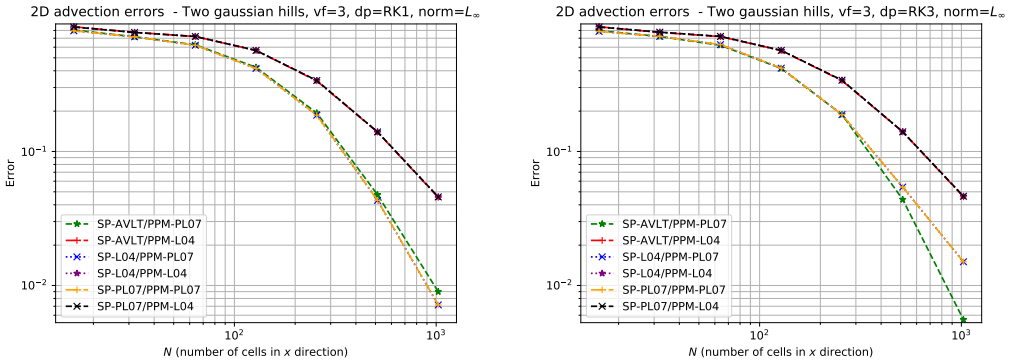
(a) *RK1.* **(b)** *RK3.*

Figure 3.6: *Convergence of the error for the schemes PPM-PL07and PPM-L04 with AVLT/L04/PL07/ splitting applied to the linear advection problem using a velocity from Equation a CFL number equal to* 0.8*, a final time of integration equal to 5 time units and the initial condition given by Equation* (3.26)*. The departure points are computed using the RK1 (left) and RK3 (right) schemes.*



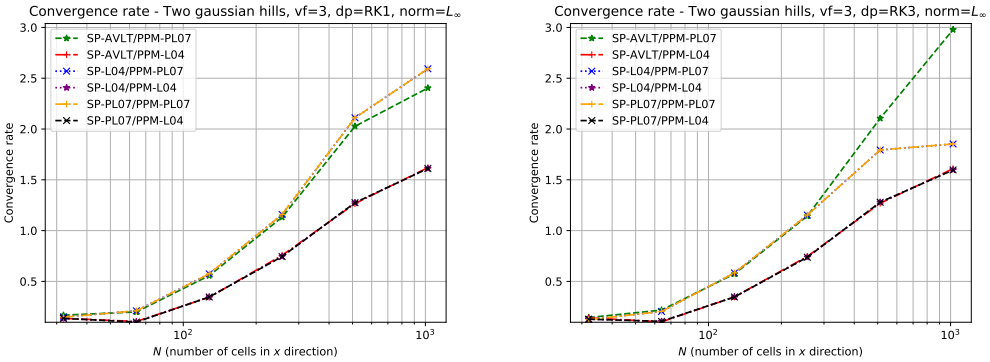
(a) *RK1.* **(b)** *RK3.*

Figure 3.7: *Similar to Figure 3.6 but considering the convergence rate.*



(a) *RK1.* **(b)** *RK3.*

Figure 3.8: *Similar to Figure 3.6 but using the velocity from Equation* (3.30)*.*



(a) *RK1.* **(b)** *RK3.*

Figure 3.9: *Similar to Figure 3.7 but using the velocity from Equation (3.30).*

3.5 Concluding remarks

In this Chapter, we introduced the dimension-splitting method which replaces the solution of the 2D advection equation by the solution of multiples 1D advection equations. The average of two Lie-Trotter splitting. which is second-order accurate, was modified to ensure the preservation of a constant scalar field with a divergence-free velocity, which the classical averaging Lie-Trotter splitting lacks, following the methodology used in FV3.

From the constant velocity simulations, we conclude that all the splitting schemes are equal and they do not introduce any splitting error. In fact, they are exact. We could observe that all conclusion for the 1D simulations were wit in the 2D case, with mass conservation and monotonicity being respected in the 2D case whenever we employ it in the 1D subproblems.

For variable velocity simulations, we investigated two deformation test cases. We observed that all splitting schemes preserved the monotonicity. The schemes PL07 and L04 yields very similar results and they introduced a first-order error, which was observed when we employed the RK3 scheme to compute the departure point. In this case, the AVLT reached third-order of accuracy, which is better than expected since this scheme is only second-order. Therefore, we can conclude that a more accurate departure point calculation benefits more the AVLT splitting. However, when using a first-order departure point computation, the splitting schemes PL07 and L04 produced slightly smaller errors.

Chapter 4

Cubed-sphere grids

The cubed-sphere grid was originally proposed by Sadourny (1972) and was reinvestigated by Ronchi et al. (1996) and Rančić et al. (1996). As is usual to Planotic grids, we start with a Platonic solid, in this case, a cube, circumscribed in a sphere and project its faces on the sphere. The original cubed-sphere proposed by Sadourny (1972), called equidistant cubed-sphere, leads to non-uniform grid; a solution to this problem was proposed with the introduction of angular coordinates, leading to a quasi-uniform grid called equiangular cubed-sphere. The cubed sphere then consists of six panels, each one having a local Cartesian coordinate system, which make it easier to extend methods from the plane to the sphere. Indeed, Putman and Lin (2007) extends the dimension splitting from Lin and Rood (1996), presented in Chapter 3, to the cubed-sphere.

There are essentially two major challenges when working on the cubed-sphere: 1) the non-orthogonal grid system; 2) the discontinuity of the coordinate system at the cube edges. Challenge 1) is more related to the appearance of metric terms in the equations, which adds more computational cost requiring, for instance, several conversions between contravariant/covariant components of a velocity field. The second challenge, perhaps the most problematic, is related to the computation of stencils along the cube edges, where the coordinate system is discontinuous. A possible way to compute the stencil at the edges is to extend the local coordinate of each panel to its neighbor panels, adding ghost cells in the so-called halo region. In this case, the equiangular cubed sphere has ghost cell values, lying on the same geodesics containing the data from its neighbor panels, which allows us to use one-dimensional high-order Lagrange interpolation (for a review of this method, see Zerroukat and Allen (2022)). This approach was already investing since the work of Ronchi et al. (1996) and it is widely used in the literature (X. Chen, 2021; Croisille, 2013; Katta et al., 2015a, 2015b). Putman and Lin (2007), on the other hand, uses extrapolation at grid values near the cube edges. Another approach that avoids the need for interpolation/extrapolation near the edges is the conformal cubed-sphere developed by Rančić et al. (1996). This grid leads to an orthogonal and continuous coordinate system near the edges, but it generates grid singularities near the cube corners, similar to the pole problem. An improved and more uniform conformal grid, called Uniform Jacobian cubed sphere, was later proposed Rančić et al. (2017). Each approach is likely to generate grid imprinting and one of our goals this work is to investigate how much grid-imprinting each method generates.

This chapter aims to review and investigated the geometrical properties of the cubed-sphere. Besides that, we also aim to investigate the process of interpolating/extrapolating near the cube edges. We start with a basic review of the cubed-sphere mappings in Section 4.1, while Section 4.2 investigates the interpolation/extrapolation near the cube edges with some numerical experiments.

4.1 Cubed-sphere mappings

4.1.1 Equidistant cubed-sphere

We consider a sphere radius $R > 0$ and $a = \frac{R}{\sqrt{3}}$ representing the half-length of the cube, and the family of maps $\Psi_p : [-a, a] \times [-a, a] \to \mathbb{S}^2_R$, $p = 1, \cdots, 6$, where:

$$\Psi_1(x, y) = \frac{R}{\sqrt{a^2 + x^2 + y^2}}(a, x, y), \tag{4.1}$$

$$\Psi_2(x, y) = \frac{R}{\sqrt{a^2 + x^2 + y^2}}(-x, a, y), \tag{4.2}$$

$$\Psi_3(x, y) = \frac{R}{\sqrt{a^2 + x^2 + y^2}}(-a, -x, y), \tag{4.3}$$

$$\Psi_4(x, y) = \frac{R}{\sqrt{a^2 + x^2 + y^2}}(x, -a, y), \tag{4.4}$$

$$\Psi_5(x, y) = \frac{R}{\sqrt{a^2 + x^2 + y^2}}(-y, x, a), \tag{4.5}$$

$$\Psi_6(x, y) = \frac{R}{\sqrt{a^2 + x^2 + y^2}}(y, x, -a). \tag{4.6}$$

The family of maps $\{\Psi_p, p = 1, \cdots, 6\}$ allow us to cover the sphere. Here p denotes a panel, and they are defined as Figure 4.1 shows. The derivative of the maps Ψ_p are given by:

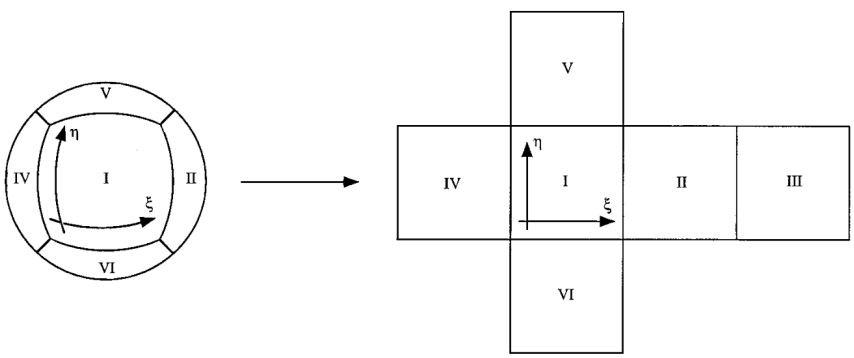


Figure 4.1: *Cubed-sphere panels definition. Figure taken from Ronchi et al. (1996).*

$$D\Psi_1(x,y) = \frac{R}{(a^2+x^2+y^2)^{3/2}} \begin{bmatrix} -ax & -ay \\ a^2+y^2 & -xy \\ -xy & a^2+x^2 \end{bmatrix}, \tag{4.7}$$

$$D\Psi_2(x,y) = \frac{R}{(a^2+x^2+y^2)^{3/2}} \begin{bmatrix} -(a^2+y^2) & xy \\ -ax & -ay \\ -xy & a^2+x^2 \end{bmatrix}, \tag{4.8}$$

$$D\Psi_3(x,y) = \frac{R}{(a^2+x^2+y^2)^{3/2}} \begin{bmatrix} ax & ay \\ -(a^2+y^2) & xy \\ -xy & a^2+x^2 \end{bmatrix}, \tag{4.9}$$

$$D\Psi_4(x,y) = \frac{R}{(a^2+x^2+y^2)^{3/2}} \begin{bmatrix} a^2+y^2 & -xy \\ ax & ay \\ -xy & a^2+x^2 \end{bmatrix}, \tag{4.10}$$

$$D\Psi_5(x,y) = \frac{R}{(a^2+x^2+y^2)^{3/2}} \begin{bmatrix} xy & -(a^2+x^2) \\ a^2+y^2 & -xy \\ -ax & -ay \end{bmatrix}, \tag{4.11}$$

$$D\Psi_6(x,y) = \frac{R}{(a^2+x^2+y^2)^{3/2}} \begin{bmatrix} -xy & a^2+x^2 \\ a^2+y^2 & -xy \\ ax & ay \end{bmatrix}. \tag{4.12}$$

With the aid of the derivative, we may define a basis of tangent vectors $\{\boldsymbol{g}_1, \boldsymbol{g}_2\}$ on each point on the sphere by:

$$\boldsymbol{g}_1(x,y;p) = D\Psi_p(x,y)\begin{bmatrix}1\\0\end{bmatrix}, \boldsymbol{g}_2(x,y;p) = D\Psi_p(x,y)\begin{bmatrix}0\\1\end{bmatrix}. \tag{4.13}$$

