

**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 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

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Motivations . . . . .	4
1.3	Goals . . . . .	5
1.4	Outline . . . . .	6
<b>2</b>	<b>One-dimensional finite-volume methods</b>	<b>7</b>
2.1	One-dimensional advection equation in integral form . . . . .	8
2.1.1	Notation . . . . .	8
2.1.2	The 1D advection equation . . . . .	11
2.2	The finite-volume Semi-Lagrangian approach . . . . .	15
2.3	Departure point computation . . . . .	16
2.3.1	RK1 scheme . . . . .	16
2.3.2	RK2 scheme . . . . .	17
2.4	Reconstruction . . . . .	18
2.4.1	The Piecewise-Parabolic Method . . . . .	20
2.4.2	Monotonization . . . . .	22
2.5	Flux . . . . .	23
2.6	Numerical experiments . . . . .	24
2.7	Concluding remarks . . . . .	28
<b>3</b>	<b>Two-dimensional finite-volume methods</b>	<b>29</b>
3.1	Two-dimensional advection equation in integral form . . . . .	30
3.1.1	Notation . . . . .	30
3.1.2	The 2D advection equation . . . . .	33
3.2	The finite-volume approach . . . . .	35
3.3	Dimension splitting . . . . .	37
3.4	Numerical experiments . . . . .	43

3.5	Concluding remarks . . . . .	45
<b>4</b>	<b>Cubed-sphere grids</b>	<b>49</b>
4.1	Cubed-sphere mappings . . . . .	50
4.1.1	Equidistant cubed-sphere . . . . .	50
4.1.2	Equiangular cubed-sphere . . . . .	52
4.1.3	Tangent vectors on the sphere . . . . .	53
4.1.4	Notation . . . . .	55
4.2	Edges treatment . . . . .	57
4.2.1	Ghost cells scalar field interpolation . . . . .	57
4.2.2	Ghost cells wind interpolation . . . . .	59
4.2.3	Edges reconstruction . . . . .	61
4.3	Concluding remarks . . . . .	65
<b>5</b>	<b>Cubed-sphere finite-volume methods</b>	<b>67</b>
5.1	Cubed-sphere advection equation in integral form . . . . .	67
5.2	Finite-volume on the cubed-sphere approach . . . . .	69
5.2.1	Mass fixer . . . . .	70
5.3	Dimension splitting . . . . .	72
5.3.1	Metric tensor treatment . . . . .	73
5.3.2	Flux at edges treatment . . . . .	75
5.4	Numerical experiments . . . . .	76
5.4.1	Divergence of the rotated zonal wind . . . . .	77
5.4.2	Advection of one Gaussian hill through the rotated zonal wind . . . . .	78
5.4.3	Steady steady . . . . .	79
5.5	Concluding remarks . . . . .	79
<b>6</b>	<b>Conclusions</b>	<b>85</b>
6.1	Future work . . . . .	87

## Appendices

<b>A</b>	<b>Numerical Analysis</b>	<b>89</b>
A.1	Finite-difference estimates . . . . .	89
A.2	Lagrange interpolation . . . . .	93
A.3	Numerical integration . . . . .	93
A.3.1	Midpoint rule . . . . .	94
A.3.2	Multi-step schemes . . . . .	95
A.4	PPM reconstruction accuracy analysis . . . . .	97

A.5	Convergence of 1D FV-SL schemes . . . . .	101
A.5.1	Consistency and convergence . . . . .	101
A.5.2	Stability . . . . .	103
A.5.3	Flux accuracy analysis . . . . .	106
A.6	Convergence, consistency and stability of 2D-FV schemes . . . . .	106
<b>B</b>	<b>Code availability</b>	<b>109</b>
<b>References</b>		<b>111</b>



# Chapter 1

## Introduction

### 1.1 Background

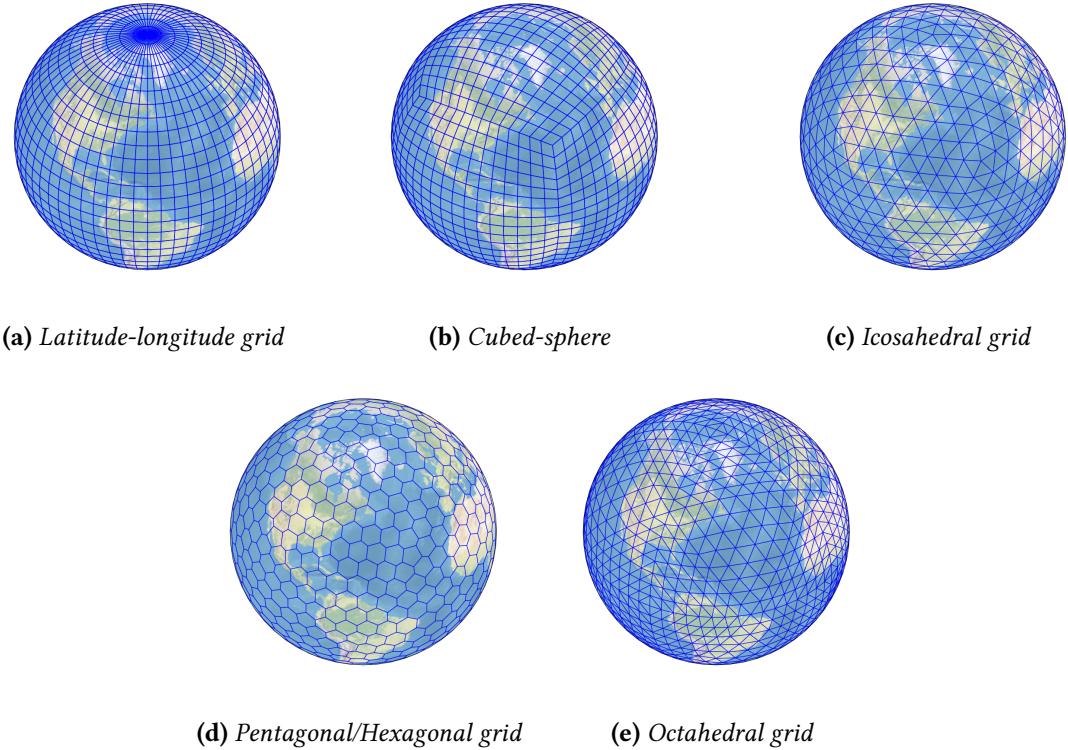
Weather and climate predictions are recognized as a good for mankind, due to the information they yield for diverse activities. For instance, short-range forecasts are useful for public use, while medium-range forecasts are helpful for industrial activities and agriculture. Seasonal forecasts (one up 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-Lowy (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).



**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 revisited 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 differential 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 adopted 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., 2023). 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 an 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 performing 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 addition of metric terms, due to non-orthogonality of the grid, and interpolation between panels to obtain ghost cells values needed for stencil computations. We give some future work perspectives in Chapter 6.

# Chapter 2

## One-dimensional finite-volume methods

The aim of this chapter is to provide a detailed description of one-dimensional (1D) finite-volume (FV) schemes within a Semi-Lagrangian (SL) framework, specifically applied to the 1D advection equation with periodic boundary conditions. These schemes are also known as flux-form Semi-Lagrangian schemes, and they allow for time steps beyond the Courant-Friedrichs-Lowy (CFL) condition while preserving the total mass. FV-SL schemes have been explored in the literature since the work of LeVeque (1985), which extended the finite-volume schemes from Godunov (1959) to accommodate larger time steps. This approach has been further investigated in the literature (c.f, e.g. . Leonard et al. (1996) and Lin and Rood (1996)).

To introduce the FV-SL schemes, we begin by discretizing the spatial and temporal domains into uniform grids. Subsequently, the FV-SL schemes involve three steps. The first step involves computing the departure points of the spatial grid edges. The second step, known as reconstruction, utilizes the grid cell average values to determine a piecewise function within each cell. This piecewise function approximates the values of the advected quantity and ensures the preservation of its local mass within each grid cell. The third step entails updating the fluxes at the grid edges by integrating the reconstruction function over a domain that extends from the departure point of the grid edge to the grid edge itself.

The first step of FV-SL schemes can be accomplished by integrating an ordinary differential equation backward in time. The second step is performed using the Piecewise-Parabolic Method (PPM) proposed by Colella and Woodward (1984). As the name suggests, PPM employs piecewise-parabolic functions. The third and final step is computed easily, as the reconstruction functions consist of parabolas that preserve the local mass.

It is worth noting that the reconstruction function can be constructed using functions other than parabolas. In fact, the Piecewise-Parabolic Method (PPM) can be seen as an extension of the Piecewise-Linear method proposed by Van Leer (1977), which, in turn, was inspired by the Piecewise-Constant method introduced by Godunov (1959). Additionally, other schemes inspired by PPM have been proposed in the literature utilizing higher-order

polynomials, such as quartic polynomials (White & Adcroft, 2008). For a comprehensive review of general piecewise-polynomial reconstruction, we recommend referring to the technical report by Engwirda and Kelley (2016), Lauritzen et al. (2011), and the references therein.

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, PPM has been implemented in the FV3 dynamical core on its latitude-longitude grid (Lin, 2004) and cubed-sphere (Putman & Lin, 2007) versions. Although many other shapes for the basis functions and higher-order schemes are available in the literature, L. Harris et al. (2021) points out that the PPM scheme suits the needs of FV3 well. It is a flexible method that can be modified to ensure low diffusivity or shape preservation, for example. Additionally, a finite-volume numerical method usually requires monotonicity constraints, which, according to Godunov's theorem, limit the order of convergence to at most 1. Therefore, a higher-order scheme needs to strike a well-balanced trade-off between increasing computational cost and potential benefits.

This chapter begins with a basic review of one-dimensional advection equation in the integral form in Section 2.1. In Section 2.2, we establish the framework for general one-dimensional finite-volume Semi-Lagrangian schemes. Section 2.3 presents methods for computing the departure point. The PPM reconstruction is described in Section 2.4, while Subsection 2.4.2 introduces different approaches to ensure the monotonicity of parabolas. Section 2.5 focuses on the description and investigation of the PPM flux computation. Section 2.6 presents numerical results using the PPM scheme for the advection equation. Finally, Section 2.7 presents some concluding remarks. The application of PPM to solve two-dimensional problems will be addressed in Chapter 3.

## 2.1 One-dimensional advection equation in integral form

### 2.1.1 Notation

Before introducing the FV-SL schemes, let us establish some notation by introducing the concepts of a  $\Delta x$ -grid, a  $\Delta t$ -temporal grid, and the  $(\Delta x, \Delta t, \lambda)$ -discretization, as well as the concept of grid function/winds. In this chapter, we will use the notation  $\Omega = [a, b]$  to represent the interval under consideration, and  $v$  to represent a non-negative integer indicating the number of ghost cell layers in each boundary. We also use the notations  $\mathbb{R}_v^N := \mathbb{R}^{N+2v}$  and  $\mathbb{R}_v^{N+1} := \mathbb{R}^{N+1+2v}$ .

**Definition 2.1** ( $\Delta x$ -grid). *For a given interval  $\Omega$  and a positive real number  $\Delta x$  such that  $\Delta x = (b - a)/N$  for some positive integer  $N$ , we say that  $\Omega_{\Delta x} = \{X_i\}_{i=-v+1}^{N+v}$  is a  $\Delta x$ -grid for  $\Omega$  if*

$$X_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] = [a + (i - 1)\Delta x, a + i\Delta x],$$

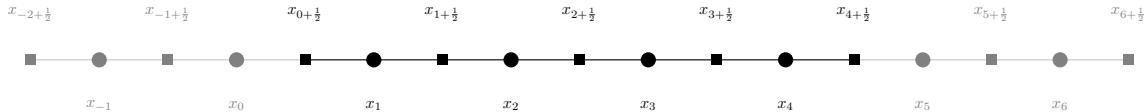
and  $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ . Each  $X_i$  is referred to as a control volume or cell, and  $x_{i-\frac{1}{2}}$  and  $x_{i+\frac{1}{2}}$  are

the edges of the control volume  $X_i$ . The cell centroid is defined by

$$x_i = \frac{1}{2}(x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}), \quad \forall i = -v+1, \dots, N+v,$$

and  $\Delta x$  is the cell length.

**Remark 2.1.** If  $1 \leq i \leq N$ , we refer to  $i$  as an interior index; otherwise,  $i$  is considered a ghost cell index and we say the  $X_i$  is a ghost cell.



**Figure 2.1:** Illustration of a  $\Delta x$ -grid with  $N = 4$  cells in its interior (in black) and  $v = 2$  ghost cell layers (in gray). The edges are denoted by squares and the cell centroids are denoted using circles.

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

$$T_n = [t^n, t^{n+1}], \quad t^n = n\Delta t, \quad \Delta t = \frac{T}{N_T}, \quad \forall n = 0, \dots, N_T.$$

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

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

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

**Definition 2.4** ( $\Delta x$ -grid function). For a  $\Delta x$ -grid, we say that  $Q$  is a  $\Delta x$ -grid function if  $Q = (Q_{-v+1}, \dots, Q_{N+v}) \in \mathbb{R}_v^N$ .

**Definition 2.5** ( $\Delta x$ -grid wind). For a  $\Delta x$ -grid, we say that  $u$  is a  $\Delta x$ -grid wind if  $u = (u_{-v+\frac{1}{2}}, \dots, u_{N+v+\frac{1}{2}}) \in \mathbb{R}_v^{N+1}$ .

The definition of a  $\Delta x$ -grid wind is based on the Arakawa grids (Arakawa & Lamb, 1977). Considering functions  $q, u : \Omega \times [0, T] \rightarrow \mathbb{R}$  and a  $(\Delta x, \Delta t, \lambda)$ -discretization of  $\Omega \times [0, T]$ , we introduce the grid functions  $q^n \in \mathbb{R}_v^N$  and  $u^n \in \mathbb{R}_v^{N+1}$ . Here,  $q_i^n = q(x_i, t^n)$  and  $u_{i+\frac{1}{2}}^n = u(x_{i+\frac{1}{2}}, t^n)$ . These grid functions represent the discrete values of  $q$  and  $u$  at the cell centroids and edges, respectively, for each time level  $t^n$  (Figure 2.2).

In this chapter, our focus lies on periodic grid functions. We define a  $\Delta x$ -grid function  $Q$  as periodic if it satisfies the following conditions:

$$\begin{aligned} Q_i &= Q_{N+i}, \quad i = -v+1, \dots, 0, \\ Q_i &= Q_{i-N}, \quad i = N+1, \dots, N+v. \end{aligned}$$

Similarly, we define a  $\Delta x$ -grid wind as periodic if it meets the following requirements:

$$\begin{aligned} u_{i-\frac{1}{2}} &= u_{N+i+\frac{1}{2}}, \quad i = -v, \dots, -1, \\ u_{i+\frac{1}{2}} &= u_{i+\frac{1}{2}-N}, \quad i = N+1, \dots, N+v. \end{aligned}$$

We use the notation  $\mathbb{P}_v^N$  and  $\mathbb{P}_v^{N+1}$  to represent the spaces of periodic  $\Delta x$ -grid functions and winds, respectively.



**Figure 2.2:** Illustration of  $\Delta x$ -grid function  $Q$  (black circles) and a  $\Delta x$ -grid wind  $u$  (blue squares) and its ghost cell values (in gray) assuming periodicity.

Given  $Q \in \mathbb{P}_v^N$ , we define the  $p$ -norm as

$$\|Q\|_{p,\Delta x} = \begin{cases} \left( \sum_{i=1}^N |Q_i|^p \right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty, \\ \max_{i=1,\dots,N} |Q_i| & \text{otherwise ,} \end{cases} \quad (2.1)$$

which is indeed a norm for periodic grid functions. Using a similar notation as in Engwirda and Kelley (2016), we define the stencil and a grid function evaluated on a stencil as follows.

**Definition 2.6** (Stencil). *For a  $\Delta x$ -grid, and each  $i = 0, \dots, N$ , we define a stencil as a set of the form  $S_{i+\frac{1}{2}} = \{i-r+1, \dots, i-1, i, i+1, \dots, i+s\} \subset \{-v+1, \dots, N+v\}$ .*

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

These definitions provide the necessary notation for describing grid functions and their evaluations on stencils. To achieve a more compact notation in some situations, we introduce the centered difference notation:

$$\delta_x g(x_i, t) = g(x_{i+\frac{1}{2}}, t) - g(x_{i-\frac{1}{2}}, t), \quad (2.2)$$

for any function  $g : \Omega \times [0, T] \rightarrow \mathbb{R}$ . Additionally, we introduce the average value of  $q$  in the  $i$ -th control volume at time  $t$ , denoted as  $Q_i(t)$ , defined by:

$$Q_i(t) = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x, t) dx. \quad (2.3)$$

Moreover, we define the  $\Delta x$ -grid function of average values as  $Q(t) = (Q_i(t))_{i=-v+1}^{N+v}$ . Here,  $Q_i(t)$  represents the average value of  $q$  in the  $i$ -th control volume at time  $t$ .

For the consideration of periodic boundary conditions, we can define spaces of periodic functions over the interval  $\Omega$  as follows:

$$\mathcal{S}_P(\Omega) = \{q : \mathbb{R} \times [0, +\infty[ \rightarrow \mathbb{R} : q(x+b-a, t) = q(x, t), \quad \forall x \in \mathbb{R}, \quad t \geq 0\}.$$

Similarly, the space of  $k$ -times periodically differentiable functions  $C_p^k(\Omega)$  can be defined as:

$$C_p^k(\Omega) = S_p(\Omega) \cap C^k(\mathbb{R} \times [0, \infty[),$$

where  $C^k(\mathbb{R} \times [0, +\infty[)$  denotes the space of functions that are  $k$  times continuously differentiable in both the spatial and temporal variables. In summary,  $S_p(\Omega)$  represents the space of periodic functions, and  $C_p^k(\Omega)$  represents the space of  $k$ -times periodically differentiable functions over the interval  $\Omega$  subject to periodic boundary conditions.

## 2.1.2 The 1D advection equation

In this section, we will derive the integral form of the 1D advection equation with periodic boundary conditions over the interval  $\Omega$ . What is going to be presented here follows LeVeque (1990, 2002) closely. The advection equation with periodic boundary conditions in its differential form is given by:

$$\begin{cases} [\partial_t q + \partial_x(uq)](x, t) = 0, & \forall (x, t) \in \mathbb{R} \times ]0, +\infty[, \\ q(a, t) = q(b, t), & \forall t \geq 0, \\ q_0(x) = q(x, 0), & \forall x \in \Omega. \end{cases} \quad (2.4)$$

Here,  $q \in C_p^1(\Omega)$  represents the advected quantity, and  $u \in C_p^1(\Omega)$  represents the velocity. We will focus on Equation (2.4) over the domain  $D = \Omega \times [0, T]$ , where  $T > 0$  is a finite time. A strong or classical solution to the advection equation is defined as a function  $q \in C_p^1(\Omega)$  and satisfies Equation (2.4). In order to deduce the integral form of Equation (2.4), we consider  $[x_1, x_2] \times [t_1, t_2] \subset D$ . Integrating Equation (2.5) over  $[x_1, x_2]$ , we obtain:

$$\frac{d}{dt} \int_{x_1}^{x_2} q(x, t) dx = -((uq)(x_2, t) - (uq)(x_1, t)), \quad (2.5)$$

and integrating Equation (2.5) over  $[t_1, t_2]$ , we get

$$\int_{x_1}^{x_2} q(x, t_2) dx = \int_{x_1}^{x_2} q(x, t_1) - \left( \int_{t_1}^{t_2} (uq)(x_2, t) dt - \int_{t_1}^{t_2} (uq)(x_1, t) dt \right). \quad (2.6)$$

The presented problem, Problem 2.1, aims to find a solution, called weak solution, to the advection equation in its integral form, considering the given initial condition  $q_0$  and velocity function  $u$ .

**Problem 2.1.** Given an initial condition  $q_0$  and a velocity function  $u$  we would like to find a weak solution  $q$  of the advection equation in the integral form:

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

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

We point out that, for Problem 2.1, the total mass in  $\Omega$  at time  $t$  defined by:

$$M_{[a,b]}(t) = \int_a^b q(x, t) dx,$$

remains constant over time, i.e.,

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

This conservation of total mass property is highly desirable for numerical schemes aiming to approximate general conservation law solutions accurately.

Applying the steps from Equation (2.4) to Equation (2.6) in reverse order, one can verify that if  $q$  is a weak solution and  $q \in C_P^1(\Omega)$ , then it satisfies Equation (2.4). Therefore, Equation (2.4) and Problem (2.1) are equivalent when  $q \in C_P^1(\Omega)$ . However, Problem (2.1) can be formulated for functions that are not  $C^1$  and have discontinuities. In fact, Problem (2.1) only requires that  $q$  and  $uq$  are locally integrable.

It is worth noting that Equation (2.6) holds for all  $x_1, x_2, t_1$ , and  $t_2$  such that  $[x_1, x_2] \times [t_1, t_2] \subset D$ . Therefore, let us consider a  $(\Delta x, \Delta t, \lambda)$ -discretization of  $D$  and rewrite Equation (2.6) in terms of this discretization. By replacing  $t_1, t_2, x_1$ , and  $x_2$  with  $t^n, t^{n+1}, x_{i-\frac{1}{2}}$ , and  $x_{i+\frac{1}{2}}$ , respectively, in Equation (2.6), we obtain:

$$Q_i(t^{n+1}) = Q_i(t^n) - \frac{1}{\Delta x} \left( \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, t) dt - \int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, t) dt \right), \quad (2.7)$$

$$\forall i = 1, \dots, N, \quad \forall n = 0, \dots, N_T - 1.$$

To achieve a more compact notation, we use the centered difference notation and then Equation (2.7) can be rewritten as:

$$Q_i(t^{n+1}) = Q_i(t^n) - \frac{1}{\Delta x} \delta_x \left( \int_{t^n}^{t^{n+1}} (uq)(x_i, t) dt \right), \quad \forall i = 1, \dots, N, \quad \forall n = 0, \dots, N_T - 1. \quad (2.8)$$

Now we can define a discretized version of Problem 2.1 as Problem 2.2.

**Problem 2.2.** Let us consider the framework of Problem 2.1 and a  $(\Delta x, \Delta t, \lambda)$ -discretization of  $\Omega \times [0, T]$ . Since we are operating within the framework of Problem 2.1, the following relationship holds:

$$Q_i(t^{n+1}) = Q_i(t^n) - \lambda \delta_x \left( \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} (uq)(x_i, t) dt \right), \quad \forall i = 1, \dots, N, \quad \forall n = 0, \dots, N_T - 1, \quad (2.9)$$

where  $Q_i(t) = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x, t) dx$ . Our objective now is to determine the values  $Q_i(t^n)$ ,  $\forall i = 1, \dots, N$ ,  $\forall n = 0, \dots, N_T - 1$ , given the initial values  $Q_i(0)$ ,  $\forall i = 1, \dots, N$ . In other words, we aim to find the average values of  $q$  in each control volume  $X_i$  at the specified time instances.

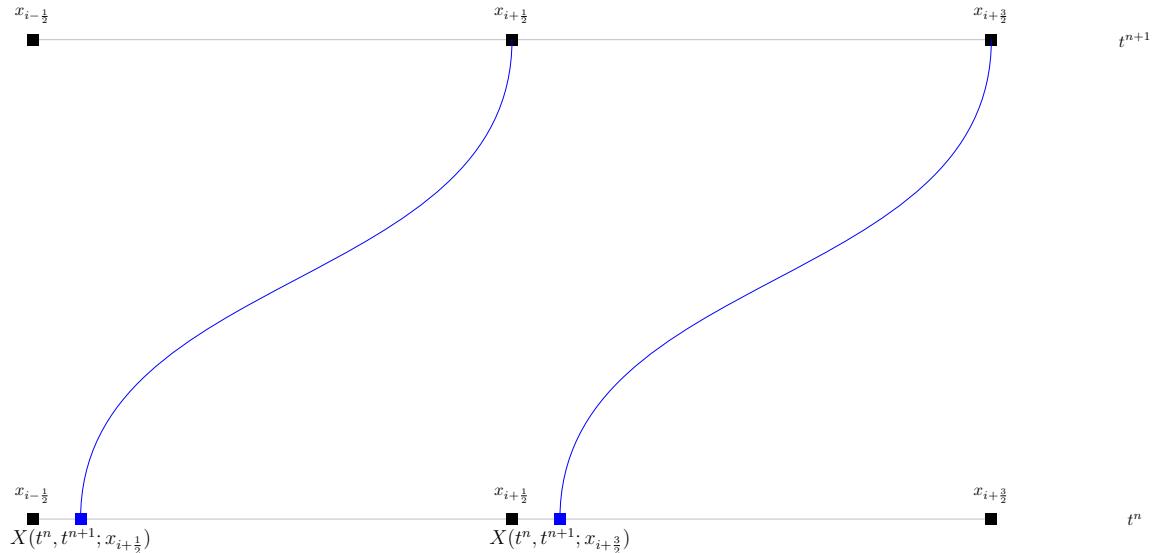
It is important to note that no approximations have been made in problems (2.1) and (2.2). In Equation (2.9), we divided and multiplied by  $\Delta t$  to interpret  $\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} (uq)(x_{i \pm \frac{1}{2}}, t) dt$  as a time-averaged flux. This interpretation is useful for deriving finite-volume schemes.

In Problem 2.2, we need to approximate the time-averaged flux at the cell edges  $x_{i\pm\frac{1}{2}}$  to derive a finite-volume scheme. This flux, in principle, requires knowledge of  $q$  over the entire interval  $[t^n, t^{n+1}]$ . To overcome this, we can express the temporal integral as a spatial integral at time  $t^n$ . This approach avoids the need for information about  $q$  throughout the entire interval  $[t^n, t^{n+1}]$ . Furthermore, this spatial integral domain is closely related to the definition of the 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:

$$\begin{cases} \frac{\partial X}{\partial t}(t, s; x_{i+\frac{1}{2}}) = u(X(t, s; x_{i+\frac{1}{2}}), t), & t \in [t^n, s] \\ X(s, s; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}}. \end{cases} \quad (2.10)$$

The point  $X(t^n, s; x_{i+\frac{1}{2}})$  is called departure point at time  $t^n$  of the point  $x_{i+\frac{1}{2}}$  at time  $s$ . In Figure 2.3 we illustrate the departure point idea.



**Figure 2.3:** Illustration of the departure point of the cell edges from time  $t^{n+1}$  to  $t^n$ .

Integrating Equation (2.10) over the interval  $[t, s]$ , we get:

$$X(t, s; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - \int_t^s u(X(\theta, s; x_{i+\frac{1}{2}}), \theta) d\theta. \quad (2.11)$$

In the following Proposition, we show how the time-averaged flux is related to a spatial integral over a interval depending on departure points.

**Proposition 2.1.** *Assume the framework of Problem 2.2. If  $q$  and  $u$  are  $C^1$  functions, then:*

$$\int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s) ds = \int_{X(t^n, t^{n+1}; x_{i+\frac{1}{2}})}^{x_{i+\frac{1}{2}}} q(x, t^n) dx \quad (2.12)$$

*Proof.* Using the Leibniz rule for integration in Equation (2.11), it follows that:

$$\begin{aligned}\frac{\partial X}{\partial s}(t, s; x_{i+\frac{1}{2}}) &= - \left( u(x_{i+\frac{1}{2}}, s) + \int_t^s \frac{du}{ds}(X(\theta, s; x_{i+\frac{1}{2}}), \theta) d\theta \right) \\ &= -u(x_{i+\frac{1}{2}}, s) - \int_t^s \frac{\partial u}{\partial x}(X(\theta, s; x_{i+\frac{1}{2}}), \theta) \frac{\partial X}{\partial s}(\theta, s; x_{i+\frac{1}{2}}) d\theta.\end{aligned}\quad (2.13)$$

Taking the derivative with respect to  $t$  of Equation (2.13), we have:

$$\frac{\partial}{\partial t} \left( \frac{\partial X}{\partial s} \right)(t, s; x_{i+\frac{1}{2}}) = \frac{\partial u}{\partial x}(X(t, s; x_{i+\frac{1}{2}}), t) \frac{\partial X}{\partial s}(t, s; x_{i+\frac{1}{2}}). \quad (2.14)$$

Using standard ordinary differential equations techniques (ODE), we get that  $X$  that solves Equations (2.13) and (2.14) is given by:

$$\frac{\partial X}{\partial s}(t, s; x_{i+\frac{1}{2}}) = - \exp \left( \int_t^s \frac{\partial u}{\partial x}(X(\theta, s; x_{i+\frac{1}{2}}), \theta) d\theta \right) u(x_{i+\frac{1}{2}}, s). \quad (2.15)$$

Computing  $q$  on the trajectory give by  $X(t, s; x_{i+\frac{1}{2}})$  and taking its time derivative, we obtain:

$$\begin{aligned}\frac{dq}{dt}(X(t, s; x_{i+\frac{1}{2}}), t) &= \frac{\partial q}{\partial t}(X(t, s; x_{i+\frac{1}{2}}), t) + u(X(t, s; x_{i+\frac{1}{2}}), t) \frac{\partial q}{\partial x}(X(t, s; x_{i+\frac{1}{2}}), t) \\ &= -\frac{\partial u}{\partial x}(X(t, s; x_{i+\frac{1}{2}}), t) q(X(t, s; x_{i+\frac{1}{2}}), t),\end{aligned}\quad (2.16)$$

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 (2.10). Using again standard ODE techniques, we get that  $q$  that solves Equation (2.16) is given by:

$$q(X(t, s; x_{i+\frac{1}{2}}), t) = \exp \left( - \int_t^s \frac{\partial u}{\partial x}(X(\theta, s; x_{i+\frac{1}{2}}), \theta) d\theta \right) q(x_{i+\frac{1}{2}}, s). \quad (2.17)$$

Notice that if  $u$  does not depend on  $x$ , then  $q$  is constant along the trajectory  $X(t, s; x_{i+\frac{1}{2}})$ .

Let us consider the mapping  $s \in [t^n, t^{n+1}] \rightarrow X(t^n, s, x_{i+\frac{1}{2}})$ . Integrating  $q$  over all departure points at time  $t^n$  from  $x_{i+\frac{1}{2}}$  at time  $s$ , we have

$$\int_{X(t^n, t^n; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}}}^{X(t^n, t^{n+1}; x_{i+\frac{1}{2}})} q(x, t^n) dx = \int_{t^n}^{t^{n+1}} q(X(t^n, s; x_{i+\frac{1}{2}}), t^n) \frac{\partial X}{\partial s}(t^n, s; x_{i+\frac{1}{2}}) ds, \quad (2.18)$$

where we are just using the variable change integration formula. Then, it follows from Equations (2.15) and (2.17) with  $t = t^n$  that:

$$\int_{x_{i+\frac{1}{2}}}^{X(t^n, t^{n+1}; x_{i+\frac{1}{2}})} q(x, t^n) dx = - \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, s) ds,$$

which is the desired formula.  $\square$

With the aid of Proposition 2.1, we can rewrite Problem 2.2 in terms of the departure

point, avoiding the need for knowledge about  $q$  over the entire interval  $[t^n, t^{n+1}]$ . This is described in Problem 2.3:

**Problem 2.3.** Assume the framework of Problem 2.1 and a  $(\Delta x, \Delta t, \lambda)$ -discretization of  $\Omega \times [0, T]$ . Since we are in the framework of Problem 2.1, it follows that:

$$Q_i(t^{n+1}) = Q_i(t^n) - \lambda \left( \frac{1}{\Delta t} \int_{X(t^n, t^{n+1}; x_{i+\frac{1}{2}})}^{x_{i+\frac{1}{2}}} q(x, t^n) dx - \frac{1}{\Delta t} \int_{X(t^n, t^{n+1}; x_{i-\frac{1}{2}})}^{x_{i-\frac{1}{2}}} q(x, t^n) dx \right), \quad (2.19)$$

$$\forall i = 1, \dots, N, \quad \forall n = 0, \dots, N_T - 1,$$

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  $Q_i(t^n)$ ,  $\forall i = 1, \dots, N$ ,  $\forall n = 0, \dots, N_T - 1$ , given the initial values  $Q_i(0)$ ,  $\forall i = 1, \dots, N$ . In other words, we would like to find the average values of  $q$  in each control volume  $X_i$  at the considered time instants.

At each time step  $t^n$ , we compute the values of  $Q_i(t^{n+1})$  based on  $Q_i(t^n)$  and the integrals of  $q(x, t^n)$  over specific intervals. These intervals are defined by the departure points  $X(t^n, t^{n+1}; x_{i+\frac{1}{2}})$  and  $X(t^n, t^{n+1}; x_{i-\frac{1}{2}})$ . To perform the computations, we need to determine the departure points from the edges of all control volumes and calculate the required integrals. This idea serves as the motivation for defining finite-volume Semi-Lagrangian schemes. These schemes involve estimating the departure points and reconstructing the function  $q$  at time  $t^n$  using its average values  $Q_i(t^n)$ , which enables us to compute the necessary integrals.

## 2.2 The finite-volume Semi-Lagrangian approach

Finally, we define the 1D FV-SL scheme problem as follows in Problem 2.3.

**Problem 2.4 (1D FV-SL scheme).** Assume the framework defined in Problem 2.3. The finite-volume Semi-Lagrangian approach of Problem 2.3 consists of finding a scheme of the form:

$$Q_i^{n+1} = Q_i^n - \lambda(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n), \quad \forall i = 1, \dots, N, \quad \forall n = 0, \dots, N_T - 1, \quad (2.20)$$

where  $Q^n \in \mathbb{P}_v^N$  is intended to be an approximation of  $Q(t^n) \in \mathbb{P}_v^N$  in some sense. We define  $Q_i^0 = Q_i(0)$  or  $Q_i^0 = q_i^0$ . The terms  $F_{i \pm \frac{1}{2}}^n$  are known as numerical flux and are given by

$$F_{i \pm \frac{1}{2}}^n = \frac{1}{\Delta t} \int_{\tilde{x}_{i \pm \frac{1}{2}}^n}^{x_{i \pm \frac{1}{2}}} \tilde{q}(x; Q^n) dx, \quad (2.21)$$

where  $\tilde{x}_{i \pm \frac{1}{2}}^n$  is an estimate of the departure point  $X(t^n, t^{n+1}; x_{i \pm \frac{1}{2}})$ , and  $\tilde{q}$  is a reconstruction function for  $q$  built with the values  $Q^n$ . Thus,  $F_{i \pm \frac{1}{2}}^n$  approximates  $\frac{1}{\Delta t} \int_{X(t^n, t^{n+1}; x_{i \pm \frac{1}{2}})}^{x_{i \pm \frac{1}{2}}} q(x, t^n) dx$ .

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

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

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

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

where we are using that  $F_{N+\frac{1}{2}}^n = F_{\frac{1}{2}}^n$ , since we are assuming periodic boundary conditions.

We would like to highlight an important relationship between the average values of  $q$  and its values at the cell centroids. In Problem 2.4, we mentioned that the initial condition can be represented as  $q_i^0$  instead of  $Q_i(0)$ . Moreover, when analyzing the convergence of a FV-SL scheme, it is useful to compare  $Q_i^n$  with  $q_i^n$  since computing  $Q_i(t^n)$  requires evaluating an analytical integral, which can be challenging in certain cases. In Proposition 2.2, we provide a simple proof that  $q_i^n$  approximates  $Q_i(t^n)$  with second-order error when  $q$  is twice continuously differentiable.

**Proposition 2.2.** *If  $q \in C_P^2(\Omega)$ , 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.3 for the function  $q(x, t^n)$ . □

The Problem of the convergence of 1D FV-SL schemes is addressed in Section A.5. Now we are going to address the problem of the departure point estimation and the reconstruction problem.

## 2.3 Departure point computation

### 2.3.1 RK1 scheme

Equation (2.11) enables us to compute or estimate the departure point. For instance, if  $u$  is constant, the departure point at time  $t^n$  for the point  $x_{i+\frac{1}{2}}$  at time  $t^{n+1}$  is given by:

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - u \Delta t. \quad (2.23)$$

In general, the estimated departure point, denoted by  $\tilde{x}_{i+\frac{1}{2}}^n$ , takes the form:

$$\tilde{x}_{i+\frac{1}{2}}^n = x_{i+\frac{1}{2}} - \tilde{u}_{i+\frac{1}{2}}^n \Delta t, \quad (2.24)$$

where  $\tilde{u}_{i+\frac{1}{2}}^n$  represents the time-averaged wind and approximates:

$$\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} u(X(\theta, t^{n+1}; x_{i+\frac{1}{2}}), \theta) d\theta. \quad (2.25)$$

The departure point  $\tilde{x}_{i+\frac{1}{2}}^n$  is said to be  $p$ -order accurate if there exists a constant  $C$  that does not depend on  $\Delta t$ , such that:

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) - \tilde{x}_{i+\frac{1}{2}}^n = C \Delta t^p. \quad (2.26)$$

One possible way of estimating the time-averaged wind is by using:

$$\tilde{u}_{i+\frac{1}{2}}^n = u_{i+\frac{1}{2}}^n, \quad (2.27)$$

as in the original PPM scheme by Colella and Woodward (1984). This scheme will be referred to as **RK1**. Our objective now is to determine the constant  $C$  and the value of  $p$  in Equation (2.26) in the following proposition. It is useful to introduce the concept of a material derivative beforehand:

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

where  $h$  is a function belonging to  $C^1$ .

