

# The ‘Recovered Space’ Advection Scheme for Lowest-Order Compatible Finite Element Methods

Thomas M. Bendall, Colin J. Cotter, Jemma Shipton

May 11, 2020

## Abstract

We present a new compatible finite element advection scheme for the compressible Euler equations. Unlike the discretisations described in Cotter and Kuzmin (2016) and Shipton et al (2018), the discretisation uses the lowest-order family of compatible finite element spaces, but still retains second-order numerical accuracy. This scheme obtains this second-order accuracy by first ‘recovering’ the function in higher-order spaces, before using the discontinuous Galerkin advection schemes of Cotter and Kuzmin (2016). As well as describing the scheme, we also present its stability properties and a strategy for ensuring boundedness. We then demonstrate its properties through some numerical tests, before presenting its use within a model solving the compressible Euler equations.

*Keywords:* Advection scheme; Discontinuous Galerkin; Compatible finite element methods; Numerical weather prediction

## 1 Introduction

Over the past few decades, technological improvements in parallel computing have driven increased performance of numerical models of weather and climate systems, allowing them to be run at increasingly finer resolutions and delivering significantly improved predictive capabilities. However, the traditional latitude-longitude grids used in these models are leading to the approach of a scalability bottleneck: improvements in potential computational power no longer lead to improved model performance, as the clustering of points around the poles limits the rate of data communication (the so-called ‘pole problem’). This has led to a search for alternative grids for these models.

However, latitude-longitude grids (coupled with Arakawa C-grid staggering) provided many properties important for accurate representation of atmospheric motion: avoiding spurious pressure modes, accurately representing geostrophic balance and not supporting spurious gravity-inertia or Rossby wave modes. These properties are described in [1]. It is therefore desirable that any alternative grid should maintain these properties.

Finite element methods have therefore been explored, as they offer the opportunity to solve the equations of motion with arbitrary mesh structures. Extending previous work into the use of mixed finite element methods for such geophysical fluid models, [2] proposed mixed finite element methods with two crucial properties. Firstly, that the finite element spaces used should be *compatible* with one another, so that the discrete versions of the vector calculus operators preserve the properties that  $\text{div}(\text{curl}) = 0$  and that  $\text{curl}(\text{grad}) = 0$ .

Secondly, the function spaces used for the velocity and pressure variables should be chosen to fix the ratio of velocity to pressure degrees of freedom per element to be 2:1. This was shown in [2] to be a necessary

condition for the avoidance of spurious inertia-gravity and Rossby wave modes. [2] also proposed two sets of finite element spaces that would meet these criteria on quadrilateral and triangular elements.

[2] has been followed by a series of works into the use of compatible finite element methods for use in numerical weather models, for instance [3], [4], [5], [6] and [7]. Most relevant to the work here are [8], which presented a full discretisation of the Euler-Boussinesq equations in the context of compatible finite element methods; and [9], which introduced an embedded discontinuous Galerkin transport scheme.

The compatible finite element framework thus suggests using particular compatible finite element spaces for the velocity  $\mathbf{v}$  and density  $\rho$ . For horizontal grids, the examples given in [2] are  $\mathbf{v} \in \text{RT}_{k+1}$  with  $\rho \in \text{Q}_k$  on quadrilateral elements, and  $\mathbf{v} \in \text{BDFM}_k$  with  $\rho \in \text{P}_k$  on triangular elements. Here  $\text{RT}_k$  is the space of  $k$ -th degree Raviart-Thomas quadrilateral elements,  $\text{Q}_k$  is the space of discontinuous  $k$ -th order polynomial elements on quadrilaterals,  $\text{BDFM}_k$  is the space of  $k$ -th degree Brezzi-Douglas-Fortin-Marini triangular elements and finally  $\text{P}_k$  is the space of  $k$ -th order polynomial elements on triangles. Most of these elements appear in the Periodic Table of Finite Elements [10].

Another set of spaces to consider are those used in 2D vertical slices, which are relevant for test cases used in developing the model. These models are constructed as the tensor product  $U \times V$  of a 1D horizontal space  $U$  with a 1D vertical space  $V$ . We will denote the space of  $k$ -th degree continuous polynomial elements by  $\text{CG}_k$ , and that of  $l$ -th degree discontinuous polynomials by  $\text{DG}_l$ . In this case horizontal velocities lie in the space  $\text{CG}_{k+1} \times \text{DG}_l$ , vertical velocities are in  $\text{DG}_k \times \text{CG}_{l+1}$ , whilst the density is in  $\text{DG}_k \times \text{DG}_l$ . In order to mimic the Charney-Philips grid used in finite difference models, the potential temperature  $\theta$  lies in the partially-continuous space,  $\text{DG}_k \times \text{CG}_{l+1}$ , i.e. temperature degrees of freedom (DOFs) are co-located with those for vertical velocity, as in [6] and [8]. The construction of tensor product finite element spaces is described by [11]. Throughout the rest of the paper we will consider the case that  $k = l$  for quadrilateral elements in this vertical slice set-up, although the advection scheme can be applied more generally.

Advection schemes for the  $k = 1$  family of spaces have been presented in [9], [8] and [7]. An advantage of the  $k = 1$  degree was that it is easy to formulate for it advection schemes with the property of second-order numerical accuracy, i.e. that the error associated with the discretisation of the advection process is proportional to  $(\Delta x)^2$ , the grid size squared. This is one of the crucial properties of discretisations for numerical weather models listed in [1]. However, in the  $k = 1$  families, coupling of the dynamics to the sub-grid physical processes (for example the effects of moisture or radiation) may be more challenging than for  $k = 0$  case. The effects of such physical processes are commonly expressed as tendencies to the prognostic variables and typically calculated pointwise at the degrees of freedom. For the fields that are piecewise constant or linear, the pointwise values can be interpreted as mean quantities for that element and the tendencies can be formulated as such. In the  $k = 1$  case on quadrilateral elements, the temperature and moisture fields are piecewise quadratic functions in the vertical, and the physical interpretation of the values at the degrees of freedom becomes less clear. It is therefore desirable to consider  $k = 0$  spaces using advection schemes that have second-order numerical accuracy. The main result of this paper is thus a presentation of such an advection scheme for this set of spaces.

This scheme has been inspired by the results of [12], which implies that it is possible to reconstruct a discontinuous zeroth-order field in a continuous first-order space, via an averaging operation that has second-order numerical accuracy.

After describing the scheme in Section 2, this paper presents several of its properties in Section 3, including a general argument of its stability and von Neumann analysis of the scheme in three particular cases. Section 4 presents the results of numerical tests demonstrating the second-order numerical accuracy of the scheme, the stability calculations of Section 3.2 and the use of a limiter within the advection scheme. Finally, the use of the advection scheme within a model of the compressible Euler equations is presented in Section 5.

## 2 The ‘Recovered Space’ Scheme

The key idea upon which this scheme is based is the family of recovered finite element methods introduced by [12]. These methods combine features of discontinuous Galerkin approaches with conforming finite element methods. They are similar to other recovery methods, such as those in [13, 14], in that they reconstruct higher-order polynomials from lower order data in a patch of cells. They differ in that they do not attempt to reproduce polynomials of a certain degree exactly. Instead, they involve mapping discontinuous finite element spaces to continuous ones, via recovery operators, relying on analysis estimates of stability and accuracy. The scheme that we will introduce involves the use of one of these operators to recover a function in a discontinuous first-order space from one in a discontinuous zeroth-order space. To do this, we first recover a first-order continuous function from the zeroth-order discontinuous function using an averaging operator described in [12] and [15]. This operator finds the values for any degree of freedom shared between elements in a continuous function space, by averaging between the values of the surrounding degrees of freedom from the discontinuous space.

Once this operator has been applied, existing transport schemes can be used to perform the advection upon the recovered field. This approach is compatible when the transport equation is in ‘advective’ form

$$\frac{\partial q}{\partial t} + \mathbf{v} \cdot \nabla q = 0, \quad (2.1)$$

or ‘conservative’ form

$$\frac{\partial q}{\partial t} + \nabla \cdot (q\mathbf{v}) = 0, \quad (2.2)$$

where  $q$  is the quantity to be transported by velocity  $\mathbf{v}$ . However most of our analysis will focus on the application of this scheme to the ‘advective’ form of the equation, under which the mass  $\int_{\Omega} q \, dx$  over the whole domain  $\Omega$  will only be necessarily conserved when the flow is incompressible,  $\nabla \cdot \mathbf{v} = 0$ .

### 2.1 The Scheme

First we will define a set of spaces that our functions will lie in. Let  $V_0(\Omega)$  be the lowest-order finite element space in which the initial field lies, where  $\Omega$  is our spatial domain\*.  $V_1(\Omega)$  is then the space of next degree, which will be fully discontinuous. We also have that  $V_0 \subset V_1$ .  $\tilde{V}_1(\Omega)$  is the fully continuous space of same degree as  $V_1(\Omega)$ , whilst  $\hat{V}_0(\Omega)$  is a broken (i.e. fully discontinuous) version of  $V_0(\Omega)$ . In many cases,  $\hat{V}_0(\Omega)$  and  $V_0(\Omega)$  will coincide.

We now define a series of operators to map between these spaces.

**Definition 1.** *The recovery operator  $\mathcal{R}$  acts upon a function in the initial space to make a function in the continuous space of higher-degree, so that  $\mathcal{R} : V_0 \rightarrow \tilde{V}_1$ . The operator has second-order numerical accuracy.*

**Assumption 1.** *The recovery operator  $\mathcal{R}$  has the property that for all  $\rho_0 \in V_0$ , there is some  $C > 0$  such that  $\|\mathcal{R}\rho_0\|_{L^2} \leq C\|\rho_0\|_{L^2}$ .*

**Definition 2.** *The injection operator  $\mathcal{I} : V \rightarrow V_1$  identifies a function in  $V_0$ ,  $\hat{V}_0$  or  $\tilde{V}_1$  as a member of  $V_1$ . This must be numerically implemented, although it does nothing else mathematically.*

**Definition 3.** *The projection operator  $\hat{\mathcal{P}} : \tilde{V}_1 \rightarrow \hat{V}_0$ , is defined to give  $\hat{u} = \hat{\mathcal{P}}\tilde{v}$ , from  $\tilde{v} \in \tilde{V}_1$ , by finding the solution  $\hat{u} \in \hat{V}_0$  to*

$$\int_{\Omega} \hat{\psi} \hat{u} \, dx = \int_{\Omega} \hat{\psi} \tilde{v} \, dx, \quad \forall \hat{\psi} \in \hat{V}_0. \quad (2.3)$$

---

\*This spatial domain can be arbitrary, but with geophysical applications in mind we anticipate the scheme being used upon rectangular or cuboid domains (with or without periodicity) or spherical shells. However the recovery operator that we consider in Section 2.3 is intended for use in flat spaces or with only scalar fields in curved spaces and we do not yet consider the application to transport of vector fields in curved spaces. Therefore in this work we will predominantly consider rectangular domains with a vertical coordinate, with rigid walls at the top and bottom edges.

**Definition 4.** *The advection operator  $\mathcal{A} : V_1 \rightarrow V_1$ , represents the action of performing one time step of a stable discretisation of the advection equation (in either advective or conservative form) and has second-order numerical accuracy in space.*

**Definition 5.** *The projection operator  $\mathcal{P} : V_1 \rightarrow V_0$  will have two forms. The first,  $\mathcal{P}_A$ , is defined to give  $u = \mathcal{P}_A v$  from  $v \in V_1$ , by finding the solution  $u \in V_0$  to*

$$\int_{\Omega} \psi u \, dx = \int_{\Omega} \psi v \, dx, \quad \forall \psi \in V_0, \quad (2.4)$$

where  $u \in V_0$  and  $v \in V_1$ .