In other words, we have $\{\boldsymbol{g}_1(x,y;p), \boldsymbol{g}_2(x,y;p)\} \subset T_{\Psi_p(x,y)}\mathbb{S}^2_R$, $\forall (x,y) \in [-a,a]\times[-a,a]$ (see Appendix B for the definition of tangent space). Notice that

$$[D\Psi_p(x,y)]^T D\Psi_p(x,y) = \frac{R^2}{(a^2+x^2+y^2)^2} \begin{bmatrix} a^2+x^2 & -xy \\ -xy & a^2+y^2 \end{bmatrix}, \tag{4.14}$$

does not depend on p. Hence, it makes sense to define the matrix $G_\Psi(x,y) = [D\Psi_p(x,y)]^T D\Psi_p(x,y)$ which is known as metric tensor. It is easy to see that:

$$G_\Psi(x,y) = \begin{bmatrix} \langle \boldsymbol{g}_1(x,y;p), \boldsymbol{g}_1(x,y;p)\rangle & \langle \boldsymbol{g}_1(x,y;p), \boldsymbol{g}_2(x,y;p)\rangle \\ \langle \boldsymbol{g}_1(x,y;p), \boldsymbol{g}_2(x,y;p)\rangle & \langle \boldsymbol{g}_2(x,y;p), \boldsymbol{g}_2(x,y;p)\rangle \end{bmatrix}, \tag{4.15}$$

and that $G_\Psi(x,y)$ is positive-definite, $\forall (x,y) \in [-a,a]\times[-a,a]$. The Jacobian of the metric tensor $G_\Psi(x,y)$ is then given by:

$$\sqrt{|\det G_\Psi(x,y)|} = \frac{R^2}{(a^2+x^2+y^2)^{3/2}} a. \tag{4.16}$$

4.1.2 Equiangular cubed-sphere

Another cubed-sphere mapping is the equiangular mapping (Ronchi et al., 1996), which leads to a more uniform grid. This mapping is a composition of equidistant mapping with angular coordinates. We consider again $a = \frac{R}{\sqrt{3}}$ and we define the family of maps $\Phi_p : [-\frac{\pi}{4}, \frac{\pi}{4}] \times [-\frac{\pi}{4}, \frac{\pi}{4}] \to \mathbb{S}^2_R$, $p = 1, \cdots, 6$, given by $\Phi_p(x, y) = \Psi_p(a \tan x, a \tan y)$. The coordinates $(a \tan x, a \tan y)$ are called angular coordinates. By the chain rule:

$$D\Phi_p(x, y) = aD\Psi_p(a \tan x, a \tan y) \begin{bmatrix} \frac{1}{\cos^2 x} & 0 \\ 0 & \frac{1}{\cos^2 y} \end{bmatrix}, \tag{4.17}$$

and therefore we can define the following tangent vectors

$$\boldsymbol{r}_1(x, y; p) = D\Phi_p(x, y) \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{a}{\cos^2 x} \boldsymbol{g}_1(\tan x, \tan y; p), \tag{4.18}$$

$$\boldsymbol{r}_2(x, y; p) = D\Phi_p(x, y) \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{a}{\cos^2 y} \boldsymbol{g}_2(\tan x, \tan y; p), \tag{4.19}$$

that is, $\{\boldsymbol{r}_1(x, y; p), \boldsymbol{r}_{2,p}(x, y; p)\} \subset T_{\Phi_p(x,y)}\mathbb{S}^2_R$, $\forall (x, y) \in [-\frac{\pi}{4}, \frac{\pi}{4}] \times [-\frac{\pi}{4}, \frac{\pi}{4}]$. Again, it makes sense to define the matrix

$$G_\Phi(x, y) = [D\Phi_p(x, y)]^T D\Phi_p(x, y) \tag{4.20}$$

$$= a^2 [D\Psi_p(a \tan x, a \tan y)]^T \begin{bmatrix} \frac{1}{\cos^4 x} & 0 \\ 0 & \frac{1}{\cos^4 y} \end{bmatrix} D\Psi_p(a \tan x, a \tan y), \tag{4.21}$$

that does not depend on p and is the metric tensor. It is easy to see that:

$$G_\Phi(x, y) = \begin{bmatrix} \langle \boldsymbol{r}_1(x, y; p), \boldsymbol{r}_1(x, y; p) \rangle & \langle \boldsymbol{r}_1(x, y; p), \boldsymbol{r}_2(x, y; p) \rangle \\ \langle \boldsymbol{r}_1(x, y; p), \boldsymbol{r}_2(x, y; p) \rangle & \langle \boldsymbol{r}_2(x, y; p), \boldsymbol{r}_2(x, y; p) \rangle \end{bmatrix}, \tag{4.22}$$

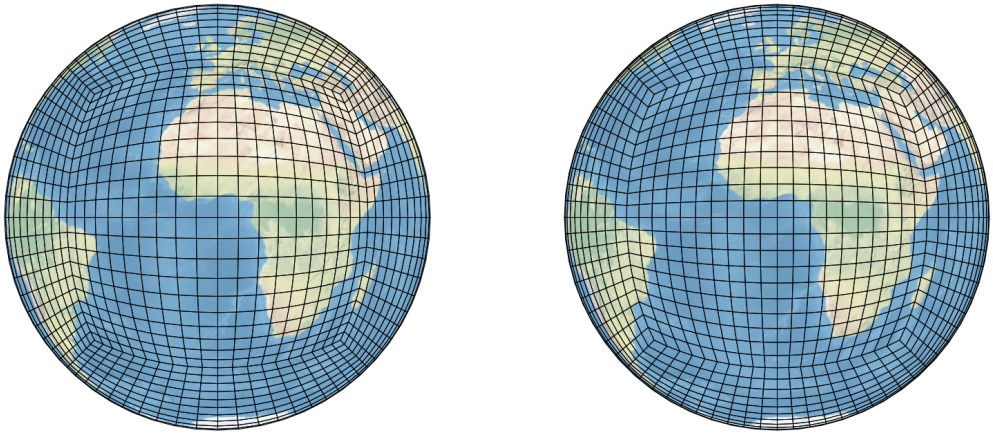
and that $G_\Phi(x, y)$ is positive-definite, $\forall (x, y) \in [-\frac{\pi}{4}, \frac{\pi}{4}] \times [-\frac{\pi}{4}, \frac{\pi}{4}]$. The Jacobian of the metric tensor $G_\Phi(x, y)$ is then given by:

$$\begin{aligned} \sqrt{|\det G_\Phi(x, y)|} &= \frac{a}{\cos^2 x \cos^2 y} \frac{R^2}{(a^2 + a^2 \tan^2 x + a^2 \tan^2 y)^{3/2}} a \\ &= \frac{R^2}{\cos^2 x \cos^2 y} \frac{1}{(1 + \tan^2 x + \tan^2 y)^{3/2}}. \end{aligned} \tag{4.23}$$

Hereafter, we shall denote $g(x, y) = |\det G_\Phi(x, y)|$.

4.1.3 Examples

To represent the cubed-sphere grid, we consider the notation of Section 3.2. We shall assume that we have a $(\Delta x, \Delta y)$−grid of $[-a, a]^2$, with $\Delta x = \Delta y$, $N = M$, where a depends on the mapping considered. Similarly to Chapter 3, we may extend the grid to ghost cells. Hence, for each $p = 1, \cdots, 6$, we apply the mapping Ψ_p or Φ_p on each local coordinate system grid to generate the cubed-sphere with $6N^2$ cells. In Figure 4.2, we depict an example of the grids generated using the equidistant and equiangular mappings for $N = 20$.



(a) *Equidistant* **(b)** *Equiangular*

Figure 4.2: *Equidistant (a) and equiangular (b) cubed-spheres generated with* $N = 20$.

From Figure 4.2, we notice that the equiangular cubed-sphere is much more uniform than the equidistant cubed-sphere. As pointed out by Rančić et al. (1996), the ratio between the maximum and minimum cell area on the equiangular cubed-sphere is approximately 1.3, while the same ratio is approximately 5.2 on the equidistant cubed-sphere.

4.2 Edges treatment

4.2.1 Ghost cells interpolation

Hereafter, we are going to use the notation $(x, y; p)$ for $p = 1, \cdots, 6$, to represent a point on the sphere obtained trough the equiangular cubed-sphere mapping. Let us assume that we have a function $q : \mathbb{S}^2_R \to \mathbb{R}$ given in the cell centroids and let us denote these values by $q_{ijp} = q(x_i, y_j; p)$, for $i, j = 1 \cdots, N$, $p = 1, \cdots, 6$. We wish to estimate these values outside of the range $1, \cdots, N$, that is, we wish to estimate them at ghost cells positions.

As we pointed out before, this can be done by noticing that the ghost cells on the local Cartesian systems are mapped onto the geodesic of a neighbor panel, which allows us to use Lagrange interpolation to obtain the ghost cell values. To illustrate this process in Panel 1, in Figure 4.3 we depict the values of q_{ijp} in Panel 1 using green circles, for $N = 8$ and black circles for the other panels. Assuming a halo size equal to 3, we also show the target value at ghost cells using yellow and magenta circles. Observe that the dashed yellow lines in Figure 4.3 illustrate how the ghost cells are in geodesics containing grid values from neighbor panels. Except for the magenta circles, all the ghost values may be obtained using 1D Lagrange interpolation using the surrounding black circles on the geodesic. This can be performed for all panels. After that, the magenta circles may be interpolated using values obtained in the first step of interpolation, (see cyan circles in Figure 4.3), preserving the order of accuracy of the interpolation, supposing this one is fixed.

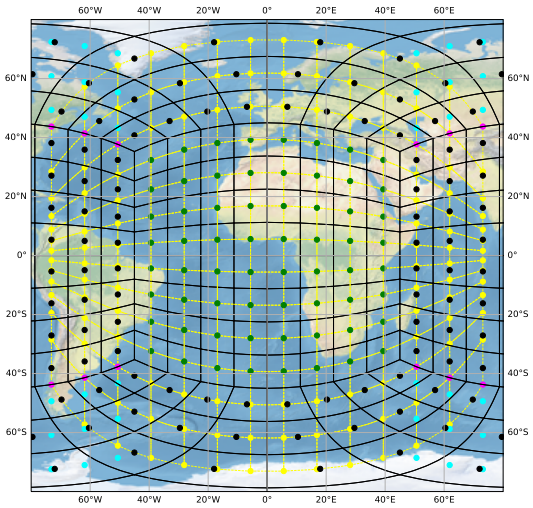


Figure 4.3: *Equiangular cubed-sphere panel 1 with $N = 8$: centroid at panel 1 (green circles) and others panel centroids (black circles), ghost cell points at panel 1 (yellow and magenta circles) and others panels (cyan circles).*

We are going to show some numerical example of this interpolation process using a halo region of size 3 and assuming the radius of the sphere equal to one. First, we consider the following function represented in $\mathbb{R}^3$ coordinates as

$$q(X, Y, Z) = \exp(-10((X - X_0)^2 + (Y - Y_0)^2 + (Z - Z_0)^2)), \tag{4.24}$$

where (X_0, Y_0, Z_0) represent the $\mathbb{R}^3$ coordinates of the latitude-longitude point $(\frac{\pi}{4}, \frac{\pi}{6})$ as in Zerroukat and Allen (2022), which consists of a Gaussian centered at a panel 1 corner as Figure 4.4 shows.

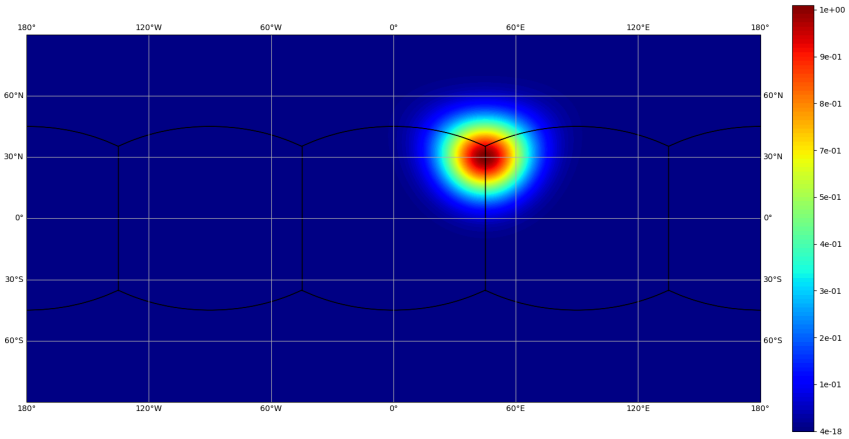


Figure 4.4: *Gaussian defined by Equation* (4.24).

We shall also consider the following trigonometric function, which is the divergence of a velocity field, as in Peixoto and Barros (2013) in our tests:

$$q(\lambda,\phi) = \frac{1}{\cos(\phi)}\Big(-2\cos^3(\phi)\sin(\lambda)\cos(\lambda) + 16\sin^2(\lambda)\cos(\lambda)\cos^3(\phi)\sin(\phi)\Big), \quad (4.25)$$

whose graph is depicted in Figure 4.5.

