# Reducing spurious diapycnal mixing in ocean circulation models

DISSERTATION

of

Margarita Smolentseva

at the University of Bremen and the Alfred Wegener Institute for Polar and Marine Research

Supervisors: Prof. Dr. Thomas Jung, Prof. Dr. Sergey Danilov
Reviewers: Prof. Dr. Thomas Jung, Prof. Dr. Jörn Behrens

Date of submission: 23.08.2020

Date of Ph.D. colloquium: 28.09.2020

## Abstract

Spurious diapycnal mixing of water masses occurs in ocean circulation models as an artifact of numerical algorithms used to advect temperature and salinity. Most of the ocean models used in climate research are based on geopotential vertical coordinates ($z$- coordinates), which intersect isopycnal surfaces. The non-alignment of coordinate surfaces with isopycnals causes spurious diapycnal mixing during horizontal advection of a water-parcel by high-order upwind transport schemes. The growth in the potential energy of the system appears without any sources. This behavior is physically incorrect and leads to an energetic inconsistency and incorrect water mass transformation. Therefore, spurious diapycnal mixing in ocean models is one of the reasons that lead to the incorrect hydrological state of the ocean basins after some integration time. Improvements are required which would reduce spurious mixing in ocean models.

There are several ways that can potentially reduce numerical mixing and spurious diapycnal mixing in particular. Three of them are considered in the current work. The first way is the design of more accurate advection schemes with the intention to achieve a reduction in truncation error which leads to a decrease in numerical mixing in a system. The second option is the stabilization of central high-order advection schemes by isoneutral diffusion. And the last approach is a choice of the right mesh. It is a question to investigate whether meshes can cause spurious mixing due to their structure, regularity, and other properties.

The current work deals with the problem of spurious diapycnal mixing. It analyses the stability of numerical implementation of isoneutral diffusivity on triangular meshes of FESOM2. It proposes a new compact advection scheme characterized by a reduced truncation error compared to other finite volume schemes in FESOM2. It shows that the application of isoneutral diffusion to stabilize central schemes can reduce spurious diapycnal mixing in models, however, it requires special tuning for every initial state of a model. It is also found out that mesh irregularity does not necessarily imply an enhanced numerical mixing in a system, however, it might depend on the type of triangles.

# Contents

# Chapter 1

# Introduction

A significant part of our knowledge about the variability of the Earth climate relies on numerical modeling. Numerical climate models help researchers to understand the climate of the past and make projections for the future. Ocean circulation models are one of the main components of Earth climate models. Ocean significantly contributes to the meridional heat transport, and ocean-atmosphere interactions provide boundary conditions for atmospheric models over a large part of the Earth's surface. Ocean transport processes are very slow. It takes centuries for water masses to complete the whole circulation cycle. Ocean has a huge impact on climate, being a part of the planetary energy cycle (Rahmstorf [2002]), and on biogeochemical cycles, exchanging gases with the atmosphere. It has a higher heat capacity than atmosphere and land, influencing climate fluctuations on daily, seasonal, annual, and interannual time scales (Clark et al. [2002]).

Despite ongoing efforts, the existing climate models demonstrate a wide range in their projections concerning the behavior of climate trajectories under different carbon dioxide increase scenarios (e.g., see Community [2020], Vivek and et [2019], Karen and Richard [2007]). Moreover, model outputs demonstrate quite large biases from historical observations which also bring particular uncertainty into their projections. Many possible reasons for such results have been investigated, and a significant part of these reasons arises from numerical errors.

Numerous modeling groups over the world make significant efforts to improve the results of the models. Nevertheless, climate models, and particularly global ocean circulation models, still suffer from insufficient resolution and inaccuracies coming from parameterizations and numerical assumptions. With the growth of computer facilities, finer computational meshes become affordable, allowing better representation of dynamical processes. However, currently, even the highest resolution that can be used in models for climate research (about 1/10 degree) is insufficient for the representation of many processes. Therefore, these processes are parameterized. The imperfection of numerical methods leads to a lack of consistency in simulated balances and energy pathways.

One of the problems that researchers face is spurious numerical mixing accompanying the simulation of ocean motion of the ocean water masses. Unlike the atmosphere, the deep ocean is nearly adiabatic, with small background mixing on the level of $10^{-5}$ $\mathrm{m}^2/\mathrm{s}$ provided largely by internal-gravity waves through their nonlinear evolution and final breaking. The spurious mixing caused by the implementation of numerical operators in ocean models can easily reach values comparable to background mixing or even exceed them as estimated by Ilicak et al. [2012] and Megann [2018]. As explained below, spurious numerical mixing generally occurs when a fluid parcel is advected in ocean models. For numerical stability or the need for positive-definite solutions for the transported quantities, advective operators are, as a rule, equipped with build-in numerical dissipation that depends on the detail of the implementation. Spurious mixing of temperature and salinity generally implies also spurious mixing of density. The biggest concern is due to a diapycnal (occurring across isopycnals) component of mixing which leads to a direct rearrangement of isopycnal surfaces. Spurious isopycnal (along isopycnal surfaces) mixing also has consequences and may affect the ocean density structure through nonlinearities of the ocean water equation of state (cabeling and thermobaricity).

Mixing in the ocean is the major source of the deep water masses transformation. For example, it is well known that the lowest part of the meridional overturning circulation is driven by mixing that is balanced by slow upwelling of the deep waters (e.g., Kuhlbrodt et al. [2007]). The fact that numerical mixing in ocean models can be as high or even sometimes higher than the physical mixing implies that spurious mixing can strongly affect water mass transformations on long time scales. That might be one of the reasons for biases developing in these models over the long integration time. Furthermore, mixing raises the center of mass and potential energy of the water, so, it cannot happen without external sources of energy. Background mixing parameterizations such as IDEMIX (Olbers and Eden [2013]) aims to take into account this kind of sources as an important part of the wider idea of energy-consistent modeling. The presence of spurious mixing also means the loss of energy balance.

Another important aspect of the spurious mixing problem is that it is highly sensitive to the intensity of simulated eddy motions (to the grid-scale Reynolds number) as it was found by Ilicak et al. [2012]. Nowadays resolution in modern climate models is becoming finer, which allows scientists to perform simulations that resolve mesoscale eddy motions. Mesoscale eddies are an indispensable component of ocean dynamics responsible for vertical and horizontal exchanges in the ocean. Resolving them also means that the simulated eddy kinetic energy becomes much larger, which also might imply spurious mixing in the deep ocean. The existing analyses for $1/4$ degree global ocean models show that it can indeed be the case (see, e.g., Ilicak et al. [2012]) but can be controlled in other cases (Gibson et al. [2018]). It remains to be seen how this behavior will be modified with finer resolution and in longer simulations.

For the reasons mentioned above, there is a considerable amount of research aimed at diagnosing spurious mixing and exploring measures to minimize it. Numerous papers devoted to the analysis of water mass transformation (see e.g. Xu et al. [2018] and references therein) concentrate on estimates of diapycnal velocities which are produced by both physical and spurious mixing. The difficulty here lies in the need to compare the diagnosed total diapycnal transport with the transport which we would have if all mixing were physical. This can be done in a simplified way (e.g., zonally averaged), as demonstrated by Lee et al. [2002] and Megann [2018], but is difficult otherwise. Nevertheless, Xu et al. [2018] presented an example of situations where the presence of spurious mixing is clearly identifiable. Klingbeil et al. [2014], based on Burchard and Rennau [2008] propose an approach that estimates the discrete variance decay associated with the implemented advective operator. These methods provide three-dimensional maps of spurious mixing, however, they do not distinguish between dia- and isopycnal directions. Approaches for estimating effective spurious diapycnal diffusivities were proposed in Griffies et al. [2000] based on adiabatic sorting of water parcels and in Getzlaf et al. [2010] and Hill et al. [2012] based on releasing passive tracers. The method of Griffies et al. [2000], similarly to Megann [2018], provides a basin-averaged view on spurious diapycnal mixing, which is not necessarily related to the mixing caused by numerically simulated water parcels (see Ilicak et al. [2012]). Finally, Ilicak et al. [2012] suggested to consider just the increase in the so-called reference potential energy (RPE) proposed by Winters et al. [1995] and also used in Griffies et al. [2000] as a part of their analysis. Since then, the concept of RPE is used by many model development teams to identify their spurious mixing (see Ilicak et al. [2012], Petersen et al. [2015], Mohammadi-Aragh et al. [2015], Gibson et al. [2018], Kärnä et al. [2018]). Its advantage is the ease of diagnostics. Once again, it is a global measure, characterizing the entire basin. Ilicak [2016] proposes an extension of this concept to quantify the spatial distribution of mixing. The concept of RPE will be used in this work and will be explained in detail in chapter 2.

Ocean modelers are mostly interested in the reduction of diapycnal spurious mixing because it affects the density structure of the ocean and, thus, the structure of the entire ocean circulation. It is also well understood that motions in nearly adiabatic deep ocean happen along isopycnal surfaces: The exchange of two water parcels along isopycnals does not change gravitational potential energy and thus does not require extra sources and sinks. One of the main reasons why the density structure of the deep ocean is not preserved by numerical ocean dynamics is that their vertical discretization often does not follow isopycnals coordinate. Most models used for global ocean simulations use geopotential or $z$-coordinates in vertical. Although isopycnals have a very small slope angle across wide regions of the ocean, they still cross $z$-model level surfaces, and the angle of crossing can be substantial in frontal areas, e.g. nearby shore in polar regions. Numerical operators in $z$-coordinate models are implemented in horizontal and vertical directions. They are causing numerical errors. There are two issues, one of which is related to diffusion. One can think of physical diffusion as a combination of mixing along isopycnals with much smaller mixing across isopycnals. Because of this difference, even though isopycnals slope at small angles to the horizontal $z$-surfaces outside the mixed layer, the approach of splitting mixing into vertical and horizontal components leads to a spurious cross-isopycnal component which is not negligible. This problem was recognized long ago, and Redi [1982] proposed to introduce a diffusivity tensor, which approaches diffusion in such a way that it occurs along isopycnals with isopycnal diffusivity $K_i$ and across isopycnals with much smaller diffusivity $K_d$ defined by physical parameterizations. Numerical implementation of these (rotated or Redi) diffusivities faced certain problems before Griffies et al. [1998] proposed a numerical solution based on the variational principle. A recent discussion of related numerics and stability analysis for quadrilateral meshes has been presented in Lemarié et al. [2012]. The appearance of models working on triangular meshes requires to reconsider these analyses. For Finite volumE Sea-ice Ocean Model (FESOM) the variational implementation was described in Danilov et al. [2017], however, its stability remained unexplored. Analysis of the stability is provided in chapter 3.

The second issue is advection. In particular, if the horizontal advection of tracers follows a high-order upwind scheme, it, as a rule, causes spurious mixing. Upwind schemes are widely used in ocean circulation models, for example, Regional Ocean Modeling System (ROMS, see Moore et al. [2011]), Nucleus for European Modelling of the Ocean (NEMO, see Madec and Team [2016]), MIT General Circulation Model (MITgcm, see Marshall et al. [1997]). Indeed, in this case, it can be shown that the numerical advection operator has the main truncation error of diffusive type (see, e.g., Lemarié et al. [2012]). To stabilize central, or so-called mixed (combination of central and upwind), schemes, additional methods such as slope or flux limiters are used. In particular, the flux corrected transport method described by Zalesak [1978] is one of them. Limiters are applied to ensure that solutions stay positive (in case of concentrations) or monotone. These methods introduce either local mixing or unmixing (Mohammadi-Aragh et al. [2015]). They are utilized in ocean circulation models such as Finite volumE Sea-ice Ocean Model (FESOM2, see Danilov et al. [2017]) or Model for Prediction Across Scales-Ocean (MPAS-O, see Ringler et al. [2013]). In all cases this mixing will have a cross-isopycnal component which is fully spurious because continuous operators of advection cannot change tracer variance, i.e., create mixing. Numerical advection operators are believed to be the major source of spurious mixing in general circulation ocean models with $z$-coordinate models. Therefore, the reduction of spurious mixing is an important problem in numerical ocean modeling, which becomes even more significant with the increase of model resolution. There are several ways of handling this problem. One is the design of more accurate advection schemes because a reduction in truncation error means the reduction of spurious effects due to the dissipativity of this error (Hill et al. [2012], Mohammadi-Aragh et al. [2015]). The other approach is the use of high-order centered schemes stabilized with isopycnal (rotated) diffusion, as advised by Lemarié et al. [2012]. A new advection scheme with a reduced truncation error is proposed and analyzed compared to other schemes in chapter 4. The application of isoneutral diffusion is analyzed in chapter 5.

The last topic considered in this thesis relates to meshes. Regular meshes and among them the meshes composed of equilateral triangles are expected to be associated with smaller numerical approximation errors than irregular or distorted meshes. However, variable-resolution meshes are always distorted to some degree. Thus, there is a question if the lack of regularity and structure of meshes can influence numerical mixing in the system and in which way. This is also investigated in the last section of chapter 5.

All the methods described in this thesis were implemented in FESOM version 2. A concise outlook of the model is provided in chapter 2.

# Chapter 2

# Methods

In this chapter, I give an overview of the required basis which the current work is founded on. I describe methods used for reducing spurious mixing and its analysis by researchers. Also, a short description of the model FESOM version 2, used in this work, is provided.

## 2.1 FESOM2

FESOM2 is a finite-volume sea ice-ocean circulation model. In the model, the equations of motion, continuity, and tracer balance are integrated with respect to time. The main principles of FESOM2 are described in this subsection. More detailed information about FESOM2 can be found in Danilov et al. [2017].

FESOM2 is based on unstructured triangular meshes of variable resolution. The mesh is defined by a grid of triangles in the horizontal directions and a system of horizontal levels in the vertical direction which partition the computational domain into triangular prisms. In the horizontal plane, the scalar discrete variables of FESOM2 are placed at vertices of triangles, discrete horizontal velocities are placed at centroids of triangles. A control volume for velocities is defined by a triangular prism in the centroid of which this velocity value is defined. Control volumes of scalars are defined by median-dual prisms. A median-dual prism for a scalar $T$ is obtained by linking centroids of all the triangles that have the vertex of $T$ through the mid-edge points. Figure 2.2 shows schematically a control volume for velocities and scalars.

In the vertical direction, the horizontal velocities and scalars are placed in the mid-layers, and the vertical velocity is at full levels. Horizontal gradients of scalar quantities are located at the centers of triangles, and vertical gradients

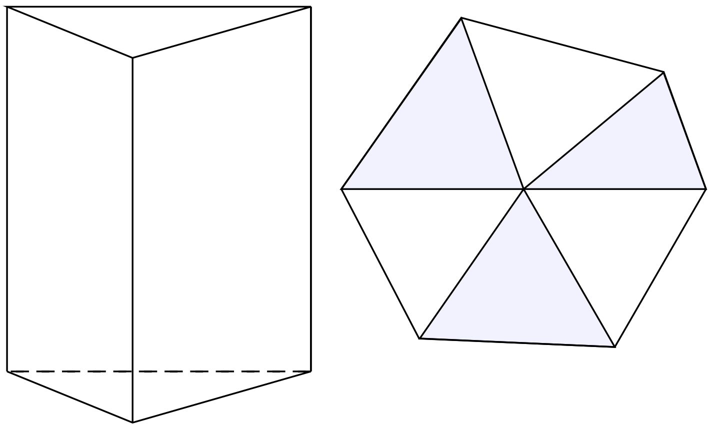


Figure 2.1: Elements of the triangular mesh: a triangular prism (left) and the horizontal part of triangular prisms (right).

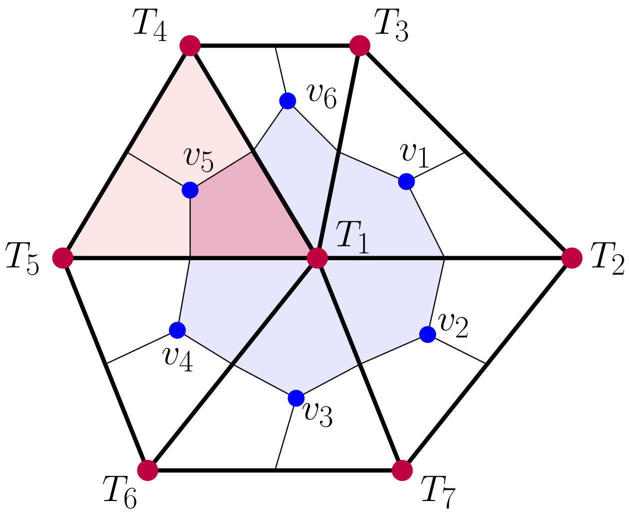


Figure 2.2: Schematic of the median-dual control volumes. Control volume for scalar $T_1$ in the horizontal plane is the median-dual volume around the respective vertex shaded with light blue colors. The control volume for velocity in the horizontal plane is represented by a triangle (indicated by light red colors for $v_5$). The light purple color is the area where both these control volumes intersect.

are at the vertices and full levels (see Fig.2.3). A triangular prism of the mesh can therefore be split into six triangular sub-prisms with unique combinations of discrete values of scalars and their derivatives (see Fig. 2.4). We will need these sub-prisms to explain the implementation of the isoneutral diffusion operator.

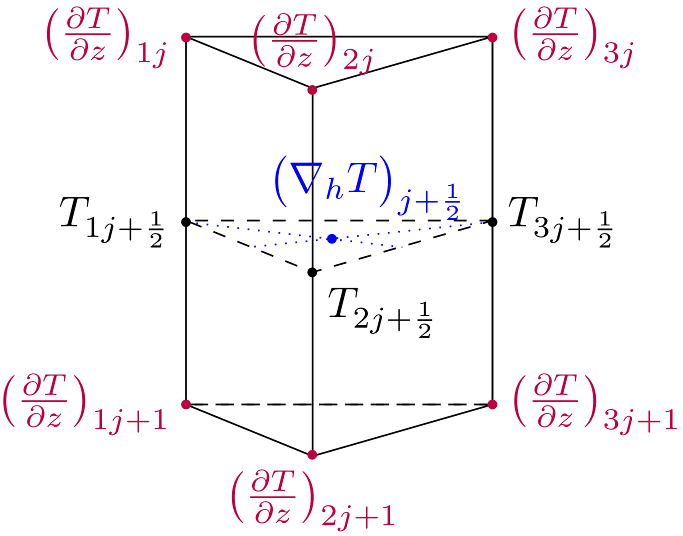

$\left(\frac{\partial T}{\partial z}\right)_{1j}$
$\left(\frac{\partial T}{\partial z}\right)_{2j}$
$\left(\frac{\partial T}{\partial z}\right)_{3j}$
$\left(\nabla_h T\right)_{j+\frac{1}{2}}$
$T_{1j+\frac{1}{2}}$
$T_{3j+\frac{1}{2}}$
$T_{2j+\frac{1}{2}}$
$\left(\frac{\partial T}{\partial z}\right)_{1j+1}$
$\left(\frac{\partial T}{\partial z}\right)_{3j+1}$
$\left(\frac{\partial T}{\partial z}\right)_{2j+1}$

Figure 2.3: A triangular prism element of the mesh. The scalar values are located in the middle of the prism edges, the horizontal gradient of the scalars is placed in the middle of the prism, and their vertical derivatives are in the prism vertices. Six sub-prisms obtained by splitting the triangular prism by the mid-plane (drawn) and faces of median-dual scalar control volumes (not shown) are characterized by triples of scalar value, scalar horizontal gradient, and scalar vertical gradient.

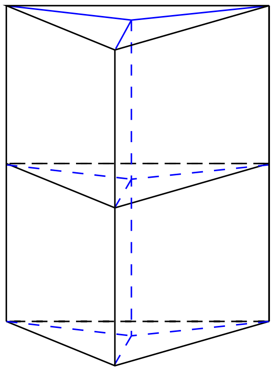


Figure 2.4: A triangular prism divided into six sub-prisms.

## Governing equations

Let us introduce vertical layer thicknesses $h_k = h_k(x, y, z)$ where $k$ changes from 1 to $K$, and $K$ is the number of layers. Then the continuity equation integrated over the vertical extent of the layer is written as

$$\partial_t h_k + \nabla_h \cdot (\mathbf{u}h)_k + (w^t - w^b)_k + W\delta_k = 0. \tag{2.1}$$

Here **u** is horizontal velocity, W is the water flux leaving the ocean at the surface, $w_k^t$ and $w_k^b$ denote transport velocities through the top (the superscript $t$) and bottom (the superscript $b$) boundaries of layer $k$, and $\nabla_h = (\partial_x, \partial_y)$. Let us denote an arbitrary tracer as $T$. The layer-integrated equation for the tracer will be written as

$$\partial_t(hT)_k + \nabla_h \cdot (\mathbf{u}hT)_k + (w^tT^t - w^bT^b)_k + WT_W\delta_k = \nabla \cdot h_k\mathbf{K}\nabla T_k. \tag{2.2}$$

In this equation $\nabla = (\partial_x, \partial_y, \partial_z)$ denotes the 3-dimensional operator, i.e., the right hand side in (2.2) is the divergence of the flux due to 3-by-3 diffusivity tensor **K** which will be described in details in this chapter. The equation for the surface elevation $\eta$ (also called the sea surface height (SSH)) is obtained by summing (2.1) over layers

$$\partial_t\eta + \nabla_h \cdot \sum_k h_k\mathbf{u}_k + W = 0. \tag{2.3}$$

The pressure field in hydrostatic approximation is expressed as

$$p = p_a + g\rho_0\eta + p_h, p_h = \int_z^0 \rho dz \tag{2.4}$$

where $p_a$ is the atmospheric pressure and $\rho$ the deviation of density from its reference value $\rho_0$, and $p_h$ is the hydrostatic pressure due to density variation $\rho$.

The momentum equation is taken directly, without layer integration, and the layer index $k$ is omitted. It is written below in a vector-invariant form:

$$\begin{aligned}\partial_t\mathbf{u} + \frac{\omega + f}{h}\mathbf{k} \times \mathbf{u}h + \Big((w\partial_z\mathbf{u})^t + (w\partial_z\mathbf{u})^b\Big)/2 \\ + \nabla_h\Big(p/\rho_0 + \mathbf{u}^2/2\Big) + g\rho\nabla_h Z/\rho_0 = D_u\mathbf{u} + \partial_z(\nu\partial_z\mathbf{u}),\end{aligned} \tag{2.5}$$

but also a flux form is available. In this equation $g\rho\nabla_h Z$ reflects the fact that the model layers may deviate from geopotential surfaces. The variable $Z$ stands for the vertical coordinate $z$ of the midplane of the layer with the thickness $h$. The second term of the equation represents the potential vorticity $q = (\omega + f)/h$.

Finally, the set of equations is completed by the equation of state,

$$\rho = \rho(T, S, z),$$

with $T$ the potential temperature, $S$ the salinity, and $x$ the depth which replaces the pressure because of the Boussinesq approximation used in the equations above.

## ALE

FESOM2 provides three options for the treatment of free-surface. One of them is the linear free surface where the change in the thickness of the upper layer due to surface elevation is ignored in the tracer and momentum equations so that the vertical layers are kept fixed in time, and volumes of the prisms stay constant throughout the simulations. Two other options represent the full free surface and use arbitrary Lagrangian-Eulerian (ALE) vertical coordinate (see Donea and Huerta [2003]) to account for the moving surface, which warrants full conservation. The full free surface is represented in FESOM2 with two options: zlevel and zstar. With the zlevel option, only the thickness of the upper layer is changing with time following the variations in SSH. With the zstar option, the thicknesses of all the layers are dynamically updated and the change in the thickness of the fluid column due to the elevation is equally distributed among all the vertical layers. The zlevel option is slightly more computationally efficient since only the thickness of the upper layers is updated. However, this option may impose limitations when the unperturbed thickness of the upper layer is too small: the thickness of the upper layer may become too close to zero, or even negative, which is not allowed. For more details see Scholz et al. [2019], Danilov et al. [2017]. The ALE and full free surface options allow one to get rid of so called virtual salinity fluxes. When the freshwater is added to the surface, the thickness of the upper layer is modified according to the thickness equation 2.1 and the salinity in the upper layer is controlled by that change. There is zero flux of salt, and total salt is conserved. The heat flux is taken into account through the boundary conditions at the surface in the temperature equation 2.2.

## Spatial discretization

First, let us consider the notation. The indices $c$, $e$ and $v$ enumerate cells, edges, and vertices respectively. The notation like $v(e)$ will be used to denote the set of vertices of edge $e$. The notation $e(v)$ means the set of all the edges coming out of the vertex $v$. This rule is applied to all other possible sets $c$, $e$ and $v$.

In order to discretize the governing equations in terms of finite volumes, these equations are integrated over the control volumes. For the velocities it means

$$A_c \mathbf{u}_c = \int_c \mathbf{u} dS,$$

and similarly for the scalar tracers

$$A_{kv} T_{kv} = \int_{kv} T dS.$$

Here $A_c$ is a cell area, and S is a contour. $A_{kv}$ is a horizontal area of a prism for the vertex $v$ at the layer $k$. The layer index $k$ is suppressed everywhere where this causes no ambiguity.

There are three options for the momentum equation in FESOM2. Two of them have a flux form and another one has a vector-invariant form. Cell-vertex discretization on triangular meshes has an excessive number of velocity degrees of freedom (compared to scalar degrees of freedom) which was taken into account in these options. Implementation of the momentum equation requires a certain amount of averaging in order to avoid noise which can occur on a grid scale.

- Vertex velocity option. Velocity at a vertex is calculated by averaging among the velocities defined at the cells which contain the vertex $v$

$$A_v \mathbf{u}_v h_v = \sum_{\overline{c}(v)} \mathbf{u}_c h_c A_c/3.$$

  Here $\overline{c}(v)$ is a set of all the prisms containing the vertex $v$, $h_c$ are prism thickness. $\overline{c}(v)$ coincides with $c(v)$ in the upper layers, but bottom topography excludes some prisms in deeper layers. Using this averaging the diversion of horizontal momentum flux is calculated

$$A_c \Big( \nabla \cdot (h\mathbf{u}\mathbf{u}) \Big)_c = \sum_{e(c)} l_e \Bigg( \sum_{v(e)} \mathbf{n}_e \cdot \mathbf{u}_v h_v \Bigg) \Bigg( \sum_{v(e)} \mathbf{u}_v/4 \Bigg).$$

  Here $\mathbf{n}_e$ is an external normal at edge $e$ of cell $c$, $l_e$ is the length of the edge $e$.

- Scalar control volumes. The horizontal part of the momentum flux divergence on scalar control volumes is computed as

$$A_v \Big( \nabla \cdot (h\mathbf{u}\mathbf{u}) \Big)_v = \sum_{e(v)} \sum_{c(v)} \mathbf{u}_c h_c \cdot \mathbf{n}_{ec} \mathbf{u}_c d_{ec}.$$

  For the vertical part, the flux going through the top boundary of the prism in layer $k$ is

$$A_v (w_v \mathbf{u}^t) = w_v \sum_{\overline{c}(v)} \mathbf{u}_c^t A_c/3.$$

  Here the superscript $t$ indicate the top surface.

- Vector-invariant form. The Coriolis parameter (see (2.5)) has to be defined on the vertices of a prism because the relative vorticity is defined there

$$\Big( (\omega + f)\mathbf{k} \times (u) \Big) = \sum_{v(c)} (\omega + f)_v \mathbf{k} \times \mathbf{u}_c/3.$$

  The kinetic energy in (2.5) is first computed at scalar locations, and then its horizontal gradient is computed at triangles.

Transport equation is discretized with high order schemes are used. FESOM2 uses upwind (3rd order), central (4th order), and mixed schemes of 3rd-4th order. The schemes are based on the estimates of the tracer gradients on an upwind or central stencil extending beyond the nearest neighbors. This and other advection schemes are described in more detail and analyzed in chapter 4.

The Gent-McWilliams (GM) parametrization of eddy stirring (Gent and McWilliams [1990]) is implemented in FESOM2 using the algorithm proposed by Ferrari et al. [2010]. The GM parametrization is always applied together with isoneutral diffusion proposed by Redi [1982]. The GM parametrization together with isoneutral diffusion is important for parametrizing eddy-scale motions and diapycnal mixing in the ocean. Isoneutral diffusion and its analysis in FESOM2 is one of the key points considered in the current work.

### Isoneutral diffusivities

Isoneutral diffusivity directs mixing of temperature, salinity, and other tracers, caused by unresolved eddies, along isoneutral surfaces in the ocean. It can also be used for the stabilization of central advection schemes.

