Journal of Computational Physics 310 (2016) 127–160

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

# Accuracy analysis of mimetic finite volume operators on geodesic grids and a consistent alternative

Pedro S. Peixoto [a,b,1]

[a] College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK
[b] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

ARTICLE INFO

*Article history:*
Received 11 August 2015
Received in revised form 1 November 2015
Accepted 17 December 2015
Available online 14 January 2016

*Keywords:*
Shallow water model
Finite volume
Staggered C grid
Spherical Voronoi grid
Mimetic discretization

ABSTRACT

Many newly developed climate, weather and ocean global models are based on quasi-uniform spherical polygonal grids, aiming for high resolution and better scalability. Thuburn et al. (2009) and Ringler et al. (2010) developed a C staggered finite volume/difference method for arbitrary polygonal spherical grids suitable for these next generation dynamical cores. This method has many desirable mimetic properties and became popular, being adopted in some recent models, in spite of being known to possess low order of accuracy. In this work, we show that, for the nonlinear shallow water equations on non-uniform grids, the method has potentially 3 main sources of inconsistencies (local truncation errors not converging to zero as the grid is refined): (i) the divergence term of the continuity equation, (ii) the perpendicular velocity and (iii) the kinetic energy terms of the vector invariant form of the momentum equations. Although some of these inconsistencies have not impacted the convergence on some standard shallow water test cases up until now, they may constitute a potential problem for high resolution 3D models. Based on our analysis, we propose modifications for the method that will make it first order accurate in the maximum norm. It preserves many of the mimetic properties, albeit having non-steady geostrophic modes on the f-sphere. Experimental results show that the resulting model is a more accurate alternative to the existing formulations and should provide means of having a consistent, computationally cheap and scalable atmospheric or ocean model on C staggered Voronoi grids.



## 1. Introduction

Aiming for very high resolution scalable models, several newly developed global atmospheric models are using quasi-uniform spherical grids [1]. In Thuburn et al. [2] and Ringler et al. [3], a finite volume/difference numerical scheme with very robust characteristics was developed for arbitrary orthogonal C grids, and later it was extended to non-orthogonal grids [4,5]. The developed scheme, hereafter named TRSK, possesses many desirable properties that mimic the continuous equations [1]. It also uses very compact stencils and does not require a mass matrix or the solution of global linear systems. For these reasons, it became popular among researchers, being used in models such as the NCAR/Los Alamos Laboratory Model for Prediction Across Scales (MPAS) for the atmosphere [6] and ocean [7], in the DYNAMICO model [8], and in many other studies [5,9–12]. Although the methodology is proving to be adequate for weather and climate global modelling, it

*E-mail address:* pedrosp@ime.usp.br.
[1] Tel.: +55 11 30916136; fax: +55 11 30916131.

http://dx.doi.org/10.1016/j.jcp.2015.12.058
0021-9991/

presents low accuracy order. Apparently, increasing the accuracy order of the model is not possible without breaking some of its properties. The aim of this work is to analyse the accuracy properties of this method and to propose a more accurate method that preserves as many of the mimetic properties as possible. The methodology is mostly suitable for orthogonal grids, and we will adopt a similar framework as used in MPAS, with quasi-uniform spherical Voronoi grids formed by hexagons and pentagons, with dual grid formed by triangles.

There are many properties that a numerical method for global models are desired to have [1]. To start with, it should be consistent with the continuous equations (the discrete system should approximate the continuous one with increasing resolution) and stable. In addition, it is desirable to mimic conservation properties, such as of mass, energy and axial angular momentum. Moreover, the model should be able to accurately represent balanced flows and adjustment processes. On a shallow water system, the TRSK scheme has many of these properties, and some others, e.g.

1. Mass conservation.
2. Accurate representation of fast waves and adjustment (C staggering).
3. Curl-free pressure gradient (Discrete version of $\nabla \times \nabla \psi = 0$).
4. Energy conservation of pressure terms (Discrete analogue of $\vec{v} \cdot \nabla h + h \nabla \cdot \vec{v} = \nabla \cdot (h \vec{v})$).
5. Energy conserving Coriolis term (Discrete analogue of $\vec{v} \cdot \vec{v}^{\perp} = 0$).
6. Conservation of total energy.
7. Steady geostrophic modes (on the f-sphere).
8. Compatible discretization of potential vorticity with its Lagrangian behaviour.

Property 2 deals with the accurate representation of fast waves, such as the inertial-gravity ones, which for well resolved Rossby radius scenarios has direct implications on the models ability to represent correctly the process of geostrophic adjustment [13]. Properties 3, 4 and 5 rely on having discrete analogues of certain continuous relations. Total energy conservation is obtained within time truncation errors. Properties 7 and 8 are interconnected and related to the ability of the model to accurately represent geostrophically balanced flows and the evolution of the potential vorticity.

The TRSK methodology is second order accurate on regular hexagons and rectangles for the shallow water equations. It also has a small constant of proportionality between accuracy and resolution due to the mimetic and other desirable numerical properties. In addition, it shows second order accuracy in the mean error norm ($L_2$) in most shallow water test cases [3,5]. Nevertheless, some test cases show that the method fails to converge in the maximum norm, and this lack of convergence can be carried to the 3D model, which happens, for example, in the MPAS model (Skamarock, 2015, personal communication).

Although consistency is not a necessary condition for convergence (since supraconvergence may occur, e.g. [14]), it is nonetheless highly desirable to have a consistent scheme, which, if stable, would then imply convergence. Consistency analysis also provides means of tracking sources of convergence problems. We will discuss in this work that, for on a shallow water system with input data given point-wisely, several discrete operators of TRSK fail to approximate the continuous ones with increasing resolution in the maximum norm. The main problems occur on the horizontal divergence, kinetic energy and Coriolis terms.

Lack of accuracy of finite volume/difference discretizations on quasi-uniform spherical grids is usually caused by grid related problems [11,15]. To improve certain grid properties, grid optimizations have been proposed (see [16] for a review). Two of them will be of major interest in our analysis: (i) Spherical Centroidal Voronoi Tesselations (SCVT), which optimize the grid by iteratively moving the cell centres to their mass centroids, and (ii) an optimization that approximates the midpoints of primal (hexagonal/pentagonal cells) and dual (triangles) edge midpoints, due to Heikes and Randall [17], hereafter HR95 optimization. We will show that the common finite volume discretization of the divergence and curl may lead to inconsistent schemes with respect to the maximum norm, with first order accuracy being attained only if a HR95 like optimization is used. On the other hand, the use of SCVT optimization improves the accuracy of some other operators, as for example, in the treatment of the Coriolis term with the TRSK scheme.

Based on previous analyses of Peixoto and Barros [18], we will propose several modifications to the TRSK scheme. The main modifications involves a different location of the normal velocities on the grid (preserving the C-staggering on Voronoi Cells), a different reconstruction algorithm for the velocities and the use of a barycentric second order linear interpolation of scalar fields. This modified scheme is a first order accurate method in the maximum norm, which has similar properties to those of TRSK, except that,

(i) total energy will no longer be conserved, due solely to a different discretization of the kinetic energy term,
(ii) the geostrophic modes will no longer be steady on the f-sphere, due a different discretization of the Coriolis term, and, for the same reason,
(iii) the potential vorticity discretization will no longer be compatible with its Lagrangian behaviour.

We will discuss how these modifications impact the models characteristics, concluding that the modified scheme provides a good trade off in losing some properties of TRSK in order to obtain first order accuracy.

Summarizing, the main goals of this paper are as follows. First, we wish to present an in depth analysis of the local truncation errors of the TRSK scheme, in order to provide awareness of its strengths and weaknesses. Also, we will show

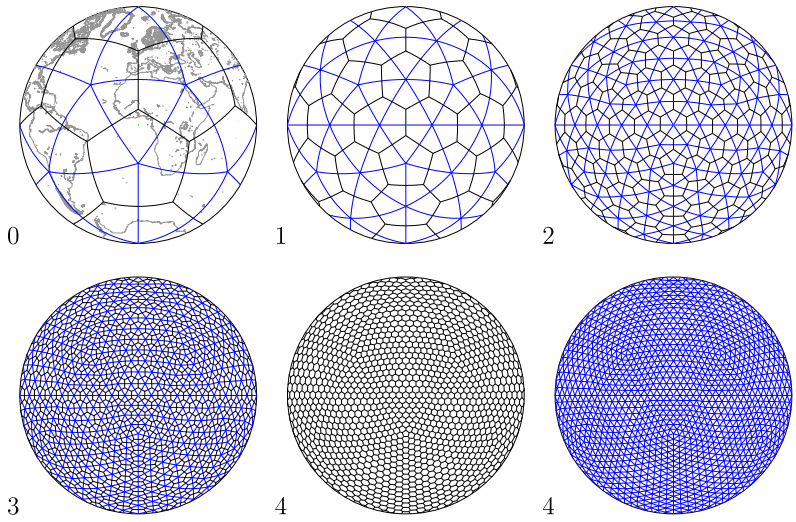


**Fig. 1.** Delaunay triangulations and/or Voronoi cells are shown on different levels (indicated by each grid) for icosahedral grids.

several possible modifications to the existing formulations to achieve consistency on each discrete operator. These can be interpreted operator-wisely, improving only certain aspects of the existing schemes, or adopted as a whole, considering all the alternative approaches suggested – to form the here proposed consistent scheme.

In section 2 we describe the basic framework of the model and grid, within a shallow water system. Section 3 is dedicated to a detailed analysis of the accuracy of each discrete operator, and of possible modifications to ensure better accuracy. In section 4, results from several shallow water tests cases are presented, comparing TRSK with the modified scheme on different grids. We also discuss about some of the TRSK properties which are lost. We finish the paper with a summary, discussing the proposed modifications and the attained results in section 5.

## 2. Model description

### 2.1. Shallow water equations

The nonlinear shallow water equations can be written for the sphere in vector invariant form as [3],

$$\frac{\partial h}{\partial t} + \nabla \cdot (h\vec{u}) = 0, \tag{1}$$

$$\frac{\partial \vec{u}}{\partial t} + qh\vec{u}^{\perp} = -g\nabla(h+b) - \nabla K, \tag{2}$$

where $h$ is the height or thickness of the fluid layer, $\vec{u}$ is the fluid velocity, which is assumed to be tangent to the sphere, $b$ is the bottom topography, $g$ is the gravity constant, $q = \eta/h$ is the potential vorticity, where $\eta$ is the absolute vorticity, defined as

$$\eta = \vec{k} \cdot \nabla \times \vec{u} + f, \tag{3}$$

where $\vec{k}$ is the unit vector pointing in the local vertical direction (such that $\vec{k} \cdot \vec{u} = 0$), $\vec{u}^{\perp} = \vec{k} \times \vec{u}$ and $f$ is the Coriolis parameter. $K$ is the kinetic energy per unit mass, defined as

$$K = \frac{\vec{u} \cdot \vec{u}}{2}. \tag{4}$$

### 2.2. Grid specifications

We will adopt an icosahedral based geodesic grid, refined by bisecting the spherical triangles edges (see Fig. 1). The triangles form a Delaunay triangulation of the sphere and their vertices will be set as grid nodes. The set of nodes may be used as generators of a spherical Voronoi tessellation. Each spherical Voronoi polygon will be viewed as a computational cell relative to its node. The resulting grid will be formed by an arbitrary number of hexagonal Voronoi cells (depending on the number of refinements done) and 12 pentagonal Voronoi cells (primal grid), with an underlying triangular grid (dual grid).

We use the convention that the grid level 0 is the original grid obtained from the icosahedron (it has distance between nodes of approximately 7054 km). The mean distance between Voronoi cell centres on the finer grids is approximately

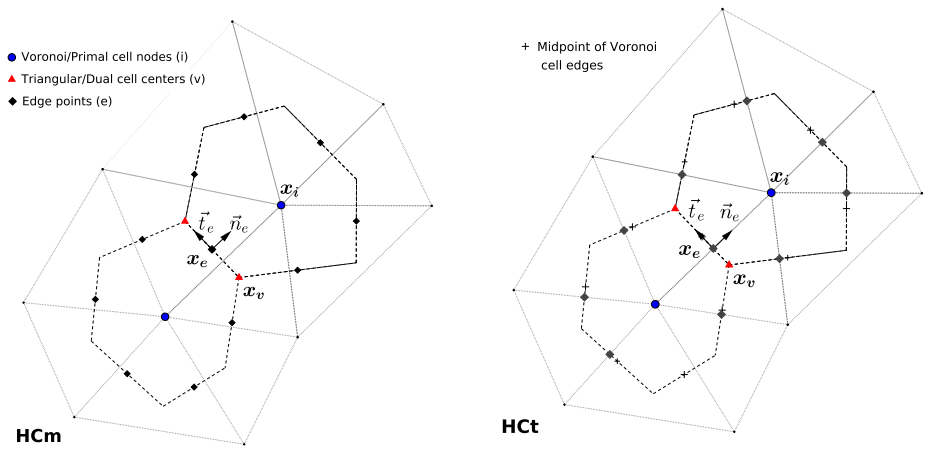


**Fig. 2.** HC grid variable collocation points.

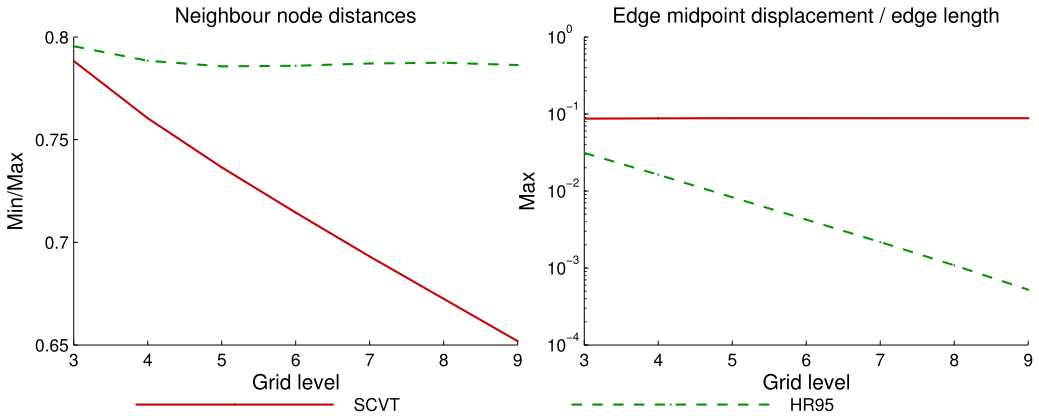


**Fig. 3.** Grid properties for SCVT and HR95 optimized icosahedral grids. Left: Minimum distance between two nodes divided by maximum distance between nodes. Right: Maximum of the ratio of distance between edge midpoints and the Voronoi edge length.

60 km for the grid level 7 (with 163,842 cells), 30 km for grid level 8 (with 655,362 cells) and 15 km for grid level 9 (with 2,621,442 cells).

We will use a C grid staggering, where the scalar field (fluid height) is stored at the Voronoi cell nodes and the velocities are stored at the edges, but only the component normal to the Voronoi cell edge. Here, two points must be made clear, as these will greatly influence the analysis that will follow:

(i) the nodes that generate the cell nodes are not necessarily the centroids (mass centres) of the Voronoi cells,
(ii) the midpoints of the Voronoi cell edges and the triangle cell edges do not coincide (see Fig. 2 for an example of how the edge midpoints might happen to be positioned).

As shown in Peixoto and Barros [15], the issue number (i) tends to reduce the order of accuracy of certain discrete operators, such as the divergence, curl or Laplacian. To avoid this problem, it is possible to optimize the position of the nodes positioning in order to obtain a grid with the property of having Voronoi nodes coinciding with the cell centroids – these are known as Spherical Centroidal Voronoi Tessellations (SCVT), see [19] and [20] for details. With respect to topic (ii), Heikes and Randall [17] showed that it causes the usual discretization of the Laplacian to be inconsistent. To enable a consistent scheme, they propose a grid optimization in which the midpoints of the triangle edges converge to the midpoints of the Voronoi cell edges with grid refinement – we will address this optimized grid as a HR95 grid. Another possible optimization applied to this kind of grid is the "spring dynamics" optimization [21], which reduces grid distortions, but does not improve the above issues as significantly as the purpose specific two above mentioned optimizations.

The analysis of Miura [16] shows that two of the above mentioned optimizations, SCVT and HR95, seem to be antagonistic, and apparently one could not obtain both properties simultaneously. We show in Fig. 3 two important properties of these kinds of optimized grids:

(i) The ratio between the maximum and minimum distance between two neighbour nodes, which is related to the degree of uniformity of the grid.
(ii) The maximum ratio between the distance from the triangle edge midpoint to the Voronoi cell edge midpoint (edge midpoint displacement) and the actual Voronoi edge length.

The first point shows that the degree of uniformity of the SCVT grid is being gradually lost with increasing resolution, whereas the HR95 grid preserves the grid uniformity with increasing resolution (see left panel of Fig. 3). We recall that our initial goal of going to non-structured grids is to use a grid as uniform as possible on the sphere, to avoid concentration of points (singularities) as happens in latitude–longitude grids.

The second point shows that the edge displacement (distance between edge midpoints) is converging to zero faster than the edge length with increasing resolution on the HR95 grid (see right panel of Fig. 3). This is a main property of this grid optimization, and will be used later to show that it is an important requisite to obtain consistent finite difference/volume methods on a staggered C grids for the shallow water equations. On the SCVT grid, the distance between edge midpoints converges to zero at the same rate as the edge length goes to zero. For a full comparison of grid optimization properties, see [16].

Our grid and discretization notation will closely follow the ones used in Ringler et al. [3], but our placement of variables will be slightly different. It is possible to position the velocities in two different (and relevant) points of the edges: (i) the midpoints of the edges of the Voronoi cells – we will call this positioning **HCm**, and (ii) the midpoint of the triangle edges, which represents a point on the Voronoi cell edge that intersects the triangle edges – we will call this positioning **HCt**. The latter is what is used in the TRSK scheme [3]. We point out that the velocity information stored is just the normal component relative to the Voronoi cell edge in both cases (see Fig. 2). Interestingly, we will show that this difference in positioning has a lot of impact over the accuracy of the discrete schemes.

In the analysis that follows we will investigate the two icosahedral optimized grids discussed, SCVT and HR95, each one with two possible variable positionings, HCm or HCt, resulting in 4 possible configurations. On a HR95 optimized grid, these two positionings (HCm and HCt) will asymptotically converge to each other, since the edge midpoint displacement will be reduced very fast with resolution. Therefore, the results for HCt and HCm grid positionings tend to be similar if the HR95 optimization is used. However, since the HR95 optimization does not ensure exact matching of the edge midpoints, some difference might still be relevant and will be analysed.

Variables stored in Voronoi cell nodes (triangle vertices) will be denoted using the index $i$. Variables stored in triangle circumcentres (Voronoi cell vertices) will be denoted using the index $v$, and variables stored on edges will be denoted using the index $e$.

### 2.3. Model initialization and interpretation of the degrees of freedom

Our analysis will be based on pointwise specification of the height ($h_i$) and wind ($u_e$) as initial conditions, relative to their exact positions on the grid. For given input wind fields, only the normal components of the pointwise velocities relative to the edges are used in the initialization. The main reason for analysing the problem this way is that, in non-idealized (real) atmospheric and ocean models, the variables (such as wind and pressure) are usually interpolated point-wisely into their grid positions.

As common in finite volume schemes, the variables may be thought in terms of cell averages and mean flux/circulation. Therefore, $h_i$ may be thought as the mean cell height average, and $u_e$ the mean edge velocity normal to the edge. This is closely related the framework proposed by Thuburn and Cotter [4]. In Thuburn and Cotter, they suggest considering the cell integrated height (geopotential in their notation) and 2 possibilities for the velocity degrees of freedom:

(i) The circulation along the dual edge (triangle edge), i.e. the integral along the dual edge of the velocity tangent to this edge.
(ii) The volume flux across the primal edge (Voronoi edge), i.e. the integral along the primal edge of the velocity normal to this edge.

Considering these interpretations of the degrees of freedom as the mean integrals (integrals divided by their respective edge lengths), then we can directly relate these with the HCt and HCm positionings as follows. The velocity degree of freedom considering the mean circulation may be pointwisely approximated to the midpoint of the triangle edge with 2nd order accuracy using the midpoint integration rule. The midpoint of the triangle edge is the reference velocity point for the HCt positioning, and therefore, the HCt positioning is a second order approximation to the mean circulation integral. The HCm positioning would lead to a first order only approximation to the mean circulation. On the other hand, the velocity degree of freedom considering the mean volume flux may be pointwisely approximated to the Voronoi edge midpoint with 2nd order accuracy using the midpoint rule. This implies that the HCm positioning is a second order approximation for the mean volume flux, and HCt would be only a first order approximation of the mean volume flux.