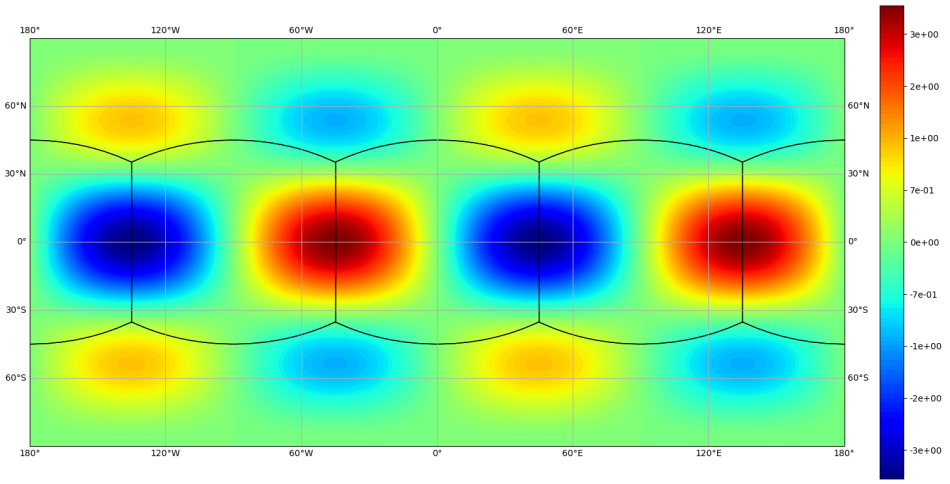
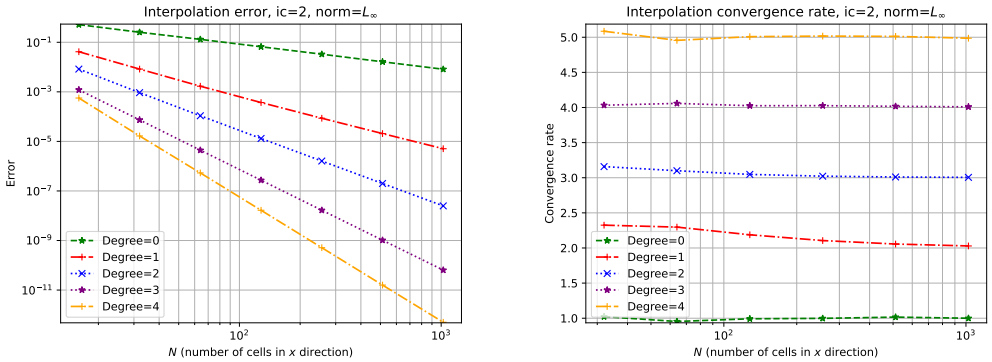


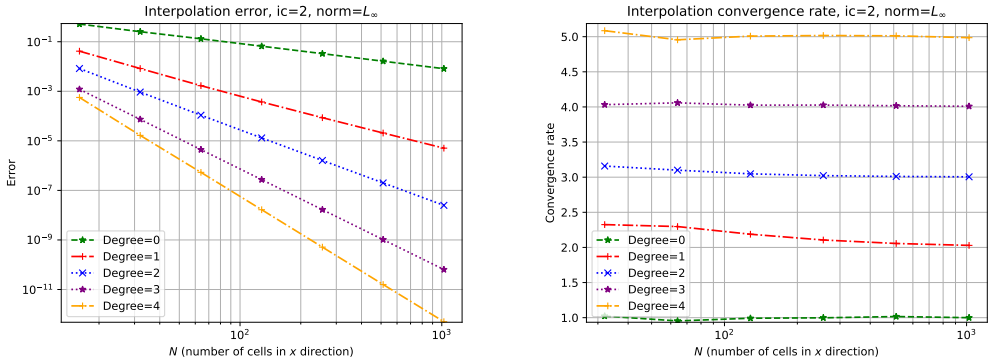
Figure 4.5: *Trigonometric function defined by Equation* (4.25).

We are going to consider the relative error in the maximum norm and the convergence rate at the host cell positions defined analogously as in Section 3.4. We also consider values os N given by 2^k, for $k = 4, \cdots, 10$, in order to compute the error and convergence rate.



(a) *Error.* **(b)** *Convergence rate.*

Figure 4.6: *Relative error convergence (a) and convergence rate (b) for the ghost cell interpolation process for different polynomial degrees, using the Gaussian function given by Equation* (4.24).



(a) *Error.* **(b)** *Convergence rate.*

Figure 4.7: *As Figure 4.6 but using the trigonometric function given by Equation* (4.25).

In Figures 4.6 and 4.7 we show the errors and convergence rate for the functions from Equations (4.24) and (4.25), respectively, considering polynomials of degrees from 0 up to 4. As both graphs show, we were able to achieve the expected order of convergence. Next, we shall use the ghost cell interpolation when computing the stencils from PPM edge reconstructions presented in Chapter 2 in each direction of a panel and also compare the results obtained with extrapolation.

4.2.2 Edges reconstruction

Let denote a control volume of the cubed-sphere by Ω_{ijp}, that is:

$$\Omega_{ijp} = \Phi_p([x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]), \quad 1 \leq i, j \leq N, \quad 1 \leq p \leq 6.$$

We define the average values of a function q with the aid of the metric tensor $g(x, y)$ (Equation 4.23):

$$Q_{ijp} = \frac{1}{|\Omega_{ijp}|} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q(x, y; p) \sqrt{g(x, y)}\, dx\, dy,$$

where $|\Omega_{ijp}|$ is the control volume area given by:

$$|\Omega_{ijp}| = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \sqrt{g(x, y)}\, dx\, dy.$$

Similar to Proposition 3.1, we may approximate the average value using the centroid value, that is

$$Q_{ijp} - q_{ijp} = O(\Delta x^2),$$

where $q_{ijp} = q(x_i, y_j; p)$, recalling that on the cubed-sphere we always assume $\Delta x = \Delta y$. In this work, we shall always approximate the average values since our schemes are expected to be at most second-order, this approximation does not deteriorate the convergence order.

Let us consider the following problem: given the values q_{ijp} we wish to find approxi-

mations of the function q at the control volume edge midpoints denoted by $q^{L,x}_{ijp} \approx q_{i-\frac{1}{2},j,p}$, $q^{R,x}_{ijp} \approx q_{i+\frac{1}{2},j,p}$, $q^{L,y}_{ijp} \approx q_{i,j-\frac{1}{2},p}$, $q^{R,y}_{ijp} \approx q_{i,j+\frac{1}{2},p}$, where we also using the notations $q_{i+\frac{1}{2},j,p} \approx q(x_{i+\frac{1}{2}}, y_j; p)$, $q_{i,j+\frac{1}{2},p} \approx q(x_i, y_{j+\frac{1}{2}}; p)$. These points are illustrated in Figure 4.8 for panel 1.

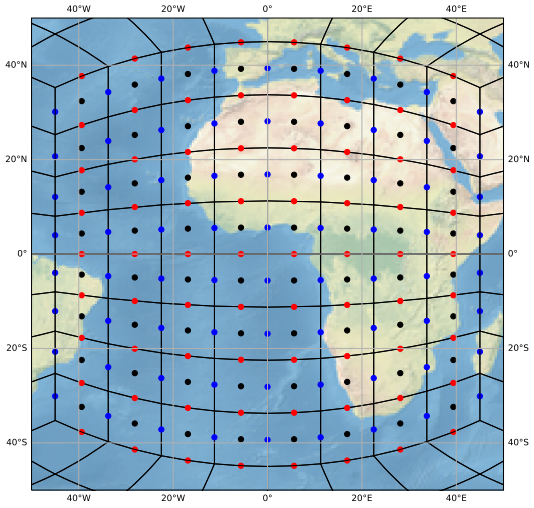


Figure 4.8: *The reconstruction problem on the cubed-sphere panel 1: we are given the centroid values of a function (black circles) and we wish to estimate these values at the edges midpoints in the x direction (blue circles) and y direction (red circles). This figure uses an equiangular cubed-sphere with $N = 8$.*

We can estimate the desired values using the one-dimensional reconstruction schemes from Sections 2.3.1 and 2.3.2 by performing the PPM reconstruction in the x and y directions independently. Notice that all the schemes from Sections 2.3.1 and 2.3.2 are expected to be second-order accurate due to centroid point approximation. The major difference here is that when we compute the stencil near the cube edges. Unlikely the previous chapters where we assumed periodic boundary conditions, the boundary conditions are related to the adjacent panel. One to overcome this problem is just to add ghost cell layers and use the process described in Section 4.2.1, hence all the stencils may be computed. This approach shall be referred to as **ET-1** (ET stands for edge treatment). Observe that the points that lie on a cube edge are computed twice, where each calculation uses one adjacent panel per time. One possible way to uniquely define this value while preserving the order is to average the two values obtained by the adjacent panels. This scheme is named **ET-2**.

An approach that avoids the use of ghost has been developed by Putman and Lin (2007) using extrapolation at the cells surrounding to the cube edge. We are going to describe

this scheme that we shall name **ET-3**. This scheme uses the extrapolation:

$$
\begin{aligned}
q^{L,x}_{1,j,p} &= \frac{1}{2}\Big(3Q_{1,j,p} - Q_{2,j,p}\Big),\\
q^{R,x}_{N,j,p} &= \frac{1}{2}\Big(3Q_{N,j,p} - Q_{N-1,j,p}\Big),\\
q^{L,y}_{i,1,p} &= \frac{1}{2}\Big(3Q_{i,1,p} - Q_{i,2,p}\Big),\\
q^{R,y}_{i,N,p} &= \frac{1}{2}\Big(3Q_{i,N,p} - Q_{i,N-1,p}\Big),
\end{aligned}
$$

at the points that are located on the cube edges. The other edge values are estimated as:

$$
\begin{aligned}
q^{R,x}_{1,j,p} &= \frac{1}{14}\Big(3Q_{1,j,p} + 11Q_{2,j,p} - 2(Q_{3,j,p} - Q_{1,j,p})\Big),\\
q^{L,x}_{2,j,p} &= q^{R,x}_{1,j,p},\\
q^{L,x}_{N,j,p} &= \frac{1}{14}\Big(3Q_{N,j,p} + 11Q_{N-1,j,p} - 2(Q_{N-2,j,p} - Q_{N,j,p})\Big),\\
q^{R,x}_{N-1,j,p} &= q^{L,x}_{N,j,p},
\end{aligned}
$$

in the x direction and in the y direction we use the formulas

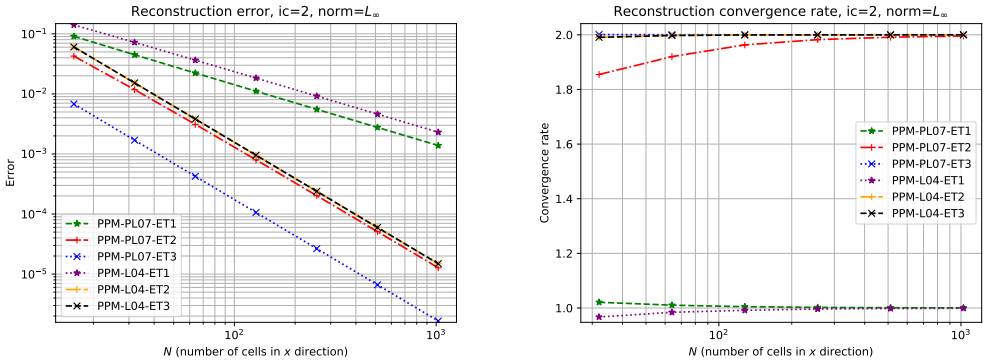
$$
\begin{aligned}
q^{R,y}_{i,1,p} &= \frac{1}{14}\Big(3Q_{i,1,p} + 11Q_{i,2,p} - 2(Q_{i,3,p} - Q_{i,1,p})\Big),\\
q^{L,y}_{i,2,p} &= q^{R,x}_{i,1,p},\\
q^{L,y}_{i,N,p} &= \frac{1}{14}\Big(3Q_{i,N,p} + 11Q_{i,N-1,p} - 2(Q_{i,N-2,p} - Q_{i,N,p})\Big),\\
q^{R,y}_{i,N-1,p} &= q^{L,y}_{i,N,p}.
\end{aligned}
$$

Again, we can average to values at the cube edges, which leads to the scheme named **ET-4**. We are going to use the trigonometric (Equation (4.24)) functions as before on the unit sphere to compare the schemes ET-1, ET-2, ET-3 and ET-4. The scheme ET-1 and ET-2 uses

cubic polynomials. We going to introduce the relative errors:

$$
\begin{aligned}
e_{i-\frac{1}{2},j,p} &= (|q_{i-\frac{1}{2},j,p} - q^{L,x}_{ijp}|)/|q_{i-\frac{1}{2},j,p}|,\\
e_{i+\frac{1}{2},j,p} &= (|q_{i+\frac{1}{2},j,p} - q^{R,x}_{ijp}|)/|q_{i+\frac{1}{2},j,p}|,\\
e_{i,j-\frac{1}{2},p} &= (|q_{i,j-\frac{1}{2},p} - q^{L,y}_{ijp}|)/|q_{i,j-\frac{1}{2},p}|,\\
e_{i,j+\frac{1}{2},p} &= (|q_{i,j+\frac{1}{2},p} - q^{R,y}_{ijp}|)/|q_{i,j+\frac{1}{2},p}|,\\
e_{ijp} &= \max\{e_{i-\frac{1}{2},j,p}, e_{i+\frac{1}{2},j,p}, e_{i,j-\frac{1}{2},p}, e_{i,j+\frac{1}{2},p}\},\\
E &= \max\{e_{ijp}\}.
\end{aligned}
$$

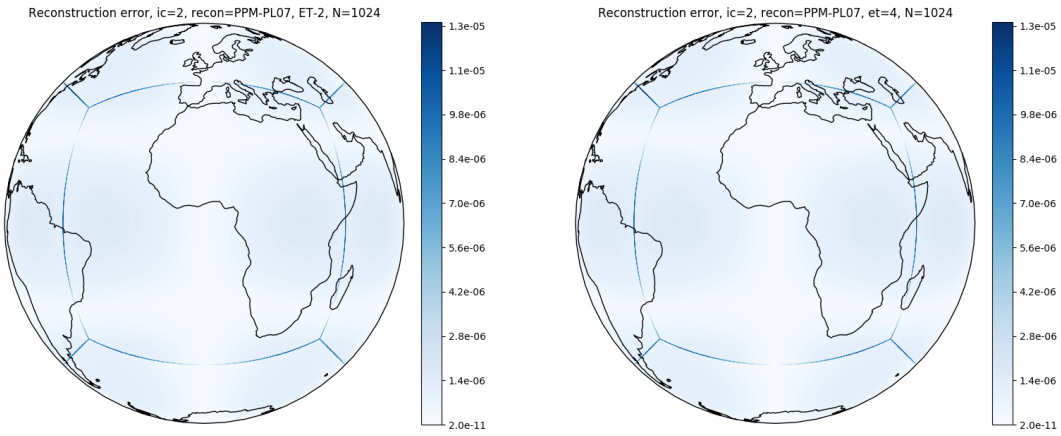
We are going to compute E for different values of N as in the numerical experiments of Section 4.2.1.