An isopycnal is commonly understood as the surface of constant potential density defined with respect (referenced) to a particular pressure level. The nonlinearity of the equation of state leads to the fact that this surface deviates from a locally referenced density surface, which is also called a neutral density surface and commonly denoted $\gamma$, even if the potential density and isoneutral density coincide at some point. Small and almost energetically neutral exchanges of fluid particles are expected to occur along neutral surfaces and be isoneutral.

The neutral slope vector is computed as

$$\nabla\gamma = -\alpha\nabla T + \beta\nabla S.$$

Here $T$ and $S$ are the potential temperature and salinity respectively, $\alpha$ is the coefficient of thermal expansion and $\beta$ is the coefficient of haline contraction. Note that the pressure appears in this equation only through the expansion and contraction coefficients, and therefore these coefficients depend on local pressure (depth in the Boussinesq approximation). $\gamma$ surfaces can be computed only approximately and this is the reason the language of isopycnal surfaces is often used to express the physics approximately. In most cases, the notions of isopycnal and diapycnal directions below have the precise meaning of isoneutral and dianeutral directions respectively.

Diffusive flux $\mathbf{F}_T$ of tracer $T$ is commonly parameterized as being proportional to the gradient of the tracer. Since the flux is a vector, a general connection between it and another vector $\nabla T$ is a diffusivity tensor given by 3 by 3 matrix $\mathbf{K}$ in the standard Cartesian representation,

$$\mathbf{F}_T = -\mathbf{K}\nabla T.$$

Here the minus sign takes care that the flux is down gradient in the simplest case when $\mathbf{K} = \kappa\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

To have the flux vector oriented along the $\gamma$ surfaces one should write $\mathbf{K} = K_i\mathbf{P}$, where $K_i$ is the isoneutral diffusivity and $\mathbf{P}$ is a projecting operator that eliminates components aligned with $\nabla\gamma$. For the matrix, $\mathbf{K}$ this operator is expressed as

$$\mathbf{P} = \mathbf{I} - \nabla\gamma\nabla\gamma/|\nabla\gamma|^2,$$

or in the component form

$$\mathbf{P}_{ij} = \delta_{ij} - n_i n_j, \quad n_i = \partial_i\gamma/|\nabla\gamma|,$$

where $i, j \in \{x, y, z\}$, and $\delta_{ij}$ is the Kronecker delta. Normally a slope vector in the direction of $\nabla\gamma$ is used instead of the unit vector $n_i$. This vector is defined as

$$\mathbf{s} = -\nabla_h\gamma/\partial_z\gamma, \quad \nabla_h = (\partial_x, \partial_y).$$

The matrix $\mathbf{K}$ in terms of the slope vector $\mathbf{s}$

$$\mathbf{K} = \frac{K_i}{1+s^2}\begin{pmatrix} 1+s_y^2 & -s_x s_y & s_x \\ -s_x s_y & 1+s_x^2 & s_y \\ s_x & s_y & s^2 \end{pmatrix}, \tag{2.6}$$

where $s_x$ and $s_y$ are components of the isoneutral slope vector $\mathbf{s}$, and $s$ is its magnitude.

We assume that the slope is small which means that the values of the vector $\mathbf{s}$ are also small, $|\mathbf{s}| \ll 1$. In this case, we can approximate $1+s^2$ by 1 up to small errors that are quadratic in small $|\mathbf{s}|$. For the same reason, all quadratic terms in the matrix can be ignored except for the diagonal $s^2$ term which corresponds to the vertical diffusion. Thus, the expression above is simplified to

$$\mathbf{K} = \begin{pmatrix} K_i & 0 & s_x K_i \\ 0 & K_i & s_y K_i \\ s_x K_i & s_y K_i & s^2 K_i \end{pmatrix},$$

which is referred to as small-angle approximation. Finally, we should add the physical dianeutral diffusion $K_d$ that is provided by vertical mixing parameterization schemes. The dianeutral direction is very close to the vertical one, so with sufficient accuracy, this addition affects the lower diagonal term. The overall form of the diffusivity tensor (in small-angle approximation) will take the form:

$$\mathbf{K} = \begin{pmatrix} K_i & 0 & s_x K_i \\ 0 & K_i & s_y K_i \\ s_x K_i & s_y K_i & s^2 K_i + K_d \end{pmatrix}.$$

In the mixed layer, the slope vector is no longer small. However, the fluid motion is diabatic in the mixed layer, and fluid is mixed horizontally across the isoneutral surfaces. For this reason, we taper the off-diagonal components of the diffusivity tensor to zero at the locations in and close to the mixed layer. Thus, $K_i$ becomes the coefficient of horizontal diffusion in regions where tapering is done. A practical criterion is the magnitude of the slope vector.

There are two important points to consider while implementing the rotation diffusivity numerically. The first one is that in finite-volume methods the components of the slope vector are naturally computed at different locations: $\alpha$ and $\beta$ are defined at the center of a scalar cell, horizontal and vertical derivatives are computed, respectively, at lateral and horizontal faces. The slope vector is not defined at a single point unless some averaging is made. However, if this averaging is made, it is not guaranteed that the diffusive operator will lead to variance decay. The rigorous approach requires using variational calculus, i.e., writing a discrete dissipation functional and obtaining discretization by taking variations (differentiating) of this functional, as first proposed by Griffies et al. [1998] (see also discussion in Lemarié et al. [2012] and the explanation below).

The second point is the numerical stability of the time integration. Because of the presence of mixed (horizontal-vertical) derivatives, the explicit time stepping of the isoneutral diffusion operator faces severe restrictions if the aspect ratio of a cell is smaller than the slope. In order to avoid this difficulty, the integration has to be split into explicit and implicit parts. We explored the stability of this split at triangular meshes where scalars are placed at vertices in chapter 3.

The time integration split, tapering, and other numerical details make the isoneutral diffusion not truly isoneutral, although the dianeutral component, introduced by them, is small compared to the horizontal diffusion.

### Analysis of isoneutral diffusion

We consider a diffusion equation

$$\frac{\partial T}{\partial t} = \nabla \mathbf{K} \nabla T. \tag{2.7}$$

Here T is a tracer field (temperature, salinity, etc.). This equation has two properties.

- Property 1.
  If tracer $T$ is the isoneutral density $\gamma$, from definitions above (see equation 2.6)
  $$\mathbf{K}\nabla\gamma = 0. \tag{2.8}$$
  This happens because $\mathbf{K}$ is the projection operator which removes components aligned with $\nabla\gamma$.
- Property 2.
  We multiply equation (2.7) with $T$ and integrate over volume. The integral on the right-hand side is integrated by parts assuming no flux conditions to obtain
  $$\partial_t \int T^2/2\, dV = \int T\partial_i \mathrm{K}_{ij}\partial_j T\, dV = -\int \partial_i T \mathrm{K}_{ij}\partial_j T\, dV \leq 0. \tag{2.9}$$
  The last equality follows from matrix $\mathrm{K}_{ij}$ of $\mathbf{K}$ being symmetric and positive if $T$ differs from $\gamma$. Here $i, j \in \{x, y, z\}$.

Let us prove that the matrix K is positive-definite as it is mentioned in the property 2. For this, let us consider Sylvester's criteria which says that a symmetric matrix is positive-definite when all its leading principal minors are positive. Indeed, $\Delta_1 = |K_{iso}| > 0$, $\Delta_2 = K_{iso}^2 > 0$, and $\Delta_3 = K_{iso}^2(s^2 K_{iso} + K_d) - s_x^2 K_{iso}^3 - s_y^2 K_{iso}^3 = s^2 K_{iso}^2 K_d > 0$. Thus, the matrix $K$ is positive-definite.

Properties 1 and 2 must be maintained at a discrete level. The difficulty is that in the discrete case $T$, its horizontal derivatives, and its vertical derivatives lie at different locations and the same concerns the quantities needed to compute the isoneutral gradient vector defining $\mathbf{K}$.

We introduce the dissipation functional

$$\mathcal{F}(T) = -\frac{1}{2}\int \nabla T \mathbf{K} \nabla T d\Omega. \tag{2.10}$$

As can be seen, the right-hand side of (2.9) is expressible as $2\mathcal{F}(T)$. One can also readily see that the right-hand side of (2.7) can be obtained by the calculus of variations as the functional derivative

$$\begin{aligned}\delta\mathcal{F} &= -(1/2)\int \nabla\delta T\mathbf{K}\nabla T d\Omega - (1/2)\int \nabla T\mathbf{K}\nabla\delta T d\Omega \\ = -\int \nabla\delta T\mathbf{K}\nabla T d\Omega &= -\int \nabla(\delta T\mathbf{K}\nabla T) d\Omega + \int \delta T(\nabla\mathbf{K}\nabla T) d\Omega \\ &= \int \delta T(\nabla\mathbf{K}\nabla T) d\Omega \qquad (2.11)\end{aligned}$$

Here $\delta T$ is a small variation of tracer field satisfying boundary conditions. The first equality in the chain of transformations follows from the symmetry of $\mathbf{K}$, and the second one is due to our assumption that no flux is coming through boundaries (fluxes coming to the ocean are associated with vertical (dianeutral) diffusivity). The chain of transformations implies that the right-hand side of (2.7) is also a multiplier of $\delta T$ in the last integrand, i.e., it is the functional derivative of the dissipation functional,

$$\frac{\delta\mathcal{F}}{\delta T} = \nabla(\mathbf{K}\nabla T). \qquad (2.12)$$

Thus, starting from negative-definite dissipation functional one arrives at the right-hand side of (2.7). Vice versa, it was shown above that the right-hand side of (2.7) leads to the variance decay.

The idea of Griffies et al. [1998] exploits this fact. To get a numerical implementation of isoneutral diffusivity that satisfies Property 2, one first needs to write the dissipation functional (2.10) in the discrete form and obtain the expression for the right-hand side of discretized (2.7) by performing a discrete analog of variation computation. Analysis of isoneutral diffusivity, and dissipational functional in particular, on triangular meshes is provided in chapter 3.

## 2.2 Variance decay and truncation errors

Consider a 1D advection–diffusion equation

$$\partial_t T + \partial_x(uT) = \partial_x\kappa\partial_x T \qquad (2.13)$$

in a domain $x \in [a, b]$ assuming periodicity of $u$ and $T$ for simplicity to eliminate the consideration of boundary effect. In the equation above $T$ is a scalar field (tracer), $u$ is the advecting velocity and $\kappa$ is the diffusivity. Integrating this equation over the domain we easily find that $\partial_t \int_a^b T\,dx = 0$, i.e. the total amount of tracer is conserved independently of the presence of diffusion. In order to see the effect of diffusion we should consider behavior of variance $T^2$. To learn about behavior of the variance, we multiply the equation above with $T$ and integrate over the interval. After simple manipulations we find:

$$\partial_t \int_a^b T^2\,dx = -2\int_a^b \kappa(\partial_x T)^2\,dx, \tag{2.14}$$

which means that the variance decay is related only to the diffusion, and not to the advection. Indeed, $\int_a^b T\partial_x(uT)\,dx = \int_a^b \partial_x(uT^2/2)\,dx = 0$ (in a more common case this result is a consequence of impermeability of boundaries). Thus in the continuous case advection does not create variance decay, i.e. it does not relate to mixing.

In the discrete case, our approximation of the advection operator always contains truncation errors. Let us discretize the interval $[a, b]$ with cells of the size $h$. The cells will be indexed with $n$, and it will be assumed that $T_n$ is the discrete value of $T$ at the center of cell $n$. Assume for simplicity that $u$ is constant and positive over the interval. With the first order upwind method, the following approximation is used

$$u\partial_x T|_n \approx u(T_n - T_{n-1})/h. \tag{2.15}$$

In this formula, we assume that the velocity $u$ is positive. We use the Taylor expansion for $T_{n-1}$ around the center of cell $n$ to obtain

$$T_{n-1} = T_n - h\partial_x T|_n + (h^2/2)\partial_{xx}T|_n + \mathcal{O}(h^3),$$

i.e.,

$$u\partial_x T|_n \approx u(T_n - T_{n-1})/h = u\partial_x T|_n - (uh/2)\partial_{xx}T|_n + \mathcal{O}(h^2). \tag{2.16}$$

The leading term in the truncation error is diffusive and will lead to mixing that is rather high (with $\kappa_A = uh/2$).

Acting in the same way we will find that for the second-order central differences

$$u\partial_x T|_n \approx u(T_{n+1} - T_{n-1})/(2h) = u\partial_x T|_n + (uh^2/6)\partial_{xxx}T|_n + \mathcal{O}(h^4). \tag{2.17}$$

Here the main truncation error is coming with odd derivative, and the leading error is dispersive. The order of the method here is associated with the power of $h$ in the leading truncation term.

The difference between the two methods lies in both their order and the type of their truncation term. This behavior is preserved for higher-order methods.

For example, the standard third-order upwind method has the truncation error $(uh^3/12)\partial_{xxxx}$ (see, e.g., Lemarié et al. [2012]), which is a biharmonic diffusion. The numerical illustrations for the errors of the fourth-order method are given in chapter 4.

The type of truncation error is important because a dispersive error does not lead to variance decay. Take, for example, the case of second-order centered differences. In this case, the contribution of the error to the variance decay is proportional to

$$\int_a^b T\partial_{xxx}T\,dx = \int_a^b \partial_x(T\partial_{xx}T)dx - \int_a^b \partial_x(T\partial_x T)dx = 0.$$

The integrals above take zero values due to the periodic boundary conditions of the flux. One will get the same zero result for higher order even methods. There might be some boundary effects in a general case, but their role is expected to be small if the domain is large.

This is the reason why one expects that even-order methods do not create mixing on their own. If velocity is variable, this statement is not necessarily correct. Nevertheless, it is natural to expect that even-order methods will be nearly non-dissipative. The drawback of even-order methods is that, because of dispersion, the numerical propagation speed of perturbations becomes a function of their wavelength even if $u$ is constant. As a result, oscillations will be developing around frontal features, leading finally to code instabilities. An even-order scheme has to be supplemented with an appropriate dissipation which will suppress oscillations. Thus, in the end, both odd-order schemes with build-in dissipation and even-order schemes with added dissipation imply some numerical dissipation, i.e. spurious mixing. The conceptual advantage of an even-order scheme is that dissipation can be treated independently from advection. In the 3D case, dissipation can be turned along isoneutral surfaces to minimize or even avoid dianeutral mixing (Lemarié et al. [2012]).

If the order of scheme is increased, the truncation term becomes asymptotically smaller (assuming that the tracer field is smooth enough), and will either introduce less dissipation or will need less dissipation to counter dispersion. For this reason, methods reducing the magnitude of truncation errors are expected to need less dissipation for stability. Chapter 4 presents a fourth-order compact scheme developed in the framework of this thesis that has a reduced truncation error.

The simple analysis above explains the motivation behind particular parts of this work. There are, however, complications in the practical realization, which may partly modify the simple arguments described above. One of the complications is the discrete time stepping. In FESOM2 advection is implemented with second-order Adams–Bashforth method. Using the 1D example above and assuming $\kappa = 0$, the procedure of FESOM2 is modeled by

$$T_n^{m+1} - T_n^m = -(u\tau/h)(T_{n+1}^{AB2} - T_{n-1}^{AB2})/2, \tag{2.18}$$

where $m$ is the time step index, $\tau$ is the time step, $u =$ const and

$$T_n^{AB2} = (3/2 + \epsilon)T_n^m - (1/2 + \epsilon)T_n^{m-1}. \tag{2.19}$$

Here, $\epsilon$ is a small parameter (0.1) needed for stability. To proceed, the fields are expanded around $(m, n)$ point in time and space in a Taylor series, and then time derivatives are expressed in terms of space derivatives with account for the order of approximation. As a result, we get a modified equation that shows what is actually solved. The procedure is called a modified equation method and is described in many textbooks (see, e.g., Donea and Huerta [2003]). The result for the case considered is

$$\partial_t T + u\partial_x T = \epsilon u^2 \tau \partial_{xx} - uh^2 \partial_{xxx} T(1/6 + (u\tau/h)^2(-5/12 - \epsilon + \epsilon^2)) + \mathcal{O}(\tau, h)^3 \tag{2.20}$$

The time stepping stabilization through $\epsilon$ introduces in the leading order diffusion with the equivalent diffusivity coefficient $\epsilon u^2 \tau$, i.e. spurious mixing. Spurious mixing can be essential in frontal regions where the ocean velocity can reach 1 m/s. Note also that the amplitude of the dispersion term is modified by time stepping. The combination $u\tau/h$ is called the Courant number, which is commonly small in ocean applications. For example, FESOM2 in global simulations on a 1/4 degree mesh is running with $\tau = 1200$ s, giving $u\tau/h \leq 0.1$. For this reason, the compensation of dispersive error by time stepping is mostly weak, and errors of spatial discretization are as a rule dominant.

It should be reminded once again that the analysis above is only valid for uniform velocity. In realistic application, velocity is variable and depends on the fields of active tracers such as temperature and salinity. However, it can be expected that the behavior of advection schemes remains qualitatively similar to that described above.

## 2.3 RPE analysis

To estimate spurious mixing in the model, the concept of reference potential energy (RPE) will be applied in a way how it was described by Winters et al.

[1995]. RPE is the minimum potential energy of the system obtained by sorting density without mixing. RPE changes only through dianeutral transport. When mixing occurs in a model, RPE generally increases with time (Ilicak et al. [2012]).

For calculation RPE all the water parcels in a model have to be sorted in a way that the heaviest parcels are located at the bottom and the lightest ones - at the top of a reservoir. The RPE is calculated as an integral by volume over density-weighted geopotential of the new state of a model with sorted parcels:

$$RPE = g \iiint \rho^* z dV. \tag{2.21}$$

Here $\rho^*$ stands for density of the state with sorted parcels. It is important to remember that after this operation when a parcel is placed at another depth under different pressure, its density will change. This behavior causes additional problems in the RPE computation for the ocean with the full equation of state, requiring new sorting after each time a parcel for the next depth is sought. However, Ilicak et al. [2012] shows that using an isopycnal density instead of in situ density in this procedure allows one to do sorting only once, for incurring errors are small. Basically speaking, RPE can be considered as unavailable potential energy of the system. It would not change with time in models with switched off explicit dissipation and forcing unless the effects of spurious mixing. When mixing exists in the system, e.g. from dissipation brought by advection schemes, RPE will grow.

Hence in an ocean domain with no boundary fluxes of buoyancy, the evolution of the RPE directly reflects a nonzero dianeutral transport (Ilicak et al. [2012]). To avoid unnecessary information about RPE due to its small changes, I consider relative changes of RPE with respect to initial state:

$$\overline{RPE} = \frac{RPE(t) - RPE(t=0)}{RPE(t=0)}. \tag{2.22}$$

Following Ilicak et al. [2012] even after adiabatic sorting we avoid diagnosing local, or isopycnally averaged, diffusivity from the sorted profile due to its noisy character and because it can be potentially misleading. Exploring the potential energy of mixing through the RPE analysis is both consistent and relevant to the global energy of the ocean.

# Chapter 3

# Isoneutral diffusion on triangular meshes

In deep ocean mixing of water mass properties by eddy motion, up to small dianeutral mixing, occurs along isoneutral surfaces. This is taken into account by the diffusivity tensor **K** defined above. Numerical implementation of isoneutral diffusivity is a challenge because it has to satisfy several properties mentioned above. The existing discrete implementations were formulated and analyzed for quadrilateral meshes (see, e.g., Griffies et al. [1998], Lemarié et al. [2012]). Additional analyses are required for triangular meshes. The description of discretization appropriate for triangular meshes of FESOM2 was provided in Danilov et al. [2017]. I worked on the implementation of isoneutral diffusivities in FESOM2 which was then tested in Scholz et al. [2019]. This chapter deals with the analysis of limitations on the time step required for numerical stability of isoneutral diffusion. First, it considers a 2D case to explain the procedure, and then continues with the analysis of 3D discretization of FESOM2.

## 3.1 2D discretization

The intention of this section is to explain computations using a 2D example. A fragment of 2D mesh is shown in Fig. 3.1. Indices $k$ and $i$ enumerate mesh cells in the vertical and horizontal directions respectively. The black dots show the placement of discrete tracer variables. Using the finite volume method, we can discretize equation 2.7 as follows

$$V_{k,i}\partial_t T_{k,i} = \int_{k,i} \nabla \mathbf{K} \nabla T d\Omega = \int_{k,i} (\mathbf{K}\nabla T)\mathrm{n} dS. \tag{3.1}$$

In this equation $V_{k,i}$ stands for the volume of the cell $\{k,i\}$ (area in two-dimensional case); $S$ stands for the surface of the volume $V_{k,i}$, and n is the

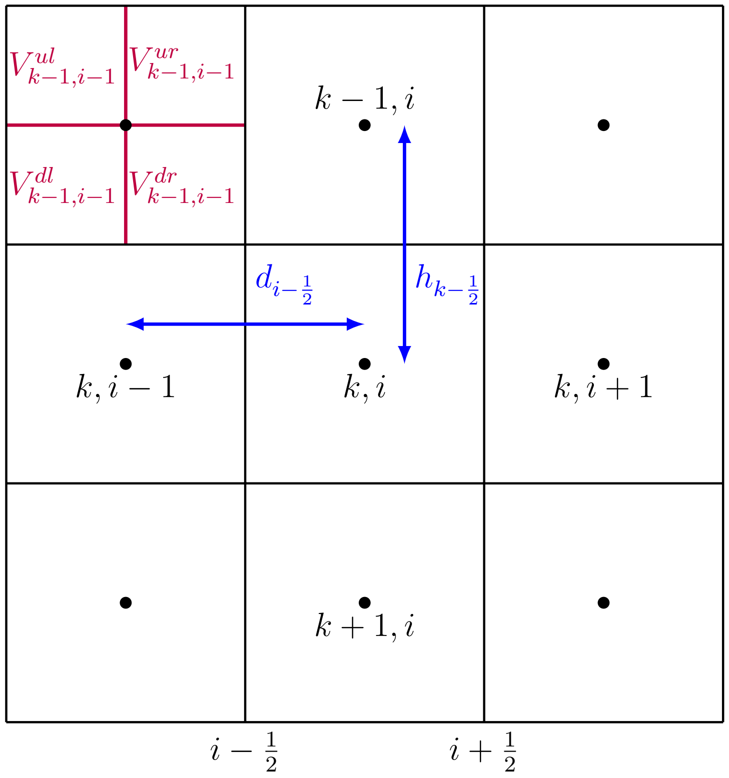


Figure 3.1: Discretization on a quadratic mesh with the step h in the vertical direction and step d in the horizontal direction. Subvolumes $V^j_{k,i}$ are shown in purple color in the left upper volume.

outer normal vector to the surface. In two-dimensional case the **K** matrix will have the following form

$$\mathbf{K} = K_{iso}\begin{pmatrix} 1 & s \\ s & s^2 \end{pmatrix}. \tag{3.2}$$

Discrete vertical derivatives of the tracer will be at horizontal faces (edges) of the cells, i.e at $k+1/2$ and so on. Horizontal derivatives will be at $i+1/2$ and so on. Since **K** combines vertical and horizontal derivatives, fluxes through the cell faces in (3.1) on each face will depend on the derivatives defined at other places. However, there is no immediately straightforward way to average vertical derivatives to vertical faces and horizontal derivatives to the horizontal faces so that it will ensure that the discrete variance decays.

First, let us introduce the discrete form of the dissipation functional $F$:

$$\mathcal{F} = -\frac{1}{2}\sum_{k,i}\int_{k,i} \nabla T \mathbf{K} \nabla T d\Omega. \tag{3.3}$$

Since horizontal and vertical derivatives are defined at different places, $V_{k,i}$ will be split in four pieces. Let us use indices u for up, d for down, l for left and r for right parts of the volume. Thus, the volume $V_{k,i}$ will be split into four smaller volumes each of which takes the fourth part of the volume of $V_{k,i}$, denoted as $V^{ul}_{k,i}, V^{ur}_{k,i}, V^{dl}_{k,i}, V^{dr}_{k,i}$ as it is shown on the Fig. 3.1.

Using these sub-volumes, we split the integral in (3.3) in four contributions

$$\mathcal{F} = -\frac{1}{2}\sum_{k,i}\sum_{j=1}^{4}(\nabla T)^j_{k,i}K^j_{k,i}(\nabla T)^j_{k,i}V^j_{k,i}. \tag{3.4}$$

Here j is array of indexes from 1 to 4 where 1 stands for the combinations ul, 2 for ur, 3 for dl and 4 for dr. Each of four sub-volumes of $V_{k,i}$ is characterized by unique combination of tracer $T_{k,i}$, and horizontal and vertical derivatives, producing a positive contribution to the sum in (3.4) because of the positivity of the matrix $K^j_{k,i}$. The small variation of the tracer now becomes the vector of perturbations of the discrete values $\delta T_{i,k}$. Since (3.4) is a quadratic form in $T_{i,k}$, the variation of (3.4) can be computed as

$$\delta\mathcal{F} = \sum_{k,i}\delta T_{k,i}\frac{\partial\mathcal{F}}{\partial T_{k,i}}. \tag{3.5}$$

On the other hand, in analogy with (2.11), we write

$$\delta\mathcal{F} = \sum_{k,i}\delta T_{k,i}R_{k,i}V_{k,i}. \tag{3.6}$$

Here $R_{k,i}$ is used to represent the right hand side of discrete diffusion equation (3.1) as $\int_{k,i}\nabla\mathbf{K}\nabla T d\Omega = V_{k,i}R_{k,i}$. Hence, we will get

$$V_{k,i}R_{k,i} = \frac{\partial\mathcal{F}}{\partial T_{k,i}}. \tag{3.7}$$

Thus, the right hand side of (3.1) is computed by differencing of the discrete dissipation functional. It can be seen that it ensures discrete variance decay (Property 2). Indeed,

$$\sum_{i,k}T_{k,i}V_{k,i}R_{k,i} = \sum_{k,i}T_{k,i}\frac{\partial\mathcal{F}}{\partial T_{k,i}} = 2\mathcal{F}.$$

The last equality is once again the consequence of $\mathcal{F}$ being a quadratic form in $T_{i,k}$. Property 1 will be ensured if the slope is computed separately in every four subvolumes by using the same expressions for vertical and horizontal derivatives as for the tracer.

Let us expand the term $(\nabla T)^j_{k,i}K^j_{k,i}(\nabla T)^j_{k,i}$ in (3.4) using directional derivatives and suppressing indices for briefness

$$\begin{aligned}(\nabla T)^j_{k,i}K^j_{k,i}(\nabla T)^j_{k,i} = (\partial_x T, \partial_z T)(K_{iso}(\partial_x T + s\partial_z T), K_{iso}(s\partial_x T + s^2\partial_z T)) = \\ K_{iso}((\partial_x T)^2 + 2\partial_x T\partial_z T s + s^2(\partial_z T)^2).\end{aligned} \tag{3.8}$$

Here $K_{iso}$ stands for isoneutral diffusivity in the matrix $\mathbf{K}$.

Let us have a closer look at the terms of the equation above. We will be writing down the contributions coming from the volume $V_{k,i}$ into the dissipation functional (3.4). For simplicity of explanations, we will assume that the slope $s$ is the same for the entire volume $V_{k,i}$. This assumption is compatible with Property 2 but violates Property 1.

Term 1: $(\partial_x T)^2 K_{iso}$.

The contribution from volume $V_{ki}$ and term 1 differs for the left and right sub-volumes; all sub-volumes lead to the expression

$$-\left[\frac{V_{ki}}{2}\left(\frac{T_{k,i+1}-T_{k,i}}{d_{i+1/2}}\right)^2+\frac{V_{ki}}{2}\left(\frac{T_{k,i}-T_{k,i-1}}{d_{i-1/2}}\right)^2\right]\frac{K_{iso}}{2}$$

Here $d_i$ is the step in $x$ (horizontal) direction. Therefore, the contribution to the right hand side according to equation 3.7 becomes

$$\begin{aligned}
V_{k,i}R_{k,i} &: K_{iso}\left[\frac{1}{d_{i+1/2}}T_{x,k,i+1/2}+\frac{1}{d_{i-1/2}}T_{x,k,i-1/2}\right]\frac{1}{2}V_{k,i},\\
V_{k,i+1}R_{k,i+1} &: -K_{iso}\frac{1}{d_{i+1/2}}T_{x,k,i+1/2}V_{k,i},\\
V_{k,i-1}R_{k,i-1} &: K_{iso}\frac{1}{d_{i+1/2}}T_{x_{k,i-1/2}}V_{k,i}.
\end{aligned}$$

In these formulas, $T_x$ is the shortcut for the discrete $x$-derivative at locations given by further subscripts. It is important to note that if $K_{iso}$ is varying as a function of $x$, it should be taken in the same way as $V_{k,i}$ (with indices $k,i$).

