Department of Mathematics and Statistics

University of Reading

CONSERVATIVE MONOTONE ADVECTION SCHEMES ALLOWING LONG TIME STEPS

James Woodfield

A thesis submitted for the degree of Doctor of Philosophy

November 2022

0.1 Statement of Originality: Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged. James Woodfield

0.2 Acknowledgements

This research was supported by the Mathematics of Planet Earth Centre for Doctoral Training program.
I would firstly like to thank my primary supervisor Dr. Hilary Weller for her encouragement and guidance throughout my PhD. In particular for her insight in numerical methods, commitment to good scientific practice, and practical advice in coding and writing. I would secondly like to express my gratitude to my secondary supervisor Professor Colin Cotter for his thoughtful comments and encouragement.
I am thankful to the staff of the MPE CDT for organising seminars, hackathons, events, the MOTR placement, training and funding through EPSRC. As well as grateful to the wider MPE CDT community.
I would like to thank my family for supporting me and I would like to thank Ruby for her encouragement and support.

Abstract

Numerical weather prediction models are simulating more passive and active tracers than anytime in history. In the next generation of dynamical cores unconditionally monotone, stable, locally mass preserving algorithms are sought. Due to the importance of local mass conservation, the historically dominant semi-Lagrangian method may see competitors from the Eulerian frame. This thesis is concerned with the development of slope and flux limiters for dynamical core advection algorithms.
The first investigation in this thesis regards the use and development of one-dimensional limiters applied to the multidimensional advection equation on a uniform grid. Two new limiter regions are derived which are sufficient for the numerical solution of the incompressible advection to retain a local maximum principle.
The second investigation in this thesis regards the use and development of truly unstructured multidimensional slope limiters suitable for a wider class of schemes and meshes. A general theory is presented, with two limiters introduced capable of preserving different local maximum principles. We then illustrate two practical examples of how this theory can be applied. The first example introduces the limiters and how slight improvements on state-of-the-art limiters can be achieved for second order finite volume methods. The second example illustrates the true use of the theory by introducing a new fourth order finite volume scheme, and how the limiting procedures can be used for discrete maximum principles.
The third contribution of this thesis consists of a numerical study in implicit linearised slope limiters and the use of implicit flux corrected transport. This approach concerns how one can achieve stability and monotonicity at large Courant numbers, whilst still retaining monotonicity and accuracy at low Courant numbers. A simple one stage flux corrected transport scheme on two implicit methods emerges as a robust method achieving many of the desired properties for a dynamical core advection algorithm.

Contents

Chapter 1

Introduction

Fukushima, Chernobyl, and above ground nuclear tests in America have highlighted the importance of tracking radioactive isotopes in the atmosphere including Cs137, Cs134, Sr90. Volcanic eruptions such as Eyjafjallajökull, highlight the importance of tracking ash in the atmosphere. The Kyoto protocol saw the importance of tracking environmentally important tracers such as CFCs, CH4, CO2, O3. The tracking of inert tracers such as SF6, CFC-11, CFC-12 have been found to be crucial in studying the flow structures in the ocean [1]. Numerical weather prediction models themselves require the modelling of active and passive tracers, such as temperature and moisture for accurate and stable weather predictions. There are an increasing number of desired properties required on the next generation of dynamical core advection algorithms [2], achieving some of these requirements is the subject of this thesis.

Advection is the transport of a quantity by the motion of a fluid and is an important aspect of weather and climate models. However, "Modelling of highly advective transport is embarrassingly difficult, even in the superficially simple case of one-dimensional constant-velocity flow" B.P. Leonard 1991 [3].

There is evidence that the choice of dynamical core algorithms fundamentally affects the simulation and prediction capabilities of operational weather centres [4]. The dynamical core is typically one of the most costly parts of a weather prediction model and the advection scheme is typically one of the most costly parts of a dynamical core. This varies model to model, but for example in the Met Office NERC Cloud model 70 percent of the total runtime is spent in the dynamical core, and 50 percent of the total runtime is spent performing advection [5].

1.1 Motivation

The need to go beyond semi-Lagrangian

The Semi-Lagrangian method is widely used in numerical weather prediction models [6]. The prevalence of the Semi-Lagrangian method can be in part credited to the increase in allowable time-step without decreasing accuracy or increasing computational cost, [7], [8], theoretical stability beyond the largest Courant numbers in the

atmosphere [9], [10], [11], and the lack of dispersive features near shocks [12]. Making the numerical scheme hugely effective and adopted in a variety of atmospheric settings, some notable examples in operational models include the ECMWF Integrated Forecasting System, RPN Canada's GEM model, Météo-France's ARPEGE model, and the Met Offices' Unified Model.

The stability of a Semi-Lagrangian method is not limited by the Courant number, however its accuracy is limited by the less severe Lipschitz criterion [appendix A in Pudikiewicz, J., R. Benoit, and A. Staniforth (1985) [13]], in which the iterative departure point calculation requires a timestep restriction for convergence. The physical interpretation of this restriction is that trajectories do not cross each other [14]. The practical consequence is that a timestep three to six times larger than an Eulerian scheme can be taken without a loss in accuracy [15], and stability is guaranteed over large Courant number ranges. The increase in stepsize and stability lead to its widespread adoption in atmospheric dynamical cores using a latitude longitude grid because it solved the "pole problem". The "pole problem" is a name given to a numerical instability commonly arising in atmospheric advection resulting from a high local Courant number located at the poles of a latitude longitude mesh, which occurs because the meridians of a latitude longitude mesh converge to a point with much higher local mesh refinement than elsewhere on the Earth.

The Semi-Lagrangian method, like all numerical methods, has some disadvantages. Although the Semi-Lagrangian method is unconditionally stable, it only remains accurate for Courant numbers typically an order of magnitude larger than explicit Eulerian schemes [13], beyond this the accuracy in finding a departure point is compromised and the Semi-Lagrangian method is subject to unusual behaviour. Hereil and Laprise found that this leads to poor representation of internal gravity waves at high Courant numbers [15]. Bartello and Thomas [16] also point out some (albeit perhaps controversial) concerns regarding the cost effectiveness of the Semi-Lagrangian method for mesoscale dynamics, particularly in the presence of deformational flow with forcing terms. Despite some concerns over efficiency and accuracy at extreme Courant numbers, the Semi-Lagrangian method has been deployed successfully in operational dynamical cores. Stability rather than accuracy was the primary concern for the pole problem. Although the Semi-Lagrangian method is not inherently monotone there exists methods to embed "quasi-monotonicity" or "shape preservation" into the Semi Lagrangian method [17], [18].

The real disadvantage for the currently operationally deployed Semi-Lagrangian scheme is the lack of any formal conservation properties. The loss of local mass conservation, even in the presence of global mass fixer algorithms can have undesirable mass drift over long time runs of the atmosphere [2]. Some researchers have proposed inherently mass preserving variants of the Semi-Lagrangian method, known as Flux Form Semi-Lagrangian. On orthogonal grids, local mass conservation can be achieved using dimensionally split Flux Form Semi-Lagrangian techniques [19, 20]. Conservative multidimensional remapping has also been proposed, this method is

locally conservative but computationally expensive at large Courant numbers due to the geometric considerations needed. Examples include SLICE [21] and FF-CSLAM [22]. This thesis concerns the alternative Eulerian method of lines approach to achieve the requirements of a dynamical core.

1.2 Background

Focus and scope:
Requirements of a transport algorithm in an atmospheric model

What are the demands on dynamical core advection algorithms? As numerical weather prediction models increase in complexity and computer architecture changes, the requirements of advection algorithms also change. A list of desired properties for the next generation of dynamical cores is presented below:

1. Stability in the presence of the largest Courant number used. Stability in the sense of Von-Neumann ensures that numerical errors do not get amplified and grow exponentially. This can lead to blow up and is disastrous for an operational numerical weather centre [2].

2. Inherent local mass conservation. The local conservation of mass is a desired property for advection algorithms [2]. This can be achieved by using locally conservative numerical flux functions Definition 3.1.2, satisfying $F_{KL} = -F_{LK}$ to ensure that the flux out of cell K into cell L through the face/edge σ_{KL} equals the flux into cell L from cell K.

3. Consistency and constancy preservation. The unit tracer reduces the evolution of a tracer density to the continuity equation [2]. Constant tracers should be preserved for incompressible flow. This can be achieved by using consistent numerical flux functions Definition 3.1.2. The consistency of fluxes and a divergence free condition ensures that the numerical scheme will preserve constant tracers under incompressible flow. This property is deemed hugely important for the design of tracer advection in the atmosphere [19, 20].

4. Positivity preservation. The positivity of tracer values should be preserved even for compressible flow [1][23]. When the flow is incompressible tracers should be Range bounded/Globally bounded/Global maximum principle satisfying.

5. Monotone or locally bounded. The shape of the tracer shouldn't be polluted by non-physical ripples, this is often referred under the umbrella terms "monotonic" or "monotone"[2] and is also deemed important for algorithms that trans-

[1] Or more generally the scheme is sign preserving (positive definite)

[2] Several notions of monotonicity have been introduced in the numerical solution of hyperbolic PDE's, for example HHLK-monotone(L1 contractive semigroup)[24], Total variation diminishing [25], Monotonicity preserving [26], Spekreijse monotone (positive coefficient) [27] or Local Extrema diminishing, [28]. Other notable examples include entropy stable methods [29], and essentially non oscillatory methodology [30] [31] [32].

port atmospheric tracers [23, 2]. In this thesis we will use the notion of being locally bounded as a general framework to impose several different discrete local maximum principles for suppressing non-physical oscillations in incompressible flow.

6. The method must be adaptable to at least one of the quasi-structured meshes proposed for dynamical cores, Lat-Long, Cubed-Sphere, Yin-Yang, Icosahedral, Reduced Grid, and unstructured [33].

7. The numerical method should be better than first order accurate whenever theoretically possible. This is determined by several linear and nonlinear order barrier theorems, which depend on the Courant number and the monotonicity properties imposed on the scheme. We have to achieve higher order accuracy > 1 at low Courant number flows, away from discontinuities and extrema. We want to sacrifice higher order accuracy in favour of stability and monotonicity on a local basis at high Courant numbers, at extrema and discontinuities.

Literature review:
Achieving accuracy and monotonicity

Higher order accuracy and monotonicity have long been at odds, we briefly introduce a historical background to this conflict. This is to set the scene and highlight where Chapter 2 and Chapter 3, fit into this ongoing conflict, and how a compromise between both high order accuracy and monotonicity can be achieved for dynamical cores.

Godunov [26] showed that the order of a linear monotonicity preserving numerical scheme is necessarily first order. Harten Hyman Lax and appendix by Keyfitz (HHLK)[24] generalised the well-known Godunov order barrier, to state nonlinear schemes with the stronger HHLK-monotone condition are necessarily first order. Harten [34] developed a nonlinear total variation diminishing (TVD) scheme, satisfying a property weaker than HHLK monotone property but stronger than the monotonicity preserving property of Godunov [26]. These TVD schemes bypass the Godunov order barrier, are second order[3] and have a convenient framework introduced by Sweby in 1984 [38] for the construction of different schemes. The total variation diminishing framework did not generalise easily into multi-dimensions well. This is in part due to the definition of total variation, and Goodman and Leveque showed that a particular two-dimensional definition of TVD would also run into the HHLK order barrier [39] implying first order accuracy, further dissuading onlookers. Amongst the many achievements in the pioneering work of Spekreijse [27] is a developed notion of monotonicity suitable for more dimensions, it is based on positivity of coefficients and enforces a discrete local maximum principle. Spekreijse showed that a wide class of flux split schemes satisfy a discrete local maximum principle

[3]The story of whether TVD schemes are second order at and near smooth extrema and sonic points has some contention in the literature since a minor inaccuracy in [35] this has been resolved by the overlooked paper by Hua-mo [36], but sometimes attributed to [37].

(strictly stronger than the total variation diminishing property) and showed less restrictions on limiter functions were required. Spekreijse's [27] work generalised well to unstructured meshes and semi-discrete frameworks and are referred to as positive coefficient schemes or local extrema diminishing schemes [28]. Spekreijse's work has been used to motivate monotonic atmospheric advection [40] with promising numerical results for a relatively simple design. It is widely believed Spekreijse's monotonicity theory can be used to carry monotone properties into multiple dimensions for the incompressible advection equation [41, 40, 37]. However, it has never been proven and to what extent that this is true has never been resolved, this is the subject of Chapter 2.

The monotonicity theory developed by Spekreijse [27] does generalise to some unstructured meshes [28], and multidimensional limiters such as Barth and Jespersen's limiter [42] can preserve certain local maximum principles for some higher order methods. It is not apparent as to how this framework can adapt to the increasingly higher order methods being deployed in atmospheric and ocean models, with rigorous guarantees for local boundedness. More recently Zhang et al. [43] modified the HHLK-monotone property to develop limiters suitable for positivity and a global maximum principle, this framework is applicable for higher order finite volume and discontinuous Galerkin (DG) finite element methods and has seen practical success for different meshes [44]. However, this framework has not been adapted for local maximum principles, this is the subject of Chapter 3.

We have been overlooking the first item in the dynamical core shopping list, stability in the presence of large Courant numbers. To propose a Eulerian competitor to the Semi-Lagrangian method, we must solve the problem of robustness in the presence of high local Courant numbers, (this is to deal with large local vertical Courant numbers rather than the pole problem). Implicit time stepping such as implicit Runge Kutta methods can be unconditionally stable. Historically implicit time-stepping has been considered too expensive, but as modern numerical linear algebra software continues its rapid development, the use of implicit time-stepping could emerge as a feasible robust method due to the parallel scaling and efficiency introduced by multigrid and preconditioning methods [45]. The NUMA model has seen improvements in efficiency over existing explicit schemes by adopting IMEX Runge Kutta methods to deal with large local Courant numbers [46].

Yee Warming and Harten [47] introduced a linearised flux limiters scheme with provable guarantees on the total variation, however their numerical method sacrifices mass preservation. They also introduced a different linearisation approach in flux form to ensure mass preservation. This approach preserves mass, but to date there are no theoretical guarantees on monotonicity, despite the numerical results being promising in one dimension [47]. To what extent the mass preserving linearisation technique is monotone has not been addressed, and if it isn't how one can correct this would need to be solved. Furthermore, implicit nonlinear methods can often be monotonicity violating not from the result of the time-stepping itself, but due to the

solving strategy. For example, when the spatial scheme is nonlinear (required by Godunov's theorem), the solving strategy of a nonlinear system requires an iterative process converging to the solution, as does the numerical linear algebra technique. These iterative processes may not converge in a strong enough norm to ensure that the approximate solution will be monotone.

High order accuracy (greater than 1) of an implicit Runge Kutta method once again comes into conflict with monotonicity at high Courant numbers. This is unresolvable and a consequence of a nonlinear order barrier theorem, arising out the strong stability preserving literature and confirmation by numerical search [48], [49], [50], [51], [52]. Flux corrected transport (FCT) emerged in the early 1970's [53], it was the first non-linear finite difference technique capable of producing high order non-oscillatory solutions for fluid dynamics bypassing Godunov's order barrier theorem [26] by being non-linear. Flux corrected transport, is an algorithm with two major stages, a monotone stage followed by an antidiffusive stage, where the antidiffusive stage is corrected as to not generate new maxima or minima in the monotone solution. This is done by limiting fluxes on a case-by-case basis to ensure that the cell mean value is not pushed beyond its neighbouring values [53]. This flux corrected transport algorithm can also be interpreted as a means of correcting a high order oscillatory solution on a lower order non-monotonic one [54]. However, FCT differs from flux limiting techniques by the direct computation of the intermediary monotone solution which is used in the subsequent limiting process.

The generality of flux corrected transport was first introduced in [55], and extended further in [56] where notions of adding and removing artificial diffusion, were applied to different schemes including Lax-Wendroff and the Leapfrog. In 1979 Zalesak extended the framework to consider truly multidimensional problems [54], allowing a truly unstructured flux corrected transport algorithm. The flux corrected transport method has since been applied to various problems and methodologies including spectral methods [57], finite element [58], finite volume and finite difference methods. For problems as varied as Plasma dynamics [59] to Wave propagation in media [60]. More recently, flux corrected transport schemes have been developed for the explicit multistep Adam Bashford third order time integration used in dynamical cores, with results showing suitability for monotonic, positive, inherently conservative tracer transport on an icosahedral grid [61]. Recently, one-step flux corrected transport schemes have been used to correct an explicit fourth order Runge Kutta method on the corner transport finite volume method [62].

Flux corrected transport methods have been used to correct high order implicit anti diffusion [55], [56], [63] on explicit first order numerical methods, with motivations in improving phase/truncation errors. Steinle and Morrow [64] have proposed correcting high order implicit fluxes on an explicit first order upwind method consisting of multiple explicit first order passes, with motivation on creating a numerical scheme robust in the presence of large Courant numbers. The low order scheme is an explicit multiple pass scheme whose cost increases with the Courant number, it was noted in

the conclusion of this paper [64] that the computational cost was dominated by the low order multiple passes at when the Courant number is high. Chapter 4 takes the conclusion from [64] seriously and proposes using implicit methods for both the high and low order fluxes. Compared to explicit FCT, implicit flux corrected transport, has been relatively unexplored in the finite volume and finite difference literature, but has emerged as a powerful tool in the finite element framework [65], [66], [67] with several different proposed algorithms.

1.3 Outline of thesis

Chapter 2 presents the derivation of a new limiter region for incompressible flow to retain a discrete maximum principle. We numerically demonstrate Spekreijse's theory doesn't directly apply to the incompressible advection equation, and explain under what circumstances it fails. We show that symmetric limiters quasi-necessarily lie in the Sweby region. The new limiter regions derived are more general than the Sweby region if one breaks the limiter symmetry requirement, and this can be used to get: more accurate limiters, more compressive limiters, or more smooth limiter functions, that will work for incompressible multidimensional flow.
Chapter 3 explains how the Zhang et al. [43] framework can be used as a basis for imposing local boundedness, and how this leads to new multidimensional slope limiters suitable for a wide variety of methods. This is demonstrated with two examples. We derive a new multidimensional slope limiter, for second order finite volume schemes and compare to the current state of the art multidimensional limiters. We then introduce a new higher order finite volume method, and show that the new multidimensional slope limiter framework is readily applicable under a decomposition of the cell average.
Chapter 4 presents a numerical investigation into what extent implicit linearised slope limiters are monotonic. We demonstrate that the mass preserving linearisation strategy in [47] for both one dimensional and multi-dimensional limiters produce small negative values at low Courant number, due to approximations introduced in the linearisation. We describe an implicit flux correction transport algorithm that allows the fix of such monotonicity violations. We also demonstrate the same flux correction technique brings a local boundedness principle without the use of any linearisation procedure and retains a local maximum principle in the presence of large Courant numbers by dropping the accuracy and order of the method. We numerically test the use of pre-limiting, and conclude that for our implicit FCT algorithm the use of pre-limiting had unfavourable effects. Chapter 4 indicates that implicit flux corrected transport algorithms can be used to create monotone(local maximum principle satisfying), positive, inherently conservative, consistent, unconditionally stable, transport algorithms capable of attaining second order or above at low Courant numbers, and unconditional nonlinear stability at large Courant numbers.

Chapter 2

Multidimensional local maximum principle limiter region

2.1 Introduction

2.1.1 Motivation in atmospheric modelling

In the field of atmospheric modelling and computational fluid dynamics there is an ongoing search for advection algorithms with an increasing number of physically reasonable properties, in particular a high order (> 1), monotonic (local discrete maximum principle satisfying), sign preserving, locally mass preserving, linear invariant, consistent transport algorithm for the incompressible advection equation is sought after. In subsection Section 2.2 we introduce such a scheme for solving the incompressible advection equation

$$\frac{\partial u}{\partial t} + \operatorname{div}(\boldsymbol{v}u) = 0, \quad \forall(\boldsymbol{x}, t) \in \Omega \times [0, T], \tag{2.1}$$

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \Omega. \tag{2.2}$$

The incompressible advection equation is a standard mathematical model of transport phenomena which occurs in physical space $\Omega \subset \mathbb{R}^d, d = 1, 2, 3$, over an arbitrary time span $t \in [0, T]$, where a tracer density $u(\boldsymbol{x}, t)$, is advected by a divergence free, bounded, continuous velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$, $\operatorname{div}(\boldsymbol{v}) = 0$. Interpretation of this equation can be pointwise or as a integral conservation law by using the Gauss divergence theorem over each cell in a mesh.

Smolarkiewicz and Rasch [68] compared explicit forward in time Eulerian techniques to Lagrangian techniques for monotone advection on the sphere, and advocate for Eulerian techniques such as MPDATA when monotonicity is essential. Hundsdorfer, Koren, van Loon and Verwer [40] propose a different Eulerian technique consisting of one-dimensional flux limiters in a multidimensional flux form finite difference spatial discretisation, which is subsequently discretised in time by the method of lines. Hundsdorfer, Koren, van Loon and Verwer [40] observe preferable accuracy and computational speed with the one-dimensional limiters when compared with the third

order MPDATA scheme in [68]. The research in [40] indicates that one-dimensional flux limiters in a method of lines framework can be a competitive scheme for positive mass preserving transport on the sphere. Wesseling [37] proposes a very similar scheme under a finite volume interpretation, and in one dimension they demonstrate stronger non-linear properties than positivity such as TVD, and they suggest some limiters in the Spekreijse region to improve accuracy. It is known using one-dimensional limiters in the Spekreijse region can create a TVD scheme, it is also widely believed that this can be generalised to the multidimensional flux form advection equation using the positive coefficient framework of Spekreijse [27], some examples of this belief can be found in [41, 37, 40]. We show that Spekreijse framework does not directly apply when the scheme is in flux form, and the extended admissible limiter region can fail. Limiters such as OSPRE, van Albada, Albada family, Hemker-Koren, local double-logarithmic reconstruction (LDLR), Cada–Torrilhon, TCDF, MPL2-κ and ENO2 will fail to have a provable discrete maximum principle and more importantly will lose positivity under certain flows. The region in which limiters are discrete maximum principle satisfying is important to the successful design of monotone advection algorithms.

2.1.2 Background material and summary: Different existing limiter regions

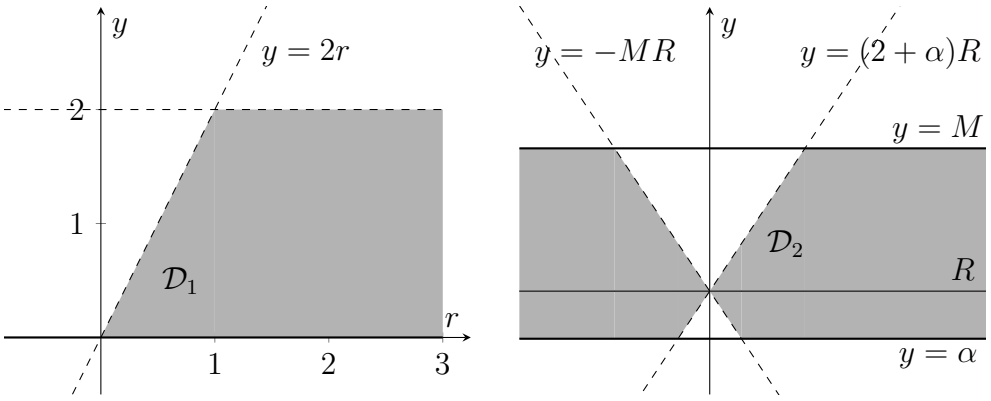
A semi discrete flux limited numerical scheme for the constant coefficient advection equation $u_t + au_x = 0$ can be written as $\frac{du_i}{dt} + \frac{F_{i+1/2}-F_{i-1/2}}{\Delta x} = 0$, where the flux is given by $F_{i+1/2} = a[u_i + \frac{1}{2}\psi(R_i)(u_i - u_{i-1})]$, ψ is the limiter function, and the ratio of successive gradients is given by $R_i = \frac{u_{i+1}-u_i}{u_i-u_{i-1}}$. A similar but different construction is given using the inverse ratio $r_i = \frac{1}{R_i}$ as $F_{i+1/2} = a[u_i + \frac{1}{2}\psi(r_i)(u_{i+1} - u_i)]$. We will later distinguish these different frameworks by the parameter $\theta = 1, 0$, respectively. The total variation of u is defined as $TV(u) = \sum_{\forall i} |u_i - u_{i-1}|$, and serves as a measure of how oscillatory u is. A desirable property of a numerical method is that the total variation does not increase with time. This property is referred to as total variation diminishing (TVD).

In 1984 Sweby [38] introduced sufficient conditions for a flux limited scheme to be total variation diminishing,

$$\psi(r) \leq \min(2r, 2), \quad \forall r > 0, \tag{2.3}$$

$$\psi(r) = 0, \quad \forall r < 0. \tag{2.4}$$

These conditions serve as bounds on acceptable limiters in a diagram known as the Sweby diagram, however it is still sometimes misunderstood that these bounds are both necessary and sufficient (rather than sufficient) for schemes to be total variation diminishing [69]. Spekreijse showed that this is not the case in the inverse



(a) Sweby (b) Spekreijse

Figure 2.1: Fig. 2.1a is the plot of the sufficient admissible limiter region in the (r, y) plane as defined by Sweby [38]. Fig. 2.1b is the plot of the admissible limiter region in the (R, y) plane as defined by Spekreijse. The Sweby region $\mathcal{D}_1$ is sufficient for a one-dimensional scheme with flux limiters to be TVD, the Spekreijse region $\mathcal{D}_2$, which has two free parameters $\alpha \in [-\infty, 0]$, $M \in (0, \infty)$, is also sufficient for the scheme to be TVD.

ratio formulation by finding a more general (TVD) admissible limiter region,

$$\psi(R) \in [\alpha, M], \quad \psi(S)/S \in [-M, 2+\alpha], \quad M \in (0, \infty), \quad \alpha \in [-\infty, 0], \tag{2.5}$$

for $R, S \in \mathbb{R}$. Spekreijse showed that flux splitting schemes with limiters in this region satisfy a local discrete maximum principle, which is strictly stronger that the total variation diminishing property in one dimension and is more convenient in multiple dimensions. The use of limiters in the extended region of Spekreijse has been successful in both flux difference splitting [70] and flux vector splitting frameworks [27].
In Fig. 2.1 we have plotted both the Sweby region $\mathcal{D}_1$ in Fig. 2.1a, and the Spekreijse region $\mathcal{D}_2$ in Fig. 2.1b. The extended limiter region has allowed a huge variety of new limiters to be introduced into the literature, which can improve accuracy [36], and are not necessarily reduced to the first order upwind scheme at extrema and can be globally smooth [27].

In this chapter we will theoretically investigate the numerical method introduced in Wesseling and Zijlema [37] and Hundsdorfer et al. [40], to find the admissible flux limiter region for the incompressible advection equation in flux form. It is widely believed that the monotonicity properties carry from 1-D to 2-D using Spekreijse's monotonicity criterion [41, 37, 40], however it has not been proven and there are a few technical points that seem to be missing in the literature as to which limiters can be carried over in this framework.

1. The first finding of our chapter is that some limiters in the Spekreijse region will no longer be discrete maximum principle satisfying or even positivity preserving, for the flux form advection equation when in multiple dimensions,

unless the flow satisfies a directional mean value theorem. This is indicated theoretically in Section 2.3.1 and demonstrated numerically in Table 2.2.

2. The main finding in this chapter is the derivation of two new limiter regions suitable for the incompressible flux form advection equation given in Fig. 2.3. These limiter regions are more general than the Sweby region [38], and maintain a local maximum principle for general incompressible flow unlike the generalised monotonicity region of Spekreijse. These limiter regions are new to the literature and are not a subset of the Convective boundedness criterion or the Sweby region.

3. We also show that Sweby's region is not only sufficient but a quasi-necessary assumption for symmetric $\psi(\frac{1}{r}) = \frac{\psi(r)}{r}$ limiters to attain a discrete maximum principle for the incompressible flux form equation. This is indicated in Section 2.3.2, and we use the term quasi-necessary due to the fact ruling out the existence of some yet to be found transform into another alternative positive coefficient representation is hard. We push the ENO2, Ospre and van Albada limiters into the Sweby region, and numerically demonstrate this fixes the monotonicity failures and has little to no consequence in terms of accuracy, convergence or peak resolution.

4. By breaking the symmetry condition on the limiter function, there is an infinite family of possible limiter implementations with a free parameter θ. We investigate when $\theta = \{0, 1\}$, and derive two new limiter regions outside the Sweby region. These limiter regions are proven sufficient for the flux form incompressible advection equation to retain a local maximum principle. We introduce the first globally differentiable limiter function contained entirely in a second order region suitable for incompressible flow Eq. (2.85), and some other limiters functions that from a preliminary glance are competitive with some of the most accurate limiter functions.

5. Hundsdorfer [71] remarks that precise theoretical support on the question of how small the timestep should be for monotonicity would be of practical importance. We include some convenient formal Courant number restrictions for practical application.

6. We prove linear invariance, and test numerically a variety of different new limiters functions.

2.2 Space discretisation: Flux form

In this sub section we introduce the (slightly unusual) notation to be used throughout the rest of this chapter. As well as describe the flux form finite volume scheme of Wesseling and Zijlema [37] or the finite difference scheme of Hundsdorfer et al [40] in the following steps.

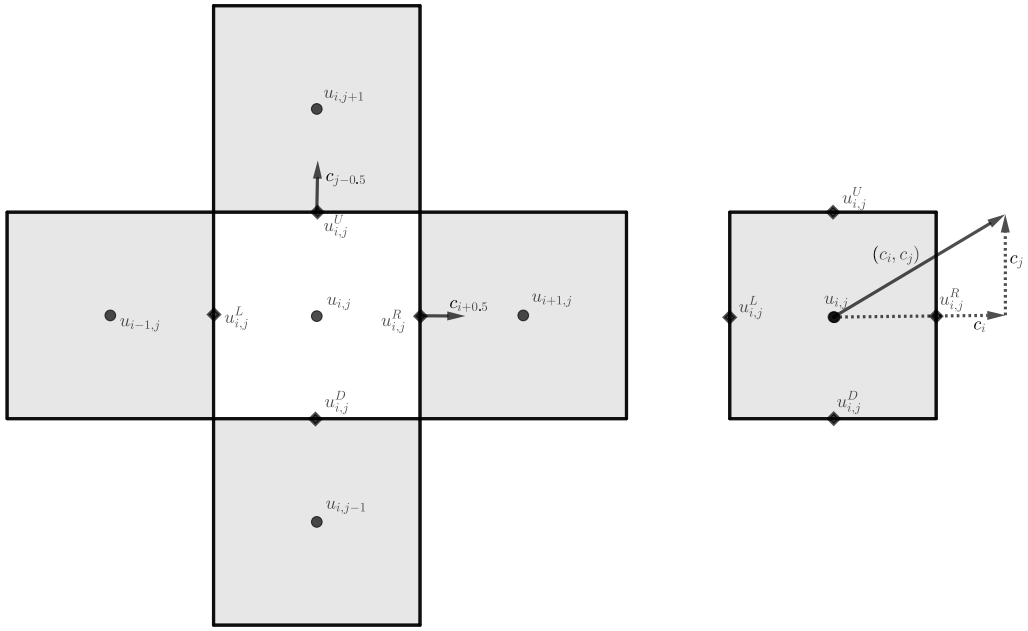


Figure 2.2: The left diagram is a flux form stencil, where the velocities $c_{i+1/2}$, $c_{j+1/2}$ are located at the faces, sometimes called a C-grid. The positions for the left, right, up, down interface values u^L, u^R, u^U, u^D are denoted by diamonds at positions $(i-1/2, i), (i+1/2, j), (i, j+1/2), (i, j-1/2)$. The diagram on the right is an advective form stencil with velocity c_i, c_j at the midpoint, and is sometimes called an A-grid stencil.

1. Let $u = u_{i,j}$ denote the cell mean or pointwise value of a tracer within a cell, where if one or other subscript is missing it is assumed to be at position i or j as appropriate.

2. Reconstruct: A reconstruction operator constructs the values attained to the right, left, up and down of cell (i, j) as follows

$$u_i^R = u_i + \frac{\theta}{2}\psi(R_i)(u_i - u_{i-1}) + \frac{(1-\theta)}{2}\psi(\frac{1}{R_i})(u_{i+1} - u_i), \tag{2.6}$$

$$u_i^L = u_i + \frac{\theta}{2}\psi(\frac{1}{R_i})(u_i - u_{i+1}) - \frac{(1-\theta)}{2}\psi(R_i)(u_i - u_{i-1}), \tag{2.7}$$

$$u_j^U = u_j + \frac{\theta}{2}\psi(R_j)(u_j - u_{j-1}) + \frac{(1-\theta)}{2}\psi(\frac{1}{R_j})(u_{j+1} - u_j), \tag{2.8}$$

$$u_j^D = u_j + \frac{\theta}{2}\psi(\frac{1}{R_j})(u_j - u_{j+1}) - \frac{(1-\theta)}{2}\psi(R_j)(u_j - u_{j-1}), \tag{2.9}$$

$$R_i = \frac{u_{i+1} - u_i}{u_i - u_{i-1}}, \quad R_j = \frac{u_{j+1} - u_j}{u_j - u_{j-1}}, \tag{2.10}$$

using upwind bias flux limiting. The right, left, up and down of cell (i, j) are denoted with R, L, U, D superscripts. To be clear we are using the ratio of successive gradients defined in [27] as R when $\theta = 1$, we call this the Roe gradient. When $\theta = 0$ we are using the inverse ratio defined in [38] $r_i = \frac{1}{R_i}$,

we call the Sweby gradient. The choice of θ does not matter when the limiter is symmetric $\psi(1/R) = \psi(R)/R$. We only investigate $\theta = \{0, 1\}$ due to the reduced computational cost in evaluating the expressions in Eqs. (2.6) to (2.9). When the symmetry condition is broken the two limiter functions should be distinguished differently, as they have their own respective limiter regions, however this should be apparent from the context of the limiter, so we do not include this notationally.

3. Riemann: the donor cell numerical flux function

$$F(u_i^R, u_{i+1}^L, c_{i+0.5}) = c_{i+0.5}^+ u_i^R + c_{i+0.5}^- u_{i+1}^L, \tag{2.11}$$

resolves the Riemann problems at the boundaries. Here we define the notation $(\cdot)^+ := \max(\cdot, 0)$, $(\cdot)^- := \min(\cdot, 0)$, to mean the positive or negative component of the argument respectively, to be used throughout this chapter. $c_{i+1/2}$ denotes the x component of the velocity at position $(i + 1/2, j)$, and $c_{j+1/2}$ denotes the y component of the velocity at position $(i, j + 1/2)$. We often absorb constants like timestep and mesh width into this constant because the numerical flux function is linear. This numerical flux function is consistent, monotone, and Lipschitz continuous as defined in [72] or Definition 3.1.2. This type of flux is known as a state interpolated flux as defined in [41, 71, 40], rather than the flux interpolated flux. In the finite volume setting we approximate the integral of flux through the face by second order gauss quadrature over each face(midpoint rule), whereas in the finite difference setting we interpret as a point valued flux.

4. Evolve: the semi discrete evolution operator is given by the flux form method

$$\begin{aligned} &\frac{\partial u}{\partial t} + [F_{i+0.5}(u_i^R, u_{i+1}^L, c_{i+0.5}) - F_{i-0.5}(u_{i-1}^R, u_i^L, c_{i-0.5})] \\ &+ [F_{j+0.5}(u_j^U, u_{j+1}^D, c_{j+0.5}) - F_{j-0.5}(u_{j-1}^U, u_j^D, c_{j-0.5})] = 0. \end{aligned} \tag{2.12}$$

We have absorbed the mesh spacing into the coefficients c by linearity of the donor cell flux function, and plan to do so with the time-step Δt.

For context see [27],[37],[40], and for an example of its generalisation to unstructured meshes see [73].

2.3 Theory

In this section we will establish the following cell mean local maximum principle,

Definition 2.3.1 (Cell mean local maximum principle)**.**

$$u_{i,j}^{n+1} \in [\min\{u_{i+1}^n, u_{i-1}^n, u^n, u_{j+1}^n, u_{j-1}^n\}, \max\{u_{i+1}^n, u_{i-1}^n, u^n, u_{j+1}^n, u_{j-1}^n\}], \quad \forall i, j \tag{2.13}$$

for the forward Euler numerical flow map $u^{n+1} = \text{Forward Euler}(u^n, c^n)$, of Section 2.2 defined as

$$\begin{aligned} u_{i,j}^{n+1} = u_{i,j}^n &- [F_{i+0.5}(u_i^R, u_{i+1}^L, c_{i+0.5}^n) - F_{i-0.5}(u_{i-1}^R, u_i^L, c_{i-0.5}^n)] \\ &- [F_{j+0.5}(u_j^U, u_{j+1}^D, c_{j+0.5}^n) - F_{j-0.5}(u_{j-1}^U, u_j^D, c_{j-0.5}^n)]. \end{aligned} \tag{2.14}$$

Where the time step, Δt, and mesh spacing, Δx, have been absorbed into the face defined velocity field c. We introduce the scaled face defined velocities $c_{i+1/2} = \frac{\Delta t}{\Delta x}(v_1)_{i+1/2}$, $c_{j+1/2} = \frac{\Delta t}{\Delta y}(v_2)_{j+1/2}$ and the cell defined Courant number

$$C_{i,j} = \frac{1}{2}\big[|c_{i+1/2}| + |c_{i-1/2}| + |c_{j+1/2}| + |c_{j-1/2}|\big]. \tag{2.15}$$

The proof will put the scheme in a positive coefficient type form Definition 2.3.2 as defined by Spekreijse [27]. This definition allows use of a Lemma 2.3.1 also given by Spekreijse [27] which can be used to prove a local maximum principle. In Section 2.3.3 we make note of the strong stability literature that allows this to be made into a higher order scheme whilst retaining some aspects of a discrete maximum principle.

Definition 2.3.2 (Spekreijse, Positive-coefficient type scheme [27])**.** The semidiscrete scheme given by

$$\begin{aligned} \frac{\partial u_{i,j}}{\partial t} &+ A_{i-1/2}(u_i - u_{i-1}) + B_{i+1/2}(u_i - u_{i+1}) \\ &+ C_{j-1/2}(u_j - u_{j-1}) + D_{j+1/2}(u_j - u_{j+1}) = 0, \end{aligned} \tag{2.16}$$

is a positive coefficient scheme when all the nonlinear leading coefficients are non-negative

$$A_{i-1/2}, B_{i+1/2}, C_{j-1/2}, D_{j+1/2} \geq 0. \tag{2.17}$$

Lemma 2.3.1 (Spekreijse [27])**.** *When the forward Euler temporal scheme of a positive coefficient type scheme (satisfying Definition 2.3.2) satisfies the time step restriction,*

$$(A_{i-1/2}^n + B_{i+1/2}^n + C_{j-1/2}^n + D_{j+1/2}^n) \leq 1, \tag{2.18}$$

then $u_{i,j}^{n+1}$ is a convex combination of $u_{i+1}^n, u_{i-1}^n, u^n, u_{j+1}^n, u_{j-1}^n$, trivially implying the following local discrete maximum principle with respect to edge sharing neighbours cell mean values,

$$u^{n+1} \in [\min\{u_{i+1}^n, u_{i-1}^n, u^n, u_{j+1}^n, u_{j-1}^n\}, \max\{u_{i+1}^n, u_{i-1}^n, u^n, u_{j+1}^n, u_{j-1}^n\}]. \tag{2.19}$$

This in turn this implies weaker properties such as boundedness $||u^{n+1}||_\infty \leq ||u^n||_\infty$, and sign preservation.

Proof. The forward Euler discretisation of Eq. (2.16) can be written as

$$\begin{aligned} u_{i,j}^{n+1} &= (1 - [A_{i-1/2}^n + B_{i+1/2}^n + C_{j-1/2}^n + D_{j+1/2}^n])u_{i,j} \\ &+ A_{i-1/2}^n u_{i-1} + B_{i+1/2}^n u_{i+1} + C_{j-1/2}^n u_{j-1} + D_{j+1/2}^n u_{j+1}. \end{aligned} \tag{2.20}$$

Which is a convex combination under the condition Eq. (2.18). □

2.3.1 Applicability of the extended Spekreijse region

In this section, we will discuss under what conditions Spekreijse's theorem 2.2 applies to the incompressible flux form advection equation. This is to discuss under what circumstances the extended limiter region in [27] is appropriate.

Spekreijse theorem 2.2 relies on putting the scheme in a positive coefficient representation. For the flux form method Section 2.2 it is most natural to write the scheme Eq. (2.12) in the following positive coefficient form

$$\begin{aligned} &\frac{\partial u_{i,j}}{\partial t} + \frac{(u_i^R c_{i+1/2}^+ - u_{i-1}^R c_{i-1/2}^+)}{(u_i^R - u_{i-1}^R)} \frac{u_i^R - u_{i-1}^R}{u_i - u_{i-1}} (u_i - u_{i-1}) \\ &- \frac{(u_{i+1}^L c_{i+1/2}^- - u_i^L c_{i-1/2}^-)}{(u_{i+1}^L - u_i^L)} \frac{u_{i+1}^L - u_i^L}{u_{i+1} - u_i} (u_i - u_{i+1}) \\ &+ \text{y-direction} = 0. \end{aligned} \tag{2.21}$$

Where we have omitted the y-directional terms. In order for this to be of positive coefficient type, we require that the leading terms are positive. This is traditionally done by a mean value theorem on the numerical flux function with respect to u. However, the numerical flux function has additional dependence on a velocity field and the mean value theorem won't necessarily apply in u. It is here one could say that the Spekreijse region Eq. (2.5) clearly is inappropriate for incompressible flow, by choosing an incompressible flow with $(u_i^R, c_{i+1/2}^+, u_{i-1}^R, c_{i-1/2}^+) = (1, 0, 0.5, 1)$ giving a negative leading coefficient. This is an indication that the theory doesn't hold, i.e. it is not sufficient, it is hard to rule out the existence of a transform in which the scheme can be put into a different positive coefficient representation. Instead, we will rely on a numerical demonstration in Table 2.2, Fig. 2.8a, to show that this region is inappropriate. This gives us the first contribution of this chapter: the extended region of Spekreijse does not give schemes that have a discrete maximum principle or even positivity when applied to the multidimensional flux form incompressible advection equation.

However, there are clearly some circumstances in which you can still use Spekreijse's more ambitious limiter region, namely when you can show a mean value theorem applies in each direction. For example, a sufficient condition for the leading coefficients in Eq. (2.21) to be positive is that the x-component of velocity v_1, is independent of x, so that $v_1 = v_1(y, t)$ and the y-component of velocity v_2, is independent of y so that $v_2 = v_2(x, t)$. This implies that the flow has a directional mean value

theorem, and in fact this implies the directionally constant $c_{i+1/2} = c_{i-1/2} = c_i$, $c_{j+1/2} = c_{j-1/2} = c_j$ property, and we can show that the leading terms will be positive directly

$$\frac{(u_i^R c_{i+1/2}^+ - u_{i-1}^R c_{i-1/2}^+)}{(u_i^R - u_{i-1}^R)} = c_i^+ \geq 0. \tag{2.22}$$

Furthermore, seemingly difficult flows such as $v_1(y,t) = y^3 + \sin(y)\cos(t)$, $v_2(x,t) = x + \sin(3x)\sin(t)$, will likely have a discrete maximum principle, with the ambitious limiters of Spekreijse without a divergence free condition. A numerical test case for incompressible advection should be chosen without this directional mean value theorem property when testing incompressible flow monotonicity, we use a sine deformation Eq. (2.87).

It is important to note that Spekreijse's theory is entirely correct as a flux splitting framework for hyperbolic PDE's, it could be misunderstood how this well-established theory translates into a flux form finite volume method when there is additional dependence on velocity defined at edges. If you were to directly use Spekreijse's flux splitting method as defined in [27] on the advection equation, you would be using an advective form method, where velocity is located at the cell centres, pictured in Fig. 2.2. In this circumstance not only will the extended limiter of Spekreijse provably work, but the monotonicity criteria can be generalised further for the advection equation as there is no need for the uniform boundedness condition in [27], since we do not need to use the mean value theorem (positive coefficient scheme property can be directly calculated using Eq. (2.22)). This leads to an even more general monotonicity criterion with 4 free parameters for the advective form equation

$$\psi(R) \in [m_1, M_1], \quad \psi(S)/S \in [m_2, M_2], \quad m_1, m_2 \leq 0, \quad M_1, M_2 \geq 1. \tag{2.23}$$

We state this preliminarily but do not investigate any further due to one of the requirements in dynamical cores is local mass conservation, and this form is not in flux form and is not locally mass preserving.

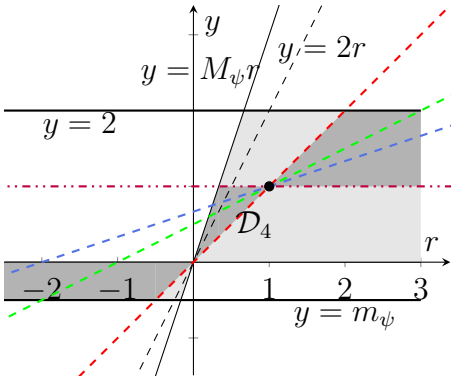
2.3.2 Limiter suitable for incompressible flow

We now derive limiters which give a positive coefficient scheme in the more general setting of incompressible advection

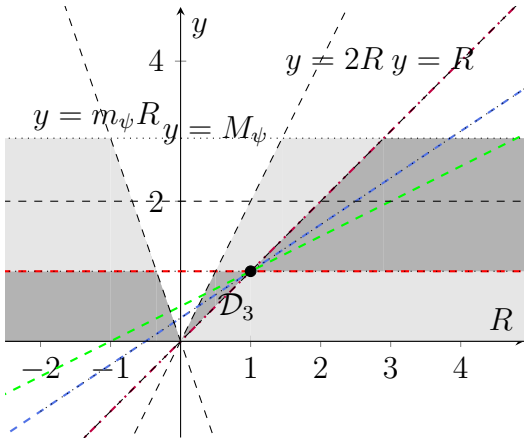
$$\partial_t u(x,y,t) + \partial_x(v_1(x,y,t)u(x,y,t)) + \partial_y(v_2(x,y,t)u(x,y,t)) = 0, \tag{2.24}$$

where the discrete maximum principle arises as a result of the incompressibility condition

$$\partial_x(v_1(x,y,t)) + \partial_y(v_2(x,y,t)) = 0. \tag{2.25}$$



(a) We plot the $\theta = 0$ schemes new limiter region in the Sweby ratio r. Light gray are admissible regions for a local maximum principle, and dark grey indicates a desirable second order region.