In Thuburn and Cotter [4] they recommend the circulation interpretation for the prognostic velocity degree of freedom. In Ringler et al. [3] they use the HCt positioning, which we discussed to be a second order approximation to the edge mean

circulation. Here, we will open room for the HCm positioning (mean volume flux interpretation), which we will later see that can help to enhance the accuracy of certain operators.

When integral forms are considered for the degrees of freedom, the initialization process requires some care. For problems with known analytical stream function, the model can be initialized with the exact integral values considering the pointwise stream function values at the ends of the edges. When no stream function is known analytically for the initial conditions, these integrals need to be approximated. A natural approximation to initialize the model would be to consider the second order approximations using HCt or HCm pointwise input data (respectively approximations to the mean circulation along the dual edge or the mean flux across the primal edges).

Alternatively, one might be given exact mass fluxes ($h\vec{u}$) as initial conditions. In this case, dividing the flux by the edge length would give the average edge flux, which is represented with second order at the edge midpoint, and is similar to what we are calling HCm grid. If the equations were to be written fully in flux form (having $h\vec{u}$ as unknown), and the initial data were to be given as fluxes, then it would not matter were the velocity is positioned within the edge, but a natural place would be at the edge midpoint (what we calling HCm grid).

### 2.4. Discrete shallow water system

We will consider a similar discrete system as the one described in Ringler et al. [3], that may be concisely stated as,

$$\frac{\partial h_i}{\partial t} = -D_i,$$
$$\frac{\partial u_e}{\partial t} = -Q_e^{\perp} - G_e,$$

where $D_i$, $G_e$ and $Q_e^{\perp}$ are respectively discretizations of the divergence term of the mass equation ($\nabla \cdot (h\vec{u})$), gradient term ($\nabla(g(h+b)+K)$) and Coriolis term ($qh\vec{u}^{\perp}$) of the momentum equation.

The discretizations will depend on the kind of positioning used (HCt or HCm) and will be described in the next section, together with an accuracy analysis of the discrete operators.

## 3. Accuracy analysis of discrete operators

In what follows, we will analyse the local truncation errors of each operator involved in the horizontal discretization of the shallow water equations. We will adopt the maximum error norm to label the consistency order of the operator. Therefore, when stating that a method defines a first or second order approximation, we will in general be considering this in the maximum norm. A method will be called inconsistent if the maximum local truncation error does not converge to zero with increasing resolution.

It is important to point out that, for stable schemes, the order of the local truncation error (consistency order) does not necessarily lead to the same convergence order. Supraconvergence may take place to enable larger convergence rates than what is observed in the consistency analysis (for an example refer to [14]). Nevertheless, the consistency order gives a potential insight to what might be causing a model to lose convergence. Therefore, although we will show that the TRSK scheme is inconsistent with respect to certain operators, this would not rule out the possibility of it still being convergent.

Also, as discussed in the previous section, one might interpret the variables as integral quantities. The analysis that follows directly applies to these cases by considering the analogy between the mean edge circulation with the HCt positioning, and the mean edge flux with the HCm positioning.

### 3.1. Divergence term

To build the discrete divergence $D_i$, that approximates the divergence of $h\vec{u}$, two steps are required. First the fluid height ($h$) must be interpolated to the edges, as it lives on the centre of the cells. Second, the divergence may be estimated using a discrete version of the divergence theorem.

*Interpolation of height*

Interpolation of fluid height to the edges is straightforward on the HCt grid, as, due to the orthogonality, the information lays exactly on the ends of the triangle edges and the interpolation point is precisely the midpoint of the triangle edge. Therefore, a simple average of the two values of fluid height from the two cells that share the edge will give a second order approximation of the fluid height on the desired edge point,

$$h_e = (h_i + h_j)/2, \tag{5}$$

with $i$ and $j$ indexes of the Voronoi cells that share the edge.

On a HCm grid, the simple averaging will result in a first order approximation only. To obtain a second order interpolation, we can use a linear interpolation using barycentric coordinates (see [18] for details), which leads to minimal extra computational effort and very accurate results on HCm positioning.

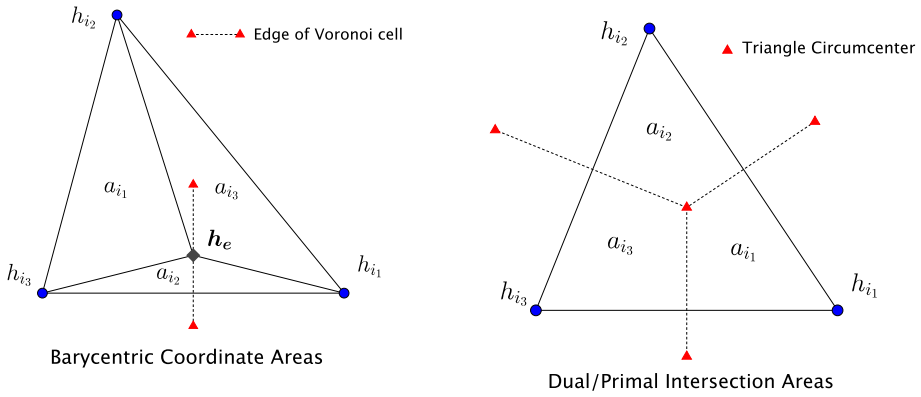


**Fig. 4.** Areas used in the barycentric coordinates (left) and for the compatible TRSK scheme (right).

The formulas for barycentric coordinate interpolation are as follows. Assume that the point of interpolation $x_e$ belongs to the triangle $T_k$, which has vertices which we will indicate using the index $i$, then, the interpolated value $h_e$ may be expressed as

$$h_e = \sum_i \lambda_i(x_e) h_i, \tag{6}$$

with the weights given by

$$\lambda_i(x_e) = \frac{a_i(x_e)}{A_k}, \tag{7}$$

where $a_i(x_e)$ is the area of the triangle formed by the two vertices of the triangle opposite $i$ and the interpolation point for $h_e$, and $A_k$ is the area of the triangle $T_k$. This is a second order approximation on the sphere and we can use the spherical areas to obtain $\lambda_i$ (see Fig. 4 on the left). It must be clear that this interpolation is different from the one used in Ringler et al. [3] to interpolate scalar values to triangle circumcentres. They use the areas relative to the intersections of the primal and dual grids as weight in the interpolation (see Fig. 4 on the right). Their approach is first order accurate and is non-interpolatory (if the interpolation point is on a vertex, it does not result in the value given at the vertex).

We point out that if the midpoints of the Voronoi cell edge and the triangle edge where to match exactly, then the linear interpolation of equation (6) would be equivalent to the simple average of equation (5).

*Divergence calculation*

Once the fluid height is known at the edge points, the divergence of $h\vec{u}$ on a Voronoi cell node $i$ can be estimated as

$$D_i = \frac{1}{A_i} \sum_e h_e u_e n_{ei} l_e, \tag{8}$$

where $l_e$ is the length of the Voronoi edge $e$, $u_e$ is the normal component of the velocity (relative to the Voronoi edge) and $n_{ei}$ is a correction term that is 1 if the normal vector to the edge $\vec{n}_e$ points outward, and $-1$ if it points inward with respect to the cell. The summation varies over the edges ($e$) of the Voronoi cell and $A_i$ is the area of the cell.

In Peixoto and Barros [15] several properties of this discrete operator were analysed. In particular, first order accuracy in the maximum norm will be obtained when: (i) the values of $h_e u_e$ are known with (at least) second order accuracy, (ii) the edge point used is the midpoint of the Voronoi cell (HCm positioning) and (iii) the grid is not too distorted (e.g. all triangles have their circumcentre inside them). For a second order approximation of the divergence (in the maximum norm), additional geometric conditions should be met (alignment of the cell edges) and the values of $h_e u_e$ must be known with greater accuracy (at least third order). For triangular cells, second order accuracy cannot be ensured (see example in [15]).

With a HCt grid, the normal velocities are not at the midpoint of the Voronoi cell edges, and therefore, this discretization is in general inconsistent (in the maximum norm), as shown in Appendix A. However, with the HR95 grid optimization, first order is attainable, since the discretization error will be of the order of the distance between the edge midpoints divided by the Voronoi edge length, and this converges linearly to zero on HR95 grids, but not on SCVT grids.

Using the HCm grid with barycentric interpolation fulfils the requirements for the discrete operator to be first order accurate in the maximum norm. The barycentric coordinates interpolation on the HCm does not affect the mimetic properties of the discrete divergence (the product rule), including energy conservation, since the same flux is used on both cells sharing an edge.

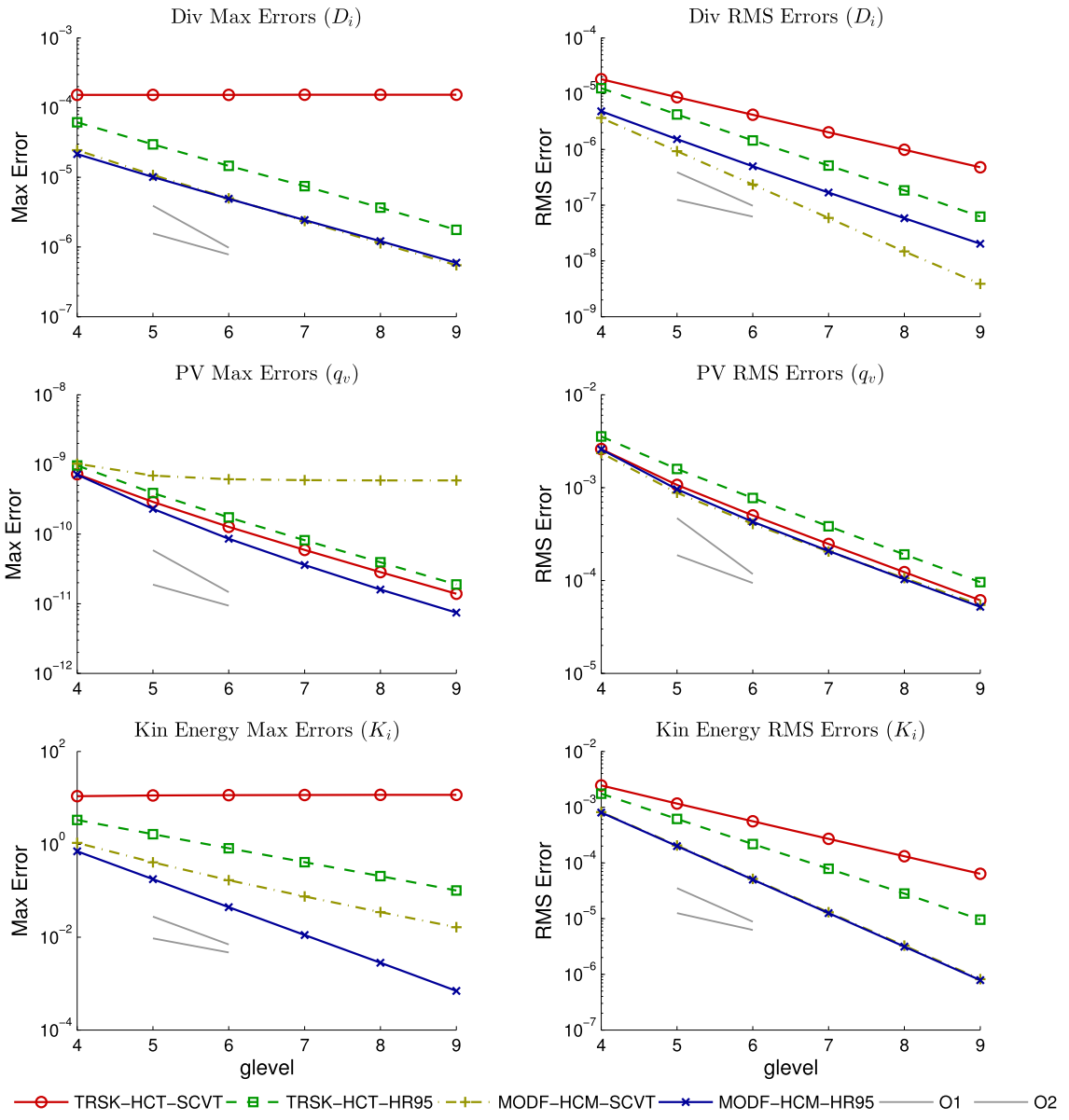


**Fig. 5.** Maximum and RMS errors of discrete operators on test case 2 initial conditions. Top line shows the divergence errors (relative to $D_i$), middle line shows the potential vorticity errors (relative to $q_v$), and the bottom line shows the kinetic energy errors (relative to $K_i$) varying with resolution. See text for a description of the methods used. Grey lines indicate first and second order reference convergence rates.

In Fig. 5, top layer, we show the errors associated with the discretization of the divergence operator ($D_i$) using a solid body rotation vector field with the fluid height defined in geostrophic balance with the velocities, as in test case 2 of Williamson et al. [22]. The maximum and root mean square (RMS) errors are shown, where the maximum error is simply the maximum absolute difference between the numerical ($h$) and the analytical solution ($\overline{h}$), and RMS errors are as described in [3],

$$E_{RMS} = \sqrt{\frac{\sum_i (h_i - \overline{h}_i)^2 A_i}{\sum_i \overline{h}_i^2 A_i}}, \tag{9}$$

where $i$ is varying over the grid cells and $A_i$ represent their areas. The maximum error was not normalized at this stage (as requested in typical test cases [22]), as the absolute errors will provide information about the overall error dominance in the shallow water equations.

Four possibilities were used to investigate the interpolation and positioning errors relative to the divergence discretization shown in the top line of Fig. 5: The simple average of $h$ with HCt positioning and SCVT or HR95 grids (respectively labelled as TRSK-HCT-SCVT and TRSK-HCT-HR95), and the linear interpolated $h$ with HCm positioning and SCVT or HR95 grids (respectively MODF-HCM-SCVT and MODF-HCM-HR95). All 4 methods were used together with the usual discrete divergence discretization of equation (8).

It is clear that with the HCt positioning on SCVT, consistency cannot be achieved (with respect to the maximum norm), and that the HR95 grid optimization makes the method first order accurate, as discussed before. The HCm positioning provides the most accurate results for the divergence, since the velocities are known in the correct position for the method to be first order accurate, without having to rely on the HR95 optimization. The two different positionings (HCt and HCm) tested with the HR95 grid do not result in the same accuracy, since the midpoints of the edges do not match exactly (the difference is of the order of the edge displacement).

Whereas the discretization on the HCt SCVT grid is inconsistent and needs the HR95 optimization to become first order accurate, on an HCm grid, the SCVT optimization gives the most accurate results in the mean error norm, almost achieving second order in the maximum norm. This could be mainly due to the fact that on this grid the cell nodes are approximately the cell mass centroid, which is a necessary condition for the method to be second order accurate (see [15] for a comprehensive discussion on this matter).

### 3.2. Vorticity

In this section we start analysing the accuracy of the discretization of the potential vorticity (PV), which is required for the perpendicular term ($Q_e^{\perp}$) of the momentum equation. The potential vorticity ($q = \eta/h$) depends on the absolute vorticity (relative vorticity plus $f$) and on the fluid height.

*Relative vorticity*

The discretization of the relative vorticity depends on the curl operator ($\vec{k} \cdot \nabla \times \vec{u}$) which may be approximated using Gauss's Theorem in the same way as the divergence operator is obtained. The resulting discrete scheme for the relative vorticity $\zeta$ at the triangles is

$$\zeta_v = \frac{1}{A_v} \sum_e u_e t_{ev} d_e, \tag{10}$$

where $d_e$ is the length of the triangle edge $e$, the summation varies over the edges ($e$) of the triangle $v$ and $A_v$ is the area of the triangle. $u_e$ is still the normal component of the velocity (relative to the Voronoi edge), but notice that now, due to the orthogonality, it may be viewed as a tangent component of the wind relative to the triangle edge (see Fig. 2). The term $t_{ei}$ is 1 if the normal vector of the Voronoi edge $\vec{n}_e$ points in the counterclockwise direction relative to the triangle and $-1$ otherwise.

The vorticity is naturally positioned at the triangle circumcentres [3], but can also be thought in similar way to live at the edges, if calculated from pairs of triangles (see [23]). The latter is equivalent to the former if the vorticity is linearly interpolated from the triangles to edge.

The accuracy analysis of [15] naturally extends to the curl operator, under the same hypothesis, but now given the information of the tangent components. It follows that on triangular cells this discretization is first order, provided that the tangent components of the velocities are at least second order accurately calculated at the midpoints of the edges of the triangles. This means that on the HCt grid, in which the velocities are know at the midpoints of the triangle edges, we have a first order approximation for the relative vorticity. On the HCm grid, the normal velocities at the midpoints of the Voronoi cells do not directly represent the tangents at the midpoints of the triangle edges (see Fig. 2). This will in general lead to an inconsistent discretization (see Appendix A for details). With a HR95 grid optimization the HCm grid positioning of the velocity tangent components approach that of the HCt grid and first order accuracy is obtained. The mimetic properties of the curl operator are preserved in the HCm grid.

As an overall view of the divergence and curl operators, we have that either one of them will be inconsistent on the HCm or HCt grids, if the HR95 optimization is not used. On HCt grids, the problem will occur on the divergence, whereas for the HCm grids, it will be in the vorticity. It is to be expected that, in general, the divergence will be more accurate on HCm positioning and the vorticity on HCt positioning, even on a HR95 grid, since the optimization does not match exactly the midpoints of the edges.

Considering the degrees of freedoms to be known in the mean integral forms (circulation or flux) leads to similar conclusions. The divergence is better represented if the degree of freedom adopted is the mean volume flux, whereas the vorticity is better represented if the adopted degree of freedom is the mean edge circulation. In this case, given one of the interpretations, if the other is obtained using a diagonal matrix with values relative to the primal and dual edge lengths [4], then they are similarly susceptible to the inconsistencies discussed for the HCm and HCt positionings.

Finally, with the HR95 optimization both the discrete divergence and curl operators are consistent with first order accuracy.

*Potential vorticity*

To obtain the potential vorticity, the fluid height is needed at the triangle circumcentres. In Ringler et al. [3], they use the areas of the intersection of the primal (Voronoi) and dual (triangular) cells as interpolation weights. The layer depth $h$ is calculated at the triangle centres as

$$h_v = \frac{1}{A_k} \sum_i a_{iv} h_i, \tag{11}$$

with areas as shown in Fig. 4 (right panel).

As mentioned before, this actually leads to a non-interpolatory first order approximation. The benefits of this formula is that it plays an important role to ensure the existence of compatible fluid height and potential vorticity auxiliary equations on the dual grid. A more accurate choice is to use linear barycentric coordinates interpolation, as described before for the divergence operator (but now we interpolate to the triangle circumcentre), resulting in a second order approximation. The final form of the potential vorticity will be given by

$$q_v = \frac{\eta_v + f_v}{h_v}, \tag{12}$$

where $f_v$ if the Coriolis parameter calculated at the cell vertices (triangle circumcentres).

In Fig. 5 we show the errors obtained for the potential vorticity ($q_v$) calculation of the initial conditions of test case 2 of [22] for 4 approaches: The first order interpolation of $h$ (eq. (11)), with HCt positioning and SCVT or HR95 grids (TRSK-HCT-SCVT and TRSK-HCT-HR95), and the second order linear interpolated $h$, using barycentric coordinates, with HCm positioning and SCVT or HR95 grids (MODF-HCM-SCVT and MODF-HCM-HR95). All 4 methods were used together with the usual discrete curl discretization of equation (10).

Except with the HCm SCVT grid, all other methods result in similar accuracy with first order convergence (in the maximum norm). The main problem with the HCm SCVT approach is that the relative vorticity will be inconsistent. Although the vorticity is expected to be more accurately calculated on HCt positioning, we see that because of the second order interpolation of $h$, the modified MODF-HCm-HR95 presents the best results. Of course the barycentric interpolation could be combined with the HCt positioning from TRSK, which should improve its accuracy.