**Proposition 2.3.** *If  $u \in C^1$  and the time-averaged wind is computed using Equation (2.27), then the departure point from Equation (2.24) satisfies:*

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) - \tilde{x}_{i+\frac{1}{2}}^n = C\Delta t^2, \quad (2.28)$$

for a constant  $C$  that depends 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 (2.11), we obtain:

$$X(t^n, t^{n+1}; x_{i+\frac{1}{2}}) = x_{i+\frac{1}{2}} - u(X(t^n, t^{n+1}; x_{i+\frac{1}{2}}), t^n)\Delta t - \frac{1}{2} \frac{Du}{Dt}(X(\theta_1, t^{n+1}; x_{i+\frac{1}{2}}), \theta_1)\Delta t^2, \quad (2.29)$$

for  $\theta_1 \in [t^n, t^{n+1}]$ . Using Taylor's expansion of  $u(X(t, t^{n+1}; x_{i+\frac{1}{2}}), t^n)$ , we have:

$$u(X(t^n, t^{n+1}; x_{i+\frac{1}{2}}), t^n) = u_{i+\frac{1}{2}}^n - \left( u \frac{\partial u}{\partial x} \right)(X(\theta_2, t^{n+1}; x_{i+\frac{1}{2}}), t^n)\Delta t, \quad (2.30)$$

for  $\theta_2 \in [t^n, t^{n+1}]$ . Substituting Equation (2.30) into Equation (2.29), we obtain the desired constant  $C$  given by:

$$C = C(\theta_1, \theta_2) = -\frac{1}{2} \frac{Du}{Dt}(X(\theta_1, t^{n+1}; x_{i+\frac{1}{2}}), \theta_1) - \left( u \frac{\partial u}{\partial x} \right)(X(\theta_2, t^{n+1}; x_{i+\frac{1}{2}}), t^n). \quad (2.31)$$

□

### 2.3.2 RK2 scheme

Before presenting a higher-order estimate for the departure point, let us recall the definition of the CFL number.

**Definition 2.8.** *For Problem 2.4, the CFL number at an edge  $x_{i+\frac{1}{2}}$  and at a time level  $t^n$  is defined by*

$$c_{i+\frac{1}{2}}^n = \frac{\Delta t}{\Delta x} u_{i+\frac{1}{2}}^n. \quad (2.32)$$

The CFL number is the maximum of the values  $c_{i+\frac{1}{2}}^n$ . The problem of estimating the

departure point is very common in Semi-Lagrangian schemes, which are quite popular in atmospheric modeling. For a review of 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 2.1 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 method to compute the departure point, which we express in terms of  $\tilde{u}_{i+\frac{1}{2}}^n$  using the following equations (D. R. Durran, 2010):

$$\begin{aligned}\tilde{x}_{i+\frac{1}{2}}^{n+\frac{1}{2}} &= x_{i+\frac{1}{2}} - u_{i+\frac{1}{2}}^n \frac{\Delta t}{2} = x_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}^n \frac{\Delta x}{2}, \\ \tilde{u}_{i+\frac{1}{2}}^n &= u\left(\tilde{x}_{i+\frac{1}{2}}^{n+\frac{1}{2}}, t^n + \frac{\Delta t}{2}\right).\end{aligned}\quad (2.33)$$

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:

$$u_{i+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{3}{2}u_{i+\frac{1}{2}}^n - \frac{1}{2}u_{i+\frac{1}{2}}^{n-1}. \quad (2.34)$$

Then we use linear interpolation in space:

$$\tilde{u}_{i+\frac{1}{2}}^n = \begin{cases} (1 - \alpha_{i+\frac{1}{2}}^n)u_{i+\frac{1}{2}-k}^{n+\frac{1}{2}} + \alpha_{i+\frac{1}{2}}^n u_{i-\frac{1}{2}-k}^{n+\frac{1}{2}} & \text{if } u_{i+\frac{1}{2}}^n \geq 0, \\ \alpha_{i+\frac{1}{2}}^n u_{i+\frac{3}{2}-k}^{n+\frac{1}{2}} + (1 - \alpha_{i+\frac{1}{2}}^n)u_{i+\frac{1}{2}-k}^{n+\frac{1}{2}} & \text{if } u_{i+\frac{1}{2}}^n < 0, \end{cases} \quad (2.35)$$

where  $\frac{c_{i+\frac{1}{2}}^n}{2} = \alpha_{i+\frac{1}{2}}^n + k$ ,  $k = \lfloor \frac{c_{i+\frac{1}{2}}^n}{2} \rfloor$ ,  $\alpha_{i+\frac{1}{2}}^n \in [0, 1[$ , and  $\lfloor \cdot \rfloor$  is the floor function. This scheme leads to a third-order error in the departure point estimate (see *e.g.* D. R. Durran (2010, Section 7.1.2)). This scheme shall be referred to as **RK2**. Notice that for this scheme, we need ghost values for the velocity, depending on how large the CFL number is. In particular, if the CFL number is less than 2, then  $k = 0$  and we need the ghost values  $u_{-1+\frac{1}{2}}^n$  and  $u_{N+\frac{3}{2}}^n$ . Finally, the impact of the departure point approximation on the time-averaged flux is investigated in Section A.5.3.

## 2.4 Reconstruction

In this section, we will review the Piecewise-Parabolic Method (PPM). The analysis of its accuracy will be presented in Section A.4. PPM was originally proposed by Colella and Woodward (1984) for gas dynamic simulations, and its applicability to atmospheric simulations has been demonstrated by Carpenter et al. (1990). This method is based on utilizing parabolas to reconstruct the function using its average values, ensuring both mass conservation and monotonicity. PPM is an extension of the Piecewise-Linear Method introduced by Van Leer (1977), and it is implemented in the FV3 model using the dimension splitting method developed by Lin and Rood (1996).

Let's consider a function  $q$  defined in  $\Omega = [a, b]$  and a  $\Delta x$ -grid covering  $\Omega$ . We assume

that we are given the average values  $Q_i = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x) dx$  for each control volume  $X_i$ , where  $i = 1, \dots, N$ . In this context, it is convenient to define the  $\Delta x$ -grid function  $Q \in \mathbb{P}_v^N$  with the entries given by  $Q_i$ . To facilitate the discussion, we introduce the indicator function  $\chi_i(x)$  for each control volume  $X_i$ , defined as:

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

Drawing inspiration from Stoer and Bulirsch (2002, Chapter 1), we consider a family of functions  $\Phi(\xi; \mu)$  defined for  $\xi \in [0, 1]$ , depending on a parameter  $\mu = (\mu_0, \mu_1, \dots, \mu_d) \in \mathbb{R}^{d+1}$ . The reconstruction problem involves finding a piecewise function:

$$\tilde{q}(x; Q) = \sum_{i=1}^N \chi_i(x) q_i(x; Q), \quad (2.36)$$

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

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

which means that  $q_i(x; Q)$  preserves the mass within each control volume  $X_i$ .

Notice that, given  $q_i(x; Q) = \Phi\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta x}; \alpha_i\right)$ , it is reasonable to expect that  $\Phi(0; \alpha_i)$  approximates  $q_i(x_{i-\frac{1}{2}})$  and  $\Phi(1; \alpha_i)$  approximates  $q_i(x_{i+\frac{1}{2}})$ . Additionally, if both  $q$  and  $\Phi$  are sufficiently differentiable,  $\Phi^{(l)}(0; \alpha_i)$  should approximate  $(\Delta x)^l q^{(l)}(x_{i-\frac{1}{2}})$  and  $\Phi^{(l)}(1; \alpha_i)$  should approximate  $(\Delta x)^l q^{(l)}(x_{i+\frac{1}{2}})$ , provided these derivatives exist.

One approach to estimating these values at the edges  $x_{i+\frac{1}{2}}$  using the average values  $Q$  is by employing a reconstruction method based on primitive functions (LeVeque, 2002, Chapter 17). It is worth noting that if we define:

$$Q(x) = \int_a^x q(\xi) d\xi, \quad (2.37)$$

we have  $Q^{(l)}(x) = q^{(l-1)}(x)$ . Specifically,  $Q^{(l)}(x_{i+\frac{1}{2}}) = q^{(l-1)}(x_{i+\frac{1}{2}})$  and  $Q(x_{i+\frac{1}{2}}) = \Delta x \sum_{k=1}^i Q_k$ , for all  $i = 0, \dots, N$ . Therefore, we can employ finite-difference schemes to estimate  $q^{(l-1)}(x_{i+\frac{1}{2}})$  using the  $\Delta x$ -grid function  $Q$ , given that it is assumed to be known.

Let us assume that the  $l$ -th derivative of  $Q$  at  $x_{i+\frac{1}{2}}$  is approximated using a stencil  $S_{i+\frac{1}{2}}^{(l)}$  and weights  $\beta_{k,i}^{(l)}$ , where  $k \in S_{i+\frac{1}{2}}^{(l)}$ . When  $d$  is odd, we can seek a parameter  $\alpha_i \in \mathbb{R}^{d+1}$  that ensures mass conservation and approximates  $q$  and its derivatives at the edges by solving the following system:

$$\begin{cases} \int_0^1 \Phi(\xi; \alpha_i) d\xi &= Q_i, \\ \Phi^{(l)}(0; \alpha_i) &= (\Delta x)^l \sum_{k \in S_{i-\frac{1}{2}}^{(l)}} \beta_{k,i}^{(l)} Q_k, \quad \text{for } l = 0, \dots, d-1. \end{cases} \quad (2.38)$$

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

$$\begin{cases} \int_0^1 \Phi(\xi; \alpha_i) d\xi = Q_i, \\ \Phi^{(l)}(0; \alpha_i) = (\Delta x)^l \sum_{k \in S_{i-\frac{1}{2}}^{(l)}} \beta_{k,i}^{(l)} Q_k, \quad \text{for } l = 0, \dots, \frac{d}{2} - 1, \\ \Phi^{(l)}(1; \alpha_i) = (\Delta x)^l \sum_{k \in S_{i+\frac{1}{2}}^{(l)}} \beta_{k,i}^{(l)} Q_k, \quad \text{for } l = 0, \dots, \frac{d}{2} - 1. \end{cases} \quad (2.39)$$

The reconstruction problem becomes linear when  $\Phi(\xi; \mu)$  can be expressed as:

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

where  $\Phi_k$  are functions defined on  $[0, 1]$ . In this case, Equation (2.38) and Equation (2.39) form  $(d+1) \times (d+1)$  linear systems. It is common to assume that the  $\Phi_k$ 's are linearly independent. Therefore, we have described a method that allows us to reconstruct a function from its average values, preserving its mass in each control volume, and approximating  $q$  at the edges. This method works for functions  $\Phi_k$  as long as they are sufficiently differentiable. For example, choosing  $d = 0$  and  $\Phi_0(\xi) = 1$  gives us piecewise constant functions, as used in Godunov (1959). If we choose  $d = 1$ ,  $\Phi_0(\xi) = 1$ , and  $\Phi_1(\xi) = \xi$ , we obtain a piecewise linear reconstruction, similar to Van Leer (1977). For polynomial reconstruction schemes, we refer to Engwirda and Kelley (2016) and the references therein.

#### 2.4.1 The Piecewise-Parabolic Method

Hereafter, we are going the focus on the piecewise parabolic method from Colella and Woodward (1984) that uses  $d = 2$ ,  $\Phi_0(\xi) = 1$ ,  $\Phi_1(\xi) = \xi$ ,  $\Phi_2(\xi) = (1 - \xi)\xi$ . 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:

$$q_i(x; Q) = q_{L,i} + \Delta q_i z_i(x) + q_{6,i} z_i(x)(1 - z_i(x)), \quad \text{where } z_i(x) = \frac{x - x_{i-\frac{1}{2}}}{\Delta x}, \quad x \in X_i, \quad (2.40)$$

where the values  $q_{L,i}$ ,  $\Delta q_i$  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  $\lim_{x \rightarrow x_{i-\frac{1}{2}}^+} q_i(x; Q) = q_{L,i}$ . If we define  $q_{R,i} = \lim_{x \rightarrow x_{i+\frac{1}{2}}^-} q_i(x; Q)$ , then we have:

$$\Delta q_i = q_{R,i} - q_{L,i}. \quad (2.41)$$

The average value of  $q_i$  is given by:

$$\frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q_i(x; Q) dx = \frac{(q_{L,i} + q_{R,i})}{2} + \frac{q_{6,i}}{6}. \quad (2.42)$$

Under the hypothesis of mass conservation, we have:

$$q_{6,i} = 6 \left( Q_i - \frac{(q_{L,i} + q_{R,i})}{2} \right). \quad (2.43)$$

Therefore, we have found the parameters  $\Delta q_i$  and  $q_{6,i}$  as functions of the parameters  $q_{L,i}$  and  $q_{R,i}$ , such that the parabola  $q_i$  from (2.36) 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 (2.37), 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):

$$q_{i+\frac{1}{2}} = \frac{1}{2} \left( Q_{i+1} + Q_i \right) - \frac{1}{6} \left( \delta Q_{i+1} - \delta Q_i \right), \quad (2.44)$$

where  $\delta Q_i$  is the average slope in the  $i$ -th control-volume:

$$\delta Q_i = \frac{1}{2} \left( Q_{i+1} - Q_{i-1} \right). \quad (2.45)$$

We notice that Formula (2.45) may be rewritten more explicitly as:

$$q_{i+\frac{1}{2}} = \frac{7}{12} \left( Q_{i+1} + Q_i \right) - \frac{1}{12} \left( Q_{i+2} + Q_{i-1} \right). \quad (2.46)$$

The Formula (2.46) is fourth-order accurate if  $q$  is at least  $C^4$  (Colella & Woodward, 1984). Indeed, we prove this later in Proposition A.1. An explicit expression for the values of  $q_{R,i}$  and  $q_{L,i}$  are given by:

$$q_{R,i} = q_{i+\frac{1}{2}} = \frac{7}{12} \left( Q_{i+1} + Q_i \right) - \frac{1}{12} \left( Q_{i+2} + Q_{i-1} \right), \quad (2.47)$$

$$q_{L,i} = q_{i-\frac{1}{2}} = \frac{7}{12} \left( Q_i + Q_{i-1} \right) - \frac{1}{12} \left( Q_{i+1} + Q_{i-2} \right). \quad (2.48)$$

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:

$$q_{R,i} = \frac{1}{60} \left( 2Q_{i-2} - 13Q_{i-1} + 47Q_i + 27Q_{i+1} - 3Q_{i+2} \right), \quad (2.49)$$

$$q_{L,i} = \frac{1}{60} \left( -3Q_{i-2} + 27Q_{i-1} + 47Q_i - 13Q_{i+1} + 2Q_{i+2} \right). \quad (2.50)$$

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. The PPM scheme, utilizing Equations (2.47) and (2.48), will be referred to as **PPM-0**. On the other hand, the PPM scheme, employing Equations (2.49) and (2.50), will be denoted as **PPM-PL07**.

## 2.4.2 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) - PPM-CW84

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  $\delta Q_i$  in Equation (2.44) by the values  $\delta_m Q_i$  given by:

$$\delta_m Q_i = \begin{cases} \max(|\delta Q_i|, 2|Q_{i+1} - Q_i|, 2|Q_i - Q_{i-1}|) \cdot \text{sgn}(\delta Q_i) & \text{if } (Q_{i+1} - Q_i)(Q_i - Q_{i-1}) > 0, \\ 0 & \text{otherwise,} \end{cases} \quad (2.51)$$

where  $\text{sgn}$  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:

$$q_{L,i} \leftarrow Q_i, \quad q_{R,i} \leftarrow Q_i, \quad \text{if } (Q_{R,i} - Q_i)(Q_i - Q_{L,i}) \geq 0 \quad (2.52)$$

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 (A.33) 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:

$$\begin{cases} q_{L,i} \leftarrow 3Q_i - 2q_{R,i} & \text{if } \Delta q_i \cdot q_{6,i} > (\Delta q_i)^2, \\ q_{R,i} \leftarrow 3Q_i - 2q_{L,i} & \text{if } -(\Delta q_i)^2 > \Delta q_i \cdot q_{6,i} \end{cases} \quad (2.53)$$

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. This scheme is referred to as **PPM-CW84**.

### Limiter from Lin (2004) - PPM-L04

Similarly to Colella and Woodward (1984), Lin (2004) reduces numerical oscillations in the parabolas by replacing the term  $\delta Q_i$  in Equation (2.44) with the values  $\delta_m Q_i$  given by:

$$\delta_m Q_i = \max(|\delta Q_i|, 2\delta Q_{\min,i}, 2\delta Q_{\max,i}) \cdot \text{sgn}(\delta Q_i), \quad (2.54)$$

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:

$$q_{L,i} \leftarrow Q_i - \max(|\delta_m Q_i|, |q_{L,i} - Q_i|) \cdot \text{sgn}(\delta_m Q_i), \quad (2.55)$$

$$q_{R,i} \leftarrow Q_i - \max(|\delta_m Q_i|, |q_{R,i} - Q_i|) \cdot \text{sgn}(\delta_m Q_i). \quad (2.56)$$

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. This scheme is referred to as **PPM-L04**.

## 2.5 Flux

Let's consider the framework outlined in Problem 2.4. Assuming that  $Q^n \in \mathbb{P}_v^N$  is known, our objective is to compute the values  $Q^{n+1}$ . To accomplish this, we utilize a scheme similar to the one presented in Problem 2.4, taking into account the presence of a reconstruction function  $\tilde{q}(x; Q^n)$  as discussed in Section 2.4, and an initial estimation  $\tilde{x}_{i+\frac{1}{2}}^n = x_{i+\frac{1}{2}} - \tilde{u}_{i+\frac{1}{2}}^n \Delta t$  for a time-averaged wind  $\tilde{u}_{i+\frac{1}{2}}^n$  as explained in Section 2.3. The numerical flux function is then expressed as:

$$F_{i+\frac{1}{2}}^n(Q^n, \tilde{u}_{i+\frac{1}{2}}^n) = \frac{1}{\Delta t} \int_{x_{i+\frac{1}{2}} - \tilde{u}_{i+\frac{1}{2}}^n \Delta t}^{x_{i+\frac{1}{2}}} \tilde{q}(x; Q^n) dx. \quad (2.57)$$

Notice that if we define the averaged CFL number,

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

where  $\tilde{c}_{i+\frac{1}{2}}^n = k + \alpha_{i+\frac{1}{2}}^n$ ,  $k = \lfloor \tilde{c}_{i+\frac{1}{2}}^n \rfloor$ ,  $\alpha_{i+\frac{1}{2}}^n \in [0, 1[$ , we can express the numerical flux as (Y. Chen et al., 2017; Lin & Rood, 1996):

$$F_{i+\frac{1}{2}}^n(Q^n, \tilde{u}_{i+\frac{1}{2}}^n) = \frac{1}{\Delta t} \begin{cases} \Delta x \sum_{l=0}^{k-1} Q_{i-l} + \int_{x_{i-k+\frac{1}{2}} - \alpha_{i+\frac{1}{2}}^n \Delta x}^{x_{i-k+\frac{1}{2}}} \tilde{q}(x; Q^n) dx, & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ \Delta x \sum_{l=0}^{k-1} Q_{i-l} - \int_{x_{i-k+\frac{1}{2}} - \alpha_{i+\frac{1}{2}}^n \Delta x}^{x_{i-k+\frac{1}{2}}} \tilde{q}(x; Q^n) dx, & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0. \end{cases} \quad (2.58)$$

where we used that  $\tilde{q}$  preserves the local mass.

We will provide explicit expressions for the integrals in Equation (2.58) when using the PPM method. For each control volume edge, denoted by  $i = 0, \dots, N$ , and  $y > 0$ , we define the following averages of the Piecewise-Parabolic approximation, as defined in Equation (2.36) for  $Q^n$  (Colella & Woodward, 1984):

$$F_{L,i+\frac{1}{2}}(Q^n, y) = \frac{1}{y} \int_{x_{i+\frac{1}{2}} - y}^{x_{i+\frac{1}{2}}} \tilde{q}(x; Q^n) dx, \quad (2.59)$$

and

$$F_{R,i+\frac{1}{2}}(Q^n, y) = \frac{1}{y} \int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}} + y} \tilde{q}(x; Q^n) dx. \quad (2.60)$$

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 (2.40) that:

$$F_{L,i+\frac{1}{2}}(Q^n, y) = \frac{1}{y} \int_{x_{i+\frac{1}{2}} - y}^{x_{i+\frac{1}{2}}} q_i(x; Q^n) dx = q_{R,i} + \frac{(q_{6,i} - \Delta q_i)}{2\Delta x} y - \frac{q_{6,i}}{3\Delta x^2} y^2, \quad (2.61)$$

and

$$F_{R,i+\frac{1}{2}}(Q^n, y) = \frac{1}{y} \int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}} + y} q_{i+1}(x; Q^n) dx = q_{L,i+1} + \frac{(q_{6,i+1} + \Delta q_{i+1})}{2\Delta x} y - \frac{q_{6,i+1}}{3\Delta x^2} y^2. \quad (2.62)$$

The numerical flux function for PPM is then defined by:

$$F_{i+\frac{1}{2}}^n(Q^n, \tilde{u}_{i+\frac{1}{2}}^n) = \frac{1}{\Delta t} \begin{cases} \Delta x \sum_{l=0}^{k-1} Q_{i-l} + \Delta x \alpha_{i+\frac{1}{2}}^n F_{L,i+\frac{1}{2}}(Q^n, \alpha_{i+\frac{1}{2}}^n \Delta x) & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ \Delta x \sum_{l=0}^{k-1} Q_{i-l} + \Delta x \alpha_{i+\frac{1}{2}}^n F_{R,i+\frac{1}{2}}(Q^n, -\alpha_{i+\frac{1}{2}}^n \Delta x) & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases} \quad (2.63)$$

where  $\tilde{u}_{i+\frac{1}{2}}^n$  is the velocity used in the departure point estimation. In particular, if the CFL number is less than one, then:

$$F_{i+\frac{1}{2}}^n(Q^n, \tilde{u}_{i+\frac{1}{2}}^n) = \begin{cases} \tilde{u}_{i+\frac{1}{2}}^n F_{L,i+\frac{1}{2}}(Q^n, \tilde{u}_{i+\frac{1}{2}}^n \Delta t) & \text{if } \tilde{u}_{i+\frac{1}{2}}^n \geq 0, \\ \tilde{u}_{i+\frac{1}{2}}^n F_{R,i+\frac{1}{2}}(Q^n, -\tilde{u}_{i+\frac{1}{2}}^n \Delta t) & \text{if } \tilde{u}_{i+\frac{1}{2}}^n < 0, \end{cases} \quad (2.64)$$

If the monotonic scheme from Lin (2004) is employed, then  $F_{i+\frac{1}{2}}$  uses the stencil  $S_{i+\frac{1}{2}} = \{i-3, i-2, i-1, i, i+1, i+2, i+3\}$  and we need  $v = 4$  layers of ghost cells. Otherwise the stencil used by  $F_{i+\frac{1}{2}}$  is given by  $S_{i+\frac{1}{2}} = \{i-2, i-1, i, i+1, i+2\}$  for all the other schemes that we presented and  $v = 3$ . The accuracy of this flux formulation is described in Section A.5.3.

## 2.6 Numerical experiments

This section is dedicated to presenting the numerical results of the PPM and its variations discussed here. We will consider the following reconstruction schemes: PPM-0, PPM-PL07, PPM-CW84, and PPM-L04, which were presented in Section 2.4, as well as the departure point schemes RK1 and RK2 described in Section 2.3. The code used in this section can be found in Appendix B.

For all the simulations presented here, we will consider the spatial domain  $[0, 1]$  and the time interval  $[0, 5]$ . The relative change at time step  $n$  in the mass is computed as:

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

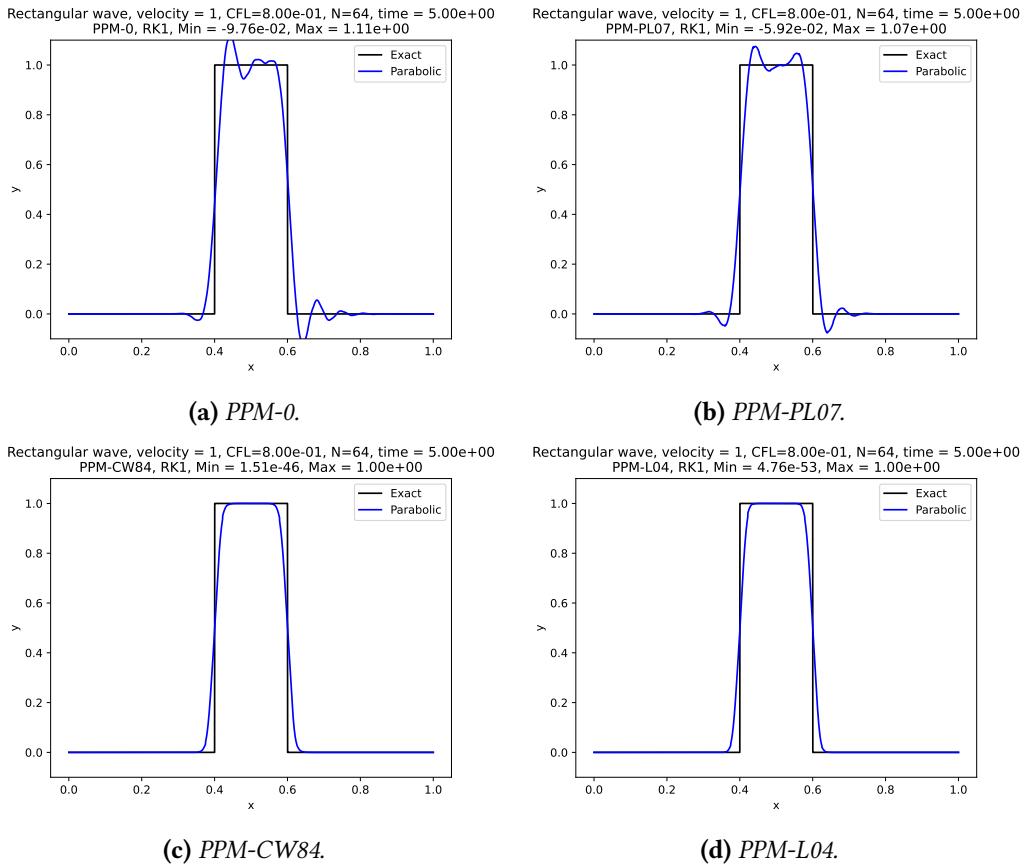
where  $M^n$  is given by Equation (2.22). For all the simulations, the mass is preserved with machine precision. Furthermore, we compute the initial average values  $Q_i(0)$  using the initial values of  $q_i^0$  at the control volume centroids for all simulations, which is second-order accurate by Proposition 2.2. In the error calculation, only when  $q_0$  is given by Equation (2.67), we replace  $Q_i(t^n)$  by its centroid value  $q_i(t^n)$ , which again gives a second-order

approximation by Proposition 2.2.

As a first numerical experiment, we consider a discontinuous initial condition given by:

$$q_0(x) = \begin{cases} 1 & \text{if } x \in [0.4, 0.6], \\ 0 & \text{otherwise.} \end{cases} \quad (2.65)$$

for the linear advection equation with constant velocity, which we adopt as  $u = 0.2$ . It is easy to check that the exact solution of Problem 2.1 is given by  $q_0(x - ut)$ . We will use a CFL number equal to 0.8. The spatial domain will be  $[0, 1]$ , and the time integration interval will be  $[0, 5]$ . Since we are assuming periodic boundary conditions, the period is equal to 5. Therefore, the simulations presented here will advect an initial profile for one time period. The departure schemes RK1 and RK2 compute the departure point exactly in this case, so we will only use the RK1 scheme. In Figure 2.4, we present the obtained results.



**Figure 2.4:** Linear advection experiment using a constant velocity equal to 0.2, a CFL number equal to 0.8,  $N = 64$  cells, and the initial condition is given by Equation (2.65). 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).

It is evident that the monotonic schemes exhibit a significant advantage. These schemes effectively prevent the strong oscillations observed in the non-monotonic schemes, as well as the generation of new extrema, which aligns with our expectations. Besides that, the scheme PPM-PL07 is less dispersive than the scheme PPM-0

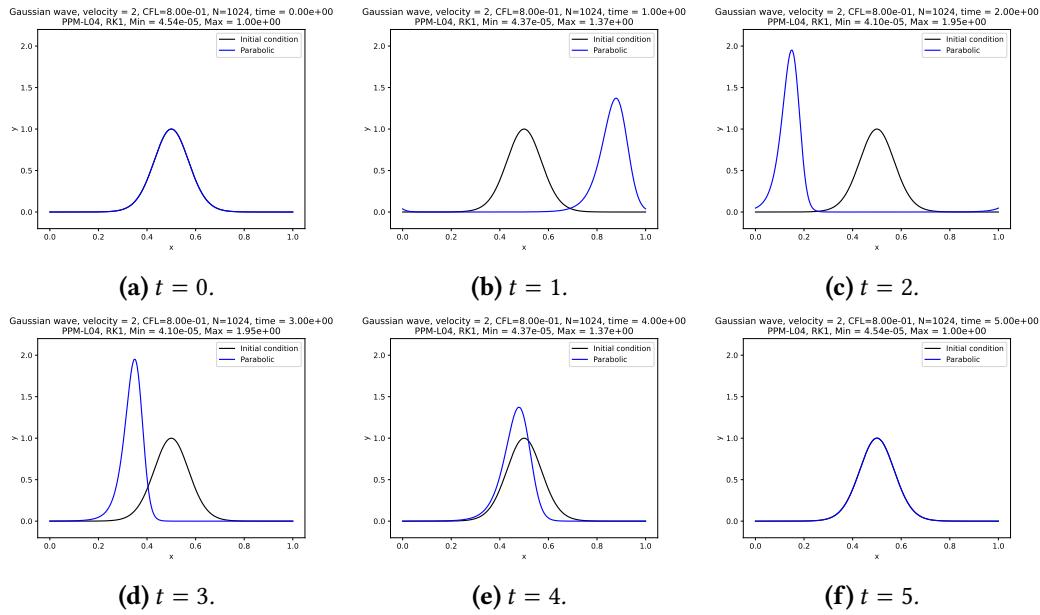
As a second experiment, we shall investigate the how the PPM schemes behave when the velocity is variable. We are going to consider the velocity

$$u(x, t) = u_0 \cos\left(\frac{\pi t}{T}\right) \sin^2\left(\pi\left(x - \frac{t}{T}\right)\right) + u_1. \quad (2.66)$$

We adopt the parameters  $u_0 = u_1 = 0.2$ , and  $T = 5$ . Following the approach in Trefethen (2000), we initialize the periodic Gaussian profile defined as:

$$q(x) = \exp(-10 \cos^2(\pi x)), \quad x \in [0, 1]. \quad (2.67)$$

The velocity function given by Equation (2.66) is based on the deformational flow test case in Nair and Lauritzen (2010). As the velocity is variable, we utilize the departure point schemes RK1 and RK2. In this case, the solution exhibits a period of 5 time units, meaning that the profile deforms and returns to its initial shape and position after 5 time units, allowing us to compute the error. Indeed, in Figure 2.5, we show how the solution behaves using a high-resolution ( $N = 1024$ ), the PPM-L04 scheme and the RK1 departure point scheme.



**Figure 2.5:** Linear advection experiment using the velocity from Equation (2.67), a CFL number equal to 0.8,  $N = 1024$  cells, and the initial condition is given by Equation (2.67) (2.5a). These figures show the advected profile at 1 (2.5b), 2 (2.5c), 3 (2.5d), 4 (2.5e), and 5 (2.5f) time units. We are using the PPM-L04 scheme with the RK1 departure point scheme.

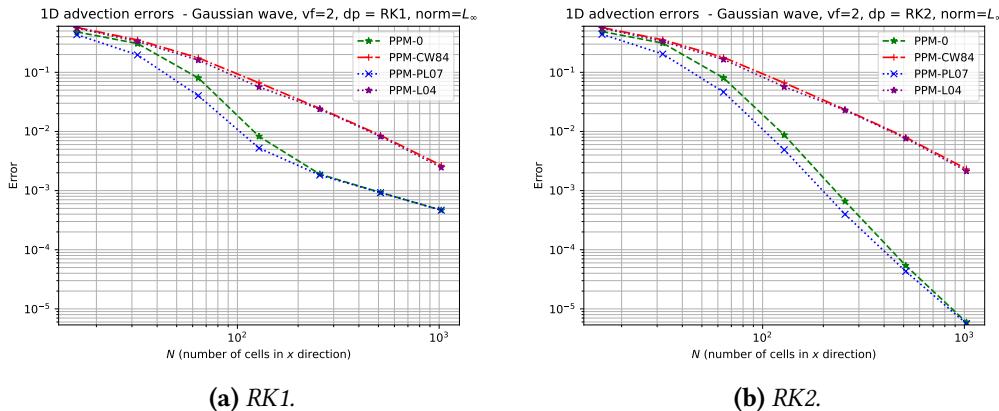
To investigate the error convergence, we employ  $(\Delta x^{(k)}, \Delta t^{(k)}, \lambda)$ -discretizations with  $\Delta x^{(k)} = 1/2^k$ ,  $\Delta t^{(k)} = 1/2^{k-2}$ , for  $k = 4, \dots, 10$  and  $\lambda = 4$ . To measure the accuracy, we consider the relative error in the maximum norm as follows:

$$E_k = \frac{\|Q^{N_T} - Q^0\|_{\infty, \Delta x}}{\|Q^0\|_{\infty, \Delta x}}.$$

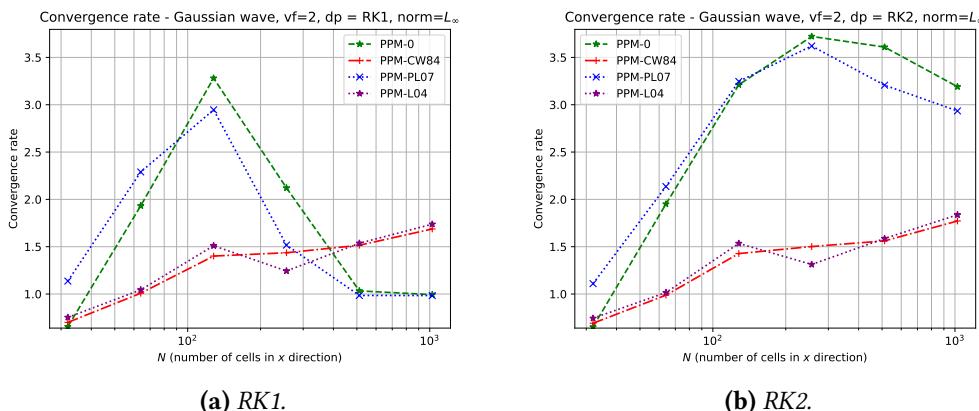
The convergence rate is defined by

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

The difference between the RK1 and RK2 schemes becomes clear when observing the relative error in Figure 2.6 and the convergence rate in Figure 2.7. The RK1 scheme results in a first-order error in the departure point, which dominates the total error for all PPM schemes. This observation is in agreement with the discussion in Section 2.3. On the other hand, when employing the RK2 scheme, we can achieve a third-order accuracy for the PPM-0 and PPM-PL07 schemes, surpassing the expected second-order accuracy of the departure point scheme. This experiment illustrates the impact of the departure point calculation error on the overall error. Furthermore, regardless of the departure point scheme, the PPM-PL07 scheme exhibits a smaller error. Among the monotonic schemes, the PPM-L04 has a slightly smaller error compared to PPM-CW84.



**Figure 2.6:** Relative error for different PPM schemes using the RK1 (left) and RK2 (right) departure point scheme for the initial condition given by Equation (2.67) and the variable velocity given by Equation (2.66).



**Figure 2.7:** As in Figure 2.6 but considering the convergence rate.

## 2.7 Concluding remarks

In this chapter, we provided a general overview of 1D FV-SL schemes for the advection equation. We discussed the three essential tasks involved in these schemes. The first task is the reconstruction of a function from its average values. We employed the PPM method introduced by Colella and Woodward (1984) and its variants. We were able to achieve third-order accuracy in the reconstruction process, even without imposing monotonicity constraints. The second task involves computing the departure point of the control volume edges. For this purpose, we utilized the first-order departure point calculation method from Colella and Woodward (1984) known as RK1. Additionally, we explored a second-order approach by employing a two-stages Runge-Kutta scheme (RK2) to integrate the departure point ordinary differential equation (ODE). Lastly, the third task entails computing the flux, which involves integrating the reconstructed function over a domain determined by the departure point.

