Design and implementation of a data-driven parameterization for mesoscale thickness fluxes

Dhruv Balwada[1], Pavel Perezhogin[2], Alistair Adcroft[4], and Laure Zanna[2,3]

[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA
[2]Courant Institute of Mathematical Science, New York University, New York, NY, USA
[3]Center for Data Science, New York University, New York, NY, USA
[4]Princeton University, Princeton, NJ, USA

Key Points:

- A data-driven mesoscale thickness flux parameterization was designed and implemented in MOM6 as an alternative to the Gent-McWilliams scheme
- The parameterization is stable and skillful across a range of resolutions including the eddy-permitting gray zone
- The parameterization is able to reduce potential energy without overly dissipating the resolved eddy energy

Corresponding author: Dhruv Balwada, `dbalwada@ldeo.columbia.edu`

Abstract

Mesoscale eddies are a major sink of available potential energy (APE) in the ocean. When these eddies are not resolved or only partially resolved in a model, this effect needs to be parameterized to simulate a realistic large-scale state. Traditionally, the Gent-McWilliams (GM) parameterization has provided this sink of APE. However, the GM parameterization, which diffuses isopycnal heights, is not accompanied by a skillful prescription for GM diffusivity rooted in data from observations or models. Also, at eddy permitting resolutions, GM diffusion can negatively impact resolved eddies, and the only scale-aware prescription is to turn GM off in regions where eddies are permitted. Here we present a novel data-driven parameterization, as a substitute for GM, that extracts APE without overly negative impacts on the resolved flow. It is both flow-aware and scale-aware, and its magnitude can be tuned using an O(1) non-dimensional number. Features like non-dimensional inputs/outputs, lateral non-locality, flow-dependent coordinates, and range limitations improve the generalization of the data-driven scheme. Functional forms are learned via a small multi-layer perceptron, ensuring low computational cost and simple implementation in ocean models. The parameterization performs skillfully in offline evaluation, especially at scales smaller than the largest eddies. It is implemented in NOAA GFDL?s MOM6 and shown to be skillful in online tests in two-layer idealized simulations of a zonal channel and wind-driven gyre, at both eddy-permitting and non-eddying resolutions. This work suggests a path towards leveraging high-resolution simulations for the reduction of structural error and improvement in the fidelity of climate simulations.

Plain Language Summary

Mesoscale (100 km) eddies are the dominant flows in the ocean and play a key role in shaping large-scale circulation features such as wind-driven gyres and the meridional overturning circulation. Since these eddies are not fully resolved in many modern ocean models – especially those that are run for long periods or include many ensemble members – their effects must be represented through parameterizations. A commonly used approach, the Gent-McWilliams (GM) parameterization, removes available potential energy (APE) from the system but lacks a data-driven way to set its strength. Moreover, at eddy-permitting resolutions, GM can interfere with the resolved flows.

We present a new data-driven parameterization designed to better represent eddy effects in ocean models. It learns a functional form from high-resolution simulations using a compact neural network and is designed to be flow-aware, scale-aware, and computationally efficient. The parameterization is implemented in the MOM6 ocean model and shows skillful performance both offline and in idealized simulations, especially at scales smaller than the largest eddies. It extracts APE from large scale flow without degrading resolved features, offering a promising alternative to GM for a wide range of ocean modeling applications.

1 Introduction

Ocean circulation models solve equations describing the motions in the ocean on a finite-size discrete grid, and are thus unable to resolve the phenomena at scales smaller than the grid scales. However, it is often the case that these *sub-grid* phenomena are more than just small-scale variability, and could have a profound impact on the characteristics of the resolved motions. To ensure fidelity of the model behavior at the resolved scales, the effects of these sub-grid phenomena need to be appropriately *parameterized*. One important and often unresolved range of scales in ocean models are the mesoscales.

Mesoscales($\sim 50-200$km) are the dominant energy-containing scales in the ocean (Ferrari & Wunsch, 2009), and subsequently play an important role in shaping the mean circulation and stratification (Gent, 2011), and in transporting tracers (Abernathey & Wortham, 2015). These eddies are believed to be largely generated as a result of baroclinic instability (K. S. Smith, 2007), which has its fastest growth rate at scales close to the first baroclinic deformation radius (scales of $\sim 10-100$km) (Chelton et al., 1998; Tulloch et al., 2011). These instabilities have a tendency to grow at the expense of the available potential energy (APE), and thus their bulk effect is to flatten isopycnals in the ocean. The dominance of rotation at these scales also usually leads to a subsequent inverse energy cascade, which is why the dominant peak of energy is usually larger than the dominant scale of the instability (Tulloch et al., 2011). Thus, to properly resolve these eddies and their effects the model grid spacing needs to be at least as small as the deformation radius (Hallberg, 2013). The inverse cascade also takes place in the vertical (K. S. Smith & Vallis, 2002), transferring energy to the barotropic and first couple of baroclinic modes, and the associated flows tend to have weak vertical shear and do not generate of small-scale turbulence or diapycnal mixing — mesoscale eddies are dominantly adiabatic processes in the interior of the ocean.

Conventionally, mesoscale eddy effects have been parameterized using the Gent-McWilliams (GM) parameterization (Gent & Mcwilliams, 1990), particularly in models that do not resolve the deformation radius. This parameterization was designed to respect two important aspects of mesoscale processes: (i) the parameterization is adiabatic, and (ii) the net effect of the parameterization is to reduce available potential energy. In models with depth as the vertical coordinate, this is achieved by representing the horizontal eddy fluxes in the form of a horizontally downgradient buoyancy diffusion and then setting the vertical component of the eddy flux to be upgradient, such that the net flux is along isopycnals and the resulting operator behaves like advection. In contrast in isopycnal models, this is achieved by diffusing the interface heights, which can also be represented as extra eddy driven advection. These recipes led to dramatic qualitative improvements in the ocean simulations, particularly the adiabatic aspect ensures that water masses were not eroded away in the interior by spurious diffusion. However, the associated eddy diffusivity has always been a major source of uncertainty and a very active topic of research for decades (e.g. Visbeck et al., 1997; Ferreira et al., 2005; Eden & Greatbatch, 2008; Marshall et al., 2012; Jansen et al., 2015). Also, as with all conventional ocean parameterizations, this scheme was designed to represent the bulk effects of eddies and not to have any skill in representing the spatial or temporal structures of the eddy effects, which leads to detrimental effects at resolutions where eddies are partially resolved (Hallberg, 2013; Mak et al., 2023).

In recent years, machine learning (ML) based methods have started to show a lot of promise in improving different aspects of computational modeling, including improving parameterizations (Bracco et al., 2025; Lai et al., 2024). While conventional parameterizations require domain scientists to develop mathematical operators that are at best able to mimic the bulk effects of sub-grid phenomena, ML methods directly learn the functional relationship between the sub-grid effects and the large-scale fields using appropriate data. In many instances it is also possible to design the ML models to obey some physical properties. These methods have led to the development of parameterizations for idealized systems (Ross et al., 2023; Srinivasan et al., 2023), thermodynamic and momentum tendencies in the atmosphere (Brenowitz & Bretherton, 2018; Yuval et al., 2021), boundary layers processes in the ocean (Sane et al., 2023; Ramadhan et al., 2020; Bodner et al., 2023), and momentum tendencies in the ocean (Zhang et al., 2023; Zanna & Bolton, 2020; Perezhogin et al., 2024). All these new parameterizations have shown improved skill over conventional parameterizations, and the potential issues raised about implementation, stability and generalization are rapidly being addressed. While some sub-grid effects of ocean eddies have been investigated using this approach, the impact of mesoscale eddies on the density or thickness field - the aspect parameterized by the GM parameterization - has not yet been addressed using data-driven parameterizations.

Here we present a new data-driven parameterization to account for the impact that mesoscale eddies have on the thickness field in the ocean. Our parameterization is designed to ensure that it maintains the important adiabatic constraint that was introduced by the GM parameterization. However, unlike GM, our parameterization has not been designed to be a local sink of APE; rather it is optimized to capture the spatial structure of the eddy effects. With an eye towards implementation, we use a small fully connected neural network - multi-layer perceptron (MLP), which can be easily added to any ocean model code. Here we implemented this into GFDL's Modular Ocean Model 6 (MOM6). In section 2 we describe the filtering framework that is used to diagnose the sub-grid impact of eddies from high resolution (HR) simulations, and in section 3 describe how to cast these sub-grid effects into a functional form that potentially has some ability to generalize to unseen data. Also, in section 3 we describe the machine learning architecture, and training process. In section 4 we describe the HR datasets and how they were processed. In section 5 we show that our parameterization is successful in both an offline and online sense, and finally, in section 6 we conclude with a discussion, potential caveats and an outlook towards future work.

2 Sub-grid thickness fluxes

The mesoscale processes can be most strictly isolated with the help of an isopycnal model, as the adiabatic and diabatic processes are clearly distinguished in this framework. In this framework (stacked shallow water or isopycnal coordinate) the flow can be modeled using momentum and thickness equations (e.g. (Vallis, 2017; Loose, Marques, et al., 2023)). The thickness equation, the primary focus of our study, can be written for each layer as,

$$\partial_t h_n + \nabla \cdot (\mathbf{u}_n h_n) = 0, \tag{1}$$

where h_n and $\mathbf{u}_n = (u_n, v_n)$ are the thickness and velocity in the n^{th} layer respectively.

This equation can simulate the flow over the full spectrum of scales where its assumptions are valid, but with a finite grid size only a limited range of scales can be resolved. Here we distinguish between scales that can be resolved and are too small to resolved (sub-grid) using spatial filtering and coarse-graining $\overline{(\cdot)}$, as is routinely done in the large eddy simulation (LES) framework (Sagaut, 2005; Aluie et al., 2018). The high-pass signal after applying this spatial operation is referred to as *sub-grid* flows in this study. Consequently, the impact of sub-grid flow, on the resolved flows, can be elucidated in the thickness equation as,

$$\partial_t \overline{h}_n + \nabla \cdot (\overline{\mathbf{u}}_n \overline{h_n}) = -\nabla \cdot (\overline{\mathbf{u}_n h_n} - \overline{\mathbf{u}}_n \overline{h_n}) \tag{2}$$

The impact of the sub-grid flow (e.g. $\mathbf{u}_n - \overline{\mathbf{u}}_n$) on the resolved flow (e.g. $\overline{h}_n$) arises as the divergence of a sub-grid flux on the RHS ($\nabla \cdot \mathbf{F}_n = \nabla \cdot (\overline{\mathbf{u}_n h_n} - \overline{\mathbf{u}}_n \overline{h_n})$). Henceforth, we shall refer to

$$\mathbf{F}_n = \overline{\mathbf{u}_n h_n} - \overline{\mathbf{u}}_n \overline{h_n} \tag{3}$$

as the sub-grid scale (SGS) thickness flux, which will be the target of the parameterizations we develop below. It is common practice to represent this SGS thickness flux in terms of an eddy-driven stream function or velocity, as described in Appendix B. Also, in this theoretical framing we have assumed that our spatial filtering and coarse-graining commutes with the derivatives, which may not be true for all filter choices and near boundaries (Moser et al., 2021); the exact choice of the operators used in this study is described in section 4.2.

Note that, considering resolved and sub-grid flows to the momentum equation would result in complementary SGS forcing terms in the momentum equation as well. However in this study, we focus our attention only on the SGS thickness fluxes , as these correspond to one of the major parameterizations - the GM parameterization - in ocean models. Data-drive parameterizations of SGS momentum forcing were considered recently

(Perezhogin et al., 2024; Zhang et al., 2023; Zanna & Bolton, 2020), and in follow-up work we plan to target both thickness and momentum SGS parameterizations simultaneously.

3 Machine learning model design and implementation

The goal of this work is to develop a data-driven parameterization for the SGS thickness fluxes in terms of the resolved variables. Here, we use a multi-layer perceptron (MLP) with a few hidden layers to approximate this functional relationship. These MLPs will be trained by using data from high-resolution (HR) simulations, which have been filtered and coarse-grained to diagnose the SGS thickness fluxes and the corresponding resolvable fields. The machine learning parameterizations will be tested in both offline and online evaluation settings.

3.1 Parameterization function design

MLPs are universal function approximators, and can represent a vast space of functions. While this flexibility is powerful, it also increases the risk of overfitting, allowing the MLP to approximate data using functions that do not generalize. To avoid overfitting and allow for a degree of generalizability, we implement certain design constraints into the MLP.

Input features: Here we will search for functions of the form,

$$\mathbf{F}_n = f_\theta(\nabla\overline{\mathbf{u}}_n, \nabla\overline{h}_n, \triangle), \tag{4}$$

where $\nabla\overline{\mathbf{u}}_n$ is the velocity gradient tensor and $\nabla\overline{h}_n$ is the thickness gradient, both for the resolved fields. $\triangle$ is a measure of the grid scale, which allows the parameterization to be scale-aware. $f_\theta(\cdot)$ is a MLP function with unknown parameters θ, which need to be learned, that represents the two components of the SGS flux vector. Note that in the offline setting the inputs to this function will be the filtered and coarse-grained fields, while in the online setting the inputs will be resolved fields from the coarse-resolution simulation.

We found that functions of the above form can be trained to get remarkable offline success, but struggle when testing offline on data with distributional shifts (not shown). Adding some additional constraints discussed below, allows the model to generalize better to many more scenarios offline (Beucler et al., 2021).

Lateral non-locality: The GM parameterization and the velocity gradient model (VGM) parameterization, a common model used in the LES literature, are horizontally local (see Appendix C and Appendix D), i.e. parameterization output depends only on inputs from the same horizontal grid box (i, j). Here, we relax this and allow for a small degree of non-locality in the horizontal, considering input information from regions surrounding the point where the prediction needs to be made. This is mathematically denoted as $\mathbf{F}_{n,(i,j)} = f_\theta(\nabla\overline{\mathbf{u}}_{n,(I,J)}, \nabla\overline{h}_{n,(I,J)}, \triangle_{i,j})$, where $I = i + p$ and $J = j + q$, and p, q are integers in the range $(-m, m)$ – thus I, J correspond to the wider stencil around i, j. For a purely local model (1×1 stencil) $m = 0$, for a model with a 3×3 stencil $m = 1$, for a model with a $5{\times}5$ stencil $m = 2$ and so on. Note, that for all input stencil sizes, the prediction is always made only at the central point (i, j). In principle, vertically non-local models can also be formulated, but these will not be considered here.

Flow dependent coordinates: We do not expect the sub-grid impacts to be coordinate dependent. However, when learning from data that comes from limited setups, a data-driven model may erroneously learn details about the coordinate. For example, if learning from data that comes from a f-plane channel simulation oriented in the x-direction, a data-driven model has the potential to learn that the SGS flux directed in the y-direction results in an available potential energy (APE) reduction. This model may fail if the setup is rotated by 90 degrees - even though we expect the impacts to not have changed.

To ward against this issue, we work in a flow-dependent, rather than a coordinate dependent, frame (Prakash et al., 2022). In particular, we rotate all our variables into a frame of reference oriented with the thickness gradients in each layer (see Appendix E for details). We denote variables in the flow dependent frame as $\widetilde{(\cdot)}$. When working with laterally non-local inputs, the rotation is done with respect to the thickness gradient at the center point of the stencil. Also in this frame of reference, we implicitly reduce the number of inputs by one, as only the magnitude of the thickness gradient at the center of the stencil is now needed to quantify $\nabla \overline{h}_n$. This frame of reference is also conceptually advantageous, as the projection of the SGS thickness flux in the direction of the thickness gradient is responsible for the dissipation of resolved thickness variance, which is linked to the mean APE dissipation (Loose, Bachman, et al., 2023).

Non-dimensionalization and range limitation: The physics and sub-grid parameterizations should be invariant to the units of measurement. However, ML models are not unit invariant by default, and this property needs to be built in. Here this is achieved by casting all inputs and outputs into non-dimensional forms. In particular, we use the following non-dimensional forms: $\frac{\mathbf{F}_n}{\triangle^2 |\nabla \overline{\mathbf{u}}_n||\nabla \overline{h}_n|}$, $\frac{\nabla \overline{\mathbf{u}}_n}{|\nabla \overline{\mathbf{u}}_n|}$, and $\frac{\nabla \overline{h}_n}{|\nabla \overline{h}_n|}$. Here $|\nabla \overline{\mathbf{u}}_n|$ is the Forbenius norm of the velocity gradient tensor; this is $\sqrt{\partial_x \overline{u}_n^2 + \partial_y \overline{u}_n^2 + \partial_x \overline{v}_n^2 + \partial_y \overline{v}_n^2}$ for a 1×1 stencil. For a larger stencil this takes the form
$\sqrt{\sum_{p=-m}^{p=m} \sum_{p=-m}^{p=m} (\partial_x \overline{u}_{n,(i+p,j+q)}^2 + \partial_y \overline{u}_{n,(i+p,j+q)}^2 + \partial_x \overline{v}_{n,(i+p,j+q)}^2 + \partial_y \overline{v}_{n,(i+p,j+q)}^2})$.