### 3.3. Kinetic energy

For the gradient term ($G_e$) the kinetic energy is required at the cell nodes. Ringler et al. [3] suggests a discretization that leads to total energy conservation, defining

$$K_i = \frac{1}{4A_i} \sum_e l_e d_e u_e^2, \tag{13}$$

where $e$ varies within the edges of the Voronoi cell $i$ and $d_e$ and $l_e$ are defined as before as the triangle and Voronoi edge lengths, respectively. Energy conserving discretizations for the kinetic energy are possible if deduced from different ways of calculating the mass flux (see [8]). For the mass flux defined in section 3.1, the form given above is unique (see [3]). In general, this discrete version of the kinetic energy may not be consistent with the analytic kinetic energy in arbitrary non-regular Voronoi cells.

Perot [24] showed that this discretization is first order accurate (in the maximum norm) on triangular staggered grids, when the dual grid edge crosses the triangle edge at its midpoint. The analysis for the Voronoi staggered grid follows from the fact that for a constant vector field ($\vec{u}_0$), the kinetic energy can be represented exactly by

$$\frac{1}{2}\vec{u}_0 \cdot \vec{u}_0 = \frac{1}{2A_i} \sum_e u_{0e}(\vec{u}_0 \cdot (\vec{x}_e - \vec{x}_i)) l_e, \tag{14}$$

where $\vec{x}_e$ is the midpoint of the Voronoi cell edge and $u_{0e} = \vec{u}_0 \cdot \vec{n}_e$. If the midpoints of the Voronoi and triangular cell edges coincide, then, $\vec{u} \cdot (\vec{x}_e - \vec{x}_i) = u_e d_e/2$, resulting in the formulae shown in (13). This indicates the importance of a HR95 optimization to obtain a first order accurate approximation of the kinetic energy. We show an example of this in Fig. 5, for a solid body rotation velocity field, where for a grid optimized with SCVT and this kinetic energy scheme (indicated as TRSK-HCT-SCVT), no convergence is obtained in the maximum error, whereas for a HR95 optimized grid (indicated as TRSK-HCT-HR95), first order is observed even in the maximum error.

To obtain a consistent kinetic energy approximation in general, the $u_e$ values should be combined taking into account cell geometry features. Notice that $u_e$ carries only partial information about the vector field and that the kinetic energy depends on the full vector defined on the cell node. Several possibilities of reconstructing the full velocities to cell nodes were investigated in Peixoto and Barros [18]. For instance, Perot's method may be used to first reconstruct the velocities at cell nodes,

$$\vec{u}_i = \frac{1}{A_i} \sum_e (\vec{x}_e - \vec{x}_i) u_e l_e, \tag{15}$$

and then the kinetic energy at the nodes can be obtained as

$$K_i = \frac{1}{2}\vec{u}_i \cdot \vec{u}_i. \tag{16}$$

On a HCt grid, or on a grid with coinciding edge midpoints, $(\vec{x}_e - \vec{x}_i) = \vec{n}_e d_e/2$, and

$$\vec{u}_i = \frac{1}{A_i}\sum_e u_e l_e d_e \vec{n}_e/2. \tag{17}$$

The resulting expression will in general not simplify to the energy conserving $K_i$ shown before.

An important point is that Perot's reconstruction is only accurate (first order in the maximum norm) if the normal velocities are given at the midpoints of the Voronoi edges (HCm positioning). This is crucial, and using HCt grids greatly deteriorates its accuracy, so it should rather be used with a HCm grid or at least with HR95 optimized grids. Although the method is formally first order accurate only, as shown in [24], it is second order accurate on specially shaped cells, called aligned cells, as shown in [18]. In fact, it attains second order except in a minority of Voronoi grid cells (e.g. the pentagons). This can be seen in Fig. 5, bottom line, where this modified scheme (indicated as MODF-HCM-SCVT and MODF-HCM-HR95, for the 2 types of grid optimizations SCVT and HR95) reveals better accuracy than the scheme developed in [3], in spite of not conserving total energy.

### 3.4. Gradient term

The gradient term ($G_e$) can be approximated directly on orthogonal grids at the midpoint of a triangle edge, as

$$G_e = \frac{g}{d_e}(h_i + b_i - h_j - b_j) + \frac{1}{d_e}(K_i - K_j), \tag{18}$$

where $i$ indicates the index of the Voronoi cell which the reference normal edge vector points to, and $j$ indicates the opposite cell. On orthogonal Voronoi grids, this differencing operation is centred and therefore is second order accurate at the midpoint of the triangle edge (which is the case for HCt grids). It is only first order accurate if the midpoint of the Voronoi cell edge and of the triangle do not coincide (which is the case for general HCm grids).

On the evaluation of the gradient term, the first term on the right hand side of equation (18) is accurately calculated because $h_i$ is known at the Voronoi cell nodes and no approximation is required. The accuracy of the second term on the right hand side also depends on how the kinetic energy is calculated. If it is computed inconsistently, one can even observe an accuracy of order $-1$ for the gradient term, due to the division by a grid length scale ($d_e$), that is, the error may grow with increasing resolution. We can verify this numerically for the energy conserving version of the kinetic energy calculation (given in (13)) when the SCVT grid is used, as shown in Fig. 6 for the maximum errors (indicated with TRSK-HCT-SCVT). If the energy conserving scheme is used with a HR95 optimization, the kinetic energy will be first order accurate in the maximum norm, but one can still obtain an inconsistent gradient, as seen in Fig. 6 for the method indicated as TRSK-HCT-HR95. In order to guarantee a first order discretization of this gradient term, the kinetic energy should be calculated with second order accuracy.

On a HCm positioning, Perot's method is almost second order accurate. Using it together with a HR95 optimization, where the gradient differencing will also approach second order, the accuracy of this term improves, as can be seen in Fig. 6 under the name MODF-HCM-HR95 scheme. The attained accuracy is more than 4 orders of magnitude better in this test than with the energy conserving scheme on an SCVT grid on finer grids and 3 orders of magnitude better than with HR95 grid optimization.

Although Perot's vector reconstruction scheme presents second order accuracy on the great majority of the hexagonal cells, the scheme may still be only first order accurate on the maximum norm depending on the test case selected (see [18] for an example). As a drawback, the gradient term would be prone to being locally inconsistent where the reconstruction is only first order accurate. We investigated the use of second order accurate reconstruction schemes such as in [18], but we have seen little impact in global accuracy, at a price of larger stencils for the reconstruction. We analysed a linear least squares method and the hybrid scheme proposed in [18], and even though the hydrid scheme has little overhead, we did not notice much advantage in using these methods. For this reason, we did not employ these methods further in the shallow water models. However, they would be required to ensure a first order accurate local truncation of the gradient term for any velocity field.

Considering the integral interpretation of the degrees of freedom, as in [4], the gradient is considered in integral form. Therefore, a first order accurate kinetic energy would be enough to obtain a first order accurate integral of gradient. Nevertheless, once the integral form is normalized to ensure the correct velocity physical units, which is done dividing the integrals by reference edge lengths (in the case of the velocity degrees of freedom), we see that it would be desired to have the integral of the gradient having second order accuracy to ensure that in physical velocity units the method is consistent.

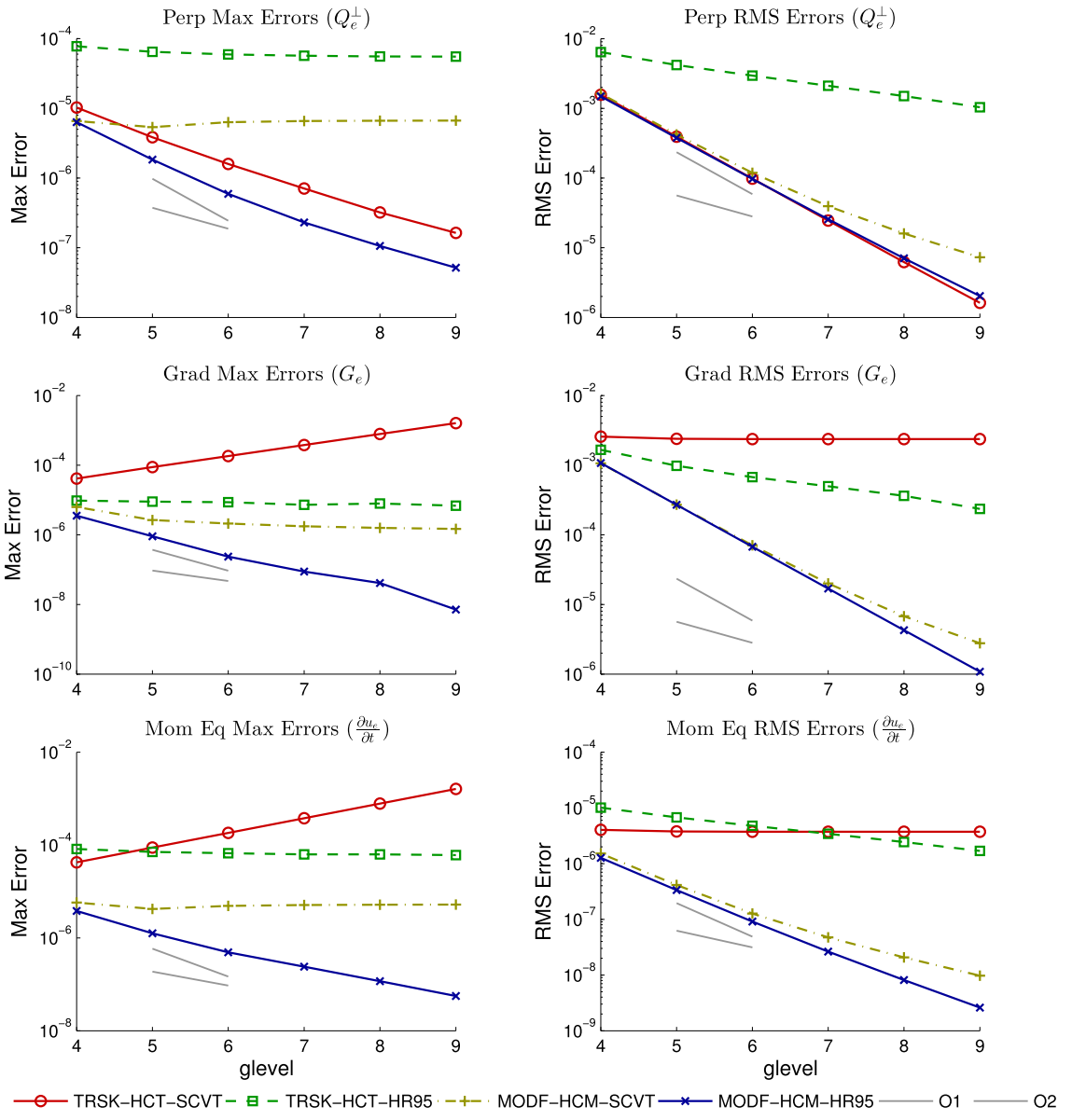


**Fig. 6.** Maximum and RMS errors of discrete operators on test case 2 initial conditions. Top line shows the perpendicular operator errors (relative to $Q_e^\perp$), middle line shows the gradient errors (relative to $G_e$), and the bottom line shows the overall momentum tendency errors (relative to $\frac{\partial u_e}{\partial t}$) varying with resolution. See text for a description of the methods used. Grey lines indicate first and second order reference convergence rates.

### 3.5. Coriolis term

Here we will analyse several aspects of the discretization of the perpendicular term ($Q^\perp$). We will start investigating the consistency conditions for the tangential velocity reconstruction ($\vec{u}^\perp$). Then we will show the reconstruction proposed by Thuburn et al. [2] and analyse its (in)consistency. The necessary conditions for energy conservation are then investigated and, finally, we will show an alternative energy conserving discretization for the perpendicular term.

*General theory for consistency*

To begin with, we consider an arbitrary planar C-grid, with the normal velocity components stored at the edge midpoints. For an edge $e$ of the grid, we denote as $\vec{n}_e$ a unit vector normal to the edge (in an arbitrary direction), and $\vec{t}_e$ the unit vector tangent to the edge, such that $\vec{t}_e = \vec{k} \times \vec{n}_e$ and $\vec{n}_e = \vec{t}_e \times \vec{k}$, where $\vec{k}$ is a unit vector normal to the plane and $\times$ is the usual cross product. For a vector field $\vec{v}$, the available information is the velocity component $u_e = \vec{v}(\vec{p}_e) \cdot \vec{n}_e$, where $\vec{p}_e$ is the midpoint of the edge $e$.

For a reconstruction method to be at least first order accurate, it is sufficient (and desirable) that it reproduces exactly constant vector fields on the plane. This means that if it is known only on 2 edges, and these are not parallel, then the vector field is uniquely defined. The key point is to notice that the normal data given at an edge ($u_e$) plays a role in reconstructing the tangent component ($u_e^\perp$), although they are defined for orthogonal vectors, as illustrated in Appendix B. This is a consequence of having only partial information given (only the normal components of the vector field known).

If we impose initially that the reconstructed tangent vector component will be obtained directly as [2],

$$u_e^\perp = \sum_{e' \in \mathrm{NB}(e)} w_{ee'} u_{e'}, \tag{19}$$

for some (yet) unknown weights $w_{ee'}$ relative to a set of neighbour edges NB($e$), then, since for a constant vector field $\vec{v} = u_e \vec{n}_e + v_e \vec{t}_e$, we must have $v_e = u_e^\perp$ (given the values of $u_{e'} = \vec{v} \cdot \vec{n}_{e'}$), the following condition results,

$$\vec{v} \cdot \vec{t}_e = u_e^\perp = \sum_{e' \in \mathrm{NB}(e)} w_{ee'} \vec{v} \cdot \vec{n}_{e'} = \vec{v} \cdot \left( \sum_{e' \in \mathrm{NB}(e)} w_{ee'} \vec{n}_{e'} \right),$$

and therefore,

$$\vec{t}_e = \sum_{e' \in NB(e)} w_{ee'} \vec{n}_{e'}. \tag{20}$$

This condition defines a unique solution for a triangular C grid case when only 2 edge informations are used to calculate a reconstructed tangent component for the remaining edge, as we have only 2 unknown weights. As it is unique, the constant reconstructed vector field is the same obtained from a lowest order (0th order) Raviart–Thomas element (see [18] for details). The triangular C-grid case has been well analysed, for example, in [25,26] and [27], and further details are provided in Appendix B.

A method satisfying the condition derived above maybe combined consistently at an edge common to cells $i$ and $j$ as

$$u_e^\perp = a_{e,i} u_{e,i}^\perp + a_{e,j} u_{e,j}^\perp, \tag{21}$$

where $u_{e,i}^\perp$ and $u_{e,j}^\perp$ are the reconstructions at edge $e$ using the $i$-th and $j$-th Voronoi cells set of edges. The weights $a_{e,i}$ and $a_{e,j}$ are positive and should add to 1 to preserve consistency. This description is appropriate to clarify certain properties, but these weights may be absorbed in the reconstruction weights and the method can be equivalently viewed as a single stencil method.

As shown in [15], spherical polygons can be locally approximated by planar polygons tangent to the sphere. Proceeding as in [15], we can extend the consistency conditions derived above (first order conditions) for the tangent reconstruction on the sphere. The spherical compact stencils may be stated as the ones in the planar grid, but now considering the normal and tangent vectors as vectors in $R^3$ tangent to the sphere, and substituting the edge lengths and polygonal areas by the geodesic ones.

*TRSK scheme*

The TRSK scheme, proposed in Thuburn et al. [2], uses two arbitrary Voronoi cells to define the reconstruction stencil. The method defines

$$u_e^\perp = u_{e,i}^\perp + u_{e,j}^\perp, \tag{22}$$

with $i$ and $j$ being two neighbour Voronoi cells sharing the edge $e$, with $u_{e,i}^\perp$ (and respectively $u_{e,j}^\perp$) calculated using the following weights,

$$w_{ee'} = c_{ee'} \frac{l_{e'}}{d_e} \left( \frac{1}{2} - \sum_v \frac{A_{iv}}{Ai} \right) n_{e'i}, \tag{23}$$

where $A_{iv}$ is the area of the quadrilateral defined by the cell $i$ centre, the vertex $v$ and two adjacent cell edge midpoints (which is also the overlapping area between the Voronoi cell and the dual triangular cell), the sum is within the vertices $v$ between edge $e$ and $e'$ and there is a sign correction for the normal components, such that they point outward ($n_{e'i} = \pm 1$) and a general sign correction ($c_{ee'} = \pm 1$) relative to the orientation of the tangent vector at $e$ with respect to the dual triangular cell (see [2] or [11] for details).

Although the scheme is consistent for rectangles and regular hexagons, in general, it does not satisfy the consistency conditions. On spherical geodesic grids, it is not possible to obtain only regular Voronoi cells, and the method is after all inconsistent in the maximum norm. On the sphere, the edge lengths and areas are geodesic.

To analyse whether the method obeys the consistency condition, we calculate the numerical tangent ($\vec{t}_e^n$) vector (using the weights (23) on condition (20)) and compare it with the actual tangent vector ($\vec{t}_e$), defining for each edge an inconsistency index $\chi_e = \|\vec{t}_e^n - \vec{t}_e\|$, with the usual Euclidean norm in $\mathbb{R}^3$. On a planar grid, consistency should be achieved if, and

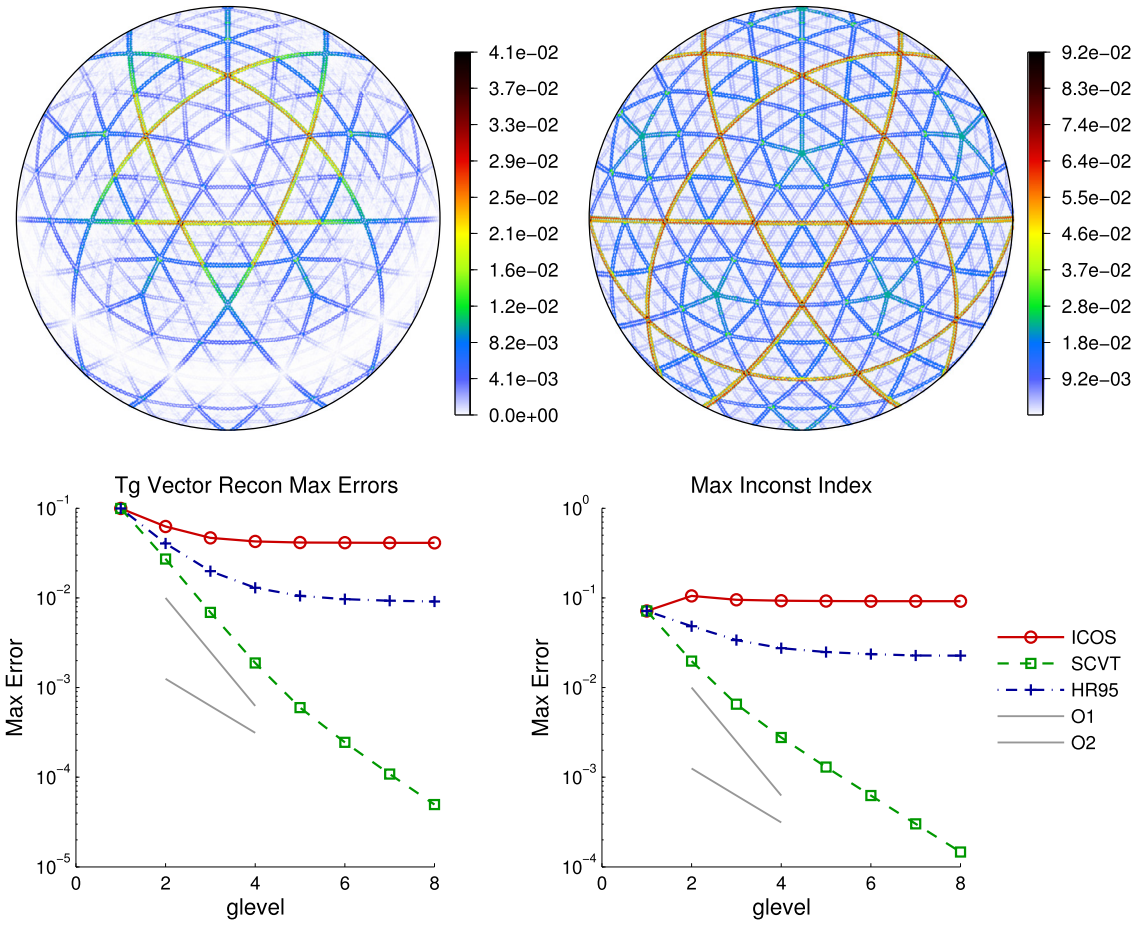


**Fig. 7.** On the top left it is shown the error on the vector reconstruction of the tangent component using the TRSK scheme on a HCt positioning of a solid body rotation vector field (test case 2 of [22]) and on the top right the inconsistency index for TRSK scheme. The plot was centred at latitude 65 degrees and zero longitude and used an icosahedral grid level 6 without optimization. On the bottom line, we show convergence rates for the maximum error (left) and the inconsistency index (right) for non-optimized grids (ICOS), SCVT and HR95 optimized grids. A HCt positioning was used in all cases.

only if, $\chi_e = 0$. On the spherical grid, $\chi_e$ should converge to zero for the method to be consistent. Therefore, it can be used to locally evaluate the potential accuracy of the method.

