

Generalizable neural-network parameterization of mesoscale eddies in idealized and global ocean models

Pavel Perezhigin¹, Alistair Adcroft³, Laure Zanna^{1,2}

¹Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

²Center for Data Science, New York University, New York, NY, USA

³Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA

Key Points:

- Physics constraints are developed for a neural-network parameterization of mesoscale eddy fluxes
- Dimensional scaling constraints improve offline generalization to unseen grid resolutions and depths
- New parameterization improves the representation of kinetic and potential energy online in coarse idealized and global ocean models

Abstract

Data-driven methods have become popular to parameterize the effects of mesoscale eddies in ocean models. However, they perform poorly in generalization tasks and may require retuning if the grid resolution or ocean configuration changes. We address the generalization problem by enforcing physics constraints on a neural network parameterization of mesoscale eddy fluxes. We found that the local scaling of input and output features helps to generalize to unseen grid resolutions and depths offline in the global ocean. The scaling is based on dimensional analysis and incorporates grid spacing as a length scale. We formulate our findings as a general algorithm that can be used to enforce data-driven parameterizations with dimensional scaling. The new parameterization improves the representation of kinetic and potential energy in online simulations with idealized and global ocean models. Comparison to baseline parameterizations and impact on global ocean biases are discussed.

Plain Language Summary

Ocean models can't directly simulate eddies that are smaller than the resolution of the computational grid. The effect of these eddies is represented by parameterizations. Machine learning offers a new way to build parameterizations directly from data, however, such parameterizations may fail when tested in new, unseen scenarios. Here, we leverage physics constraints to mitigate this, generalization, problem. Specifically, we found that method of dimensional analysis can be used to constrain data-driven parameterizations to enhance their accuracy in new scenarios without the need for retraining. New parameterization is tested in a realistic ocean model and brings us closer to robust, data-driven methods for ocean and climate models.

1 Introduction

Numerical ocean models rely on parameterizations to represent the effects of physical processes smaller than the model grid spacing, which are unresolved (Fox-Kemper et al., 2019; Hewitt et al., 2020; Christensen & Zanna, 2022). Recently, there has been a growing interest in applying machine learning methods to parameterize these subgrid physics in ocean models (Bolton & Zanna, 2019; Zanna & Bolton, 2020; Guillaumin & Zanna, 2021; Zhang et al., 2023; Sane et al., 2023; Yan et al., 2024; Perezhogin, Zhang, et al., 2024; Maddison, 2024). However, developing data-driven parameterizations for ocean models is still in its early stages, and their application is often limited to idealized configurations. Deploying data-driven parameterizations in the global ocean presents several challenges, one of which is addressed in this study – the problem of generalization to unseen scenarios.

Data-driven parameterizations rely heavily on sets of training data, and their successful implementation often requires tuning when applied to a new grid resolution (Zhang et al., 2023), flow regime (Ross et al., 2023), model configuration (Perezhogin, Zhang, et al., 2024), depth, or geographical region (Gultekin et al., 2024). However, in practice, it would be desirable to have a single parameterization that performs effectively across a variety of scenarios without requiring retuning. The ability of a data-driven model to work on new (testing) data, which is distinct from the training data, is measured by the *generalization error* (Bishop & Nasrabadi, 2006; Hastie et al., 2009). Data-driven methods work best when the testing data is drawn from the same distribution as the training data. However, in geophysical applications, the distribution of physical variables can vary vastly across different scenarios—a phenomenon referred to as a *distribution shift* (Beucler et al., 2024; Gultekin et al., 2024). In this case, domain knowledge and physics constraints can be leveraged to mitigate the generalization error of data-driven models (Kashinath et al., 2021).

In this work, we demonstrate how physics constraints can be leveraged to enhance the generalization of an Artificial Neural Network (ANN) parameterization of the ocean mesoscale eddy fluxes. Following Beucler et al. (2024), we rescale features of the ANN to minimize the distribution shift. To identify a suitable normalization technique for eddy fluxes, we apply dimensional analysis and Buckingham (1914)'s Pi theorem. Specifically, we introduce a local dimensional scaling constructed from the grid spacing and velocity gradients (Prakash et al., 2022). The local scaling improves offline generalization of the ANN parameterization to unseen grid resolutions and depths, as found in the global ocean dataset CM2.6 (Griffies et al., 2015). Our findings are formulated as a general algorithm that can be used to incorporate the dimensional scaling in future applications. Additional physics constraints for the ANN parameterization are enforced following Guan et al. (2022) and Srinivasan et al. (2024). We present an online evaluation of the new ANN parameterization in the GFDL MOM6 ocean model (Adcroft et al., 2019) in idealized and global configurations.

2 A Method to Constrain Neural Network with Dimensional Scaling

Here we introduce the concept of physical dimensionality and demonstrate how it can be used to constrain data-driven parameterizations. We start with a trivial example, followed by a general algorithm. Finally, we draw connections to existing approaches.

2.1 Trivial Example

Consider the case where a scalar momentum flux T (units of m^2s^{-2}) can be predicted using a length scale Δ (units of m) and inverse time scale X (units of s^{-1}):

$$T = f(\Delta, X). \quad (1)$$

Eq. (1) must remain invariant under rescaling the units of time and length, that is for any $\alpha, \beta > 0$, the equality must hold: $f(\alpha\Delta, \beta X) = \alpha^2\beta^2 f(\Delta, X)$. However, the unit invariance can be violated when f is parameterized by neural networks. One way to enforce it is by leveraging Buckingham (1914)'s Pi theorem, which states that the dimensional equation (such as Eq. (1)) can be rewritten in non-dimensional form. Specifically, for a set of three dimensional variables (T, Δ, X) with two independent dimensions (length and time), there is only one (three minus two) non-dimensional variable ($\pi_1 = T/(\Delta^2 X^2)$). Thus, Eq. (1) transforms to $\pi_1 = \text{const}$, or equivalently:

$$T = \Delta^2 X^2 \theta, \quad (2)$$

where θ can be interpreted as a non-dimensional Smagorinsky (1963) coefficient. A data-driven parameterization in the form of Eq. (2) with a trainable parameter θ , which is constant, follows the dimensional scaling as a hard constraint, in contrast to Eq. (1), which does not guarantee dimensional consistency. Eq. (2) promotes generalization as it explicitly accounts for the change in the magnitude of independent variables (Δ and X), constraining the learnable part of the mapping (θ) to be on the order of unity.

2.2 General Algorithm

Extending the example above, we suggest an algorithm to enforce dimensional scaling in ANN parameterizations by preprocessing input and output features:

1. Identify the input features that contribute significantly to the accurate prediction of the output features;
2. Construct non-dimensional input and output features from a *combined* set of identified input and output features;
3. Verify that a traditional known parameterization is a special case of the constructed non-dimensional mapping.

Step 1 follows standard dimensional analysis textbooks (Bridgman, 1922; Barenblatt, 1996). Specifically, a relevant set of input features can be identified by physical intuition or through ablation studies by evaluating the gain in offline performance from including additional dimensional features in the input set. Constructing non-dimensional features is a common approach in physics-constrained data-driven parameterizations (Ling et al., 2016; Schneider et al., 2024). However, the normalization of input features is often considered separately from the normalization of output features (Xie et al., 2020; Kang et al., 2023; Beucler et al., 2024; Christopoulos et al., 2024), unlike our proposed method (step 2 above). Additionally, the emphasis in these works is often placed on identifying normalization factors that minimize the distribution shift, while we suggest starting with identifying features responsible for the prediction (step 1). Finally, traditional parameterizations are often used to propose efficient normalization factors (Xie et al., 2020; Kang et al., 2023; Connolly et al., 2025), while we instead advocate for having traditional parameterizations as a special case (step 3, Prakash et al. (2022, 2024)).

3 Physics Constraints for Ocean Mesoscale Parameterization

Our goal is to predict the subfilter momentum fluxes of mesoscale eddies using an Artificial Neural Network (ANN) parameterization, see schematic in Figure 1(a). Various physical invariances were imposed to promote generalization.

3.1 Learning Subfilter Fluxes

We consider the acceleration produced by subfilter ocean mesoscale eddies (subfilter forcing, Bolton & Zanna, 2019):

$$\partial_t \bar{\mathbf{u}} = \mathbf{S} = (\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}} - \overline{(\mathbf{u} \cdot \nabla) \mathbf{u}}, \quad (3)$$

where $\mathbf{u} = (u, v)$ is the horizontal ocean velocity, $\nabla = (\partial_x, \partial_y)$ is the horizontal gradient, and (\cdot) is the horizontal filter. The subfilter forcing can be approximated (Loose, Marques, et al., 2023) as a divergence of the momentum flux:

$$\mathbf{S} \approx \nabla \cdot \mathbf{T} \quad (4)$$

where

$$\mathbf{T} = \begin{pmatrix} T_{xx} & T_{xy} \\ T_{yx} & T_{yy} \end{pmatrix} = \begin{pmatrix} \bar{u}\bar{u} - \bar{u}\bar{u} & \bar{u}\bar{v} - \bar{u}\bar{v} \\ \bar{u}\bar{v} - \bar{u}\bar{v} & \bar{v}\bar{v} - \bar{v}\bar{v} \end{pmatrix}. \quad (5)$$

We predict the three components of \mathbf{T} , namely T_{xx}, T_{xy}, T_{yy} , rather than \mathbf{S} directly, similarly to Zanna and Bolton (2020) (ZB20 hereafter) to impose momentum conservation as a hard constraint. We enforce symmetry of the tensor \mathbf{T} by predicting T_{xy} and sharing its prediction with T_{yx} , which guarantees angular momentum conservation (Griffies, 2018, Section 17.3.3). We also promote rotational and reflection invariances via data augmentation (Guan et al., 2022), independently rotating each training snapshot by 90° and reflecting it along the x and y axes, resulting in $8 = 2^3$ augmented snapshots per original one.

We learn the components of \mathbf{T} by minimizing the mean squared error (MSE) loss function:

$$\mathcal{L}_{\text{MSE}} = \|(\mathbf{S} - \nabla \cdot \hat{\mathbf{T}}) \cdot m\|_2^2 / \|\mathbf{S} \cdot m\|_2^2, \quad (6)$$

where m is the mask of wet points and $\hat{\mathbf{T}}$ is the neural network prediction of the subfilter flux as discussed below. See SI for further details.

3.2 Input Features

Here, we identify the input features relevant for the prediction of momentum fluxes (step 1 of the algorithm presented in Section 2.2).