(a) *Error.* **(b)** *Convergence rate.*

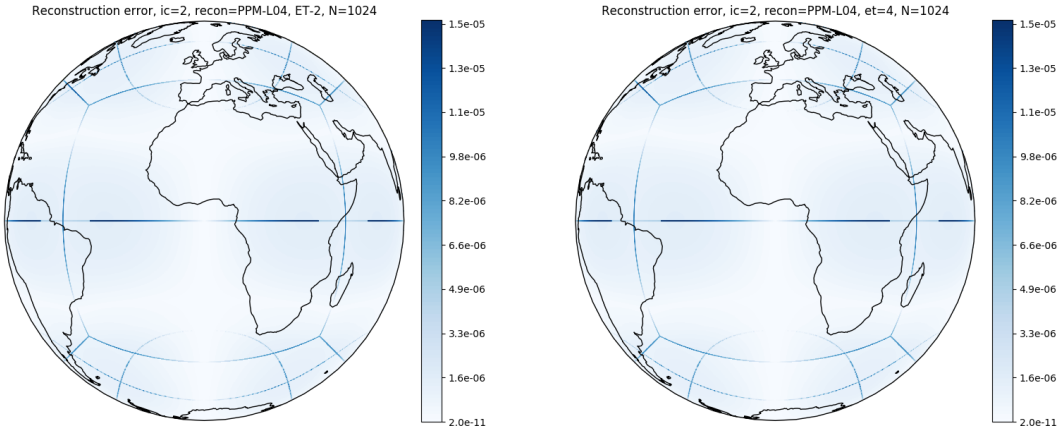
Figure 4.9: *Relative error convergence (a) and convergence rate (b) for the reconstruction problem for different edge treatments (ET), using the trigonometric function given by Equation* (4.25).

In Figure 4.9 we show the errors and the convergence rate using the PPM-PLO07 and PPM-L04 reconstruction schemes. We can observe that all the schemes converge to zero with second-order. Besides that, the scheme ET-1 and ET-2 produce essentially the same error, and so are the schemes ET-3 and ET-4. The difference between the ghost cell interpolation-based schemes and the extrapolation-based schemes may be observed in Figure 4.10 and Figure 4.11, for the PPM-PL07 and PPM-L04 reconstruction schemes, respectively. We notice that the cube edges appear on the error graph when we use the ET-4 scheme. This is an example of grid imprinting. Although all schemes are second-order, the ghost cell interpolation-based schemes do seem to produce grid imprinting in the reconstruction problem. The graphs of ET1 are very similar to the graph of ET-2 and the graphs of ET-3 are very similar to the graph of ET4 (not shown).



(a) *ET-2.* **(b)** *ET-4.*

Figure 4.10: *Error for the edge midpoint values considering the trigonometric function (Equation (4.25)) with edges treatment schemes ET-2 (a) and ET-4 (b) for the cubed-sphere with* $N = 1024$ *using the reconstruction PPM-PL07.*



(a) *ET-2.* **(b)** *ET-4.*

Figure 4.11: *As Figure in 4.10 but using the PPM-L04 scheme.*

Chapter 5

Cubed-sphere finite-volume methods

5.1 Advection finite-volume scheme

In this Chapter, we show how we can use the dimension splitting method presented in Chapter 3 to solve the advection equation on the cubed-sphere with base on Putman and Lin (2007).

We denote by $\Psi_p : [-a,a]\times[-a,a] \to \mathbb{S}^2_R$, $p = 1,\cdots,6$, as a cubed-sphere mapping introduce in Chapter 4. We introduce the notations:

- $(x, y; p)$ represents a point on the cubed-sphere using a cubed-sphere mapping;
- $[-a,a]^2 = \bigcup_{i,j=1}^{N} [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$;
- $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$;
- $\Omega_{ijp} = \Psi_p([x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}])$ are the cubed-sphere control-volumes;
- $\boldsymbol{g}_1(x,y;p) = D\Psi_p(x,y)\begin{bmatrix}1\\0\end{bmatrix}$ and $\boldsymbol{g}_2(x,y;p) = D\Psi_p(x,y)\begin{bmatrix}0\\1\end{bmatrix}$ are the tangent vectors;
- $g_\Psi(x,y) = \begin{bmatrix}\langle \boldsymbol{g}_1(x,y;p), \boldsymbol{g}_1(x,y;p)\rangle & \langle \boldsymbol{g}_1(x,y;p), \boldsymbol{g}_2(x,y;p)\rangle \\ \langle \boldsymbol{g}_1(x,y;p), \boldsymbol{g}_2(x,y;p)\rangle & \langle \boldsymbol{g}_2(x,y;p), \boldsymbol{g}_2(x,y;p)\rangle\end{bmatrix}$ is the metric tensor;
- $\sqrt{\det g_\Psi(x,y)}$ is the metric tensor Jacobian;
- $$|\Omega_{ijp}| = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{i-\frac{1}{2}}}^{y_{i+\frac{1}{2}}} \sqrt{\det g_\Psi(x,y)}\, dx\, dy$$
  are the control-volume areas
- $$Q_{ijp}(t) = \frac{1}{|\Omega_{ijp}|} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q(x,y,t;p)\sqrt{\det g_\Psi(x,y)}\, dx\, dy$$

are the averages of q on the control-volumes;

- $u^n_{i+\frac{1}{2},j,p} = u(x_{i+\frac{1}{2}}, y_j, t_n; p)$;
- $v^n_{i,j+\frac{1}{2},p} = v(x_i, y_{j+\frac{1}{2}}, t_n; p)$.

Given a tangent velocity field $\boldsymbol{u}$ on the sphere, we denote its contravariant components by $\tilde{u}$ and $\tilde{v}$. For a give a detailed discussion on contravariant representations in Appendix B. The advection equation on panel the p of the cubed-sphere is given by:

$$\frac{\partial}{\partial t}q + \frac{1}{\sqrt{\det g_\Psi}}\left(\frac{\partial}{\partial x}(\tilde{u}\sqrt{\det g_\Psi}q) + \frac{\partial}{\partial y}(\tilde{v}\sqrt{\det g_\Psi}q)\right) = 0,$$

$\forall (x, y, t) \in [-a, a]^2 \times [0, T]$, $q = q(x, y, t; p)$. Its integral form is given by:

$$\begin{aligned} Q_{ijp}(t_{n+1}) = Q_{ijp}(t_n) &- \frac{\Delta t}{|\Omega_{ijp}|}\delta_x\left(\frac{1}{\Delta t}\int_{t_1}^{t_2}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (\tilde{u}\sqrt{\det g_\Psi}q)(x_i, y, t; p)\, dy\, dt\right) \\ &- \frac{\Delta t}{|\Omega_{ijp}|}\delta_y\left(\frac{1}{\Delta t}\int_{t_1}^{t_2}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\tilde{v}\sqrt{\det g_\Psi}q)(x, y_j, t; p)\, dx\, dt\right), \end{aligned}$$

Hence, we can use the dimension splitting presented in Chapter 3 to the variable $\sqrt{\det g_\Psi}q$. However, when computing the stencils near to the cube edges, we need to approximate the values of q in the ghost cells in order to compute the stencils.

Appendix A

Numerical Analysis

A.1 Finite-difference estimates

This Section aims to prove all finite-difference error estimations used throughout this text. All the proves are very simple and consist of applying Taylor's expansions, as it is usual when computing the accuracy order of many numerical schemes.

Lemma A.1. *Let $F \in C^5(\mathbb{R})$, $x_0 \in \mathbb{R}$ and $h > 0$. Then, the following identity holds:*

$$F'(x_0) = \frac{4}{3}\left(\frac{F(x_0+h) - F(x_0-h)}{2h}\right) - \frac{1}{3}\left(\frac{F(x_0+2h) - F(x_0-2h)}{4h}\right) + C_1 h^4, \qquad \text{(A.1)}$$

where C_1 is a constant that depends only on F and h.

Proof. Given $\delta \in]0, 2h]$, then $x_0 + \delta \in]x_0, x_0 + 2h]$ and $x_0 - \delta \in]x_0 - 2h, x_0]$. Then, we get using the Taylor expansion of F:

$$F(x_0+\delta) = F(x_0) + F'(x_0)\delta + F^{(2)}(x_0)\frac{\delta^2}{2} + F^{(3)}(x_0)\frac{\delta^3}{3!} + F^{(4)}(x_0)\frac{\delta^3}{4!} + F^{(5)}(\theta_\delta)\frac{\delta^5}{5!} \quad \theta_\delta \in [x_0, x_0+\delta],$$
$$F(x_0-\delta) = F(x_0) - F'(x_0)\delta + F^{(2)}(x_0)\frac{\delta^2}{2} - F^{(3)}(x_0)\frac{\delta^3}{3!} + F^{(4)}(x_0)\frac{\delta^4}{4!} - F^{(5)}(\theta_{-\delta})\frac{\delta^5}{5!}, \quad \theta_{-\delta} \in [x_0-\delta, x_0].$$

Thus:

$$\frac{F(x_0+\delta) - F(x_0-\delta)}{2\delta} = F'(x_0) + F^{(3)}(x_0)\frac{\delta^2}{3!} + \left(F^{(5)}(\theta_\delta) + F^5(\theta_{-\delta})\right)\frac{\delta^4}{2\cdot 5!}, \qquad \text{(A.2)}$$

Applying Equation (A.2) for $\delta = h$ and $\delta = 2h$, we get, respectively:

$$\frac{F(x_0+h) - F(x_0-h)}{2h} = F'(x_0) + F^{(3)}(x_0)\frac{h^2}{3!} + \left(F^{(5)}(\theta_h) + F^{5)}(\theta_{-h})\right)\frac{h^4}{2\cdot 5!}, \quad \theta_h \in [x_0, x_0+h], \quad \theta_{-h} \in [x_0-h, x_0], \qquad \text{(A.3)}$$

and

$$\frac{F(x_0+2h)-F(x_0-2h)}{4h} = F'(x_0) + F^{(3)}(x_0)\frac{4h^2}{3!} + \left(F^{(5)}(\theta_{2h}) + F^{(5)}(\theta_{-2h})\right)\frac{16h^4}{2\cdot 5!}, \quad \theta_{2h} \in [x_0, x_0+2h], \quad \theta_{-2h} \in [x_0-2h, x_0]. \tag{A.4}$$

Using Equations (A.3) and (A.4), we obtain:

$$\frac{4}{3}\left(\frac{F(x_0+h)-F(x_0-h)}{2h}\right) = \frac{4}{3}F'(x_0) + F^{(3)}(x_0)\frac{4h^2}{3\cdot 3!} + \left(F^{(5)}(\theta_h) + F^{(5)}(\theta_{-h})\right)\frac{h^4}{2\cdot 5!}, \tag{A.5}$$

$$\frac{1}{3}\left(\frac{F(x_0+2h)-F(x_0-2h)}{4h}\right) = \frac{1}{3}F'(x_0) + F^{(3)}(x_0)\frac{4h^2}{3\cdot 3!} + \left(F^{(5)}(\theta_{2h}) + F^{(5)}(\theta_{-2h})\right)\frac{16h^4}{3\cdot 2\cdot 5!} \tag{A.6}$$

Subtracting Equation (A.6) from Equation (A.5) we get the desired Equation (A.1) with

$$C_1 = \frac{1}{720}\left(3F^{(5)}(\theta_h) + 3F^{(5)}(\theta_{-h}) - 16F^{(5)}(\theta_{2h}) - 16F^{(5)}(\theta_{-2h})\right), \tag{A.7}$$

where $\theta_h \in [x_0, x_0+h], \theta_{-h} \in [x_0-h, x_0]$, $\theta_{2h} \in [x_0, x_0+2h], \theta_{-2h} \in [x_0-2h, x_0]$. Using the intermediate value theorem, we can express C_1 in a more compact way as

$$C_1 = \frac{1}{720}\left(6F^{(5)}(\eta_1) - 32F^{(5)}(\eta_2)\right), \tag{A.8}$$

where $\eta_1, \eta_2 \in [x_0-2h, x_0+2h]$, which concludes the proof. □

Lemma A.2. *Let $F \in C^4(\mathbb{R})$, $x_0 \in \mathbb{R}$ and $h > 0$. Then, the following identity holds:*

$$F''(x_0) = \frac{-2F(x_0-2h) + 15F(x_0-h) - 28F(x_0) + 20F(x_0+h) - 6F(x_0+2h) + F(x_0+3h)}{6h^2} + C_2h^2, \tag{A.9}$$

where C_2 is a constant that depends only on F and h.