In Fig. 7 we display $\chi_e$ on an icosahedral grid level 6, without optimization (top right panel) with clear patterns of grid imprinting. We can also observe in Fig. 7 (top left) that the inconsistency is related to the error of the reconstruction of the tangent velocity component of a simple solid body rotation from Williamson [22] test case 2, presenting larger errors where the approximation of tangent vectors is worse.

On the bottom line of Fig. 7 we show how the errors of the tangential reconstruction and the values of the inconsistency index vary on different grid levels. Inconsistency is clear for the non-optimized (ICOS) and HR95 grids, but surprisingly, on the SCVT grids, the method not only converges for this test case, but the inconsistency index indicates that it should converge for any smooth vector field (at least up until the grid level tested here). For these tests we used a HCt positioning, but similar results were observed for the HCm positioning (not shown).

*Energy conservation conditions*

The discrete system must satisfy an analogous condition to $h\vec{u} \cdot (qh\vec{u}^{\perp}) = 0$ to enable energy conservation with respect to the Coriolis term. On usual C grids, the velocity is known on the edges, while $q$ and $h$ are not. Because $\vec{u}^{\perp}$ needs to be calculated from neighbour edges, this conditions cannot be fulfilled on each edge. Nevertheless, the integral of the quantity $h\vec{u} \cdot (qh\vec{u}^{\perp})$ should vanish over the entire grid to ensure energy conservation.

A natural way of considering this integral over the discrete grid is to assume constant piecewise values over edge areas, resulting in the following condition for energy conservation,

$$\sum_e A_e q_e h_e^2 u_e u_e^{\perp} = 0, \tag{24}$$

where $A_e$ is an area associated with edge $e$, and $q_e$, $h_e$ are respectively the potential vorticity and fluid depth at the edge, and the sum is over all grid edges.

Using a reconstruction scheme to calculate $u_e^\perp$, leads to the following condition,

$$\sum_e \sum_{e' \in NB(e)} w_{ee'} A_e q_e h_e^2 u_e u_{e'} = 0. \tag{25}$$

With the above condition, it is clear that $u_e$ should not participate in the calculation of $u_e^\perp$, otherwise a $u_e^2$ term will appear and will not allow cancellation for arbitrary vector fields. A sufficient condition to satisfy equation (25) is

$$w_{ee'} A_e q_e h_e^2 = -w_{e'e} A_{e'} q_{e'} h_{e'}^2. \tag{26}$$

The fluid depth is usually stored in the cell centre on C-grids, and the potential vorticity will naturally be defined on the dual grid cell centres (primal grid cell vertices). Therefore, $q_e$ and $h_e$ have to be approximated from their given values to the edges. Since the weights of the reconstruction ($w_{ee'}$) should depend only on the geometry of the grid, this implies that the approximation of $q$ and $h$ at the edges $e$ and $e'$ should be equal for any pair of edges belonging to same stencil (or cell), which leads to a constant $q$ and $h$ for all grid edges.

An alternative is to define

$$\vec{v} = hq\vec{u} \tag{27}$$

and restate the energy conservation condition for the Coriolis force as

$$\sum_e A_e u_e h_e v_e^\perp = 0. \tag{28}$$

Since $hq$ is a scalar field, in the continuous case these two conditions are equivalent, but will be different on the discrete system. In this case, $\vec{v}_e^\perp$ can be estimated analogously to $u_e^\perp$ as

$$v_e^\perp = \sum_{e' \in NB(e)} w_{ee'} q_{ee'} h_{e'} u_{e'}, \tag{29}$$

where $q_{ee'}$ refers to the value of $q$ at edge $e'$ used in the calculation of their values at the edge $e$, and $h_{e'}$ refers to value used in the calculation of $h$ at edge $e'$ (independently of $e$). Now a sufficient condition to conserve energy may be given as

$$w_{ee'} A_e q_{ee'} = -w_{e'e} A_{e'} q_{e'e}. \tag{30}$$

To obtain reconstruction weights independent of $q$, $q_{ee'}$ should be equal to $q_{e'e}$. This is easily obtained with a linear combination of $q_e$ and $q_{e'}$, as

$$q_{ee'} = q_{e'e} = \gamma_e q_e + \gamma_{e'} q_{e'}, \tag{31}$$

with $\gamma_e + \gamma_{e'} = 1$ to preserve consistency. Ringler et al. [3] use $\gamma_e = 1/2$.

Consequently, the conditions for the reconstruction weights to ensure energy conservation are that

$$w_{ee'} A_e = -w_{e'e} A_{e'}. \tag{32}$$

If the reconstruction method is decomposed in two stencils, with a reconstruction for each cell adjacent to the edge, then this condition may be restated as

$$a_{e,i} w_{ee'} A_e = -a_{e',i} w_{e'e} A_{e'}, \tag{33}$$

which gives some degrees of freedom to choose $a_{e,i}$ so that this condition is satisfied.

It is straightforward to verify that the TRSK scheme of [2,3] satisfies this condition if the perpendicular term is discretized as

$$Q_e^\perp = \sum_{e'} w_{ee'} h_{e'} u_{e'} q_{ee'}, \tag{34}$$

with the sum over the relevant $e'$ edges of the two neighbour Voronoi cells with weights given by equation (23). $q_e$ can be obtained by linear interpolation from the circumcentres of the triangles that define this Voronoi edge. In Ringler et al. [3], this is done using equal weights (1/2), which results in a first order interpolation in general, since the interpolation point does not coincide with the midpoint of the edge on HCt grids, but simple linear interpolation could be used instead. On the other hand, on HCm grids, this will always be second order accurate.

*Alternative scheme*

The consistency conditions together with the energy conserving conditions do not seem to provide a direct way to derive consistent energy conserving methods, although they give key insights for specific polygonal shapes, such as triangles and rectangles (not shown). We will discuss an approach derived from some ideas in [27], essentially generalizing them using Perot's reconstruction method [24] for arbitrary polygons. Another possibility would be to consider the reconstruction on the dual triangular grid, for which the solution is known and unique, and interpolating to the edge. This latter approach leads to a similar method as in [28], which is known not to perform well with respect to keeping balanced flows [2], and it is more inaccurate the former scheme, even though it is first order accurate (the method is described in details in the Appendix B). Therefore, for now, we will proceed with the first approach only.

Perot's [24] method to reconstruct the velocity to the Voronoi cell nodes (equation (15)) can be modified to fulfil the energy conservation conditions, including the potential vorticity and the layer thickness into the method in the following way. Let $e$ be the edge at which we wish to calculate the perpendicular operator $Q_e$. Then, the quantities

$$\overrightarrow{(hqu)}_i^e = \frac{1}{A_i}\sum_{e'}(\vec{x}_{e'} - \vec{x}_i)u_{e'}h_{e'}q_{ee'}l_{e'}, \tag{35}$$

are calculated for each cell $(i)$ that shares the edge $e$, with the quantities $q_{ee'}$ and $h_{e'}$ defined as stated before for the TRSK scheme. Then the perpendicular term can be obtained as

$$Q_e^{\perp} = \frac{1}{2}(\overrightarrow{(hqu)}_i^e + \overrightarrow{(hqu)}_j^e)\cdot\vec{t}_e, \tag{36}$$

for $i$ and $j$ cells that share the edge $e$. All operations performed here preserve the first order accuracy of Perot's original reconstruction method.

The method maybe restated explicitly in the notation used before with the weights

$$w_{ee'} = \frac{1}{2}\frac{l_{e'}}{A_i}(\vec{x}_{e'} - \vec{x}_i)\cdot\vec{t}_e, \tag{37}$$

for $e'$ ranging over the edges of both cells that share the edge $e$.

To have energy conservation, we need that $w_{ee'}A_e = -w_{e'e}A_{e'}$. On a triangular C grid, with $\vec{x}_i$ defined on the circumcentres, $(\vec{x}_{e'} - \vec{x}_i) = d_{e'}\vec{n}_{e'}/2$. Therefore, assuming $A_e = d_e l_e/2$, and using that $\vec{n}_{e'}\cdot\vec{t}_e = -\vec{n}_e\cdot\vec{t}_{e'}$, energy conservation is obtained. On arbitrary HCt grids, the orthogonality and Voronoi properties help to obtain the same cancellations; energy conservation is again achieved. Nevertheless, Perot's method is inconsistent on HCt grids, since it requires the velocities to be given at the midpoints of the Voronoi edges, and better accuracy is obtained on HCm grids. In this case, the orthogonality property will hold only approximately, $(\vec{x}_{e'} - \vec{x}_i) = d_{e'}\vec{n}_{e'}/2 + \epsilon_{e'}\vec{t}_{e'}$, where $\epsilon_{e'}$ is a constant that depends on the distance between the midpoints of the Voronoi and triangle edges with index $e'$. On HR95 optimized grids, $\epsilon$ is very small and tends to zero with increasing resolution, so that the energy conserving condition is met within an acceptable error bound. We will see numerically that in fact this approximation will have very small impact on energy conservation.

We compare the accuracy of the TRSK scheme with this alternative scheme in Fig. 6 (top line). The methods investigated are as follows: TRSK-HCT-SCVT and TRSK-HCT-HR95 use the TRSK tangent reconstruction with first order interpolations of $h$ for $q$, on HCt SCVT and HR95 grids, respectively. MODF-HCM-SCVT and MODF-HCM-HR95 employ the alternative tangent reconstruction with second order interpolations of $h$ for $q$, on HCm SCVT and HR95 grids.

We see that both the TRSK scheme on an HR95 grid and the alternative scheme (MODF) on an SCVT grid do not converge in the maximum norm. The first happens due to the inconsistency of the tangent reconstruction, and the latter, due to the inconsistency of the vorticity calculation (and consequently of the PV calculation). The TRSK scheme converges on the SCVT grid if a HCt grid is used, as predicted before by the inconsistency index. The best accuracy was noticed with the new scheme (MODF) on a HCm HR95 grid, where the vorticity (and consequently the PV) is consistent and so is the tangent velocity reconstruction.

### 3.6. Overall local truncation errors

To summarize the results obtained so far we have the following requirements for each operator, where we are always referring to the accuracy order with respect to the maximum norm.

- Divergence ($D_i$): Needs HCm grid for consistency, or a HR95 optimized grid.
- Vorticity ($\eta_v$): Needs HCt grid for consistency, or a HR95 optimized grid.
- Kinetic energy ($K_i$): Has to be 2nd order accurate to provide first order accurate gradient. First order accuracy may be obtained using the energy conserving method on a HR95, whereas on SCVT grids it is inconsistent. Second order accurate methods may be used instead.
- Gradient ($G_e$): Relies on the accuracy of the kinetic energy to be first order accurate.
- Perpendicular ($Q_e^{\perp}$): Depends on 2 main points:

**Table 1**
Approximate accuracy order using maximum and RMS norms (indicated by MAX/RMS) of each approach with respect to each operator investigated (based on the finer grids).

| Operator | TRSK | | MODF | |
|---|---|---|---|---|
| | HCt-SCVT | HCt-HR95 | HCm-SCVT | HCm-HR95 |
| Divergence ($D_i$) | 0 / 1 | 1 / 2 | 1 / 2 | 1 / 2 |
| P. vorticity ($q_v$) | 1 / 1 | 1 / 1 | 0 / 1 | 1 / 1 |
| Kin. energy ($K_i$) | 0 / 1 | 1 / 1 | 1 / 2 | 2 / 2 |
| Gradient ($G_e$) | −1 / 0 | 0 / 1 | 0 / 1 | 1 / 2 |
| Tang Vel ($u_e^{\perp}$) | 1 / 2 | 0 / 1 | 1 / 2 | 1 / 2 |
| Perp ($Q_e^{\perp}$) | 1 / 2 | 0 / 1 | 0 / 1 | 1 / 2 |
| Overall Mass ($\frac{\partial h_i}{\partial t}$) | 0 / 1 | 1 / 1 | 1 / 2 | 1 / 2 |
| Overall Mom ($\frac{\partial u_e}{\partial t}$) | −1 / 0 | 0 / 1 | 0 / 1 | 1 / 2 |

- Requires the vorticity to be consistent, therefore a HCt grid or an HR95 optimized grid.
- Requires the tangent velocity reconstruction to be consistent. This may be obtained with the TRSK method on an SCVT grid or with the new alternative scheme.

We show in Fig. 6 the errors for the overall truncation error of horizontal discretization of the shallow water model for the momentum equation for 4 sets of schemes, combining what has been discussed so far. The TRSK-HCT-SCVT and TRSK-HCT-HR95 are using exactly the methodology proposed in [2] and [3] on these 2 optimized grids with the HCt positioning. The MODF-HCM-SCVT and MODF-HCM-HR95 schemes use the linear interpolation formulas for the layer depth interpolation, Perot's method for the kinetic energy reconstruction and the new scheme for the tangential component reconstruction with a HCm positioning on both SCVT and HR95 grids. Many other combinations of methods and grids were investigated, but this set is representative of the accuracy properties of the analysed schemes.

The truncation errors of the horizontal mass equation are dominated by the divergence term ($D_i$). Therefore, it should be first order accurate in the maximum norm if either a HCm grid is used, or if a HR95 optimization is used. The kinetic energy problem leads to a severe accuracy problem for the SCVT grid TRSK scheme, with increasing errors on finer grids. Both the TRSK scheme on the HR95 grid and the MODF scheme on the SCVT grid are overall inconsistent (0th order accurate in the maximum norm), mainly due to the inconsistency of the perpendicular operator, but also due to the gradient term.

We summarize the local truncation error results in Table 1, where we show the accuracy order with respect to the maximum norm, known theoretically from our previous analysis, and also the RMS norm, approximated from the experimental results.

The only scheme showing first order accuracy in the overall maximum norm is the MODF scheme on the HR95 grid with the HCm positioning. We again would like to point out that by using Perot's reconstruction method for the kinetic energy term, second order is achieved almost everywhere in the grid, but it is possible to find test cases where it reveals its locally first order accuracy, which could lead to an inconsistency for the gradient term. This is simple to be corrected, using the second order methods analysed in Peixoto and Barros [18], but we present here only results using Perot's method as it already sufficiently improves the accuracy and uses a more compact stencil.

The main inconsistency problems are grid related, and therefore are likely to be only dominant in certain grid regions and could be cancelled out in time evolving problems. Nevertheless, looking at the RMS overall errors (Fig. 6), we notice that even then the TRSK method fails to be first order accurate on an SCVT grid, and is slightly under first order accuracy for the HR95 grid. The new scheme shows RMS errors of second order accuracy and several orders of magnitude smaller errors, with essentially the same computational costs, since it uses similar stencils.

Other derived discrete operators not analysed here may play important roles in the overall accuracy of the scheme. For example the Laplacian of the fluid depth, which is obtained taking the discrete divergence of the gradient of $h$, is important when analysing the gravity waves. This discrete Laplacian operator was investigated in Heikes and Randall [17]. They concluded that to obtain first order local truncation errors it is important to have the midpoints of the triangle edges and the Voronoi edges converging to each other, and therefore they derived the HR95 grid optimization. Thus, on the SCVT grid it is expected to be 0th order accurate, and 1st order accurate on the HR95 grid, independently of the velocity positioning. This once again shows the importance of the HR95 grid optimization to obtain consistent local truncation errors.

If one considers the degrees of freedom as integral forms, as in [4], similar conclusions may be drawn if the analogies discussed in section 2.3 are taken into account, but some extrapolation of the analysis may be required if different equation formulations are used.

We repetitively used the orthogonality grid property in the analysis, therefore we do not expect this analysis to be valid in nonorthogonal grids. Nevertheless, it is possible to infer from this analysis some potential additional consistency problems one might encounter in nonorthogonal grids. For example, the divergence and vorticity would both no longer be consistent in nonorthogonal grids for input data given point-wisely.

As a final point for this section, we stress that the consistency orders observed here are not necessarily those that we will observe as convergence rates in the time evolving model. Cancellation of errors may take place as time evolves, hiding

possible inconsistencies. Also, supraconvergence effects may happen and we may see higher convergence orders than those observed in the consistency analysis. Therefore, the fact that the TRSK scheme, or the modified scheme in SCVT grids, are inconsistent with respect to some operators does not rule it being convergent. The convergence properties of the methods are the matter to be discussed in the next section.

## 4. Nonlinear shallow water model analysis

The TRSK scheme [2,3] has gone through several tests and the shallow water results have been reported in many papers (e.g. [2–5,11]), therefore we will address here only test cases where the modifications are of relevance.

We will use the test cases of Williamson et al. [22] and the nonlinear barotropic instability test case of Galewsky et al. [29] to verify the effects of the modifications with respect to accuracy and the desirable properties.

Test case 2 of [22] consists of a solid body rotation velocity field, and a fluid height that ensures a steady state to the nonlinear problem. Test case 5 consists of a zonal flow over an isolated mountain. Test case 6 addresses the Rossby–Haurwitz wave problem. The Galewsky et al. [29] test cases will be used in two different settings: (i) a zonal jet with balanced flow and no perturbation, and (ii) the same zonal jet but with perturbation in the layer thickness to trigger the instability.

Some numerical methodological problems can only be fully identified on a 3D model. This is the case, for example, for the triangular grid ICON model [30], where the checkerboard pattern arises due to a mode in the divergence discretization that passes onto the vertical velocity. To detect such problems earlier, Gassmann [31] suggests to use shallow water tests with smaller equivalent depths, and therefore slower gravity wave speeds but unchanged horizontal velocities. The proposed test is built for the f-plane shallow water equations and it does not seem to be easily transported to the rotating sphere. Following a discussion from Bell et al. [32], we will use a modification of test case 2 of Williamson et al. [22] that mimics the small equivalent depths suggested by Gassmann.

All tests were performed with a four stages fourth order Runge–Kutta method, with time steps small enough so that the dominant error is mainly due to the horizontal discretization, which is the main objective under investigation. The time steps range from 200 s on the coarser grid to 50 s on the finest grid tested (grid level 9, with approximately 15 km between cell centres).

As reference solution, we used a shallow water version of the MetOffice dynamical core ENDGame [33] (a semi-Lagrangian, semi-implicit, latitude–longitude grid model). We ran the model on a $2048 \times 1024$ grid, with an approximate resolution of 20 km near the equator. A small time step of 50 s was chosen in order to avoid slowing down the gravity waves (John Thuburn, personal communication). Therefore, at the highest resolution tested for the icosahedral grids, both models have similar resolution and time step.

The error norms used here are the same as in Ringler et al. [3], where the RMS error (or L2 error) was described in equation (9), and the maximum error is now normalized (divided by the maximum analytical value).

We choose 4 combinations of methods and grids to compare:

- TRSK-HCT-SCVT: The TRSK scheme as proposed in Ringler et al. [3] on a SCVT grid.
- TRSK-HCT-HR95: Same as TRSK-HCT-SCVT but on a HR95 grid.
- MODF-HCM-SCVT: The new scheme as described in the previous section using linear layer depth interpolation, second order kinetic energy reconstruction and the alternative scheme for the tangential component reconstruction with a HCm positioning on a SCVT grid.
- MODF-HCM-HR95: Same as MODF-HCM-SCVT but on a HR95 grid.

Other combinations of methods were also tested, and comments on these results will be depicted when relevant.

### 4.1. Preserved properties

Mass conservation is not affected by the modifications, even though the flux calculation has changed, since the same flux value is used in both cells sharing an edge. Therefore, local mass cancellation will still occur in the modified scheme. Numerical experiments with shallow water tests cases confirm mass errors within round off machine errors (not shown).

The curl and gradient operators are still the same as in TRSK, so the pressure gradient term should not produce any spurious sources of vorticity.