ANN schematic

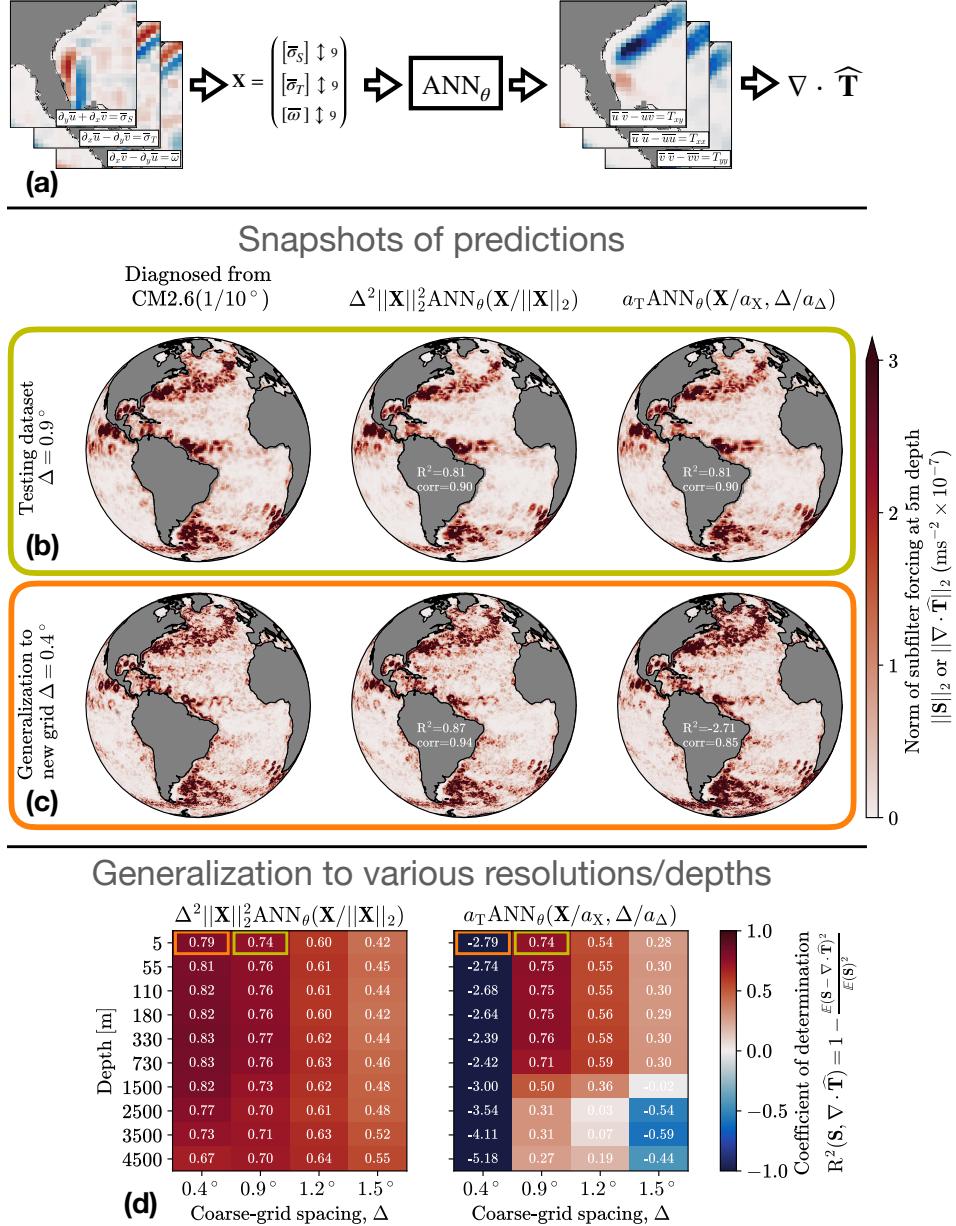


Figure 1. (a) Artificial neural network (ANN) parameterization predicting the divergence of subfilter fluxes given the velocity gradients on the horizontal stencil of 3×3 points. (b) Snapshots of predictions by two ANNs: with local dimensional scaling (Eq. (11), center column) or with fixed normalization coefficients (Eq. (10), right column) at the resolution (0.9°) and depth (5m) used for training (testing data is separated by 10 years). (c) Prediction at the unseen resolution (0.4°) and the same depth (5m). (d) Coefficient of determination (R^2) in prediction of subfilter forcing for various resolutions and depths, different from that used for training (0.9° , 5m). The R^2 is averaged over 2 years of held-out data and excludes 2 grid points adjacent to the coastline, where green and orange boxes correspond to panels (b) and (c), respectively.

Following ZB20, we consider the components of the strain-rate tensor and vorticity as input features:

$$\begin{aligned}\bar{\sigma}_S &= \partial_y \bar{u} + \partial_x \bar{v} && \text{shearing strain,} \\ \bar{\sigma}_T &= \partial_x \bar{u} - \partial_y \bar{v} && \text{horizontal tension/stretch,} \\ \bar{\omega} &= \partial_x \bar{v} - \partial_y \bar{u} && \text{relative vorticity.}\end{aligned}\quad (7)$$

These input features exclude explicit dependence on the velocity, guaranteeing Galilean invariance of the parameterization (Srinivasan et al., 2024; Pope, 1975; Lund & Novikov, 1993; Ling et al., 2016). When using the input features (Eq. (7)) pointwise, the resulting ANN parameterization highly correlates with the ZB20 equation-discovery model. Thus, we decided to include the non-local contribution of these features (Srinivasan et al., 2024). To do so, the input vector to the ANN consists of velocity gradients, each defined on a 3×3 horizontal stencil and flattened into a vector of length 9 (denoted as $[\cdot] \uparrow\downarrow 9$):

$$\mathbf{X} = \begin{pmatrix} [\bar{\sigma}_S] \uparrow\downarrow 9 \\ [\bar{\sigma}_T] \uparrow\downarrow 9 \\ [\bar{\omega}] \uparrow\downarrow 9 \end{pmatrix} \in \mathbb{R}^{27}. \quad (8)$$

Including information about velocity gradients from the closest neighboring points is a common approach in subgrid modeling (Maulik & San, 2017; Maulik et al., 2019; Pawar et al., 2020; Wang et al., 2021, 2022; Gultekin et al., 2024).

To facilitate generalization across different resolutions (scale-aware or grid-aware parameterization, Bachman et al., 2017), we account for the local grid spacing of the coarse resolution model $\Delta = \sqrt{\Delta x \Delta y}$, resulting in the following functional form of the parameterization (Lund & Novikov, 1993; Li et al., 2025):

$$\mathbf{T} \approx \hat{\mathbf{T}}(\mathbf{X}, \Delta). \quad (9)$$

Accounting for grid spacing is physically justified as velocity gradients and momentum fluxes differ in dimensionality and require a length scale to be invoked.

3.3 Neural Network Parameterizations

We consider a baseline data-driven parameterization of eddy fluxes with fixed normalization coefficients, following the form of Eq. (9):

$$\hat{\mathbf{T}}(\mathbf{X}, \Delta) = a_T \text{ANN}_\theta(\mathbf{X}/a_X, \Delta/a_\Delta), \quad (10)$$

where ANN_θ is the neural network with trainable parameters θ , coefficients $a_T = 10^{-2} \text{m}^2 \text{s}^{-2}$ and $a_X = 10^{-6} \text{s}^{-1}$ approximate the standard deviations of eddy fluxes and velocity gradients in our dataset, and $a_\Delta = 50 \text{km}$. Using fixed normalization coefficients in parameterizations similar to Eq. (10) is a common practice (Srinivasan et al., 2024). Below, we contrast this approach to a normalization that follows solely from dimensional analysis presented in Section 2.2.

A combined set of input and output features $(\mathbf{T}, \mathbf{X}, \Delta)$ is used to construct non-dimensional input $(\mathbf{X}/\|\mathbf{X}\|_2)$ and output $(\mathbf{T}/(\Delta \|\mathbf{X}\|_2)^2)$ features (step 2 in Section 2.2), where $\|\mathbf{X}\|_2 = \sqrt{\sum_i X_i^2}$. This normalization of features is local, that is computed separately for each grid point. By designing the ANN to operate on non-dimensional variables, we propose a parameterization with the local dimensional scaling (Reissmann et al., 2021; Prakash et al., 2022):

$$\hat{\mathbf{T}}(\mathbf{X}, \Delta) = \Delta^2 \|\mathbf{X}\|_2^2 \text{ANN}_\theta(\mathbf{X}/\|\mathbf{X}\|_2). \quad (11)$$

According to the Buckingham (1914)'s Pi theorem, there is freedom in constructing non-dimensional variables. We opt to use the non-dimensional vector $\mathbf{X}/\|\mathbf{X}\|_2$ to constrain

183 the range of its components between -1 and 1 , thereby reducing the distribution shift
 184 in ANN inputs.

185 Following step 3 in Section 2.2, we show that the model form (Eq. (11)) admits ZB20,
 186 Smagorinsky, biharmonic Smagorinsky, and Leith (1996) parameterizations as special
 187 cases, with well-behaved functional representations (see Text S1 in SI for details). Fur-
 188 thermore, Eq. (11) guarantees that the predicted fluxes vanish ($\hat{\mathbf{T}} \rightarrow \mathbf{0}$) as velocity gra-
 189 dients diminish ($\mathbf{X} \rightarrow \mathbf{0}$), similarly to known parameterizations, see Text S2 in SI. We
 190 experimentally verified that the spatial variability of the normalization factor ($\|\mathbf{X}\|_2$)
 191 in Eq. (11) does not amplify the parameterization errors ($\mathbf{S} - \nabla \cdot \hat{\mathbf{T}}$) compared to Eq.
 192 (10).

193 4 Experimental setups

194 4.1 Training dataset

195 The training dataset is created using the climate model CM2.6 (Griffies et al., 2015),
 196 which has a nominal ocean resolution of 0.1° . Velocity gradients (Eq. (7)), used as in-
 197 put features, and subfilter forcing (\mathbf{S} , Eq. (3)), used as output, are diagnosed using hor-
 198 izontal filtering followed by coarse-graining, which avoids the inclusion of discretization
 199 errors (Guillaumin & Zanna, 2021; Christensen & Zanna, 2022; Agdestein & Sanderse,
 200 2025). The filtering is applied by sliding a Gaussian kernel with a filter width three times
 201 the width of the target coarse grid box, using Grooms et al. (2021); Loose et al. (2022).
 202 Subsequent coarse-graining is done by averaging over the fine grid boxes contained within
 203 each coarse grid box. The filtering and coarse-graining are done for 4 coarse resolutions
 204 and 10 depth levels (Figure 1(d) and Table S1 in SI).

205 4.2 ANN architecture

206 For the offline analysis of parameterizations (Eqs. (10), (11)), we found that an ANN
 207 with two hidden layers, 32 neurons each, was sufficiently large to effectively learn from
 208 the input features. See Text S3 in SI for details.

209 4.3 Online implementation

210 We implement the ANN mesoscale eddy parameterization (Eq. (11)) in two con-
 211 siderably different configurations of the GFDL MOM6 ocean model (Adcroft et al., 2019)
 212 at eddy-permitting ($1/4^\circ$) resolution. To ensure that ANN inference remains computa-
 213 tionally efficient, we retrain a smaller network with only one hidden layer and 20 neu-
 214 rons, which keeps the ANN inference time below 10% of the ocean model runtime (Text
 215 S3 in SI for details). While our goal was to implement the ANN parameterization with-
 216 out further modifications, minor adjustments were necessary for numerical stability, see
 217 Text S4 in the SI.

218 The idealized ocean configuration, NeverWorld2 (NW2, Marques et al. (2022)), in-
 219 cludes 15 stacked shallow water layers, featuring a single basin ocean with a reentrant
 220 channel. The circulation is driven by a steady wind forcing, giving rise to a circumpo-
 221 lar current and gyres. Coarse simulations are initialized from rest, and run for 30000 days,
 222 similar to Marques et al. (2022) and Perezhogin, Zhang, et al. (2024).