Using the non-dimensionalization of the inputs using the magnitudes also ensures that all normalized input variables are limited in range between -1 and 1, helping constrain the input domain of the samples. There can still be gaps inside this multi-dimensional unit-sphere that were not sampled in the input data, but range limiting is still better than having unconstrained inputs. While not explicitly apparent, normalizing by the norm reduces the degree of freedom in input variable group by one. Apart from providing unit-invariance and range limiting, non-dimensionalization can potentially also provide generalization across some regimes where the energy levels are different but the underlying dynamics are similar.

Final functional form: The above considerations, result in the following :

$$\widetilde{\mathbf{F}}_{n,(i,j)} = \triangle_{i,j}^2 |\nabla \overline{\mathbf{u}}_{n,(I,J)}||\nabla \overline{h}_{n,(I,J)}| f_\theta \left(\frac{\widetilde{\nabla \overline{\mathbf{u}}}_{n,(I,J)}}{|\nabla \overline{\mathbf{u}}_{n,(I,J)}|}, \frac{\widetilde{\nabla \overline{h}}_{n,(I,J)}}{|\nabla \overline{h}_{n,(I,J)}|} \right), \tag{5}$$

where subscripts i, j and I, J have been included to make the non-locality of the model explicitly clear. No rotation is needed for the norms of the inputs, as the norm is invariant to coordinate rotation. While the above function choice is more restrictive than equation 4, it was chosen after some trial and error and we found that the functions estimated under these constraints are skillful.

3.2 Neural network architecture, hyper-parameters, and software

As mentioned above, we use a MLP to learn the function $f_\theta(.)$. In this architecture the input layer is linked to the output layer through N_H hidden layers. Each hidden layer can have a different width (W_s, where $s \in (1, N_H)$), such that there are $\sum_{s=1}^{N_H} W_s$ hidden nodes. Each node applies a linear transformation to its inputs, followed by a nonlinear activation function, which was chosen to be ReLU. The outputs of one layer serve as inputs to the next, enabling hierarchical feature learning. The final layer produces the output through a linear transformation.

We conducted a comprehensive sensitivity study on various MLP design and training choices, as detailed in Appendix F. Among all hyperparameters tested, we found model skill to be most sensitive to the total number of trainable parameters. For a given stencil size and input/output configuration, performance improved with model size up to a threshold, beyond which additional parameters did not lead to further gains. Consequently, for discussion purposes in the results section, we restrict attention to models with approximately the minimum number of parameters required to achieve maximal skill for each stencil size. We evaluate three MLP models that differ only in stencil size: 1X1, 3X3, and 5X5. Each model uses two hidden layers with 48 nodes per layer. As stencil size increases, the number of input features and thus the number of trainable parameters increases: from 2,786 (1X1), to 5,090 (3X3), to 9,698 (5X5). The models take non-dimensionalized velocity and thickness gradients as input and predict non-dimensionalized SGS thickness fluxes as output. Both inputs and outputs are rotated into local thickness gradient coordinates. Also, in addition to the non-dimensionalization, all input and output features were also normalized by the order of magnitude of their standard deviations.

The details of the Double Gyre (DG) and Phillips 2 Layer (P2L) HR simulations are presented in the Section 4, and the details of the training set choices are the following. Training was performed on the first 2,048 snapshots from both the DG and P2L simulations, sampled every 10 days and including spin-up. For each snapshot, data from multiple filter and coarsening scales (details presented later) were used simultaneously. To give equal weight to each scale during training, data from finer filters were sub-sampled to align with the grid points of the coarsest filter. Offline evaluation was conducted on snapshots 2,4000 to 3,6000.

We used Python and JAX (`https://docs.jax.dev/`) for all our machine learning pipelines. Specifically, we used Flax-Linen library (`https://flax-linen.readthedocs.io/`) for the design of our MLPs and used the Optax library (`https://optax.readthedocs.io/`) for optimization. The loss function was the mean absolute error in the non-dimensionalized outputs. Models were trained using the Adam optimizer with a learning rate of 0.01, and training was stopped when the validation loss failed to improve by more than 0.1% over 10 consecutive epochs.

3.3 Implementation in MOM6

The MLP-based SGS thickness flux parameterization was implemented in MOM6 via two new modules: an MLP module and a thickness flux prediction module. The MLP module, as the name suggests, reads a NetCDF file containing the model architecture, trained weights, and normalization factors, and performs inference as would be done by a standard feedforward MLP. This module is general-purpose and can be called from anywhere within MOM6, enabling the integration of multiple MLP-based data-driven models into the codebase.

The thickness flux prediction module incorporates the design choices described in Section 3.1. To keep the implementation simple and in light of the limited guidance in the literature regarding appropriate numerical schemes for such models we interpolate the necessary input fields to the centers of grid cells. For a 3X3 model, this means that each input in the 3X3 stencil surrounding the prediction point is evaluated at the 3X3 grid cell centers. After predicting the two components of the thickness flux at the central point, the fluxes are then interpolated to the appropriate edge locations. We also introduced a non-dimensional coefficient (C_{ANN}) which can be used to adjust the strength of the parameterized flux if needed.

Fluxes at solid boundaries are set to zero to ensure volume conservation. We also observed that regions with very thin fluid layers could produce numerical artifacts. To suppress these, we modified the computation of the thickness gradient magnitude that multiplies the MLP prediction in Equation (5). Specifically, we replaced $|\nabla h_n|$ with $\left|h_n^2 \nabla \left(\frac{1}{h_n+\epsilon}\right)\right|$, which preserves the overall scaling in well-resolved regions but naturally drives the flux toward zero in layers thinner than ϵ. This adjustment maintains the desired magnitude in most regions while ensuring numerical stability in thin layers.

While the P2L simulation does not contain vanishing layers, the lower layer in the DG simulation can vanish, as discussed in the next section. We tested the sensitivity of the results to different values of ϵ between 1 to 20 m and found the simulation outcome to be relatively insensitive to this choice.

While, in principle, the boundary conditions described above should suffice for layered models (Killworth, 2001), MOM6 additionally enforces the constraint that the barotropic sum of the SGS thickness fluxes within a water column must be zero. This is accomplished by expressing the SGS fluxes in terms of a streamfunction (see Appendix Appendix B) and setting the streamfunction to zero at the boundaries using sophisticated tapering techniques (Ferrari et al., 2008, 2010). In our case, the situation is simpler, as we consider only two layers in the simulations evaluated in this study. Accordingly, we predict the lower-layer flux and satisfy this constraint by setting the upper-layer flux to be equal and opposite to the lower-layer flux. We also tested the parameterization without this constraint; while those simulations remained numerically stable, they frequently exhibited noisy solutions. Although enforcing a zero barotropic component has minimal impact on the overall energetics of the simulation (see appendix G4), it could play a role in lateral tracer transport. This effect is not studied here.

4 Data

4.1 Ocean Model Simulations

In this study, we work with two different idealized simulations in MOM6: Phillips 2 Layer (P2L) - a 2 layer model of Phillips baroclinic instability, described in Hallberg (2013), and Double Gyre (DG) - 2 layer wind driven double gyre, described in Zhang et al. (2023). Both these setups have the minimum ingredients needed for the development of a rich baroclinic mesoscale field, while having a very simple vertical structure (Figure 1). Some physical characteristics of the simulations are described in section 4.3 below. Simulations using both these setups were run over a range of resolutions. The highest resolution output was filtered and coarsened for generating data to train and evaluate the ML model offline, while the lower resolution simulations were used during online evaluation.

4.2 Processing of HR simulation data for ML training and evaluation

Filtering and Coarsening: To diagnose the input and output fields we processed the data using both filtering and coarse-graining. First for simplicity, we interpolated all the simulation prognostic variables onto the grid centers. The layer thickness and interface are already computed on the grid center, and u and v velocity components were linearly interpolated using the xgcm package (`https://xgcm.readthedocs.io/`). Also regions with layer thickness smaller than 20 m were masked and treated as land, which only impacted the lower layer in the DG simulation where the layer thickness vanishes at incropping locations. These centered and masked data were then filtered using a Gaussian filter, using the gcm-filters package (Loose et al., 2022). Specifically, we used the 'simple fixed factor filter' in gcm-filters (`https://gcm-filters.readthedocs.io/en/latest/examples/example_filter_types.html#simple-fixed-factor-filter`), which does an area weighting. In the P2L simulation, with a Cartesian grid of size $\triangle_g = 4$ km, we used filter scales $L_f = 48, 100, 200$ and 400 km, where fixed filter factors of 12, 25, 50 and 100 are used. In the DG simulation, with a non-uniform grid of size $\triangle_g = 1/20^o$ this leads to variable filter scales of sizes $L_f = 0.55^o$, 1.10^o, 2.20^o and 4.40^o, which are approximately equal to the scales used for filtering P2L.

After all the filtered variables and the corresponding SGS fluxes were computed, the data was further coarsened using box averages. This step only leads to a data reduction, and for convenience we do not account for the SGS fluxes resulting from this operation. This is sufficient since the fields were already filtered, and the additional fluxes resulting from this coarsening operation are found to be very small (not shown). We decided to choose the ratio between the filtering scale and the coarsening scale, to be 5. For example, a filter scale $L_f = 200$ km was combined with a coarsening scale of $\triangle_c =$ 40 km. This choice of the filter to grid ratio (FGR=$L_f/\triangle_c$) is based partially on considering the spectra of ocean models, which often show a damped energy level or numerical artifacts emerging at scales larger than the grid scale - often upto 5 times in size (e.g. notice the small-scale bump in the EKE spectrum in Figure 7). This is a heuristic choice, and can be explored in more detail in future work. For the rest of this study we indicated the four filter scales nominally with L_f =50, 100, 200, and 400 km, and corresponding coarse grid scales nominally with $\triangle_c =$ 10, 20, 40 and 80 km. However this is only for the convenience of presentation, and in the computations, where $\triangle_c$ is needed (e.g. equation 5), the actual coarse grid scales were used.

Currently there is no prescribed way to process data from a HR simulation to make it match a LR simulation in some objective sense. Hence, our approach to diagnosing filtered and coarsened data is relatively adhoc, and based on pragmatism. We believe and hope that the impact of these choices would likely be minimal, with the acknowledgement that a big shortcomings of these ML models in online settings would likely arise from the fact that the distribution of a low resolution simulation would be shifted relative to a simplistically filtered version of a high resolution simulation, and some degree of tuning may be required to address this. Also, in future work more care can be taken for dealing with staggered grids, precisely accounting for the separate contributions from filtering and coarsening operations, and for handling boundaries differently when the filters do not commute with the gradients.

Layer thickness decomposition: When computing the thickness fluxes, special care was taken when the bottom topography (η_b) was not flat (in the DG case). We did not want to filter the bottom topography when filtering thickness, since the topography present in a coarse model is not a filtered version of the high resolution topography. So we chose to use the condition that $\overline{\eta_b} = \eta_b$. Thus, thickness fluxes were dealt with by filtering interface heights only. To be more precise, in the bottom layer the filtered thickness would be $\overline{h}_N = \overline{\eta}_{N-1/2} - \eta_b$ and the filtered advection would be $\overline{\mathbf{u}h}_N = \overline{\mathbf{u}\eta}_{N-1/2} - \overline{\mathbf{u}}\eta_b$, where $\eta_{N-1/2}$ is the upper interface height of the bottom layer N.

Further, we decomposed thickness gradients into a steady and a deformable parts ($\nabla\overline{h}_n = \overline{h}_N^{deformable} + \nabla\overline{h}_n^{steady}$). When applied to a layer where the lower interface is topography, (e.g. $\overline{h}_N = \overline{\eta}_{N-1/2} - \eta_b$) we get $\nabla\overline{h}_N^{steady} = -\nabla\eta_b$ and $\nabla\overline{h}_N^{deformable} = \nabla\overline{\eta}_{N-1/2}$. In other layers, the steady contribution is zero ($\nabla\overline{h}_n^{steady} = 0$) and the full layer thickness contributes to the deformable part ($\nabla\overline{h}_n^{deformable} = \nabla\overline{\eta}_{n-1/2} - \nabla\overline{\eta}_{n+1/2} = \nabla\overline{h}_n$). This decomposition allows us to distinguish the impact of dynamic layer thickness variations and bottom topography on the SGS fluxes. In this study, we only use the deformable contribution as inputs to our MLP, and henceforth replace the notation for the deformable part $\nabla\overline{h}_N^{deformable}$ by $\nabla\overline{h}_N$ for simplicity in most places, unless otherwise noted.

4.3 Physical characteristics of the simulations and the filtered data

4.3.1 Data distributions

Both the P2L and DG simulations produce a turbulent flow field, with a rich array of eddies and jets (Figure 1). The magnitude of the SGS fluxes are greater in the top layer than the bottom layer, and are larger in P2L than DG (Figure 2). Also the SGS fluxes vary by three to four orders of magnitude in each layer, and the magnitude of these fluxes increase by about an order of magnitude from filter scales of 50 km to 400 km (coarse grid scales of 10 km to 80 km). In contrast, the non-dimensionalized fluxes have a much narrower distribution, with little variation of the median across filter scales and layers – indicating that the non-dimensionalization choices made here are quite successful at collapsing the data distribution and may help with generalization. However, note that the width of non-dimensionalized flux distribution does increases slightly with filter scale, indicating that there may still be some room for further improved non-dimensionalization factors in the future, which may be able to address this scale dependence.

Parameterizations in Z-level models and MOM6 impose that the eddy driven stream function takes a boundary condition of the zero at the surface, which is equivalent to saying that there is no barotropic (depth integrated) SGS flux (see Appendix B). This condition is not naturally satisfied by the diagnosed data (red distributions in Figure 2), and is not even expected based on the eddy-mean decomposition of the thickness equation in layered models (Killworth, 2001). In the diagnosed data, we notice that the fluxes in the two layer have a very slight opposing tendency, which slightly increases with larger filter scales. Only in the long time average, and if the mean flow is weak, would we expect the vertical sum of the eddy thickness fluxes to go to zero based on volume conservation. However, it is worth noting that even though the depth integrated SGS thickness flux is far from negligible, its impact on the APE tendency, which is the primary target of our parameterization, is very small (see Appendix G4).

The distributions of other model fields, particularly those relevant for our neural network design, are shown in Figure 3. The velocity gradients have a large range across scales, and their magnitude decreases with increasing filter scale. Generally, the upper layer has stronger velocity gradients than lower layer, and the P2L simulation has stronger velocity gradients than the DG simulation. In contrast to velocity gradients, the deformable thickness gradients vary less with filter scale. Additionally, since the surface variations are much weaker than the interface variations (Figure 3c), the deformable thickness gradients are essentially the same for the two layers. Also, the deformable thickness gradients are slightly weaker in the DG simulation relative to the P2L simulation. The bottom topography slopes in the DG simulation are very strong relative to the interface variations, which was one reason for us to decompose the thickness gradients into its steady and deformable contributions.

4.3.2 Bulk properties

The rich turbulent eddies impact the mean or large spatial and temporal scale flow, and the target of traditional parameterizations has been to skillfully model some of these effects. Here we describe what some of these feedback are when the eddies are resolved.

In the P2L simulation the mean state is a zonal jet, which is sustained by slowly relaxing the middle interface to a sloping state and thus the relaxation works as a source of APE (Hallberg, 2013). The eddies in this simulation work to flatten this interface, thus removing the APE that is input by the relaxation forcing. To achieve this the eddies flux volume to the north in the upper layer and to the south in the lower layer, generating an eddy driven overturning circulation. This overturning is sustained because the relaxation also leads to a diapycnal transformation of watermasses from one layer to the other. When the eddies are not resolved or only partially resolved, this eddy-driven overturning circulation weakens and a parameterization is needed to ensure that the appropriate levels of APE are removed and the overturning circulation is maintained (Hallberg, 2013).

In the DG simulation (Zhang et al., 2023), the mean state is maintained by forcing with a steady wind stress. The wind stress peaks at the intermediate latitude (40N) and generates a region of Ekman downwelling to the south and Ekman upwelling to the north, which pushes and pulls the middle interface ("thermocline") to create an interface slope. The sloping interface has APE that is siphoned out by the mesoscale eddies, which work to reduce the APE and thus counteract the impact of the wind.

5 Results

Data-driven models may be evaluated along two interconnected aspects: offline skill and online skill. Offline skill refers to the accuracy in predicting SGS thickness fluxes for a given set of inputs, with the reference or "truth" diagnosed from high-resolution simulations. In contrast, online skill assesses the ability of a data-driven parameterization to improve the fidelity of a lower-resolution simulation, where the truth is defined in terms of large-scale behavior - either from high-resolution models or observations.

Here, we first evaluate the offline skill of the machine learning model on filtered and coarsened data. We then turn to the arguably more important question: how well does the model perform online, when it is embedded within a simulation and actively interacts with and modifies the resolved state?

5.1 Offline Evaluation

Here we discuss three particular MLP models that differ only in stencil size: 1X1, 3X3, and 5X5 (details in section 3.2). We also contrast these models against the GM parameterization and the VGM parameterization, where the free parameters in these conventional schemes were estimated using least-squares fitting to the SGS thickness fluxes for each parameterization separately at each filter scale and for each simulation setup. This is in contrast to the MLPs, which were trained across the entire range of filter scales and simulations simultaneously. Given the large variation in the estimated parameters for the conventional parameterizations, their performance would be worse if they were trained across the entire range of data simultaneously.