Proof. From the Taylor's expansion, we have:

$$
\begin{aligned}
F(x_0-2h) &= F(x_0) - 2F'(x_0)h + 2F^{(2)}(x_0)h^2 - \frac{8}{6}F^{(3)}(x_0)h^3 + \frac{16}{24}F^{(4)}(\theta_{-2h})h^4,\\
F(x_0-h) &= F(x_0) - F'(x_0)h + \frac{1}{2}F^{(2)}(x_0)h^2 - \frac{1}{6}F^{(3)}(x_0)h^3 + \frac{1}{24}F^{(4)}(\theta_{-h})h^4,\\
F(x_0+h) &= F(x_0) + F'(x_0)h + \frac{1}{2}F^{(2)}(x_0)h^2 + \frac{1}{6}F^{(3)}(x_0)h^3 + \frac{1}{24}F^{(4)}(\theta_{h})h^4,\\
F(x_0+2h) &= F(x_0) + 2F'(x_0)h + 2F^{(2)}(x_0)h^2 + \frac{8}{6}F^{(3)}(x_0)h^3 + \frac{16}{24}F^{(4)}(\theta_{2h})h^4,\\
F(x_0+3h) &= F(x_0) + 3F'(x_0)h + \frac{9}{2}F^{(2)}(x_0)h^2 + \frac{27}{6}F^{(3)}(x_0)h^3 + \frac{81}{24}F^{(4)}(\theta_{3h})h^4,
\end{aligned}
$$

where $\theta_{-2h} \in [x_0-2h, x_0-h], \theta_{-h} \in [x_0-h, x_0], \theta_h \in [x_0, x_0+h], \theta_{2h} \in [x_0+h, x_0+2h], \theta_{3h} \in [x_0+2h, x_0+3h]$. Multiplying these equations by their respective coefficients given in Equation (A.9), one get:

$$
\begin{aligned}
-2F(x_0-2h) &= -2F(x_0) + 4F'(x_0)h - 4F^{(2)}(x_0)h^2 + \frac{16}{6}F^{(3)}(x_0)h^3 - \frac{32}{24}F^{(4)}(\theta_{-2h})h^4,\\
15F(x_0-h) &= 15F(x_0) - 15F'(x_0)h + \frac{15}{2}F^{(2)}(x_0)h^2 - \frac{15}{6}F^{(3)}(x_0)h^3 + \frac{15}{24}F^{(4)}(\theta_{-h})h^4,\\
-28F(x_0) &= -28F(x_0),\\
20F(x_0+h) &= 20F(x_0) + 20F'(x_0)h + 10F^{(2)}(x_0)h^2 + \frac{20}{6}F^{(3)}(x_0)h^3 + \frac{20}{24}F^{(4)}(\theta_{h})h^4,\\
-6F(x_0+2h) &= -6F(x_0) - 12F'(x_0)h - 12F^{(2)}(x_0)h^2 - 8F^{(3)}(x_0)h^3 - \frac{96}{24}F^{(4)}(\theta_{2h})h^4,\\
F(x_0+3h) &= F(x_0) + 3F'(x_0)h + \frac{9}{2}F^{(2)}(x_0)h^2 + \frac{27}{6}F^{(3)}(x_0)h^3 + \frac{81}{24}F^{(4)}(\theta_{3h})h^4.
\end{aligned}
$$

Summing all these equations, we get the desired Formula (A.9) with C_2 given by:

$$
C_2 = \frac{1}{24}\left(32F^{(4)}(\theta_{-2h}) - 15F^{(4)}(\theta_{-h}) - 20F^{(4)}(\theta_h) + 96F^{(4)}(\theta_{2h}) - 81F^{(4)}(\theta_{3h})\right). \tag{A.10}
$$

Using the intermediate value theorem, we can express C_2 in a more compact way as

$$
C_2 = \frac{1}{24}\left(128F^{(5)}(\eta_1) - 116F^{(5)}(\eta_2)\right), \tag{A.11}
$$

where $\eta_1, \eta_2 \in [x_0-2h, x_0+3h]$, which concludes the proof. □

Lemma A.3. *Let $F \in C^4(\mathbb{R})$, $x_0 \in \mathbb{R}$ and $h > 0$. Then, the following identity holds:*

$$F^{(3)}(x_0) = \frac{F(x_0-2h) - 7F(x_0-h) + 16F(x_0) - 16F(x_0+h) + 7F(x_0+2h) - F(x_0+3h)}{2h^3} + C_3 h, \tag{A.12}$$

where C_3 is a constant that depends only on F and h.

Proof. From the Taylor's expansion, we have:

$$\begin{aligned}
F(x_0-2h) &= F(x_0) - 2F'(x_0)h + 2F^{(2)}(x_0)h^2 - \frac{8}{6}F^{(3)}(x_0)h^3 + \frac{16}{24}F^{(4)}(\theta_{-2h})h^4,\\
F(x_0-h) &= F(x_0) - F'(x_0)h + \frac{1}{2}F^{(2)}(x_0)h^2 - \frac{1}{6}F^{(3)}(x_0)h^3 + \frac{1}{24}F^{(4)}(\theta_{-h})h^4,\\
F(x_0+h) &= F(x_0) + F'(x_0)h + \frac{1}{2}F^{(2)}(x_0)h^2 + \frac{1}{6}F^{(3)}(x_0)h^3 + \frac{1}{24}F^{(4)}(\theta_{h})h^4,\\
F(x_0+2h) &= F(x_0) + 2F'(x_0)h + 2F^{(2)}(x_0)h^2 + \frac{8}{6}F^{(3)}(x_0)h^3 + \frac{16}{24}F^{(4)}(\theta_{2h})h^4,\\
F(x_0+3h) &= F(x_0) + 3F'(x_0)h + \frac{9}{2}F^{(2)}(x_0)h^2 + \frac{27}{6}F^{(3)}(x_0)h^3 + \frac{81}{24}F^{(4)}(\theta_{3h})h^4,
\end{aligned}$$

where $\theta_{-2h} \in [x_0-2h, x_0-h], \theta_{-h} \in [x_0-h, x_0], \theta_h \in [x_0, x_0+h], \theta_{2h} \in [x_0+h, x_0+2h], \theta_{3h} \in [x_0+2h, x_0+3h]$. Multiplying these equations by their respective coefficients given in Equation (A.12), one get:

$$\begin{aligned}
F(x_0-2h) &= F(x_0) - 2F'(x_0)h + \frac{4}{2}F^{(2)}(x_0)h^2 - \frac{8}{6}F^{(3)}(x_0)h^3 + \frac{16}{24}F^{(4)}(\theta_{-2h})h^4,\\
-7F(x_0-h) &= -7F(x_0) + 7F'(x_0)h - \frac{7}{2}F^{(2)}(x_0)h^2 + \frac{7}{6}F^{(3)}(x_0)h^3 - \frac{7}{24}F^{(4)}(\theta_{-h})h^4,\\
16F(x_0) &= 16F(x_0),\\
-16F(x_0+h) &= -16F(x_0) - 16F'(x_0)h - \frac{16}{2}F^{(2)}(x_0)h^2 - \frac{16}{6}F^{(3)}(x_0)h^3 - \frac{16}{24}F^{(4)}(\theta_{h})h^4,\\
7F(x_0+2h) &= 7F(x_0) + 14F'(x_0)h + \frac{28}{2}F^{(2)}(x_0)h^2 + \frac{56}{6}F^{(3)}(x_0)h^3 + \frac{112}{24}F^{(4)}(\theta_{2h})h^4,\\
-F(x_0+3h) &= -F(x_0) - 3F'(x_0)h - \frac{9}{2}F^{(2)}(x_0)h^2 - \frac{27}{6}F^{(3)}(x_0)h^3 - \frac{81}{24}F^{(4)}(\theta_{3h})h^4.
\end{aligned}$$

Summing all these equations, we have:

$$F(x_0-2h) - 7F(x_0-h) + 16F(x_0) - 16F(x_0+h) + 7F(x_0+2h) - F(x_0+3h) = 2F^{(3)}(x_0)h^3 - 2C_3h^4,$$

we get the desired Formula (A.12) with C_3 given by:

$$C_3 = \frac{1}{48}\left(-16F^{(4)}(\theta_{-2h}) + 7F^{(4)}(\theta_{-h}) + 16F^{(4)}(\theta_h) - 112F^{(4)}(\theta_{2h}) + 81F^{(4)}(\theta_{3h}) \right). \tag{A.13}$$

Using the intermediate value theorem, we can express C_3 in a more compact way as

$$C_3 = \frac{1}{48}\left(104F^{(5)}(\eta_1) - 128F^{(5)}(\eta_2)\right), \tag{A.14}$$

where $\eta_1, \eta_2 \in [x_0 - 2h, x_0 + 3h]$, which concludes the proof. □

A.2 Lagrange interpolation

Given real numbers, called nodes, $x_0 < x_1 < \cdots < x_m$, we define the k-th Lagrange polynomial by

$$L_k(x) = \prod_{j=0, j\neq k}^{m} \frac{x - x_j}{x_k - x_j}.$$

They satisfy $L_k(x_j) = \delta_{kj}$, where δ_{kj} is the Kronecker delta. Given a function f defined at the nodes x_j, its interpolating polynomial of degree m is given by:

$$P_m(x) = \sum_{k=0}^{m} f(x_k)L_k(x).$$

Indeed, this polynomial interpolates f since $P_m(x_j) = f(x_j)$. It is well known that P_m always exists and is unique. Besides that, we have the following error formula for Lagrange interpolation.

Theorem A.1. *Let $f \in C^{m+1}(\mathbb{R})$. Then, then there is ξ in the smallest interval containing $x_0, \cdots, x_m, x$ such that:*

$$f(x) - P_m(x) = \omega(x)\frac{f^{(m+1)}(\xi)}{(m+1)!}, \tag{A.15}$$

where $\omega(x) = (x - x_0)(x - x_1)\cdots(x - x_m)$.

Proof. See Stoer and Bulirsch (2002, Theorem 2.1.4.1. on p. 49). □

A.3 Numerical integration

The following mean value theorem for integrals is a very useful tool when working with numerical integration errors.

Theorem A.2 (Mean value theorem for integrals)**.** *If $f \in C([a, b])$, and g is a integrable function in $[a, b]$ whose sign does not change in $[a, b]$, then there exists $c \in]a, b[$ such that*

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx.$$

Proof. See Courant and John (1999, p. 143). □

A.3.1 Multi-step schemes

Let us consider the following problem: given a function $f \in C^{m+1}([0,T])$, a discretization of $[0,T]$ given by $t^n = n\Delta t$, $\Delta t = \frac{T}{N_T}$, for some $N_T \in \mathbb{N}$, we wish to estimate $\int_{t^n}^{t^{n+1}} f(t)\,dt$ using the values $f(t_{n-k})$, for $k = 0, \cdots, m$. This kind of problem arises, for instance, when we are interested in computing departure points as in Equation 2.74. We can estimate the desired integral by computing the interpolating polynomial of $f(t_{n-k})$, for $k = 0, \cdots, m$ and then integrating this polynomial. This approach is exactly what is used in multi-step Adams-Bashforth methods. On the next theorem, we give an expression the error of this approach.

Theorem A.3. *If $f \in C^{m+1}([0,T])$, $t^n = n\Delta t$, $n = 0, \cdots, N_T$, $\Delta t = \frac{T}{N_T}$ for some $N_T \in \mathbb{N}$, then:*

$$\int_{t^n}^{t^{n+1}} f(t)\,dt = \Delta t \sum_{k=0}^{m} \left(\int_0^1 L_k(s)\,ds \right) f(t_{n-k}) + \frac{(\Delta t)^{k+1}}{(m+1)!} f^{(m+1)}(\eta) \int_0^1 \omega(s)\,ds, \tag{A.16}$$

where $w(s) = s(s+1)\cdots(s+m)$, $\eta \in [t^{n-m}, t^n]$.