223 The second configuration, OM4 (Adcroft et al., 2019), is a coupled ocean-sea-ice
 224 model forced at the air-sea interface by prescribing the atmosphere state according to
 225 the CORE-II interannual forcing (IAF) protocol (Large & Yeager, 2009). The simula-
 226 tions span 60 years (1948-2007) and were initialized with a state of the Control model
 227 after 270 years of spin-up.

228 The biharmonic Smagorinsky scheme for gridscale dissipation is used with viscosity
 229 coefficient $\nu_4 = 0.06\sqrt{\bar{\sigma}_S^2 + \bar{\sigma}_T^2}\Delta^4$ (Adcroft et al., 2019), applied in control and parameterized
 230 (mixed modeling, Meneveau and Katz (2000)) simulations.

231 5 Results

232 5.1 Offline generalization

233 Our primary goal is to demonstrate that the local dimensional scaling promotes
 234 the generalization of the eddy parameterization to unseen grid resolutions and depths.
 235 Limited generalization in such scenarios has been reported for previous machine-learning
 236 models of mesoscale eddies (Zhang et al., 2023; Gultekin et al., 2024; Ross et al., 2023)
 237 and traditional physics-based parameterizations (Yankovsky et al., 2024).

238 We compare two ANNs: one incorporating local dimensional scaling (Eq. (11)) and
 239 a baseline ANN with fixed normalization coefficients (Eq. (10)). To explore generaliza-
 240 tion, we let the ANNs learn based solely on data from one combination of depth (5m)
 241 and coarse grid resolution (0.9°) during training. The local grid spacing varies accord-
 242 ing to the tripolar grid used in the ocean component of the CM2.6 climate model. In par-
 243 ticular, at the nominal resolution of 0.9° , the coarse grid spacing $\Delta = \sqrt{\Delta x \Delta y}$ is in a
 244 range from 50km to 100km for non-polar latitudes ($60S^\circ$ – $60N^\circ$). Spatially varying grid
 245 spacing provides essential information for effective learning by the baseline ANN. The
 246 offline evaluation of ANNs on held-out data similar to that used for training is shown
 247 in Figure 1(b). Both ANNs exhibit high and equal pattern correlation (0.90) and R^2 (0.81)
 248 in the prediction of the norm of subfilter forcing.

249 We now consider generalization to a different grid resolution, which is finer (0.4°)
 250 compared to that used for training (0.9°), see Figure 1(c). The range of grid spacings
 251 in this case is beyond the range seen by a baseline ANN during training, resulting in a
 252 distribution shift between the testing and training data. The baseline ANN parameter-
 253 ization (Eq. (10)) predicts the norm of subfilter forcing at a new grid resolution with a
 254 reasonably high pattern correlation (0.85). However, the magnitude of the prediction is
 255 too large, resulting in a low R^2 (-2.71). Instead, the ANN with dimensional scaling (Eq.
 256 (11)) offers improved generalization capability. The proposed ANN naturally accounts
 257 for the reduction of the grid spacing and reduces the magnitude of the prediction, re-
 258 sulting in high pattern correlation (0.94) and R^2 (0.87) metrics (Figure 1(c)).

259 The generalization to both finer and coarser grids, and different depths, is summa-
 260 rized in Figure 1(d). At coarser grid spacings (1.2° , 1.5°) compared to that used for train-
 261 ing (0.9°), the local dimensional scaling again helps to achieve higher R^2 by increasing
 262 the magnitude of the prediction.

263 In the deep layers, the subfilter forcing and velocity gradients are approximately
 264 one order of magnitude smaller than near the surface. Thus, a baseline ANN, trained
 265 on much larger values near the surface (such as here, at depth 5m), can lead to inaccu-
 266 rate predictions at depth (Figure S1 in SI). However, the local dimensional scaling ef-
 267 ffectively rescales the input features, thereby improving generalization to deep, unseen
 268 layers, as summarized in Figure 1(d).

269 The major reason why baseline ANN has a poor skill at unseen resolutions and depths
 270 is the lack of training data. We verified that the generalization of the baseline ANN to
 271 various resolutions and depths can be restored if these resolutions and depths are included
 272 in the training dataset. Incorporating local dimensional scaling as done in this work, there-
 273 fore, requires less training data and improves out-of-distribution generalization.

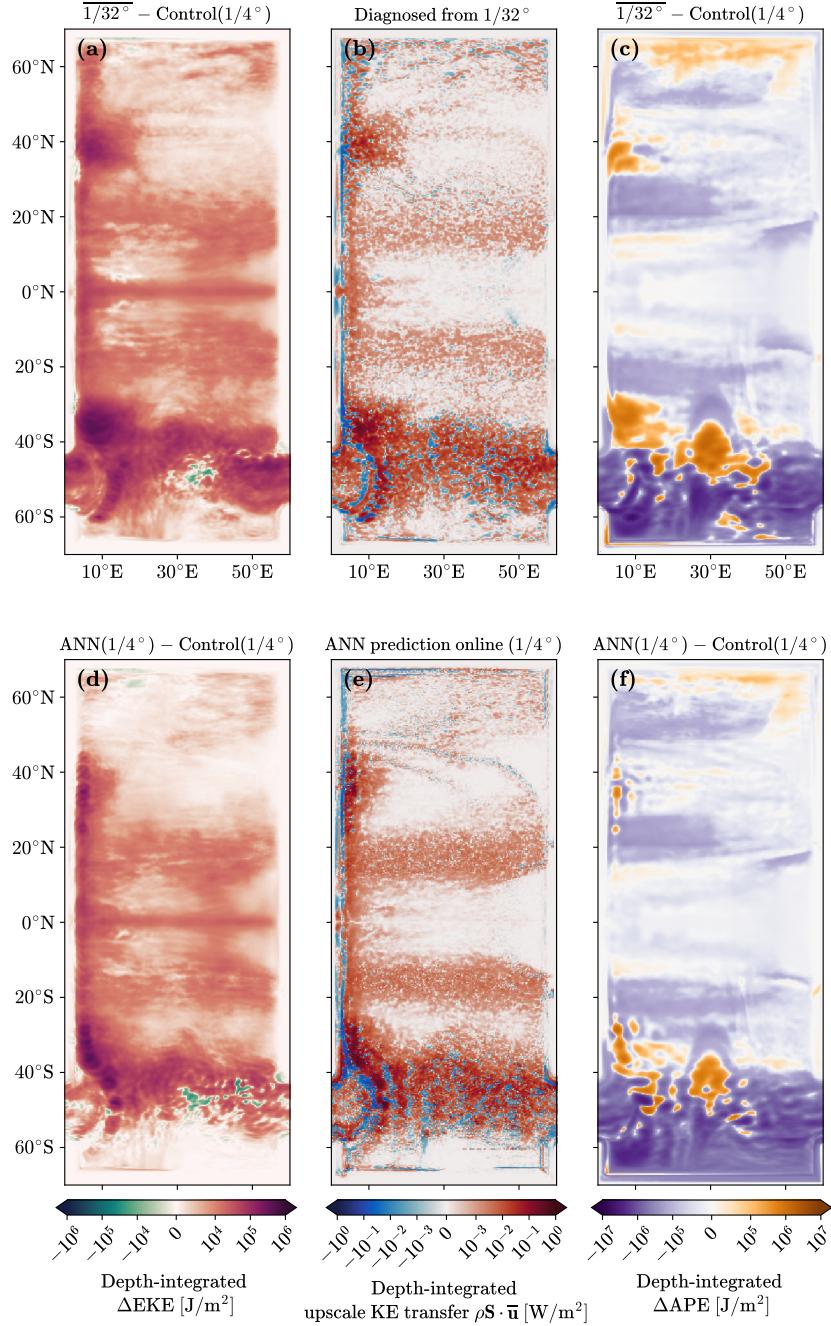


Figure 2. Effect of increasing resolution in idealized configuration NeverWorld2 (Marques et al., 2022) in the upper row, where $\overline{1/32^\circ}$ represents filtered and coarse-grained high-resolution simulation. Impact of the ANN parameterization with local dimensional scaling (Eq. (11)) online at resolution $1/4^\circ$ in the lower row. We consider three depth-integrated metrics: difference (denoted as Δ) in Eddy Kinetic Energy (EKE) (left column); diagnosed and predicted online upscale kinetic energy transfer (center column) (positive values represent backscatter); difference in Available Potential Energy (APE) (right column). All metrics are averaged over 160 snapshots corresponding to the last 800 days of the simulations.

274 **5.2 Online evaluation in the MOM6 ocean model**

275 We use all available depths and grid resolutions shown in Figure 1(d) for training
 276 to make the implemented ANN parameterization (Eq. (11)) less tied to any specific res-
 277 olution or depth. The retrained, smaller ANN (see Section 4.3) exhibits a slightly lower
 278 offline skill (R^2) than the version described earlier, on average, by 0.1 (Figure S5 in SI).

279 **5.2.1 Idealized configuration NeverWorld2**

280 We first consider an idealized adiabatic ocean configuration NW2, which generates
 281 various circulation patterns similar to the global ocean but allows us to isolate the ef-
 282 fect of mesoscale eddies. Our goal is to show that the impact of the ANN parameteri-
 283 zation on the flow is similar to that of increasing the horizontal resolution.

284 Mesoscale eddies extract available potential energy, $\text{APE} = \frac{\rho}{2} \sum_k g'_k (\eta_k^2 - (\eta_k^{\text{ref}})^2)$,
 285 from the mean flow, which is then converted into the eddy kinetic energy, $\text{EKE} = \frac{\rho}{2} (\overline{|\mathbf{u}^t|^2}) -$
 286 $|\overline{\mathbf{u}^t}|^2$ (Salmon, 1980). Here, $(\cdot)^t$ is the temporal averaging, ρ is the density, g'_k is the re-
 287 duced gravity of k -th isopycnal interface, η_k is the interface height and η_k^{ref} is the state
 288 of rest with flat isopycnals. At an eddy-permitting resolution ($1/4^\circ$), this energy path-
 289 way is partially unresolved (Jansen & Held, 2014; Mana & Zanna, 2014; Juricke et al.,
 290 2019; Loose, Bachman, et al., 2023). As a result, the coarse ocean model has too low EKE
 291 and too large APE when compared to the filtered and coarse-grained high-resolution sim-
 292 ulation, denoted as $1/32^\circ$ (Figure 2(a,c)). However, Figure 2(a) suggests that the miss-
 293 ing eddies can be nominally resolved on the coarse grid. Traditional backscatter param-
 294 eterizations are designed to directly reduce this EKE bias by energizing the resolved ed-
 295 dries, resulting in additional extraction of APE (Jansen & Held, 2014; Yankovsky et al.,
 296 2024).

297 Eddy backscatter is diagnosed when the kinetic energy transfer produced by the
 298 subfilter forcing (Eq. (3)) is predominantly positive (upscale), as shown in Figure 2(b).
 299 We verified that our ANN parameterization accurately predicts the eddy backscatter of-
 300 fline (Figure S2 in SI), suggesting reasonable generalization capabilities of our approach.
 301 In Figure 2(e), we show a more challenging task – the prediction of the eddy backscat-
 302 ter online once the ANN is coupled to the coarse ocean model. The online prediction of
 303 eddy backscatter grossly resembles the diagnosed data shown in Figure 2(b), although
 304 there are slight differences caused by the difference in distributions of input features.