(b) We plot the $\theta = 1$ schemes new limiter region in the Roe ratio R. Light gray are admissible regions, and dark grey indicates a desirable second order region.

Figure 2.3: We have plotted dotted coloured lines of well-known linear schemes on top of the new limiter regions. Red denotes second order upwind (SOU), green denotes the Fromm scheme, blue denotes the cubic upwind interpolation scheme (CUI), and purple denotes second order central differencing (CDS).

of the velocity field $\boldsymbol{v} = (v_1, v_2)$, rather than the directional mean value theorem. In this section, we show that this changes the region of acceptable one-dimensional limiters.

Assumptions 2.3.1 (Theorem assumptions)**.**

1. *The mesh scaled velocity satisfies a discrete divergence free condition*

$$c_{i+1/2} - c_{i-1/2} + c_{j+1/2} - c_{j-1/2} = 0, \quad \forall (i,j). \tag{2.26}$$

2. *When $\theta = 1$, the limiter function satisfies*

$$\psi(R) \in [0, M_\psi], \quad \psi(S)/S \in [m_\psi, 2]. \quad M_\psi \in [0,\infty), \quad m_\psi \in (-\infty, 0]. \tag{2.27}$$

When $\theta = 0$, the limiter function satisfies

$$\psi(1/R) \in [m_\psi, 2], \quad S\psi(\frac{1}{S}) \in [0, M_\psi], \quad M_\psi \in [0,\infty), \quad m_\psi \in (-\infty, 0]. \tag{2.28}$$

3. *The timestep restriction*

$$C \leq C_{FE} = \frac{2}{2 + M_\psi - m_\psi} \tag{2.29}$$

holds. (written in terms of the Courant number)

Theorem 2.3.1. *The Forward Euler discretisation of the numerical method described in Section 2.2, can be written as convex combination of neighbour cell mean values as in Eq. (2.20) and as a result satisfies the discrete maximum principle Eq. (2.19) when the Assumptions 2.3.1 hold.*

Before the proof what does this theorem look like practically?
Before we proceed with the proof, what does this theorem look like practically? We first point the reader to the diagrams Fig. 2.3, to see the new limiter regions, Figs. 2.5a and 2.5b for new limiters plotted in these regions, Fig. 2.4b for some limiters out of the newly defined region. Table 2.1, for theoretical time step restrictions for some different scheme using the result from Assumptions 2.3.1. This theorem and proof does generalise to arbitrary dimensions, with the appropriate modification of the Courant number.

Proof. of Theorem 2.3.1. Expand both the method in Section 2.2

$$\begin{aligned}\frac{\partial \bar{u}}{\partial t} &+ c^{+}_{i+0.5,j}u^{R}_{i} + c^{-}_{i+0.5}u^{L}_{i+1} - c^{+}_{i-0.5}u^{R}_{i-1} - c^{-}_{i-0.5}u^{L}_{i} \\ &+ c^{+}_{j+0.5}u^{U}_{j} + c^{-}_{j+0.5}u^{D}_{j+1} - c^{+}_{j-0.5}u^{U}_{j-1} - c^{-}_{j-0.5}u^{D}_{j} = 0,\end{aligned} \tag{2.30}$$

and a discrete form of the divergence free condition

$$\begin{aligned}&c^{+}_{i+0.5}u + c^{-}_{i+0.5}u - c^{+}_{i-0.5}u - c^{-}_{i-0.5}u \\ &+ c^{+}_{j+0.5}u + c^{-}_{j+0.5}u - c^{+}_{j-0.5}u - c^{-}_{j-0.5}u = 0,\end{aligned} \tag{2.31}$$

in terms of their positive and negative components, for $\theta = 1$. Taking away the divergence free condition (2.31) from (2.30) gives

$$\begin{aligned}\frac{\partial \bar{u}}{\partial t} &+ c^{+}_{i+0.5,j}(u^{R}_{i} - u_i) + c^{-}_{i+0.5}(u^{L}_{i+1} - u_i) - c^{+}_{i-0.5}(u^{R}_{i-1} - u_i) - c^{-}_{i-0.5}(u^{L}_{i} - u_i) \\ &+ c^{+}_{j+0.5}(u^{U}_{j} - u_i) + c^{-}_{j+0.5}(u^{D}_{j+1} - u_i) - c^{+}_{j-0.5}(u^{U}_{j-1} - u_i) - c^{-}_{j-0.5}(u^{D}_{j} - u_i) = 0.\end{aligned} \tag{2.32}$$

Using the expressions

$$u^{R}_{i} - u_i = 1/2\psi(R_i)(u_i - u_{i-1}), \tag{2.33}$$

$$u^{L}_{i} - u_i = 1/2\psi(\frac{1}{R_i})(u_i - u_{i+1}), \tag{2.34}$$

$$u^{R}_{i-1} - u_i = -[1 - \frac{\psi(R_{i-1})}{2R_{i-1}}](u_i - u_{i-1}), \tag{2.35}$$

$$u^{L}_{i+1} - u_i = -[1 - \frac{R_{i+1}}{2}\psi(\frac{1}{R_{i+1}})](u_i - u_{i+1}), \tag{2.36}$$

and similar expressions in the other directions, allows us to write down the scheme

in the following positive coefficient type form

$$\frac{\partial u}{\partial t} + \left[c^+_{i+1/2}\frac{\psi(R_i)}{2} + c^+_{i-1/2}[1 - \frac{\psi(R_{i-1})}{2R_{i-1}}]\right](u_i - u_{i-1}) + \left[- c^-_{i+1/2}[1 - \frac{1}{2}R_{i+1}\psi(\frac{1}{R_{i+1}})] - c^-_{i-1/2}\frac{1}{2}\psi(\frac{1}{R_i})\right](u_i - u_{i+1}) + \text{y-direction} = 0 \tag{2.37}$$

where we read off the leading term coefficients

$$A_{i-1/2} = \left(\frac{1}{2}\left[c^+_{i+0.5}\psi(R_i)\right] + c^+_{i-0.5}\left[1 - \frac{1}{2}\psi(R_{i-1})/R_{i-1}\right]\right), \tag{2.38}$$

$$B_{i+1/2} = -\left(c^-_{i+0.5}\left[1 - \frac{1}{2}\psi(\frac{1}{R_{i+1}})R_{i+1}\right] + \frac{1}{2}\left[c^-_{i-0.5}\psi(R_i^{-1})\right]\right), \tag{2.39}$$

$$C_{j-1/2} = (i \mapsto j) \circ (A_{i-1/2}), \quad D_{j+1/2} = (i \mapsto j) \circ (B_{i+1/2}). \tag{2.40}$$

We have introduced unorthodox notation $(i \mapsto j) \circ (f(i))$ to mean the same expression but with i replaced with j. Clearly

$$\psi(R) \geq 0 \quad \text{and} \quad \psi(S)/S \leq 2, \tag{2.41}$$

are sufficient, for the scheme to be of positive coefficient type. It is also imposed as a quasi-necessary assumption on the limiter functions because of the arbitrary nature of the velocity field as follows. Suppose that $c_{i+1/2} \geq 0$, then for the scheme to be of positive coefficient type at both $u_{i,j}$ and $u_{i+1,j}$ we must require that,

$$1/2c^+_{i+1/2}\psi(R_i) \geq 0, \quad c^+_{i+1/2}(1 - 1/2\psi(R_i)R_i^{-1}) \geq 0. \tag{2.42}$$

We say quasi-necessary because there could always exists a different positive coefficient representation of the scheme under some yet to be found transform or rearrangement.
We now attempt to find a sensible sufficient timestep restriction, using Lemma 2.3.1, the conditions

$$0 \leq \psi(R) \leq M_\psi, \qquad m_\psi \leq \psi(S)/S \leq 2, \tag{2.43}$$

$$m_\psi \leq \psi(1/T)T \leq 2, \qquad 0 \leq \psi(1/r) \leq M_\psi, \quad \forall R, S, T \in \mathbb{R}, \tag{2.44}$$

are sufficient for the following bounds

$$\begin{aligned} A_{i-1/2} &\in [0, c^+_{i+0.5}M_\psi/2 + c^+_{i-0.5}(1 - m_\psi/2)], \\ B_{i+1/2} &\in [0, -c^-_{i+0.5}[1 - m_\psi/2] - c^-_{i-0.5}M_\psi/2], \\ C_{j-1/2} &\in [0, c^+_{j+0.5}M_\psi/2 + c^+_{j-0.5}(1 - m_\psi/2)], \\ D_{j+1/2} &\in [0, -c^-_{j+0.5}[1 - m_\psi/2] - c^-_{j-0.5}M_\psi/2], \end{aligned} \tag{2.45}$$

where we are yet to define the constants $M_\psi \geq 0$, and $m_\psi \leq 2$.

The time step restriction for the Lemma 2.3.1 is

$$A_{i-1/2} + B_{i+1/2} + C_{i-1/2} + D_{i+1/2} \leq 1, \tag{2.46}$$

which can be satisfied when

$$\begin{aligned} &c^+_{i+0.5}M/2 + c^+_{i-0.5}(1-m/2) - c^-_{i+0.5}(1-m/2) - c^-_{i-0.5}M/2 \\ &c^+_{j+0.5}M/2 + c^+_{j-0.5}(1-m/2) - c^-_{j+0.5}(1-m/2) - c^-_{j-0.5}M/2 \leq 1. \end{aligned} \tag{2.47}$$

We lose some generality for a more convenient sufficient time step restriction. Define flow in and out Courant numbers by the following definitions

$$C^{in}_{i,j} := c^+_{i+0.5} - c^-_{i-0.5} + c^+_{j+0.5} - c^-_{j-0.5}, \tag{2.48}$$

$$C^{out}_{i,j} := -c^-_{i+0.5} + c^+_{i-0.5} - c^-_{j+0.5} + c^+_{j-0.5}. \tag{2.49}$$

where these definitions are chosen based on how the flows effect the solution du/dt. The time step restriction can be written as

$$C^{in}_{i,j}\frac{M_\psi}{2} + C^{out}_{i,j}(1 - \frac{m_\psi}{2}) \leq 1, \tag{2.50}$$

using incompressibility $C^{in}_{i,j} = C^{out}_{i,j}$ we can write this as the following Courant number restriction

$$C \leq C_{FE} = \frac{1}{(1 + \frac{M_\psi - m_\psi}{2})}. \tag{2.51}$$

So far we have proven the $\theta = 1$ case, the general form is given by

$$\begin{aligned} \frac{\partial u}{\partial t} &+ \left(c^+_{i+1/2}\big[\frac{\theta\psi(R_i)}{2} + \frac{(1-\theta)}{2}R_i\psi(\frac{1}{R_i})\big] + c^+_{i-1/2}\big[1 - \frac{\theta\psi(R_i)}{2R_i} - \frac{(1-\theta)}{2}\psi(\frac{1}{R_{i-1}})\big] \right)[u_i - u_{i-1}] \\ &+ \left(- c^-_{i+1/2}\big[1 - \frac{\theta}{2}\psi(\frac{1}{R_{i+1}}) - \frac{1-\theta}{2}\psi(R_{i+1})\big] - c^-_{i-1/2}\big[\frac{\theta}{2}\psi(\frac{1}{R_i}) + \frac{1-\theta}{2}\frac{\psi(R_i)}{R_i}\big] \right)[u_i - u_{i+1}] \\ &+ \text{y-direction}. \end{aligned} \tag{2.52}$$

We now repeat the previous argument for $\theta = 0$, omitting some of the details. The below conditions

$$\frac{1}{2}R_i\psi(\frac{1}{R_i}) \geq 0, \tag{2.53}$$

$$1 - \frac{1}{2}\psi(\frac{1}{R_{i-1}}) \geq 0, \tag{2.54}$$

$$1 - \frac{1}{2}\psi(R_{i+1}) \geq 0, \tag{2.55}$$

$$\frac{\psi(R_i)}{2R_i} \geq 0, \tag{2.56}$$

are sufficient for the positivity of the coefficients. This can be written more conve-

niently as

$$m_\psi \leq \psi(1/R) \leq 2, \quad 0 \leq S\psi(\frac{1}{S}) \leq M_\psi. \tag{2.57}$$

The timestep restriction for the discrete maximum principle for the $\theta = 0$ scheme is

$$C \leq \frac{2}{2 + M_\psi - m_\psi}. \tag{2.58}$$

We have yet to put some additional sensible constraints on the limiter functions region. $m_\psi \leq 0$ is a design principle that should be adhered to, otherwise $\psi(S) > m_\psi S$, runs into conflict with both the $\psi(R) \leq M_\psi$, $\psi(R) \leq 0$ conditions, when M_ψ is finite and R is large. Good design requires passing through $\psi(1) = 1$ for second order accuracy, so that $M_\psi \in [1, \infty)$ is a sensible construction. $\psi(0) = 0$ is also a common assumption on limiter functions. □

We have come up with a slight distinction from the literature. Here we point out why this is worthwhile in a practical sense.

Remark (Consequence 0). We have found several limiters are not appropriate for incompressible flow. The diagrams Fig. 2.3, allow you to see if the limiter you are using is ok for multidimensional incompressible flow.

Remark (Consequence 1). Having a more general limiter region, is good news to those people who require and design problem specific limiters. We introduce a couple of novel limiters Eq. (2.83), Eq. (2.84) in the new region.

Remark (Consequence 2). Symmetric limiters suitable for incompressible flow quasi-necessarily lie in the Sweby region. This fact can be straightforwardly derived, however Fig. 2.3 can be used to prove this fact graphically. Limiters with the flux limiter symmetry property look identical plotted in r or R, and are invariant under θ, meaning symmetric limiters must lie in both acceptable regions in Fig. 2.3. The overlap of both $\theta = \{0, 1\}$ regions is always contained by the Sweby diagram. We arrive at the conclusion we should push any symmetric limiter into the Sweby diagram, for incompressible flow. This means limiters such as Ospre should be pushed into the Sweby diagram if they are to be used in incompressible flow situations and symmetry is demanded, this defines the following symmetric limiters Eqs. (2.79) to (2.81) we will test these in Section 2.4.1.

Remark (Consequence 3). There are no globally differentiable limiters entirely contained in the second order region when $\theta = 1$ which give a discrete maximum principle for incompressible flow. This is a result of the graph near $R = 0$ and is a disappointing result for those designing implicit methods for which differentiability of the limiter is required. However, when implementing in the $\theta = 0$ form, one can construct analytic $\psi = \tanh(r)e^r \in C^\infty$ limiter functions in the second order region, this may be an important fact for those dealing with implicit systems of the form

$$u_t + (v_1(x, y, t)f(u))_x + (v_2(x, y, t)g(u))_x = 0, \tag{2.59}$$

that have difficulty converging. We introduce a new differentiable limiter Fig. 2.5a which will be more accurate than the analytic one previously defined.

2.3.3 Temporal discretisation

We can define forward Euler numerical flow map $u^{n+1} = \text{Forward Euler}(u^n, c^n)$, as

$$\begin{aligned} u_{i,j}^{n+1} = u_{i,j}^n &- [F_{i+0.5}(u_i^R, u_{i+1}^L, c_{i+0.5}^n) - F_{i-0.5}(u_{i-1}^R, u_i^L, c_{i-0.5}^n)] \\ &- [F_{j+0.5}(u_j^U, u_{j+1}^D, c_{j+0.5}^n) - F_{j-0.5}(u_{j-1}^U, u_j^D, c_{j-0.5}^n)]. \end{aligned} \tag{2.60}$$

Where the time step, Δt, and mesh spacing, Δx, have been absorbed into the face defined velocity c. The SSP33 Runge Kutta scheme, which can be implemented in the following memory efficient (2 register) Shu Osher representation

$$k^1 = \text{Forward Euler}(u^n, c^n), \tag{2.61}$$

$$k^2 = 3/4 \cdot u^n + 1/4 \cdot \text{Forward Euler}(k^1, c^{n+1}), \tag{2.62}$$

$$u^{n+1} = 1/3 \cdot u^n + 2/3 \cdot \text{Forward Euler}(k^2, c^{n+1/2}), \tag{2.63}$$

is a third order, three stage Runge Kutta method with radius of monotonicity of 1. This means that this method preserves convex semi-norms (such as $||\cdot||_\infty$) under the same timestep restriction forward Euler does. The local maximum principle is not preserved in exactly the same way, we actually ensure that each substage satisfies a local maximum principle with respect to the previous substage Appendix A.2.
We will also test the standard Runge Kutta 4 method, as it is well established for both its accuracy and efficiency, and is proposed for this spatial scheme by both [41] and [40]. Koren [41] indicates monotonicity is supposed to be guaranteed by small timesteps, this is not the case, the RK4 algorithm has no Shu Osher representation (without perturbation techniques and downwind/upwind biasing) and has no(zero) radius of monotonicity. Hundsdorfer et al. [40] notes this but indicates the SSP literature is of little practical importance. In [71], similar experiments are performed and quantifies this more specifically with the observation of 10^{-6} negative values [71]. These negatives would be considered not just significant but very large in our application where the positivity of density should not be negative. We consider roughly -10^{-14} the threshold of acceptable monotonicity failure, due to machine precision error in constructing a divergence free vector field, and other accumulation of machine precision errors.

2.3.4 Properties of the scheme.

In this section we will collect and show some known properties of the numerical scheme, this is for convenience of the reader and completeness, as a couple of minor results have previously stated incorrectly in the literature.
When discussing the order of the numerical scheme the analysis depends on the

interpretation of the scheme and the underlying equation being modelled and even the number of dimensions used. Hundsdorfer et al. [40] viewed the scheme as a finite difference method evolving pointwise values $u_{i,j}$ solving the singular point value equation

$$(u)_t + (cu)_x + (\text{y-direction}) = 0. \tag{2.64}$$

where as Zijlema and Wesseling [37] proposed the scheme as a finite volume method on cell averaged quantities $\bar{u}$ so approximates the integral form of the equation

$$\bar{u}_t + \frac{1}{\Delta x}[c(x_{i+1/2})u(x_{i+1/2}) - c(x_{i-1/2})u(x_{i-1/2})] + (\text{y-direction}) = 0. \tag{2.65}$$

The semi-discrete method Section 2.2 interpreted as a pointwise finite difference method approximates the singular point equation Eq. (2.64) to second order accuracy, [40] shows that the method solves the following modified equation

$$\begin{aligned} &u_t + (v_1 u)_x = \\ &-\frac{\Delta x^2}{24}\left[u(v_1)_{xxx} + 3u_x(v_1)_{xx} + 2u_{xx}(v_1)_x\right] + O(\Delta x^3) + (\text{y-direction}). \end{aligned} \tag{2.66}$$

So the scheme is a formally second order accuracy finite difference method approximating the singular point equation in one and two dimensions, unless the flow is directionally constant in which case the scheme can be third order. This should be expected as only two points of velocity are used in each direction for the state interpolated construction.

However, local truncation error analysis as a finite difference method does not tell the full story. When the method is interpreted as a finite volume scheme approximating the integral form of the equation, the formal truncation error analysis becomes different see [74, 75, 76] for explanation. In which case in one dimension the finite volume interpretation of this scheme becomes truly third order with respect to the integral form of the equation (Aleksandar Donev has uploaded lecture notes proving this fact using the symbolic algebra package Mathematica [77]). However, this once again changes in two dimensions because the exact flux is now represented by the one-dimensional flux integral. For example, the right edge flux

$$F_{i+1/2,j} = \frac{1}{\Delta y}\int_{y=y_{j-1/2}}^{y=y_{j+1/2}} c(x_{i+1/2}, y)u(x_{i+1/2}, y)dy. \tag{2.67}$$

is approximated in the finite volume scheme described in Section 2.2 by the midpoint rule (second order Gauss quadrature). The finite volume scheme is a formally second order accurate method solving the integral form in of the equation Eq. (2.65) when in two dimensions.

The formal truncation analysis of the full method is tedious in the more general setting, and instead one designs limiters using simplifications. If it is assumed the flow is uniform constant, and the equation is considered in the finite difference sense.

Then the linear scheme associated with

$$\psi(R) = aR + b, \tag{2.68}$$

has the semi-discrete truncation error

$$\frac{u_i^R - u_{i-1}^R}{\Delta x} - (u_x)_i = [a+b-1]\frac{\Delta x u_{xx}}{2!} + [1-3b]\frac{\Delta x^2 u_{xxx}}{3!} + (a+7b-1)\frac{\Delta x^3 u_{xxxx}}{4!}. \tag{2.69}$$

So that $u \in C^2$ and $a + b = 1$ is sufficient for second order, such that $\psi(R) = aR + (1-a)$ passes through $(1,1)$. $u \in C^3$, $a = 2/3, b = 1/3$ is sufficient for third order, and is a desirable limiter region. These second order straight lines through $(1,1)$ are given in the pictures Fig. 2.3 these lines are more well-known under the name Kappa schemes. This truncation analysis is known and serves as a hugely important design criterion for limiters. We plot the numerical results for the solid body rotational test case for $(a,b) = (0,1), (2/3,1/3), (1,0), (1/2,1/2), (0,0)$ in Figs. 2.9a to 2.9e respectively, under their more common names, second order upwind (SOU), cubic upwind interpolated (CUI), central difference scheme (CDS), Fromm, and first order upwind (FOU).

Spekreijse, showed that sufficient conditions on the limiter function $\psi(R)$ for second order accuracy are $\psi(1) = 1$, $\psi \in C^2$ near 1, [27]. Sufficient and necessary conditions on the limiter function $\psi(r)$ were shown by [36] to be $\psi(1) = 1$, and ψ is Lipschitz continuous.

We also want to understand the effect limiting has on the truncation error particularly in the neighbourhood of smooth extrema where the gradient changes. It is also often mistaken that TVD schemes are first order at noncritical extrema because of a slight oversight in the truncation analysis in Osher 1984 [35], and persists even amongst experts [78]. However, this is not quite the case as explained by Hua-mo [36] where formal truncation analysis of the $\theta = 0$ scheme is done by expanding the higher order correction in the neighbourhood of a critical point $x_\alpha = x_i + \alpha\Delta x$. We summarise the results from [36] below.

Theorem 2.3.2 (constant flow in one dimension near the critical points $\theta = 0$ [36].)**.** *TVD schemes $\theta = 0$ may have second order accuracy at critical points if $\psi(r=3) + \psi(r=-1) = 2$. But cannot be uniformly second-order accurate in the whole neighbourhood of critical points. If $\psi(1) = 1$, then these TVD schemes have second-order accuracy in the region sufficiently far from the critical points of smooth solutions.*

Proof. [36]. □

Similar conclusions for the $\theta = 1$ form is done in [37] however both theorem 3.1 and proof of theorem 3.1, have a few technical inaccuracies and unnecessary assumptions. Nevertheless, all the arguments are correct in spirit, second order at

extrema is possible but second order accuracy is impossible to achieve in the local neighbourhood of extrema. We recommend following the analysis methods of [36] instead.

Theorem 2.3.3 ([36]-for $\theta = 1$)**.** *Let τ_i denote the truncation error about point x_i, let $x_\xi = x_i + \xi\Delta x, \quad \xi \in [0, \infty)$, be a point of interest. TVD schemes may have second order accuracy at critical points if $3\psi(R = 1/3) - \psi(R = -1) = 2$. But cannot be uniformly second-order accurate in the whole neighbourhood of critical points. If $\psi(1) = 1$, then TVD schemes have second-order accuracy in the region sufficiently far from the critical points of smooth solutions.*

Proof. We follow the analysis methods of [36] in the appendix □

The reason this theoretical work is revisited by us here, is because some authors suggest using the second order at extrema condition in the design of limiters [36, 37] and others [79] suggest that satisfying the theoretical second order at extrema condition does not have practical consequences. In our work we have suggested various modifications to limiters including pushing into the Sweby region, this breaks the $3\psi(R = 1/3) - \psi(R = -1) = 2$ condition for the ENO2 limiter.
We now turn to a very important property of the numerical scheme, linear invariance.

Definition 2.3.3 (Linear invariant scheme)**.** The map $u_i \mapsto \alpha w_i + \beta$ leaves the numerical method unchanged.

One consequence of linear invariance is the preservation of constants under incompressible flow and normally desired in atmospheric advection [20], [19]. Linear invariance is more general still and allows temperature to be modelled identically whether in Kelvin or Celsius. In [40] it is incorrectly stated that the scheme is not linear invariant whether in state or flux interpolated form, we show below that the state interpolated form is linear invariant for an incompressible flow.

Theorem 2.3.4 (Linear invariance)**.** *The method described in Section 2.2 is linear invariant.*

Proof. As in [40] the transform $u_i \mapsto \alpha w_i + \beta$ is investigated for linear invariance. However, we do the analysis after the use of a discrete divergence free condition,

$$\frac{\partial u}{\partial t} + \left(\frac{1}{2}\left[c^+_{i+0.5}\psi(R_i)\right] + c^+_{i-0.5}\left[1 - \frac{1}{2}\psi(R_{i-1})/R_{i-1}\right]\right)(u_i - u_{i-1}) \tag{2.70}$$

$$- \left(c^-_{i+0.5}\left[1 - \frac{1}{2}\psi(\frac{1}{R_{i+1}})R_{i+1}\right] + \frac{1}{2}\left[c^-_{i-0.5}\psi(R_i^{-1})\right]\right)(u_i - u_{i+1}) \tag{2.71}$$

$$+ \text{same expression with } i \mapsto j = 0, \tag{2.72}$$

and observe

$$R_i = \frac{(u_{i+1} - u_i)}{(u_i - u_{i-1})} \mapsto \frac{(\alpha w_{i+1} + \beta) - (\alpha w_i + \beta)}{(\alpha w_i + \beta) - (\alpha w_{i-1} + \beta)} = \frac{(w_{i+1} - w_i)}{(w_i - w_{i-1})}, \tag{2.73}$$

$$u_i - u_{i-1} \mapsto \alpha(w_i - w_{i-1}), \tag{2.74}$$

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial w}{\partial t}, \tag{2.75}$$

giving exact linear invariance of the numerical method. So constants are preserved exactly, and the equation is scaling invariant. This holds for both $\theta = 0, 1$. □

2.3.5 Time step restrictions

Our numerical scheme Section 2.2 has a time step restriction for a discrete maximum principle that depends on which limiter is used, whether a directional mean value theorem holds and what temporal discretisation is used. We collate a table Table 2.1 of the maximum cell defined Courant number the scheme has a discrete maximum principle predicted by Theorem 2.3.1, for some of the limiters introduced in the next subsection. This table can be used to conveniently check whether the scheme should be producing positive or discrete maximum principle satisfying results.

Scheme	Limiter	Theorem 2.3.1-C	Spekreijse-C
SSP33/FE	$\psi(R) = 0$	1	1
SSP33/FE	$\psi(R) = aR + b$	0	0
SSP33/FE	vanAlbada(R)	0	$2 - \sqrt{2}$
SSP33/FE	Ospre(R)	0	1/2
SSP33/FE	ENO2(R)	0	1/2
SSP33/FE	vanAlbada$_P$(R)	$\frac{4}{5+\sqrt{2}}$	$\frac{4}{5+\sqrt{2}}$
SSP33/FE	Ospre$_P$(R)	4/7	4/7
SSP33/FE	minmod(R)/ENO2$_P$(R)	2/3	2/3
SSP33/FE	Koren(R)	1/2	1/2
SSP33/FE	Woodfield(R,M,m)	$2/(2 + M - m)$	$2/(2 + M - m)$
SSP33/FE	Differentiable(r)	$\frac{2}{4+\sqrt{\frac{5\sqrt{5}}{2}-\frac{11}{2}}} \approx 0.4651$	$\frac{2}{4+\sqrt{\frac{5\sqrt{5}}{2}-\frac{11}{2}}} \approx 0.4651$
RK4	Koren(R)	0	0

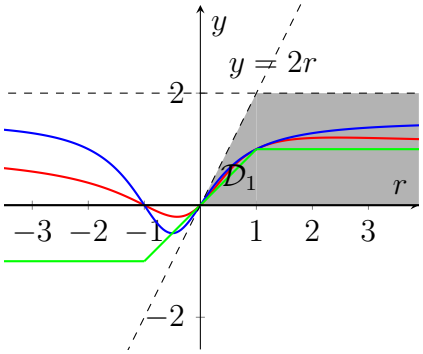
Table 2.1: This table contains the sufficient Courant number restrictions for temporal discretisation of Section 2.2 to satisfy a local discrete maximum principle for different limiters. Column one is the sufficient Courant number limit for incompressible flow and column two is the sufficient Courant number limit for directional mean value theorem satisfying flows. These time stepping criteria hold in more dimensions; however the definition of the Courant number changes.

2.3.6 Symmetric limiters: Old and New

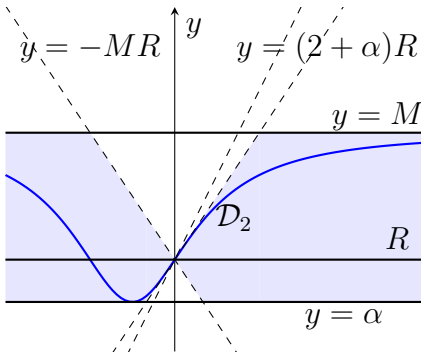
We first introduce some important symmetric flux limiter functions in Speckreijse admissible region, but not in the Sweby admissible region

$$\text{van-Albada}(R) = \frac{R^2 + R}{R^2 + 1}, \tag{2.76}$$

$$\text{Ospre}(R) = \frac{3}{2}\frac{R^2 + R}{R^2 + R + 1}, \tag{2.77}$$



(a) Sweby admissible region, as well as the ENO2 scheme in green, Ospre in blue, and van Albada in red



(b) Spekreijse limiter region for the Ospre scheme.

Figure 2.4: $\mathcal{D}_1$ is the Sweby region, $\mathcal{D}_2$, is the Spekreijse region, which has two free parameters $\alpha \in [-\infty, 0]$, $M \in (0, \infty)$. We can see that the limiters, van Albada (red) Eq. (2.76), Ospre (blue) Eq. (2.77), and ENO2 (green) Eq. (2.78) are not contained in the Sweby region $\mathcal{D}_1$ (or even the new incompressible flow limiter region $\mathcal{D}_4$ Fig. 2.3), but are in the Spekreijse region $\mathcal{D}_2$ with values for (M, α) given by $[1.5, -0.5]$, $[1/2(1-\sqrt{2}), 1/2(1+\sqrt{2})]$, $[-1, 1]$ respectively. We have only plotted this region for the Ospre limiter in Fig. 2.4b.

$$\text{ENO2}(R) = \begin{cases} R & \text{where} \quad |R| \leq 1, \\ 1 & \text{where} \quad |R| \geq 1. \end{cases} \tag{2.78}$$

These are introduced in the respective papers [80, 79, 30] and are plotted in Fig. 2.4a for convenience. We also introduced the following subscript P limiters, as those positive coefficient limiters who are "pushed" into being in the Sweby diagram by the removal of the tail, these limiters are designed keep the symmetry property and fix the monotonicity problems of the previous limiters in accordance with Section 2.3.2.

$$\text{van-Albada}_P(R) = \begin{cases} \frac{R^2+R}{R^2+1}, & R \geq 0 \\ 0, & R < 0 \end{cases}, \tag{2.79}$$

$$\text{Ospre}_P(R) = \begin{cases} \frac{3}{2}\frac{R^2+R}{R^2+R+1}, & R \geq 0 \\ 0, & R < 0 \end{cases} \tag{2.80}$$

$$\text{ENO}_P(R) = \text{minmod}(R) = \max(0, \min(R, 1)) \tag{2.81}$$

2.3.7 Non symmetric limiters and symmetry breaking

We also introduce the Koren [41] limiter,

$$\text{Koren}(R) = \max(0, \min(2, 2R, 2R/3 + 1/3)). \tag{2.82}$$

It is not symmetric, but is an accurate limiter consisting of restricting the third order upwind region to the Sweby region. We introduce a new limiter with free

parameters M_ψ, m_ψ defined by

$$\text{Woodfield}(R, M_\psi, m_\psi) = \begin{cases} \left\{R \leq -1 : \frac{R+R^2}{1+R^2}\mathcal{I})\right\} \\ \left\{-1 < R < -\frac{1}{2} : 0, \right\} \\ \left\{-\frac{1}{2} < R < -\frac{1}{-3m_\psi - 2} : \ \frac{2}{3}R + \frac{1}{3}\right\} \\ \left\{\frac{1}{-3m_\psi - 2} < R < 0 : \ m_\psi R\right\} \\ \left\{0 < R < \frac{1}{4} : \ 2R\right\} \\ \left\{\frac{1}{4} < R < \frac{3M_\psi - 1}{2} : \ \frac{2}{3}R + \frac{1}{3}\right\} \\ \left\{\frac{3M_\psi - 1}{2} < R : \ M_\psi\right\} \end{cases} \tag{2.83}$$

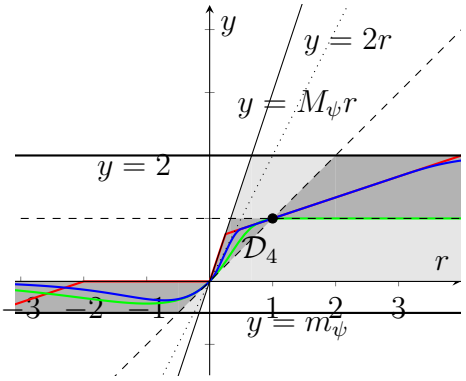
consisting of restricting the third order upwind region to the new more general limiter region. We also add on the tail of the van-Albada scheme as an optional test using an indicator function $\mathcal{I}$, this is arbitrary and for numerical experiments only. We also introduce the SuperbeeR limiter,

$$\text{SuperbeeR}(R, d, M_\psi = 3, m_\psi = -1) = \begin{cases} \max(0, \max(\min(2R, 1), \min(R, M_\psi))) : R \geq 0 \\ \min(m_\psi R, 1) \quad \text{where} : (R < 0) \end{cases} \tag{2.84}$$

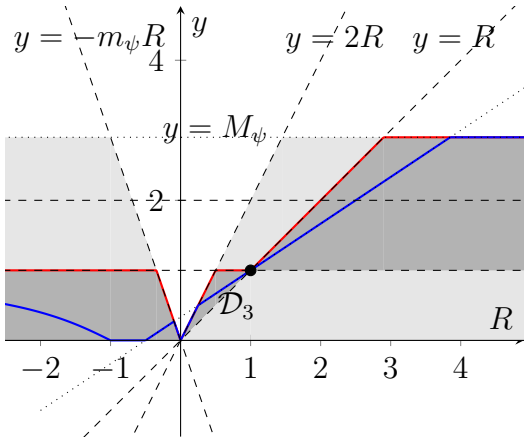
which serves as an extension of traditional Superbee limiter to the new limiter region. This is to test our limiter region but could have application to front tracking and free surface flows. Both these limiters are plotted for convenience in Fig. 2.5b.
We can define a similar extension to the Koren limiter by restricting the third order upwind region to the $\theta = 0$ maximum principle limiter region for some M, m creating the Woodfield(r, M, m) limiter, plotted in Fig. 2.5a.
We introduce the first(to our knowledge) globally differentiable limiter function contained entirely within the second order region, suitable for incompressible flow. It touches the third order region for accuracy.

$$\text{Differentiable}(r) = \begin{cases} \tanh(r)\exp(r), \quad \text{where} \quad r \leq 0, \\ -8r^3 + 16/3r^2 + r, \quad 0 < r \leq 1/2 \\ 1/3r + 2/3, \quad 1/2 < r \leq 3 \\ 1/3\tanh(r-3) + 5/3, \quad r > 3 \end{cases} \tag{2.85}$$

plotted in Fig. 2.5a. This limiter is a placeholder limiter that could easily be improved on/replaced with the proposal of piecewise polynomials with the same properties.
We are not advocating for such limiters at this point, just demonstrating our theory is correct and may have uses.



(a) When we are using the Sweby implementation $\theta = 0$, we plot the new Differentiable(r) limiter in blue, the analytic limiter $\tanh(r)\exp(r)$ in green, and the Woodfield($r, M = 3, m = -1$) limiter in red.



(b) Woodfield($R, M = 2.9, m = 2.9$) blue, and SuperbeeR($R, M = 2.9, m = 2.9$) red.

Figure 2.5: New limiters in their respective $\theta = 0, 1$ regions.

2.4 Numerical Demonstrations: Test setup and results

We have two distinct types of flow (Mean value theorem satisfying, and Mean value theorem violating), three different limiter regions (Spekreijse, Sweby, and our new limiter region(s)Fig. 2.3) and we also have two types of time stepping algorithms to test, SSP33, and RK4. We introduce two monotonicity tests, solid body rotation in which a mean value theorem holds in each direction and all limiters should work for by [27], and a deformational one used to verify that the Spekreijse limiter region is not appropriate for incompressible flow, but our new limiter regions are. We also introduce the solid body rotation test case of the LeVeque initial conditions to compare the accuracy results of the new limiter functions.

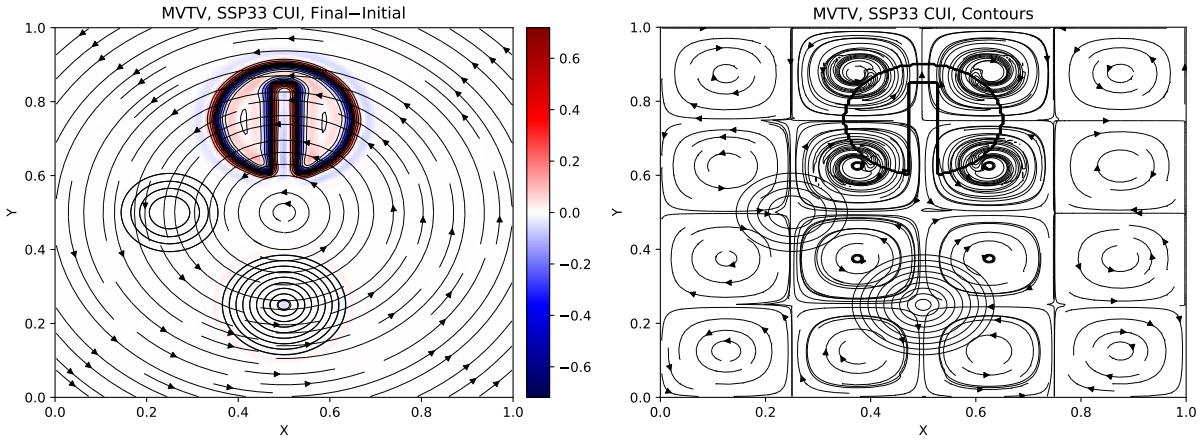
2.4.1 Setup: monotonicity tests

The numerical domain is $\Omega = [0, 1] \times [0, 1]$ discretised by 200×200 cells, with periodic boundary conditions. We run the scheme with 4000 timesteps resulting in a Courant number maximum of around 0.3. The velocities and fluxes are located at the midpoints of faces $(i + 1/2, j), (i, j + 1/2)$, and we only consider normal components. We locate the stream function $\Psi(x, y)$ at the cell vertices and then use a discrete form of the curl operator to create a divergence free vector field (to machine precision). For the initial stream functions, we use the following functions

$$\Psi = -\pi((x - x_c)^2 + (y - y_c)^2), \tag{2.86}$$

$$\Psi = \frac{1}{2}\sin(4\pi x)\sin(4\pi y). \tag{2.87}$$

The well-known solid body rotation flow Eq. (4.32) has a directional mean value theorem applying (with the directionally constant property with $c_i = c_{i+1/2,j} =$



(a) Solid Body rotation of the Le-Veque initial conditions, this flow satisfies the directionally constant property. This flow returns to the inital conditions, this is observed, and the difference is plotted.

(b) Sin deformation of the Le-Veque initial conditions. This flow coils up and deforms the initial conditions of Le-Veque far from their initial configurations. This test will be used for testing positivity preservation of different limiters when the flow does not satisfy the directional mean value theorem property. As will be shown in Fig. 2.8a.

Figure 2.6: These figures are generated from the two monotonicity tests and have used the SSP33 initial conditions and the CUI scheme. Plotted are the initial and final contours at levels $0.1, 0.2, 0.3, ...0.9, 1.0$, the streamlines are plotted, and the colours denote the difference between the final and initial condition.

$c_{i-1/2,j} = -\frac{\Delta t}{\Delta x}\frac{\pi}{2}y_j$, $c_j = c_{i,j+1/2} = c_{i,j-1/2} = \frac{\Delta t}{\Delta y}\frac{\pi}{2}x_i)$, the sinusoidal deformationEq. (2.87) does not have a directional mean value theorem. For the initial condition of the tracer we use the challenging LeVeque initial conditions [81] , defined below

$$q = \begin{cases} 1 & \sqrt{((x-0.5)^2+(y-0.75)^2))} \leq 0.15, \text{and}(x \leq 0.475) \\ 1 & \sqrt{((x-0.5)^2+(y-0.75)^2)} \leq 0.15, \text{and}(x > 0.525) \\ 1 & \sqrt{((x-0.5)^2+(y-0.75)^2)} \leq 0.15, \text{and}(y \geq 0.85), \\ & \text{and}(0.475 < x \leq 0.525) \\ (1-\frac{R_{cone}}{0.15}) & \text{for} \quad (R_{cone} = \sqrt{((x-0.5)^2+(y-0.25)^2)} \leq 0.15) \\ \frac{1}{2}(1+\cos(\pi\frac{R_{cos}}{0.15}) & \text{for} \quad (R_{cos} = \sqrt{((x-0.25)^2+(y-0.5)^2)} \leq 0.15) \\ 0 & else, \end{cases} \tag{2.88}$$

These convergence tests have been visualised for the unlimited SSP33 CUI scheme in Fig. 2.6a, Fig. 2.6b at 128×128 resolution and the unlimited scheme is sufficiently accurate.

2.4.2 Setup: convergence tests

We define the convergence test suite by constructing four fields and a sufficiently smooth initial condition. Convergence test one is diagonally constant flow with doubly periodic boundary conditions Eq. (2.89). Convergence test two is a time reversing quadratic deformation Eq. (2.90). Convergence test three is a time reversing sine deformation Eq. (2.91). Convergence test four is solid body rotation Eq. (2.92). These velocity fields are respectively defined in terms of their stream functions as follows

$$\Psi = (y - x), \tag{2.89}$$
$$\Psi = 8\pi x(x-1)y(y-1)\cos(\pi t), \tag{2.90}$$
$$\Psi = \frac{1}{2}\sin(2\pi x)\sin(2\pi y)\cos(\pi t), \tag{2.91}$$
$$\Psi = -\pi((x-x_c)^2 + (y-y_c)^2). \tag{2.92}$$

The initial conditions used to test convergence is the following compact C^4 widened cosine bump located at $0.5, 0.75$

$$\begin{aligned} u_0 &= \frac{1}{4}(1+\cos(\pi r))^2, \\ r &= \frac{1}{0.25}\min(((x-0.5)^2 + (y-0.75)^2)^{0.5}, 0.25). \end{aligned} \tag{2.93}$$

The initial condition streamlines and several contours for each numerical simulation of the SSP33 CUI unlimited scheme are plotted in Fig. 2.7 at resolution 128×128 where the colour map denotes the difference between the initial and final initial condition. All convergence tests are done at 16×16, 32×32, 64×64, 128×128 resolution, where the Courant number is held constant around 0.2 by decreasing the time step proportionally to the mesh refinement.

2.4.3 Numerical results: Description and Conclusions

In Fig. 2.8b we plot the minimum values produced $\min_{i,j} u^n_{i,j}$ under the solid body rotation flow of the LeVeque initial conditions Fig. 2.6a. This is done for the limiters, Ospre, Van Albada, ENO2, OspreP, Van AlbabaP, ENO2P, Koren, Woodfield, and the new differentiable limiter, (Eqs. (2.76) to (2.83) and (2.85)) with the SSP33 scheme. We also plot the minimum values produced for the RK4 scheme with the Koren limiter Eq. (2.82). We observe an initial negative value of order 10^{-8} for the RK4 scheme with the Koren limiter. This gives evidence that the SSP literature is of practical importance, this is contrary to the conclusion in [40]. We observe all the schemes with SSP timestepping including Ospre and Van Albada remain positive for the directionally constant flow test case (solid body rotation), this provides evidence that Spekreijse's theory does hold for directionally mean value theorem satisfying (MVTS) test cases as indicated in Section 2.3.1.