We can proceed by putting the result as written to the right-hand sides of three cells $\{k,i-1\}$, $\{k,i\}$ and $\{k,i+1\}$. We can also combine the contributions from the cells $\{k,i-1\}$, $\{k,i\}$ and $\{k,i+1\}$ into the right-hand side of the equation for $T_{k,i}$ as follows

$$\begin{aligned}
V_{k,i}R_{k,i} &: \frac{K_{iso}V_{k,i}}{2}\left[\frac{1}{d_{i+1/2}}T_{x,k,i+1/2}-\frac{1}{d_{i-1/2}}T_{x,k,i-1/2}\right]\\
&+\frac{K_{iso}V_{k,i+1}}{2}\frac{1}{d_{i+1/2}}T_{x,k,i+1/2}-\frac{K_{iso}V_{k,i-1}}{2}\frac{1}{d_{i-1/2}}T_{x,k,i-1/2}=\\
&K_{iso}\left[\frac{T_{x,k,i+1/2}}{2d_{i+1/2}}\Big(V_{k,i}+V_{k,i+1}\Big)-\frac{T_{x,k,i-1/2}}{2d_{i-1/2}}\Big(V_{k,i}+V_{k,i-1}\Big)\right].
\end{aligned}\tag{3.9}$$

This equation expresses the result in a flux form. We see the flux entering the cell $\{k,i\}$ through its right and left faces. The fluxes are, however, weighted with volumes of cells on both sides of faces.

Term2: $2\partial_x T \partial_z T s K_{iso}$.

The contribution of the cell $\{k,i\}$ to the $\mathcal{F}$ is:

$$-\frac{sV_{k,i}K_{iso}}{4}\left[\frac{T_{k-1,i}-T_{k,i}}{h_{k-1/2}}\frac{T_{k,i}-T_{k,i-1}}{d_{i-1/2}}+\frac{T_{k-1,i}-T_{k,i}}{h_{k-1/2}}\frac{T_{k,i+1}-T_{k,i}}{d_{i+1/2}}\right.$$
$$\left.+\frac{T_{k,i}-T_{k+1,i}}{h_{k+1/2}}\frac{T_{k,i}-T_{k,i-1}}{d_{i-1/2}}+\frac{T_{k,i}-T_{k+1,i}}{h_{k+1/2}}\frac{T_{k,i+1}-T_{k,i}}{d_{i+1/2}}\right].$$

In this equation $h_{k+1/2}, h_{k-1/2}$ stand for the distance in the $z$-direction between the respective cell centers. Note that here all four sub-volumes contribute differently. The slope $s$ is a common multiplier because of our simplification. It is computed separately for each sub-volume otherwise. As can be seen, five discrete tracer values are involved, so, that this term will generate five contributions to the right-hand sides. They are computed by taking derivatives

$$\begin{aligned}
V_{k,i}R_{k,i}:\ & \frac{sK_{iso}V_{k,i}}{4}\left[\frac{1}{h_{k-1/2}}T_{x,k,i-1/2}-\frac{1}{d_{i-1/2}}T_{z,k-1/2,i}\right.\\
& \frac{1}{h_{k-1/2}}T_{x,k,i+1/2}+\frac{1}{d_{i+1/2}}T_{z,k-1/2,i}\\
& -\frac{1}{h_{k+1/2}}T,x_{k,i-1/2}-\frac{1}{d_{i-1/2}}T,z_{k+1/2,i}\\
& \left.-\frac{1}{h_{k+1/2}}T,x_{k,i+1/2}+\frac{1}{d_{i+1/2}}T_{z,k+1/2,i}\right],\\
V_{k-1,i}R_{k-1,i}:\ & \frac{sK_{iso}V_{k,i}}{4}\left[-\frac{1}{h_{k-1/2}}T_{x,k,i-1/2}-\frac{1}{h_{k-1/2}}T_{x,k,i+1/2}\right],\\
V_{k+1,i}R_{k+1,i}:\ & \frac{sK_{iso}V_{k,i}}{4}\left[\frac{1}{h_{k+1/2}}T_{x,k,i-1/2}+\frac{1}{h_{k+1/2}}T_{x,k,i+1/2}\right],\\
V_{k,i-1}R_{k,i-1}:\ & \frac{sK_{iso}V_{k,i}}{4}\left[\frac{1}{d_{i-1/2}}T_{z,k+1/2,i}+\frac{1}{d_{i-1/2}}T_{z,k-1/2,i}\right],\\
V_{k,i+1}R_{k,i+1}:\ & \frac{sK_{iso}V_{k,i}}{4}\left[-\frac{1}{d_{i+1/2}}T_{z,k+1/2,i}-\frac{1}{d_{i+1/2}}T_{z,k-1/2,i}\right].
\end{aligned}$$

Here $T_z$ is the shortcut for vertical derivative, and indices point to its location. Similarly to computations above, the contributions from five cells into $R_{k,i}$ can be grouped to reveal that one deals with weighted fluxes through the faces of the cell $\{k,i\}$. I do not write them down here.

Term3: $(\partial_z T)^2 K_{iso}$.

The contribution of the cell $\{k,i\}$ to the functional $\mathcal{F}$ for the term 3 is

$$-\frac{s^2K_{iso}V_{k,i}}{4}\left[\left(\frac{T_{k-1,i}-T_{k,i}}{h_{k-1/2}}\right)^2+\left(\frac{T_{k,i}-T_{k+1,i}}{h_{k+1/2}}\right)^2\right].$$

In terms of variational calculation it will contribute to the term 3 as

$$\begin{aligned}
V_{k,i}R_{k,i} &: \frac{s^2K_{iso}V_{k,i}}{2}\left[\frac{1}{h^2_{k-1/2}}\Big(T_{k-1,i}-T_{k,i}\Big)-\frac{1}{h^2_{k+1/2}}\Big(T_{k,i}-T_{k+1,i}\Big)\right],\\
V_{k-1,i}R_{k-1,i} &: -\frac{s^2K_{iso}V_{k,i}}{2}\frac{T_{k-1,i}-T_{k,i}}{h^2_{k-1/2}},\\
V_{k+1,i}R_{k+1,i} &: \frac{s^2K_{iso}V_{k,i}}{2}\frac{T_{k,i}-T_{k+1,i}}{h^2_{k+1/2}}.
\end{aligned}$$

Here the derivative abbreviations are skipped because, as it turns out, this term has to be treated implicitly. For the same reason we have to assemble all the parts of the term 3 which contribute to $V_{k,i}R_{k,i}$ together:

$$\begin{aligned}
V_{k,i}R_{k,i} : \frac{s^2K_{iso}}{h^2_{k-1/2}}\Big(T_{k-1,i}-T_{k,i}\Big)\left[\frac{V_{k,i}}{2}+\frac{V_{k-1,i}}{2}\right]&\\
-\frac{s^2K_{iso}}{h^2_{k+1/2}}\Big(T_{k,i}-T_{k+1,i}\Big)\left[\frac{V_{k,i}}{2}+\frac{V_{k+1,i}}{2}\right]&.
\end{aligned}$$

Now all the terms are considered. The consideration of the 3D case on triangular meshes follows the same procedure. The difference is that each triangular prism will be split into six sub-prisms and that many triangular prisms contribute to a given scalar point.

## Stability analysis

The time stepping of isoneutral diffusion will be analyzed in this section. As it is common in such studies, our final intention will be to apply the Fourier analysis. For this reason, I will do some further simplifications. Let us assume the slope scalar $s$ and isoneutral diffusivity $K_{iso}$ from the matrix $\mathbf{K}$ defined in 3.2 to be constant. The mesh will be considered to be uniform with $V_{i,k} = hd$, where $h$, $d$ are the vertical and horizontal mesh steps respectively. Traditionally, the horizontal diffusivity in ocean codes is considered explicitly. Although the isoneutral diffusion is a replacement for the horizontal diffusion, it needs explicit-implicit treatment, as demonstrated in Lemarié et al. [2012]. Here I explain why. Explicit discretization of equation 2.7 in time can be written either as

$$T^{n+1}_j - T^n_j = K_{iso}\Delta t[\partial_x(\partial_x + s\partial_z)T^n + \partial_z(s\partial_x + s^2\partial_z T^n)], \tag{3.10}$$

where $n$ is the time step index and the time step length is denoted as $\Delta t$. In the explicit-implicit form everything concerning the horizontal operator will be explicit, and vertical term $\partial_{zz}$ will be computed implicitly

$$T^{n+1}_j - T^n_j = K_{iso}\Delta t\Big[(\partial_x\partial_x + \partial_x s\partial_z + \partial_z s\partial_x)T^n + s^2\partial_z\partial_z T^{n+1}\Big]. \tag{3.11}$$

For the analysis let us represent the spatial dependence of tracer $T$ as a Fourier harmonic

$$T = \tilde{T}e^{ikx+imz}. \tag{3.12}$$

In the continuous case for such a choice of $T$, $\partial_x T = ikT$ and $\partial_z T = imT$. In the discrete case considered above, as it can be readily seen, they should be replaced with $\partial_x T = (2i/d)\sin(kd/2)T$ and $\partial_z T = (2i/h)\sin(mh/2)T$. Furthermore, there will be additional factors related to averaging that is present in the expressions for fluxes obtained by variational method. For briefness we will omit them here, but full expressions will be used further in the 3D case. Since in the discrete case $|k| \leq \pi/d$ and $|m| \leq \pi/h$, the difference here will not be large to change the answer qualitatively. If we insert this representation in equations (3.10) and (3.11), we will get the relations

$$T_j^{n+1} - T_j^n = K_{iso}\Delta t\Big[(-k^2 - 2mks)T_j^n - m^2 s^2 T_j^n\Big] \tag{3.13}$$

and

$$T_j^{n+1} - T_j^n = K_{iso}\Delta t\Big[(-k^2 - 2mks)T_j^n - m^2 s^2 T_j^{n+1}\Big]. \tag{3.14}$$

For stability analysis I introduce $\lambda = T^{n+1}/T^n$. In order solutions do not diverge with time, i.e., for the scheme stability, it should be $|\lambda| \leq 1$. Expressing $\lambda$ we will get

$$\lambda_e = 1 - K_{iso}\Delta t(k^2 + 2mks + m^2 s^2) \tag{3.15}$$

and

$$\lambda_i = \frac{1 - K_{iso}\Delta t(k^2 + 2mks)}{1 + K_{iso}\Delta t m^2 s^2}. \tag{3.16}$$

Let us start from the explicit case. As it was mentioned before, $|\lambda|$ has to be less than 1. This brings us to the following relation

$$K_{iso}\Delta t|k^2 + 2mks + m^2 s^2| \leq 2 \Rightarrow K_{iso}\Delta t\frac{\pi^2}{d^2}(1+S)^2 \leq 2, \tag{3.17}$$

where $S = sd/h$.

For typical $K_{iso}$ and $d$ it does not cause severe limitations on time step if $S$ is small. The limitation on the time step $\Delta t$ becomes much stronger if $S$ is large. For example, if we take $d = 10$ km, $h = 10$ m and the slope $s = 10^{-3}$, then $S = 1$. With $K_{iso} = 300 m^2/s$ we can evaluate the limiting time step as $\Delta t \leq 8$ hours. However, if $d$ is larger, or $h$ is smaller, or $s$ is larger, we can get a strong limitation implying that $\Delta t$ gets smaller in $S^2$ times. For instance, with $S = 10$, the time step approximately has to be $\Delta t \leq 9$ min which is already quite small. With $S = 20$ it gets really limiting. In reality, the problem will happen earlier, as diffusion will be combined with advection, and the latter will contribute to the time step limitation as well.

Now let us go back to the explicit-implicit case. We rewrite $\lambda_i$ as

$$\lambda_i = \frac{1 - C(1+2S)}{1 + CS^2} = 1 - C\frac{(1+S)^2}{1+CS^2}, \tag{3.18}$$

where $C = K_{iso}\Delta t k^2$ and $S = s\frac{m}{k}$. Note that $S$ is different here from the previous case, yet it has similar sense. The condition $|\lambda_i| \leq 1$ implies that $C(1+S)^2/(1+CS^2) \leq 2$. Considering the left hand side of this inequality as a function of $S$, we find that its extremum is achieved at $S = -1$ or $CS = 1$.

Putting $S = -1$ in (3.18) we see that $\lambda = 1$ which means that there are no limitations in this case. For the other extremum we will get $\lambda = -C$. Therefore we have to require that $C = K_{iso}\Delta t k^2 < 1$. We see that in the explicit-implicit case the limitation does not depend on $S$ (and thus on using $ik$ and $im$ as spectral symbols of derivatives). If we want to avoid oscillations as well, we have to require that $\lambda$ stays positive:

$$1 - C(1+2S) > 0 \Rightarrow C < \frac{1}{1+2S} \quad \text{if} \quad 1 + 2S > 0.$$

Negative $S$ will not cause any problems, but positive values of $S$ may limit the admissible time step for a large $S$. Even this limitation is less strong than the limitation of the explicit case.

In summary, this implies that isoneutral diffusion has to be treated in an explicit-implicit way.

## 3.2 3D discretization on triangular meshes

A 3D mesh used by FESOM2 is based on a triangular surface mesh and a set of horizontal level surfaces as it was described in chapter 2. Each surface triangle defines a triangular prism going vertically down to the bottom. It is cut by the horizontal level surfaces into smaller prisms. A dual set of prisms is created by first constructing the so-called median-dual cells around each surface vertex (see Fig. 3.2). Such a cell is formed by connecting centers of triangles with mid-points at their edges. For a regular surface mesh made of equilateral triangles, these cells are hexagonal cells of dual mesh. Scalar degrees of freedom are located in the middle of such dual prisms in the vertical, and under the vertices in horizontal directions.

Vertical discretization is therefore similar to the 2D case above. The horizontal discretization needs special treatment. When discussing it, we will be using a horizontal plane and saying that scalar degrees of freedom are at vertices.

The main aspects of the discretization of isoneutral diffusion are based on the variational approach on meshes used by FESOM2 (see details in Danilov et al. [2017]). I implemented this discretization in FESOM2 and complemented it by the analysis of its numerical stability presented further. Some aspects of the performance of the isoneutral diffusion implementation are tested in Scholz et al. [2019].

The vertex placement of scalar degrees of freedom in the horizontal plane implies that the natural positions of horizontal gradients are at the centers of triangular prisms. Vertical gradients are therefore located at vertices of triangular prisms (see Fig.2.3). To have a unique combination of vertical and horizontal derivative, each triangular prism is split into six sub-prisms cut by the mid-plane and the faces of a dual prism (see Fig. 2.4).

As the result, the dissipation functional is obtained as a double sum over triangular prisms and six sub-prisms. The resulting right-hand side of the tracer equation is, however, obtained in the same way as in the 2D case above by differentiation of dissipation functional. The expressions are even bulkier than the expressions for the 2D case above and are not presented here.

The goal of the rest of this section is the time stepping stability analysis for the 3D discretization of FESOM2.

Let us consider a regular triangular surface mesh consisting of equilateral triangles. In FESOM2 scalar degrees of freedom are placed at the vertices of a triangular mesh. As it was mentioned above, the natural location for horizontal gradients is at centroids of triangles. However, a triangular mesh includes triangles of two different orientations, as shown in Fig.3.2. We will distinguish between triangles pointing upward ($u$-triangles) and downward ($d$-triangles). This is needed because one type of triangles cannot be obtained from the other type by translations, and this has to be taken into account in the Fourier analysis. To carry out such an analysis, we assume that the slope vector $\mathbf{s}$ is constant, and that the mesh is uniform in horizontal and vertical direction. We also assume for the horizontal part

$$T_j = \tilde{T} e^{ikx_j + ily_j}, \tag{3.19}$$

where $j$ enumerates all vertices of surface mesh, $k, l$ are the zonal and meridional wavenumbers respectively, and $x_j, y_j$ are the coordinates of vertex $j$ in the horizontal plane. The length of the triangle side will be denoted as $d$. I suppressed the vertical index in the discussion of the horizontal part. In Fig. 3.2 discrete tracer values $T_i$ are defined at the nodes of a triangles, and the notation $\mathbf{n_i}$ stand for normal vectors.

If one makes calculations on a uniform mesh and with the constant $\mathbf{s}$ through the variational method, they will lead to the standard expressions for fluxes. Exception is the case of vertical and horizontal averaging in computations of mixed terms. I therefore start from the continuous expressions

$$\nabla\mathbf{K}\nabla T = K_{iso}\Big[\partial_x(\partial_x+s_x\partial_z)+\partial_y(\partial_y+s_y\partial_z)+\partial_z(s_x\partial_x+s_y\partial_y)+\partial_z s^2\partial_z\Big]T, \quad (3.20)$$

and replace derivatives with spectral symbols for the discrete case, taking into account averaging, as explained below.

First, let us consider a $d$-triangle. Here $\mathbf{n_1} = (\frac{\sqrt{3}}{2}, -\frac{1}{2})$, $\mathbf{n_2} = (0,1)$ and $\mathbf{n_3} = (-\frac{\sqrt{3}}{2}, -\frac{1}{2})$ are normal vectors to the sides of the triangle. Using equation (3.19) and taking into account that the length of the triangle side is $d$, the tracer values will take the following form

$$\begin{aligned} T_1 &= \tilde{T}e^{-ik\frac{d}{2}+il\frac{\sqrt{3}}{6}d}, \\ T_2 &= \tilde{T}e^{-il\frac{\sqrt{3}}{3}d}, \\ T_3 &= \tilde{T}e^{ik\frac{d}{2}+il\frac{\sqrt{3}}{6}d} \end{aligned} \quad (3.21)$$

if coordinates are measured from the triangle center. If linear interpolation is assumed on a triangle, each its vertex contributes to the horizontal gradient of $T$ as $T_j\mathbf{n}_j/h$, where $h = d\sqrt{3}/2$ is the triangle height. Combining these contributions, we will obtain expressions for $x$ and $y$ components of the horizontal gradient on $d$ triangle in the spectral form (spectral symbols)

$$\begin{aligned} G_x^d &= \frac{2i}{d}\sin\Big(\frac{kd}{2}\Big)e^{il\frac{\sqrt{3}}{6}}d, \\ G_y^d &= \frac{2}{\sqrt{3}d}\Big[\cos\Big(\frac{kd}{2}\Big)e^{il\frac{\sqrt{3}}{6}d} - e^{-il\frac{\sqrt{3}}{3}d}\Big]. \end{aligned} \quad (3.22)$$

Similarly, we can get the expressions for a $u$-triangle which are minus complex conjugate of those for $d$ triangle (3.22).

$$\begin{aligned} G_x^u &= \frac{2i}{d}\sin\Big(\frac{kd}{2}\Big)e^{-il\frac{\sqrt{3}}{6}}d = (-G_x^d)^*; \\ G_y^u &= \frac{2}{\sqrt{3}d}\Big[-\cos\Big(\frac{kd}{2}\Big)e^{-il\frac{\sqrt{3}}{6}d} + e^{il\frac{\sqrt{3}}{3}d}\Big] = (-G_y^d)^*. \end{aligned} \quad (3.23)$$

The outer derivatives in expression (3.20) are related to the divergence. Therefore, we have to find the expression for the discrete divergence. In Fig. 3.2

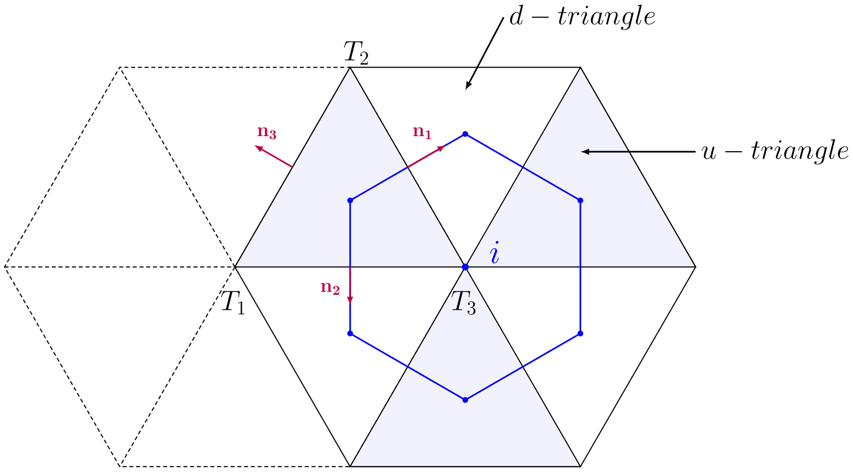


Figure 3.2: The full horizontal stencil of a scalar control volume involving of six triangles. Neighbouring elements are shown with the dotted lines. The values of tracers ($T_1$, $T_2$ and $T_3$) are located in the vertices of a triangle. Vectors $\mathbf{n_1}$, $\mathbf{n_2}$ and $\mathbf{n_3}$ are outer normals to the sides of the triangle.

the hexagon shows the scalar control cell around vertex $i$. If there is a vector field $\mathbf{u} = (u, v)$ with discrete values defined at the triangle centers, the divergence will be obtained by summing the contributions from segments of the boundary. One readily obtains that in the Fourier representation

$$(\partial_x u + \partial_y v)_i = \frac{1}{2}[G_x^d u^u + G_x^u u^d + G_y^d v^u + G_y^u v^d].$$

The vertical component of the gradient of tracer $T$ is defined in the middle points of levels as

$$T_{z,k-1/2,j} = \frac{T_{(k-1),j} - T_{k,j}}{h}. \tag{3.24}$$

For the Fourier analysis the vertical dependence is selected proportional to $e^{imz}$. In this case the spectral symbol of the vertical gradient component becomes

$$G_z = \frac{2i}{h} \sin \frac{mh}{2}. \tag{3.25}$$

The vertical contribution to the divergence is $-G_z^* = G_z$.

We have now all partial derivatives in the Fourier representation and can proceed with the isoneutral diffusion operator. The diagonal terms of the operator take the form

$$\begin{aligned}
\partial_x \partial_x &= \frac{1}{2}(G_x^d G_x^u + G_x^u G_x^d) = -G_x^u (G_x^u)^*, \\
\partial_y \partial_y &= \frac{1}{2}(G_y^d G_y^u + G_y^u G_y^d) = -G_y^u (G_y^u)^*, \\
\partial_z \partial_z &= \frac{1}{2}(G_z^d G_z^u + G_z^u G_z^d) = -G_z G_z^* s^2.
\end{aligned} \tag{3.26}$$

For the mixed derivatives horizontal and spatial averaging has to be taken into account.

$$\partial_x(s_x\partial_z) = \frac{1}{2}\Big(G_x^d(s_x\partial_z)^u + G_x^u(s_x\partial_z)^d\Big). \tag{3.27}$$

Averaging of $s_x\partial_z$ to $u$-triangle location introduces the factor

$$f_h = \frac{1}{3}\Big(e^{2il\frac{\sqrt{3}d}{6}} + e^{-ik\frac{d}{2}-ild\frac{\sqrt{3}}{6}} + e^{ik\frac{d}{2}-ild\frac{\sqrt{3}}{6}}\Big), \tag{3.28}$$

and vertical averaging to mid-layer will introduce the factor

$$f_v = \cos(mh/2).$$

We combine both factors to $f = f_h f_v$. Averaging of $s_x\partial_z$ to $d$-triangle location will lead to the complex conjugate factor. Same averaging coefficients will appear in the averaging $(s_x\partial_x)$ to vertex locations and levels. As the result, mixed derivatives are expressed as follows

$$\begin{aligned}\partial_x(s_x\partial_z) &= \frac{1}{2}\Big(G_x^d f + G_x^u f^*\Big)s_x G_z,\\ \partial_z(s_x\partial_x) &= \frac{1}{2}\Big(f^* G_x^u + f G_x^d\Big)s_x(-G_z^*).\end{aligned} \tag{3.29}$$

The expressions involving meridional direction are obtained by exchanging $x$-indices with $y$-indices.

## Stability analysis

Knowing from the 2D case that the explicit-implicit time stepping is the most essential for better stability, I consider this case only. Acting in the same way as in 2D case above, we get

$$\lambda_i = \frac{1 - K_{iso}\Delta t[G_x^{u*}G_x^u + G_y^{u*}G_y^u + [(G_x^d f + G_x^u f^*)s_x + (G_y^d f + G_y^u f^*)s_y]G_z]}{1 + K_{iso}\Delta t G_z^* G_z s^2}. \tag{3.30}$$

The structure of this expression is similar to that of two dimensional case, and stability implies

$$K_{iso}\Delta t\frac{G_x^{u*}G_x^u + G_y^{u*}G_y^u + [(G_x^d f + G_x^u f^*)s_x + (G_y^d f + G_y^u f^*)s_y]G_z + G_z^* G_z s^2}{1 + K_{iso}\Delta t G_z^* G_z s^2} \le 2. \tag{3.31}$$

The combination $G_x^{u*}G_x^u + G_y^{u*}G_y^u$ is the discrete analog of the horizontal wavenumber squared, and I denote it $K^2$. I also denote $S = s|G_z|/K$ and $C = K_{iso}\Delta t K^2$. The inequality above will be then rewritten as

$$C\frac{1 + 2bS + S^2}{1 + CS^2} \le 2,$$

where

$$b = [(G_x^d f + G_x^u f^*)s_x + (G_y^d f + G_y^u f^*)s_y]/(2sK).$$

Since $f$ is due to averaging, $|f| \leq 1$, and therefore $|b| \leq 1$. Indeed, in the absence of $f$ $b$ is a dot product of two vectors divided on their amplitudes, and $f$ can only reduce its amplitude. Solving the inequality with respect to $C$ we find

$$C \leq \frac{2}{1 + 2bS - S^2}.$$

Denominator reaches maximum for $S = b$ which equals $1 + b^2$. To be valid in the worst case, we must require $C \leq 1$. As in the 2D case we see that the limitation on the time step does not depend on $S$ as would be the case for fully explicit time stepping. The limitation obtained is only twice worse than the limitation on purely horizontal diffusion ($C \leq 2$). Such limitation is as a rule much less restrictive than other limitations in ocean codes (see, e.g., Lemarié et al. [2015]).

Once again, one is not necessarily satisfied with formal stability and may require absence of oscillations, i.e. that $\lambda_i$ stays positive. In this case the answer will be $C \leq 1/\max(1 + 2bS, 0 + \epsilon)$. For large $S$ the condition can be essentially more demanding, yet it is less restrictive than for purely explicit time stepping.

It turns out that for stability analysis the precise form of spectral symbols for derivatives is not important. However, it is needed to convert the stability criterion on $C$ into stability criterion on $\Delta t$. For this let us define the maximum value of $K^2d^2$ (we use $K^2d^2$ in place of $K^2$ because it is dimensionless). From the definition of $K^2$ we will get

$$K^2d^2 = 4\sin^2\frac{kd}{2} + \frac{4}{3}\Big(\cos^2\frac{kd}{2} - \cos\frac{kd}{2}\cos\frac{ld}{2} + 1\Big).$$

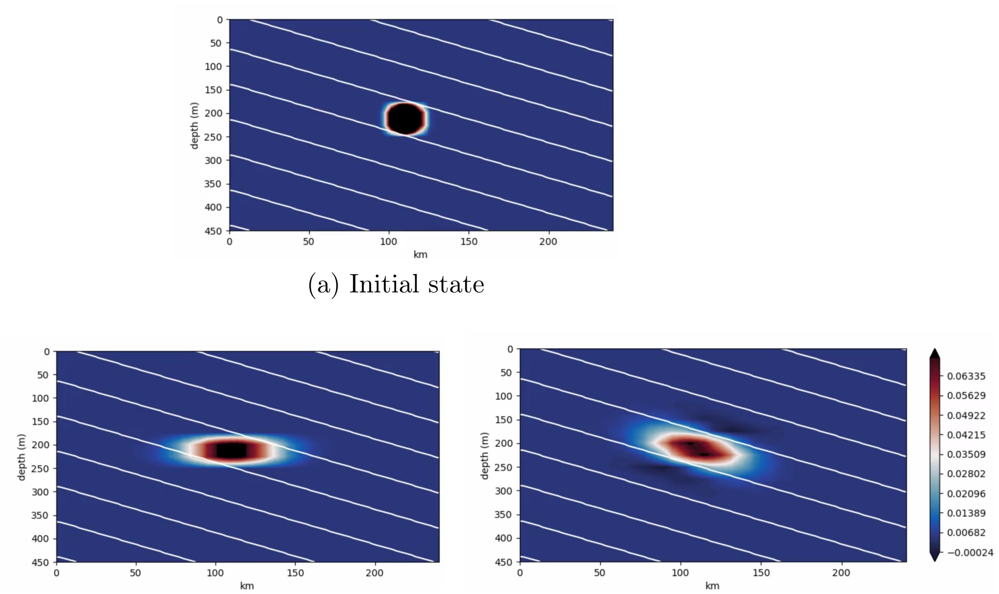
Maximum value of $K^2d^2$ on triangular grid for vertex placement of scalars turns out to be 16/3, implying that $\Delta t \leq 3d^2/(16K_{iso})$. The condition on the absence of oscillations implies $\Delta t \leq 1/(K_{iso}K^2(1 + 2bS))$. Taking $b = 1$, the worst value of $|G_z| = 2/h$ and maximum value of $K^2$, we find $\Delta t \leq 3d^2/(16K_{iso}(1 + \sqrt{3}sd/h))$. This is more limiting if $sd/h$ is large, but even in this case it is less limiting as the explicit time stepping. For $d = 10^3$ m and $K_{iso} = 300$ m$^2$/s the time stepping will be stable for $\Delta t \leq 17.4$ h, and will show no oscillations for $sd/h = 10$ if $\Delta t \leq 1$ h. Larger values of $K_{iso}$ or slope can make limitations stronger, but still unlikely dominant.

