



RESEARCH ARTICLE

10.1029/2025MS005245

Key Points:

- A data-driven parameterization for mesoscale eddy-induced thickness fluxes is developed and implemented in NOAA GFDL's MOM6
- This parameterization replaces the diffusive closure in traditional Gent-McWilliams scheme with an ANN
- In two-layer simulations the parameterization reduces potential energy without overly dissipating the kinetic energy

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

D. Balwada,
dbalwada@ldeo.columbia.edu

Citation:

Balwada, D., Perezhogan, P., Adcroft, A., & Zanna, L. (2026). Design and implementation of a data-driven parameterization for mesoscale eddy-induced transport: Idealized two-layer configurations. *Journal of Advances in Modeling Earth Systems*, 18, e2025MS005245. <https://doi.org/10.1029/2025MS005245>

Received 19 MAY 2025

Accepted 14 AUG 2026

Author Contributions:

Conceptualization: Dhruv Balwada, Alistair Adcroft, Laure Zanna

Data curation: Dhruv Balwada

Formal analysis: Dhruv Balwada

Funding acquisition: Alistair Adcroft, Laure Zanna

Investigation: Dhruv Balwada

Methodology: Dhruv Balwada, Pavel Perezhogan

Project administration: Alistair Adcroft, Laure Zanna

© 2026 The Author(s). Journal of Advances in Modeling Earth Systems published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](#), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Design and Implementation of a Data-Driven Parameterization for Mesoscale Eddy-Induced Transport: Idealized Two-Layer Configurations

Dhruv Balwada¹ , Pavel Perezhogan² , Alistair Adcroft³ , and Laure Zanna^{2,4} 

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

Abstract Mesoscale eddies contribute to ocean transport and are a major sink of available potential energy (APE). When unresolved or only partially resolved, their effects must be parameterized to simulate a realistic large-scale state. Traditionally, this is achieved via the Gent-McWilliams (GM) parameterization, which represents the eddy-induced transport as an adiabatic advection built from a downgradient diffusive closure for the horizontal flux. While effective at capturing the bulk APE removal by eddies, the downgradient diffusive framework limits the structural accuracy of the parameterized fluxes, which in reality can contain significant upgradient and cross-gradient components. Additionally, at eddy permitting resolutions, GM can negatively impact the resolved eddies, and is often turned off in regions where eddies are partially resolved. Here we present a novel data-driven parameterization that is not constrained to be diffusive, and can capture the full structure of the eddy-induced thickness fluxes while extracting APE and 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 limitation improve the generalization of the data-driven scheme. A small neural network is used, ensuring low cost and ease of implementation. The parameterization performs skillfully in offline evaluation, especially at scales smaller than the largest eddies. It is implemented in NOAA GFDL's MOM6 and is skillful in online tests in two-layer idealized simulations of a zonal channel and a wind-driven gyre, at eddy-permitting and non-eddy-resolving resolutions.

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 ocean models—especially those that are run for long periods or including many ensemble members—their effects must be represented through parameterizations. We present a new data-driven parameterization designed to better represent mesoscale eddy effects in ocean models. The parameterization is trained using data from high-resolution simulations, and is formed of a small neural network that 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 online in idealized simulations. It extracts APE from large scale flow without degrading resolved features, offering a promising enhancement to the existing Gent-McWilliams (GM) parameterization.

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 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. An important and often unresolved or partially resolved range of scales in ocean models are the mesoscales.

Mesoscales (~50–200 km) are the dominant energy-containing scales in the ocean (Ferrari & Wunsch, 2009), playing an important role in shaping the mean circulation and stratification (Gent, 2011), and in transporting tracers (Meredith & Naveira Garabato, 2021). These eddies are believed to be largely generated as a result of baroclinic instability (Killworth, 1997; K. S. Smith, 2007), which has its fastest growth rate at scales close to the

Resources: Dhruv Balwada,
Alistair Adcroft, Laure Zanna
Software: Dhruv Balwada,
Pavel Perezhogin
Supervision: Alistair Adcroft,
Laure Zanna
Validation: Dhruv Balwada,
Pavel Perezhogin
Visualization: Dhruv Balwada
Writing – original draft: Dhruv Balwada
Writing – review & editing:
Dhruv Balwada, Pavel Perezhogin,
Alistair Adcroft, Laure Zanna

first baroclinic deformation radius (scales of ~ 10 – 100 km) (Chelton et al., 1998; Tulloch et al., 2011). Also, these instabilities have a tendency to grow at the expense of the available potential energy (APE), and thus a bulk effect is to flatten isopycnals in the ocean. The dominance of rotation effects at these scales also usually leads to a subsequent inverse kinetic energy (KE) cascade, which is why the dominant peak of KE 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 smaller than 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 small-scale turbulence or diapycnal mixing—mesoscale flow is predominantly adiabatic in the ocean interior.

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. The GM parameterization was designed to respect two important aspects of mesoscale processes: (a) the parameterization is adiabatic, and (b) the net effect of the parameterization is to reduce APE. In both depth-coordinate and isopycnal models, the mesoscale eddy effects are represented as an eddy-induced advection defined through a streamfunction. In depth-coordinate models, this streamfunction is constructed from downgradient diffusion of horizontal eddy buoyancy fluxes, while in isopycnal models it arises from a diffusion of interface heights—in both cases the horizontal flux closure is diffusive, but the resulting transport is adiabatic by construction, and is not a diffusion of buoyancy across isopycnals. The GM parameterization has led to dramatic improvements in the ocean simulations, the adiabatic aspect ensures that water masses were not eroded away in the interior by spurious diffusion. However, the choice of isotropic diffusion as a closure and the associated eddy diffusivity has always been a major source of uncertainty and is an active area of research.

Since the original work of Gent and McWilliams (1990), research has advanced along several fronts: improved estimation of eddy diffusivity, extensions of the diffusive operator itself, and complementary approaches that address processes GM does not represent. Observations, models, and theory all indicate that the eddy diffusivity must vary in space and time (e.g., Ferreira et al., 2005; Kusters et al., 2025), motivating approaches based on mixing-length arguments (Eden & Greatbatch, 2008; Visbeck et al., 1997), linear stability analysis (Eden, 2012; Eden & Dettmer, 2025), and energy-budget constraints from the MEKE (Jansen et al., 2015, 2019) and GEO-METRIC (D. P. Marshall et al., 2012; Mak et al., 2018) closures. The GM operator itself has also been extended: R. D. Smith and Gent (2004) introduced an anisotropic formulation, and Eden (2023) showed that biharmonic PV mixing is equivalent to biharmonic friction plus biharmonic thickness mixing, providing a form of GM appropriate for eddy-permitting models. A separate but related challenge is that GM, by design, removes APE without returning any KE to the resolved flow, leaving coarse models under-energized. This has motivated backscatter parameterizations that couple energy return through the momentum equations, either deterministically (Bachman, 2019; Jansen et al., 2019) or stochastically (Grooms et al., 2025). At eddy-permitting resolutions, standard GM is known to excessively damp the partially resolved eddies (Hallberg, 2013), which has led to the use of resolution functions to taper or disable GM (Adcroft et al., 2019; Hallberg, 2013), as well as the proposal by Mak et al. (2023) to apply GM only to the large-scale field via a splitting approach. These advances have substantially improved the physical realism of the conventional diffusive closure in GM. However, all remain rooted in the assumption of downgradient diffusion, which limits the structural accuracy of the predicted fluxes—the eddy fluxes diagnosed from high-resolution simulations contain significant upgradient and cross-gradient components that cannot be captured by a diffusive operator (e.g., Bachman et al., 2020; Kamenkovich et al., 2021; Uchida et al., 2023). More broadly, most of these approaches do not fully leverage the information available in high-resolution simulations to move beyond the diffusive framework.

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, for example isotropic diffusion or biharmonic friction, that are at best able to mimic the bulk effects of sub-grid phenomena, ML methods directly learn a numerical mapping between the sub-grid effects and the large-scale fields using appropriate data. It is also possible to design ML models to obey physical properties through modifications to the architectural design or the loss function (Beucler, Pritchard, Rasp, et al., 2021; Zanna & Bolton, 2020). These methods have led to the development of parameterizations for idealized models (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 (Bodner et al., 2025; Ramadhan

et al., 2020; Sane et al., 2023), and momentum tendencies in the ocean (Perezhogin et al., 2024; Zanna & Bolton, 2020; Zhang et al., 2023). Many of these new parameterizations have demonstrated improvements over conventional parameterizations in various metrics, including offline prediction accuracy, reduced biases in mean state and variability, and improved energy spectra in online simulations, though the specific baselines and metrics vary across studies. The potential issues with implementation, stability and generalization remain an area of rapidly evolving research (Zanna et al., 2025). While many 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 or density field in the ocean. Our parameterization is designed to ensure that it maintains the important adiabatic constraint that was introduced by the GM parameterization, and is thus framed as an eddy-induced transport in terms of a streamfunction. However, unlike GM, our parameterization does not rely on diffusion and is not designed to be a local sink of APE; rather it is optimized to accurately capture the pointwise patterns of the eddy-induced fluxes, including their magnitude and direction, which can locally be upgradient or cross-gradient relative to the large-scale density fields. With an eye toward implementation, we use a small artificial neural network (ANN), which can be easily added to any ocean model code. We implemented this ANN into GFDL's Modular Ocean Model 6 (MOM6), and show its effects in idealized two-layer simulations where mesoscale eddies are the dominant turbulent process. In a follow up study, we will extend this work to a hierarchy of more realistic simulations. In Section 2 we describe the filtering framework that is used to diagnose the sub-grid impact of eddies from high-resolution simulations, and in Section 3 describe how to cast these sub-grid effects into a numerical mapping that potentially has some ability to generalize to unseen data. Also, in Section 3 we describe the ML architecture, and training process. In Section 4 we describe the high-resolution data sets and how they were processed. In Section 5 we show that our parameterization is successful in both offline and online tests, and finally, in Section 6 we conclude with a discussion, potential caveats and an outlook toward 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., Loose, Marques, et al., 2023; Vallis, 2017). The thickness equation, the primary focus of our study, can be written for each layer as,

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

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

These equations 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 that are too small to be resolved (sub-grid) using a combination of spatial filtering and coarse-graining $\overline{(\cdot)}$, as is routinely done in the large eddy simulation (LES) framework (Aluie et al., 2018; Sagaut, 2005). 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 n th layer thickness equation as,

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

where for theoretical convenience we have assumed that our spatial filtering and coarse-graining commutes with the derivatives.

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

$$\mathbf{F}_n = \overline{\mathbf{u}_n h_n} - \overline{\mathbf{u}_n} \overline{h_n} \quad (3)$$

as the sub-grid thickness flux, which will be the target of the parameterizations we develop below. In fact, it is common practice to represent this sub-grid thickness flux in terms of an eddy-driven streamfunction or velocity (described in Section S2 in Supporting Information S1), which is the approach taken here as well.

By parameterizing the flux itself, rather than its divergence, thickness conservation is ensured: the divergence is computed by the model's own numerical operators, and integrates to zero over the domain given appropriate boundary conditions. Targeting the flux rather than its divergence also has practical advantages for the optimization problem, as the flux field is smoother than its divergence (Yan et al., 2024).

Note that, considering resolved and sub-grid contribution in the momentum equation would result in complementary sub-grid forcing terms in the momentum equation as well. However in this study, we focus our attention only on the sub-grid thickness fluxes, as these correspond to one of the major parameterizations—the GM parameterization—in ocean models. Data-driven parameterizations of sub-grid momentum forcing were considered recently (Perezhogin et al., 2024; Zanna & Bolton, 2020; Zhang et al., 2023), and in a follow-up work we plan to target both thickness and momentum sub-grid parameterizations simultaneously.

3. Machine Learning Model Design and Implementation

The goal of this work is to develop a data-driven parameterization for the sub-grid thickness fluxes in terms of the resolved variables. Here, we use artificial neural networks (ANNs) with a few hidden layers to approximate this numerical mapping. These ANNs will be trained by using data from high-resolution simulations, which have been filtered and coarse-grained to diagnose the sub-grid thickness fluxes and the corresponding resolvable fields. The ML parameterizations will be tested in both offline and online evaluation settings.