The second form,  $\mathcal{P}_B$ , is composed of two operations:  $\mathcal{P}_I : V_1 \rightarrow \hat{V}_0$ , interpolation into the broken space by pointwise evaluation at degrees of freedom, and  $\mathcal{P}_R : \hat{V}_0 \rightarrow V_0$ , recovery from the broken space to the original space, restoring continuity via the reconstruction operator from [12].  $\mathcal{P}_B$  can thus be written as  $\mathcal{P}_B = \mathcal{P}_R \mathcal{P}_I$ .

In the case that  $V_0$  is fully discontinuous,  $\mathcal{P}_A$  and  $\mathcal{P}_B$  will be identical operations. However for fully or partially continuous  $V_0$ ,  $\mathcal{P}_B$  prevents the formation of any new maxima and minima, whereas  $\mathcal{P}_A$  does not. We may thus use  $\mathcal{P}_B$  as the projection operator when trying to bound the transport, such as for a moisture species. Further discussion can be found in Section 2.4. The drawback is that whilst  $\mathcal{P}_A$  preserves the mass (setting  $\psi = 1$  gives  $\int_{\Omega} u \, dx = \int_{\Omega} v \, dx$ ),  $\mathcal{P}_B$  does not necessarily do so.

**Definition 6.** *The ‘recovered space’ scheme then takes the function  $\rho_0^n \in V_0$  at the  $n$ -th time step and returns the function  $\rho_0^{n+1} \in V_0$  at the  $(n+1)$ -th time step by performing the following series of operations:*

$$\rho_0^{n+1} = \mathcal{P} \mathcal{A} \mathcal{I} (\mathcal{R} - \hat{\mathcal{P}} \mathcal{R} + 1) \rho_0^n, \quad (2.5)$$

where  $\mathcal{P}$  could be either  $\mathcal{P}_A$  or  $\mathcal{P}_B$ .

An important property of this scheme is that in the absence of flow, the field being advected must remain unchanged. In this case  $\mathcal{A}$  will be the identity, and since  $\mathcal{P} \mathcal{I} \mathcal{R} \equiv \mathcal{P} \mathcal{I} \hat{\mathcal{P}} \mathcal{R}$ , then  $\rho_0^{n+1} = \mathcal{P} \mathcal{I} \rho_0^n = \rho_0^n$ . In practice, mass will be only be conserved up to the precision used by the numerical solver for  $\mathcal{P}$ .

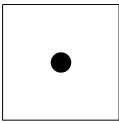
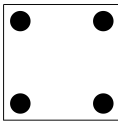
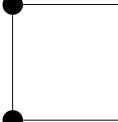
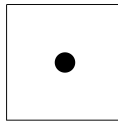
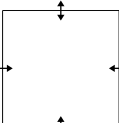
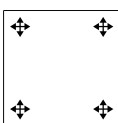
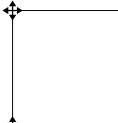
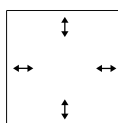
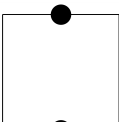
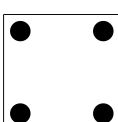
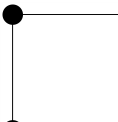
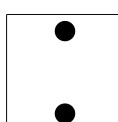
## 2.2 Example Spaces

In this section we give an example set of spaces  $\{V_0, V_1, \tilde{V}_1, \hat{V}_0\}$  on quadrilateral elements that can be used for this scheme, in the context of 2D vertical slice problems. The variables that we will consider are the density  $\rho$ , velocity  $\mathbf{v}$  and potential temperature  $\theta$ .

For each variable the space  $V_0$  is the normal space in which the variable lies. Since we are motivated by using the lowest-order family of compatible finite element spaces on quadrilateral elements, that is what we will use here. For  $\mathbf{v}$  this is  $V_0 = \text{RT}_0$ , the lowest-order Raviart-Thomas space with vector DOFs that have normal components continuous over cell boundaries. The density  $\rho$  lies  $\text{DG}_0 \times \text{DG}_0$ , which has a single DOF at the centre of the cell, and  $\theta \in \text{DG}_0 \times \text{CG}_1$ , i.e. discontinuous constant values in the horizontal but continuous linear in the vertical. The  $\theta$  DOFs are co-located with those for vertical velocity.

For the advection operator  $\mathcal{A}$  to have second-order numerical accuracy, the advection should take place in spaces that are at least linear in each direction. We therefore choose  $V_1$  to be the smallest entirely discontinuous space that is linear in both directions. For  $\rho$  and  $\theta$ , this is  $\text{DG}_1 \times \text{DG}_1$ , whilst for  $\mathbf{v}$  this is the vector  $\text{DG}_1 \times \text{DG}_1$  space, (and so is  $\text{DG}_1 \times \text{DG}_1$  for each component).

The space  $\tilde{V}_1$  is then formed by taking the completely continuous form of  $V_1$ , whilst  $\hat{V}_0$  is formed from the completely broken or discontinuous version of  $V_0$ . The full set of spaces is listed in Table 1, whilst also shows the spaces diagrammatically, representing scalar DOFs by dots and vector DOFs by arrows.

Variable	$V_0$	$V_1$	$\tilde{V}_1$	$\hat{V}_0$
$\rho$	$DG_0 \times DG_0$ 	$DG_1 \times DG_1$ 	$CG_1 \times CG_1$ 	$DG_0 \times DG_0$ 
$\mathbf{v}$	$RT_0$ 	Vector $DG_1 \times DG_1$ 	Vector $CG_1 \times CG_1$ 	Broken $RT_0$ 
$\theta$	$DG_0 \times CG_1$ 	$DG_1 \times DG_1$ 	$CG_1 \times CG_1$ 	$DG_0 \times DG_1$ 

**Table 1:** An example set of spaces  $\{V_0, V_1, \tilde{V}_1, \hat{V}_0\}$  to be used in the recovered advection scheme, for a vertical slice model on quadrilateral elements. These spaces are given for the density  $\rho$ , velocity  $\mathbf{v}$  and potential temperature  $\theta$ .  $V_0$  represents the original space of the function whilst  $V_1$  is the space in which the advection takes place.  $\tilde{V}_1$  the fully-continuous space of same degree as  $V_1$  and  $\hat{V}_0$  is the fully-discontinuous version of  $V_0$ . The diagrams represent the locations of the degrees of freedom for each of these spaces. The space  $RT_k$  is the  $k$ -th Raviart-Thomas space, while  $V_h \times V_v$  represents the tensor product of the horizontal space  $V_h$  with the vertical space  $V_v$ .

## 2.3 The Recovery Operator

Here we discuss the details of the recovery operator that we will use, which is very similar to the ‘weighted averaging’ operator used in [12]. Our recovery operator reconstructs  $\rho_0 \in V_0$  into  $\tilde{V}_1$  using the following procedure. Let  $i$  be a degree of freedom in the space  $\tilde{V}_1$ . The value of the field in  $\tilde{V}_1$  at  $i$  is determined to be the value of  $\rho_0$  at the location of  $i$ . However as  $\tilde{V}_1$  is continuous,  $i$  may be shared between a set of multiple elements  $\{e_i\}$ . In this case the value in  $\tilde{V}_1$  is the average of the values over  $\{e_i\}$ . Such an operator is found in [12] to possess second-order convergence in the  $L^2$ -norm when  $V_0$  is the discontinuous constant space  $DG_0 \times DG_0$  and  $\tilde{V}_1$  is  $CG_1 \times CG_1$ , the space of continuous linear functions over cells. These spaces correspond to those listed in Section 2.2 that we will use for our advection schemes. **This operator is intended only for use with fields on flat meshes, and must be extended for transport of vector fields on curved meshes.**

However, this operation for  $DG_0 \times DG_0$  to  $CG_1 \times CG_1$  does not have second-order convergence when representing fields with non-zero gradient at the boundaries of the domain. Whilst the second-order convergence holds for the interior of the domain, at the boundaries degrees of freedom are shared by fewer elements and may not necessarily accurately represent the gradient. We therefore extend the recovery operation at the boundaries by finding the field  $\rho_1 \in \hat{V}_1$  that minimises the curvature  $\int_e |\nabla \rho_1|^2 dx$  over a given element  $e$  subject to the constraints:

1.  $\rho_1 = \rho_r$  at the interior degrees of freedom of the element  $e$ , where  $\rho_r \in \tilde{V}_1$  is the first recovered field.

2.  $\int_e \rho_1 dx = \int_e \rho_0 dx$  for the element  $e$ , where  $\rho_0 \in V_0$  is the original field.

Once this  $\rho_1$  has been found, the final recovered field  $\rho_{\mathcal{R}} \in \tilde{V}_1$  is given by applying the original ‘averaging’ recovery operator again to  $\rho_1$ .

Finally, we will now show that the specific recovery operator that we have defined in this section satisfied Assumption 1, i.e. for all  $\rho_0 \in V_0$  there is some  $C > 0$  such that  $\|\mathcal{R}\rho_0\| \leq C\|\rho_0\|$ , where we take  $V_0 = \text{DG}_0 \times \text{DG}_0$  and  $\tilde{V}_1 = \text{CG}_1 \times \text{CG}_1$ . This property is used in Section 3.1.

**Theorem 1.** *Consider the action of the specific recovery operator  $\mathcal{R} : V_0 \rightarrow \tilde{V}_1$  defined in Section 2.3 upon a field  $\rho_0 \in V_0$  for  $V_0 = \text{DG}_0 \times \text{DG}_0$  and  $\tilde{V}_1 = \text{CG}_1 \times \text{CG}_1$ . There is some  $C > 0$  such that  $\|\mathcal{R}\rho_0\| \leq C\|\rho_0\|$  for all  $\rho_0 \in V_0$ , where  $\|\cdot\|$  denotes the  $L^2$  norm.*