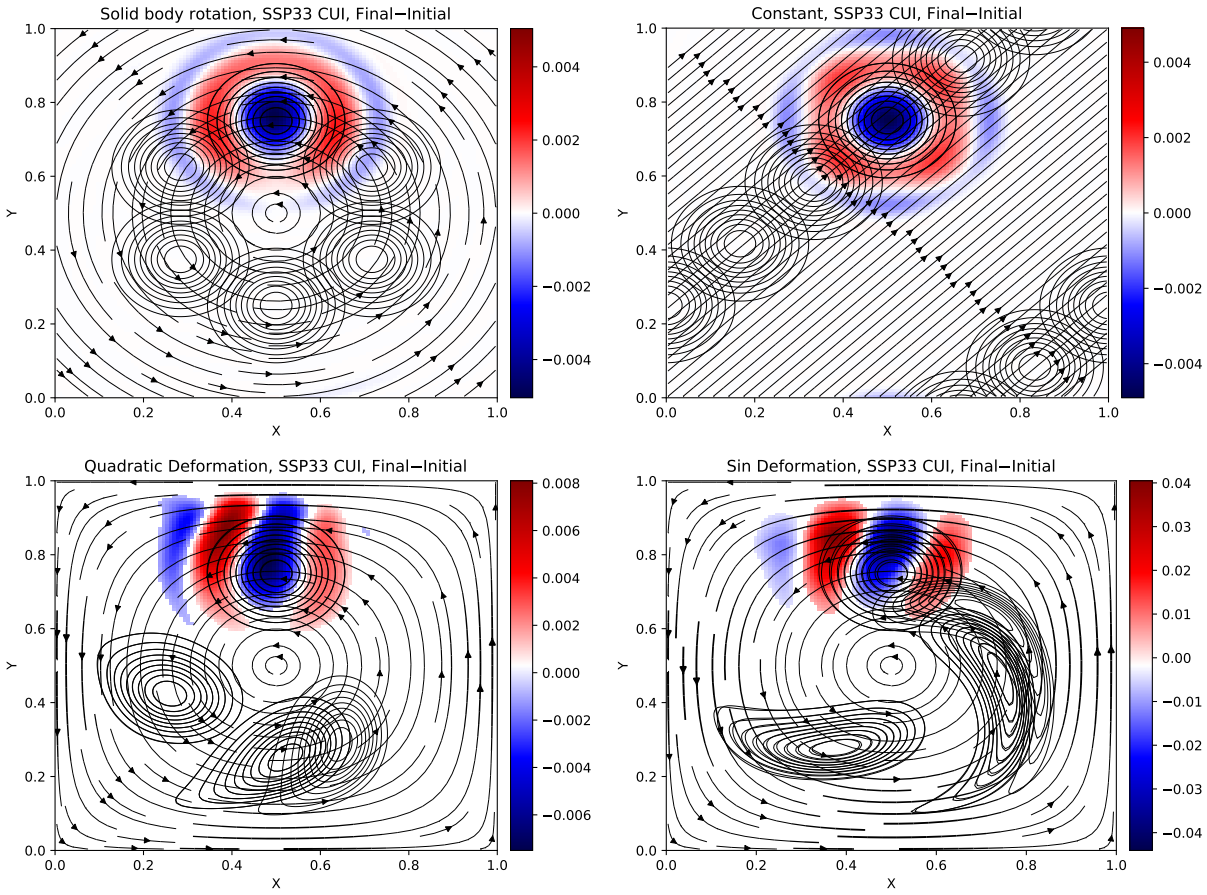
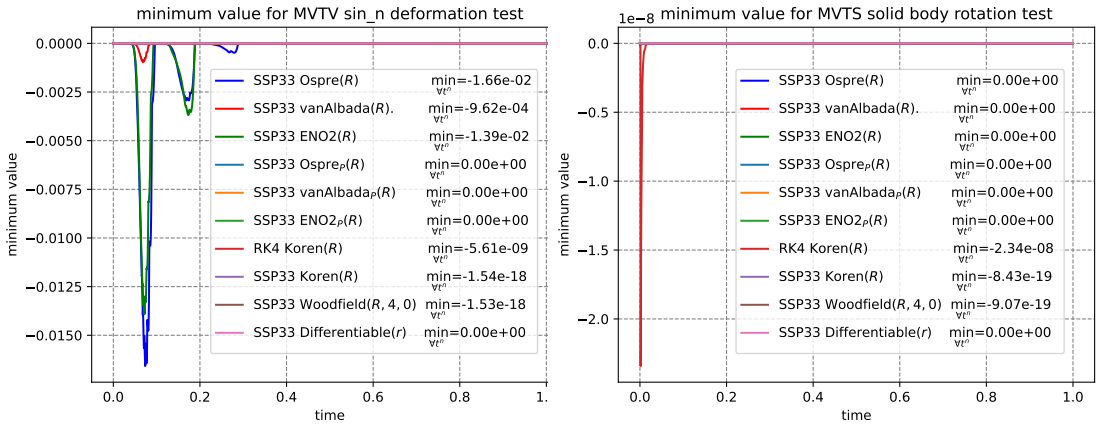


Figure 2.7: SSP33 CUI scheme under the different convergence test suites at 128×128 resolution, we plot tracer contours at every 1/6th of the test case at the tracer values $0.1, 0.2, 0.3, ..., 1$. We plot the initial condition streamlines, the colour represents the difference between initial and the final timestep. The unlimited scheme produces accurate enough results to propose it for a potential advection algorithm for dynamical cores.

As discussed in Section 2.3.1 we need to test a flow which is not directionally constant and may be mean value theorem violating (MVTV). This is to assess the suitability of limiters for more general incompressible flow. We use the sinusoidal incompressible flow defined in Eq. (2.87) on the same LeVeque initial conditions. The minimum values for several limiters are plotted in time in Fig. 2.8a. The ENO2, Ospre, and the van-Albada limiters Eqs. (2.76) to (2.78) all fail to preserve positivity with significant negative values of order 10^{-2} appearing. The negative values produced in Fig. 2.8a for the ENO, Ospre, and Van Albada limiter Eqs. (2.76) to (2.78) are one of the most important numerical results in this chapter, and is evidence that the Spekreijse region is not appropriate for truly incompressible flows. We also see in Fig. 2.8a and the Table 2.2 that versions of these limiters pushed into the Sweby region (with subscripts P) Eqs. (2.79) to (2.81) have no observed negative values at any point in time. These numerical results align with the theoretical conclusion that the Sweby region is appropriate for incompressible flow as explained in Section 2.3.2. We also see in Fig. 2.8a and the Table 2.2, that the new limiters Woodfield$(R, 4, 0)$, Differentiable(r) also remain positive over all time. This gives evidence that limiters which are strictly not contained in the Sweby region but are in the newly derived larger limiter regions Fig. 2.3 are suitable for incompressible flow and gives numerical evidence that the Theorem 2.3.1 holds.
We also collate the minimum value over all space and time for both the sinusoidal flow and the solid body rotation case in the Table 2.2. This is to check for minimum values too small to be visible in Figures 2.8a and 2.8b and to introduce a few extra limiters. If one compares Table 2.2 to Table 2.1, we see that all the theoretical predictions made by the strong stability preserving literature and predictions made in this chapter regarding suitable limiter regions in Section 2.3.2 and Theorem 2.3.1 are observed numerically to machine precision. More specifically we observe for strong stability preserving schemes, that the positivity is preserved to machine precision for the Sweby region limiters, and both of the new limiter regions, but not for some traditional limiters in the Spekreijse region, unless the flow is directionally constant.

In Fig. 2.9 we plot the results of the solid body rotation Eq. (2.92) test case of the LeVeque initial conditions Eq. (2.88), for the five unlimited schemes, associated with SOU, CUI, CDS, Fromm and FOU (first order upwind). These methods are plotted in the limiter regions in Fig. 2.3 and defined explicitly in Eq. (2.68). We observe that all linear schemes with order greater than one have visible negative values. First order upwind does not produce negative values but it is not accurate enough to resolve the various shapes. Furthermore, there is a noticeable enhancement in accuracy and shape preservation using the CUI method. The plots of the unlimited schemes under solid body rotation in Fig. 2.9 are useful design criteria for designing and motivating a new limiter. This observation lead to the design improvement found in the new limiters Eq. (2.85), Eq. (2.83) by trying to attain the accuracy of the unlimited CUI scheme. The fact that the CUI scheme performs more accurately



(a) Minimum values attained in the non-reversing sinusoidal deformational case Eq. (2.87).

(b) Minimum values attained in the LeVeque solid body rotational test case Eq. (2.92).

Figure 2.8

is perhaps explained by Eq. (2.66), where the method can be interpreted as a finite difference method approximating the pointwise equation Eq. (2.64) with a vanishing third order truncation error when the flow is directionally constant.

In Fig. 2.10 we plot the results of the solid body rotation test case of the LeVeque initial conditions at the final timestep, for the Ospre(R) limiter, Ospre$_P(R)$ limiter and the new Differentiable(r) limiter, defined by Eq. (2.77), Eq. (2.80) and Eq. (2.85). All results are bounded and shape preserving. The difference between Ospre(R) limiter and the Ospre$_P(R)$ limiter is negligible. This was a surprise, as some literature suggests that the use of limiters out of the Sweby region could resolve peaks better because the limiter is non zero when the gradient changes [36, 37]. In Fig. 2.10 we observe that the new Differentiable(r) limiter resolves the cone and the slotted cylinder notably more accurately, than both the Ospre(R) and the Ospre$_P(R)$ limiter, and unlike Ospre$_P(R)$ the Differentiable(r) limiter is differentiable, and unlike the Ospre(R) limiter the Differentiable(r) limiter is suitable for truly incompressible flow.

In Fig. 2.11 we plot the numerical solution at the final timestep of the solid body rotation test case of the LeVeque initial conditions for the new limiters Woodfield$(R, 4, 0)$, Woodfield$(R, 2, -2)$ alongside the well-known Koren(R) limiter. We observe boundedness and shape preservation in all three cases. There is preferable resolution of the slotted cylinder and cone peak in the Woodfield$(R, 4, 0)$ limiter. This gives numerical evidence that there may be accuracy benefits in using the newly defined limiter regions in Fig. 2.3. There is no discernable difference between the Woodfield$(R, 2, -2)$ and the Koren Limiter. This, provides more numerical evidence that the negative gradient $\theta = 1$ region was not particularly helpful in gaining additional accuracy.

In Fig. 2.12 we plot the final time step solution of the LeVeque initial conditions under the solid body rotation flow for the novel limiter SuperbeeR$(R, 4, 0)$, alongside the well-known Superbee(R) limiter. We do not observe any unboundedness larger than machine precision in Fig. 2.12. In Fig. 2.12 we see boundedness and some shape

preserving properties for both limiters; however, we also see the famous terracing and over compressive behaviour typically associated to the Superbee limiter on both the cone and on the cosine bell. This results in poor representation of smooth peaks, and the slotted cylinder being represented suspiciously well. The extended Superbee$R$ limiter is even more compressive than the traditional Superbee limiter, as can be seen at the peak of the cone, but can also be observed at the slotted cylinder when zoomed in. Both Superbee limiters have been tested for monotonicity and positivity under the sin deformational flow and also observe boundedness for all time in Fig. 2.8a. Both are too compressive to be seriously suggested in dynamical core algorithms, but may have applications in front tracking.

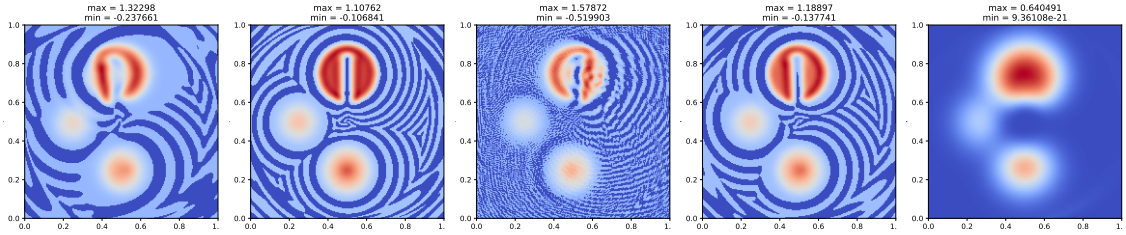
In Table 2.3 we display the computed relative error norms of the solution u compared to the exact solution u_e defined by $re(u) := \frac{||u-u_e||_p}{||u_e||_p}$ for $p = 1, 2, \infty$ for the limiters in Eqs. (2.76) to (2.78) and their modified pushed forms in Eqs. (2.79) to (2.81) for the solid body rotation of the LeVeque initial conditions. In Fig. 2.13 we plot horizontal cross sections through the midpoint of each advected cosine, cone and slotted cylinder shape after the solid body rotation of the LeVeque initial conditions. Out of all these limiter functions Ospre and Ospre$_P$ are the most accurate. The push into the Sweby region does not appear to decrease the accuracy of these traditional limiters as observed in L^1, L^2, L^∞ in Table 2.3 and to the eye in Fig. 2.13. Indicating the push of Spekreijse limiters into the Sweby region does not result in a loss in accuracy. The ENO2 and the minmod(ENO2$_P$) scheme perform similarly at the peaks, despite the fact that the ENO2 scheme satisfies the second order at smooth extrema condition $3\psi(1/3)-\psi(-1) = 2$, but the minmod(ENO2$_P$) does not. This indicates that the theoretical result regarding second order accuracy at critical extrema does not translate to better accuracy at peaks, which could be attributed to the degradation of accuracy within some region of the extrema condition [36].

In Fig. 2.14 we plot the horizontal cross sections through the midpoint of each advected shape; cosine, cone and slotted cylinder, at the final time step of the solid body rotation test of the LeVeque initial conditions. We do this for the limiters Eq. (2.82), Eq. (2.77), two versions of the Eq. (2.83) limiter, as well as the new globally differentiable limiter Eq. (2.85), all under the SSP33 scheme. In Fig. 2.14 we observe that the new differentiable limiter is almost as accurate as the Koren limiter, it suffers a very slight loss of accuracy at the peaks. The Woodfield$(R, 2, -2)$ performs practically indiscernibly to the eye to the Koren(R) limiter. The Woodfield$(R, 4, 0)$ limiter has slightly better accuracy than the Koren limiter, most notable at the peaks. This indicates that additional accuracy can be gained using the new limiter regions.

We plot the results of all convergence test cases in Fig. 2.15, defined in Section 2.4.2 and visualised in Fig. 2.7. The results are arranged in a 2 by 4 grid of sub-figures, where each row corresponds to a different flow test case, and figures in the same row but different column differ only with regard to the schemes and limiters being used. Row one contains the convergence of the cosine squared bump Eq. (2.93) under the

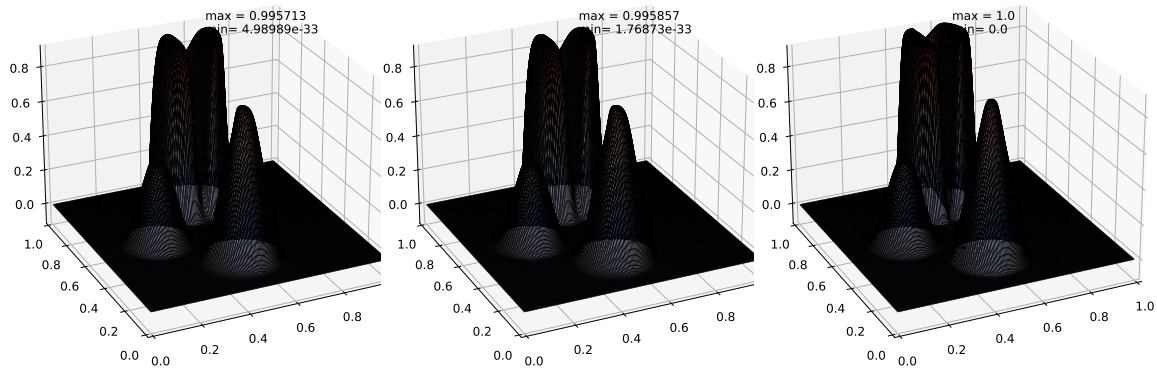
Scheme	Limiter	MVTV-sin $\min_{\forall n,i,j} u^n_{i,j}$	MVTS-sbr $\min_{\forall n,i,j} u^n_{i,j}$
SSP33	vanAlbada(R)	$\mathbf{-9.62151 \times 10^{-4}}$	0.0
SSP33	Ospre(R)	$\mathbf{-1.65800 \times 10^{-2}}$	0.0
SSP33	ENO2(R)	$\mathbf{-1.39113 \times 10^{-2}}$	0.0
SSP33	vanAlbada$_P$(R)	0.0	0.0
SSP33	Ospre$_P$(R)	0.0	0.0
SSP33	minmod(R)	0.0	0.0
SSP33	Koren(R)	-2.36110×10^{-18}	-1.54498×10^{-18}
SSP33	Woodfield(R)	-2.66384×10^{-18}	-2.01525×10^{-18}
SSP33	Differentiable(r)	0.0	0.0
RK4	Koren(R)	$\mathbf{-5.32587 \times 10^{-9}}$	$\mathbf{-2.29303 \times 10^{-8}}$

Table 2.2: This table contains the minimum values attained over all points over all time of the experiments. For both the Mean Value Theorem Violating flow (MVTV), and Mean Value Theorem Satisfying flow(MVTS). The important points are: the RK4 scheme is not positivity preserving but performs quite well. Limiters in the Spekreijse region strictly out of the new limiter regions Fig. 2.3 are no longer appropriate for incompressible flow, see the first 3 values. Bold indicates significant positivity violation.



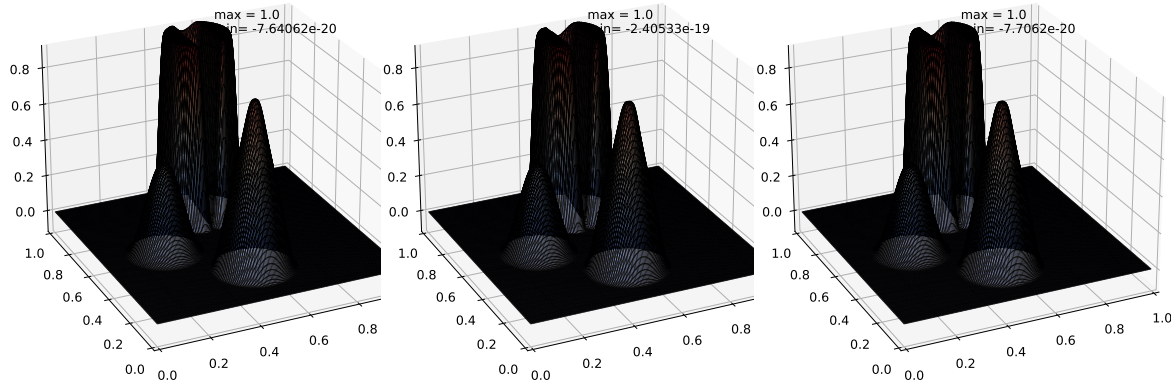
(a) SSP33:SOU (b) SSP33:CUI (c) SSP33:CDS (d) SSP33:Fromm (e) SSP33:FOU

Figure 2.9: Solid body rotation of LeVeque initial conditons for SOU, CUI, CDS, Fromm and first order upwind, at the final time of a $200 \times 200 \times 2000$ resolution simulation. Any negative values below negative $-1e-14$ will be plotted as if -0.5 and will shift the entire colour range, so that the colour scheme highlights negatives should they appear.



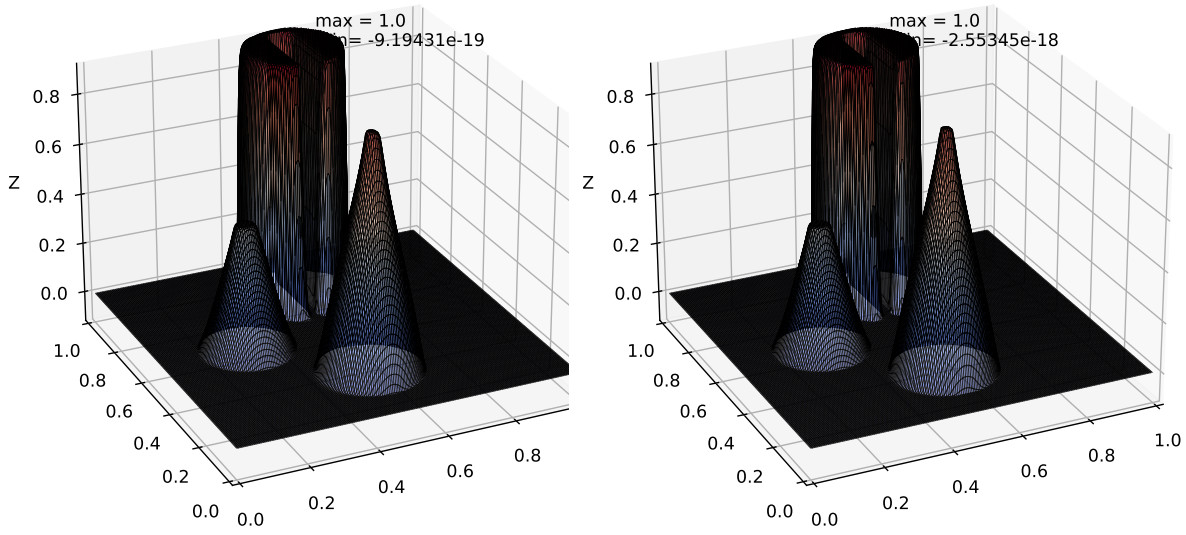
(a) SSP33:Ospre(R) (b) SSP33:Ospre$_P$(R) (c) SSP33:Differentiable(r)

Figure 2.10: Solid body rotation of the LeVeque initial conditions, at 200×200 resolution for the SSP33 scheme with the Ospre(R), Ospre$_P$(R)and the Differentiable(r) limiters respectively.



(a) SSP33:Woodfield$(R, 4, 0)$ (b) SSP33:Woodfield$(R, 2, -2)$ (c) SSP33:Koren(R)

Figure 2.11: Solid body rotation of the LeVeque initial conditions, at 200×200 resolution for the SSP33 scheme with the Woodfield$(R, 4, 0)$, Woodfield$(R, 2, -2)$ and the Koren(R) limiters respectively.



(a) SSP33:Superbee (b) SSP33:SuperbeeR$(R, 3, -1)$ limiter

Figure 2.12: Solid body rotation of the LeVeque initial conditions, at 200×200 resolution for the SSP33 scheme with the Superbee(R) and the SuperbeeR$(R, 3, -1)$ limiters respectively.

diagonally constant flow Eq. (2.89) flow, row two uses the quadratic time reversing deformational flow Eq. (2.90), row three uses the sinusoidal time reversing deformational flow Eq. (2.91), and row four uses solid body rotation Eq. (2.92). The first column contains the Ospre(R),van Albada(R) and the ENO2(R) one-dimensional limiters, alongside their pushed counterparts Ospre$_P(R)$, van Albada$_P(R)$ and the ENO2$_P(R)$ and the new Differentiable(r) limiter all run using the SSP33 timestepping scheme. The second column of Fig. 2.15 contains different schemes, including the Koren limiter with both RK4 and SSP33, the linear schemes CUI, Fromm, the first order upwind scheme and the Woodfield(R, M, m), Eq. (2.83) limiters.
The convergence results in Fig. 2.15 are perhaps not run far enough into the asymp-

Scheme	Limiter	L_1	L_2	L^∞
SSP33	van Albada(R)	0.254469	0.309882	0.811324
SSP33	van Albada$_P$(R)	**0.254296**	**0.309748**	**0.811290**
SSP33	Ospre(R)	0.231790	0.295968	**0.804238**
SSP33	Ospre$_P$(R)	**0.231324**	**0.295734**	0.804449
SSP33	ENO2(R)	0.350092	0.366133	0.819102
SSP33	ENO2$_P$(R)	**0.349999**	**0.366052**	**0.818404**

Table 2.3: This table contains the relative error norms of the symmetric limiters, and there pushed counterparts for the solidbody rotation of the LeVeque initial conditions, at 200×200 resolution for the SSP33 scheme. We see that the new limiters do about the same in most error norms. Bold font is used to distinguish the smaller of the error when comparing the limiter and its pushed counterpart.

totic regime to conclude the formal order of all the methods, however convergence rates can be compared amongst each other, and second order is certainly being approached. Conveniently, because in Fig. 2.15 no lines significantly cross each other, the final gradient computed by $\log(re_{L^2}(u_{128^2})/re_{L^2}(u_{64^2}))\log(2)^{-1}$ serves as a simple measure of the order and informally indicates the accuracy of the respective numerical methods. We collate the convergence rate of all schemes under all test cases in Table 2.4, as a more convenient method of comparing accuracy and convergence order between schemes.

In the convergence plots Fig. 2.15b, Fig. 2.15d, Fig. 2.15f and Fig. 2.15h, and line 15 in Table 2.4 we observe that the first order upwind numerical scheme convergences slowly and doesn't reach near first order convergence in our test cases. We observe the Fromm scheme without any limiting procedure observes near second order convergence as expected, but has not reached this asymptotic convergence for the more challenging test case of sinusoidal time reversing deformational flow, which requires more resolution. However, the CUI scheme without limiting appears to have achieved second order in these flow cases, and near third order convergence behaviour for the diagonal and solid body rotation case, in line 13 of Table 2.4, and remains the most accurate scheme for all convergence plots, as seen in Fig. 2.15b, Fig. 2.15d, Fig. 2.15f and Fig. 2.15h. The success of the CUI's convergence further indicates the utility of designing limiters near the third order region when possible. For the Sin Deformation flow, the true second order behaviour is observed, perhaps exposing the formal order limitations of using one gauss point for the numerical flux (yet to be determined). In Figure 2.15 we occasionally see preferable convergence and error for the Koren/Woodfield/Differentiable limited schemes, over the unlimited Fromm scheme.

In the first 6 lines in Table 2.4 and also the plots of relative error in Figures 2.15a, 2.15c, 2.15e and 2.15g(column one), we notice that there is no discernible change to the order of convergence or relative error in all four convergence test cases Eqs. (2.90), (2.91), (4.32) and (4.33), between the traditional limiters (ENO2(R), Ospre(R), van-Albada(R)) Eqs. (2.76) to (2.78) in the Spekreijse region and the

pushed limiters ENO2$_P(R)$, Ospre$_P(R)$, van-Albada$_P(R)$) Eqs. (2.79) to (2.81) strictly in the Sweby diagram. We also observe the Ospre limiters outperform the van Albada limiters, which in turn outperforms the ENO2 limiters. This is seen in the plots of error in Fig. 2.15a, Fig. 2.15c, Fig. 2.15e, Fig. 2.15g, as well as the order displayed in the first 6 lines in the table Table 2.4. In the 7th line in Table 2.4 and over all convergence plots in Fig. 2.15, we observe that the Differentiable limiter Eq. (2.85) has better convergence and error than the ENO2(R), Ospre(R), van-Albada(R) Eqs. (2.76) to (2.78) limiters and approaches the order and accuracy of the Koren(R) limiter.

In Fig. 2.15b, Fig. 2.15d, Fig. 2.15f and Fig. 2.15h and lines 8,9 in Table 2.4 we observe that there is little difference in accuracy or convergence order between the RK4 and SSP33 timestepping methods. This indicates that the spatial error dominates the total numerical error, and little is gained by using a higher order in time integration scheme for the choice of resolution and tests described here. Because the Courant number is near the Maximum permissible time step this observation may be useful in optimising computational cost, it indicates the cheaper SSP33 scheme gets similar levels of accuracy to RK4, and the cheaper SSP22 scheme may be worth investigating.

In the convergence plots Fig. 2.15b, Fig. 2.15d, Fig. 2.15f and Fig. 2.15h, and lines 9,10,11,12 in Table 2.4 we observe that Woodfield$(R, 4, 0)$ limiter can gain accuracy and better convergence order than the Koren limiter but Woodfield$(R, 2, -2)$ does not. This implies there can be improved convergence and accuracy using methods in the new extended limiter region, but requires careful design of the limiter, adding on part of a van-Albada tail arbitrarily was not particularly beneficial, but touching the third order region for more of the limiter region (R, y) was helpful. In the $\theta = 1$ framework the Woodfield$(R, 4, 0)$ did produce better results than the Koren limiter, however the improvement in accuracy comes at a reduced timestep Table 2.1. It is an open question to what value Woodfield$(R, M, 0)$, $M \in [1, \infty]$ is most computationally efficient as there is a trade-off between accuracy/order and time step.

2.5 Conclusion

In a lot of applications users of flux limiters are not concerned about violating condition Eq. (2.4) ($\psi(r) = 0, r \leq 0$) by appealing to practicality [82], [83]. More formal justification, of violating condition (2.4) and Eq. (2.3) can be invoked using Spekreijse's extended limiter region and using Spekreijse's monotonicity theory. However, this only applies when one uses the flux difference splitting or the flux vector splitting frameworks. When the equation has an explicit dependence on an incompressible velocity field, and one uses a flux form method, we have demonstrated that violating Eq. (2.4) and using the Spekreijse's extended limiter region for some commonly used limiter functions can cause serious problems in practice.

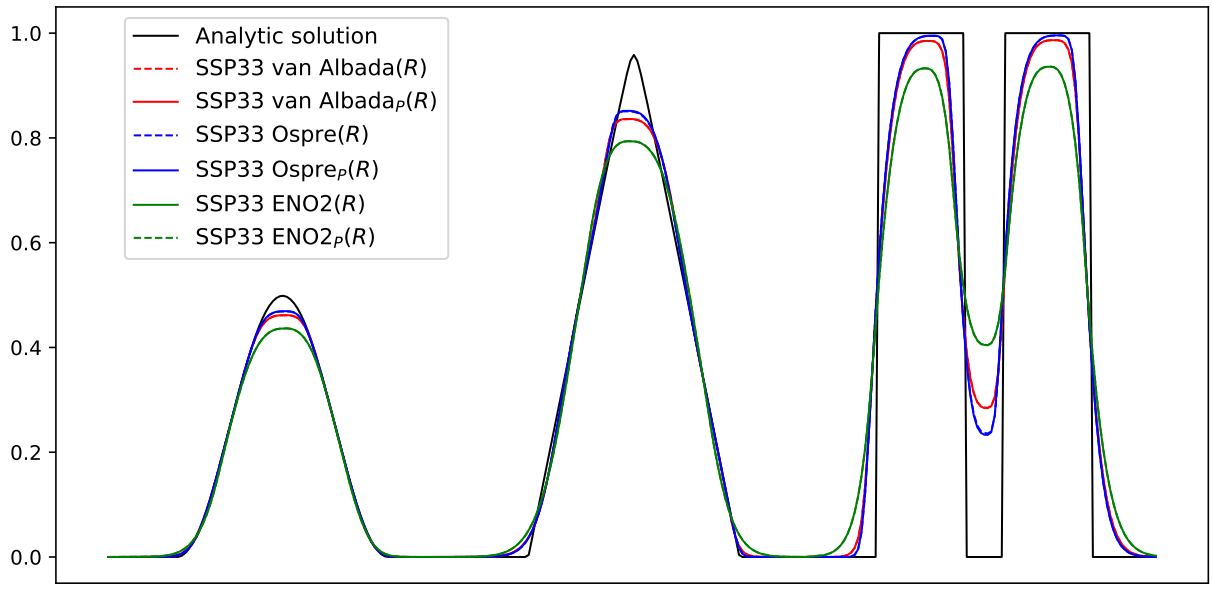


Figure 2.13: Symmetric Spekreijse limiters and their Symmetric push into the Sweby region. We use the horizontal cross sections through the middle of the three LeVeque shapes; cosine, cone and the Zalesak slotted cylinder. Conclusion, all the pushed limiters are almost indistinguishable to the eye from the original limiters.

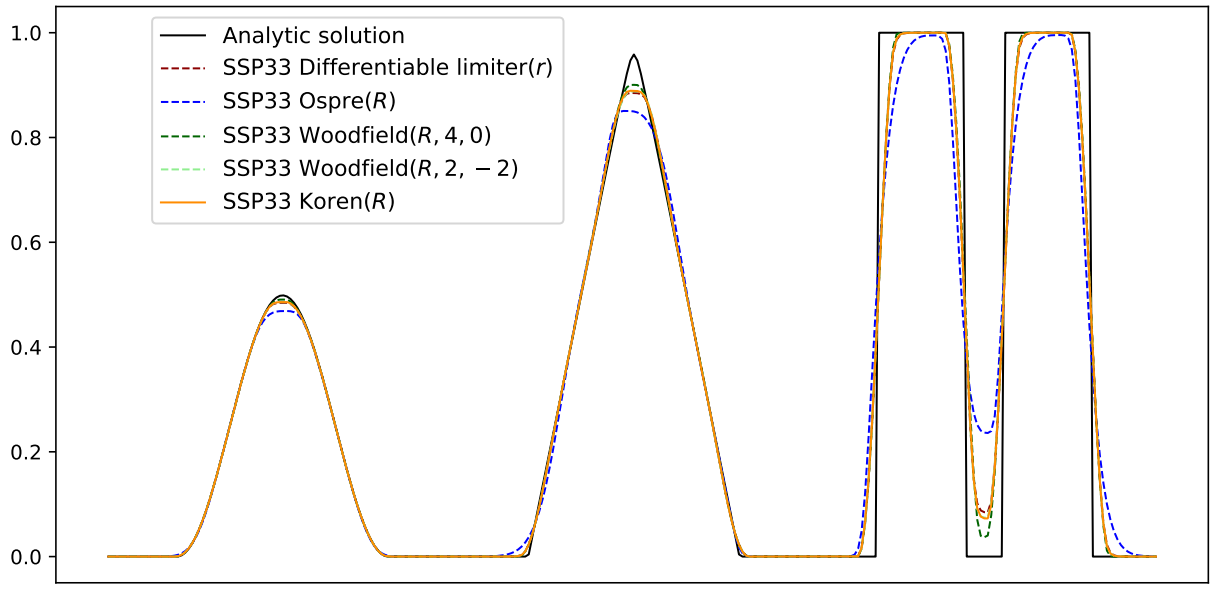


Figure 2.14: New differentiable limiter has results very similar to the Koren limiter. The differentiable limiter is more accurate than the Ospre limiter and is suitable for incompressible flow. The Woodfield(R,2,-2) limiter has similar accuracy to the Koren scheme and runs at a reduced timestep, the Woodfield(R,4,0) limiter is more accurate than the Koren limiter, however it must be run at a reduced timestep.

Convergence		Test cases	Observed	Order	
Scheme	Limiter	Diag	Quad	Sin	Sbr
SSP33	minmod(R)	1.473	1.465	1.005	1.560
SSP33	ENO2(R)	1.475	1.465	1.005	1.561
SSP33	vanAlbada$_P$(R)	1.522	1.711	1.366	1.716
SSP33	vanAlbada(R)	1.523	1.711	1.365	1.716
SSP33	Ospre$_P$(R)	1.590	1.875	1.472	1.767
SSP33	Ospre(R)	1.586	1.868	1.464	1.764
SSP33	Differentiable(r)	2.082	2.354	1.783	2.364
RK4	Koren(R)	2.125	2.396	1.816	2.424
SSP33	Koren(R)	2.125	2.396	1.816	2.424
SSP33	Woodfield($R, 2, -2$)	2.115	2.404	1.813	2.394
SSP33	Woodfield($r, 3, -1$)	2.288	2.487	1.888	2.565
SSP33	Woodfield($R, 4, 0$)	2.333	2.516	1.904	2.581
SSP33	CUI	2.880	2.519	1.881	2.868
SSP33	Fromm	1.962	2.476	1.789	1.947
SSP33	FOU	0.412	0.354	0.236	0.404

Table 2.4: This table contains the convergence rate of relative L^2 error between running at 64×64 as compared to 128×128 resolution for the four seperate flow cases.

The main contribution is the derivation of two new limiter regions, more general than the Sweby region, sufficient for multidimensional incompressible flow to maintain a discrete local maximum principle. More generally we show that the Spekreijse limiter region is not appropriate for the flux form incompressible advection equation unless a directional mean value theorem can be proven for each direction.

The reason this could have been missed is for many reasons; people commonly have more ambitious equations in mind and use flux vector and flux difference splitting techniques. Another reason this is missed is that for lots of simpler tests a mean value theorem does apply, for example in [83], the advection equation is tested numerically using seven different flows. However, in all examples they picked flows in which a directional mean value theorem held by coincidence (or possibly shrewdly chosen tests knowing the true intent for gas dynamics), and consequently they did not see any monotonicity violations. Another reason this is missed is because people often use the Sweby region rather than the Spekreijse's limiter region because they care about the internal subcell representation remaining bounded by its neighbour's cell mean values.

We have proven that the Sweby region is sufficient for a discrete maximum principle, and demonstrated quasi-necessity when the limiter is demanded symmetric in Section 2.3.2. Furthermore, pushing limiters in the Spekreijse region into the Sweby region limiters is not only easy and can preserve the symmetry of the limiter, but we have numerically tested that the new modification preserves accuracy (Fig. 2.13), convergence order (Table 2.4), peak preservation(Fig. 2.13), and makes the limiters suitable for incompressible flow Theorem 2.3.1, not only this, it also improves the Courant number restriction (Table 2.1). However, there is a small price for such a

modification, is that the push into the Sweby region can make the limiter function no longer differentiable at zero.

The main advantage to limiter functions such as Ospre and Van Albada arises not from their accuracy consideration, but from their global differentiability, which allows improved convergence for methods such as Newton iteration. So, although we have modified them to be discrete maximum principle satisfying, the loss of globally differentiability is highly undesirable for some applications. By sacrificing the symmetry requirement on the limiter, and using our new $\theta = 0$ limiter region, we have found the first globally differentiable limiter function contained within a second order limiter region suitable for truly incompressible flow. Our numerical experiments demonstrate preferable accuracy, convergence order and peak resolution that both the Ospre and van Albada limiters, whilst also being suitable for incompressible flow.

The newly introduced limiter regions are new and relatively unexplored, the careful design of limiters in these regions could be used to improve discrete maximum principle satisfying schemes. We have managed to increase the compressible properties of the Superbee limiter with the new SuperbeeR limiter, which may be useful for front tracking applications. Compressive one-dimensional limiters such as Superbee do have uses in physical oceanography [84], amongst the more common limiters used in this field [85]. As in [40, 41] we also conclude that the Koren [41] limiter is a robust and accurate limiter to use for numerical transport on the sphere. The Woodfield$(R, M, 0)$ limiter could be used to increase the accuracy of the Koren limiter, however additional work needs to go into the trade-off between accuracy and time-step for the parameter M.

Contrary to [40, 41] we come to the conclusion that the SSP literature is of practical importance, and SSP33 is recommended for this spatial discretisation, showing reduced cost and monotonicity over the RK4 method without noticeable loss of accuracy, due to the dominance of spatial errors. The RK4 scheme does perform surprisingly monotonic, but not sufficiently enough for our application, and has no theoretical guarantees it will continue to behave this way. It is likely the SSP22 scheme (Heun) will perhaps be useful in reducing the computational cost, and enforcing a more strict one step maximum principle in light of(Appendix A.2) than the SSP33 method.

Using one dimensional limiters in each direction of a multidimensional flux form advection scheme can achieve stability up to the Courant numbers defined in Table 2.1. They have inherent local mass conservation due to the flux form description. We newly identify that the method is exactly linear invariant in state interpolated form for incompressible flow, this gives constancy preservation. We identify new sufficient conditions on the limiter for positivity of the cell mean value at the next timestep. We have identified new sufficient conditions on the limiter for a discrete local cell mean maximum principle to hold. The method has been tested to be suitable for a tensor product mesh, but would require additional adaption for the cubed sphere.

The method has been shown to approach better than first order accuracy for low Courant numbers.

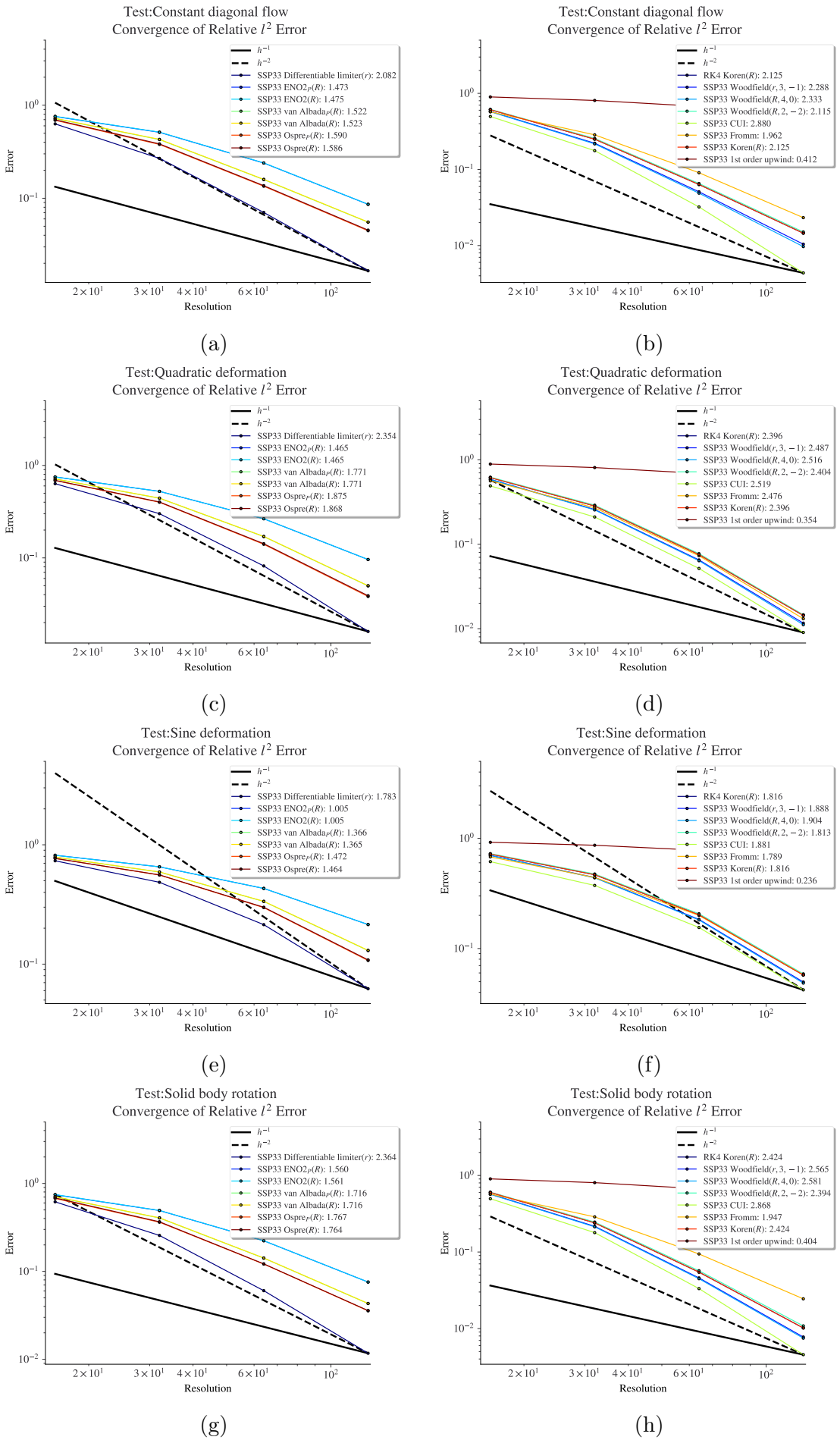


Figure 2.15: All convergence results plotted here, are described in Section 2.4.3.

Chapter 3

Local boundedness principles for multidimensional slope limiters

3.1 Introduction

3.1.1 Motivation

One dimensional limiting procedures introduced in Chapter 2 may not always be appropriate for all meshes proposed for dynamical cores [86] without additional adaption [87]. Furthermore, one dimensional limiting and even some multidimensional limiting procedures may also not be adaptable to truly higher order methods such those emerging in finite volume NWP models such as MPAS [88] and MCORE [89] as well as those emerging in discontinuous Galerkin finite element methods such as the NUMA [90]. Zhang et al. [43], has formulated a slope limiter framework applicable for a wide class of higher order methods on more general meshes. This framework currently has been used to maintain global boundedness principles such as positivity or range boundedness [43]. We will slightly extend the framework of Zhang et al. [43] from global boundedness principles to local boundedness principles. Barth and Jespersen in [42] introduced unstructured multidimensional limiters that have shown to be suitable for second order methods on unstructured and structured grids. We show that using the new approach to local boundedness limiters in this chapter leads to a new multidimensional limiter strictly more accurate than the Barth and Jespersen limiter, whilst still satisfying the same cell mean maximum principle. We then define a new 4th order finite volume scheme on a uniform structured grid. We demonstrate that this method it is indeed 4th order for some of the test cases from the other chapters. We use the new theory introduced to create a limiter function suitable for maintaining a local maximum principle with respect to edge sharing neighbour cell mean values. We then use our newly constructed slope limiter, under the LeVeque solid body rotation test case to demonstrate that the scheme produces acceptable results with regards to boundedness, and accuracy, and a local maximum principle. This is all introduced in a framework suitable for extension. But we first go back to 1976 where Harten Hyman Lax and Keyfitz introduce

a powerful notion of monotonicity which makes all this possible.

3.1.2 Background material: Forward Euler Upwind HHLK-monotonicity for unstructured advection.

We first establish the monotonicity of a forward Euler scheme in an unstructured HHLK [24] sense, and discuss out how the explicit dependence on an arbitrary velocity field fits into the notion of sign preservation, and a discrete local maximum principle. In this section we review this historical example with the more modern unstructured notation aligning with [72] but we introduce additional dependence on the velocity field, rather than separate out the averaged flow through an edge as in [72] this is to ensure the generalisation to higher order finite volume schemes is straightforward.

Definition 3.1.1 (**Forward Euler Upwind**)**.** The forward Euler first order upwind scheme on an unstructured mesh, is the simplest finite volume method. This method consists of approximating the compact subcell reconstruction p_K within cell K by the constant cell mean value $\bar{u}_K$ and the flux through a face is approximated using second order Gauss quadrature at the midpoint of each face, and takes the following form

$$\bar{u}_K^{n+1} = \bar{u}_K^n - \Delta t \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{|K|} f_{KL}^n(\bar{u}_K, \bar{u}_L, \boldsymbol{v}(\boldsymbol{x}_{KL}) \cdot \boldsymbol{n}_{KL}), \quad \forall K \in \mathcal{M}. \tag{3.1}$$

We sketch an element of the mesh in Fig. 3.1. The face belonging to the boundary of cell K and L is defined by the intersection of the cell boundaries $\sigma_{KL} := \partial K \cup \partial L = \partial K \cap L$. $N(K) := \{L \in \mathcal{M} \big| |\sigma_{KL}| > 0\}$ denotes the set of face-sharing neighbours of cell K. The midpoint of face σ_{KL} is denoted by the position vector $\boldsymbol{x}_{KL}$. The positive and negative superscript denotes $(\cdot)^+ := \max(0, \cdot), (\cdot)^- := \min(0, \cdot)$ the positive and negative component of an input. $|K|$ denotes the volume (Lebesgue measure) of the cell K and $|\sigma_{KL}|$ denotes the volume/area (Lebesgue measure) of the face σ_{KL} which are assumed positive. We denote $p_K(\boldsymbol{x})$, as the subcell representation of cell K. f_{KL} denotes the flux from cell K into the cell L. $\boldsymbol{n}_{KL}$ is the outward unit normal from cell K into cell L. $\boldsymbol{v}(\boldsymbol{x})$ denotes the velocity. For the advection equation, the Riemann problem is tractable and given by the upwind/donor cell numerical flux function

$$f_{KL} = f_{KL}(a_K, b_L, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) = [\boldsymbol{v} \cdot \boldsymbol{n}_{KL}]^+ a_K + [\boldsymbol{v} \cdot \boldsymbol{n}_{KL}]^- b_L. \tag{3.2}$$

The definition of a Consistent Conservative Monotone Numerical Flux Function defined in [72], can be trivially extended to schemes with a faced defined velocity field as follows.