Proof. We introduce auxiliary functions $\theta(s) = s\Delta t + t_n$, $s \in [-m, 1]$ and $g(s) = f(\theta(s))$. It is clear that $f(t_{n-k}) = g(-k)$, for $k = -1, 0, \cdots, m$. Hence, we can write:

$$\int_{t^n}^{t^{n+1}} f(t)\,dt = \Delta t \int_0^1 f(\theta(s))\,ds = \Delta t \int_0^1 g(s)\,ds. \tag{A.17}$$

Defining the nodes $s_k = -k$ for $k = 0, \cdots, m$, it follows from Theorem A.1 that the interpolating polynomial P_m of $g(s_k)$ satisfies:

$$g(s) - P_m(s) = \omega(s) \frac{g^{(m+1)}(\xi)}{(m+1)!}, \tag{A.18}$$

where $\xi \in [-m, 1]$. Substituting Equation (A.18) in Equation (A.17), we obtain

$$\int_{t^n}^{t^{n+1}} f(t)\,dt = \Delta t \sum_{k=0}^{m} \left(\int_0^1 L_k(s)\,ds \right) g(-k) + \frac{\Delta t}{(m+1)!} \int_0^1 g^{(m+1)}(\xi)\omega(s)\,ds. \tag{A.19}$$

Since $w(s)$ does not change its sign in $[0,1]$ it follows from Theorem A.2 that:

$$\int_{t^n}^{t^{n+1}} f(t)\,dt = \Delta t \sum_{k=0}^{m} \left(\int_0^1 L_k(s)\,ds \right) g(-k) + \frac{\Delta t}{(m+1)!} g^{(m+1)}(\overline{\xi}) \int_0^1 \omega(s)\,ds, \tag{A.20}$$

for some $\overline{\xi} \in [-m, 1]$. Notice that by the chain rule we get $g^{(m+1)}(s) = (\Delta t)^k f^{(m+1)}(\theta(s))$, therefore Equation (A.20) in terms of f reads:

$$\int_{t^n}^{t^{n+1}} f(t)\,dt = \Delta t \sum_{k=0}^{m} \left(\int_0^1 L_k(s)\,ds \right) f(t_{n-k}) + \frac{(\Delta t)^{k+1}}{(m+1)!} f^{(m+1)}(\eta) \int_0^1 \omega(s)\,ds, \tag{A.21}$$

where $\eta \in [t^{n-m}, t^n]$, which is the desired identity. □

In the following corollaries, we give the explicit formulas for Equation (A.21) for $m = 0, m = 1, m = 2$. This is achieved by computing the terms $\int_0^1 L_k(s)\,ds$ and $\int_0^1 \omega(s)\,ds$, which are trivial to be computed.

Corollary A.1. *If $f \in C^1([0,T])$, $t^n = n\Delta t$, $n = 0, \cdots, N_T$, $\Delta t = \frac{T}{N_T}$ for some $N_T \in \mathbb{N}$, then:*

$$\int_{t^n}^{t^{n+1}} f(t)\,dt = \Delta t f(t_n) + \frac{\Delta t^2}{2} f'(\bar{t}), \tag{A.22}$$

for some $\bar{t} \in [t^n, t^{n+1}]$.

Corollary A.2. *If $f \in C^2([0,T])$, $t^n = n\Delta t$, $n = 0, \cdots, N_T$, $\Delta t = \frac{T}{N_T}$ for some $N_T \in \mathbb{N}$, then:*

$$\int_{t^n}^{t^{n+1}} f(t)\,dt = \frac{\Delta t}{2}(3f(t_n) - f(t_{n-1})) + \frac{5\Delta t^3}{12} f^{(2)}(\bar{t}), \tag{A.23}$$

for some $\bar{t} \in [t^{n-1}, t^{n+1}]$.

Corollary A.3. *If $f \in C^3([0,T])$, $t^n = n\Delta t$, $n = 0, \cdots, N_T$, $\Delta t = \frac{T}{N_T}$ for some $N_T \in \mathbb{N}$, then:*

$$\int_{t^n}^{t^{n+1}} f(t)\,dt = \frac{\Delta t}{12}(23f(t_n) - 16f(t_{n-1}) + 5f(t_{n-2})) + \frac{3\Delta t^4}{8} f^{(3)}(\bar{t}), \tag{A.24}$$

for some $\bar{t} \in [t^{n-2}, t^{n+1}]$.

When using these schemes for and ODE written in its integral form, $m = 0$ gives the classical Euler method; for $m = 1$ we get the second-order Adams-Bashforth scheme and for $m = 2$ we have the third-order Adams-Bashforth scheme.

A.3.2 Midpoint rule

When considering finite-volume schemes, it is useful to compare the average value on a control volume of a function with its value at the control volume centroid. In the following theorems, for the one and two dimensional cases, respectively, we show that the value of a function at the centroid of a control volume given a second-order approximation to its average value on the control volume.

Theorem A.4. *If $f \in C^2([x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}])$, then*

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx - f(x_i) = C_1 \Delta x^2, \tag{A.25}$$

where C_1 is a constant that depends only on f, and $x_i = \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2}$, $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$.

Proof. From Taylor's expansion, it follows that, for $x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, we have:

$$f(x) = f(x_i) + f(x_i)(x - x_i) + f''(\xi)\frac{(x - x_i)^2}{2}, \tag{A.26}$$

for some ξ between x and x_i. Therefore:

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx - f(x_i) = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(f'(x_i)(x-x_i) + f''(\xi)\frac{(x-x_i)^2}{2}\right) dx$$
$$= \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f''(\xi)\frac{(x-x_i)^2}{2}\,dx.$$

Using the mean value theorem for integrals (see Theorem A.2), we have:

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx - f(x_i) = f''(\eta)\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{(x-x_i)^2}{2}\,dx = f''(\eta)\frac{\Delta x^2}{24}$$

for some $\eta \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, from which the proposition follows with

$$C_1 = \frac{1}{24}f''(\eta). \tag{A.27}$$

□

Theorem A.5. *If $f \in C^2([x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}])$, then*

$$\left|\frac{1}{\Delta x \Delta y}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(x,y)\,dx\,dy - f(x_i,y_j)\right| \le C_1\Delta x^2 + C_2\Delta x\Delta y + C_3\Delta y^2, \tag{A.28}$$

where C_1, C_2 and C_3 are constants that depend only on f, and $x_i = \frac{x_{i+\frac{1}{2}}+x_{i-\frac{1}{2}}}{2}$, $y_i = \frac{y_{j+\frac{1}{2}}+y_{j-\frac{1}{2}}}{2}$, $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, $\Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$.

Proof. From Taylor's expansion, it follows that, for $x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, $y \in [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, we have:

$$f(x,y) = f(x_i,y_j) + \frac{\partial f}{\partial x}(x_i,y_j)(x-x_i) + \frac{\partial f}{\partial y}(x_i,y_j)(y-y_j)$$
$$+ \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}(\xi,\theta)(x-x_i)^2 + 2\frac{\partial^2 f}{\partial x\partial y}(\xi,\theta)(x-x_i)(y-y_j)\frac{\partial^2 f}{\partial y^2}(\xi,\theta)(y-y_j)^2\right)$$

for some ξ between x and x_i, and θ between y and y_j. Therefore:

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(x,y)\,dy\,dx - \Delta x\Delta y f(x_i,y_j) = \frac{1}{2}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial^2 f}{\partial x^2}(\xi,\theta)(x-x_i)^2\,dy\,dx +$$
$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial^2 f}{\partial x\partial y}(\xi,\theta)(x-x_i)(y-y_j)\,dy\,dx + \frac{1}{2}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial^2 f}{\partial y^2}(\xi,\theta)(y-y_j)^2\,dy\,dx$$

Using the mean value theorem for integrals (see Theorem A.2), we have:

$$\frac{1}{\Delta x \Delta y}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(x,y)\,dx\,dy - f(x_i,y_j) = \frac{\partial^2 f}{\partial x^2}(\eta_1,\lambda_1)\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\frac{(x-x_i)^2}{2\Delta x\Delta y}\,dx\,dy$$
$$+\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\frac{\partial^2 f}{\partial x\partial y}(\xi,\theta)\frac{(x-x_i)(y-y_j)}{\Delta x\Delta y}\,dx\,dy + \frac{\partial^2 f}{\partial x^2}(\eta_2,\lambda_2)\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\frac{(y-y_j)^2}{2\Delta x\Delta y}\,dx\,dy$$
$$=\frac{\partial^2 f}{\partial x^2}(\eta_1,\lambda_1)\frac{\Delta x^2}{24}+\frac{\partial^2 f}{\partial y^2}(\eta_2,\lambda_2)\frac{\Delta y^2}{24}+\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\frac{\partial^2 f}{\partial x\partial y}(\xi,\theta)\frac{(x-x_i)(y-y_j)}{\Delta x\Delta y}\,dx\,dy$$

for $\eta_1,\eta_2 \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, $\lambda_1,\lambda_2 \in [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, from which the proposition follows. □

Appendix B

Spherical coordinates and geometry

Given $R > 0$, we denote the sphere of radius R centered at the origin of $\mathbb{R}^3$:

$$\mathbb{S}^2_R = \{(X, Y, Z) \in \mathbb{R}^3 \ : \ X^2 + Y^2 + Z^2 = R^2\}.$$

The tangent space at $P \in \mathbb{S}^2_R$ by $T_P\mathbb{S}^2$. It is easy to see that:

$$T_P\mathbb{S}^2_R = \{Q \in \mathbb{R}^3 \ : \ \langle P, Q\rangle = 0\},$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product of $\mathbb{R}^3$. The tangent bundle is denoted by:

$$T\mathbb{S}^2_R = \bigcup_{P \in \mathbb{S}^R_2} T_P\mathbb{S}^2_R.$$

We are going to consider three ways to represent an element of $\mathbb{S}^2_R$: using (X, Y, Z) coordinates, or using (λ, ϕ) latitude-longitude coordinates, or, at last, using the cubed-sphere coordinates (x, y, p), where (x, y) are the cube face coordinates and $p \in \{1, 2, \cdots, 6\}$ stands for a cube panel, as presented in Chapter 4.

B.1 Conversions between latitude-longitude and contravariant coordinates

We consider the latitude-longitude mapping $\Psi_{ll} \ : \ [0, 2\pi] \times [-\frac{\pi}{2}, \frac{\pi}{2}] \ \rightarrow \ \mathbb{S}^2_R$, given by:

$$X(\lambda, \phi) = R \cos\phi \cos\lambda, \tag{B.1}$$
$$Y(\lambda, \phi) = R \cos\phi \sin\lambda, \tag{B.2}$$
$$Z(\lambda, \phi) = R \sin\phi, \tag{B.3}$$

The derivative or Jacobian matrix of the mapping Ψ_{ll} is given by:

$$D\Psi_{ll}(\lambda,\phi) = R\begin{bmatrix} -\cos\phi\sin\lambda & -\sin\phi\cos\lambda \\ \cos\phi\cos\lambda & \sin\phi\sin\lambda \\ 0 & \cos\phi \end{bmatrix} \tag{B.4}$$

Using this matrix columns, we can define the tangent vectors:

$$\boldsymbol{g}_\lambda(\lambda,\phi) = D\Psi_{ll}(\lambda,\phi)\begin{bmatrix}1\\0\end{bmatrix}, \quad \boldsymbol{g}_\phi(\lambda,\phi) = D\Psi_{ll}(\lambda,\phi)\begin{bmatrix}0\\1\end{bmatrix}, \tag{B.5}$$

We normalize the vectors $\boldsymbol{g}_\lambda$ and $\boldsymbol{g}_\phi$ and we obtain unit tangent vectors on the sphere at $\Phi_{ll}(\lambda,\phi)$:

$$\boldsymbol{e}_\lambda(\lambda,\phi) = \begin{bmatrix}-\sin\lambda\\ \cos\lambda\\ 0\end{bmatrix}, \quad \boldsymbol{e}_\phi(\lambda,\phi) = \begin{bmatrix}-\sin\phi\cos\lambda\\ -\sin\phi\sin\lambda\\ \cos\phi\end{bmatrix}, \tag{B.6}$$

Let us consider a tangent vector field $\boldsymbol{u} : \mathbb{S}^2_R \to T\mathbb{S}^2_R$ on the sphere, *i.e.*, $\boldsymbol{u}(P) \in T_P\mathbb{S}^2_R$, $\forall P \in \mathbb{S}^2_R$. We may express this vector fields in latitude-longitude coordinates as:

$$\boldsymbol{u}(\lambda,\phi) = u_\lambda(\lambda,\phi)\boldsymbol{e}_\lambda(\lambda,\phi) + v_\phi(\lambda,\phi)\boldsymbol{e}_\phi(\lambda,\phi). \tag{B.7}$$

Or, we may also represent this vector field using the basis obtained by cubed-sphere coordinates:

$$\boldsymbol{u}(x,y;p) = \tilde{u}(x,y;p)\boldsymbol{e}_x(x,y;p) + \tilde{v}(x,y;p)\boldsymbol{e}_y(x,y;p). \tag{B.8}$$

This representation is known as contravariant representation. In order to relate the latitude-longitude representation with the contravariant representation, we notice that:

$$\boldsymbol{e}_x(x,y;p) = \langle \boldsymbol{e}_x, \boldsymbol{e}_\lambda\rangle \boldsymbol{e}_\lambda(\lambda,\phi) + \langle \boldsymbol{e}_x, \boldsymbol{e}_\phi\rangle \boldsymbol{e}_\phi(\lambda,\phi), \tag{B.9}$$

$$\boldsymbol{e}_y(x,y;p) = \langle \boldsymbol{e}_y, \boldsymbol{e}_\lambda\rangle \boldsymbol{e}_\lambda(\lambda,\phi) + \langle \boldsymbol{e}_y, \boldsymbol{e}_\phi\rangle \boldsymbol{e}_\phi(\lambda,\phi), \tag{B.10}$$

which holds since the vectors $\boldsymbol{e}_\lambda(\lambda,\phi)$ and $\boldsymbol{e}_\phi(\lambda,\phi)$ are orthogonal. Replacing Equations (B.9) and (B.10) in Equation (B.8), we obtain the values (u_λ, v_ϕ) in terms of the contravariant components $(\tilde{u},\tilde{v})$ as the following matrix equation:

$$\begin{bmatrix} u_\lambda(\lambda,\phi) \\ v_\phi(\lambda,\phi)\end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{e}_x, \boldsymbol{e}_\lambda\rangle & \langle \boldsymbol{e}_y, \boldsymbol{e}_\lambda\rangle \\ \langle \boldsymbol{e}_x, \boldsymbol{e}_\phi\rangle & \langle \boldsymbol{e}_y, \boldsymbol{e}_\phi\rangle \end{bmatrix} \begin{bmatrix} \tilde{u}(x,y;p) \\ \tilde{v}(x,y;p)\end{bmatrix}. \tag{B.11}$$