*Proof.* We begin by defining  $\rho_{\mathcal{R}} := \mathcal{R}\rho_0$ . We consider the  $L^2$  norm of  $\rho_{\mathcal{R}}$  over an individual element  $e_i$ ,  $\|\rho_{\mathcal{R}}\|_{e_i}^2$ . The element  $e_i$  and cells that it shares DOFs with form a patch  $\mathcal{P}_i$ , and we introduce a coordinate scaling  $\mathbf{x} \rightarrow \mathbf{x}/h$  under which  $e_i$  becomes  $e'_i$ , which has unit area. If  $\{\phi_j(\mathbf{x})\}$  are the  $M$  basis functions spanning  $\tilde{V}_1$ , then  $\rho_{\mathcal{R}}$  can be written as

$$\rho_{\mathcal{R}} = \sum_j^M \rho_{\mathcal{R},j} \phi_j(\mathbf{x}), \quad (2.6)$$

where  $\rho_{\mathcal{R},j}$  is the value of  $\rho_{\mathcal{R}}$  at the  $j$ -th DOF. We now consider  $\|\rho_{\mathcal{R}}\|_{e'_i}^2$

$$\|\rho_{\mathcal{R}}\|_{e'_i}^2 = \int_{e'_i} \sum_j^M \sum_k^M \rho_{\mathcal{R},j} \rho_{\mathcal{R},k} \phi_j(h\mathbf{x}) \phi_k(h\mathbf{x}) dx \equiv \|\boldsymbol{\rho}_{\mathcal{R}}\|_{\Phi', e'_i}^2, \quad (2.7)$$

where  $\boldsymbol{\rho}_{\mathcal{R}}$  is the vector of values of  $\rho_{\mathcal{R}}$  at DOFs and  $\|\cdot\|_{\Phi'} \equiv \|\mathbf{y}^T \Phi' \mathbf{y}\|$  denotes the norm with the mass matrix  $\Phi' := \int \phi_j(h\mathbf{x}) \phi_k(h\mathbf{x}) dx$  acting upon some  $M$ -dimensional vector  $\mathbf{y}$ . From norm equivalence, we know that for some  $C > 0$ ,  $\boldsymbol{\rho}_{\mathcal{R}}$  evaluated in the  $\Phi'$  norm can be bounded from above by evaluation with the vector norm, and we can write this as the sum of components

$$\|\boldsymbol{\rho}_{\mathcal{R}}\|_{e'_i}^2 \leq C_i \sum_{j \in e_i} (\rho_{\mathcal{R},j})^2, \quad (2.8)$$

where  $C_i$  is a constant that depends upon the size of the element. If the  $j$ -th DOF in  $\tilde{V}_1$  is shared between  $N_j$  cells, then  $\rho_{\mathcal{R},j}$  is the average values of  $\rho_0$  in those cells, and hence  $\rho_{\mathcal{R},j} = \frac{1}{N_j} \sum_k^{N_j} \rho_{0,k}$ , giving

$$\begin{aligned} \|\boldsymbol{\rho}_{\mathcal{R}}\|_{e'_i}^2 &\leq C_i \sum_{j \in e'_i}^M \left( \frac{1}{N_j} \sum_k^{N_j} \rho_{0,k} \right)^2 \leq C_i \sum_{j \in e'_i}^M \left( \sum_k^{N_j} \rho_{0,k} \right)^2 \\ &\leq C_i \sum_{j \in e'_i}^M \sum_k^{N_j} (\rho_{0,k})^2 \leq C_i \|\rho_0\|_{\mathcal{P}'_i}^2, \end{aligned} \quad (2.9)$$

where the equalities follow as  $N_j$  is a positive integer, from the Cauchy-Schwarz inequality and from the definition of the  $L^2$  norm in  $V_0$ . The constant  $C_i$  has absorbed the effect of double-counting of cells over the patch. Under some regularity assumptions about the shape of the mesh,

$$\|\boldsymbol{\rho}_{\mathcal{R}}\|_{e_i}^2 \leq C^* \|\rho_0\|_{\mathcal{P}_i}^2 \quad (2.10)$$

where  $C^*$  is the maximum of  $C_i$  over the mesh. Now, considering the  $L^2$  norm over the whole domain,

$$\|\rho_{\mathcal{R}}\|^2 \leq \sum_i \|\boldsymbol{\rho}_{\mathcal{R}}\|_{e_i}^2 \leq \sum_i C_i \|\rho_0\|_{\mathcal{P}_i}^2 \leq C \|\rho_0\|^2, \quad (2.11)$$

where the constant again takes double-counting into account. This also holds for the procedure at the boundaries, where the values of the reconstructed field are a linear function of the interior and original field values. Thus we arrive at the conclusion that for some  $C > 0$ ,

$$\|\rho_{\mathcal{R}}\| \leq C\|\rho_0\|. \quad (2.12)$$

□

## 2.4 Limiting

In numerical weather or climate models, there may be many additional prognostic variables representing moisture or chemical species. These variables will typically lie in **either the same space as the density  $\rho$  or the potential temperature  $\theta$** . In this paper we will consider only moisture variables, which will lie in the space of  $\theta$ , which can simplify the thermodynamics associated with phase changes. However it may come at the cost of sacrificing conservation of the mass of water, although this could be remedied by solving the transport equation in ‘conservative’ form (2.2) for  $r\rho$  rather than in ‘advective’ form for tracer  $r$ .

The continuous equations describing the advection of these **tracer** variables have monotonicity and shape-preserving properties; however the discrete representation may not replicate these properties, which may lead to unphysical solutions such as negative concentrations. This can be avoided by the application of slope limiters.

In the recovered scheme, both the advection operator  $\mathcal{A}$  and final projection operator  $\mathcal{P}$  may produce spurious overshoots and undershoots, and so need limiting. In the case of the projection operator, we do this by using the second projection operator  $\mathcal{P}_B$ . **This prevents the formation of new maxima and minima as it is composed of two bounded operations: the projection into the broken lower-order space and then the recovery of continuity.** For the set of spaces proposed in Section 2.2, to limit the advection operator  $\mathcal{A}$ , we use the vertex-based limiter outlined in [16]. **This limiter divides the field in each element into a constant mean part and a linear perturbation. Considering the values of neighbouring elements at shared vertices gives upper and lower bounds. The limited field is then the mean part plus a constant times the perturbation, so that the field remains bounded.** This limiter is applied to the field before the advection operator begins, and after each stage of  $\mathcal{A}$ .

## 3 Properties of the Numerical Scheme

### 3.1 Stability

Here we will show the stability of the ‘recovered space’ scheme, following a similar argument to that used in [9]. First, we will need the following Lemma.

**Lemma 1.** *Let the operator  $\mathcal{J} : V_0 \rightarrow V_1$  be defined by*

$$\mathcal{J} := \mathcal{I}(\mathcal{R} - \hat{\mathcal{P}}\mathcal{R} + 1), \quad (3.1)$$

*so that the ‘recovered space’ scheme can be written as*

$$\rho_0^{n+1} = \mathcal{P}\mathcal{A}\mathcal{J}\rho_0^n. \quad (3.2)$$

*Denote by  $\|\cdot\|$  the  $L^2$  norm. Then  $\|\mathcal{J}\rho_0\| \leq \kappa\|\rho_0\|$  for some  $\kappa > 0$  for all  $\rho_0 \in V_0$ .*

*Proof.* From the definition of  $\mathcal{J}$ ,

$$\|\mathcal{J}\rho_0\| = \|\rho_0 + \mathcal{R}\rho_0 - \hat{\mathcal{P}}\mathcal{R}\rho_0\|. \quad (3.3)$$

By applying the triangle inequality,

$$\|\mathcal{J}\rho_0\| \leq \|\rho_0\| + \|\mathcal{R}\rho_0\| + \|\hat{\mathcal{P}}\mathcal{R}\rho_0\|. \quad (3.4)$$

We will now inspect the  $\|\hat{\mathcal{P}}\mathcal{R}\rho_0\|$  term. The definition of  $\hat{\mathcal{P}}$  is that  $\int_{\Omega} \hat{\psi} \tilde{\rho} \, dx = \int_{\Omega} \hat{\psi} \hat{\mathcal{P}} \tilde{\rho} \, dx$  for all  $\hat{\psi} \in \hat{V}_0$ , where  $\tilde{\rho} \in \tilde{V}$ . Since  $\hat{\mathcal{P}} \tilde{\rho} \in \hat{V}_0$ , then it must be true that

$$\int_{\Omega} (\hat{\mathcal{P}} \tilde{\rho})^2 \, dx = \int_{\Omega} (\hat{\mathcal{P}} \tilde{\rho}) \tilde{\rho} \, dx. \quad (3.5)$$

Now we consider the integral  $\int_{\Omega} (\hat{\mathcal{P}} \tilde{\rho} - \tilde{\rho})^2 \, dx$ , which cannot be negative. Expanding this integral and using the result (3.5) gives  $\int_{\Omega} \tilde{\rho}^2 \, dx \geq \int_{\Omega} (\hat{\mathcal{P}} \tilde{\rho})^2 \, dx$ , in other words that  $\|\hat{\mathcal{P}}\mathcal{R}\rho_0\| \leq \|\mathcal{R}\rho_0\|$ . Hence, returning to considering  $\|\mathcal{J}\rho_0\|$ , we obtain

$$\|\mathcal{J}\rho_0\| \leq \|\rho_0\| + 2\|\mathcal{R}\rho_0\|. \quad (3.6)$$

Finally, we use the property of  $\mathcal{R}$  that  $\|\mathcal{R}\rho_0\| \leq C\|\rho_0\|$  for some  $C > 0$ , and so letting  $\kappa = 1 + 2C$  we arrive at

$$\|\mathcal{J}\rho_0\| \leq \kappa\|\rho_0\|. \quad (3.7)$$

This completes the proof.  $\square$

**Theorem 2.** *Let the advection operator  $\mathcal{A}$  have a stability constant  $\alpha$ , such that*

$$\|\mathcal{A}\| := \sup_{\rho_1 \in V_1, \|\rho_1\| > 0} \frac{\|\mathcal{A}\rho_1\|}{\|\rho_1\|} \leq \alpha. \quad (3.8)$$

*Then the stability constant  $\alpha^*$  of the ‘recovered space’ scheme on  $V_0$  satisfies  $\alpha^* = \kappa\alpha$  for some constant  $\kappa$ .*

*Proof.* Since from Lemma 1,  $\|\mathcal{J}\rho_0\| \leq \kappa\|\rho_0\|$  for some  $\kappa > 1$ ,

$$\sup_{\rho_0 \in V_0, \|\rho_0\| > 0} \frac{\|\mathcal{P}\mathcal{A}\mathcal{J}\rho_0\|}{\|\rho_0\|} \leq \sup_{\rho_0 \in V_0, \|\rho_0\| > 0} \kappa \frac{\|\mathcal{P}\mathcal{A}\mathcal{J}\rho_0\|}{\|\mathcal{J}\rho_0\|}. \quad (3.9)$$

As  $V_0 \subset V_1$ , the supremum over elements in  $V_1$  cannot be smaller than the supremum over elements in  $V_0$ . Recognising that  $\mathcal{J}\rho_0 \in V_1$ ,

$$\sup_{\rho_0 \in V_0, \|\rho_0\| > 0} \frac{\|\mathcal{P}\mathcal{A}\mathcal{J}\rho_0\|}{\|\mathcal{J}\rho_0\|} \leq \sup_{\rho_1 \in V_1, \|\rho_1\| > 0} \frac{\|\mathcal{P}\mathcal{A}\rho_1\|}{\|\rho_1\|}. \quad (3.10)$$

For the final step, we must consider both cases  $\mathcal{P}_A$  and  $\mathcal{P}_B$  for the projection operator. In the case of  $\mathcal{P}_A$ , we can use a similar argument to that of the projection operator in Lemma 1 to obtain that  $\|\mathcal{P}\mathcal{A}\rho_1\| \leq \|\mathcal{A}\rho_1\|$ . For  $\mathcal{P}_B = \mathcal{P}_R\mathcal{P}_I$ , each step maps a function into a space that is smaller; i.e.  $\hat{V}_0 \subset V_1$  and  $V_0 \subset \hat{V}_0$ , so that the supremum of  $\|\mathcal{P}\mathcal{A}\rho_1\|$  must be smaller than the supremum of  $\|\mathcal{A}\rho_1\|$ . In both cases we obtain that