305 The kinetic energy injection from the ANN parameterization leads to an increase
 306 in EKE that aligns with the high-resolution data, in particular in the ACC (Antarctic
 307 circumpolar current) region (40°S – 60°S), near the western boundaries, and in the sub-
 308 tropics (20°S – 20°N), see Figure 2(d). However, the EKE increase in western bound-
 309 ary current extension (40°N , 10°E – 20°E) and subpolar gyre (50°N – 70°N) is smaller
 310 than expected from the high-resolution simulation. The spatial pattern of APE reduc-
 311 tion in the parameterized simulation is close to that produced by increasing horizontal
 312 resolution (Figure 2(f)). The APE is predominantly reduced in the Southern Ocean and
 313 ACC regions (40°S – 70°S), followed by APE reduction in gyres (20°N – 60°N , 20°S –
 314 40°S). Local patches of APE increase in the higher resolution model (Figure 2(c)) cor-
 315 respond to enhanced horizontal recirculation and are reproduced by the ANN param-
 316 eterization, but less accurately compared to the diagnosed APE reduction.

317 Consequences of APE reduction include flattening of isopycnals and improving the
 318 structure of isopycnal interfaces across multiple cross-sections (Figure S3 in SI), along
 319 with a weakening of ACC transport through the Drake Passage (Table S3 in SI).

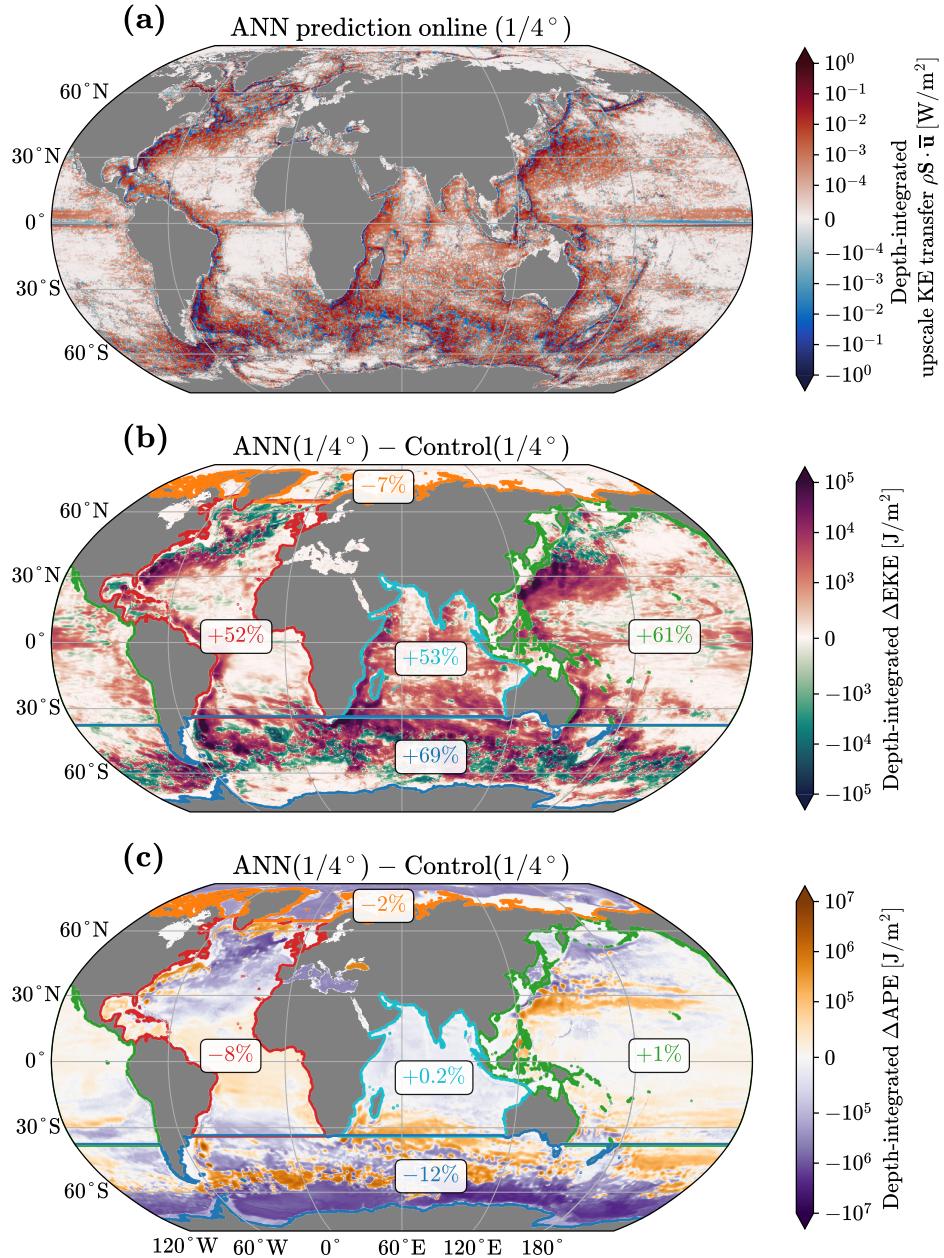


Figure 3. Online evaluation of the ANN parameterization in the global ocean-ice model OM4 (Adcroft et al., 2019) at eddy-permitting resolution ($1/4^\circ$). The following depth-integrated diagnostics are averaged over one year (2003): (a) upscale kinetic energy transfer predicted by the ANN parameterization online, (b) difference in Eddy kinetic energy (EKE), (c) difference in Available potential energy (APE). The integrated percentage change in EKE and APE relative to the control simulation is shown for five ocean basins.

5.2.2 Global ocean-sea-ice model OM4

We next evaluate the ANN parameterization in the global ocean model OM4. Unlike in the idealized configuration, the interaction of many physical processes in driving the circulation in the global ocean model impede our ability to directly isolate the ef-

324 effect of mesoscale eddies (Ferrari & Wunsch, 2009; Lévy et al., 2010). Building on the dy-
 325 namical expectations established in the idealized NW2 configuration, our goal is to as-
 326 sess whether the global ocean model exhibits similar response patterns to the eddy pa-
 327 rameterization.

328 The prediction of the kinetic energy injection by the ANN parameterization on-
 329 line is shown in Figure 3(a). Similarly to the idealized configuration, the kinetic energy
 330 is injected in the subtropical gyres, near the western boundaries, and occasionally in the
 331 ACC region. The energy injection is accompanied by an increase of the EKE in the same
 332 locations, see Figure 3(b). However, compared to the pattern found in an idealized con-
 333 figuration, the EKE decrease appears more frequently: along topographic features in the
 334 subpolar gyres of the North Atlantic and North Pacific oceans, and occasionally in the
 335 ACC region. The decrease of EKE in these regions is due to a shift or weakening of the
 336 mean currents, potentially as a result of the removal of kinetic energy by the ANN pa-
 337 rameterization along the lateral boundaries, changes in deep water formation, and/or
 338 changes in global overturning circulation. The complexity of the model prohibits us from
 339 identifying a single mechanism.

340 Similarly to the idealized configuration, APE is primarily reduced in the South-
 341 ern Ocean (-12%), with minor APE reductions observed in the Subpolar Gyres of the
 342 North Atlantic and North Pacific oceans (Figure 3(c)). APE is additionally reduced in
 343 the Arctic Ocean despite the lack of increased eddy activity in this region. However, its
 344 relative change is moderately small (-2%).

345 We assessed whether the offline performance of the ANN parameterization corre-
 346 lates with the online results (Figure S5 in SI). We found that using spatially non-local
 347 features on a 3×3 stencil, as in this study, is important for achieving higher offline skill
 348 in CM2.6 and improved energetics in OM4 compared to a pointwise ANN parameter-
 349 ization. However, increasing the number of neurons, which also contributes to the offline
 350 skill, has a smaller impact on the energetics.

351 We note that the evaluation presented in this section is qualitative and can be strength-
 352 ened by comparing the parameterized global ocean model to filtered and coarse-grained
 353 higher-resolution simulations. Such evaluation can be performed in future studies by im-
 354 plementing the proposed parameterization in the hierarchy of GFDL climate models, CM4X,
 355 which differ in the horizontal resolution of the ocean component (Griffies et al., 2024).

356 5.2.3 Comparison to an anti-viscosity parameterization

357 We confront our ANN parameterization to a traditional anti-viscosity parameter-
 358 ization representing mesoscale eddy effects (Jansen et al., 2015) and already tested in
 359 OM4 by Chang et al. (2023). Repeating their analysis, we found that both ANN and anti-
 360 viscosity parameterizations reduce the regional biases in the Gulf Stream region, see Fig-
 361 ure S4 in SI for sea surface temperature and salinity biases. The response in other global
 362 ocean circulation metrics is remarkably similar for both parameterizations as well (Fig-
 363 ure 4). Specifically, both parameterizations increase the globally integrated kinetic en-
 364 ergy by roughly the same percentage and reduce the APE by nearly the same percent-
 365 age. The restratification effect of mesoscale eddies leads to the reduction of the globally-
 366 averaged potential temperature (Griffies et al., 2015; Adcroft et al., 2019). As previously
 367 discussed, the transport through the Drake Passage is reduced in both parametrized sim-
 368 ulations, see also Grooms et al. (2024). Unlike in Chang et al. (2023), both parameter-
 369 izations weaken the Atlantic meridional circulation (AMOC). This suggests that the AMOC
 370 response depends on the ocean model state, perhaps to a greater extent than the details
 371 of mesoscale eddy parameterizations. The response in some global metrics (ACC, AMOC,
 372 globally-averaged potential temperature) does not appear to project onto the existing
 373 ocean model biases. That is, both parameterized ocean simulations are less consistent
 374 with the observational data than the control simulation (Figure 4). We note that our

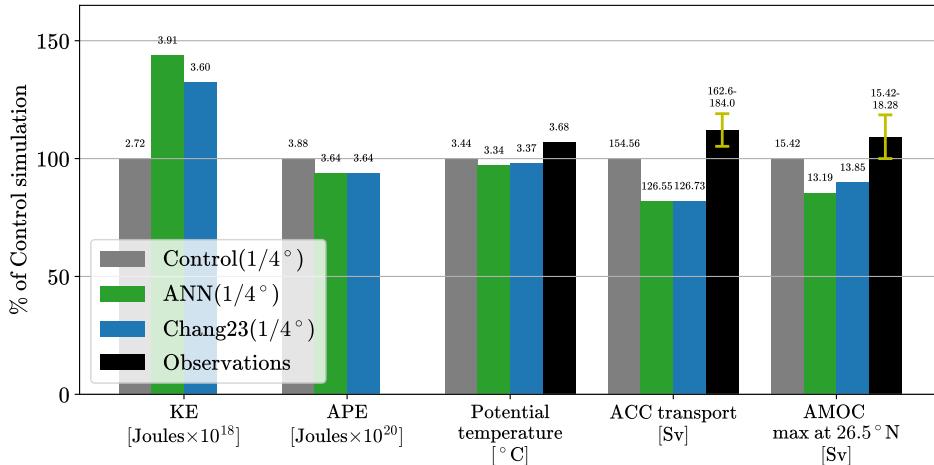


Figure 4. Comparison of the ANN parameterization to the negative viscosity backscatter parameterization (Chang et al., 2023) in the global ocean-ice model OM4. Kinetic energy (KE) and Available potential energy (APE) are integrated globally. Potential temperature is averaged globally. ACC transport is computed at the Drake Passage section at 70°W, and AMOC is computed as the maximum over depth streamfunction at 26.5°N. Model output is averaged over the years 1981-2007. Observational data for potential temperature is given by World Ocean Atlas 2018 (WOA18, Locarnini et al. (2018)), for ACC transport with error bar is given by cDrake (Donohue et al., 2016), and for AMOC is given by RAPID (Cunningham et al., 2007) averaged over 2004-2021 years with error bar showing interannual standard deviation.

goal was to improve the representation of mesoscale eddy processes. Bias reduction is not guaranteed due to compensating model errors from other parameterizations and remains an important direction for future work. A full recalibration of the ocean model may be necessary, particularly for physical processes competing with mesoscale eddies in determining average potential temperature and the strength of the ACC and AMOC.