Definition 3.1.2. A consistent conservative monotone numerical flux function satisfies the following properties. The semi-discrete numerical flux function $f_{KL}(a, b, \boldsymbol{v} \cdot$

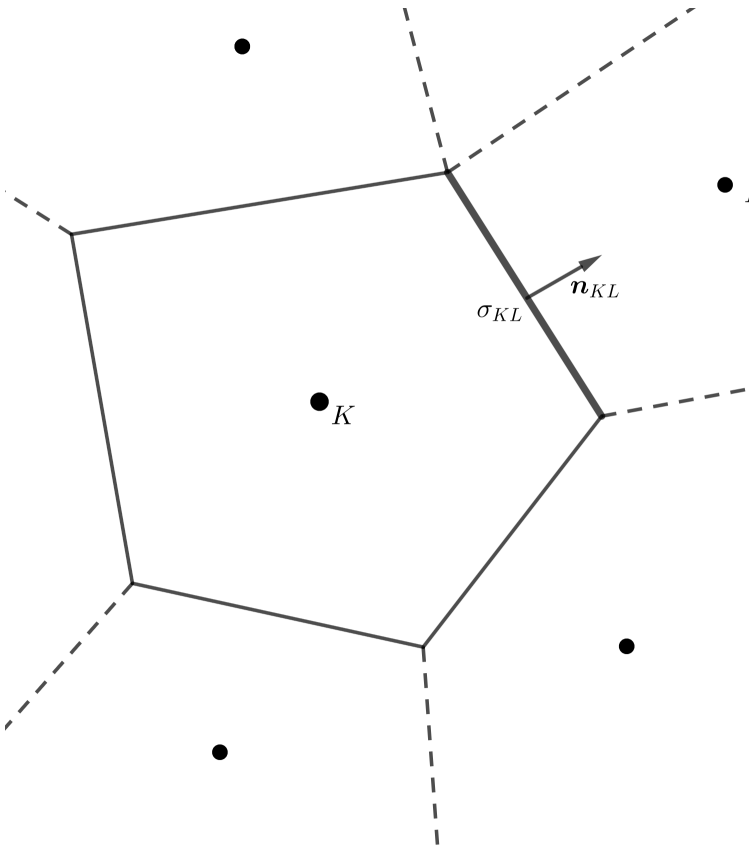


Figure 3.1: Diagram of cell K, and the face σ_{KL} of a face sharing neighbour $L \in N(K)$, with outward unit normal $\boldsymbol{n}_{KL}$.

$\boldsymbol{n}_{KL}$) reconstructs the face value such that it is consistent with the boundary flux. The numerical flux should also inherit the conservative properties of the continuous flux, $f_{KL}(a, b, c_{KL}) = -f_{LK}(b, a, c_{LK})$. The map defined by the numerical flux is f_{KL} is a monotonic flux function in the sense that it is non-decreasing with respect to the first argument and non-increasing with respect to the second argument in the following sense $\partial_a f_{KL}(a, b, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) \geq 0$, $\partial_b f_{KL}(a, b, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) \leq 0$.

Theorem 3.1.1. *The upwind numerical flux* $f_{KL}(a_K, b_L, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) = [\boldsymbol{v} \cdot \boldsymbol{n}_{KL}]^+ a_K + [\boldsymbol{v} \cdot \boldsymbol{n}_{KL}]^- b_L$*, is a consistent conservative monotone numerical flux function satisfying Definition 3.1.2 for the flux form advection equation.*

Direct computation. The numerical flux $f_{KL}(a_K, b_L, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) = [\boldsymbol{v} \cdot \boldsymbol{n}_{KL}]^+ a_K + [\boldsymbol{v} \cdot \boldsymbol{n}_{KL}]^- b_L$ is consistent with respect to the physical value at the boundary since it satisfies the following condition

$$f_{KL}(a, a, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) = [\boldsymbol{v} \cdot \boldsymbol{n}_{KL}]^+ a + [\boldsymbol{v} \cdot \boldsymbol{n}_{KL}]^- a = a(\boldsymbol{v} \cdot \boldsymbol{n}_{KL}). \tag{3.3}$$

The numerical flux is conservative since

$$\begin{aligned} f_{KL}(a, b, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) &= a(\boldsymbol{v} \cdot \boldsymbol{n}_{KL})^+ + b(\boldsymbol{v} \cdot \boldsymbol{n}_{KL})^- && (3.4)\\ &= a(\boldsymbol{v} \cdot -\boldsymbol{n}_{LK})^+ + b(\boldsymbol{v} \cdot -\boldsymbol{n}_{LK})^- && (3.5)\\ &= -a(\boldsymbol{v} \cdot \boldsymbol{n}_{LK})^- - b(\boldsymbol{v} \cdot \boldsymbol{n}_{LK})^+ && (3.6)\\ &= -f_{LK}(b, a, \boldsymbol{v} \cdot \boldsymbol{n}_{LK}). && (3.7) \end{aligned}$$

The numerical flux is monotone in the classical sense by direct computation

$$\partial_a f_{KL}(a, b, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) = (\boldsymbol{v} \cdot \boldsymbol{n}_{KL})^+ \geq 0, \tag{3.8}$$

$$\partial_b f_{KL}(a, b, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) = (\boldsymbol{v} \cdot \boldsymbol{n}_{KL})^- \leq 0. \tag{3.9}$$

□

Theorem 3.1.2 (Forward Euler HHLK monotone [24])**.** *Given a numerical flux of form [Definition 3.1.2], the forward Euler scheme Definition 3.1.1 is a monotone function of surrounding cell mean values. This is sufficient for sign preservation under compressible flow, under the Courant number restriction*

$$C_K = \Delta t \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{|K|} \partial_a f_{KL}(a, b, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}) \leq 1. \tag{3.10}$$

If in addition a discrete divergence free condition $0 = \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{|K|} \boldsymbol{v} \cdot n_{KL}$ *is also satisfied, then a local maximum principle with respect to neighbouring (face sharing) cell mean values holds* $\min_{L \in N(K) \cup K} \bar{u}_L^n \leq u_K^{n+1} \leq \max_{L \in N(K) \cup K} \bar{u}_L^n$.

Proof. By differentiating the function

$$\bar{u}_K^{n+1} = H(\bar{u}_K, \{\bar{u}_L\}_{\forall L \in N(K)}, \{\boldsymbol{v} \cdot n_{KL}\}_{\forall L \in N(K)}), \tag{3.11}$$

$$= \bar{u}_K^n - \Delta t \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{|K|} f_{KL}^n(\bar{u}_K, \bar{u}_L, \boldsymbol{v} \cdot \boldsymbol{n}_{KL}), \tag{3.12}$$

with respect to each cell mean argument

$$\frac{\partial H}{\partial \bar{u}_L} = -\Delta t \frac{|\sigma_{KL}|}{|K|} \partial_{\bar{u}_L} f_{KL} \geq 0, \quad \forall L \in N(K), \tag{3.13}$$

$$\frac{\partial H}{\partial \bar{u}_K} = 1 - \Delta t \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{|K|} \partial_{\bar{u}_K} f_{KL} \geq 0, \tag{3.14}$$

we verify the scheme is a monotone function of surrounding cell mean values under the following definition of a local cell defined Courant number

$$C_K = \Delta t \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{|K|} \partial_{\bar{u}_K} f_{KL} \leq 1. \tag{3.15}$$

This notion of monotonicity trivially gives the sign preservation property for arbitrary velocity fields. From a discrete divergence free condition $0 = \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{|K|} \boldsymbol{v} \cdot n_{KL}$ and consistency of the numerical fluxes one can derive that the equation is constancy preserving as follows

$$0 = \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{|K|} f_{KL}(c, c, \boldsymbol{v} \cdot n_{KL}) \quad \forall c \in \mathbb{R} \tag{3.16}$$

$$c = H(\bar{u}_K = c, \{\bar{u}_L = c\}_{\forall L \in N(K)}, \{\boldsymbol{v} \cdot \boldsymbol{n}_{KL}\}_{\forall L \in N(K)}), \quad \forall c \in \mathbb{R}. \tag{3.17}$$

By temporarily setting the local minima m_K and maxima M_K to be the neighbour inclusive cell mean values

$$m_K = \min_{L \in N(K) \cup K} \bar{u}_L, \quad M_K = \max_{L \in N(K) \cup K} \bar{u}_L, \tag{3.18}$$

we can observe the inclusive face sharing local maximum principle from consistency and HHLK monotonicity of the function H as follows

$$m_K = H(m_K, \{m_K\}, \{\boldsymbol{v} \cdot \boldsymbol{n}_{KL}\}_{\forall L \in N(K)}) \leq \bar{u}_K^{n+1} \leq H(M_K, \{M_K\}, \{\boldsymbol{v} \cdot \boldsymbol{n}_{KL}\}_{\forall L \in N(K)}) = M_K. \tag{3.19}$$

□

We have seen how the monotonicity theory of Harten Hyman Lax and Keyfitz gives sign preservation, and how the addition of a divergence free vector-field is required for a discrete local maximum principle.

Remark. The monotonicity of the flux function may sometimes be relaxed to the weaker Lipschitz continuity property [72, 91]. However, this is no longer alone sufficient for a discrete maximum principle because we have introduced the face defined velocity field which is also required divergence free.

We have introduced the background material for the simplest unstructured finite volume method, and what it means for such a scheme to retain monotone properties under a variety of flows, both the sign preservation property of compressible flow and the discrete face sharing maximum principle arising from the incompressibility assumption. In the next section we introduce some higher order finite volume methods and develop sufficient conditions on multidimensional slope limiters for the preservation of a local maximum principle of the following form

$$\bar{u}_K^n \in [m_K, M_K], \quad \forall K. \tag{3.20}$$

Where m_K, M_K may depend on some local quantities such as a stencil of local cell means $\{\bar{u}_i\}_{i \in S_K}$. To do so, we rely on the theoretical slope limiting framework of Zhang et al. [43], who modified the notion of HHLK-monotone to higher order methods. In this powerful framework Zhang, Xiangxiong and Chi-Wang [43] constructed a framework to create higher order schemes for scalar conservation law that satisfy a global maximum principle (bounded in L^∞). Which satisfy the global maximum principle of the following form

$$\bar{u}_K \in [m, M], \quad \forall K, \quad m, M \in \mathbb{R}. \tag{3.21}$$

In Section 3.2 we will introduce Theorem 3.2.1 indicating sufficient conditions for an arbitrary order scheme to retain a local boundedness principle on an unstructured mesh. We introduce some stencil notation in Section 3.2 this will help specify where we will impose maximum principles. In Section 3.2.1 we design two new limiters

based on the ideas of Section 3.2. This concludes the main mathematical contribution, and the next two sections are applications/examples consisting of pseudo code type methods with introductions on how the theory in Section 3.2 applies to the finite volume technique and how limiters in Section 3.2.1 simplify and can be put into practice.
The first example Section 3.3 consists of introducing a common second order finite volume method on the simplest grid. We explain how Theorem 3.2.1 can be used for the Courant number restrictions. In Section 3.3 we also introduce two of the most commonly used multidimensional slope limiters, and compare with the two new limiter's we have introduced.
The second example Section 3.4.1 introduces a new fourth order advection algorithm, we explain how Theorem 3.2.1 can be used once a new decomposition of the cell average is found.

3.2 High order, multidimensional slope limiting for arbitrary meshes, and arbitrary flow.

The framework of Zhang et al. [43] has been used to create positivity preserving solutions to the compressible Euler equations for arbitrary order Discontinuous Galerkin finite element methods [44] and has seen practical success for both high order DG and high order finite volume methods for more triangular meshes [92]. In this section we use this framework to derive sufficient conditions for a higher order DGFE or finite volume method to preserve the strictly stronger local boundedness principle. We do so in a general way, as to design a theoretical limiter capable of preserving a local maximum principle.
Choosing the test function to be a piecewise constant over a DG finite element method, or by doing Gauss divergence theorem for a finite volume method, we arrive at the forward Euler cell mean evolution equation, of the following form

$$\bar{u}_K^{n+1} = \bar{u}_K^n - \Delta t \frac{1}{|K|} \sum_{L \in N(K)} |\sigma_{KL}| \sum_{q \in \sigma_{KL}} w_q^{\sigma_{KL}} f_{KL}(p_K(x_q), p_L(x_q), \boldsymbol{v} \cdot \boldsymbol{n}_{KL}), \quad (3.22)$$

where $p_K(\boldsymbol{x})$ denotes a high order polynomial[1] approximating the true solution $u_K(\boldsymbol{x})$ in the cell K. $\{w_q^{\sigma_{KL}}\}_{\forall q \in \sigma_{KL}}$, denotes the set of quadrature weights associated with the corresponding set of quadrature nodes $\{x_q\}_{\forall q \in \sigma_{KL}}$ on a face σ_{KL} used to approximate the flux through a face

$$\int_{\boldsymbol{x} \in \sigma_{KL}} f_{KL}(u_K(\boldsymbol{x}), u_L(\boldsymbol{x}), \boldsymbol{v}_{KL}(\boldsymbol{x}) \cdot \boldsymbol{n}_{KL}) ds \approx |\sigma_{KL}| \sum_{q \in \sigma_{KL}} w_q^{\sigma_{KL}} f_{KL}(u_K(x_q), u_L(x_q), \boldsymbol{v} \cdot \boldsymbol{n}_{KL}). \quad (3.23)$$

[1] This polynomial could be solved for as in the finite element method or alternatively constructed from other cell average values as in the finite volume method, or even reconstructed in a more abstract setting to satisfy certain properties [92].

The Eq. (3.22) scheme is no longer a monotone function of surrounding cell mean values in the HHLK [24] sense, however [43, 92] point to the fact that under some decompositions of the cell average, the scheme is a monotone function of quadrature point evaluations. The key to this interpretation relies on the assumption that the cell mean $\bar{u}_K$ can be decomposed in terms of a positive weighting of flux contributing quadrature points. This is non-trivial and depends on the method used. One such cell mean decomposition proposed in [43] involves the fact that the numerical quadrature of a k-exact polynomial reconstruction over a cell is exact and uses positive quadrature weights, this is also available for unstructured finite volume methods [93]. We will simply assume an abstract cell mean decomposition as follows

$$\bar{u}_K = \frac{1}{|K|}\int_K p_K(\boldsymbol{x})d\boldsymbol{x} = \sum_{q\in K^{fc}\cup K^{nfc}} p_K(\boldsymbol{x}_q)w_q^K = \sum_{q\in K^{nfc}} p_K(\boldsymbol{x}_q)w_q^K + \sum_{q\in K^{fc}} p_K(\boldsymbol{x}_q)w_q^K, \tag{3.24}$$

where $\{w_q^K\}_{\forall q\in K}$ are the set of non-negative quadrature weights associated with the total set $\{x_q\}_{\forall q\in K}$ of quadrature points used decompose the cell average. The quadrature points associated to the cell mean decomposition can be split into the flux contributing quadrature points K^{fc} and the non-flux contributing quadrature points K^{nfc}.

We will consider the cell mean decomposition as Zhang-acceptable when all flux contributing quadrature points from Eq. (3.22) are captured with strictly positive weighting $w_q > 0$, $\forall q \in K^{fc}$.

The numerical scheme is then written as a finite positive sum of three-point HHLK-monotone schemes, which resolve the local Riemann problems at the face defined quadrature points. Theorem 3.2.1 below describes the sufficient conditions for a local cell mean boundedness principle. We have assumed that there are no corner defined flux contributing quadrature points to simplify the presentation and time-step restriction.

Theorem 3.2.1 (Monotone DG and FV schemes (with flux contributing vertex exclusion))**.** *The cell mean value at the next time-step $\bar{u}_K^{n+1}$ evolving by the cell mean evolution equation Eq. (3.22) with a flux of Definition 3.1.2 (with no vertex contributing quadrature points) can be expressed as a monotone function of quadrature point evaluations [43, 92]. If all quadrature point evaluations arising from a Zhang-acceptable cell mean decomposition are non-negative $p_K(x_q) \geq 0, \quad \forall q \in K, \quad \forall K \in \mathcal{M}$, and all the edge defined Riemann problems Courant number restrictions are satisfied:*

$$\Delta t \frac{w_q^{\sigma_{KL}}|\sigma_{KL}|}{w_q^K|K|}\frac{\partial f_{KL}}{\partial p_K} \leq 1, \quad \forall q \in \sigma_{KL}, \quad \forall L \in N(K), \quad \forall K \in \mathcal{M} \tag{3.25}$$

then the scheme is positivity preserving $\bar{u}_K^{n+1} \geq 0$ [43, 92]. If in addition the vector

field satisfies the following discrete divergence free condition,

$$\frac{1}{|K|}\sum_{L\in N(K)}|\sigma_{KL}|\sum_{q\in\sigma_{KL}}w_q^{\sigma_{KL}}(\boldsymbol{v}\cdot\boldsymbol{n}_{KL})=0, \tag{3.26}$$

as well as local boundedness of quadrature point evaluations,

$$p_K(x_q)\in[m_K,M_K],\quad \forall x_q\in K^{nfc} \tag{3.27}$$

$$p_K(x_q),p_L(x_q)\in[m_K,M_K],\quad \forall x_q\in\sigma_{KL},\quad \forall L\in N(K) \tag{3.28}$$

$$\Delta t\frac{w_q^{\sigma_{KL}}|\sigma_{KL}|}{w_q^K}\frac{\partial f_{KL}}{\partial p_K}\leq 1,\quad \forall q\in\sigma_{KL},\quad \forall L\in N(K) \tag{3.29}$$

then the next time level will satisfy a $\bar{u}_K^{n+1}\in[m_K,M_K]$ *local boundedness principle.*

Remark. This can be extended for the case in which when there are flux contributing vertex points. We avoid this technicality.

Proof. [Theorem 3.2.1] Use the Zhang-acceptable cell mean decomposition, to write the scheme as a positive sum of non-flux contributing quadrature point evaluations and Riemann problems at the flux contributing quadrature points as follows

$$\bar{u}_K^{n+1}=\frac{1}{|K|}\sum_{x_q\in K^{fc}\cup K^{nfc}}w_q^Kp_K(x_q)-\Delta t\frac{1}{|K|}\sum_{L\in N(K)}|\sigma_{KL}|\sum_{q\in\sigma_{KL}}w_q^\sigma f_{KL}(p_K(x_q),p_L(x_q),\boldsymbol{v}(x_q,t)\cdot\boldsymbol{n}_{KL} \tag{3.30}$$

$$=\frac{1}{|K|}\sum_{x_q\in K^{nfc}}w_q^Kp_K(x_q)+\frac{1}{|K|}\sum_{L\in N(K)}\sum_{q\in\sigma_{KL}}w_q^K\Big(Rie(p_K(x_q),p_L(x_q),\boldsymbol{v}(x_{KL},t)\cdot\boldsymbol{n}_{KL})\Big), \tag{3.31}$$

where the flux contributing quadrature point Riemann problems are solved by the three point classically HHLK-monotone scheme,

$$Rie(p_K(x_q),p_L(x_q),\boldsymbol{v}(x_{KL},t)\cdot\boldsymbol{n}_{KL}):=p_K(x_q)-\Delta t\frac{w_q^\sigma|\sigma_{KL}|}{w_q^K}f_{KL}(p_K(x_q),p_L(x_q),\boldsymbol{v}(x_q,t)\cdot\boldsymbol{n}_{KL}). \tag{3.32}$$

The derivative of the numerical method Eq. (3.22) with respect to each quadrature point evaluation is given by

$$\frac{\partial\bar{u}_K^{n+1}}{\partial p_K(x_q)}=\frac{w_q^K}{|K|},\quad \forall q\in K^{nfc}, \tag{3.33}$$

$$\frac{\partial\bar{u}_K^{n+1}}{\partial p_K(x_q)}=w_q^K\big[1-\frac{\Delta t|\sigma_{KL}|w_q^{\sigma_{KL}}}{|K|w_q^K}\frac{\partial f_{KL}}{\partial p_K(x_q)}\big],\quad \forall q\in K^{fc}, \tag{3.34}$$

$$\frac{\partial\bar{u}_K^{n+1}}{\partial p_L(x_q)}=-\frac{\Delta t|\sigma_{KL}|w_q^{\sigma_{KL}}}{|K|}\frac{\partial f_{KL}}{\partial p_L(x_q)},\quad \forall q\in K^{fc}\cap\sigma_{KL},\quad \forall L\in N(K). \tag{3.35}$$

The weight properties ($w_q^K\geq 0,\forall q\in K^{nfc}$), ($w_q^K>0,\forall q\in K^{fc}$), ($w_q^{\sigma_{KL}}>0,\forall q\in K^{fc}\cap\sigma_{KL}$), the monotone property of the flux $\partial_a f_{KL}(a,b,c)\geq 0,\partial_b f_{KL}(a,b,c)\leq 0$,

and the flux contributing time-step restrictions $\frac{\Delta t|\sigma_{KL}|w_q^\sigma}{|K|}\frac{\partial f_{KL}}{\partial p_K(x_q)} \leq 1, \quad \forall q \in K^{fc}$ imply all derivatives are non-negative. So, the scheme is a monotone function of quadrature point evaluations. This means that for an arbitrary velocity field, the following conditions

$$p_K(x_q) \geq 0, \quad \forall x_q \in K^{nfc} \tag{3.36}$$

$$p_K(x_q), p_L(x_q) \geq 0, \quad \forall x_q \in \sigma_{KL}, \quad \forall L \in N(K) \tag{3.37}$$

$$\Delta t \frac{w_q^\sigma |\sigma_{KL}|}{w_q^K |K|}\frac{\partial f_{KL}}{\partial p_K} \leq 1, \quad \forall q \in \sigma_{KL}, \quad \forall L \in N(K) \tag{3.38}$$

are sufficient for the scheme to be positivity preserving, the negativity preservation is similar and gives sign preservation of the numerical scheme. If in addition, we suppose that the following discrete divergence free condition holds

$$\frac{1}{|K|}\sum_{L\in N(K)} |\sigma_{KL}| \sum_{q\in\sigma_{KL}} w_q^{\sigma_{KL}}(\boldsymbol{v}\cdot\boldsymbol{n}_{KL}) = 0, \tag{3.39}$$

and that the numerical fluxes are consistent. We can derive equation consistency, from

$$\frac{1}{|K|}\sum_{L\in N(K)} |\sigma_{KL}| \sum_{q\in\sigma_{KL}} w_q^K \Big(f_{KL}(c,c,\boldsymbol{v}(x_q,t)\cdot\boldsymbol{n}_{KL})\Big) = 0, \tag{3.40}$$

as it implies the preservation of constants of the scheme $H(c,c,c,c,...,c,c,c,\boldsymbol{v}) = c$. If in addition we assume that the internal and boundary quadrature points are locally bounded by the constants m_K, M_K in the following way

$$p_K(x_q) \in [m_K, M_K], \quad \forall x_q \in K^{int}, \tag{3.41}$$

$$p_K(x_q), p_L(x_q) \in [m_K, M_K], \quad \forall x_q \in \sigma_{KL}, \quad \forall L \in N(K), \tag{3.42}$$

$$\Delta t \frac{w_q^\sigma |\sigma_{KL}|}{w_q^K}\frac{\partial f_{KL}}{\partial p_K} \leq 1, \quad \forall q \in \sigma_{KL}, \quad \forall L \in N(K). \tag{3.43}$$

We can then deduce the following maximum principle, by the monotonicity and the consistency of the numerical method. For the cell K

$$m_K = H(m_K, m_K, m_K, \boldsymbol{v}) \leq \bar{u}^{n+1} \leq H(M_K, M_K, M_K, \boldsymbol{v}) = M_K \tag{3.44}$$

□

The main distinction from the Zhang et al. [43] theory is that the flux contributing quadrature points at an edge σ_{KL} must satisfy two local boundedness principles

$$p_L(x_q), p_K(x_q) \in [m_K, M_K], \quad \text{if} \quad x_q \in \sigma_{KL}, \tag{3.45}$$

$$p_L(x_q), p_K(x_q) \in [m_L, M_L], \quad \text{if} \quad x_q \in \sigma_{KL}, \tag{3.46}$$

when the requirements of Theorem 3.2.1 are viewed from the perspective of cells K, L

respectively. This has important consequences on the design of multidimensional limiter functions. It implies that both $p_L(x_q), p_K(x_q)$ could be limited based on the same but extended edge defined maximum principle

$$p_L(x_q), p_K(x_q) \in [\min\{m_K, m_L\}, \max\{M_K, M_L\}] \quad \text{if} \quad x_q \in \sigma_{KL}, \tag{3.47}$$

and satisfy the maximum principle

$$u_K^{n+1} \in [\min_{i \in N(K) \cup K} m_i, \max_{i \in N(K) \cup K} M_i], \tag{3.48}$$

based on the union of flux contributing edge defined maximum principles.

Mesh neighbourhood notation

Before introducing our new limiter, we first remark that several different local maximum principles are already proposed to control non-physical oscillations. The one-dimensional limiting procedures in [27] satisfy a maximum principle based on the inclusive face sharing neighbourhood, the Barth and Jespersen limiter [42] satisfies a maximum principle based on the "squared" inclusive face sharing neighbourhood. The Kuzmin limiter [94] satisfies a maximum principle based on the inclusive vertex sharing neighbours. These neighbourhoods and more are defined below

- $N(K)$ denotes the face neighbours of cell K,
- $N(K) \cup K$ is the inclusive face sharing neighbourhood,
- $N^2(K) \cup N(K)$ is the set of "squared" inclusive face sharing neighbourhood,
- $N(v)$ is the cell neighbourhood of a vertex,
- $VN(K)$ is the inclusive vertex neighbourhood of cell K, $VN(K) := \cup_{v \in K} N(v)$.

We introduce the convenient diagram Fig. 3.2 to help with the visualisation of these different regions. In the next section we will introduce the $N(K) \cup K$-MP limiter, it is based on an edge sharing maximum principle $K \cup L$ for quadrature points on edges, and whose resulting cell mean value satisfies a maximum principle on the inclusive face sharing neighbourhood $N(K) \cup K$. We introduce the $N^2(K) \cup N(K)$-MP limiter based on an edge sharing maximum principle $N(K) \cup N(L)$, whose cell mean satisfies a maximum principle on the "squared" inclusive face sharing neighbourhood like the Barth and Jespersen limiter. However, these limiters must also take into account some non-flux contributing quadrature points as will be described. We have preliminarily indicated how vertex defined flux contributing quadrature points do require additional care, but we will be seeking methods without flux contributing vertex points to simplify the presentation.

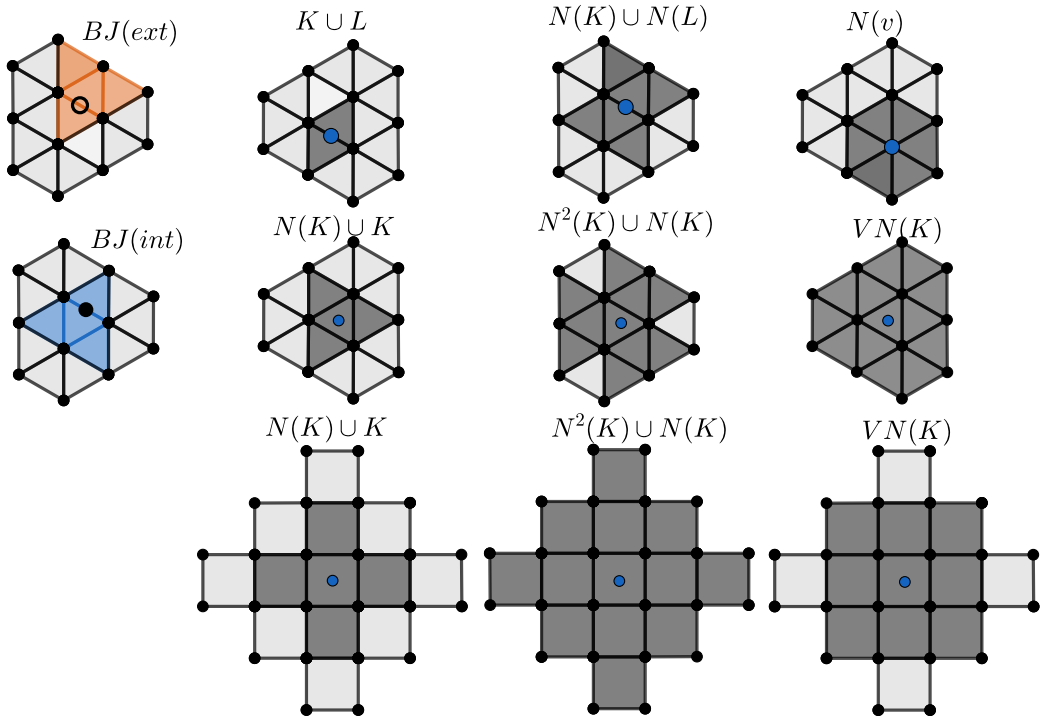


Figure 3.2: Visualisations of some neighbourhoods. Blue dot is an informal representation of the "middle" of the neighbourhood, representing cell K or the middle of K and L or even the v vertex. Blue and orange are specific to the Barth and Jespersen limiter.

3.2.1 New local boundedness slope limiters

We will use the theoretical results established in Theorem 3.2.1, to create a local maximum principle limiter capable of preserving

$$\min_{L \in N(K) \cup K} \bar{u}_L^n \leq \bar{u}_K^{n+1} \leq \max_{L \in N(K) \cup K} \bar{u}_L^n, \tag{3.49}$$

and

$$\min_{L \in N^2(K) \cup N(K)} \bar{u}_L^n \leq \bar{u}_K^{n+1} \leq \max_{L \in N^2(K) \cup N(K)} \bar{u}_L^n. \tag{3.50}$$

These new limiter functions are called the $N(K) \cup K$-MP limiter, and $N^2(K) \cup N(K)$-MP limiter. They are straightforwardly generalisable to include a maximum principle with stencil of arbitrary size $N^{s+1}(K) \cup N^s(K)$, such that the limit $s \to \infty$ recovers the global boundedness limiter of [43, 92] with globally defined bounds $M = \max_{\forall K \in \mathcal{M}} \bar{u}_K^n$, $m = \min_{\forall K \in \mathcal{M}} \bar{u}_K^n$.
The non-flux contributing quadrature points must satisfy a regular local maximum principle, and each flux contributing quadrature point has its own stencil maximum principle. We have also included a preliminary explanation into to the flux contributing vertex extension in the appendix.

Definition 3.2.1 ($N(K) \cup K$-MP limiter)**.**

1. Per face σ_{KL}, we compute and associate the local face defined maximum prin-

ciple bounds

$$[m_{\sigma_{KL}}, M_{\sigma_{KL}}] = [\min_{M \in L \cup K} \bar{u}_M^n, \max_{M \in L \cup K} \bar{u}_M^n], \tag{3.51}$$

this is associated to each flux contributing $x_q \in \sigma_{KL}$ not on a vertex.

2. Per cell K we associate the desired maximum principle

$$[m_{K^{nfc}}, M_{K^{nfc}}] = [\min_{L \in N(K) \cup K} \bar{u}_L^n, \max_{L \in N(K) \cup K} \bar{u}_L^n], \tag{3.52}$$

this is associated to each non-flux contributing quadrature point $x_q \in K^{nfc}$.

3. Per vertex of K, with two faces σ_{KL}, σ_{KM} we compute the local vertex maximum principle bounds

$$[m_{v_{KLM}}, M_{v_{KLM}}] = [\min_{i \in L \cup K \cup M} \bar{u}_i^n, \max_{i \in L \cup K \cup M} \bar{u}_i^n]. \tag{3.53}$$

This extends to a vertex with more than two faces connected as one would expect. This maximum principle is associated with flux contributing quadrature points at vertices.

4. We then per cell compute all the Barth and Jespersen quadrature corrections factors α_q, to ensure $\tilde{p}_K(x) = \alpha(p_K(x) - \bar{u}_K) + \bar{u}_K$, satisfies the conditions in Theorem A.4.1.

$$\tilde{p}_K(x_q) \in [m_{K^{nfc}}, M_{K^{nfc}}], \quad \forall q \in K^{nfc}, \tag{3.54}$$
$$\tilde{p}_K(x_q) \in [m_{\sigma_{KL}}, M_{\sigma_{KL}}], \quad \forall q \in \sigma_{KL} \cap K^{fc}, \quad \forall L \in N(K), \tag{3.55}$$
$$\tilde{p}_K(x_q) \in [m_{v_{KLM}}, M_{v_{KLM}}], \quad \forall q \in VN(K) \cap K^{fc}. \tag{3.56}$$
$$\tag{3.57}$$

by choosing the smallest value

$$\alpha = \min_{\forall q \in K} \alpha_q. \tag{3.58}$$

Definition 3.2.2 ($N^2(K) \cup N(K)$-MP limiter)**.**

1. Per face σ_{KL}, we compute and associate the local face defined maximum principle bounds

$$[m_{\sigma_{KL}}, M_{\sigma_{KL}}] = [\min_{M \in N(L) \cup N(K)} \bar{u}_M^n, \max_{M \in N(L) \cup N(K)} \bar{u}_M^n], \tag{3.59}$$

this is associated to each flux contributing $x_q \in \sigma_{KL}$ not on a vertex.

2. Per cell K we associate the desired maximum principle

$$[m_{K^{nfc}}, M_{K^{nfc}}] = [\min_{L \in N^2(K) \cup N(K)} \bar{u}_L^n, \max_{L \in N^2(K) \cup N(K)} \bar{u}_L^n], \tag{3.60}$$

this is associated to each non-flux contributing quadrature point $x_q \in K^{nfc}$.

3. Per vertex of K, with two faces σ_{KL}, σ_{KM} we compute the local vertex maximum principle bounds

$$[m_{v_{KLM}}, M_{v_{KLM}}] = [\min_{i \in N(L) \cup N(K) \cup N(M)} \bar{u}_i^n, \max_{i \in N(L) \cup N(K) \cup N(M)} \bar{u}_i^n] \tag{3.61}$$

This extends to a vertex with more than two faces connected as one would expect. This maximum principle is associated with flux contributing quadrature points at vertices.

4. We then per cell compute all the Barth and Jespersen quadrature corrections factors α_q, to ensure $\tilde{p}_K(x) = \alpha(p_K(x) - \bar{u}_K) + \bar{u}_K$, satisfies the conditions in Theorem A.4.1.

$$\tilde{p}_K(x_q) \in [m_{K^{nfc}}, M_{K^{nfc}}], \quad \forall q \in K^{nfc}, \tag{3.62}$$
$$\tilde{p}_K(x_q) \in [m_{\sigma_{KL}}, M_{\sigma_{KL}}], \quad \forall q \in \sigma_{KL} \cap K^{fc}, \quad \forall L \in N(K), \tag{3.63}$$
$$\tilde{p}_K(x_q) \in [m_{v_{KLM}}, M_{v_{KLM}}], \quad \forall q \in VN(K). \tag{3.64}$$
$$\tag{3.65}$$

by choosing the smallest value

$$\alpha = \min_{\forall q \in K \cup L} \alpha_q. \tag{3.66}$$

Remark. Practical limiters would have various speed ups to the above implementation as described in [92], and depend on the method used.

3.3 Application 1: Second order finite volume

We will consider the conditions for a second order finite volume scheme to have a local maximum principle on a uniform square mesh, of cell width Δx and height Δy respectively. The interpolating polynomial aligns with a linear subcell representation

$$p_{i,j}(x, y) = \bar{u}_{i,j} + \alpha(u_x)_{i,j}(x - x_i) + \alpha(u_y)_{i,j}(y - y_j), \tag{3.67}$$
$$(u_x)_{i,j} = \frac{\bar{u}_{i+1} - \bar{u}_{i-1}}{2\Delta x}, \quad (u_y)_{i,j} = \frac{\bar{u}_{j+1} - \bar{u}_{j-1}}{2\Delta y}, \tag{3.68}$$

where α arises from the slope limiter. This subcell representation satisfies the conservation property $\frac{1}{\Delta x \Delta y} \int_{x_{i-1/2}}^{x_{i+1/2}} \int_{y_{j-1/2}}^{y_{j+1/2}} p_{i,j}(x, y) = \bar{u}_{i,j}$. The flux contributing quadrature points are at the midpoint of each face, and the quadrature point evaluations for cell (i, j) are the right left up and down values defined below

$$u_{i,j}^R = p_{i,j}(x_{i+1/2}, y_j), \tag{3.69}$$
$$u_{i,j}^L = p_{i,j}(x_{i-1/2}, y_j), \tag{3.70}$$

$$u^U_{i,j} = p_{i,j}(x_i, y_{j+1/2}), \tag{3.71}$$
$$u^D_{i,j} = p_{i,j}(x_i, y_{j-1/2}). \tag{3.72}$$

The decomposition of the cell average can be conveniently found in terms of the cell quadrature points as $\bar{u}_{i,j} = \frac{1}{4}(u^R_{i,j}+u^L_{i,j}+u^U_{i,j}+u^D_{i,j})$, there are no non-flux contributing quadrature points. The right-hand flux through face $(i+1/2, j)$ is computed by the second order Gauss quadrature at these points, $\frac{|\Delta y|}{|\Delta x \Delta y|} F_{i,i+1}(u^R_{i,j}, u^L_{i+1,j}, \boldsymbol{v}\cdot n_{i,i+1})$. So that the method can be written as the sum of 4 local Riemann problems

$$Rei_{i+1/2,j} = \frac{1}{4}[u^R_{i,j} - \frac{4\Delta t}{\Delta x} F(u^R_{i,j}, u^L_{i+1,j}, \boldsymbol{v}\cdot n_{i,i+1})], \tag{3.73}$$

there are no corner defined flux contributing quadrature points, and only one face defined flux contributing quadrature point located at the midpoint of each face. Therefore the cell mean evolution equation for cell (i, j)

$$\bar{u}^{n+1}_{i,j} = \bar{u}^n_{i,j} - \frac{\Delta t}{\Delta x} F_{i,i+1} - \frac{\Delta t}{\Delta x} F_{i,i-1} - \frac{\Delta t}{\Delta y} F_{j,j+1} - \frac{\Delta t}{\Delta y} F_{j,j-1}, \tag{3.74}$$

is a monotonic function of the edge defined quadrature points $u^R_{i,j}$, $u^L_{i+1,j}$,$u^L_{i,j}$, $u^R_{i-1,j}$, $u^U_{i,j}$, $u^D_{i,j+1}$, $u^D_{i,j}$, $u^U_{i,j-1}$, when the following local Courant number conditions holds $\frac{\Delta t}{\Delta x}(\boldsymbol{v}\cdot n_{i,i+1})^+$, $\frac{\Delta t}{\Delta x}(\boldsymbol{v}\cdot n_{i,i-1})^+$, $\frac{\Delta t}{\Delta y}(\boldsymbol{v}\cdot n_{j,j+1})^+$, $\frac{\Delta t}{\Delta y}(\boldsymbol{v}\cdot n_{j,j-1})^+ \leq \frac{1}{4}$.
The Courant number is now a concept to be interpreted on edges

$$\inf_{K\in\mathcal{M}} \inf_{L\in N(K)} \frac{\Delta t|\sigma_{KL}|(\boldsymbol{v}\cdot n_{KL})^+}{|K|} \leq 1/4, \tag{3.75}$$

but one can pessimistically write this in terms of a more convenient cell defined Courant number as

$$C_K = \sum_{L\in N(K)} \frac{\Delta t|\sigma_{KL}|(\boldsymbol{v}\cdot n_{KL})^+}{|K|} \leq 1/4, \quad \forall K \in \mathcal{M}, \tag{3.76}$$

if one assumes incompressibility 1/4 becomes 1/2.
For a well-defined triangular discretisation, one could expect Courant number restrictions of $1/3, 2/3$ respectively for compressible and incompressible flow [91]. This type of argument can be extended to more general meshes using the geometric shape parameter of Barth [78, 91].
We have defined the method and shown that it is a monotone function of quadrature points, equivalent to using Theorem 3.2.1 with $w_q^{\sigma_{KL}} = 1$, $w_q^K = 1/4$ and identifying no non-flux contributing quadrature points, no vertex defined flux contributing quadrature points, and one flux contributing quadrature point per face of the cell. However, we have not introduced how the new $N^{s+1}(K) \cup N^s(K)$-MP limiters $s = 0, 1$, will locally limit u^R, u^L, u^D, u^U. Before we do so we introduce the Barth and Jespersen multidimensional limiter.
Barth and Jespersen [42], introduce a computationally convenient slope limiter de-

signed on a different notion of monotonicity, where the subcell reconstruction values within each cell are required not to exceed its local neighbours cell mean values. It is defined as follows,

Definition 3.3.1 (Barth and Jespersen limiter)**.**

1. Compute the local neighbours cell mean for local bounds of cell K

$$[m_K, M_K] := [\min_{L \in N(K) \cup K} \bar{u}_L^n, \max_{L \in N(K) \cup K} \bar{u}_L^n]. \tag{3.77}$$

2. Compute a quadrature point correction factor α_q,

$$\alpha_q = \begin{cases} \min\{1, \frac{M_K - \bar{u}_K}{p_K(x_q) - \bar{u}_K}\} & \text{where} \quad p_K(x_q) - \bar{u}_K > 0, \\ \min\{1, \frac{m_K - \bar{u}_K}{p_K(x_q) - \bar{u}_K}\} & \text{where} \quad p_K(x_q) - \bar{u}_K < 0, \\ 1 \quad p_K(x_q) - \bar{u}_K = 0. \end{cases} \tag{3.78}$$

   to ensure the subcell reconstruction at x_q is locally bounded by $[m_K, M_K]$.

3. Limit the entire subcell representation based on the worst violator of the local bounds

$$\alpha_K = \min_{L \in N(K)} \min_{q \in \sigma_{KL}} \alpha_q, \tag{3.79}$$

   so that $\tilde{p}_K(\boldsymbol{x}) = \bar{u}_K + \alpha_K(p_K(\boldsymbol{x}) - \bar{u}_K) \in [m_K, M_K]$ is locally bounded for all quadrature points.

The Barth and Jespersen limiter only ensures that the limited subcell representation satisfies $p_K(x_{KL}) \in [m_K, M_K] = [\min_{L \in N(K) \cup K} \bar{u}_L^n, \max_{L \in N(K) \cup K} \bar{u}_L^n]$, $p_L(x_{KL}) \in [m_L, M_L] = [\min_{M \in N(L) \cup L} \bar{u}_M^n, \max_{M \in N(L) \cup L} \bar{u}_M^n]$. Therefore, the Barth and Jespersen limiter does not satisfy the sufficient conditions of Theorem 3.2.1 for a local boundedness principle of the form

$$m_K = \max_{L \in N(K) \cup K} u_L \leq u_K^{n+1} \leq \max_{L \in N(K) \cup K} u_L = M_K. \tag{3.80}$$

Instead, the Barth and Jespersen limiter satisfies the following principle

$$\max_{L \in N^2(K) \cup N(K)} u_L \leq u_K^{n+1} \leq \max_{L \in N^2(K) \cup N(K)} u_L, \tag{3.81}$$

with respect to cell neighbours. This fact is understated in the literature, but can be found in figure 5 of Park, Yoon and Kim [91] and follows directly from Theorem 3.2.1. There are also different type of limiters based on vertex sharing neighbourhood limiting principles, this is beyond the scope of this work but will be introduced for numerical work. We introduce the Park/Kuzmin vertex-based limiter [91, 94] which satisfies the vertex sharing neighbour maximum principle, and benefits from

(specifically exploits) the fact that linear subcell extrema are contained at the vertex of a cell. It is defined as follows,

Definition 3.3.2 (Kuzmin Vertex Limiter/ Park Yoon Kim MLP limiter)**.**

1. Compute the vertex defined local bounds for a maximum principle

$$[m_v, M_v] := [\min_{i \in N(v)} \bar{u}_i^n, \max_{i \in N(v)} \bar{u}_i^n], \tag{3.82}$$

   where $N(v) := \{L | L \cap v \neq \emptyset\}$ denotes the set of cells which share the vertex v.

2. Compute a vertex correction factor α_v,

$$\alpha_v = \begin{cases} \min\{1, \frac{M_v - \bar{u}_K}{p_K(x_v) - \bar{u}_K}\} & \text{if} \quad p_K(x_v) - \bar{u}_K > 0, \\ \min\{1, \frac{m_v - \bar{u}_K}{p_K(x_v) - \bar{u}_K}\} & \text{if} \quad p_K(x_v) - \bar{u}_K < 0, \\ 1 \quad \text{if} \quad p_K(x_v) - \bar{u}_K = 0, \end{cases} \tag{3.83}$$

   to ensure that all vertex points are locally bounded by their local vertex sharing neighbours $[m_v, M_v]$.

3. Then the entire subcell representation is limited based on the worst violator of the local maximum principle,

$$\alpha_K = \min_{v \in K} \alpha_v, \tag{3.84}$$

   so that $\tilde{p}_K = \bar{u}_K + \alpha_K(p_K(\boldsymbol{x}) - \bar{u}_K) \in [m_{VN}, M_{VN}]$ is locally bounded between the union of all vertex defined quadrature bounds. $VN(K)$ denotes the set of vertex neighbours of K and defines the local maximum principle

$$[m_{VN}, M_{VN}] = [\min_{L \in VN(K)} \bar{u}_L, \max_{L \in VN(K)} \bar{u}_L]. \tag{3.85}$$

Park [91] describes how this pertains to the following maximum principle

$$u_K^{n+1} \in [\min_{L \in VN(K)} \bar{u}_L, \max_{L \in VN(K)} \bar{u}_L]. \tag{3.86}$$

Note that the method still uses the midpoint method for the flux contributing quadrature points, and no corner points are used directly in the fluxes. This results in the computation of points such as the upper right u^{UR} corner, to determine the correction factor α_K to be used on the whole cell.
We now reintroduce the new $N(K) \cup K$-MP limiter which preserves the strictly stronger local maximum principle

$$\min_{L \in N(K) \cup K} \bar{u}_L^n \leq \bar{u}_K^{n+1} \leq \max_{L \in N(K) \cup K} \bar{u}_L^n, \tag{3.87}$$

and for this simple second order finite volume method reduces to the following procedure.