### 4.2. Test case 2 – steady state zonal geostrophic flow

The test case 2 (TC2) of Williamson et al. [22] defines a solid body rotation velocity field, with

$$u = u_0 \cos(\theta)$$

$$v = 0$$

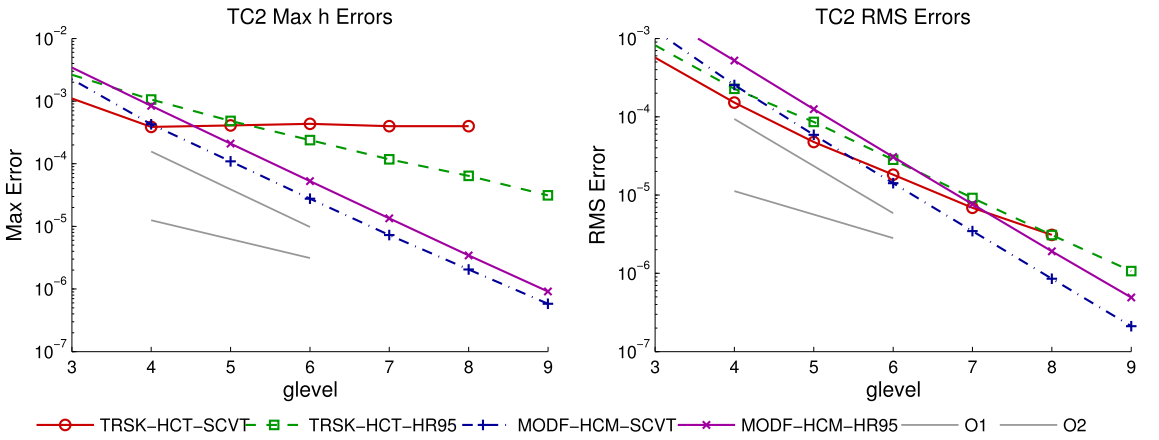


**Fig. 8.** Maximum and RMS errors of TC2 at day 12. TRSK indicates the methodology developed in [3] on HCt grids with either SCVT or HR95 optimizations. The MODF scheme indicates the proposed scheme with velocities given at midpoints of the Voronoi cell edges (HCm grid) with the modifications described in the text with either SCVT or HR95 grids.

where $\theta$ is the latitude, and $u$, $v$ are the zonal and meridional wind components. $u_0 = 2\pi a/1{,}036{,}800$, where $a = 6.37122 \times 10^6$. The height field is defined to be in balance with the wind,

$$h = h_0 - \frac{1}{g}\left(a\Omega u_0 + \frac{u_0^2}{2}\right)\sin^2(\theta), \tag{38}$$

where $h_0 = 2.94 \times 10^4/g$, $g = 9.80616$, $\Omega = 7.292 \times 10^{-5}$. The initial condition should remain over all time integration, as the problem is steady state, therefore the analytical solution is known for this test case. The dynamical regime under investigation here has, approximately, a grid Rossby number of $Ro = \frac{u_0}{f\Delta x} = 8.8$ and Froude number of $Fo = \frac{u_0}{\sqrt{gh_0}} = 7 \times 10^{-2}$, where we consider a grid resolution of $\Delta x = 30$ km (grid level 8) and Coriolis parameter of $f = 2\Omega$.

In Fig. 8 we show error measurements of the fluid height field for test case 2. For the 12th day of integration, the modifications increase the accuracy up to an order with respect to TRSK scheme on HR95 optimized grids and several orders of magnitude with respect to the SCVT TRSK scheme. The differences are obviously larger on very fine grids. These results are comparable to the ones in Fig. 7 of [3]. The only method with first order local truncation errors tested is the MODF-HCM-HR95. Nevertheless, we notice that the MODF-HCM-SCVT shows second order convergence rates. The main term with inconsistency for the modified scheme on SCVT grids is the vorticity discretization, which is not a dominant error term in this test case.

Interestingly, the lack of convergence of the TRSK scheme on SCVT grids is not due to the discretization of the perpendicular term ($Q_e^{\perp}$), because we showed in section 3.5 that, on SCVT optimized grids, this term is in fact first order accurate. Therefore, the convergence problem that shows up in test case 2 with SCVT grid optimization could only be due to the divergence term ($D_i$) or the kinetic energy term ($K_e$) problems (or both). In Ringler et al. [34], they used a SCVT HCt grid but initialized the normal velocity from the stream function, which ensures that the discrete divergence is initially zero and is similar (second order approximation) to the initialization on a HCm grid. This would potentially reduce the errors in the divergence term, but would lead to more inaccurate vorticity calculations. Nevertheless, they still observe non-convergence of the method, even in the RMS norm (see Fig. 8 of [34]), indicating that the problem could be related to the kinetic energy term. We confirmed this result in our experiments by using HCm positioning with the TRSK scheme and SCVT grids.

We performed additional experiments to understand the nature of the lack of convergence in the TRSK scheme using combinations of methods. First, we analysed the use of the modified kinetic energy ($K_e$) together with the TRSK scheme for the perpendicular term ($Q_e^{\perp}$) on a HCm SCVT grid. This leads to errors very similar to those of the MODF-HCM-SCVT scheme shown in Fig. 8, and shows that using a more accurate representation of the kinetic energy indeed removes the problem of lack of convergence (at least for the grid levels tested). It is interesting to note that the tangent velocity reconstruction of TRSK is first order accurate in this case (because we are on SCVT grids), so this could be a way to preserve the property of having stationary geostrophic modes on the f-sphere and still gain first order accuracy – at least for this test case and levels tested. We highlight, however, that both the gradient of kinetic energy and the vorticity are still inconsistent (see Table 1), since the HR95 grid optimization is not used.

We also analysed the use of the modified kinetic energy ($K_e$) with the TRSK perpendicular term ($Q_e^{\perp}$) scheme now on a HCm HR95 grid. In this case there, all terms are consistent except the perpendicular one (due to the TRSK scheme on HR95 grids). The errors are very similar to those shown for the TRSK-HCT-HR95 case in Fig. 8. This shows that what is stopping the TRSK scheme in HR95 grids from having better convergence rates is mainly the lack of consistency in the tangent velocity reconstruction method.

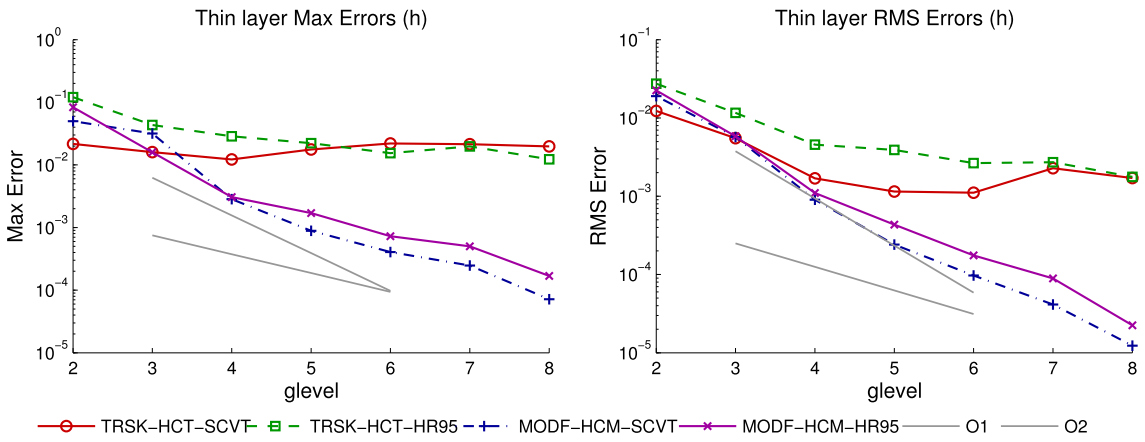


**Fig. 9.** Maximum and RMS errors of modified TC2 for a thin layer (small equivalent depth of 100 m) at day 12. Same methods used in Fig. 8.

Summarizing, we have observed that the lack of convergence of TRSK scheme on SCVT grids seems to be mainly due to the lack of consistency in the kinetic energy term ($K_e$). Whereas on HR95 grids, the slow convergence rates seem to be mainly attributed to the lack of consistency in the tangential velocity reconstruction scheme ($u_e^{\perp}$).

### 4.3. Steady state with thin layer

In the test case 2 results, we see that the TRSK scheme exhibits first order accuracy on HR95 grids, even though it has local truncation error of 0th order. The reason seems to be that the most inaccurately calculated terms, which are mainly nonlinear, such as the kinetic energy term, do not dominate the overall accuracy of the method as time evolves. Slowing down the gravity waves, but maintaining the horizontal wind speeds, would potentially reveal what is happening with the nonlinear term errors, and would also provide insights to what could happen on a 3D model with small equivalent depths.

To investigate the case when there is dominance of nonlinear term errors, Bell et al. [32] suggest reducing the equivalent depth of problem.[2] We will do this by modifying test case 2 in the following way. The velocity field is the same, but now the balanced layer thickness field is defined as the bottom topography $b$ (as defined in equation (38)), and a constant thin layer is defined for $h$. The bottom topography will make sure that the problem is still steady state, whereas the thin layer constant $h$ will simulate the small equivalent depth. We will adopt a constant layer with 100 m of height, which, when compared to the previous test case, gives a much higher Froude number ($Fo = 1.2$), but the same Rossby number. This indicates that the gravity waves are now slower than the previous test (test case 2), possibly making more visible the errors due to other terms in the equations (mainly nonlinear ones). Since the problem is steady state, we again have an analytical solution to evaluate the errors.

In Fig. 9 we show the error results for the layer of 100 m. The TRSK scheme clearly loses convergence, in both maximum and RMS errors, as now the errors in the inconsistent terms seem to dominate the convergence. The new scheme maintains first to second order accuracy, and 2 orders of magnitude smaller errors for the finer grids.

We notice in Fig. 10 that the errors are not just concentrated around the pentagons, but spread in a large global belt. The scales used in these plots had to be made separately, since the errors are orders of magnitude different.

As in test case 2, we also experimented using the TRSK scheme considering the HCm positioning and modified kinetic energy, as a way to preserve all its mimetic properties except the energy conservation due to the more accurate kinetic energy term calculation. On SCVT grids, the errors are very similar to those of the MODF-HCM-SCVT shown in Fig. 9, which shows convergence with increasing resolution. This indicates that the possible error dominating the lack of convergence is due to the inconsistency in the kinetic energy calculation. On the HR95 grids, the method still lacks convergence, and has errors similar to those indicated as TRSK-HCT-HR95 of Fig. 9. Since in this case the gradient of kinetic energy is consistent, as are the divergence and vorticity, this indicates that the convergence problem is possibly happening due to the inconsistency in the perpendicular term calculation ($Q_e^{\perp}$).

Although this is an idealized test case, it should be a fare way to capture the dominance of the nonlinear terms simulating what happens in 3D models with small equivalent depths. It therefore points out that possible convergence problems of the TRSK scheme could emerge in 3D models. We highlight that this could imply errors of several orders of magnitude higher in very fine grids when compared to the modified scheme (or other converging schemes).

[2] The main purpose of [32] is to investigate an instability related to the nonlinear terms. Here, we will use a layer deep enough so that we will not get the instability problem, but thin enough so that we can have dominance of nonlinear errors.

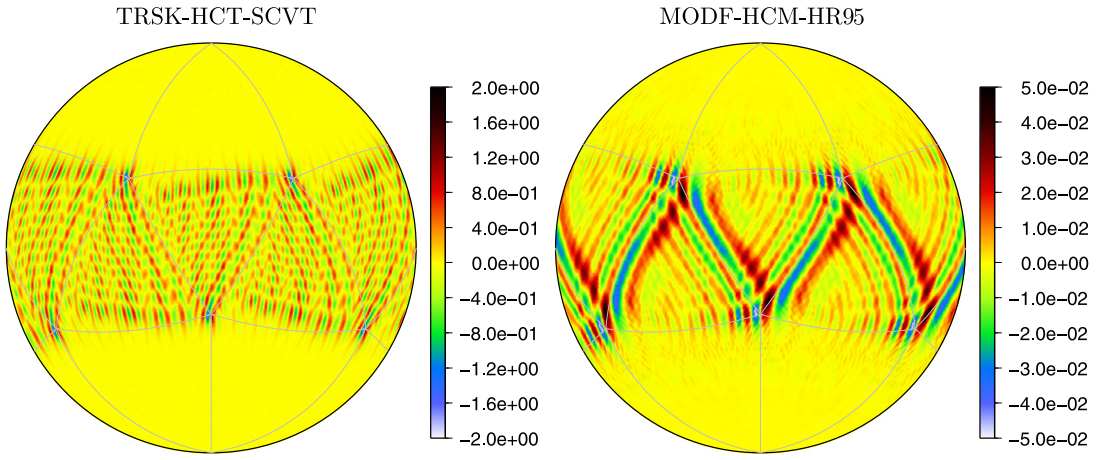


**Fig. 10.** Cellwise errors of the fluid height for the modified TC2 for a thin layer (small equivalent depth of 100 m) at day 12. On the left, the errors for the TRSK scheme on a HCt-SCVT grid, and, on the right, the errors for the modified scheme on a HCm-HR95 grid. Both cases use a grid level 7 (approximately 60 km). Please note that different scales are used for each plot, due to the difference in magnitude of the errors.

### *4.4. Normal mode linear analysis*

The modified scheme still relies on a hexagonal C grid, so it is expected to maintain accurate representation of inertial-gravity waves. To further investigate this aspect we calculated the normal modes for this scheme using the same approach as in [11]. We linearised the equation about a state of rest and a constant mean layer thickness of $gh_0 = 10^5$ m$^2$ s$^{-2}$. Two scenarios were tested: (i) constant $f$ (f-sphere), with $f = 1.4584 \times 10^{-4}$, and (ii) a rotating sphere with variable $f = 2\Omega \sin(\theta)$. This was analysed on a HR95 optimized grid level 5 (approximately 240 km grid resolution). The first scenario has Rossby deformation radius of $R_d = \frac{\sqrt{gh_0}}{f} \approx 2168$ km (slightly smaller than the Earth radius $a = 6.371$ km, but with same order of magnitude), and the chosen Coriolis parameter is relative to the north pole, therefore the highest Coriolis frequency.

Fig. 11 shows the frequencies obtained for both the geostrophic and inertial-gravity modes. On the f-sphere, the number of geostrophic and inertial-gravity modes is known (see [2]), so we ordered the frequencies to match these same number of modes. On the rotating sphere (variable $f$), these numbers are not known, but we assumed the same separation just for an easier comparison.

The higher frequency inertial-gravity modes for the different schemes are indistinguishable, in both the f-sphere and the rotating sphere, which means that the representation of the faster inertial-gravity waves accomplished by the new scheme is at least as adequate as the TRSK scheme. However, we notice that the modified scheme does not have all geostrophic modes stationary for the f-sphere as in TRSK, as expected, with spurious modes having frequency of the order of $10^{-5}$ s$^{-1}$. When we consider the scenario of variable $f$, the frequencies of the modes labelled as "geostrophic" are reduced, and one of the reasons for this is that now the Coriolis parameter is on average smaller than that used on the f-sphere (since the f-sphere uses the maximum Coriolis parameter). It is difficult, however, to judge from this if the frequencies of the spurious modes of the on f-sphere, and analogously in the rotating sphere, are low enough to be under control and therefore not cause significant harm to the numerical model.

Thuburn [35] analysed the spurious geostrophic modes of another scheme, due to Nickovic et al. [28], that uses a stencil of size 4 to reconstruct the tangential velocity. He concluded that these spurious modes are related to smaller scales (higher wavenumbers). This should also be the case for the modified scheme developed here, specially it being convergent and more accurate than the Nickovic et al. [28] scheme (not shown). Therefore, these spurious modes should mainly affect the ability of preserving small scale balanced flows. Since it is difficult to analyse this analytically, we will challenge the new method experimentally with a small scale balanced test case in the next section.

It is known that the hexagonal grids carry an extra branch of Rossby modes, due to the imbalance of degrees of freedom (see [31,35]). Therefore, the Rossby modes shown for the TRSK scheme in Fig. 11 cannot be assumed to be the correct (physical) modes, and therefore are not necessarily an adequate reference to check if the modes on modified scheme are close or not the physical ones. It is possible to analyse the Rossby modes analytically on the large Rossby deformation radius regime ($R_d >> a$), where the slow and fast modes separate, and we would therefore be able to depict how well both the TRSK and new schemes represent the physical Rossby modes. Nevertheless, this is a rather dense topic (see e.g. [10]), and is still under investigation, so it is left to be shown in a future work.

To summarize, we see that the modified scheme has good representation of the fast waves (inertial-gravity), but lacks the ability to preserve steady state geostrophic modes on the f-sphere. The latter issue will be explored experimentally in the next section. Further issues concerning Rossby modes will be investigated in future efforts, so, for now, we will rely on the numerical results presented in this paper.

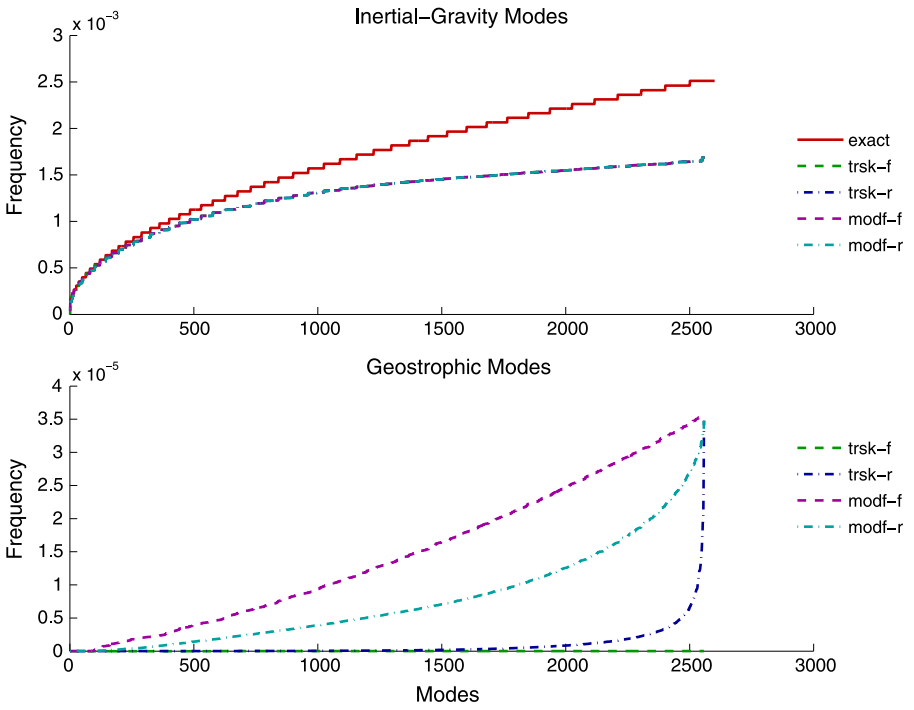


**Fig. 11.** Normal mode frequencies of the shallow water model linearised about a rest state and constant mean layer thickness. "exact" shows the analytical solution for the f-sphere, "trsk-f" and "modf-f" indicates the TRSK and modified methods, respectively, on an f-sphere, whereas "trsk-r" and "modf-r" indicates the TRSK and modified methods, respectively, on a rotating sphere (variable $f$). A grid level 5 with HR95 optimization was used.

### *4.5. Localized balanced flow*

Test case 2 shows the ability of the model to maintain a geostrophically balanced steady state solution with respect to planetary length scales (small wavenumbers). In this section, we will investigate the ability of the model to preserve geostrophically balanced states in small length scales (high wavenumbers). This is particularly important since one of the broken properties with the modified scheme is that it does not have steady geostrophic modes on the f-sphere, and, as discussed in the previous section, this can have direct impact on how the model behaves on small scale balanced flows.

To investigate how the modified scheme behaves in this situation, we propose the following test case for a f-sphere with constant Coriolis parameter denoted as $f_0$. Define the fluid depth to be a local depression at the poles,

$$h = h_0(2 - \sin^n(\theta)), \tag{39}$$

where $h_0$ is a constant, $n$ is an even integer, which we will express as $n = 2k + 2$, and $\theta$ is the latitude. Assuming zero meridional velocity ($v = 0$), we can now derive the zonal velocity, $u$, to maintain steady state from the shallow water equation written in spherical coordinates as,

$$u = \frac{-F + \sqrt{F^2 + 4C}}{2}, \tag{40}$$

where

$$F = \frac{\cos\theta}{\sin\theta} f_0 a, \tag{41}$$

$$C = g h_0 n \sin^{n-2}(\theta)\cos^2(\theta), \tag{42}$$

and we assume null wind at the pole. For large values of $n$, the depression will be strictly limited to a region near the pole, so the wind exists only near the poles and is zero elsewhere in the globe. The geodesic grids used in this work have a pentagon and radially symmetric hexagons near the pole. To avoid this symmetry, and really challenge the ability of a scheme to preserve balance, we rotate these initial conditions to an arbitrary point of the sphere, where no obvious grid symmetry would occur.[3]

[3] This can be easily done converting $u$ and $v$ to its $\mathbb{R}^3$ vector tangent to the sphere ($\vec{u}$) and using 3D rotation matrices.