6 Discussion

We address the generalization issue of ANN parameterizations of mesoscale eddies by embedding physics constraints into the inputs, outputs, and parameterization itself. The Buckingham (1914)'s Pi-theorem and dimensional analysis are invoked to obtain local normalization coefficients. The ANN parameterization with local dimensional scaling significantly outperforms the ANN with fixed normalization coefficients offline, demonstrating superior generalization to unseen grid resolutions and depths in the global ocean data CM2.6. A general algorithm for constructing dimensional scaling, which can be applied to other neural-network parameterizations, is presented.

The proposed ANN parameterization with dimensional scaling is successfully tested online in the GFDL MOM6 ocean model. It accurately predicts upscale kinetic energy transfer, despite many challenges presented by online implementation. The parameterization improves the energy pathways by energizing the resolved eddies and reducing APE, consistent with the expected restratification effects of mesoscale eddies. These improvements hold across idealized (NW2) and global ocean (OM4) configurations, with the most pronounced APE reduction occurring in the Southern Ocean. The ANN achieves comparable online performance to an existing backscatter parameterization (Jansen et al.,

397 2015; Chang et al., 2023) in OM4, and does not require significant retuning between idealized
 398 and global setups.

399 We demonstrate the improved or similar performance of the ANN parameterization
 400 in NW2 compared to existing backscatter schemes (Yankovsky et al., 2024; Perezhigin,
 401 Zhang, et al., 2024) across different resolutions ($1/3^\circ$ – $1/6^\circ$, see Table S2 in SI). At $1/2^\circ$,
 402 however, the ANN offers no clear improvement compared to the control simulation, likely
 403 due to less resolved eddies and stronger viscosity. At coarser resolutions ($\sim 1^\circ$), the sub-
 404 filter momentum fluxes vanish as the Rossby radius is unresolved (Figure S6 in SI). At
 405 such coarse resolutions, combining the ANN with bulk parameterizations or online learn-
 406 ing approaches may help (Maddison, 2024; Shankar et al., 2025), along with parameter-
 407 izations explicitly extracting APE (Bachman, 2019; Jansen et al., 2019; Grooms et al.,
 408 2024; Perezhigin, Balakrishna, & Agrawal, 2024; Balwada et al., 2025).

409 Additional work is needed to enhance data-driven parameterizations beyond the
 410 performance of traditional parameterizations in realistic global configurations. Both pa-
 411 rameterization approaches exhibit substantial departures from observations and contribute
 412 comparably to persistent model biases. This highlights the potential for improving pa-
 413 rameterization schemes, evaluation metrics, and model calibration in ocean modeling.
 414 Looking ahead, the generalization issue addressed in this study has immediate implica-
 415 tions for climate models, where parameterizations must remain reliable under changing
 416 conditions.

417 Open Research Section

418 The training algorithm, plots, ANN weights, implemented parameterization and
 419 MOM6 setups are available at Perezhigin (2025). The training dataset, offline skill, and
 420 simulation data are available at Perezhigin et al. (2025). For high-resolution NW2 sim-
 421 ulation data, see Marques et al. (2022). Observational products can be found: WOA18
 422 (Garcia et al., 2019) and RAPID (Moat et al., 2025).

423 Acknowledgments

424 This project is supported by Schmidt Sciences, LLC. LZ was also partially funded through
 425 NOAA NA19OAR4310364-T1-01. This research was also supported in part through the
 426 NYU IT High Performance Computing resources, services, and staff expertise. This re-
 427 search used resources of the National Energy Research Scientific Computing Center, a
 428 DOE Office of Science User Facility supported by the Office of Science of the U.S. De-
 429 partment of Energy under Contract No. DE-AC02-05CH11231 using NERSC award BER-
 430 ERCAP0032655. The authors would like to thank the M²LInES team, in particular, Dhruv
 431 Balwada, Alex Connolly, and Nora Loose, and also Wenda Zhang and Aviral Prakash
 432 for helpful comments and discussion. We also thank the anonymous reviewers for their
 433 helpful suggestions, which have helped improve the manuscript.

434 References

- 435 Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., ...
 436 others (2019). The GFDL global ocean and sea ice model OM4.0: Model
 437 description and simulation features. *Journal of Advances in Modeling Earth
 438 Systems*, 11(10), 3167–3211. doi: <https://doi.org/10.1029/2019MS001726>
 439 Agdestein, S. D., & Sanderse, B. (2025). Discretize first, filter next: Learn-
 440 ing divergence-consistent closure models for large-eddy simulation. *Jour-
 441 nal of Computational Physics*, 522, 113577. doi: <https://doi.org/10.1016/j.jcp.2024.113577>
 442 Bachman, S. D. (2019). The GM + E closure: A framework for coupling backscatter
 443 with the Gent and McWilliams parameterization. *Ocean Modelling*, 136, 85–

- 445 106. doi: <https://doi.org/10.1016/j.ocemod.2019.02.006>
- 446 Bachman, S. D., Fox-Kemper, B., & Pearson, B. (2017). A scale-aware subgrid
447 model for quasi-geostrophic turbulence. *Journal of Geophysical Research: Oceans*, 122(2), 1529–1554. doi: <https://doi.org/10.1002/2016JC012265>
- 448 Balwada, D., Perezhigin, P., Adcroft, A., & Zanna, L. (2025). Design and imple-
449 mentation of a data-driven parameterization for mesoscale thickness fluxes.
450 *Authorea Preprints*.
- 451 Barenblatt, G. I. (1996). *Scaling, self-similarity, and intermediate asymptotics: di-
452 mensional analysis and intermediate asymptotics* (No. 14). Cambridge Univer-
453 sity Press.
- 454 Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., ... others (2024).
455 Climate-invariant machine learning. *Science Advances*, 10(6), eadj7250. doi:
456 <https://doi.org/10.1126/sciadv.adj7250>
- 457 Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learn-
458 ing* (Vol. 4) (No. 4). Springer.
- 459 Bolton, T., & Zanna, L. (2019). Applications of Deep Learning to Ocean Data In-
460 ference and Subgrid Parameterization. *Journal of Advances in Modeling Earth
461 Systems*, 11(1), 376–399. doi: <https://doi.org/10.1029/2018MS001472>
- 462 Bridgman, P. W. (1922). *Dimensional analysis*. Yale university press.
- 463 Buckingham, E. (1914). On physically similar systems; illustrations of the use of di-
464 mensional equations. *Physical review*, 4(4), 345. doi: [https://doi.org/10.1103/PhysRev.4.345](https://doi.org/10.1103/
465 PhysRev.4.345)
- 466 Chang, C.-Y., Adcroft, A., Zanna, L., Hallberg, R., & Griffies, S. M. (2023). Remote
467 versus local impacts of energy backscatter on the North Atlantic SST biases in
468 a global ocean model. *Geophysical Research Letters*, 50(21), e2023GL105757.
469 doi: <https://doi.org/10.1029/2023GL105757>
- 470 Christensen, H., & Zanna, L. (2022). Parametrization in Weather and Climate
471 Models. Oxford University Press. doi: [https://doi.org/10.1093/acrefore/9780190228620.013.826](https://doi.org/10.1093/acrefore/
472 9780190228620.013.826)
- 473 Christopoulos, C., Lopez-Gomez, I., Beucler, T., Cohen, Y., Kawczynski, C., Dun-
474 bar, O. R., & Schneider, T. (2024). Online learning of entrainment closures
475 in a hybrid machine learning parameterization. *Journal of Advances in Mod-
476eling Earth Systems*, 16(11), e2024MS004485. doi: [https://doi.org/10.1029/2024MS004485](https://doi.org/10.1029/
477 2024MS004485)
- 478 Connolly, A., Cheng, Y., Walters, R., Wang, R., Yu, R., & Gentine, P. (2025). Deep
479 learning turbulence closures generalize best with physics-based methods. *Au-
480 thorea Preprints*. doi: [https://doi.org/10.22541/essoar.173869578.80400701/v1](https://doi.org/10.22541/essoar.173869578.80400701/
481 v1)
- 482 Cunningham, S. A., Kanzow, T., Rayner, D., Baringer, M. O., Johns, W. E.,
483 Marotzke, J., ... others (2007). Temporal variability of the Atlantic merid-
484 ional overturning circulation at 26.5 N. *science*, 317(5840), 935–938. doi:
485 <https://doi.org/10.1126/science.1141304>
- 486 Donohue, K. A., Tracey, K., Watts, D. R., Chidichimo, M. P., & Chereskin,
487 T. (2016). Mean Antarctic Circumpolar Current transport measured
488 in Drake Passage. *Geophysical Research Letters*, 43(22), 11–760. doi:
489 <https://doi.org/10.1002/2016GL070319>
- 490 Ferrari, R., & Wunsch, C. (2009). Ocean circulation kinetic energy: Reservoirs,
491 sources, and sinks. *Annual Review of Fluid Mechanics*, 41, 253–282. doi:
492 <https://doi.org/10.1146/annurev.fluid.40.111406.102139>
- 493 Fox-Kemper, B., Adcroft, A., Böning, C. W., Chassignet, E. P., Curchitser, E., Dan-
494 abasoglu, G., ... others (2019). Challenges and prospects in ocean circulation
495 models. *Frontiers in Marine Science*, 6, 65. doi: [https://doi.org/10.3389/fmars.2019.00065](https://doi.org/10.3389/
496 fmars.2019.00065)
- 497 Garcia, H. E., Boyer, T. P., Baranova, O. K., et al. (2019). *World Ocean Atlas
498 2018: Product Documentation [Dataset]*. Retrieved from [https://www.ncei.gov](https://www.ncei.
499 gov)