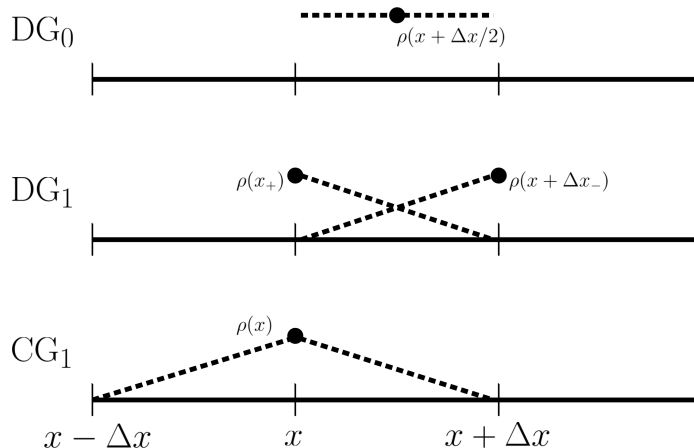
$$\sup_{\rho_1 \in V_1, \|\rho_1\| > 0} \frac{\|\mathcal{P}\mathcal{A}\rho_1\|}{\|\rho_1\|} \leq \sup_{\rho_1 \in V_1, \|\rho_1\| > 0} \frac{\|\mathcal{A}\rho_1\|}{\|\rho_1\|} \leq \alpha, \quad (3.11)$$

where the final inequality defines the stability of  $\mathcal{A}$ . Combining these arguments together gives

$$\sup_{\rho_0 \in V_0, \|\rho_0\| > 0} \frac{\|\mathcal{P}\mathcal{A}\mathcal{J}\rho_0\|}{\|\rho_0\|} \leq \kappa\alpha, \quad (3.12)$$

and so the stability constant of the ‘recovered space’ scheme is  $\alpha^* = \kappa\alpha$ .  $\square$





**Figure 1:** The degrees of freedom and basis functions for each of the spaces used in the Von Neumann analysis.

## 3.2 Von Neumann Analysis

Now we will attempt to identify the stability constant for three one-dimensional examples, by performing Von Neumann stability analysis. This can also be used with the Courant-Friedrich-Lewy (CFL) condition to give upper limits to stable Courant numbers.

The three cases that will be considered are that of  $V_0 = \text{DG}_0$  (which might represent the advection of  $\rho$ ), and the two cases of  $V_0 = \text{CG}_1$  with  $\mathcal{P}_A$  and  $\mathcal{P}_B$  as the projection operators (for advection of velocity and moisture respectively). The same advection operator **discretising the advective form (2.1) of the transport equation** will be used for all three cases, with the advection taking place in  $V_1 = \text{DG}_1$ .

In each case we consider a periodic domain of length  $L$ , divided into  $N$  cells, each of length  $\Delta x$ . We will make the assumption that our function  $\rho_n(x)$  at the  $n$ -th time step can be written as a sum of Fourier modes,

$$\rho^n(x) = \sum_k A_k^n e^{ikx}. \quad (3.13)$$

Then for the  $k$ -th mode  $\rho_{n,k}(x + \Delta x) = \rho_{n,k}(x)e^{ik\Delta x}$ .

Three spaces are relevant to this analysis:  $\text{DG}_0$ , which is piecewise constant and whose DOFs are in the centre of cells;  $\text{CG}_1$ , which is continuous piecewise linear and has one DOF per cell at the cell boundary; and  $\text{DG}_1$ , which is linear within a cell but discontinuous between cells and has two DOFs per cell – one at each cell boundary. These spaces are shown in Figure 1.

### 3.2.1 Advection Operator

First we described the advection operator  $\mathcal{A}_k$  acting upon the  $k$ -th mode of a function in  $\text{DG}_1$ . In each cell, the function can be described by two components: evaluation of the field at each cell boundary. For the advection, we use a simple upwinding scheme with a forward Euler time discretisation, within the framework of a three-step Runge-Kutta scheme. We describe the action of a single forward Euler step with the operator  $\mathcal{L}_k$ . This is determined by discretising the one dimensional advection equation with

constant  $u > 0$ ,

$$\frac{\partial q}{\partial t} + u \frac{\partial q}{\partial x} = 0, \quad (3.14)$$

for  $q \in V_1$  by integrating with the test function  $\psi \in V_1$ . This gives

$$\int_0^L \psi q_{n+1} dx = \int_0^L \psi q_n dx + u \Delta t \int_0^L \frac{\partial \psi}{\partial x} q_n dx - u \Delta t \sum_j \llbracket \psi q_n \rrbracket_j, \quad (3.15)$$

where  $\llbracket \cdot \rrbracket_j$  denotes the jump in field between the  $j$ -th cell and the  $(j+1)$ -th cell. Making the assumption that  $q(x) = q(x + \Delta x)e^{-ik\Delta x}$ , and using that  $q$  is piecewise linear, we can write down a representation of  $\mathcal{L}_k$  for the degrees of freedom on either side of a given cell:

$$\mathcal{L}_k = \begin{pmatrix} 1 - 3c & 4ce^{-ik\Delta x} - c \\ 3c & 1 - c - 2ce^{-ik\Delta x} \end{pmatrix}. \quad (3.16)$$

We then obtain the full advection operator by using the three-step Runge-Kutta scheme outlined in [17]:

$$q^{(1)} := q_{n,k} + \mathcal{L}_k q_{n,k}, \quad (3.17)$$

$$q^{(2)} := \frac{3}{4}q_{n,k} + \frac{1}{4}q^{(1)} + \frac{1}{4}\mathcal{L}_k q^{(1)}, \quad (3.18)$$

$$q_{n+1,k} = \frac{1}{3}q_{n,k} + \frac{2}{3}q^{(2)} + \frac{2}{3}\mathcal{L}_k q^{(2)}. \quad (3.19)$$

The overall advection operator is then

$$\mathcal{A}_k = \mathbb{1} + \mathcal{L}_k + \frac{1}{2}\mathcal{L}_k^2 + \frac{1}{6}\mathcal{L}_k^3, \quad (3.20)$$

where  $\mathbb{1}$  is the identity operator. We omit the matrix representation of  $\mathcal{A}_k$  here for brevity.

### 3.2.2 Case A: DG<sub>0</sub>

This represents the advection of density  $\rho$ , or the velocity  $\mathbf{v}$  perpendicular to its direction. The set of spaces  $\{V_0, V_1, \tilde{V}_1, \hat{V}_0\}$  is  $\{\text{DG}_0, \text{DG}_1, \text{CG}_1, \text{DG}_0\}$ .

Then, for a given cell and Fourier mode, the operators can be represented in the following matrix forms

$$\hat{\mathcal{P}} = \mathcal{P} = \frac{1}{2} \begin{pmatrix} 1 & 1 \end{pmatrix}, \quad \mathcal{R}_k = \frac{1}{2} \begin{pmatrix} 1 + e^{-ik\Delta x} \\ e^{ik\Delta x} + 1 \end{pmatrix}, \quad \mathcal{I} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad (3.21)$$

Combining these operators, the advection scheme for the  $k$ -th mode of  $\rho_n(x)$  is then expressed as

$$\rho_{n+1,k} = \frac{1}{4} \begin{pmatrix} 1 & 1 \end{pmatrix} \mathcal{A}_k \begin{pmatrix} 2 - i \sin(k\Delta x) \\ 2 + i \sin(k\Delta x) \end{pmatrix} \rho_{n,k}. \quad (3.22)$$

Following the analysis through and writing  $\phi = k\Delta x$  gives a stability constant

$$\begin{aligned} |A_k|^2 &= c^2 \left[ c^3 \left( \frac{13}{4} \cos \phi - \frac{5}{3} \cos 2\phi + \frac{1}{12} \cos 3\phi + \frac{1}{6} \cos 4\phi - \frac{11}{6} \right) \right. \\ &+ c^2 \left( \frac{29}{12} \sin \phi - \frac{5}{3} \sin 2\phi + \frac{1}{12} \sin 3\phi + \frac{1}{6} \sin 4\phi - \frac{7}{4} \cos \phi + \frac{3}{4} \cos 2\phi - \frac{1}{4} \cos 3\phi + \frac{5}{4} \right) \\ &\left. + c \left( -\frac{3}{4} \sin \phi + \frac{3}{4} \sin 2\phi - \frac{1}{4} \sin 3\phi - \cos \phi + \frac{1}{4} \cos 2\phi + \frac{3}{4} \right) - \frac{3}{2} \sin \phi + \frac{1}{4} \sin^2 2\phi - 1 \right]^2 \end{aligned} \quad (3.23)$$

### 3.2.3 Case B: CG<sub>1</sub> with $\mathcal{P} = \mathcal{P}_A$

In this case the set of spaces  $\{V_0, V_1, \tilde{V}_1, \hat{V}_0\}$  is  $\{\text{CG}_1, \text{DG}_1, \text{CG}_1, \text{DG}_1\}$ . This describes advection of velocity parallel to its direction, or of potential temperature without bounding the final projection step. The operators can be represented by

$$\mathcal{P}_A = \frac{1}{4 + 2 \cos k \Delta x} (2 + e^{-ik\Delta x} \quad 1 + 2e^{-ik\Delta x}), \quad \hat{\mathcal{P}} = \mathcal{I} = \begin{pmatrix} 1 \\ e^{ik\Delta x} \end{pmatrix}, \quad (3.24)$$

where the projection operator  $\mathcal{P}_A$  has been determined by solving equation (2.4). The recovery operator is the identity, and since the injection  $\mathcal{I}$  and the projection  $\hat{\mathcal{P}}$  are equivalent in this case the scheme acting upon  $v_n$  becomes  $v_{n+1} = \mathcal{P}_A \mathcal{A} \mathcal{I} v_n$ . Following through the analysis gives

$$\begin{aligned} |A_k|^2 &= \left( \frac{c}{\cos \phi + 2} \right)^2 (c^2 \cos \phi - c^2 + 3)^2 \sin^2 \phi \\ &+ \left( \frac{c}{\cos \phi + 2} \right)^2 \left( -2c^3 \cos \phi + \frac{1}{2}c^3 \cos 2\phi + \frac{3}{2}c^3 + 3c^2 \cos \phi - 3c^2 + \cos \phi + 2 \right)^2 \end{aligned} \quad (3.25)$$

### 3.2.4 Case C: CG<sub>1</sub> with $\mathcal{P} = \mathcal{P}_B$

For this case, the set of spaces are the same as in the second case. The only difference is that the projection operator  $\mathcal{P}$  is now  $\mathcal{P}_B = \mathcal{P}_R \mathcal{P}_I$ . As  $V_1 = \hat{V}_0 = \text{DG}_1$ , the interpolation  $\mathcal{P}_I$  is the identity, and  $\mathcal{P}_B = \mathcal{P}_R$ . The operators are

$$\mathcal{P}_B = \frac{1}{2} (1 \quad e^{-ik\Delta x}), \quad \hat{\mathcal{P}} = \mathcal{I} = \begin{pmatrix} 1 \\ e^{ik\Delta x} \end{pmatrix}, \quad (3.26)$$

which gives leads to the amplification factor

$$\begin{aligned} |A_k|^2 &= \left( \frac{2}{3}c^3 \sin \phi + \frac{1}{6}c^3 \sin 2\phi - \frac{1}{3}c^3 \sin 3\phi - c^2 \sin \phi + \frac{1}{2}c^2 \sin 2\phi - c \sin \phi \right)^2 \\ &+ \left( -\frac{7}{3}c^3 \cos \phi - \frac{1}{6}c^3 \cos 2\phi + \frac{1}{3}c^3 \cos 3\phi + \frac{13}{6}c^3 + 3c^2 \cos \phi - \frac{1}{2}c^2 \cos 2\phi - \frac{5}{2}c^2 + 1 \right)^2 \end{aligned} \quad (3.27)$$

## 3.3 Critical Courant Numbers

The Courant-Friedrich-Lewy (CFL) criterion says that an advection scheme with amplification factor  $|A_k| > 1$  may not be stable. The critical Courant number  $c^*$  is the lowest Courant number  $c = u\Delta t/\Delta x$  such that the amplification factor is greater than unity. We numerically measured the critical Courant numbers for the three cases laid out in Section 3.2, and these are displayed in Table 2. Although case C has a significantly lower critical Courant number, the intention is to run this scheme with a limiter or with a subcycling time discretisation, allowing it to be used at higher Courant numbers. Instances of work using these kind of limiters are [9] and [16].