Point-wise skill: Mostly, all the MLP models demonstrate relatively high skill in predicting the SGS thickness fluxes point-wise. As an example, the true and predicted SGS fluxes from the 3X3 model at a filter scale of 100 km are shown in Figure 4. The MLP model does extremely well, and produces the right patterns and magnitudes of both the along and across thickness gradient components of the SGS flux in both layers. Note in the figure that the error is so small that it had to be multiplied by 5 to bring it to the same color scale as the SGS flux.

In a more quantitative sense, we find that the MLP skill, compared across all MLP configurations considered here and quantified using R2 or correlation score (Appendix A), depends on the input stencil size and the filter scale (Figure 5); the skill also depends on other factors as discussed in Appendix G but those sensitivities can be alleviated with enough data or free parameters. The ML model skill increases with stencil size, with a large improvement in going from 1X1 to 3X3 and a smaller improvement when going further to 5X5. Also, model skill decays as filter scales get larger, which is more rapid in the case of DG relative to P2L. We think that this might have to do with the different deformation radii, which is on average 40 km in P2L and 20 km in DG (Figure H1), and ML models may be less skillful as the filter scale gets much larger than the deformation radius or larger than the largest resolved eddies.

In figure 5, we also contrasted the skill of the MLPs against the conventional parameterizations. The VGM parameterization, which is sometimes used in LES studies, has essentially the same point-wise skill as our 1X1 ML model. The GM parameterization in contrast has very little point-wise skill (both R2 and correlation are usually less than 0.2), which only marginally increases at larger filter scales. This is to be expected, as the GM parameterization is a bulk model and not designed to have skill in producing the right local structural patterns in the SGS fluxes. Notice that in contrast, the skill of the 3X3 and 5X5 models is significantly higher than both the 1X1 ML model and the VGM parameterization.

Bulk (time-averaged) skill: While pointwise skill is a useful metric, we are often (also) interested in ensuring that our parameterizations produce the appropriate bulk effects. One way to quantify the bulk effects is in terms of time averages. In Figure 6 we consider the skill over different temporal averaging windows, here the skill is quantified for both layers and flux components.

We expect that the GM parameterization skill may improve in this bulk sense as the time averaging duration is increased, and as expected this is found to be the case. In Figure 6 last column, this effect is seen quite clearly for the P2L data but not for the DG data. Even for the P2L case the correlation skill only rises to 0.5, rather than 1, which is because GM only predicts the along gradient fluxes, and so when averaging skill over along and across gradient fluxes, we can only achieve a maximum of 0.5. The reason for the discrepancy between DG and P2L arises because GM did not turn out to be a good model for upper layer fluxes in the DG case (even when quantified just in terms of correlation), which might be a result of mean flows and inhomogeneity in the turbulent statistics. The skill of the GM model on the lower layer along gradient fluxes in the DG data is higher, but still not as large as for P2L data (not shown).

The 1X1 MLP, and similarly the VGM parameterization (Khani & Dawson, 2023), have R2 skill decrease with increasing temporal averaging. In contrast, the correlation is not impacted, suggesting that the decrease in skill has less to do with the functional form and more to do with the parameterization coefficient or amplitude. In contrast the 3X3 and 5X5 MLPs have almost no impact on either R2 or correlation skill with averaging, if anything there is a very slight increase in skill at longer temporal averaging. The fact that the skill score is usually close to 1, except for larger filter sales in DG, also shows that these models do very well in predicting both the along and across fluxes.

Overall, we found that the offline skill of the MLP models, particularly those with wider stencils, is very promising. Also, we show in Appendix Figure G4 that a MLP trained on data from one simulation shows relatively high skill when tested on the unseen simulation. Thus, these new models are scale and context aware, and no retraining or tuning is needed when testing over different datasets. This is contrast to the traditional models, which are unable to match the MLP skill even after the corresponding coefficients were estimated separately for each scale and dataset. Thus, we are compelled to evaluate the performance of these MLPs in an online setting.

In the online setting we only evaluate the skill of the 3X3 MLP discussed above, as this model provides a good compromise between computational cost and offline skill (skill of the 5X5 MLP is only marginally better than the 3X3 MLP). Also since the offline performance of the 1X1 model (and VGM) is worse and degrades in a bulk sense, we chose to not evaluate it either.

5.2 Online Evaluation - Phillips 2 Layer

We first test the MLP in the P2L simulation, which is the simpler of the two simulation setups considered in this study. To assess the sensitivity of this setup to model resolution and parameterizations, we performed a suite of simulations (see Appendix H).

In this section, we focus on the 20 km simulations – both parameterized and unparameterized – and compare them with a 4 km HR simulation. The deformation radii in this setup range from 25 to 50 km, placing the 20 km grid in the “gray zone” where mesoscale eddies are only partially resolved, whereas the 4 km grid resolves them more fully. The parameterized simulations were selected to approximately match the total overturning transport of the HR case. For the GM parameterization, we include two configurations: one with low diffusivity (1000 m^2/s), often cited as a canonical value in the literature, and another with high diffusivity (8000 m^2/s), chosen because it yields an overturning close to the HR simulation. These values are not fine-tuned to exactly match the HR overturning, but instead represent two qualitatively distinct regimes. Notably, 1000 m^2/s also marks a threshold beyond which the GM scheme begins to strongly damp the resolved eddy field (also see Appendix H). In contrast, the MLP-based parameterization with a tuning coefficient (C_{ANN}) of 1 produced overturning transport closely aligned with the HR simulation. All simulations reached a spun-up state within 2 years, and time-averaged statistics are computed over years 2 to 10.

Snapshots of upper layer relative vorticity and the EKE spectrum averaged temporally and meridionally in all simulations evaluated in this section are shown in Figure 7. The variability in the 20 km simulations without any parameterization and with the MLP is very similar, while the addition of the GM parameterization leads to a substantial reduction in flow variability. The low GM diffusivity simulation permits some eddies, while the eddies are entirely suppressed in the high GM diffusivity simulation. Also, the HR filtered and coarsened simulation state matches the low-resolution unparameterized simulation at large scales, but has lower energy levels at smaller scales. This is the result of the specific properties of the Gaussian filter that was chosen to filter the simulation, and more refined filters could definitely be employed if needed.

The peak overturning transport in the upper layer is shown in Figure 8a (the lower layer's overturning is identical but with the opposite sign). The HR simulation produces approximately 13 Sv of transport, with 9.5 Sv attributed to scales larger than the filter scales and 3.5 Sv from SGS fluxes. In contrast, the unparameterized 20 km simulation produces only about 10.5 Sv of transport. The 20 km simulation with the MLP sustains about 12.5 Sv of transport, reducing the resolved component to 9.5 Sv while adding around 3 Sv from parameterized fluxes. The low GM diffusivity case similarly reduces the resolved transport marginally, but is unable to generate enough parameterized flux to match the total overturning transport of the high-resolution simulation. As the GM diffusivity is increased, the resolved transport is drastically reduced, with only a slight increase in the total transport. When the GM diffusivity becomes large enough, the total transport can match that of the high-resolution simulation, but at the expense of completely eliminating the resolved contribution. Note that we can further fine tune the MLP and the high GM diffusivity case to exactly match the HR simulation. However, this is not possible for the low GM diffusivity case because the resolved transport drops much more rapidly than the rate at which the parameterized transport increases with changing the coefficient (can also see Figure 7 in Hallberg (2013)).

While the parameterizations were tuned to approximately match the overturning, here we focus on contrasting their impact on other relevant metrics (details of these metrics are described in Appendix G). The kinetic energy (KE) and available potential energy (APE) from various contributions are shown in Figures 8b and c. The unparameterized 20 km simulation exhibits lower KE and APE than the HR simulation, as expected, and the LR simulation KE and APE are close to the filtered KE and APE from the filtered HR data. The 20 km simulation with the MLP results in a slight reduction in the EKE and EAPE, but the overall KE and APE is roughly in line with the KE and APE from the filtered HR simulation. The low GM diffusivity simulation shows lower KE and APE than both the unparameterized and MLP-based simulations, while the high GM diffusivity case has no EKE or EAPE.Since in this simulation setup the interface height is restored to a prescribed state, the MKE and MAPE remain nearly unchanged across the different setups. Only in the high GM diffusivity case does the parameterization forcing become large enough to cause a very small reduction in MKE and MAPE, and the EKE and EAPE are completely wiped out.

The tendency of the APE arising from the SGS or parameterized fluxes is shown in Figure 8d. We display the contribution to both the MAPE tendency and the EAPE tendency. In the filtered HR simulation, the SGS fluxes contribute just over half of their tendency towards reducing the MAPE. The 20 km simulation with the MLP parameterization produces an APE tendency that is relatively close to the HR filtered case, with a slightly larger impact on the MAPE. In contrast, the low GM diffusivity case leads to a disproportionately large impact on the EAPE, without sufficiently reducing the MAPE. The high GM diffusivity case produces a similar MAPE tendency to the low diffusivity case but has no impact on the EAPE, as no eddies remain to be damped out.

While the APE tendency analysis is illustrative, it has some limitations due to the changes in the simulation state across different cases. To further emphasize this point, we also evaluated the APE tendency that would arise from the MLP in the 20 km simulation, where the MLP was not actually coupled to the resolved state of the simulation (see bar labeled 20 km in Figure 8d). In this scenario, the APE tendency is much larger than in the filtered case, likely because this unparameterized simulation has a higher eddy kinetic energy (EKE) than the filtered EKE. The non-linear interaction between the parameterized and resolved flow makes it difficult to predict a priori how the system will respond to the parameterization. This non-linearity is also evident in the sensitivity study plots shown in Appendix H, where the response to the parameterization coefficients is non-monotonic. Note that at eddy-permitting resolutions, even the response to the GM parameterization is non-trivial and non-monotonic.

In summary, while both the MLP and GM parameterizations can be tuned to produce approximately the correct overturning circulation in the P2L simulation, only the MLP is able to achieve this without significantly damaging the resolved flow and eddies. In contrast, as also shown by (Hallberg, 2013), the GM parameterization excessively dissipates the eddies.

5.3 Online Evaluation - Double Gyre

Next, we test the MLP parameterization in the DG simulation, a canonical system for studying wind-driven gyre dynamics. In contrast to the P2L setup discussed earlier, the DG exhibits strong boundary currents and pronounced spatial inhomogeneity in eddy statistics. Additionally, unlike the P2L system, the mean state here is not maintained through relaxation, but is rather a result of balance between winds and eddies. These differences lead to two key consequences that are different from P2L: (i) mean-state biases can emerge as resolution and parameterizations are varied, and (ii) the system is purely adiabatic, with no overturning circulation.

Similar to the P2L case, we conducted a suite of simulations to assess the sensitivity of the DG setup to both resolution and parameterization coefficients (see Appendix H). Here, we focus on the $1/5°$ ($\sim$20 km) simulations, which only marginally resolve the deformation radius - ranging between 5 and 30 km in this configuration. In addition to the unparameterized baseline, we analyze simulations that employ MLP and GM parameterizations. The coefficients for these parameterized runs were selected to minimize the mean state error in sea surface height (SSH), which is strongly correlated with thermocline depth in this system.

The mean sea surface height (SSH) and kinetic energy (KE) fields, averaged over years 3-13 of the simulations, are shown in Figure 9. In the $1/5°$ simulation without any parameterization, a standing eddy forms just downstream of the boundary current separation point - a region that also exhibits stronger flow than in the high-resolution reference simulation. This eddy feature vanishes, and the mean state bias is reduced, in both the MLP- and GM-parameterized simulations, since the parameterizations coefficients were explicitly chosen to reduce this bias. However, consistent with the P2L results, the

MLP achieves this correction without the substantial loss of KE observed in the GM simulation.

This result is quantified in the KE and APE metrics shown in Figure 10. The mean state of the $1/5°$ simulation without parameterization is overly energetic compared to the HR filtered simulation, as evident in both the MKE and MAPE. While both parameterizations reduce this excess energy, the GM parameterization does so at the cost of a much larger reduction in EKE and EAPE compared to the MLP parameterization.

To further assess the role of the parameterizations in influencing the mean flow and eddies, we examine the APE tendency induced by SGS fluxes – both in the filtered HR simulation and in the parameterized fluxes of the coarser simulations (Figure 11). The APE tendencies affecting both the MAPE and EAPE exhibit qualitative similarities between the filtered HR and the MLP parameterized simulations, generally acting to reduce APE. However, this impact is not spatially or temporally uniform and in many instances there is even APE gain, resulting in localized regions of APE gain even in the 10-year mean shown here. In contrast, the GM parameterization acts as a sign-definite sink of APE, producing a much stronger and more widespread reduction in both MAPE and EAPE compared to the MLP parameterization or the diagnosed tendencies from the filtered HR simulation.

In summary, as in the P2L case, the MLP parameterization outperforms the GM parameterization in the DG simulation. It effectively reduces the mean state bias without causing an excessive suppression of eddy energy.

6 Discussion and Conclusions

In this work we developed and implemented a data-driven parameterization for subgrid scale (SGS) thickness fluxes produced by mesoscale eddies. This was achieved by training a relatively small multi-layer perceptron (MLP) to learn a functional relationship between the gradients of the large-scale/resolved fields and the SGS fluxes, using data from high resolution (HR) simulations. By introducing features like lateral non-locality, coordinate invariance, and non-dimensionalization into the MLP design, we were able to produce a more generalizable and stable data-driven parameterization (Perezhogin et al., submitted). Of these features, the lateral non-locality and the non-dimensionalization, particularly the aspect that produces range-limited inputs, were found to be the most important design choices. The trained models have very high offline skill (Figure 5), even when testing on data coming from unseen simulations (Figure G4). The skill relatively degrades at scales larger than the largest eddies, or as we transition from eddy permitting to non-eddying resolutions, but even at these scales the offline skill is comparable or higher than traditional approaches like an appropriately tuned Gent-McWillims (GM) parameterization (Figure 6).

This new data-driven parameterization was implemented into GFDL's Modular Ocean Model 6 (MOM6) and tested in two idealized simulation setups: Phillps 2 Layer (P2L) and Double Gyre (DG), where baroclinic mesoscale eddies play a first order role in the dynamics. In both these setups the MLP enhanced the simulation performance at coarse resolutions (grid scales on the order of the deformation radius or coarser), reducing biases in aspects like the meridional overturning transport and mean state. In the MOM6 implementation, we also introduced a non-dimensional tuning parameter that controls the global amplitude of the parameterization. Sensitivity studies showed that O(1) values of this parameter, values of 1 for P2L and 0.5 for DG, were optimal in online setting across all eddy permitting resolutions. At coarser resolutions the values needed to be slightly adjusted, a value of 2 for P2L and 0.75 for DG seemed optimal. In contrast, the GM diffusivity had to be adjusted for every resolution and setup individually, with optimal values of the diffusivity ranging between three orders of magnitude (10^2 to 10^4 m^2/s) across resolutions and setups.

The development of the GM parameterization more than three decades ago provided a step change in the quality and fidelity of ocean simulations. However, it has been clear since the beginning that while GM parameterization is phenomenologically appropriate, reduction of APE and adiabatic conservation of watermass volume are appropriate bulk expectations for the effects of mesoscale eddies, it has structural shortcomings. Much of the work over the past decades has gone towards improved parameter estimation (Visbeck et al., 1997), but progress towards reducing the structural errors has been very limited (R. D. Smith & Gent, 2004). Our work attempts to address this gap, and provide a data-driven model that seems to have lower structural errors and provides a path towards improving the GM parameterization along a new axis.

One of the well known drawbacks of using the GM parameterization is its tendency to dissipate the resolved eddies, reducing the effective resolution of a simulation (Hallberg, 2013). This has led to development of adhoc fixes, like turning off the GM parameterization using a resolution function, or not using the GM parameterization at all in simulations that run in the gray zone resolution (Adcroft et al., 2019). Our parameterization does not suffer from this problem, and is able to improve properties of the large scale without being overly detrimental towards the resolved eddies. One major advantage of the GM parameterization is that it is essentially guaranteed to be stable, which can not be claimed unequivocally for our data-driven parameterization. All the simulations we tested were stable, but this may not hold to be true if the tuning coefficient is pushed to larger values.

Our data-driven parameterization has produced very promising results, but a few aspects can likely be improved. Firstly, as is common with machine learning models, there is a spectral bias in the predictions. This means that the MLP's offline performance degrades at scales where the signal variance is low (Figure G3 and G4, and also implicit in Figure 4). This did not end up being an issue for model stability in our online tests, likely because the thickness variance has a tendency to cascade down-scale and the errors get cascaded to model dissipation scales. Improved loss functions or experimenting with smoother non-linearities could help alleviate this issue. Secondly, it came almost as a surprise that the model performed so well across resolutions and setups with such limited class of inputs. We think that at eddy permitting resolutions this is the case because the structure of largest eddies that are resolved already contain a lot of information about dynamically important environmental conditions, which shape the mesoscale eddy field. Reduction in offline skill at scales larger than the size of the largest eddies and change in tuning parameters at non-eddying resolutions led us to this hypothesis. This suggests that in future work it may be worth paying more attention to these scales, and potentially introducing more model inputs or vertical non-locality to be able to perform well over a larger range of scales without much tuning. Thirdly, our parameterization mainly acts on reducing the APE in the system, and is not designed to improve the MKE or EKE of the resolved state that may arise if the removed APE was appropriately cascade upscale in an inverse KE cascade. Coupling our parameterization with a backscatter parameterization in an energetically consistent manner could be investigated to produce further improvements in ocean models. Lastly, we chose to use a reasonably small MLP to keep the computational burden due to the parameterization low in principle. However, we have not performed a comprehensive testing and optimization to get the best possible computational performance for our model in MOM6, yet.