## 3.3 Implementation in FESOM2

The isoneutral diffusivity as described above was implemented in FESOM2.

For the test, a 20° wide channel centered at the latitude 38° and having the depth of 1600 metres was considered. All the forcing and dianeutral vertical mixing were switched off in the system. The system evolution is only due to diffusion. The full equation of state is used. The surface mesh obtained by splitting reqular quadrilaterals with a side 18 km in two triangles was taken. The vertical resolution is smoothly varying from 10 m to 50 down to 200 m depth and then is set to 100 m.

Salinity is constant and set to 35 PSU and temperature is set to form with tilted isotherms, so, the density stratification is due to temperature only. A passive tracer is set to zero everywhere but for a spherical patch in the middle of the channel (see Fig. 3.3). Only the equation for passive tracer was integrated. In this case isopycnals do not change with time, and we can observe diffusion of the passive tracer. The Fig. 3.3 demonstrates how diffusion ignores slope of density layers when we switch off the isoneutral diffusion, and how it becomes aligned with isopycnals with time otherwise.



(a) Initial state

(b) Isoneutral diffusion method is not applied

(c) Isoneutral diffusion method is applied

Figure 3.3: The passive tracer concentration initially (a) and after 5 years of integration with isoneutral diffusion (c) and without it (b). White lines indicate isopycnals.

As we can see from the figure 3.3, change in the system happen only due to diffusion. Isopycnals are formed by changes in temperature. When isonuetral diffusion is not applied, diffusion of the passive tracer goes along vertical layers ignoring isopycnals while with this method passive tracer diffuses along the density layers.

## 3.4 Conclusion

The method described in the current chapter is well known and used in models working on quadrilateral meshes. Its realization and stability analysis on rectangular meshes were described by Lemarié et al. [2012]. However, it was not analysed and fully implemented on triangular meshes. In this chapter I explained the implementation of isoneutral diffusion on triangular meshes and made stability analysis to get conditions when it is stable and do not produce oscillations. As a result, it is proven that the explicit-implicit treatment, same as on quadrilateral meshes, does not create severe limitations on triangular meshes of FESOM2, similarly to the conclusion of Lemarié et al. [2012] concerning quadrilateral meshes.

Also, isoneutral diffusion method was implemented in FESOM2. Numerical analysis of the method is provided in chapter 5. However, implementation of biharmonic isoneutral diffusion in FESOM2 did not bring desired effect on aligning the diffusion of the passive tracer along isoneutral surfaces. With biharmonic isoneutral diffusion I could not reach stable behavior for the passive tracer diffusion. This case requires further investigations which are not covered by the current work.

# Chapter 4

# Numerical advection schemes

This chapter is a compilation of the paper "Comparison of several high-order advection schemes for vertex-based triangular discretization" written by the author together with S. Danilov and published in Ocean Dynamics, in 2020 (see Smolentseva and Danilov [2020]).

## 4.1 Introduction

The third-order upwind advection schemes are a common choice in many ocean circulation models, formulated on structured meshes (see, e.g., Lemarié et al. [2015], Soufflet et al. [2016]). The large-scale unstructured-mesh ocean circulation models MPAS (Ringler et al. [2013]) and FESOM2 (Danilov et al. [2017]) use the unstructured-mesh analogs of this approach, as a rule in a version blending the third (upwind) and fourth (centered) order estimates, combined with the FCT (flux corrected transport) procedure. It is believed that the residual biharmonic diffusion introduced by the third-order upwind schemes presents a good compromise between the high accuracy and the need to damp perturbations at grid scales with unphysical behavior (see, e.g., Soufflet et al. [2016]). A question is whether there are advection schemes for vertex-based scalars on unstructured meshes that have a comparable computational cost but introduce even smaller dissipation in eddy-rich regimes of high-resolution global ocean simulations. We limit ourselves to the schemes that are sufficiently cheap in the sense that advection of temperature and salinity takes about or less than 50% of the time step of full 3D primitive-equation general ocean circulation model.

Our interest in the question is motivated by the development of the new finite-volume (FV) dynamical core of FESOM2 (Danilov et al. [2017]), which uses vertex placement of scalar variables, similar to the finite-element FESOM1.4 (Wang et al. [2014]). It turned out that it is rather difficult to propose an inexpensive FV advection scheme that compares in performance to the Taylor-Galerkin (TG) scheme of FESOM1.4, augmented with the flux corrected transport (FCT) limiter and using a consistent mass matrix (i.e., the FE FCT method of Löhner et al. [1987]). In spite of its nominally second order (linear polynomial representation on triangles), it easily outperforms the standard third and fourth order FV schemes tested by us (see below). The reason is that the inversion of consistent mass matrix in the TG method removes the leading spatial dispersive error, resulting in an effectively fourth-order method with a rather small residual term (see, e.g. Donea and Huerta [2003]). The question therefore is whether similar behavior can be reached in a FV code where mass matrices do not appear, and even if introduced, would destroy the locality of fluxes at interfaces of control volumes. We propose an elementary solution suitable for the FV method, which in idealized tests generally shows lower errors than the common third-fourth order methods, requiring lower computational resources. For uniform meshes, the idea of the method resembles that of the compact fourth-order method discussed, for example, in Lemarié et al. [2015] and Shchepetkin [2015], or in Zerroukat et al. [2006] as related to the parabolic spline method. For this reason we will refer to the proposed scheme as the compact scheme. Its description presents the main goal of present work.

We note from the very beginning that because the advecting velocity field in the ocean varies on the same spatial scale as temperature and salinity, the practical accuracy of the advection will be only second-order. The search for higher-order transport algorithms is motivated by the observation that they generally provide smaller residual errors and spurious mixing (see, e.g., Mohammadi-Aragh et al. [2015]).

We will discuss the schemes in two dimensions. The unstructured meshes used in ocean modeling are vertically aligned, so that the vertical dimension is not different from that on structured meshes. The modifications needed to extend the descriptions to 3D case are straightforward and consist in adding fluxes through the top and bottom faces of the respective control volumes. The transport equation will be taken in a flux form

$$\partial_t T + \nabla \cdot (\mathbf{u}T) = 0, \tag{4.1}$$

where $\mathbf{u}$ is the given velocity, and $T$ an arbitrary tracer (like mass fraction if $\nabla \cdot \mathbf{u} = 0$ or like density otherwise). The solution is sought in some domain $\Gamma$

subject to the condition of no flux through its impermeable lateral boundaries. Initial conditions and the domain will be specified below.

Since the vertex placement of scalar variables on triangular meshes is equivalent to the cell placement of variables on hexagonal meshes if the meshes are uniform, their FV transport algorithms can be easily shared upon minimal adjustments. Several transport algorithms have been proposed recently for the hexagonal meshes. One group is based on polynomial reconstructions. The simplest choice is the linear upwind reconstruction scheme proposed by Miura [2007], which has been generalized by Skamarock and Menchaca [2010] to quadratic, cubic and quartic reconstructions, and later by Miura and Skamarock [2013] to a different variant of quadratic reconstruction based on a wider stencil and showing generally better accuracy. A related third-order scheme based on quadratic reconstruction has been proposed earlier by Chen et al. [2012]. It differs by the implementation of upwind fluxes, but essentially relies on the same stencil as Miura and Skamarock [2013]. Another group relies, in essence, on the standard finite-difference algorithm used to construct the third-order upwind or fourth-order centered implementation (see, e.g., Webb et al. [1998]). Skamarock and Gassmann [2011] adapt it to the third/fourth-order scheme on the Voronoi hexagons, and Miura [2013] discusses its further generalizations. Miura [2013] also proposes a procedure based on the gradient estimate (see further) which leads to the equivalent results. FESOM2 prototype implementation in Danilov [2012] uses the quadratic polynomial reconstruction (third-order) and gradient estimate schemes (of the third and fourth order) that are analogous to those of Skamarock and Menchaca [2010] and Skamarock and Gassmann [2011] respectively. The conclusion of Skamarock and Gassmann [2011] same as Danilov [2012] is that the methods based on the quadratic polynomial reconstruction are already less numerically efficient than the other group of methods, and do not demonstrate better accuracy. However, in our present implementation they turn out to be similarly numerically efficient.

Although these approaches are new to the ocean or atmosphere modeling, they are well known in computational fluid dynamics, see, e.g., Barth and Frederickson [1990], Ollivier-Gooch and Van Altena [2002] as concerns the polynomial reconstruction, and Abalakin et al. [2002] as concerns the gradient estimates. There is extensive literature on high-order unstructured-mesh weighted essentially non-oscillatory schemes (WENO, see, e.g., Dumbser and Käser [2007] and references therein) which can be very accurate, but are computationally more demanding than the methods mentioned above, except for the third-order, and it remains to be seen how to make them fast enough. For the cell-based tracers Ye et al. [2019] show that the third-order WENO scheme is an affordable and robust performer.

Among these methods, the gradient estimate allows obvious generalizations toward the methods of the fifth and sixth order at a moderate additional cost (Abalakin et al. [2002]), and we explore below whether this possibility can be beneficial for ocean models.

The main goal of this work is to describe the compact scheme and compare its performance to that of the traditional 3rd/4th order algorithms and 5th/6th order extension of the gradient-based scheme described in Abalakin et al. [2002]. In contrast to the high-order algorithms based on the polynomial reconstruction, which are expected to preserve their order even on distorted meshes, other algorithms discussed here (including the compact and 5th/6th order schemes) do this formally only on uniform meshes. Since unstructured meshes used in the ocean modeling are of variable resolution, a practical question is how the methods relying on the mesh uniformity compared to the reconstruction-based methods on general non-uniform meshes.

The structure of this work is as follows. In section 4.2 we explain the proposed compact method, whereas a brief summary of well-known methods based on polynomial reconstruction and gradient estimates is presented in the Appendix, including the description of the fifth and sixth order extension. Section 4.3 compares the advection schemes based on a simple 2D test case of a circular shear flow and the Stommel gyre flow, and also discusses their functioning in FESOM2. These sections are followed by short discussion and conclusions.

## 4.2 Elementary considerations

Figure 4.1 shows schematically a control volume around the vertex $v_1$ obtained by connecting mid-edges with triangle centroids. Its area $S_{v_1}$ is given by the sum $S_{v_1} = \sum_{c\in C(v_1)} S_c/3$, where $C(v_1)$ is the set of cells (triangles) containing $v_1$, and $S_c$ is the cell area. Introducing $T_i = (1/S_i)\int_{S_i} TdS$ where $i$ is the vertex index and $S_i$ is the area of the median-dual control volume associated to the vertex $i$, the discrete representation of (4.1) becomes

$$S_i\partial_t T_i + \sum_{e\in E(i)} \sum_{s\in S(e)} \int_{l_s} (\mathbf{u}\cdot\mathbf{n}T)_s dl = 0, \qquad (4.2)$$

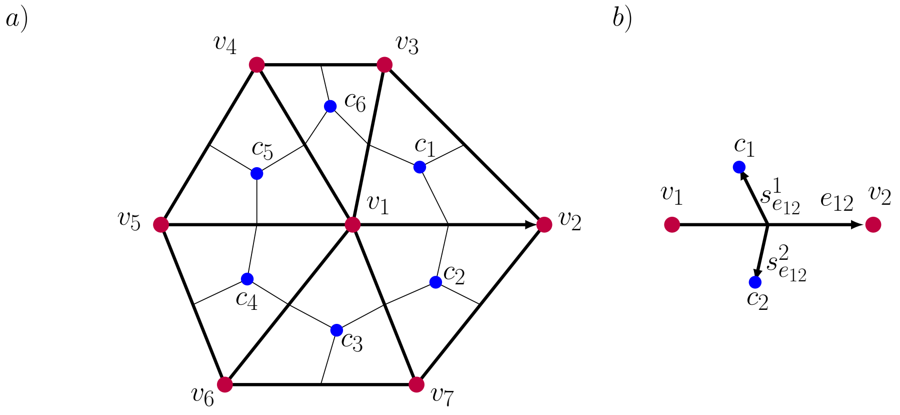


Figure 4.1: Schematic of the median-dual control volumes. a) For each edge fluxes are computed through the two segments of boundary (connecting mid-edge point to cell centroids). Here $v_i$ stands for vertices and $c_i$ - for the centres of triangles. b) $e_{12}$ is an edge coming from the vertex $v_1$ to the vertex $v_2$, and $s_{e_{12}}$ are two segments of this edge.

where the first summation is over edges $e$ emanating from the vertex $i$ ($E(i)$ is the set of such edges), the second one is over the segments $s$ connecting mid-point of the edge $e$ with the centers of cells on both its sides ($S(e)$ is the set of these segments), and integration is over the segment length. In Fig. 4.1 the edge with vertices $v_1, v_2$ contributes with the flux through segments connecting its midpoint to $c_1$ and $c_2$, and similarly for other edges. The remaining notation is the outer normal $\mathbf{n}$ to the segment and its length element $dl$. For a given velocity the spatial discretization reduces to specifying the estimate of a tracer being advected through the boundary segments.

## Polynomial reconstruction versus gradient estimate

In FV schemes based on field reconstruction, one writes a polynomial reconstruction around vertex $i$

$$\mathcal{T}_i = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 y^2 + a_5 xy + ...,$$

where $x, y$ are the horizontal coordinates in the local reference frame associated to vertex $i$, the coefficients $a_n$, $n$ =0, 1, 2, ... are found by imposing a strong constraint $\int_{S_i} \mathcal{T} dS = T_i S_i$ at vertex $i$ and similar constraints at a sufficient number of neighboring locations (given by index $j$), but in a weak sense. The resultant least square problem

$$\mathcal{L} = \sum_{j \in N(i)} w_j^2 |S_j^{-1} \int_{S_j} \mathcal{T} dS - T_j|^2 + \lambda (S_i^{-1} \int_{S_i} \mathcal{T} dS - T_i) = min$$

is solved for the coefficients $a_n$, and their values are used to estimate fields and thus flux leaving the control volume. Here $N(i)$ is a set containing a sufficient number of neighbor vertices; the weights $w_j$ are (commonly) the inverse distances from vertex $i$ to a neighbor vertex $j$, and $\lambda$ is the Lagrange multiplier. In this polynomial expansion $a_0$ is the tracer value at location $i$ which generally differs from the area averaged value $T_i$. The estimates of the fluxes leaving control volumes should be carried out at properly selected Gaussian quadrature points at the boundary segments to ensure accuracy. All matrices needed to compute the coefficients $a_n$ in terms of $T_i$ and surrounding $T_j$ (for $j \in N(i)$) are computed in advance and stored. This is the essence of schemes based on high-order reconstructions as proposed by Barth and Frederickson [1990] and developed further in many works that followed. Instead of neighboring locations one may select any additional locations as, for example, in Chen et al. [2012] or Miura and Skamarock [2013]. Generalizations may rely on several reconstruction stencils and WENO procedure (see, e.g. Dumbser and Käser [2007]).

On good quality triangular meshes, a vertex has 6 nearest neighbors, which is sufficient for a quadratic reconstruction (QR). A scheme based on quadratic reconstruction was implemented in prototype FESOM2 (Danilov [2012]) for median-dual control volumes. In its standard version, a reconstruction from the upwind vertex is used, but any combination of upwind and centered versions is possible. On 3D $z$-coordinate meshes the number of neighbors might vary with depth, and quadratic reconstruction is replaced by the linear one when bottom topography is encountered. In practice it turns out that QR schemes are nearly same accurate as the third and fourth order schemes based on the gradient estimate (GE) to be discussed below. This statement is similar to the conclusion in Skamarock and Gassmann [2011], who proposed a scheme that can be reformulated in terms of gradient estimates (see Miura [2013]). In early implementation in FESOM the QR schemes were about twice as expensive as the GE schemes explained further, which was an argument in favor of the latter. They are even better now.

A wider stencil is needed for higher-order polynomial reconstructions, and, although such reconstructions open unlimited possibilities even on strongly distorted meshes (see examples in Dumbser and Käser [2007]), their computational cost increases, as well as the halo size in parallel implementations. According to Skamarock and Menchaca [2010], the QR is optimal judged by accuracy per computational cost in standard tests where the convergence rate is about the second order.

The gradient estimate schemes achieve high-order only on uniform meshes, and stay second-order otherwise. We begin with a 1D simplification to explain them. Let the velocity $u$ be positive and uniform. Traditionally to estimate fluxes leaving a control volume around vertex $i$ one needs to estimate the tracer at $i+1/2$ based on $T_i$ and $T_j$ of neighboring control volumes. The procedure can be cast in terms of gradients.

One first introduces the upwind and downwind estimates

$$T^-_{i+1/2} = T_i + (h/2)G^-_{i+1/2}, \quad T^+_{i+1/2} = T_{i+1} - (h/2)G^+_{i+1/2},$$

where $h$ is the uniform grid spacing. The simplest high-order choice is to estimate the gradients $G^\pm_{i+1/2}$ as

$$G^-_{i+1/2} = (2/3)G^c_{i+1/2} + (1/3)G^u_{i+1/2}, \quad G^+_{i+1/2} = (2/3)G^c_{i+1/2} + (1/3)G^d_{i+1/2}.$$

Here $G^c_{i+1/2} = (T_{i+1} - T_i)/h$, $G^u_{i+1/2} = (T_i - T_{i-1})/h$ and $G^d_{i+1/2} = (T_{i+2} - T_{i+1})/h$ are the centered, upwind and downwind estimates of gradient respectively. Writing the flux

$$F_{i+1/2} = (1/2)(F^-_{i+1/2} + F^+_{i+1/2}) + (\lambda/2)(F^-_{i+1/2} - F^+_{i+1/2})sign(u),$$

one obtains the standard third-order upwind method for $\lambda = 1$ (GE3) or the forth-order centered method for $\lambda = 0$ (GE4). Intermediate $\lambda$ can be used to reduce dissipation and increase accuracy ($\lambda = 0.15 - 0.25$ works well in practice). One readily sees that the estimates $T^\pm_{i+1/2}$ can be rewritten as

$$T^-_{i+1/2} = (1/2)(T_i + T_{i+1}) - (1/6)h^2\delta^2 T|_i, \tag{4.3}$$

and

$$T^+_{i+1/2} = (1/2)(T_i + T_{i+1}) - (1/6)h^2\delta^2 T|_{i+1}, \tag{4.4}$$

where $\delta^2$ stands for the operator of second derivative. They are the familiar estimates used to construct the standard third-order finite-difference method.

The advantage of the GE is that it can be generalized to arbitrary triangular meshes and also to higher orders (see, for example, Abalakin et al. [2002]) by extending estimates to wider stencils. Consider the arrangement shown in Fig. 4.1 and Fig. 4.2. We denote the vector connecting vertices $i$ and $j$ of edge $e$ in Fig. 4.2 as $\mathbf{x}_{ij}$. For each edge $e$ we store the indices to up-edge ($u$) and down-edge triangles ($d$) that contain the continuation of the edge line. Such storage is sufficient for the third and fourth order schemes. For the fifth and sixth order one additionally stores the indices of vertices forming the edges intersected by the

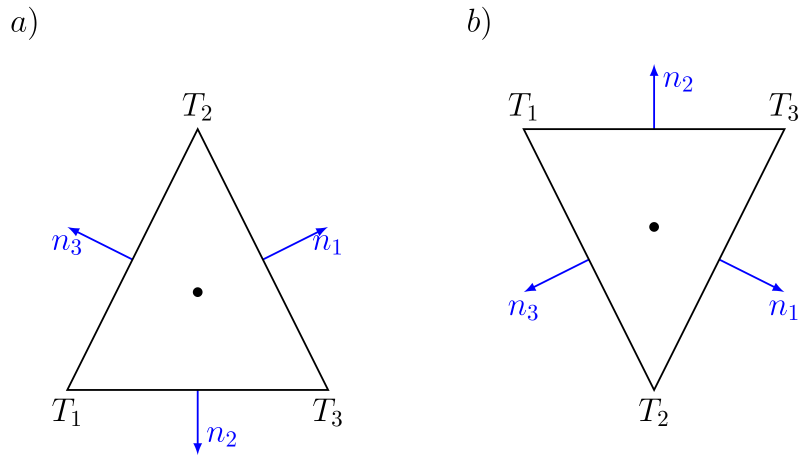


Figure 4.2: Schematic of the arrangement. Edge $e$ with vertices $i$ and $j$ is characterized by the edge vector $\mathbf{x}_{ij}$. $u$ and $d$ are up- and down-edge triangles, black circles are points where the continuation of edge $e$ intersects the sides of $u$ and $d$ triangles, $u_1, u_2$ and $d_1, d_2$ are the vertices related to these sides. The gradients on triangles are computed based on three vertex values, and the gradients at vertices are obtained by averaging over neighboring triangles.

continuation of edge $e$ in $u$ and $d$ triangles ($u_1, u_2$ and $d_1, d_2$) and the coefficients to interpolate the vertex gradients at these vertices to the intersection points. Similarly to the 1D case one writes

$$T_{ij} = T_i + (1/2)\mathbf{x}_{ij}\mathbf{G}_{ij}, \quad T_{ji} = T_j - (1/2)\mathbf{x}_{ji}\mathbf{G}_{ji}.$$

Here we use pairs $ij$ or $ji$ to indicate up-edge or down-edge reconstructions. The reconstruction is done to the edge mid-point, because the boundary of control volume passes through it for median-dual control volumes on triangular meshes. As an alternative to median-dual control volumes one can use Voronoi polygons of dual mesh. According to Abalakin et al. [2002] the following estimate for the gradients can be used

$$\mathbf{G}_{ij} = (1-\beta)\mathbf{G}^c + \beta\mathbf{G}^u + \delta_c(\mathbf{G}^u + \mathbf{G}^d - 2\mathbf{G}^c) + \delta_d(\mathbf{G}^j + \mathbf{G}^{u*} - 2\mathbf{G}_i)$$

and

$$\mathbf{G}_{ji} = (1-\beta)\mathbf{G}^c + \beta\mathbf{G}^d + \delta_c(\mathbf{G}^u + \mathbf{G}^d - 2\mathbf{G}^c) + \delta_d(\mathbf{G}^i + \mathbf{G}^{d*} - 2\mathbf{G}_j).$$

In these expressions $\mathbf{G}^u$ and $\mathbf{G}^d$ are the gradients on triangles $u$ and $d$, $\mathbf{x}_{ij}\mathbf{G}^c = T_j - T_i$ is the centered estimate, $\mathbf{G}^i$ and $\mathbf{G}^j$ are the estimate at vertices $i$ and $j$, and $\mathbf{G}^{u*}$ and $\mathbf{G}^{d*}$ are the vertex gradients interpolated to the edge continuation intersection points (see Fig. 4.2). The gradients on triangles are computed assuming linear interpolation. The gradients on vertices are computed as area-weighted averages of the gradients over neighboring triangles, or equivalently, by applying the divergence theorem. The selection $\beta = 1/3$ and $\delta_c = \delta_d = 0$ leads to a third/forth-order methods depending on the upwind parameter $\lambda$ in the expression for fluxes. Keeping $\beta = 1/3$ but taking $\delta_c = -1/30$ and $\delta_d = -2/15$ leads to a fifth/sixth order method, once again, depending on the value of the upwind parameter. On 3D z-coordinate meshes any of or both $u$ and $d$ triangles can be absent for edges touching boundaries or bottom topography. We replace $u$ and $d$ gradients by vertex gradients at $i$ and $j$ in this case. Technically, the fifth/sixth order method adds computations of vertex gradients and increases the amount of computations needed to estimate fluxes. However logistics related to the $z$-coordinate bottom is expensive for the third/fourth order, and even more so for the fifth/sixth order method. The third/fourth order method requires an extended halo for triangles, and the fifth/sixth order method needs additionally an extended halo for vertices including neighbors of neighbors. We note that if we were only interested in the methods of third or fourth order, the implementation of Skamarock and Gassmann [2011], based on computing second-order derivatives, as in (4.3) and (4.4), is the same convenient as operations on gradients. The results are not equivalent, but similar. The coefficients and indices to vertices needed to compute the second-order derivatives can be computed before, but the $z-$coordinate bottom introduces complications as in the other case.

## The role of mass matrix

It is well-known from the FE literature (see, e.g., Donea and Huerta [2003]) that transport schemes based on linear continuous finite elements become the fourth order if used with consistent mass matrices on uniform meshes. They are only second-order if used without them. Since linear continuous FE are analogous to the vertex-based FV discretization it is natural to ask whether similar improvement is possible with FV. We begin with a brief summary of the FE case.

We will use the same notation $T$ for the continuous and discrete representations. In the case of linear continuous finite elements a scalar field is represented as $T = T_j(t)N_j(x, y)$ (summation is implied over the repeating indices), where $j$ is the vertex index, $N_j$ is the linear basis function (equal to one at the location of vertex $j$ and decaying to zero at neighboring vertices) and $T_j(t)$ is the discrete vertex value of tracer field. Substituting this representation in eq. (4.1), then multiplying with $N_i$ and integrating over the entire area occupied by flow, we get the matrix formulation

$$\mathrm{M}_{ij}\partial_t T_j + \mathrm{A}_{ij} T_j = 0, \tag{4.5}$$

where

$$\mathrm{M}_{ij} = \int N_i N_j dS$$

are the components of mass matrix M, and

$$\mathrm{A}_{ij} = -\int \nabla \cdot N_i \mathbf{u} N_j dS$$

are the components of the advection matrix $\mathbf{A}$. The boundary conditions of the zero flux through the lateral boundaries are already taken into account. Since $N_i$ have a finite support, integration is limited to triangles containing vertex $i$. A triangle $c$ containing the vertex $i$ contributes with $S_c/6$, to the mass matrix entry $M_{ii}$ and with $S/12$ to the off-diagonal entries $M_{ij}$ for $j$ that correspond to other vertices in triangle $c$. The row sum of entries in the mass matrix for the row $i$ is therefore the area of the median-dual control volume around the vertex $i$. The mass matrix appearing here provides weighting of the time derivative whereby the time derivative and advection operator are approximated on the same stencil, which reduces dispersion. It can be readily seen that the advection operator $\mathbf{A}$ corresponds to the centered tracer estimate at mid-edge as $T_{ij} = T_{ji} = (T_i + T_j)/2$, where $i$ and $j$ are the indices of edge vertices, in the FV implementation (4.2).