3.1. Parameterization Function Design

ANNs are highly flexible function approximators, which makes them prone to overfitting—they can fit the data with functions or mappings that do not generalize. To avoid overfitting and allow for a degree of generalizability, we implement certain design constraints into the ANN.

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

$$\mathbf{F}_n = \mathbf{f}_\theta(\nabla \bar{\mathbf{u}}_n, \nabla \bar{h}_n, \Delta), \quad (4)$$

where $\nabla \bar{\mathbf{u}}_n$ is the velocity gradient tensor and $\nabla \bar{h}_n$ is the thickness gradient, both for the resolved fields. Δ is a measure of the grid scale, which allows the parameterization to be scale-aware. $\mathbf{f}_\theta()$ is an ANN with unknown parameters θ , which need to be learned, that represents the two components of the sub-grid thickness flux vector. Note that in the offline setting the inputs to this ANN will be the filtered and coarse-grained fields from the high-resolution simulation, which serve as a proxy for the resolved state of an ideal coarse model. In the online setting, no high-resolution data is available, and the inputs are instead the resolved fields produced by the coarse-resolution simulation itself.

The choice of input features—velocity gradients, thickness gradients, and grid scale—was motivated by the observation that existing parameterizations like GM and the velocity gradient model (VGM) can be expressed as functions of these same quantities (see Sections S3 and S4 in Supporting Information S1). This represents a minimal set of inputs that proved sufficient for the configurations tested here. In future work, additional inputs could be explored if the parameterization appears to consistently fail in a wider range of regimes.

Lateral non-locality/stencil-size: The GM parameterization and the VGM parameterization, a common model used in the LES literature, are horizontally local (see Sections S3 and S4 in Supporting Information S1), that is parameterization output at a given horizontal grid point depends only on inputs from that same grid point. Denoting horizontal grid indices as (i, j) , this locality means the parameterized flux depends only on inputs at (i, j) .

Here, we relax this and allow for a small degree of non-locality by considering a wider horizontal stencil, which allows input information to come from regions surrounding the point where the prediction needs to be made. This is mathematically denoted as $\mathbf{F}_{n,(i,j)} = \mathbf{f}_\theta(\nabla \bar{\mathbf{u}}_{n,(I,J)}, \nabla \bar{h}_{n,(I,J)}, \Delta_{i,j})$, 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 local model (1×1

stencil) $m = 0$, for a model with a 3×3 stencil $m = 1$, for a model with a 5×5 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) . For points near domain boundaries where the stencil extends outside the domain, the out-of-bounds inputs are set to zero. In principle, vertically non-local models can also be formulated, but these will not be considered here.

Using a wider input stencil here is distinct from employing higher-order finite difference schemes for computing the spatial derivatives. Higher-order schemes use neighboring values to improve the approximation of a derivative at the central point—the output remains a function of local differential quantities. Here, the values from the wider stencil are provided directly as inputs to the ANN, which is free to learn arbitrary nonlinear mappings from the spatially distributed input field. The ANN is thus not constrained to only approximate local derivatives more accurately; rather, it can capture dependencies on spatial patterns across the stencil that are not representable by any single-point derivative. In this sense, the extended stencil expands the functional space of the parameterization beyond what is achievable with purely local models, even those using high-order numerics.

We found that ANNs of the above form (Equation 4) with wide stencils can be trained to achieve high offline skill, but struggle when tested on data with distributional shifts relative to the training data (not shown). To improve generalization, we introduce a set of additional constraints on the ANN's inputs and outputs, described below. These constraints are motivated by physical reasoning and principles from past literature (Beucler, Pritchard, Yuval, et al., 2021; Perezhogin et al., 2025), but are ultimately validated empirically—we retain them because they measurably improved offline generalization across a range of test scenarios, not because they are provably necessary or sufficient.

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 sub-grid thickness flux directed in the y -direction results in an APE reduction. This model may fail if the setup is rotated by 90° —even though we expect the dynamics 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 Section S5 in Supporting Information S1 for details). We denote variables in the flow dependent frame as $\tilde{(\cdot)}$. 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 $\nabla \tilde{h}_n$. This frame of reference may also be conceptually advantageous, as the projection of the sub-grid 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: Casting the inputs and outputs into non-dimensional forms provides several practical benefits for the ANN (Perezhogin et al., 2025). Here we use the following non-dimensional forms:

$$\mathbf{F}_n / (\Delta^2 |\nabla \tilde{\mathbf{u}}_n| |\nabla \tilde{h}_n|), \nabla \tilde{\mathbf{u}}_n / |\nabla \tilde{\mathbf{u}}_n|, \nabla \tilde{h}_n / |\nabla \tilde{h}_n|.$$

Here $|\nabla \tilde{\mathbf{u}}_n|$ is the Frobenius norm of the velocity gradient tensor. For a 1×1 stencil this is

$$|\nabla \tilde{\mathbf{u}}_n| = \sqrt{(\partial_x \tilde{u}_n)^2 + (\partial_y \tilde{u}_n)^2 + (\partial_x \tilde{v}_n)^2 + (\partial_y \tilde{v}_n)^2},$$

and for a larger stencil this takes the form

$$|\nabla \tilde{\mathbf{u}}_n| = \sqrt{\sum_{p=-m}^{p=m} \sum_{q=-m}^{q=m} [(\partial_x \tilde{u}_{n,(i+p,j+q)})^2 + (\partial_y \tilde{u}_{n,(i+p,j+q)})^2 + (\partial_x \tilde{v}_{n,(i+p,j+q)})^2 + (\partial_y \tilde{v}_{n,(i+p,j+q)})^2]},$$

where m is the stencil size.

Normalizing inputs by their magnitudes ensures that all normalized input variables are bounded in the range $[-1, 1]$, constraining the input domain and reducing the risk of extrapolation. 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 Frobenius norm also reduces the degrees of freedom in each input variable group by one.

The non-dimensionalization for the output ensures that the parameterization is invariant to the choice of units, as the ANN only operates on dimensionless ratios and the dimensional scales are absorbed into the pre-factor $\Delta^2 |\nabla \bar{\mathbf{u}}_n| |\nabla \bar{h}_n|$. Because the dimensional scaling is fixed by this pre-factor, it is embedded in the construction by design: for a given set of (non-dimensional) input profiles, the ANN learns only the dimensionless mapping $\mathbf{f}_\theta$ and is neither required nor able to learn the dimensional scaling from data. Non-dimensionalization may potentially aid generalization across regimes where the energy levels differ but the underlying dynamics are similar.

Final constrained form. The above considerations, result in the following constrained ANN:

$$\tilde{\mathbf{F}}_{n,(i,j)} = \Delta_{i,j}^2 |\nabla \bar{\mathbf{u}}_{n,(I,J)}| |\nabla \bar{h}_{n,(I,J)}| \mathbf{f}_\theta \left(\frac{\tilde{\nabla} \bar{\mathbf{u}}_{n,(I,J)}}{|\nabla \bar{\mathbf{u}}_{n,(I,J)}|}, \frac{\tilde{\nabla} \bar{h}_{n,(I,J)}}{|\nabla \bar{h}_{n,(I,J)}|} \right), \quad (5)$$

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 these choices make the ANN more restrictive than Equation 4, they were made after significant amount of trial and error. We found that the mappings estimated under these constraints are skillful and generalized better.

3.2. Neural Network Architecture, Hyper-Parameters, and Software

As mentioned above, we use an ANN to learn the numerical mapping $\mathbf{f}_\theta()$. These ANNs are specifically multi-layer perceptrons (MLPs)—fully-connected, feed-forward networks applied independently at each grid point (or over a local stencil)—in contrast to the convolutional architectures used in some previous ocean parameterization studies; we retain the broader term ANN for brevity throughout. 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 an affine transformation to its inputs, followed by a non-linear 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 ANN design and training choices, as detailed in Section S6 in Supporting Information S1. 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 ANN models that differ only in stencil size: 1×1 , 3×3 , and 5×5 . 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 (1×1), to 5,090 (3×3), to 9,698 (5×5). The models take non-dimensionalized velocity and thickness gradients as input and predict non-dimensionalized sub-grid thickness fluxes as output. Both inputs and outputs are rotated into local thickness gradient coordinates. Also, in addition to the non-dimensionalization—which removes units and normalizes each gradient group to unit magnitude, but leaves the individual features within a group with systematically different typical sizes—all input and output features were further normalized by the order of magnitude of their standard deviations. This ensures that all features seen by the ANN during training are $O(1)$, which improves the conditioning of the loss landscape and helps the optimizer learn effectively from all features.

The details of the Double-Gyre and Phillips-2-Layer high-resolution simulations are presented in Section 4, and the details of the training set choices are the following. Both these simulations were run for 100 years, and a

snapshot was saved every 10 days—producing 3,600 snapshots for each simulation. Training was performed on the first 2,048 snapshots from both the Double-Gyre and Phillips-2-Layer simulations, and included 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 2,400–3,600.

We used Python and JAX (Bradbury et al., 2018) for all our ML pipelines. Specifically, we used Flax-Linen library (Heek et al., 2024) for the design of our ANNs and used the Optax library (DeepMind et al., 2020) 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 ANN-based sub-grid thickness flux parameterization was implemented in MOM6 via two new modules: an ANN module and a thickness flux prediction module. The ANN 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 ANN. This module is general-purpose and can be called from anywhere within MOM6, enabling the integration of multiple ANN-based data-driven models into the codebase. All ANN inference is performed inline within MOM6's Fortran on CPU, with no GPU or external machine-learning runtime; the simulations themselves are likewise run on CPU.

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 3×3 model, this means that each input in the 3×3 stencil surrounding the prediction point is evaluated at the 3×3 grid cell centers. At each central point, the input gradients are first rotated into the local thickness-gradient frame and non-dimensionalized before being passed to the ANN. The two flux components predicted by the ANN are produced in this rotated, non-dimensional frame; they are then rotated back to grid (x, y) coordinates and re-dimensionalized (via the pre-factor in Equation 5) before being interpolated to the appropriate edge locations. The divergence of these edge fluxes is subsequently computed by MOM6's standard operators in grid coordinates—the flow-dependent rotation therefore enters only in the pointwise network evaluation, and the interpolation to edges is performed after the dimensional scaling has been restored.

We also introduced a non-dimensional coefficient (C_{ANN}) which can be used to adjust the strength of the parameterized flux. The sub-grid thickness fluxes depend on the choice of filter scale, which determines the split between resolved and parameterized contributions. It is not obvious what filter scale best corresponds to a particular coarse grid resolution, as this depends on the numerics of the coarse model and the properties of the filter used to diagnose the sub-grid thickness fluxes—even in the LES literature this remains a matter of pragmatic choices (Lund, 2003). Due to this mismatch, the resolved fields in a coarse simulation differ from the filtered high-resolution fields used during training, and C_{ANN} provides a simple way to compensate for this distributional shift without retraining. In practice, we found that modest adjustments to C_{ANN} could improve online performance relative to the default $C_{ANN} = 1$ case, making it a useful hyperparameter. However, the fact that optimal values are $O(1)$ across resolutions and setups (Section 5) indicates that the ANN captures the correct magnitude of the sub-grid thickness fluxes to leading order, and only a minor adjustment is needed.

Fluxes at solid boundaries are set to zero to ensure volume conservation, which together with the fact that the ANN predicts the flux rather than its divergence, ensures that no net thickness is created or destroyed in any layer. This conservation property is intrinsic to the parameterization approach of predicting the flux and is independent of the additional constraint that the depth-integrated thickness flux vanishes (discussed below).

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 ANN 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 Phillips-2-Layer simulation does not

contain vanishing layers, the lower layer in the Double-Gyre simulation can vanish, as discussed in the next section. We tested the sensitivity of the results to different values of ϵ between 1 and 20 m and found the simulation outcome to be relatively insensitive to this choice. In all the simulations discussed below we chose $\epsilon = 20$ m.