From the numerical experiments, we observed that the PPM-PL07 (Putman & Lin, 2007), which uses fifth-order reconstruction at the edges, leads to third-order accuracy but is more accurate than the PPM-0 (Colella & Woodward, 1984) scheme, which uses fourth-order reconstruction at the edges. Regarding the monotonic schemes, we observed that both schemes were able to avoid overshoots, with PPM-L04 (Lin, 2004) being more accurate than the scheme PPM-CW84 proposed by Colella and Woodward (1984).

The difference between the departure point schemes became apparent when we performed a test with variable velocity. The simulation using the RK1 scheme resulted in a final first-order error, despite the scheme having third-order accuracy in space. However, the RK2 scheme preserved third-order accuracy despite being only second-order accurate. We expect that, in general, combining PPM with the RK2 scheme should result in at least second-order accuracy.

Clearly, the RK2 scheme is more computationally expensive since it requires time extrapolation and linear interpolation of the velocity field. One possible way to reduce its cost would be to use larger CFL numbers allowed by the FV-SL schemes, as discussed in Section 2.5.

# Chapter 3

## Two-dimensional finite-volume methods

In Chapter 2, we addressed the problem of solving the one-dimensional linear advection equation using the finite-volume method based on PPM. In this chapter, our focus shifts to solving the two-dimensional linear advection equation using the finite-volume method. This step is crucial in our work since, as we will explore in Chapter 5, solving the linear advection equation on the cubed-sphere relies on solving two-dimensional linear advection equations at each cube face, with interpolation between adjacent panels.

A natural approach to develop a finite-volume method for the two-dimensional linear advection equation would involve extending PPM to two dimensions. Indeed, Rančić (1992) proposed a piecewise bi-parabolic extension of PPM using a semi-Lagrangian temporal discretization. Further, this type of method can be extended to the cubed-sphere (Lauritzen et al., 2010). However, this method suffers from a significant drawback—its computationally expensive nature. As a popular alternative, dimension-splitting methods are often used, which replace the two-dimensional problem with a sequence of one-dimensional problems. For example, we can solve the two-dimensional linear advection equation by solving a series of one-dimensional linear advection equations using the PPM from Chapter 2. Moreover, in principle, we can employ any numerical method that solves the one-dimensional linear advection equation.

A comparison between two-dimensional and dimension-splitting semi-Lagrangian schemes on a plane was investigated by Y. Chen et al. (2017), utilizing the PPM as the one-dimensional solver and distorted two-dimensional grids. Their main conclusion was that dimension-splitting schemes are more sensitive to grid distortions, but they are computationally cheaper and more accurate than two-dimensional methods, particularly when dealing with large CFL numbers.

The primary objective of this chapter is to provide a comprehensive explanation of the dimension splitting method proposed by Lin and Rood (1996). This method is currently utilized in the FV3 dynamical core and is applied to the two-dimensional linear advection equation using the one-dimensional finite-volume schemes described in Chapter 2. Similar to Chapter 2, we start this chapter with a review of the integral form of the two-dimensional

advection equation in Section 3.1. In Section 3.2, we establish the framework for general two-dimensional finite-volume schemes. The dimension splitting method is presented in Section 3.3, and we showcase numerical experiments in Section 3.4.

## 3.1 Two-dimensional advection equation in integral form

### 3.1.1 Notation

This section is dedicated to extending the notation of Section 2.1.1. Based on definitions 2.1 and 2.3, we introduce the concepts of a  $(\Delta x, \Delta y)$ -grid and  $(\Delta x, \Delta y, \Delta t, \lambda)$  discretization. Throughout this chapter, we will use the notation  $\Omega = [a, b] \times [c, d]$  and  $v$  to represent a non-negative integer indicating the number of ghost cell layers in each boundary. We also use the notations  $\mathbb{R}_v^{N \times M} := \mathbb{R}^{(N+2v) \times (M+2v)}$  and  $\mathbb{R}_v^{(N+1) \times M} := \mathbb{R}^{(N+1+2v) \times (M+2v)}$ ,  $\mathbb{R}_v^{N \times (M+1)} := \mathbb{R}^{(N+2v) \times (M+1+2v)}$ .

**Definition 3.1** ( $(\Delta x, \Delta y)$ -grid). *Given  $\Omega$  and positive real numbers  $\Delta x$  and  $\Delta y$  such that  $\Delta x = (b - a)/N$ ,  $\Delta y = (d - c)/M$ , for positive integers  $N$  and  $M$ , we say that  $\Omega_{\Delta x, \Delta y} = (\Omega_{ij})_{i=-v+1, \dots, N+v}^{j=-v+1, \dots, M+v}$  is a  $(\Delta x, \Delta y)$ -grid for  $\Omega$  if*

$$\Omega_{ij} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}] = [a + (i - 1)\Delta x, a + i\Delta x] \times [c + (j - 1)\Delta y, c + j\Delta y],$$

$\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ ,  $\Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$ . Each  $\Omega_{ij}$  is called control volume or cell. The cell centroids  $(x_i, y_j)$  are defined by

$$x_i = \frac{1}{2}(x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}), y_j = \frac{1}{2}(y_{j+\frac{1}{2}} + y_{j-\frac{1}{2}}).$$

**Remark 3.1.** If  $1 \leq i \leq N, 1 \leq j \leq M$ , we refer to  $(i, j)$  as an interior index; otherwise,  $(i, j)$  is considered a ghost cell index and we say the  $\Omega_{ij}$  is a ghost cell.

**Definition 3.2** ( $(\Delta x, \Delta y, \Delta t, \lambda)$ -discretization). *Given  $\Omega \times [0, T]$ , and positive real numbers  $\Delta x, \Delta y$  and  $\Delta t$ , we say that  $(\Omega_{\Delta x, \Delta y}, T_{\Delta t})$  is a  $(\Delta x, \Delta y, \Delta t, \lambda)$ -discretization of  $\Omega \times [0, T]$  if  $\Omega_{\Delta x, \Delta y}$  is a  $(\Delta x, \Delta y)$  grid for  $\Omega$  and  $T_{\Delta t}$  is a  $\Delta t$ -temporal grid for  $[0, T]$ ,  $\frac{\Delta t}{\Delta x} = \lambda$  and  $\frac{\Delta t}{\Delta y} = \lambda$ .*

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

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, we say that  $Q = (Q_{ij})_{i=-v+1, \dots, N+v}^{j=-v+1, \dots, M+v} \in \mathbb{R}_v^{N \times M}$  is a  $(\Delta x, \Delta y)$ -grid function.*

**Definition 3.4** ( $(\Delta x, \Delta y)$ -C grid wind). *For a  $(\Delta x, \Delta y)$ -grid, we say that  $(u, v)$  is a  $(\Delta x, \Delta y)$ -C grid wind if  $u = (u_{i+\frac{1}{2}, j})_{i=-v, \dots, N+v}^{j=-v+1, \dots, M+v} \in \mathbb{R}_v^{(N+1) \times M}$ ,  $v = (v_{i, j+\frac{1}{2}})_{i=-v+1, \dots, N+v}^{j=-v, \dots, M+v} \in \mathbb{R}_v^{N \times (M+1)}$ .*

Considering a function  $q : \Omega \times [0, T] \rightarrow \mathbb{R}$ , a vector field  $\mathbf{u} : \Omega \times [0, T] \rightarrow \mathbb{R}$ ,  $\mathbf{u} = (u, v)$ , a  $(\Delta x, \Delta y, \Delta t, \lambda)$ -discretization of  $\Omega \times [0, T]$ , we introduce the grid functions

## 3.1 | TWO-DIMENSIONAL ADVECTION EQUATION IN INTEGRAL FORM

$q^n \in \mathbb{R}_v^{N \times M}$ ,  $u^n \in \mathbb{R}_v^{(N+1) \times M}$ ,  $v^n \in \mathbb{R}_v^{N \times (M+1)}$ . Here,  $q_{ij}^n = q(x_i, y_j, t^n)$ ,  $u_{i+\frac{1}{2},j}^n = u(x_{i+\frac{1}{2}}, y_j, t^n)$ ,  $v_{i,j+\frac{1}{2}}^n = v(x_i, y_{j+\frac{1}{2}}, t^n)$ . These grid functions represent the discrete values of  $q$  and  $\mathbf{u}$  at the cell centroids and edges, respectively, for each time level  $t^n$  (Figure 2.2). We shall also use the notations  $q_{i+\frac{1}{2},j}^n = q(x_{i+\frac{1}{2}}, y_j, t^n)$  and  $q_{i,j+\frac{1}{2}}^n = q(x_i, y_{j+\frac{1}{2}}, t^n)$ .



**Figure 3.1:** Illustration of  $(\Delta x, \Delta y)$ -grid function  $Q$  (black circles) and a  $(\Delta x \Delta y)$ -C grid wind  $u$  (blue squares) and  $v$  (red squares) and its ghost cell values (in gray) assuming biperiodicity.

We denote by  $\nabla \cdot (qu)$  the divergence operator:

$$\nabla \cdot (qu)(x, y, t) = [\partial_x(uq) + \partial_y(vq)](x, y, t). \quad (3.1)$$

We recall that we say the  $\mathbf{u}$  is **non-divergent** if  $\nabla \cdot \mathbf{u} = 0$ . We define the  $(\Delta x, \Delta y)$ -grid function  $\delta^n$  as the exact divergence of  $qu$  at the cell centers, namely

$$\delta_{ij}^n = \nabla \cdot (qu)(x_i, y_j, t^n). \quad (3.2)$$

In this chapter, our focus also lies on periodic grid functions. We define a  $(\Delta x, \Delta y)$ -grid

function  $Q$  as periodic if it satisfies the following conditions:

$$\begin{aligned} Q_{i,j} &= Q_{N+i,j}, & i = -v + 1, \dots, 0, & j = -v + 1, \dots, M + v, \\ Q_{i,j} &= Q_{i-N,j}, & i = N + 1, \dots, N + v, & j = -v + 1, \dots, M + v, \\ Q_{i,j} &= Q_{i,M+j}, & j = -v + 1, \dots, 0, & i = -v + 1, \dots, N + v, \\ Q_{i,j} &= Q_{i,j-M}, & j = M + 1, \dots, M + v, & i = -v + 1, \dots, N + v. \end{aligned}$$

We use the notation  $\mathbb{P}_v^{N \times M}$  represent the spaces of periodic  $(\Delta x, \Delta y)$ -grid functions. Similarly, we define a  $(\Delta x, \Delta y)$ -grid wind  $(u, v)$  as periodic if it meets the following requirements:

$$\begin{aligned} u_{i-\frac{1}{2},j} &= u_{N+i+\frac{1}{2},j}, & i = -v, \dots, -1, & j = -v + 1, \dots, M + v, \\ u_{i+\frac{1}{2},j} &= u_{i+\frac{1}{2}-N,j}, & i = N + 1, \dots, N + v, & j = -v + 1, \dots, M + v, \\ u_{i+\frac{1}{2},j} &= u_{i+\frac{1}{2},M+j}, & i = -v, \dots, N + 1 + v, & j = -v + 1, \dots, 0, \\ u_{i+\frac{1}{2},j} &= u_{i+\frac{1}{2},j-M}, & i = -v, \dots, N + 1 + v, & j = M + 1, \dots, M + v, \\ v_{i,j-\frac{1}{2}} &= v_{i,M+j+\frac{1}{2}}, & j = -v, \dots, -1, & i = -v + 1, \dots, N + v, \\ v_{i,j+\frac{1}{2}} &= v_{i,j+\frac{1}{2}-M}, & j = M + 1, \dots, M + v, & i = -v + 1, \dots, N + v, \\ v_{i,j+\frac{1}{2}} &= v_{N+i,j+\frac{1}{2}}, & j = -v, \dots, M + 1 + v, & i = -v + 1, \dots, 0, \\ v_{i,j+\frac{1}{2}} &= c_{i-N,j+\frac{1}{2}}, & j = -v, \dots, N + 1 + v, & i = N + 1, \dots, N + v. \end{aligned}$$

In this case, we use the notation  $u \in \mathbb{P}_v^{(N+1) \times M}$ ,  $v \in \mathbb{P}_v^{N \times (M+1)}$ .

For a grid function  $Q$  we also use the notations:

$$\begin{aligned} Q_{\times,j} &:= (Q_{-v+1,j}, \dots, Q_{N+v,j}) \in \mathbb{R}_v^N, \\ Q_{i,\times} &:= (Q_{i,-v+1}, \dots, Q_{i,M+v}) \in \mathbb{R}_v^M. \end{aligned}$$

Given  $Q = (Q_{ij}) \in \mathbb{P}_{v,P}^{N \times M}$ , we define the  $p$ -norm by

$$\|Q\|_{p,\Delta x \times \Delta y} = \begin{cases} \left( \sum_{i=1}^N \sum_{j=1}^M |Q_{ij}|^p \right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty, \\ \max_{i=1, \dots, N, j=1, \dots, M} |Q_{ij}| & \text{otherwise.} \end{cases} \quad (3.3)$$

We also introduce the centered difference notation:

$$\delta_x h(x_i, y, t) = h(x_{i+\frac{1}{2}}, y, t) - h(x_{i-\frac{1}{2}}, y, t), \quad (3.4)$$

$$\delta_y h(x, y_j, t) = h(x, y_{j+\frac{1}{2}}, t) - h(x, y_{j-\frac{1}{2}}, t), \quad (3.5)$$

for any function  $h : \Omega \times [0, T] \rightarrow \mathbb{R}$ . Additionally, we introduce the average value of  $q$  in the control volume  $\Omega_{ij}$  at time  $t$ , denoted as  $Q_{ij}(t)$ , defined by:

$$Q_{ij}(t) = \frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q(x, y, t) dx. \quad (3.6)$$

Moreover, we define the  $(\Delta x, \Delta y)$ -grid function of average values as  $Q(t) = (Q_{ij}(t))_{i=-v+1, \dots, N+v}^{j=-v+1, \dots, M+v}$ .

For the consideration of periodic boundary conditions, we can define spaces of periodic functions over the interval  $\Omega$  as follows:

$$\mathcal{S}_P(\Omega) = \{q : \mathbb{R}^2 \times [0, +\infty[ \rightarrow \mathbb{R} : q(x + b - a, y + d - c, t) = q(x, y, t), \quad \forall x, y \in \mathbb{R}, \quad t \geq 0\}.$$

Similarly, the space of  $k$ -times periodically differentiable functions  $C_P^k(\Omega)$  can be defined as:

$$C_P^k(\Omega) = \mathcal{S}_P(\Omega) \cap C^k(\mathbb{R}^2 \times [0, \infty[),$$

where  $C^k(\mathbb{R}^2 \times [0, +\infty[)$  denotes the space of functions that are  $k$  times continuously differentiable in both the spatial and temporal variables. In summary,  $\mathcal{S}_P(\Omega)$  represents the space of periodic functions, and  $C_P^k(\Omega)$  represents the space of  $k$ -times periodically differentiable functions over  $\Omega$  subject to periodic boundary conditions.

### 3.1.2 The 2D advection equation

Let us consider a velocity field given by  $\mathbf{u} = (u, v)$ , where  $u$  is the velocity in  $x$ -direction and  $v$  is the velocity in  $x$  and  $y$  direction and  $u, v \in C_P^1(\Omega)$ . The two-dimensional advection equation in its differential form in a domain  $\Omega$  associated to the velocity field  $\mathbf{u}$  and assuming biperiodic boundary conditions is given by:

$$\begin{cases} [\partial_t q + \partial_x(uq) + \partial_y(vq)](x, y, t) = 0, & \forall (x, y, t) \in \mathbb{R}^2 \times ]0, +\infty[, \\ q(a, y, t) = q(b, y, t), & \forall y \in [c, d], \quad \forall t \geq 0, \\ q(x, c, t) = q(x, d, t), & \forall x \in [a, b], \quad \forall t \geq 0, \\ q_0(x) = q(x, y, 0), & \forall (x, y) \in \Omega. \end{cases} \quad (3.7)$$

A classical or strong solution to the two-dimensional advection equation is a  $C_P^1(\Omega)$  function  $q$  satisfying Equation (3.7). As we did in Section 2.1, our goal is to deduce an integral form of Equation (3.7). Thus, let us consider  $[x_1, x_2] \times [y_1, y_2] \subset \Omega$  and  $[t_1, t_2] \subset [0, +\infty[$ . Integrating Equation (3.7) over  $[x_1, x_2] \times [y_1, y_2]$  yields:

$$\begin{aligned} \frac{d}{dt} \left( \int_{x_1}^{x_2} \int_{y_1}^{y_2} q(x, y, t) dx dy \right) &= - \int_{y_1}^{y_2} \left( (uq)(x_2, y, t) - (uq)(x_1, y, t) \right) dy \\ &\quad - \int_{x_1}^{x_2} \left( (vq)(x, y_2, t) - (vq)(x, y_1, t) \right) dx. \end{aligned} \quad (3.8)$$

Integrating Equation (3.8) over the time interval  $[t_1, t_2]$ , we have:

$$\begin{aligned} \int_{x_1}^{x_2} \int_{y_1}^{y_2} q(x, y, t_{n+1}) dx dy &= \int_{x_1}^{x_2} \int_{y_1}^{y_2} q(x, y, t_n) dx dy \\ &\quad - \int_{t_1}^{t_2} \int_{y_1}^{y_2} \left( (uq)(x_2, y, t) - (uq)(x_1, y, t) \right) dy dt \\ &\quad - \int_{t_1}^{t_2} \int_{x_1}^{x_2} \left( (vq)(x, y_2, t) - (vq)(x, y_1, t) \right) dx dt. \end{aligned} \quad (3.9)$$

Equation (3.9) is the integral form of Equation (3.7). We say that  $q$  is a weak solution to the advection equation (3.7) if  $q$  satisfies the integral form (3.9),  $\forall [x_1, x_2] \times [y_1, y_2] \subset \Omega^\circ$  and  $\forall [t_1, t_2] \subset [0, +\infty[$ . We summarize the weak version of Equation (3.7) in Problem (3.1).

**Problem 3.1.** Given an initial condition  $q_0$  and a velocity function  $\mathbf{u} = (u, v)$  we would like to find a weak solution  $q$  of the two-dimensional advection equation in its integral form:

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

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

Similarly to Section 2.1, Equation (3.7) and Problem (3.1) are equivalent when  $q, \mathbf{u} \in C_P^1(\Omega)$ . For Problem 3.1, the total mass in  $\Omega$  is defined by:

$$M_\Omega(t) = \int_{\Omega} q(x, y, t) dx dy, \quad \forall t \in [0, T], \quad (3.10)$$

and is conserved within time:

$$M_\Omega(t) = M_\Omega(0), \quad \forall t \in [0, T]. \quad (3.11)$$

Considering a  $(\Delta x, \Delta y, \Delta t, \lambda)$  discretization of  $D = \Omega \times [0, T]$  and 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 (3.9), we obtain:

$$\begin{aligned} Q_{ij}(t_{n+1}) &= Q_{ij}(t_n) - \frac{\Delta t}{\Delta x \Delta y} \delta_x \left( \frac{1}{\Delta t} \int_{t_1}^{t_2} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (uq)(x_i, y, t) dy dt \right) \\ &\quad - \frac{\Delta t}{\Delta x \Delta y} \delta_y \left( \frac{1}{\Delta t} \int_{t_1}^{t_2} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (vq)(x, y_j, t) dx dt \right), \end{aligned} \quad (3.12)$$

where we are using the centered finite-difference notation. Now we can define a discretized version of Problem 3.1 as Problem 3.2.

**Problem 3.2.** Assume the framework of Problem 3.1 and consider a  $(\Delta x, \Delta y, \Delta t, \lambda)$ -

discretization of  $\Omega \times [0, T]$ . Since we are in the framework of Problem 3.1, it follows that:

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

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  $Q_{ij}(t_n)$ ,  $\forall i = 1, \dots, N$ ,  $\forall j = 1, \dots, M$ ,  $\forall n = 0, \dots, N_T - 1$ , given the initial values  $Q_{ij}(0)$ ,  $\forall i = 1, \dots, N$ ,  $\forall j = 1, \dots, M$ . In other words, we aim to find the average values of  $q$  in each control volume  $\Omega_{ij}$  at the specified time instances.

It is important to note that no approximations have been made in Problems (3.1) and (3.2).

## 3.2 The finite-volume approach

Finally, we define the 2D-FV scheme problem as follows in Problem 3.3.

**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:

$$\begin{aligned} Q_{ij}^{n+1} &= Q_{ij}^n - \lambda \delta_i F_{ij}^n - \lambda \delta_j G_{ij}^n, \\ \forall i &= 1, \dots, N, \quad \forall j = 1, \dots, M, \quad \forall n = 0, \dots, N_T - 1, \end{aligned} \tag{3.13}$$

where  $\delta_i F_{ij}^n = F_{i+\frac{1}{2},j}^n - F_{i-\frac{1}{2},j}^n$ ,  $\delta_j G_{ij}^n = G_{i,j+\frac{1}{2}}^n - G_{i,j-\frac{1}{2}}^n$  and  $Q^n \in \mathbb{P}_v^{N \times M}$  is intended to be an approximation of  $Q(t_n) \in \mathbb{P}_v^{N \times M}$  in some sense. We define  $Q_{ij}^0 = Q_{ij}(0)$  or  $Q_{ij}^0 = q_{ij}^0$ .

The term  $F_{i+\frac{1}{2},j}^n$  is known as numerical flux in the  $x$  direction and it approximates  $\frac{1}{\Delta t \Delta y} \int_{t_n}^{t_{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (uq)(x_{i+\frac{1}{2}}, y, t) dy dt$ ,  $\forall i = 0, 1, \dots, N$ , and  $G_{i,j+\frac{1}{2}}^n$  is known as numerical flux in the  $y$  direction and it approximates  $\frac{1}{\Delta t \Delta x} \int_{t_n}^{t_{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (vq)(x, y_{j+\frac{1}{2}}, t) dx dt$ ,  $\forall j = 0, 1, \dots, M$ , or, in other words, they estimate the time-averaged fluxes at the control volume  $\Omega_{ij}$  boundaries.

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

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

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

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

straightforward computation:

$$\begin{aligned} M^{n+1} &= \Delta x \sum_{i=1}^N \sum_{j=1}^M Q_{ij}^{n+1} = M^n - \Delta t \sum_{i=1}^N \sum_{j=1}^M (F_{i+\frac{1}{2},j}^n - F_{i-\frac{1}{2},j}^n) - \Delta t \sum_{i=1}^N \sum_{j=1}^M (G_{i,j+\frac{1}{2}}^n - G_{i,j-\frac{1}{2}}^n) \\ &= M^n - \Delta t \sum_{j=1}^M (F_{N+\frac{1}{2},j}^n - F_{\frac{1}{2},j}^n) - \Delta t \sum_{i=1}^N (G_{i,M+\frac{1}{2}}^n - G_{i,\frac{1}{2}}^n) = M^n, \end{aligned}$$

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

As we mentioned in Problem 3.3, the initial condition may be assumed as  $q_{ij}^0$  or  $Q_{ij}(0)$ . For two-dimensional simulations, we are going to assume  $q_{ij}^0$  as initial data to avoid the computation of integrals. Furthermore, the errors will be calculated using the values  $q_{ij}^n$  instead of  $Q_{ij}(t_n)$ . Similarly to Proposition 2.2, 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  $|Q_{ij}(t^n) - q_{ij}^n| \leq C_1 \Delta x^2 + C_2 \Delta x \Delta y + C_3 \Delta y^2$ , where  $C_1, C_2$  and  $C_3$  are constants.*

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

In order to check the consistency of 2D-FV, it is useful to use the notion of discrete divergence.

**Definition 3.5** (Discrete divergence). *For Problem 3.3, we define the discrete divergence as a  $(\Delta x, \Delta y)$ -grid function  $\mathbb{D}^n(Q^n, u^n, v^n) \in \mathbb{P}_v^{N \times M}$  given by:*

$$\mathbb{D}_{ij}^n(Q^n, u^n, v^n) = \frac{1}{\Delta t} \left( \frac{\delta_i F_{ij}^n}{\Delta x} + \frac{\delta_j G_{ij}^n}{\Delta y} \right), \quad i = 1, \dots, N, \quad j = 1, \dots, M. \quad (3.14)$$

With the aid of the discrete divergence, we may rewrite Equation (3.13) as:

$$Q^{n+1} = Q^n - \Delta t \mathbb{D}^n(Q^n, u^n, v^n), \quad (3.15)$$

Notice that if we replace  $Q^n$  by the exact solution  $Q(t^n)$  in Equation (3.15), we have

$$Q(t^{n+1}) = Q(t^n) - \Delta t \mathbb{D}^n(Q(t^n), u^n, v^n) - \Delta t \tau^n, \quad (3.16)$$

where  $\tau^n \in \mathbb{P}_v^{N \times M}$  is the local truncation error (LTE). Rearranging the terms of Equation (3.16), we obtain:

$$\tau^n = \frac{Q(t^{n+1}) - Q(t^n)}{\Delta t} - \mathbb{D}^n(Q(t^n), u^n, v^n). \quad (3.17)$$

We define the consistency of the 2D-FV scheme as follows.

**Definition 3.6** (Consistency). *Let us consider the framework of Problem 3.3. A 2D-FV scheme is said to be consist in the  $p$ -norm if for any sequence of  $(\Delta x^{(k)}, \Delta y^{(k)}, \Delta t^{(k)}, \lambda)$ -discretizations,*

$k \in \mathbb{N}$ , with  $\lim_{k \rightarrow \infty} \Delta x^{(k)} = \lim_{k \rightarrow \infty} \Delta y^{(k)} = \lim_{k \rightarrow \infty} \Delta t^{(k)} = 0$ , we have:

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

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

$$\max_{1 \leq n \leq N_T^{(k)}} \|\tau^n\|_{p, \Delta x^{(k)} \times \Delta y^{(k)}} = O(\Delta x^d).$$

The relationship between consistency and convergence is explained in Section A.6. If  $q$  satisfies Equation (3.7), it can be observed that consistency is equivalent to the following:

$$\max_{1 \leq n \leq N_T^{(k)}} \|\delta^n - \mathbb{D}^n(Q^n, u^n, v^n)\|_{p, \Delta x^{(k)} \times \Delta y^{(k)}} = O(\Delta x^d),$$

where  $\delta^n \in \mathbb{P}_v^{N \times M}$  is defined in Equation (3.2). Therefore, we can determine whether a 2D-FV scheme is consistent by comparing the discrete divergence to the exact divergence.

### 3.3 Dimension splitting

This section aims to demonstrate how a 2D-FV scheme, such as the one presented in Problem 3.3, can be constructed using 1D-FV schemes through a technique known as dimension splitting. Before introducing the dimension splitting scheme proposed by Lin and Rood (1996), it is helpful to examine general operator splitting schemes, as the dimension splitting technique is a specific instance of operator splitting methods.

For a given time interval  $[0, T]$ , we utilize a  $\Delta t$ -temporal grid. Let us consider the abstract Cauchy problem.

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

for  $n = 0, \dots, N_T - 1$ , where  $q(t) \in \mathcal{B}$  for some Banach space  $\mathcal{B}$ , and  $A : \mathcal{B} \rightarrow \mathcal{B}$  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  $A_1, A_2 : \mathcal{B} \rightarrow \mathcal{B}$ , 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  $q(t_0 + \Delta t)$  as  $q^{21}(t^n + \Delta t)$  with an error of  $O(\Delta t)$  if  $A_1$  and  $A_2$

do not commute. Otherwise, this method is exact. This approach is known as Lie-Trotter splitting. It's worth noting that the Lie-Trotter splitting can also be performed in reverse order when solving the sub-problems:

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

and

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

and again we estimate  $q(t^{n+1})$  by  $q^{12}(t^{n+1})$  with error  $O(\Delta t)$ . As noted by Strang (1968), we can consider the following equation to approximate  $q(t^{n+1})$  using a second-order ( $O(\Delta t^2)$ ) symmetric scheme:

$$q^*(t^{n+1}) = \frac{q^{21}(t^{n+1}) + q^{12}(t^{n+1})}{2}, \quad (3.18)$$

This scheme is referred to as the average Lie-Trotter splitting (Holden et al., 2010). The process of averaging two Lie-Trotter splittings is a specific case of methods known as weighted sequential splitting methods in the literature. Furthermore, this scheme averaging process can be extended to achieve higher-order schemes (Jia & Li, 2011). For an analysis of the accuracy of weighted sequential splitting methods, we recommend referring to Csomós et al. (2005).

It is worth noting that one of the most commonly used second-order splitting schemes in the literature is the Strang splitting (Strang, 1968). This scheme requires solving three sub-problems per time-step, with one of them at time  $t_n + \frac{\Delta t}{2}$ . In contrast, the average Lie-Trotter splitting requires solving four sub-problems per time-step. Consequently, the Strang splitting is computationally more efficient. However, as we will observe in this chapter, when applied to the linear advection equation, the average Lie-Trotter splitting allows for a modification that eliminates a splitting error arising from considering a constant scalar field and non-divergent velocity (Lin & Rood, 1996).

To move towards the scheme from Lin and Rood (1996), let us consider Problem 3.1 in its differential form (Equation (3.7)). We are going to consider  $N + 2v$  one-dimensional advection equations in the  $x$ -direction:

$$[\partial_t q^x + \partial_x(uq^x)](x, y_j, t) = 0,$$

and the  $N + 2v$  one-dimensional advection equations in the  $y$ -direction

$$[\partial_t q^y + \partial_y(vq^y)](x_i, y, t) = 0.$$

We shall assume that these problems are solved using a 1D-FV scheme as in Problem 2.4 with numerical flux functions  $F_{i+\frac{1}{2},j}^n(Q_{x,j}^n, \tilde{u}_{i+\frac{1}{2},j}^n)$  and  $G_{i,j+\frac{1}{2}}^n(Q_{i,x}^n, \tilde{v}_{i,j+\frac{1}{2}}^n)$ , respectively where  $\tilde{u}_{i+\frac{1}{2},j}^n$  is the time-averaged used in the departure point estimation in the  $x$  direction and  $\tilde{v}_{i,j+\frac{1}{2}}^n$  is the time-averaged used in the departure point estimation in the  $y$  direction. In this

work, we assume that the fluxes  $F_{i+\frac{1}{2},j}^n(Q_{\times,j}^n, \tilde{u}_{i+\frac{1}{2},j}^n)$  and  $G_{i,j+\frac{1}{2}}^n(Q_{i,\times}^n, \tilde{v}_{i,j+\frac{1}{2}}^n)$  are computed using the PPM flux assuming that the CFL number is less than one (see Equation (2.64)).

More explicitly, as in Section 2.4.1, we have a piecewise-parabolic approximation in the  $x$  direction:

$$\begin{cases} q_{ij}^x(x; Q_{\times,j}^n) = q_{L,ij}^x + \Delta q_{ij}^x z_i(x) + q_{6,ij}^x z_i(x)(1 - z_i(x)), \\ z_i(x) = \frac{x - x_{i-\frac{1}{2}}}{\Delta x}, \quad x \in X_i, \\ q_{L,ij}^x = q_{i-\frac{1}{2},j}^n + O(\Delta x^2), \\ q_{R,ij}^x = q_{i+\frac{1}{2},j}^n + O(\Delta x^2), \\ \Delta q_{ij}^x = q_{R,ij}^x - q_{L,ij}^x, \\ q_{6,ij}^x = 6 \left( Q_{ij}^n - \frac{(q_{L,ij}^x + q_{R,ij}^x)}{2} \right), \end{cases} \quad (3.19)$$

for  $i = 1, \dots, N$ ,  $j = -v + 1, \dots, M + v$ , and we also construct a piecewise-parabolic approximation in the  $y$  direction:

$$\begin{cases} q_{ij}^y(y; Q_{i,\times}^n) = q_{L,ij}^y + \Delta q_{ij}^y z_j(y) + q_{6,ij}^y z_j(y)(1 - z_j(y)), \\ z_j(y) = \frac{y - y_{j-\frac{1}{2}}}{\Delta y}, \quad y \in Y_j, \\ q_{L,ij}^y = q_{i,j-\frac{1}{2}}^n + O(\Delta y^2), \\ q_{R,ij}^y = q_{i,j+\frac{1}{2}}^n + O(\Delta y^2), \\ \Delta q_{ij}^y = q_{R,ij}^y - q_{L,ij}^y, \\ q_{6,ij}^y = 6 \left( Q_{ij}^n - \frac{(q_{L,ij}^y + q_{R,ij}^y)}{2} \right), \end{cases} \quad (3.20)$$

for  $i = -v + 1, \dots, N + v$ ,  $j = 1, \dots, M$ . The values  $q_{L,ij}^x$ ,  $q_{R,ij}^x$ ,  $q_{L,ij}^y$ , and  $q_{R,ij}^y$ , which approximate the values of  $q$  at C-grid wind positions, are computed using one of the schemes PPM-0, PPM-PL07, PPM-CW84, or PPM-L04, as described in Sections 2.4.1 and 2.4.2. These approximations are expected to be second-order accurate because the given average values are computed on the 2D control volume  $\Omega_{ij}$  instead of the 1D control volumes  $X_i$  or  $Y_j$ . Then, we may express the fluxes as in Equation (2.64), namely:

$$F_{i+\frac{1}{2},j}^n(Q_{\times,j}^n, \tilde{u}_{i+\frac{1}{2},j}^n) = \tilde{u}_{i+\frac{1}{2},j}^n \times \begin{cases} q_{R,ij}^x + \frac{1}{2}(q_{6,ij}^x - \Delta q_{ij}^x) \tilde{c}_{i+\frac{1}{2},j}^{x,n} + \frac{1}{3} q_{6,ij}^x (\tilde{c}_{i+\frac{1}{2},j}^{x,n})^2, & \text{if } \tilde{u}_{i+\frac{1}{2},j}^n > 0, \\ q_{L,i+1,j}^x - \frac{1}{2}(q_{6,i+1,j}^x + \Delta q_{i+1,j}^x) \tilde{c}_{i+\frac{1}{2},j}^{x,n} - \frac{1}{3} q_{6,i+1,j}^x (\tilde{c}_{i+\frac{1}{2},j}^{x,n})^2, & \text{if } \tilde{u}_{i+\frac{1}{2},j}^n \leq 0, \end{cases} \quad (3.21)$$

for  $i = 0, \dots, N$ ,  $j = -v + 1, \dots, M + v$ , and

$$G_{i,j+\frac{1}{2}}^n(Q_{i,\times}^n, \tilde{v}_{i,j+\frac{1}{2}}^n) = \tilde{v}_{i,j+\frac{1}{2}}^n \times \begin{cases} q_{R,ij}^y + \frac{1}{2}(q_{6,ij}^y - \Delta q_{ij}^y) \tilde{c}_{i,j+\frac{1}{2}}^{y,n} + \frac{1}{3} q_{6,ij}^y (\tilde{c}_{i,j+\frac{1}{2}}^{y,n})^2, & \text{if } \tilde{v}_{i,j+\frac{1}{2}}^n > 0, \\ q_{L,i,j+1}^y - \frac{1}{2}(q_{6,i,j+1}^y + \Delta q_{i,j+1}^y) \tilde{c}_{i,j+\frac{1}{2}}^{y,n} - \frac{1}{3} q_{6,i,j+1}^y (\tilde{c}_{i,j+\frac{1}{2}}^{y,n})^2, & \text{if } \tilde{v}_{i,j+\frac{1}{2}}^n \leq 0, \end{cases} \quad (3.22)$$

for  $i = -v + 1, \dots, N + v$ ,  $j = 0, \dots, M$ , and

$$\begin{aligned}\tilde{c}_{i+\frac{1}{2},j}^{x,n} &= \tilde{u}_{i+\frac{1}{2},j}^{x,n} \frac{\Delta t}{\Delta x}, \\ \tilde{c}_{i,j+\frac{1}{2}}^{y,n} &= \tilde{v}_{i,j+\frac{1}{2}}^{y,n} \frac{\Delta t}{\Delta y},\end{aligned}$$

are the time-averaged CFL numbers in the  $x$  and  $y$  directions, respectively, which are assumed to have absolute value less than one. The time-averaged winds are computed using the RK1 and RK2 schemes from Section 2.3. When we use the PPM-L04 scheme, we set  $v = 4$ ; otherwise,  $v = 3$ .