If we consider, for example, a uniform mesh obtained by splitting quads regularly and apply the standard von Neumann analysis to (4.5), writing $T_j = T_0 \exp(-i\omega t + ikx_j + ily_j)$ where $(x_j, y_j)$ are the coordinates of the vertex $j$, $\omega$ is the frequency, and $k, l$ are the wave numbers, the matrix system of equations (4.5) reduces to a single equation

$$(-i\omega \mathrm{M} + \mathrm{A})T_0 = 0,$$

with the matrices replaced by respective spectral symbols M and A. For the mesh formed by splitting quads with side $h$ with SW–NE diagonals, the spectral symbols are $\mathrm{M} = (h^2/6)(3+\cos kh+\cos(k+l)h+\cos lh)$, and $\mathrm{A} = u\mathrm{G}_x + v\mathrm{G}_y$ with $\mathrm{G}_x = (ih/3)(2\sin kh - \sin lh + \sin(k+l)h)$ and $\mathrm{G}_y = (ih/3)(2\sin lh - \sin kh + \sin(k+l)h)$. In the continuous case the phase speed is $c_p = \omega/(k \cdot n) = \mathbf{u} \cdot n$, where $\mathbf{k} = (k, l)$, in any direction $\mathbf{n} = \mathbf{k}/|\mathbf{k}|$. In the discrete equations this result is only recovered for sufficiently small $kh, lh$. A simple way to see the role of mass matrix is to do the Taylor expansion in this limit to find that $\mathrm{G}_x/ik$, $\mathrm{G}_y/il$ and M are $\approx h^2(1 - h^2(k^2 + l^2 + kl)/6 + O(h^4))$ for small wavenumbers. Thus, the inversion of the mass matrix will remove the third-order derivatives present in $\mathrm{G}_x$ and $\mathrm{G}_y$ raising the accuracy to the fourth order! This result holds for other regular meshes (using equilateral or isosceles triangles). For example, on equilateral meshes, the common factor in the mass matrix and $\mathrm{G}_x/ik$ and $\mathrm{G}_y/il$ is approximately $1 - h^2(k^2 + l^2)/6$, with $h$ the height of triangle. It will be lost on general unstructured meshes. However, even in that case the account for the mass matrix effectively removes a part of dispersive errors.

In practice an approximate inversion of mass matrix is performed. To solve $\mathbf{Ma} = \mathbf{b}$, where $\mathbf{a}$ and $\mathbf{b}$ are vectors composed of vertex values, one writes

$$\mathbf{M}_L\mathbf{a}^p = \mathbf{b} + (\mathbf{M}_L - \mathbf{M})\mathbf{a}^{p-1},$$

where $\mathbf{M}_L$ is the diagonal (lumped) approximation to the mass matrix (obtained by summing row entries and placing the result at the diagonal, i.e., placing the areas of median-dual control volumes) and does just two iterations ($p$=1,2) with $\mathbf{a}^0 = \mathbf{M}_L^{-1}\mathbf{b}$. Further iterations will not change the leading-order error term, as explained further.

## A FV analog of the FE method with consistent mass matrix

If $\mathbf{M}$ in (4.5) were replaced with $\mathbf{M}_L$, it would be equivalent to (4.2) for the centered tracer estimate mentioned above. Replacing the time derivative term in (4.2) with $\mathrm{M}_{ij}\partial_t T_j$ and approximately inverting the mass matrix will destroy the conservation in terms of fluxes leaving control volumes, which is not desirable. Instead we propose to correct the tracer to be advected with anti-dispersion, i.e., to advect the discrete tracer field $\tilde{\mathbf{T}}$ such that

$$\mathbf{M}\tilde{\mathbf{T}} = \mathbf{M}_L\mathbf{T},$$

where $\tilde{\mathbf{T}}$ and $\mathbf{T}$ are vectors with components $\tilde{T}_i$ and $T_i$ respectively. In matrix notation the method is

$$\mathbf{M}_L\partial_t\mathbf{T} + \mathbf{A}\tilde{\mathbf{T}} = 0.$$

Since $\tilde{\mathbf{T}} = \mathbf{M}^{-1}\mathbf{M}_L\mathbf{T}$, the method differs from the finite-element one by the order the operators $\mathbf{A}$ and $\mathbf{M}^{-1}$ are applied. These operators commute only if the advective velocity is uniform, so the proposed method is generally different from the FE method. It, however, inherits the main advantage of the consistent mass matrix in FE: $\tilde{\mathbf{T}}$ contains anti-dispersion to the second-order centered advection. The method can be used, for example, with the second or third order Adams–Bashforth (AB2 or AB3) time stepping. To obtain the equivalent of the TG method (i.e., the Lax–Wendroff (LW) method), $\tilde{T}_{ij}$ should be taken at a point offset by $-\mathbf{u}\Delta t/2$ from the center of a boundary segment in flux estimates for each segment. Linear interpolation on triangles is used in such cases.

We do approximate inversion as explained above to obtain

$$\tilde{\mathbf{T}} = \mathbf{T} + (\mathbf{I} - \mathbf{M}_L^{-1}\mathbf{M})\mathbf{T} + (\mathbf{I} - \mathbf{M}_L^{-1}\mathbf{M})^2\mathbf{T},$$

where $\mathbf{I}$ is the identity matrix. Computations include two cycles over the nearest neighbors (and two halo exchanges in parallel implementation, but again involving only the nearest neighbors). Next, we compute fluxes leaving median-dual control volumes using $T_{ij} = (\tilde{T}_i + \tilde{T}_j)/2$ for the value of the tracer advected through the two segments of boundary associated to the edge between the vertices $i$ and $j$ for AB2 and AB3, or using the offset for the LW method.

An elementary 1D analysis below explains why the approximate inversion is sufficient. Further statements are valid on uniform meshes only.

In 1D, the equation connecting $\tilde{T}_i$ and $T_i$ on an equidistant mesh is

$$(1/6)(\tilde{T}_{i-1} + 4\tilde{T}_i + \tilde{T}_{i+1}) = T_i,$$

or, in the Fourier representation,

$$\left(1 - \frac{1 - \cos kh}{3}\right)\tilde{T}_0 = T_0,$$

where $h$ is the cell size. The index 0 is used for Fourier amplitudes and will be suppressed further. The iterative procedure implies that instead of the exact solution $\tilde{T} = T/(1-(1-\cos kh)/3)$ one takes the Taylor expansion of the power leaving two terms in the case of two iterations: $\tilde{T} = (1 + (1 - \cos kh)/3 + (1 - \cos kh)^2/9)T$. In the limit of small wavenumbers it becomes $\tilde{T} \approx T(1 + [(kh)^2/6 - (kh)^4/72] + (kh)^4/36 + O((kh)^6))$. Here the group of terms in square brackets comes from the first iteration.

The spectral symbol of advective operator is written as

$$h^{-1}\mathrm{A}\tilde{T} = (iu/h)\tilde{T}\sin kh.$$

Expanding the sine we obtain $h^{-1}\mathrm{A}\tilde{T} \approx iukT(1-(kh)^4/30)$ if only one iteration is made and $h^{-1}\mathrm{A}\tilde{T} \approx iukT(1-(kh)^4/180)$ if the second iteration is done. One can readily see that the result after one iteration is identical to the fourth-order centered differences (see, e.g., Shchepetkin [2015]), and the second iteration reduces the remaining dispersive error by a factor of 6. Further iterations will not change the leading error, which is the reason why they are generally not needed.

Returning to the 1D equation on $\tilde{T}$ one sees an analogy with the so-called compact fourth-order method (e.g., Shchepetkin [2015], Lemarié et al. [2015]), and the implementation through parabolic splines as in Zerroukat et al. [2006]). While in 1D the approximate inversion with two iterations is not cheaper than direct solution of the three-diagonal system of linear equations, it is the only practically affordable way in the horizontal plane on unstructured meshes.

We call the method proposed above the compact (fourth-order) method (C4) because of this analogy with the 1D method and because we always deal with the nearest vertices. By construction, the method is equivalent to the FE method with a consistent mass matrix if advection velocity is uniform, so it has the same order as the FE method. However, it has the flux form (4.2), and thus is locally and globally conserving. It needs either an FCT limiter or diffusion for stability. However, it can be generalized to include upwind dissipation.

Write

$$T_{ij} = \frac{\tilde{T}_i + \tilde{T}_j}{2} = \frac{T_i + T_j}{2} + \frac{\tilde{T}_i - T_i}{2} + \frac{\tilde{T}_j - T_j}{2}$$

as a sum of the second-order centered estimate and two corrections. If, instead of the sum of corrections, we take twice the first correction if the transport is out of the control volume around vertex $i$ ($\mathbf{u} \cdot n > 0$ for $\mathbf{n}$ directed from vertex $i$) and twice the second one if $\mathbf{u} \cdot n < 0$, we will get a third-order method. Indeed, it is easy to see that these corrections do a similar job as the corrections in (4.3) and (4.4). The final form is

$$T_{ij}\mathbf{u} \cdot n = \frac{\tilde{T}_i + \tilde{T}_j}{2}\mathbf{u} \cdot n + \lambda|\mathbf{u} \cdot n|\frac{\delta T_i - \delta T_j}{2}, \quad (\delta T_i = \tilde{T}_i - T_i),$$

where $0 \leq \lambda \leq 1$ is the parameter to blend the orders, giving the C34 scheme. However, it turns out that already small non-zero $\lambda$ degrades the extra accuracy of C4 toward that of the finite-volume scheme GE34 based on the gradient estimates.

A remark is due on implementation: in FESOM2, the mass matrix is not assembled. The multiplication of the mass matrix with the vector of the tracer vertex values is done in a cycle over triangular prisms of the three-dimensional mesh. This automatically takes into account the difference in the neighborhood occurring because of the bottom topography in deep layers. (The bottom topography is constant on triangles in FESOM2). Despite the computations implied by the approximate inversion of the mass matrix, the algorithm turns out to be noticeably faster than GE34 in realistic 3D applications. First, there is no need to compute gradients on triangles, and second, there is no up- and down-edge logistics.

## 4.3 Results

### Preliminary remarks

Our main intention here is to compare C34 and GE56 schemes to the commonly used GE34 schemes on triangular meshes. Since all these schemes achieve their accuracy on uniform meshes, an important question is also their performance on distorted meshes. This comparison will be done with an idealized 2D test case of the circular shear flow described below. For the compact scheme we also carry out the Stommel gyre test by Hecht et al. [2000].

The even-order schemes ($\lambda = 0$) have only dispersive errors. Simulations with these schemes were possible in the idealized 2D test below except for the higher resolutions, where small-scale noise contaminates solutions of some of them. They need some dissipation in a general case, and FESOM2 relies on the FCT limiting. We therefore include C4 augmented with the standard FCT procedure (Zalesak [1979]), to show that the reduction in accuracy is rather moderate. The simulations showing excessive noise were repeated with small biharmonic diffusion. The blended schemes combining odd and even-order flux estimates (with $\lambda$ depending on applications) can be tuned to be stable without FCT or explicit diffusion. We include the QR scheme to illustrate that it performs very closely to GE. We also add the FE TG and FEM-FCT scheme by Löhner et al. [1987], used in FESOM1.4, which will provide a benchmark we are striving to achieve. The FV and FE implementation of the FCT limiting differs in their low-order part. The FV FCT relies on the first-order upwind whereas the FEM-FCT scheme uses artificial damping, the level of which can be tuned. For FV, antidiffusive fluxes are computed at edges in a single pass of the procedure that computes the low-order solution and the difference in fluxes between the high order and the low order (antidiffusive fluxes). The high-order is either the compact scheme or the GE4 scheme.

As a second part, we compare the performance of GE34, C34, GE56 and

QR34 in a 3D baroclinic channel test case run with FESOM2 and either using $\lambda > 0$ or FCT. There are no obvious criteria of accuracy in this test case and we only illustrate that the level of eddy kinetic energy energy is not very sensitive to the dissipation in the schemes.

Finally, time stepping methods also contribute to the accuracy. We will be using AB2 in 3D simulations with FESOM2, and only briefly illustrate the effect of AB3 or the LW methods in 2D test cases. The AB methods are implemented by interpolating the advected $T$ as $T^{AB2} = (3/2+\epsilon)T^n - (1/2+\epsilon)T^{n-1}$ and $T^{AB3} = (23T^n - 16T^{n-1} + 5T^{n-2})/12$, where $n$ denotes the time step and $\epsilon$ is a small offset needed for stability. We set $\epsilon = 0.01$ in 2D tests and $\epsilon = 0.1$ in 3D simulations. The LW method is used only with C34. It is implemented by computing $T^{LW}$ at a displaced point as mentioned above.

## 2D test case

We deal here with an idealized test case, which is only intended to rank the schemes in terms of their accuracy for smooth perturbations. Different from the atmospheric community, there is no well-agreed suite of test cases in ocean modeling, except, perhaps, for the Stommel-gyre configuration proposed by Hecht et al. [2000] (see also Budgell et al. [2007]). We use it here only to demonstrate the performance of the compact scheme. Because of its boundary layer character, this test in reality requires a variable mesh resolving the western boundary current. As consequence errors depend also on how the resolution varies, the aspect we were willing to avoid. Most of the atmospheric tests (see examples in Miura and Skamarock [2013]) are formulated for global spherical geometry and are of little relevance for ocean models.

The domain is a box with sides $L_x = L_y$ that correspond to 10 degrees at the equator, taken in plane geometry for simplicity. The velocity field is described by the stream function

$$\psi = (2\pi/\tau)(-Rr\cos(\pi r/R)/\pi + (R/\pi)^2\sin(\pi r/R)) \tag{4.6}$$

for $r < R = L_y/2$ and $\psi = R^2/\pi$ otherwise. Here $r$ is counted from the center of domain and $\tau$ sets the period of rotation at $r = R/2$. The stream function corresponds to a circular shear flow with the azimuth velocity $u_\phi = (2\pi r/\tau)\sin(\pi r/R)$ shown in Fig. 4.3a. Triangle-based velocities are obtained by computing derivatives of $\psi$ on triangles. Such discrete velocities have zero discrete divergence, i.e., (4.2) is exactly satisfied for $T = 1$.

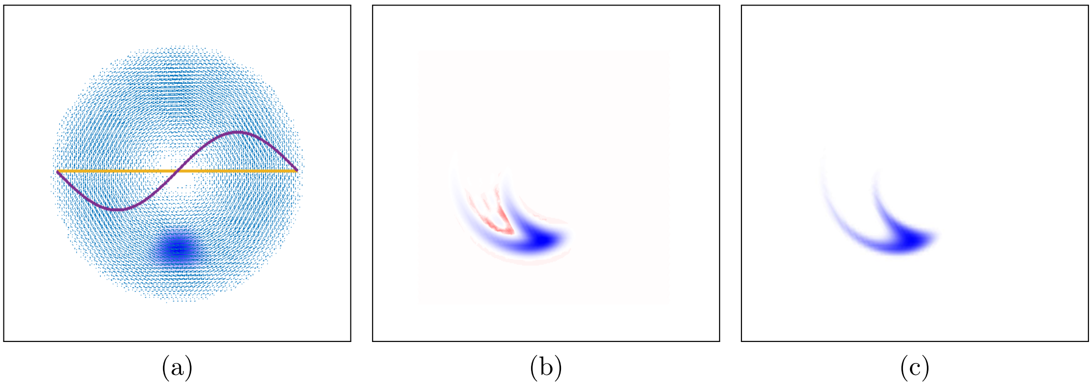
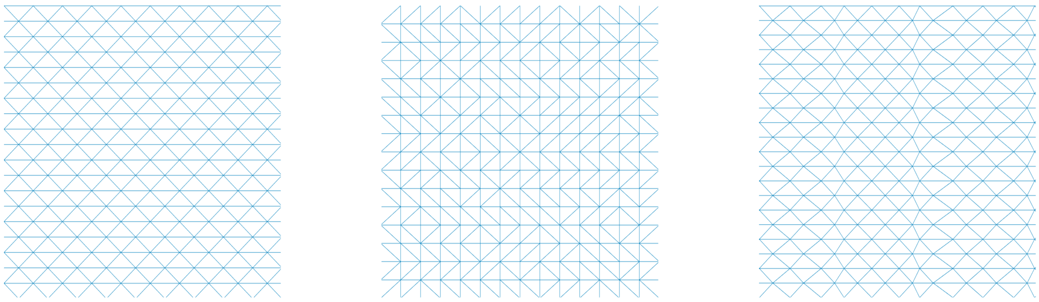


Figure 4.3: The geometry of 2D experiments: a) the initial tracer distribution and the shear velocity field with superimposed schematic profile; b) the final tracer distribution after one full rotation on a 1/6 degree mesh simulated with the GE56 method; c) the exact solution.

The initial perturbation is taken in a separable form $T_0 = 0.25(1+\cos(4\pi(r/R-0.5)))(1+\cos(6(\phi+\pi/2)))$ which facilitates computations of the analytical solution. Here $\phi$ is the angular coordinate with respect to the center of the domain, and perturbation is absent if $|r/R-0.5| > 0.25$ or $|\phi+\pi/2| > \pi/6$. The analytical solution $T^a(r,\phi,t)$ is given by $T_0(r,\phi-\omega t)$, where $\omega = \omega(r) = (2\pi/\tau)\sin(\pi r/R)$. The integration is carried out for 30 days, which also is the value of $\tau$. The perturbation makes one full rotation, but is sheared in the course of rotation, as shown in Fig. 4.3b, c. The tracer flux trough lateral boundaries is zero because of the velocity choice.

The deviations between the simulated $T$ and analytical $T_i^a$ discrete distributions in the idealized test case are quantified in $L_2$ and $L_1$ norms. The $L_2$ norm is computed in a finite-element sense (assuming linear interpolation on triangles) as $L_2 = (\sum_c \int (T-T^a)^2 dS_c / \sum_c S_c)^{1/2}$, and $L_1 = \sum_c \int |T-T^a| dS_c / \sum_c S_c$ with summation over cells (triangles). Both norms demonstrate very similar behavior, so only $L_2$ is discussed further. Note that the discrete representation of velocity and initial tracer distribution may contribute to such errors as well as the finite-element sense of computations. We ignore these details here.

We use four meshes for the main set of tests. The first one is made of equilateral triangles and is referred to as ET. The second one is obtained by the division of quads in a random (irregular) way (IT). It is characterized by irregular neighborhood (from 4 to 8 neighbors). The area of its control volumes varies by a factor of two, so it is anticipated to lead to increased grid-scale noise. The third one is derived from the ET mesh by smoothly distorting it in the zonal direction (DT). The fourth one is unstructured triangular mesh (UT) where the resolution is varied by a factor of 3 in the meridional direction, with the resolution at the center coinciding with that of other meshes. The coarser half of UT mesh is anticipated to lead to increased errors. The IT mesh is also anticipated to show reduced accuracy because areas of its median-dual control volumes may differ by a factor of two, triggering grid-scale noise. In contrast, triangles vary smoothly on the DT mesh and only in one direction, so errors should be close to ET. Mesh fragments are shown in Fig. 4.4 and Fig. 4.5 where UT mesh is shown in more details. Note also that the discrete velocity field is less smooth on UT meshes due to the varying resolution.



(a) ET

(b) IT

(c) DT

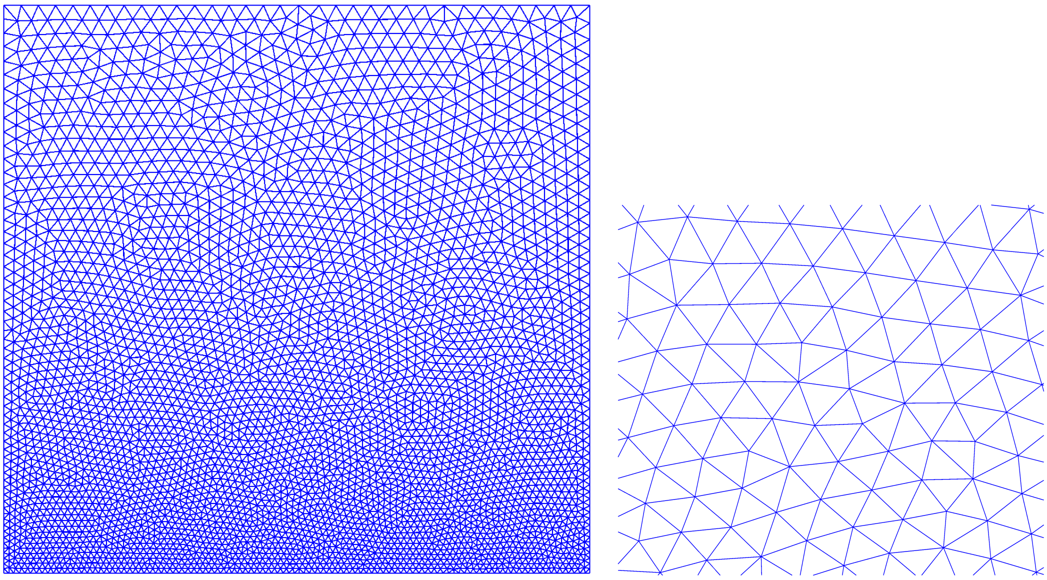
Figure 4.4: Fragments of meshes : a) equilateral mesh (ET), b) irregular mesh (IT), c) horizontally distorted mesh (DT).

We carried simulations at meshes with the triangle side of 1/6, 1/12 and 1/24 degree. The time step $\Delta t$ is 900 s for the resolution of 1/12 degree and is adjusted proportionally to the resolution considered. For the UT mesh only simulations on meshes with nominal resolution of 1/6 and 1/12 degree have been performed.

For convenience, we summarize abbreviations used to denote the schemes:

- Schemes using gradient estimate GExx;
- Compact schemes Cxx;
- Scheme with FCT limiting C4FCT and GE4FCT;
- QR34 scheme based on quadratic polynomial reconstruction.

Here xx stands for the scheme order or combination of orders.



(a) UT mesh

(b) Fragment of UT mesh

Figure 4.5: Unstructured mesh: a) Full mesh with central resolution of 1/6 degree, b) its fragment.

## 2D circular flow

### Sensitivity to time stepping

Errors of advection schemes depend on both temporal and spatial discretizations, and a question arises on their relative contributions. To assess the errors due to the temporal discretization, we run the 2D circular flow test with various time steps $\Delta t$ and AB2, AB3, and LW as time stepping methods. The compact scheme on a 1/6 degree ET mesh was taken for this assessment. The resulting errors are summarized in the Table 4.1. As can be seen, the errors vary only slightly with a time step and a time stepping method. We conclude that the errors are dominated by the contribution from spatial discretization. This is not surprising because the Courant number $C = |\mathbf{u}|\Delta t/h$, where $h$ is the side of the triangle, is less than 0.17 for the largest $\Delta t$. Simulations below are carried out with the time step $\Delta t$ of 1800, 900 and 450 s for the resolutions of 1/6, 1/12 and 1/24 degree, respectively, so the Courant number ($< 0.085$) is the same. The time step selected here is about the time step used in large-scale circulation models with a similar resolution (see, e.g., Lemarié et al. [2015] for limiting factors). Only AB2 and AB3 methods are used further to illustrate that errors remain insensitive to the method when a resolution is varied. In these methods, the tracer field is first AB extrapolated, and then the spatial scheme is applied.

Table 4.1: The errors of AB2, AB3 and LW time stepping for time steps 450, 900, 1800 and 3600 s on the ET mesh with the resolution of 1/6 degree.

| Method | $\Delta t(s)$ | $L_2$ error $(\cdot 10^{-2})$ |
|---|---|---|
| C34 + AB2 | 450 | 1.00 |
| | 900 | 1.00 |
| | 1800 | 0.99 |
| | 3600 | 0.96 |
| C34 + AB3 | 450 | 1.01 |
| | 900 | 1.01 |
| | 1800 | 1.01 |
| | 3600 | 1.00 |
| C34 + LW | 450 | 0.98 |
| | 900 | 0.96 |
| | 1800 | 0.93 |
| | 3600 | 0.91 |

Accuracy of schemes

Figure 4.6 shows the errors of the tested advection schemes after one complete rotation. They are also summarized in the Table 4.2. The schemes were run with $\lambda = 0$ and $\lambda = 0.25$ (values in the parentheses in Table 4.2). Since the velocity field varies in space, the measured accuracy is expected to be the second-order. (The higher order will be seen only for a uniform velocity field, which is of no interest for the ocean and atmosphere where the velocity varies on the same scale as temperature and salinity.) As follows from Fig. 4.6, the convergence indeed stays close to the second order, yet it is systematically lower on IT and UT meshes for some methods. The size of perturbation chosen here is about 150 km, which corresponds to typical midlatitude eddies in the ocean. The resolution of 1/6 degree is already sufficient to represent them but is, perhaps, on the lower end of the resolutions needed. Some high order methods run with $\lambda = 0$ demonstrated unexpectedly high errors, especially on IT and DT meshes with the resolution 1/24 for GE6 and GE4 methods. This happened due to noise reflection from the boundaries where no sponge zone was provided, leading to noise accumulation with time. A small biharmonic diffusion (with the diffusion coefficient $\kappa_i = -\kappa_0 S_i$, where $\kappa_0 = 16, 4, 1$ m$^2$/s for the resolution of 1/6, 1/12 and 1/24 degree, and $S_i$ the control volume area) was added for stability in such cases. It was not needed for ET meshes or in runs with nonzero $\lambda$. We also checked that adding biharmonic diffusion on ET meshes leaves the $L_2$ errors nearly the same.

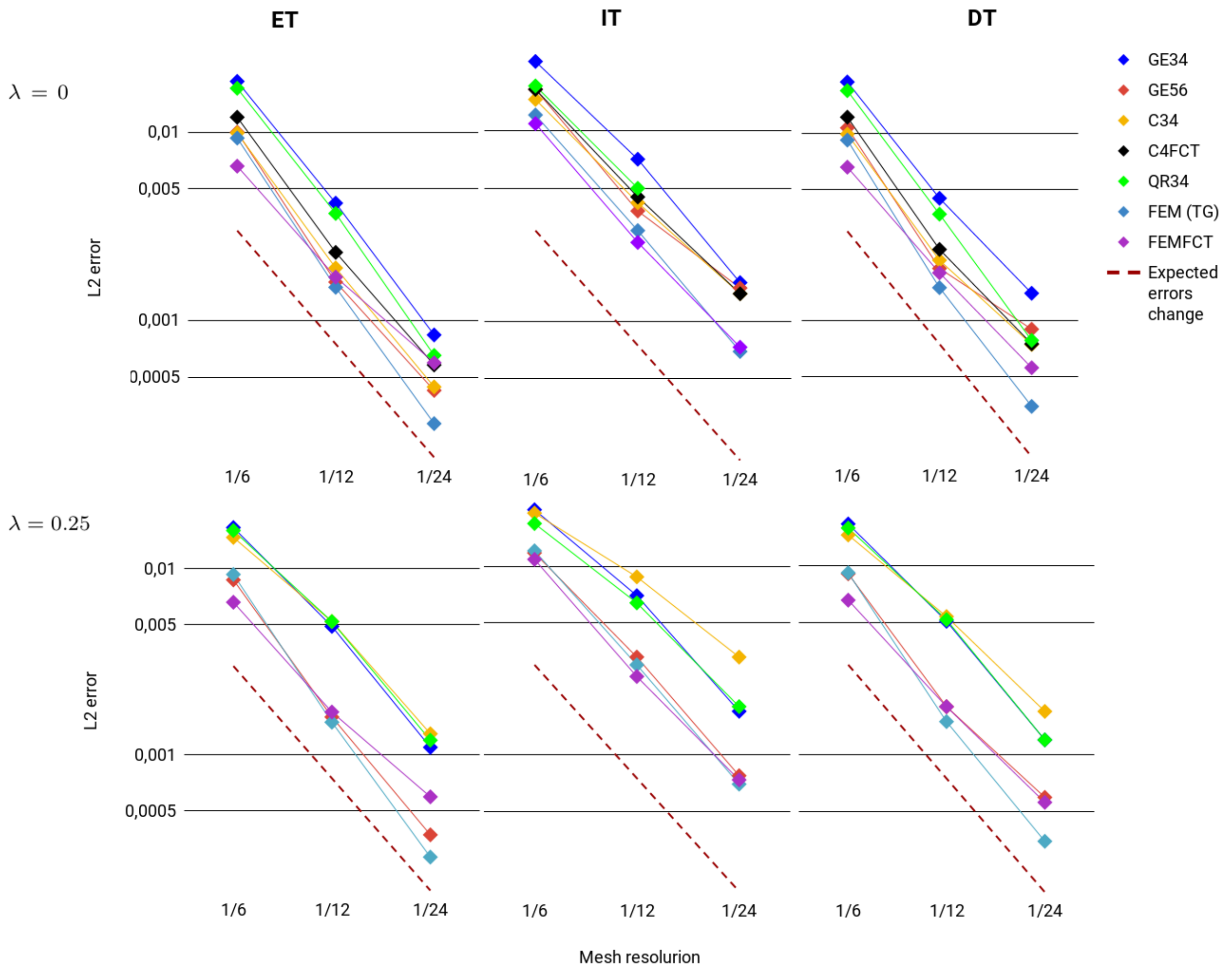


Figure 4.6: The errors on the ET (left column), IT (middle column) and DT (right column) meshes as a function of resolution for $\lambda = 0$ (above) and $\lambda = 0.25$ (below). The dashed line represents the $L_2$ error change for a second-order method. The axes are logarithmic.