While in principle, the boundary conditions described above should suffice for layered models (Killworth, 2001), MOM6 and other layered models will often additionally enforce the constraint that the vertical sum of the sub-grid thickness fluxes within a water column must be zero. This is equivalent to imposing the physical constraint that the parameterization does not directly influence the barotropic mode—perturbations to the sea surface height (SSH). This constraint ensures that the parameterization does not interact with the fast modes, an interaction that is physically expected to be weak for mesoscale flows. In both, layered or Z-level models, this is accomplished by expressing the sub-grid thickness fluxes in terms of a streamfunction (see Section S2 in Supporting Information S1) and setting the streamfunction to zero at the top and bottom boundaries using some 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 Section S7.4 in Supporting Information S1), 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—two-layer model of Phillips baroclinic instability, described by Hallberg (2013); and Double-Gyre—two-layer wind driven double gyre, described by Zhang et al. (2023).

The Phillips-2-Layer simulation uses a zonally-reentrant channel of 1,200 km $\times$ 1,600 km with solid walls at the northern and southern boundaries and a flat bottom at 2,000 m depth. The density difference between layers is $\Delta\rho = 0.002\rho_0$, and the Coriolis parameter varies linearly from $6.49 \times 10^{-5} \text{ s}^{-1}$ to $9.69 \times 10^{-5} \text{ s}^{-1}$ following a β -plane approximation. The simulation is forced by restoring the zonal-mean interface height toward a specified profile that corresponds to a baroclinically unstable zonal jet, on a timescale of 10 days. The deformation radius ranges from approximately 45 km at the southern boundary to 20 km at the northern boundary.

The Double-Gyre simulation uses a bowl-shaped basin spanning 22° in longitude and 20° in latitude (from 30°N to 50°N) on a sphere, with a maximum depth of 2,000 m and an interface initially at 1,000 m depth. The reduced gravity is $g' = 0.0098 \text{ m/s}^2$, and the deformation radius ranges from approximately 30 km in the south to 15 km in the north. The simulation is forced by a steady zonal wind stress that varies sinusoidally with latitude, with a maximum of $\tau_0 = 0.1 \text{ N/m}^2$ at the center latitude (40°N) and zero stress at the northern and southern boundaries. The circulation reaches statistical equilibrium after approximately 5 years.

Both these idealized 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 High-Resolution Simulation Data for ML Training and Evaluation

Filtering and Coarsening: Here we describe how we diagnosed the input and output fields 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 (Abernathy et al., 2022). Also regions with layer thickness smaller than 20 m were masked and treated as land, which only impacted the lower layer in the Double-Gyre 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

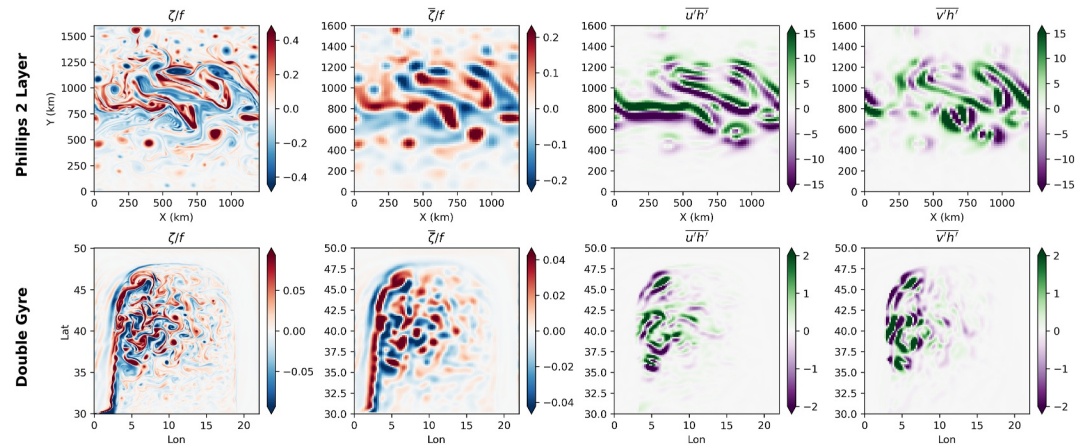


Figure 1. Snapshots exemplifying the high-resolution vorticity (left), filtered vorticity (second column), and sub-grid thickness fluxes (last two columns) from the Phillips-2-Layer (top) and Double-Gyre (bottom). The filtered fields are shown from the case of coarse-graining scale of 20 km. Only the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

“simple fixed factor filter” option in gcm-filters, which does an area weighting. In the Phillips-2-Layer simulation, with a Cartesian grid of size $\Delta_g = 4$ 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 Double-Gyre simulation, with a non-uniform grid of size $\Delta_g = 1/20^\circ$ this leads to variable filter scales of sizes $L_f = 0.55^\circ, 1.10^\circ, 2.20^\circ$ and 4.40° , which are approximately equal to the scales used for filtering Phillips-2-Layer.

After all the filtered variables and the corresponding sub-grid thickness 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 sub-grid thickness 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 $\Delta_c = 40$ km. This choice of the filter to grid ratio ($FGR = L_f/\Delta_c$) 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 up to 5 times in size (e.g., notice the small-scale bump in the EKE spectrum in Figure 7). However, note that this is a heuristic choice, and can be varied and 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 $\Delta_c = 10, 20, 40$, and 80 km. However this is only for the convenience of presentation, and in the computations, where Δ_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 high-resolution simulation to make it match a low-resolution simulation in some objective sense. Hence, our approach to diagnosing filtered and coarsened data is relatively adhoc, and based on pragmatism. One deliberate choice, however, is the use of explicit filtering (Gullbrand & Chow, 2003; Lund, 2003) to define the sub-grid thickness fluxes: the training targets are computed entirely from the filtered high-resolution fields and do not involve the coarse model's numerical operators. This contrasts with implicit filtering approaches, where the sub-grid forcing is defined as the residual between the filtered high-resolution tendency and the tendency computed by the coarse model's numerical scheme applied to coarse-grained fields (Ghosal, 1996), and thus includes discretization errors. We believe and hope that the impact of these choices would likely be minimal, with the acknowledgment that a big shortcoming 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—this is the reason to have introduced the tuning parameter C_{ANN} .

Layer thickness decomposition: When computing the thickness fluxes, special care was taken when the bottom topography (η_b) was not flat (only matters in the Double-Gyre 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

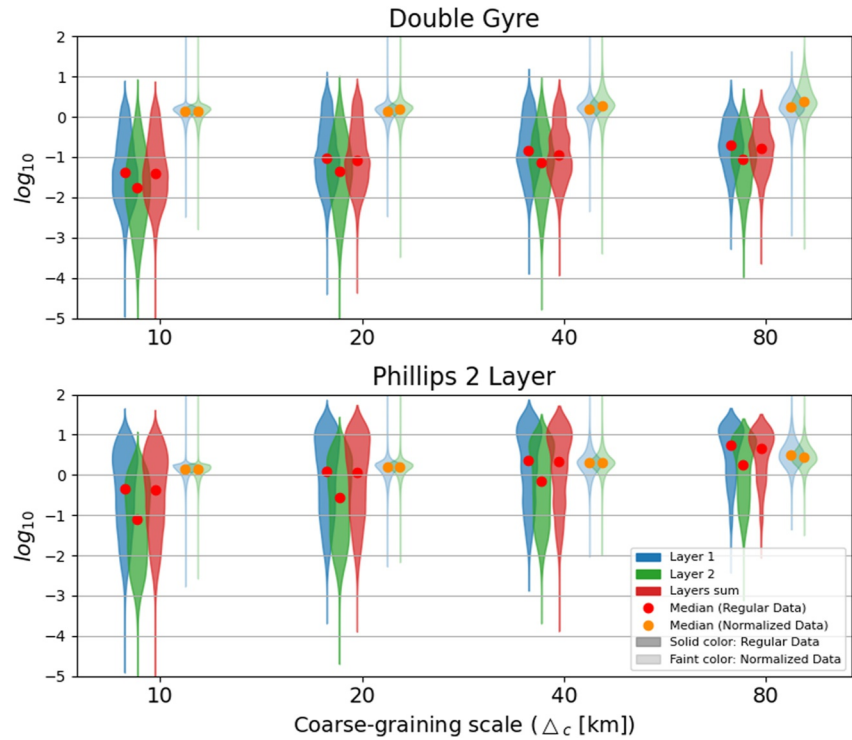


Figure 2. Distribution of output data: Distributions of the logarithm of the thickness flux magnitudes ($|\mathbf{F}_n|$) in both layers and their sum (barotropic contribution), and the normalized thickness flux ($\frac{|\mathbf{F}_n|}{\Delta_c^2 \|\nabla \mathbf{u}_n\| \|\nabla \mathbf{h}_n\|}$) in 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.

high-resolution topography. So we chose to use the condition that $\bar{\eta}_b = \eta_b$. Thus, thickness fluxes were dealt with by filtering interface heights only. To be more precise, in the bottom layer the filtered thickness would be $\bar{h}_N = \bar{\eta}_{N-1/2} - \eta_b$ and the filtered advection would be $\bar{\mathbf{u}}_N = \bar{\mathbf{u}}_{N-1/2} - \bar{\mathbf{u}}_b$, where half-integers correspond to the interfaces around the n^{th} layer—e.g. $\eta_{N-1/2}$ is the upper interface height of the bottom layer N .

Furthermore, we decomposed thickness gradient into a deformable and a non-deformable part associated with topography: $\bar{h}_n = \bar{h}_n^{\text{deformable}} + \bar{h}_n^{\text{topo}}$. When applied to a layer where the lower interface is topography, for example $\bar{h}_N = \bar{\eta}_{N-1/2} - \eta_b$, we get $\nabla \bar{h}_N^{\text{topo}} = -\nabla \eta_b$ and $\nabla \bar{h}_N^{\text{deformable}} = \nabla \bar{\eta}_{N-1/2}$. In other layers, the topographic contribution is zero ($\nabla \bar{h}_n^{\text{topo}} = 0$) and the full layer thickness contributes to the deformable part ($\nabla \bar{h}_n^{\text{deformable}} = \nabla \bar{\eta}_{n-1/2} - \nabla \bar{\eta}_{n+1/2} = \nabla \bar{h}_n$). Any unchanging or slowly changing flow features generated due to topography, like standing waves, will be included in the deformable component as they have a signature on the layer interfaces.