We introduce the auxiliary grid functions  $\mathbf{F}$  and  $\mathbf{G}$ , both belonging to  $\mathbb{R}_v^{N \times M}$ , given by:

$$\mathbf{F}_{ij}(Q^n, \tilde{u}^n) = -\lambda \left( F_{i+\frac{1}{2},j}^n(Q_{\times,j}^n, \tilde{u}_{i+\frac{1}{2},j}^n) - F_{i-\frac{1}{2},j}^n(Q_{\times,j}^n, \tilde{u}_{i-\frac{1}{2},j}^n) \right),$$

for  $i = 1, \dots, N$ ,  $j = -v + 1, \dots, M + v$ , and

$$\mathbf{G}_{ij}(Q^n, \tilde{v}^n) = -\lambda \left( G_{i,j+\frac{1}{2}}^n(Q_{i,\times}^n, \tilde{v}_{i,j+\frac{1}{2}}^n) - G_{i,j-\frac{1}{2}}^n(Q_{i,\times}^n, \tilde{v}_{i,j-\frac{1}{2}}^n) \right),$$

for  $i = -v + 1, \dots, N + v$ ,  $j = 1, \dots, M$ , which are the numerical flux update of the 1D-FV schemes in the  $x$  and  $y$  direction, respectively. The Lie-Trotter splitting is obtained by solving the advection in the  $x$  direction

$$Q^{x,n} = Q^n + \mathbf{F}(Q^n, \tilde{u}^n),$$

for  $j = v + 1, \dots, M + v$  (Figure 3.2a), and then we advect in the  $y$  direction with initial data  $Q^{x,n}$

$$Q^{yx,n} = Q^{x,n} + \mathbf{G}(Q^{x,n}, \tilde{v}^n),$$

for  $i = -v + 1, \dots, N + v$  (Figure 3.2b). To get the average Lie-Trotter splitting we repeat the process in the reverse order by solving the advection equation in the  $y$  direction

$$Q^{y,n} = Q^n + \mathbf{G}(Q^n, \tilde{v}^n),$$

for  $i = -v + 1, \dots, N + v$  (Figure 3.3a), and then we advect in the  $x$ -direction with initial data  $Q^{y,n+1}$

$$Q^{xy,n} = Q^{y,n} + \mathbf{F}(Q^{y,n}, \tilde{u}^n),$$

for  $j = -v + 1, \dots, M + v$ , (Figure 3.3b) and thus we have the average Lie-Trotter solution:

$$\begin{aligned}Q^{n+1} &= \frac{(Q^{xy,n} + Q^{yx,n})}{2} \\ &= Q^n + \frac{1}{2} \mathbf{F}(Q^n, \tilde{u}^n) + \frac{1}{2} \mathbf{G}(Q^n, \tilde{v}^n) + \frac{1}{2} \mathbf{F} \left( Q^n + \mathbf{G}(Q^n, \tilde{v}^n), \tilde{u}^n \right) + \frac{1}{2} \mathbf{G} \left( Q^n + \mathbf{F}(Q^n, \tilde{u}^n), \tilde{v}^n \right),\end{aligned}$$

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):

$$Q^{n+1} = Q^n + \mathbf{F}\left(Q^n + \frac{1}{2}\mathbf{G}(Q^n, \tilde{v}^n), \tilde{u}^n\right) + \mathbf{G}\left(Q^n + \frac{1}{2}\mathbf{F}(Q^n, \tilde{u}^n), \tilde{v}^n\right). \quad (3.23)$$



**Figure 3.2:** Illustration of the Lie-Trotter splitting applied in the  $x$  direction (operator  $\mathbf{F}$ ) and then in the  $y$  direction (operator  $\mathbf{G}$ ). Interior cells are depicted using black lines, while ghost cells are depicted using gray lines. All the winds shown are the ones used in the RK1 departure point scheme. If the RK2 scheme is used, an additional layer of wind ghost values should be added at each boundary in (a) and (b).



**Figure 3.3:** Similar to Figure 3.2 but considering the Lie-Trotter splitting in reverse order.

The numerical flux functions defined in Chapter 2 are indeed linear the input  $Q$  if there are no monotonic constrain, but we are going to consider this scheme even when there are monotonic constraints since it requires fewer operations. Further, if we assume that  $q = \bar{q}$  is constant and  $\nabla \cdot \mathbf{u} = 0$  then the solution remains constant and then, assuming also

that  $\mathbf{u}$  does not depend on  $t$ , then  $\mathbf{F}$  and  $\mathbf{G}$  are given by

$$\begin{aligned}\mathbf{F}_{ij}(Q^n, \tilde{u}^n) &= -\bar{q}\lambda\delta_x u(x_i, y_j), \\ \mathbf{G}_{ij}(Q^n, \tilde{v}^n) &= -\bar{q}\lambda\delta_y v(x_i, y_j).\end{aligned}$$

However, if we compute the updated solution using Equation (3.23), we have that the error is given by

$$Q_{ij}^{n+1} - \bar{q} = -\Delta t \left( \frac{\delta_x u(x_i, y_j)}{\Delta x} + \frac{\delta_y v(x_i, y_j)}{\Delta y} \right) - \Delta t^2 \bar{q} \left( \frac{\delta_y v \delta_x u(x_i, y_j) + \delta_x u \delta_y v(x_i, y_j)}{2\Delta x \Delta y} \right) \quad (3.24)$$

$$= \Delta t(O(\Delta x^2) + O(\Delta y^2)) - \Delta t^2 \bar{q} \left( \frac{\delta_y v \delta_x u(x_i, y_j) + \delta_x u \delta_y v(x_i, y_j)}{2\Delta x \Delta y} \right). \quad (3.25)$$

Thus, the terms in the equation above multiplied by  $\Delta t^2$  are related to a 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

$$Q^{n+1} = Q^n + \mathbf{F} \left( Q^n + \frac{1}{2} \mathbf{g}(Q^n, \tilde{v}^n), \tilde{u}^n \right) + \mathbf{G} \left( Q^n + \frac{1}{2} \mathbf{f}(Q^n, \tilde{u}^n), \tilde{v}^n \right), \quad (3.26)$$

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  $\Delta t^2$  that appeared in Equation (3.24) when we apply these operators for a constant grid function  $Q^n$  and a non-divergent velocity field in Equation (3.26). Therefore, these inner advective operators eliminate the splitting error for a constant field and a non-divergent velocity field.

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

**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:

$$\mathbf{D}^n = -\frac{1}{\Delta t} \left[ \mathbf{F} \left( Q^n + \frac{1}{2} \mathbf{g}(Q^n, \tilde{v}^n), \tilde{u}^n \right) + \mathbf{G} \left( Q^n + \frac{1}{2} \mathbf{f}(Q^n, \tilde{u}^n), \tilde{v}^n \right) \right], \quad (3.27)$$

and as pointed out in Section A.6, we may use the discrete divergence to check the scheme consistency.

## 3.4 Numerical experiments

To assess the dimension splitting schemes introduced 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 FV-SL schemes PPM-0, PPM-PL07, PPM-CW84, PPM-L04 with the departure point schemes RK1 and RK2 as in Section 2.6. For all simulations, the CFL number is set to 0.8, and the time integration interval is  $[0, 5]$ , which represents the period of the exact solution for all tests. We employ  $(\Delta x^{(k)}, \Delta y^{(k)}, \lambda)$ -discretizations with  $\Delta x^{(k)} = \Delta y^{(k)} = 1/2^k$ ,  $\Delta t^{(k)} = 0.05/2^{k-4}$  for  $k = 4, \dots, 10$ , and  $\lambda = 0.8$ . We introduce the relative error in the maximum norm:

$$E_k = \frac{\|Q^n - Q^0\|_{\infty, \Delta x \times \Delta y}}{\|Q^0\|_{\infty, \Delta x \times \Delta y}}.$$

The convergence rate is defined as in Section 2.6, as well as the total mass variation, which is preserved with machine precision in all experiments presented here. Notice that in the error computation, we use the centroid values instead of the exact average values to avoid computations of analytical integrals. As we pointed out in Proposition 3.1, this approximation introduces a second-order error. As a first test, we consider a constant velocity  $\mathbf{u} = (0.2, -0.2)$ . We are going to consider as initial condition a rectangular profile (Figure 3.4e) given by:

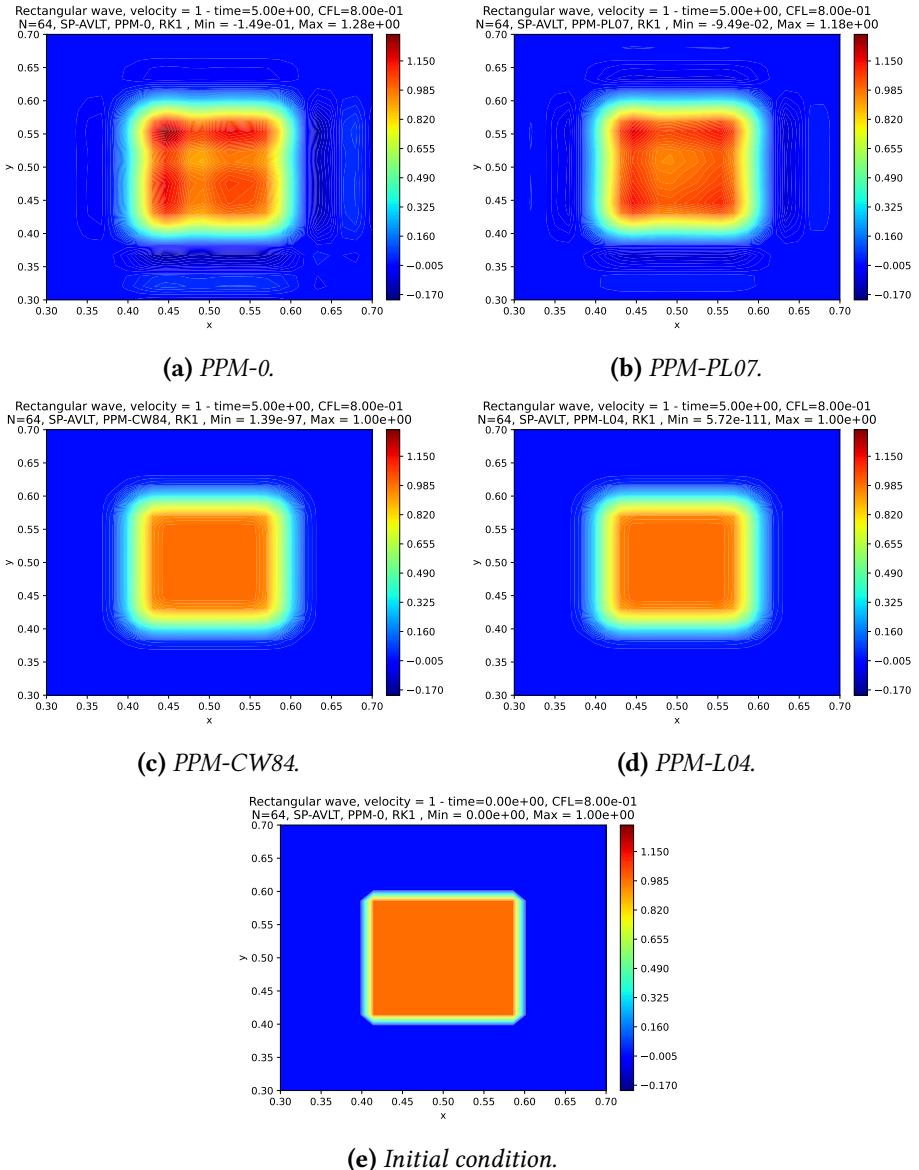
$$q_0(x, y) = \begin{cases} 1 & \text{if } (x, y) \in [0.4, 0.6] \times [0.4, 0.6], \\ 0 & \text{otherwise.} \end{cases} \quad (3.28)$$

The exact solution of Problem 3.1 in this case is  $q_0((x, y) - \mathbf{u}t)$ . Since the velocity field is constant, all the splitting schemes introduced in Section 3.3 are the same, so we will only consider the AVLT splitting. Additionally, it can be easily seen that the Lie-Trotter splitting is exact in this case (cf. eg. LeVeque, 1990, p. 202-203), meaning no splitting error is introduced. For the 1D schemes, we use RK1 to compute the departure point since this scheme is exact when the velocity is constant.

The conclusions for this test are very similar to the first 1D test from Section 2.6. This behavior is due to the fact that no splitting error is introduced when the velocity is constant. From Figure 3.4, it can be observed that the AVLT splitting preserves monotonicity when we use the monotonic 1D schemes PPM-CW84 and PPM-L04. For the non-monotonic schemes, as seen in Figure 2.4, PPM-PL07 produces less numerical dispersion than PPM-0, similar to the first 1D test.

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

$$q_0(x, y) = \exp(-10 \cos^2(\pi(x - 0.1)) \exp(-10 \cos^2(\pi y)) + \exp(-10 \cos^2(\pi(x + 0.1)) \exp(-10 \cos^2(\pi y))), \quad (3.29)$$



**Figure 3.4:** Linear advection experiment using a constant velocity equal to  $\mathbf{u} = (0.2, -0.2)$ , a CFL number equal to 0.8,  $N = M = 64$ , and the initial condition is given by Equation (3.28). 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).

defined in  $[0, 1] \times [0, 1]$ , whose graph is shown in Figure 3.5a.

We 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:

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

where  $L_x = 2\pi$ ,  $L_y = \pi$ ,  $T = 5$ ,  $c = \frac{10}{2\pi}(\frac{L_x}{2\pi})^2$ ,  $\alpha_1 = 2\pi\left(\frac{X}{L_x} - \frac{t}{T}\right)$ ,  $\alpha_2 = \frac{\pi Y}{L_y}$ ,  $\alpha_3 = \frac{\pi t}{T}$ ,  $X = -\pi + 2\pi x$ ,  $Y = -\frac{\pi}{2} + \pi y$ . Y. Chen et al. (2017) uses periodic boundary conditions in the  $x$ -direction and



(a) Initial condition from Equation (3.29).

zero-gradient in the  $y$ -direction. However, we will 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, allowing us to compute the error.

In Figure 3.6, the results obtained using two Gaussian hills and the velocity field from Equation (3.30) are depicted. The PPM-L04 scheme with AVLT, L04, and PL07 splitting, along with RK1 as the departure point scheme, is used. It can be observed how the scalar field deforms and eventually returns to its initial position. Comparing the schemes, it is noticeable that the PL07 and L04 schemes produce almost identical results and are less diffusive than AVLT.

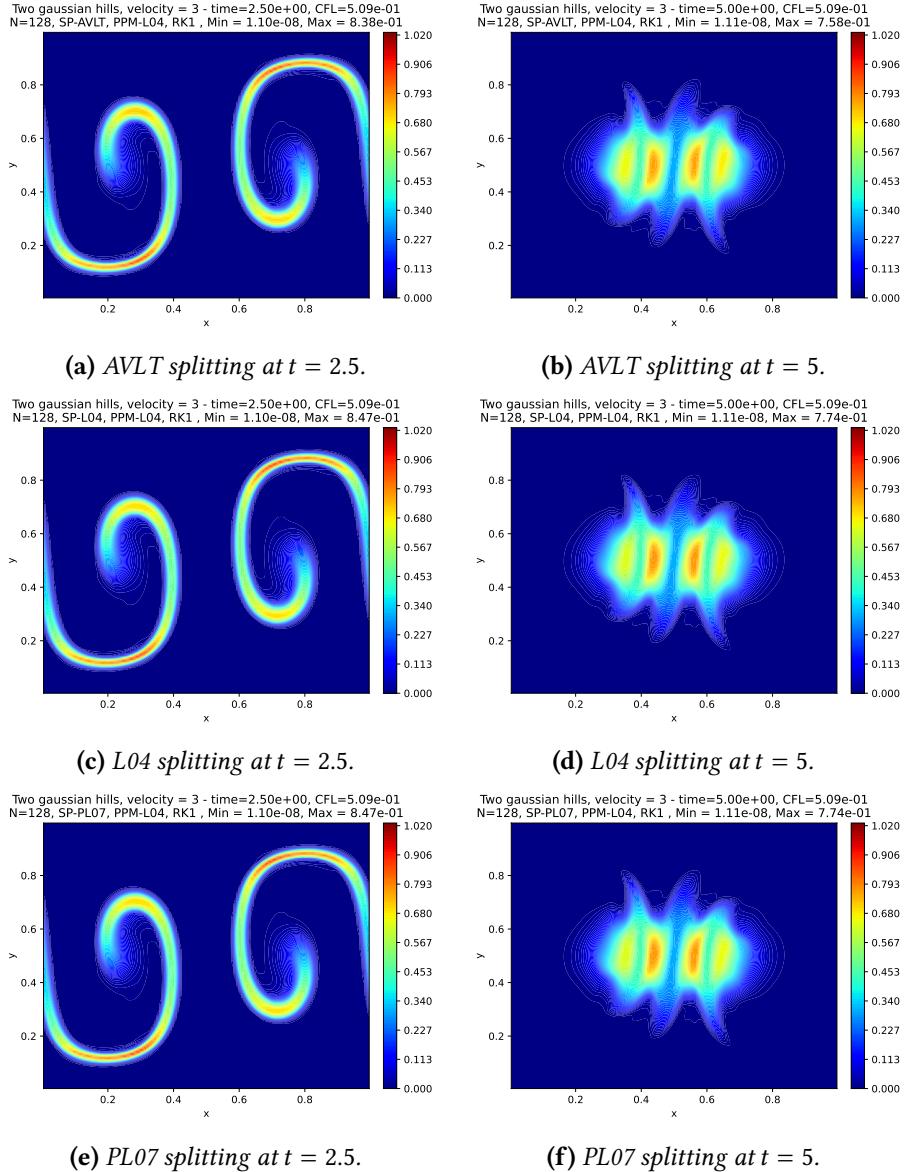
From Figure 3.7a, it can be observed that when using the RK1 schemes, the PL07 and L04 schemes are slightly more accurate than AVLT for the 1D PPM-PL07 schemes. All these schemes exhibit a convergence order greater than two (Figure 3.8). When the 1D scheme PPM-L04 is used, all the splitting methods yield very similar results, regardless of the departure point scheme. However, when the RK2 schemes are used, AVLT is the most accurate and achieves a third-order convergence (Figure 3.7b).

## 3.5 Concluding remarks

In this chapter, we introduced the dimension-splitting method, which replaces the solution of the 2D advection equation with the solution of multiple 1D advection equations, resulting in more cost-effective 2D-FV schemes. For our simulations, we adopted the 1D FV-SL scheme based on PPM to solve the 1D equations.

We modified the average of two Lie-Trotter splittings, which is second-order accurate, to ensure the preservation of a constant scalar field with a divergence-free velocity. This modification addresses the limitation of the classical averaging Lie-Trotter splitting and follows the methodology used in FV3.

Based on the simulation with constant velocity, we concluded that all the splitting schemes are equivalent and do not introduce any splitting errors. In fact, the splittings are exact in this case. We observed that all the conclusions from the 1D simulations hold true in the 2D case as well, with mass conservation and monotonicity being preserved when

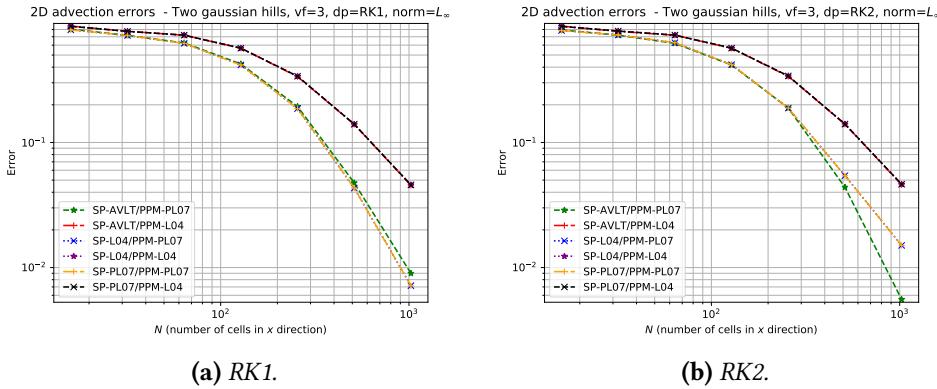


**Figure 3.6:** Linear advection experiment using the velocity from Equation (3.30), a CFL number equal to 0.8,  $N = M = 128$ , and the initial condition is given by Equation (3.29). We use the scheme PPM-L04 with AVL T (3.6a and 3.6b), L04 (3.6c and 3.6d) and PL07 (3.6e and 3.6f) splitting and RK1 departure point scheme. These figures show the advected profile after 2.5 (left) and 5 (right) time units (one time period).

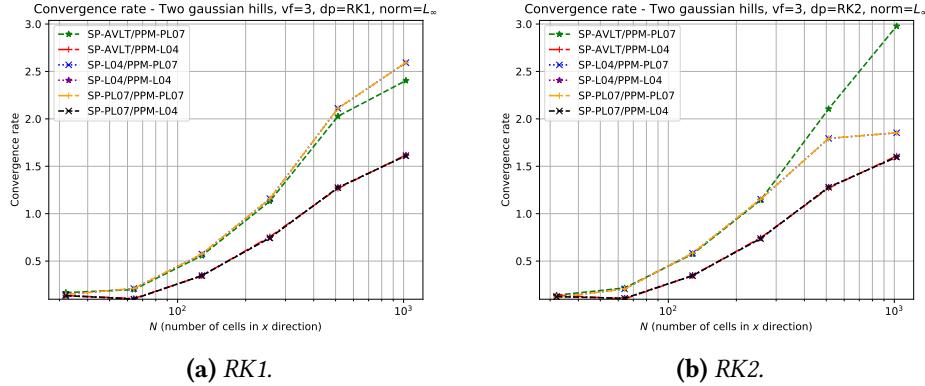
using the monotonic limiters in the 1D subproblems.

In the simulation with variable velocity, we conducted a flow deformation test case. We observed that all splitting schemes preserved monotonicity in all simulations. The PL07 and L04 schemes yielded very similar results and introduced a first-order error, which was observed when using the RK2 scheme to compute the departure point. Surprisingly, the AVL T scheme achieved third-order accuracy, surpassing its expected second-order accuracy. This indicates that a more accurate departure point calculation benefits the AVL T splitting. However, when using a first-order departure point computation, the splitting

## 3.5 | CONCLUDING REMARKS



**Figure 3.7:** Convergence of the error for the schemes PPM-PL07 and PPM-L04 with AVLT/L04/PL07 splitting applied to the linear advection problem using a velocity from Equation (3.30), a CFL number equal to 0.8, a final time of integration equal to 5 time units and the initial condition given by Equation (3.28). The departure points are computed using the RK1 (left) and RK2 (right) schemes.



**Figure 3.8:** Similar to Figure 3.7 but considering the convergence rate.

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 for Planotic grids, we start with a Platonic solid, in this case, a cube, which is circumscribed in a sphere. We then project its faces onto the sphere. The original cubed-sphere, called the equidistant cubed-sphere, was proposed by Sadourny (1972) but resulted in a non-uniform grid. To address this issue, a solution was proposed by introducing angular coordinates, leading to a quasi-uniform grid known as the equiangular cubed-sphere. The cubed sphere consists of six panels, each one having a local Cartesian coordinate system. This makes it easier to extend methods from the plane to the sphere. In fact, Putman and Lin (2007) extends the dimension splitting technique from Lin and Rood (1996), as presented in Chapter 3, to the cubed-sphere.

There are essentially two major challenges when working with the cubed-sphere grid:

1. The non-orthogonal grid system: This challenge is primarily related to the appearance of metric terms in the equations. It adds computational cost and often requires conversions between contravariant and covariant components of a velocity field.
2. The discontinuity of the coordinate system at the cube edges: This is perhaps the most problematic challenge. Computing stencils along the cube edges becomes challenging due to the discontinuous nature of the coordinate system.

One possible approach to compute stencils at the edges is to extend the local coordinate of each panel to its neighboring panels, adding ghost cells in the halo region. In the case of the equiangular cubed-sphere, ghost cell values lie on the same geodesics containing the data from the neighboring panels. This allows for the use of one-dimensional high-order Lagrange interpolation to compute the stencils at the edges. This approach has been extensively used in the literature (X. Chen, 2021; Croisille, 2013; Katta et al., 2015a, 2015b) and was initially introduced by Ronchi et al. (1996). This approach is referred to as duo grid, as named by X. Chen (2021). Alternatively, Putman and Lin (2007) uses extrapolation for grid values near the cube edges. Another approach that avoids the need for interpolation or extrapolation near the edges is the conformal cubed-sphere developed by Rančić et al. (1996). While this grid leads to an orthogonal and continuous coordinate system near the

edges, it generates grid singularities near the cube corners, similar to the pole problem. An improved and more uniform conformal grid, called the Uniform Jacobian cubed sphere, was later proposed by Rancić et al. (2017). Each approach is likely to generate grid imprinting, and one of the goals of this work is to investigate the amount of grid imprinting produced by each method.

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 how we can apply 1D Lagrange interpolation using the adjacent panels data to obtain values of a scalar/vector field on ghost cells, as well different edge treatments when computing stencils near to the cube edges. Final thoughts are presented in Section 4.3.

## 4.1 Cubed-sphere mappings

### 4.1.1 Equidistant cubed-sphere

We start this chapter by introducing the equidistant cubed-sphere proposed by Sadourny (1972). Given  $R > 0$ , we denote the sphere of radius  $R$  centered at the origin of  $\mathbb{R}^3$  as:

$$\mathbb{S}_R^2 = \{P = (X, Y, Z) \in \mathbb{R}^3 : X^2 + Y^2 + Z^2 = R^2\}.$$

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

$$\begin{aligned}\Psi_1(x, y) &= \frac{R}{\sqrt{a^2 + x^2 + y^2}}(a, x, y), \\ \Psi_2(x, y) &= \frac{R}{\sqrt{a^2 + x^2 + y^2}}(-x, a, y), \\ \Psi_3(x, y) &= \frac{R}{\sqrt{a^2 + x^2 + y^2}}(-a, -x, y), \\ \Psi_4(x, y) &= \frac{R}{\sqrt{a^2 + x^2 + y^2}}(x, -a, y), \\ \Psi_5(x, y) &= \frac{R}{\sqrt{a^2 + x^2 + y^2}}(-y, x, a), \\ \Psi_6(x, y) &= \frac{R}{\sqrt{a^2 + x^2 + y^2}}(y, x, -a).\end{aligned}$$

The set of 6 maps  $\{\Psi_p, p = 1, \dots, 6\}$  allow us to cover the sphere. Here  $p$  denotes a panel, and they are defined and orientated as Figure 4.1 shows. Then, we can represent a point on the sphere using the cubed-sphere coordinates  $(x, y, p)$ .

The derivative of the maps  $\Psi_p$  are given by:

$$\begin{aligned} D\Psi_1(x, y) &= \frac{R}{(a^2 + x^2 + y^2)^{3/2}} \begin{bmatrix} -ax & -ay \\ a^2 + y^2 & -xy \\ -xy & a^2 + x^2 \end{bmatrix}, \\ D\Psi_2(x, y) &= \frac{R}{(a^2 + x^2 + y^2)^{3/2}} \begin{bmatrix} -(a^2 + y^2) & xy \\ -ax & -ay \\ -xy & a^2 + x^2 \end{bmatrix}, \\ D\Psi_3(x, y) &= \frac{R}{(a^2 + x^2 + y^2)^{3/2}} \begin{bmatrix} ax & ay \\ -(a^2 + y^2) & xy \\ -xy & a^2 + x^2 \end{bmatrix}, \\ D\Psi_4(x, y) &= \frac{R}{(a^2 + x^2 + y^2)^{3/2}} \begin{bmatrix} a^2 + y^2 & -xy \\ ax & ay \\ -xy & a^2 + x^2 \end{bmatrix}, \\ D\Psi_5(x, y) &= \frac{R}{(a^2 + x^2 + y^2)^{3/2}} \begin{bmatrix} xy & -(a^2 + x^2) \\ a^2 + y^2 & -xy \\ -ax & -ay \end{bmatrix}, \\ D\Psi_6(x, y) &= \frac{R}{(a^2 + x^2 + y^2)^{3/2}} \begin{bmatrix} -xy & a^2 + x^2 \\ a^2 + y^2 & -xy \\ ax & ay \end{bmatrix}. \end{aligned}$$

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

$$\mathbf{g}_x(x, y, p) = D\Psi_p(x, y) \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{g}_y(x, y, p) = D\Psi_p(x, y) \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Notice that

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

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:

$$G_\Psi(x, y) = \begin{bmatrix} \langle \mathbf{g}_x(x, y, p), \mathbf{g}_x(x, y, p) \rangle & \langle \mathbf{g}_x(x, y, p), \mathbf{g}_y(x, y, p) \rangle \\ \langle \mathbf{g}_x(x, y, p), \mathbf{g}_y(x, y, p) \rangle & \langle \mathbf{g}_y(x, y, p), \mathbf{g}_y(x, y, p) \rangle \end{bmatrix},$$

where  $\langle \cdot, \cdot \rangle$  denotes the standard inner product of  $\mathbb{R}^3$ , 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:

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



**Figure 4.1:** Cubed-sphere panels definition and orientation. Figure taken from Jung et al. (2019).

### 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  $\Phi_p : [-\frac{\pi}{4}, \frac{\pi}{4}] \times [-\frac{\pi}{4}, \frac{\pi}{4}] \rightarrow \mathbb{S}_R^2$ ,  $p = 1, \dots, 6$ , 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:

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

and therefore we can define the following tangent vectors

$$\mathbf{r}_x(x, y, p) = D\Phi_p(x, y) \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{a}{\cos^2 x} \mathbf{g}_x(\tan x, \tan y, p), \quad (4.1)$$

$$\mathbf{r}_y(x, y, p) = D\Phi_p(x, y) \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{a}{\cos^2 y} \mathbf{g}_y(\tan x, \tan y, p), \quad (4.2)$$

Again, it makes sense to define the matrix

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

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

$$G_\Phi(x, y) = \begin{bmatrix} \langle \mathbf{r}_x(x, y, p), \mathbf{r}_x(x, y, p) \rangle & \langle \mathbf{r}_x(x, y, p), \mathbf{r}_y(x, y, p) \rangle \\ \langle \mathbf{r}_x(x, y, p), \mathbf{r}_y(x, y, p) \rangle & \langle \mathbf{r}_y(x, y, p), \mathbf{r}_y(x, y, p) \rangle \end{bmatrix}, \quad (4.3)$$

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:

$$\begin{aligned} \sqrt{|\det G_\Phi(x, y)|} &= \frac{a}{\cos^2 x \cos^2 y} \frac{R^2}{(a^2 + a^2 \tan^2 x + a^2 \tan^2 y)^{3/2}} a \\ &= \frac{R^2}{\cos^2 x \cos^2 y} \frac{1}{(1 + \tan^2 x + \tan^2 y)^{3/2}}. \end{aligned} \quad (4.4)$$

### 4.1.3 Tangent vectors on the sphere

The tangent space at  $P \in \mathbb{S}_R^2$  is denoted by  $T_P \mathbb{S}^2$ . It is easy to see that:

$$T_P \mathbb{S}_R^2 = \{P_0 \in \mathbb{R}^3 : \langle P, P_0 \rangle = 0\}.$$

We are going to consider three ways to represent an element of  $\mathbb{S}_R^2$ : using  $(X, Y, Z)$  coordinates, or using  $(\lambda, \phi)$  latitude-longitude coordinates, or, at last, using the cubed-sphere coordinates  $(x, y, p)$ , where  $(x, y)$  are the cube face coordinates and  $p \in \{1, 2, \dots, 6\}$  stands for a cube panel. We say that a vector field  $\mathbf{u} : \mathbb{S}_R^2 \rightarrow \mathbb{R}^3$  is tangent on the sphere if  $\mathbf{u}(P) \in T_P \mathbb{S}_R^2, \forall P \in \mathbb{S}_R^2$ .

### Conversions between latitude-longitude and contravariant coordinates

We consider the latitude-longitude mapping  $\Psi_{ll} : [0, 2\pi] \times [-\frac{\pi}{2}, \frac{\pi}{2}] \rightarrow \mathbb{S}_R^2$ , given by:

$$X(\lambda, \phi) = R \cos \phi \cos \lambda, \quad (4.5)$$

$$Y(\lambda, \phi) = R \cos \phi \sin \lambda, \quad (4.6)$$

$$Z(\lambda, \phi) = R \sin \phi. \quad (4.7)$$

The derivative or Jacobian matrix of the mapping  $\Psi_{ll}$  is given by:

$$D\Psi_{ll}(\lambda, \phi) = R \begin{bmatrix} -\cos \phi \sin \lambda & -\sin \phi \cos \lambda \\ \cos \phi \cos \lambda & \sin \phi \sin \lambda \\ 0 & \cos \phi \end{bmatrix}. \quad (4.8)$$

Using this matrix columns, we can define the tangent vectors:

$$\mathbf{r}_\lambda(\lambda, \phi) = D\Psi_{ll}(\lambda, \phi) \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{r}_\phi(\lambda, \phi) = D\Psi_{ll}(\lambda, \phi) \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \quad (4.9)$$

We normalize the vectors  $\mathbf{r}_\lambda$  and  $\mathbf{r}_\phi$  and we obtain unit tangent vectors on the sphere at  $\Phi_{ll}(\lambda, \phi)$ :

$$\mathbf{e}_\lambda(\lambda, \phi) = \begin{bmatrix} -\sin \lambda \\ \cos \lambda \\ 0 \end{bmatrix}, \quad \mathbf{e}_\phi(\lambda, \phi) = \begin{bmatrix} -\sin \phi \cos \lambda \\ -\sin \phi \sin \lambda \\ \cos \phi \end{bmatrix}. \quad (4.10)$$

Let us consider a tangent vector field  $\mathbf{u} : \mathbb{S}_R^2 \rightarrow \mathbb{R}^3$  on the sphere, represented as

$$\mathbf{u}(\lambda, \phi) = u_\lambda(\lambda, \phi)\mathbf{e}_\lambda(\lambda, \phi) + v_\phi(\lambda, \phi)\mathbf{e}_\phi(\lambda, \phi). \quad (4.11)$$

Or, we may also represent this vector field using the basis obtained by cubed-sphere coordinates:

$$\mathbf{u}(x, y, p) = u(x, y, p)\mathbf{r}_x(x, y, p) + v(x, y, p)\mathbf{r}_y(x, y, p). \quad (4.12)$$

This representation is known as contravariant representation. In order to relate the latitude-longitude representation with the contravariant representation, we notice that:

$$\mathbf{r}_x(x, y, p) = \langle \mathbf{r}_x, \mathbf{e}_\lambda \rangle \mathbf{e}_\lambda(\lambda, \phi) + \langle \mathbf{r}_x, \mathbf{e}_\phi \rangle \mathbf{e}_\phi(\lambda, \phi), \quad (4.13)$$

$$\mathbf{r}_y(x, y, p) = \langle \mathbf{r}_y, \mathbf{e}_\lambda \rangle \mathbf{e}_\lambda(\lambda, \phi) + \langle \mathbf{r}_y, \mathbf{e}_\phi \rangle \mathbf{e}_\phi(\lambda, \phi), \quad (4.14)$$

which holds since the vectors  $\mathbf{e}_\lambda(\lambda, \phi)$  and  $\mathbf{e}_\phi(\lambda, \phi)$  are orthogonal. Replacing Equations (4.13) and (4.14) in Equation (4.12), we obtain the values  $(u_\lambda, v_\phi)$  in terms of the contravariant components  $(u, v)$  as the following matrix equation:

$$\begin{bmatrix} u_\lambda(\lambda, \phi) \\ v_\phi(\lambda, \phi) \end{bmatrix} = \begin{bmatrix} \langle \mathbf{r}_x, \mathbf{e}_\lambda \rangle & \langle \mathbf{r}_y, \mathbf{e}_\lambda \rangle \\ \langle \mathbf{r}_x, \mathbf{e}_\phi \rangle & \langle \mathbf{r}_y, \mathbf{e}_\phi \rangle \end{bmatrix} \begin{bmatrix} u(x, y, p) \\ v(x, y, p) \end{bmatrix}. \quad (4.15)$$

Conversely, we may express the contravariant components in terms of latitude-longitude components by inverting Equation (4.15):

$$\begin{bmatrix} u(x, y, p) \\ v(x, y, p) \end{bmatrix} = \frac{1}{\langle \mathbf{r}_x, \mathbf{e}_\lambda \rangle \langle \mathbf{r}_y, \mathbf{e}_\phi \rangle - \langle \mathbf{r}_y, \mathbf{e}_\lambda \rangle \langle \mathbf{r}_x, \mathbf{e}_\phi \rangle} \begin{bmatrix} \langle \mathbf{r}_y, \mathbf{e}_\phi \rangle & -\langle \mathbf{r}_y, \mathbf{e}_\lambda \rangle \\ -\langle \mathbf{r}_x, \mathbf{e}_\phi \rangle & \langle \mathbf{r}_x, \mathbf{e}_\lambda \rangle \end{bmatrix} \begin{bmatrix} u_\lambda(\lambda, \phi) \\ v_\phi(\lambda, \phi) \end{bmatrix}. \quad (4.16)$$

## Covariant/contravariant conversion

Let us consider again a tangent vector field  $\mathbf{u} : \mathbb{S}_R^2 \rightarrow \mathbb{R}^3$  on the sphere. Its contravariant representation is given by Equation (4.12). The covariant components  $(U, V)$  are given by:

$$U(x, y, p) = \langle \mathbf{u}(x, y, p), \mathbf{r}_x(x, y, p) \rangle, \quad (4.17)$$

$$V(x, y, p) = \langle \mathbf{u}(x, y, p), \mathbf{r}_y(x, y, p) \rangle. \quad (4.18)$$

Replacing Equation (4.12) in Equations (4.17) and (4.18) we obtain the relation covariant components in terms of the contravariant terms:

$$\begin{bmatrix} U(x, y, p) \\ V(x, y, p) \end{bmatrix} = \begin{bmatrix} \langle \mathbf{r}_x, \mathbf{r}_x \rangle & \langle \mathbf{r}_x, \mathbf{r}_y \rangle \\ \langle \mathbf{r}_x, \mathbf{r}_y \rangle & \langle \mathbf{r}_y, \mathbf{r}_y \rangle \end{bmatrix} \begin{bmatrix} u(x, y, p) \\ v(x, y, p) \end{bmatrix}. \quad (4.19)$$

We may express the contravariant components in terms of the covariant terms inverting Equation (4.19):

$$\begin{bmatrix} u(x, y, p) \\ v(x, y, p) \end{bmatrix} = \frac{1}{\langle \mathbf{r}_x, \mathbf{r}_x \rangle \langle \mathbf{r}_y, \mathbf{r}_y \rangle - \langle \mathbf{r}_x, \mathbf{r}_y \rangle^2} \begin{bmatrix} \langle \mathbf{r}_y, \mathbf{r}_y \rangle & -\langle \mathbf{r}_x, \mathbf{r}_y \rangle \\ -\langle \mathbf{r}_x, \mathbf{r}_y \rangle & \langle \mathbf{r}_x, \mathbf{r}_x \rangle \end{bmatrix} \begin{bmatrix} U(x, y, p) \\ V(x, y, p) \end{bmatrix}. \quad (4.20)$$

Notice that combining Equations (4.19) and (4.20) with Equations (4.15) and (4.16) one may get relations between the latitude-longitude components and the covariant components.