Examples of critical Courant numbers for other upwinding schemes can be found in Table 2.2 of [18]. The most relevant comparison that can be made from this is to that of polynomials of degree 1 with a Runge-Kutta method of order 3, which has a critical Courant number of 0.409. A space of discontinuous linear polynomials has the same number of degrees of freedom as a space of discontinuous constants but with half the grid size, and thus improvements are made if the critical Courant number is more than twice that of the transport scheme for the linear functions. We do therefore observe that the critical Courant numbers for cases A and B are improvements on the discontinuous upwinding scheme applied just to discontinuous linear functions.

Case	A	B	C
$c^*$	0.8506	0.9930	0.3625

**Table 2:** The critical Courant numbers for the three cases of the advection scheme outlined in Section 3.2.

## 4 Numerical Tests

The numerical implementation of this scheme was performed using the Firedrake software of [19], and relied heavily on the tensor product element functionality on extruded meshes, descriptions of which can be found in [11], [20] and [21].

### 4.1 Numerical Accuracy

To verify the numerical accuracy of the scheme, we performed a series of convergence tests. The aim is to find how the error due to advection changes with the grid spacing  $\Delta x$ . We used tests that have an analytic solution in the limit that  $\Delta x \rightarrow 0$ , and compare the final advected profile  $q$  with the ‘true’ profile  $q_h$ , which is the analytic solution projected into the relevant function space. This gives an error  $\|q - q_h\|$  (where  $\|\cdot\|$  denotes the  $L^2$  norm) which is calculated for the same problem at different resolutions, and the errors are plotted as a function of the grid spacing  $\Delta x$ . The order of the numerical accuracy is the number  $n$  such that  $\|q - q_h\| \sim \mathcal{O}(\Delta x^n)$ , which can be measured from the **slope** of a plot of  $\ln(\|q - q_h\|)$  against  $\ln(\Delta x)$ . For simplicity, the tests we used are designed so that the ‘true’ profile is the same as the initial condition.

The initial conditions were obtained by pointwise evaluation of the expressions into higher order fields (we used CG<sub>3</sub>). These were then projected into the correct fields. The advecting velocity used lay in the RT<sub>1</sub> space. To mimic how the scheme might be used in a numerical weather model, we performed some of the tests on the different sets of spaces laid out in Section 2.2 and the configurations described in Section 3.2. Each set of spaces is labelled by the variable name in Table 1, with  $\{V_0, V_1, \tilde{V}_1, \hat{V}_0\}$  for the fields  $\rho, \mathbf{v}, \theta$  and  $r$  given by

$$\rho \in \{\mathbf{DG}_0 \times \mathbf{DG}_0, \mathbf{DG}_1 \times \mathbf{DG}_1, \mathbf{CG}_1 \times \mathbf{CG}_1, \mathbf{DG}_0 \times \mathbf{DG}_0\}, \quad (4.1)$$

$$\mathbf{v} \in \{\mathbf{RT}_1, \mathbf{DG}_1 \times \mathbf{DG}_1, \mathbf{CG}_1 \times \mathbf{CG}_1, \text{broken RT}_1\}, \quad (4.2)$$

$$\theta \in \{\mathbf{DG}_0 \times \mathbf{CG}_1, \mathbf{DG}_1 \times \mathbf{DG}_1, \mathbf{CG}_1 \times \mathbf{CG}_1, \mathbf{DG}_0 \times \mathbf{DG}_1\}, \quad (4.3)$$

$$r \in \{\mathbf{DG}_0 \times \mathbf{CG}_1, \mathbf{DG}_1 \times \mathbf{DG}_1, \mathbf{CG}_1 \times \mathbf{CG}_1, \mathbf{DG}_0 \times \mathbf{DG}_1\}, \quad (4.4)$$

where the bold font recognises that the space has vector valued nodes. While the scheme labelled  $\theta$  uses the projection operator  $\mathcal{P}_A$ , the scheme labelled  $r$  represents a moisture variable, so uses the same spaces as  $\theta$  but the projection operator  $\mathcal{P}_B$  and the vertex-based limiter of [16]. **All tests solve the transport equation in the ‘advective’ form of (2.1).**

The first three tests involve advection around a 2D domain representing a vertical slice that is periodic along its side edges but rigid walls at the top and bottom. The final test is performed over the surface of a sphere. All the vertical slice tests use a domain of height and width 1 m and advect the profile with time steps of  $\Delta t = 10^{-4}$  s for a total time of  $T = 1$  s.

#### 4.1.1 Rotational Convergence Test

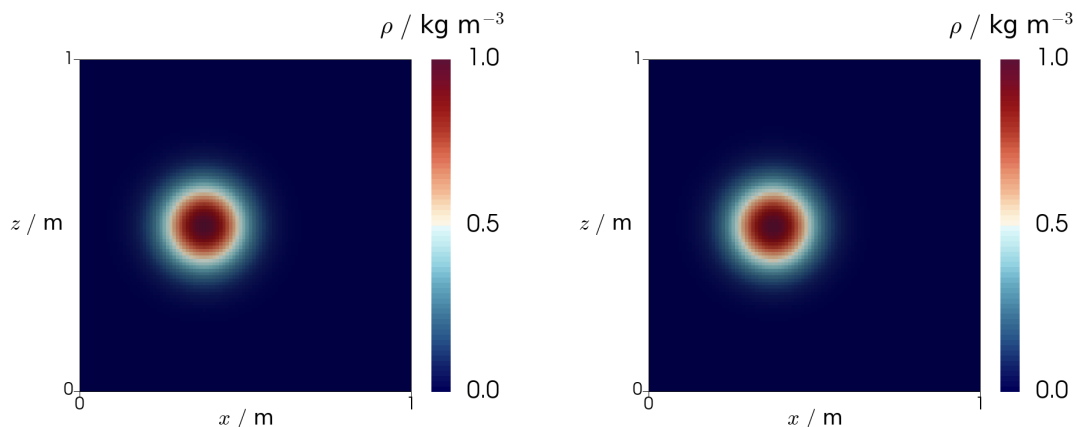
The first test involves a rigid body rotation of a Gaussian profile around the centre of the domain. Using  $x$  and  $z$  as the horizontal and vertical coordinates, defining  $r^2 = (x - x_0)^2 + (z - z_0)^2$  for  $x_0 = 0.375$  m,  $z_0 = 0.5$  m and using  $r_0 = 1/8$  m, the initial condition used for all fields was

$$q = e^{-(r/r_0)^2}. \quad (4.5)$$

For the velocity variable, this initial profile was used for each component of the field. The advecting velocity is generated from a stream function  $\psi$  via  $\mathbf{v} = (-\partial_z \psi, \partial_x \psi)$ . Defining  $r_v^2 = (x - x_v)^2 + (z - z_v)^2$  with  $x_v = 0.5$  m and  $z_v = 0.5$  m, the stream function used was:

$$\psi(\mathbf{x}) = \begin{cases} \pi(r_v^2 - 0.5), & r_v < r_1, \\ Ar_v^2 + Br_v + C, & r_1 \leq r_v < r_2, \\ Ar_2^2 + Br_2 + C, & r_v \geq r_2. \end{cases} \quad (4.6)$$

This is designed to be a rigid body rotation for  $r_v < r_1$ , with no velocity for  $r_v \geq r_2$  to prevent spurious noise being generated from the edge of the domain. The stream function and its derivative vary smoothly for  $r_1 \leq r_v < r_2$ . We use  $r_1 = 0.48$  m and  $r_2 = 0.5$  m, with  $A = \pi r_1 / (r_1 - r_2)$ ,  $B = -2Ar_2$  and  $C = \pi(r_1^2 - 0.5) - Ar_1^2 - Br_1$ . Results showing second order numerical accuracy can be found in Figure 4 (left). Initial and final fields for the density in the lowest resolution run ( $\Delta x = 0.01$  m) are displayed in Figure 2.



**Figure 2:** The initial (left) and final (right) fields in the  $DG_0 \times DG_0$  space from the rotational convergence test of Section 4.1.1, showing the field labelled  $\rho$  in Figure 4 (left) at the lowest resolution ( $\Delta x = 0.01$  m). Almost no difference is visible between the fields.

#### 4.1.2 Deformational Convergence Test

The second test is a more challenging convergence test, based on the deformational flow experiment described in [9]. The initial profiles were the same as used in the rotational advection test, but with  $x_0 = 0.5$  m. The advecting velocity was that of [9]:

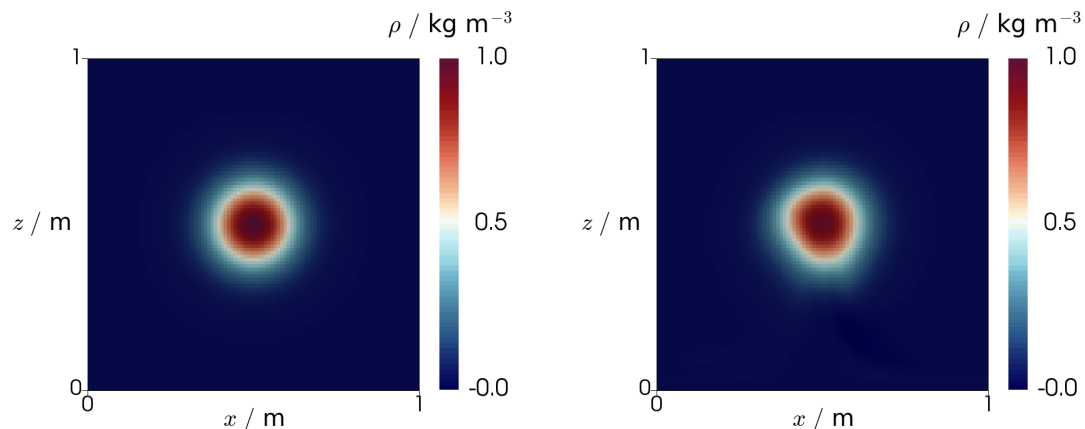
$$\mathbf{v}(\mathbf{x}, t) = \begin{pmatrix} 1 - 5(0.5 - t) \sin(2\pi(x - t)) \cos(\pi z) \\ 5(0.5 - t) \cos(2\pi(x - t)) \sin(\pi z) \end{pmatrix}. \quad (4.7)$$

Figure 4 (right) plots the results of this test, with each variable measuring second order numerical accuracy. Initial and final fields for the density in the lowest resolution run ( $\Delta x = 0.01$  m) are displayed in Figure 3.