This decomposition allows us to distinguish the impact of dynamic layer thickness variations and bottom topography on the sub-grid thickness fluxes. In this study, we only use the deformable contribution as inputs to our ANN. Empirically, we found that using only the deformable component led to better offline skill than including the full thickness gradient. Physically, the interaction between mesoscale eddies and topography is complex and energetically constrained—for example, eddies do not simply act to homogenize layer thickness over topography (Adcock & Marshall, 2000)—and adequately representing these effects likely requires additional considerations beyond what is attempted here. Furthermore, in these two-layer configurations the upper layer flux is set to ensure a zero barotropic streamfunction (Section 3.3), so the parameterization effectively relies on the bottom layer alone. Using only the deformable thickness gradient in this layer mirrors the spirit of traditional parameterizations, which diffuse the interface height rather than the full layer thickness. A more complete treatment of topographic effects on sub-grid thickness fluxes is left for future work. We henceforth

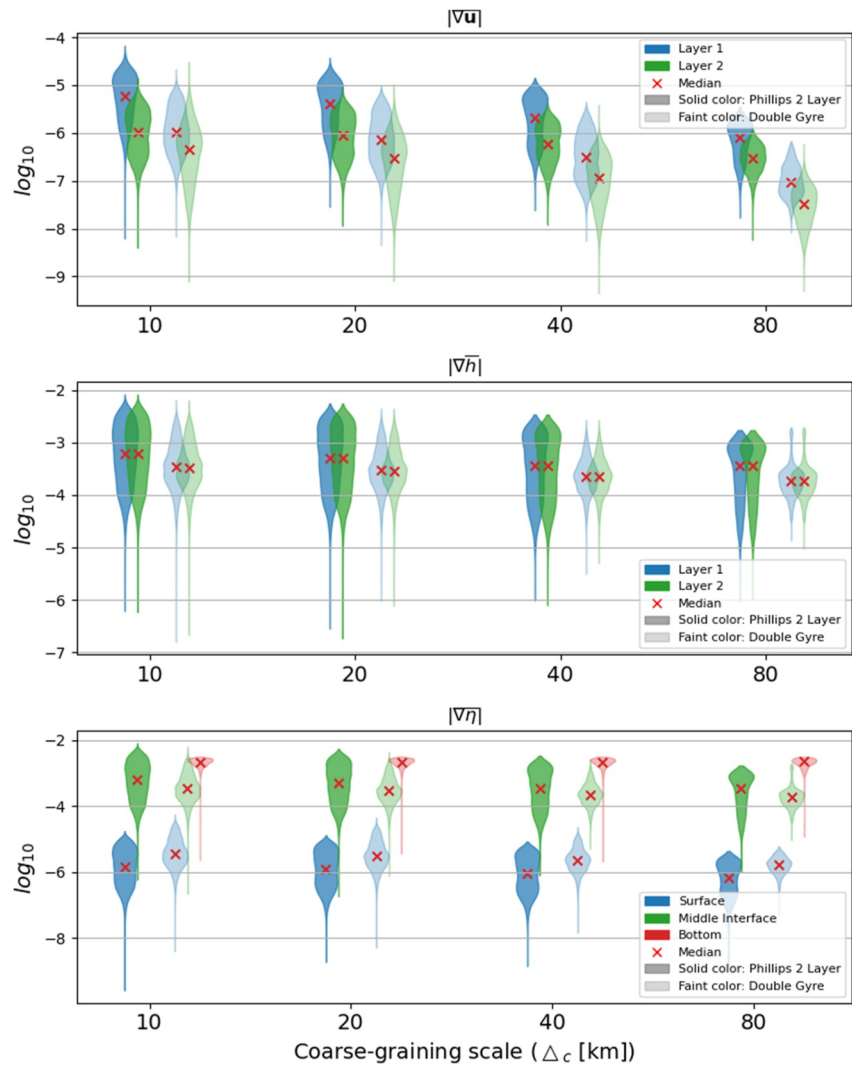


Figure 3. Distribution of input data: Distributions showing the logarithmic magnitude range of different filtered fields, which may be used as input variables for neural network design from both simulations and at different layers and interfaces. Similar to Figure 2 the width of the patch corresponds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

replace the notation for the deformable part $\nabla \bar{h}_N^{\text{deformable}}$ by $\nabla \bar{h}_N$ 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 Phillips-2-Layer and Double-Gyre simulations produce a turbulent flow field, with a rich array of eddies and jets (Figure 1). The magnitude of the sub-grid thickness fluxes are greater in the top layer than the bottom layer, and are larger in Phillips-2-Layer than Double-Gyre (Figure 2). Also the sub-grid thickness 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 50–400 km (coarse grid scales of 10–80 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 ML model training and generalization. However, note that the width of non-dimensionalized flux distribution does increase 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 streamfunction takes a boundary condition of the zero at the surface, which is equivalent to saying that there is no barotropic (depth integrated) sub-grid thickness flux. 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 sub-grid thickness flux is far from negligible, its impact on the APE tendency, which is the primary target of our parameterization, is very small (see Section S7.4 in Supporting Information S1).

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 Phillips-2-Layer simulation has stronger velocity gradients than the Double-Gyre 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 3), the deformable thickness gradients are essentially the same for the two layers. Also, the deformable thickness gradients are slightly weaker in the Double-Gyre simulation relative to the Phillips-2-Layer simulation. The bottom topography slopes in the Double-Gyre 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

This section describes the bulk physical properties of the simulations that will help set the stage for the idealized systems under consideration here, and serve as benchmarks for evaluating online parameterization skill in Section 5 (also see Section S7 in Supporting Information S1 for definitions of the metrics considered here and in later sections). The 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 the bulk or large scale behavior that emerges in this simulations, which is a function of the simulation setup, external forcing, and the feedback from the eddies.

In the Phillips-2-Layer 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 on average 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 level of the overturning circulation is maintained and APE is removed (Hallberg, 2013).

In the Double-Gyre 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 (40°N) 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. Unlike Phillips-2-Layer, there is no eddy driven overturning circulation here since there are no diabatic effects.

5. Results

Data-driven models may be evaluated along two interconnected aspects: offline skill and online skill. Here offline skill refers to the accuracy in predicting sub-grid 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

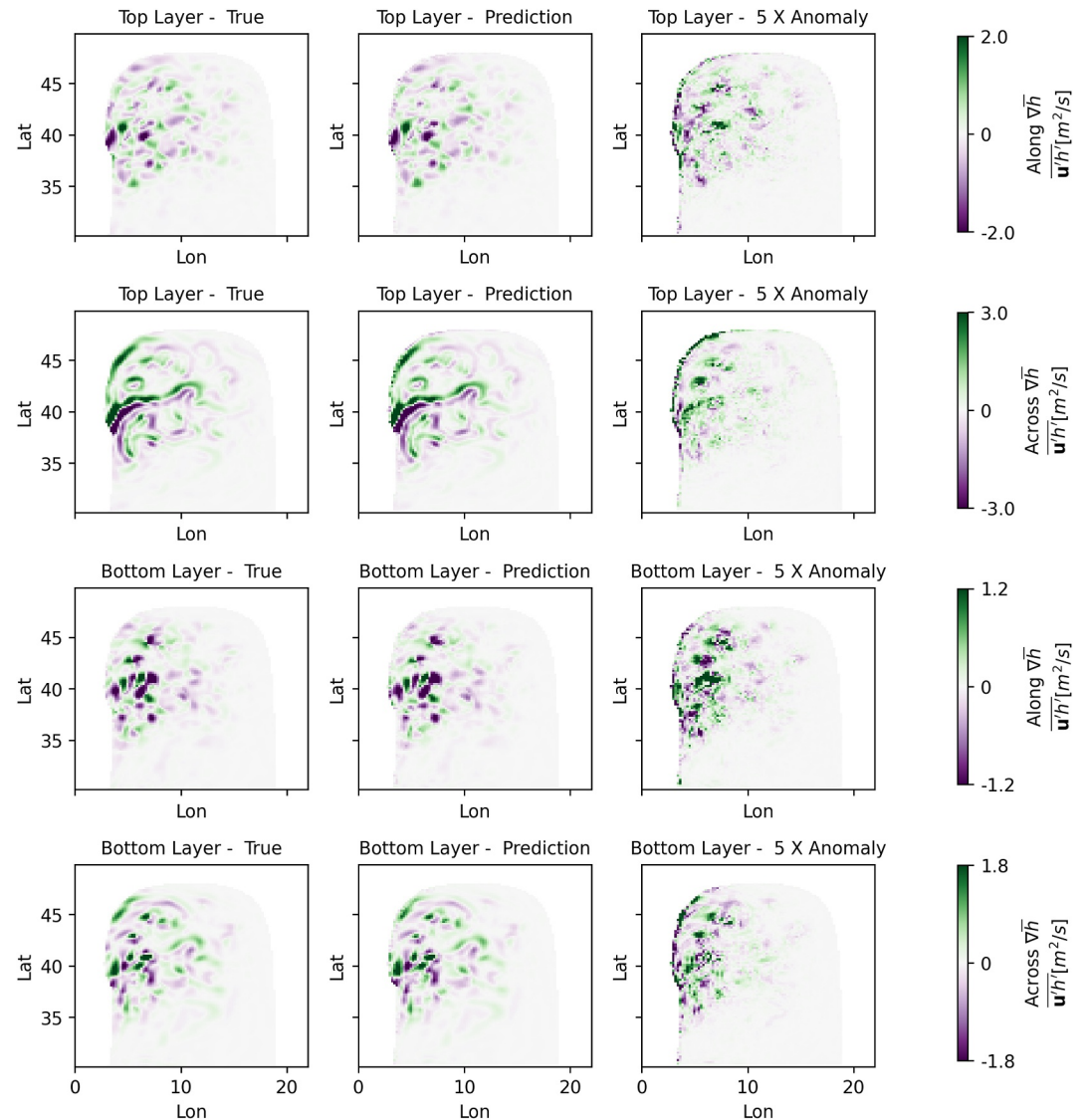


Figure 4. The true (first column), predicted (second column), and prediction error (third column) in the along (first and third row) and across (second and fourth row) thickness gradient fluxes for the top (first and second row) and bottom (third and fourth row) layers of the Double-Gyre simulations. Here the results are shown for the ML model of the following configuration: 3×3 stencil, trained on data from Double-Gyre + Phillips-2-Layer, and 5090 learnable parameters. These offline skill results are shown for the filter scale of 100 km. 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.

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.

Online skill is ultimately what we care about: does the parameterization improve the resolved scales in the simulation? Traditional parameterizations are typically designed with only this question in mind, and may achieve good online performance in some metrics through relatively simple closures that capture the bulk effects of the sub-grid processes without necessarily reproducing their detailed structure. Offline skill, on the other hand, measures whether the parameterization produces the right flux patterns in the right places for the right reasons—essentially quantifying the degree to which it faithfully represents the detailed, local nature of the sub-grid effects. A data-driven approach offers the possibility of achieving good online skill through high offline skill, that is by getting the large-scale behavior right because the underlying sub-grid processes are being represented more faithfully.

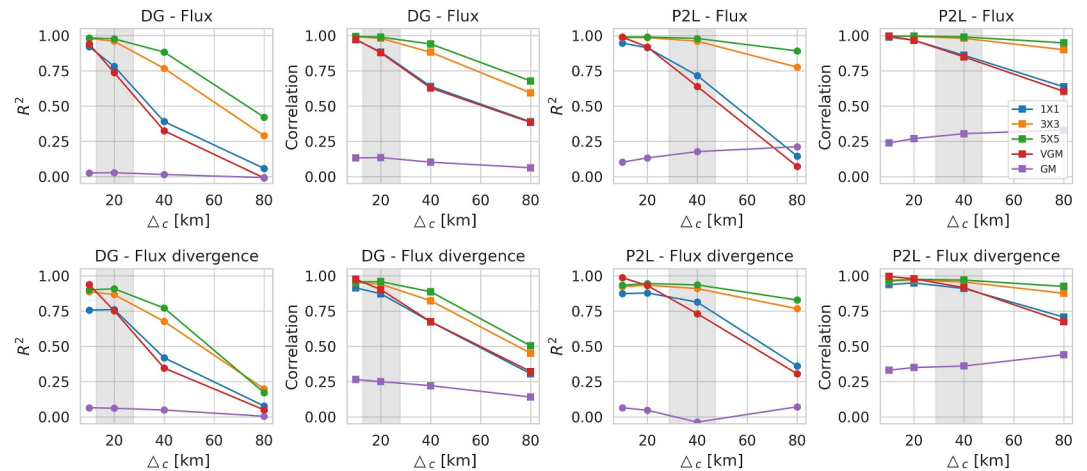


Figure 5. Pointwise offline skill in terms of the R^2 skill score (first and third columns) and correlation (second and fourth columns) for 5 different models (indicated in the legend) at different scales (x-axis) in the Double-Gyre (DG, first and second columns) and Phillips-2-Layer (P2L, third and fourth columns) simulations. The top row shows skill for the sub-grid thickness flux and the bottom row shows skill for its divergence. The metrics were evaluated over the bottom layer and across the full test data set spanning 33 years (1,200 snapshots). The gray shaded area indicates the deformation radius range (10th–90th percentile) in the simulation.

In both the offline and online evaluations below, we compare our ANN models against the GM parameterization with a constant diffusivity and the VGM parameterization (Khani & Dawson, 2023). We emphasize that these represent the simplest forms of these conventional approaches. More advanced specifications of the GM coefficient—including energy-budget-based closures like MEKE (Jansen et al., 2019) and GEOMETRIC (Mak et al., 2018), as well as scale-aware formulations (Mak et al., 2023)—have been proposed and shown to provide improved state-awareness. Comparisons with these more advanced schemes are beyond the scope of the present work, which aims to introduce the data-driven framework and demonstrate its potential. We chose constant-coefficient GM as the baseline because it remains a commonly used form in ocean models, and because the structural limitation we aim to address—the restriction to downgradient diffusive fluxes—is shared by all diffusive closures regardless of how the coefficient is specified.