#### 4.1.4 Notation

Let us denote by  $\Sigma_p : \Omega \rightarrow \mathbb{S}_R^2$ ,  $p = 1, \dots, 6$ ,  $\Omega = [-a, a]^2$ , a cubed-sphere mapping like the equiangular or the equidistant mappings introduced in Section 4.1. With these mappings, we denote by  $\mathbf{r}_x(x, y, p), \mathbf{r}_y(x, y, p)$  the basis of tangent vectors at  $(x, y, p)$ . The metric tensor is denoted by  $\sigma(x, y) = \sqrt{|\det G_\Sigma(x, y, p)|}$ , where

$$G_\Sigma(x, y, p) = \begin{bmatrix} \langle \mathbf{r}_x(x, y, p), \mathbf{r}_x(x, y, p) \rangle & \langle \mathbf{r}_x(x, y, p), \mathbf{r}_y(x, y, p) \rangle \\ \langle \mathbf{r}_x(x, y, p), \mathbf{r}_y(x, y, p) \rangle & \langle \mathbf{r}_y(x, y, p), \mathbf{r}_y(x, y, p) \rangle \end{bmatrix}. \quad (4.21)$$

We will utilize the notation introduced in Section 3.1.1 throughout this chapter. The parameter  $v$  represents a non-negative integer indicating the number of ghost cell layers in each panel boundary, called halo size.

To introduce the cubed-sphere, we consider a  $(\Delta x, \Delta y)$ -grid denoted by  $\Omega_{\Delta x, \Delta y} = (\Omega_{ij})_{i,j=-v+1, \dots, N+v}$ , where  $\Delta x = \Delta y$ , and it covers the domain  $\Omega$ . A control volume of the cubed-sphere is denoted by  $\Omega_{ijp}$ , defined as follows:

$$\Omega_{ijp} = \Sigma_p(\Omega_{ij}) \quad -v + 1 \leq i, j \leq N + v, \quad 1 \leq p \leq 6.$$

The cubed-sphere grid refers to the collection of control volumes  $(\Omega_{ijp})_{i,j=-v+1, \dots, N+v}^{p=1, \dots, 6}$ . In Figure 4.2, an example of the cubed-sphere grid is depicted, excluding the ghost cells. This grid is generated using the equidistant and equiangular mappings for  $N = 20$ .

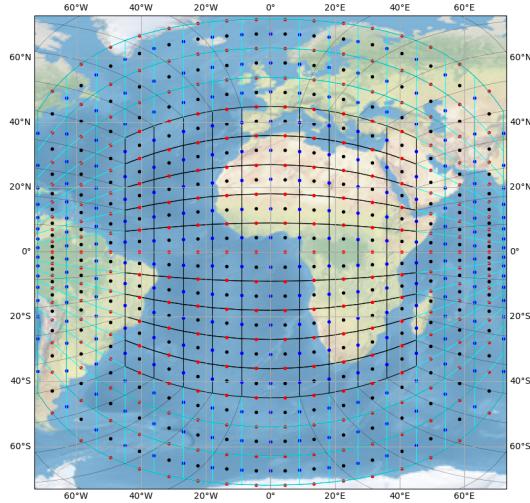
From Figure 4.2, it is evident that the equiangular cubed-sphere exhibits a higher degree of uniformity compared to the equidistant cubed-sphere. As noted by Rančić et al. (1996), the ratio between the maximum and minimum cell areas on the equiangular cubed-sphere is approximately 1.3, whereas the same ratio is approximately 5.2 on the equidistant cubed-sphere.

We also utilize the notation  $\mathcal{CS}_N = \mathbb{R}^{(N+v) \times (N+v) \times 6}$  to represent grid functions on the cubed-sphere at cell centers. Let's assume we have a function  $q : \mathbb{S}_R^2 \times [0, T] \rightarrow \mathbb{R}$ , and we have a  $(\Delta x, \Delta y, \Delta t, \lambda)$ -discretization of  $\Omega \times [0, T]$ . We introduce  $q^n \in \mathcal{CS}^N$ , which represents the grid function  $q$  evaluated at the discrete points. In other words,  $q_{ijp}^n = q(x_i, y_j, p, t^n)$ , where  $i, j = -v + 1, \dots, N + v$ , and  $p = 1, \dots, 6$ . Furthermore, we use the notations  $q_{i+\frac{1}{2}, j, p}^n = q(x_{i+\frac{1}{2}}, y_j, t^n)$  for  $i = -v, \dots, N + v$  and  $j = -v + 1, \dots, N + v$  to represent  $q$  at the midpoint of edges in the  $x$  direction. Similarly, we use  $q_{i, j+\frac{1}{2}, p}^n = q(x_i, y_{j+\frac{1}{2}}, t^n)$  for  $i = -v + 1, \dots, N + v$  and  $j = -v, \dots, N + v$  to represent  $q$  at the midpoint of edges in the  $y$  direction. When  $q$



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

does not depend on the time variable  $t$ , we can omit the index  $n$ . In Figure 4.3, we depict a grid function at centers, edge midpoints in the  $x$  direction and edge midpoints in the  $y$  direction for the equiangular cubed-sphere considering the halo size equal to three. For a



**Figure 4.3:** Grid function at centers (black dots), edge midpoints in the  $x$  direction (blue dots) and edge midpoints in the  $y$  direction (red dots) at panel 1 for the equiangular cubed-sphere using three layers of ghost cells and  $N = 10$ . Ghost cells of panel 1 are drawn using cyan colors.

grid function  $Q$  we also use the notations:

$$\begin{aligned} Q_{\times,j,p} &:= (Q_{-v+1,j,p}, \dots, Q_{N+v,j,p}) \in \mathbb{R}_v^N, \\ Q_{i,\times,p} &:= (Q_{i,-v+1,p}, \dots, Q_{i,M+v,p}) \in \mathbb{R}_v^N. \end{aligned}$$

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

$$|\Omega_{ijp}| = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \sigma(x, y) dx dy. \quad (4.22)$$

We define the average values of a function  $q$  with the aid of the metric tensor  $\sigma(x, y)$  at time  $t$ :

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

In this context, we define  $Q(t) \in \mathcal{CS}_N$  given by  $Q(t) = (Q_{ijp}(t))_{i,j=-v+1,\dots,N+v}^{p=1,\dots,6}$ . Similar to Proposition 3.1, we may approximate the average value using the centroid value, that is

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

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.

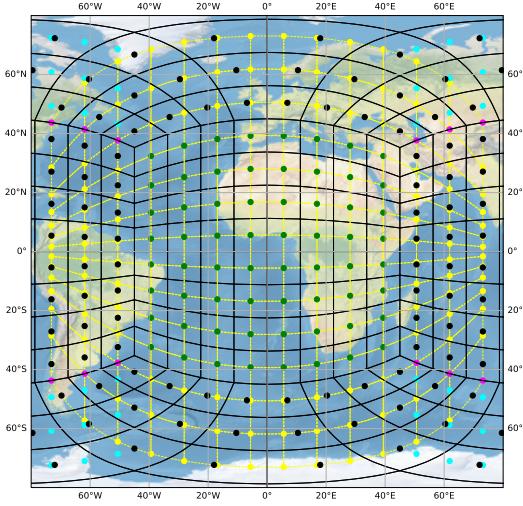
## 4.2 Edges treatment

### 4.2.1 Ghost cells scalar field interpolation

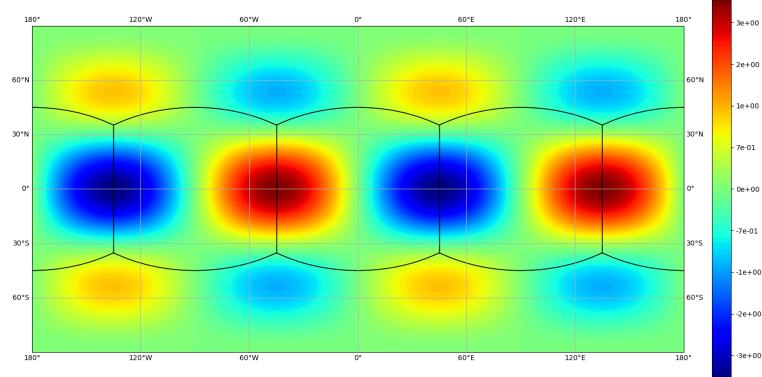
Let's consider a function  $q : \mathbb{S}_R^2 \rightarrow \mathbb{R}$  given at the cell centroids, denoted by  $q_{ijp}$ , where  $i, j = 1 \dots, N$  and  $p = 1, \dots, 6$ . Our objective is to estimate these values at positions outside the range  $1, \dots, N$ , specifically at ghost cell positions.

To solve this problem, we will employ the strategy outlined in Zerroukat and Allen (2022), named duo grid by X. Chen (2021). As previously mentioned, the ghost cells in the local Cartesian systems are mapped onto the geodesics of adjacent panels, which enables us to use Lagrange interpolation to obtain the values of ghost cells.

To illustrate this process in Panel 1, we depict the values of  $q_{ijp}$  in Figure 4.4. The green circles represent the values in Panel 1, while the black circles represent the values in the other panels, for a given  $N = 8$ . Assuming a halo size of 3, we also indicate the target values at the ghost cell positions using yellow and magenta circles. It is worth noting that the dashed yellow lines in Figure 4.4 illustrate how the ghost cell points lie on geodesics containing grid positions from adjacent panels. With the exception of the magenta circles, all the ghost cell values can be obtained using 1D Lagrange interpolation, utilizing the surrounding black circles on the geodesic. This interpolation procedure can be performed for all panels. Subsequently, the magenta circles can be interpolated using the values obtained in the first step of interpolation (depicted as cyan circles in Figure 4.4), while preserving the order of accuracy of the interpolation, assuming it is fixed.



**Figure 4.4:** 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).



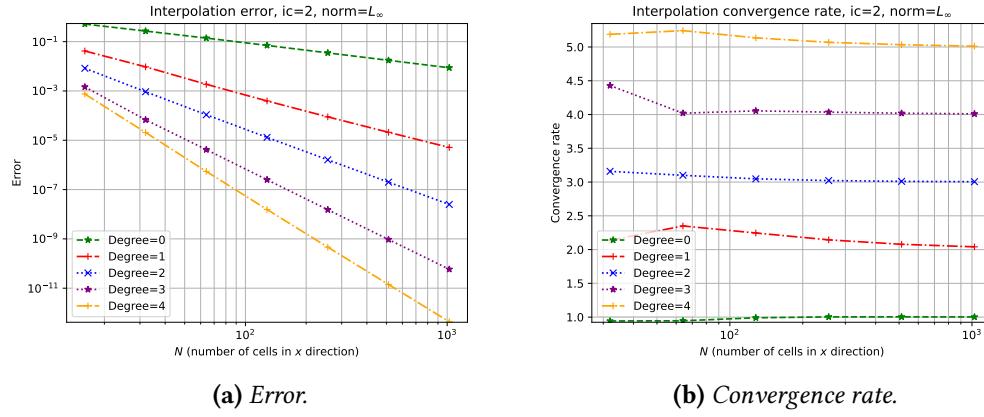
**Figure 4.5:** Trigonometric function defined by Equation (4.23).

We are going to show a numerical example of this interpolation process using a halo region of size 4 and assuming the radius of the sphere is equal to one. We shall consider the following trigonometric function, which is the divergence of a velocity field, as in Peixoto and Barros (2013) in our tests:

$$q(\lambda, \phi) = \frac{1}{\cos(\phi)} \left( -2 \cos^3(\phi) \sin(\lambda) \cos(\lambda) + 16 \sin^2(\lambda) \cos(\lambda) \cos^3(\phi) \sin(\phi) \right), \quad (4.23)$$

whose graph is depicted in Figure 4.5.

We are going to consider the relative error in the maximum norm and the convergence



**Figure 4.6:** Relative error convergence (a) and convergence rate (b) for the ghost cell interpolation process for different polynomial degrees, using the trigonometric function given by Equation (4.23).

rate at the ghost cell positions defined analogously as in Section 3.4. We also consider values os  $N$  given by  $2^k$ , for  $k = 4, \dots, 10$ , in order to compute the error and convergence rate. In Figures 4.6 we show the error and convergence rate for the trigonometric function (Equation (4.23)), respectively, considering polynomials of degrees from 0 up to 4. As both graphs show, we were able to achieve the expected order of convergence.

## 4.2.2 Ghost cells wind interpolation

Let us consider the following problem: Assume that we are given a tangent vector field of the sphere, denoted as  $\mathbf{u} : \mathbb{S}_R^2 \rightarrow \mathbb{R}^3$ . We also have its contravariant components at the control-volume midpoints, namely  $u_{i+\frac{1}{2},j,p}$  for  $i = 0, \dots, N$  and  $j = 1, \dots, N$ , as well as  $v_{i,j+\frac{1}{2},p}$  for  $i = 1, \dots, N$  and  $j = 0, \dots, N$ .

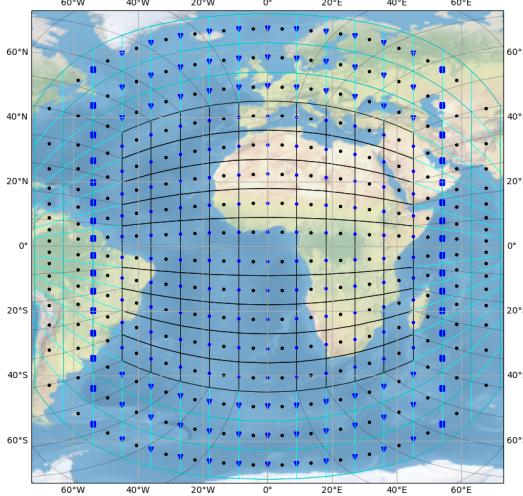
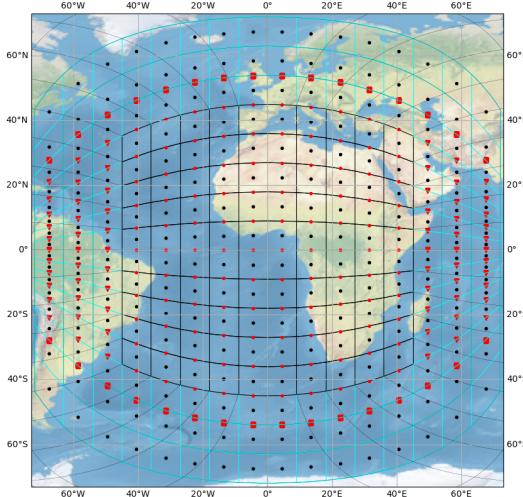
Our objective is to obtain the values

$$\begin{aligned} u_{i+\frac{1}{2},j,p} &\quad \text{for } i = -1, \dots, N+1, & j = -\nu + 1, \dots, 0, \quad j = N, \dots, N+\nu, \\ v_{i,j+\frac{1}{2},p} &\quad \text{for } j = 0, \dots, N, & i = -\nu + 1, \dots, 0, \quad i = N, \dots, N+\nu, \end{aligned}$$

as shown in Figure 4.7. This problem arises when we apply the dimension splitting method on each panel of the cubed-sphere.

This problem can be solved by using the interpolation process described earlier for a scalar field. To apply the interpolation process, we first need to interpolate the values of  $u$  and  $v$  from the edges to the centers at the points required for the ghost cells interpolation. Specifically, we need the values:

$$\begin{aligned} u_{1+k,j,p}, v_{1+k,j,p} &\quad \text{for } j = 1, \dots, N, \quad k = 0, \dots, \nu - 1, \\ u_{N-k,j,p}, v_{N-k,j,p} &\quad \text{for } j = 1, \dots, N, \quad k = 0, \dots, \nu - 1, \\ u_{i,1+k,p}, v_{i,1+k,p} &\quad \text{for } i = 1, \dots, N, \quad k = 0, \dots, \nu - 1, \\ u_{i,N-k,p}, v_{i,N-k,p} &\quad \text{for } i = 1, \dots, N, \quad k = 0, \dots, \nu - 1. \end{aligned}$$

(a) Contravariant component  $u$ .(b) Contravariant component  $v$ .

**Figure 4.7:** Illustration of contravariant components  $u$  (a) and  $v$  (b) of a vector field at edge midpoints. Ghost cell points are denoted by triangles and squares. The extra points needed for the RK2 scheme are denoted by squares.

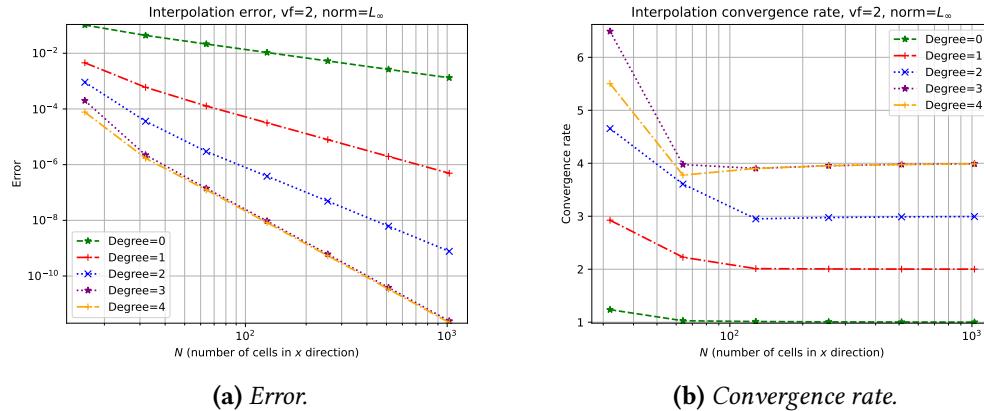
We apply a third-order interpolation in the  $x$  direction of each panel to interpolate  $u$  from the edges to the centers (see Figure 4.7a). Similarly, we use a third-order interpolation to obtain the values of  $v$  at the centers in the  $y$  direction (see Figure 4.7b). Once these interpolated values are computed, we convert the contravariant values  $u_{ijp}, v_{ijp}$  to their latitude-longitude components  $(u_\lambda)_{ijp}, (v_\phi)_{ijp}$  using Equation (4.16). This conversion avoid any coordinate system discontinuity. Then, we can use the ghost cell centers interpolation procedure described before for the latitude-longitude components to recover the wind at

the ghost cell centers using any polynomial degree. Finally, we can use the values at the ghost cell centers to obtain the values at the ghost cell edges by employing a third-order interpolation once again. Subsequently, the contravariant components can be obtained by using Equation (4.15).

We will consider the following rotated zonal field on the unit sphere, as a numerical test, based on D. Williamson et al. (1992):

$$\begin{cases} u_\lambda(\lambda, \phi, t) = u_0(\cos(\phi) \cos(\alpha) + \sin(\phi) \cos(\lambda) \sin(\alpha)), \\ v_\phi(\lambda, \phi, t) = -u_0 \sin(\lambda) \sin(\alpha). \end{cases} \quad (4.24)$$

Here,  $u_0 = \frac{2\pi}{5}$  and  $\alpha = -\frac{45\pi}{180}$ . We will adopt the same grids as in Section 4.2.1. Next, we will compute the relative errors of the contravariant components at the midpoint edges, as depicted in Figure 4.7a and Figure 4.7b. The errors are presented in Figure 4.8, along with the convergence rate for different degrees employed in the ghost cell center interpolation. We emphasize that, since we utilize a third-order interpolation to retrieve the wind at centers from the edges, as well as in the interpolation from the ghost cell centers to ghost cell edges, the maximum attainable scheme order is 4. Indeed, from Figure 4.8, we observe that when employing a fourth-order interpolation in the ghost cell interpolation step, the final order achieved is 4.



**Figure 4.8:** Relative error convergence (a) and convergence rate (b) for the ghost cell interpolation of a vector field process for different polynomial degrees, using the vectorial function given by Equation (4.24).

### 4.2.3 Edges reconstruction

Let us consider the following problem: given the values  $q_{ijp}$  we wish to find approximations of the function  $q$  at the control volume edge midpoints denoted by

$$q_{ijp}^{L,x} \approx q_{i-\frac{1}{2},j,p}, \quad q_{ijp}^{R,x} \approx q_{i+\frac{1}{2},j,p}, \quad q_{ijp}^{L,y} \approx q_{i,j-\frac{1}{2},p}, \quad q_{ijp}^{R,y} \approx q_{i,j+\frac{1}{2},p}.$$

These points are illustrated in Figure 4.9 for panel 1.

We can estimate the desired values by using the one-dimensional reconstruction



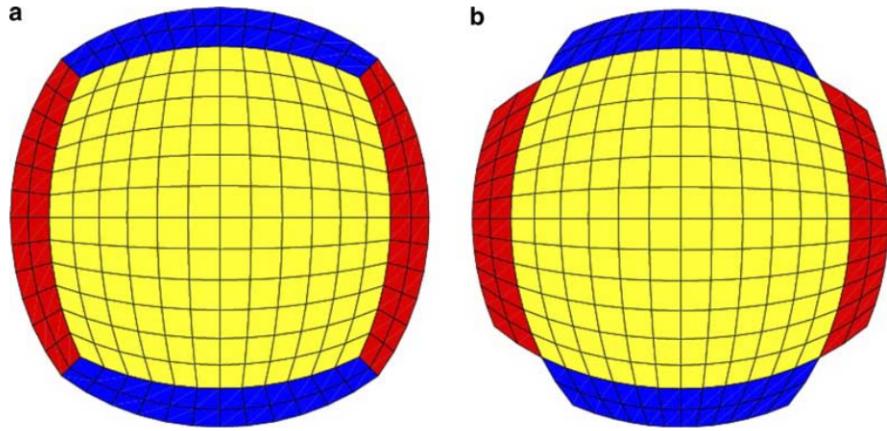
**Figure 4.9:** The reconstruction problem on the cubed-sphere panel 1: we are given the centroid values of a function (black circles) and we wish to estimate these values at the edges midpoints in the  $x$  direction (blue circles) and  $y$  direction (red circles). This figure uses an equiangular cubed-sphere with  $N = 8$ .

schemes described in Sections 2.4 and 2.4.2, performing PPM reconstruction independently in the  $x$  and  $y$  directions. It is worth noting that all the schemes discussed in those sections are expected to be second-order accurate due to the centroid point approximation.

However, there are differences in the computation of the stencil near the cube edges. Unlike in the previous chapters, where periodic boundary conditions were assumed, the boundary conditions in this context are related to the adjacent panels. One way to address this issue is by adding ghost cell layers and utilizing the process described in Section 4.2.1 to compute the stencils. This approach is referred to as **ET-DG** (Edge Treatment - duo grid), as it extends the grid lines and employs Lagrange interpolation, as described by Zerroukat and Allen (2022).

Another approach, employed in Sadourny (1972), involves ignoring the discontinuity of the coordinate system and simply using the values of the cells in the adjacent panels as the ghost cell values. This scheme is labeled **ET-S72** (Edge Treatment - S72).

Additionally, an alternative method that avoids the use of ghost cells was developed by Putman and Lin (2007), which entails extrapolation at the cells surrounding the cube edge. We will refer to this scheme as **ET-PL07** (Edge Treatment - PL07). This scheme uses



**Figure 4.10:** Different approaches for defining the halo region: (a) using adjacent cell values (used in ET-S72 and ET-PL07); (b) extending the panel grid lines (used in ET-DG). Figure taken from Rossmanith (2006).

the following extrapolations:

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

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

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

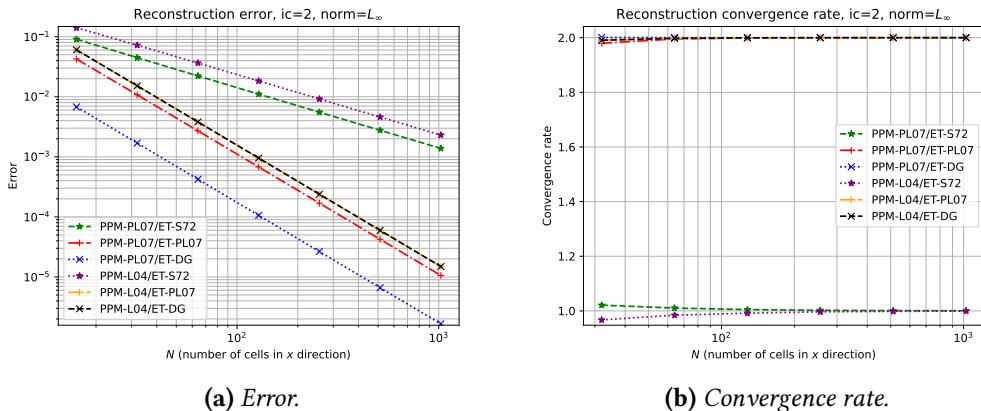
in the  $x$  direction and in the  $y$  direction we use the formulas

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

We are going to use the trigonometric function (Equation (4.23)) as before on the unit sphere to compare the schemes ET-DG, ET-S72 and ET-PL07. The scheme ET-DG uses cubic polynomials. We introduce the relative errors:

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

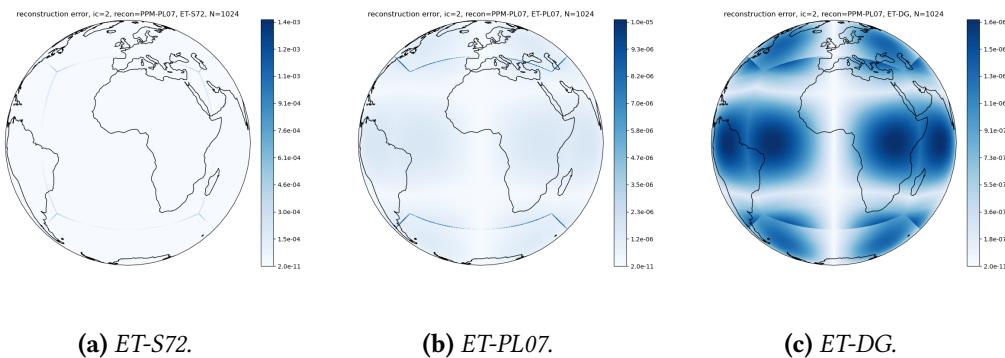
We are going to compute  $E$  for different values of  $N$  as in the numerical experiments of Section 4.2.1.



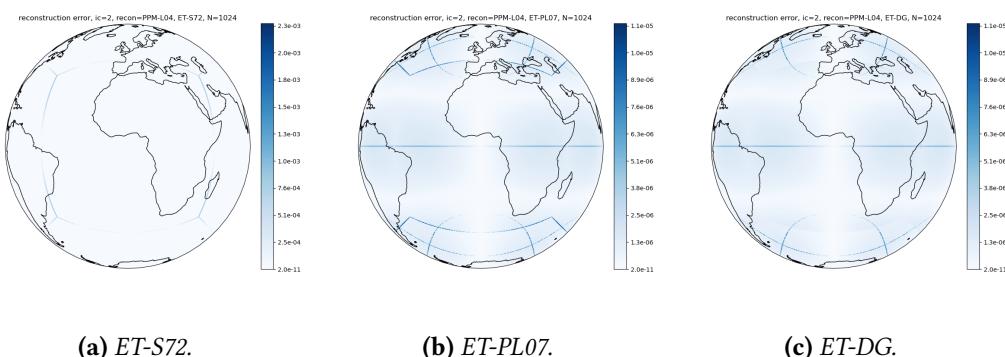
**Figure 4.11:** Relative error convergence (a) and convergence rate (b) for the reconstruction problem for different edge treatments (ET), using the trigonometric function given by Equation (4.23).

In Figure 4.11 we show the errors and the convergence rate using the PPM-PL07 and PPM-L04 reconstruction schemes. We can observe that all the schemes converge to zero with second-order. The difference between the edge treatments ET-S72, ET-PL07 and ET-DG schemes may be observed in Figure 4.12 and Figure 4.13, 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-S72 and ET-PL07 schemes. This is an example of grid imprinting. Although all schemes are second-order, the scheme ET-DG do not seem to produce grid imprinting in the reconstruction problem.



**Figure 4.12:** Error for the edge midpoint values considering the trigonometric function (Equation (4.23)) with edges treatment schemes ET-S72 (a), ET-PL07 (b) and ET-DG for the cubed-sphere with  $N = 1024$  using the reconstruction PPM-PL07.



**Figure 4.13:** As Figure in 4.12 but using the PPM-L04 scheme.

### 4.3 Concluding remarks

In this chapter, we reviewed cubed-sphere mappings with a special focus on the equiangular mapping, which leads to a more uniform cubed-sphere grid. We observed that the equiangular cubed-sphere ghost cells, obtained by extending the gridlines, have a nice property: their edge and center ghost points are located on a common geodesic that contains the edge and center points of the adjacent panels. This property allows us to use 1D Lagrange interpolation to obtain the values of scalar and vector fields in the ghost cells. In fact, we demonstrated the accuracy of this method through numerical examples in Sections 4.2.1 and 4.2.2.

Afterward, in Section 4.2.3, we investigated different methods for computing stencils near the cube edges. We considered the scheme based on 1D Lagrange interpolation

(ET-DG), a scheme that ignores the coordinate system (ET-S72), and a scheme based on extrapolations (ET-PL07). The scheme ET-PL07 was proposed by Putman and Lin (2007) and is used in the FV3 core. Through numerical examples, we demonstrated that the reconstruction at cell edges using PPM schemes generates grid imprinting near the cubed edges when using ET-S72. This grid imprinting is reduced when we apply the extrapolation scheme ET-PL07, but the error still persists to be larger on the cube edges. However, this problem is eliminated when we adopt the scheme ET-DG. Furthermore, the error produced by ET-DG is the smallest observed. Therefore, we concluded that the scheme ET-DG is the most accurate.

# Chapter 5

## Cubed-sphere finite-volume methods

In this chapter, we demonstrate the application of the dimension splitting method, presented in Chapter 3, to solve the advection equation on the cubed-sphere based on Putman and Lin (2007). One significant difference is that on the cubed-sphere, special attention must be given to the stencils near the cube edges. Additionally, when employing ghost cell layers, the flux at the cube edges is computed twice, requiring treatment to ensure a unique value in order to achieve mass conservation.

This chapter is organized as follows: Section 5.1 introduces the advection equation on the cubed-sphere. Section 5.2 presents its finite-volume discretization with a focus on the extension of dimension splitting (Section 5.3) as presented in Section 3.3. Numerical experiments are presented in Section 5.4, where we use dimension splitting to assess its accuracy in computing the divergence of a given vector field to check its numerical consistency, as well as to solve the advection equation. In particular, we explore different treatments for the cube edges. Section 5.5 presents the final thoughts.

### 5.1 Cubed-sphere advection equation in integral form

Given a tangent velocity field  $\mathbf{u}$  on the sphere, we denote its contravariant components by  $u$  and  $v$ . We shall use all the notations introduced in Section 4.1.4. The advection equation on panel the  $p$  of the cubed-sphere with initial condition  $q_0$  is given by:

$$\begin{cases} \left[ \partial_t q + \frac{1}{\sigma} \left( \partial_x(uq\sigma) + \partial_y(vq\sigma) \right) \right] (x, y, p, t) = 0, \\ q(x, y, p, 0) = q_0(x, y, p), \end{cases} \quad (5.1)$$

$\forall (x, y) \in [-a, a]^2, t \in [0, T]$ . We denote by  $\nabla \cdot (q\mathbf{u})$  the divergence operator:

$$\nabla \cdot (q\mathbf{u})(x, y, p, t) = \frac{1}{\sigma} [\partial_x(uq\sigma) + \partial_y(vq\sigma)](x, y, p, t). \quad (5.2)$$

We recall that we say the  $\mathbf{u}$  is **non-divergent** if  $\nabla \cdot \mathbf{u} = 0$ . We define the  $\mathcal{CS}_N$  grid function  $\delta^n$  as the exact divergence of  $q\mathbf{u}$  at the cell centers, namely

$$\delta_{ijp}^n = \nabla \cdot (q\mathbf{u})(x_i, y_j, p, t^n). \quad (5.3)$$

Since the metric tensor does not depend on  $t$ , we may rewrite Equation (5.1) as

$$\left[ \partial_t(q\sigma) + \partial_x(uq\sigma) + \partial_y(vq\sigma) \right](x, y, p, t) = 0. \quad (5.4)$$

Therefore, as in Problem (3.1), the integral form of Equation (5.1) is stated in Problem (5.1).

**Problem 5.1.** *Given an initial condition  $q_0$  and a velocity on the sphere  $\mathbf{u}$ , with contravariant components  $(u, v)$  on the cubed-sphere coordinate system, we would like to find a weak solution  $q$  of the cubed-sphere advection equation in its integral form:*

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

$\forall [x_1, x_2] \times [y_1, y_2] \times [t_1, t_2] \subset \Omega \times [0, T]$ , and  $q(x, y, p, 0) = q_0(x, y, p)$ .

Similarly to Section 2.1, Equation (5.1) and Problem (5.1) are equivalent when  $q, \mathbf{u} \in \mathcal{C}^1(\mathbb{S}_R^2)$ . For Problem 5.1, the total mass in  $\mathbb{S}_R^2$  is defined by:

$$M_{\mathbb{S}_R^2}(t) = \sum_{p=1}^6 \int_{\Omega} (q\sigma)(x, y, p, t) dx dy, \quad \forall t \in [0, T], \quad (5.5)$$

and is conserved within time:

$$M_{\mathbb{S}_R^2}(t) = M_{\mathbb{S}_R^2}(0), \quad \forall t \in [0, T]. \quad (5.6)$$

We define a discretized version of Problem (5.1) as Problem (5.2).

**Problem 5.2.** *Assume the framework of Problem 5.1 and consider a  $(\Delta x, \Delta y, \Delta t, \lambda)$ -discretization of  $\Omega \times [0, T]$ , with  $\Delta x = \Delta y$ . Since we are in the framework of Problem 5.1, it*

follows that:

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

where

$$Q_{ijp}(t) = \frac{1}{|\Omega_{ijp}|} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (q \sigma)(x, y, p, t) dx dy. \quad (5.7)$$

Our problem now consists of finding the values  $Q_{ijp}(t_n)$ ,  $\forall i = 1, \dots, N$ ,  $\forall j = 1, \dots, M$ ,  $\forall n = 0, \dots, N_T - 1$ , given the initial values  $Q_{ijp}(0)$ ,  $\forall i = 1, \dots, N$ ,  $\forall j = 1, \dots, M$ . In other words, we aim to find the average values of  $q$  in each control volume  $\Omega_{ijp}$  at the specified time instances.

It is important to note that no approximations have been made in Problems (5.1) and (5.2). In practice, the term  $\frac{\Delta x \Delta y}{|\Omega_{ijp}|}$  is estimated using the second-order formula obtained by applying the midpoint rule in Equation (4.22):

$$\frac{\Delta x \Delta y}{|\Omega_{ijp}|} = \frac{1}{\sigma_{ijp}} + O(\Delta x^2). \quad (5.8)$$

## 5.2 Finite-volume on the cubed-sphere approach

We are ready to introduce the finite-volume scheme on the cubed-sphere (CS-FV). A CS-FV scheme problem as follows in Problem 5.3.