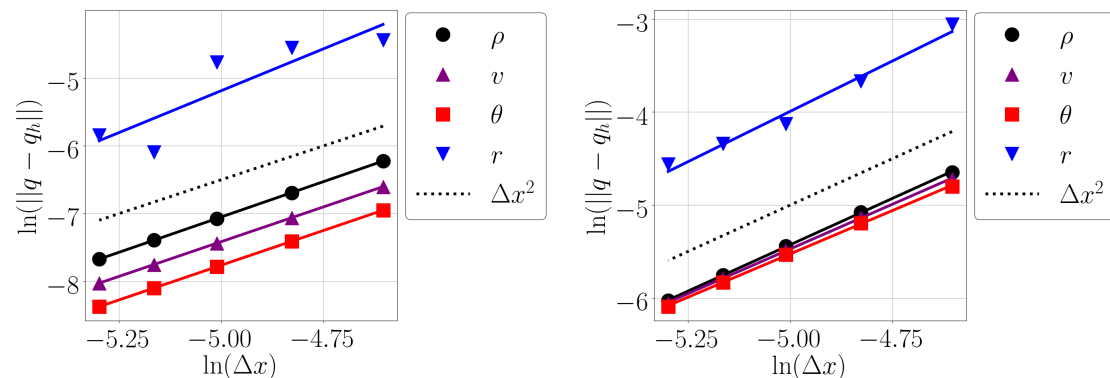
#### 4.1.3 Boundary Convergence Test

The third test was intended to investigate the integrity of the advection scheme at the boundaries of the domain. We use the following reversible flow, which squashes the advected material into the boundary before recovering it:

$$\mathbf{v}(\mathbf{x}, t) = \begin{cases} (1, -\sin(2\pi z)) & t < 0.5 \\ (1, \sin(2\pi z)) & t \geq 0.5 \end{cases}. \quad (4.8)$$



**Figure 3:** The initial (left) and final (right) fields in the  $DG_0 \times DG_0$  space from the deformational convergence test of Section 4.1.2, showing the field labelled  $\rho$  in Figure 4 (right) at the lowest resolution ( $\Delta x = 0.01$  m).

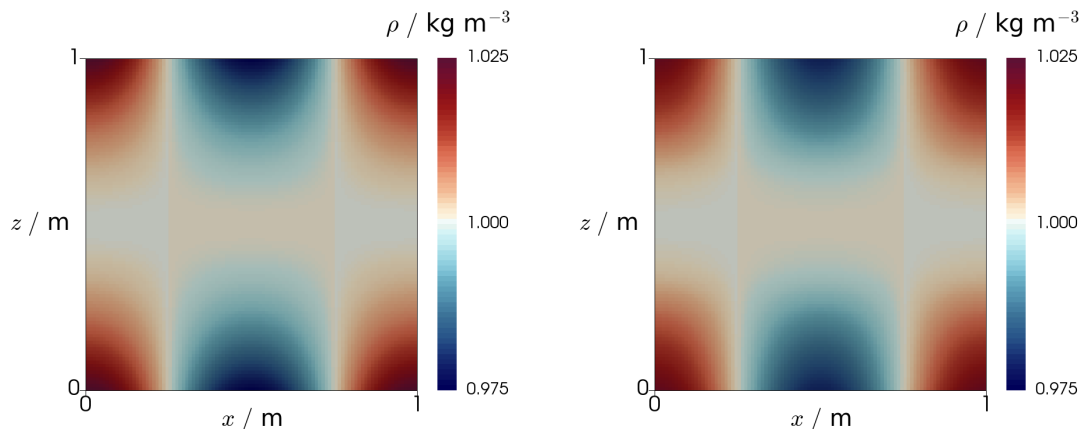


**Figure 4:** Results from convergence tests for the recovered space scheme, plotting, as a function of grid spacing  $\Delta x$ , the error in an advected solution  $q$  against the true solution  $q_h$ . The different lines labelled  $\rho$ ,  $v$  and  $\theta$  represent performing the test in each of the different sets of spaces laid out in Section 2.2 and using the projection operator  $\mathcal{P}_A$ , whilst the line labelled  $r$  represents advection with the same spaces as  $\theta$ , but with the projection operator  $\mathcal{P}_B$ . (Left) The test represents a rigid body rotation. In all cases, the slopes are around 2, indicating second order numerical accuracy. (Right) A more difficult convergence test featuring deformational flow. Accuracy is approaching second order for each case.

The initial condition was

$$q = 1 + \frac{1}{10} \left( z - \frac{1}{2} \right)^2 \cos(2\pi x). \quad (4.9)$$

To see the effect of the extra recovery performed at the boundary described in Section 2.3, this extra recovery was turned off for the variables labelled with an asterisk. The results in Figure 7 (left) demonstrate that without doing extra recovery at the boundaries, the whole recovery process does not have second order numerical accuracy. Initial and final fields for the density in the lowest resolution run ( $\Delta x = 0.01$  m) are displayed in Figure 5.



**Figure 5:** Initial (left) and final (right) fields in the  $DG_0 \times DG_0$  space of the boundary convergence test of Section 4.1.3, showing the field labelled  $\rho$  in Figure 7 (left) at the lowest resolution ( $\Delta x = 0.01$  m).

#### 4.1.4 Spherical Convergence Test

The final convergence test was performed on the surface of a sphere. In this case we used a cubed sphere mesh of a sphere of radius 100 m. The advecting velocity field used was  $\mathbf{v} = U \sin \lambda$ , for latitude  $\lambda$  and  $U = \pi/10$  m s<sup>-1</sup>, which gave a constant zonal rotation rate about the sphere. We took time steps of  $\Delta t = 0.5$  s up to a total time of 2000 s so that the initial profile should be equal to the ‘true’ profile. The initial profile that we used was very similar to that used in the first test case of [22]:

$$q = \begin{cases} \frac{1}{2} [1 + \cos(\frac{\pi r}{R})], & r < R, \\ 0, & \text{otherwise,} \end{cases} \quad (4.10)$$

where  $R = 100/3$  m and for latitude  $\lambda$  and longitude  $\varphi$  with  $\lambda_0 = 0$  and  $\varphi_0 = -\pi/2$ , and where  $r$  is now given by

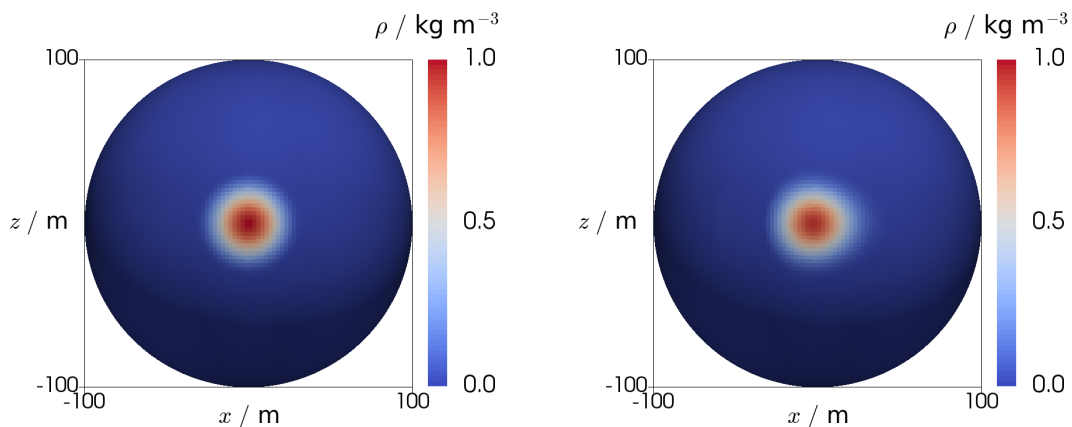
$$r = 100 \text{ m} \times \cos^{-1} [\sin \lambda_0 \sin \lambda + \cos \lambda_c \cos \lambda \cos(\varphi - \varphi_0)]. \quad (4.11)$$

The errors of this test as a function of resolution are plotted in Figure 7 (right). This also appears to show second order accuracy. We found that at lower resolutions, the errors due to the advective scheme were obscured by those from the imperfect discretisation of the surface of the sphere. [The initial and final fields of this test are plotted for the coarsest resolution in Figure 6.](#)

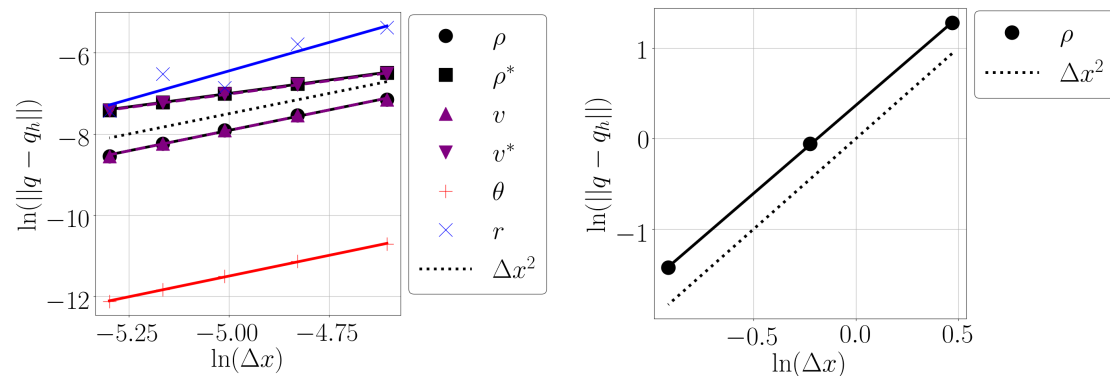
## 4.2 Stability

We tested the formulas (3.23), (3.25) and (3.27) by advecting sine and cosine waves for each of the cases defined. The domain used was a square vertical slice of length 120 m with grid spacing  $\Delta x = 1$  m. The amplification factor for a given wavenumber  $k$  and Courant number  $c$  was measured by advecting a sine and cosine wave of wavenumber  $k$  by a constant horizontal velocity  $c$  for a single time step of  $\Delta t = 1$  s. As before, the domain had periodic boundary conditions on the vertical walls. The amplification factor was then found by measuring the amplitude of the sine and cosine components after the first time step. This was done for several values of  $c$ .

The measured values are compared with those from the formula in Figure 8 which shows agreement for each of the cases considered in Section 3.2.



**Figure 6:** The initial (left) and final (right) fields in the  $DG_0 \times DG_0$  space from the spherical convergence test of Section 4.1.2, showing the field labelled  $\rho$  in Figure 7 (right) at the lowest resolution (the sixth refinement level of the cubed sphere, with  $\Delta x \approx 1.6$  m).