Definition 3.3.3 (simplification of the $N(K) \cup K$-MP-limiter)**.** In pseudo code the $N(K) \cup K$-MP-limiter admits the following simplification for the second order finite volume scheme.

1. Per face σ_{KL}, we compute and associate the local face defined maximum principle bounds

$$m_{\sigma_{KL}}, M_{\sigma_{KL}} = \min\{\bar{u}_K^n, \bar{u}_L^n\}, \max\{\bar{u}_K^n, \bar{u}_L^n\} \tag{3.88}$$

   this is associated to each $x_q \in \sigma_{KL}$.

2. We then per cell compute all the Barth and Jespersen quadrature corrections factors α_q to ensure

$$\bar{u}_K + \alpha_q(p_K(x_q) - \bar{u}_K) \in [m_{\sigma_{KL}}, M_{\sigma_{KL}}], \quad \forall q \in \sigma_{KL} \quad \forall L \in N(K). \tag{3.89}$$

3. Choose the smallest value,

$$\alpha = \min_{\forall q \in K} \alpha_q \tag{3.90}$$

   this ensures that the internal subcell representation $\tilde{p}_K(x) = \alpha(p_K(x) - \bar{u}_K) + \bar{u}_K$, satisfies the required edge sharing maximum principle at flux contributing quadrature points.

This is sufficient to use Theorem 3.2.1, to prove the local inclusive face sharing maximum principle. We now reintroduce the new $N^2(K) \cup N(K)$-MP limiter in Section 3.2.1 to this second order finite volume method, we see there is no need to do the corner or non-flux contributing limiting steps 2,3 and the new limiting function takes a more compact definition. It is more accurate than the Barth and Jespersen limiter and preserves the same cell mean maximum principle

$$\min_{L \in N(K) \cup K} \bar{u}_L^n \leq \bar{u}_K^{n+1} \leq \max_{L \in N(K) \cup K} \bar{u}_L^n. \tag{3.91}$$

This is achieved by dropping the assumption that the subcell reconstruction values need be bounded by the local cell means, we instead rely on the theoretical prediction of Theorem 3.2.1 and instead choose to limit both $p_L(x_q), p_K(x_q)$ based on extended edge defined maximum principles plotted in Fig. 3.3, whose union over a cell K defines the inclusive "squared" neighbourhood maximum principle. The limiter is defined as follows:

Definition 3.3.4 (simplification of $N^2(K) \cup N(K)$-MP-limiter)**.** In pseudo code the $N^2(K) \cup N(K)$-MP-limiter admits the following simplification for the second order finite volume scheme.

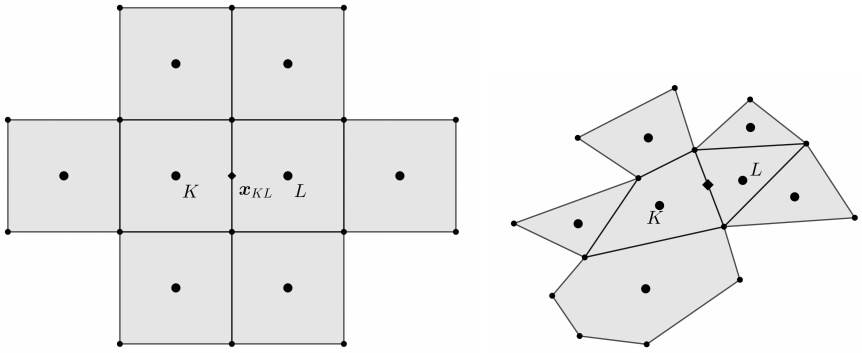


Figure 3.3: The stencil $N(K) \cup N(L)$ for a structured and unstructured mesh. In particular this region is employed by the $N^2(K) \cup N(K)$-MP limiter for the second order finite volume scheme when ensuring that both $p_K(x_{KL})$ and $p_L(x_{KL})$ are locally bounded by surrounding cell means.

1. Per face σ_{KL}, we compute and associate the local face defined maximum principle bounds

$$m_{\sigma_{KL}}, M_{\sigma_{KL}} = \min_{M \in N(L) \cup N(K)} \bar{u}_M^n, \max_{M \in N(L) \cup N(K)} \bar{u}_M^n \tag{3.92}$$

   this principle is associated to the quadrature point $x_{KL} \in \sigma_{KL}$.

2. We then per cell K compute all the Barth and Jespersen quadrature corrections factors α_q to ensure

$$\tilde{p}_K(x_q) = \bar{u}_K + \alpha_q(p_K(x_q) - \bar{u}_K) \in [m_{\sigma_{KL}}, M_{\sigma_{KL}}], \quad \forall q \in \sigma_{KL}, \forall \sigma_{KL} \in K. \tag{3.93}$$

3. Choose the smallest value,

$$\alpha = \min_{\forall q \in K} \alpha_q \tag{3.94}$$

   this ensures the limited internal subcell representation $\tilde{p}_K(x) = \alpha(p_K(x) - \bar{u}_K) + \bar{u}_K$, satisfies the required edge sharing quadrature maximum principles.

Once this is done for all cells this is sufficient to use Theorem 3.2.1, to prove the local inclusive "squared" face sharing neighbour maximum principle.

$$\min_{L \in N^2(K) \cup N(K)} \bar{u}_L^n \leq \bar{u}_K^{n+1} \leq \max_{L \in N^2(K) \cup N(K)} \bar{u}_L^n. \tag{3.95}$$

3.3.1 Factors affecting accuracy

The Barth and Jespersen limiter ensures that the subcell representation does not exceed its neighbouring cell mean values, but still allows for discontinuities at the cell boundary,

$$\tilde{p}_K^{BJ}(x_{KL}) \in [\min_{i \in N(K) \cup K} \bar{u}_i, \max_{i \in N(K) \cup K} \bar{u}_i], \tag{3.96}$$

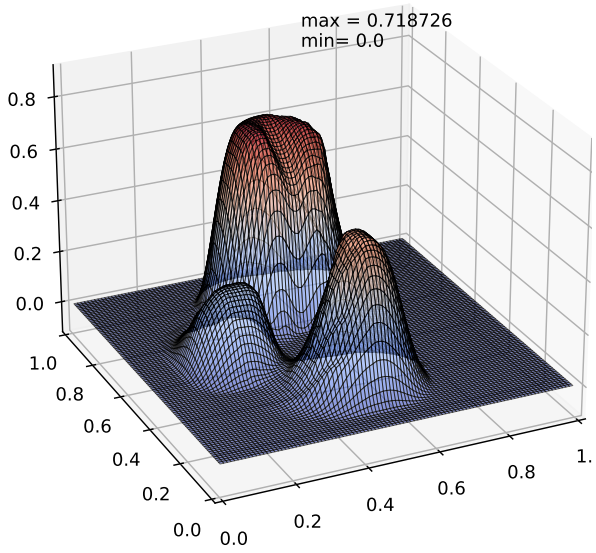
$$\tilde{p}_L^{BJ}(x_{KL}) \in [\min_{i \in N(L) \cup L} \bar{u}_i, \max_{i \in N(L) \cup L} \bar{u}_i]. \tag{3.97}$$

See $BJ(int)$, and $BJ(ext)$ in Fig. 3.2 for a visualisation of these neighbourhoods. Whereas the new $N^2(K) \cup N(K)$-MP limiter ensures the edge maximum principle

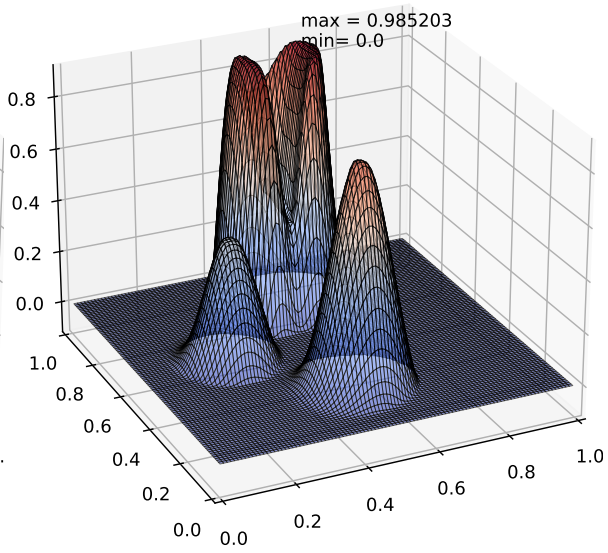
$$\tilde{p}_K^{new}(x_{KL}), \tilde{p}_L^{new}(x_{KL}) \in [\min_{i \in N(K) \cup N(L)} \bar{u}_i, \max_{i \in N(K) \cup N(L)} \bar{u}_i]. \tag{3.98}$$

See $N(K) \cup N(L)$ in Fig. 3.2 for a visualisation of this neighbourhood.
Since $N(K) \cup K \subset N(K) \cup N(L)$, and $N(L) \cup L \subset N(K) \cup N(L)$, the allowable variation is larger in the new $N^2(K) \cup N(K)$-MP limiter, therefore all possible correction factors are less severe(or equal) to those of the Barth and Jespersen limiter $\alpha_q^{BJ} \leq \alpha_q^{new}$. They both satisfy the same maximum principle on cell means, and the new $N^2(K) \cup N(K)$-MP limiter uses fluxes more similar to that of the higher order flux. This result holds for all meshes and is to be expected as the Barth and Jespersen demands different properties of the subcell reconstruction [42]. It should be noted that the $N(K) \cup K$-MP and $N^2(K) \cup N(K)$-MP limiters still demand local boundedness of the subcell reconstructed quadrature points which ensures a positivity of these reconstructions.
Park et al. [91] do similar analysis to show that the Barth and Jespersen limiter is worse than the Kuzmin/MPL limiter. However, the Kuzmin/MPL limiter enforces a different maximum principle to the Barth and Jespersen Limiter, and this is a mesh dependent result. This can be seen in Fig. 3.2 that on triangles $N^2(K) \cup N(K) \subset VN(K)$, but on rectangles $VN(K) \subset N^2(K) \cup N(K)$. Heuristically we expect the Barth and Jespersen limiter to be less accurate on the 3 sided meshes, but more accurate on 4 sided meshes. Both clearly have advantages and disadvantages, and enforce different properties.

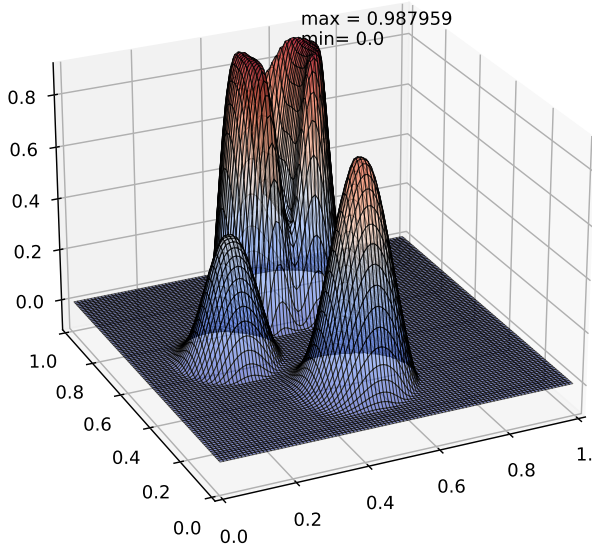
3.3.2 Numerical results



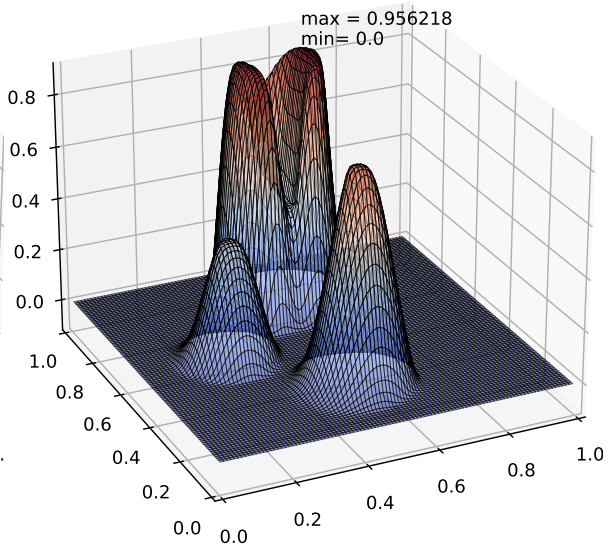
(a) SSP22 $N(K) \cup K$-MP unstructured limiter.



(b) SSP22 Barth and Jespersen limiter.



(c) SSP22 $N^2(K) \cup N(K)$-MP limiter.



(d) SSP22 KUZ limiter.

Figure 3.4: Solid body rotation of the LeVeque initial conditions at 100×100 resolution, using SSP22 timestepping with limiters at each internal substage of the Shu Osher representation. Fig. 3.4a is the $N(K) \cup K$-MP limiter. Fig. 3.4b is the Barth and Jespersen limiter. Fig. 3.4c is the $N^2(K) \cup N(K)$-MP limiter. Fig. 3.4d is the Kuzmin/Park vertex limiter.

In this subsection we will present the results of all four multidimensional limiter functions, after the solid body rotation test case of the LeVeque initial conditions in Fig. 3.4 and Table 3.1. We also present convergence plots and table for the multidimensional limiters under four different flows in Fig. 3.5 and Table 3.2.
We plot the final time step for the $N(K) \cup K$-MP multidimensional limiter in Fig. 3.4a, the Barth and Jespersen limiter in Fig. 3.4b, the $N^2(K) \cup N(K)$-MP multidimensional limiter in Fig. 3.4c and the Kuzmin limiter in Fig. 3.4d, after the solid body rotation test case of the LeVeque initial conditions. The relative errors compared to the analytic solution and the peak value at the final time-step are extracted from the solid body rotation test case an put in Table 3.1, this is done

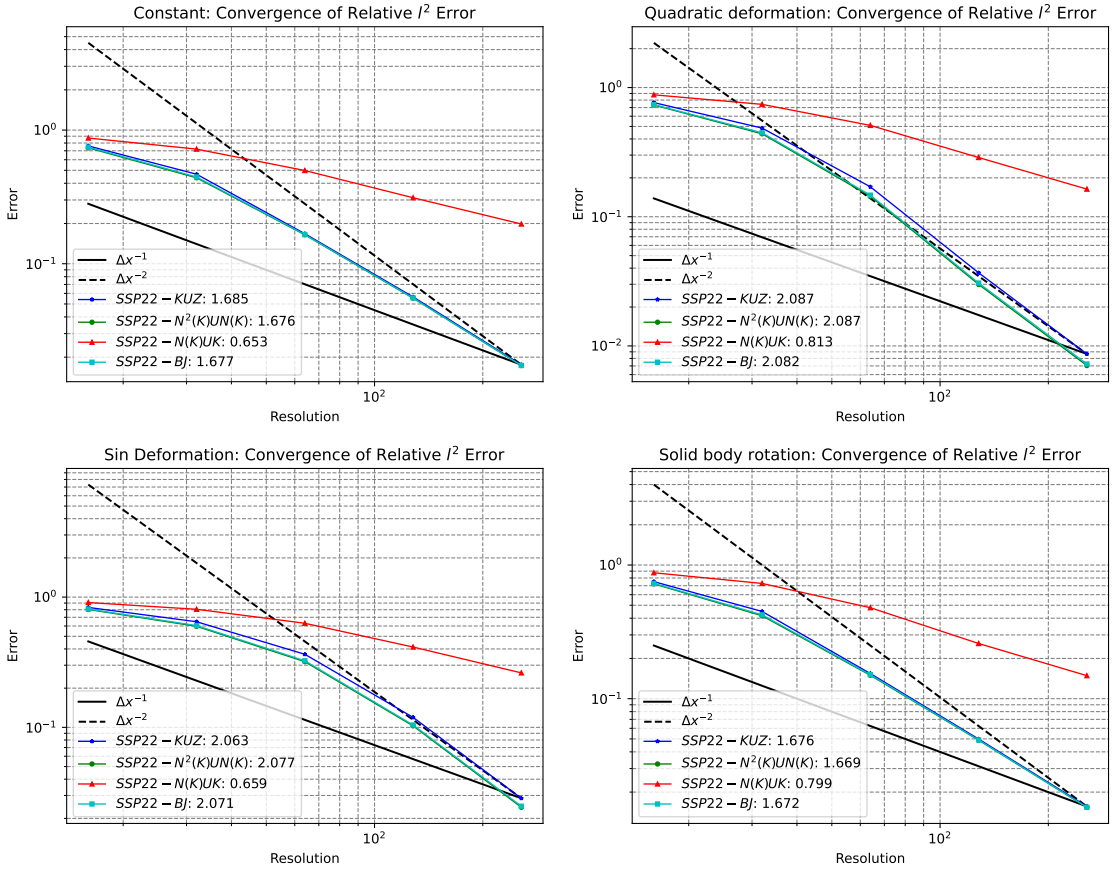


Figure 3.5: Log-log plot of relative error of the SSP22 multidimensional limiters in L^2 for the smooth cosine bell initial conditions but different velocity fields. The velocity fields are defined in Chapters 2 and 4 by the stream functions Equations (2.89) to (2.92).

to compare the Barth and Jespersen, $N^2(K) \cup N(K)$, and Kuzmin limiter whose performance is similar. For all limiters and all tests, the timestepping is the optimal two stage second order strong stability preserving Runge Kutta method SSP22, and the limiting procedure is employed at each substage in the optimal Shu Osher representation. The LeVeque initial conditions are directly sampled from Eq. (4.37) and undergo the solid body rotation test case defined by the stream-function Eq. (4.32). The solid body rotation test case is performed with 100×100 resolution with 1256 timesteps, with a Courant number maximum near 0.5.

The results of Fig. 3.4 indicate the $N(K) \cup K$-MP for the solid body rotation of the LeVeque initial condition Fig. 3.4a, is noticeably less accurate than the other multidimensional limiters. The Barth and Jespersen limiter, the Kuzmin limiter and the new $N^2(K) \cup N(K)$-MP limiter all produce similar visual results for the LeVeque solid body rotation test in Fig. 3.4. We have extracted the relative error norms in L^1, L^2, L^∞, as well as the height of the maximum value at final time-step from the Solid body rotation test case in Table 3.1. We can see that the $N^2(K) \cup N(K)$-MP is slightly more accurate than the Barth and Jespersen limiter(in $L2$) , and the Barth and Jespersen limiter is slightly more accurate(in $L2$) than the Kuzmin limiter. This is consistent with the theoretical prediction from Section 3.3.1, where we predicted the $N^2(K) \cup N(K)$-MP limits the subcell representation less than the Barth and

Jespersen limiter, and explain that for four sided meshes the Barth and Jespersen limiter will likely limit the subcell representation less than the Kuzmin limiter.
Fig. 3.5 contains four convergence plots for each limiter, each convergence plot has used the same C^1 compact cosine bump defined by Eq. (4.36), but use the four different velocity fields defined by the stream-functions Equations (2.89) to (2.92), these are the same flow test cases in Chapter 4. The relative error norm in L^p is computed using $re_{L^p}(u) := \frac{||u-u_e||_p}{||u_e||_p}$ where u_e denotes the analytic solution. To approximate the order of the methods in Table 3.2, the relative error in L^2 is computed at the resolutions 128×128, and 256×256 at a fixed Courant number with maximum value near 1/2. We then use the log-log-gradient to approximate the order of the method $\log(re_{L^2}(u_{256^2})/re_{L^2}(u_{128^2}))\log(2)^{-1}$. To see how convergence is changing with resolution see the log log plot of relative L^2 limiter Fig. 3.5.
The Barth and Jespersen limiter, the Kuzmin limiter and the new $N^2(K) \cup N(K)$-MP limiter all produce similar convergence results for the 4 convergence tests in Table 3.2 between order 1.6 and 2.1. For the same convergence test case (when the velocity field is the same) the Barth and Jespersen limiter, the Kuzmin limiter and the $N^2(K) \cup N(K)$-MP limiter have less than a 0.01 difference in observed order. The $N(K) \cup K$-MP limiter, observed a drop in order of convergence Table 3.2 and Fig. 3.5 and shows worse accuracy in Fig. 3.5 for all test cases.

metric	$N^2(K) \cup N(K)$	BJ	KUZ
Relative error L^1	**0.321384**	0.323794	0.334256
Relative error L^2	**0.368622**	0.369762	0.372376
Relative error L^∞	0.849103	0.847545	**0.813771**
$\max_{\forall i,j} u_{i,j}^{1256}$	**0.987959**	0.985203	0.956218
$\min_{\forall i,j} u_{i,j}^{1256}$	0	0	0

Table 3.1: This table contains error norms and the maxima and minima at the final time-step from the solid body rotation case for the $N^2(K) \cup N(K)$, Barth and Jespersen and Kuzmin limiter. Bold values indicate the smallest error norms, or the least clipped maxima.

Convergence		Test cases	Observed	Order	
Scheme	Limiter	Diag	Quad	Sin	Sbr
ssp22	$N(K) \cup K$	0.653	0.813	0.659	0.799
ssp22	BJ	1.677	2.082	2.071	1.672
ssp22	$N^2(K) \cup N(K)$	1.676	2.087	2.077	1.669
ssp22	KUZ	1.685	2.087	2.063	1.676

Table 3.2: This table contains the convergence rate of relative L^2 error between running at 128×128 as compared to 256×256 resolution for the four flow cases with the limiter activated.

3.4 Application2: Higher order limiting

3.4.1 FV4: Fourth order finite volume

We define a fourth order finite volume method; it is directly applicable for a 2d orthogonal grid and bears some similarity to the MCORE [89] finite volume dynamical core, but does not use a convolution and deconvolution strategy for the fluxes. Instead the scheme uses direct evaluations at Gauss points from the high order subcell representation.
It can be defined by a sequence of compositions

$$\bar{u}^{n+1} = (\mathcal{E} \circ \mathcal{R} \circ \mathcal{Q} \circ \mathcal{G} \circ \mathcal{P}) \circ \bar{u} \tag{3.99}$$

in pseudo code format as follows.

1. We use the following fourth order projection map $\mathcal{P}_4 : \bar{u}_{i,j} \mapsto u_{i,j} + O(\Delta x^4 + \Delta y^4)$ to approximate point values from cell mean values. It is consistent with respect to constants.

$$u_{i,j} = \bar{u}_{i,j} - \frac{1}{24}[\bar{u}_{i+1,j} - 2\bar{u}_{i,j} + \bar{u}_{i-1,j}] - \frac{1}{24}[\bar{u}_{i,j+1} - 2q\bar{u}_{i,j} + \bar{u}_{i,j-1}] \quad \forall(i,j) \tag{3.100}$$

2. We use the gradient map

$$\mathcal{G}_3 : u \mapsto u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}, u_{xxx}, u_{xxy}, u_{xyy}, u_{yyy} \tag{3.101}$$

defined by the 4th order centred finite difference weights

$$w1 = 1/12([-1, 8, 0, -8, 1]) \tag{3.102}$$
$$w2 = 1/12([-1, 16, -30, 16, -1]) \tag{3.103}$$
$$w3 = 1/8([-1, 8, -13, 0, 13, -8, 1]) \tag{3.104}$$

to construct u_x, u_{xx}, u_{xxx} and u_y, u_{yy}, u_{yyy} from the newly computed point values. We use these newly computed values, and the finite difference stencil $1/12([-1, 8, 0, -8, 1])$, to compute all the missing cross term derivatives u_{xy}, u_{xxy}, u_{xyy} within each subcell representation.

3. $\mathcal{Q}_4 : (x_q, y_q) \mapsto p_K(x_q, y_q)$, $\forall(x_q, y_q) \in K$, We compute a set of 4th order accurate quadrature point evaluations at (x_q, y_q) for all cells. We do so by evaluating the following formula of the subcell representation

$$P_{i,j}(x,y) = \bar{u}_{i,j} + (x - x_i)u_x + (y - y_j)u_y + \tag{3.105}$$
$$\frac{1}{2}[\{(x - x_i)^2 - \frac{\Delta x^2}{12}\}u_{xx} + 2(x - x_i)(y - y_i)u_{xy} + \{(y - y_i)^2 - \frac{\Delta y^2}{12}\}u_{yy}] \tag{3.106}$$
$$\frac{1}{3!}[(x - x_i)^3 u_{xxx} + 3(x - x_i)^2(y - y_i)u_{xxy} \tag{3.107}$$

$$+ 3(x - x_i)(y - y_i)^2 u_{xyy} + (y - y_i)^3 u_{yyy}]. \tag{3.108}$$

4. $\mathcal{R}_4$(Resolve Riemann Reconstruct flux) We resolve the local quadrature defined Riemann problems

$$F(x_{i+1/2}, y_q, u(x_q, y_q)) = p_K(x_{i+1/2}, y_q)u(x_{i+1/2}, y_q)^+ + p_L(x_{i+1/2}, y_q)u(x_{i+1/2}, y_q)^-, \tag{3.109}$$

using the local state interpolated evaluations of quadrature velocity. The flux is computed by a 4th order Gauss quadrature, for example the right edge is computed using

$$F_{i,i+1} = \sum_{q_k \in \sigma_{i,i+1}} w_{q_k}[F(x_{i+1/2}, y_{q_k}, u(x_{i+1/2}, y_{q_k}))] \tag{3.110}$$

$$\text{where} \quad w_{q_1}, w_{q_2} = [1/2, 1/2], y_{q_1}, y_{q_2} = [y_{j+1/2} - \frac{\Delta y}{2\sqrt{3}}, y_{j+1/2} + \frac{\Delta y}{2\sqrt{3}}] \tag{3.111}$$

which is a weighted sum of the computed quadrature point defined Riemann problems.

5. The final stage involves the normal cell mean evolution procedure, where the fluxes on each face are used to update the solution.

$$\bar{u}_{i,j}^{n+1} = \bar{u}_{i,j}^{n+1} - \frac{\Delta t}{|K|} \sum_{L \in N(K)} |\sigma_{KL}| F_{K,L}. \tag{3.112}$$

3.4.2 $N^2(K) \cup N(K)$-MP limiter for FV4

Based on Theorem 3.2.1. We wish to employ the $N^2(K) \cup N(K)$-MP limiter introduced in Section 3.2.1, to the new fourth order finite volume method. We first remark on some non-trivial facts about this specific finite volume construction and how the limiter interacts with the scheme non uniquely.

1. The 8 flux contributing quadrature points $(\boldsymbol{x}_q, q \in K^{fc})$ for cell (i, j) are located at the positions

$$[x_i \pm \frac{\Delta x}{2}, y_j \pm \frac{\Delta y}{2\sqrt{3}}], \quad [x_i \pm \frac{\Delta x}{2\sqrt{3}}, y_j \pm \frac{\Delta y}{2}]. \tag{3.113}$$

These are limited by an edge defined quadrature maximum principle Fig. 3.3. There are two points per face sharing the same maximum principle.

2. The subcell representation is cell mean preserving.

$$\frac{1}{h^2} \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} P_{i,j}(x, y) dx dy = \bar{u}_{i,j} \tag{3.114}$$

3. There exists a convex Zhang-acceptable decomposition of the cell average onto

flux contributing quadrature points, it is not unique, the following representation has been found to be convenient

$$\bar{q}_{i,j} = \frac{1}{2}P(x_i, y_j) + \frac{1}{16}\sum_{q\in K^{fc}} P(\boldsymbol{x}_q). \tag{3.115}$$

The cell midpoint (x_i, y_j) is not flux contributing and must satisfy the traditional maximum principle associated with non-flux contributing step 2 in Section 3.2.1 on $N(K)\cup K$ or the larger $N^2(K)\cup N(K)$ stencil.

4. The local Riemann problem at the upper quadrature point on the right face takes the form.

$$Rei_{i+1/2,j+\frac{\Delta y}{2\sqrt{3}}} = \frac{1}{16}[u^{R1}_{i,j} - \frac{8\Delta t}{\Delta x}F(u^{R1}_{i,j}, u^{L1}_{i+1,j}, \boldsymbol{v}_{i+1/2,j+\frac{\Delta y}{2\sqrt{3}}}\cdot n_{i,i+1})], \tag{3.116}$$

5. The Courant number limit is 1/8 for compressible flow

$$C_K = \sum_{L\in N(K)} \frac{\Delta t|\sigma_{KL}|(\boldsymbol{v}\cdot n_{KL})^+}{|K|} \leq 1/8, \quad \forall K\in\mathcal{M}, \tag{3.117}$$

and 1/4 for incompressible flow. This can be identified by making the associations $w_q^{\sigma_{KL}} = 1/2$ and $w_q^K = 1/16$.

We have stated enough about the scheme to use the $N^2(K)\cup N(K)$-MP limiter.

Definition 3.4.1 (simplification of $N^2(K)\cup N(K)$-MP-limiter)**.** We point to the Fig. 3.6 and captions in Fig. 3.6.

1. Per face σ_{KL}, we compute and associate the local face defined maximum principle bounds

$$m_{\sigma_{KL}}, M_{\sigma_{KL}} = \min_{M\in N(L)\cup N(K)} \bar{u}^n_M, \max_{M\in N(L)\cup N(K)} \bar{u}^n_M. \tag{3.118}$$

this principle is associated to both quadrature points $x_q\in\sigma_{KL}$ at the face.

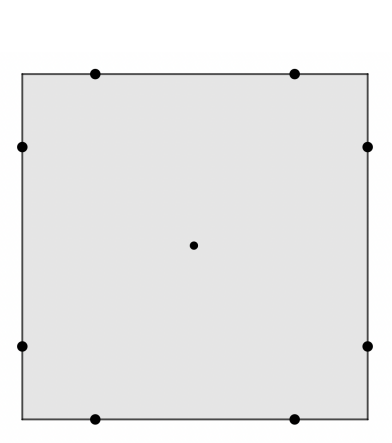
2. Per cell K we associate the desired maximum principle

$$[m_{K^{nfc}}, M_{K^{nfc}}] = [\min_{L\in N^2(K)\cup N(K)} \bar{u}^n_L, \max_{L\in N^2(K)\cup N(K)} \bar{u}^n_L], \tag{3.119}$$

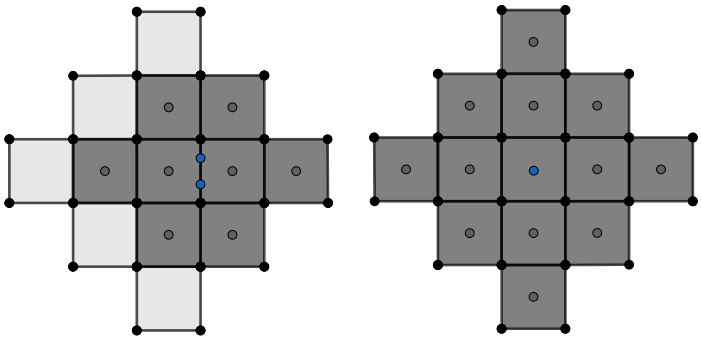
this is associated to the one non-flux contributing quadrature point $x_q\in K^{nfc}$ located at the cell midpoint.

3. We then per cell compute all the Barth and Jespersen quadrature corrections factors α_q, to ensure $\tilde{p}_K(x) = \alpha(p_K(x) - \bar{u}_K) + \bar{u}_K$, satisfies the conditions in Theorem A.4.1.

$$\tilde{p}_K(x_q)\in[m_{K^{nfc}}, M_{K^{nfc}}], \quad \forall q\in K^{nfc}, \tag{3.120}$$



(a) Points used in the Zhang-acceptable cell mean decomposition of FV4 as in Equation (3.115), there are two flux contributing quadrature points per face at Gauss nodes and one cell cell midpoint evaluation.



(b) Flux contributing quadrature points at the edge σ_{KL} are limited based on being bounded by the cell mean values $\bar{u}$ in the $N(K) \cup N(L)$ region (darker grey left diagram). Non flux contributing quadrature point evaluation of the midpoint $u(x_i, y_j)$ is limited by based on being locally by the cell mean values $\bar{u}$ in the $N^2(K) \cup N(K)$ region (darker grey right diagram).

Figure 3.6: Points from the FV4 cell mean decomposition, and interaction with the $N^2(K) \cup N(K)$-MP limiter.

$$\tilde{p}_K(x_q) \in [m_{\sigma_{KL}}, M_{\sigma_{KL}}], \quad \forall q \in \sigma_{KL} \cap K^{fc}, \quad \forall L \in N(K), \tag{3.121}$$

by choosing the smallest value

$$\alpha = \min_{\forall q \in K} \alpha_q. \tag{3.122}$$

that ensures the limited internal subcell representation $\tilde{p}_K(x) = \alpha(p_K(x) - \bar{u}_K) + \bar{u}_K$, satisfies the required edge sharing quadrature maximum principles for both flux contributing quadrature points and the cell midpoint satisfies a non-flux contributing quadrature point maximum principle Fig. 3.6.

Remark. There exists other Zhang-acceptable decompositions of the cell mean such as

$$\bar{q}_{i,j} = \frac{p}{2} P(x_i, y_j) + \frac{1-p}{8} [P(x_i, y_{j+1/2}) + P(x_i, y_{j-1/2}) + P(x_{i+1/2}, y_j) + P(x_{i-1/2}, y_j)] \tag{3.123}$$

$$+ \frac{1}{16} \sum_{q \in K^{fc}} P(\boldsymbol{x}_q) \quad p \in [0, 1]. \tag{3.124}$$

Such that the free parameter p could be locally varied to minimise the Barth and Jespersen correction factors arising from the non-flux contributing quadrature principle, this could be used for increased accuracy. We take $p = 1$ and move on.

3.4.3 Numerical demonstration of order

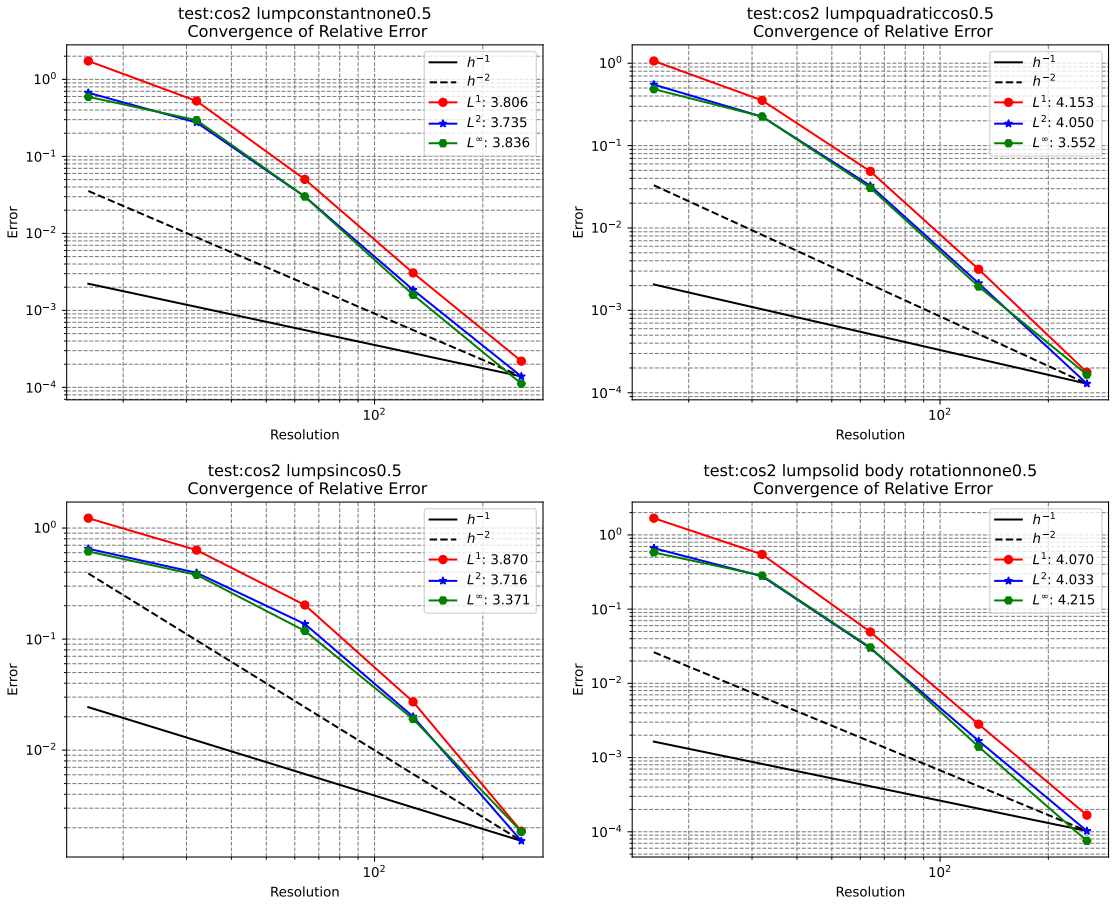


Figure 3.7: Log-log plot of relative error of the SSP33 FV4 scheme without limiting in L^1, L^2, L^∞ for the smooth cosine bell initial conditions but different velocity fields. This is done up to 256×256 resolution, at Courant number maximum near $1/2$. It appears between third and fourth order for a variety of test cases. All these tests use a smooth cosine bell for initial conditions, but use the four different velocity fields defined inEquations (2.89) to (2.92) but directly sample the velocity functions.

We quickly check the unlimited scheme is indeed 4th order as it has not been proposed before in this exact formulation. We use a compact cosine bump

$$q = [\frac{1}{2}(1 + \cos(\pi \min(\frac{r}{0.15}, 1))]^2, \quad \text{where} \quad r = \sqrt{(x - 0.5)^2 + (y - 0.75)^2}), \tag{3.125}$$

as the initial condition, and test for convergence using the previously defined incompressible flow fields Eqs. (2.90) to (2.92) and (4.33), the velocities evaluated directly at quadrature points (not using the stream function formulation). When we use the SSP33 time stepping algorithm Eq. (2.63) without limiting we get the theoretical predicted convergence behaviour of between 3 and 4 in the three lines of Table 3.3, in L^1, L^2, L^∞ norms and for all the test cases Eqs. (2.90), (2.92) and (4.33). The log log plot of the same results are included in Fig. 3.7 where 3rd/4th order is observed.

Convergence			Test cases	Observed	Order	
Scheme	limiter	norm	Diag	Quad	Sin	Sbr
SSP33 FV4	none	L^1	3.806	4.153	3.870	4.070
SSP33 FV4	none	L^2	3.735	4.050	3.716	4.033
SSP33 FV4	none	L^∞	3.836	3.552	3.371	4.215

Table 3.3: This table contains the convergence rate of relative L^1, L^2, L^∞ errors between running at 128×128 as compared to 256×256 resolution for the four separate flow cases.

3.4.4 Numerical demonstration of new limiters

Solid body rotation of the LeVeque initial conditions for the new finite volume method described at the start of Section 3.4.1 is performed with four different limiting procedures described in section Section 3.2.1 and plotted in Fig. 3.8. Where the timestepping is the optimal three stage third order strong stability preserving Runge Kutta method SSP33, and the limiting procedure is employed at each substage in the optimal Shu Osher representation. For the solid body rotation test case we evaluate the solid body rotational velocity field at the Gauss quadrature points. We also directly sample the LeVeque initial conditions.

The first row of solid body rotation results in Fig. 3.8 show our new finite volume method without any limiter. In the second row our new limiter $N(K) \cup K$-MP is applied at each stage of the Shu Osher representation. In the third row our new limiter $N^2(K) \cup N(K)$-MP is applied at each stage of the Shu Osher representation. In the last row we use a boundedness limiter by the old timestep maxima and minima, which can be thought of as $N^{s+1}(K) \cup N^s(K)$ with s large enough to cover the entire domain. Column one corresponds to maximum Courant number 0.5 with 100×100 resolution with a ghost of the initial condition, and the trace of error on the bottom contour. Column 2 corresponds to a maximum Courant number 0.3 with 200×200 resolution and we have plotted a boundedness violation contour at $z = -0.1$. In the spirit of disclosure, we are running at over what the theoretical maximum Courant number should run at, this because the solid body rotation test runs at a smaller Courant (1/4) number regime near the centre of rotation where the tracer initial conditions are defined, and the tracer is zero out of this region. No violations of maximum and minima have been observed even at machine precision. The new unlimited finite volume scheme with SSP33 timestepping in the first row of Fig. 3.8, observes good resolution of the cone and cosine bell but general unboundedness and unphysical oscillations near the slotted cylinder. The second row in Fig. 3.8 involves the same experiment but with our new limiter $N(K) \cup K$, it observes boundedness to machine precision, but is heavily diffusive. The third row in Fig. 3.8 involves the same experiment but with our new limiter $N^2(K) \cup N(K)$, it observes boundedness to machine precision, it clips the extrema of the cone at both resolutions, the back wall of the slotted cylinder is degraded slightly at the 100×100 low resolution, the high-resolution slotted cylinder does still have some

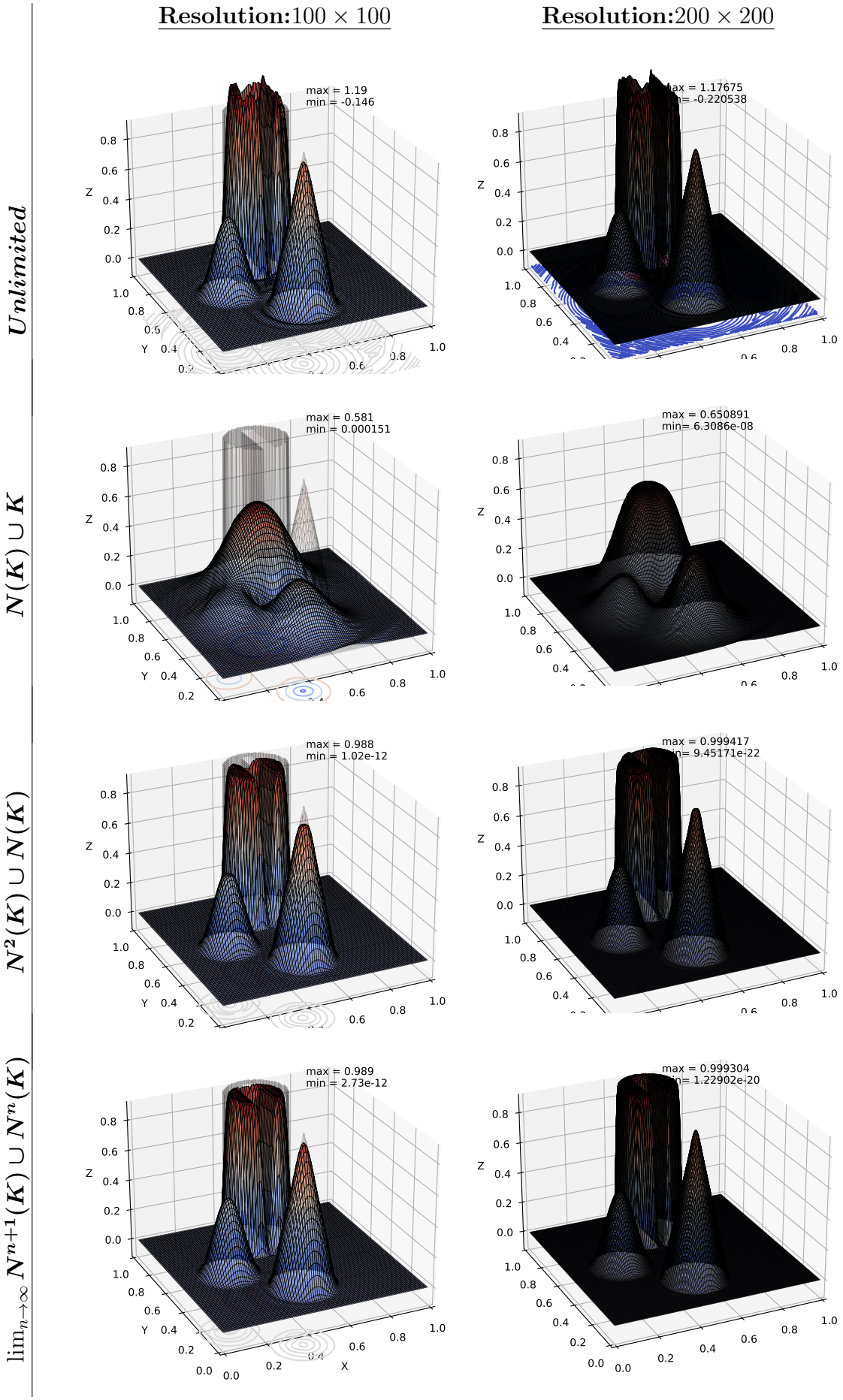
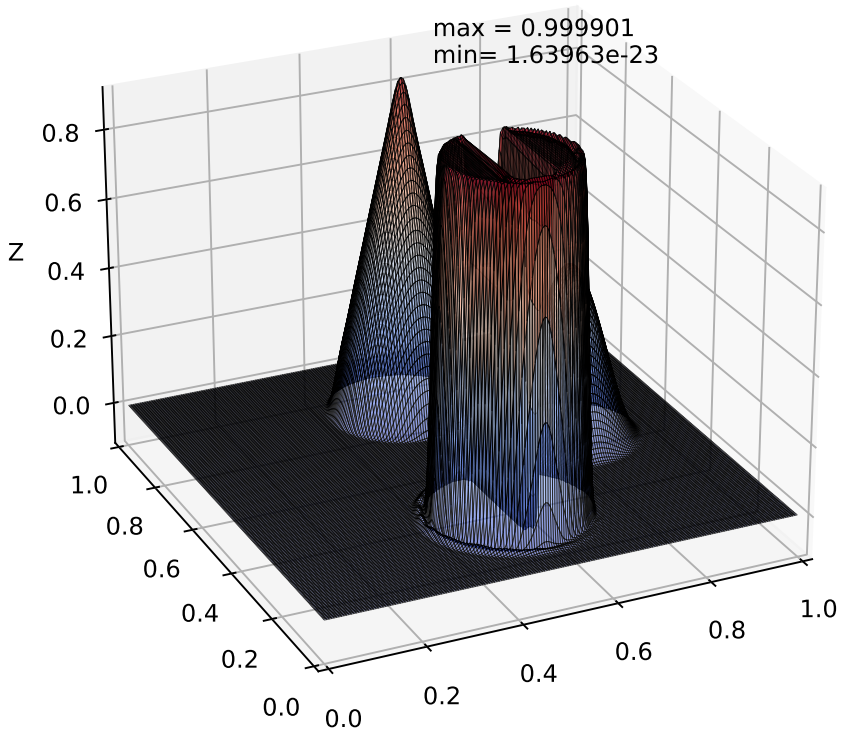


Figure 3.8: Final timestep of solid body rotation of the SSP33 FV4 scheme, with the Unlimited, $N(K) \cup K$, $N^2(K) \cup N(K)$, and boundedness limiters in each row. Each column corresponds to a different resolution.