Table 4.2: $L_2$ errors in the 2D test case. Tests were run with $\lambda = 0$ and $\lambda = 0.25$ (in parentheses). Runs on the UT mesh were only carried out for two lower resolutions. Runs with $\lambda = 0$ showing tendency to noise were stabilized with small biharmonic diffusion. The last column shows the convergence of errors which are calculated taking into account errors of the results on 1/6 and 1/12 resolutions: $\log_2 \frac{error_{1/6}}{error_{1/12}}$.

| Method | Mesh type | $\Delta t$: 1800, resolution: 1/6, $L_2 \cdot 10^{-2}$ | $\Delta t$: 900, resolution: 1/12, $L_2 \cdot 10^{-3}$ | $\Delta t$: 450, resolution: 1/24, $L_2 \cdot 10^{-4}$ | Convergence |
|---|---|---|---|---|---|
| GE34 + AB2 | ET | 1.86 (1.66) | 4.2 (4.9) | 8.38 (11) | 2.15 (1.76) |
| | IT | 2.31 (2.00) | 7.1 (7.0) | 16.00 (17) | 1.70 (1.51) |
| | DT | 1.88 (1.67) | 4.5 (5.1) | 14.00 (12) | 2.06 (1.71) |
| | UT | 2.52 (2.27) | 7.1 (7.1) | – | 1.83 (1.56) |
| GE34 + AB3 | ET | 1.89 | 4.3 | 8.38 | 2.14 |
| | IT | 2.33 | 7.1 | 16.00 | 1.71 |
| | DT | 1.92 | 4.6 | 14.00 | 2.06 |
| GE56 + AB3 | ET | 1.00 (0.87) | 1.6 (1.6) | 4.25 (3.73) | 2.64 (2.44) |
| | IT | 1.65 (1.18) | 3.8 (3.3) | 15.00 (7.73) | 2.12 (1.84) |
| | DT | 1.07 (0.91) | 1.9 (1.8) | 9.00 (5.96) | 2.35 (2.34) |
| | UT | 1.41 (1.31) | 3.1 (3.1) | – | 2.19 (2.08) |
| C34 + AB2 | ET | 0.99 (1.47) | 1.9 (5.2) | 4.44 (13) | 2.38 (1.50) |
| | IT | 1.46 (1.92) | 4.2 (8.8) | 13.00 (33) | 1.8 (1.13) |
| | DT | 0.99 (1.46) | 2.1 (5.4) | 7.46 (17) | 2.24 (1.43) |
| | UT | 1.47 (1.88) | 3.2 (7.2) | – | 2.20 (1.38) |
| C34 + AB3 | ET | 1.01 | 1.9 | 3.96 | 2.41 |
| | IT | 1.46 | 4.2 | 13.00 | 1.8 |
| | DT | 1.00 | 2.0 | 7.25 | 2.32 |
| C4FCT + AB2 | ET | 1.20 | 2.3 | 5.81 | 2.38 |
| | IT | 1.65 | 4.5 | 14.00 | 1.87 |
| | DT | 1.22 | 2.4 | 7.51 | 2.35 |
| | UT | 1.83 | 4.0 | – | 2.19 |
| C4FCT + AB3 | ET | 1.28 | 2.7 | 6.61 | 2.25 |
| | IT | 1.73 | 4.7 | 14.00 | 1.88 |
| | DT | 1.27 | 2.8 | 7.75 | 2.18 |
| QR34 | ET | 1.71 (1.60) | 3.7 (2.3) | 6.52 (12) | 2.21 (1.62) |
| | IT | 1.72 (1.69) | 5.0 (6.4) | 13.00 (18) | 1.78 (1.40) |
| | DT | 1.69 (1.59) | 3.7 (5.2) | 7.84 (12) | 2.19 (1.61) |
| | UT | 2.18 (2.05) | 5.5 (7.1) | – | 1.96 (1.53) |
| FEM (TG) | ET | 0.93 | 1.5 | 2.83 | 2.63 |
| | IT | 1.21 | 3.0 | 6.97 | 2.01 |
| | DT | 0.92 | 1.5 | 3.49 | 2.62 |
| | UT | 1.46 | 2.8 | – | 2.38 |
| FEMFCT (TG) | ET | 0.66 | 1.7 | 5.96 | 1.96 |
| | IT | 1.09 | 2.6 | 7.34 | 2.07 |
| | DT | 0.66 | 1.8 | 5.60 | 1.87 |
| | UT | 1.19 | 2.9 | – | 2.04 |

Indeed, we see that with $\lambda = 0.25$ errors of GE34 and GE56 are smaller than for GE4 and GE6. This happens because the dissipation caused by the lower order method filters the small-scale noise before it reaches the boundaries. However, for the C34 method, the $L_2$ errors are higher for $\lambda = 0.25$ than for $\lambda = 0$, and the scheme becomes even worse than the GE34 even though it was essentially more accurate for $\lambda = 0$. To see the dependence of $L_2$ errors on $\lambda$, we carried out experiments with different values of $\lambda$ on the ET mesh with the resolution of 1/6 degree. As can be seen from Fig. 4.7, the lowest errors for the C34 method are observed at $\lambda = 0.05$ (0.0097) which is close to the 4th order of these methods. Then with the growth of dissipation, the errors also increase. Meanwhile, for the method GE56 the lowest error is at $\lambda = 0.25$ (0.0087). Moreover, the difference between the errors of GE56 for the 5th order and the 6th order is low: 0.0100 for GE6 and 0.0112 for GE5. For GE56 method dissipation does not have a big influence on the result. Also, GE56 demonstrates the lowest errors among these methods. The errors change smoother for GE34 than for C34. Similar behavior was also observed for IT and DT meshes. We can conclude that the compact methods are more sensitive to dissipation than GE56 and that they show optimal results in a combination with the FCT limiting instead of high-order upwind dissipation.

Another observation is that the QR34 method does not behave as expected. By construction of the method, we supposed that its error will be less sensitive to the type of the mesh than for other methods. However, for the IT mesh with higher resolutions, the error significantly exceeds errors that we obtain on DT and ET meshes. Although the difference between errors with the parameter $\lambda = 0.25$ reduces for the resolution 1/24, it still stays essential. The situation changes with the change of parameter $\lambda$ and the best results were obtained with $\lambda = 0.35$ which are represented in the Table 4.3. Now we can observe a smaller difference in the errors, and the behavior becomes close to GE34 with $\lambda = 0.25$. The reason for the loss of accuracy on the IT meshes can be the noise alluded to above and the presence of vertices that have only four triangle neighbors, in which case quadratic reconstruction is replaced by a linear one, and grid structure may be imprinted in solutions through this.

Figure 4.8 demonstrates noise which occurs when we use C34 schemes with different upwind parameter $\lambda$. Even though the amplitude of the error is less for C4 than for C3, we can see that C4 produces a lot of small-scale noise which completely disappears in the case of C3. Even small dissipation with $\lambda = 0.05$ significantly improves the situation. The scheme with FCT also demonstrates

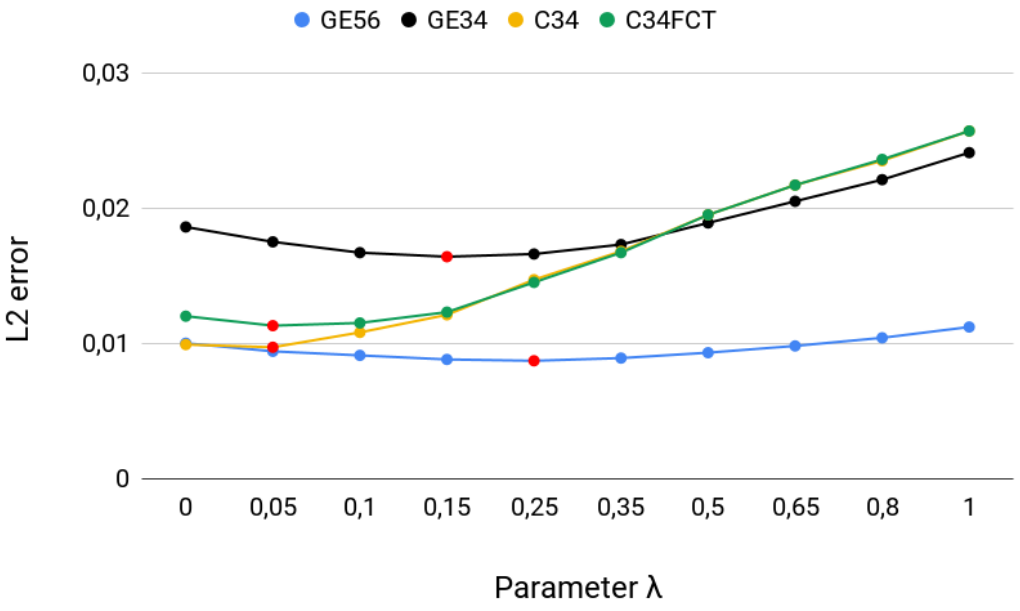


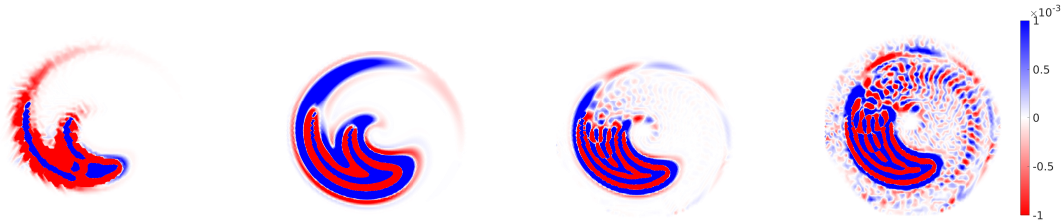
Figure 4.7: $L_2$ errors as a function of the parameter $\lambda$ for GE34, GE56, and C34 (with FCT and without). The experiments were held on the ET mesh with a resolution of 1/6 degree. The red dots highlight the minimum errors for every method.

Table 4.3: $L_2$ errors of QR34 method with parameter $\lambda = 0.35$.

| Mesh type | 1/6<br>$L_2 \cdot 10^{-2}$ | 1/12<br>$L_2 \cdot 10^{-3}$ | 1/24<br>$L_2 \cdot 10^{-3}$ |
|---|---|---|---|
| ET | 1.73 | 6.1 | 1.5 |
| IT | 1.89 | 7.5 | 2.2 |
| DT | 1.72 | 6.1 | 1.5 |

less small-scale noise. In addition, we can see how FCT cuts the maximums.

The FE TG and FEM-FCT schemes outperform all other FV schemes explored here in almost all cases. This is partly due to low errors of the TG scheme with a consistent mass matrix (used as high-order scheme) and partly due to tuning of the FCT limiter in our implementation of FEM-FCT (we adjust the admissible bounds with account for the solution from the previous time step and use a relatively low value of 0.25 for the parameter controlling the dissipation in the low-order solution, see Löhner et al. [1987], making it not necessarily monotone). The error of FEM-FCT increases nearly two-fold for the lowest resolution if information about fields on the previous time step is omitted in computations of admissible bounds, becoming worse than TG, but the TG result is recovered



(a) C4FCT (b) C3 (c) C34 $\lambda = 0.05$ (d) C4

Figure 4.8: Difference between the analytical and numerical solutions for the compact schemes a) C4FCT, b) C3, c) C34 with $\lambda = 0.05$, and d) C4. The colorbar limits are selected to visualize the noise.

for higher resolutions. The proposed C4 scheme performs worse than TG despite it being designed using analogy to TG. The errors of C4FCT are always higher than in the absence of FCT, whereas for FEMFCT we can see even partial improvement at the coarsest resolution of $1/6$ degree. It should be reminded that neither TG nor FEM-FCT can be directly transferred to a FV code because of mass matrices.

The GE56 in this test case shows an approximately two-fold error reduction with respect to its lower-order counterpart GE34. With upwind dissipation added, it outperforms C4 and approaches the accuracy of FEM TG.

The errors simulated on the IT and UT meshes are the largest, and the order of convergence is deteriorated, being generally lower than two. The reason is the excitation of grid-scale perturbations as a tracer is advected through the irregular mesh (IT) or additionally through a coarse mesh part (UT). We note that the loss of accuracy on the IT mesh is partly due to the larger size of triangles than on the ET mesh. If the side of the triangle is $a$, the triangle area on mesh ET is $\sqrt{3}a^2/4$ compared to $a^2/2$. Assuming the second-order scaling we should expect that errors on ET meshes are by a factor $\sqrt{3}/2 \approx 0.87$ smaller for the same triangle side. This factor explains most of the error difference for the coarsest resolution but fails to do so for finer resolutions.

## The Stommel gyre test

Since the performance of many unstructured-mesh schemes in the test case by Hecht et al. [2000] was documented by Budgell et al. [2007], we carried it out with the new compact scheme proposed here. The test and its parameters are described in Budgell et al. [2007] and are not repeated here except for mentioning

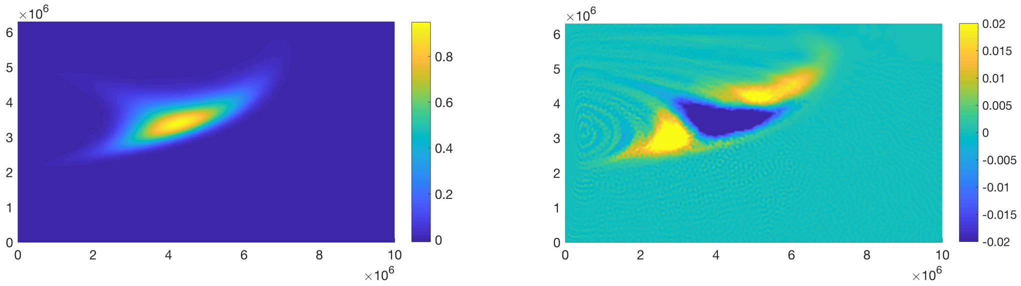


Figure 4.9: The final state (left) and error distribution in the Stommel gyre test for C34 with $\lambda = 0$. The colorbar in the right panel is saturated to visualize small-scale errors behind the tracer and south to it.

that the basin size is 10 by 6.3 thousand kilometers, and the western boundary current is concentrated within less than 100 km. The mesh resolution (the side of triangle) in our run follows the scaling $h = h_0(5 + 4\tanh{(x - x^*)/b^*})$, where $h_0 = 10$ km, $x^* = 300$ km and $b^* = 150$ km, $x$ is counted from the western basin wall. It varies from approximately 10 km to 90 km at the eastern wall, leading to a mesh with 20847 vertices. Although 10 km is coarser than the finest resolution used in Budgell et al. [2007], our mesh is twice larger, with generally better resolved western part of the basin. We used $\lambda = 0$ but added a small harmonic diffusion of 10 m$^2$/s during the first three years of simulations when the tracer is passing through the boundary layer. Such diffusion can only lead to a diffusive spreading of about 30 km, which is negligible compared to the size of initial and final tracer distribution. It is however necessary to eliminate small-scale wavy perturbations appearing because of leading dispersive error of the compact scheme. The presence of small elements and relatively high velocities in the western boundary current limit the admissible time step to 1800 s. Figure 4.9 shows the simulated final state after $1.5\times10^8$ s (left) and error distribution (right) with respect to the semianalytical solution obtained by numerical integration along streamlines. In the absence of diffusion, the errors behind the tracer spot would be of smaller scale and higher amplitude, generally increasing the error norms. The $l_2, l_1$ and $l_\infty$ errors (defined same as in Hecht et al. [2000] and Budgell et al. [2007]) are 0.05, 0.07 and 0.055 respectively. The errors at maximum and minimum are -0.05 and -0.009 respectively. These errors are smaller than the errors reported in Budgell et al. [2007] except (unsurprisingly) for VELA and RKDG3 schemes, yet these schemes cannot be adapted to vertex-based finite volume discretization. The TG scheme explored in Budgell et al. [2007] is the closest analog to the compact scheme. Its errors in Budgell et al. [2007] are noticeably larger than for the compact scheme and we suppose that this is the consequence of small-scale noise contaminating the TG solution in that study and, possibly, of the generally lower resolution of the western part of the basin.

### 3D simulations

To illustrate the performance of C34 and GE56 we simulate a baroclinically unstable flow and measure the level of eddy kinetic energy and the variance of eddy vertical velocity. One expects that a scheme with less dissipation and better accuracy will introduce less damping and will lead to higher eddy kinetic energy because of more effective conversions from the available potential energy. A strongly turbulent flow is characterized by a cascade of scalar variance to the grid scale, which is the reason it cannot be run without dissipation.

The compact scheme and GE56 schemes have been implemented in FESOM2 and tested by simulating an idealized baroclinically unstable flow in a zonally reentrant channel described in Soufflet et al. [2016]. The channel is 500 by 2000 km in zonal and meridional directions respectively, and 4 km deep. We used an equilateral mesh with a triangle side of 10 km, and 41 unevenly spaced vertical levels.

Turbulence is triggered with a small perturbation added to the initial buoyancy distribution. To maintain the instability, the zonal mean buoyancy and zonal velocity are relaxed to climatology as described in Soufflet et al. [2016], which keeps isopycnals inclined in the presence of eddies. The applied forcing would drive a surface-intensified zonal jet flow confined to the central part of the channel with maximum velocity about 0.2 m/s. The horizontal viscosity is biharmonic, with a flow-dependent viscosity coefficient. We are willing to explore the effect of horizontal scalar advection. Therefore vertical tracer and momentum advection are computed with the (non-dissipative) 4th-order centered scheme GE4 same for all the cases. The horizontal momentum advection uses a 2nd-order centered flux scheme based on scalar control volumes (see Danilov et al. [2017]). The integration length is 5 years. The first year is discarded, and the rest is used to compute mean eddy kinetic energy and vertical velocity variance. The boundary conditions are of zero flux at all the boundaries. In FESOM2 the advection is implemented without operator split, by combining the horizontal and vertical fluxes through the faces of control volumes. The vertical and horizontal fluxes can rely on different methods. If FCT is used, 3D admissible bounds are also sought without a directional split.

We present simulations with GE34, GE4FCT, and similar simulations with

Table 4.4: Mean values of energy and velocities in the test case for 3D.

| Method | Mean values of EKE ($m^2/s^2$) ($\cdot 10^{-3}$) | RMS of vertical velocities ($\cdot 10^{-8}$) |
|---|---|---|
| C34 | 1.74 | 1.81 |
| C4FCT | 2.01 | 1.81 |
| GE34 | 1.84 | 1.82 |
| GE4FCT | 1.64 | 1.81 |
| GE56 | 1.83 | 1.80 |
| QR34 | 1.76 | 1.79 |

C34, C4FCT, QR34. GE34, C34 and QR34 were run with $\lambda = 0.25$. Concerning GE56, higher $\lambda$ ($\geq 0.5$) is needed in order to control the noise in vertical velocity. Adding FCT allows the schemes to be run with $\lambda = 0$. We also tried the compact fourth-order scheme for vertical tracer advection (see, e.g., Soufflet et al. [2016]), but found that it leads to noise in vertical velocity if run in a combination with C34 unless the limiter of the Piecewise Parabolic Method (Colella and Woodward [1984]) is applied. It works stably in other situations, but there is no apparent indication that it is better than the 4th-order centered scheme.

The mean values of eddy kinetic energy (EKE) of all the schemes are very close to each other which can be clearly seen from Table 4.4. Even though we can see some differences, they are smaller than the EKE variability shown in Fig. 4.10. The variance of vertical velocity in all cases is nearly the same, indicating that the flow stays in approximately the same dynamical regime. We conclude that, concerning the effect of explored advection schemes on eddy dynamics, it is relatively weak. However, it can be observed that the mean value of EKE for C4FCT is the highest among other methods. GE34 and GE56 also demonstrate relatively high results and outperform other schemes, including C34. The latter can be linked to a higher sensitivity of C34 to the upwind parameter $\lambda$. The very similar behavior of the simulations with GE34 and GE56 is at variance with noticeably higher accuracy (smaller dissipation) of GE56. EKE for GE4FCT is the lowest. We can observe that, for example, peaks in the results of simulations using C4FCT are higher than for GE4FCT. However, a small difference in mean values of EKE together with the fluctuations demonstrates that in our case the choice of an advection scheme does not impact much on the simulated energy content. Given this behavior, the choice of the optimal scheme is largely defined by its computational efficiency, and less so by its accuracy.

Figure 4.11 shows the time it takes to simulate two tracers with different

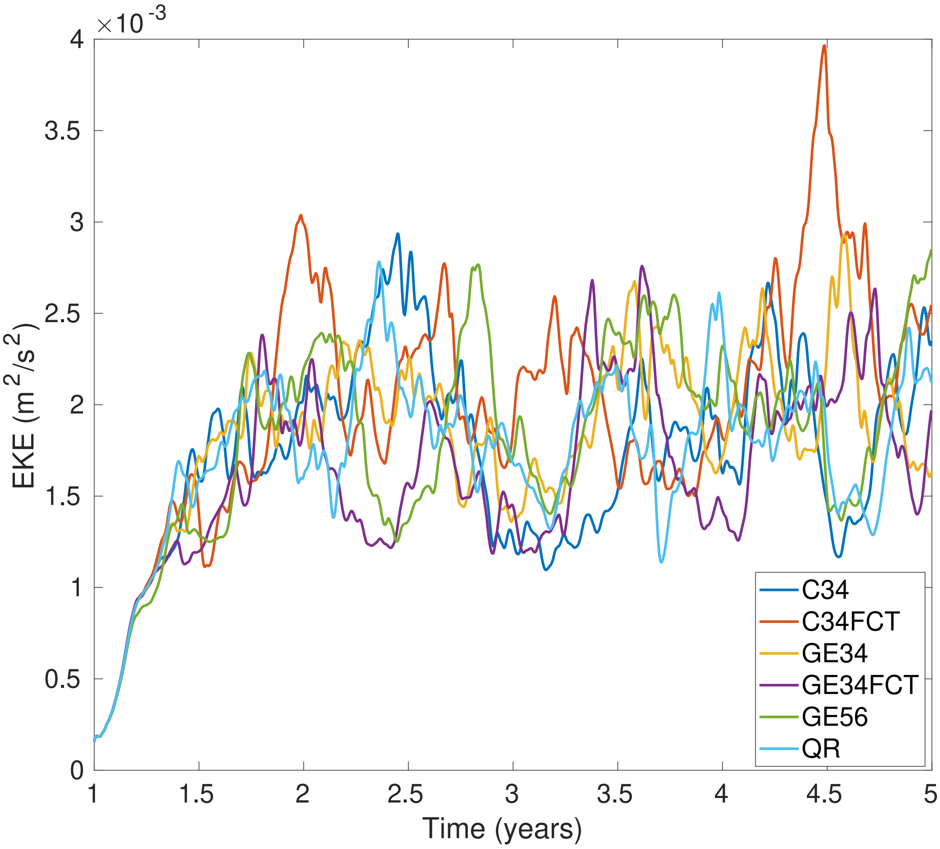


Figure 4.10: Basin-mean EKE in simulations with different advective schemes.

schemes in FESOM2 per simulated day on 36 cores of CRAY CS400 in the setup used here. Although the EKE level simulated with GE56 is one of the highest (the second after the C4FCT scheme), this scheme is the slowest. We admit that its implementation is suboptimal as concerns exchanges over the extended halo, and there is a potential for improvement. However, it will remain more expensive than GE34 which uses a smaller stencil. The C34 scheme is the fastest one. The use of FCT is expensive, and for the compact scheme, costs much more than the scheme proper. Surprisingly, QR34 shows the fastest results after C34. We note that such numbers should be taken with caution because they depend on machine, parallelization, and compiler options. Furthermore, they also depend on implementation detail and may change if, for example, a better algorithm will be proposing for GE schemes in the vicinity of topography.

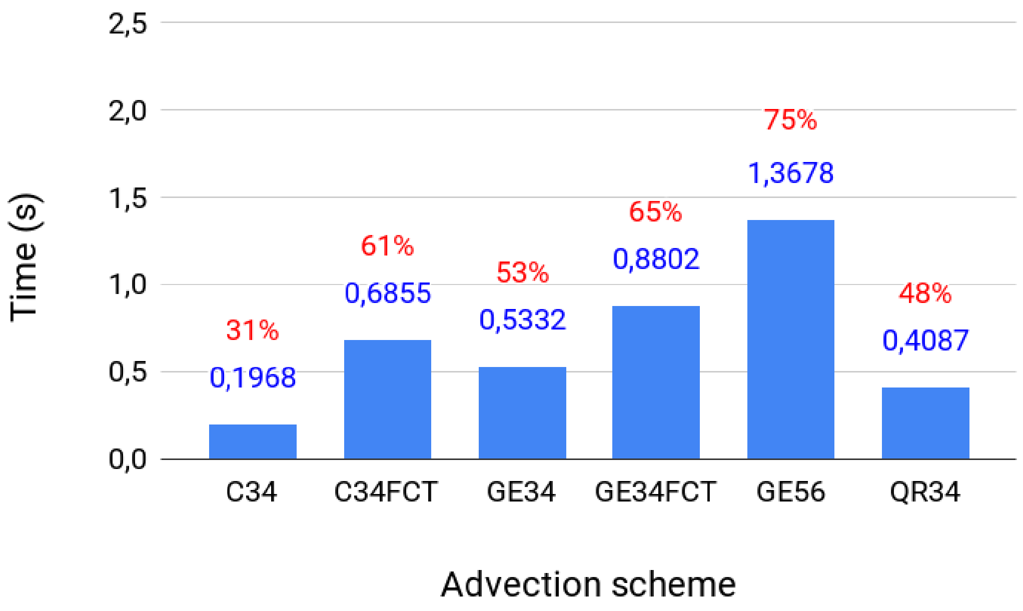


Figure 4.11: Time needed to perform advection, per one simulated day, in full 3D simulations. Its fraction in the total simulations time is shown in percents above.

## 4.4 Discussions

The initial motivation of this study was the observation that vertex-based FV advection methods like GE34 used in FESOM2 are less accurate compared to the FE TG and FEM FCT methods used in FESOM1.4. The compact method designed by analogy to the TG method with a consistent mass matrix is more accurate than the standard GE34 method, but still demonstrates larger errors than TG despite its conceptual closeness. The extension of GE34 to the higher order, GE56, approaches the TG method in terms of accuracy, but at a relatively high computational cost. The advantage of the compact scheme is its computational efficiency, and disadvantage is the sensitivity of its errors to the presence of upwind fluxes, meaning that it loses its extra accuracy relative to GE34 for rather small $\lambda$ (see Fig. 4.7). This implies that using the upwind version of the compact scheme can only be recommended if small $\lambda$ is sufficient to control noise. We guess that the stronger loss of accuracy, in this case, is related to a less directed stencil than in the case of GE methods. A natural question is whether the upwind flux computation proposed here can be improved; it is left for the future. It also remains to be seen whether the cost of GE56 can be reduced to approach that of GE34. In our present implementation, the expensive part is the processing of different situations on the $z$-coordinate bottom where a part of the numerical stencil can be absent. It is more expensive than for GE34 where it already takes almost half of the computational cost.

From Table 4.2 we can observe that for the methods GE34, C34, and QR34 the order of convergence deteriorates notably for $\lambda = 0.25$. Also, convergence order for these methods is lower on the IT and UT meshes than on the other meshes. As expected it is the lowest on the IT meshes where areas of scalar control volumes may vary twofold for two neighbor vertices, and the reduction in the convergence order is weaker on UT meshes because the areas of control volumes vary smoothly in this case. In contrast, the convergence order on DT meshes characterized by smooth distortions of triangles is nearly the same as on ET meshes which are uniform. This demonstrates that a mesh structure is an important factor influencing accuracy. Note that the FE methods show a weaker sensitivity to mesh quality.

In practice, FESOM2 uses FCT limiting in most cases as a mean to warrant stability of integration on general meshes under realistic forcing. In this case, using C4 as a high-order scheme instead of GE4 should lead to improved accuracy and reduced computational cost. However, FCT is computationally expensive, and a question is whether it can be avoided by adding upwind dissipation in realistic simulations. The channel simulations here are stable with $\lambda$ as low as 0.15, but higher dissipation can be needed in realistic ocean simulations. The optimal practical strategy is a question for future research.