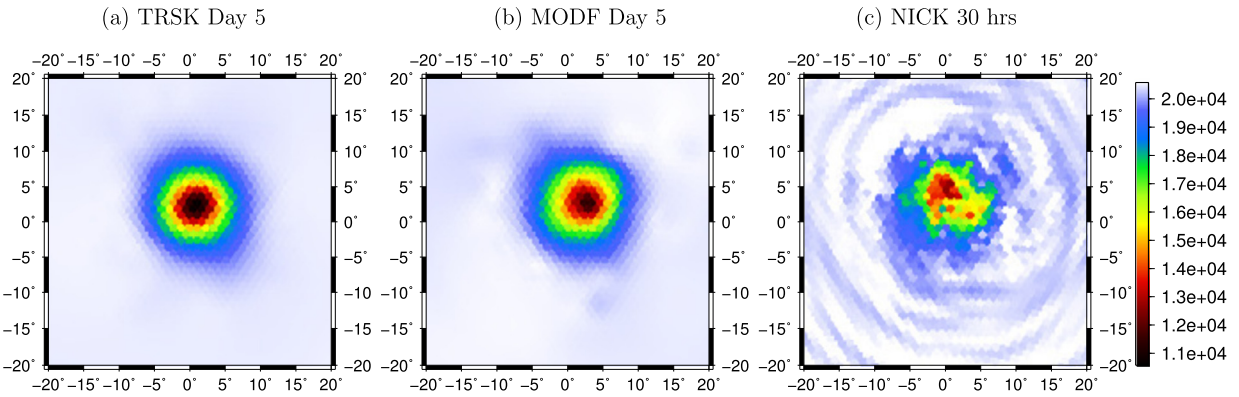


**Fig. 12.** Height field for the localized balanced flow for (a) TRSK scheme at day 5, (b) modified scheme at day 5, (c) the same configurations as the TRSK scheme but using a tangential velocity reconstruction with only 4 edge informations, due to Nickovic et al. [28] and described in Appendix B, at the time of 30 h. All simulations used a HR95 grid of level 6.

We will assume $gh_0 = 10^5\ \mathrm{m^2\,s^{-2}}$ and $k = 160$, which leads to a very localized depression with radius covering less than 10 times the mean grid cell size on a grid level 6 (approximatelly 120 km resolution). The depression is very abrupt, ranging to half of the mean fluid depth in only a very limited number of cells. With $f_0 = 1.4584 \times 10^{-4}$, the dominant wind speed comes from the term $C$, which can enable unrealistic winds of more than 1000 $\mathrm{m\,s^{-1}}$. The Rossby deformation radius for this problem is approximately $R_d = \frac{\sqrt{gh_{\max}}}{f_0} \approx 3000$ km, and we have a grid Rossby number of approximately $Ro \approx 50$, considering a grid resolution of about 120 km (grid level 6), and a Froude number approximately $Fo \approx 2$. This is an exceptionally challenging test case, and we hope that a method able to preserve balance for a few days in this test case would adequate to sufficiently preserve small scale balance in realistic cases.

To avoid grid symmetries, we rotated the localized depression to be centred at the point of longitude 1° and latitude 3°. Although the continuous equations are in balance, nothing was made to ensure discrete balance, so the model has to be able to adjust to its discrete balanced state.

The TRSK scheme can preserve balance very well for a couple of weeks, and we show in Fig. 12(a) the height field at day 5. The modified scheme, although not designed for this purpose, preserves well the balance for approximately one week (see Fig. 12(b) for the height field at day 5).

Nickovic et al. [28] showed that using a stencil with four edges to reconstruct the tangential velocities on a planar hexagonal C grid leads to non-steady state geostrophic modes. Thuburn [35] and Thuburn et al. [2] discusses how Nickovic's scheme is inadequate to preserve geostrophic balanced flows. This method is a special case to the dual triangle formulation which we describe in Appendix B. We indeed verified that, with this approach, after around 1 day of integration a break in balance starts to occur (see the height field at time of 30 h in Fig. 12(c)).

These results indicate that, although the modified scheme does not have steady geostrophic modes on the f-sphere, it is still able to represent balance on small scale flows for adequate time periods, which we hope is enough to make the method useful for practical applications. Also, since the method is convergent, better representation of geostrophic balance is expected to happen with increasing resolution.

### *4.6. Test case 5 – flow over mountain*

This test consists in a solid body rotation flow over an isolated mountain at mid-latitude. The errors in this test are very similar for the TRSK and modified schemes, both qualitatively and quantitatively, for the 15 days tested (not shown). The model shows near second order convergence (see Fig. 8 of [3]), indicating no problems with respect to having inconsistent discrete operators.

The mountain has a non-smooth transition at its base and peak, therefore this test would lead to potentially a non-smooth solution, which would not be adequate to investigate accuracy order of discrete schemes. We investigated the use of a smoother mountain (Gaussian with similar dimensions), but this again revealed similar results for the tested methods, except for very early integration times, where the modified was more accurate (not shown).

Due to the similarity of the results between methods, and since the results for the TRSK are already existing in the literature, we do not present them here. Nevertheless, we will return to this test case in section 4.9 to investigate the energy budget of the new scheme.

### *4.7. Test case 6 – Rossby–Haurwitz wave*

Test case 6 has as initial condition a wavenumber 4 Rossby–Haurwitz wave. As initially proposed in [22], the wave should zonally propagate maintaining its shape. However, this case was shown to be unstable [36], even for wave number 4, so that small grid errors could trigger such instability. Therefore, ideally it is adequate for shorter time scale analysis only, even

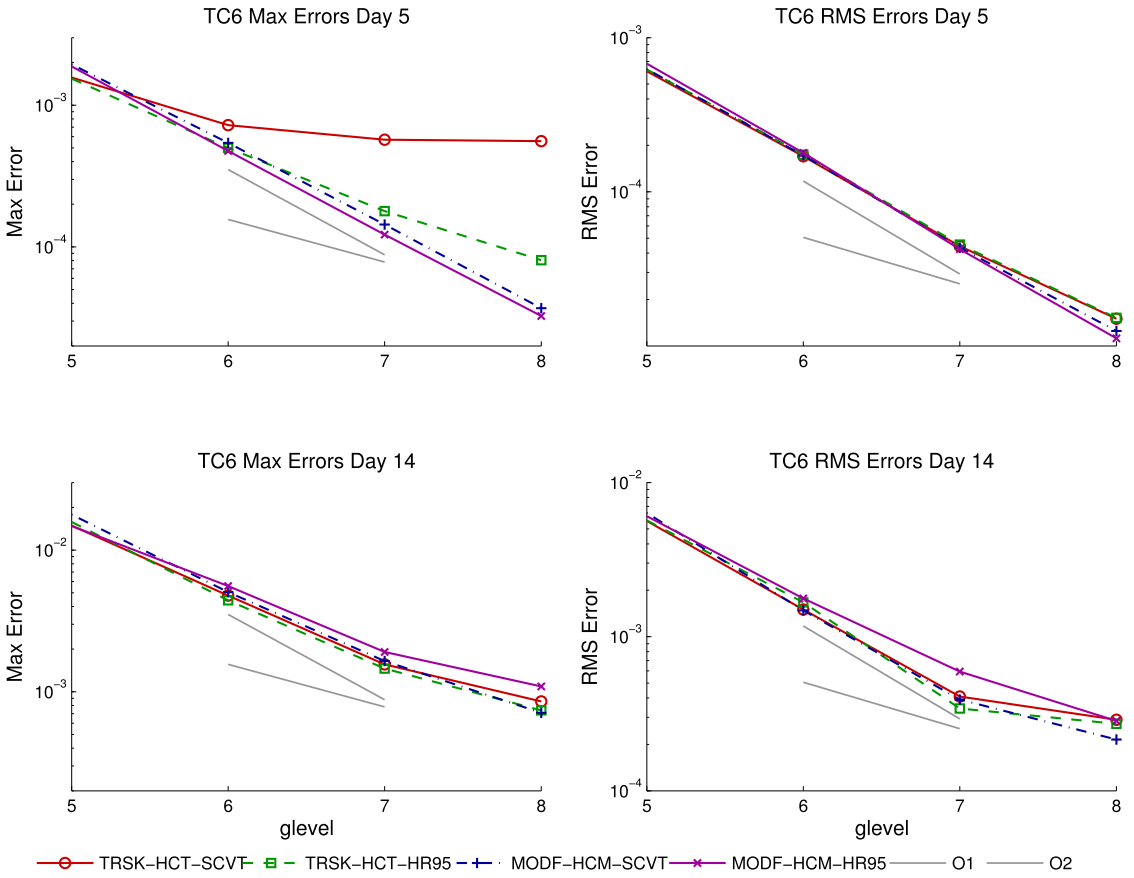


**Fig. 13.** Maximum and RMS errors of TC6 at day 5 (on top) and day 14 (on the bottom) for the layer thickness field ($h$). TRSK indicates the methodology developed in [3], optimized with either SCVT or HR95 methods, on a HCt grid. The MODF scheme indicates the proposed scheme with velocities given at midpoints of the Voronoi cell edges (HCm grid) with the modifications described in the text.

though the model is able to run stably for longer periods. We ran it for 5 and 14 days, with results shown in Fig. 13. At day 5, the difference in the convergence properties of the methods is evident. For the longer run of 14 days, little difference is noticeable between the methods. On the 14th day, the dominant error is a phase error (not shown), common to all schemes.

### *4.8. Barotropically unstable jet test case*

We first use the test case proposed in [29] as a barotropically unstable jet without the perturbation, to see how the modified scheme behaves in preserving the initial balanced state. Although the initial state is balanced, the flow is unstable, and even small features, such grid imprinting errors, are enough to trigger the instability in the jet. We ran the test case for 15 days on a grid level 7 (approximately 60 km resolution) and we show in Fig. 14 how the error in $h$ evolves. The modified scheme is able to maintain initial balance as well as the TRSK scheme, and the errors have very similar patterns. Although the schemes tested revealed very similar maximum errors, they have some qualitative differences, which we will discuss using the test case with initial perturbation.

For the barotropically unstable jet with the perturbation proposed in [29], it is usually expected that grid related wave patterns should not show up in the solution (see [11] for a comparison between grid effects in this test case). Spectral and latitude–longitude grid methods tend to perform well on this test case due to the alignment of the jet and the grid. They usually show a flatter pattern from 100° to 160° East, and 4 well defined vortices bellow the jet, as shown in Fig. 15(d) for the reference ENDGame solution (see also Fig. 4 of [29]). With the TRSK scheme, we clearly see a wave number 6 pattern in the potential vorticity field at day 6 (Fig. 15(a)), due to the underlying grid. The modified scheme was designed to achieve better accuracy taking into account grid related issues, and it therefore tends to reduce grid related effects. We notice in Fig. 15(b) that the modified scheme indeed has a flatter pattern near the plot boundaries, and vortex formations closer to what the expected solution should be.

We notice in Fig. 15 that both the TRSK scheme and the modified scheme show small scale oscillations. We will discuss this further in section 4.10.

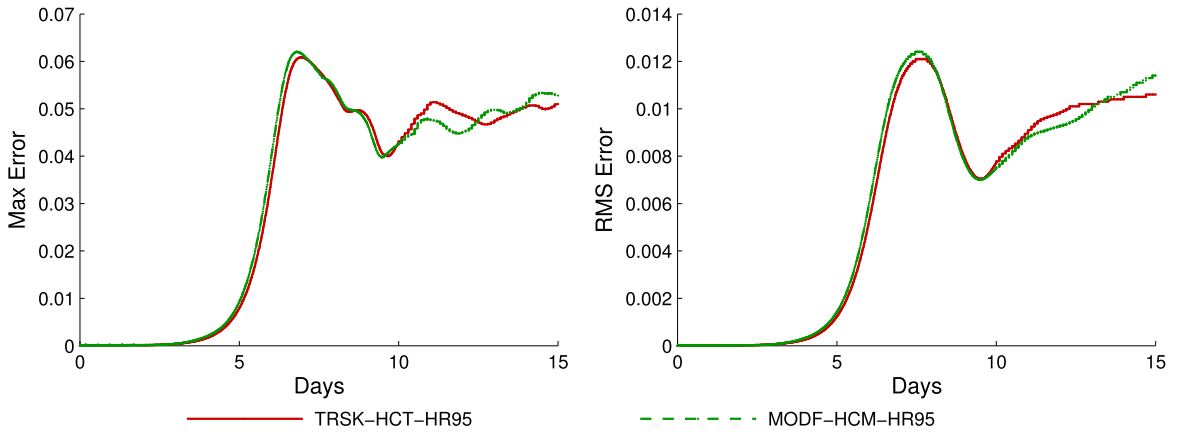


**Fig. 14.** Barotropically unstable jet without perturbation error of $h$ with time. On the left, the maximum errors, on the right the RMS errors. A grid level 7 with HR95 optimization was used.

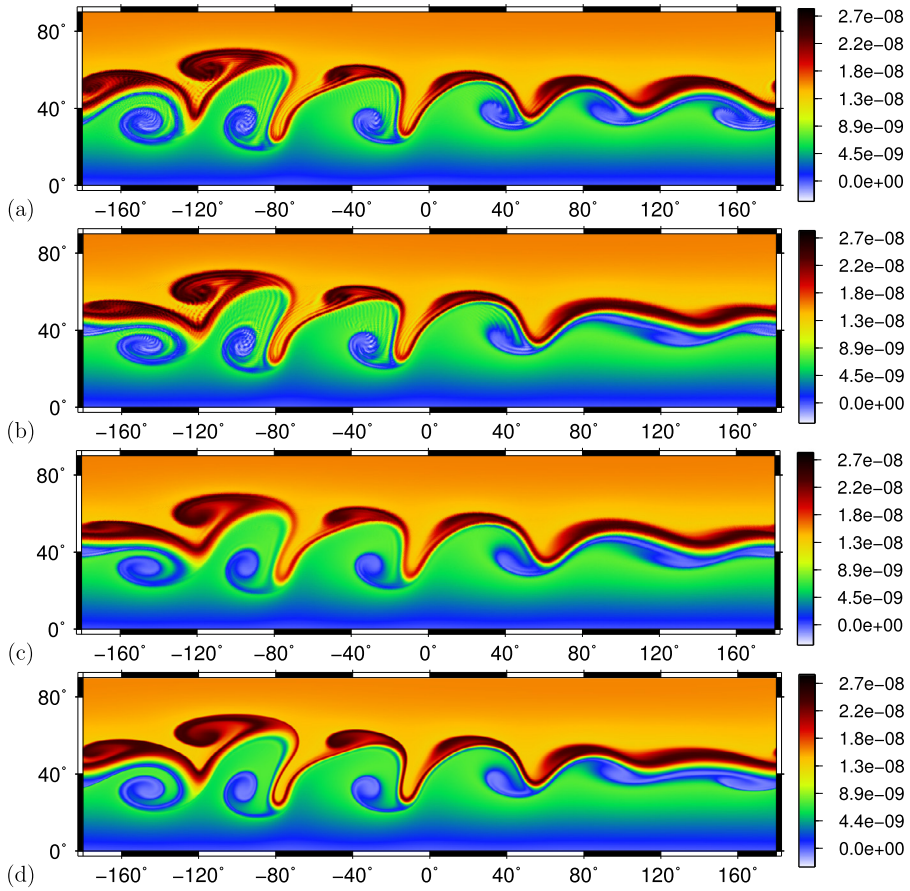


**Fig. 15.** Potential vorticity field at day 6 for the barotropically unstable jet test case with perturbation. (a) TRSK scheme, (b) modified scheme, (c) modified scheme with APVM stabilization and (d) ENDGame reference solution. The plots show the potential vorticity at the edges of the Voronoi cells. A HR95 optimized grid level 7 (approximately 60 km resolution) was used for panels (a), (b) and (c). The reference solution in panel (d) used 2048 × 1024 longitude–latidude grid points (approximately 20 km at the equator).

### *4.9. Energy conservation*

The divergence and gradient operators are still the same in the modified scheme as in TRSK, except for the calculation of the fluxes. This will not affect the product rule identity, since the same flux is used on both cells that share an edge. Therefore, the pressure terms should be energy conserving as in TRSK.

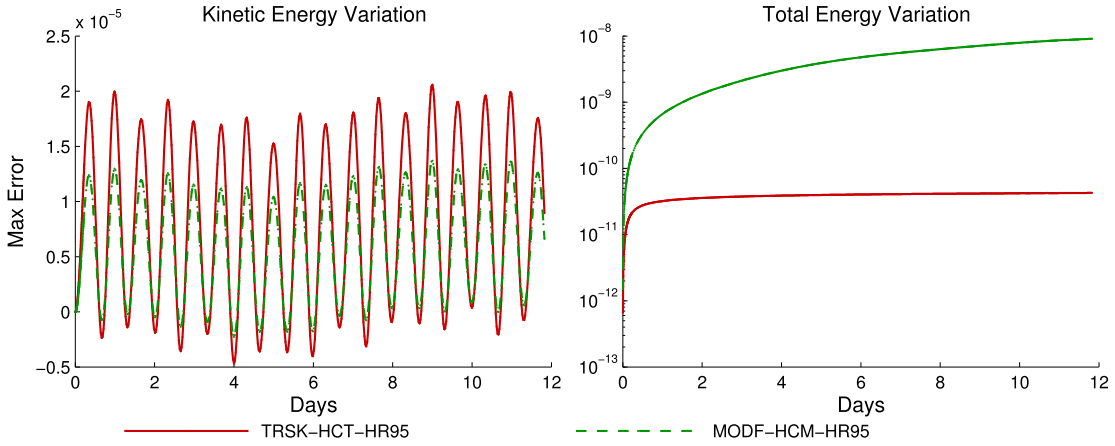


**Fig. 16.** Relative variation of kinetic (left) and total (right) energy for test case 2 with a grid level 7 HR95 optimized grid.

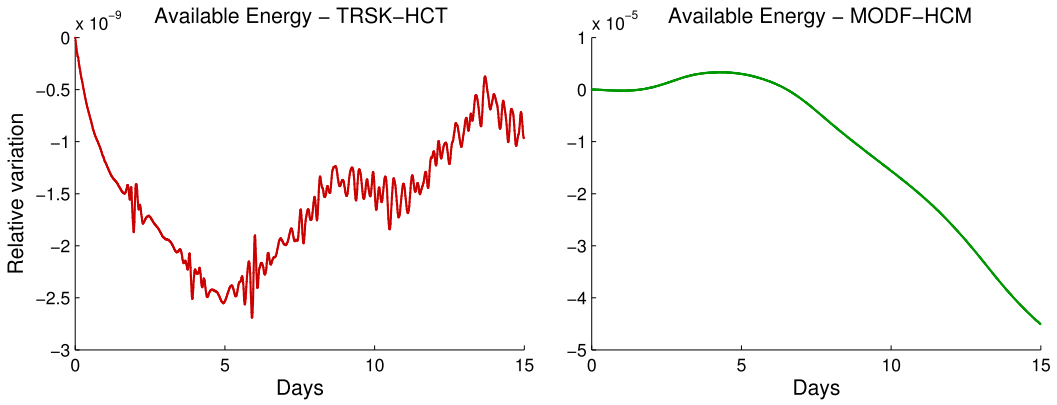


**Fig. 17.** Relative variation of total available energy with the TRSK (left) and modified (right) schemes using test case 5 with a grid level 7 with HR95 optimization.

The Coriolis term (perpendicular term $Q_e^{\perp}$) discretization of the modified scheme was built to ensure that it did not contribute to the energy budget. On an HCm grid, this conservation will not be exact, but related to the distance between the midpoints of the Voronoi cell and triangle edges, which should converge to zero on HR95 grids. We evaluated the energy budget contribution of the Coriolis term, as given in (28), for the initial conditions of test case 2, and the results show that the quantity contributed negatively with the energy budget and is of an order of $10^{-7}$ $m^3 s^{-2}$, which is neglectable with respect to the existing amounts of available energy in the atmosphere.