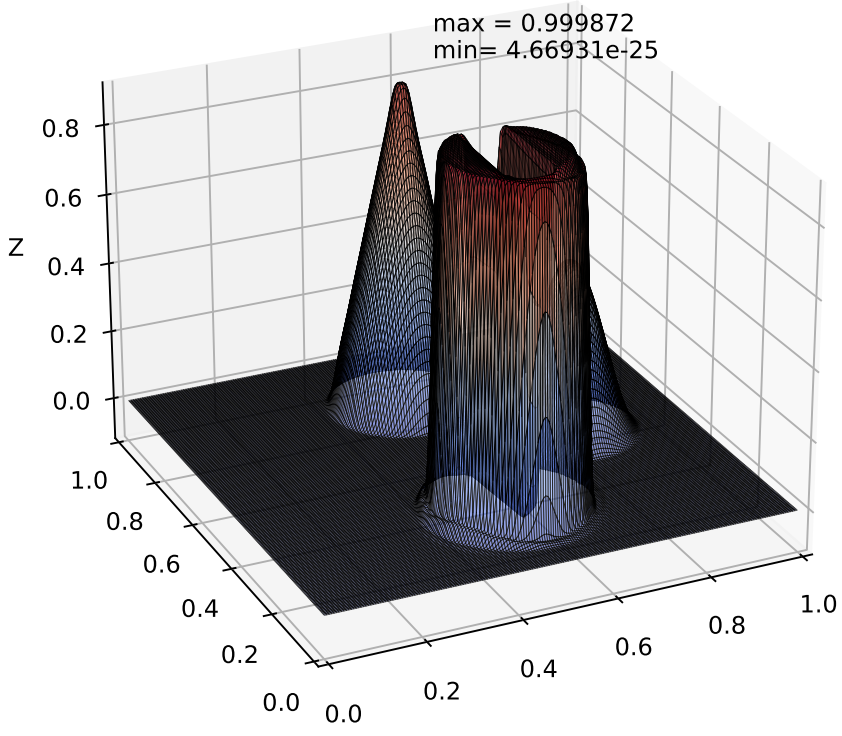
degradation on the left slope. In the final row of Fig. 3.8, we produce the results of the traditional form of the limiter [92] which enforces a global boundedness principle based on the last time-step maximum and minimum. There is a clear improvement in accuracy over the local maximum principle, the peak of the cone is well resolved and has not been limited at both resolutions, the back wall of the slotted cylinder is accurately represented as compared with the local maximum principle limiters.
In Fig. 3.9 we have plotted the solution after 1/2 a rotation of the unlimited and the $N^2(K) \cup N(K)$-MP limiter. For the global maximum principle limiter, we see on the top of the slotted cylinder there is an indent in both the left and right halves, this is a local minimum generation. We also see a ring of local maxima at the base of the slotted cylinder, this is local maxima generation. Whereas the $N^2(K) \cup N(K)$-MP limiter has suppressed these extrema to some extent, by joining the ring of local maxima to the slotted cylinder. This coalescing does appear to have larger error as expected. The $N^2(K) \cup N(K)$-MP limiter has degraded the wall on the slotted cylinder by mild "landsliding", the indent is no longer a local minimum.

3.4.5 Conclusion

This new limiter framework and extension of the theoretical work in [43] [24], allows for many schemes to maintain a local maximum principle. We have followed the general approach introduced in [43] closely enough so that this method could be adopted for a large class of hyperbolic PDE's, for both finite volume and discontinuous Galerkin methods. This direction will be of direct interest for higher order finite volume cores. Already the FV4 scheme bears some similarity to the MCORE dynamical core [89], but more generally one could extend this theory to be used in the unstructured finite volume K-exact reconstruction process which can be found in [93]. It is also relevant for DGFE methods, which will see further development in climate modelling with the introduction of GPU accelerated supercomputers. The limiting techniques provide theoretical guarantees on local boundedness and is likely applicable for a wide variety of schemes.
However, the methodology and limiting procedure requires a decomposition of the cell average onto flux contributing quadrature points, this can be difficult or expensive, fortunately several methods have already been proposed in [92]. Our new FV4 introduced a new kind of cell mean decomposition deduced by symmetry of the mesh, this cell mean decomposition only uses one additional point. The non-uniqueness of such a cell mean decomposition is likely of practical consequence to the accuracy of the limiter and is an open problem. Comparisons of the new limiters to the vertex-based limiters of [91, 94] should not be drawn so readily, these limiters rely on and use additional assumptions to find correction factors which enforce different maximum principles entirely. The extension of the work presented here to different neighbourhoods is also open to further study.
The $N(K) \cup K$-MP limiter, is a multidimensional limiter capable of preserving a cell mean local maximum principle on the stencil of face sharing neighbours, this is



(a) global boundedness limiter, $\lim_{s\to\infty} N^{s+1}(K) \cup N^s(K)$.



(b) $N^2(K) \cup N(K)$

Figure 3.9: We see new local extrema are generated in the global boundedness limiter trailing the slotted cylinder and within it the top of it. The $N^2(K) \cup N(K)$-MP limiter enforces a maximum principle which does smooth out these features, and also smooths out the cone peak.

new but has shown to be overly diffusive for both the fourth order method and the second order method. The $N(K) \cup K$-MP limiter reduces the order of convergence of the second order method. It could be concluded that this local maximum principle seems to be too strong when using a multidimensional limiter which does not exploit geometric properties of the mesh or components of velocity and flux contributions. The $N^2(K) \cup N(K)$-MP limiter is a multidimensional limiter capable of preserving a cell mean local maximum principle on the stencil of face sharing neighbours. For the second order method we have theoretical and numerical evidence to suggest that it is marginally preferable to that of the Barth and Jespersen limiter. Our unlimited FV4 finite volume scheme is fourth order, our new limiter $N^2(K) \cup N(K)$-MP is sufficient to satisfy a local discrete maximum principle with respect to "squared" edge sharing neighbour cell mean values.

Chapter 4

Implicit monotone time-stepping

4.1 Introduction

Explicit slope limiters can create monotone, mass preserving schemes out of structured and unstructured finite volume and finite element methods. These methods have been used in a variety of industrial and academic settings [85], and are used for atmospheric advection in dynamical cores since they are applicable to the variety of quasi-structured meshes proposed for the earth, for historical developments of such meshes see [33]. Global atmospheric models based on explicit slope limiting techniques suffer from global computational bottlenecks. In particular the time-step taken must be reduced globally to accommodate the local Courant numbers arising from large local vertical velocities. At the expense of increasing numerical diffusion, Li and Zhang [95] showed that using adaptively implicit methods in regions of locally high vertical Courant numbers improves the robustness and stability of atmospheric climate models, particularly when resolving strong vertical advection of moisture. Implicit time stepping methods are often overlooked because of the increased computational cost historically associated with them, however they are computationally competitive when the problem is stiff and with modern numerical linear algebra techniques the solve time is better than it used to be. Sometimes implicit methods are preferable entirely due the fact preconditioning can help decouple the computational cost from the Courant number, and can lead to the creation of highly scalable numerics [45],[96].

In this chapter we explore the application of different implicit linearised slope limiters to the advection equation for a variety of different incompressible flows, to see if they can be used to create an advection algorithm satisfying several discrete properties. The linearisation of one-dimensional slope limiters in mass preserving form was proposed in [47], under the acronym LCI (Linearised conservative implicit). However, despite promising numerical results for the backward Euler temporal discretisation, the LNI (Linearised non-conservative implicit) method is used instead of the LCI method because of the provable total variation diminishing properties (historically useful for convergence proofs [97, 98, 99, 25]), and the application to steady state problems.

4.1.1 Background: implicit linearised limiters

Consider the constant advection equation $u_t + au_x = 0$, with positive wind $a \geq 0$, over a periodic domain $\Omega = [0, 1]$, on a uniform grid with spacing Δx, with the backward Euler scheme in time, with time-step Δt, and the Courant number defined as $c = \frac{a\Delta t}{\Delta x}$. The resulting non-linear system can be written as

$$u_i^{n+1} - u_i^n + c(u_i^R - u_{i-1}^R)^{n+1} = 0, \quad \forall i \in \mathbb{N}, \tag{4.1}$$

$$u_i^R = u_i + \frac{1}{2}\psi(R_i)(u_i - u_{i-1}), \quad R_i = \frac{u_{i+1} - u_i}{u_i - u_{i-1}}, \tag{4.2}$$

where u_i^R denotes the value attained on the right-hand edge of cell i, and ψ is a non-linear function on the ratio of successive gradients R_i. This nonlinear implicit scheme satisfies sufficient conditions for an implicit version of Harten's lemma to apply (lemma 6.1 [100]) this can prove the total variation diminishing property for an arbitrary Courant number. This can alternatively be proven unconditionally TVD by using the SSP literature [101]. However, such schemes do not exist in practice, the above description is a nonlinear system of equations and must be approximated with a sequence of linear problems.

Yee Harten and Warming [47] propose two manners to linearise an implicit slope or flux limited method. The first linearises the scheme in flux form giving the following

$$u_i^{n+1} + c\Big[u_i^{n+1} + \frac{1}{2}\psi(R_j^n)(u_i^{n+1} - u_{i-1}^{n+1}) - u_{i-1}^{n+1} - \frac{1}{2}\psi(R_{i-1}^n)(u_{i-1}^{n+1} - u_{i-2}^{n+1})\Big] = u_i^n, \tag{4.3}$$

flux form method. This linearisation ensures local mass preservation and is denoted (LCI-Linearised Conservative Implicit). However, this method has not been proven TVD. Yee, Harten and Warning instead rearrange the scheme into a non flux form before linearising to give the following form

$$[1 + c + \frac{c}{2}\psi(R_i^n) - \frac{c}{2}\frac{\psi(R_{i-1}^n)}{R_{i-1}^n}]u_i^{n+1} - [c + \frac{c}{2}\psi(R_i^n) - \frac{c}{2}\frac{\psi(R_{i-1}^n)}{R_{i-1}^n}]u_{i-1}^{n+1} = u_i^n, \tag{4.4}$$

this (LNI-Linearised Non-Conservative Implicit) method has the M-matrix representation sufficient (not necessary) for a provable local maximum principle (TVD) under standard assumptions on the limiter function ψ. This is suitable for iteratively converging to the steady state solution of hyperbolic PDE's [47], however the loss of local mass conservation is not suitable for dynamical cores.

Despite the fact that LCI schemes are not proven to be TVD, counter examples to this claim are sparse or non-existent to our search of the literature. Perhaps due to the lack of counter examples, this mass preserving linearisation technique of various slope limiters in conservative form has become a commonly adopted technique in the monotone solution of steady state hyperbolic systems since the pioneering work in [27]. Strict monotonicity is typically relaxed, usually because the Barth and Jesperson's multidimensional slope limiter [42] is modified for differentiability [102] or

higher regularity [103, 104, 105], to improve the convergence of Newton's method. However, for numerically solving the linear advection equation, only one iteration of a non-linear method is required for convergence and additional smoothing does not seem to be necessary. This leads to the question whether the linearised conservative implicit methods are TVD, or more generally whether implicit linearised multidimensional limiters maintain a discrete local maximum principle when a mass preserving linearisation is taken. This will be the subject of this chapter, and will consist of numerical work which will be assessing, to what extent does the mass preserving linearisation and matrix solve effect the monotonicity. This will answer the broader research question: are linearised slope limiters useful for improving the robustness and stability of atmospheric climate models and for what range of Courant numbers?

4.2 Implicit Advection schemes

In this section we establish the unstructured notation, and summarise the semi discrete finite volume approach for solving the following mathematical model of transport phenomena. A tracer $u(\boldsymbol{x}, t)$, is advected by a divergence free ($\text{div}(\boldsymbol{v}) = 0$), bounded ($||\boldsymbol{v}||_\infty < C$), continuous velocity field $\boldsymbol{v}(\boldsymbol{x}, t) \in C^p(\Omega \times [0, T]), p \geq 0$.

$$\frac{\partial u}{\partial t} + \text{div}(\boldsymbol{v}u) = 0, \quad \forall (\boldsymbol{x}, t) \in \Omega \times [0, T], \tag{4.5}$$

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \Omega, \tag{4.6}$$

where the initial conditions $u_0(\boldsymbol{x})$ are assumed to be in $L^\infty(\Omega)$, $\Omega \subset \mathbb{R}^d, d = 1, 2, 3$ denotes the domain, and $t \in [0, T]$ denotes the time span of interest.

4.2.1 Semi-discrete form: Spatial discretisation

In this section we will continue establishing notation and framework with emphasis on the spatial discretisation used for arbitrary meshes.
We assume a finite volume method for the spatial discretisation, so that the rate of change of the cell mean value u_K in cell K, for every cell K in the mesh $\mathcal{M}$ evolves according to the semi discrete flux form,

$$\frac{du_K}{dt} = - \sum_{L \in N(K)} f_{K,L}, \quad \forall K \in \mathcal{M}, \tag{4.7}$$

arising from a finite volume spatial discretisation. The one-dimensional numerical flux f_{KL} through cell K into cell $L \in N(K)$ is generated by numerically integrating a monotone, conservative, Lipschitz continuous numerical flux function [72] over the cell boundary σ_{KL}. The notation $N(K) := \{L \in \mathcal{M} | \partial L \cap \partial K \neq \emptyset\}$ denotes the set of face-sharing neighbours of cell K, depicted in an unstructured arbitrary mesh diagram Fig. 4.1. Typically, the flux is resolved by numerical quadrature and

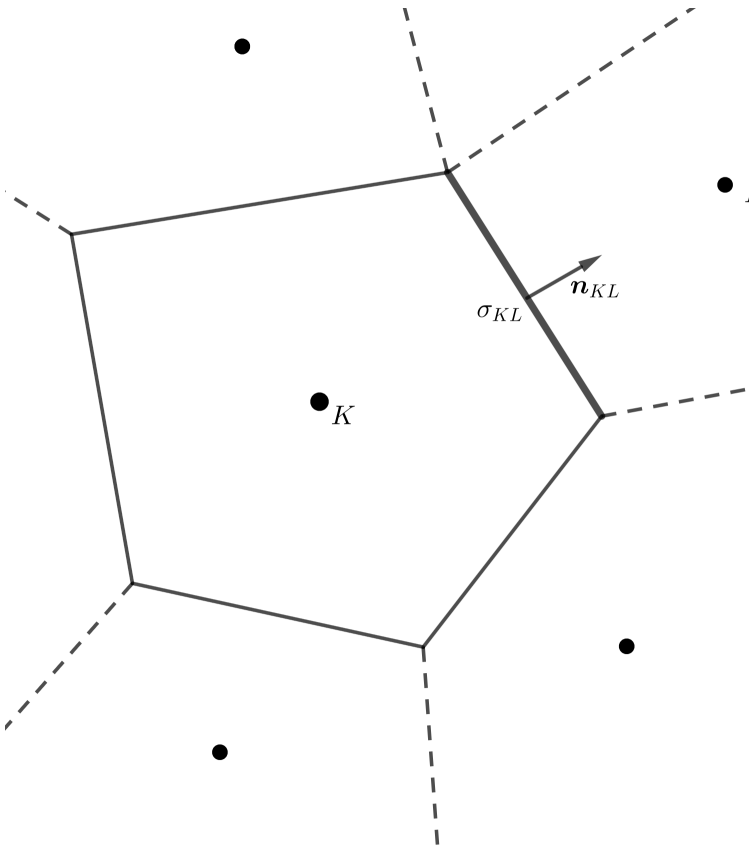


Figure 4.1: Diagram of Cell K, the face σ_{KL}, to a face sharing neighbour $L \in N(K)$, and the unit normal $\boldsymbol{n}_{KL}$.

throughout the rest of this chapter we use the midpoint rule in space (2nd order Gauss quadrature) for all flux calculations. The midpoint of edge σ_{KL} is denoted by the position vector $\boldsymbol{x}_{KL}$, for example see the position on a uniform grid Fig. 4.2. The solution of the Riemann problem for advection is tractable and gives a numerical flux function known as donor cell or upwind,

$$f_{K,L} = \gamma_{KL} p_K(\boldsymbol{x}_{KL}) - \gamma_{LK} p_L(\boldsymbol{x}_{KL}), \tag{4.8}$$

where $\gamma_{KL} := \frac{|\sigma_{KL}|}{|K|}(\boldsymbol{v}(\boldsymbol{x}_{KL}) \cdot \boldsymbol{n}_{KL})^+$, is the positive component of flux out of cell K into cell L. We define the following notation $(\cdot)^+ := \max(0, \cdot), (\cdot)^- := \min(0, \cdot)$, $|K|$ denotes the volume (Lebesgue measure) of the cell K and $|\sigma_{KL}|$ denotes the volume/area (Lebesgue measure) of the face σ_{KL} which are assumed positive. We also denote $p_K(\boldsymbol{x})$, as either the subcell representation of cell K, or a polynomial defined on a wider set of cell mean values. Typically, we will use a linear polynomial construction from the surrounding cell mean values, but other methods can be used to construct flux contributions (ENO [30], WENO [32], LSQ [106, 107], UTOPIA [19]).

Given the representation of a flux Eq. (4.8) one can construct the well-known first order upwind scheme

$$f^l_{K,L} = \gamma_{KL} u_K - \gamma_{LK} u_L, \tag{4.9}$$

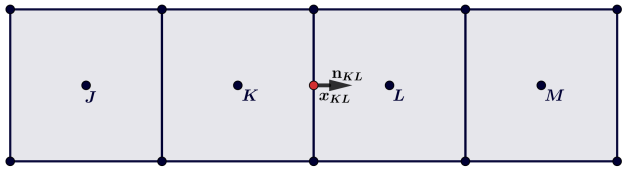


Figure 4.2: One dimensional upwind stencil. If $\boldsymbol{v}_{KL} \cdot n_{KL} \geq 0$ we use cells $[L, K, J]$ to reconstruct the flux. If $\boldsymbol{v}_{KL} \cdot n_{KL} < 0$ we use $[K, L, M]$ to reconstruct the flux. The ratio of upwind sucessive gradient are given by $R_K = \frac{u_L - u_K}{u_K - u_J}$

by letting the subcell polynomial representations p_K, p_L be represented by the constant cell mean value within each cell, and evaluated at the midpoint x_{KL}, gives $p_K(x_{KL}) = u_K$, $p_L(x_{KL}) = u_L$.
The efficient construction of high order flux contributions should depend on and exploit the particular mesh geometry and is beyond the scope of this work. On an orthogonal uniform grid Fig. 4.2 with a directional stencil J, K, L, M, well defined slope limited numerical fluxes Eq. (4.8) will often reduce to the following forms

$$f^h_{K,L} = \gamma_{K,L}[u_K + \frac{\phi(R_K)}{4}(u_L - u_J)] - \gamma_{L,K}[u_L + \frac{\phi(R_L)}{4}(u_K - u_M)], \tag{4.10}$$

$$f^h_{K,L} = \gamma_{K,L}[u_K + \frac{\alpha_K}{4}(u_L - u_J)] - \gamma_{L,K}[u_L + \frac{\alpha_L}{4}(u_K - u_M)]. \tag{4.11}$$

The first form corresponds to when a one dimensional slope limiter is used, and the second when a multidimensional slope limiter is used. The above flux representations Equation (4.10),Eq. (4.11) are the particular constructions used for the numerical results in this chapter, and are a subcase of the more general case. One can derive these particular semi discrete flux expressions Equations (4.10) and (4.11) as particular instances of using the "High-Order Accurate Solution Reconstruction" formulas in [108] or the least squares gradient in [106] and choosing linear subcell representation with 2nd order Gauss quadrature over a face, on a uniform orthogonal mesh. In the above formulas Eq. (4.10), Eq. (4.11), we have additionally introduced two different types of slope limiters denoted by α, and ϕ. α represents a multidimensional slope limiter and ϕ represents a directional slope limiter used in each coordinate direction. We use the ratio of successive gradients $R_K := \frac{u_L - u_K}{u_K - u_J}$, the non-linear flux limiter function $\psi(r)$ defined by Koren [41] (contained in the TVD region of Sweby [38]) and the relationship $\phi(R) = \frac{2\psi(R)}{1+R}$ to upwind bias slope limiters in [27] to define our one-dimensional directional slope limiter $\phi(R_K)$. (This particular choice does not have a well defined subcell representation however gains additional accuracy with little extra computation cost and makes the use of other schemes operationally unlikely)
The cell defined correction α is between zero and one and is defined by a sequence of functions that ensures that the values at all flux contributing quadrature points remain bounded by some locally determined bounds $p_K(x_q) \in [m_K, M_K]$, $\forall q \in \sigma_{KL}$, $\forall L \in N(K)$. This multidimensional slope limiter limits all the flux contributions

from a cell by a single correction factor, but does so using multidimensional bounds. When a discrete divergence free condition holds, both multidimensional slope limiting and one-dimensional slope limiting can be proven sufficient for a discrete maximum principle (albeit slightly different maximum principles as explained in Chapter 3). Barth and Jespersen [42] defined the following sequence of non-linear function evaluations to define the cell correction factor α_K.

1. Compute the old time step inclusive neighbour bounds.

$$[m_K^n, M_K^n] = \min(u_K^n, \min_{L \in N(K)} u_L^n), \max(u_K^n, \max_{L \in N(K)} u_L^n). \tag{4.12}$$

2. Compute quadrature point correction factors α_q,

$$\alpha_q = \begin{cases} \min\{1, \frac{M_K - \bar{u}_K}{p_K(x_q) - \bar{u}_K}\} & \text{where} \quad p_K(x_q) - \bar{u}_K > 0, \\ \min\{1, \frac{m_K - \bar{u}_K}{p_K(x_q) - \bar{u}_K}\} & \text{where} \quad p_K(x_q) - \bar{u}_K < 0, \\ 1 \quad \text{where} \quad p_K(x_q) - \bar{u}_K = 0. \end{cases} \tag{4.13}$$

   to ensures $\alpha_q p_K(\boldsymbol{x}_q)$ is locally bounded by m_K^n, M_K^n.

3. Limit the entire subcell representation based on the worst violator of the local quadrature maximum principle,

$$\alpha_K = \min_{L \in N(K)} \min_{q \in \sigma_{KL}} \alpha_q. \tag{4.14}$$

   So that $\alpha_K p_K(\boldsymbol{x}_q)$ is locally bounded between m_K^n, M_K^n, for all quadrature points on each face $\forall q \in \sigma_{KL}$, $\forall L \in N(K)$.

We will test numerically whether truly multidimensional slope limiters of Barth and Jespersen [42], the Kuzmin vertex limiter [94], or the one-dimensional slope limiter of [41], will retain monotonicity when combined with a mass preserving implicit linearisation technique. We do not improve the regularity of the minimum function within the Barth and Jespersen limiter sequence, using methods such as those in Venkatakrishnan, Ollivier-Gooch [102, 103] because we do not have difficulty with convergence, and these smoothing procedures can degrade the solution accuracy and even worse can degrade monotonicity properties. This completely defines the spatial aspect of the finite volume method used for the numerical section, however the linearisation and timestepping needs to be defined. Non uniform tensor product meshes have one-dimensional slope limiters potentially suitable for linearisation using the framework of [109], and both the Barth and Jespersen and the Kuzmin limiters [42, 94] are suitable for truly unstructured grids.

4.2.2 Flux Corrected Transport for implicit schemes

It will turn out that implicit linearised slope limiters are not sufficient to guarantee strictly monotonic solutions particularly when run beyond the radius of monotonicity

of the time-stepping method. We will require an additional implicit flux corrected transport procedure to fix this problem. Flux corrected transport originated from Boris and Book's [53] one-dimensional limiter. This was later generalised into multi-dimensions by Zalesak [54]. We use Zalesak's algorithm, however we will differ in the choice of both prelimiting and the choice of the local bounds Eq. (4.22) because the method is implicit. Although flux correction has been applied extensively in the finite volume literature in multiple dimensions for explicit schemes, the use of flux correction for implicit linearised scheme's had not been done before so we also present the derivation of both the high and low order fluxes later in Section 4.2.4, Section 4.2.3.

We suppose that there exists a monotone method and a high order method, with the one stage flux form representations

$$u_K^d = u_K^n - \sum_{L \in N(K)} F_{K,L}^l, \quad u_K^h = u_K^n - \sum_{L \in N(K)} F_{K,L}^h. \tag{4.15}$$

Where u_K^d denotes the monotone solution which is typically diffusive, and u_K^h is the higher order solution which is typically oscillatory. The capitalised F_{KL}^l, F_{KL}^h denote one stage "transportative flux" [54] contributions through the boundary of cells K and L over one time step. These transportative fluxes are different from the "semi discrete" fluxes f_{KL}^l, f_{KL}^h defined earlier in Eq. (4.7), as they account for time-stepping. The existence of such representations is always possible when using a method of lines technique with a flux form finite volume method. We introduce examples and the construction of such fluxes using Runge Kutta methods in Section 4.2.1. In this work we will use the backward Euler first order upwind scheme to create a transportative flux $F_{K,L}^l$, given explicitly by formula Eq. (4.26). We will test a variety of high order transportative flux contributions to construct $F_{K,L}^h$, constructed from the implicit midpoint in time and various linearised slope limiters in space Eqs. (4.11) and (4.31). This assumption/representation of a numerical method is not always possible, schemes such as the semi-Lagrangian method typically do not possess a flux form representation.

It is yet to be determined if the high order method using linearised slope limiters will be monotone, it is for this reason that we introduce the Flux corrected transport procedure of [54] as a safeguard. Flux corrected transport blends the high order and low order schemes in the following way,

$$u_K^{n+1} = u_K^n - \sum_{L \in N(K)} F_{KL}^{low} - \sum_{L \in N(K)} C_{KL}(F_{KL}^{high} - F_{KL}^{low}). \tag{4.16}$$

The anti-diffusive flux corrections on each face are defined as $AF_{KL} := F_{KL}^{High} - F_{LK}^{Low}$. If $C_{KL} = 0$ we recover the low order monotonic method, and if $C_{KL} = 1$ we recover the high order non monotonic method Eq. (4.15). Flux corrected transport can be used to determine face defined corrections C_{KL} to the anti diffusive fluxes $AF_{KL} :=$

$F_{KL}^{High} - F_{LK}^{Low}$ such that the one step numerical scheme Eq. (4.16) remains bounded by two locally defined constants $u_K^{n+1} \in [m_K, M_K]$. The correction factors C_{KL} are determined by the following algorithm.

1. Calculate the sum of antidiffusive flux into and out of cell K

$$P^p = \sum_{L \in N(K)} AF_{KL}^+, \quad P^m = \sum_{L \in N(K)} AF_{KL}^-, \tag{4.17}$$

   where the notation $AF_{KL}^+ := \max(0, AF_{KL})$ and $AF_{KL}^- := \min(0, AF_{KL})$ denotes the positive and negative component of the anti diffusive flux.

2. Compute the maximal and minimal allowable change to the monotone solution

$$\Gamma_K^M = M_K - u_K^d, \quad \Gamma_K^m = m_K - u_K^d, \tag{4.18}$$

   where m_K and M_K are the locally defined minima and maxima respectively.

3. Determine the outflow and inflow corrections

$$RP_K := \min(1, \frac{\Gamma_K^M}{P^p}) \quad \text{for} \quad P^p > 0, \tag{4.19}$$

$$RM_K := \min(1, \frac{\Gamma_K^m}{P^m}) \quad \text{for} \quad P^m > 0. \tag{4.20}$$

4. Determine the face defined correction factor

$$C_{KL} := \begin{cases} \min(RP_L, RM_K) & AF_{KL} > 0, \\ \min(RM_L, RP_K) & AF_{KL} < 0. \end{cases} \tag{4.21}$$

This procedure can be used for both implicit and explicit methods, to enforce different bounds. We impose a local boundedness principle with respect to the monotone solution, so use the coefficients

$$M_K^d := \max\{u_K^d, \max_{L \in N(K)} u_L^d\} \quad m_K^d := \min\{u_K^d, \min_{L \in N(K)} u_L^d\}, \tag{4.22}$$

to ensure that the solution satisfies a local maximum principle with respect to a known monotone solution $u_K^{n+1} \in [m_K^d, M_K^d]$. This is different from the bounds used by Zalesak [54] because the old time step neighbours cease to provide physically motived bounds at large Courant numbers. The local maximum principle employed is strictly stronger that the discrete maximum principle

$$u_K^{n+1} \in [\min\{u_K^n, \min_{L \in N(K)} u_L^n\}, \max\{u_K^n, \max_{L \in N(K)} u_L^n\}]$$

typically imposed of explicit flux correction algorithms, and is more appropriate for high Courant number flow.

Although the solution satisfies a local maximum principle with respect to both the old timestep u_K^n and the monotone solution u_K^d, it does not prevent the enhancement of directional extrema including saddle points [110]. It was for this reason Devore [110] recommended using the one-dimensional Boris and Book limiter

$$AF_{KL} = S_{KL}\big(\min\{|AF_{KL}|, S_{KL}(u_K^d - u_J^d), S_{KL}(u_M^d - u_L^d)\}\big)^+. \tag{4.23}$$

where $S_{KL} = \text{sign}(u_L^d - u_K^d)$ [53], [55] to prelimit the anti diffusive flux contributions before the flux correction procedure. We will numerically investigate the use of this pre limiter Method 4.2.10 in row 4 of figure Fig. 4.5, to determine whether this will improve the quality of the solution in the context of implicit flux correction transport algorithms.

4.2.3 Low order transportative fluxes

Flux corrected transport requires the computation of lower order fluxes, which produce the monotone (typically diffusive) solution u^d. This is done for backward Euler as follows:

1. Construct and solve the linear matrix system associated with backward Euler in time, with semi discrete numerical fluxes given by first order upwind.

$$u_K^{n+1,d} = u_K^n - \Delta t \sum_{L \in N(K)} f_{K,L}^{n+1,d}, \tag{4.24}$$

$$f_{K,L}^{n+1,d} = \gamma_{K,L}^{n+1} u_K^{n+1,d} - \gamma_{L,K}^{n+1} u_L^{n+1,d}, \tag{4.25}$$

   for the diffusive solution $u_K^{n+1,d}$.

2. Reconstruct the low order "transportative" fluxes using the newly computed $u_K^{n+1,d}$,

$$F_{K,L}^{n+1,d} = \Delta t(\gamma_{K,L}^{n+1} u_K^{n+1,d} - \gamma_{L,K}^{n+1} u_L^{n+1,d}). \tag{4.26}$$

This method and resulting fluxes will be denoted $BE1$ later in this text.

4.2.4 High order transportative fluxes

Flux corrected transport requires the computation of higher order fluxes. We derive the high order flux derivation in algorithmic form for the 1D slope limiter Eq. (4.10), for the implicit midpoint rule in time using its compositional (extended)Shu Osher representation.

1. Construct and solve the linearised slope limiter system

$$u_K^{n+1/2} = u_K^n - \frac{\Delta t}{2} \sum_{L \in N(K)} f_{K,L}^{n+1/2,n,h}, \tag{4.27}$$

for $u_K^{n+1/2}$ associated with a backward Euler step of half a time-step. Where the numerical flux above is linearised at the old time step in mass conserving form as follows

$$f_{K,L}^{n+1/2,n,h} = \gamma_{K,L}^{n+1/2}[u_K^{h,n+1/2} + 1/4\phi(R_K^n)(u_L^{h,n+1/2} - u_J^{h,n+1/2})], \tag{4.28}$$
$$- \gamma_{L,K}^{n+1/2}[u_L^{h,n+1/2} + 1/4\phi(R_K^n)(u_K^{h,n+1/2} - u_M^{h,n+1/2})]. \tag{4.29}$$

and put into the matrix left hand side.

2. Using the newly computed substage $u_K^{n+1/2}$ as well as R_K^n and $\gamma_{K,L}^{n+1/2}$ we explicitly rebuild the previously defined fluxes $f_{K,L}^{n+1/2,n,h}$.

3. We now take the new solution, $u_K^{n+1/2}$ and compute the forward Euler fluxes,

$$f_{K,L}^{n+1/2,h} = \gamma_{K,L}^{n+1/2}[u_K^h + 1/4\phi(R_K^{n+1/2})(u_L^h - u_J^h)] - \gamma_{L,K}^{n+1/2}[u_L^h + 1/4\phi(R_K^{n+1/2})(u_K^h - u_M^h)]. \tag{4.30}$$

4. We reconstruct the the one stage transportative second order numerical flux,

$$F_{KL}^{high} = \frac{\Delta t}{2}(f_{K,L}^{n+1/2,n,h} + f_{K,L}^{n+1/2,h}). \tag{4.31}$$

The above choice of linearisation aligns with the (extended) Shu Osher representation, and was chosen such that any monotonicity violations, must come from either the linearisation or solver tolerances of the backward Euler scheme or the forwards part being run at too high a Courant number. In fact, one can linearise the implicit midpoint rule directly and reduce the number of fluxes needed to be calculated, and use the scheme described in the appendix Appendix A.3 for improved computational efficiency. For the multidimensional slope limiter version of the above method choosing a replacing the one-dimensional limiters with a multidimensional limiter $\phi \mapsto \alpha$, characterises the change to the algorithm.

4.2.5 Proposed Schemes

We need to test the extent the linearisation of one dimensional limiters and multidimensional slope limiters cause monotonicity violations at a range of Courant numbers. To do so we introduce the following methods: IMkoren, IM2BJ, IM2KUZ, these schemes are associated with mass preserving linearisations of the one-dimensional Koren [41] limiter, the multidimensional Kuzmin [94] limiter, and the multidimensional Barth and Jesperson [42] limiter. We also introduce the unlimited implicit midpoint third order upwind method $IM3$ and the backward Euler first order upwind unlimited $BE1$ method. We completely characterise these methods in terms of fluxes as this will be convenient for when we use flux correction.

Method 4.2.1 (*BE1*). *Backwards Euler first order upwind. Fluxes can be attained from Eq.* (4.26).

Method 4.2.2 (***IMKoren***). *Implicit midpoint rule is used for temporal discretisation and the spatial discretisation aligns with using one-dimensional slope limiters in each direction. By using the slope limiter relation* $\phi(r) = 2\psi(r)/(1+r)$ *[27] we use the more accurate one-dimensional Koren limiter [41]* $\psi(r) = \min\{1/3 + 2/3r, 2r, 2\}^+$. *The linearisation procedure aligns with the mass preserving method of [47] and the high order transportative flux Eq. (4.31) is computed using the method of Section 4.2.4.*

Method 4.2.3 (***IM2BJ***). *Implicit Midpoint rule in time with the linearised Barth and Jespersen [42] limiter, the transform* $\phi(R_K) \mapsto \alpha_K$ *and Section 4.2.4 completely defines the high order flux computation Eq. (4.31).*

Method 4.2.4 (***IM2KUZ***). *Implicit Midpoint rule in time with the linearised vertex based slope limiter of Kuzmin [108], the transform* $\phi(R_K) \mapsto \alpha_K$ *and Section 4.2.4 completely defines the high order flux computation Eq. (4.31).* [1]

Method 4.2.5 (***IM3***). *We will use the third order linear upwind in space by choosing a linear slope limiter* $\phi(r) \mapsto \frac{2}{1+r}(1/3 + 2/3r)$, *this method has interpretation aligning with creating upwind second order polynomials, over the interpolating cells* LKJ, KLM *respectively, and the resulting flux is* $f^h_{K,L} = \gamma_{K,L}[2u_L + 5u_K - u_J]/6 - \gamma_{L,K}[2u_K + 5u_L - 1u_M]/6$. *which is third order accurate when the flow is constant, but is truly second order. Section 4.2.4 completely defines the high order flux computation Eq. (4.31). No linearisation is required because the scheme is linear and the flux can be computed cheaply using Eq. (A.45).*

We also introduce the four methods IMKoren∘fct∘BE1, IM2BJ ∘fct∘ BE1, IM2KUZ ∘ fct ∘ BE1, IM3 ∘ fct ∘ BE1 which include a flux corrected transport step.

Method 4.2.6 (***IMKoren*** $\circ$ ***fct*** $\circ$ ***BE1***). *We will use the same high order fluxes as from IMkoren but use the flux correction transport procedure Eq. (4.22) to correct on the low order fluxes given by backwards Euler first order upwind Eq. (4.26).*

Method 4.2.7 (***IM2BJ*** $\circ$ ***fct*** $\circ$ ***BE1***). *We will use the same high order fluxes as from IM2BJ but use the flux correction transport procedure Eq. (4.22) to correct on the low order fluxes given by backwards Euler first order upwind Eq. (4.26).*

Method 4.2.8 (***IM2KUZ*** $\circ$ ***fct*** $\circ$ ***BE1***). *We will use the same high order fluxes as from IM2KUZ but use the flux correction transport procedure Eq. (4.22) to correct on the low order fluxes given by backwards Euler first order upwind Eq. (4.26).*

Method 4.2.9 (***IM3*** $\circ$ ***fct*** $\circ$ ***BE1***). *We will use the same high order fluxes as from IM3 but use the flux correction transport procedure Eq. (4.22) to correct on the low order fluxes given by backwards Euler first order upwind Eq. (4.26).*

[1] This limiter only differs from the Barth and Jespersen algorithm [42] by the choice of the locally defined maxima and minima, and the points used to create the correction factor. We simply evaluate the subcell representation for the corner defined quadrature points and ensure they are bounded by the corner sharing neighbours. See Chapter 3 or [108, 91] for a more detailed explanation.

We also introduce the following methods $IM3\circ BBP+fct\circ BE1$, $IM3\circ UP+fct\circ BE1$ associated with doing additional one-dimensional pre-limiting on the $IM3\circ fct\circ BE1$ scheme.

Method 4.2.10 ($\boldsymbol{IM3\circ BBP+fct\circ BE1}$). *We take the Method 4.2.9 method, and introduce the one-dimensional pre limiter of Boris and Book [53] as recommended by [110] to reverse the direction of some flux contributions in accordance with a directional FCT procedure.*

Method 4.2.11 ($\boldsymbol{IM3\circ UP+fct\circ BE1}$). *We take the Method 4.2.9 method, and introduce the unstructured pre limiter of [66]. Which sets anti-diffusive flux contributions to zero rather than reversing them.*

4.3 Numerical Results

4.3.1 Test Suite

The numerical domain is $\Omega = [0,1]\times[0,1]$, with nx, ny grid points in the x and y dimensions. The mesh consists of (nx, ny) cell centers, and $nx+1, ny+1$ faces, with periodic boundary conditions. The velocities and fluxes are located at the face midpoints $(i+1/2, j), (i, j+1/2)$. We locate the stream function $\psi(x,y)$ at the cell vertices and then use a discrete form of the curl operator to create a divergence free (machine precision) vector field $\boldsymbol{v}$. For the stream functions $\psi(\boldsymbol{x},t)$ we use the following four streamfunctions defining the velocity fields,

$$\psi = -\pi((x-x_c)^2+(y-y_c)^2), \qquad \text{solid body rotation} \tag{4.32}$$

$$\psi = y-x, \qquad \text{constant diagonal} \tag{4.33}$$

$$\psi = 2\sin(x)\sin(y)\cos(\frac{t}{T}\pi), \qquad \text{sine deformation} \tag{4.34}$$

$$\psi = 8\pi x(x-1)y(y-1)\cos(\frac{t}{T}\pi). \qquad \text{quadratic deformation} \tag{4.35}$$

For the initial condition of the tracer, we use a compact C^1 cosine bump Eq. (4.36),

$$q = \frac{1}{2}(1+\cos(\pi\min(\frac{r}{0.15},1))), \quad \text{where} \quad r=\sqrt{(x-0.5)^2+(y-0.75)^2}), \qquad \text{cosine bump} \tag{4.36}$$

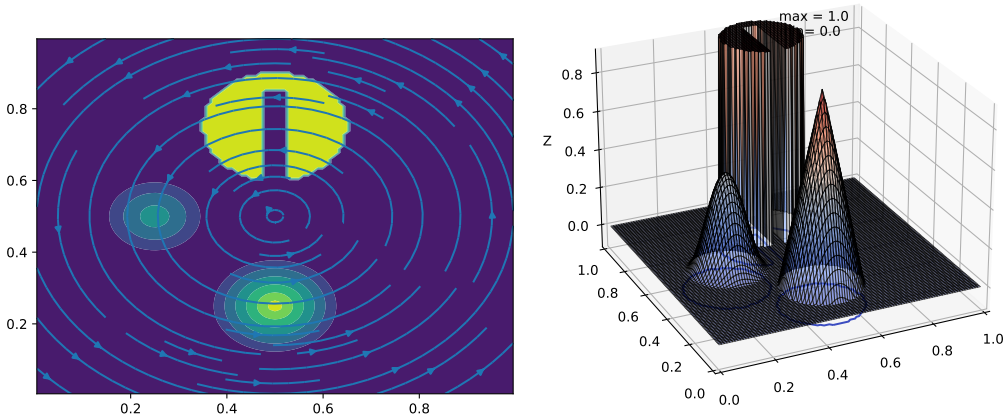


Figure 4.3: Flow visual and direct sampling of the initial condition.

or the LeVeque initial condition Eq. (4.37) [81], Fig. 4.3

$$u_0 = \begin{cases} 1 & \text{for} \sqrt{((x-0.5)^2+(y-0.75)^2)} \leq 0.15, \text{and}(x \leq 0.475), \\ 1 & \text{for} \sqrt{((x-0.5)^2+(y-0.75)^2)} \leq 0.15, \text{and}(x > 0.525), \\ 1 & \text{for} \sqrt{((x-0.5)^2+(y-0.75)^2)} \leq 0.15, \text{and}(y \geq 0.85), \text{and} \\ & (0.475 < x \leq 0.525), \\ (1 - \frac{r_{cone}}{0.15}) & \text{for} \quad (r_{cone} = \sqrt{((x-0.5)^2+(y-0.25)^2)} \leq 0.15), \\ \frac{1}{2}(1+\cos(\pi \frac{r_{cos}}{0.15}) & \text{for} \quad (r_{cos} = \sqrt{((x-0.25)^2+(y-0.5)^2)} \leq 0.15), \\ 0 & \text{otherwise.} \end{cases} \tag{4.37}$$

4.3.2 First test case: Solid body rotation of the LeVeque initial conditions

Solid body rotation test case Eq. (4.32), of the LeVeque initial conditions Eq. (4.37) [81], is considered a challenging test of monotonicity and shape preservation ([54],[81]), we have included a visualisation in Fig. 4.3. The initial conditions consist of a Zalesak slotted cylinder, a compact smooth cosine bell of half height, and a cone with a sharp peak. We sample initial conditions, rather than integrate over the control volumes, this is to keep the initial conditions as discontinuous as possible. We use the coarser Zalesak [54] 100×100 resolution, and choose the number of time-steps as $nt = 1256, 157, 19$, to give maximum Courant numbers of approximately 0.5, 4 and 33.

Results: Solid body rotation with implicit linearised slope limiters

Fig. 4.4 shows the results of the solid body rotation of the LeVeque initial conditions at 100×100 resolution for the BE1, IMKoren, IM2BJ, IM3 and the IM2KUZ methods at maximum Courant numbers 1/2, 4, and 33.

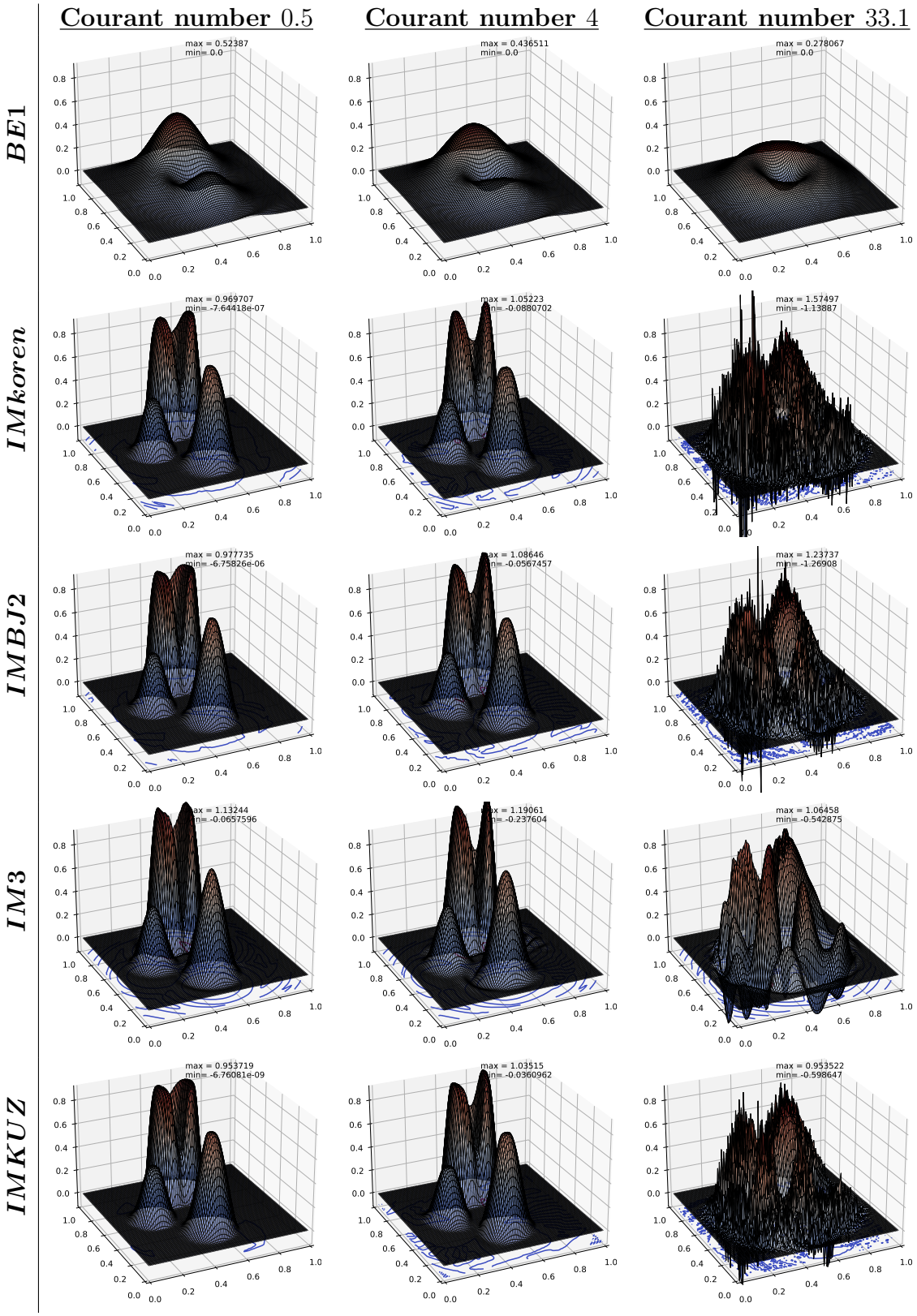


Figure 4.4: LeVeque Initial conditions Eq. (4.37), solid body rotation Eq. (4.32) test case at three Courant numbers for the first five methods described in Section 4.2.5 without flux corrected transport.