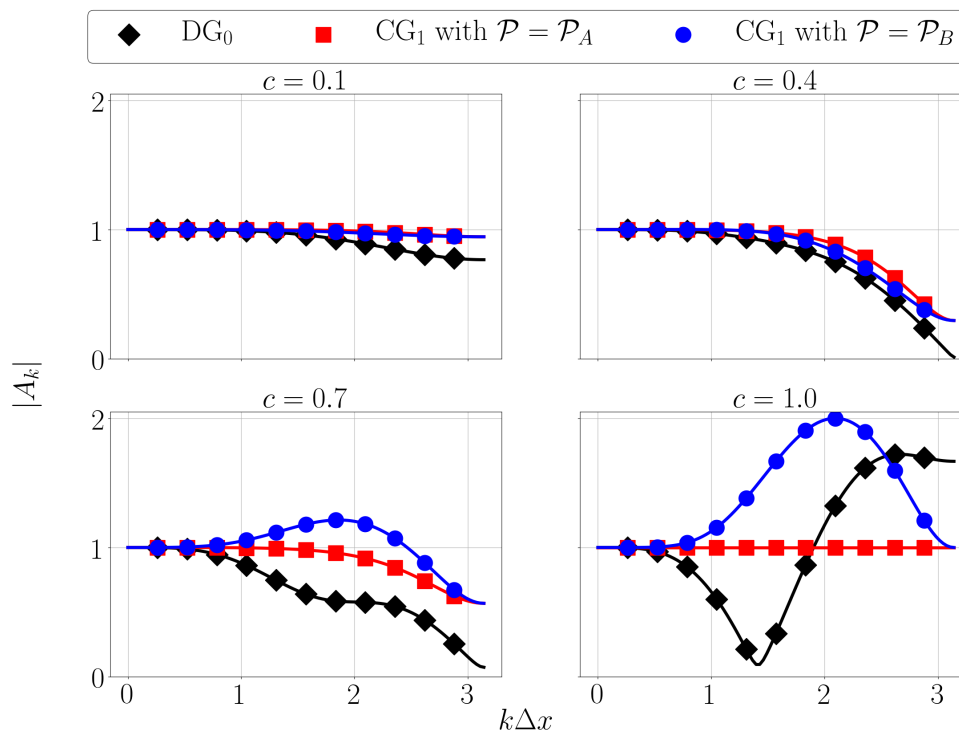


**Figure 7:** More results from convergence tests for the recovered space scheme, plotting, as a function of grid spacing  $\Delta x$ , the error in an advected solution  $q$  against the true solution  $q_h$ . (Left) A test demonstrating the need for the extra recovery at the boundaries, by comparing the scheme with and without this extra recovery process. The schemes without extra recovery at the boundaries are denoted with an asterisk. They not only display a larger error, but also show lower accuracy. As  $\theta$  and  $r$  are linear in the vertical, they are accurately represented at the boundary by the recovery scheme without performing any additional recovery at the boundary. However if rigid walls were present on the side of the domain, these fields would require additional recovery at these boundaries. (Right) The test performed on a cubed sphere mesh. The slope here is very close to 2, again supporting the claim that the advection scheme has second order numerical accuracy.

### 4.3 Limiting

The efficacy of the limiting scheme was tested by using the LeVeque slotted-cylinder, hump, cone set-up originally defined in [23] and used in both [16] and [9]. The advected field was initialised lying with this condition, lying in the  $DG_0 \times CG_1$  space to mimic moisture variables, before a solid-body rotation was completed. This was performed for the bounded case of the scheme defined in Section 2, using the projection operator  $\mathcal{P}_B$  and the vertex-based limiter of [16] for the advection. The resulting field is





**Figure 8:** The results from testing the validity of the expressions (3.23), (3.25) and (3.27) for the amplification factors in the 1D advection cases presented in Section 3.2. The markers denote measurements of the amplification factor by advecting sine and cosine wave profiles, whilst the lines plot the expressions derived in Section 3.2. All plots show agreement between the expressions and the measured amplification factors. *These results also agree with the critical Courant numbers found in Table 2, including for  $c = 0.4$  in which the values are all below unity for the  $\mathcal{P}_A$  case, but with some values above unity for the  $\mathcal{P}_B$  case.*

shown in Figure 9 where it is also compared to the rotation of a field in the  $DG_1 \times DG_1$  space, using the same limited advection scheme, but without the ‘recovered’ parts of the scheme. The field does indeed remain bounded, suggesting that the limiter has worked well.

## 5 Compressible Euler Model

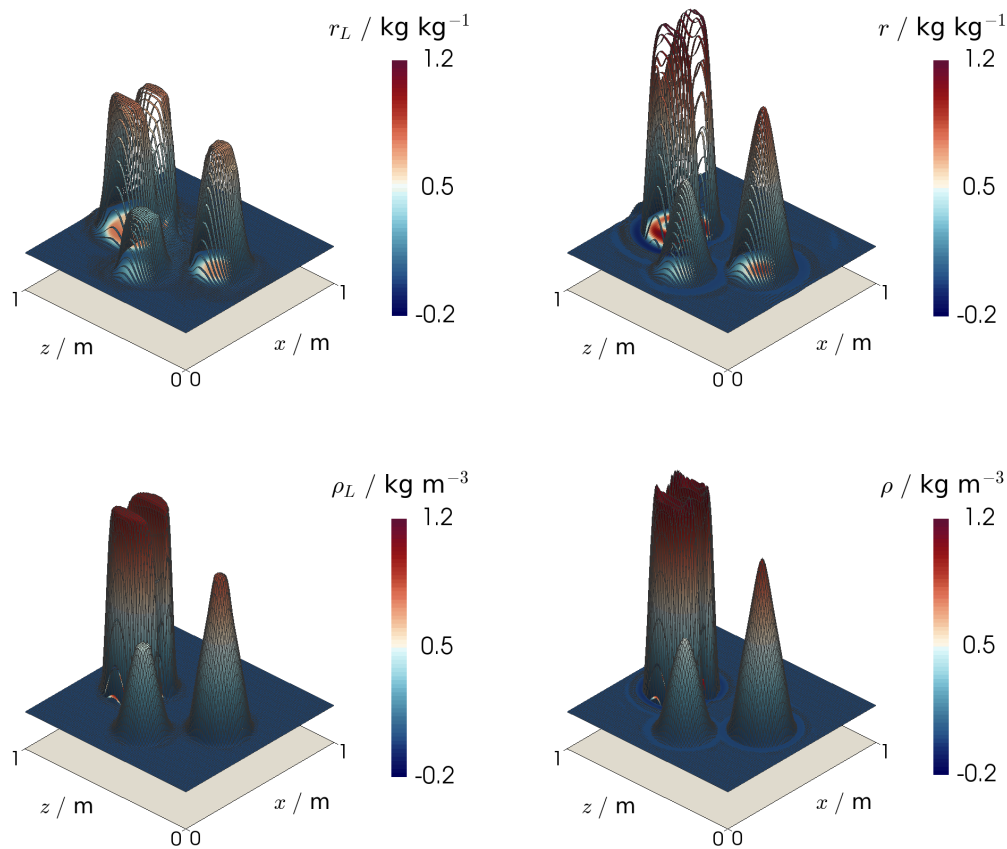
We have used this advection scheme in a numerical model of the compressible Euler equations. The continuous equations we used are

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} + c_p \theta \nabla \Pi - \mathbf{g} = \mathbf{0}, \quad (5.1)$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \quad (5.2)$$

$$\frac{\partial \theta}{\partial t} + \mathbf{v} \cdot \nabla \theta = 0, \quad (5.3)$$

$$\Pi = \left( \frac{R \theta \rho}{p_0} \right)^{\frac{\kappa}{1-\kappa}} \quad (5.4)$$



**Figure 9:** The resulting fields from one revolution of the solid-body rotation case of [23]. (Top left) A hypothetical moisture field in the  $DG_0 \times CG_1$  space, advected using the limited ‘recovered’ space scheme, compared with (top right) the same field advected using the non-limited scheme. Although overshoots and undershoots are prevented, conservation of mass is compromised. (Bottom left) A density field in the  $DG_1 \times DG_1$  space, using the same advection operator  $\mathcal{A}$  as in the ‘recovered’ scheme and limited by the vertex-based limiter of [16], with (bottom right) the same solution but without a limiter applied. This shows the effectiveness of the limiter in preventing overshoots and undershoots.

where  $\mathbf{g} = -g\hat{\mathbf{k}}$  with  $g = 9.81 \text{ m s}^{-2}$  is the uniform gravitational acceleration towards the Earth’s surface,  $c_p = 1004.5 \text{ J kg}^{-1} \text{ K}^{-1}$  is the specific heat capacity at constant pressure of a dry ideal gas,  $R = 287 \text{ J kg}^{-1} \text{ K}^{-1}$  is the specific gas constant of a dry ideal gas,  $\kappa = R/c_p$  and  $\Pi$  is the Exner pressure, found at reference pressure  $p_0 = 1000 \text{ hPa}$ .

The general strategy to solve these equations is based upon that used in the UK Met Office Endgame model, and is very similar to that described in [8], which discretised the Boussinesq equations.

The overall structure of our model can be described by the performance of various operations to the state variables, which we denote together by  $\boldsymbol{\chi} = (\mathbf{v}, \rho, \theta)$ . In a time step, the first stage is to apply a ‘forcing’  $\mathcal{F}(\boldsymbol{\chi})$  to  $\boldsymbol{\chi}$ , the details of which are explained below. At this point the algorithm enters an outer iterative loop, in which an advecting velocity  $\bar{\mathbf{u}}$  is determined, and an advection step is performed by an

advection operator  $\mathcal{V}_{\bar{\mathbf{u}}}$ . There is then an inner loop, in which the forcing is reapplied and a residual is calculated between the newly forced state and the best estimate for the state at the next time step. The state is corrected by solving a linear problem (here we denote the linear operator by  $\mathcal{S}$ ) for the residual. The scheme is balanced between being explicit and implicit by the off-centering parameter  $\alpha$ , which we will take to be  $1/2$ . This process is summarised by the following pseudocode:

1. FORCING:  $\boldsymbol{\chi}^* = \boldsymbol{\chi}^n + (1 - \alpha)\Delta t\mathcal{F}(\boldsymbol{\chi}^n)$
2. SET:  $\boldsymbol{\chi}_p^{n+1} = \boldsymbol{\chi}^n$
3. OUTER:
  - (a) UPDATE:  $\bar{\mathbf{u}} = \frac{1}{2}(\mathbf{v}_p^{n+1} + \mathbf{v}^n)$
  - (b) ADVECT:  $\boldsymbol{\chi}_p = \mathcal{V}_{\bar{\mathbf{u}}}(\boldsymbol{\chi}^*)$
  - (c) INNER:
    - i. FIND RESIDUAL:  $\boldsymbol{\chi}_{\text{rhs}} = \boldsymbol{\chi}_p + \alpha\Delta t\mathcal{F}(\boldsymbol{\chi}_p^{n+1}) - \boldsymbol{\chi}_p^{n+1}$
    - ii. SOLVE:  $\mathcal{S}(\Delta\boldsymbol{\chi}) = \boldsymbol{\chi}_{\text{rhs}}$  for  $\Delta\boldsymbol{\chi}$
    - iii. INCREMENT:  $\boldsymbol{\chi}_p^{n+1} = \boldsymbol{\chi}_p^{n+1} + \Delta\boldsymbol{\chi}$
4. ADVANCE TIME STEP:  $\boldsymbol{\chi}^n = \boldsymbol{\chi}_p^{n+1}$

In our case, the forcing operator acts only upon the velocity. It is the solution  $\mathbf{v}_{\text{trial}}$  to the following problem:

$$\int_{\Omega} \boldsymbol{\psi} \cdot \mathbf{v}_{\text{trial}} \, dx = \int_{\Omega} c_p \nabla \cdot (\theta \boldsymbol{\psi}) \, dx - \int_{\Gamma} c_p \llbracket \theta \boldsymbol{\psi} \cdot \hat{\mathbf{n}} \rrbracket \langle \Pi \rangle \, dS - \int_{\Omega} g \boldsymbol{\psi} \cdot \hat{\mathbf{k}} \, dx, \quad \forall \boldsymbol{\psi} \in V_{\mathbf{v}}, \quad (5.5)$$

where  $\Omega$  is the domain,  $\Gamma$  is the set of all interior facets, the angled brackets  $\langle \cdot \rangle$  denote the average value on either side of a facet and  $V_{\mathbf{v}}$  is the function space in which the velocity field lies.