The relatively efficient performance of the QR34 indicates that extending it to a quadratic-reconstruction WENO scheme might be a good idea for future work too. A quadratic-reconstruction WENO scheme is already available in SCHISM (Ye et al. [2019]), showing a much more improved behavior compared to other versions of advection available in SCHISM. The Stommel gyre test, applied in several publications (e.g. Budgell et al. [2007]), has been used here only for the compact scheme to show that its errors are only worse than errors of schemes based on discontinuous Galerkin or Lagrangian methods explored in Budgell et al. [2007]. The western boundary layer present in this test needs to be resolved on the mesh used, which requires variable-resolution meshes and leads to results that depend on how precisely the resolution is selected. This is why a simpler configuration has been selected to explore accuracy.

The real motivation for using high-order advection schemes in large-scale ocean models is the hope that their higher accuracy leads to reduced dissipation which also implies reduced spurious diapycnal mixing concerning their less accurate counterparts. However, a scheme has to be equipped with sufficient dissipation, and answering how spurious diapycnal mixing is affected is far from being straightforward, especially on highly variable meshes. This aspect is not touched in the present study and presents an obvious direction for future studies.

## 4.5 Conclusions

I describe the compact scheme for the horizontal advection that is based on an approximate mass-matrix inversion and show that it is performing similarly to the already known (GE34, QR34) schemes and demonstrates better accuracy in the limit of small upwind fluxes. I also test the performance of the high-order GE-based method and demonstrate that augmenting it to GE56 leads to a substantial reduction in the simulated errors compared to GE34. The improved accuracy has only a small impact on the simulated EKE levels in the 3D test runs here. The compact scheme is more numerically efficient that GE and is recommended instead of GE schemes.

# Chapter 5

# RPE analysis

In this section, I describe the experiments which were performed with FESOM2 to investigate the extent to which spurious diapycnal mixing can be reduced. All the experiments were simulating 3D baroclinically unstable flows in channels of different sizes on meshes of different resolutions.

Advection schemes in ocean models use either explicit or implicit horizontal diffusion to prevent the accumulation of tracer variance at grid scales. Diffusion is introduced implicitly in schemes using upwind fluxes, including the schemes with high-order upwind fluxes. It is also created by limiters such as FCT (flux corrected transport). Because $z$-vertical levels cross isopycnal surfaces, any horizontal diffusion will be creating spurious dianeutral mixing in numerical simulations, seen as spurious water mass transformations. The RPE method (see chapter 2) allows one to classify the schemes used concerning spurious dissipation they create, and this is done here for the first time for the schemes used in FESOM2.

Ocean models routinely use isoneutral diffusion parameterizing mixing due to unresolved eddies. This diffusion is implemented in the codes with the help of rotated diffusivities described in chapter 3. Since it will generally create quasi-horizontal mixing of scalar quantities such as temperature and salinity, the idea is to study whether it can be used to stabilize advection schemes with centered flux estimates which have no built-in diffusion and do not create spurious mixing on their own. This idea is a simplified version of the approach proposed by Lemarié et al. [2012], and it is explored here for the advection schemes available in FESOM2. Here and further for the experiments, I consider the scheme based on gradient estimation of triangles (GE) of 4th (centered), 3rd (upwind), or mixed 3rd-4th order, which are implemented and used in the FESOM2. The notation of the used schemes is:

- GE3 - upwind scheme of the 3rd order,
- GE4 - central scheme of the 4th order,
- GE34 - mixed upwind-central scheme with $\lambda = 0.25$,
- GE4 FCT - GE4 scheme with stabilization by FCT,
- GE4 Redi - GE4 scheme with stabilization by isoneutral diffusion.

Finally, it can be assumed that spurious mixing depends on the quality of meshes, and is sensitive to the deviation of mesh triangles from the equilateral one. This question is also addressed here. The experiments were considered on three types of meshes as they were described in chapter 4: ET, QT, DT and IT.

Common to all questions considered here is the use of the RPE method as a diagnostic tool.

## 5.1 Setups

In this chapter for the experiments, three initial setups were used. I considered various channels, initial states of physical properties, and resolution. All these were needed for testing different approaches and analyzing what can play role in growth of RPE. First, I wanted to compare the state of FESOM2 to other models. For this I used Setup1 which was described by Ilicak et al. [2012] and also used by others (e.g., Kärnä et al. [2018], Petersen et al. [2015], Gibson et al. [2018]). However, with this setup, it was impossible to assess the use of isoneutral diffusion because it requires to have changes in density due to both, temperature and salinity, while in the Setup1 salinity is constant. For this reason, I used Setup2 described below. Finally, as it will be seen below, different schemes and meshes can perform differently depending on initial conditions. To see these differences I used Setup3.

Setup1. I consider a zonally re-entrant basin of 160 km in zonal direction and 500 km in the meridional direction centered at the latitude of 54° and with the depth 1000 m. The stratification is due to temperature only. The stratification changes from the warm water at the top to the cold water at the bottom, and also from the warmer water in the north to the colder one in the south (see Fig.5.1). Salinity is constant and set to 35 PSU. The mesh horizontal resolution is 4 km, and the vertical spacing is 25 m. The Coriolis parameter $f = 1.2 \cdot 10^{-4} s^{-1}$. The full equation of state is used. All forcing and explicit dissipation are switched off. The duration of the integration is one year.

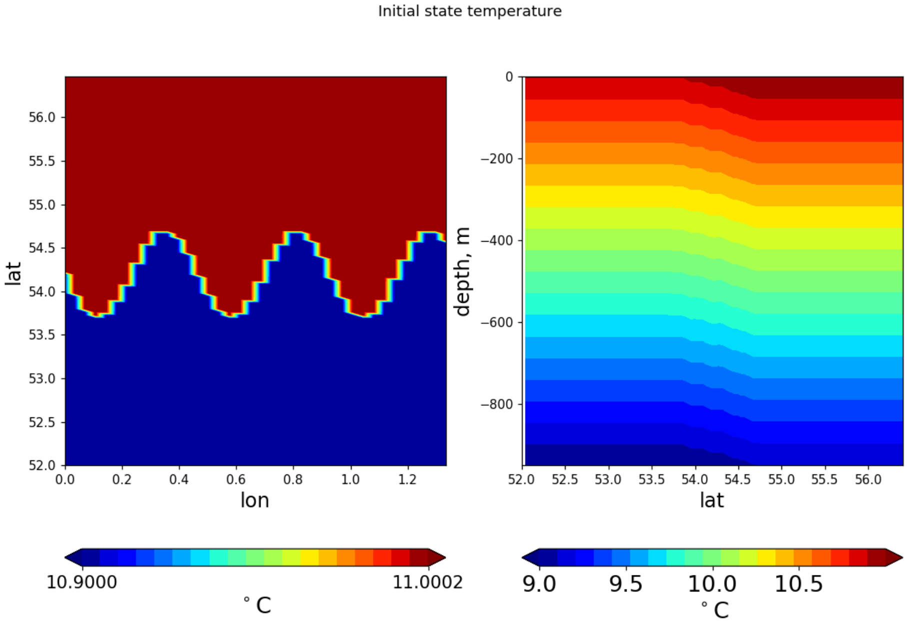


Figure 5.1: Initial state of the temperature in 3D basin of 160 km wide and 1000 m depth.

Setup2. Let us consider the same channel at the latitude 54° with linear stratification by salinity from 34 PSU at the top to 35 PSU at the bottom of the channel. Temperature changes from 22° at the top to 13° at the bottom. The mesh horizontal resolution is 4km, and vertical spacing is as previously 25 m. The isopycnals are tilted in the meridional direction, but they now depart from isothermal surfaces. The initial state is shown in Fig. 5.2, its right two panels show the temperature distribution with the superimposed initial perturbation. The full equation of state is used. All forcing and explicit dissipation are switched off. The duration of the integration is one year.

Setup3. Let us take a bigger 20° wide channel at the latitude 38° with a depth of 1600 meters. By salinity, the stratification was made flat in such a way that the water was changing from 34 PSU at the surface to 35 PSU at the bottom. The temperature was changing from 25 degrees at the surface to the 6 degrees at the bottom and isotherms had a slope $10^{-3}$. Sinusoidal perturbation was added to the temperature field (see Fig. 5.3). All the forcing and vertical mixing are switched off in the system. The full equation of state is used. ALE vertical coordinates condition with the zstar option is used (see details Scholz et al. [2019]). A QT mesh with 18 km with the horizontal resolution was taken.

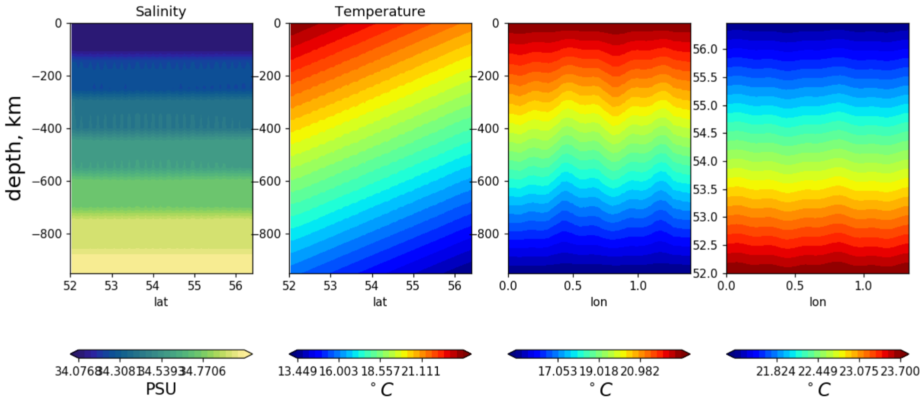


Figure 5.2: Initial state of Setup2. Left to right: the stratification by salinity and stratification by temperature in the vertical meridional section; stratification by temperature with imposed initial perturbations in the vertical meridional section and on the surface.

The vertical resolution varies from 10 m to 50 m until 200 m depth is reached and then is set to 100 m.

## 5.2 Assessment of the advection schemes of FESOM2

First of all, to assess the performance of advection schemes in FESOM2, let us consider the known benchmark initial conditions, first described by Ilicak et al. [2012] and also appearing in Ilicak [2016], Kärnä et al. [2018], Gibson et al. [2018], Petersen et al. [2015]. For these experiments Setup1 was used on the QT mesh. Fig. 5.4 plots the mean RPE of the system over one year. It is seen that over the first few days RPE dramatically jumps up, then grows over around 120 days, and then stabilizes or the increase goes very slowly. This behavior happens since in the beginning the system is extremely unstable, and RPE grows extremely fast during the first days. Then we can observe the mixing of the water with different temperatures and constant salinity, the appearance of eddies and decay of motion in the end. The mean RPE stabilizes at the values between $0.6 \cdot 10^{-6}$ and $3.5 \cdot 10^{-6}$ for different schemes. These values correspond to the ones in the range of viscosity 50 $m^2/s$ and 20 $m^2/s$ of the coastal model by Kärnä et al. [2018], and to the RPE values of viscosity between 1 $m^2/s$ and 5 $m^2/s$ of the MPAS-Ocean model (Petersen et al. [2015]). Unfortunately, I cannot say what is exact viscosity in my system as in FESOM2 it is implemented in a dynamic way

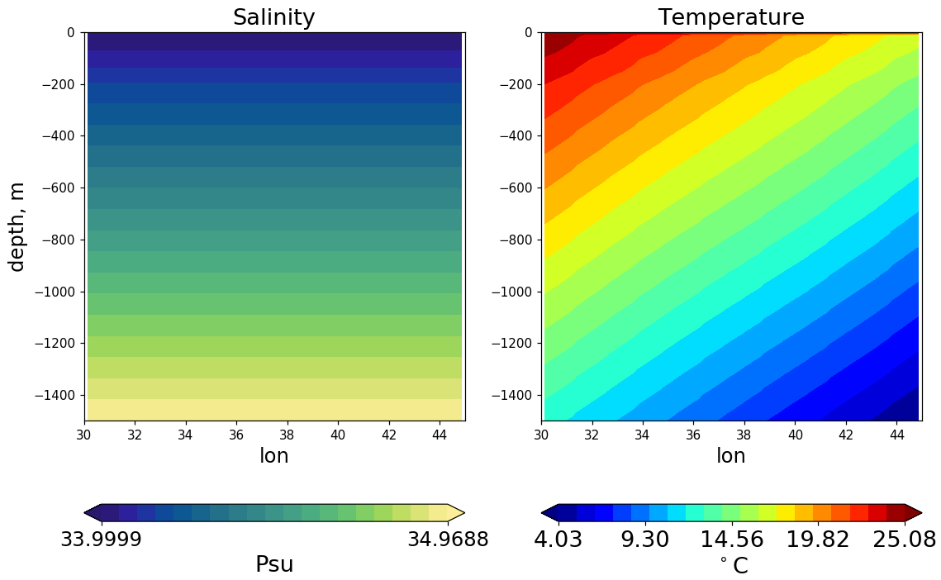


Figure 5.3: Initial state of the temperature and salinity in 20° 3D basin of 1600 m depth.

and changes in time and space. Nevertheless, I calculated the mean viscosity of the system at 200 days of the simulations and it stands at an approximate value of 3 $m^2/s$. Thus, it is possible to say that the output of our model is comparable to others.

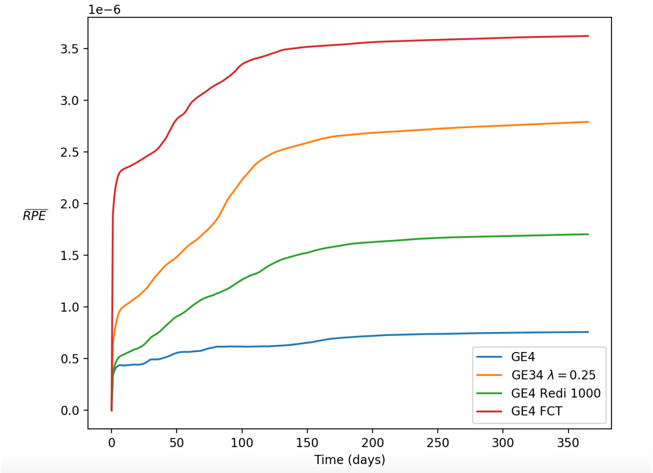
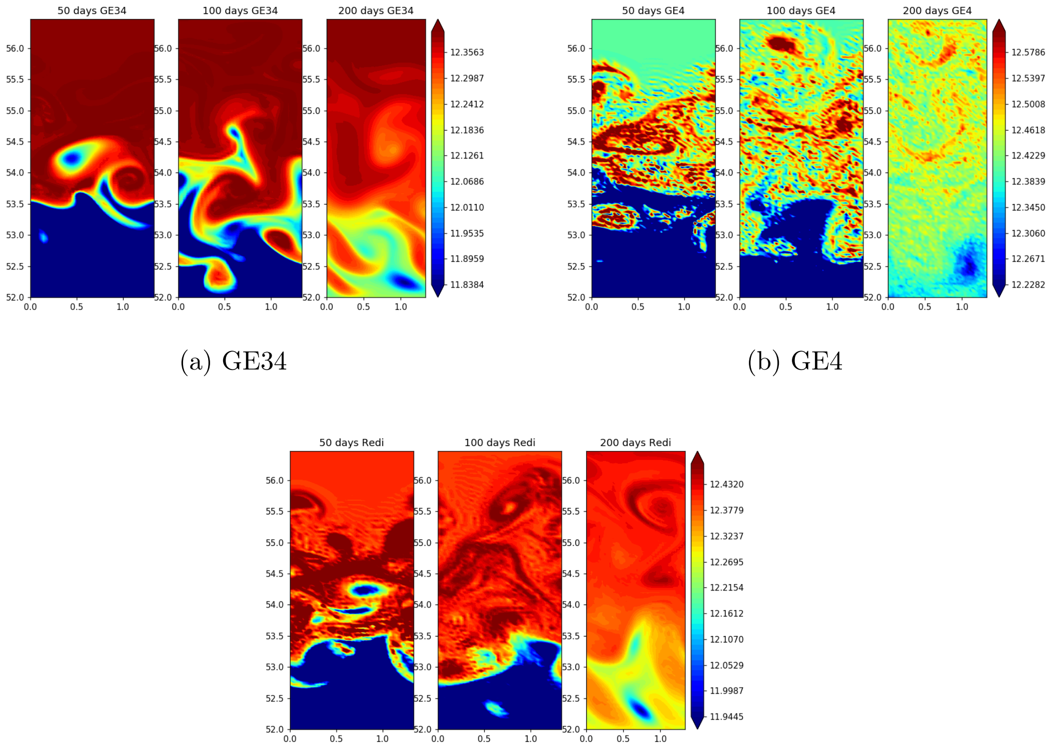


Figure 5.4: Mean RPE value for the integration period over one year of the different schemes: central GE4, GE4 Redi with isoneutral diffusivity $K_{iso} = 1000$ $m^2/s$, GE4 FCT and mixed GE34 with the parameter $\lambda = 0.25$. Setup1.

As it was expected, the mixed scheme GE34 and the central scheme with FCT (GE4 FCT) produce the largest mixing. This effect appears due to dissipation which is brought by both upwind scheme and FCT. Apparently, FCT creates even more spurious mixing than upwinding. Meanwhile, RPE of the central scheme stays almost as flat as it is supposed to be because there is no implicit dissipation involved. RPE of the scheme with isoneutral diffusion, GE4 Redi, is the second smallest after the GE4. Although in the current experiment I expected it to stay similar to the GE4 due to density changes only by temperature, it still shows some growing values. This expectation goes from the definition of rotated (isoneutral) diffusion (see chapter 3) because isoneutral and isothermal surfaces are nearly identical in this case so that the observed behavior appears from the implementation of the scheme. We should remember that the vertical part of the GE4 Redi scheme is computed implicitly. Also, to avoid blow-up of the model when the isopycnals slope becomes too steep (e.g., on approaching the mixed layer), the isopycnal slope is tapered, which introduces horizontal diffusion. Furthermore, in my implementation of the isoneutral diffusivity in FESOM2 the slope vector is smoothed over scalar prisms to enhance model stability.

If we look at Fig. 5.5, we will see that the schemes GE4 and GE4 Redi cause oscillations due to their instability. This instability occurs because these schemes either do not have built-in dissipation or have only isoneutral dissipation and in the described experiments density layers are set only by temperature while salinity stays constant. Since the GE4 Redi shows less oscillations, it can be guessed that it comes from the implicit time stepping of the vertical part of rotated diffusion. If we want to compare the behavior of these schemes better, we should set density layers in such a way that their density changes due to both fields, temperature and salinity.

Let us now consider Setup2. With this initial state of the system, one can see from Fig. 5.6 that the schemes GE34 and GE4 Redi produce more similar mixing than before. The difference in the produced mixing by GE4 Redi differs from the mixing produced by the scheme GE4 the way more than in the previous case. Also, the scheme GE4 Redi does not lead to oscillations anymore, while the GE4 still does (see Fig. 5.7). This behavior is expected as now the density layers deviate from isothermal and isohaline surfaces so that there is isoneutral mixing of both temperature and salinity which stabilizes the scheme.



(a) GE34

(b) GE4

(c) GE4 Redi

Figure 5.5: Surface temperature after 50, 100 and 200 days of integration of the system with the different schemes: a)GE34, b) GE4, and c) GE4 Redi.

## Comparison of GE4 Redi with other schemes

To avoid noise due to instability in the GE4 Redi scheme, let us further consider a new set up with stratification induced by both salinity and temperature. Further Setup3 was used.

Fig. 5.8 demonstrates five cases: isonetral diffusion with $K_{hor} = 3000$ and $K_{hor} = 500$ (GE4 Redi), FCT correction of the GE4 method, mixed upwind-central scheme GE34 with $\lambda = 0.25$ and the pure central scheme GE4. Vertical parameter $K_{ver} = 10^{-6}$. As we can see, the largest mixing comes from the GE34 scheme, and the GE4 is characterized by the least spurious mixing as expected. To see the behaviour of central scheme with isoneutral diffusion, the system was integrated for a longer period. From Fig. 5.9 we can see that RPE grows smoothly. Still, there is a question of why it grows so much for the isoneutral diffusion. To test the correctness of the implementation, the ALE condition was switched off and a linear free surface was used instead of ALE. Also, all the velocities in the system were removed, and small vertical

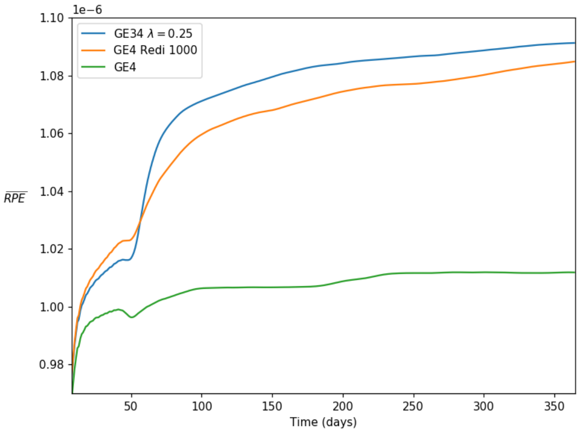
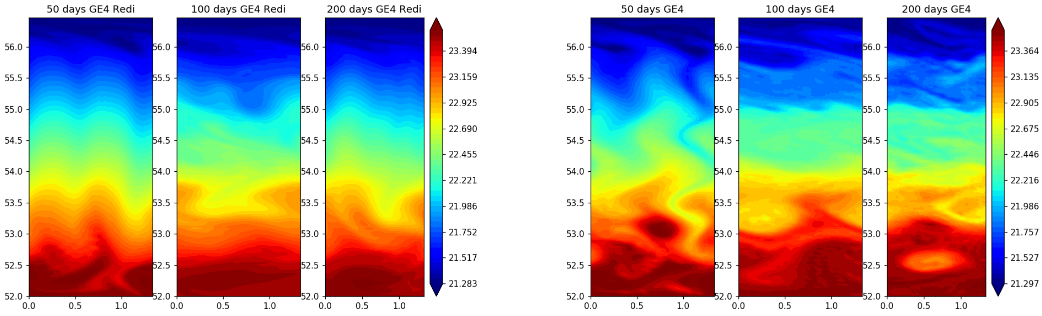


Figure 5.6: RPE of the system with the schemes GE34, GE4 Redi, GE4 after one year of integration. Setup2.



(a) GE4 Redi

(b) GE4

Figure 5.7: Surface temperature after 50, 100 and 200 days of integration with the schemes GE4 Redi and GE4. Density stratification is due to both temperature and salinity.

mixing with vertical diffusivity $K_v = 10^{-6}$ m$^2$/s was added. Experiments were run with the isoneutral diffusion, FCT, and pure central scheme. The scheme GE4 FCT was used together with the benchmark GE4 to assess the behavior of isoneutral diffusivity. Fig. 5.10 demonstrates that RPE growth is the same for all cases. The growth of RPE is only due to the parameter $K_v$. It means that the spurious mixing identified above for the scheme GE4 Redi stabilized with isoneutral diffusivity comes from the presence of motion. In order to see it better, let us consider the case with different horizontal viscosities. The higher viscosity, the lower velocity fluctuations in the system, and vice versa. Indeed, reducing viscosity in two times, we get higher mixing, and making viscosity

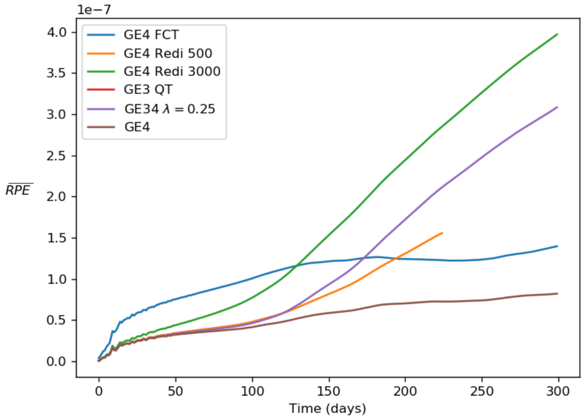


Figure 5.8: Mean RPE of the GE based methods in FESOM2 in the 20° channel with QT mesh with resolution of 1/6 degree.

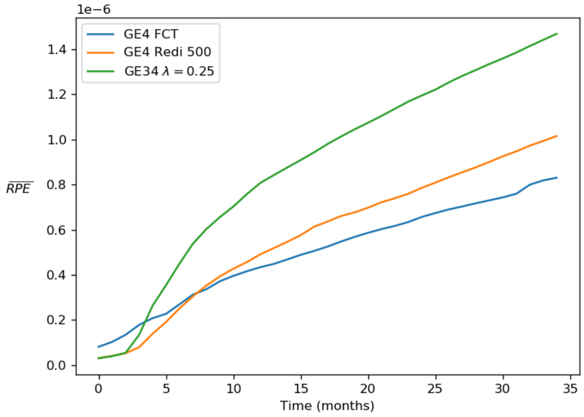


Figure 5.9: Evolution of the mean RPE of the GE based methods in FESOM2 throughout 35 days in the 20° channel with QT mesh with resolution of 1/6 degree.

five times higher, we reduce mixing in the system (see Fig. 5.11). However, this correlation is not linear. It means that with default viscosity, velocity fluctuations are already moderate; with the increased viscosity, they are still further reduced but the effect is weaker than the growth in RPE created by increasing velocity fluctuations for reduced viscosities.

## 5.3 Analysis of different types of meshes

Triangular meshes can have different triangles on their basis: equilateral, right, random, and so on. They can be unstructured, irregular, distorted (changing

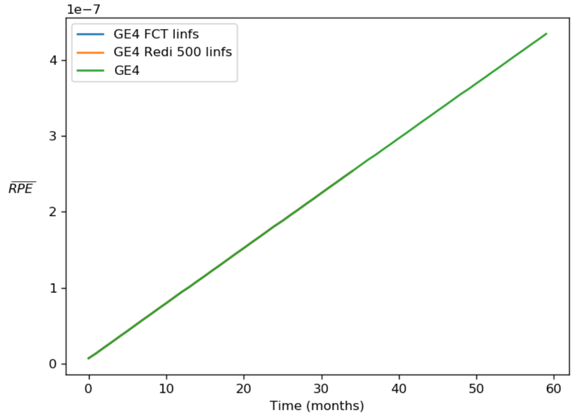


Figure 5.10: Mean RPE of the system without velocities. Results are the same for integration with different schemes.