5.1. Offline Evaluation

Here we discuss three particular ANN models that differ only in stencil size: 1×1 , 3×3 , and 5×5 (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 sub-grid thickness fluxes for each parameterization separately at each filter scale and for each simulation setup. This is in contrast to the ANNs, 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 (see Table S1 in Supporting Information S1), the performance of these conventional parameterizations would be worse if their parameters were estimated across the entire range of data simultaneously.

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

In a more quantitative sense, we find that the ANN skill, compared across all ANN configurations considered here and quantified using R^2 or correlation score (see Section S1 in Supporting Information S1 for definitions), depends on the input stencil size and the filter scale (top row of Figure 5); the skill also depends on other factors as discussed in Section S6 in Supporting Information S1 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 1×1 to

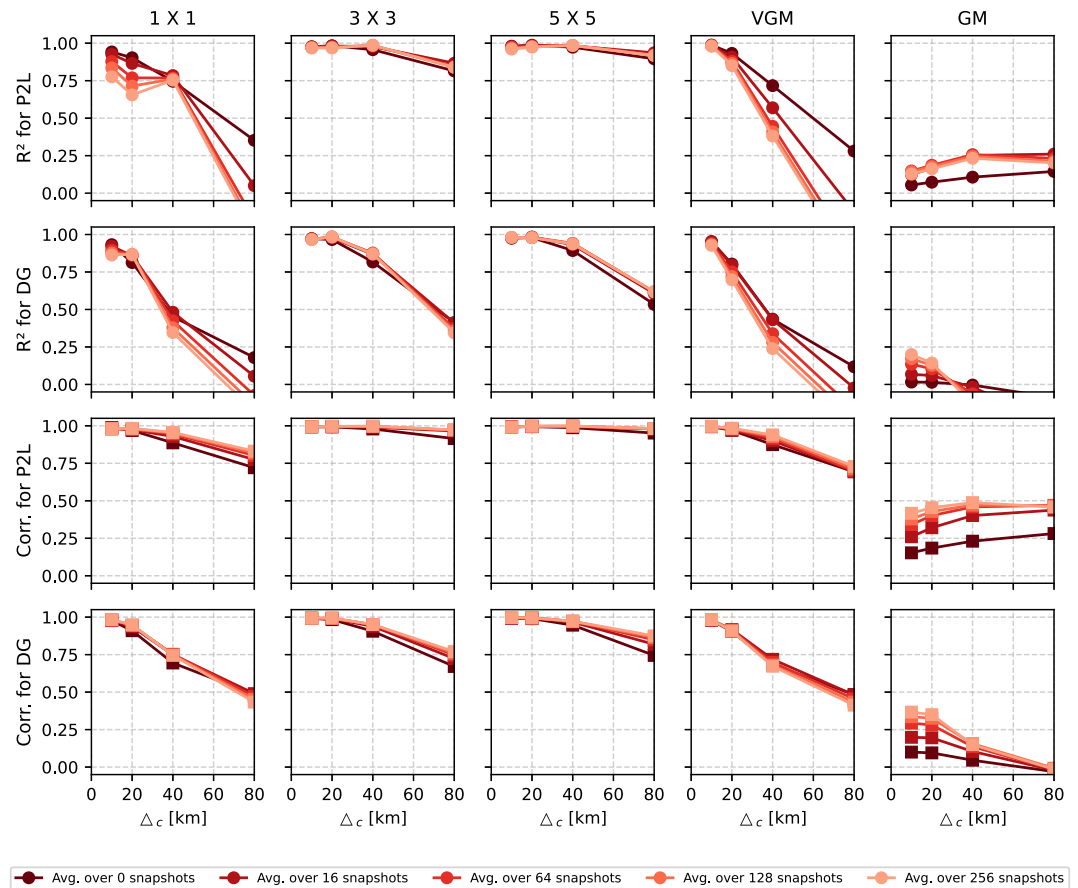


Figure 6. Offline skill in predicting the time average of the subgrid thickness fluxes, using five different models (columns) for the Phillips-2-Layer (P2L) and Double-Gyre (DG) simulations. For all models the skill in predicting the time averaged flux is quantified as the average skill over the two layers and for both the along and across thickness gradient direction.

3×3 and a smaller improvement when going further to 5×5 . Also, model skill decays as filter scales get larger, which is more rapid in the case of Double-Gyre relative to Phillips-2-Layer. We suspect that this might be due to different deformation radii in the different simulations, on average 40 km in Phillips-2-Layer and 20 km in Double-Gyre (Figure S5 in Supporting Information S1), and ML models seem to be less skillful offline 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 ANNs against the conventional parameterizations. The VGM parameterization, which is sometimes used in LES studies, has essentially the same point-wise skill as our 1×1 ML model. The GM parameterization in contrast has very little point-wise skill (both R^2 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 sub-grid thickness fluxes. Notice that in contrast, the skill of the 3×3 and 5×5 models is significantly higher than both the 1×1 ML model and the VGM parameterization.

It is not the sub-grid thickness flux itself but its divergence that enters as a forcing in the thickness equation (Equation 2). The flux can contain large rotational (non-divergent) contributions that have zero divergence and therefore no dynamical impact, so high skill in predicting the flux does not guarantee high skill in predicting what matters dynamically (J. Marshall & Shutts, 1981). It is therefore important to verify that the models also predict the flux divergence well. When evaluated on the flux divergence (bottom row of Figure 5), all models show slightly reduced absolute skill compared to the flux itself, which is expected as the divergence operator amplifies small-scale errors in the flux field. However, the relative ranking of the models is largely preserved: the 3×3 and 5×5 ANN models continue to outperform the 1×1 model, VGM, and GM across most scales. The GM parameterization shows some improvement in correlation for the flux divergence relative to the flux, particularly

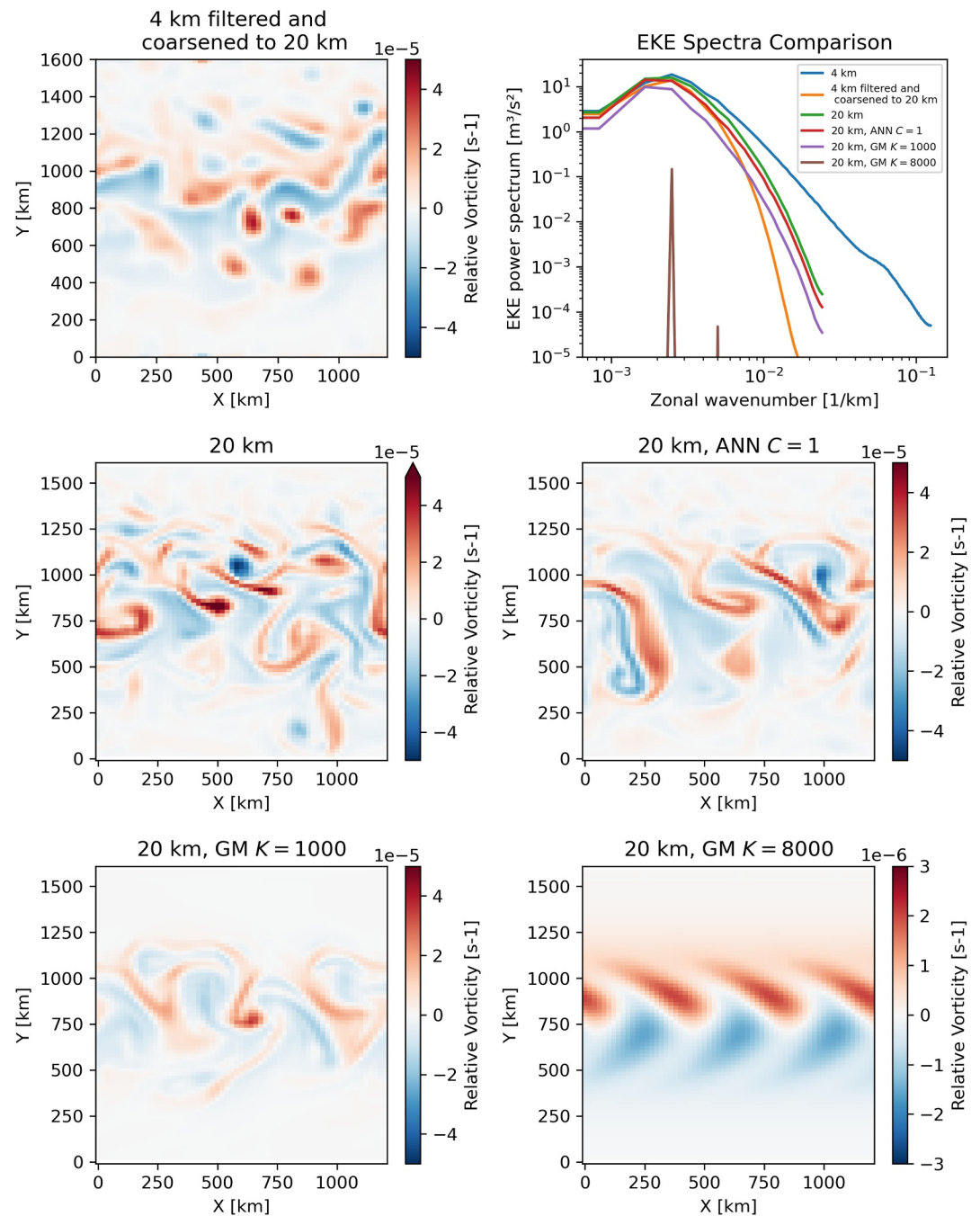


Figure 7. Snapshots of relative vorticity and the EKE spectra from different Phillips-2-Layer simulations. Note that the colorbar on the bottom right panel has been adjusted to show the range of values in that simulation.

at coarser scales. Nevertheless, its R^2 remains low, indicating that while it may get the sign of the divergence somewhat right, it does not capture its magnitude or patterns well. These results confirm that the high offline skill of the ANN models is not an artifact of fitting dynamically irrelevant rotational flux components.

Bulk (time-averaged) offline 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. The skill for the divergence of flux follows the same patterns.

The temporal averaging is started at snapshot 2,400, and extended longer using the snapshots that are available every 10 days.

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 Phillips-2-Layer data but not for the Double-Gyre data. Even for the Phillips-2-Layer 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 skill of 0.5. The reason for the discrepancy between Double-Gyre and Phillips-2-Layer arises because GM did not turn out to be a good model for upper layer fluxes in the Double-Gyre 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 Double-Gyre data is higher, but still not as large as for Phillips-2-Layer data (not shown).

The 1×1 ANN, and similarly the VGM parameterization, have R^2 skill decrease with increasing temporal averaging. This indicates that while these models are able to learn some aspects of the distribution, the mean is poorly captured. The correlation is not impacted as severely by the averaging, suggesting that the decrease in skill has less to do with the spatial patterns and more to do with the amplitude. In contrast the 3×3 and 5×5 ANNs have almost no impact on either R^2 or correlation skill with averaging, and 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 the largest filter scales in Double-Gyre, also shows that these models do very well in predicting both the along and across gradient fluxes.

Overall, we found that the offline skill of the ANN models, particularly those with wider stencils, is very promising. Also, we show in Figure S4 in Supporting Information S1 that an ANN trained on data from one simulation shows relatively high skill when tested on the unseen simulation, which is not the case if we directly use an ANN (Equation 4) rather than the non-dimensionalized form (Equation 5). Thus, these new models are scale and context aware, and seem to show some degree of generalization when testing over different data sets. This is in contrast to the traditional models, which are unable to match the ANN skill even after the corresponding coefficients were estimated separately for each scale and data set.

In the online setting next we only evaluate the skill of the 3×3 ANN discussed above, as this model provides a good compromise between computational cost and offline skill (skill of the 5×5 ANN is only marginally better than the 3×3 ANN). Also since the offline performance of the 1×1 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 ANN in the Phillips-2-Layer 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 Section S8.1 in Supporting Information S1).