**Problem 5.3** (CS-FV scheme). *Assume the framework defined in Problem 5.2. The finite-volume approach of Problem 5.1 consists of a finding a scheme of the form:*

$$Q_{ijp}^{n+1} = Q_{ijp}^n - \frac{\lambda}{\sigma_{ijp}} \delta_i F_{ijp}^n - \frac{\lambda}{\sigma_{ijp}} \delta_j G_{ijp}^n, \quad (5.9)$$

$$\forall i = 1, \dots, N, \quad \forall j = 1, \dots, M, \quad p = 1, \dots, 6, \quad \forall n = 0, \dots, N_T - 1,$$

where  $\delta_i F_{ijp}^n = F_{i+\frac{1}{2},j,p}^n - F_{i-\frac{1}{2},j,p}^n$ ,  $\delta_j G_{ijp}^n = G_{i,j+\frac{1}{2},p}^n - G_{i,j-\frac{1}{2},p}^n$  and  $Q^n \in \mathcal{CS}_N$  is intended to be an approximation of  $Q(t_n) \in \mathcal{CS}_N$  in some sense. We define  $Q_{ijp}^0 = Q_{ijp}(0)$  or  $Q_{ijp}^0 = q_{ijp}^0$ .

The term  $F_{i+\frac{1}{2},j,p}^n$  is known as numerical flux in the  $x$  direction and it approximates  $\frac{1}{\Delta t \Delta y} \int_{t_n}^{t_{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (u q \sigma)(x_{i+\frac{1}{2}}, y, p, t) dy dt$ ,  $\forall i = 0, 1, \dots, N$ , and  $G_{i,j+\frac{1}{2},p}^n$  is known as numerical flux in the  $y$  direction and it approximates  $\frac{1}{\Delta t \Delta x} \int_{t_n}^{t_{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (v q \sigma)(x, y_{j+\frac{1}{2}}, p, t) dx dt$ ,  $\forall j = 0, 1, \dots, M$ , or, in other words, they estimate the time-averaged fluxes at the control volume  $\Omega_{ijp}$  boundaries.

**Remark 5.1.** For Problem 5.3, we define the CFL number in the  $x$  and  $y$  direction by  $\max\{|u_{i+\frac{1}{2},j}^n|\} \frac{\Delta t}{\Delta x}$  and  $\max\{|v_{i,j+\frac{1}{2}}^n|\} \frac{\Delta t}{\Delta y}$ , respectively. The CFL number is maximum between these numbers and we say that the CFL condition is satisfied if the CFL number is less than one.

As we mentioned in Problem 5.3, the initial condition may be assumed as  $q_{ijp}^0$  or  $Q_{ijp}(0)$ . We are going to assume  $q_{ijp}^0$  as initial data to avoid the computation of integrals. Furthermore, the errors will be calculated using the values  $q_{ijp}^n$  instead of  $Q_{ijp}(t_n)$ . As in Section 3.2 this approximation leads to a second-order error.

As in Section 3.2 we introduce the notion of discrete divergence, which allow us to check the consistency of CS-FV schemes.

**Definition 5.1** (Discrete divergence). *For Problem 5.3, we define the discrete divergence as a CS<sub>N</sub>-grid function  $\mathbb{D}^n(Q^n, u^n, v^n)$  given by:*

$$\mathbb{D}_{ijp}^n(Q^n, u^n, v^n) = \frac{1}{\Delta t \sigma_{ijp}} \left( \frac{\delta_i F_{ijp}^n}{\Delta x} + \frac{\delta_j G_{ijp}^n}{\Delta y} \right), \quad i = 1, \dots, N, \quad j = 1, \dots, M. \quad (5.10)$$

With the aid of the discrete divergence, Equation (5.9) becomes:

$$Q^{n+1} = Q^n - \Delta t \mathbb{D}^n(Q^n, u^n, v^n). \quad (5.11)$$

For a CS-FV scheme the discrete total mass at the time-step  $n$  is given by

$$M^n = \sum_{p=1}^6 \sum_{i,j=1}^N Q_{ijp}^n \sigma_{ijp} \Delta x \Delta y$$

It follows from Equation (5.11) that:

$$M^{n+1} = M^n - \sum_{p=1}^6 \sum_{i,j=1}^N \mathbb{D}_{ijp}^n \sigma_{ijp} \Delta x \Delta y.$$

Hence, to ensure mass conservation, we must ensure that

$$\sum_{p=1}^6 \sum_{i,j=1}^N \mathbb{D}_{ijp}^n \sigma_{ijp} \Delta x \Delta y = 0.$$

This property is discrete version of

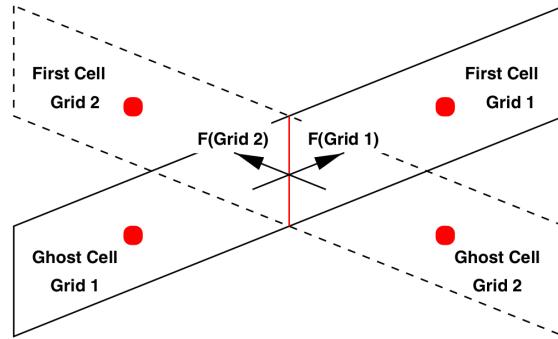
$$\int_{S_R^2} \nabla \cdot (\mathbf{u} q) dS = 0,$$

which follows from the divergence theorem and the fact of the sphere has no boundary, where  $dS$  is the surface measure of the sphere.

### 5.2.1 Mass fixer

When computing the flux, if we ignore the discontinuity in the cubed sphere coordinate system and use values from adjacent panels (as in the ET-S72 scheme from Chapter 4) to compute stencils, we can ensure mass conservation because the flux at points lying on the cube edge will be the same. However, if we consider ghost cell layers by extending the gridlines (as in the ET-DG scheme from Chapter 4), the flux is computed twice at points

lying on the cube edge. Therefore, in this case, some modification is needed to ensure mass conservation (Figure 5.1).



**Figure 5.1:** Figure that illustrates the flux being computed twice on the cube edge, breaking the total mass conservation. Figure taken from Rossmanith (2006).

### Average flux at cube interfaces

One common alternative used in the literature to handle the issue of values being defined twice at points on the cube edges is to simply average the values (as seen in works such as C. Chen and Xiao (2008) and Rossmanith (2006)). This approach will be explored in this chapter as one of the mass fixers and is referred to as **MF-AF** (AF representing average flux and MF representing mass fixer).

In order to investigate the impact of using the average flux, consider the Figure 5.1 and let us denote by  $Q_{1,j,p_1}^n$  the average value at the first cell on grid 1 and  $Q_{N,j,p_2}^n$  the average value at the first cell on grid 2. Also, we denote the numerical flux functions from grid 1 by  $F_{i+\frac{1}{2},j,p_1}$  and  $G_{i,j+\frac{1}{2},p_1}$ , and the fluxes from grid 2 is given by  $F_{i+\frac{1}{2},j,p_2}$  and  $G_{i,j+\frac{1}{2},p_2}$ . Therefore, the average values depicted in Figure 5.1 are updated as follows:

$$Q_{1,j,p_1}^{n+1} = Q_{1,j,p_1}^n - \frac{\Delta t}{\sigma_{1,j,p_1}} \left( \frac{F_{1+\frac{1}{2},j,p_1} - F_{\frac{1}{2},j,p_1}}{\Delta x} + \frac{G_{1,j+\frac{1}{2},p_1} - G_{1,j-\frac{1}{2},p_1}}{\Delta y} \right), \quad (5.12)$$

$$Q_{N,j,p_2}^{n+1} = Q_{N,j,p_2}^n - \frac{\Delta t}{\sigma_{N,j,p_2}} \left( \frac{F_{N+\frac{1}{2},j,p_2} - F_{N-\frac{1}{2},j,p_2}}{\Delta x} + \frac{G_{N,j+\frac{1}{2},p_2} - G_{N,j-\frac{1}{2},p_2}}{\Delta y} \right). \quad (5.13)$$

We also shall assume the following regarding the numerical fluxes consistency:

$$F_{i+\frac{1}{2},j,p} = \frac{1}{\Delta t \Delta x} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{t^n}^{t^{n+1}} u \sigma q(x_{i+\frac{1}{2}}, y, p, t) dt dy + O(\Delta x^P),$$

$$G_{i,j+\frac{1}{2},j,p} = \frac{1}{\Delta t \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{t^n}^{t^{n+1}} v \sigma q(x, y_{j+\frac{1}{2}}, p, t) dt dx + O(\Delta x^P),$$

for any panel  $p$ . That is, they are consistent with order  $P$ . In particular, at the cube interfaces, we have:

$$F_{N+\frac{1}{2},j,p_2} - F_{\frac{1}{2},j,p_1} = O(\Delta x^P). \quad (5.14)$$

The average flux at the cube interface is denoted by:

$$F_{\frac{1}{2},j,p_1}^* = \frac{F_{\frac{1}{2},j,p_1} + F_{N+\frac{1}{2},j,p_2}}{2}. \quad (5.15)$$

Then, if we replace  $F_{\frac{1}{2},j,p_1}^*$  by  $F_{\frac{1}{2},j,p_1}$  in Equation (5.12), we obtain:

$$\begin{aligned} Q_{1,j,p_1}^{n+1} &= Q_{1,j,p_1}^n - \frac{\Delta t}{\sigma_{1,j,p_1}} \left( \frac{F_{1+\frac{1}{2},j,p_1} - F_{\frac{1}{2},j,p_1}^*}{\Delta x} + \frac{G_{1,j+\frac{1}{2},p_1} - G_{1,j-\frac{1}{2},p_1}}{\Delta y} \right) \\ &= Q_{1,j,p_1}^n - \frac{\Delta t}{\sigma_{1,j,p_1}} \left( \frac{F_{1+\frac{1}{2},j,p_1} - F_{\frac{1}{2},j,p_1}}{\Delta x} + \frac{G_{1,j+\frac{1}{2},p_1} - G_{1,j-\frac{1}{2},p_1}}{\Delta y} \right) + \frac{\Delta t}{\sigma_{1,j,p_1}} \frac{(F_{\frac{1}{2},j,p_1} - F_{N+\frac{1}{2},j,p_2})}{2\Delta x} \end{aligned}$$

Therefore, it follows from Equation (5.14) that:

$$\frac{\sigma_{1,j,p_1}(Q_{1,j,p_1}^{n+1} - Q_{1,j,p_1}^n)}{\Delta t} + \left( \frac{F_{1+\frac{1}{2},j,p_1} - F_{\frac{1}{2},j,p_1}}{\Delta x} + \frac{G_{1,j+\frac{1}{2},p_1} - G_{1,j-\frac{1}{2},p_1}}{\Delta y} \right) = O(\Delta x^{P-1}).$$

Hence, the average flux reduces the local truncation error order by one at the cube interfaces and we expect higher errors at this region. In particular, if  $P = 1$ , the local truncation error does not converge to zero.

### Average flux at cube interfaces

Another alternative approach, based on Semi-Lagrangian mass corrections (for a review of these methods, we refer to Diamantakis (2014)), involves computing the orthogonal projection of the grid function  $\mathbb{D}^n$  onto the following space:

$$V_0 = \{\delta \in \mathcal{CS}_N : \sum_{p=1}^6 \sum_{i,j=1}^N \delta_{ijp} \sigma_{ijp} \Delta x \Delta y = 0\},$$

where  $V_0$  represents the space of functions with a zero average. The element of  $V_0$  that minimizes its distance to  $\mathbb{D}^n$  in the norm 2 is given by:

$$\mathbb{D}_{ijp}^{cor,n} = \mathbb{D}_{ijp}^n - \left( \frac{\sum_{p=1}^6 \sum_{r,s=1}^N \mathbb{D}_{rsp}^n \sigma_{rsp} \Delta x \Delta y}{\sum_{p=1}^6 \sum_{r,s=1}^N \sigma_{rsp}^2 \Delta x \Delta y} \right) \sigma_{ijp}.$$

When employing the projection scheme, we refer to it as **MF-PR** (PR representing projection).

## 5.3 Dimension splitting

In this section, we will utilize the operator splitting method described in Section 3.3 to obtain a CS-FV scheme. We will focus on 1D fluxes, specifically the PPM fluxes introduced in Section 2.5, denoted as  $F_{i+\frac{1}{2},j,p}^n$  and  $G_{i,j+\frac{1}{2},p}^n$ . To facilitate the scheme, we introduce the

auxiliary grid functions  $\mathbf{F}$  and  $\mathbf{G}$  belonging to  $\mathcal{CS}_N$ , defined as follows:

$$\mathbf{F}_{ijp}(Q^n, \tilde{u}^n) = -\lambda \left( F_{i+\frac{1}{2},j,p}^n(Q_{\times,j,p}^n, \tilde{u}_{i+\frac{1}{2},j,p}^n) - F_{i-\frac{1}{2},j,p}^n(Q_{\times,j,p}^n, \tilde{u}_{i-\frac{1}{2},j,p}^n) \right),$$

for  $i = 1, \dots, N$ ,  $j = -v + 1, \dots, N + v$ , and

$$\mathbf{G}_{ijp}(Q^n, \tilde{v}^n) = -\lambda \left( G_{i,j+\frac{1}{2},p}^n(Q_{i,\times,p}^n, \tilde{v}_{i,j+\frac{1}{2},p}^n) - G_{i,j-\frac{1}{2},p}^n(Q_{i,\times,p}^n, \tilde{v}_{i,j-\frac{1}{2},p}^n) \right),$$

for  $i = -v + 1, \dots, N + v$ ,  $j = 1, \dots, N$ . The terms  $\tilde{u}_{i+\frac{1}{2},j,p}^n$  and  $\tilde{v}_{i,j+\frac{1}{2},p}^n$  represent the time-averaged winds used in the computation of departure points in the  $x$  and  $y$  directions, respectively. These averages are calculated using either RK1 or RK2 from Section 2.3. To facilitate the description of the 1D fluxes  $F_{i+\frac{1}{2},j,p}^n$  and  $G_{i,j+\frac{1}{2},p}^n$ , we introduce the time-averaged CFL numbers defined as follows:

$$\tilde{c}_{i+\frac{1}{2},j,p}^{x,n} = \tilde{u}_{i+\frac{1}{2},j,p}^{x,n} \frac{\Delta t}{\Delta x}, \quad \tilde{c}_{i,j+\frac{1}{2},p}^{y,n} = \tilde{v}_{i,j+\frac{1}{2},p}^{y,n} \frac{\Delta t}{\Delta y}.$$

The discrete divergence is then obtained as:

$$\mathbb{D}_{ijp}^n = -\frac{1}{\Delta t \sigma_{ijp}} \left[ \mathbf{F}_{ijp} \left( Q^n + \frac{1}{2} \mathbf{g}(Q^n, \tilde{v}^n), \tilde{u}^n \right) + \mathbf{G}_{ijp} \left( Q^n + \frac{1}{2} \mathbf{f}(Q^n, \tilde{u}^n), \tilde{v}^n \right) \right], \quad (5.16)$$

where the inner advective operators  $\mathbf{f}$  and  $\mathbf{g}$  are given in Table 3.1.

Now, our objective is to describe the 1D fluxes  $F_{i+\frac{1}{2},j,p}^n$  and  $G_{i,j+\frac{1}{2},p}^n$ . It is important to note that these fluxes depend on the edge treatment, as well as the computation of stencils and metric tensor treatment, as we will see in the following section.

### 5.3.1 Metric tensor treatment

#### MT-0

Let us assume that we are given  $Q^n$ , with its ghost cells filled using some edge treatment. When solving Equation (5.1) in the  $x$  direction, we need to reconstruct the function  $\psi = q\sigma$ . Then, we have a piecewise-parabolic approximation in the  $x$  direction:

$$\begin{cases} \psi_{ijp}^x(x; Q_{\times,j,p}^n) = \psi_{L,ijp}^x + \Delta\psi_{ijp}^x z_i(x) + \psi_{6,ijp}^x z_i(x)(1 - z_i(x)), \\ z_i(x) = \frac{x - x_{i-\frac{1}{2}}}{\Delta x}, \quad x \in X_i, \\ \psi_{L,ijp}^x = q_{R,ijp}^x \sigma_{i-\frac{1}{2},j,p}, \quad \psi_{R,ijp}^x = q_{L,ijp}^x \sigma_{i+\frac{1}{2},j,p}, \\ q_{L,ijp}^x = q_{i-\frac{1}{2},j,p}^n + O(\Delta x^2), \quad q_{R,ijp}^x = q_{i+\frac{1}{2},j,p}^n + O(\Delta x^2), \\ \Delta\psi_{ijp}^x = \psi_{R,ijp}^x - \psi_{L,ijp}^x, \quad \psi_{6,ijp}^x = 6 \left( \sigma_{ijp} Q_{ijp}^n - \frac{(\psi_{L,ijp}^x + \psi_{R,ijp}^x)}{2} \right), \end{cases} \quad (5.17)$$

for  $i = 1, \dots, N$ ,  $j = -v + 1, \dots, M + v$ ,  $p = 1, \dots, 6$ , and we also construct a piecewise-parabolic approximation in the  $y$  direction:

$$\begin{cases} \psi_{ijp}^y(y; Q_{i,x,p}^n) = \psi_{L,ijp}^y + \Delta\psi_{ijp}^y z_j(y) + \psi_{6,ijp}^y z_j(y)(1 - z_j(y)), \\ z_j(y) = \frac{y - y_{j-\frac{1}{2}}}{\Delta y}, \quad y \in Y_j, \\ \psi_{L,ijp}^y = q_{R,ijp}^y \sigma_{i,j-\frac{1}{2},p}, \quad \psi_{R,ijp}^y = q_{L,ijp}^y \sigma_{i,j+\frac{1}{2},p}, \\ q_{L,ijp}^y = q_{i,j-\frac{1}{2},p}^n + O(\Delta y^2), \quad q_{R,ijp}^y = q_{i,j+\frac{1}{2},p}^n + O(\Delta y^2), \\ \Delta\psi_{ijp}^y = q_{R,ijp}^y - q_{L,ijp}^y, \quad \psi_{6,ijp}^y = 6 \left( \sigma_{ijp} Q_{ijp}^n - \frac{(\psi_{L,ijp}^y + \psi_{R,ijp}^y)}{2} \right), \end{cases} \quad (5.18)$$

for  $i = -v + 1, \dots, N + v$ ,  $j = 1, \dots, M$ ,  $p = 1, \dots, 6$ . The values  $q_{L,ijp}^x$ ,  $q_{R,ijp}^x$ ,  $q_{L,ijp}^y$ , and  $q_{R,ijp}^y$  are computed using the PPM stencils as in Section 4.2.3. Then, integrating the parabolas with the aid of Equation (2.57), we may express the fluxes as:

$$\begin{aligned} F_{i+\frac{1}{2},j,p}^n(Q_{i,x,j,p}^n, \tilde{u}_{i+\frac{1}{2},j,p}^n) &= \int_{x_{i+\frac{1}{2}} - \tilde{u}_{i+\frac{1}{2},j,p}^n}^{x_{i+\frac{1}{2}}} \psi_{ijp}^x(x; Q_{i,x,p}^n) dx = \\ \tilde{u}_{i+\frac{1}{2},j,p}^n \times \begin{cases} \psi_{R,ijp}^x + \frac{1}{2}(\psi_{6,i,j}^x - \Delta\psi_{ijp}^x) \tilde{c}_{i+\frac{1}{2},j,p}^{x,n} + \frac{1}{3}\psi_{6,ijp}^x (\tilde{c}_{i+\frac{1}{2},j,p}^{x,n})^2, & \text{if } \tilde{u}_{i+\frac{1}{2},j,p}^n > 0, \\ \psi_{L,i+1,j,p}^x - \frac{1}{2}(\psi_{6,i+1,j,p}^x + \Delta\psi_{i+1,j,p}^x) \tilde{c}_{i+\frac{1}{2},j,p}^{x,n} - \frac{1}{3}\psi_{6,i+1,j,p}^x (\tilde{c}_{i+\frac{1}{2},j,p}^{x,n})^2, & \text{if } \tilde{u}_{i+\frac{1}{2},j,p}^n \leq 0, \end{cases} \end{aligned} \quad (5.19)$$

for  $i = 0, \dots, N$ ,  $j = -v + 1, \dots, M + v$ ,  $p = 1, \dots, 6$  and

$$\begin{aligned} G_{i,j+\frac{1}{2},p}^n(Q_{i,y,j,p}^n, \tilde{v}_{i,j+\frac{1}{2},p}^n) &= \int_{y_{j+\frac{1}{2}} - \tilde{v}_{i,j+\frac{1}{2},p}^n}^{y_{j+\frac{1}{2}}} \psi_{ijp}^y(y; Q_{i,y,p}^n) dy = \\ \tilde{v}_{i,j+\frac{1}{2},p}^n \times \begin{cases} \psi_{R,ijp}^y + \frac{1}{2}(\psi_{6,i,j}^y - \Delta\psi_{ijp}^y) \tilde{c}_{i,j+\frac{1}{2},p}^{y,n} + \frac{1}{3}\psi_{6,ijp}^y (\tilde{c}_{i,j+\frac{1}{2},p}^{y,n})^2, & \text{if } \tilde{v}_{i,j+\frac{1}{2},p}^n > 0, \\ \psi_{L,i,j+1,p}^y - \frac{1}{2}(\psi_{6,i,j+1,p}^y + \Delta\psi_{i,j+1,p}^y) \tilde{c}_{i,j+\frac{1}{2},p}^{y,n} - \frac{1}{3}\psi_{6,i,j+1,p}^y (\tilde{c}_{i,j+\frac{1}{2},p}^{y,n})^2, & \text{if } \tilde{v}_{i,j+\frac{1}{2},p}^n \leq 0, \end{cases} \end{aligned} \quad (5.20)$$

for  $i = -v + 1, \dots, N + v$ ,  $j = 0, \dots, M$ ,  $p = 1, \dots, 6$ . This flux formulation that incorporates the metric tensor in the parabolic approximation is referred to as **MT-0** (MT stands for metric tensor).

## MT-PL07

Let us assume again that we are given  $Q^n$ , with its ghost cells filled using some edge treatment. Another approach to compute the 1D fluxes, employed in Putman and Lin (2007) and Lin (2004), involves disregarding the metric tensor and constructing a piecewise-parabolic approximation of  $q$  in the  $x$  direction:

$$\begin{cases} q_{ijp}^x(x; Q_{i,x,j,p}^n) = q_{L,ijp}^x + \Delta q_{ijp}^x z_i(x) + q_{6,ijp}^x z_i(x)(1 - z_i(x)), \\ z_i(x) = \frac{x - x_{i-\frac{1}{2}}}{\Delta x}, \quad x \in X_i, \\ q_{L,ijp}^x = q_{i-\frac{1}{2},j,p}^n + O(\Delta x^2), \quad q_{R,ijp}^x = q_{i+\frac{1}{2},j,p}^n + O(\Delta x^2), \\ \Delta q_{ijp}^x = q_{R,ijp}^x - q_{L,ijp}^x, \quad q_{6,ijp}^x = 6 \left( Q_{ijp}^n - \frac{(q_{L,ijp}^x + q_{R,ijp}^x)}{2} \right), \end{cases} \quad (5.21)$$

for  $i = 1, \dots, N$ ,  $j = -v+1, \dots, M+v$ ,  $p = 1, \dots, 6$  and we also construct a piecewise-parabolic approximation in the  $y$  direction:

$$\begin{cases} q_{ijp}^y(y; Q_{i,x}^n) = q_{L,ijp}^y + \Delta q_{ijp}^y z_j(y) + q_{6,ijp}^y z_j(y)(1 - z_j(y)), \\ z_j(y) = \frac{y - y_{j-\frac{1}{2}}}{\Delta y}, \quad y \in Y_j, \\ q_{L,ijp}^y = q_{i,j-\frac{1}{2},p}^n + O(\Delta y^2), \quad q_{R,ijp}^y = q_{i,j+\frac{1}{2},p}^n + O(\Delta y^2), \\ \Delta q_{ijp}^y = q_{R,ijp}^y - q_{L,ijp}^y, \quad q_{6,ijp}^y = 6 \left( Q_{ijp}^n - \frac{(q_{L,ijp}^y + q_{R,ijp}^y)}{2} \right), \end{cases} \quad (5.22)$$

for  $i = -v+1, \dots, N+v$ ,  $j = 1, \dots, M$ ,  $p = 1, \dots, 6$ . Then, we integrate the  $\sigma q_{ijp}^x$  and  $\sigma q_{ijp}^y$  assuming that metric tensor is constant on each integration domain, and, with the aid of Equation (2.57), we get

$$\begin{aligned} F_{i+\frac{1}{2},j,p}^n(Q_{x,j,p}^n, \tilde{u}_{i+\frac{1}{2},j,p}^n) &= \int_{x_{i+\frac{1}{2}} - \tilde{u}_{i+\frac{1}{2},j,p}^n}^{x_{i+\frac{1}{2}}} \sigma(x, y_j, p) q_{ijp}^x(x; Q_{x,j,p}^n) dx = \\ \sigma_{i+\frac{1}{2},j,p} \tilde{u}_{i+\frac{1}{2},j,p}^n &\times \begin{cases} q_{R,ijp}^x + \frac{1}{2}(q_{6,ijp}^x - \Delta q_{ijp}^x) \tilde{c}_{i+\frac{1}{2},j,p}^{x,n} + \frac{1}{3} q_{6,ijp}^x (\tilde{c}_{i+\frac{1}{2},j}^{x,n})^2, & \text{if } \tilde{u}_{i+\frac{1}{2},j,p}^n > 0, \\ q_{L,i+1,j,p}^x - \frac{1}{2}(q_{6,i+1,j,p}^x + \Delta q_{i+1,j,p}^x) \tilde{c}_{i+\frac{1}{2},j,p}^{x,n} - \frac{1}{3} q_{6,i+1,j,p}^x (\tilde{c}_{i+\frac{1}{2},j,p}^{x,n})^2, & \text{if } \tilde{u}_{i+\frac{1}{2},j,p}^n \leq 0, \end{cases} \end{aligned} \quad (5.23)$$

for  $i = 0, \dots, N$ ,  $j = -v+1, \dots, M+v$ ,  $p = 1, \dots, 6$  and

$$\begin{aligned} G_{i,j+\frac{1}{2},p}^n(Q_{x,j,p}^n, \tilde{v}_{i,j+\frac{1}{2},p}^n) &= \int_{y_{j+\frac{1}{2}} - \tilde{v}_{i,j+\frac{1}{2},p}^n}^{y_{j+\frac{1}{2}}} \sigma(x_i, y, p) q_{ijp}^y(y; Q_{i,x,p}^n) dy = \\ \sigma_{i,j+\frac{1}{2},p} \tilde{v}_{i,j+\frac{1}{2},p}^n &\times \begin{cases} q_{R,ijp}^y + \frac{1}{2}(q_{6,ijp}^y - \Delta q_{ijp}^y) \tilde{c}_{i,j+\frac{1}{2},p}^{y,n} + \frac{1}{3} q_{6,ijp}^y (\tilde{c}_{i,j+\frac{1}{2},p}^{y,n})^2, & \text{if } \tilde{v}_{i,j+\frac{1}{2},p}^n > 0, \\ q_{L,i,j+1,p}^y - \frac{1}{2}(q_{6,i,j+1}^y + \Delta q_{i,j+1,p}^y) \tilde{c}_{i,j+\frac{1}{2},p}^{y,n} - \frac{1}{3} q_{6,i,j+1,p}^y (\tilde{c}_{i,j+\frac{1}{2},p}^{y,n})^2, & \text{if } \tilde{v}_{i,j+\frac{1}{2},p}^n \leq 0, \end{cases} \end{aligned} \quad (5.24)$$

for  $i = -v+1, \dots, N+v$ ,  $j = 0, \dots, M$ ,  $p = 1, \dots, 6$ . This flux formulation is referred to as **MT-PL07** and introduces a first-order error because of the assumption of a constant metric tensor on each integration domain.

### 5.3.2 Flux at edges treatment

If we choose the ET-DG scheme, we will calculate the required wind values for ghost positions as described in Section 4.2.2. However, this scheme requires modifying the flux at the edges to maintain mass conservation. To achieve this, we will apply the mass fixer schemes discussed earlier. Using these flux values, we can compute the discrete divergence (Equation (5.16)) for the interior cells and update the solution for the next time step.

On the other hand, if we utilize the ET-PL07 scheme, we will populate the scalar field and velocity field contravariant components at ghost cell positions with the adjacent panel values, similar to the ET-S72 scheme. This allows us to compute  $\mathbf{F}$  and  $\mathbf{G}$  at all necessary positions and we can update the solution. In the PPM reconstruction, we employ extrapolation near the cubed edges that defines the ET-PL07 scheme. When combining MT-PL07 with ET-PL07, a first-order error is introduced. However, we can ensure that the PL07 splitting scheme (Table 3.1) eliminates the splitting error when  $q$  is constant.

## 5.4 Numerical experiments

This section is dedicated to present the numerical experiments for the advection equation on the sphere. In Table 5.1 we present the initial conditions (IC) and in Table 5.2 we present the velocity fields (VF) considered.

IC name	$q_0$
IC1	$\exp(b_0((X - X_0)^2 + (Y - Y_0)^2 + (Z - Z_0)^2))$
IC2	$\exp(b_0[(X - X_1)^2 + (Y - Y_1)^2 + (Z - Z_1)^2]) + \exp(b_0[(X - X_2)^2 + (Y - Y_2)^2 + (Z - Z_2)^2])$
IC3	$1 + \cos(\lambda) \cos(\phi) \sin(\alpha) - \sin(\phi) \cos(\alpha),$

**Table 5.1:** Initial conditions considered in the numerical experiments (Figure 5.2).

VF name	$u_\lambda(\lambda, \phi, t)$	$v_\phi(\lambda, \phi, t)$	$\Delta t^{(0)}$
VF1	$u_0(\cos(\phi) \cos(\alpha) + \sin(\phi) \cos(\lambda) \sin(\alpha))$	$-u_0 \sin(\lambda) \sin(\alpha)$	0.025
VF2	$k \sin^2(\lambda_p + \pi) \sin(2\phi) \cos(\frac{\pi t}{T}) + \frac{2\pi}{T} \cos \phi$	$k \sin(2(\lambda_p + \pi)) \cos(\phi) \cos(\frac{\pi t}{T})$	0.0125
VF3	$-k \sin^2(\frac{\lambda+\pi}{2}) \sin(2\phi) \cos^2(\phi) \cos(\frac{\pi t}{T}),$	$\frac{k}{2} \sin(\lambda + \pi) \cos^3(\phi) \cos(\frac{\pi t}{T})$	0.00625

**Table 5.2:** Velocity fields considered in the numerical experiments and its initial time step  $\Delta t^{(0)}$ .

In Table 5.1,  $(X_0, Y_0, Z_0)$ ,  $(X_1, Y_1, Z_1)$  and  $(X_2, Y_2, Z_2)$  are the Cartesian coordinates of the latitude-longitude points  $(\lambda_0, \phi_0) = (\frac{\pi}{4}, \frac{\pi}{6})$ ,  $(\lambda_1, \phi_1) = (-\frac{\pi}{6}, 0)$  and  $(\lambda_2, \phi_2) = (\frac{\pi}{6}, 0)$ . IC1 represents a Gaussian hill centered at a cube corner and we set  $b_0 = -10$ . IC2 represents two Gaussian hills as suggested by Nair and Lauritzen (2010) and we set  $b_0 = -5$ . IC3 is based on the steady flow presented in D. Williamson et al. (1992) and we adopt  $\alpha = -\frac{45\pi}{180}$ . The initial conditions are shown in Figure 5.2.

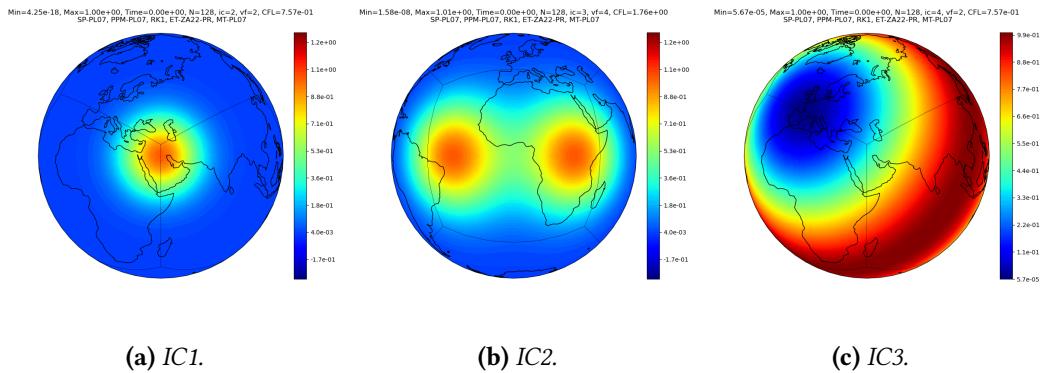
For the velocities provided in Table 5.2, we adopt the following parameter values:  $u_0 = \frac{2\pi}{5}$ ,  $\alpha = -\frac{45\pi}{180}$ ,  $\lambda_p = \lambda - \frac{2\pi t}{T}$ ,  $T = 5$ . In this context, VF1 represents the non-divergent rotated zonal field introduced in D. Williamson et al. (1992). VF2 corresponds to the non-divergent deformational flow described in Nair and Lauritzen (2010), and VF3 represents the divergent flow also presented in Nair and Lauritzen (2010). Additionally, VF2 utilizes  $k = 2$  and VF3 uses  $k = 1$ . For all velocity fields presented here, the initial condition is equal to the final solution after 5 time units. Also, we consider the unit sphere.

To compute the convergence, consider cubed-sphere grids with value of  $N_k = 2^{k+4}$ , and  $\Delta t^{(k)} = \frac{\Delta t^{(0)}}{2^k}$ ,  $k = 0, \dots, 6$ , where the value of  $\Delta t^{(0)}$  in Table 5.2 for each VF. The relative error in the maximum norm is computed as

$$E_k = \frac{\max |Q^{N_T} - Q^0|}{\max |Q^0|}, \quad (5.25)$$

and the convergence rate is defined as

$$CR_k = \frac{\ln \left( \frac{E_k}{E_{k-1}} \right)}{\ln 2}, \quad \text{for } k = 1, \dots, 6.$$



**Figure 5.2:** Illustration of the initial conditions considered in this chapter (Table 5.1).

In Table 5.3, we present all four schemes that we are considering. We shall always assume that the PPM-PL07 is employed. The main distinction between Scheme 1 and Scheme 2 lies in their ghost cell treatment. In Scheme 1, we adopt the approach proposed by Putman and Lin (2007) for the ghost cell formulation, while Scheme 2 utilizes the MF-PR scheme. Both Scheme 1 and Scheme 2 employ the splitting PL07 and the metric tensor formulation MT-PL07 with the RK1 departure point scheme. On the other hand, Scheme 3 and Scheme 4 utilize the splitting, the metric tensor formulation MT-0, and the RK2 departure point scheme. Scheme 1 and Scheme 2 are expected to achieve first-order accuracy, and so is schemes 3, since as we have seen in Section 5.2.1, the local truncation error when using MF-AF is reduced by one at the cube interfaces. Scheme 4 is designed to attain second-order accuracy. Finally, the key difference between Scheme 3 and Scheme 4 lies in the mass fixer.