The advection operators use the scheme defined in Section 2 with  $\mathcal{A}$  as the simple upwinding and three-step Runge-Kutta method described in Section 3.2. The spaces used in the recovered scheme are listed in Section 2.2. **We do not use any limiting strategy for  $\theta$ , and use the projection operator  $\mathcal{P}_{\mathcal{A}}$  in all schemes.**

Finally, the strategy for the linear solve step is to first analytically eliminate  $\theta$ . The resulting problem for  $\mathbf{u}$  and  $\rho$  defines the operator  $\mathcal{S}$ , which we solve using a Schur complement preconditioner. Then  $\theta$  is reconstructed from the result.

## 5.1 Rising Bubble Test Case

Here we show some results of using the recovered space advection scheme within a full model of the compressible Euler equations in the context of a vertical slice model. The example test case that we use is the dry bubble test of [24]. The initial state is  $\theta_b = 300$  K and zero velocity everywhere, while  $\rho$  is determined via from solving for hydrostatic balance using the procedure described in [6]. The following perturbation to  $\theta$  was then applied:

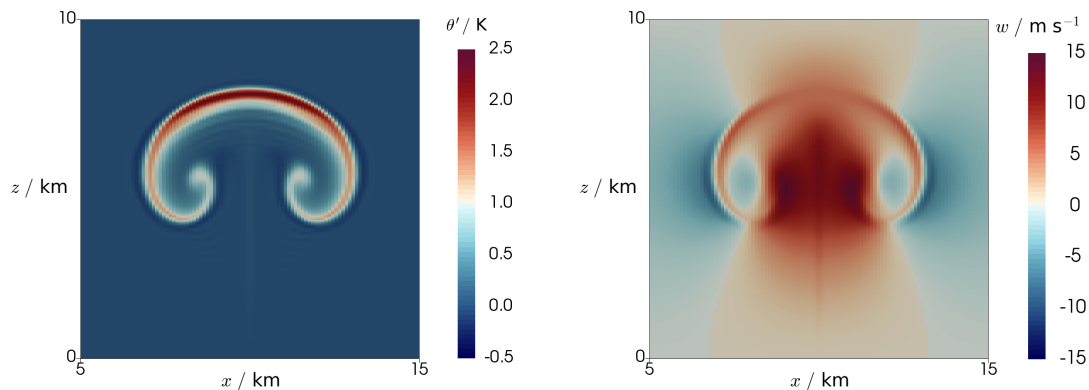
$$\theta' = \begin{cases} 2 \cos^2(\pi r / (2r_c)) \text{ K}, & r < r_c, \\ 0, & r \geq r_c, \end{cases} \quad (5.6)$$

so that  $\theta = \theta_b + \theta'$ , with

$$r = \sqrt{(x - x_c)^2 + (z - z_c)^2}, \quad (5.7)$$

where  $x_c = 10$  km,  $z_c = r_c = 2$  km. In the model used in [24], the Exner pressure is a prognostic variable, rather than the density  $\rho$ . To ensure that our initial pressure is unchanged by the perturbation, we found the initial density state by solving for  $\rho_{\text{trial}}$ :

$$\int_{\Omega} \phi \rho_{\text{trial}} \, dx = \int_{\Omega} \phi \rho_b \theta_b / \theta \, dx, \quad \forall \phi \in V_{\rho}. \quad (5.8)$$



**Figure 10:** Fields at  $t = 1000$  s from a run at resolution  $\Delta x = 100$  m of the dry bubble case from [24], representing a rising thermal. We used the lowest-order family of spaces with the ‘recovered’ space scheme as the advection method. (Left) the perturbation  $\theta'$  to the background constant state of 300 K. (Right) The vertical velocity field.

where  $V_\rho$  is the function space that  $\rho$  lives in, and  $\theta_b$  and  $\rho_b$  are the hydrostatically balanced background states. The domain used had a width of 20 km and a height of 10 km. The top and bottom boundaries had rigid lid boundary conditions ( $\mathbf{v} \cdot \hat{\mathbf{n}} = 0$ ) whilst there were periodic boundary conditions on the left and right sides. The perturbed potential temperature field at the final time  $t = 1000$  s is shown in Figure 10 for a simulation with grid spacing  $\Delta x = 100$  m and time steps of  $\Delta t = 1$  s. **Although there is good agreement between the fields shown in Figure 10 and those presented in [24], there are some visible errors in the solution: chiefly some ringing artifacts surrounding the mushroom-shaped perturbation that develops in the  $\theta$  field. In the authors’ experience, such numerical simulations of rising or sinking thermals can often exhibit instability at the leading edge of the moving bubble, capturing an inherent physical instability. In this case, the underlying equations may have multiple solutions and the numerical solution may not converge to one as the resolution is refined.**

## 6 Summary and Outlook

We have presented a new ‘recovered’ advection scheme for use in numerical weather prediction models. This scheme is a form of the embedded DG advection described in [9], but in which higher-degree spaces are recovered via averaging operators, as described in [12]. It is intended for use with the compatible finite element set-up laid out in [2], and in particular can be used with the zeroth-degree set of spaces. With these spaces, the scheme has second-order numerical accuracy. We have also presented a bounded version of this scheme, which can be used for moisture variables to preserve monotonicity or to prevent negative values. Stability properties of the scheme have also been provided. **Future work should extend this scheme to cover vector functions that lie in curved domains. This will require careful averaging of vectors that lie in different tangent spaces.**

## Acknowledgements

TMB was supported by the EPSRC Mathematics of Planet Earth Centre for Doctoral Training at Imperial College London and the University of Reading. CJC was supported by EPSRC grant EP/L000407/1, while JS was supported by the EPSRC EP/L000407/1 and NERC NE/R008795/1 grants. The authors would also like to thank the developers of the Firedrake software which was used extensively for this work, and in particular Thomas H. Gibson who wrote the original piece of code for the recovery operator.

## References

- [1] A. Staniforth and J. Thuburn, “Horizontal grids for global weather and climate prediction models: a review,” *Quarterly Journal of the Royal Meteorological Society*, vol. 138, no. 662, pp. 1–26, 2012.
- [2] C. J. Cotter and J. Shipton, “Mixed finite elements for numerical weather prediction,” *Journal of Computational Physics*, vol. 231, no. 21, pp. 7076–7091, 2012.
- [3] A. Staniforth, T. Melvin, and C. Cotter, “Analysis of a mixed finite-element pair proposed for an atmospheric dynamical core,” *Quarterly Journal of the Royal Meteorological Society*, vol. 139, no. 674, pp. 1239–1254, 2013.
- [4] C. J. Cotter and J. Thuburn, “A finite element exterior calculus framework for the rotating shallow-water equations,” *Journal of Computational Physics*, vol. 257, pp. 1506–1526, 2014.
- [5] A. T. McRae and C. J. Cotter, “Energy-and enstrophy-conserving schemes for the shallow-water equations, based on mimetic finite elements,” *Quarterly Journal of the Royal Meteorological Society*, vol. 140, no. 684, pp. 2223–2234, 2014.
- [6] A. Natale, J. Shipton, and C. J. Cotter, “Compatible finite element spaces for geophysical fluid dynamics,” *Dynamics and Statistics of the Climate System*, vol. 1, no. 1, 2016.
- [7] J. Shipton, T. Gibson, and C. Cotter, “Higher-order compatible finite element schemes for the non-linear rotating shallow water equations on the sphere,” *Journal of Computational Physics*, vol. 375, pp. 1121–1137, 2018.
- [8] H. Yamazaki, J. Shipton, M. J. Cullen, L. Mitchell, and C. J. Cotter, “Vertical slice modelling of nonlinear Eady waves using a compatible finite element method,” *Journal of Computational Physics*, vol. 343, pp. 130–149, 2017.
- [9] C. J. Cotter and D. Kuzmin, “Embedded discontinuous Galerkin transport schemes with localised limiters,” *Journal of Computational Physics*, vol. 311, pp. 363–373, 2016.
- [10] D. N. Arnold and A. Logg, “Periodic table of the finite elements,” *SIAM News*, vol. 47, no. 9, p. 212, 2014.
- [11] A. T. McRae, G.-T. Bercea, L. Mitchell, D. A. Ham, and C. J. Cotter, “Automated generation and symbolic manipulation of tensor product finite elements,” *SIAM Journal on Scientific Computing*, vol. 38, no. 5, pp. S25–S47, 2016.
- [12] E. H. Georgoulis and T. Pryer, “Recovered finite element methods,” *Computer Methods in Applied Mechanics and Engineering*, vol. 332, pp. 303–324, 2018.
- [13] V. A. Titarev and E. F. Toro, “Ader: Arbitrary high order godunov approach,” *Journal of Scientific Computing*, vol. 17, no. 1-4, pp. 609–618, 2002.
- [14] B. Van Leer and S. Nomura, “Discontinuous galerkin for diffusion,” in *17th AIAA Computational Fluid Dynamics Conference*, p. 5108, 2005.
- [15] O. A. Karakashian and F. Pascal, “Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems,” *SIAM Journal on Numerical Analysis*, vol. 45, no. 2, pp. 641–665, 2007.
- [16] D. Kuzmin, “A vertex-based hierarchical slope limiter for p-adaptive discontinuous Galerkin methods,” *Journal of computational and applied mathematics*, vol. 233, no. 12, pp. 3077–3085, 2010.
- [17] C.-W. Shu and S. Osher, “Efficient implementation of essentially non-oscillatory shock-capturing schemes,” *Journal of Computational physics*, vol. 77, no. 2, pp. 439–471, 1988.
- [18] B. Cockburn and C.-W. Shu, “Runge–kutta discontinuous galerkin methods for convection-dominated problems,” *Journal of scientific computing*, vol. 16, no. 3, pp. 173–261, 2001.
- [19] F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. McRae, G.-T. Bercea, G. R. Markall, and P. H. Kelly, “Firedrake: automating the finite element method by composing abstractions,” *ACM Transactions on Mathematical Software (TOMS)*, vol. 43, no. 3, p. 24, 2017.

## REFERENCES

---

- [20] G.-T. Bercea, A. T. McRae, D. A. Ham, L. Mitchell, F. Rathgeber, L. Nardi, F. Luporini, and P. H. Kelly, “A structure-exploiting numbering algorithm for finite elements on extruded meshes, and its performance evaluation in firedrake,” *arXiv preprint arXiv:1604.05937*, 2016.
- [21] M. Homolya and D. A. Ham, “A parallel edge orientation algorithm for quadrilateral meshes,” *SIAM Journal on Scientific Computing*, vol. 38, no. 5, pp. S48–S61, 2016.
- [22] D. L. Williamson, J. B. Drake, J. J. Hack, R. Jakob, and P. N. Swarztrauber, “A standard test set for numerical approximations to the shallow water equations in spherical geometry,” *Journal of Computational Physics*, vol. 102, no. 1, pp. 211–224, 1992.
- [23] R. J. LeVeque, “High-resolution conservative algorithms for advection in incompressible flow,” *SIAM Journal on Numerical Analysis*, vol. 33, no. 2, pp. 627–665, 1996.
- [24] G. H. Bryan and J. M. Fritsch, “A benchmark simulation for moist nonhydrostatic numerical models,” *Monthly Weather Review*, vol. 130, no. 12, pp. 2917–2928, 2002.