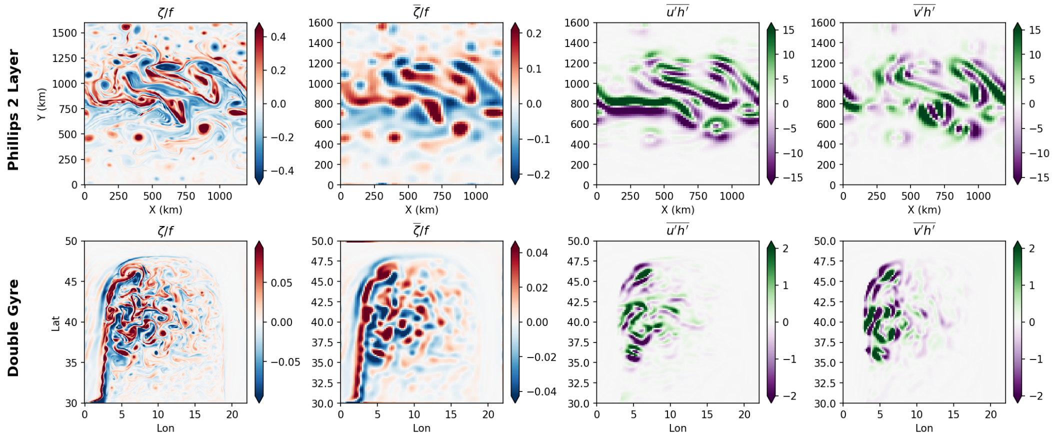

Figure 1. Snapshots exemplifying the high-resolution vorticity (left), filtered vorticity (second column), and sub-grid fluxes (last two columns) from the Phillips 2 layer (top) and Double Gyre (bottom). The filtered fields are shown from the case of coarse-graining scale of 20 km. Only the upper layer data is shown, and different panels have different color ranges.

Open Research

All the code to generate the simulation data and machine learning data, do the machine learning training, and do the analysis and figure generation can be found at `https://github.com/dhruvbalwada/mesoscale_buoyancy_param_ML`.

Acknowledgments

This project is supported by Schmidt Sciences, LLC. DB also wants to acknowledges Ryan Abernathey, now at EarthMover PBC, for providing him advice and guidance, and helping him develop his research interests in the area of mesoscale parameterizations for numerical models. This work and DB's career likely would not have shaped out the way they did without Ryan's constant support, encouragement, and enthusiasm.

Appendix A Offline Skill Metrics

To evaluate the skill of our ML model in an offline setting we use two main skill metrics. The first is the coefficient of determination or R2 skill, defined as,

$$R^2 = 1 - \frac{\overline{(y-\hat{y})^2}}{\overline{(y-\overline{y})^2}}, \tag{A1}$$

y is the truth value and $\hat{y}$ is the prediction. The $\overline{(\cdot)}$ corresponds to average over all samples being considered, which is chosen to be over both flux components from both layers, full spatial domain, and all temporal snapshots in the test data (unless indicated otherwise). The R2 skill will be 1 when prediction is perfect and reduces as prediction gets worse.

The second metric is the Pearson correlation coefficient,

$$C = \frac{\overline{(y-\overline{y})(\hat{y}-\overline{\hat{y}})}}{\sqrt{\overline{(y-\overline{y})^2}\,\overline{(\hat{y}-\overline{\hat{y}})^2}}}, \tag{A2}$$

which is 1 when the truth and the prediction are perfectly correlated, -1 for inverse correlation, and 0 for no correlation.

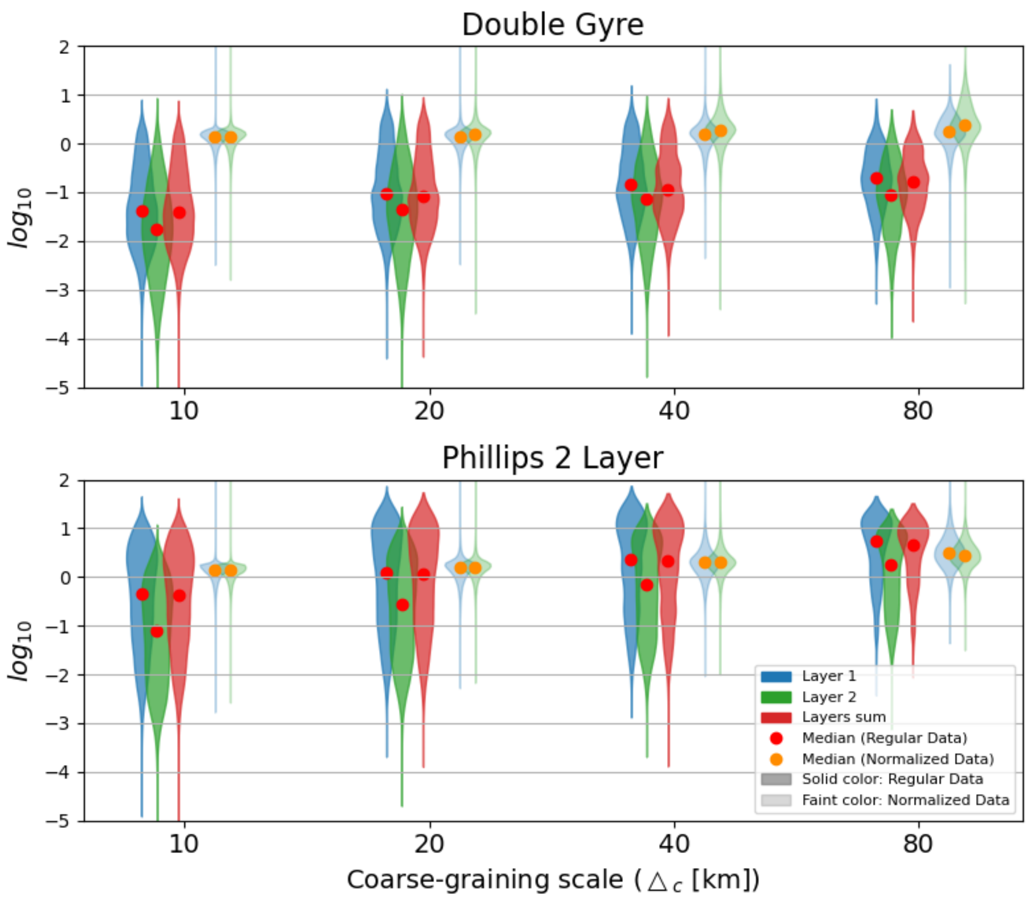

Figure 2. Distribution of output data: Distributions of the logarithm of the thickness flux magnitudes ($|\mathbf{F}_n|$) in both layers and their sum (barotropic contribution), and the normalized thickness flux ($\frac{|\mathbf{F}_n|}{\triangle_c^2|\nabla\bar{\mathbf{u}}_n||\nabla\bar{h}_n|}$) in both layers for the different experiments (top and bottom panel) and different coarse-graining scales (indicated on the x-axis). The legend in the lower panel is used to indicate the different elements in both the panels. The density of the distribution is indicated by the width of each patch, with wider regions, usually in the middle, indicating a higher concentration of data points near those values.

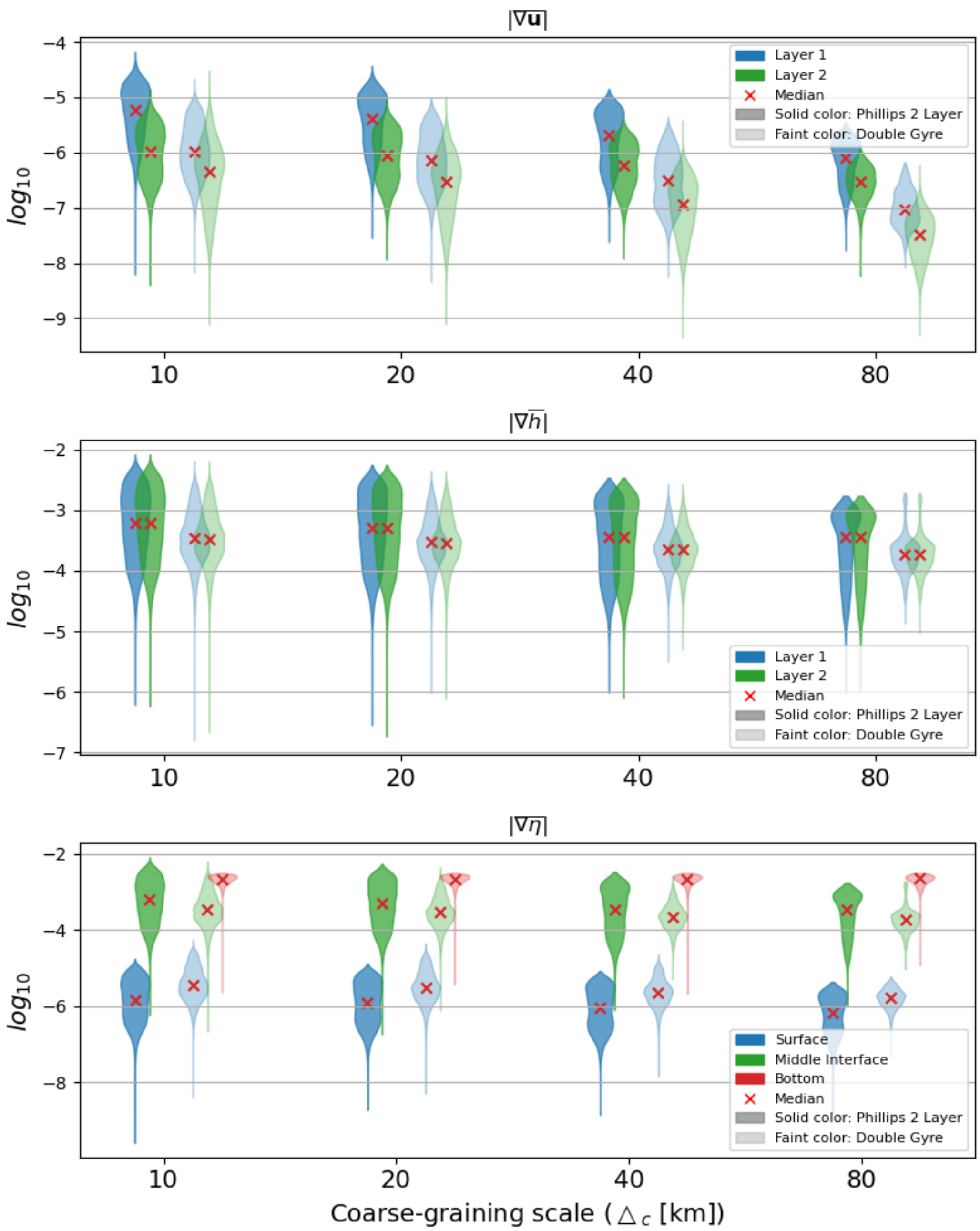

Figure 3. Distribution of input data: Distributions showing the logarithmic magnitude range of different filtered fields, which may be used as input variables for neural network design from both simulations and at different layers and interfaces. Similar to Figure 2 the width of the patch corresponds to values with a higher probability of occurence.

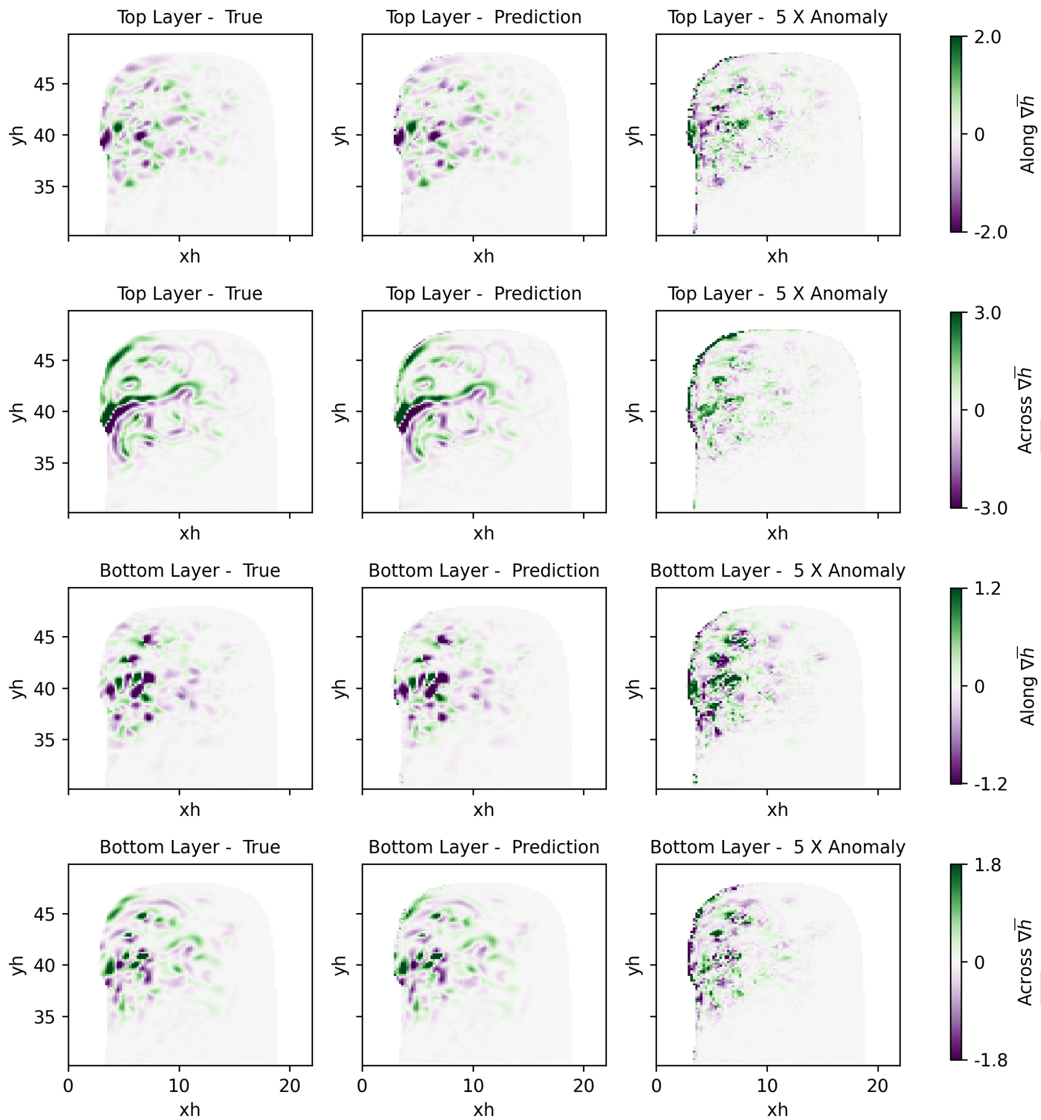

Figure 4. The true (1st column), predicted (2nd column), and prediction error (3rd column) in the along (1st and 3rd row) and across (2nd and 4th row) thickness gradient fluxes for the top (1st and 2nd row) and bottom (3rd and 4th row) layers of the Double Gyre simulations. Here the results are shown for the ML model of the following configuration: 3X3 stencil, trained on data from DG+P2L, and 5090 learnable parameters. These offline skill results are shown for the filter scale of 100km. Note that the prediction error has been multiplied by a factor of 5 to be easily visible on the same scale as the true and predicted fluxes.