The modified scheme was not built to conserve total energy, due to modification on the kinetic energy term. However, energy conservation is expected to be bounded by the spatial local truncation error, which will reduce with resolution since the modified scheme is consistent (convergent).

We evaluated the ability of the model to preserve total energy according to the discrete energy equation of [3] (their equation (70)), and the relative energy variation with time of test case 2 is shown in Fig. 16. The variation was calculated as the absolute difference between actual energy and initial energy divided by the initial energy. It must be made clear that the energy conservation properties of TRSK are only within time truncation errors, so it is not expected to conserve total energy at machine round off precision. The modified scheme exhibits an error in total energy conservation greater than those of the TRSK schemes, but still within reasonable bounds. The kinetic energy evolution pattern is very similar in both schemes, as shown in Fig. 16.

In test case 2, most of the energy is unavailable for dissipation, as the flow is already minimizing a combination of total energy, mass and angular momentum. Therefore, the analysis with test case 2 provides little insight to what is happening with the energy budget, apart from stating that there are some spurious sources of energy – that is, total energy is not exactly conserved.

To further investigate the energy budget, we analysed energy sources/sinks using test case 5 of Williamson et al. [22], which, from the initial conditions, already has available potential energy due to the disturbance imposed by the mountain. We used the same formulas to calculate the total available energy as in Ringler et al. [34], which removes the unavailable potential energy, given for a fluid at rest state, from the total energy. Fig. 17 shows the evolution of the available energy with the TRSK and modified schemes. The TRSK scheme conserves available energy within $10^{-9}$ relative bounds, with slight

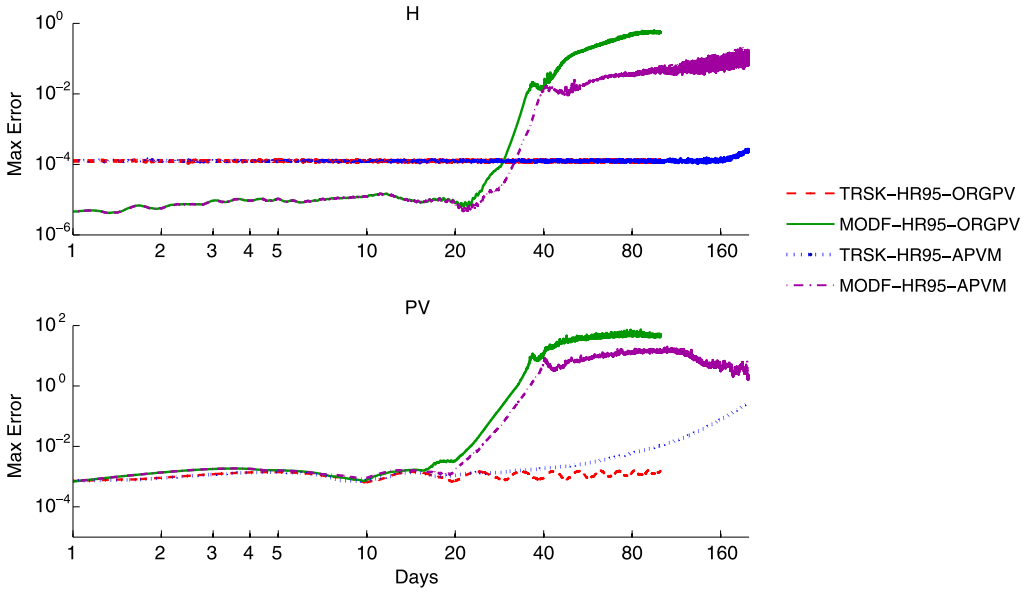


**Fig. 18.** Evolution of the maximum error the layer depth field (top) and PV field (bottom) for the balanced zonal flow (test case 2 of [22]) for the investigated schemes with the APVM PV stabilization scheme and without (ORGPV). A grid level 7 with HR95 optimization was used in all cases. A log scale in the $x$-axis was used to highlight the behaver in the beginning of the time integration.

loss of energy. The modified scheme loses more energy, specially after 1 week of integration, but still has errors within very reasonable bounds ($10^{-5}$). Since the modified scheme is asymptotically consistent (convergent), these bounds should decrease with increasing resolution.

In models with added physical parametrizations, total energy is not expected to be conserved, due to, for example, cascading of kinetic energy towards grid scale turbulence. Although it would be desired that this energy cascade should be solely due to physical processes, it is common that weather forecasting systems include some numerical sources of energy dissipation. Therefore, we assume that the loss of total energy conservation, due solely to the kinetic energy term, should not be considered a major drawback of the modifications. With this in mind, we recall that the kinetic energy term can be one of the main sources of inconsistencies, and by modifying the discretization of this term, a lot of accuracy can be gained.

### *4.10. PV compatibility*

Both the new scheme and the TRSK scheme suffer from a grid scale computational mode (checker board pattern) on the triangles. This pattern occurs due to the discretization of the vorticity on the triangles due to under and overshooting of the analytical vorticity for each triangle pair (see [15] and [11] for details). Since the potential vorticity is averaged before being used in the perpendicular term, the computational mode remains unseen by the momentum equations, and therefore cannot be mitigated.

TRSK possesses the property of compatibility with the Lagrangian form of the PV advection equation, which ensures that this computational mode will keep the method stable even on very long time runs. This compatibility is closely related to having steady geostrophic modes on the f-sphere, which is not achieved in the new formulation. Therefore, the modified scheme is prone to require some treatment of this computational mode for long time scale runs. On short and medium ranges, the PV on the new method evolves in a very similar way to TRSK (e.g. see Figs. 18 and 19 – until near day 20), and no stabilization was required on any of the standard shallow water test cases.

It is desirable that a discrete method for the shallow water equations should dissipate some enstrophy, but, because of this computational mode in the PV, the original TRSK scheme tends to build up enstrophy with time (see [10]). In Ringler et al. [3], they propose an enstrophy dissipating scheme that works as an Anticipated Potential Vorticity Method (APVM). Later, Weller [10] investigated other approaches and proposed the Continuous Linear-Upwind Stabilized Transport (CLUST) scheme. The modified method requires this kind of enstrophy dissipating scheme for stabilization reasons, whereas the TRSK scheme requires them as a way to obtain enstrophy dissipation. Nevertheless, both methods would be generally coupled to such kind of stabilization procedures.

We experimented the APVM of [3] and the CLUST scheme of Weller [10]. Both schemes stabilize the vorticity mode adequately and the new scheme was able to run for long times (e.g. test case 2 was simulated for over a year). We show in Fig. 15(c) how the APVM acts on the new scheme to reduce the oscillations observed in the barotropically unstable jet test case.

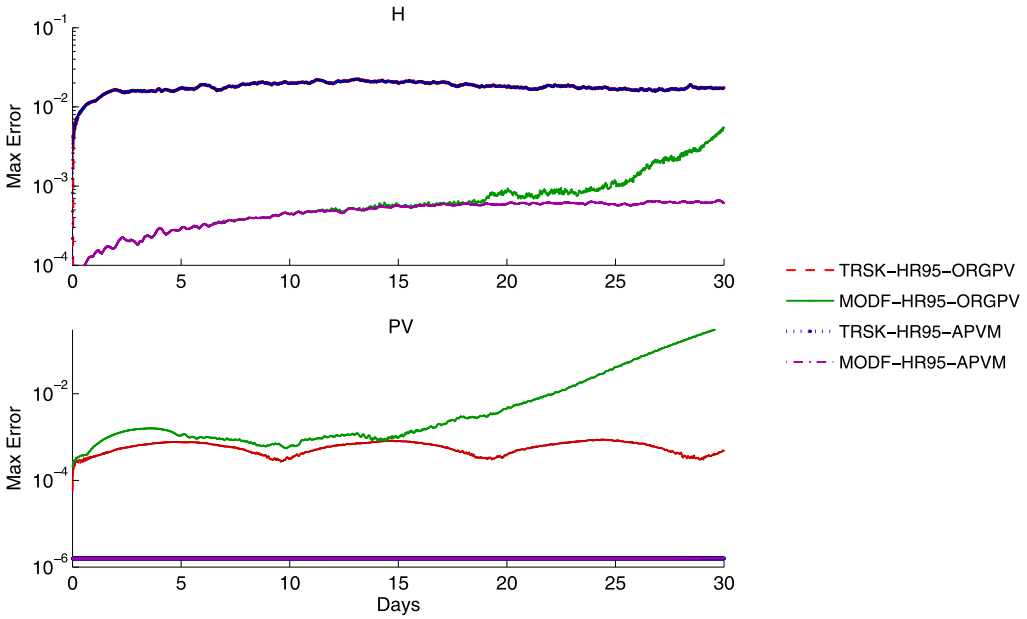


**Fig. 19.** Evolution of the maximum error the layer depth field (top) and PV field (bottom) for the thin layer balanced zonal flow (see section 4.3) for the investigated schemes with and without the APVM scheme. A grid level 7 with HR95 optimization was used in all cases.

Both stabilizing schemes, APVM and CLUST, only change the way the potential vorticity is interpolated to the edge. We will focus on the APVM, for which the new edge PV ($\tilde{q}_e$) is obtained using the original linear interpolated edge PV ($q_e$) plus a stabilization term that depends on the wind and the gradient of the PV, resulting in

$$\tilde{q}_e = q_e - 0.5dt\,\vec{u}_e \cdot [\nabla q]_e, \tag{43}$$

where $\vec{u}_e$ is the full velocity reconstructed to the edge midpoint, $[\nabla q]_e$ is the gradient of the PV calculated at the edge midpoint, and $dt$ is the time step. To ensure that the overall method would still be first order accurate, we calculate the gradient in the following way. First the normal gradients are calculated with centred differences to the triangle edges. Then, the full gradient is recovered at the triangle centre with a Perot's reconstruction method (as in equation (15) but for triangles). The velocity is reconstructed to the edges using a linear interpolation of Perot reconstructions on the Voronoi cells.

Results for the evolution of the errors of test case 2 (steady state balanced flow) are shown in Fig. 18. We see that the errors of the modified scheme without APVM grows after 20 days, and the model will eventually blow up near day 100. The APVM stabilization allowed us to run the model for very long periods (over 1 year), but the results became more inaccurate after 20 days. Even though the APVM is expected to remove most of the grid scale PV mode, after 20 days the error in the PV is still related to high wave numbers (not shown). The TRSK scheme is able to maintain its initial error for a long period, with or without the APVM. Eventually though, the PV in the TRSK scheme with APVM starts to become less accurate with time.

We have discussed that test case 2 does not challenge much the accuracy of the nonlinear terms, so we again will investigate the errors in the thin layer test case described in section 4.3. In this case, the APVM method is able to preserve the good accuracy of the modified scheme for longer periods (see Fig. 19). We also notice that the APVM is able to reduce the grid scale errors in the PV to very low levels, in both the TRSK and modified schemes.

## 5. Summary and discussion

Although some of the usual shallow water test cases do not evidence the inaccuracy problems of the TRSK method developed in [2] and [3], these constitute a potential problem for the method. In the present work we analysed where and why the problems might occur, and we propose alternative ways to mitigate the detected inaccuracy issues.

We show that TRSK even has local truncation errors of order $-1$ on a SCVT grid (the error grows with increasing resolution), due to the inconsistent discretization of the kinetic energy term ($K_i$). On a HR95 grid, the local truncation errors are of 0th order accuracy, again mainly due to low accuracy of the kinetic energy discretization, but also due to inaccuracies in the discretization of the perpendicular term ($Q_e^{\perp}$) on this grid.

An important outcome of the here performed analysis is that the tangent velocity reconstruction used in TRSK ($u_e^{\perp}$) was proven to be first order accurate for any vector field on SCVT grids (at least for the grid levels tested). This result can be used if one wishes to not withdraw from having steady state geostrophic modes on the f-sphere (consequently

preserving PV compatibility). Nevertheless, model developers should be aware of potential inaccuracies due to other sources of inconsistencies in SCVT grids. On HR95 grids, the reconstruction was shown to be 0th order accurate in the maximum norm.

Some of the detected inconsistencies could be avoided by interpreting the degrees of freedom differently (e.g. integrals), or formulating the system in different ways (e.g. flux form). Nevertheless, it seems that this only changes where the inconsistency is prone to happen, but they might still occur. Also, since in non-idealized simulations the initialization of the wind and scalar fields are usually done via interpolation to grid points, the method would still be subject to the analysis performed in this paper.

With respect to the convergence properties observed for the TRSK scheme, we notice lack of convergence using the SCVT grid (test case 2, thin layer and test case 6). Nevertheless, our results indicate that using the HCm positioning (or similarly using the mean velocity fluxes as degrees of freedom) and improving the accuracy of the kinetic energy term to a first order scheme (for example using the proposed modified scheme for $K_e$) is enough to observe convergence on SCVT grids, at least for the tests done here. This maintains all the mimetic properties of TRSK except total energy conservation due to the kinetic energy term. Although convergence is observed for the shallow water test cases performed here, there is no guarantee that the method will be in general convergent, since on HCm SCVT grids it will still be inconsistent with respect to the vorticity and, at least for the modified kinetic energy tested here, also with respect to the gradient of the kinetic energy. This convergence does not seem to be achievable with the HR95 grid by changing only the kinetic energy calculation, since in this case the perpendicular term ($Q_e^{\perp}$) is inconsistent.

As an overall view of the TRSK scheme analysis, if model developers wish to maintain the mimetic property of having steady geostrophic modes on the f-sphere, it is recommended that at least a more accurate kinetic energy calculation should be adopted and the model should run on SCVT grids with HCm positioning (or similarly use mean velocity fluxes as degrees of freedom).

We propose an alternative first order method based on: (i) the use of linear barycentric interpolation for the scalar terms, (ii) a second order reconstruction method for the kinetic energy term, and (iii) an alternative way to reconstruct the tangential component of the wind. The stencils used are similar to the ones in TRSK, so they will have similar computational costs.

With the proposed methodology, we still preserve many of the good properties of the original TRSK scheme, as listed below.

1. Mass conservation.
2. C staggering.
3. Curl-free pressure gradient.
4. Energy conservation of pressure terms.
5. Energy conserving Coriolis term.
6. Local operators – no mass matrix or global systems required.

Some properties, however, are lost. (i) Total energy conservation is no longer met for the kinetic energy term. This is not seen as a major drawback, since in actual applications, with physical parametrizations, one would expect to have some cascading of kinetic energy towards grid scale turbulence, for example. (ii) The geostrophic modes on the f-sphere are no longer steady, but we showed that the method is able to avoid spontaneous breakdown of geostrophic balance in a challenging scenario (see section 4.5). (iii) The PV discretization is no longer compatible with its Lagrangian formulation, which means that it would require the use of enstrophy dissipating schemes to control the vortical computational mode on triangles. This is not seen as a major drawback, since it would be desirable to have enstrophy dissipation in the model anyway. Overall, no major issue was observed with the modified scheme, specially in short and medium time ranges.

Very importantly, we have gained first order accurate local truncation errors (consistency), which, for stable runs, implies in first order convergence. In some test cases this means orders of magnitude smaller errors on fine grids. For even finer grids, as required in cloud resolving models, and specially important when subgrid scale models of physical processes are coupled with the dynamics, this will be even more relevant. The thin layer version of test case 2 (section 4.3) shows how potential inaccuracies in TRSK could appear in 3D models with small equivalent depths, and the benefits of using a more accurate scheme such as the one here proposed.

The suggested modifications can be easily incorporated in many of the existing codes that use the TRSK scheme, but some robustness will be lost. The proposed scheme relies on properties of a HR95 grid optimization to attain first order accuracy, which has been devised for globally quasi-uniform grids only. We point out, however, that even though the method is not formally first order accurate on other grids, such as SCVT grids, it provides a gain in accuracy when compared with the TRSK scheme (e.g. see Fig. 9). So, although it would not be theoretically first order accuracy on locally refined SCVT grids, such as the ones discussed in [7], it could still be used and an accuracy improvement would be expected. The matter of having locally refined grids with HR95 like grid optimizations is currently under investigation.

Overall, our results suggest that the modified scheme could be a more accurate alternative to TRSK in existing atmospheric models, specially if only short or medium range forecasts is desired.

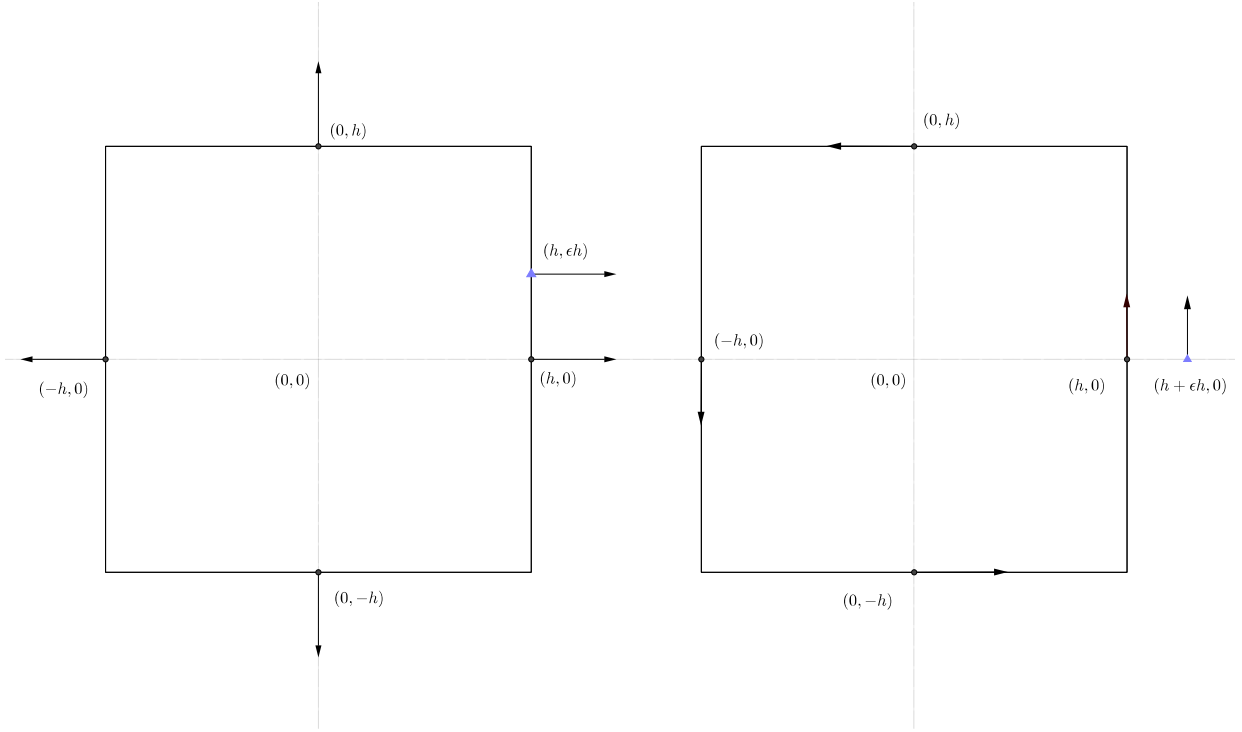


**Fig. A.20.** Schematic positioning for vector values for the divergence discretization (left) and the curl discretization (right). The vectors indicated with triangular basis are positioned to represent the mis-positioning that occurs on spherical grids where the midpoints of the Voronoi and triangle edges do not coincide.

## Acknowledgements

We would like to acknowledge John Thuburn for his suggestions regarding test cases and careful reading. We also thank Saulo Barros and Hilary Weller for comments and suggestions. Two anonymous reviewers contributed with important feedbacks to improve the paper, and are sincerely acknowledged. Financial support from the Sao Paulo Research Foundation (FAPESP), under the grant number 2014/10750-0, is also acknowledged.

## Appendix A. Inconsistency of the divergence and curl discrete operators

We show here that if the discretization uses the finite volume discrete divergence and curl operators for the Voronoi cells and the triangular cells, respectively, then either one of them will be inconsistent unless a grid optimization that approximates the midpoints of the Voronoi edges and of the triangle edges, such as HR95, is used. In general, on a HCt grid, the divergence term will be inconsistent and the curl will be first order. On the HCm grid, the divergence will be first order and the curl will be inconsistent.