In the first column of Fig. 4.4 the linearised slope limiter schemes $IMKoren$, $IM2BJ$, and $IM2KUZ$ are run below Courant number 0.5, we see some small negative values of order 10^{-6}, 10^{-5}, 10^{-8} respectively. This essentially forms a counter example to the statement that implicit schemes with linearised conservative slope limiters can preserve a discrete maximum principle. The implicit linearised schemes $IMKoren$, $IM2BJ$, $IM2KUZ$ clearly possess shape preserving properties as compared with the unlimited scheme $IM3$, which has negatives of order 10^{-1} for the same test.
In the second column of Fig. 4.4 at Courant number 4, all of the schemes aside from the first order method (BE1) produce monotonicity violations. When comparing the linearised limiter schemes IMkoren, IMBJ2 and IMKUZ to the unlimited IM3 scheme, in the second column of Fig. 4.4 one can see the suppression of ripples behind the real features, and surprisingly the linearised limiters still do go some way towards shape preservation. However, the negative values produced for the linearised slope limiters are significantly larger and are of order $10^{-1}, 10^{-2}$. These negative values can be attributed to the effect of time-stepping beyond the radius of monotonicity of the implicit midpoint method and the previous non-linear solving strategies. In [111], implicit linearised WENO methods were investigated numerically, it was also deemed that implicit schemes of order 2 were not non-linearly stable at Courant numbers twice the explicit stability criterion. Our numerical results agree with this conclusion, and provides further evidence that the non-linear order barrier for implicit strong stability [49, 48] has numerical consequences almost immediately.
In the third column of Fig. 4.4 at Courant number 33, the slope limiters have not suppressed monotonicity violations and have subjectively created unfavourable unphysical noisy features as compared with the unlimited IM3 method. We have not observed numerical blow-up or crashing for our particular test cases, however this cannot be ruled out as a possibility using linear stability theory.

4.3.3 Problems, Outlooks, Explanations

We are aware of three main reasons why implicit slope limiters may lead to non-monotonic solutions:

1. The first cause of potential monotonicity failure arises from the linear matrix solving strategy. Where numerical linear algebra techniques iteratively approximates for the solution within a given solver tolerance, the interim solutions generated by these inner iterations are not guaranteed to be monotone.

2. The second major cause of potential monotonicity failure arises from the non-linear solving strategy. Fixed point methods such as Newton and Picard iteratively converge to the non-linear solution using outer iterations until a solver tolerance is reached. The interim solutions generated by the outer iterations are not guaranteed to be monotone.

3. The third major cause of monotonicity failure arises from the time stepping method being run beyond its radius of monotonicity.

The linear matrix solving strategy was not suspected to be the major source of monotonicity error for our numerical experiments, because the observed 10^{-6} monotonicity violations at low Courant numbers (first column of Fig. 4.4) were not found to be dependent on a wide range of sensible solver tolerances. However, from a theoretical perspective Item 1 is not to be ruled out as a potential cause of monotonicity error, as injudicious choices for the solver and solver tolerances can create additional monotonicity failures. Instead, we hypothesise that the small 10^{-6} monotonicity failures in the first column of Fig. 4.4 are caused by the first step of Picard iteration (mass preserving linearisation) by process of elimination and because we have found that increasing the number of Picard iterations, did minutely change the location of the monotonicity errors. We hypothesise that the significantly larger 10^{-2} monotonicity violations at Courant number 4 and 33 in both column 2 and 3 of Fig. 4.4 are a result of running the implicit midpoint scheme beyond its theoretical time step restriction (Item 3).

Potential solutions to Items 1 and 2 could include using, L^{∞} or more stringent stopping criterion. However, this will involve potentially never converging since both Picard and Newton iterations typically converge in L^2, as do most numerical linear algebra techniques. Customised solvers with monotone convergence properties, would involve the creation of new solvers likely slower than current ones. The method of deferred correction [112] has been previously proposed as an alternative method to perform an implicit linearised solve in [37], by ensuring that the only matrices inverted are M-matrices and limiting is on explicit terms. However, this method does not intrinsically have good monotonicity properties (one can analyse it using the ARK extension of the strong stability preserving literature and compare to a traditional IMEX method). Furthermore, the right hand side corrections for this problem are also not strong stability preserving under any time-step so would require an explicit flux corrected transport procedure to be used on the explicit corrections within the iterative process [67]. It is often claimed that this method is cheap because all the left-hand side matrices are diagonally dominant M-matrices so can be solved fast, whilst this is true, high order spatial operators for the advection equation may also be solved similarly fast when appropriate solving strategies are employed. We will instead use a one-step implicit flux corrected transport procedure as a means of solving these small 10^{-6} monotonicity errors, and hypothesise that the FCT algorithm will not activate much, preserving the order and accuracy.

Concerns Items 1 and 2 are a consequence of implicit solving strategies having convergence in L^2 towards the monotone solution. Whilst this is a problem, there is still some control as to the size of the monotonicity violation. Concern Item 3, is more serious still, the temporal discretisation itself poses a challenge to monotonicity at high Courant numbers, and we have suggested this is the cause of the much larger (10^{-2} and above) monotonicity failures in columns two and three of Fig. 4.4

Potential solutions to concern Item 3 include,

1. Using a different high order implicit Runge Kutta method, with greater radius of monotonicity,
2. Use the theta scheme based on the globally highest Courant number,
3. Use a locally defined theta scheme, or a multi-rate algorithm,
4. Implicit Flux correction.

Solution 1 is not possible due to a fundamental nonlinear order barrier theorem. Spijker considers unconditional monotonicity preservation in the non-linear case in [49] under the notion of contractivity. Spijker notes that the upper bound for the order of unconditional non-linear stable methods is 1, which follows from the earlier linear positivity theory done by Bolley and Crouziex [48], in which the last of Kraaijevanger conditions [113] is required of a rational function and shown incompatible with second order. This analysis does not rule out the possibility of a second order implicit scheme with a very large radius of monotonicity. However, for a large class of implicit methods with order greater than or equal to 2, the largest effective radius of monotonicity has been verified and searched for both numerically and theoretically, using the connection between the optimal (extended) Shu Osher representation and the radius of monotonicity [51], [52]. The largest stage scaled radius of monotonicity is disenchantingly small, it is 2, and is attained by the implicit midpoint method. One can attain a better absolute radius of monotonicity, by increasing the number of stages but this linearly increases the computational cost (near equivalent to running at a smaller time step). Therefore, if we want true monotonicity at large Courant number we must drop the formal order of the method at large Courant number.
One could use a global theta scheme based on the globally largest Courant number. This reduces the order of the model globally, and will be deemed unviable because there is inevitably a region of high flow in a transport model, so will suffer a bottle neck affecting global accuracy. The local theta scheme [114] and multi-rate methods [115] are potentially a viable strategy, however these are not method of lines schemes and fall out of the scope of this discussion. We instead decide to fix both of these concerns by applying the flux corrected transport algorithm Eq. (4.22) on backward Euler scheme's fluxes.

4.3.4 Results: Solid body rotation with additional flux corrected transport

In Fig. 4.5 we plot the final time-step of the solid body rotation test case of the LeVeque initial conditions. We do so for the six flux corrected transport defined in Section 4.2.5. We make the following observations. The flux corrected procedure has generated bounded results, that are accurate at low Courant number (column one Fig. 4.5) and non-linearly stable at high Courant numbers (column two and three

Fig. 4.5), for all choices of space discretisation Fig. 4.5. At low Courant numbers, flux corrected transport smooths the peaks and degrades the thin back wall of the slotted cylinder. At high Courant numbers the scheme remains monotone but is very diffusive.

Out of all the rows in Fig. 4.5, flux correcting the unlimited scheme IM3∘fct∘BE1 seems to produce the most accurate results, the IMKoren∘fct∘BE1, IM2BJ∘fct∘BE1 and IM2KUZ∘fct∘BE1 schemes all produce similar acceptable results with additional limiting behaviour. There seems to be little benefit in using the linearised slope limiter methods for the incompressible advection equation, if in addition the implicit flux correction is used.

The largest values attained at the final timestep in Fig. 4.5 are achieved by the IM3∘BBP+fct∘BE1 scheme, followed by the IM3∘UP+fct∘BE1, followed by the $IM3 \circ fct \circ BE$ scheme. This increased one directional compressive behaviour seems to arise out of the 1D flux prelimiting and is observed in row five of Fig. 4.5, the scheme remains bounded, but the accuracy of the shape is questionable particularly at Courant number 4 when using the Boris and Book prelimiting FCT algorithm IM3∘fct+BBP∘BE1. The less ambitious unstructured pre-limiter introduced in [116] has produced less compressive 1D effects as can be seen in row six of Fig. 4.5, but at this stage it is still not obvious whether any prelimiting is worth using.

4.3.5 Second test case: accuracy, boundedness and errors.

The total tracer values after 0, 1/6, 2/6, 3/6, 4/6, 5/6 and one solid body rotation are contoured in Fig. 4.6 for the four schemes IMkoren∘FCT∘BE1, IM2BJ∘FCT∘BE1, IM3∘FCT∘BE1, IM3∘BBP+FCT∘BE1. In this same figure Fig. 4.6 we also colour errors whose absolute value exceeds 0.02 and we also display the maxima and minima of the tracer after 0, 1/6, 2/6, 3/6, 4/6, 5/6 and one full rotation. We use 128×128 spatial resolution and 1536 time-steps, so the Courant number remains approximately below 1/2, and we use the C^1 compact cosine initial condition Eq. (4.36) under solid body rotation Eq. (4.32).

The results from all 4 of the FCT methods remain bounded up to machine precision, the most accurate method is the IM3∘fct∘BE1 method whose primary error occurs at the extrema being degraded. Both the linearised 1d and multidimensional limiter methods IMKoren∘fct∘BE1 and IM2BJ∘fct∘BE1 retain sufficient accuracy but have additional dispersive errors in the direction of flow compared with IM3∘fct∘BE1.

The IM3∘BBP+fct∘BE1 method differs from the IM3∘fct∘BE1 method by using the one-dimensional pre-limiter of Boris and Book. It is known [110] that this pre-limiter step reverses the direction of unphysical anti-diffusive flux contributions in a monotone manner to prevent the growth of directional extrema (saddle points). We have observed some quite unfavourable "squaring" effects, apparent in the colour plots of numerical error in Fig. 4.6.

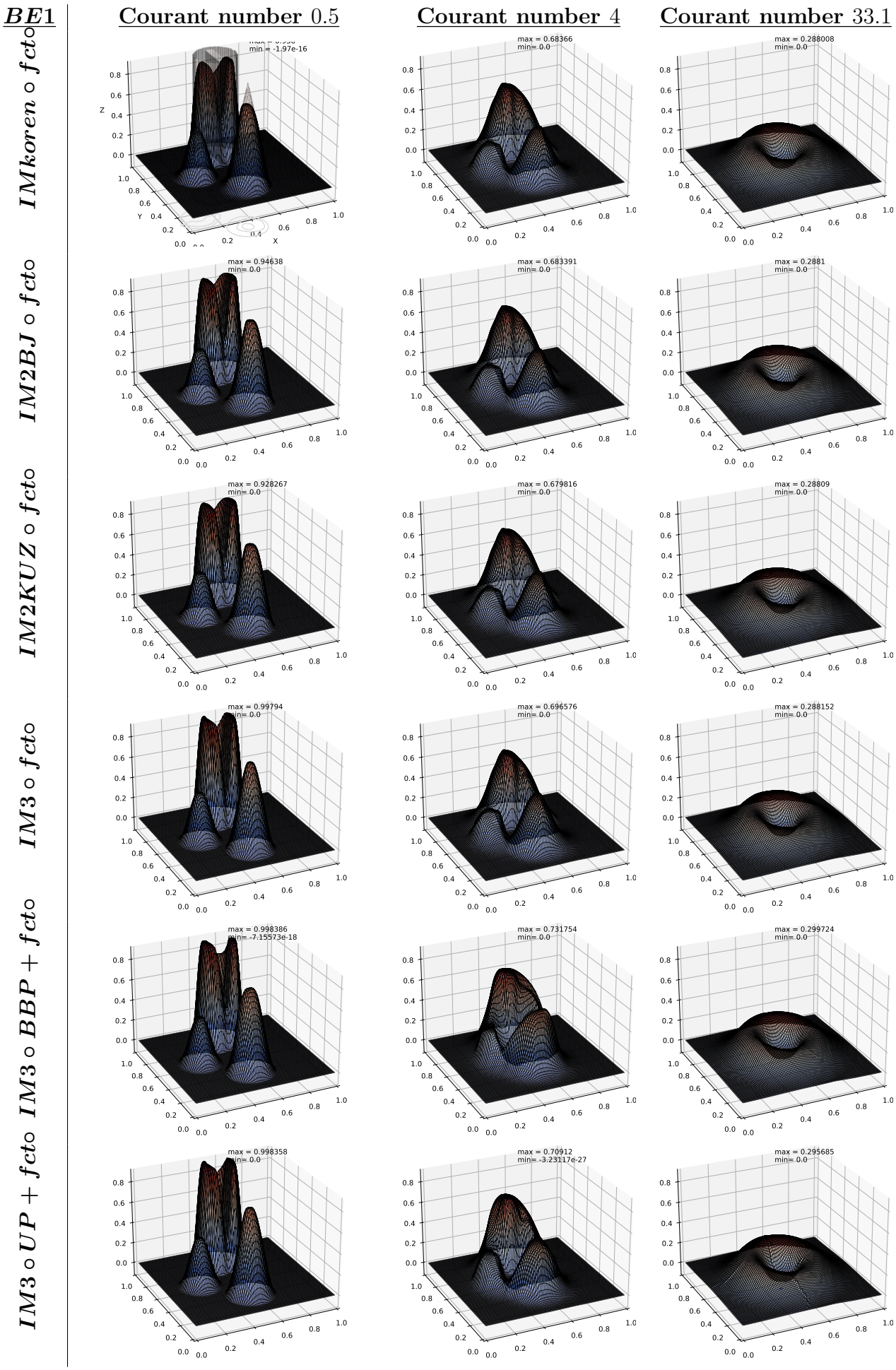


Figure 4.5: Final timestep of the LeVeque Initial conditions Eq. (4.37), after the solid body rotation test case Eq. (4.32) at Courant number maximum of: 0.5, 4, 33, for the six flux corrected transport methods defined in Section 4.2.5.

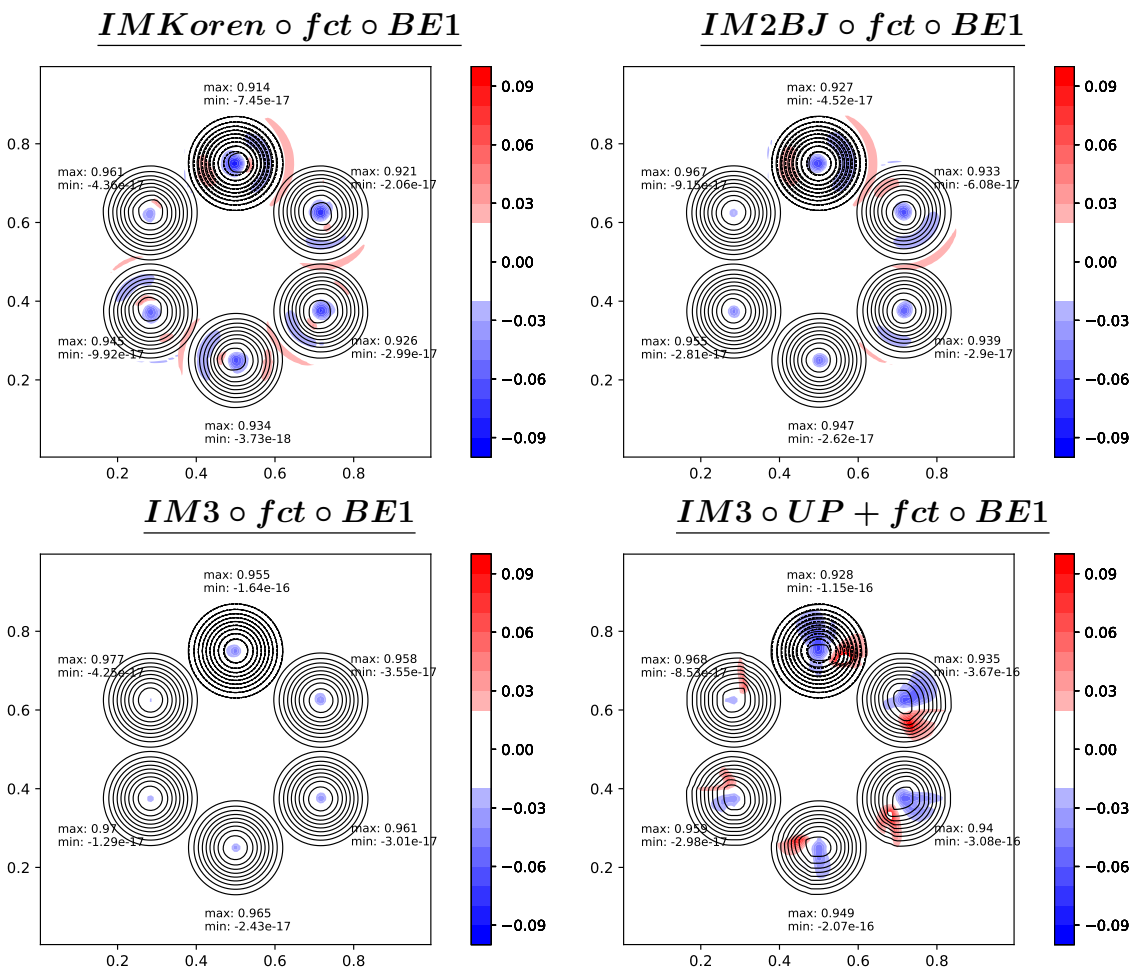


Figure 4.6: We test the accuracy, boundedness and errors of the flux corrected transport schemes $IMkoren \circ FCT \circ BE1$, $IMBJ2 \circ FCT \circ BE1$, $IM3 \circ FCT \circ BE1$, $IM3 \circ BBP + FCT \circ BE1$. We colour errors exceeding absolute value of 0.02, and plot contours of tracer value the interval between contours is 0.1, this is plotted at every 1/6th of a revolution. The use of implicit flux correction without any linearised slope limiting is the most accurate, followed by the multidimensional limiter of Barth and Jespersen, then followed by the one-dimensional limiting, however all these three results are acceptable. The last figure indicates the prelimiting step has an unusual perhaps negative consequence on the distribution of error.

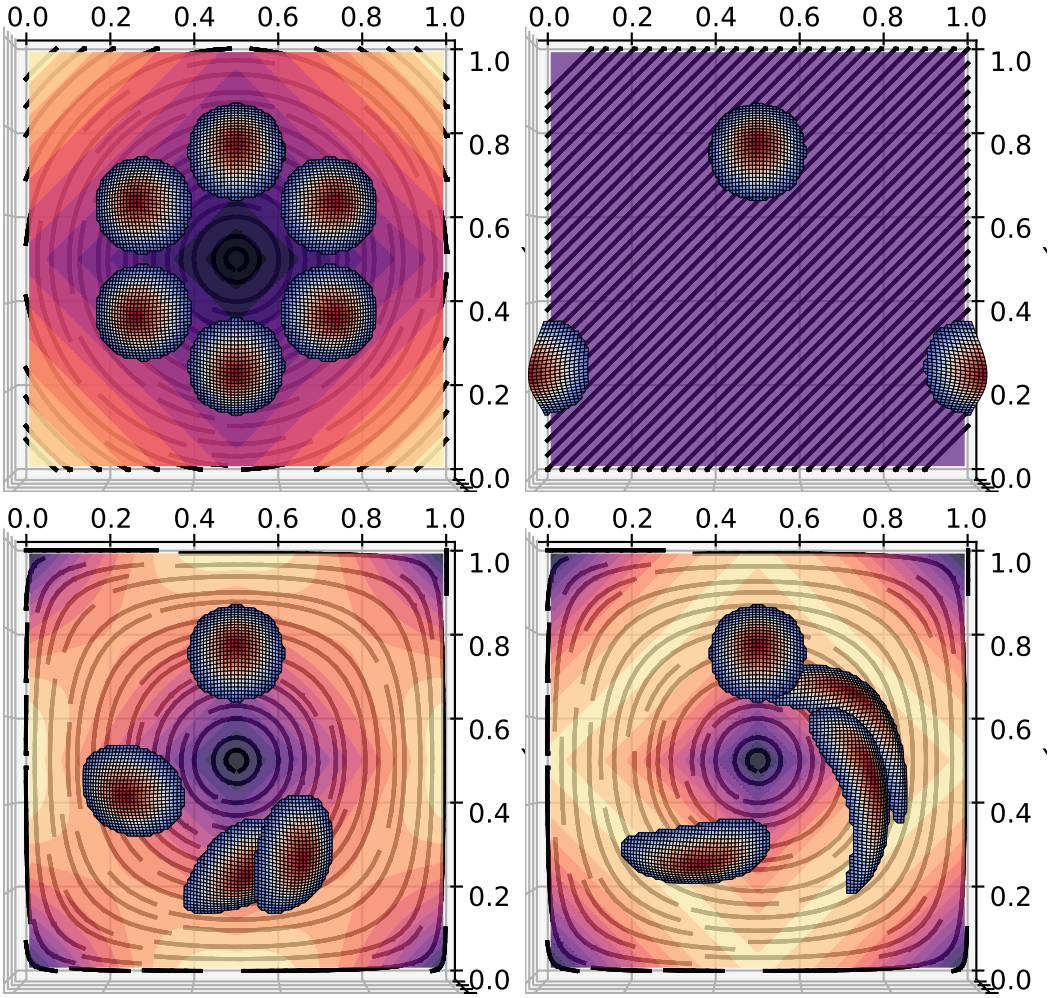


Figure 4.7: Eq. (4.36), under the Eq. (4.32), Eq. (4.33), Eq. (4.35), Eq. (4.34) deformations. For the solid body rotation Eq. (4.32) we plot the tracer value at every 1/6th of the counter-clockwise rotation. The constant diagonal flow Eq. (4.33), with periodic boundary conditions plotted are the initial condition and the half timestep. For the quadratic and sinusoidal time reversing deformational flows, we only show the tracer values who have corner defined neighbours all above 0.1. We show the solution being deformed to its maximum deformation counter-clockwise, before being reversed clockwise to the initial condition. Also plotted are the streamlines and Courant number contours. The solution of the tracer was created with the IM3 scheme at 128×128 resolution.

4.3.6 Third test case: Convergence

For the convergence testing we use 16×16, 32×32, 64×64 and 128×128 resolution at maximum Courant number near $\frac{1}{2}$, with the C^1 cosine initial condition Eq. (4.36) under the four velocity fields Eq. (4.33), Eq. (4.35), Eq. (4.34), Eq. (4.32). The Courant number is held near $\frac{1}{2}$, by using $nt = 64, 128, 256, 512$ for the diagonal constant flow Eq. (4.33), and by using $nt = 192, 384, 768, 1536$ for the quadratic time reversing deformational test case Eq. (4.35), the sinusoidal time reversing deformational test case Eq. (4.34), and the solid body test case Eq. (4.32). These tests are defined in Section 4.3.1 and visualised in Fig. 4.7. The plots of relative L^2 error at Courant number 0.5 against resolution are plotted in Fig. 4.8, this figure consists of four sub-figures, each of which correspond to one of the four different velocity fields defined in Section 4.2.5. Each sub-figure of Fig. 4.8 contains the convergence of 10 different schemes, the order of the scheme is calculated and displayed in the figure legend using the errors at resolution 64 and 128.

4.3.7 Results: Convergence tests at Courant number 0.5

In Fig. 4.8 we observe nearly all of the second order schemes introduced have achieved near the theoretical second order accuracy for all test cases. The simplest unconditionally monotone scheme IM3∘fct∘ BE1 stands out as the most accurate of all the limited methods with highest convergence. The IM∘ BBP+fct∘BE1 method employing the one-dimensional Boris and Book pre-limiter has performed consistently worse than the other higher order methods, and in the directionally constant flow test case has observed a drop in convergence order. Interestingly the addition of the flux corrected transport step did not significantly decrease the accuracy of the implicit linearised schemes, and in some cases the use of FCT actually increased the accuracy and convergence order of the method.

4.3.8 Results: Convergence tests at Courant number 2

In Fig. 4.9 we plot the relative L^2 error of the same schemes against resolution at a maximum Courant number approximately 2. The Courant number is held near 2, by using $nt = 16, 32, 64, 128$ for the diagonal constant flow Eq. (4.33), and by using $nt = 48, 96, 192, 384$ for the quadratic time reversing deformational test case Eq. (4.35), the sinusoidal time reversing deformational test case Eq. (4.34), and the solid body test case Eq. (4.32). We observe that all the flux corrected schemes drop to first order accuracy, but still have smaller relative error than $BE1$. This is intended and in line with the theoretical non-linear order barrier [48, 49]. The drop in order was not observed for the solid body rotation case, because the flow did not reach the high Courant number regime. The implicit linearised limiters such as IMKoren, IM2BJ and IM2KUZ attain second order convergence beyond the radius of monotonicity and aligns with the fact that these methods will be monotonicity violating at these Courant numbers.

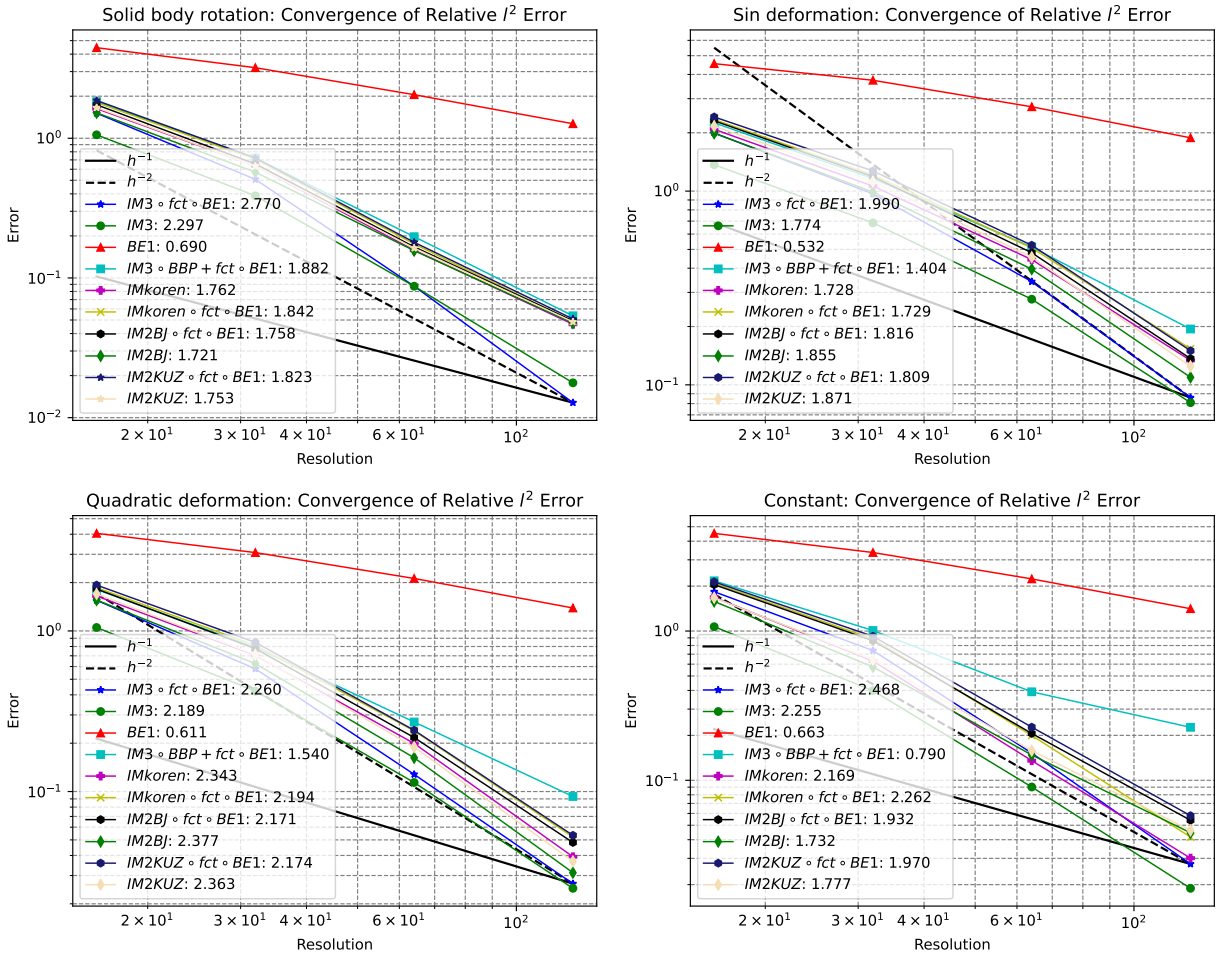


Figure 4.8: Convergence of 10 different schemes introduced in Section 4.2.5 under 4 different test cases at Courant number 0.5. All schemes (aside from $BE1$)achieve near the theoretical order of accuracy for all test cases, with the notable exception of the $IM3 \circ BBP + fct \circ BE1$ scheme in the constant diagonal flow test case.

4.3.9 Discussion on solvers

Numerical results were generated using one outer iteration of Picard/Newton, and the numerical scheme is linearised in mass conserving form. The linear solver used was ILU preconditioned GMRES to solver tolerance 10^{-10}. The first step of Picard and Newton methods are identical, and it is also often the most important stage, in terms of amount of error corrected, more iterations do not mitigate SSP concerns or drastically improve accuracy. We have observed convergence with and without preconditioning, we have observed convergence from both multigrid solvers(Ruge Stuben) and multigrid pre-conditioners(Ruge Stuben, smoothed aggregation) from the PyAMG software [117]. The vast computational time was spent preconditioning, leading to the promise of multi tracer efficiency. We have plotted the solver time against Courant number for the variety of matrices used, and it scales better than linear when preconditioning is used. We have plotted the solver time for higher order interpolating polynomials and observed better than linear scaling for solver time against spatial order, for a variety of flows and Courant numbers. The average number of inner iterations for GMRES for CFL 0.5, 4 are two and three, we have

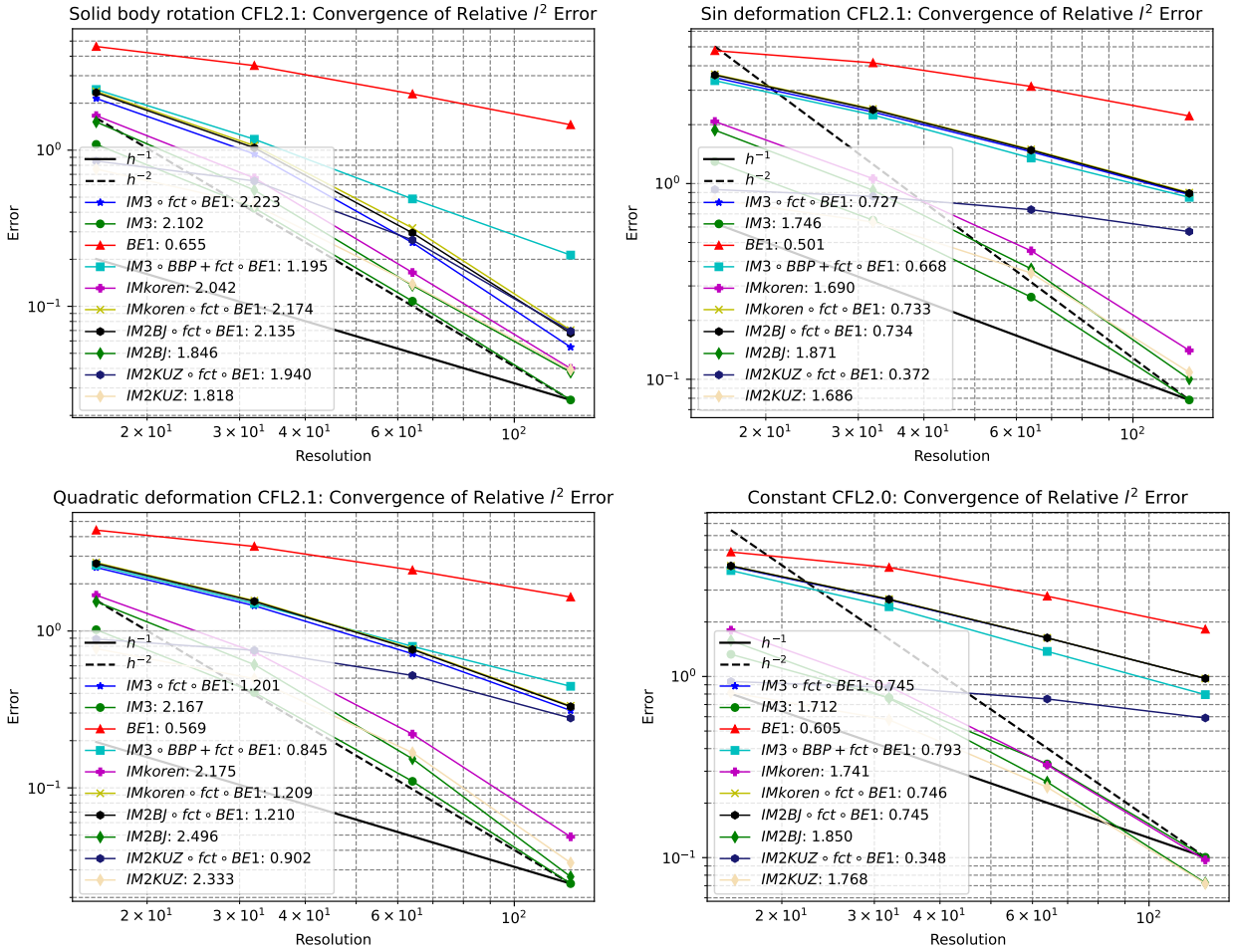


Figure 4.9: Convergence of 10 different schemes under 4 different test cases at CFL 2, beyond the theoretical radius of monotonicity FCT methods drop the methods to first order. (Solid body rotation being a exception because the initial condition remains within a certain CFL range.) Second order is still attainable for monotonicity violating schemes.

used one outer iteration.

4.4 Conclusions

Following the result of Higueras [118] the monotonicity/SSP property of the Forward Euler method under any timestep restriction, implies that the backward Euler method is monotonic for all time steps. Despite this, numerically this is a contentious assumption for discrete higher order spatial operators when different Krylov/multigrid solvers, and nonlinear solving strategies are employed. We have shown that the mass preserving implicit linearisation of 1d slope limiters [47], and mass preserving linearisations of both the multidimensional slope limiters of Kuzmin [94] and Barth and Jespersen [42], have introduced small boundedness violations, typically of order 10^{-6} even when the implicit midpoint timestepping scheme is run below its radius of monotonicity. These observations are new, slightly disappointing, but not all that surprising. The implicit flux correction algorithm proposed can fix these violations by forcing a discrete maximum principle by acting on rebuilt fluxes,

this has little impact on the error or convergence order at low Courant numbers Fig. 4.8.

We observed the that implicit linearised slope limiters will not recover monotonicity when the timestepping scheme used is run beyond the radius of monotonicity. The one stage implicit flux correction method proposed in this chapter can fix these monotonicity violations and embeds a provable discrete maximum principle. It does so for arbitrary large Courant number but unavoidably lowers the convergence order of the scheme in accordance with the non-linear order barrier theorem of [48, 49].

Devore [110] recommended the one-dimensional Boris and Book limiter [53] to be used as a directional flux pre-limiter to create a truly monotone numerical method. This method treats saddle points as directional extrema and subsequently limits them [110]. In our tests we observe heavy clipping phenomena in each direction resulting in the Boris and Book pre-limiter reducing the order of the method by at least 0.5 in the convergence tests we performed. Extreme squaring behaviour occurred along the constant diagonal flow test case and strange errors appeared in the solid body rotation test. This is surprising as many consider prelimiting as an essential stage in flux corrected transport and advocacy for pre-limiting is common [110]. Interestingly, Chaplin and Colella [62] also suggest that pre-limiting did not give good results for their flux corrected transport advection algorithm. We have tried the more conservative unstructured mesh prelimiting [66] which does not produce such extreme clipping, but it is not clear whether this is worth using.

The one stage flux correction of implicit midpoint third order upwind fluxes on backward Euler first order upwind fluxes using the FCT transport algorithm (4.22), can achieve second order convergence on a variety of deformational time reversing flows, provided the CFL number is small enough. The solution remains discrete maximum principle satisfying in the presence of large CFL numbers and discontinuous initial conditions but lowers the order of the method at large Courant number. The solution remains bounded, consistent, discrete maximum principle satisfying and mass preserving to machine precision due to the flux form representation.

This is an easy scheme to use and attains all required properties of the next generational dynamical core. Furthermore, both the high order and low order matrices are linear and independent of the tracer so the same preconditioner could be reused for the hundreds of tracers used in dynamical cores, this will likely lead to very high multi tracer efficiency.

Chapter 5

Conclusions

In this thesis we have contributed to the development of discrete maximum principle satisfying advection algorithms.

We have studied the multidimensional advection equation under incompressible flow when one-dimensional slope limiters are used in each direction on a structured mesh, and found two new limiter regions suitable for this scheme to maintain a local maximum principle. This numerical method has been posed as an Eulerian scheme for transport on the sphere in [40], where it has previously been indicated [41, 40, 37] but not proven that Spekreijse's monotonicity theory [27] allows one-dimensional limiters to be used in more than one dimension. We have indicated that Spekreijse's monotonicity theory does not directly apply because a face defined velocity may not result in a mean value theorem applying in each direction. We have numerically demonstrated the creation of significant 10^{-2} negative values for the van-Albada and Ospre limiters. We have shown that by using the divergence free condition we can still prove a local maximum principle, but the admissible limiter region is different. We have shown that the Sweby region is a sufficient, quasi-necessary assumption for symmetric limiters to maintain the local maximum principle, and have numerically tested that pushing limiters into the region of Sweby has little to no effect on: convergence order, accuracy, or peak preservation (in spite of some theoretical predictions) whilst retaining symmetry and in some cases can increases the permissible time-step.

We have also shown that by breaking the symmetry condition of the limiter, we can define limiters outside the Sweby diagram in two new derived flux limiter regions and still have a maximum principle for incompressible flow. This extended limiter region has been used to create limiters more compressive (SuperbeeR(R, m, M)) and more accurate (Woodfield(R, m, M)) than existing limiter functions whilst being suitable for incompressible flow. We have also used the symmetry breaking property to expose two frameworks of limiting, the $\theta = 0$ framework has been used to create the first globally differentiable limiter function (Differentiable(r)) entirely contained within a second order accurate region that will preserve a local maximum principle for incompressible flow.

We combined the one-dimensional limiters with the optimal three stage third order

strong stability preserving Runge Kutta method, and demonstrated second order accuracy for a variety of deformational and non-deformational flows when the limiter is appropriately chosen. Similar to Hundsdorfer et al. [40] we come to the same conclusion that the Koren limiter [41] is a robust accurate limiter function that would be suitable for advection on the sphere. However, contrary to [40] our experiments conclude the strong stability literature is of practical value, and contrary to [40] prove that the state interpolated scheme is linear invariant when the flow is incompressible. We conclude that this numerical method is locally conservative, consistent, positive definite, (substage) discrete maximum principle satisfying, second order, and accurate. However, it is not apparent how this is adaptable for more arbitrary meshes such as some proposed for dynamical cores without additional adaption. One future extension of this work would be to study whether one-dimensional limiters could be modified to work on an orthogonal grid of non constant resolution; this has been approached in [109] for symmetric slope limiters. Our numerical results indicate that the spatial error dominate for the tests we have chosen, this indicates investigating the SSP22 Runge Kutta scheme is worth investigated from a computational cost perspective and also in light of the hidden maximum principle perspective A.2. The new one-dimensional limiters introduced have mostly been for demonstration purposes, further accuracy and efficiency are likely to be gained. For example, the Woodfield$(R, M, 0)$ limiter has a trade-off between accuracy and permissible time-step. The determination of a computationally efficient M would be of practical importance. The differentiable(r) limiter could be made more computationally efficient to evaluate by using piecewise polynomial functions. The framework introduced has a free parameter θ, we have only studied θ equalling zero or one, the possibility of extending the theory for θ not equalling zero or one is another possible direction of research.

Zhang et al. [43, 44, 92] introduced a framework in which slope limiters can be used to impose global boundedness, and has been shown effective on arbitrarily high order methods for a variety of meshes. We have slightly extended the framework of Zhang et al. to instead impose different local boundedness principles, and use this to design two new local maximum principle limiters. The $N(K) \cup K$-MP limiter imposes a local face sharing maximum principle on cell mean values, and the $N^2(K) \cup N(K)$-MP limiter imposes a local "squared" face sharing maximum principle on cell means. We have introduced a classic second order finite volume scheme on a structured grid and have shown that the $N(K) \cup K$-MP limiter is not sufficiently accurate. We have shown that the $N^2(K) \cup N(K)$-MP limiter imposes the same local maximum principle on cell mean values to the Barth and Jespersen [42] but has less severe correction factors due to different edge defined maximum principles. The application of the $N^2(K) \cup N(K)$-MP limiter has been shown to be more accurate than the Barth and Jespersen limiter (and the Kuzmin/Park limiter on square meshes), however the improvement is only marginal. The strength of the $N^2(K) \cup N(K)$-MP limiter is that it can be used on truly higher order methods with rigorous local boundedness. To

demonstrate this fact, we have introduced a new fourth order finite volume method, and explained how the decomposition of the cell average can lead to application of both the $N(K)\cup K$-MP limiter and the $N^2(K)\cup N(K)$-MP limiter. The $N(K)\cup K$-MP limiter was not accurate enough, but the $N^2(K)\cup N(K)$-MP limiter produced accurate locally bounded results. We have compared the $N^2(K)\cup N(K)$-MP limiter to a global boundedness limiter $N^{s+1}\cup N^s(K)$ for very large s, and seen that the local limiter does suppress unphysical oscillations, but does so at the expense of accuracy and peak preservation.

This new local boundedness slope limiting framework could be applicable for a wide variety of the meshes proposed in atmospheric advection and could be used on increasingly higher order methods. This method could be adaptable to different equations by modifying the local Riemann problems, studying the effectiveness of this local boundedness approach would be interesting for problems with shocks. For high order schemes on unstructured meshes the problem of efficient slope limiting for a local boundedness principle is likely dependent on the method and mesh used. The new theory introduced reformulates this problem into finding efficient decompositions of the cell average onto flux contributing quadrature points and limiting these points based on the desired local maximum principle. It is still unsolved as to what size and shape the desired local maximum principle should be.

The resulting higher order schemes have inherent local mass conservation, consistency and constancy preservation. We have derived sufficient conditions on the limiter function and the time-step for positivity preservation and a local maximum principle. The method described is adaptable to a wider class of numerical methods on more arbitrary meshes such as those proposed for dynamical core advection, but the limiter depends on the exact method used and a decomposition of the cell mean. We have shown that the $N^2(K)\cup N(K)$-mp limiter is accurate at low Courant number.

We tested the mass preserving linearisation of Yee Warming and Harten [47], for traditional one-dimensional slope limiters as well as both the multidimensional limiters of Barth and Jespersen and Kuzmin [42, 94] when using the implicit midpoint rule in time. At low Courant numbers all these were monotone to the eye and shape preserving, however small negative values $\approx 10^{-6}$ were obtained in all cases, and attributed to the mass preserving linearisation technique. At small Courant numbers (less than 0.5) the implicit flux corrected transport procedure proposed fixed all the small $\approx 10^{-6}$ monotonicity violations and attained second order accuracy and local boundedness. We saw large monotonicity failures when the high order implicit scheme was run at Courant numbers beyond the radius of monotonicity. The implicit flux corrected transport procedure fixed the larger monotonicity violations, but degraded the solution accuracy and convergence order at these larger Courant numbers.

We found that the one-dimensional flux prelimiting procedures [110] often employed in flux corrected transport had a serious negative consequence on the accuracy,

even at small Courant number. We also found that using implicit flux correction to correct the implicit midpoint scheme with linear third order upwind in space on the backward Euler first order upwind scheme. Was more accurate than flux correcting the implicit linearised slope limiters whilst being simpler and cheaper. The FCT algorithm also imposes the original maximum principle in light of its one stage methodology Appendix A.2, and uses the local velocity in the determination of the limiting allowing improved accuracy.