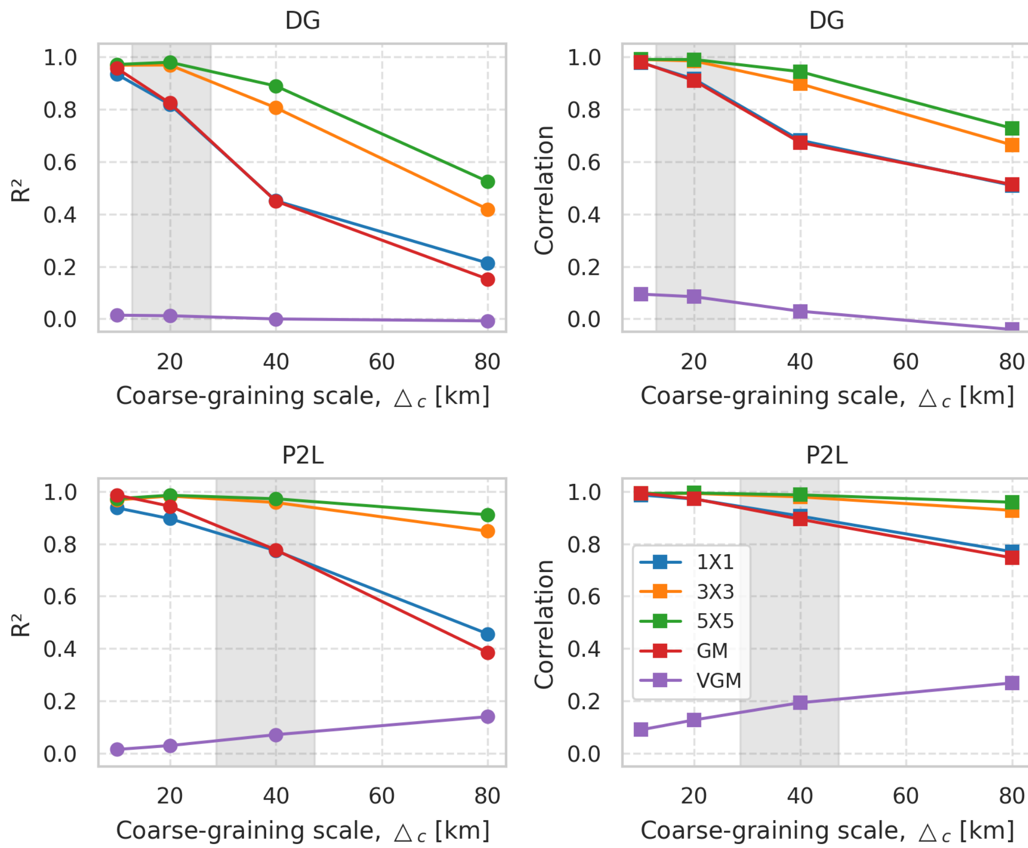

Figure 5. Pointwise offline skill in terms of the R2 skillscore (left column) and correlation (right column) for 5 different models (indicated in the legend) at different scales (x-axis) and in the double gyre (top) and Phillips 2 layer (bottom) simulations. The metrics were evaluated over both layers and across the full test dataset spanning X years. The gray shaded area indicates the deformation radius range (10 to 90^{th} percentile) in the simulation.

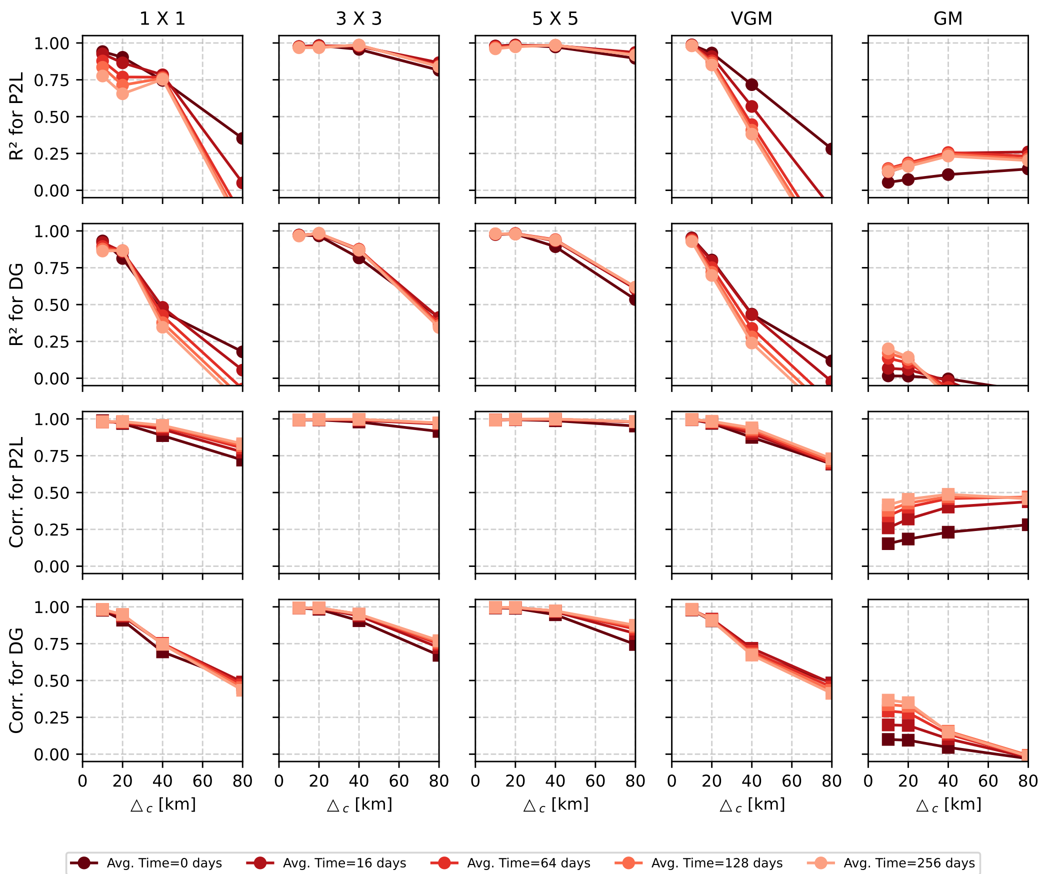

Figure 6. Offline skill in predicting the time average of the subgrid fluxes, using five different models (columns). For all models the skill in predicting the time averaged flux is quantified as the average skill over the two layers and for both the along and across thickness gradient direction.

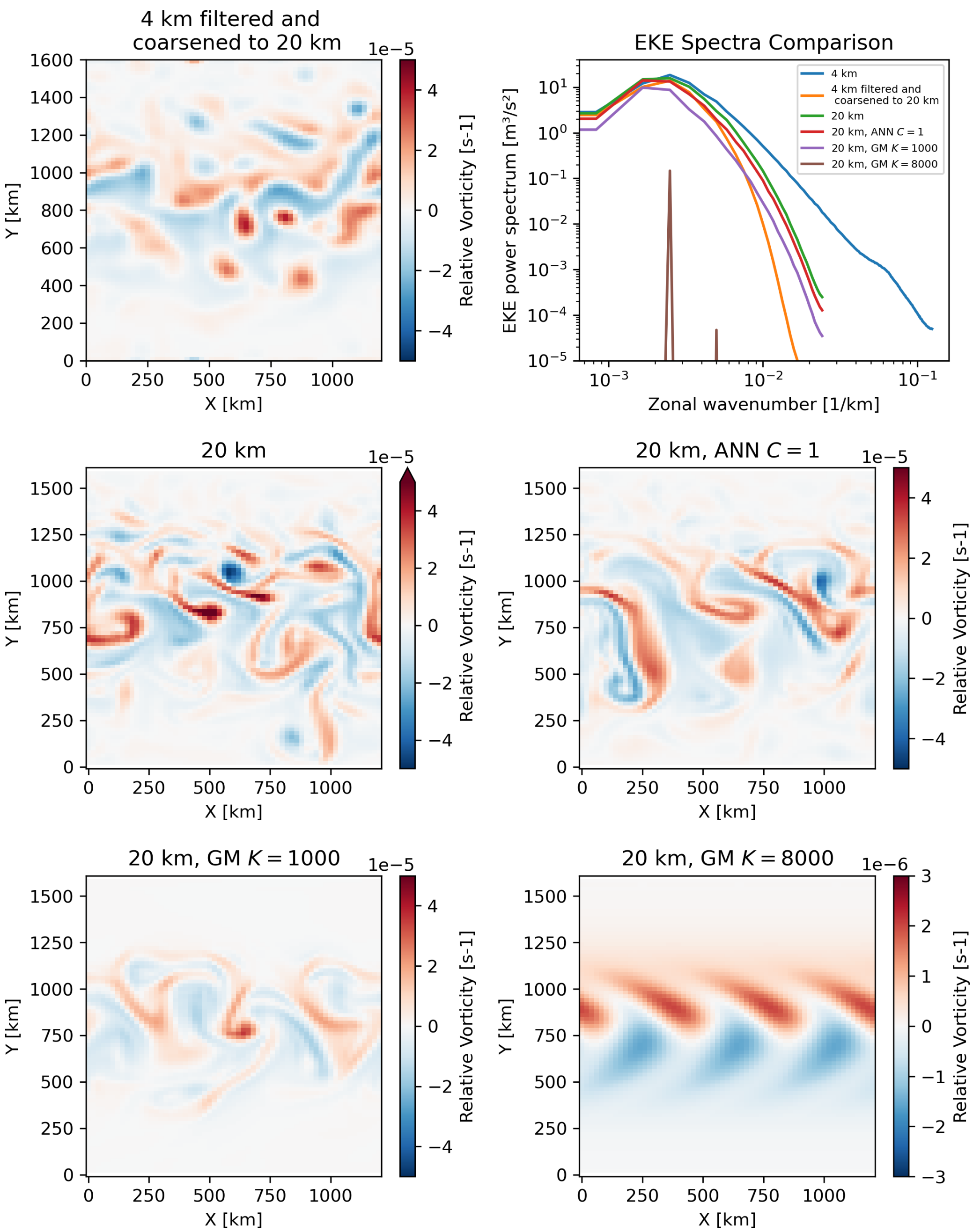

Figure 7. Snapshots of relative vorticity and the EKE spectra from different Phillips 2 layer simulations discussed in section 4.2. Note that the colorbar on the bottom right panel has been adjusted to show the range of values in that simulation.

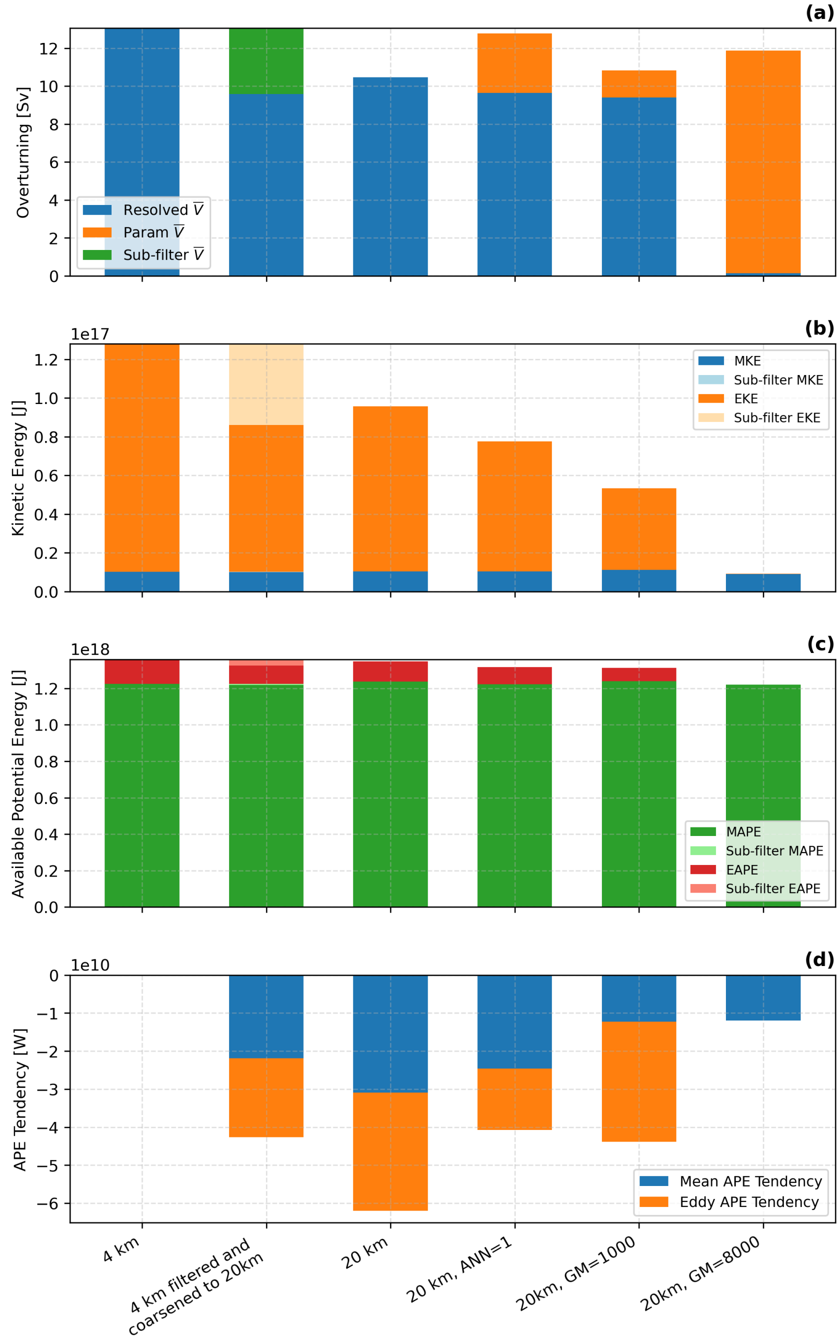

Figure 8. Bulk metrics to evaluate the online skill of the parameterizations in different experiments of Phillips 2 layer simulation (indicated along the x-axis). (a) Peak value of the total meridional transport in the upper layer for the resolved, parameterized and sub-filter components. (b) Kinetic energy and (c) available potential energy of the mean and eddy flow coming from the resolved and SGS contributions. (d) The impact of the parameterized or SGS fluxes on the mean and eddy APE tendency; no bar plot is shown for the high-resolution simulation as in this instance there are no sub-grid or parameterized thickness flux, and the bar for the 20km simulation is calculated using SGS fluxes predicted by a MLP that was not coupled with the resolved fields of the simulation.

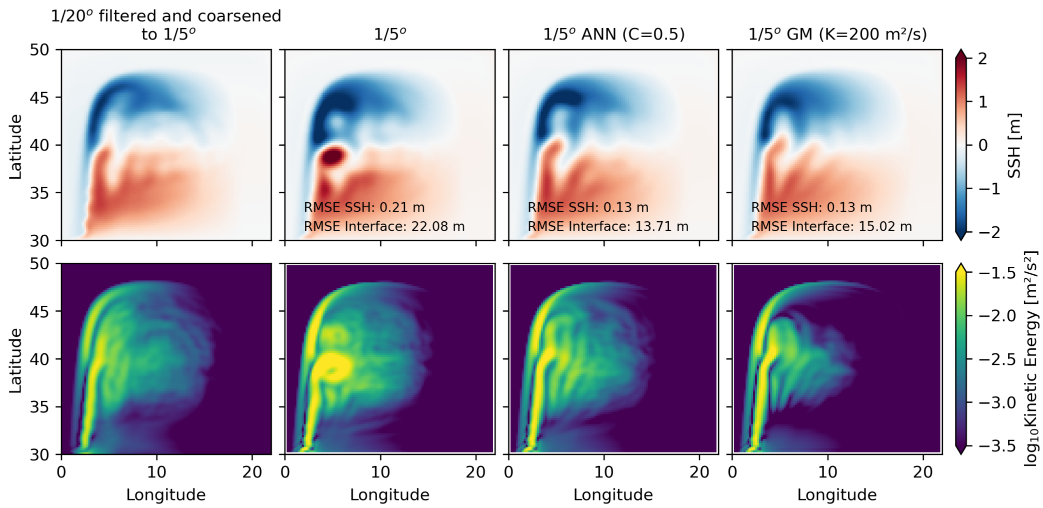

Figure 9. Mean SSH (top row) and EKE (bottow row) the for the Double gyre simulations discussed in section 4.3. The RMSE in the SSH and middle interface height are indicated for the three $1/5^o$ resolution simulations.

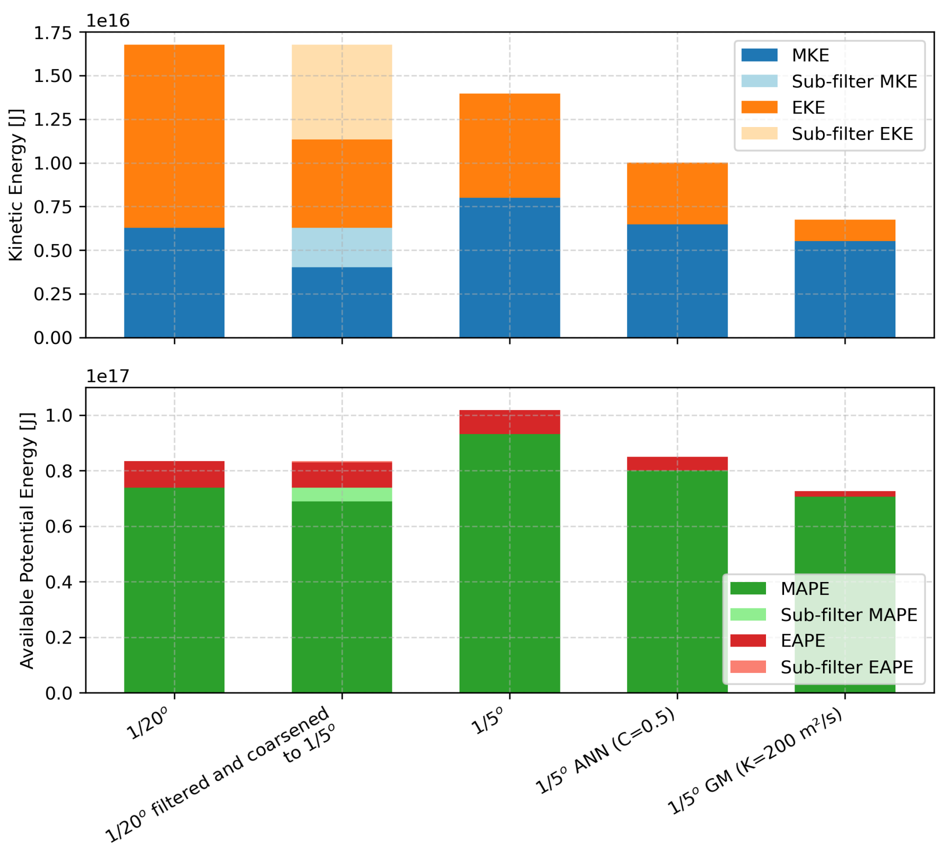

Figure 10. The volume integrated KE (top) and APE (bottom) for the Double Gyre simulations (indicated along x-axis) discussed in section 4.3. The mean, eddy, and sub-filter contributions are indicated.