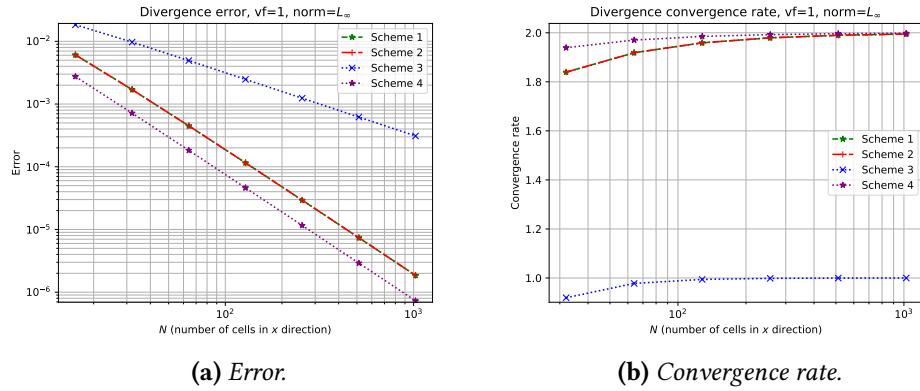
Scheme	Splitting	Departure point	Edge Treatment	Mass fixer	Metric tensor
1	PL07	RK1	ET-PL07	None	MT-PL07
2	PL07	RK1	ET-PL07	None	MT-PL07
3	AVLT	RK2	ET-DG	AF	MT-0
4	AVLT	RK2	ET-DG	PR	MT-0

**Table 5.3:** Schemes considered for this Chapter.

### 5.4.1 Divergence of the rotated zonal wind

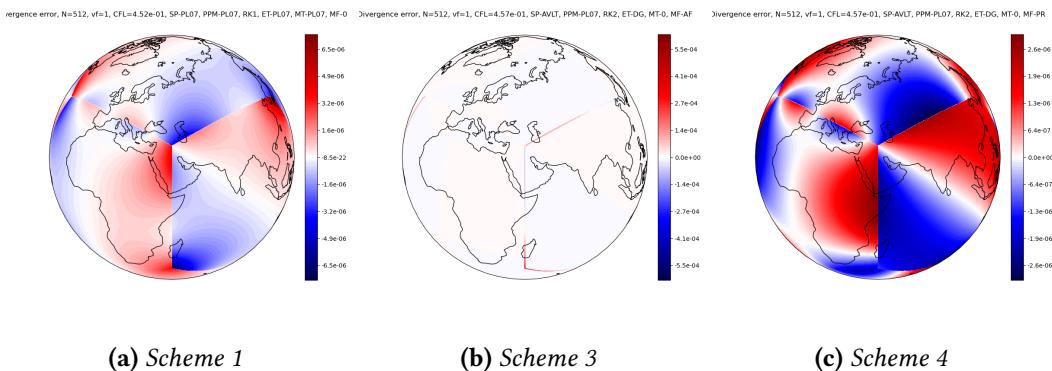
For our initial test, we will examine the velocity field VF1, which is non-divergent with  $q = 1$ . We will calculate its discrete divergence using equation (5.16). The resulting errors and convergence rates are illustrated in Figure 5.3.

From the analysis of Figure 5.3, we can observe that both Scheme 1 and Scheme 2 yield the same error, which is of second order. This result aligns with our expectations, as the PL07 splitting scheme possesses the property of eliminating splitting errors when  $q$  is constant, regardless of the edge treatment scheme. Additionally, we can notice that Scheme 3 exhibits a first-order error, while Scheme 4 demonstrates a second-order error. Hence, we can conclude that averaging the flux at the cubed edges reduces the scheme's



**Figure 5.3:** Convergence of the error for the schemes from table 5.3 obtained when computing the divergence of rotated zonal wind VF1. Green line represents scheme 1; Red line represents scheme 2; Purple line represents scheme 3; Blue line represents scheme 4.

order, while the divergence projection maintains the order. Examining Figure 5.4, which displays the obtained errors on the sphere, it becomes evident that the averaging flux introduces a grid-imprinting pattern, wherein the cube edges are observable (see Figure 5.4b). The Scheme 1,2 and 3 does not present grid-imprinting pattern, but their errors have a strong discontinuity on the panel edges.

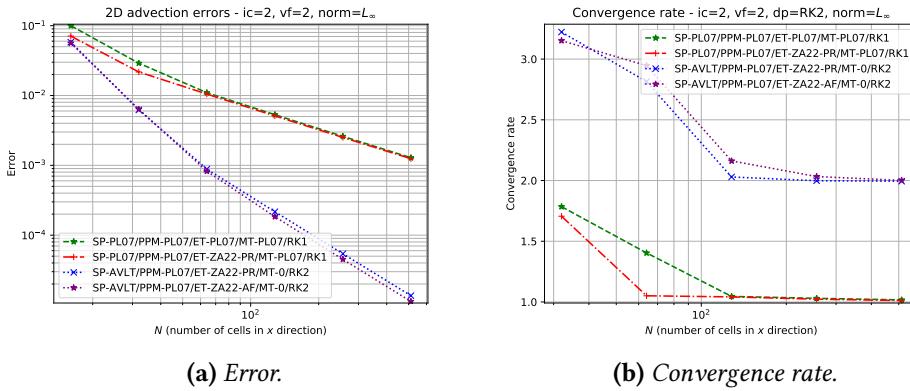


**Figure 5.4:** Divergence error obtained for all schemes from table using the VF1 velocity field from table 5.2. The error of scheme 2 is equal to the error of scheme 1 (Figure 2) This figure considers a cubed-sphere with  $N = 512$ .

### 5.4.2 Advection of one Gaussian hill through the rotated zonal wind

As a second test case, we consider the advection of the Gaussian hill given by IC1 using the rotated zonal wind VF1. In Figure 5.5, we depict the final errors obtained for all schemes from Table 5.2. We can observe that schemes 1 and scheme 2 are first-order accurate, while schemes 3 and 4 are second-order accurate. Furthermore, it is worth noting that although scheme 3 showed to be only first-order accurate in computing the divergence of VF1, the final error after simulating the advection equation is second-order accurate.

The major difference between schemes 1 and 2 is evident from Figures 5.6a and 5.6b, where we can observe that scheme 1 generates a grid imprinting pattern, while scheme 2 does not. Similarly, scheme 3 generates grid imprinting while scheme 4 does not, show in Figures 5.6c and 5.6d. We can conclude that the averaging of fluxes and the edge treatment ET-PL07 result in more grid imprinting. However, the usage of Lagrange interpolation to fill the ghost cell values eliminates grid imprinting when combined with the divergence projection method.



**Figure 5.5:** Convergence of the error for the schemes from table 5.3 obtained when computing the advection of IC1 with the rotated zonal wind VF1 after 5 time units.

### 5.4.3 Steady steady

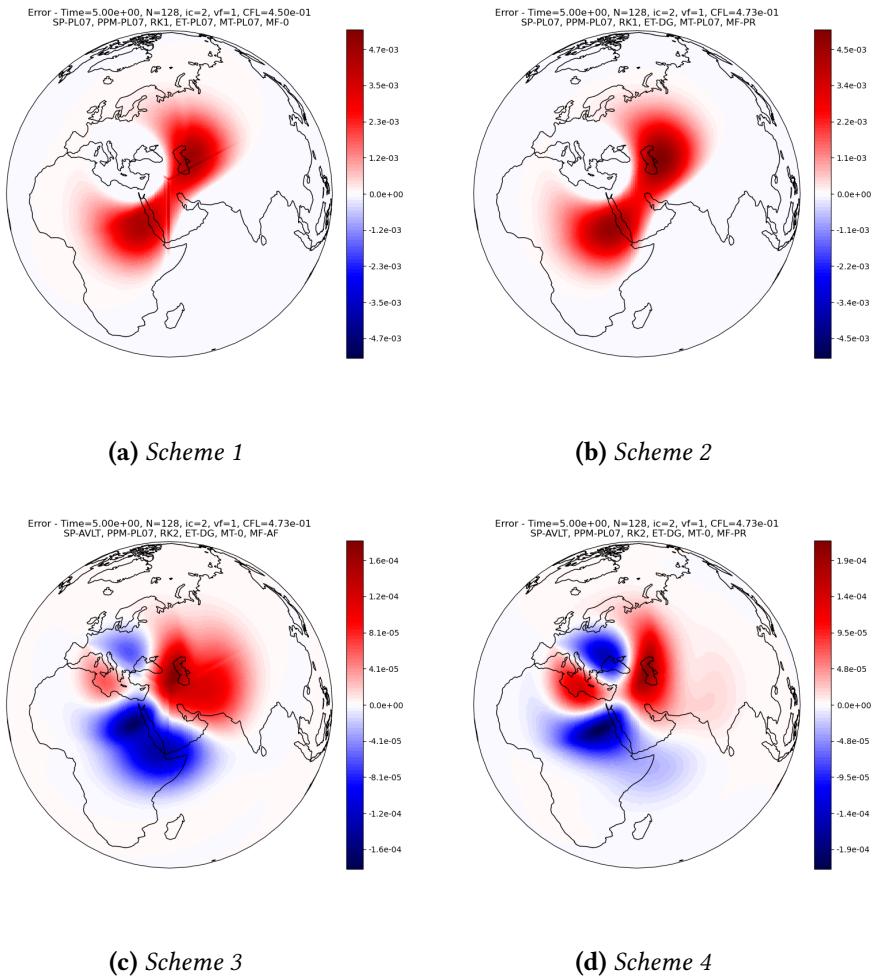
The third test case investigated focuses on IC3 with VF1. In this test, the exact solution remains constant and does not change over time.

Figure 5.5 illustrates the final errors obtained for all schemes listed in Table 5.2. Once again, schemes 1 and 2 demonstrate first-order accuracy, with scheme 2 being more accurate and producing no grid imprinting, as evident in Figures 5.8a and 5.8b. This further emphasizes the advantage of using Lagrange interpolation for ghost cells as opposed to the PL07 treatment. We also observe that scheme 3 exhibits second-order accuracy, while scheme 4 demonstrates third-order accuracy, surpassing its expected order. However, unlike the previous test, scheme 4 also produces grid imprinting, although much lesser compared to scheme 3.

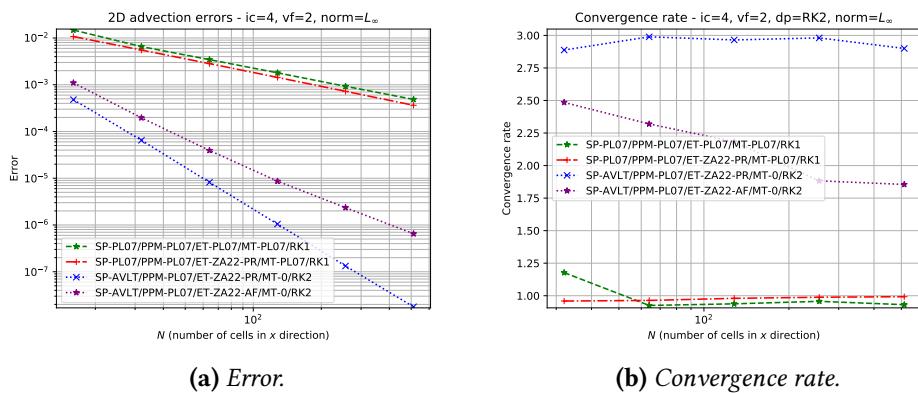
## 5.5 Concluding remarks

In this chapter, we extended the dimension splitting technique from the plane to the cubed-sphere. There were two essential differences in this case: 1) the presence of the metric tensor, and 2) the treatment of the ghost cells, which now depend on the neighboring panels.

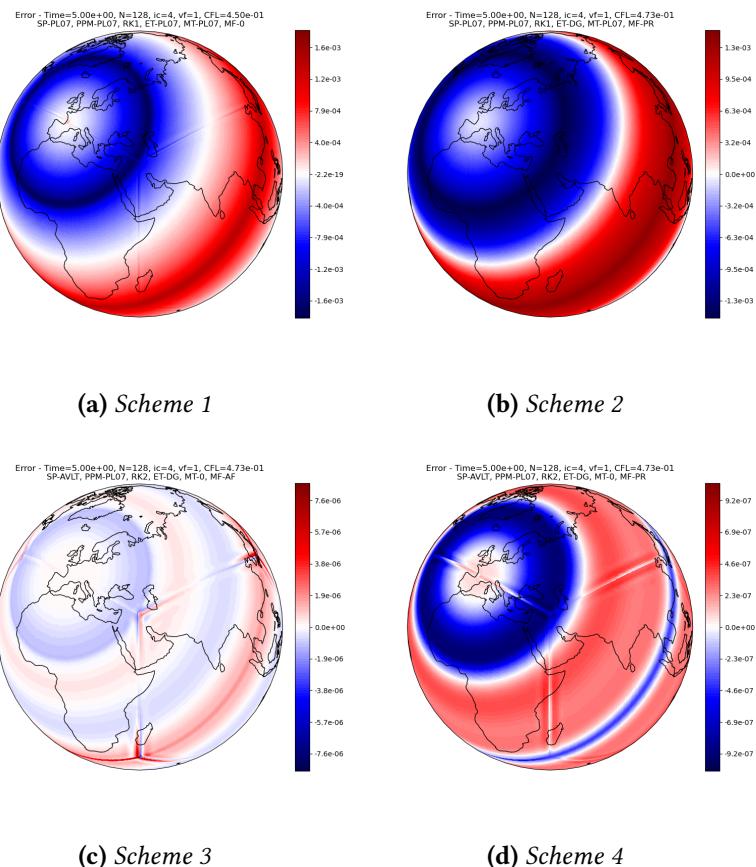
Regarding the metric tensor, we explored two different approaches. One approach, as used in Putman and Lin (2007), involved applying the metric tensor only after integrating the PPM parabolas. This approach introduced a first-order error, but it facilitated the



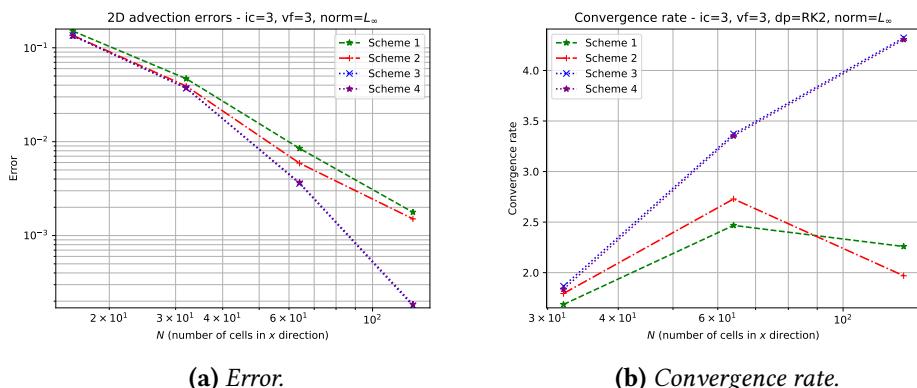
**Figure 5.6:** Error after 5 time units obtained for all schemes from table using the VF1 velocity field from table 5.2 with the initial condition IC1 from table 5.1 This figure considers a cubed-sphere with  $N = 128$ .



**Figure 5.7:** Convergence of the error for the schemes from table 5.3 obtained when computing the advection of IC3 with the rotated zonal wind VF1 after 5 time units.

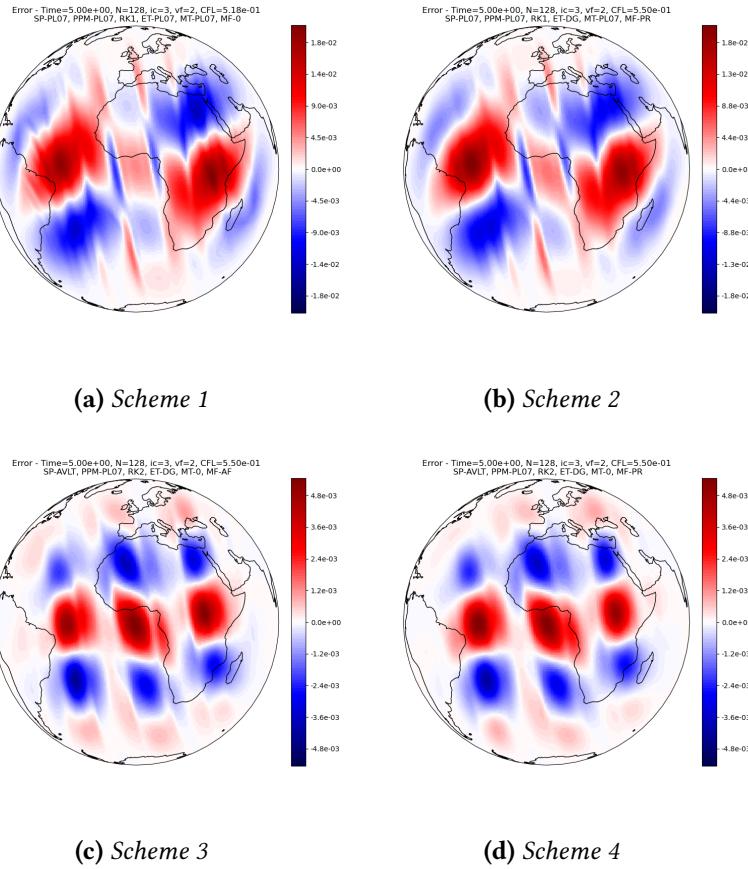


**Figure 5.8:** Error after 5 time units obtained for all schemes from table using the VF1 velocity field from table 5.2 with the initial condition IC3 from table 5.1 This figure considers a cubed-sphere with  $N = 128$ .

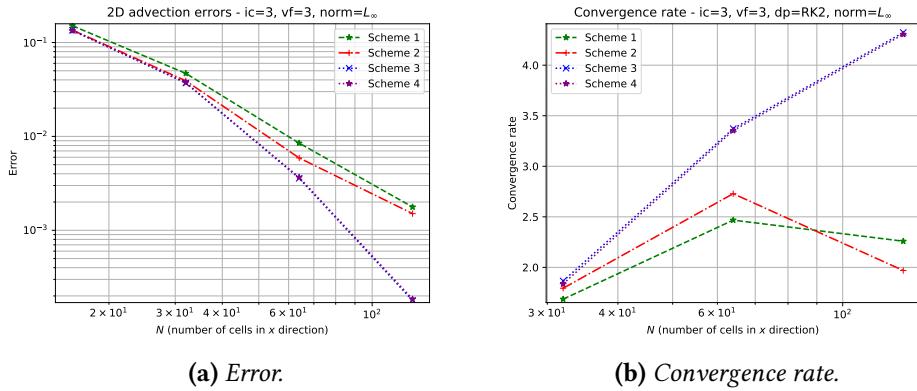


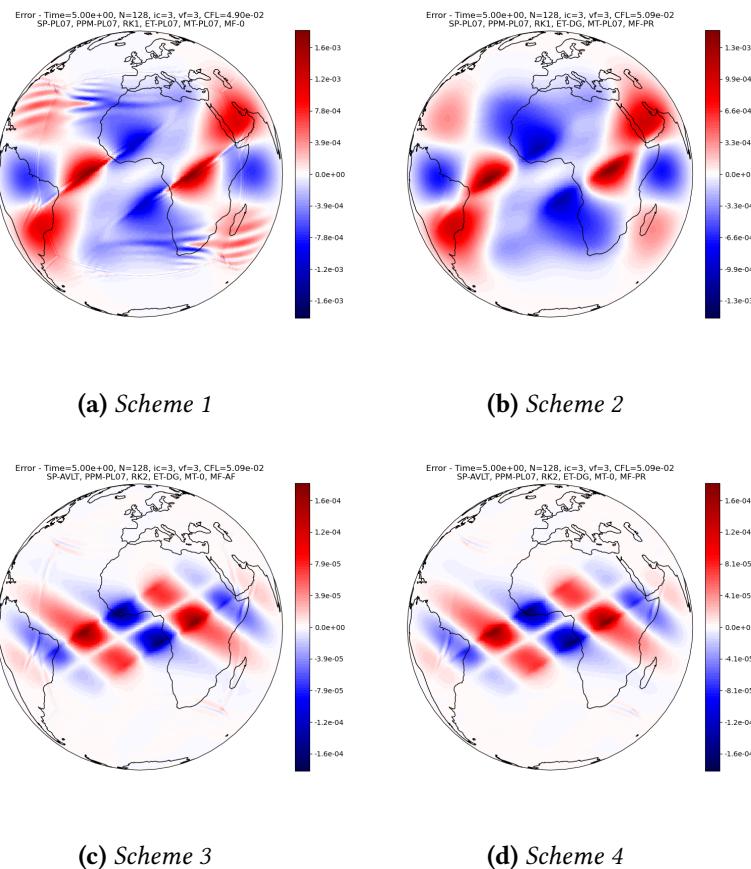
**Figure 5.9:** Convergence of the error for the schemes from table 5.3 obtained when computing the advection of IC2 with the rotated zonal wind VF3 after 5 time units.

elimination of the splitting error for the PL07 scheme. The second approach incorporated the metric tensor in the PPM reconstruction and introduced no first-order error, which was crucial for achieving a second-order scheme.



**Figure 5.10:** Error after 5 time units obtained for all schemes from table using the VF3 velocity field from table 5.2 with the initial condition IC2 from table 5.1 This figure considers a cubed-sphere with  $N = 128$ .





**Figure 5.12:** Error after 5 time units obtained for all schemes from table using the VF3 velocity field from table 5.2 with the initial condition IC2 from table 5.1 This figure considers a cubed-sphere with  $N = 128$ .

the main challenge with ET-DG was devising a strategy to ensure mass conservation at the cubed-edges.

We noticed that employing an averaging flux approach at the cube-edges led to a reduction in the scheme order and resulted in grid imprinting. As an alternative, we explored a second approach based on projecting the discrete divergence onto the space of grid functions with zero mass. This approach proved to be much more accurate than the averaging flux method. However, in a steady test case, we observed that the divergence still caused a minor grid-imprinting issue.



# Chapter 6

## Conclusions

In this report, we explore the development of finite volume methods from one dimension to two dimensions, examining their application on both the plane and the cubed-sphere. We elucidate that on the cubed-sphere, the finite volume approach for solving the advection equation simplifies to solving the two-dimensional advection equation independently on each panel, akin to the plane. However, it incorporates the inclusion of the metric tensor and boundary conditions that depend on the adjacent panels. The key disparity between methods on the plane and the sphere lies in the formulation of the metric tensor term and the computation of stencils at ghost cell positions.

In Chapter 2, we present an overview of 1D finite volume schemes in a Semi-Lagrangian perspective (FV-SL) for the advection equation. We discuss three essential tasks: function reconstruction, determination of control volume edges' departure points, and flux computation. For accurate reconstruction, we employed the PPM method and its variations to ensure monotonicity. Departure point calculation methods include RK1 and a two-stage Runge-Kutta scheme (RK2). Flux computation involves integrating the reconstructed function over a domain determined by the departure point. Numerical experiments demonstrate that a PPM variant, namely PPM-PL07 from Putman and Lin (2007), which utilizes fifth-order reconstruction, outperforms the original PPM (Colella & Woodward, 1984). Monotonic schemes, such as PPM-L04 and PPM-CW84, prevent overshoots, with PPM-L04 being more accurate. When subjected to variable velocity tests, the RK1 scheme exhibits a first-order error despite its third-order accuracy in space, while the RK2 scheme maintains third-order accuracy. Combining PPM-PL07 with RK2 ensures at least second-order accuracy.

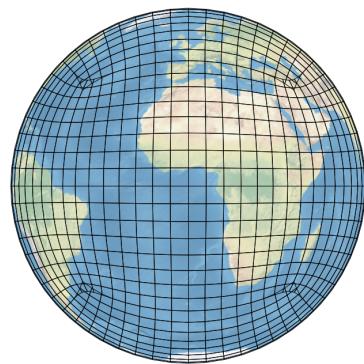
In Chapter 3, we introduce the dimension-splitting method as a cost-effective approach for solving the 2D advection equation. This method replaces the 2D problem with multiple 1D advection equations using a 1D finite volume scheme with slope limiters based on PPM. We modified the average of two Lie-Trotter splittings (AVLT) to ensure the preservation of a constant scalar field with a divergence-free velocity, resulting in the PL07 splitting proposed by Putman and Lin (2007) and the L04 splitting proposed by Lin (2004). Our conclusion from this chapter is that the PL07 splitting combined with the RK1 schemes yields the most accurate results but only first order, while the AVLT splitting achieves second-order when employed with the RK2 scheme.

In Chapter 4, we explored cubed-sphere mappings, particularly the equiangular mapping, which generates a more uniform grid. A key feature of equiangular cubed-sphere ghost cells is that their edge and center ghost points align with a common geodesic, alongside adjacent panel points. This property enables us to employ 1D Lagrange interpolation for obtaining field values in the ghost cells, as demonstrated through numerical examples. We compared this ghost cell treatment, referred to as ET-DG, with an extrapolation method called ET-PL07 proposed by Putman and Lin (2007), as well as a scheme that disregards panel discontinuities. We observed that among these methods, only ET-DG, utilizing edge reconstruction from PPM, avoided the creation of grid-imprinting patterns. The PL07 scheme exhibited second-order accuracy but generated significant grid imprinting patterns.

In Chapter 5, we extended the dimension splitting technique from Chapter 3 to each cubed-sphere panel by incorporating the edge treatment schemes discussed in Chapter 4. We highlighted that the ET-DG scheme breaks mass conservation and proposed two approaches to address this issue: one involving averaging the flux at the cube edges (MF-AF), and the other involving projecting the divergence onto the space of grid functions with zero mass (MF-PR). Numerical experiments revealed that MF-AF reduced the order near the edges and produced a grid imprinting pattern, whereas MF-PR preserved the scheme order and eliminated any grid imprinting pattern. We also showed theoretically that the MF-AF reduces the local truncation error by one order at the cube interfaces, leading to the observed grid imprinting. Further, we also proved that the MF-PR preserves the order of the original divergence. Additionally, we investigated two formulations of the metric tensor in the flux, one with its incorporation in the PPM reconstruction (MT-0) and the other without it (MT-PL07). For experiments with the metric tensor formulation MT-PL07, we employed the PL07 splitting with RK1 and utilized the edge treatments ET-DG (with MF-PR) and ET-PL07. It became evident that ET-DG and MF-PR improved the accuracy of the method and reduced grid imprinting. However, in the steady test case, we notice that the MF-PR may still generate a small grid-imprinting.

## 6.1 Future work

The next steps of this work can be divided into two main aspects. Firstly, we propose to expand upon the analysis presented in Chapter 5, where our objective is to implement the FV3 shallow-water core on the cubed-sphere based on the approach described by Lin and Rood (1997). The computational implementation of this step is expected to be relatively straightforward, as the shallow-water model already includes flux operators that have been implemented using standard finite-difference operators. Our goal is to assess the accuracy of this scheme, investigate grid imprinting at the edges, and explore potential methods for enhancing the performance of the scheme.



**Figure 6.1:** Overlapping cubed-sphere.

In our second proposal, we aim to extend the FV3 shallow-water core to the overlapping cubed-sphere grid, a grid proposed by Purser (2017), shown in Figure 6.1. This grid offers several advantages, including an orthogonal coordinate system and an abundance of cells at the cubed-edge, which enables the utilization of neighboring panel information without interpolation or extrapolation, potentially reducing grid imprinting. Also, the overlapping region is very small. The main challenge lies in effectively dealing with the overlapped region of the grid. To tackle this issue, we plan to explore methods used in the Yin-Yang grids, which also have overlapping regions and are currently implemented in the Global Environmental Multiscale (GEM) model by Canada (Husain et al., 2019; Qaddouri & Lee, 2011). Additionally, we intend to investigate existing methods developed for overlapping grids, which have been widely employed in the finite-volume literature for representing complex geometries (see for instance Hadzic (2005)). This research will aid the development of effective techniques for overlapping cubed-sphere grids.



# Appendix A

## Numerical Analysis

### A.1 Finite-difference estimates

This Section aims to prove all finite-difference error estimations used throughout this appendix. 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:*

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

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  $F$ :

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

Thus:

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

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

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

and

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

$$\theta_{2h} \in [x_0, x_0 + 2h], \quad \theta_{-2h} \in [x_0 - 2h, x_0].$$

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

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

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

Subtracting Equation (A.6) from Equation (A.5) we get the desired Equation (A.1) with

$$C_1 = \frac{1}{720} \left( 3F^{(5)}(\theta_h) + 3F^{(5)}(\theta_{-h}) - 16F^{(5)}(\theta_{2h}) - 16F^{(5)}(\theta_{-2h}) \right), \quad (\text{A.7})$$

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

$$C_1 = \frac{1}{720} \left( 6F^{(5)}(\eta_1) - 32F^{(5)}(\eta_2) \right), \quad (\text{A.8})$$

where  $\eta_1, \eta_2 \in [x_0 - 2h, x_0 + 2h]$ , which concludes the proof.  $\square$

**Lemma A.2.** *Let  $F \in C^4(\mathbb{R})$ ,  $x_0 \in \mathbb{R}$  and  $h > 0$ . Then, the following identity holds:*

$$F''(x_0) = \frac{-2F(x_0 - 2h) + 15F(x_0 - h) - 28F(x_0) + 20F(x_0 + h) - 6F(x_0 + 2h) + F(x_0 + 3h)}{6h^2} + C_2 h^2, \quad (\text{A.9})$$

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:

$$C_2 = \frac{1}{24} \left( 32F^{(4)}(\theta_{-2h}) - 15F^{(4)}(\theta_{-h}) - 20F^{(4)}(\theta_h) + 96F^{(4)}(\theta_{2h}) - 81F^{(4)}(\theta_{3h}) \right). \quad (\text{A.10})$$

Using the intermediate value theorem, we can express  $C_2$  in a more compact way as

$$C_2 = \frac{1}{24} \left( 128F^{(5)}(\eta_1) - 116F^{(5)}(\eta_2) \right), \quad (\text{A.11})$$

where  $\eta_1, \eta_2 \in [x_0 - 2h, x_0 + 3h]$ , which concludes the proof.  $\square$

**Lemma A.3.** Let  $F \in C^4(\mathbb{R})$ ,  $x_0 \in \mathbb{R}$  and  $h > 0$ . Then, the following identity holds:

$$F^{(3)}(x_0) = \frac{F(x_0 - 2h) - 7F(x_0 - h) + 16F(x_0) - 16F(x_0 + h) + 7F(x_0 + 2h) - F(x_0 + 3h)}{2h^3} + C_3 h, \quad (\text{A.12})$$

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:

$$F(x_0 - 2h) - 7F(x_0 - h) + 16F(x_0) - 16F(x_0 + h) + 7F(x_0 + 2h) - F(x_0 + 3h) = 2F^{(3)}(x_0)h^3 - 2C_3 h^4,$$

we get the desired Formula (A.12) with  $C_3$  given by:

$$C_3 = \frac{1}{48} \left( -16F^{(4)}(\theta_{-2h}) + 7F^{(4)}(\theta_{-h}) + 16F^{(4)}(\theta_h) - 112F^{(4)}(\theta_{2h}) + 81F^{(4)}(\theta_{3h}) \right). \quad (\text{A.13})$$

Using the intermediate value theorem, we can express  $C_3$  in a more compact way as

$$C_3 = \frac{1}{48} \left( 104F^{(5)}(\eta_1) - 128F^{(5)}(\eta_2) \right), \quad (\text{A.14})$$

where  $\eta_1, \eta_2 \in [x_0 - 2h, x_0 + 3h]$ , which concludes the proof.  $\square$

## A.2 Lagrange interpolation

Given real numbers, called nodes,  $x_0 < x_1 < \dots < x_m$ , 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  $\delta_{kj}$  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, there is  $\xi$  in the smallest interval containing  $x_0, \dots, x_m, x$  such that:*

$$f(x) - P_m(x) = \omega(x) \frac{f^{(m+1)}(\xi)}{(m+1)!}, \quad (\text{A.15})$$

where  $\omega(x) = (x - x_0)(x - x_1) \dots (x - x_m)$ .

*Proof.* See Stoer and Bulirsch (2002, Theorem 2.1.4.1. on p. 49).  $\square$

## 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).  $\square$

### A.3.1 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.3.** *If  $f \in C^2([x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}])$ , then*

$$\frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x) dx - f(x_i) = C_1 \Delta x^2, \quad (\text{A.16})$$

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:

$$f(x) = f(x_i) + f'(x_i)(x - x_i) + f''(\xi) \frac{(x - x_i)^2}{2}, \quad (\text{A.17})$$

for some  $\xi$  between  $x$  and  $x_i$ . Therefore:

$$\begin{aligned} \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x) dx - f(x_i) &= \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left( f'(x_i)(x - x_i) + f''(\xi) \frac{(x - x_i)^2}{2} \right) dx \\ &= \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f''(\xi) \frac{(x - x_i)^2}{2} dx. \end{aligned}$$

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

$$C_1 = \frac{1}{24} f''(\eta). \quad (\text{A.18})$$

□

**Theorem A.4.** *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*

$$\left| \frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(x, y) dx dy - f(x_i, y_j) \right| \leq C_1 \Delta x^2 + C_2 \Delta x \Delta y + C_3 \Delta y^2, \quad (\text{A.19})$$

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:

$$\begin{aligned} f(x, y) &= f(x_i, y_j) + \frac{\partial f}{\partial x}(x_i, y_j)(x - x_i) + \frac{\partial f}{\partial y}(x_i, y_j)(y - y_j) \\ &\quad + \frac{1}{2} \left( \frac{\partial^2 f}{\partial x^2}(\xi, \theta)(x - x_i)^2 + 2 \frac{\partial^2 f}{\partial x \partial y}(\xi, \theta)(x - x_i)(y - y_j) \frac{\partial^2 f}{\partial y^2}(\xi, \theta)(y - y_j)^2 \right) \end{aligned}$$

for some  $\xi$  between  $x$  and  $x_i$ , and  $\theta$  between  $y$  and  $y_j$ . Therefore:

$$\begin{aligned} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(x, y) dy dx - \Delta x \Delta y f(x_i, y_j) &= \frac{1}{2} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial^2 f}{\partial x^2}(\xi, \theta)(x - x_i)^2 dy dx + \\ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial^2 f}{\partial x \partial y}(\xi, \theta)(x - x_i)(y - y_j) dy dx + \frac{1}{2} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial^2 f}{\partial y^2}(\xi, \theta)(y - y_j)^2 dy dx \end{aligned}$$

Using the mean value theorem for integrals (see Theorem A.2), we have:

$$\begin{aligned} \frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(x, y) dx dy - f(x_i, y_j) &= \frac{\partial^2 f}{\partial x^2}(\eta_1, \lambda_1) \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{(x - x_i)^2}{2 \Delta x \Delta y} dx dy \\ &\quad + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial^2 f}{\partial x \partial y}(\xi, \theta) \frac{(x - x_i)(y - y_j)}{\Delta x \Delta y} dx dy + \frac{\partial^2 f}{\partial x^2}(\eta_2, \lambda_2) \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{(y - y_j)^2}{2 \Delta x \Delta y} dx dy \\ &= \frac{\partial^2 f}{\partial x^2}(\eta_1, \lambda_1) \frac{\Delta x^2}{24} + \frac{\partial^2 f}{\partial y^2}(\eta_2, \lambda_2) \frac{\Delta y^2}{24} + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial^2 f}{\partial x \partial y}(\xi, \theta) \frac{(x - x_i)(y - y_j)}{\Delta x \Delta y} dx dy \end{aligned}$$

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.  $\square$

### A.3.2 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  $k = 0, \dots, m$ . This kind of problem arises, for instance, when we are interested in computing departure points as in Equation 2.11. We can estimate the desired integral by computing the interpolating polynomial of  $f(t_{n-k})$ , for  $k = 0, \dots, m$  and then integrating this polynomial. This approach is exactly what is used in multi-step Adams-Basforth methods. On the next theorem, we give an expression the error of this approach.

**Theorem A.5.** If  $f \in C^{m+1}([0, T])$ ,  $t^n = n\Delta t$ ,  $n = 0, \dots, N_T$ ,  $\Delta t = \frac{T}{N_T}$  for some  $N_T \in \mathbb{N}$ , then:

$$\int_{t^n}^{t^{n+1}} f(t) dt = \Delta t \sum_{k=0}^m \left( \int_0^1 L_k(s) ds \right) f(t_{n-k}) + \frac{(\Delta t)^{k+1}}{(m+1)!} f^{(m+1)}(\eta) \int_0^1 \omega(s) ds, \quad (\text{A.20})$$

where  $\omega(s) = s(s+1) \cdots (s+m)$ ,  $\eta \in [t^{n-m}, t^n]$ .

*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  $k = -1, 0, \dots, m$ . Hence, we can write:

$$\int_{t^n}^{t^{n+1}} f(t) dt = \Delta t \int_0^1 f(\theta(s)) ds = \Delta t \int_0^1 g(s) ds. \quad (\text{A.21})$$

Defining the nodes  $s_k = -k$  for  $k = 0, \dots, m$ , it follows from Theorem A.1 that the interpolating polynomial  $P_m$  of  $g(s_k)$  satisfies:

$$g(s) - P_m(s) = \omega(s) \frac{g^{(m+1)}(\xi)}{(m+1)!}, \quad (\text{A.22})$$

where  $\xi \in [-m, 1]$ . Substituting Equation (A.22) in Equation (A.21), we obtain

$$\int_{t^n}^{t^{n+1}} f(t) dt = \Delta t \sum_{k=0}^m \left( \int_0^1 L_k(s) ds \right) g(-k) + \frac{\Delta t}{(m+1)!} \int_0^1 g^{(m+1)}(\xi) \omega(s) ds. \quad (\text{A.23})$$

Since  $\omega(s)$  does not change its sign in  $[0, 1]$  it follows from Theorem A.2 that:

$$\int_{t^n}^{t^{n+1}} f(t) dt = \Delta t \sum_{k=0}^m \left( \int_0^1 L_k(s) ds \right) g(-k) + \frac{\Delta t}{(m+1)!} g^{(m+1)}(\bar{\xi}) \int_0^1 \omega(s) ds, \quad (\text{A.24})$$

for some  $\bar{\xi} \in [-m, 1]$ . Notice that by the chain rule we get  $g^{(m+1)}(s) = (\Delta t)^k f^{(m+1)}(\theta(s))$ , therefore Equation (A.24) in terms of  $f$  reads:

$$\int_{t^n}^{t^{n+1}} f(t) dt = \Delta t \sum_{k=0}^m \left( \int_0^1 L_k(s) ds \right) f(t_{n-k}) + \frac{(\Delta t)^{k+1}}{(m+1)!} f^{(m+1)}(\eta) \int_0^1 \omega(s) ds, \quad (\text{A.25})$$

where  $\eta \in [t^{n-m}, t^n]$ , which is the desired identity.  $\square$

In the following corollaries, we give the explicit formulas for Equation (A.25) 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  $f \in C^1([0, T])$ ,  $t^n = n\Delta t$ ,  $n = 0, \dots, N_T$ ,  $\Delta t = \frac{T}{N_T}$  for some  $N_T \in \mathbb{N}$ , then:*

$$\int_{t^n}^{t^{n+1}} f(t) dt = \Delta t f(t_n) + \frac{\Delta t^2}{2} f'(\bar{t}), \quad (\text{A.26})$$

for some  $\bar{t} \in [t^n, t^{n+1}]$ .