Conversely, we may express the contravariant components in terms of latitude-longitude components by inverting Equation (B.11):

$$\begin{bmatrix} \tilde{u}(x,y;p) \\ \tilde{v}(x,y;p)\end{bmatrix} = \frac{1}{\langle \boldsymbol{e}_x, \boldsymbol{e}_\lambda\rangle\langle \boldsymbol{e}_y, \boldsymbol{e}_\phi\rangle - \langle \boldsymbol{e}_y, \boldsymbol{e}_\lambda\rangle\langle \boldsymbol{e}_x, \boldsymbol{e}_\phi\rangle} \begin{bmatrix} \langle \boldsymbol{e}_y, \boldsymbol{e}_\phi\rangle & -\langle \boldsymbol{e}_y, \boldsymbol{e}_\lambda\rangle \\ -\langle \boldsymbol{e}_x, \boldsymbol{e}_\phi\rangle & \langle \boldsymbol{e}_x, \boldsymbol{e}_\lambda\rangle \end{bmatrix} \begin{bmatrix} u_\lambda(\lambda,\phi) \\ v_\phi(\lambda,\phi)\end{bmatrix}. \tag{B.12}$$

B.2 Covariant/contravariant conversion

Given Equation Let us consider again a tangent vector field $\boldsymbol{u} : \mathbb{S}^2_R \to T\mathbb{S}^2_R$ on the sphere, the contravariant representation of $\boldsymbol{u}$ is given by Equation (B.8). The covariant components (u, v) are given by:

$$u(x, y; p) = \langle \boldsymbol{u}(x, y; p), e_x(x, y; p) \rangle, \tag{B.13}$$

$$v(x, y; p) = \langle \boldsymbol{u}(x, y; p), e_y(x, y; p) \rangle. \tag{B.14}$$

Replacing Equation (B.8) in Equations (B.13) and (B.14) we obtain the relation covariant components in terms of the contravariant terms:

$$\begin{bmatrix} u(x, y; p) \\ v(x, y; p) \end{bmatrix} = \begin{bmatrix} 1 & \langle \boldsymbol{e}_x, \boldsymbol{e}_y \rangle \\ \langle \boldsymbol{e}_x, \boldsymbol{e}_y \rangle & 1 \end{bmatrix} \begin{bmatrix} \tilde{u}(x, y; p) \\ \tilde{v}(x, y; p) \end{bmatrix}. \tag{B.15}$$

Denoting the angle between $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$ by α, we have $\langle \boldsymbol{e}_x, \boldsymbol{e}_y \rangle = \cos\alpha$. Thus, we may express the contravariant components in terms of the covariant terms inverting Equation (B.15):

$$\begin{bmatrix} \tilde{u}(x, y; p) \\ \tilde{v}(x, y; p) \end{bmatrix} = \frac{1}{\sin^2 \alpha} \begin{bmatrix} 1 & -\cos\alpha \\ -\cos\alpha & 1 \end{bmatrix} \begin{bmatrix} u(x, y; p) \\ v(x, y; p) \end{bmatrix}. \tag{B.16}$$

Notice that combining Equations (B.15) and (B.16) with Equations (B.11) and (B.12) one may get relations between the latitude-longitude components and the covariant components.

Appendix C

Code avaibility

The codes needed for this work have been built openly at GitHub. The PPM implementation for the one-dimensional advection equation used in Chapter 2 is available at https://github.com/luanfs/py-ppm. The dimension-splitting implementation for the advection equation on the plane used in Chapter 3 is available at https://github.com/luanfs/py-dimension-splitting. At last, all the grid tools for the cubed sphere used Chapters 4 and 5, including the finite volume model on this grid, is available in a Python version at https://github.com/luanfs/py-cubed-sphere and in a Fortran 90 version at https://github.com/luanfs/cubed-sphere.

References

Arakawa, A., & Lamb, V. R. (1977). Computational design of the basic dynamical processes of the ucla general circulation model. In *General circulation models of the atmosphere* (pp. 173–265). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-460817-7.50009-4. (Cit. on pp. 4, 13)

Barros, S., Dent, D., Isaksen, L., Robinson, G., Mozdzynski, G., & Wollenweber, F. (1995). The ifs model: A parallel production weather code. *Parallel Computing*, *21*(10), 1621–1638. https://doi.org/https://doi.org/10.1016/0167-8191(96)80002-0 (cit. on p. 3)

Benacchio, T., & Wood, N. (2016). Semi-implicit semi-lagrangian modelling of the atmosphere: A met office perspective. *Communications in Applied and Industrial Mathematics*, *7*(3), 4–25. https://doi.org/doi:10.1515/caim-2016-0020 (cit. on p. 1)

Carpenter, R. L., Droegemeier, K. K., Woodward, P. R., & Hane, C. E. (1990). Application of the piecewise parabolic method (ppm) to meteorological modeling. *Monthly Weather Review*, *118*(3), 586–612. https://doi.org/10.1175/1520-0493(1990)118<0586:AOTPPM>2.0.CO;2 (cit. on pp. 4, 7, 19)

Chen, X. (2021). The lmars based shallow-water dynamical core on generic gnomonic cubed-sphere geometry [e2020MS002280 2020MS002280]. *Journal of Advances in Modeling Earth Systems*, *13*(1), e2020MS002280. https://doi.org/https://doi.org/10.1029/2020MS002280 (cit. on p. 69)

Chen, Y., Weller, H., Pring, S., & Shaw, J. (2017). Comparison of dimensionally split and multi-dimensional atmospheric transport schemes for long time steps. *Quarterly Journal of the Royal Meteorological Society*, *143*(708), 2764–2779. https://doi.org/https://doi.org/10.1002/qj.3125 (cit. on pp. 33, 48, 49, 64)

Colella, P., & Woodward, P. R. (1984). The piecewise parabolic method (ppm) for gas-dynamical simulations. *Journal of Computational Physics*, *54*(1), 174–201. https://doi.org/https://doi.org/10.1016/0021-9991(84)90143-8 (cit. on pp. 4, 7, 19, 21, 22, 27, 28, 31, 32, 39, 48)

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, *19*(90), 297–301. http://www.jstor.org/stable/2003354 (cit. on p. 2)

Courant, R., & John, F. (1999). In *Introduction to calculus and analysis i*. Springer Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-58604-0. (Cit. on p. 87)

Croisille, J.-P. (2013). Hermitian compact interpolation on the cubed-sphere grid. *Journal of Scientific Computing*, *57*. https://doi.org/10.1007/s10915-013-9702-3 (cit. on p. 69)

Csomós, P., Faragó, I., & Havasi, Á. (2005). Weighted sequential splittings and their analysis [Numerical Methods and Computational Mechanics]. *Computers and Mathematics with Applications*, *50*(7), 1017–1031. https://doi.org/https://doi.org/10.1016/j.camwa.2005.08.004 (cit. on p. 58)

Dennis, J., Edwards, J., Evans, K., Guba, O., Lauritzen, P., Mirin, A., St-Cyr, A., Taylor, M., & Worley, P. (2012). Cam-se: A scalable spectral element dynamical core for the community atmosphere model. *Internat. J. High Perf. Comput. Appl.*, *26*, 74–89. https://doi.org/10.1177/1094342011428142 (cit. on p. 4)

Durran, D. (2011). Time discretization: Some basic approaches. In *Numerical techniques for global atmospheric models* (pp. 75–104). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11640-7_5. (Cit. on p. 31)

Durran, D. R. (2010). Semi-lagrangian methods. In *Numerical methods for fluid dynamics: With applications to geophysics* (pp. 357–391). Springer New York. https://doi.org/10.1007/978-1-4419-6412-0_7. (Cit. on p. 32)

Eliasen, E., Machenhauer, B., & Rasmussen, E. (1970). On a numerical method for integration of the hydrodynamical equations wirh a spectral representation of the horizontal fields. https://doi.org/10.13140/RG.2.2.13894.88645 (cit. on p. 2)

Engwirda, D., & Kelley, M. (2016). A weno-type slope-limiter for a family of piecewise polynomial methods. https://doi.org/10.48550/ARXIV.1606.08188. (Cit. on pp. 7, 13, 21)

Figueroa, S., Bonatti, J., Kubota, P., Grell, G., Morrison, H., R. M. Barros, S., Fernandez, J., Ramirez-Gutierrez, E., Siqueira, L., Luzia, G., Silva, J., Silva, J., Pendharkar, J., Capistrano, V., Alvim, D., Enore, D., Diniz, F., Satyamurty, P., Cavalcanti, I., & Panetta, J. (2016). The brazilian global atmospheric model (bam): Performance for tropical rainfall forecasting and sensitivity to convective scheme and horizontal resolution. *Weather Forecast.*, *31(5)*, 1547–1572. https://doi.org/10.1175/WAF-D-16-0062.1 (cit. on p. 3)

Giraldo, F. X., Kelly, J. F., & Constantinescu, E. M. (2013). Implicit-explicit formulations of a three-dimensional nonhydrostatic unified model of the atmosphere (numa). *SIAM Journal on Scientific Computing*, *35*(5), B1162–B1194. https://doi.org/10.1137/120876034 (cit. on p. 4)

Godunov, S. (1959). A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb.*, *47(89):3*, 271–306 (cit. on pp. 7, 21).

Guo, W., Nair, R. D., & Qiu, J.-M. (2014). A conservative semi-lagrangian discontinuous galerkin scheme on the cubed sphere. *Monthly Weather Review*, *142*(1), 457–475. https://doi.org/10.1175/MWR-D-13-00048.1 (cit. on p. 31)

Harris, L., Chen, X., Putman, W., Zhou, L., & Chen, J.-H. (2021). A scientific description of the gfdl finite-volume cubed-sphere dynamical core. *Series : NOAA technical memorandum OAR GFDL ; 2021-001*. https://doi.org/https://doi.org/10.25923/6nhs-5897 (cit. on pp. 7, 27)

Harris, L. M., & Lin, S.-J. (2013). A two-way nested global-regional dynamical core on the cubed-sphere grid. *Monthly Weather Review*, *141*(1), 283–306. https://doi.org/10.1175/MWR-D-11-00201.1 (cit. on pp. 4, 5)

Jia, H., & Li, K. (2011). A third accurate operator splitting method. *Mathematical and Computer Modelling*, *53*(1), 387–396. https://doi.org/https://doi.org/10.1016/j.mcm.2010.09.005 (cit. on p. 58)

Katta, K. K., Nair, R. D., & Kumar, V. (2015a). High-order finite volume shallow water model on the cubed-sphere: 1d reconstruction scheme. *Applied Mathematics and Computation*, *266*, 316–327. https://doi.org/https://doi.org/10.1016/j.amc.2015.04.053 (cit. on p. 69)

Katta, K. K., Nair, R. D., & Kumar, V. (2015b). High-order finite-volume transport on the cubed sphere: Comparison between 1d and 2d reconstruction schemes. *Monthly Weather Review*, *143*(7), 2937–2954. https://doi.org/https://doi.org/10.1175/MWR-D-13-00176.1 (cit. on p. 69)

Kent, J., Melvin, T., & Wimmer, G. A. (2022). A mixed finite element discretisation of the shallow water equations. *Geoscientific Model Development Discussions*, *2022*, 1–17. https://doi.org/10.5194/gmd-2022-225 (cit. on p. 4)

Krishnamurti, T., Hardiker, V., Bedi, H., & Ramaswamy, L. (2006). *An introduction to global spectral modeling* (Vol. 35). https://doi.org/10.1007/0-387-32962-5. (Cit. on p. 2)

Lauritzen, P. H., Ullrich, P. A., & Nair, R. D. (2011). Atmospheric transport schemes: Desirable properties and a semi-lagrangian view on finite-volume discretizations. In P. Lauritzen, C. Jablonowski, M. Taylor, & R. Nair (Eds.), *Numerical techniques for global atmospheric models* (pp. 185–250). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11640-7_8. (Cit. on p. 7)

LeVeque, R. J. (1990). *Numerical methods for conservation laws*. Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-5116-9. (Cit. on pp. 8–10, 63)

LeVeque, R. J. (2002). *Finite volume methods for hyperbolic problems*. Cambridge University Press. https://doi.org/10.1017/CBO9780511791253. (Cit. on pp. 8, 11, 15, 17, 18, 20)