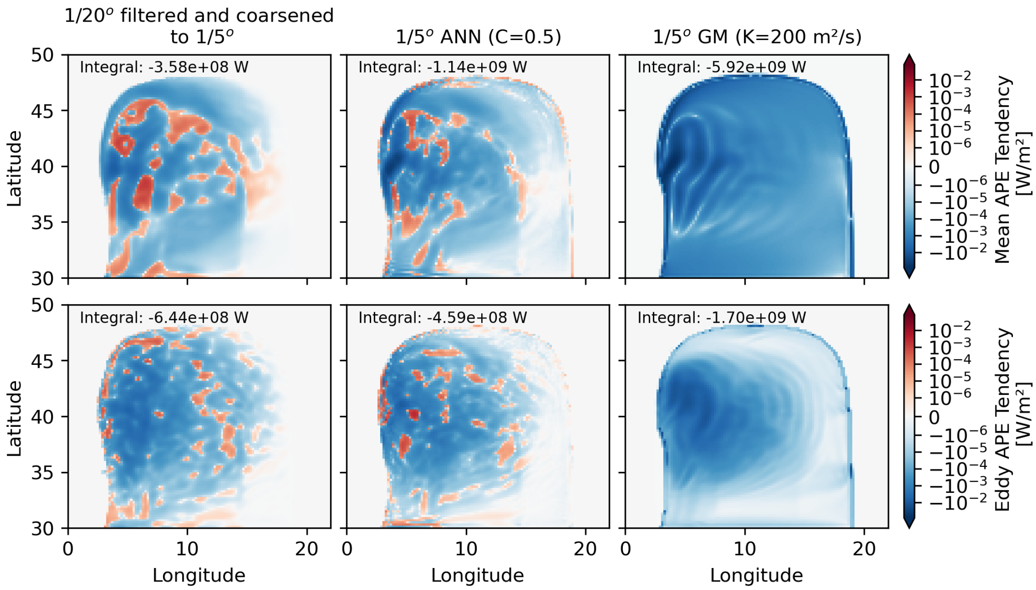

Figure 11. APE tendency exerted on the mean (top) and eddy (bottom) APE by the subgrid or parameterized fluxes in the Double Gyre simulations discussed in section 4.3. The volume integrated value for each panel is indicated at the top left.

Appendix B Eddy driven stream function

The SGS thickness flux corresponds to a volume flux in a layer ($\mathbf{F}_n$), and the net SGS volume flux below a certain interface can be represented as a stream function (McDougall & McIntosh, 2001):

$$\mathbf{\Psi}_{n-1/2} = \sum_{i=N}^{n} \mathbf{F}_i. \tag{B1}$$

Inversely, the SGS thickness flux in any layer can be expressed in terms of these SGS stream function as:

$$\begin{aligned} \mathbf{F}_n &= \overline{\mathbf{u}_n h_n} - \overline{\mathbf{u}}_n \overline{h_n} && \text{(B2)} \\ &= \mathbf{\Psi}_{n-1/2} - \mathbf{\Psi}_{n+1/2} && \text{(B3)} \\ &= \delta_n \mathbf{\Psi} && \text{(B4)} \\ & && \text{(B5)} \end{aligned}$$

This streamfunction represents a 2D divergent bolus velocity, $\mathbf{u}_n^* = \delta_n \mathbf{\Psi} / \overline{h_n}$.

While we have presented the problem entirely in terms of thickness fluxes to be used in layered models (section 2), a comparable representation of SGS buoyancy fluxes in terms of a stream function can be done for depth-level models. When representing the SGS buoyancy fluxes in depth-level models, a 3D non-divergent velocity field is constructed using this streamfunction - the quasi-stokes velocity - with no-flow boundary conditions at the top and bottom may be used.

Appendix C Interface Diffusion and Gent-McWilliams (GM) Parameterization

This SGS thickness flux in MOM6 is parameterized by the interface diffusion parameterization, which prescribes the eddy driven streamfunction as,

$$\boldsymbol{\Psi}^{GM}_{n-1/2} = -\kappa^{GM}\nabla\eta_{n-1/2}, \quad 2 \leq n \leq N. \tag{C1}$$

In this scheme, the stream function at the top and bottom of the water column are prescribed to be zero ($\boldsymbol{\Psi}^{GM}_{1/2} = \boldsymbol{\Psi}^{GM}_{N+1/2} = 0$), which ensures that the sub-grid thickness fluxes do not result in any barotropic SGS volume transport. This parameterization always flattens isopycnal surfaces and reduces APE.

This parameterization is a close cousin of the Gent-McWilliams parameterization (Gent & Mcwilliams, 1990; Gent et al., 1995), which is designed for use in z-level models, and also reduces APE of the resolved state.

Appendix D Velocity Gradient Model (VGM)

The velocity gradient model is a structural model, where the goal is to accurately predict the patterns in the SGS forcing. It is derived based on a Taylor series expansion, and by keeping only the first term of the expansion (see derivation in Aluie et al. (2022) or appendix B of Khani and Dawson (2023)).

For thickness fluxes, the VGM predicted flux is:

$$\mathbf{F}^{VGM}_n = C_{VGM}\triangle_c^2\nabla\overline{\mathbf{u}}_n\nabla\overline{h}_n \tag{D1}$$

where

$$\nabla\overline{\mathbf{u}}_n = \begin{bmatrix}\partial_x\overline{u}_n & \partial_y\overline{u}_n \\ \partial_x\overline{v}_n & \partial_y\overline{v}_n\end{bmatrix}, \tag{D2}$$

$$\nabla\overline{h}_n = \begin{bmatrix}\partial_x\overline{h}_n \\ \partial_y\overline{h}_n\end{bmatrix}, \tag{D3}$$

$\triangle_c$ is a coarse-graining scale, and C_{VGM} is a scaling coefficient corresponding to the filter scale. Note that in the presence of topography, we defined our filters such that $\overline{\eta}_b = \eta_b$. So, the sub-grid thickness flux in the bottom layer is just the sub-grid deformable thickness flux.

Unlike the GM parameterization, the VGM based expression is not guaranteed to the APE reducing, and the impact depends non-linearly on the resolved flow.

Appendix E Rotation to thickness gradient frame

In this work we often rotate all directional quantities: SGS flux vectors, the velocity gradient tensor and all thickness gradients, to a thickness gradient frame. This thickness gradient frame is defined using a set of orthogonal vectors in the horizontal plane. The first of these vectors,

$$\hat{\mathbf{T}} = \frac{\nabla\overline{h}_n}{|\nabla\overline{h}_n|} = \frac{\partial_x\overline{h}_n}{|\nabla\overline{h}_n|}\hat{\mathbf{i}} + \frac{\partial_y\overline{h}_n}{|\nabla\overline{h}_n|}\hat{\mathbf{j}}, \tag{E1}$$

points down the thickness gradient. The orthogonal (second) vector, following the right hand rule, is defined as,

$$\hat{\mathbf{N}} = \hat{\mathbf{k}} \times \frac{\nabla\overline{h}_n}{|\nabla\overline{h}_n|} = \frac{-\partial_y h_n}{|\nabla\overline{h}_n|}\hat{\mathbf{i}} + \frac{\partial_x h_n}{|\nabla\overline{h}_n|}\hat{\mathbf{j}}. \tag{E2}$$

The corresponding rotation matrix is defined as

$$\mathbf{R}_n = \begin{bmatrix}\hat{\mathbf{T}} & \hat{\mathbf{N}}\end{bmatrix} = \frac{1}{|\nabla \overline{h}_n|}\begin{bmatrix}\partial_x \overline{h}_n & -\partial_y \overline{h}_n \\ \partial_y \overline{h}_n & \partial_x \overline{h}_n\end{bmatrix}. \tag{E3}$$

This matrix can be used to rotate vector components and tensor components into the thickness gradient frame. Example, the SGS flux vector components in rotated frame can be diagnosed as

$$\widetilde{\mathbf{F}_n} = \mathbf{R}_n^T \mathbf{F}_n, \tag{E4}$$

and the velocity gradient tensor components in can be rotated as,

$$\widetilde{\nabla \overline{\mathbf{u}}_n} = \mathbf{R}_n^T (\nabla \overline{\mathbf{u}}_n)\mathbf{R}. \tag{E5}$$

When the operation in Equation E4 is applied to the thickness gradient itself, we get the expected result

$$\widetilde{\nabla \overline{h}_n} = |\nabla \overline{h}_n|\hat{\mathbf{T}} + 0\hat{\mathbf{N}}, \tag{E6}$$

since this vector does not have any projection orthogonal to the thickness gradient direction by definition.

Appendix F ML Model Design Sensitivity

There is a vast range of design choices that need to be made when working with machine learning models and designing parameterizations. For example there can be sensitivity and interdependence on (i) ML model size and architecture (e.g. we chose MLP here), (ii) training data size, (iii) training data source (which idealized simulation is used to train), (iv) learning rate and optimizers, (v) random parameter initialization, (vi) training targets, (vii) norms optimized in loss functions, (viii) input stencil/ domain of influence, (ix) selection of input features etc. It is infeasible to do a complete search over the entire parameter space and some human intuition is often used to guide the design, here we show the impact of some choices that helped guide our decisions.

F1 Impact of different network sizes

We want the most skillful predictions at the lowest cost (least number of operations per evaluation of the MLP). We expect that the skill will increase with increasing the number of trainable parameters, but likely saturate beyond a certain point as there may be no more predictive power in the input features left to be extracted. In Figure F1 we quantify this behavior for one particular class of models (these were trained with following choices: trained using data from the double gyre experiment, where all the data was rotated to the thickness gradient frame. Non-dimensional velocity and thickness gradients were used as inputs. Mean absolute error was used for the loss, with dimensional fluxes (equation 4) as output targets. Both inputs and outputs were normalized using order of magnitude estimates. We used the model snapshots 0-640 for training, 672-736 for evaluation and 736-800 for testing. Adam optimizer was used with a learning rate of 0.01, and training was continued till the relative improvement in the loss saturated within a relative tolerance of 0.01 (1% error) for at least 10 epochs).

The model skill, measured as the R2 value, improves as the number of parameters increase, and this is particularly apparent at larger filter scales and when the model stencil is wider. However, since skill is already very high at filter scales of 50 and 100 km this effect is minor, and at larger filter scales the skill seems to asymptote to its maximum value approximately around 10K parameters.

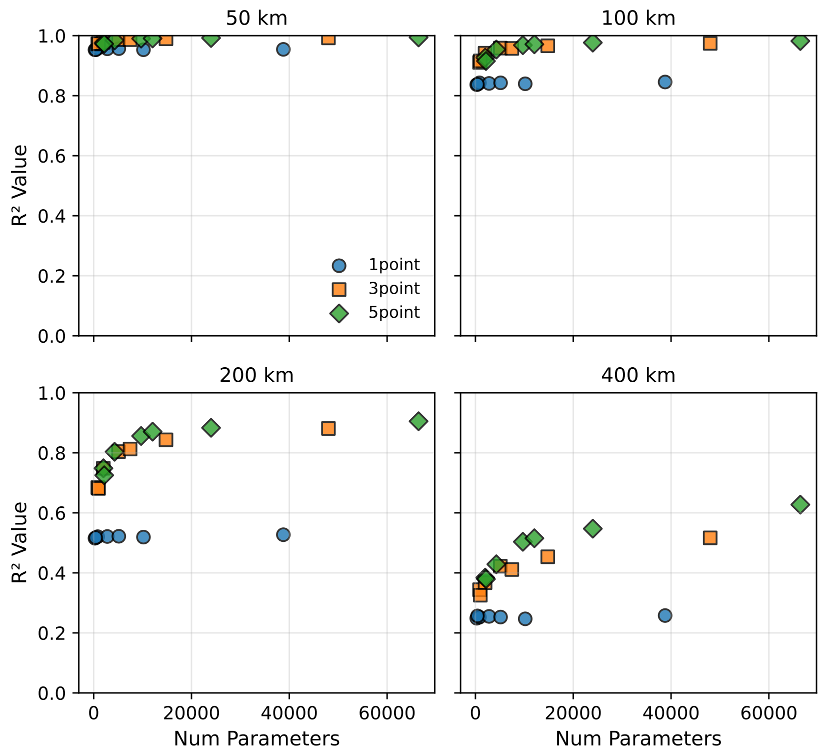

Figure F1. ML model skill, defined as the R2 value, as a function of number of parameters for different filter scales. The details ML model being tested is described in F1.

F2 Impact of training data size

To test the impact that the amount of training data has, we experimented with the same setup as the one used in the previous section with varying amount of data. We fixed all aspects, and performed the test in the case where the stencil size is 3X3 and ML model has 3386 parameters (model shape 54,36,36,2). We created batches of 16 model snapshots, and tested the effect that increasing the number of batches had. As shown in Figure F2 the model skill increases upto about 128 batches (2048 model snapshots), but saturates past that point. This led us to use this data volume for training the model used in the main text.

F3 Impact of different loss functions and targets

Non-dimensionalization of the sub-grid fluxes leads to a significant collapse in the distributions (see Figure 2), but also produces outliers (due to possibility of division by small velocity or thickness gradients). We have two choices of loss during training, either training on the dimensional fluxes $\mathcal{L}^{dim} = ||\mathbf{F} - \mathbf{F}^{pred}|| = ||\mathbf{F} - (\triangle^2 |\nabla \mathbf{u}||\nabla h|) f_\theta(.)||$ or training on non-dimensional fluxes $\mathcal{L}^{non-dim} = ||\mathbf{F}/(\triangle^2|\nabla \mathbf{u}||\nabla h|) - f_\theta(.)||$. Both these forms should give us the same functional representation if a unique function $f_\theta(.)$ exists, but in the more realistic situation where $f_\theta(.)$ is an approximation and we want it to equally balance data coming from many regimes (different simulations, varying energy levels at different depths, etc) it would seem that $\mathcal{L}^{non-dim}$ might be a better choice. Along with this we also compared the use of mean square error (MSE) vs mean absolute error (MAE), since MAE is generally considered to be more tolerant to outliers.

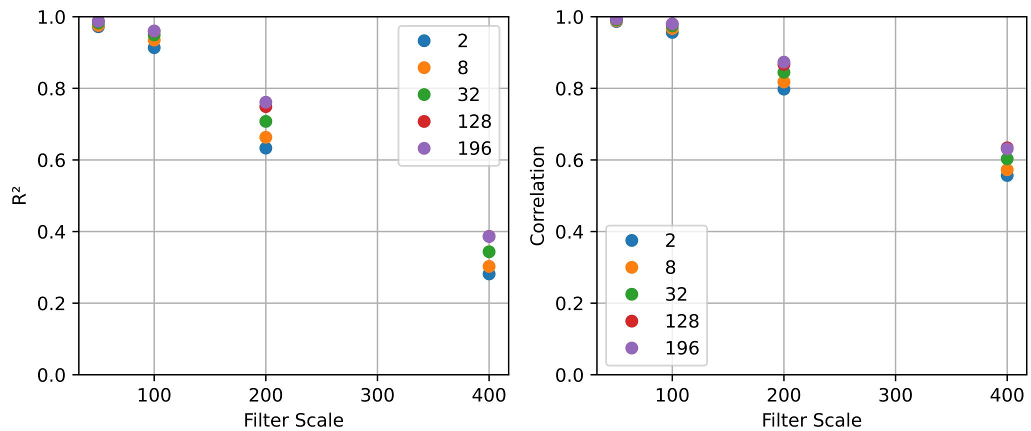

Figure F2. ML model skill, defined as the R2 value, as a function of filter scales for different number of data batches used for training.

As seen in Figure F3, the impact of these choices is relatively minor in the model considered in the previous section. Generally, the models trained using MAE seem to go better, except at the largest filter scales - where the model trained using non-dimensionalized outputs and MSE does better. However, this model is the worst performer at all other scales. The models trained using MAE also generally perform well across a wide range of scales (bottom panel). To our surprise, the impact of this choice was smaller than we had expected. For the main study we used $\mathcal{L}^{non-dim}$ along with MAE.

F4 Impact of choice of training data

We want to build ML models that are easily generalizable and training data agnostic. For example, a model trained on data from the P2L simulation should be able to make good predictions on data from DG experiment, and vice versa. Developing appropriate non-dimensionalizations, as highlighted in section 3, is a step in this direction. In Figure F4 we show that the model trained using data from both simulations simultaneously generally performs the best or close to the best, which is why we chose this training strategy for the model in the main text.