### *A.1. Divergence*

We will show with a simple example that having the velocities defined on the edges, but not on its the midpoint, leads to an inconsistent discretization of the divergence operator. To do so, we will use a square shaped cell with one of the velocities mispositioned with respect to the edge midpoint, as shown in Fig. A.20. Although this is a simplified analysis, it easily extends to more complexed shaped polygonal cells, but with more tedious calculations required. We will assume a constant unitary fluid height field, as in this case, the dominant error will come from the velocity field.

Let $\vec{u}$ be a linear vector field on the plane of the form

$$\vec{u}(\vec{x}) = \vec{a_0} + A\vec{x} = \vec{a_0} + \vec{a}_1 x + \vec{a}_2 y, \tag{A.1}$$

where $\vec{x} = (x, y)$, $\vec{a}_0 = (a_0^x, a_0^y)$, $\vec{a}_1 = (a_1^x, a_1^y)$ and $\vec{a}_2 = (a_2^x, a_2^y)$ with exact divergence given by $a_1^x + a_2^y$.

Now consider a square cell with edge lengths given by $2h$ and the normal velocities known at the midpoints of the edges, except for the rightmost edge, where the velocity is known at a different point, of distance $\epsilon h$ from the midpoints northward, as illustrated in Fig. A.20. Direct calculation of the discrete divergence results in an error of $\frac{\epsilon}{2} a_2^x$.

For the spherical grids, $\epsilon$ represents the distance from the midpoints of the triangle edge to the midpoint of the Voronoi cell edge. For non-optimized icosahedral grids, spherical centroidal Voronoi grids and spring dynamics optimized grids, it is known that $\epsilon$ is asymptotically constant (see [16]), and therefore the discrete divergence is inconsistent (the error does not converge to zero with increasing resolution). For the HR95 grid, $\epsilon$ has the order of $h$, and therefore converges to zero with increasing resolution and the method is first order accurate. For a detailed proof of the accuracy of the discrete divergence operator on general polygons see [15].

### A.2. Curl

The same example may be used for the curl operator. Again, although our practical problem lays on triangles, we use a square to illustrate the problem, as it simplifies the calculation. Now we assume that the velocities are tangent to the edges and that one of the velocities is given at $\epsilon h$ distance from the edge, in the normal direction, as in Fig. A.20. The exact curl in this case is $a_1^y - a_2^x$, and the discrete curl leads to an error of $\frac{\epsilon}{2} a_1^y$. The same considerations with respect to $\epsilon$ convergence on spherical grids are valid here. Consequently, the HR95 grid optimization is again required to obtain consistency.

## Appendix B. Coriolis term discretization

### B.1. On a triangular C grid

In this section we comment on some conditions to achieve at least first accuracy. The notation is the same used in section 3.5, and we consider only the planar case.

Consider that, for an unknown constant vector field $\vec{v}$, we know $u_e = \vec{v} \cdot \vec{n}_e$ and $u_{e'} = \vec{v} \cdot \vec{n}_{e'}$, and that we want to reconstruct the constant vector field at the edge $e$, using these 2 given data. As we already know the normal component at the edge $e$, it remains only to find $u_e^\perp = \vec{v} \cdot \vec{t}_e$, as $B_e = \{\vec{t}_e, \vec{n}_e\}$ define an orthogonal basis for the plane.

Expressing $\vec{n}_{e'}$ in the base $B_e$ as $\vec{n}_{e'} = \langle \vec{n}_{e'}, \vec{n}_e \rangle \vec{n}_e + \langle \vec{n}_{e'}, \vec{t}_e \rangle \vec{t}_e$, leads us to

$$\begin{aligned} u_{e'} = \vec{v} \cdot \vec{n}_{e'} &= \langle \vec{n}_{e'}, \vec{n}_e \rangle \langle \vec{v}, \vec{n}_e \rangle + \langle \vec{n}_{e'}, \vec{t}_e \rangle \langle \vec{v}, \vec{t}_e \rangle \\ &= \langle \vec{n}_{e'}, \vec{n}_e \rangle u_e + \langle \vec{n}_{e'}, \vec{t}_e \rangle u_e^\perp, \end{aligned}$$

and if $\langle \vec{n}_{e'}, \vec{t}_e \rangle$ is not zero, that is, $\vec{n}_e$ is not parallel to $\vec{n}_{e'}$, then we can calculate the tangent component of edge $e$, using the information given at edge $e'$, as

$$u_e^\perp = \frac{1}{\langle \vec{n}_{e'}, \vec{t}_e \rangle} \left( u_{e'} - \langle \vec{n}_{e'}, \vec{n}_e \rangle u_e \right). \tag{B.1}$$

Notice that $u_e$ plays a role in reconstructing $u_e^\perp$, although they are defined for orthogonal vectors.

If we have not one, but a set of edges to be used to reconstruct the tangent component at edge $e$ (which we will denote as $NB(e)$), and that each neighbour edge $e'$ of $e$ reconstructs exactly the constant vector field with $u_{ee'}^\perp$, as given in equation (B.1), then the final tangent component at edge $e$ will be given by

$$u_e^\perp = \sum_{e' \in NB(e)} \alpha_{ee'} u_{ee'}^\perp, \quad \text{with} \quad \sum_{e' \in NB(e)} \alpha_{ee'} = 1. \tag{B.2}$$

The consistency conditions derived in section 3.5 define a unique solution for a triangular C grid case, in which only 2 edge informations are used to calculate a reconstructed tangent component for the remaining edge, as we have only 2 unknown weights. Next, we derive the unique coefficients.

Consider a triangle defined by the edges $\{e, e', e''\}$, as in Fig. B.21 for triangle $T_i$, with the tangent vectors defined in anticlockwise direction. If we directly look for solutions of the form

$$u_e^\perp = \sum_{e' \in \mathrm{NB}(e)} w_{ee'} u_{e'}, \tag{B.3}$$

then the weights may be calculated to result in

$$w_{ee'} = \frac{1}{\langle \vec{n}_{e'}, \vec{t}_e \rangle} + \frac{l_{e'}}{l_e} \frac{\langle \vec{n}_{e'}, \vec{n}_e \rangle}{\langle \vec{n}_{e'}, \vec{t}_e \rangle}, \tag{B.4}$$

where we used the general relations for polygons (assuming tangents given anti-clock-wisely)

$$\sum_{e'} l_{e'} \vec{t}_{e'} = 0,$$

and

$$\sum_{e'} l_{e'} \vec{n}_{e'} = 0,$$

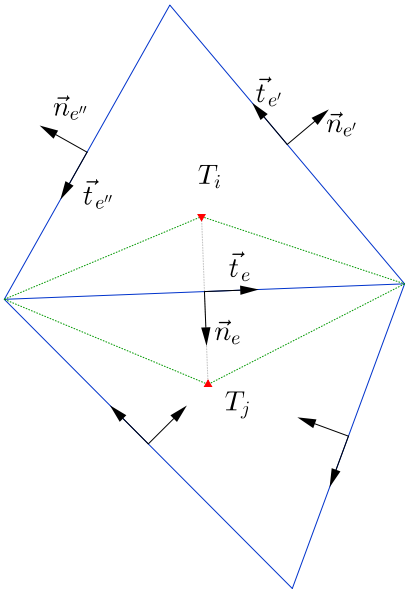


**Fig. B.21.** Triangle notation.

summing over all edges ($e'$) of the polygon, with $l_{e'}$ being the length of edge ($e'$). Equivalently, it is shown in [26], that the weights must also be given by,

$$w_{ee'} = \frac{d_{ie'} l_{e'}}{A_i} \langle \vec{n}_{e'}, \vec{t}_e \rangle, \tag{B.5}$$

where $d_{ie'}$ is the orthogonal distance from the triangle ($T_i$) circumcentre to the edge $e'$, and $A_i$ is the area of the triangle. To verify that (B.5) satisfy the consistency conditions, one may use that, for acute triangles,

$$d_{ie'} = \frac{l_{e'}}{2} \frac{|\langle \vec{n}_{e'}, \vec{n}_e \rangle|}{|\langle \vec{n}_{e'}, \vec{t}_e \rangle|},$$

with $|\langle \vec{n}_{e'}, \vec{t}_e \rangle| = \sin(\theta_{ee'})$ and that $|\langle \vec{n}_{e'}, \vec{n}_e \rangle| = \cos(\theta_{ee'})$ where $\theta_{ee'}$ is the interior triangle angle relative to the intersection of edges $e$ and $e'$. In [26], consistency is proven using consistency conditions specific for triangles.

For a planar triangular C grid, $w_{ee'}$ are uniquely defined for each triangle. The only degrees of freedom to ensure energy conservation are given by the coefficients averaging two neighbour triangle reconstructions ($a_{e,i}$). Condition (33) results in a $3 \times 3$ singular system (for each triangular cell $i$) for the corresponding coefficients $a_{e,i}$. An additional constraint for consistency, discussed in the previous section, is that $\sum_e a_{e,i} = 1$. This leads to a unique solution for the systems. The resulting coefficients are

$$a_{e,i} = \frac{A_{e,i}}{A_e} \tag{B.6}$$

where $A_{e,i}$ is the area of the subtriangle given by the circumcentre of the triangle $i$ and the edge $e$, given by

$$A_{e,i} = \frac{l_e d_{ie}}{2}.$$

$A_e$ is defined to be the edge element area as $A_{e,i} + A_{e,j}$, or simply $A_e = l_e d_e / 2$, with $d_e$ being the distance between the circumcentres of the two triangles that share edge $e$, which simplifies the formula to

$$a_{e,i} = \frac{d_{e,i}}{d_e}. \tag{B.7}$$

This method is precisely what was used in Ham et al. [26].

Simple substitution shows that the method is energy conserving, as equation (33) combined with the weights (B.5) leads to

$$\frac{A_{e,i}}{A_e} \frac{d_{ie'} l_{e'}}{A_i} \langle \vec{n}_{e'}, \vec{t}_e \rangle A_e + \frac{A_{e',i}}{A_{e'}} \frac{d_{ie} l_e}{A_i} \langle \vec{n}_e, \vec{t}_{e'} \rangle A_{e'} = 0,$$

since $\langle \vec{n}_e, \vec{t}_{e'} \rangle = -\langle \vec{n}_{e'}, \vec{t}_e \rangle$.

On a spherical triangular C grid, defining the area $A_{e,i}$ as $l_e d_{ie}/2$ results in a first order approximation of the actual geodesic area of the subtriangle $A_{e,i}$, and will provide energy conservation if $A_e$ is defined as $A_e = A_{e,i} + A_{e,j}$, with $i$ and $j$ being the triangles that share the edge $e$.

### *B.2. Dual triangle formulation*

A simple way to perform consistent energy conserving reconstructions on arbitrary Voronoi grids with a very compact stencil is to use the dual triangular grid. On a planar grid, the normal directions of the Voronoi cell edges coincide with the tangent directions of the edges in the dual triangular cells. Although their directions are the same, the midpoints of Voronoi cell edges do not coincide with those of the triangle cell edges. The information about the normal components of a vector field at the Voronoi grid edges may be seen as a first order approximation to the tangent components at the triangular grid (see [18] for further discussion on the matter). This approximation is exact for constant fields, therefore, the whole reconstruction problem can be restated as finding the normal components at the triangular grid edges given the tangent components. The unique solution for constant vector fields, obtained before for triangular grids, can be derived also in this case.

Considering $u_e$ as given tangent components relative to edges of a triangle (obtained from the normal components of the Voronoi cell edges), a constant vector field may be recovered for a triangle $i$ using Perot's method [24] as

$$\vec{u}_i = \frac{1}{A_i} \sum_{e'} d_{ie'} l_{e'} u_{e'} \vec{t}_{e'}, \tag{B.8}$$

with the tangent vector $\vec{t}_{e'}$ assumed in a counter-clock-wise direction, $A_i$ is the area of the triangle, $d_{ie'}$ is the distance from the triangle circumcentre to the midpoint of edge $e'$, $l_{e'}$ is the length of edge $e'$. This vector may be projected onto the normal vector of edge $e$ (which is approximately the tangent edge direction on the Voronoi cell edge), preserving consistency, resulting in

$$u^{\perp}_{e,i} = \frac{1}{A_i} \sum_{e'} d_{ie'} l_{e'} u_{e'} \langle \vec{t}_{e'}, \vec{n}_e \rangle. \tag{B.9}$$

Two triangle estimates can be joined to give the energy conserving scheme as

$$u^{\perp}_e = \frac{A_{e,i}}{A_e} u^{\perp}_{e,i} + \frac{A_{e,j}}{A_e} u^{\perp}_{e,j}, \tag{B.10}$$

with $A_{e,i}$ being the area of the subtriangle defined by the circumcentre of the triangle $i$ and the edge $e$.

## References


[1] A. Staniforth, J. Thuburn, Horizontal grids for global weather and climate prediction models: a review, Q. J. R. Meteorol. Soc. 138 (662) (2012) 1–26.
[2] J. Thuburn, T.D. Ringler, W.C. Skamarock, J.B. Klemp, Numerical representation of geostrophic modes on arbitrarily structured C-grids, J. Comput. Phys. 228 (2009) 8321–8335.
[3] T.D. Ringler, J. Thuburn, J.B. Klemp, W.C. Skamarock, A unified approach to energy conservation and potential vorticity dynamics for arbitrarily-structured C-grids, J. Comput. Phys. 229 (9) (2010) 3065–3090.
[4] J. Thuburn, C. Cotter, A framework for mimetic discretization of the rotating shallow-water equations on arbitrary polygonal grids, SIAM J. Sci. Comput. 34 (2012) B203–B225.
[5] H. Weller, Non-orthogonal version of the arbitrary polygonal C-grid and a new diamond grid, Geosci. Model Dev. 7 (3) (2014) 779–797.
[6] 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, Mon. Weather Rev. 140 (2012) 3090–3105.
[7] T. Ringler, M. Petersen, R.L. Higdon, D. Jacobsen, P.W. Jones, M. Maltrud, A multi-resolution approach to global ocean modeling, Ocean Model. 69 (0) (2013) 211–232.
[8] T. Dubos, S. Dubey, M. Tort, R. Mittal, Y. Meurdesoif, F. Hourdin, DYNAMICO-1.0, an icosahedral hydrostatic dynamical core designed for consistency and versatility, Geosci. Model Dev. 8 (10) (2015) 3131–3150.
[9] J. Thuburn, C.J. Cotter, T. Dubos, A mimetic, semi-implicit, forward-in-time, finite volume shallow water model: comparison of hexagonal–icosahedral and cubed-sphere grids, Geosci. Model Dev. 7 (3) (2014) 909–929.
[10] H. Weller, Controlling the computational modes of the arbitrarily structured C grid, Mon. Weather Rev. 140 (10) (2012) 3220–3234.
[11] H. Weller, J. Thuburn, C.J. Cotter, Computational modes and grid imprinting on five quasi-uniform spherical C-grids, Mon. Weather Rev. 140 (8) (2012) 2734–2755.
[12] T. Dubos, N.K.-R. Kevlahan, A conservative adaptive wavelet method for the shallow-water equations on staggered grids, Q. J. R. Meteorol. Soc. 139 (677) (2013) 1997–2020.
[13] A. Arakawa, V. Lamb, Computational design of the basic dynamical processes of the UCLA general circulation model, in: Methods in Computational Physics, vol. 17, 1977, pp. 173–265.
[14] B. Cockburn, P.-A. Gremaud, A priori error estimates for numerical methods for scalar conservation laws. Part II: flux-splitting monotone schemes on irregular Cartesian grids, Math. Comput. 66 (218) (1997) 547–572.
[15] P.S. Peixoto, S.R.M. Barros, Analysis of grid imprinting on geodesic spherical icosahedral grids, J. Comput. Phys. 237 (2013) 61–78.
[16] H. Miura, M. Kimoto, A comparison of grid quality of optimized spherical hexagonal–pentagonal geodesic grids, Mon. Weather Rev. 133 (10) (2005) 2817–2833.
[17] R. Heikes, D.A. Randall, Numerical integration of the shallow-water equations on a twisted icosahedral grid. Part I: basic design and results of tests, Mon. Weather Rev. 123 (6) (1995) 1862–1880.

[18] P.S. Peixoto, S.R. Barros, On vector field reconstructions for semi-Lagrangian transport methods on geodesic staggered grids, J. Comput. Phys. 273 (0) (2014) 185–211.
[19] Q. Du, M.D. Gunzburger, L. Ju, Constrained centroidal Voronoi tessellations for surfaces, SIAM J. Sci. Comput. 24 (2003) 1488–1506.
[20] L. Ju, T.D. Ringler, M. Gunzburger, Voronoi tessellations and their application to climate and global modeling, in: P. Lauritzen, C. Jablonowski, M. Taylor, R. Nair (Eds.), Numerical Techniques for Global Atmospheric Models, in: Lecture Notes in Computational Science and Engineering, vol. 80, Springer, Berlin, Heidelberg, 2011, pp. 313–342.
[21] H. Tomita, M. Satoh, K. Goto, An optimization of the icosahedral grid modified by spring dynamics, J. Comput. Phys. 183 (1) (2002) 307–331.
[22] D.L. Williamson, J.B. Drake, J.J. Hack, R. Jakob, P.N. Swarztrauber, A standard test set for numerical approximations to the shallow water equations in spherical geometry, J. Comput. Phys. 102 (1) (1992) 211–224.
[23] A. Gassmann, A global hexagonal C-grid non-hydrostatic dynamical core (ICON-IAP) designed for energetic consistency, Q. J. R. Meteorol. Soc. 139 (670) (2013) 152–175.
[24] B. Perot, Conservation properties of unstructured staggered mesh schemes, J. Comput. Phys. 159 (1) (2000) 58–89.
[25] L. Bonaventura, T.D. Ringler, Analysis of discrete shallow-water models on geodesic Delaunay grids with C-type staggering, Mon. Weather Rev. 133 (8) (2005) 2351–2373.
[26] D.A. Ham, S.C. Kramer, G.S. Stelling, J. Pietrzak, The symmetry and stability of unstructured mesh C-grid shallow water models under the influence of Coriolis, Ocean Model. 16 (1–2) (2007) 47–60.
[27] O. Kleptsova, J. Pietrzak, G. Stelling, On the accurate and stable reconstruction of tangential velocities in C-grid ocean models, in: The Sixth International Workshop on Unstructured Mesh Numerical Modelling of Coastal, Shelf and Ocean Flows, Ocean Model. 28 (1–3) (2009) 118–126.
[28] S. Nickovic, M.B. Gavrilov, I.A. Tosic, Geostrophic adjustment on hexagonal grids, Mon. Weather Rev. 130 (2002) 668–683.
[29] J. Galewsky, R.K. Scott, L.M. Polvani, An initial-value problem for testing numerical models of the global shallow-water equations, Tellus A 56 (5) (2004) 429–440.
[30] H. Wan, M.A. Giorgetta, G. Zängl, M. Restelli, D. Majewski, L. Bonaventura, K. Fröhlich, D. Reinert, P. Rípodas, L. Kornblueh, J. Förstner, The ICON-1.2 hydrostatic atmospheric dynamical core on triangular grids. Part 1: formulation and performance of the baseline version, Geosci. Model Dev. 6 (3) (2013) 735–763.
[31] A. Gassmann, Inspection of hexagonal and triangular C-grid discretizations of the shallow water equations, J. Comput. Phys. 230 (7) (2011) 2706–2721.
[32] M.J. Bell, P.S. Peixoto, J. Thuburn, Numerical instabilities of vector invariant momentum equations on rectangular C-grids, manuscript in preparation, 2016. Preprint available from: http://www.ime.usp.br/~pedrosp/wp/wordpress/en/papers/ [Accessed: 15 Jan 2016].
[33] J. Thuburn, M. Zerroukat, N. Wood, A. Staniforth, Coupling a mass-conserving semi-Lagrangian scheme (SLICE) to a semi-implicit discretization of the shallow-water equations: minimizing the dependence on a reference atmosphere, Q. J. R. Meteorol. Soc. 136 (646) (2010) 146–154.
[34] T.D. Ringler, D. Jacobsen, M. Gunzburger, L. Ju, M. Duda, W. Skamarock, Exploring a multiresolution modeling approach within the shallow-water equations, Mon. Weather Rev. 139 (11) (2011) 3348–3368.
[35] J. Thuburn, Numerical wave propagation on the hexagonal C-grid, J. Comput. Phys. 227 (11) (2008) 5836–5858.
[36] J. Thuburn, Y. Li, Numerical simulations of Rossby–Haurwitz waves, Tellus A 52 (2) (2000) 181–189.