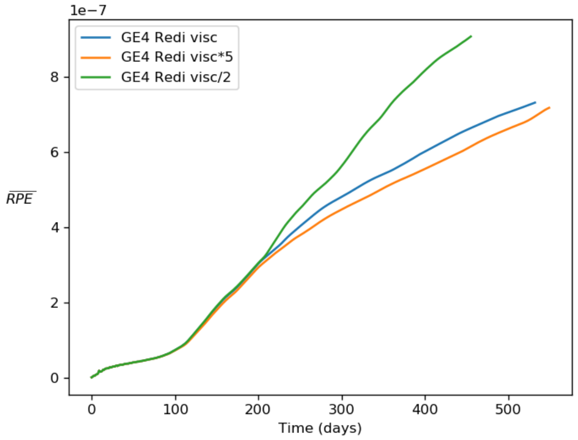
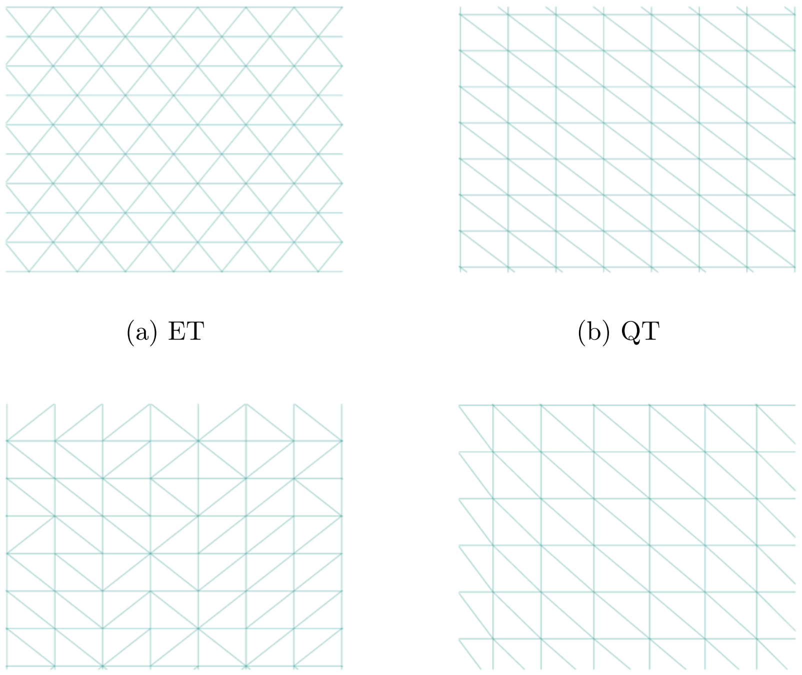


Figure 5.11: System with ALE zstar conditions, $K_{iso} = 1000$.

spatial step in the direction of distortion). Interesting question if the structure of a mesh can bring unwanted numerical mixing. Sometimes, it is hard to avoid irregularity in meshes, therefore it is important to know how much spurious mixing such meshes can cause. As in chapter 4, several meshes will be considered: ET, IT, DT, and QT (see Fig. 5.12). The experiments were held in the channel at the 54° described above. For the initial state Setup1 was taken. Meshes resolution was taken 4 km. For QT and IT meshes it means that the catheti of the triangles are 4 km. For the DT mesh, it means that in the meridional direction cathetus is 4 km, and in the zonal direction mean length of a cathetus is 4 km (my DT mesh is distorted only in the longitudinal direction). It is a little bit harder to compare these meshes to an ET mesh. For this, I considered two



(a) ET

(b) QT

(c) IT

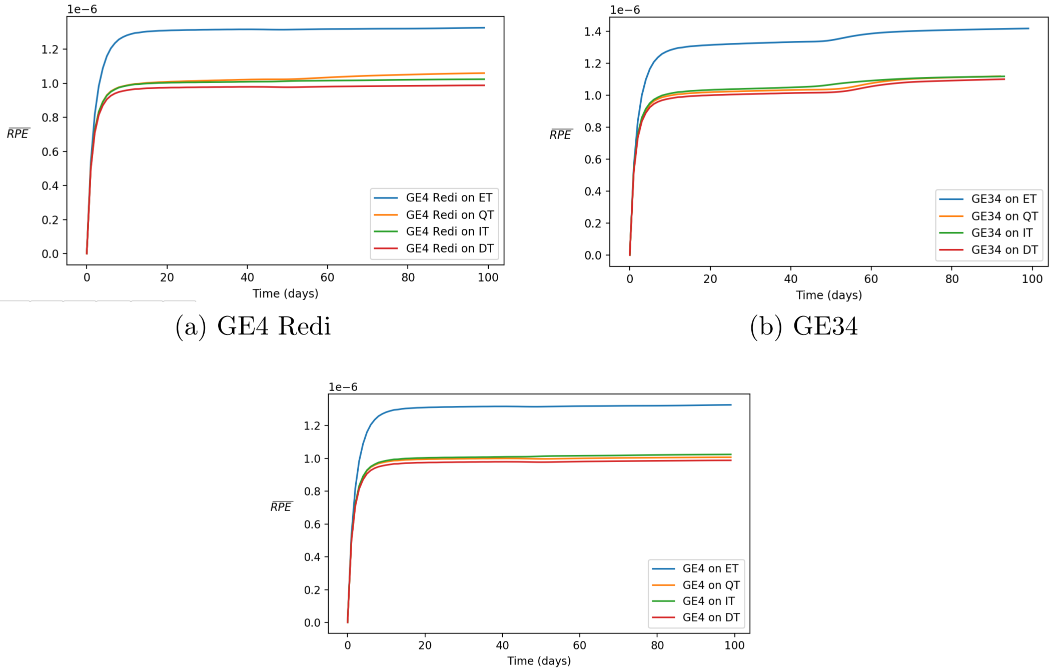
(d) DT

Figure 5.12: Fragments of meshes : a) equilateral mesh (ET), b) quadrilateral mesh (QT), c) irregular mesh (IT), d) horizontally distorted mesh (DT).

types of ET meshes. First, when the height of the equilateral triangle is taken as 4 km, its side is equal to ≈ 4.6 km. Thus, this ET mesh has ≈ 1.15 times more elements than the QT mesh. Another way to obtain a comparable resolution for an ET mesh is to take a side in such a way that its area is equal to the area of the right-angled triangle. Thus, we will get the same number of mesh elements. However, due to numbers rounding while calculating a side of the equilateral triangle, the number of the elements of the ET mesh differs slightly from the QT mesh. When I talk about this kind of ET meshes, I will mention it as ET mesh with bigger elements because its elements are bigger than triangles of the ET mesh derived from the first way.

The methods GE4, GE4 Redi and GE34 were taken to see the different behavior on the different meshes.

The results of the RPE experiments on the different meshes are shown in Fig. 5.13. As we can see, experiments on the meshes QT, IT and DT show almost the same amount of spurious mixing in the system. Surprisingly, ET mesh brings more mixing to the system than other meshes. The reason for it might be lying in the way of calculating horizontal viscosity in FESOM2.



(a) GE4 Redi

(b) GE34

(c) GE4

Figure 5.13: RPE of three schemes depending on different types of meshes: a) GE4 Redi scheme, b) GE34 scheme, c) GE4 scheme.

To get this value, harmonic Leith viscosity is calculated and combined with biharmonic background viscosity. Both of them are scaled to the mesh resolution and recalculated with every time step. Thus, we cannot talk about a particular numerical number related to viscosity and have to take into account that it relies on a mesh resolution. As long as it depends on the area of a mesh element, we can think that an ET mesh with bigger elements will produce the amount of mixing which is closer to the QT-type meshes. Another reason can be that before the initial condition with constant salinity and high instability due to abrupt temperature change were used. Let us further consider cases with stratification by both salinity and temperature and less initial instability with the Setup2.

As it was already mentioned, viscosity in FESOM2 is scaled according to the mesh resolution. Thus, knowing, that area of a triangle of the ET mesh is 1.15 times smaller than the area of a triangle of the QT mesh, we can try to increase viscosity in 1.15 times for ET mesh. From Fig. 5.14 we can see that RPE with the ET scheme with increasing of viscosity in 1.15 times is actually lower than before, however, the difference is not that high and the values of RPE in such a case are still way larger than ones for the QT and IT meshes. If we want to get comparable values of RPE for this ET mesh, viscosity has to be increased at

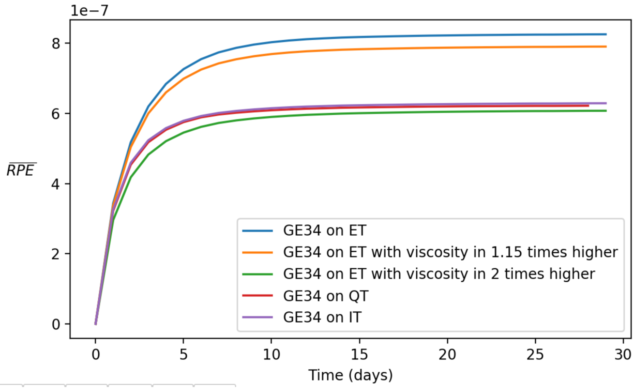


Figure 5.14: Mean RPE of the scheme GE34 on the different meshes: QT, IT, ET depending on viscosity (for the ET mesh).

least two times. From Fig. 5.15 it is seen that with enlargement of the element

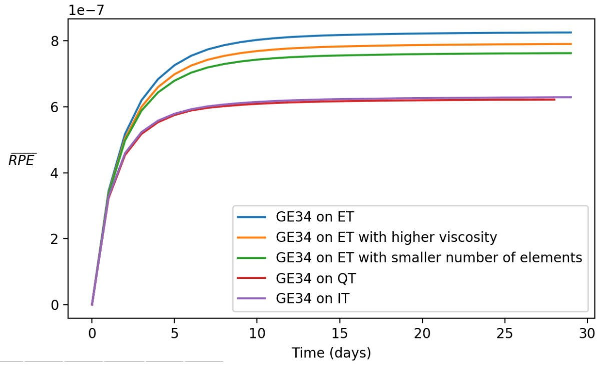
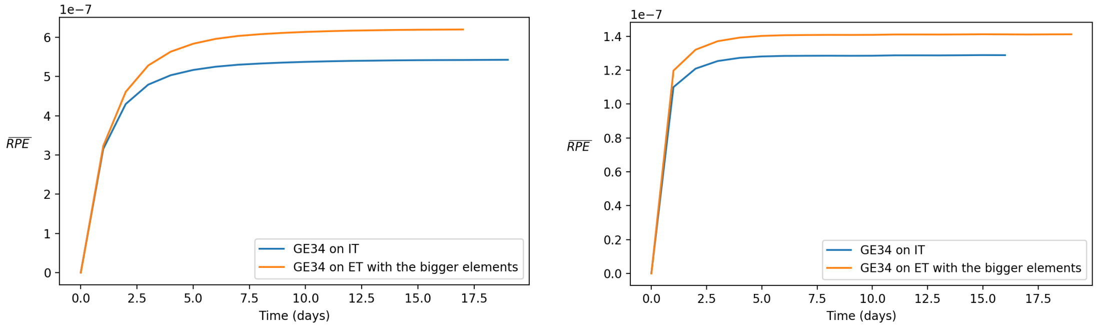


Figure 5.15: Mean RPE of the scheme GE34 on the different meshes: QT, IT, ET with bigger elements, ET depending on viscosity (for the ET mesh).

area of the ET mesh, the system produces less mixing. However, the amount of it is still larger than for the system with QT based meshes. To avoid variable horizontal viscosity, I considered cases with the viscosity set to 20 $m^2/s$ and 200 $m^2/s$. Fig. 5.16 demonstrates that with the growth of viscosity, mixing reduces for both types of meshes, IT and ET. Also, the difference in the RPE values for the both meshes decreases. However, RPE calculated on the ET mesh has never reached the same RPE values derived on the IT mesh. This behavior can appear due to the way the gradients are calculated in FESOM2. Even though the constant viscosity is set, because of the different geometry in QT and ET based meshes, they are calculated differently. Thus, it is not possible to rigorously solve the problem of comparison of such types of meshes in FESOM2.



(a) Viscosity $20m^2/s$

(b) Viscosity $200m^2/s$

Figure 5.16: Mean RPE of the system on ET mesh with bigger elements and IT mesh with different horizontal viscosity.

## 5.4 Conclusion

In this chapter, I examined spurious mixing in the model FESOM2 through the RPE analysis. The following concepts were investigated:

- Whether isoneutral diffusion can lead to the reduction of diapycnal spurious mixing if used to stabilize central schemes;
- Whether the structure of meshes can influence the amount of mixing.

From the RPE analysis of the GE4 Redi scheme compared to the GE34 and GE4 FCT, we can conclude that, indeed, the GE4 Redi scheme brings less mixing than others. However, it requires additional conditions for stable work. Under particular circumstances, the GE4 Redi might outperform other schemes. Nevertheless, there are also conditions when the GE34 produces similar mixing or even less than the GE4 Redi. The same thing we can say about the GE4 FCT though in some cases FCT brings the largest amount of mixing into the system among other schemes. Also, the performance of the GE4 Redi scheme depends on the several things one should keep in mind while using it for reduction of spurious mixing:

- Choice of the value of the isoneutral diffusivity parameter $K_{iso}$. The less this parameter is, the less mixing GE4 Redi produces. However, it also becomes more unstable because with $K_{iso} = 0$ we will get the common central scheme GE4. Thus, it is important to choose the parameter $K_{iso}$ wisely.
- Isoneutral diffusion can perform differently depending on model settings such as initial state, choice of a mesh, or parameterization.
- It is also important to notice that in the current work harmonic isoneu-

tral diffusion was used. Biharmonic isoneutral diffusion stays beyond the scope of the current work and needs an additional analysis of models with triangular meshes.

Concluding, isoneutral diffusion can help to reduce spurious mixing in ocean models but it requires particular tuning of a model.

The second conclusion of the current chapter is about the performance of different meshes. Even though it was expected that IT and DT meshes produce the largest spurious mixing, the RPE analysis shows that spurious mixing is not necessarily sensible to distortion and irregularity of the meshes. Moreover, despite the better accuracy of ET meshes, they do not lead to the reduction of spurious mixing. As it was seen from the experiments with viscosity, it might be a possible reason for it. As Ilicak et al. [2012] showed, RPE is sensitive to the viscosity of a system; we also observe similar behavior in the provided experiments. We see that changes in viscosity lead to changes in RPE of the system which have a larger impact on the system mixing than any other change caused by other actions such as parametrizations, choice of a mesh type, or diffusion scheme. Thus, a type of a mesh does not have a noticeable influence on spurious mixing in ocean models.

# Chapter 6

# Conclusion and Outlook

In the current work, I pursued a goal to investigate ways to reduce spurious diapycnal mixing in ocean circulation models. Numerical climate models are important for understanding and exploration of various processes influencing the ocean and climate behavior. However, unwanted numerical mixing occurring in these models brings uncertainty to their output. Spurious mixing is created by numerics whenever we explicitly or implicitly introduce unphysical mixing. One may hope to reduce it by using more accurate advection schemes and by replacing the dissipative part with isoneutral diffusion, as well as through the use of a "good" mesh that locally tends to equilateral triangles or regularly split squares. These directions were examined.

- Attempting to solve the problem of spurious mixing, the new compact advection scheme was introduced. The experiments and the analysis showed that although in a 2D case it and the GE56 reach the highest accuracy, in 3D case all the schemes lead to only small differences in the EKE. The compact scheme demonstrates high accuracy and in some cases outperforms other schemes. Though the difference in EKE in the 3D case is not too big, due to the higher numerical efficiency of the compact scheme it is recommended upon the others.
- Isoneutral diffusion scheme was for the first time analyzed on triangular meshes for vertex-based finite volume discretization of FESOM2, and estimates of stability limits were provided. Isoneutral diffusion scheme was implemented in FESOM2.
- RPE diagnosis has been applied to explore whether isoneutral diffusion can help to stabilize advection schemes with no build-in dissipation. It has been found that the residual dissipation of the isoneutral operator in

FESOM2 still creates some spurious mixing. Although spurious mixing can be made lower than for schemes with 3-rd order upwind or FCT, it is not as small as hoped. Furthermore, additional attention is required in choosing the isoneutral diffusivity parameter $K_{iso}$.

- RPE diagnostics has been also applied to study the effect of mesh type and quality on spurious mixing. Even though in 2D case it is clear that the results are more accurate on the meshes with equilateral triangles, in 3D case this type of meshes brought more spurious mixing in the system. Taking into account dynamically calculated horizontal viscosity depending on a mesh type, it did not seem to be possible to compare different types of meshes rigorously. The comparison of irregular, distorted, and normally organized meshes showed that irregularity and distortion do not bring any additional spurious mixing into the system.

Concluding, the new compact scheme is recommended to be chosen among other advection schemes considered in this work. Isoneutral diffusion can be applied in models with triangular meshes to stabilize advection schemes without build-in dissipation providing lower spurious mixing under particular circumstances than high-order upwind or FCT. The type of meshes does not play a big role in numerical mixing reduction.

There are still ways to investigate what can influence the reduction of spurious mixing. One of them is a biharmonic form of isoneutral diffusion. Lemarié et al. [2012] showed that it reduces spurious mixing on rectangular meshes, therefore, it has to be analyzed on triangular meshes as well. Another point to pay attention to is the implementation of a terrain-following coordinate system at overflow sites that would help to resolve plumes and minimize their spurious mixing. Further, the coordinates that follow isopycnals in the deep ocean can be used. With this type of coordinates, advection will be naturally within the isopycnal layers. In addition, the compact scheme introduced in this work was not analyzed systematically with the RPE method together with isoneutral diffusion. The reason is that it is still not included in the released version of FESOM2 which should be done in the future. It is also important to notice that the experiments in the current work were held for idealized cases. Analysis of spurious mixing for real ocean conditions will be another interesting topic to investigate. There are different processes in ocean circulation models which can cause unwanted numerical mixing, and it is crucial to see how much the reduction of spurious diapycnal mixing improves the whole system output.

# Bibliography

I. Abalakin, A. Dervieux, and T. Kozubskaya. A vertex-centered high-order MUSCL scheme applying to linearized Euler acoustics. Rapport de recherche 4459, INRIA, 2002.

T. J. Barth and P. O. Frederickson. Higher order solutions of the Euler equations on unstructured grids using quadratic reconstruction. Paper 90-0013, AIAA, 1990.

W. P. Budgell, A. Oliveira, and M. D. Skogen. Scalar advection schemes for ocean modelling on unstructured triangular grids. Ocean Dynamics, 57:339–361, 2007.

H. Burchard and H. Rennau. Comparative quantification of physically and numerically induced mixing in ocean models. Ocean Modell., 20:293–311, 2008.

C. Chen, J. Bin, and F. Xiao. A global multimoment constrained finite-volume scheme for advection transport on the hexagonal geodesic grid. Mon. Wea. Rev., 140:941–955, 2012.

P. Clark, N. Pisias, T. Stocker, and A. Weaver. The role of the thermohaline circulation in abrupt climate change. Nature, 415:863–869, 2002.

P. Colella and P. R. Woodward. The piecewise parabolic method (PPM) for gas-dynamical simulations. J. Comput. Phys., 54:174–201, 1984.

S. Community. Arctic sea ice in cmip6. Geophysical Research Letters, 47, 2020. doi: https://doi.org/10.1029/2019GL086749.

S. Danilov. Two finite-volume unstructured mesh models for large-scale ocean modeling. Ocean Modell., 47:14—-25, 2012.

S. Danilov, D. Sidorenko, Q. Wang, and T. Jung. FESOM2: from finite elements to finite volumes. Geosci. Mod. Dev., page Submitted, 2017.

J. Donea and A. Huerta. Finite element methods for flow problems. Willey, 2003.

M. Dumbser and M. Käser. Arbitrary high-order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems. J. Comput. Phys., 221:693–723, 2007.

R. Ferrari, S. Griffies, G. Nurser, and G. Vallis. A boundary-value problem for the parameterized mesoscale eddy transport. Ocean. Model., 32:143–156, 2010.

P. Gent and J. McWilliams. Isopycnal mixing in ocean circulation models. J. Phys. Oceanogr., 20:150–155, 1990.

J. Getzlaf, G. Nurser, and A. Oschlies. Diagnostics of diapycnal diffusivity in z-level ocean models. part i: 1-dimentional case studies. Ocean Modell., 35: 173–186, 2010.

A. Gibson, A. McC. Hog, A. Kiss, S. C.J., and A. Adcroft. Attribution of horizontal and vertical contributions to spurious mixing in an arbitrary lagrangian–eulerian ocean model. Elsevier. Ocean Modelling, 119:45–56, 2018. doi: https://doi.org/10.1016/j.ocemod.2017.09.008.

S. Griffies, A. Gnanadesikan, R. Pakanowski, V. Larichev, J. Dukowicz, and R. Smith. Isoneutral diffusion in a z-coordinate ocean model. J. Phys. Oceanogr., 28:805–830, 1998.

S. Griffies, R. Pakanowski, and R. Hallberg. Spurious diapycnal mixing associated with advection in a z-coordinate ocean model. Mon. Wea. Rev., 128: 538–564, 2000.

M. W. Hecht, B. A. Wingate, and P. Kassis. A better, more discriminating test problem for ocean tracer transport. Ocean Modell., 2:1–15, 2000.

C. Hill, D. Ferreira, J.-M. Campin, J. Marshall, A. Abernathey, and N. Barrie. Controlling spurious diapycnal mixing in eddy-resolving height-coordinates ocean models - insights from virtual deliberate tracer release experiments. Ocean Modell., 45–46:14–26, 2012.

M. Ilicak. Quantifying spatial distribution of spurious mixing in ocean models. Ocean Modell., 108:30–38, 2016. doi: https://doi.org/10.1016/j.ocemod.2016.11.002.

M. Ilicak, A. Adcroft, S. Griffies, and R. Hallberg. Spurious dianeutral mixing and the role of momentum closure. Elsevier. Ocean Modelling, 45–46:37–58, 2012. URL http://www.elsevier.com/locate/ocemod.

M. Karen and S. Richard. Scenarios of carbon dioxide emissions from aviation. Global Environmental Change, 20:65–73, 2007. doi: https://doi.org/10.1016/j.gloenvcha.2009.08.001.

T. Kärnä, S. Kramer, L. Mitchell, D. Ham, M. Piggott, and A. Baptista. Thetis coastal ocean model: discontinuous galerkin discretization for the three-dimensional hydrostatic equations. Geosci. Model Dev., 11:4359–4382, 2018. doi: https://doi.org/10.5194/gmd-11-4359-2018.

K. Klingbeil, M. Mohammadi-Aragh, U. Gräwe, and H. Burchard. Quantification of spurious dissipation and mixing discrete variance decay in a finite-volume framework. Ocean Modell., 81:49–64, 2014.

T. Kuhlbrodt, A. Griesel, M. Montoya, A. Levermann, M. Hofmann, and S. Rahmstorf. On the driving processes of the atlantic meridional overturning circulation. Reviews of Geophysics, 20:65–73, 2007. doi: https://doi.org/10.1029/2004RG000166.

M.-M. Lee, A. Coward, and A. Nurser. Spurious diapycnal mixing of deep waters in an eddy-permitting global ocean model. J. Phys. Oceanogr., 32:1522–1535, 2002.

F. Lemarié, L. Debreu, A. Shchepetkin, and J. McWilliams. On the stability and accuracy of the harmonic and biharmonic isoneutral mixing operators in ocean models. Elsevier. Ocean Modelling, 52–53(27):9–35, 2012. URL http://www.elsevier.com/locate/ocemod.

F. Lemarié, J. Kurian, A. F. Shchepetkin, M. J. Molemaker, F. Colas, and J. C. McWilliams. Are there inescapable issues prohibiting the use of terrain-following coordinates in climate models? Ocean Modell., 42:57–79, 2012.

F. Lemarié, L. Debreu, G. Madec, J. Demange, J. Molines, and M. Honnorat. Stability constraints for oceanic numerical models: implications for the formulation of time and space discretizations. Ocean Modelling, 92:124–148, 2015.

R. Löhner, K. Morgan, J. Peraire, and M. Vahdati. Finite-element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations. Int. J. Num. Meth. Fluids, 7:1093–1109, 1987.

G. Madec and N. S. Team. NEMO ocean engine, 2016.

J. Marshall, A. Adcroft, C. Hill, L. Perelman, , and C. Heisey. A finite-volume, incompressible navier stokes model for studies of the ocean on parallel computers. J. Geophys., 102:5753–5766, 1997.

A. Megann. Estimating the numerical diapycnal mixing in an eddy-permitted ocean model. Ocean Modell., 121:19–33, 2018.

H. Miura. An upwind-biased conservative advection scheme for spherical hexagonal-pentagonal grids. Mon. Wea. Rev., 135:4038–4044, 2007.

H. Miura. An upwind-biased conservative transport scheme for multistage temporal integrations on spherical icosahedral grids. Mon. Wea. Rev., 141:4049–4068, 2013.

H. Miura and W. C. Skamarock. An upwind-biased transport scheme using a quadratic reconstruction on spherical icosahedral grids. Mon. Wea. Rev., 141: 832–847, 2013.

M. Mohammadi-Aragh, K. Klingbeil, N. Brüggemann, C. Eden, and H. Burchard. The impact of advection schemes on restratifiction due to lateral shear and baroclinic instabilities. Ocean Modell., 2015. doi: http://dx.doi.org/10.1016/j.ocemod.2015.07.021.

A. Moore, H. Arango, G. Broquet, B. Powell, A. Weaver, and J. Zavala-Garay. The regional ocean modeling system (roms) 4-dimensional variational data assimilation systems: Part i – system overview and formulation. Progress in Oceanography, 91:34–49, 2011. doi: https://doi.org/10.1016/j.pocean.2011.05.004.

D. Olbers and C. Eden. A global model for the diapycnal diffusivity induced by internal gravity waves. J. Phys. Oceanogr., 43:1759–1779, 2013.

C. Ollivier-Gooch and M. Van Altena. A high-order-accurate unstructured mesh finite-volume scheme for the advection/diffusion equation. J. Comput. Phys., 181:729–752, 2002.

M. R. Petersen, D. W. Jacobsen, T. D. Ringler, M. W. Hecht, and M. E. Maltrud. Evaluation of the arbitrary lagrangian–eulerian vertical coordinate method in the mpas-ocean model. Ocean Modell., 86:93–113, 2015. doi: http://dx.doi.org/10.1016/j.ocemod.2014.12.004.

S. Rahmstorf. Ocean circulation and climate during the past 120,000 years. Nature, 419:207–214, 2002.

M. Redi. Oceanic isopycnal mixing by coordinate rotation. J. Phys. Oceanogr., 12:1154–1158, 1982.

T. Ringler, M. Petersen, R. L. Higdon, D. Jacobsen, P. W. Jones, and M. Maltrud. A multi-resolution approach to global ocean modeling. Ocean Modell., 69:211–232, 2013.

P. Scholz, D. Sidorenko, O. Gurses, S. Danilov, N. Koldunov, Q. Wang, D. Sein, M. Smolentseva, N. Rakowsky, and T. Jung. Assessment of the finite-volume sea ice-ocean model (fesom2.0) – part 1: Description of selected key model elements and comparison to its predecessor version. Geosci. Model Dev., 12: 4875–4899, 2019. doi: doi.org/10.5194/gmd-12-4875-2019.

A. F. Shchepetkin. An adaptive, courant-number-dependent implicit scheme for vertical advection in oceanic modeling. Ocean Modelling, 91:38–69, 2015.

W. C. Skamarock and A. Gassmann. Conservative transport schemes for spherical geodesic grids: high-order flux operators for ode-based time integration. Mon. Wea. Rev., 2011. doi: http://dx.doi.org/101175/MWR-D-10-05056.1.

W. C. Skamarock and M. Menchaca. Conservative transport schemes for spherical geodesic grids: high-order reconstructions for forward-in-time schemes. Mon. Wea. Rev., 138:4497–4508, 2010.

M. Smolentseva and S. Danilov. Comparison of several high-order advection schemes for vertex-based triangular discretization. Ocean Dynamics, 2020. doi: https://doi.org/10.1007/s10236-019-01337-4.

Y. Soufflet, P. Marchesiello, F. Lemarié, J. Jouanno, X. Capet, L. Debreu, and R. Benshila. On effective resolution in ocean models. Ocean Modelling, 98: 36–50, 2016.

K. Vivek and a. et. Carbon-concentration and carbon-climate feedbacks in cmip6 models, and their comparison to cmip5 models. Biogeoscience, 2019. doi: https://doi.org/10.5194/bg-2019-473.

Q. Wang, S. Danilov, D. Sidorenko, R. Timmermann, C. Wekerle, X. Wang, T. Jung, and J. Schröter. The finite element sea ice-ocean model (fesom) v.1.4: formulation of an ocean general circulation model. Geosci. Model Dev., 7:663—693, 2014.

D. Webb, B. A. de Cuevas, and C. Richmond. Improved advection schemes for ocean models. J. Atm. Ocean. Tech., 15:1171–1187, 1998.

K. Winters, P. Lombard, J. Riley, and E. D'Asaro. Available potential energy and mixing in density-stratified fluids. Fluid Mechanics, 289:115–128, 1995.

X. Xu, P. B. Rhines, and E. P. Chassignet. On mapping the diapycnal water mass transformation of the upper North Atlantic Ocean. J. Phys. Oceanogr., 48:2233–2258, 2018. doi: https://doi.org/10.1175/JPO-D-17-0223.1.

F. Ye, Y. J. Zhang, R. He, Z. Wang, H. V. Wang, and J. Du. Third-order WENO transport scheme for simulating the baroclinic eddying ocean on an unstructured grid. Ocean Modelling, 2019. doi: https://doi.org/10.1016/j.ocemod.2019.101466.

S. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. J. Comput. Phys., 31:335–362, 1978.

S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. J.Comput. Phys., 31:335–362, 1979.

M. Zerroukat, N. Wood, and A. Staniforth. The Parabolic Spline Method (PSM) for conservative transport problems. Int. J. Numer. Meth. Fluids, 51:1297–1318, 2006.

# Acknowledgements

I would like to thank Prof. Dr. Sergey Danilov for his constant support, answering all my questions, and helping me to get to the right way when I was lost in my Ph.D. I am thankful to Lorenzo Zampieri and Vera Fofonova for their help in improving this manuscript. I want to thank Nikolay Koldunov and Dmitry Sidorenko for their help, advice, and special humor. I am grateful to my office-mates Deniz Aydin and Damien Ringeisen who were always around and made the time I was working on my Ph.D. full of interesting conversations.

And last but not least, I would like to thank all the important people around me, my family, my partner, and my friends. With their support, I was able to finish my work.

The work has been supported by projects M5 and S2 of the Collaborative Research Centre TRR 181 "Energy Transfer in Atmosphere and Ocean" funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) under project number 274762653.

Милая мама, без тебя я бы не смогла написать эту диссертацию. Спасибо тебе за образование, заботу, веру в меня и твою любовь.