In this section, we focus the discussion on the 20 km eddy-permitting resolution simulations—both parameterized and unparameterized—and compare them with a 4 km high-resolution 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 fully resolves them. The parameterized simulations were selected to approximately match the total overturning transport of the high-resolution case. For the GM parameterization, we include two configurations: one with low diffusivity ($1,000 \text{ m}^2/\text{s}$), often cited as a canonical value in the literature, and another with high diffusivity ($8,000 \text{ m}^2/\text{s}$), chosen because it yields an overturning close to the high-resolution simulation. These values are not fine-tuned to exactly match the high-resolution overturning, but instead represent two qualitatively distinct regimes. Notably, $1,000 \text{ m}^2/\text{s}$ also marks a threshold beyond which the GM scheme begins to strongly damp the resolved eddy field (also see Figure S6 in Supporting Information S1). In contrast, the ANN-based parameterization with a tuning coefficient (C_{ANN}) of 1 produced overturning transport closely aligned with the high-resolution simulation. All simulations reached a spun-up state within 2 years, and time-averaged statistics are computed over years 2–10.

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. Broadly, the variability in the 20 km simulations without any parameterization and with the ANN 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 high-resolution filtered and coarsened simulation state matches the low-resolution unparameterized simulation at large scales, but has lower energy levels at smaller scales (see EKE spectral plot in Figure 7). This is the result of the specific properties of the Gaussian filter that was chosen to filter the simulation, and more refined filters could 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 high-resolution simulation produces approximately 13 Sv of transport, with 9.5 Sv attributed to scales larger than the filter scales and 3.5 Sv from sub-grid thickness fluxes. In contrast, the unparameterized 20 km simulation produces only about 10.5 Sv of transport. The 20 km simulation with the ANN 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 ANN and the high GM diffusivity case to exactly match the high-resolution 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 Section S7 in Supporting Information S1). An excellent parameterization should improve on multiple metrics simultaneously, rather than only working on the single metric it was tuned on. The KE and APE from various contributions are shown in Figures 8b and 8c. The unparameterized 20 km simulation exhibits lower KE and APE than the high-resolution simulation, as expected, and the low-resolution simulation KE and APE are close to the filtered KE and APE from the filtered high-resolution data. The 20 km simulation with the ANN 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 high-resolution simulation. The low GM diffusivity simulation shows lower KE and APE than both the unparameterized and ANN-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 sub-grid or parameterized fluxes is shown in Figure 8d. We display the contribution to both the MAPE tendency and the EAPE tendency. In the filtered high-resolution simulation, the sub-grid thickness fluxes contribute just over half of their tendency toward reducing the MAPE. The 20 km simulation with the ANN parameterization produces an APE tendency that is relatively close to the high-resolution 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 ANN in the 20 km simulation, where the ANN 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 KE (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 Figure S6 in Supporting Information S1, where the response to the parameterization coefficients is non-monotonic. Note

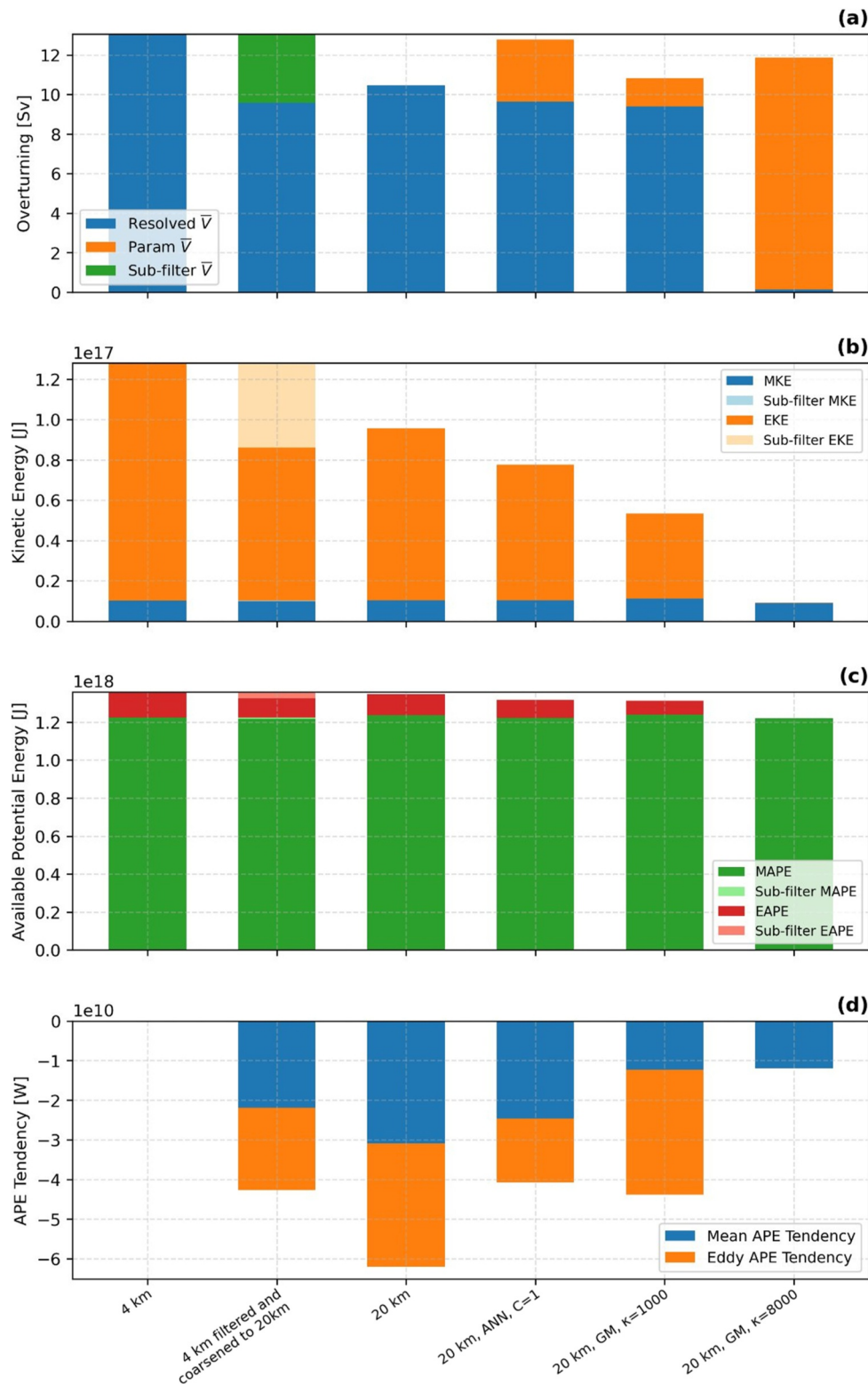


Figure 8. Bulk metrics to evaluate the online skill of the parameterizations in different experiments of Phillips-2-Layer simulation (indicated along the x-axis). (a) Peak value of the total meridional transport in the upper layer for the resolved, parameterized and sub-filter components. (b) Kinetic energy and (c) available potential energy (APE) of the mean and eddy flow coming from the resolved and sub-grid contributions. (d) The impact of the parameterized or sub-grid thickness 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 20 km simulation is calculated using sub-grid thickness fluxes predicted by a ANN that was not coupled with the resolved fields of the simulation.

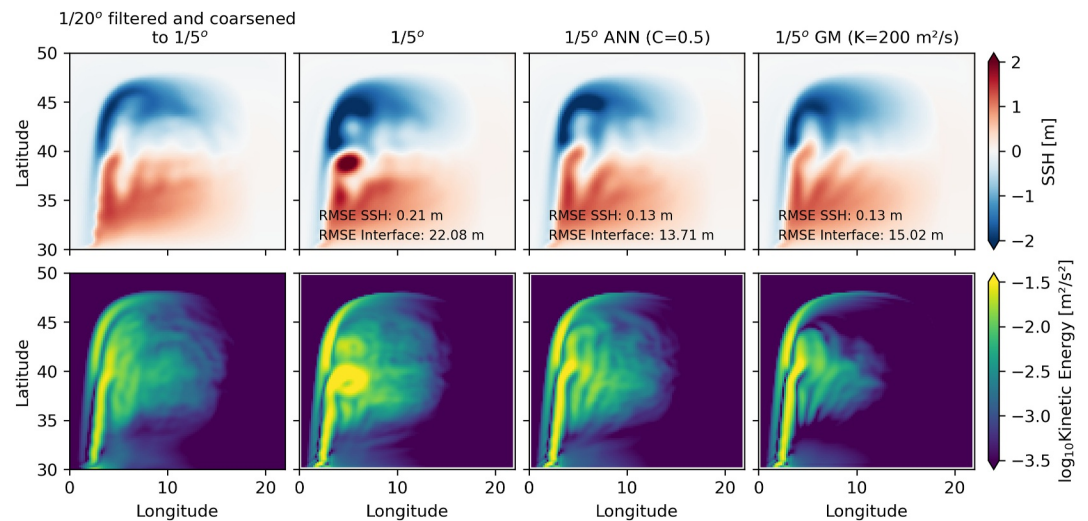


Figure 9. Mean sea surface height (SSH) (top row) and EKE (bottom row) for the Double-Gyre simulations. The RMSE in the SSH and middle interface height are indicated for the three $1/5^\circ$ resolution simulations.

that at eddy-permitting resolutions, even the response to the GM parameterization is non-trivial and non-monotonic.

In summary, while both the ANN and GM parameterizations can be tuned to produce approximately the correct overturning circulation in the Phillips-2-Layer simulation, only the ANN 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 ANN parameterization in the Double-Gyre simulation, a canonical system for studying wind-driven gyre dynamics. In contrast to the Phillips-2-Layer setup discussed earlier, the Double-Gyre exhibits strong boundary currents and pronounced spatial inhomogeneity in eddy statistics. Additionally, unlike the Phillips-2-Layer 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 Phillips-2-Layer: (a) mean-state biases can emerge as resolution and parameterizations are varied, and (b) the system is purely adiabatic, with no overturning circulation.

Similar to the Phillips-2-Layer case, we conducted a suite of simulations to assess the sensitivity of the Double-Gyre setup to both resolution and parameterization coefficients (see Section S8.2 in Supporting Information S1). Here, we focus on the $1/5^\circ$ (~ 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 ANN and GM parameterizations. The coefficients for these parameterized runs were selected to minimize the mean state error in SSH, which is strongly correlated with thermocline depth in this system.

The mean SSH and KE fields, averaged over years 3–13 of the simulations, are shown in Figure 9. In the $1/5^\circ$ 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 ANN- and GM-parameterized simulations, since the parameterizations coefficients were explicitly chosen to reduce this bias. However, consistent with the Phillips-2-Layer results, the ANN 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 $1/5^\circ$ simulation without parameterization is overly energetic compared to the high-resolution 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 ANN parameterization.

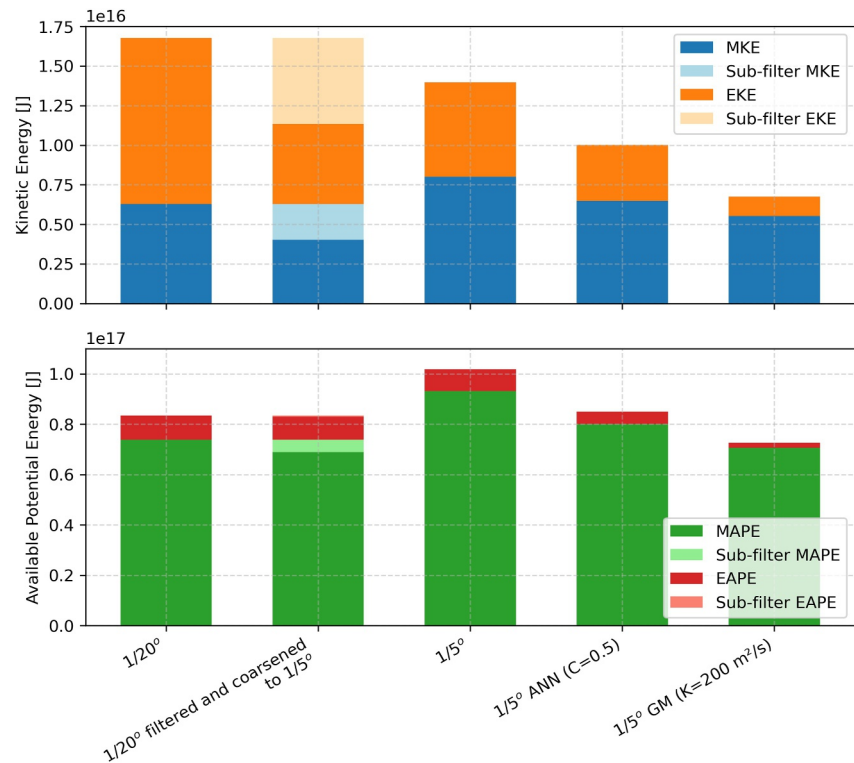


Figure 10. The volume integrated KE (top) and available potential energy (bottom) for the Double-Gyre simulations (indicated along x -axis). The mean, eddy, and sub-filter contributions are indicated.