Lin, S.-J. (2004). A "vertically lagrangian" finite-volume dynamical core for global models. *Monthly Weather Review*, *132*(10), 2293–2307. https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2 (cit. on pp. 4, 7, 27, 28, 33, 39, 48, 60)

Lin, S.-J., Chao, W. C., Sud, Y. C., & Walker, G. K. (1994). A class of the van leer-type transport schemes and its application to the moisture transport in a general circulation model. *Monthly Weather Review*, *122*(7), 1575–1593. https://doi.org/10.1175/1520-0493(1994)122<1575:ACOTVL>2.0.CO;2 (cit. on p. 4)

Lin, S.-J., & Rood, R. B. (1996). Multidimensional flux-form semi-lagrangian transport schemes. *Monthly Weather Review*, *124*(9), 2046–2070. https://doi.org/10.1175/1520-0493(1996)124<2046:MFFSLT>2.0.CO;2 (cit. on pp. 4, 19, 49, 57–60, 69)

Lin, S.-J., & Rood, R. B. (1997). An explicit flux-form semi-lagrangian shallow-water model on the sphere. *Quarterly Journal of the Royal Meteorological Society*, *123*(544), 2477–2498. https://doi.org/https://doi.org/10.1002/qj.49712354416 (cit. on p. 4)

Lu, F., Zhang, F., Wang, T., Tian, G., & Wu, F. (2022). High-order semi-lagrangian schemes for the transport equation on icosahedron spherical grids. *Atmosphere*, *13*(11). https://doi.org/10.3390/atmos13111807 (cit. on p. 31)

Müller, A., Deconinck, W., Kühnlein, C., Mengaldo, G., Lange, M., Wedi, N., Bauer, P., Smolarkiewicz, P. K., Diamantakis, M., Lock, S.-J., Hamrud, M., Saarinen, S., Mozdzynski, G., Thiemert, D., Glinton, M., Bénard, P., Voitus, F., Colavolpe, C., Marguinaud, P., … New, N. (2019). The escape project: Energy-efficient scalable algorithms for weather prediction at exascale. *Geoscientific Model Development*, *12*(10), 4425–4441. https://doi.org/10.5194/gmd-12-4425-2019 (cit. on p. 3)

Nair, R. D., & Lauritzen, P. H. (2010). A class of deformational flow test cases for linear transport problems on the sphere. *Journal of Computational Physics*, *229*(23), 8868–8887. https://doi.org/https://doi.org/10.1016/j.jcp.2010.08.014 (cit. on pp. 46, 64)

Nair, R. D., Levy, M. N., & Lauritzen, P. H. (2011). Emerging numerical methods for atmospheric modeling. In *Numerical techniques for global atmospheric models* (pp. 251–311). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11640-7_9. (Cit. on p. 10)

Orszag, S. A. (1970). Transform method for the calculation of vector-coupled sums: Application to the spectral form of the vorticity equation. *Journal of Atmospheric Sciences*, *27*(6), 890–895. https://doi.org/10.1175/1520-0469(1970)027<0890:TMFTCO>2.0.CO;2 (cit. on p. 2)

Peixoto, P. (2016). Accuracy analysis of mimetic finite volume operators on geodesic grids and a consistent alternative. *J. Comput. Phys.*, *310*, 127–160. https://doi.org/10.1016/j.jcp.2015.12.058 (cit. on p. 5)

Peixoto, P., & Barros, S. R. M. (2013). Analysis of grid imprinting on geodesic spherical icosahedral grids. *J. Comput. Phys.*, *237*, 61–78. https://doi.org/10.1016/j.jcp.2012.11.041 (cit. on pp. 5, 75)

Putman, W. M. (2007). *Development of the finite-volume dynamical core on the cubed-sphere* (Doctoral dissertation). Florida State University. Florida, US. http://purl.flvc.org/fsu/fd/FSU_migr_etd-0511. (Cit. on p. 4)

Putman, W. M., & Lin, S.-J. (2007). Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, *227*(1), 55–78. https://doi.org/https://doi.org/10.1016/j.jcp.2007.07.022 (cit. on pp. 4, 5, 7, 22, 34, 39, 48, 60, 69, 77, 81)

Rančić, M., Purser, R. J., & Mesinger, F. (1996). A global shallow-water model using an expanded spherical cube: Gnomonic versus conformal coordinates. *Quarterly Journal of the Royal Meteorological Society*, *122*(532), 959–982. https://doi.org/https://doi.org/10.1002/qj.49712253209 (cit. on pp. 69, 73)

Rančić, M. (1992). Semi-lagrangian piecewise biparabolic scheme for two-dimensional horizontal advection of a passive scalar. *Monthly Weather Review*, *120*(7), 1394–1406. https://doi.org/10.1175/1520-0493(1992)120<1394:SLPBSF>2.0.CO;2 (cit. on p. 49)

Rančić, M., Purser, R. J., Jović, D., Vasic, R., & Black, T. (2017). A nonhydrostatic multiscale model on the uniform jacobian cubed sphere. *Monthly Weather Review*, *145*(3), 1083–1105. https://doi.org/10.1175/MWR-D-16-0178.1 (cit. on pp. 4, 69)

Randall, D. A., Bitz, C. M., Danabasoglu, G., Denning, A. S., Gent, P. R., Gettelman, A., Griffies, S. M., Lynch, P., Morrison, H., Pincus, R., & Thuburn, J. (2018). 100 years of earth system model development. *Meteorological Monographs*, *59*, 12.1–12.66. https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0018.1 (cit. on pp. 1, 3)

Richtmyer, R. D., & Morton, K. W. (1968). Difference methods for initial-value problems. *SIAM Review*, *10*(3), 381–383. https://doi.org/10.1137/1010073 (cit. on p. 57)

Ringler, T., Thuburn, J., Klemp, J., & Skamarock, W. (2010). A unified approach to energy conservation and potential vorticity dynamics on arbitrarily structured C-grids. *J. Comput. Phys.*, *229*, 3065–3090. https://doi.org/10.1016/j.jcp.2009.12.007 (cit. on p. 5)

Ronchi, C., Iacono, R., & Paolucci, P. (1996). The "cubed sphere": A new method for the solution of partial differential equations in spherical geometry. *Journal of Computational Physics*, *124*(1), 93–114. https://doi.org/https://doi.org/10.1006/jcph.1996.0047 (cit. on pp. 4, 69, 70, 72)

Sadourny, R. (1972). Conservative finite-difference approximations of the primitive equations on quasi-uniform spherical grids. *Monthly Weather Review*, *100*(2), 136–144. https://doi.org/10.1175/1520-0493(1972)100<0136:CFAOTP>2.3.CO;2 (cit. on pp. 4, 69)

Samenow, J. (2019). *National weather service launches upgraded, improved global forecast model*. Retrieved July 29, 2022, from https://www.washingtonpost.com/weather/2019/06/12/national-weather-service-launches-upgraded-improved-global-forecast-model/. (Cit. on p. 4)

Santos, L. F., & Peixoto, P. S. (2021). Topography-based local spherical voronoi grid refinement on classical and moist shallow-water finite-volume models. *Geoscientific Model Development*, *14*(11), 6919–6944. https://doi.org/10.5194/gmd-14-6919-2021 (cit. on p. 5)

Skamarock, W., Klemp, J., Duda, M., Fowler, L., Park, S.-H., & Ringler, T. (2012). A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tesselations and C-grid staggering. *Mon. Weather. Rev.*, *140(09)*, 3090–3105. https://doi.org/10.1175/MWR-D-11-00215.1 (cit. on p. 5)

Staniforth, A., & Thuburn, J. (2012). Horizontal grids for global weather and climate prediction models: A review. *Q. J. Roy. Meteor. Soc.*, *138*, 1–26. https://doi.org/10.1002/qj.958 (cit. on p. 3)

Stoer, J., & Bulirsch, R. (2002). In *Introduction to numerical analysis*. Springer New York, NY. https://doi.org/https://doi.org/10.1007/978-0-387-21738-3. (Cit. on pp. 20, 87)

Strang, G. (1968). On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, *5*(3), 506–517. https://doi.org/10.1137/0705041 (cit. on p. 58)

Strikwerda, J. C. (2004). *Finite difference schemes and partial differential equations, second edition*. Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9780898717938. (Cit. on p. 18)

Suresh, A., & Huynh, H. (1997). Accurate monotonicity-preserving schemes with runge–kutta time stepping. *Journal of Computational Physics*, *136*(1), 83–99. https://doi.org/https://doi.org/10.1006/jcph.1997.5745 (cit. on pp. 11, 22)

Thuburn, J. (2011). Conservation in dynamical cores: What, how and why? In *Numerical techniques for global atmospheric models* (pp. 345–355). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11640-7_11. (Cit. on p. 3)

Thuburn, J., Ringler, T., Skamarock, W., & Klemp, J. (2009). Numerical representation of geostrophic modes on arbitrarily structured C-grids. *J. Comput. Phys.*, *228*, 8321–8335. https://doi.org/10.1016/j.jcp.2009.08.006 (cit. on p. 5)

Trefethen, L. N. (2000). *Spectral methods in matlab*. Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9780898719598. (Cit. on pp. 18, 39)

Tumolo, G. (2011). *A semi-implicit, semi-lagrangian, p-adaptative discontinuous galerkin method for the rotating shallow-water equations: Analysis and numerical experiments* (Doctoral dissertation). University of Trieste. https://core.ac.uk/download/pdf/41173373.pdf. (Cit. on p. 31)

Ullrich, P. A., Jablonowski, C., Kent, J., Lauritzen, P. H., Nair, R., Reed, K. A., Zarzycki, C. M., Hall, D. M., Dazlich, D., Heikes, R., Konor, C., Randall, D., Dubos, T., Meurdesoif, Y., Chen, X., Harris, L., Kühnlein, C., Lee, V., Qaddouri, A., … Viner, K. (2017). Dcmip2016: A review of non-hydrostatic dynamical core design and intercomparison of participating models. *Geoscientific Model Development*, *10*(12), 4477–4509. https://doi.org/10.5194/gmd-10-4477-2017 (cit. on p. 3)

Van Leer, B. (1977). Towards the ultimate conservative difference scheme. iv. a new approach to numerical convection. *Journal of Computational Physics*, *23*(3), 276–299. https://doi.org/https://doi.org/10.1016/0021-9991(77)90095-X (cit. on pp. 4, 7, 19, 21)

Weller, H. (2012). Controlling the computational modes of the arbitrarily structured c grid, *Mon. Weather. Rev., 140(10),* 3220–3234. https://doi.org/doi.org/10.1175/MWR-D-11-00221.1 (cit. on p. 5)

Whitaker, J. (2015). *Hiwpp non-hydrostatic dynamical core tests: Results from idealized test cases*. Retrieved November 5, 2022, from https://www.weather.gov/media/sti/nggps/HIWPP_idealized_tests-v8%20revised%2005212015.pdf/. (Cit. on p. 4)

White, L., & Adcroft, A. (2008). A high-order finite volume remapping scheme for nonuniform grids: The piecewise quartic method (pqm). *Journal of Computational Physics*, *227*(15), 7394–7422. https://doi.org/https://doi.org/10.1016/j.jcp.2008.04.026 (cit. on p. 7)

Williamson, D., Drake, J., Hack, J., Jakob, R., & Swarztrauber, P. (1992). A standard test set for numerical approximations to the shallow water equations in spherical geometry. *J. Comput. Phys., 102*, 211–224. https://doi.org/10.1016/S0021-9991(05)80016-6 (cit. on p. 5)

Williamson, D. L. (2007). The evolution of dynamical cores for global atmospheric models. *Journal of the Meteorological Society of Japan. Ser. II*, *85B*, 241–269. https://doi.org/10.2151/jmsj.85B.241 (cit. on pp. 1, 2)

Wood, N., Staniforth, A., White, A., Allen, T., Diamantakis, M., Gross, M., Melvin, T., Smith, C., Vosper, S., Zerroukat, M., & Thuburn, J. (2014). An inherently mass-conserving semi-implicit semi-lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Quarterly Journal of the Royal Meteorological Society*, *140*(682), 1505–1520. https://doi.org/https://doi.org/10.1002/qj.2235 (cit. on p. 1)

Woodward, P. R. (1986). Piecewise-parabolic methods for astrophysical fluid dynamics. In K.-H. A. Winkler & M. L. Norman (Eds.), *Astrophysical radiation hydrodynamics* (pp. 245–326). Springer Netherlands. https://doi.org/10.1007/978-94-009-4754-2_8. (Cit. on p. 7)

Zerroukat, M., & Allen, T. (2022). On the corners of the cubed-sphere grid. *Quarterly Journal of the Royal Meteorological Society*, *148*(743), 778–783. https://doi.org/https://doi.org/10.1002/qj.4230 (cit. on pp. 69, 74)

Zheng, Y., & Marguinaud, P. (2018). Simulation of the performance and scalability of message passing interface (mpi) communications of atmospheric models running on exascale supercomputers. *Geoscientific Model Development*, *11*(8), 3409–3426. https://doi.org/10.5194/gmd-11-3409-2018 (cit. on p. 3)