- 500 .noaa.gov/data/oceans/woa/WOA18/DATA/ doi: <https://doi.org/10.25923/tzyw-rp36>
- 501 Griffies, S. M. (2018). *Fundamentals of ocean climate models*. Princeton University
502 press. doi: <https://doi.org/10.2307/j.ctv301gzg>
- 503 Griffies, S. M., Adcroft, A., Beadling, R. L., Bushuk, M., Chang, C.-Y., Drake,
504 H. F., ... others (2024). The GFDL-CM4X climate model hierarchy,
505 Part I: model description and thermal properties. *Authorea Preprints*. doi:
506 <https://doi.org/10.22541/essoar.173282145.53065190/v1>
- 507 Griffies, S. M., Winton, M., Anderson, W. G., Benson, R., Delworth, T. L., Dufour,
508 C. O., ... others (2015). Impacts on ocean heat from transient mesoscale
509 eddies in a hierarchy of climate models. *Journal of Climate*, 28(3), 952–977.
510 doi: <https://doi.org/10.1175/JCLI-D-14-00353.1>
- 511 Grooms, I., Agarwal, N., Marques, G. M., Pégion, P., & Yassin, H. (2024). The
512 Stochastic GM+ E closure: A framework for coupling stochastic backscatter
513 with the Gent and McWilliams parameterization. *Authorea Preprints*. doi:
514 <https://doi.org/10.22541/essoar.172118408.85625257/v1>
- 515 Grooms, I., Loose, N., Abernathey, R., Steinberg, J., Bachman, S. D., Marques,
516 G. M., ... Yankovsky, E. (2021). Diffusion-based smoothers for spatial filtering
517 of gridded geophysical data. *Journal of Advances in Modeling Earth Systems*,
518 13(9), e2021MS002552. doi: <https://doi.org/10.1029/2021MS002552>
- 519 Guan, Y., Subel, A., Chattopadhyay, A., & Hassanzadeh, P. (2022). Learning
520 physics-constrained subgrid-scale closures in the small-data regime for sta-
521 ble and accurate LES. *Physica D: Nonlinear Phenomena*, 133568. doi:
522 <https://doi.org/10.1016/j.physd.2022.133568>
- 523 Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization
524 of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*,
525 13(9), e2021MS002534. doi: <https://doi.org/10.1029/2021MS002534>
- 526 Gultekin, C., Subel, A., Zhang, C., Leibovich, M., Perezhogin, P., Adcroft, A.,
527 ... Zanna, L. (2024). An analysis of deep learning parameterizations
528 for ocean subgrid eddy forcing. *arXiv preprint arXiv:2411.06604*. doi:
529 <https://doi.org/10.48550/arXiv.2411.06604>
- 530 Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The ele-
531 ments of statistical learning: data mining, inference, and prediction* (Vol. 2).
532 Springer.
- 533 Hewitt, H. T., Roberts, M., Mathiot, P., Biastoch, A., Blockley, E., Chassignet,
534 E. P., ... others (2020). Resolving and parameterising the ocean mesoscale in
535 earth system models. *Current Climate Change Reports*, 6(4), 137–152. doi:
536 <https://doi.org/10.1007/s40641-020-00164-w>
- 537 Jansen, M. F., Adcroft, A., Khani, S., & Kong, H. (2019). Toward an energetically
538 consistent, resolution aware parameterization of ocean mesoscale eddies. *Jour-
539 nal of Advances in Modeling Earth Systems*, 11(8), 2844–2860. doi: <https://doi.org/10.1029/2019MS001750>
- 540 Jansen, M. F., & Held, I. M. (2014). Parameterizing subgrid-scale eddy effects using
541 energetically consistent backscatter. *Ocean Modelling*, 80, 36–48. doi: <https://doi.org/10.1016/j.ocemod.2014.06.002>
- 542 Jansen, M. F., Held, I. M., Adcroft, A., & Hallberg, R. (2015). Energy budget-based
543 backscatter in an eddy permitting primitive equation model. *Ocean Modelling*,
544 94, 15–26. doi: <https://doi.org/10.1016/j.ocemod.2015.07.015>
- 545 Juricke, S., Danilov, S., Kutsenko, A., & Oliver, M. (2019). Ocean kinetic
546 energy backscatter parametrizations on unstructured grids: Impact on
547 mesoscale turbulence in a channel. *Ocean Modelling*, 138, 51–67. doi:
548 <https://doi.org/10.1016/j.ocemod.2019.03.009>
- 549 Kang, M., Jeon, Y., & You, D. (2023). Neural-network-based mixed subgrid-scale
550 model for turbulent flow. *Journal of Fluid Mechanics*, 962, A38. doi: <https://doi.org/10.1017/jfm.2023.260>

- 555 Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., ...
 556 others (2021). Physics-informed machine learning: case studies for weather
 557 and climate modelling. *Philosophical Transactions of the Royal Society A*,
 558 379(2194), 20200093. doi: <https://doi.org/10.1098/rsta.2020.0093>
- 559 Large, W., & Yeager, S. (2009). The global climatology of an interannually varying
 560 air-sea flux data set. *Climate dynamics*, 33, 341–364. doi: <https://doi.org/10.1007/s00382-008-0441-3>
- 561 Leith, C. (1996). Stochastic models of chaotic systems. *Physica D: Nonlinear Phe-*
 562 *nomena*, 98(2-4), 481–491. doi: [https://doi.org/10.1016/0167-2789\(96\)00107-8](https://doi.org/10.1016/0167-2789(96)00107-8)
- 563 Lévy, M., Klein, P., Tréguier, A.-M., Iovino, D., Madec, G., Masson, S., & Taka-
 564 hashi, K. (2010). Modifications of gyre circulation by sub-mesoscale
 565 physics. *Ocean Modelling*, 34(1-2), 1–15. doi: <https://doi.org/10.1016/j.ocemod.2010.04.001>
- 566 Li, H., Xie, J., Zhang, C., Zhang, Y., & Zhao, Y. (2025). A transformer-based
 567 convolutional method to model inverse cascade in forced two-dimensional tur-
 568 bulence. *Journal of Computational Physics*, 520, 113475. doi: <https://doi.org/10.1016/j.jcp.2024.113475>
- 569 Ling, J., Kurzawski, A., & Templeton, J. (2016). Reynolds averaged turbulence
 570 modelling using deep neural networks with embedded invariance. *Journal of*
 571 *Fluid Mechanics*, 807, 155–166. doi: <https://doi.org/10.1017/jfm.2016.615>
- 572 Locarnini, M., Mishonov, A., Baranova, O., Boyer, T., Zweng, M., Garcia, H., ...
 573 others (2018). World ocean atlas 2018, volume 1: Temperature. *NOAA Atlas*
 574 *NESDIS*.
- 575 Loose, N., Abernathey, R., Grooms, I., Busecke, J., Guillaumin, A., Yankovsky, E.,
 576 ... others (2022). GCM-filters: A Python package for diffusion-based spatial
 577 filtering of gridded data. *Journal of Open Source Software*, 7(70).
- 578 Loose, N., Bachman, S., Grooms, I., & Jansen, M. (2023). Diagnosing scale-
 579 dependent energy cycles in a high-resolution isopycnal ocean model. *Jour-*
 580 *nal of Physical Oceanography*, 53(1), 157–176. doi: <https://doi.org/10.1175/JPO-D-22-0083.1>
- 581 Loose, N., Marques, G. M., Adcroft, A., Bachman, S., Griffies, S. M., Grooms, I.,
 582 ... Jansen, M. F. (2023). Comparing two parameterizations for the re-
 583 stratification effect of mesoscale eddies in an isopycnal ocean model. *Jour-*
 584 *nal of Advances in Modeling Earth Systems*, 15(12), e2022MS003518. doi:
 585 <https://doi.org/10.1029/2022MS003518>
- 586 Lund, T. S., & Novikov, E. (1993). Parameterization of subgrid-scale stress by the
 587 velocity gradient tensor. *Annual Research Briefs*, 1992.
- 588 Maddison, J. R. (2024). Online learning in idealized ocean gyres. *arXiv preprint*
 589 *arXiv:2412.06393*. doi: <https://doi.org/10.48550/arXiv.2412.06393>
- 590 Mana, P. P., & Zanna, L. (2014). Toward a stochastic parameterization of ocean
 591 mesoscale eddies. *Ocean Modelling*, 79, 1–20. doi: <https://doi.org/10.1016/j.jocemod.2014.04.002>
- 592 Marques, G. M., Loose, N., Yankovsky, E., Steinberg, J. M., Chang, C.-Y., Bhamidi-
 593 pati, N., ... others (2022). NeverWorld2: An idealized model hierarchy to in-
 594 vestigate ocean mesoscale eddies across resolutions. *Geoscientific Model Devel-*
 595 *opment*, 15(17), 6567–6579. doi: <https://doi.org/10.5194/gmd-15-6567-2022>
- 596 Marques, G. M., et al. (2022). *Simulation data in idealized ocean configuration Nev-*
 597 *erWorld2 [Dataset]*. UCAR/NCAR - CISL - CDP. doi: <https://doi.org/10.26024/f130-ev71>
- 598 Maulik, R., & San, O. (2017). A neural network approach for the blind deconvo-
 599 lution of turbulent flows. *Journal of Fluid Mechanics*, 831, 151–181. doi:
 600 <https://doi.org/10.1017/jfm.2017.637>
- 601 Maulik, R., San, O., Rasheed, A., & Vedula, P. (2019). Subgrid modelling for two-
 602 dimensional turbulence using neural networks. *Journal of Fluid Mechanics*,

- 610 858, 122–144. doi: <https://doi.org/10.1017/jfm.2018.770>
- 611 Meneveau, C., & Katz, J. (2000). Scale-invariance and turbulence models for large-
612 eddy simulation. *Annual Review of Fluid Mechanics*, 32(1), 1–32. doi: <https://doi.org/10.1146/annurev.fluid.32.1.1>
- 613 Moat, B., DA, S., et al. (2025). *Atlantic meridional overturning circulation
614 observed by the RAPID-MOCHA-WBTS (RAPID-Meridional Overturn-
615 ing Circulation and Heatflux Array-Western Boundary Time Series) ar-
616 ray at 26N from 2004 to 2023 (v2023.1a) [Dataset]*. British Oceanographic
617 Data Centre - Natural Environment Research Council, UK. doi:
618 <https://doi.org/10.5285/33826d6e-801c-b0a7-e063-7086abc0b9db>
- 619 Pawar, S., San, O., Rasheed, A., & Vedula, P. (2020). A priori analysis on deep
620 learning of subgrid-scale parameterizations for kraichnan turbulence. *Theoretical
621 and Computational Fluid Dynamics*, 34(4), 429–455. doi: <https://doi.org/10.1007/s00162-019-00512-z>
- 622 Perezhogin, P. (2025). *Generalizable neural-network parameterization of mesoscale
623 eddies in idealized and global ocean models [Software]*. Zenodo. doi: <https://doi.org/10.5281/zenodo.16056926>
- 624 Perezhogin, P., Adcroft, A., & Zanna, L. (2025). *Generalizable neural-network
625 parameterization of mesoscale eddies in idealized and global ocean models
626 [Dataset]*. Zenodo. doi: <https://doi.org/10.5281/zenodo.16058005>
- 627 Perezhogin, P., Balakrishna, A., & Agrawal, R. (2024). Large eddy simulation of
628 ocean mesoscale eddies. In *Proceedings of the Summer Program 2024, Center
629 for Turbulence Research, Stanford University* (p. 507–516). Retrieved from
630 <https://arxiv.org/abs/2501.05357>
- 631 Perezhogin, P., Zhang, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2024).
632 A stable implementation of a data-driven scale-aware mesoscale parameteriza-
633 tion. *Journal of Advances in Modeling Earth Systems*, 16(10), e2023MS004104.
634 doi: <https://doi.org/10.1029/2023MS004104>
- 635 Pope, S. B. (1975). A more general effective-viscosity hypothesis. *Journal of Fluid
636 Mechanics*, 72(2), 331–340. doi: <https://doi.org/10.1017/S0022112075003382>
- 637 Prakash, A., Jansen, K. E., & Evans, J. A. (2022). Invariant data-driven sub-
638 grid stress modeling in the strain-rate eigenframe for large eddy simulation.
639 *Computer Methods in Applied Mechanics and Engineering*, 399, 115457. doi:
640 <https://doi.org/10.1016/j.cma.2022.115457>
- 641 Prakash, A., Jansen, K. E., & Evans, J. A. (2024). Invariant data-driven subgrid
642 stress modeling on anisotropic grids for large eddy simulation. *Computer Meth-
643 ods in Applied Mechanics and Engineering*, 422, 116807. doi: <https://doi.org/10.1016/j.cma.2024.116807>
- 644 Reissmann, M., Hasslberger, J., Sandberg, R. D., & Klein, M. (2021). Application
645 of gene expression programming to a-posteriori les modeling of a
646 taylor green vortex. *Journal of Computational Physics*, 424, 109859. doi:
647 <https://doi.org/10.1016/j.jcp.2020.109859>
- 648 Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2023).
649 Benchmarking of machine learning ocean subgrid parameterizations in an
650 idealized model. *Journal of Advances in Modeling Earth Systems*, 15(1),
651 e2022MS003258. doi: <https://doi.org/10.1029/2022MS003258>
- 652 Salmon, R. (1980). Baroclinic instability and geostrophic turbulence. *Geophys-
653 ical & Astrophysical Fluid Dynamics*, 15(1), 167–211. doi: <https://doi.org/10.1080/03091928008241178>
- 654 Sane, A., Reichl, B. G., Adcroft, A., & Zanna, L. (2023). Parameterizing vertical
655 mixing coefficients in the ocean surface boundary layer using neural networks.
656 *Journal of Advances in Modeling Earth Systems*, 15(10), e2023MS003890. doi:
657 <https://doi.org/10.1029/2023MS003890>
- 658 Schneider, T., Leung, L. R., & Wills, R. C. (2024). Opinion: Optimizing
659 climate models with process knowledge, resolution, and artificial intelli-
660 gence.