To further assess the role of the parameterizations in influencing the mean flow and eddies, we examine the APE tendency induced by sub-grid thickness fluxes—both in the filtered high-resolution simulation and in the parameterized fluxes of the coarser simulations (Figure 11). The APE tendencies affecting both the MAPE and

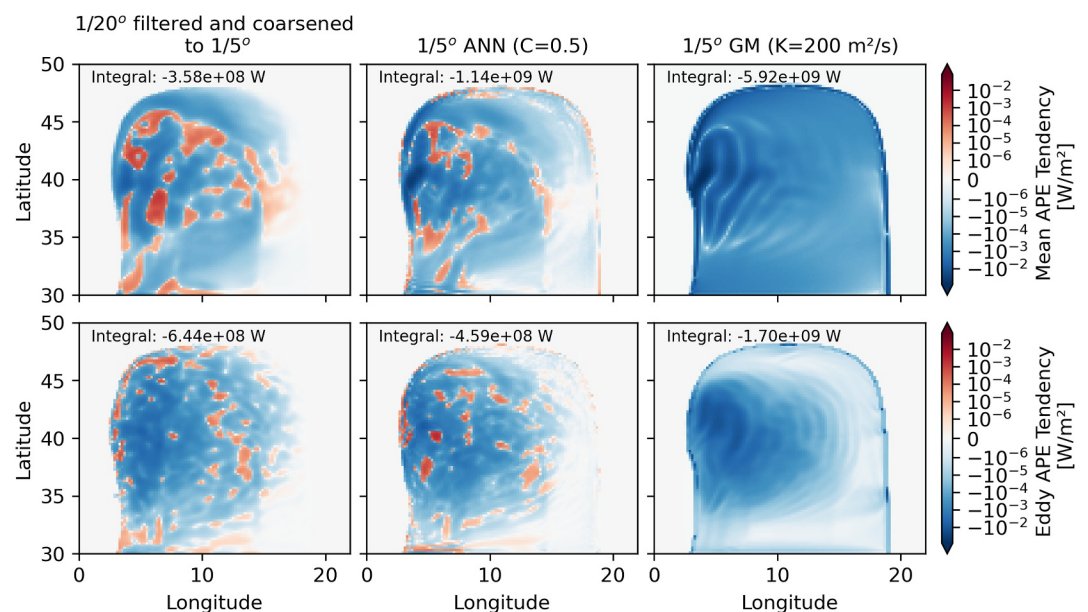


Figure 11. Available potential energy (APE) tendency exerted on the mean (top) and eddy (bottom) APE by the sub-grid or parameterized fluxes in the Double-Gyre simulations. The volume integrated value for each panel is indicated at the top left.

EAPE exhibit qualitative similarities between the filtered high-resolution and the ANN 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 ANN parameterization or the diagnosed tendencies from the filtered high-resolution simulation.

In summary, as in the Phillips-2-Layer case, the ANN parameterization outperforms the GM parameterization in the Double-Gyre 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 sub-grid thickness fluxes produced by mesoscale eddies. This was achieved by training a relatively small artificial neural network (ANN) to learn a numerical mapping between the gradients of the large-scale/resolved fields and the sub-grid thickness fluxes, using data from high-resolution simulations. By introducing features like lateral non-locality, coordinate invariance, and non-dimensionalization into the ANN design, we were able to produce a more generalizable and stable data-driven parameterization (similar to Perezhugin et al. [2025]). 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 S4 in Supporting Information S1). The skill relatively degrades at scales larger than the largest eddies, or as we transition from eddy permitting to non-eddy resolving resolutions, but even at these scales the offline skill is higher than traditional approaches, like an appropriately tuned Gent-McWilliams (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: Phillips-2-Layer and Double-Gyre, where baroclinic mesoscale eddies play a leading order role in the dynamics. In both these setups the ANN 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 ANN parameterization. Sensitivity studies showed that $O(1)$ values of this parameter, values of 1 for Phillips-2-Layer and 0.5 for Double-Gyre, were optimal in online setting across all eddy permitting resolutions. At coarser resolutions, where no eddies are resolved, the values needed to be slightly adjusted, a value of 2 for Phillips-2-Layer and 0.75 for Double-Gyre 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 m²/s) across resolutions and setups. We note that more advanced, state-aware specifications of the GM coefficient (Jansen et al., 2019; Mak et al., 2018) could likely reduce this sensitivity, though comparisons with such schemes are left for future work in more realistic configurations.

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 the 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. The largest issue with the traditional GM parameterization is the choice of isotropic diffusion as a closure for horizontal fluxes. Some work has tried to address this through anisotropic (R. D. Smith & Gent, 2004) and biharmonic (Eden, 2023) extensions, but the underlying diffusive framework remains, limiting the degree to which the local structure of the sub-grid thickness fluxes can be captured. Our work addresses this from a different angle, using a data-driven closure that is not constrained to be diffusive, and can in principle represent the upgradient and cross-gradient flux components that are present in high-resolution simulations. We see this as complementary to ongoing improvements in the traditional framework, rather than as a replacement.

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; Mak et al., 2023). This has led to development of ad hoc 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 toward the resolved eddies. One major advantage of the diffusive GM parameterization is that it is generally 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.

A natural question is whether our ML model is learning the physics of sub-grid thickness fluxes or numerical discretization errors. As discussed in Section 4.2, our sub-grid thickness fluxes are defined through filtering of the high-resolution fields, independently of any coarse model numerics, and so the ML model is trained on physical sub-grid thickness fluxes by construction. Additionally, by using velocity gradients rather than velocities as input features, the parameterization is Galilean invariant and structurally cannot represent a purely advective operator, providing a further distinction from numerical truncation errors of advection schemes, which are velocity-dependent and not Galilean invariant. We note, however, that when deployed online, the parameterization may incidentally compensate for some cumulative numerical biases, since discretization errors and sub-grid thickness fluxes can be correlated (Lund, 2003).

Our proposed data-driven parameterization has produced very promising results, but many aspects of the design can likely be improved. First, as is common with ML models, there is a spectral bias in the predictions. This means that the ANN's offline performance degrades at scales where the signal variance is low (Figure S3 and S4 in Supporting Information S1, 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 in the ANN architecture could help alleviate this issue. Second, it came almost as a surprise that the model performed so well across resolutions and setups with such limited choice of input features. 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-eddy permitting resolutions led us to this hypothesis. This suggests that in future work it may be worth paying more attention to these larger scales (larger than the deformation scale or the size of the dominant eddies), 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. Finally, we chose to use a reasonably small ANN to keep the computational burden due to the parameterization low in principle. In our work the cost of the low-resolution simulations with the ANN was about 50% greater than the simulations without the ANN, but still significantly faster than the high-resolution simulations. However, we have not performed a comprehensive testing and optimization to get the best possible computational performance for our parameterization in MOM6, yet.

We note that the parameterization developed here is trained entirely offline. Online learning approaches, where the ANN parameters are updated during the course of a simulation (Maddison, 2026; Yan et al., 2025), could potentially reduce the impact of distributional shift between offline training data and the online resolved state. However, online learning requires tight coupling between the training framework and the ocean model, which is not straightforward in traditional Fortran-based models like MOM6. Exploring online or approximately online learning strategies (Ouala et al., 2024), as well as architectures that incorporate temporal correlations, is also left for future work.

In terms of the large scale effects, our parameterization mainly acts on changing the APE in the system, and is not designed to change the KE of the resolved state—changes that may arise if the removed APE was appropriately cascaded upscale in an inverse KE cascade. Coupling our parameterization with a parameterization that produces backscatter effects in an energetically consistent manner could be investigated to produce further improvements in ocean models. We are already working on this, by combining our parameterization with the parameterizations developed for momentum fluxes (Perezhogin et al., 2024, 2025).

The idealized two-layer configurations used here were chosen to isolate the mesoscale eddy effects and to rigorously test the parameterization design in a controlled setting. Extending this work to more realistic ocean simulations introduces additional challenges, including the transition from thickness equation in layered models to buoyancy equation in z-coordinate models, and the parameterization must be re-framed to account for this. In a follow up study, we will address these challenges by developing and testing an expanded version of the

parameterization that works in terms of horizontal buoyancy fluxes, but follows the same design principles. We aim to show that this approach can work across a hierarchy of progressively more realistic configurations spanning any general vertical coordinate, moving toward its eventual application in global ocean and climate simulations.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

All the code to generate the simulation data and ML training data, do the ML training, analysis and figure generation can be found in the GitHub repository https://github.com/dhruvbalwada/mesoscale_buoyancy_param_ML, with an archived version available on Zenodo (Balwada 2025). The implementation of the parameterization in NOAA GFDL's MOM6 is archived on Zenodo (Balwada 2026); the corresponding source code is the tagged release `thickness-flux-ann-part1-v1.0` of our MOM6 fork (<https://github.com/dhruvbalwada/MOM6/tree/thickness-flux-ann-part1-v1.0>).

Acknowledgments

This project is supported by Schmidt Sciences. DB also acknowledges Dr. Ryan Abernathy, now at EarthMover PBC, for providing advice and guidance, and for helping shape his research interests in mesoscale parameterizations. This work would not have taken its current form without Ryan's constant support, encouragement, and contagious enthusiasm.