It is worth noting that in fact even models trained on a single experiment, perform relatively well when tested on the other experiment. This is a result of the design choices we made. During the initial phase of our development, when we trained models to have the form of equation 4, the skill on unseen data was extremely poor (not shown), also see Perezhogin et al. (submitted).

F5 Impact of other aspects

The random seed, which sets **the random initializations** of the ML model weights, seems to have a small impact on the skill. The model skill averaged across all filter scales, for the case discussed in the above section, varied between values of 0.7 - 0.725 over different random seeds. Since this effect seems small, in the main text we use only a single trained model.

The **learning rate**, similarly has a minor impact on the final skill but impacts the nature of decrease in the loss. Large learning rates (~ 0.1) lead to noisy training, while smaller rates (< 0.005) take too many epochs to train. We found than a learning rate

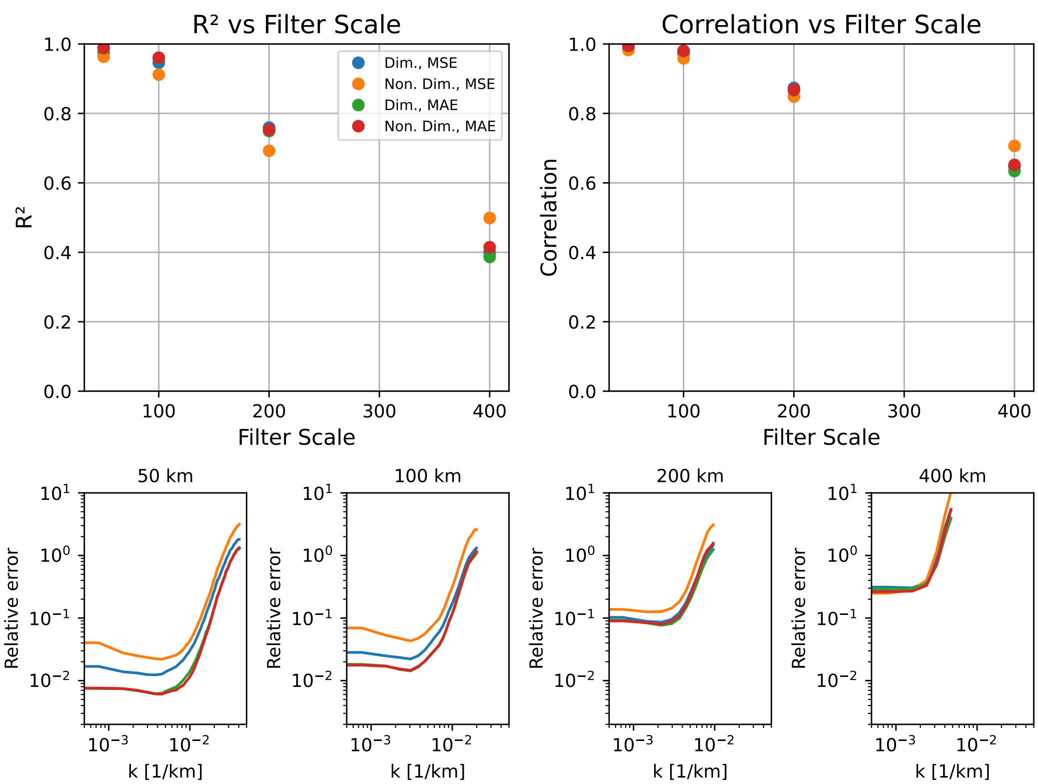

Figure F3. Top row shows the model skill in terms of R2 and correlation as a function of filter scale for different choices of loss function. Bottom two rows show the relative error, which is defined as the power spectrum of the (truth - prediction) divided by the power spectrum of the truth; value of greater than 1 implies that the variance in the anomaly field is greater than in the truth.

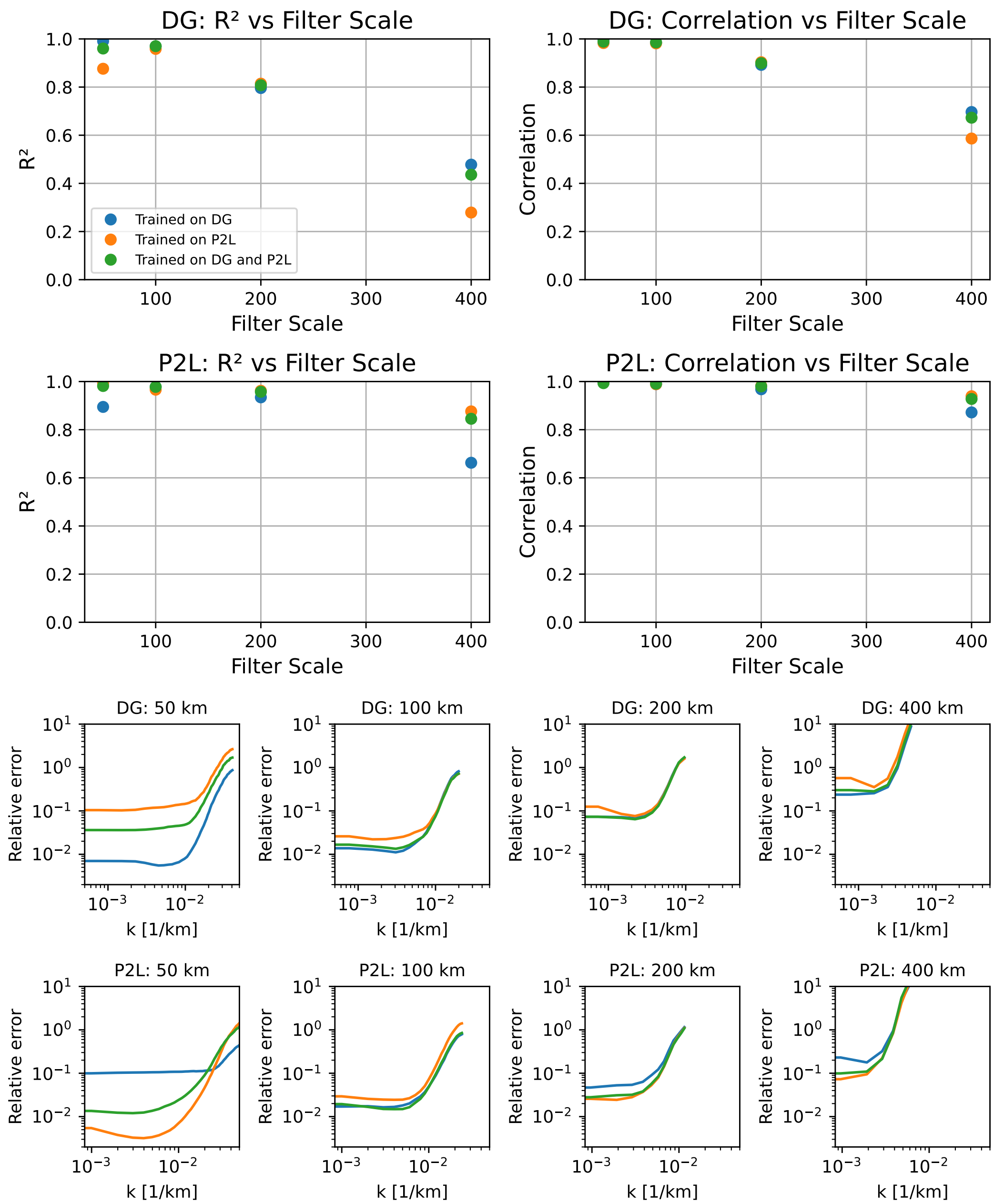

Figure F4. Top two rows show the model skill in terms of R2 and correlation as a function of filter scale for different choices of training datasets. Bottom two rows show the relative error, which is defined as the power spectrum of the (truth - prediction) divided by the power spectrum of the truth; value of greater than 1 implies that the variance in the anomaly field is greater than in the truth. The title of the panels indicate the dataset on which testing is done, and the legend indicated the dataset that is used for training.

around 0.01 was optimal in not taking too many epochs and not being too noisy, which is what we used for the models discussed in the main text.

Appendix G Simulation evaluation metrics

Here we describe the metrics used to assess the physical properties of our simulations.

G1 Overturning Circulation

The overturning circulation ($\overline{V}_n(y)$) is defined as the volume transport across a longitude band in a particular layer,

$$\overline{V}_n(y) = \oint \overline{v_n h_n}^t dx + \oint \overline{F_n^y}^t dx, \tag{G1}$$

where $F_n^y(x, y, z, t)$ is the sub-grid or sub-filter meridional flux in layer n, $\oint(.)dx$ corresponds to a zonal integral over the full longitudinal domain, and $\overline{(.)}^t$ indicates a time average. The first term on the RHS corresponds to the resolved overturning and includes the contribution from both the mean and the variable parts of the flow, and the second term corresponds to the parameterized overturning.

G2 Eddy Kinetic Energy Spectrum

The EKE spectrum provides an effective measure of the flow variability at different scales. Here, we define it as the zonal wavenumber (k) power spectrum of the $\mathbf{u}'_n(x, y, t) = \mathbf{u}_n(x, y, t) - \overline{\mathbf{u}}_n^t(x, y)$,

$$EKE_n(k) = \frac{1}{2}(\overline{|u'_n(k, y, t)|^2}^{t,y} + \overline{v'_n(k, y, t)|^2}^{t,y}), \tag{G2}$$

where $\overline{(.)}^{t,y}$ is a time and meridional average and $u'_n(k, y, t)$ is the zonal Fourier transform of $u'_n(x, y, t)$. By Parseval's theorem we have

$$\sum_k EKE_n(k) = \frac{1}{2}\int(\overline{u'_n(x, y, t)^2}^{t,y} + \overline{v'_n(x, y, t)^2}^{t,y})dx, \tag{G3}$$

which shows the scale-wise decomposition aspect of the zonal wavenumber EKE spectrum in the x-direction. We used the xrft package (`https://xrft.readthedocs.io/`) for this analysis.

G3 Integral Kinetic and Available Potential Energies

The volume integrated KE in Joules is defined as:

$$KE = \frac{1}{2}\sum_k \int \rho_0 |\mathbf{u}_k|^2 h_k dxdy \tag{G4}$$

The KE of the mean flow ($\overline{\mathbf{u}_k}^t$, $\overline{h_k}^t$) is referred to as MKE, and the EKE is defined as $EKE = \overline{KE}^t - MKE$.

The volume integrated APE in Joules is defined as:

$$APE = \frac{1}{2}\sum_k \int \rho_0 g'_{k-1/2}(\eta_{k-1/2} - \eta^{ref}_{k-1/2})^2 dxdy \tag{G5}$$

The $MAPE$ is the APE of the mean state ($\overline{\eta_{k-1/2}}^t$), and the $EAPE$ is given by $EAPE = \overline{APE}^t - EAPE$

For filtered and coarsened flow ($\overline{\mathbf{u}_k}^{\triangle}$, $\overline{h_k}^{\triangle}$) we can also define the total kinetic energy ($KE^{\triangle}$), which has its corresponding mean ($MKE^{\triangle}$, for $\overline{\mathbf{u}_k}^{\triangle,t}$, $\overline{h_k}^{\triangle,t}$) and eddy ($EKE^{\triangle}$) components. Accordingly, we also define the SGS kinetic energies for the total ($KE - KE^{\triangle}$), mean ($MKE - MKE^{\triangle}$), and eddy ($EKE - EKE^{\triangle}$) flow. Similarly APE and its components can be defined for the filtered and coarsened flow and the SGS contribution.

G4 Available potential energy tendency due to sub-filter or parameterized fluxes

Following (Loose, Marques, et al., 2023), we can compute the impact that the SGS fluxes have on volume integrated APE tendency as,

$$(\partial_t APE)^{SF} = \sum_{n=1}^{N} \rho_0 \mathbf{F}_n \cdot \nabla M_n, \tag{G6}$$

where $\mathbf{F}_n$ are the SGS fluxes and M_n is the dynamic pressure of the resolved.

In a 2 layer fluid we have,

$$(\partial_t APE)^{SF} = \rho_0 g^r_{1/2}(\mathbf{F}_1 + \mathbf{F}_2) \cdot \nabla \eta_{1/2} + \rho_0 g^r_{3/2} \mathbf{F}_2 \cdot \nabla \eta_{3/2}. \tag{G7}$$

Notably, only the last term on the RHS, which arises due to SGS fluxes in the bottom layer, makes a significant contribution to the APE tendency, since $|\nabla \eta_{3/2}| >> |\nabla \eta_{1/2}|$. The barotropic contribution to APE, first term on RHS, is small.

We decompose this tendency into the time mean and eddy contribution, by defining the contribution from the mean as

$$(\partial_t APE)^{SF,mean} = \rho_0 g^r_{3/2}(\overline{\mathbf{F}_2}^t \cdot \nabla \overline{\eta_{3/2}}^t), \tag{G8}$$

where the barotropic contribution is neglected. The eddy contribution is defined as $(\partial_t APE)^{SF,eddy} = \overline{(\partial_t APE)^{SF}}^t - (\partial_t APE)^{SF,mean}$.

Appendix H Online sensitivity to grid-size and parameterization coefficients

Simulation of oceanic mesoscale turbulence is relatively sensitive to how well the first baroclinic deformation radius (shown in Figure H1 for the two simulations considered here) is resolved. In this study we chose our coarsening scale and simulation grid-size to lie in a range that encompasses the deformation radius, so that skill of the new parameterization at both the eddy permitting and non-eddying resolutions can be investigated. Here we provide details of the sensitivity of the coarse simulations to both grid spacing and parameterization coefficients.

H1 Phillips 2 Layer

For P2L simulation we tested the sensitivity of the overturning transport to the grid size and parameterization coefficients, shown in Figure H2.

In this case, the unparameterized low resolution simulation always has lower overturning transport relative to the HR simulation. The response of the parameterized flux contribution to the GM diffusivity is almost linear and insensitive to grid size, and adjusting the GM diffusivity allows us to increase the total overturning transport to the appropriate value. However, this is only achieved when the diffusivity is relatively large ($\kappa_{GM} \sim 8,000$–$10,000\ m^2/s$), which is also a parameter regime in which the resolved eddying flow has been entirely suppressed. Consequently the contribution of the resolved flow to the overturning has been completely eliminated and the entire contribution comes from parameterized fluxes (Figure H2 second column). There seems to be no lower diffusivity value at which the appropriate overturning can be achieved with only marginal impact to the resolved flow.

In contrast, the simulations with the ANN based parameterization are usually able to achieve the appropriate level of overturning with minimal damage to the resolved flow. The response of the parameterized flux to the grid size and parameterization amplification coefficient (C_{ANN}) is non-linear. At grid sizes of 10 and 20 km, which are smaller than the deformation radius everywhere in the domain, the appropriate overturning is achieved at $C_{ANN} = 1$ and with minimal damage to the resolved overturning. At a grid size of 80 km, which is larger than the deformation radius everywhere in the domain, the appropriate overturning is achieved at $C_{ANN} = 2$. At the intermediate grid size of 40 km, the ANN parameterization does produce improvements to overall overturning with little damage to the resolved flow (also at $C_{ANN} = 1$), but is unable to achieve similar overturning to the HR simulation. We anticipate that combining the thickness flux parameterization with a momentum flux parameterization may be able to produce further improvements at these intermediate resolutions.

H2 Double Gyre

Unlike the P2L simulation, in the DG case there is no overturning circulation. In the context of the thickness flux parameterization, the error in the MAPE is the most relevant, which is also linked to the error in mean SSH. We tested the sensitivity of these quantities and a few others to the parameterization coefficients and grid-size (Figure H3).

For the GM parameterization, a diffusivity of about 200 m^2/s is able to achieve the appropriate MAPE at all resolutions, which similar to the P2L also comes an almost complete damping of the resolved eddies. The ANN is able to achieve the best MAPE with $C_{ANN} \sim 0.5 - 0.75$, and this happens with relatively less deterioration in the resolved eddies.

$\triangle_c$ [km]	κ_{GM} [m^2/s]		C_{VGM} [nondim]	
	P2L	**DG**	**P2L**	**DG**
10	137	40	0.110	0.075
20	596	105	0.077	0.072
40	2287	192	0.077	0.067
80	6266	108	0.091	0.048

Table H1. Estimated GM diffusivity (κ_{GM}) and VGM coefficient (C_{VGM}) from the data using least squares fitting, for both the Phillips 2-Layer (P2L) and Double Gyre (DG) setups.

References

Abernathey, R., & Wortham, C. (2015). Phase speed cross spectra of eddy heat fluxes in the eastern pacific. *Journal of Physical Oceanography*, *45*(5), 1285–1301.

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., ... others (2019). The gfdl global ocean and sea ice model om4. 0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3167–3211.