- gence. *Atmospheric Chemistry and Physics*, 24(12), 7041–7062. doi: <https://doi.org/10.5194/acp-24-7041-2024>
- Shankar, V., Chakraborty, D., Viswanathan, V., & Maulik, R. (2025). Differentiable turbulence: Closure as a partial differential equation constrained optimization. *Physical Review Fluids*, 10(2), 024605.
- Smagorinsky, J. (1963). General circulation experiments with the primitive equations: I. the basic experiment. *Monthly weather review*, 91(3), 99–164. doi: [https://doi.org/10.1175/1520-0493\(1963\)091%3C0099:GCEWTP%3E2.3.CO;2](https://doi.org/10.1175/1520-0493(1963)091%3C0099:GCEWTP%3E2.3.CO;2)
- Srinivasan, K., Chekroun, M. D., & McWilliams, J. C. (2024). Turbulence closure with small, local neural networks: Forced two-dimensional and β -plane flows. *Journal of Advances in Modeling Earth Systems*, 16(4), e2023MS003795. doi: <https://doi.org/10.1029/2023MS003795>
- Wang, Y., Yuan, Z., Wang, X., & Wang, J. (2022). Constant-coefficient spatial gradient models for the sub-grid scale closure in large-eddy simulation of turbulence. *Physics of Fluids*, 34(9). doi: <https://doi.org/10.1063/5.0101356>
- Wang, Y., Yuan, Z., Xie, C., & Wang, J. (2021). Artificial neural network-based spatial gradient models for large-eddy simulation of turbulence. *AIP Advances*, 11(5). doi: <https://doi.org/10.1063/5.0053590>
- Xie, C., Wang, J., & E, W. (2020). Modeling subgrid-scale forces by spatial artificial neural networks in large eddy simulation of turbulence. *Physical Review Fluids*, 5(5), 054606. doi: <https://doi.org/10.1103/PhysRevFluids.5.054606>
- 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. doi: <https://doi.org/10.1029/2023MS003915>
- Yankovsky, E., Bachman, S., Smith, K. S., & Zanna, L. (2024). Vertical structure and energetic constraints for a backscatter parameterization of ocean mesoscale eddies. *Journal of Advances in Modeling Earth Systems*, 16(7), e2023MS004093. doi: <https://doi.org/10.1029/2023MS004093>
- Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17), e2020GL088376. doi: <https://doi.org/10.1029/2020GL088376>
- Zhang, C., Perezhigin, 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. doi: <https://doi.org/10.1029/2023MS003697>

Supporting Information for "Generalizable neural-network parameterization of mesoscale eddies in idealized and global ocean models"

Pavel Perezhigin¹, Alistair Adcroft³, Laure Zanna^{1,2}

¹Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

²Center for Data Science, New York University, New York, NY, USA

³Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA

Contents of this file

1. Text S1 to S4
2. Tables S1 to S3
3. Figures S1 to S6

Text S1. Known parameterizations as a special case of dimensional scaling

Here, we show that enforcing the dimensional scaling constraint to the ANN parameterization is not too restrictive and admits multiple known parameterizations as a special case with continuous functional representations (see Prakash, Jansen, and Evans (2022) for discussion). We also show that these functional representations do not depend on the normalization factor ($\|\mathbf{X}\|_2^2$) explicitly. This suggests that the choice of the normalization factor primarily affects the range of inputs to the neural network, but not the function to be learnt.

We denote the components of the predicted momentum fluxes as follows:

$$\widehat{\mathbf{T}}(\mathbf{X}, \Delta) = \Delta^2 \|\mathbf{X}\|_2^2 \text{ANN}_\theta(\mathbf{X}/\|\mathbf{X}\|_2) \equiv \quad (1)$$

$$\Delta^2 \|\mathbf{X}\|_2^2 \begin{pmatrix} \text{ANN}_\theta^{xx}(\mathbf{x}) & \text{ANN}_\theta^{xy}(\mathbf{x}) \\ \text{ANN}_\theta^{xy}(\mathbf{x}) & \text{ANN}_\theta^{yy}(\mathbf{x}) \end{pmatrix}, \quad (2)$$

where the vector of input features is

$$\mathbf{X} = \begin{pmatrix} [\bar{\sigma}_S] \uparrow_9 \\ [\bar{\sigma}_T] \uparrow_9 \\ [\bar{\omega}] \uparrow_9 \end{pmatrix} \in \mathbb{R}^{27} \quad (3)$$

and $\mathbf{x} = \mathbf{X}/\|\mathbf{X}\|_2$.

Smagorinsky parameterization

We first consider a Smagorinsky (1963) subgrid parameterization:

$$\widehat{\mathbf{T}} = C_S \Delta^2 \sqrt{\bar{\sigma}_S^2 + \bar{\sigma}_T^2} \begin{pmatrix} \bar{\sigma}_T & \bar{\sigma}_S \\ \bar{\sigma}_S & -\bar{\sigma}_T \end{pmatrix}. \quad (4)$$

This subgrid model can be given in the form of Eq. (2) if ANN parameterizes the following functions:

$$\text{ANN}_\theta^{xx}(\mathbf{x}) = C_S x_{14} \sqrt{x_5^2 + x_{14}^2} \quad (5)$$

$$\text{ANN}_{\theta}^{yy}(\mathbf{x}) = -\text{ANN}_{\theta}^{xx}(\mathbf{x}) \quad (6)$$

$$\text{ANN}_{\theta}^{xy}(\mathbf{x}) = C_S x_5 \sqrt{x_5^2 + x_{14}^2}, \quad (7)$$

where x_5 and x_{14} represent components of the non-dimensional vector \mathbf{x} which are equal to $\bar{\sigma}_S/\|\mathbf{X}\|_2$ and $\bar{\sigma}_T/\|\mathbf{X}\|_2$ in the center of 3×3 spatial stencil, respectively. The derived functions are continuous on a bounded domain ($|x_i| \leq 1$), and thus they can be easily learned with the ANN. The functional representations of the parameterizations derived below are continuous as well.

Zanna-Bolton 2020 parameterization

Similarly, we can show that Zanna and Bolton (2020) parameterization

$$\hat{\mathbf{T}} = -\gamma \Delta^2 \begin{pmatrix} -\bar{\omega} \bar{\sigma}_S & \bar{\omega} \bar{\sigma}_T \\ \bar{\omega} \bar{\sigma}_T & \bar{\omega} \bar{\sigma}_S \end{pmatrix} - \frac{1}{2} \gamma \Delta^2 (\bar{\omega}^2 + \bar{\sigma}_T^2 + \bar{\sigma}_S^2) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (8)$$

can be represented as follows:

$$\text{ANN}_{\theta}^{xx}(\mathbf{x}) = \gamma x_5 x_{23} - \frac{1}{2} \gamma (x_5^2 + x_{14}^2 + x_{23}^2), \quad (9)$$

$$\text{ANN}_{\theta}^{yy}(\mathbf{x}) = -\gamma x_5 x_{23} - \frac{1}{2} \gamma (x_5^2 + x_{14}^2 + x_{23}^2), \quad (10)$$

$$\text{ANN}_{\theta}^{xy}(\mathbf{x}) = -\gamma x_{14} x_{23}. \quad (11)$$

Leith 1996 parameterization

Next, we consider Leith (1996) parameterization:

$$\hat{\mathbf{T}} = C_L \Delta^3 |\nabla \bar{\omega}| \begin{pmatrix} \bar{\sigma}_T & \bar{\sigma}_S \\ \bar{\sigma}_S & -\bar{\sigma}_T \end{pmatrix}. \quad (12)$$

By approximating the gradient with central differences and assuming an isotropic and uniform grid, we obtain:

$$\text{ANN}_{\theta}^{xx}(\mathbf{x}) = \frac{1}{2} C_L x_{14} \sqrt{(x_{24} - x_{22})^2 + (x_{26} - x_{20})^2}, \quad (13)$$

:

$$\text{ANN}_{\theta}^{yy}(\mathbf{x}) = -\text{ANN}_{\theta}^{xx}(\mathbf{x}), \quad (14)$$

$$\text{ANN}_{\theta}^{xy}(\mathbf{x}) = \frac{1}{2}C_L x_5 \sqrt{(x_{24} - x_{22})^2 + (x_{26} - x_{20})^2} \quad (15)$$

Biharmonic Smagorinsky parameterization

The biharmonic Smagorinsky subgrid model has the form:

$$\hat{\mathbf{T}} = -C_S \Delta^4 \sqrt{\bar{\sigma}_S^2 + \bar{\sigma}_T^2} \nabla^2 \begin{pmatrix} \bar{\sigma}_T & \bar{\sigma}_S \\ \bar{\sigma}_S & -\bar{\sigma}_T \end{pmatrix}. \quad (16)$$

By approximating the ∇^2 operator on an isotropic and uniform grid, we obtain:

$$\text{ANN}_{\theta}^{xx}(\mathbf{x}) = -C_S(x_{15} + x_{13} + x_{17} + x_{11} - 4x_{14}) \sqrt{x_5^2 + x_{14}^2}, \quad (17)$$

$$\text{ANN}_{\theta}^{yy}(\mathbf{x}) = -\text{ANN}_{\theta}^{xx}(\mathbf{x}), \quad (18)$$

$$\text{ANN}_{\theta}^{xy}(\mathbf{x}) = -C_S(x_6 + x_4 + x_8 + x_2 - 4x_5) \sqrt{x_5^2 + x_{14}^2}. \quad (19)$$

Text S2. Robustness of division by small numbers

The robustness of the parameterization with the dimensional scaling,

$$\hat{\mathbf{T}}(\mathbf{X}, \Delta) = \Delta^2 \|\mathbf{X}\|_2^2 \text{ANN}_{\theta}(\mathbf{X}/\|\mathbf{X}\|_2), \quad (20)$$

at $\mathbf{X} = \mathbf{0}$ is achieved as follows. We first identify that most known parameterizations, such as Smagorinsky (1963) and Leith (1996), predict zero fluxes when the velocity gradients are zero. We enforce the same property for our parameterization by extending Eq. (20) with:

$$\hat{\mathbf{T}}(\mathbf{X} = \mathbf{0}, \Delta) = \mathbf{0}. \quad (21)$$

Numerically, this property is implemented by adding a very small number (10^{-30}s^{-1}) to the denominator in Eq. (20).