**Corollary A.2.** *If  $f \in C^2([0, T])$ ,  $t^n = n\Delta t$ ,  $n = 0, \dots, N_T$ ,  $\Delta t = \frac{T}{N_T}$  for some  $N_T \in \mathbb{N}$ , then:*

$$\int_{t^n}^{t^{n+1}} f(t) dt = \frac{\Delta t}{2} (3f(t_n) - f(t_{n-1})) + \frac{5\Delta t^3}{12} f^{(2)}(\bar{t}), \quad (\text{A.27})$$

for some  $\bar{t} \in [t^{n-1}, t^{n+1}]$ .

**Corollary A.3.** If  $f \in C^3([0, T])$ ,  $t^n = n\Delta t$ ,  $n = 0, \dots, N_T$ ,  $\Delta t = \frac{T}{N_T}$  for some  $N_T \in \mathbb{N}$ , then:

$$\int_{t^n}^{t^{n+1}} f(t) dt = \frac{\Delta t}{12}(23f(t_n) - 16f(t_{n-1}) + 5f(t_{n-2})) + \frac{3\Delta t^4}{8}f^{(3)}(\bar{t}), \quad (\text{A.28})$$

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.4 PPM reconstruction accuracy analysis

In this Section, we are going to investigate the accuracy of the PPM reconstruction process. As we pointed out in Section 2.4.1, the approximation of  $q$  at the control volumes edges given by Equation (2.46) is fourth-order accurate when  $q \in C^4(\mathbb{R})$ . This is proved as a Corollary of the following Proposition A.1.

**Proposition A.1.** Let  $q \in C^4(\mathbb{R})$ ,  $\bar{x} \in \mathbb{R}$  and  $h > 0$ . Then, the following identity holds:

$$q(\bar{x}) = \frac{7}{12} \left( \frac{1}{h} \int_{\bar{x}}^{\bar{x}+h} q(x) dx + \frac{1}{h} \int_{\bar{x}-h}^{\bar{x}} q(x) dx \right) - \frac{1}{12} \left( \frac{1}{h} \int_{\bar{x}+h}^{\bar{x}+2h} q(x) dx + \frac{1}{h} \int_{\bar{x}-2h}^{\bar{x}-h} q(x) dx \right) + C_1 h^4, \quad (\text{A.29})$$

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 (2.37). It follows that:

$$\begin{aligned} \int_{\bar{x}}^{\bar{x}+h} q(\xi) d\xi + \int_{\bar{x}-h}^{\bar{x}} q(\xi) d\xi &= Q(\bar{x} + h) - Q(\bar{x} - h), \\ \int_{\bar{x}+h}^{\bar{x}+2h} q(\xi) d\xi + \int_{\bar{x}-2h}^{\bar{x}-h} q(\xi) d\xi &= Q(\bar{x} + 2h) - Q(\bar{x} - 2h) - (Q(\bar{x} + h) - Q(\bar{x} - h)). \end{aligned}$$

Using these identities, Equation (A.29) may be rewritten as:

$$q(\bar{x}) = \frac{4}{3} \left( \frac{Q(\bar{x} + h) - Q(\bar{x} - h)}{2h} \right) - \frac{1}{3} \left( \frac{Q(\bar{x} + 2h) - Q(\bar{x} - 2h)}{4h} \right) + C_1 h^4, \quad (\text{A.30})$$

which consists of finite-difference approximations. Thus, Equation (A.29) follows from Lemma A.1 with:

$$C_1 = C_1(\mu_1, \mu_2) = \frac{1}{720} \left( 6q^{(4)}(\mu_1) - 32q^{(4)}(\mu_2) \right), \quad (\text{A.31})$$

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

**Corollary A.4.** It follows from Proposition A.1 with  $\bar{x} = x_{i+\frac{1}{2}}$  and  $h = \Delta x$  that  $q_{i+\frac{1}{2}}$  given by

Equation (2.46) satisfies:

$$q(x_{i+\frac{1}{2}}) - q_{i+\frac{1}{2}} = C_1 \Delta x^4, \quad (\text{A.32})$$

with  $C_1$  given by Equation (A.31), whenever  $q \in C^4(\mathbb{R})$ .

**Remark A.1.** Similarly, one can show that the formulas are given by Equation (2.49) and Equation (2.49) are fifth-order accurate.

The parabolic function from (2.40) 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 (2.40) as:

$$q_i(x; Q) = q_{L,i} + \frac{(\Delta q_i + q_{6,i})}{\Delta x} (x - x_{i-\frac{1}{2}}) - \frac{q_{6,i}}{\Delta x^2} (x - x_{i-\frac{1}{2}})^2 \quad (\text{A.33})$$

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

$$q(x) = q(x_{i-\frac{1}{2}}) + q'(x_{i-\frac{1}{2}})(x - x_{i-\frac{1}{2}}) + \frac{q''(x_{i-\frac{1}{2}})}{2}(x - x_{i-\frac{1}{2}})^2 + \frac{q^{(3)}(\theta_i)}{6}(x - x_{i-\frac{1}{2}})^3, \quad (\text{A.34})$$

where  $\theta_i \in X_i$ . Comparing Equation (A.33) with Equation (A.34), it is reasonable to seek to some bound to the expressions:

$$q'(x_{i-\frac{1}{2}}) - \frac{(\Delta q_i + q_{6,i})}{\Delta x}, \quad (\text{A.35})$$

and:

$$\frac{q''(x_{i-\frac{1}{2}})}{2} - \left( -\frac{q_{6,i}}{\Delta x^2} \right). \quad (\text{A.36})$$

We have seen that term  $q_{L,i}$  gives a fourth-order approximation to  $q(x_{i-\frac{1}{2}})$ . The Corollary A.5 shall prove that the term (A.35) has a bound proportional to  $\Delta x^2$ , and the Corollary A.6 shall prove that the term (A.36) is bounded by a constant times  $\Delta x$ .

Before proving the desired bounds, it is useful to rewrite some terms explicitly as functions of the values of the  $\Delta x$ -grid function  $Q$ . Combining Equation (2.43) with Equations (2.47) and (2.48), we may write  $q_{6,i}$  as:

$$q_{6,i} = \frac{1}{4} \left( Q_{i-2} - 6Q_{i-1} + 10Q_i - 6Q_{i+1} + Q_{i+2} \right). \quad (\text{A.37})$$

Recalling the definition of  $\Delta q_i$  from Equation (2.41), and applying Equations (2.47) and (2.48), we may express  $\Delta q_i$  as:

$$\Delta q_i = \frac{1}{12} \left( Q_{i-2} - 8Q_{i-1} + 8Q_{i+1} - Q_{i+2} \right). \quad (\text{A.38})$$

Finally, we combine Equations (A.37) and (A.38) and write their sum as:

$$\frac{(\Delta q_i + q_{6,i})}{\Delta x} = \frac{2Q_{i-2} - 13Q_{i-1} + 15Q_i - 5Q_{i+1} + Q_{i+2}}{6\Delta x}. \quad (\text{A.39})$$

The next Proposition A.2 proves that Equation (A.39) approximates  $q'(x_{i-\frac{1}{2}})$  with order

2.

**Proposition A.2.** Let  $q \in C^3(\mathbb{R})$ ,  $\bar{x} \in \mathbb{R}$  and  $h > 0$ . Then, the following identity holds:

$$q'(\bar{x}) = \frac{1}{6h} \left( \frac{2}{h} \int_{\bar{x}-2h}^{\bar{x}-h} q(x) dx - \frac{13}{h} \int_{\bar{x}-h}^{\bar{x}} q(x) dx + \frac{15}{h} \int_{\bar{x}}^{\bar{x}+h} q(x) dx - \frac{5}{h} \int_{\bar{x}+h}^{\bar{x}+2h} q(x) dx + \frac{1}{h} \int_{\bar{x}+2h}^{\bar{x}+3h} q(x) dx \right) + C_2 h^2, \quad (\text{A.40})$$

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 (2.37). Like in Proposition A.2, we have:

$$\begin{aligned} & \frac{1}{6h} \left( \frac{2}{h} \int_{\bar{x}-2h}^{\bar{x}-h} q(x) dx - \frac{13}{h} \int_{\bar{x}-h}^{\bar{x}} q(x) dx + \frac{15}{h} \int_{\bar{x}}^{\bar{x}+h} q(x) dx - \frac{5}{h} \int_{\bar{x}+h}^{\bar{x}+2h} q(x) dx + \frac{1}{h} \int_{\bar{x}+2h}^{\bar{x}+3h} q(x) dx \right) \\ &= \frac{1}{6h} \left( \frac{2}{h} (Q(\bar{x} - h) - Q(\bar{x} - 2h)) - \frac{13}{h} (Q(\bar{x}) - Q(\bar{x} - h)) + \frac{15}{h} (Q(\bar{x} + h) - Q(\bar{x})) \right. \\ &\quad \left. - \frac{5}{h} (Q(\bar{x} + 2h) - Q(\bar{x} + h)) + \frac{1}{h} (Q(\bar{x} + 3h) - Q(\bar{x} + 2h)) \right) \\ &= \frac{1}{6h^2} \left( -2Q(\bar{x} - 2h) + 15Q(\bar{x} - h) - 28Q(\bar{x}) + 20Q(\bar{x} + h) - 6Q(\bar{x} + 2h) + Q(\bar{x} + 3h) \right), \end{aligned}$$

which consists of the finite-difference scheme from Lemma A.2. Therefore, Equation (A.40) follows from Lemma A.2 with:

$$C_2 = C_2(\mu_1, \mu_2) = \frac{1}{24} \left( 128q^{(3)}(\mu_1) - 116q^{(3)}(\mu_2) \right), \quad (\text{A.41})$$

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

**Corollary A.5.** It follows from Proposition A.2 with  $\bar{x} = x_{i-\frac{1}{2}}$  and  $h = \Delta x$  that  $\Delta q_i$  given by Equation (A.38) and  $q_{6,i}$  given by Equation (A.37) satisfy:

$$q'(x_{i-\frac{1}{2}}) - \frac{(\Delta q_i + q_{6,i})}{\Delta x} = C_2 \Delta x^2, \quad (\text{A.42})$$

with  $C_2$  given by Equation (A.41), whenever  $q \in C^3(\mathbb{R})$ .

Now, we analyse the following expression:

$$-\frac{2q_{6,i}}{\Delta x^2} = -\frac{1}{2\Delta x^2} \left( Q_{i-2} - 6Q_{i-1} + 10Q_i - 6Q_{i+1} + Q_{i+2} \right). \quad (\text{A.43})$$

deduced from Equation (A.37) and we prove in Proposition A.3 that Equation (A.43) approximates  $q''(x_{i-\frac{1}{2}})$  with order 1.

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

$$\begin{aligned} q''(\bar{x}) = \frac{1}{2h^2} \left( -\frac{1}{h} \int_{\bar{x}-2h}^{\bar{x}-h} q(x) dx + \frac{6}{h} \int_{\bar{x}-h}^{\bar{x}} q(x) dx - \frac{10}{h} \int_{\bar{x}}^{\bar{x}+h} q(x) dx \right. \\ \left. + \frac{6}{h} \int_{\bar{x}+h}^{\bar{x}+2h} q(x) dx - \frac{1}{h} \int_{\bar{x}+2h}^{\bar{x}+3h} q(x) dx \right) + C_3 h, \end{aligned} \quad (\text{A.44})$$

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

*Proof.* Similarly to Proposition A.2 using the same function  $Q$ , we have:

$$\begin{aligned} \frac{1}{2h^2} \left( -\frac{1}{h} \int_{\bar{x}-2h}^{\bar{x}-h} q(x) dx + \frac{6}{h} \int_{\bar{x}-h}^{\bar{x}} q(x) dx - \frac{10}{h} \int_{\bar{x}}^{\bar{x}+h} q(x) dx + \frac{6}{h} \int_{\bar{x}+h}^{\bar{x}+2h} q(x) dx - \frac{1}{h} \int_{\bar{x}+2h}^{\bar{x}+3h} q(x) dx \right) \\ = \frac{1}{2h^2} \left( -\frac{1}{h} (Q(\bar{x} - h) - Q(\bar{x} - 2h)) + \frac{6}{h} (Q(\bar{x}) - Q(\bar{x} - h)) - \frac{10}{h} (Q(\bar{x} + h) - Q(\bar{x})) \right. \\ \left. + \frac{6}{h} (Q(\bar{x} + 2h) - Q(\bar{x} + h)) - \frac{1}{h} (Q(\bar{x} + 3h) - Q(\bar{x} + 2h)) \right) \\ = \frac{1}{2h^3} \left( Q(\bar{x} - 2h) - 7Q(\bar{x} - h) + 16Q(\bar{x}) - 16Q(\bar{x} + h) + 7Q(\bar{x} + 2h) - Q(\bar{x} + 3h) \right), \end{aligned}$$

which consists of the finite-difference scheme from Lemma A.3. Therefore, Equation (A.44) follows from Lemma A.3 with:

$$C_3 = C_3(\mu_1, \mu_2) = \frac{1}{48} \left( 104q^{(3)}(\mu_1) - 128q^{(3)}(\mu_2) \right), \quad (\text{A.45})$$

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

**Corollary A.6.** It follows from Proposition A.3 with  $\bar{x} = x_{i-\frac{1}{2}}$  and  $h = \Delta x$  that  $q_{6,i}$  given by Equation (2.46) satisfies:

$$q''(x_{i-\frac{1}{2}}) - \left( -\frac{2q_{6,i}}{\Delta x^2} \right) = C_3 \Delta x, \quad (\text{A.46})$$

with  $C_3$  given by Equation (A.45), whenever  $q \in C^3(\mathbb{R})$ .

With the aid of Corollaries A.4, A.5, and A.6, we are able to prove that the PPM reconstruction approximates  $q$  with order 3. Indeed, we prove this on the follow up Proposition A.4.

**Proposition A.4.** Let  $q \in C^4([a, b])$ . Then, the Piecewise-Parabolic function given by Equation (2.40) with the parameters  $q_{R,i}$  and  $q_{L,i}$  obeying Equations (2.47) and (2.48) gives a third-order approximation to  $q$  on the control volume  $X_i$ . Namely, there exist constants  $M_1$  and  $M_2$  such that

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

*Proof.* For  $x \in X_i$ , from Equations (A.34) and (A.33), we have:

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

Using this fact with Corollaries A.4, A.5, and A.6, 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 (A.31), (A.41) and (A.45), respectively, and

$$C_4 = C_4(\theta_i) = \frac{q^{(3)}(\theta_i)}{6}. \quad (\text{A.47})$$

For  $x \in X_i$ , we have  $|x - x_{i-\frac{1}{2}}| \leq \Delta x$ , thus:

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

where

$$\begin{aligned} M_1 &= \frac{38}{720} \sup_{\xi \in [a,b]} |q^{(4)}(\xi)|, \\ M_2 &= \left( \frac{244}{24} + \frac{232}{96} + \frac{1}{6} \right) \sup_{\xi \in [a,b]} |q^{(3)}(\xi)| = \frac{143}{12} \sup_{\xi \in [a,b]} |q^{(3)}(\xi)|, \end{aligned}$$

which concludes the proof.  $\square$

**Remark A.2.** Replacing the formulas for  $q_{R,i}$  and  $q_{L,i}$  given by Equations (2.47) and (2.48) by the formulas given by Equations (2.49) and (2.50), does not change the order of convergence of the parabolic approximation.

## A.5 Convergence of 1D FV-SL schemes

### A.5.1 Consistency and convergence

Hereafter, we are going to use the notations introduced in Section 2.1.1. To move towards the convergence of 1D-FV schemes, for Problem 2.4 we introduce the local truncation error (LTE hereafter)  $\tau_i^n$  following LeVeque (2002):

$$Q_i(t^{n+1}) = Q_i(t^n) - \lambda \left( F_{i+\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i+\frac{1}{2}}^n) - F_{i-\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i-\frac{1}{2}}^n) \right) + \Delta t \tau_i^n. \quad (\text{A.48})$$

We define  $\tau^n \in \mathbb{P}_v^N$ , which represent the LTEs at the time-step  $n$ . Notice the LTE is obtained by replacing the exact solution in Equation (2.20). Since  $Q_i(t^n)$  is the exact solution

of Equation (2.9), the LTE may be rewritten as

$$\tau_i^n = \frac{1}{\Delta x} \left[ \left( \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} (uq)(x_{i+\frac{1}{2}}, t) dt - F_{i+\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i+\frac{1}{2}}^n) \right) + \left( \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} (uq)(x_{i-\frac{1}{2}}, t) dt - F_{i-\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i-\frac{1}{2}}^n) \right) \right]. \quad (\text{A.49})$$

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. Now we can define consistency.

**Definition A.1** (Consistency). *Let us consider the framework of Problem 2.4. A 1D-FV scheme is said to be consistent in the  $p$ -norm if for any sequence of  $(\Delta x^{(k)}, \Delta t^{(k)}, \lambda)$ -discretizations,  $k \in \mathbb{N}$ , with  $\lim_{k \rightarrow \infty} \Delta x^{(k)} = \lim_{k \rightarrow \infty} \Delta t^{(k)} = 0$ , we have:*

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

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

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

From Equation (A.49), it follows that we basically need to ensure that the numerical flux function  $F_{i+\frac{1}{2}}^n$  converges to the time-averaged flux at edges when  $\Delta x \rightarrow 0$  in order to guarantee consistency. In Section 2.5 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:

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

and we define the vector of errors by  $E^n \in \mathbb{P}_v^N$  with entries  $E_i^n$ .

**Definition A.2** (Convergence). *Let us consider the framework of Problem 2.4. 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  $\lim_{k \rightarrow \infty} \Delta x^{(k)} = \lim_{k \rightarrow \infty} \Delta t^{(k)} = 0$ , we have:*

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

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

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

Subtracting Equation (2.20) from Equation (A.48) we get the following equation for

the error:

$$\begin{aligned} E_i^{n+1} = & E_i^n - \lambda \left[ \left( F_{i+\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i+\frac{1}{2}}^n) - F_{i+\frac{1}{2}}^n(Q^n, \tilde{u}_{i+\frac{1}{2}}^n) \right) \right. \\ & \left. - \left( F_{i-\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i-\frac{1}{2}}^n) - F_{i-\frac{1}{2}}^n(Q^n, \tilde{u}_{i-\frac{1}{2}}^n) \right) \right] + \tau_i^n \Delta t. \end{aligned} \quad (\text{A.50})$$

Notice that if  $q, u \in C^3$ , we can rewrite Equation (A.49) as:

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

Using the midpoint rule for integration (Theorem A.3) and the mean value theorem for integrals (Theorem A.2), we have:

$$\begin{aligned} \tau_i^n = & \left[ \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \left( \frac{\partial(uq)}{\partial x}(x_i, t) + \frac{\Delta x^2}{24} \frac{\partial^3(uq)}{\partial x^3}(\xi, t) \right) dt - \left( \frac{F_{i+\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i+\frac{1}{2}}^n) - F_{i-\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i-\frac{1}{2}}^n)}{\Delta x} \right) \right] \\ = & \left[ \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \frac{\partial(uq)}{\partial x}(x_i, t) dt - \left( \frac{F_{i+\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i+\frac{1}{2}}^n) - F_{i-\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i-\frac{1}{2}}^n)}{\Delta x} \right) \right] + \frac{\Delta x^2}{24} \frac{\partial^3(uq)}{\partial x^3}(\xi, \bar{t}), \end{aligned} \quad (\text{A.51})$$

for  $\xi \in X_i$  and  $\bar{t} \in [t^n, t^{n+1}]$ . Therefore, if  $q, u \in C^3$  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  $\frac{F_{i+\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i+\frac{1}{2}}^n) - F_{i-\frac{1}{2}}^n(Q(t^n), \tilde{u}_{i-\frac{1}{2}}^n)}{\Delta x}$ . 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.

## A.5.2 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 2.4, we define the operators  $\mathcal{H}_{\Delta x, n} : \mathbb{P}_v^N \rightarrow \mathbb{P}_v^N$  whose  $i$ -th entry is given by:

$$[\mathcal{H}_{\Delta x, n}(Q)]_i = Q_i - \lambda \left( F_{i+\frac{1}{2}}^n(Q, \tilde{u}_{i+\frac{1}{2}}^n) - F_{i-\frac{1}{2}}^n(Q, \tilde{u}_{i-\frac{1}{2}}^n) \right), \quad (\text{A.52})$$

for  $i = 1, \dots, N$ ,  $n = 0, \dots, N_T - 1$ . 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 (2.20) may be rewritten in a vector form by

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

and Equation (A.48) 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 (A.50) is given by

$$E^{n+1} = \mathcal{H}_{\Delta x, n}(Q(t^n)) - \mathcal{H}_{\Delta x, n}(Q^n) + \Delta t \tau^n. \quad (\text{A.53})$$

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 A.3** (Stability). *In the context of Problem 2.4, 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:*

$$\|\mathcal{H}_{\Delta x, n}(Q) - \mathcal{H}_{\Delta x, n}(P)\|_{p, \Delta x} \leq (1 + \alpha \Delta t) \|Q - P\|_{p, \Delta x}, \quad (\text{A.54})$$

for all  $Q, P \in \mathbb{R}_v^N$  and  $\alpha$  is a constant that does not depend neither on  $\Delta x$  nor on  $\Delta t$ .

Assuming that the scheme is stable in the  $p$ -norm, then it follows from Equation (A.53) that:

$$\begin{aligned} \|E^{n+1}\|_{p, \Delta x} &\leq \|\mathcal{H}_{\Delta x, n}(Q(t^n)) - \mathcal{H}_{\Delta x, n}(Q^n)\|_{p, \Delta x} + \Delta t \max_{n=1, \dots, N_T} \|\tau^n\|_{p, \Delta x} \\ &\leq (1 + \alpha \Delta t) \|E^n\|_{p, \Delta x} + \Delta t \max_{n=1, \dots, N_T} \|\tau^n\|_{p, \Delta x} \\ &\leq (1 + \alpha \Delta t)^n \|E^0\|_{p, \Delta x} + \Delta t \max_{n=1, \dots, N_T} \|\tau^n\|_{p, \Delta x} \sum_{k=0}^{n-1} (1 + \alpha \Delta t)^k \\ &\leq e^{\alpha T} (\|E^0\|_{p, \Delta x} + T \max_{n=1, \dots, N_T} \|\tau^n\|_{p, \Delta x}), \end{aligned} \quad (\text{A.55})$$

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 2.2. 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 2.5, 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  $\theta_i = i \frac{2\pi}{N}$ ,  $i = 1, \dots, N$ ,  $\Delta\theta = \frac{2\pi}{N}$ ,  $\theta = (\theta_1, \theta_2, \dots, \theta_N)$ . The imaginary unit is denoted by  $i$ . We define  $\mathbb{C}_v^N$  similarly as  $\mathbb{P}_v^N$ . The Fourier modes  $e^{ik\theta}] \in \mathbb{C}_v^N$  for  $k = 1, \dots, N$ ,

have entries are given by:

$$[e^{ik\theta}]_i = e^{ik\theta_i}, \quad \text{for } i = 1, \dots, N.$$

Each  $k$  is referred to wavenumber and  $\theta_k$  is called dimensionless wavenumber. The Fourier modes form an orthogonal basis of  $\mathbb{C}_v^N$  with respect to the inner product

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

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

$$Q = \sum_{k=1}^N a_k e^{ik\theta},$$

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:

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

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:

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

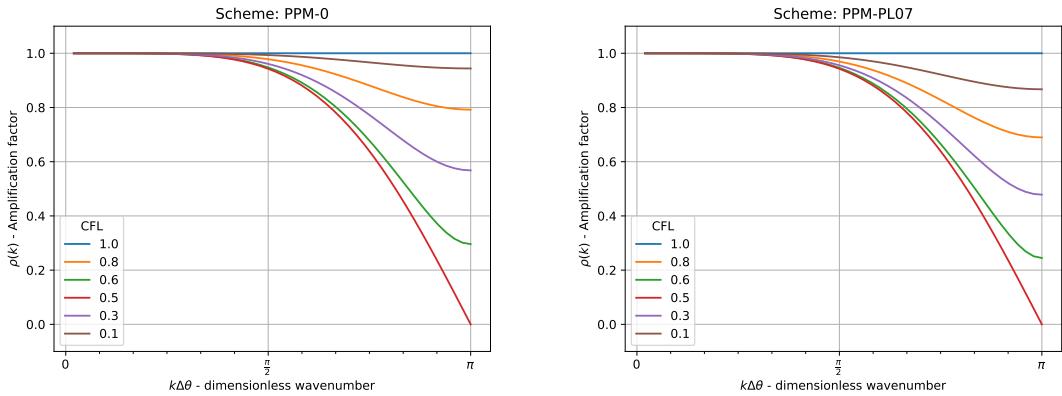
Therefore:

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

If we show that  $\max_{k=1,\dots,N} |\rho(k)| \leq 1 + \alpha \Delta t$ , with  $\alpha$  independent of  $\Delta t$ ,  $N$  and  $n$ , then we ensure the stability of  $\mathcal{H}_{\Delta x,n}$ .

### A.5.3 Flux accuracy analysis

With the PPM operator, we can compute the amplification factor by applying it on each Fourier mode considering 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 A.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. For an analysis of stability for larger time-steps, we refer to Lauritzen (2007).



**Figure A.1:** Amplification factor for the PPM (left) and hybrid PPM (right) schemes for different CFL numbers.

## A.6 Convergence, consistency and stability of 2D-FV schemes

The notions of convergence, consistency and stability for a 2D-FV schemes are straightforward from these notions for 1D-FV schemes (see Subsections A.5.1 and A.5.2). Indeed, in the context of Problem 3.3, we define the operators  $\mathcal{H}_{\Delta x, \Delta y, n} : \mathbb{R}^{N \times M} \rightarrow \mathbb{R}^{N \times M}$  whose  $(i, j)$  entry is given by:

$$[\mathcal{H}_{\Delta x, \Delta y, n}(Q)]_{ij} = Q_{ij} - \Delta t \mathbb{D}_{ij}^n$$

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

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

The local error truncation  $\tau^n \in \mathbb{R}^{N \times M}$  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

$$E^{n+1} = \mathcal{H}_{\Delta x, \Delta y, n}(Q(t^n)) - \mathcal{H}_{\Delta x, \Delta y, n}(Q^n) + \Delta t \tau^n. \quad (\text{A.56})$$

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

**Definition A.4.** A 2D-FV scheme is stable in the  $p$ -norm if

$$\|\mathcal{H}_{\Delta x, \Delta y, n}(Q) - \mathcal{H}_{\Delta x, \Delta y, n}(P)\|_{p, \Delta x \times \Delta y} \leq (1 + \alpha \Delta t) \|Q - P\|_{p, \Delta x \times \Delta y}, \quad (\text{A.57})$$

for all  $Q, P \in \mathbb{R}^{N \times M}$  and  $\alpha$  is a constant that does not depend neither on  $\Delta x$ ,  $\Delta y$ ,  $\Delta t$  nor on  $n$ .

If a 2D-FV scheme is stable in the  $p$ -norm, similarly to Equation (A.55) we have:

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

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 A.1 and convergence is defined as in Definition A.2.

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  $\theta_i = i \frac{2\pi}{N}$ ,  $i = 1, \dots, N$ ,  $\Delta\theta = \frac{2\pi}{N}$ ,  $\theta_i = (\theta_1, \theta_2, \dots, \theta_N)$ ,  $\phi_j = j \frac{2\pi}{M}$ ,  $j = 1, \dots, M$ ,  $\Delta\phi = \frac{2\pi}{M}$ ,  $\phi = (\phi_1, \phi_2, \dots, \phi_M)$ . For  $k_1 = 1, \dots, N$ ,  $k_2 = 1, \dots, M$ , the two-dimensional Fourier mode  $\mathbf{k} = (k_1, k_2)$  from  $\mathbb{C}^{N \times M}$  has its  $(i, j)$  entry given by  $[e^{i\mathbf{k}\theta}]_{ij} = e^{ik_1\theta_i} e^{ik_2\phi_j}$ . For an analysis of stability for the dimension splitting method, we refer to Lauritzen (2007) and Lin and Rood (1996).

Notice that if  $q, u, v \in C^3$ , we can rewrite the LTE as:

$$\tau_{ij}^n = \left[ \frac{1}{\Delta x \Delta y \Delta t} \int_{t^n}^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \nabla \cdot (\mathbf{u}q)(x, y, t) dy dx dt - \mathbb{D}_{ij}^n \right].$$

Using the midpoint rule for integration (Theorem A.4), the mean value theorem for integrals (Theorem A.2) and recalling the discrete divergence (Definition 3.5), we have:

$$\tau_{ij}^n = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \nabla \cdot (\mathbf{u}q)(x_i, y_j, t) dt - \mathbb{D}_{ij}^n + O(\Delta x^2) + O(\Delta y^2). \quad (\text{A.58})$$

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



# Appendix B

## Code availability

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, 9)
- 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](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](https://doi.org/10.1175/1520-0493(1990)118<0586:AOTPPM>2.0.CO;2) (cit. on pp. 4, 8, 18)
- Chen, C., & Xiao, F. (2008). Shallow water model on cubed-sphere by multi-moment finite volume method. *Journal of Computational Physics*, 227(10), 5019–5044. <https://doi.org/https://doi.org/10.1016/j.jcp.2008.01.033> (cit. on p. 71)
- 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 pp. 49, 57)
- 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. 23, 29, 44)
- 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](https://doi.org/https://doi.org/10.1016/0021-9991(84)90143-8) (cit. on pp. 4, 7, 17, 18, 20–23, 28, 85)
- 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. 93)

- 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. 49)
- 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. 38)
- 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)
- Diamantakis, M. (2014). The semi-lagrangian technique in atmospheric modelling: Current status and future challenges. *Seminar on Recent Developments in Numerical Methods for Atmosphere and Ocean Modelling*, 2-5 September 2013, 183–200 (cit. on p. 72).
- 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](https://doi.org/10.1007/978-3-642-11640-7_5). (Cit. on p. 18)
- 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](https://doi.org/10.1007/978-1-4419-6412-0_7). (Cit. on p. 18)
- Eliasen, E., Machenhauer, B., & Rasmussen, E. (1970). On a numerical method for integration of the hydrodynamical equations with 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. 8, 10, 20)
- 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, 20).
- 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. 18)
- Hadzic, H. (2005). *Development and application of finite volume method for the computation of flows around moving bodies on unstructured, overlapping grids* (Doctoral dissertation). Technische Universität Hamburg. Hamburg, Germany. (Cit. on p. 87).

## REFERENCES

- 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. 8, 22)
- 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)
- Holden, H., Karlsen, K., Lie, K.-A., & Risebro, H. (2010). *Splitting methods for partial differential equations with rough solutions: Analysis and matlab programs*. <https://doi.org/10.4171/078>. (Cit. on p. 38)
- Husain, S. Z., Girard, C., Qaddouri, A., & Plante, A. (2019). A new dynamical core of the global environmental multiscale (gem) model with a height-based terrain-following vertical coordinate. *Monthly Weather Review*, 147(7), 2555–2578. <https://doi.org/https://doi.org/10.1175/MWR-D-18-0438.1> (cit. on p. 87)
- 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. 38)
- Jung, J.-H., Konor, C. S., & Randall, D. (2019). Implementation of the vector vorticity dynamical core on cubed sphere for use in the quasi-3-d multiscale modeling framework. *Journal of Advances in Modeling Earth Systems*, 11(3), 560–577. <https://doi.org/https://doi.org/10.1029/2018MS001517> (cit. on p. 52)
- 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. 49)
- 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. 49)
- Kent, J., Melvin, T., & Wimmer, G. A. (2023). A mixed finite-element discretisation of the shallow-water equations. *Geoscientific Model Development*, 16(4), 1265–1276. <https://doi.org/10.5194/gmd-16-1265-2023> (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., Nair, R. D., & Ullrich, P. A. (2010). A conservative semi-lagrangian multi-tracer transport scheme (cslam) on the cubed-sphere grid. *Journal of Computational Physics*, 229(5), 1401–1424. <https://doi.org/https://doi.org/10.1016/j.jcp.2009.10.036> (cit. on p. 29)
- 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](https://doi.org/10.1007/978-3-642-11640-7_8). (Cit. on p. 8)
- Lauritzen, P. H. (2007). A stability analysis of finite-volume advection schemes permitting long time steps. *Monthly Weather Review*, 135(7), 2658–2673. <https://doi.org/https://doi.org/10.1175/MWR3425.1> (cit. on pp. 106, 107)

- Leonard, B. P., Lock, A. P., & MacVean, M. K. (1996). Conservative explicit unrestricted-time-step multidimensional constancy-preserving advection schemes. *Monthly Weather Review*, 124(11), 2588–2606. [https://doi.org/https://doi.org/10.1175/1520-0493\(1996\)124<2588:CEUTSM>2.0.CO;2](https://doi.org/https://doi.org/10.1175/1520-0493(1996)124<2588:CEUTSM>2.0.CO;2) (cit. on p. 7)
- LeVeque, R. J. (1985). A large time step generalization of godunov's method for systems of conservation laws. *SIAM Journal on Numerical Analysis*, 22(6), 1051–1073. <https://doi.org/10.1137/0722063> (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. 11, 43)
- LeVeque, R. J. (2002). *Finite volume methods for hyperbolic problems*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511791253>. (Cit. on pp. 11, 19, 101, 104)
- 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](https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2) (cit. on pp. 4, 8, 22–24, 28, 42, 74, 85)
- 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](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](https://doi.org/10.1175/1520-0493(1996)124<2046:MFFSLT>2.0.CO;2) (cit. on pp. 4, 7, 18, 23, 29, 37, 38, 41, 42, 49, 107)
- 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 pp. 4, 87)
- 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. 18)
- 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., Clinton, 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. 26, 44, 76)
- 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](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)

## REFERENCES

- 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, 58)
- Purser, J. (2017). Conformal localized-overset polyhedral global grids based on riemann surfaces. [https://wgne.net/bluebook/uploads/2017/docs/03\\_Purser\\_R.\\_James\\_global\\_grids.pdf](https://wgne.net/bluebook/uploads/2017/docs/03_Purser_R._James_global_grids.pdf). (Cit. on p. 87)
- 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](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/10.1016/j.jcp.2007.07.022> (cit. on pp. 4, 5, 8, 21, 28, 42, 49, 62, 66, 67, 74, 77, 79, 85, 86)
- Qaddouri, A., & Lee, V. (2011). The canadian global environmental multiscale model on the yin-yang grid system. *Quarterly Journal of the Royal Meteorological Society*, 137(660), 1913–1926. <https://doi.org/https://doi.org/10.1002/qj.873> (cit. on p. 87)
- 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. 49, 55)
- 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:SLPBF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1992)120<1394:SLPBF>2.0.CO;2) (cit. on p. 29)
- 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, 50)
- 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/AMSMONOGRAPHSD-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. 37)
- 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, 49, 52)
- Rossmannith, J. A. (2006). A wave propagation method for hyperbolic systems on the sphere. *Journal of Computational Physics*, 213(2), 629–658. <https://doi.org/https://doi.org/10.1016/j.jcp.2005.08.027> (cit. on pp. 63, 71)
- 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](https://doi.org/10.1175/1520-0493(1972)100<0136:CFAOTP>2.3.CO;2) (cit. on pp. 4, 49, 50, 62)

- 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. 19, 93)
- 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. 38)
- 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. 105)
- 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 p. 21)
- 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](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. 26, 104)
- 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. 18)
- 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)

## REFERENCES

- 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](https://doi.org/https://doi.org/10.1016/0021-9991(77)90095-X) (cit. on pp. 4, 7, 18, 20)
- 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/](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. 8)
- 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](https://doi.org/10.1016/S0021-9991(05)80016-6) (cit. on pp. 5, 61, 76)
- 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](https://doi.org/10.1007/978-94-009-4754-2_8). (Cit. on p. 8)
- 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. 57, 62)
- 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)