The IM3∘fct∘BE1 one step scheme could be recommended as a numerical algorithm that achieves all the requirements of a dynamical core advection algorithm. It is unconditional stable, is inherently mass preserving, consistent, positivity preserving, unconditionally local maximum principle satisfying, adaptable to a variety of meshes proposed for dynamical cores, and can achieve second order accuracy at low Courant number. It reduces to first order accuracy at large Courant numbers and near extrema by design. The FCT algorithm in general showed remarkable flexibility, and could be used to achieve unconditional stability and monotonicity for different higher order models.

Appendix A

Appendix

A.1 On the order of tvd schemes

Wesseling and Zijlema [37] despite proposing the scheme Eq. (2.12) as a finite volume method, also use the more convenient finite difference truncation analysis to show no conflict with second order accuracy at extrema. However, their theorem 3.1 and proof of theorem 3.1 has some technical inaccuracies, and it is recommended that the method of analysis in [36] be followed instead. Hua-mo does truncation error analysis for the $\theta = 1$ case [36], we present some of those arguments here for the $\theta = 0$ case.

The semi-discrete finite difference local truncation error τ of the spatial derivative u_x at position x_i is defined as

$$\begin{aligned}\tau_i := \frac{u(x_i) - u(x_{i-1})}{\Delta x} + \frac{\psi(R_u(x_{i+1}, x_i, x_{i-1}))(u(x_i) - u(x_{i-1}))}{2\Delta x} \\ - \frac{\psi(R_u(x_i, x_{i-1}, x_{i-2}))(u(x_{i-1}) - u(x_{i-2}))}{2\Delta x} - u'(x_i)\end{aligned} \tag{A.1}$$

where $R_u(a, b, c) := \frac{u(a)-u(b)}{u(b)-u(c)}$ and the prime notation $u'(x_i)$ denotes the spatial derivative u_x at x_i. As in [36] we wish to write this local truncation error at x_i as a function of an arbitrary point arbitrary point $x_\xi = x_i + \xi\Delta x$, to determine the order of accuracy within a distance of a critical point.

By first Taylor expanding $u(x_{i-1})$ about x_i to remove $u_x(x_i)$, then re-expanding again about the arbitrary point $x_\xi = x_i + \xi\Delta x$ as in [36]. We can deduce the first order upwind scheme has the following expression of local truncation error

$$\frac{u(x_i) - u(x_{i-1})}{\Delta x} - u'(x_i) = -\Delta x \frac{u''(x_i)}{2!} + \Delta x^2 \frac{u'''(x_i)}{3!} ... \tag{A.2}$$

$$= -\frac{\Delta x}{2!}[u''(x_\xi) - \xi\Delta x u'''(x_\xi) + \frac{\xi^2\Delta x^2 u''''(x_\xi)}{2!}] \tag{A.3}$$

$$+ \frac{\Delta x^2}{3!}[u'''(x_\xi) - \xi\Delta x u''''(x_\xi)...]... \tag{A.4}$$

$$= -\frac{\Delta x}{2!}u''(x_\xi) + (3\xi + 1)\frac{\Delta x^2 u'''(x_\xi)}{3!} + O(\Delta x^3). \tag{A.5}$$

Now for the higher order flux limited correction in Eq. (A.1) we use the following Taylor expansions about x_ξ

$$u_{i+1} - u_i = \Delta x u'(x_\xi) - (\xi - 1/2)\Delta x^2 u''(x_\xi) - \frac{(\xi-1)^3 - \xi^3}{3!}\Delta x^3 u'''(x_\xi) - ... \tag{A.6}$$

$$u_i - u_{i-1} = \Delta x u'(x_\xi) - (\xi + 1/2)\Delta x^2 u''(x_\xi) - \frac{(\xi)^3 - (\xi+1)^3}{3!}\Delta x^3 u'''(x_\xi) - ... \tag{A.7}$$

$$u_{i-1} - u_{i-2} = \Delta x u'(x_\xi) - (\xi + 3/2)\Delta x^2 u''(x_\xi) - \frac{(\xi+1)^3 - (\xi+2)^3}{3!}\Delta x^3 u'''(x_\xi) - ... \tag{A.8}$$

where we have introduced the notation $u'(x_\xi)$ to be a short hand notation for $u_x(x_\xi)$. This allows us to write down an expression for the local truncation error in terms of x_ξ

$$\begin{aligned}\tau_i := &-\frac{\Delta x}{2!}u''(x_\xi) + (3\xi+1)\frac{\Delta x^2 u'''(x_\xi)}{3!} \\ &+ \frac{1}{2}\psi(R_i)[u'(x_\xi) - (\xi+1/2)\Delta x u''(x_\xi) - \frac{(\xi-1)^3 - \xi^3}{3!}\Delta x^2 u'''(x_\xi)] \\ &- \frac{1}{2}\psi(R_{i-1})[u'(x_\xi) - (\xi+3/2)\Delta x u''(x_\xi) - \frac{(\xi+1)^3 - (\xi+2)^3}{3!}\Delta x^2 u'''(x_\xi)] + O(\Delta x^3).\end{aligned} \tag{A.9}$$

We collect terms of the appropriate order to define the truncation error as

$$\begin{aligned}\tau_i := &\frac{1}{2}[\psi(R_i) - \psi(R_{i-1})]u'(x_\xi) \\ &+ [-1 - \psi(R_i)(\xi+1/2) + \psi(R_{i-1})(\xi+3/2)]\frac{\Delta x}{2!}u''(x_\xi) \\ &+ \left[(3\xi+1) - \frac{1}{2}\psi(R_i)[(\xi-1)^3 - \xi^3] + \frac{1}{2}\psi(R_{i-1})[(\xi+1)^3 - (\xi+2)^3]\right]\frac{\Delta x^2 u'''(x_\xi)}{3!},\end{aligned} \tag{A.10}$$

here the gradient R is also Taylor expanded about the arbitrary point x_ξ,

$$R_i = \frac{u'(x_\xi) - (\xi-1/2)\Delta x u''(x_\xi) - \frac{(\xi-1)^3-\xi^3}{3!}\Delta x^2 u'''(x_\xi) - ...}{u'(x_\xi) - (\xi+1/2)\Delta x u''(x_\xi) - \frac{(\xi)^3-(\xi+1)^3}{3!}\Delta x^2 u'''(x_\xi) - ...} \tag{A.11}$$

$$R_{i-1} = \frac{u'(x_\xi) - (\xi+1/2)\Delta x u''(x_\xi) - \frac{(\xi)^3-(\xi+1)^3}{3!}\Delta x^2 u'''(x_\xi) - ...}{u'(x_\xi) - (\xi+3/2)\Delta x u''(x_\xi) - \frac{(\xi+1)^3-(\xi+2)^3}{3!}\Delta x^2 u'''(x_\xi) - ...}. \tag{A.12}$$

This allows us to state the requirements of the scheme to be first second and third order.

Definition A.1.1 (First order requirement near x_ξ)**.** First order accuracy requires

$$[\frac{\psi(R_i)}{2} - \frac{\psi(R_{i-1})}{2}]u_x(x_\xi) = O(\Delta x). \tag{A.13}$$

Definition A.1.2 (Second order requirement near x_ξ)**.** Second order accuracy requires

$$\begin{aligned}&\big[\frac{\psi(R_i)}{2} - \frac{\psi(R_{i-1})}{2}\big]u_x(x_\xi)\\&+\big[-1-(\xi+\frac{1}{2})\psi(R_i)+(\xi+\frac{3}{2})\psi(R_{i-1})\big]\frac{\Delta x u_{xx}(x_\xi)}{2!} = O(\Delta x^2).\end{aligned} \tag{A.14}$$

Definition A.1.3 (Third order order requirement near x_ξ)**.** Third order accuracy requires

$$\begin{aligned}&\big[\frac{\psi(R_i)}{2} - \frac{\psi(R_{i-1})}{2}\big]u_x(x_\xi)\\&\big[-1-(\xi+\frac{1}{2})\psi(R_i)+(\xi+\frac{3}{2})\psi(R_{i-1})\big]\frac{\Delta x u_{xx}(x_\xi)}{2!}\\&\big[(3\xi+1)-\frac{\psi(R_i)}{2}[(\xi)^3-(\xi+1)^3]+[(\xi+1)^3-(\xi+2)^3]\frac{\psi(R_{i-1})}{2}\big]\frac{\Delta x^2(u_{xxx})(x_\xi)}{3!} = O(\Delta x^3).\end{aligned} \tag{A.15}$$

There are now three cases to consider. We want to know the error at a critical point x_i(maxima/minima), at a non critical point x_i, and the error at x_i within the vicinity of a critical point x_ξ.

Case A.1.1. *Case 1: We are at a critical point in position i, so that*

$$\xi = 0 \tag{A.16}$$

$$u_x(x_\xi) = u_x(x_i) = 0 \tag{A.17}$$

$$R_i = -1 - \frac{2\Delta x}{3}\frac{u_{xxx}(x_i)}{u_{xx}(x_i)} + O(\Delta x^2) \tag{A.18}$$

$$R_{i-1} = \frac{1}{3} + \frac{2\Delta x}{3}\frac{u_{xxx}(x_i)}{u_{xx}(x_i)} + O(\Delta x^2) \tag{A.19}$$

where $u_{xx}(x_i) \neq 0$.

Case A.1.2. *Case 2: We are at a non critical point i, so that*

$$\xi = 0 \tag{A.20}$$

$$u_x(x_\xi) = u_x(x_i) \neq 0 \tag{A.21}$$

$$R_i = 1 + \Delta x u''(x_i)/u'(x_i) + O(\Delta x^3) \tag{A.22}$$

$$R_{i-1} = 1 + \Delta x u''(x_i)/u'(x_i) + \Delta x^2(4/3 u_{xxx}x_i - 3u_{xx}(x_i)^2) + O(\Delta x^3) \tag{A.23}$$

Case A.1.3. *Case 3: We are at x_i in the vicinity of a critical point at x_ξ*

$$\xi \neq 0 \tag{A.24}$$

$$u_x(x_\xi) = 0 \tag{A.25}$$

$$R_i = \frac{(\xi - 1/2)}{(\xi + 1/2)} + O(\Delta x) \tag{A.26}$$

$$R_{i-1} = \frac{(\xi + 1/2)}{(\xi + 3/2)} + O(\Delta x) \tag{A.27}$$

We satisfy first order accuracy requirements in case 1 and case 3 trivially because $u_x(x_\xi) = 0$. In case 2 at a non-critical x_i the first order requirement is

$$\psi(1 + \Delta x u''(x_i)/u'(x_i) + O(\Delta x^3)) - \psi(1 + \Delta x u''(x_i)/u'(x_i) + O(\Delta x^2)) = O(\Delta x). \tag{A.28}$$

A sufficient and necessary condition for this to be satisfied is Lipshitz continuity of the limiter function.

When is the scheme 2nd order? We first note that all inflection points will be second order because $u_x(x_i), u_{xx}(x_i) = 0$.

In case 1 at critical, non inflection points we have second order accuracy when

$$-1 - 1/2\psi(-1 - a\Delta x + O(\Delta x^2)) + 3/2\psi(1/3 + a\Delta x + O(x^2)) = O(\Delta x), \tag{A.29}$$

where $a = \frac{2}{3}\frac{u_{xxx}(x_i)}{u_{xx}(x_i)}$. If one first assumes

$$1 - 3/2\psi(1/3) + 1/2\psi(-1) = 0 \tag{A.30}$$

and then adds this to Eq. (A.29). One can deduce that Lipschitz continuity and $1-3/2\psi(1/3)+1/2\psi(-1) = 0$ are sufficient and necessary for second order, without the requirement of differentiability.

In case 2 at a non critical, non inflection point we have second order accuracy when both the conditions

$$\psi(1 + a\Delta x + O(\Delta x^3)) - \psi(1 + a\Delta x + O(\Delta x^2) = O(\Delta x^2) \tag{A.31}$$

$$-1 - 1/2\psi(1 + a\Delta x + O(\Delta x^3)) + 3/2\psi(1 + a\Delta x + b\Delta x^2) = O(\Delta x) \tag{A.32}$$

are satisfied for $a = u''(x_i)/u'(x_i)$. A sufficient and necessary condition for the first condition to hold is Lipschitz continuity. If this assumption is used on the second condition we can rewrite the second condition as

$$-1 + \psi(1 + a\Delta x + O(\Delta x^2) = O(\Delta x) \tag{A.33}$$

Which is satisfied for differentiable ψ satisfying $\psi(1) = 1$. If one assumes $\psi(1) = 1$ and instead adds $1 - \psi(1) = 0$ to the above expression Eq. (A.33) one can see that Lipschitz continuity and $\psi(1) = 1$ are sufficient and necessary, without the requirement of differentiability. We now consider second order accuracy in case 3, within a neighbourhood of a critical point which requires

$$-\left[1 - (\xi + \frac{1}{2})\psi(R_i) - (\xi + \frac{3}{2})\psi(R_{i-1})\right]\frac{u_{xx}(x_\xi)}{2!} = O(\Delta x). \tag{A.34}$$

Leading to the condition

$$\left[1+(\frac{1}{2}+\xi)\psi(\frac{\xi-1/2}{\xi+1/2}+O(\Delta x))-(\xi+\frac{3}{2})\psi(\frac{\xi+1/2}{\xi+3/2}+O(\Delta x))\right]=O(\Delta x), \tag{A.35}$$

assuming differentiability of the limiter then a sufficient condition for the above to hold is the following condition

$$1+(\xi+\frac{1}{2})\psi(\frac{\xi-1/2}{\xi+1/2})-(\xi+\frac{3}{2})\psi(\frac{\xi+1/2}{\xi+3/2})=0 \tag{A.36}$$

If the above property Eq. (A.36) is assumed and it is subtracted from Eq. (A.35), the differentiability is relaxed to the Lipschitz continuity and proven both necessary and sufficient.

The second order in the neighbourhood of an extrema condition Eq. (A.36) is satisfied for any second order linear scheme $\psi(R)=aR+b$ when $1-a-b=0$. But the only solution that satisfies $\psi(0)=0$ is given by $\psi(R)=R$ which is not within monotonicity constraints for large gradients [36]. This Eq. (A.36) condition reduces to the second order $\psi(1)=1$ non-extrema condition in the limit $\xi\to\pm\infty$ sufficiently away from extrema. The Eq. (A.36) condition reduces to the previously derived second order at extrema condition $2=3\psi(1/3)-\psi(-1)$ in the limit $\xi\to 0$.

A.2 Hidden maximum principles from internal strong stability

When one requires that the substages satisfy a discrete maximum principle, and limits at every stage in the Shu Osher representation, one implies substage maximum principles. For example our SSP33 scheme satisfies the (hidden) maximum principles

$$k^1\in[\min\{u^n_{i+1},u^n_{i-1},u^n,u^n_{j+1},u^n_{j-1}\},\max\{u^n_{i+1},u^n_{i-1},u^n,u^n_{j+1},u^n_{j-1}\}], \tag{A.37}$$

$$k^2\in[\min\{k^1_{i+1},k^1_{i-1},k^1,k^1_{j+1},k^1_{j-1}\},\max\{k^1_{i+1},k^1_{i-1},k^1,k^1_{j+1},k^1_{j-1}\}], \tag{A.38}$$

$$u^{n+1}\in[\min\{k^2_{i+1},k^2_{i-1},k^2,k^2_{j+1},k^2_{j-1}\},\max\{k^2_{i+1},k^2_{i-1},k^2,k^2_{j+1},k^2_{j-1}\}], \tag{A.39}$$

instead of the original maximum principle

$$u^{n+1}_{i,j}\in[\min\{u^n_{i+1},u^n_{i-1},u^n,u^n_{j+1},u^n_{j-1}\},\max\{u^n_{i+1},u^n_{i-1},u^n,u^n_{j+1},u^n_{j-1}\}]. \tag{A.40}$$

This implies a strictly weaker maximum principle,

$$u^{n+1}_{i,j}\in[\min_{\forall k,l\in\{-3,-2,-1,0,1,2,3\}}u_{i+k,j+l},\max_{\forall k,l\in\{-3,-2,-1,0,1,2,3\}}u_{i+k,j+l}]. \tag{A.41}$$

but enforces positivity and local boundedness of the substages. Which could be essential if the substages are used in other parts of the dynamical core.

A.3 Improved speed implicit midpoint fluxes

1. Construct and solve the linearised slope limiter system

$$u_K^{n+1/2} = u_K^n - \frac{\Delta t}{2} \sum_{L \in N(K)} f_{K,L}^{n+1/2,n,h}, \tag{A.42}$$

for $u_K^{n+1/2}$. Where the numerical flux above is linearised at the old time step in mass conserving form as follows

$$f_{K,L}^{n+1/2,n,h} = \gamma_{K,L}^{n+1/2}[u_K^{h,n+1/2} + 1/4\phi(r_K^n)(u_L^{h,n+1/2} - u_J^{h,n+1/2})], \tag{A.43}$$

$$- \gamma_{L,K}^{n+1/2}[u_L^{h,n+1/2} + 1/4\phi(r_K^n)(u_K^{h,n+1/2} - u_M^{h,n+1/2})]. \tag{A.44}$$

and put into the matrix left hand side.

2. Using the newly computed substage $u_K^{n+1/2}$ as well as r_K^n and $\gamma_{K,L}^{n+1/2}$ we explicitly rebuild the previously defined fluxes $f_{K,L}^{n+1/2,n,h}$.

3. We reconstruct the the one stage transportative second order numerical flux,

$$F_{KL}^{high} = \Delta t(f_{K,L}^{n+1/2,n,h}). \tag{A.45}$$

Although this formulation is not based on the extended Shu-Osher representation, the difference in accuracy and monotonicity has been barely discernable and has fewer flux calculations making the scheme cheaper with reduced complexity.

A.4 Vertex inclusion

The theory in chapter three can be extended to include schemes with corner defined flux contributing quadrature points, but one has to account for the number of fluxes a corner contributing quadrature point could be contributing to. The theorem is the same as before, but with corners contributing to many different fluxes and unless a better weighted decomposition of the cell average can be found this will likely need a smaller Courant number restriction. We sketch some details below.

Theorem A.4.1 (Monotone DG and FV schemes with vertex inclusion)**.** *The cell mean value at the next timestep $\bar{u}_K^{n+1}$ is described by cell mean evolution*

$$\bar{u}_K^{n+1} = \bar{u}_K^n - \Delta t \frac{1}{|K|} \sum_{L \in N(K)} |\sigma_{KL}| \sum_{q \in \sigma_{KL}} w_q^{\sigma_{KL}} F(p_K(x_q), p_L(x_q), \boldsymbol{v} \cdot \boldsymbol{n}_{KL}), \tag{A.46}$$

and can be expressed as a monotone function of quadrature point evaluations. If all quadrature point evaluations arising from the cell mean decomposition and the quadrature point evaluations are also all positive $p_K(x_q) \geq 0 \quad \forall q \in K \quad \forall K$, and all the strict interior face defined Riemann problems Courant number restrictions are

satisfied:

$$\Delta t \frac{w_q^{\sigma}|\sigma_{KL}|}{w_q^K|K|}\frac{\partial F}{\partial p_K} \leq 1, \quad \forall q \in \sigma_{KL}, \quad \forall L \in N(K), \quad \forall K \in \mathcal{M}_h \tag{A.47}$$

and the vertex defined quadrature Courant numbers hold, supposing the vertex only has 3 neighbouring cells K, L, M, then this will look like the following:

$$\Delta t \frac{w_q^{\sigma_{KM}}|\sigma_{KM}|}{w_q^K|K|}\frac{\partial F}{\partial p_K}, \Delta t \frac{w_q^{\sigma_{KL}}|\sigma_{KL}|}{w_q^K|K|}\frac{\partial F}{\partial p_K} \leq 1/2, \tag{A.48}$$

then the scheme is positivity preserving $\bar{u}_K^{n+1} \geq 0$ irregardless of the compressibility of the vector field. If the vector field satisfies the following discrete divergence free condition,

$$\frac{1}{|K|}\sum_{L \in N(K)} |\sigma_{KL}| \sum_{q \in \sigma_{KL}} w_q^{\sigma_{KL}} (\boldsymbol{v} \cdot \boldsymbol{n}_{KL}) = 0, \tag{A.49}$$

as well as local boundedness of quadrature point evaluations,

$$p_K(x_q) \in [m_K, M_K], \quad \forall x_q \in K^{int} \tag{A.50}$$

$$p_K(x_q), p_L(x_q) \in [m_K, M_K], \quad \forall x_q \in \sigma_{KL}, \quad \forall L \in N(K) \tag{A.51}$$

$$\Delta t \frac{w_q^{\sigma}|\sigma_{KL}|}{w_q^K}\frac{\partial F}{\partial p_K} \leq 1, \quad \forall q \in \sigma_{KL}, \quad \forall L \in N(K) \tag{A.52}$$

then the next time level will satisfy a $\bar{u}_K^{n+1} \in [m_K, M_K]$ local boundedness principle.

A.4.1 New local boundedness slope limiter

We will use the theoretical results established in Theorem 3.2.1, to create a local maximum principle limiter capable of preserving

$$\min_{L \in N^2(K) \cap N(K)} \bar{u}_L^n \leq \bar{u}_K^{n+1} \leq \max_{L \in N^2(K) \cap N(K)} \bar{u}_L^n. \tag{A.53}$$

Description of new Multidimensional slope limiters. The internal quadrature points must satisfy the regular local maximum principle, and each edge has its own larger stencil maximum principle.

1. Per face σ_{KL}, we compute and associate the local face defined maximum principle bounds

$$m_{\sigma_{KL}}, M_{\sigma_{KL}} = \min_{M \in N(L) \cap N(K)} \bar{u}_M^n, \max_{M \in N(L) \cap N(K)} \bar{u}_M^n \tag{A.54}$$

   this is associated to each $x_q \in \sigma_{KL}$ not on a vertex.

2. Per cell K we associate the traditional maximum principle

$$m_{K_{int}}, M_{K_i nt} = \min_{L \in N(K) \cap K} \bar{u}_L^n, \max_{L \in N(K) \cap K} \bar{u}_L^n, \tag{A.55}$$

to each internal quadrature point associated with the cell mean decomposition formula.

3. Per vertex of K, belonging to σ_{KL}, σ_{KM} we compute the local vertex maximum principle bounds

$$m_{v_{KLM}}, M_{v_{KLM}} = \min_{i \in N(L) \cap N(K) \cap N(M)} \bar{u}_i^n, \max_{i \in N(L) \cap N(K) \cap N(M)} \bar{u}_i^n \tag{A.56}$$

4. We then per cell compute all the Barth and Jespersen quadrature corrections factors using [42], α_q to ensure $\tilde{p}_K(x) = \alpha(p_K(x) - \bar{u}_K) + \bar{u}_K$, satisfies the conditions in Theorem A.4.1.

$$\tilde{p}_K(x_q) \in [m_{K_{int}}, M_{K_{int}}] \forall q \in K_{int} \tag{A.57}$$
$$\tilde{p}_K(x_q) \in [m_{\sigma_{KL}}, M_{\sigma_{KL}}] \forall q \in \sigma_{KL} \forall L \in N(K) \tag{A.58}$$
$$\tilde{p}_K(x_q) \in [m_{v_{KLM}}, M_{v_{KLM}}] \forall q \in V(K) \tag{A.59}$$
$$\tag{A.60}$$

then choose the smallest value,

$$\alpha = \min_{\forall q \in K \cap L} \alpha_q \tag{A.61}$$

Bibliography

[1] I. M. Jánosi, A. Padash, J. A. Gallas, H. Kantz, Passive tracer advection in the equatorial pacific region: statistics, correlations and a model of fractional brownian motion, Ocean Science 18 (2) (2022) 307–320.

[2] P. J. Rasch, D. L. Williamson, Computational aspects of moisture transport in global models of the atmosphere, Quarterly Journal of the Royal Meteorological Society 116 (495) (1990) 1071–1090.

[3] B. Leonard, The ultimate conservative difference scheme applied to unsteady one-dimensional advection, Computer Methods in Applied Mechanics and Engineering 88 (1) (1991) 17–74. doi:https://doi.org/10.1016/0045-7825(91)90232-U.
URL https://www.sciencedirect.com/science/article/pii/004578259190232U

[4] S.-Y. Jun, S.-J. Choi, B.-M. Kim, Dynamical core in atmospheric model does matter in the simulation of arctic climate, Geophysical Research Letters 45 (6) (2018) 2805–2814.

[5] N. Brown, Exploring the acceleration of the met office nerc cloud model using fpgas, in: International Conference on High Performance Computing, Springer, 2019, pp. 567–586.

[6] H. Ritchie, C. Temperton, A. Simmons, M. Hortal, T. Davies, D. Dent, M. Hamrud, Implementation of the semi-lagrangian method in a high-resolution version of the ecmwf forecast model, Monthly Weather Review 123 (2) (1995) 489–514.

[7] J. S. Sawyer, A semi-lagrangian method of solving the vorticity advection equation, Tellus 15 (4) (1963) 336–342.

[8] M. Kwizak, A. J. Robert, A semi-implicit scheme for grid point atmospheric models of the primitive equations, Monthly Weather Review 99 (1) (1971) 32–36.

[9] J. Bates, A. McDonald, Multiply-upstream, semi-lagrangian advective schemes: Analysis and application to a multi-level primitive equation model, Monthly Weather Review 110 (12) (1982) 1831–1842.

[10] J. Pudykiewicz, A. Staniforth, Some properties and comparative performance of the semi-lagrangian method of robert in the solution of the advection-diffusion equation, Atmosphere-Ocean 22 (3) (1984) 283–308.

[11] M. Falcone, R. Ferretti, Convergence analysis for a class of high-order semi-lagrangian advection schemes, SIAM Journal on Numerical Analysis 35 (3) (1998) 909–940.

[12] H.-C. Kuo, R. Williams, Semi-lagrangian solutions to the inviscid burgers equation, Monthly Weather Review 118 (6) (1990) 1278–1288.

[13] J. Pudykiewicz, R. Benoit, A. Staniforth, Preliminary results from a partial lrtap model based on an existing meteorological forecast model, Atmosphere-ocean 23 (3) (1985) 267–303.

[14] P. K. Smolarkiewicz, J. A. Pudykiewicz, A class of semi-lagrangian approximations for fluids, Journal of Atmospheric Sciences 49 (22) (1992) 2082–2096.

[15] P. Héreil, R. Laprise, Sensitivity of internal gravity waves solutions to the time step of a semi-implicit semi-lagrangian nonhydrostatic model, Monthly weather review 124 (5) (1996) 972–999.

[16] P. Bartello, S. J. Thomas, The cost-effectiveness of semi-lagrangian advection, Monthly weather review 124 (12) (1996) 2883–2897.

[17] R. Bermejo, A. Staniforth, The conversion of semi-lagrangian advection schemes to quasi-monotone schemes, Monthly Weather Review 120 (11) (1992) 2622–2632.

[18] D. L. Williamson, P. J. Rasch, Two-dimensional semi-lagrangian transport with shape-preserving interpolation, Monthly Weather Review 117 (1) (1989) 102–129.

[19] B. Leonard, A. Lock, M. MacVean, Conservative explicit unrestricted-time-step multidimensional constancy-preserving advection schemes, Monthly Weather Review 124 (11) (1996) 2588–2606.

[20] S.-J. Lin, R. B. Rood, Multidimensional flux-form semi-lagrangian transport schemes, Monthly Weather Review 124 (9) (1996) 2046–2070.

[21] M. Zerroukat, T. Allen, A three-dimensional monotone and conservative semi-lagrangian scheme (slice-3d) for transport problems, Quarterly Journal of the Royal Meteorological Society 138 (667) (2012) 1640–1651.

[22] P. H. Lauritzen, R. D. Nair, P. A. Ullrich, A conservative semi-lagrangian multi-tracer transport scheme (cslam) on the cubed-sphere grid, Journal of Computational Physics 229 (5) (2010) 1401–1424.

[23] A. Staniforth, N. Wood, Aspects of the dynamical core of a nonhydrostatic, deep-atmosphere, unified weather and climate-prediction model, Journal of Computational Physics 227 (7) (2008) 3445–3464.

[24] A. Harten, J. Hyman, P. Lax, On finite-difference approximations and entropy conditions for shocks, Comm. Pure Appl. Math. 29 (1976) 297–322.

[25] A. Harten, High resolution schemes for hyperbolic conservation laws, Journal of Computational Physics 49 (3) (1983) 357 – 393. doi:https://doi.org/10.1016/0021-9991(83)90136-5.
URL http://www.sciencedirect.com/science/article/pii/0021999183901365

[26] S. K. Godunov, I. O. Bohachevsky, Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics, 1959.

[27] S. Spekreijse, Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws, Mathematics of computation 49 (179) (1987) 135–155.

[28] A. Jameson, Positive schemes and shock modelling for compressible flows, International Journal for Numerical Methods in Fluids 20 (8-9) (1995) 743–776.

[29] S. Osher, Riemann solvers, the entropy condition, and difference approximations, SIAM Journal on Numerical Analysis 21 (2) (1984) 217–235.
URL http://www.jstor.org/stable/2157297

[30] A. Harten, B. Engquist, S. Osher, S. R. Chakravarthy, Uniformly high order accurate essentially non-oscillatory schemes, iii, in: Upwind and high-resolution schemes, Springer, 1987, pp. 218–290.

[31] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, Journal of computational physics 77 (2) (1988) 439–471.

[32] X.-D. Liu, S. Osher, T. Chan, Weighted essentially non-oscillatory schemes, Journal of computational physics 115 (1) (1994) 200–212.

[33] D. L. Williamson, The evolution of dynamical cores for global atmospheric models, Journal of the Meteorological Society of Japan. Ser. II 85 (2007) 241–269.

[34] A. Harten, P. D. Lax, B. V. Leer, On upstream differencing and godunov-type schemes for hyperbolic conservation laws, SIAM Review 25 (1) (1983) 35–61.
URL http://www.jstor.org/stable/2030019

[35] S. Osher, S. Chakravarthy, High resolution schemes and the entropy condition, SIAM Journal on Numerical Analysis 21 (5) (1984) 955–984.

[36] W. Hua-mo, On the possible accuracy of tvd schemes, Journal of Computational Mathematics (1992) 245–252.

[37] M. ZIJLEMA, P. WESSELING, Higher-order flux-limiting schemes for the finite volume computation of incompressible flow, International Journal of Computational Fluid Dynamics 9 (2) (1998) 89–109. arXiv:https://doi.org/10.1080/10618569808940844, doi:10.1080/10618569808940844.
URL https://doi.org/10.1080/10618569808940844

[38] P. K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws, SIAM journal on numerical analysis 21 (5) (1984) 995–1011.

[39] J. Goodman, R. Leveque, On the accuracy of stable schemes for 2d scalar conservation laws, Mathematics of Computation 45 (171) (1985) 15–21. doi:10.1090/S0025-5718-1985-0790641-4.

[40] W. Hundsdorfer, B. Koren, J. Verwer, et al., A positive finite-difference advection scheme, Journal of computational physics 117 (1) (1995) 35–46.

[41] B. Koren, A robust upwind discretization method for advection, diffusion and source terms, Centrum voor Wiskunde en Informatica Amsterdam, 1993.

[42] T. Barth, D. Jespersen, The design and application of upwind schemes on unstructured meshes, in: 27th Aerospace sciences meeting, 1989, p. 366.

[43] X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, Journal of Computational Physics 229 (9) (2010) 3091–3120.

[44] X. Zhang, C.-W. Shu, On positivity-preserving high order discontinuous galerkin schemes for compressible euler equations on rectangular meshes, Journal of Computational Physics 229 (23) (2010) 8918–8934.

[45] C. Yang, X.-C. Cai, A scalable fully implicit compressible euler solver for mesoscale nonhydrostatic simulation of atmospheric flows, SIAM Journal on Scientific Computing 36 (5) (2014) S23–S47. arXiv:https://doi.org/10.1137/130919167, doi:10.1137/130919167.
URL https://doi.org/10.1137/130919167

[46] D. S. Abdi, F. X. Giraldo, E. M. Constantinescu, L. E. Carr, L. C. Wilcox, T. C. Warburton, Acceleration of the implicit–explicit nonhydrostatic unified model of the atmosphere on manycore processors, The International Journal of High Performance Computing Applications 33 (2) (2019) 242–267.

[47] H. C. Yee, R. F. Warming, A. Harten, Implicit total variation diminishing (tvd) schemes for steady-state calculations, 1983.

[48] C. Bolley, M. Crouzeix, Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques, RAIRO. Analyse numérique 12 (3) (1978) 237–245.

[49] M. N. Spijker, Contractivity in the numerical solution of initial value problems, Numerische Mathematik 42 (3) (1983) 271–290. doi:10.1007/BF01389573.
URL https://doi.org/10.1007/BF01389573

[50] J. F. B. M. Kraaijevanger, Contractivity of runge-kutta methods, BIT Numerical Mathematics 31 (3) (1991) 482–528.

[51] L. Ferracina, M. Spijker, Strong stability of singly-diagonally-implicit runge–kutta methods, Applied Numerical Mathematics 58 (11) (2008) 1675–1686.

[52] D. I. Ketcheson, C. B. Macdonald, S. Gottlieb, Optimal implicit strong stability preserving runge–kutta methods, Applied Numerical Mathematics 59 (2) (2009) 373–392.

[53] J. Boris, D. Book, Flux-corrected transport: I. shasta, a fluid transport algorithm that works, J. Comput. Phys. 18, cited By 18 (1973).

[54] S. T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, Journal of Computational Physics 31 (3) (1979) 335 – 362. doi:https://doi.org/10.1016/0021-9991(79)90051-2.
URL http://www.sciencedirect.com/science/article/pii/0021999179900512

[55] D. L. Book, J. P. Boris, K. Hain, Flux-corrected transport ii: Generalizations of the method, Journal of Computational Physics 18 (3) (1975) 248–283.

[56] J. P. Boris, D. L. Book, Flux-corrected transport. iii. minimal-error fct algorithms, Journal of Computational Physics 20 (4) (1976) 397–431.

[57] B. McDonald, Flux-corrected pseudospectral method for scalar hyperbolic conservation laws, Journal of Computational Physics 82 (2) (1989) 413–428.

[58] D. Kuzmin, Explicit and implicit fem-fct algorithms with flux linearization, Journal of Computational Physics 228 (7) (2009) 2517–2534. doi:https://doi.org/10.1016/j.jcp.2008.12.011.
URL https://www.sciencedirect.com/science/article/pii/S0021999108006475

[59] K. Chatterjee, R. W. Schunk, A flux-corrected transport based hydrodynamic model for the plasmasphere refilling problem following geomagnetic storms, in: AGU Fall Meeting Abstracts, Vol. 2017, 2017, pp. SA43B–2649.

[60] D. Yang, E. Liu, Z. Zhang, J. Teng, Finite-difference modelling in two-dimensional anisotropic media using a flux-corrected transport technique, Geophysical Journal International 148 (2) (2002) 320–328.

[61] J.-L. Lee, R. Bleck, A. E. MacDonald, A multistep flux-corrected transport scheme, Journal of Computational Physics 229 (24) (2010) 9284–9298.

[62] C. Chaplin, P. Colella, A single-stage flux-corrected transport algorithm for high-order finite-volume methods, Communications in Applied Mathematics and Computational Science 12 (1) (2017) 1–24.

[63] G. Patnaik, R. Guirguis, J. Boris, E. Oran, A barely implicit correction for flux-corrected transport, Journal of Computational Physics 71 (1) (1987) 1–20.

[64] P. Steinle, R. Morrow, An implicit flux-corrected transport algorithm, Journal of computational physics 80 (1) (1989) 61–71.

[65] D. Kuzmin, S. Turek, Flux correction tools for finite elements, Journal of Computational Physics 175 (2) (2002) 525–558.

[66] D. Kuzmin, M. Moeller, Algebraic flux correction i. scalar conservation laws in r. loehner, d. kuzmin und s. turek (hg.), flux–corrected transport: Principles, algorithms and applications (2005).

[67] D. Kuzmin, Explicit and implicit fem-fct algorithms with flux linearization, Journal of Computational Physics 228 (7) (2009) 2517–2534.

[68] P. K. Smolarkiewicz, P. J. Rasch, Monotone advection on the sphere: An eulerian versus semi-lagrangian approach, Journal of atmospheric sciences 48 (6) (1991) 793–810.

[69] H. K. Versteeg, W. Malalasekera, An introduction to computational fluid dynamics: the finite volume method, Pearson education, 2007.

[70] H. Deconinck, P. L. Roe, R. Struijs, A multidimensional generalization of roe's flux difference splitter for the euler equations, Computers & fluids 22 (2-3) (1993) 215–222.

[71] W. Hundsdorfer, R. Trompert, Method of lines and direct discretization: a comparison for linear advection, Applied numerical mathematics 13 (6) (1994) 469–490.

[72] R. Eymard, T. Gallouët, R. Herbin, Finite Volume Methods, in: J. L. Lions, P. Ciarlet (Eds.), Solution of Equation in Rn (Part 3), Techniques of Scientific Computing (Part 3), Vol. 7 of Handbook of Numerical Analysis, Elsevier, 2000, pp. 713–1020. doi:10.1016/S1570-8659(00)07005-8.
URL https://hal.archives-ouvertes.fr/hal-02100732

[73] P. Wilders, G. Fotia, A positive spatial advection scheme on unstructured meshes for tracer transport, Journal of computational and applied mathematics 140 (1-2) (2002) 809–821.

[74] H. Nishikawa, A truncation error analysis of third-order muscl scheme for nonlinear conservation laws, International Journal for Numerical Methods in Fluids 93 (4) (2021) 1031–1052.

[75] B. Leonard, Order of accuracy of quick and related convection-diffusion schemes, Applied Mathematical Modelling 19 (11) (1995) 640–653.

[76] H. Nishikawa, The quick scheme is a third-order finite-volume scheme with point-valued numerical solutions, International Journal for Numerical Methods in Fluids 93 (7) (2021) 2311–2338.

[77] [link].
URL https://cims.nyu.edu/~donev/Teaching/CFD/Lectures/ThirdOrderUpwind.pdf

[78] T. Barth, M. Ohlberger, Finite volume methods: foundation and analysis (2003).

[79] N. Waterson, H. Deconinck, A unified approach to the design and application of bounded higher-order convection schemes, Numerical methods in laminar and turbulent flow. 9 (1995) 203–214.

[80] G. D. Van Albada, B. v. Leer, W. Roberts, A comparative study of computational methods in cosmic gas dynamics, in: Upwind and high-resolution schemes, Springer, 1997, pp. 95–103.

[81] R. J. Leveque, High-resolution conservative algorithms for advection in incompressible flow, SIAM Journal on Numerical Analysis 33 (2) (1996) 627–665.
URL http://www.jstor.org/stable/2158391

[82] D. Zhang, C. Jiang, D. Liang, L. Cheng, A review on tvd schemes and a refined flux-limiter for steady-state calculations, Journal of Computational Physics 302 (2015) 114–154.

[83] F. Kemm, A comparative study of tvd-limiters—well-known limiters and an introduction of new ones, International Journal for Numerical Methods in Fluids 67 (4) (2011) 404–440.

[84] L. Lin, Z. Liu, Tvdal: Total variation diminishing scheme with alternating limiters to balance numerical compression and diffusion, Ocean Modelling 134 (2019) 42–50.

[85] J. Pietrzak, The use of tvd limiters for forward-in-time upstream-biased advection schemes in ocean modeling, Monthly Weather Review 126 (3) (1998) 812–830.

[86] D. Williamson, The evolution of dynamical cores for global atmospheric models, J. Royal Met. Soc. Japan 85B (2007) 241–269. doi:10.2151/jmsj.85B.241.

[87] M. Darwish, F. Moukalled, Tvd schemes for unstructured grids, International Journal of heat and mass transfer 46 (4) (2003) 599–611.

[88] W. C. Skamarock, J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, T. D. Ringler, A multiscale nonhydrostatic atmospheric model using centroidal voronoi tesselations and c-grid staggering, Monthly Weather Review 140 (9) (2012) 3090–3105.

[89] P. A. Ullrich, C. Jablonowski, Mcore: A non-hydrostatic atmospheric dynamical core utilizing high-order finite-volume methods, Journal of Computational Physics 231 (15) (2012) 5078–5108.

[90] J. F. Kelly, F. X. Giraldo, Continuous and discontinuous galerkin methods for a scalable three-dimensional nonhydrostatic atmospheric model: Limited-area mode, Journal of Computational Physics 231 (24) (2012) 7988–8008.

[91] J. S. Park, S.-H. Yoon, C. Kim, Multi-dimensional limiting process for hyperbolic conservation laws on unstructured grids, Journal of Computational Physics 229 (3) (2010) 788–812.

[92] X. Zhang, Y. Xia, C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high order discontinuous galerkin schemes for conservation laws on triangular meshes, Journal of Scientific Computing 50 (1) (2012) 29–62.

[93] C. Ollivier-Gooch, M. Van Altena, A high-order-accurate unstructured mesh finite-volume scheme for the advection–diffusion equation, Journal of Computational Physics 181 (2) (2002) 729–752.

[94] D. Kuzmin, A vertex-based hierarchical slope limiter for p-adaptive discontinuous galerkin methods, Journal of computational and applied mathematics 233 (12) (2010) 3077–3085.

[95] J. Li, Y. Zhang, Enhancing the stability of a global model by using an adaptively implicit vertical moist transport scheme, Meteorology and Atmospheric Physics 134 (3) (2022) 1–12.

[96] T. B. Nguyen, H. De Sterck, L. Freret, C. P. Groth, High-order implicit time-stepping with high-order central essentially-non-oscillatory methods for unsteady three-dimensional computational fluid dynamics simulations, International Journal for Numerical Methods in Fluids 94 (2) (2022) 121–151.

[97] N. N. Kuznetsov, On stable methods of solving a quasi-linear equation of first order in a class of discontinuous functions., Sov. Math., Dokl. 16 (1975) 1569–1573.

[98] A. Kolmogoroff, Ueber kompaktheit dr funktionenmengen bei der konvergenz im mittel, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse 1931 (1931) 60–63.
URL http://eudml.org/doc/59336

[99] S. N. Kruzhkov, Nichtlineare parabolische Gleichungen mit zwei unabhängigen Veränderlichen., Tr. Mosk. Mat. O.-va 16 (1967) 329–346.

[100] S. Gottlieb, C.-W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods, SIAM review 43 (1) (2001) 89–112.

[101] I. Higueras, On strong stability preserving time discretization methods, Journal of Scientific Computing 21 (2) (2004) 193–223.

[102] V. Venkatakrishnan, Convergence to steady state solutions of the euler equations on unstructured grids with limiters, Journal of computational physics 118 (1) (1995) 120–130.

[103] C. Michalak, C. Ollivier-Gooch, Accuracy preserving limiter for the high-order accurate solution of the euler equations, Journal of Computational Physics 228 (23) (2009) 8693–8711.

[104] K. Michalak, C. O. Gooch, Differentiability of slope limiters on unstructured grids, in: Proceedings of fourteenth annual conference of the computational fluid dynamics society of Canada, 2006.

[105] H. Nishikawa, New Unstructured-Grid Limiter Functions. arXiv: https://arc.aiaa.org/doi/pdf/10.2514/6.2022-1374, doi:10.2514/6.2022-1374.
URL https://arc.aiaa.org/doi/abs/10.2514/6.2022-1374

[106] K. Michalak, C. Ollivier-Gooch, Matrix-explicit gmres for a higher-order accurate inviscid compressible flow solver, in: 18th AIAA Computational Fluid Dynamics Conference, 2007, p. 3943.

[107] H. Nishikawa, Efficient gradient stencils for robust implicit finite-volume solver convergence on distorted grids, Journal of Computational Physics 386 (2019) 486–501.

[108] D. Kuzmin, A vertex-based hierarchical slope limiter for p-adaptive discontinuous galerkin methods, Journal of computational and applied mathematics 233 (12) (2010) 3077–3085.

[109] M. Berger, M. Aftosmis, S. Muman, Analysis of slope limiters on irregular grids, in: 43rd AIAA Aerospace Sciences Meeting and Exhibit, 2005, p. 490.

[110] C. R. DeVore, Flux-corrected transport techniques for multidimensional compressible magnetohydrodynamics, Journal of Computational Physics 92 (1) (1991) 142–160.

[111] S. Gottlieb, J. S. Mullen, S. J. Ruuth, A fifth order flux implicit weno method, Journal of Scientific Computing 27 (1) (2006) 271–287.

[112] P. Khosla, S. Rubin, A diagonally dominant second-order accurate implicit scheme, Computers & Fluids 2 (2) (1974) 207 – 209. doi:https://doi.org/10.1016/0045-7930(74)90014-0.
URL http://www.sciencedirect.com/science/article/pii/0045793074900140

[113] J. F. B. M. Kraaijevanger, Contractivity of runge-kutta methods, BIT Numerical Mathematics 31 (3) (1991) 482–528. doi:10.1007/BF01933264.
URL https://doi.org/10.1007/BF01933264

[114] P. Van Slingerland, M. Borsboom, C. Vuik, A local theta scheme for advection problems with strongly varying meshes and velocity profiles, Reports of the Department of Applied Mathematical Analysis, 08-17 (2008).

[115] D. I. Ketcheson, C. B. MacDonald, S. J. Ruuth, Spatially partitioned embedded runge–kutta methods, SIAM Journal on Numerical Analysis 51 (5) (2013) 2887–2910.

[116] D. Kuzmin, M. Möller, S. Turek, High-resolution fem–fct schemes for multidimensional conservation laws, Computer Methods in Applied Mechanics and Engineering 193 (45-47) (2004) 4915–4946.

[117] N. Bell, L. N. Olson, J. Schroder, PyAMG: Algebraic multigrid solvers in python, Journal of Open Source Software 7 (72) (2022) 4142. doi:10.21105/joss.04142.
URL https://doi.org/10.21105/joss.04142

[118] I. Higueras, Representations of runge-kutta methods and strong stability preserving methods, SIAM Journal on Numerical Analysis 43 (3) (2006) 924–948.
URL http://www.jstor.org/stable/4101273