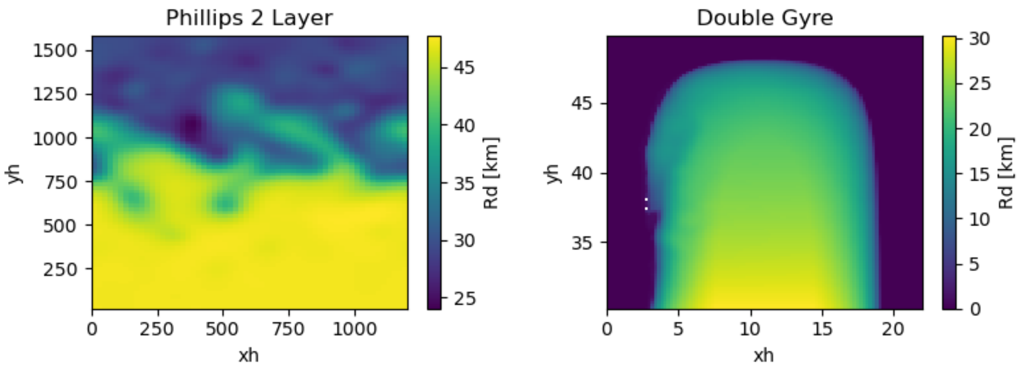

Figure H1. First baroclinic deformation radii for P2L (left) and DG (right). Note the different range of scales for the two colobars.

Aluie, H., Hecht, M., & Vallis, G. K. (2018). Mapping the energy cascade in the north atlantic ocean: The coarse-graining approach. *Journal of Physical Oceanography*, *48*(2), 225–244.

Aluie, H., Rai, S., Yin, H., Lees, A., Zhao, D., Griffies, S. M., ... Shang, J. K. (2022). Effective drift velocity from turbulent transport by vorticity. *Physical Review Fluids*, *7*(10), 104601.

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., ... others (2021). Climate-invariant machine learning. *arXiv preprint arXiv:2112.08440*.

Bodner, A., Balwada, D., & Zanna, L. (2023). A data-driven approach for parameterizing submesoscale vertical buoyancy fluxes in the ocean mixed layer. *arXiv preprint arXiv:2312.06972*.

Bracco, A., Brajard, J., Dijkstra, H. A., Hassanzadeh, P., Lessig, C., & Monteleoni, C. (2025). Machine learning for the physics of climate. *Nature Reviews Physics*, *7*(1), 6–20.

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298.

Chelton, D. B., DeSzoeke, R. A., Schlax, M. G., El Naggar, K., & Siwertz, N. (1998). Geographical variability of the first baroclinic rossby radius of deformation. *Journal of Physical Oceanography*, *28*(3), 433–460.

Eden, C., & Greatbatch, R. J. (2008). Towards a mesoscale eddy closure. *Ocean Modelling*, *20*(3), 223–239.

Ferrari, R., Griffies, S. M., Nurser, A. G., & Vallis, G. K. (2010). A boundary-value problem for the parameterized mesoscale eddy transport. *Ocean Modelling*, *32*(3-4), 143–156.

Ferrari, R., McWilliams, J. C., Canuto, V. M., & Dubovikov, M. (2008). Parameterization of eddy fluxes near oceanic boundaries. *Journal of Climate*, *21*(12), 2770–2789.

Ferrari, R., & Wunsch, C. (2009). Ocean circulation kinetic energy: Reservoirs, sources, and sinks. *Annual Review of Fluid Mechanics*, *41*, 253–282.

Ferreira, D., Marshall, J., & Heimbach, P. (2005). Estimating eddy stresses by fitting dynamics to observations using a residual-mean ocean circulation model and its adjoint. *Journal of Physical Oceanography*, *35*(10), 1891–1910.

Gent, P. R. (2011). The gent–mcwilliams parameterization: 20/20 hindsight. *Ocean Modelling*, *39*(1-2), 2–9.

Gent, P. R., & Mcwilliams, J. C. (1990). Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, *20*(1), 150–155.

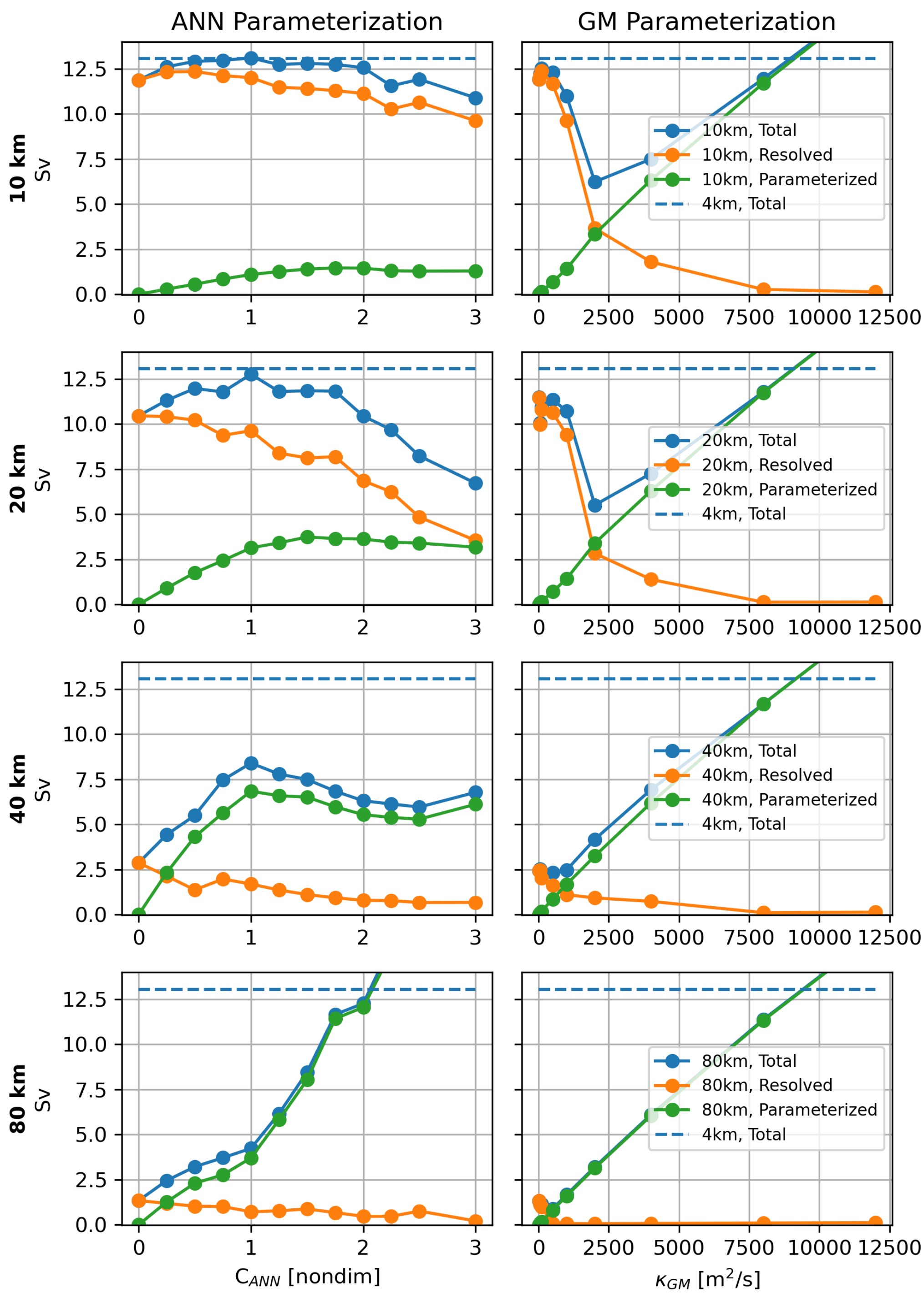

Figure H2. Sensitivity of the overturning circulation and its resolved and parameterized components to the MLP coefficient (left) and the GM diffusivity (right) for the Phillip 2 layer simulations at different grid-sizes.

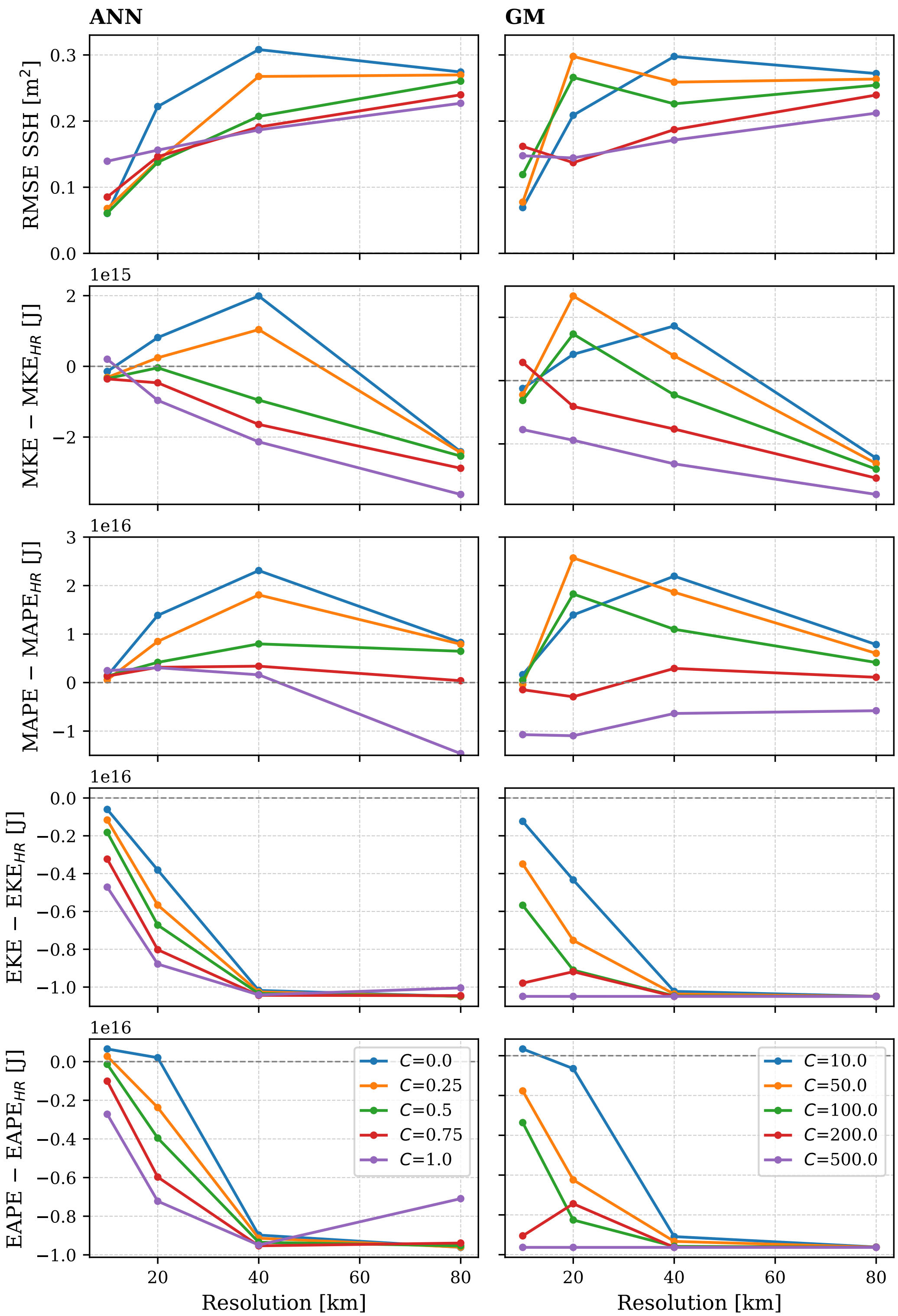

Figure H3. Sensitivity to the MLP coefficient (left) and the GM diffusivity (right) for the Double Gyre simulations at different resolutions.

Gent, P. R., Willebrand, J., McDougall, T. J., & McWilliams, J. C. (1995). Parameterizing eddy-induced tracer transports in ocean circulation models. *Journal of physical oceanography*, *25*(4), 463–474.

Hallberg, R. (2013). Using a resolution function to regulate parameterizations of oceanic mesoscale eddy effects. *Ocean Modelling*, *72*, 92–103.

Jansen, M. F., Adcroft, A. J., Hallberg, R., & Held, I. M. (2015). Parameterization of eddy fluxes based on a mesoscale energy budget. *Ocean Modelling*, *92*, 28–41.

Khani, S., & Dawson, C. N. (2023). A gradient based subgrid-scale parameterization for ocean mesoscale eddies. *Journal of Advances in Modeling Earth Systems*, *15*(2), e2022MS003356.

Killworth, P. D. (2001). Boundary conditions on quasi-stokes velocities in parameterizations. *Journal of physical oceanography*, *31*(4), 1132–1155.

Lai, C.-Y., Hassanzadeh, P., Sheshadri, A., Sonnewald, M., Ferrari, R., & Balaji, V. (2024). Machine learning for climate physics and simulations. *Annual Review of Condensed Matter Physics*, *16*.

Loose, N., Abernathey, R., Grooms, I., Busecke, J., Guillaumin, A., Yankovsky, E., ... others (2022). Gcm-filters: A python package for diffusion-based spatial filtering of gridded data. *Journal of Open Source Software*, *7*(70).

Loose, N., Bachman, S., Grooms, I., & Jansen, M. (2023). Diagnosing scale-dependent energy cycles in a high-resolution isopycnal ocean model. *Journal of Physical Oceanography*, *53*(1), 157–176.

Loose, N., Marques, G. M., Adcroft, A., Bachman, S., Griffies, S. M., Grooms, I., ... Jansen, M. F. (2023). Comparing two parameterizations for the restratification effect of mesoscale eddies in an isopycnal ocean model. *Journal of Advances in Modeling Earth Systems*, *15*(12), e2022MS003518.

Mak, J., Maddison, J. R., Marshall, D. P., Ruan, X., Wang, Y., & Yeow, L. (2023). Scale-awareness in an eddy energy constrained mesoscale eddy parameterization. *Journal of Advances in Modeling Earth Systems*, *15*(12), e2023MS003886.

Marshall, D. P., Maddison, J. R., & Berloff, P. S. (2012). A framework for parameterizing eddy potential vorticity fluxes. *Journal of Physical Oceanography*, *42*(4), 539–557.

McDougall, T. J., & McIntosh, P. C. (2001). The temporal-residual-mean velocity. part ii: Isopycnal interpretation and the tracer and momentum equations. *Journal of Physical Oceanography*, *31*(5), 1222–1246.

Moser, R. D., Haering, S. W., & Yalla, G. R. (2021). Statistical properties of subgrid-scale turbulence models. *Annual Review of Fluid Mechanics*, *53*, 255–286.

Perezhogin, P., Adcroft, A., & Zanna, L. (submitted). Generalizable neural-network parameterization of mesoscale eddies in idealized and global ocean models. *Geophysical Research Letters*.

Perezhogin, P., Zhang, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2024). A stable implementation of a data-driven scale-aware mesoscale parameterization. *Journal of Advances in Modeling Earth Systems*, *16*(10), e2023MS004104.

Prakash, A., Jansen, K. E., & Evans, J. A. (2022). Invariant data-driven subgrid stress modeling in the strain-rate eigenframe for large eddy simulation. *Computer Methods in Applied Mechanics and Engineering*, *399*, 115457.

Ramadhan, A., Marshall, J., Souza, A., Wagner, G. L., Ponnapati, M., & Rackauckas, C. (2020). Capturing missing physics in climate model parameterizations using neural differential equations. *arXiv preprint arXiv:2010.12559*.

Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2023). Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. *Journal of Advances in Modeling Earth Systems*, *15*(1),

e2022MS003258.
Sagaut, P. (2005). *Large eddy simulation for incompressible flows: an introduction.* Springer Science & Business Media.
Sane, A., Reichl, B. G., Adcroft, A., & Zanna, L. (2023). Parameterizing vertical mixing coefficients in the ocean surface boundary layer using neural networks. *arXiv preprint arXiv:2306.09045*.
Smith, K. S. (2007). The geography of linear baroclinic instability in earth's oceans. *Journal of Marine Research*, *65*(5), 655–683.
Smith, K. S., & Vallis, G. K. (2002). The scales and equilibration of midocean eddies: Forced–dissipative flow. *Journal of Physical Oceanography*, *32*(6), 1699–1720.
Smith, R. D., & Gent, P. R. (2004). Anisotropic gent–mcwilliams parameterization for ocean models. *Journal of Physical Oceanography*, *34*(11), 2541–2564.
Srinivasan, K., Chekroun, M. D., & McWilliams, J. C. (2023). Turbulence closure with small, local neural networks: Forced two-dimensional and β-plane flows. *arXiv preprint arXiv:2304.05029*.
Tulloch, R., Marshall, J., Hill, C., & Smith, K. S. (2011). Scales, growth rates, and spectral fluxes of baroclinic instability in the ocean. *Journal of Physical Oceanography*, *41*(6), 1057–1076.
Vallis, G. K. (2017). *Atmospheric and oceanic fluid dynamics.* Cambridge University Press.
Visbeck, M., Marshall, J., Haine, T., & Spall, M. (1997). Specification of eddy transfer coefficients in coarse-resolution ocean circulation models. *Journal of physical oceanography*, *27*(3), 381–402.
Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophysical Research Letters*, *48*(6), e2020GL091363.
Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, *47*(17), e2020GL088376.
Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2023). Implementation and evaluation of a machine learned mesoscale eddy parameterization into a numerical ocean circulation model. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003697.