References

- Abernathy, R. P., Busecke, J. J. M., Smith, T. A., Deauna, J. D., Banihirwe, A., Nicholas, T., et al. (2022). *xgcm* (Version v0.8.1) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.7348619>
- Adcock, S. T., & Marshall, D. P. (2000). Interactions between geostrophic eddies and the mean circulation over large-scale bottom topography. *Journal of Physical Oceanography*, 30(12), 3223–3238. [https://doi.org/10.1175/1520-0485\(2000\)030<3223:ibgeat>2.0.co;2](https://doi.org/10.1175/1520-0485(2000)030<3223:ibgeat>2.0.co;2)
- Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., et al. (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. <https://doi.org/10.1029/2019ms001726>
- 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. <https://doi.org/10.1175/jpo-d-17-0100.1>
- Bachman, S. D. (2019). The GM+ E closure: A framework for coupling backscatter with the Gent and McWilliams parameterization. *Ocean Modelling*, 136, 85–106. <https://doi.org/10.1016/j.ocemod.2019.02.006>
- Bachman, S. D., Fox-Kemper, B., & Bryan, F. O. (2020). A diagnosis of anisotropic eddy diffusion from a high-resolution global ocean model. *Journal of Advances in Modeling Earth Systems*, 12(2), e2019MS001904. <https://doi.org/10.1029/2019ms001904>
- Balwada, D. (2025). Code for: Design and implementation of a data-driven parameterization for mesoscale eddy-induced transport. Part 1: Idealized two-layer configurations [Software]. Zenodo. <https://doi.org/10.5281/zenodo.19054967>
- Balwada, D. (2026). MOM6 implementation of the data-driven parameterization for mesoscale eddy-induced transport. Part 1: Idealized two-layer configurations [Software]. Zenodo. <https://doi.org/10.5281/zenodo.20735811>
- Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentile, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9), 098302. <https://doi.org/10.1103/physrevlett.126.098302>
- Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., et al. (2021). Climate-invariant machine learning. *arXiv preprint arXiv:2112.08440*.
- Bodner, A., Balwada, D., & Zanna, L. (2025). A data-driven approach for parameterizing ocean submesoscale buoyancy fluxes. *Journal of Advances in Modeling Earth Systems*, 17, e2025MS004991. <https://doi.org/10.1029/2025MS004991>
- 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. <https://doi.org/10.1038/s42254-024-00776-3>
- Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., et al. (2018). JAX: Composable transformations of Python+NumPy programs (Version v0.3.13) [Software]. *GitHub*. Retrieved from <http://github.com/google/jax>
- Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45(12), 6289–6298. <https://doi.org/10.1029/2018gl078510>
- Chelton, D. B., DeZoeke, 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. [https://doi.org/10.1175/1520-0485\(1998\)028<0433:gvoftb>2.0.co;2](https://doi.org/10.1175/1520-0485(1998)028<0433:gvoftb>2.0.co;2)
- DeepMind, Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., et al. (2020). The DeepMind JAX ecosystem [Software]. *GitHub*. Retrieved from <https://github.com/google-deeppmind>
- Eden, C. (2012). Implementing diffusivities from linear stability analysis in a three-dimensional general circulation ocean model. *Ocean Modelling*, 57, 15–28. <https://doi.org/10.1016/j.ocemod.2012.08.001>
- Eden, C. (2023). A simple eddy potential vorticity flux closure. *Journal of Advances in Modeling Earth Systems*, 15(10), e2023MS003711. <https://doi.org/10.1029/2023ms003711>
- Eden, C., & Dettmer, J. (2025). ROSSMIX 2.0: A simplified meso-scale eddy closure applied to a realistic ocean model. *Journal of Advances in Modeling Earth Systems*, 17(5), e2024MS004769. <https://doi.org/10.1029/2024ms004769>
- Eden, C., & Greatbatch, R. J. (2008). Towards a mesoscale eddy closure. *Ocean Modelling*, 20(3), 223–239. <https://doi.org/10.1016/j.ocemod.2007.09.002>
- 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. <https://doi.org/10.1016/j.ocemod.2010.01.004>
- 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. <https://doi.org/10.1175/2007jcli1510.1>
- Ferrari, R., & Wunsch, C. (2009). Ocean circulation kinetic energy: Reservoirs, sources, and sinks. *Annual Review of Fluid Mechanics*, 41(1), 253–282. <https://doi.org/10.1146/annurev.fluid.40.111406.102139>

- 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. <https://doi.org/10.1175/jpo2785.1>
- Gent, P. R. (2011). The Gent–McWilliams parameterization: 20/20 hindsight. *Ocean Modelling*, 39(1–2), 2–9. <https://doi.org/10.1016/j.ocemod.2010.08.002>
- Gent, P. R., & McWilliams, J. C. (1990). Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, 20(1), 150–155. [https://doi.org/10.1175/1520-0485\(1990\)020<0150:imiocm>2.0.co;2](https://doi.org/10.1175/1520-0485(1990)020<0150:imiocm>2.0.co;2)
- Ghosal, S. (1996). An analysis of numerical errors in large-eddy simulations of turbulence. *Journal of Computational Physics*, 125(1), 187–206. <https://doi.org/10.1006/jcph.1996.0088>
- Grooms, I., Agarwal, N., Marques, G., Pegion, P., & Yassin, H. (2025). The stochastic GM+ E closure: A framework for coupling stochastic backscatter with the Gent and McWilliams parameterization. *Journal of Advances in Modeling Earth Systems*, 17(5), e2024MS004560. <https://doi.org/10.1029/2024ms004560>
- Gullbrand, J., & Chow, F. K. (2003). The effect of numerical errors and turbulence models in large-eddy simulations of channel flow, with and without explicit filtering. *Journal of Fluid Mechanics*, 495, 323–341. <https://doi.org/10.1017/s0022112003006268>
- Hallberg, R. (2013). Using a resolution function to regulate parameterizations of oceanic mesoscale eddy effects. *Ocean Modelling*, 72, 92–103. <https://doi.org/10.1016/j.ocemod.2013.08.007>
- Heek, J., Levskaya, A., Oliver, A., Ritter, M., Rondepierre, B., Steiner, A., & van Zee, M. (2024). Flax: A neural network library and ecosystem for JAX (Version v0.12.5) [Software]. *GitHub*. Retrieved from <http://github.com/google/flax>
- Jansen, M. F., Adcroft, A., Khani, S., & Kong, H. (2019). Toward an energetically consistent, resolution aware parameterization of ocean mesoscale eddies. *Journal of Advances in Modeling Earth Systems*, 11(8), 2844–2860. <https://doi.org/10.1029/2019ms001750>
- 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. <https://doi.org/10.1016/j.ocemod.2015.05.007>
- Kamenkovich, I., Berloff, P., Haigh, M., Sun, L., & Lu, Y. (2021). Complexity of mesoscale eddy diffusivity in the ocean. *Geophysical Research Letters*, 48(5), e2020GL091719. <https://doi.org/10.1029/2020gl091719>
- 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. <https://doi.org/10.1029/2022ms003356>
- Killworth, P. D. (1997). On the parameterization of eddy transfer Part I. Theory. *Journal of Marine Research*, 55(6), 1171–1197. <https://doi.org/10.1357/0022240973224102>
- Killworth, P. D. (2001). Boundary conditions on quasi-Stokes velocities in parameterizations. *Journal of Physical Oceanography*, 31(4), 1132–1155. [https://doi.org/10.1175/1520-0485\(2001\)031<1132:bcoqsv>2.0.co;2](https://doi.org/10.1175/1520-0485(2001)031<1132:bcoqsv>2.0.co;2)
- Kusters, N., Balwada, D., & Groeskamp, S. (2025). Global observational estimates of mesoscale eddy-driven quasi-stokes velocity and buoyancy diffusivity. *Geophysical Research Letters*, 52(12), e2025GL115802. <https://doi.org/10.1029/2025gl115802>
- 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(1), 343–365. <https://doi.org/10.1146/annurev-conmatphys-043024-114758>
- Loose, N., Abernathy, R., Grooms, I., Busecke, J., Guillaumin, A., Yankovsky, E., et al. (2022). GCM-filters: A python package for diffusion-based spatial filtering of gridded data. *Journal of Open Source Software*, 7(70), 3947. <https://doi.org/10.21105/joss.03947>
- 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. <https://doi.org/10.1175/jpo-d-22-0083.1>
- Loose, N., Marques, G. M., Adcroft, A., Bachman, S., Griffies, S. M., Grooms, I., et al. (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. <https://doi.org/10.1029/2022ms003518>
- Lund, T. (2003). The use of explicit filters in large eddy simulation. *Computers & Mathematics with Applications*, 46(4), 603–616. [https://doi.org/10.1016/s0898-1221\(03\)90019-8](https://doi.org/10.1016/s0898-1221(03)90019-8)
- Maddison, J. R. (2026). Online learning in idealized ocean gyres. *Journal of Advances in Modeling Earth Systems*, 18(2), e2024MS004883. <https://doi.org/10.1029/2024ms004883>
- Mak, J., Maddison, J. R., Marshall, D. P., & Munday, D. R. (2018). Implementation of a geometrically informed and energetically constrained mesoscale eddy parameterization in an ocean circulation model. *Journal of Physical Oceanography*, 48(10), 2363–2382. <https://doi.org/10.1175/jpo-d-18-0017.1>
- 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. <https://doi.org/10.1029/2023ms003886>
- 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. <https://doi.org/10.1175/jpo-d-11-048.1>
- Marshall, J., & Shutts, G. (1981). A note on rotational and divergent eddy fluxes. *Journal of Physical Oceanography*, 11(12), 1677–1680. [https://doi.org/10.1175/1520-0485\(1981\)011<1677:anorad>2.0.co;2](https://doi.org/10.1175/1520-0485(1981)011<1677:anorad>2.0.co;2)
- Meredith, M., & Naveira Garabato, A. C. (2021). *Ocean mixing: Drivers, mechanisms and impacts*. Elsevier.
- Ouala, S., Chapron, B., Collard, F., Gaultier, L., & Fablet, R. (2024). Online calibration of deep learning sub-models for hybrid numerical modeling systems. *Communications Physics*, 7(1), 402. <https://doi.org/10.1038/s42005-024-01880-7>
- Perezhogin, P., Adcroft, A., & Zanna, L. (2025). Generalizable neural-network parameterization of mesoscale eddies in idealized and global ocean models. *Geophysical Research Letters*, 52(19), e2025GL117046. <https://doi.org/10.1029/2025gl117046>
- 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. <https://doi.org/10.1029/2023ms004104>
- 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. <https://doi.org/10.1016/j.cma.2022.115457>
- Ramadhan, A., Marshall, J., Souza, A., Wagner, G. L., Ponnappati, 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. <https://doi.org/10.1029/2022ms003258>
- 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. <https://doi.org/10.1357/002224007783649484>

- 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. [https://doi.org/10.1175/1520-0485\(2002\)032<1699:tsaeom>2.0.co;2](https://doi.org/10.1175/1520-0485(2002)032<1699:tsaeom>2.0.co;2)
- Smith, R. D., & Gent, P. R. (2004). Anisotropic Gent–McWilliams parameterization for ocean models. *Journal of Physical Oceanography*, 34(11), 2541–2564. <https://doi.org/10.1175/jpo2613.1>
- 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. <https://doi.org/10.1175/2011jpo4404.1>
- Uchida, T., Balwada, D., Jamet, Q., Dewar, W. K., Deremble, B., Penduff, T., & Le Sommer, J. (2023). Cautionary tales from the mesoscale eddy transport tensor. *Ocean Modelling*, 182, 102172. <https://doi.org/10.1016/j.ocemod.2023.102172>
- 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. [https://doi.org/10.1175/1520-0485\(1997\)027<0381:soetci>2.0.co;2](https://doi.org/10.1175/1520-0485(1997)027<0381:soetci>2.0.co;2)
- Yan, F. E., Frezat, H., Le Sommer, J., Mak, J., & Otness, K. (2025). Adjoint-based online learning of two-layer quasi-geostrophic baroclinic turbulence. *Journal of Advances in Modeling Earth Systems*, 17(7), e2024MS004857. <https://doi.org/10.1029/2024ms004857>
- Yan, F. E., Mak, J., & Wang, Y. (2024). On the choice of training data for machine learning of geostrophic mesoscale turbulence. *Journal of Advances in Modeling Earth Systems*, 16(2), e2023MS003915. <https://doi.org/10.1029/2023ms003915>
- 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. <https://doi.org/10.1029/2020gl091363>
- Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17), e2020GL088376. <https://doi.org/10.1029/2020gl088376>
- Zanna, L., Gregory, W., Perezhogin, P., Sane, A., Zhang, C., Adcroft, A., et al. (2025). A framework for hybrid physics-AI coupled ocean models. *arXiv preprint arXiv:2510.22676*.
- 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. <https://doi.org/10.1029/2023ms003697>

References From the Supporting Information

- Aluie, H., Rai, S., Yin, H., Lees, A., Zhao, D., Griffies, S. M., et al. (2022). Effective drift velocity from turbulent transport by vorticity. *Physical Review Fluids*, 7(10), 104601. <https://doi.org/10.1103/physrevfluids.7.104601>
- 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. [https://doi.org/10.1175/1520-0485\(1995\)025<0463:peitti>2.0.co;2](https://doi.org/10.1175/1520-0485(1995)025<0463:peitti>2.0.co;2)
- 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. [https://doi.org/10.1175/1520-0485\(2001\)031<1222:trmvp>2.0.co;2](https://doi.org/10.1175/1520-0485(2001)031<1222:trmvp>2.0.co;2)
- Uchida, T., Rokem, A., Nicholas, T., Abernathy, R., Vanderplas, J., & Mayer, A. (2018). xrft: Fourier transforms on xarray data structures (Version v0.1.0) [Computer software]. *Zenodo*. <https://doi.org/10.5281/zenodo.1402635>