Additionally, we ensure that Eq. (20) is continuous at $\mathbf{X} = \mathbf{0}$, that is, the corresponding limit exists and is equal to the function value (zero):

$$\lim_{\mathbf{X} \rightarrow \mathbf{0}} \widehat{\mathbf{T}}(\mathbf{X}, \Delta) = \widehat{\mathbf{T}}(\mathbf{X} = \mathbf{0}, \Delta) = \mathbf{0}. \quad (22)$$

The function ANN_θ is continuous as a composition of continuous activation functions (ReLU). Furthermore, for any $\|\mathbf{X}\|_2 > 0$, the function ANN_θ is evaluated on a unit sphere, which is a compact set. Therefore, the continuous function ANN_θ is bounded on the compact set by some constant $A(\theta)$ that depends only on the trainable parameters θ .

We verify the limit (Eq. (22)) by inequality:

$$|\Delta^2 \|\mathbf{X}\|_2^2 \text{ANN}_\theta(\mathbf{X}/\|\mathbf{X}\|_2)| \leq A(\theta) \Delta^2 \|\mathbf{X}\|_2^2 \rightarrow 0 \text{ as } \mathbf{X} \rightarrow \mathbf{0}. \quad (23)$$

Text S3. Details of the training algorithm

The training dataset is created using four coarse-graining factors, selected to be similar to those used in Gultekin et al. (2024), and 10 depths (extending Gultekin et al. (2024)), see Table S1.

ANN model architecture

For offline analysis, we use an ANN, also known as a multilayer perceptron (MLP), with two hidden layers, 32 neurons each, in a total of 2051 parameters, both for parameterizations with and without dimensional scaling (see Table S1). For online implementation, the ANN model is chosen to be smaller (see Table S1): it has only a single hidden layer with 20 neurons, as in Prakash et al. (2022), with a total of 623 trainable parameters. We verified that reducing the number of neurons for online implementation does not significantly impact the response in kinetic and potential energy and the time-mean sea surface tem-

perature in short 5-year simulations in the global ocean model OM4 (Figure S5). Thus, we keep the smaller ANN for online implementation to bound its computational cost to within $\approx 10\%$ of the global ocean model runtime.

Training algorithm and boundary conditions

We train the ANN model on data from the full globe, similarly to Gultekin et al. (2024). The loss function is defined to optimize for the divergence ($\nabla \cdot \mathbf{T}$) of subfilter fluxes (\mathbf{T}) similarly to Zanna and Bolton (2020) and Srinivasan, Chekroun, and McWilliams (2024). The mean squared error (MSE) loss is minimized on every 2D snapshot of subfilter forcing \mathbf{S} and normalized by the corresponding l_2 -norm of \mathbf{S} (Agdestein & Sanderse, 2025):

$$\mathcal{L}_{\text{MSE}} = \|(\mathbf{S} - \nabla \cdot \hat{\mathbf{T}}) \cdot m\|_2^2 / \|\mathbf{S} \cdot m\|_2^2, \quad (24)$$

where m is the mask of wet points. The input features (velocity gradients, \mathbf{X}) and predicted subfilter fluxes are set to zero on the land as well: $\hat{\mathbf{T}} \equiv m \cdot \hat{\mathbf{T}}(m \cdot \mathbf{X})$. That is, we impose zero Neumann boundary condition (Zhang et al., 2024) and free-slip boundary condition. We found that including the grid points adjacent to the land to the loss function is essential for ensuring the numerical stability of online runs. Another important design choice for online numerical stability is performing the ANN inference on the collocated, rather than on the staggered grid, similarly to Guillaumin and Zanna (2021) and Agdestein and Sanderse (2025).

The loss function (Eq. (24)) is evaluated and minimized for a total of 16000 two-dimensional snapshots during training, see Table S1. We do not use any regularizations, such as weight decay, during the training of ANNs because the size of the dataset is much bigger relative to the number of trainable parameters. We verified that the offline skill on

training and testing data is very similar, suggesting that there is no overfitting, and there is no need for regularization.

Sensitivity to the random seed

The R-squared of the offline predictions of the ANN is almost insensitive to the random seed used to initialize the training algorithm. In addition, the prediction errors $\mathbf{S} - \nabla \cdot \hat{\mathbf{T}}$ are highly correlated between different seeds (as in Srinivasan et al. (2024)). We also confirmed that the kinetic energy is nearly unchanged in online two-layer Double Gyre experiments, using ANNs generated from different initializations of the training algorithm, similarly to Zhang et al. (2024). However, there is some sensitivity to the training algorithm initialization for the mean flow prediction: the response pattern in the mean flow is similar, but the response magnitude can vary by 50%. The sensitivity of the mean fields to the random seed is not apparent in the global ocean configuration OM4.

Choice of the filter scale in the training dataset

The filtering operator used is a Gaussian filter implemented in the package GCM-Filters (Grooms et al., 2021; Loose et al., 2022) with width $\bar{\Delta}$, chosen in relation to the coarse grid spacing Δ . The filter-to-grid width ratio parameter ($FGR = \bar{\Delta}/\Delta$) represents the strength of the subfilter parameterization: relatively low value of FGR ($\bar{\Delta}/\Delta = 1$) in the training dataset results in a learned parameterization that has negligible effect in online simulations. On the other hand, a relatively large value ($FGR = 4$) results in over-energized grid-scale features. The value used here ($FGR = 3$) corresponds to the strongest parameterization effect without generating grid-scale noise. Note that the optimal FGR parameter depends on the numerical and physical dissipation schemes present in the

ocean model, as the ANN subfilter parameterization alone does not produce enough grid-scale dissipation. For the discussion of how to choose FGR parameter see Perezhigin, Balakrishna, and Agrawal (2024); Perezhigin, Zhang, Adcroft, Fernandez-Granda, and Zanna (2024); Perezhigin and Glazunov (2023).

Text S4. Online implementation and numerical stability in MOM6

The trained parameters of the ANN subfilter model are saved to a NetCDF file and read by the numerical ocean model during initialization. The neural network inference is implemented using the Fortran module of Sane, Reichl, Adcroft, and Zanna (2023). The ANN inference takes $\approx 10\%$ of the ocean model runtime, for a neural network with one hidden layer and 20 neurons. However, the inference can be further accelerated, as we found that the inference in Python is generally faster than in Fortran.

The implemented ANN parameterization works stably (free of NaNs in prognostic fields) in idealized Double Gyre and global ocean OM4 configurations, without any tuning, in part because the biharmonic Smagorinsky model provides the dissipation. In the idealized configuration NeverWorld2 (NW2, Marques et al. (2022)), however, tuning is required to improve the numerical stability even when a backscatter parameterization (whether our ANN or more traditional parametrization) is used together with biharmonic Smagorinsky model; e.g., Yankovsky, Bachman, Smith, and Zanna (2024). We have modified the ANN parameterization to achieve stability, without optimizing for online metrics, using a set of minimal changes. Our tuning includes attenuating the magnitude of the ANN parameterization in high-strain regions following Perezhigin, Zhang, et al. (2024) and

allowing the MOM6 dynamical core to truncate velocities if they are too big. Additionally, at resolution $1/6^\circ$ in NW2, we had to reduce the time stepping interval.

Table S1. Parameters of the training data and artificial neural network (ANN) model

Category	Value
Training Data Parameters	
High-resolution data	CM2.6 (Griffies et al., 2015), 0.1° ocean grid
Diagnosed features	$\bar{\sigma}_S$, $\bar{\sigma}_T$, $\bar{\omega}$, \mathbf{S} , \mathbf{T}
Layer Depths (m)	5, 55, 110, 180, 330, 730, 1500, 2500, 3500, 4500
Horizontal grid type	Tripolar
Horizontal extent	All globe including polar latitudes
Coarse Grid Spacing, Δ (nominal)	0.4° , 0.9° , 1.2° , 1.5°
Coarse Grid Spacing, Δ (km, $60S^\circ - 60N^\circ$)	22-44, 50-100, 67-134, 85-167
Gaussian Filter Width, $\bar{\Delta}$	1.2° , 2.7° , 3.6° , 4.5° ; i.e., $\bar{\Delta}/\Delta = 3$
Training / Validation / Test Splitting (years)	181 – 188 / 194 / 199 – 200
Snapshot Averaging Interval	5 days
Time Separation Between Snapshots	1 month
Number of 2D Snapshots used for Training	$10 \times 4 \times 8 \times 12 = 3840$
Number of Training Iterations	16000 (each iteration randomly selects 2D snapshot)
ANN Parameters	
Input Size	3×3
ANN type	Multilayer Perceptron (MLP)
ANN used for offline analysis	2 hidden layers, 32 neurons each, 2051 parameters
ANN used in online implementation	1 hidden layer with 20 neurons, 623 parameters
Activation Function	ReLU
Note	Regularization is not applied during training

RMSE	0°E	15°E	30°E	45°E
Control(1/3°)	51.3	46.8	49.1	36.8
Yankovsky24(1/3°)	33.5	31.6	29.9	27.4
ZB20-Reynolds(1/3°)	32.7	25.4	26.1	21.6
ANN(1/3°)	35.0	24.3	30.9	28.6
Control(1/4°)	52.1	42.1	40.3	34.6
Yankovsky24(1/4°)	27.8	23.0	20.7	21.3
ZB20-Reynolds(1/4°)	26.9	21.0	18.4	18.5
ANN(1/4°)	29.2	20.0	16.7	19.7
Control(1/6°)	42.7	30.7	31.8	26.7
Yankovsky24(1/6°)	26.2	22.3	16.1	16.8
ZB20-Reynolds(1/6°)	27.7	24.5	18.9	18.4
ANN(1/6°)	23.5	18.8	13.8	14.6

Table S2. Online results in idealized configuration NeverWorld2 at three coarse resolutions (1/3°, 1/4° and 1/6°). The root mean squared errors (RMSE) in 1000-day averaged position of interfaces over four meridional transects at longitudes 0°E, 15°E, 30°E and 45°E. RMSE units are metres. The interfaces for Control and ANN-parameterized runs at resolution 1/4° at longitudes 0°E and 45°E are also shown in Figure S3. The error is computed w.r.t. 1/32° model. Yankovsky24 stands for parameterization of Yankovsky et al. (2024), ZB20-Reynolds stands for Zanna and Bolton (2020) parameterization implemented and modified by Perezhigin, Zhang, et al. (2024). Parameterizations are not retuned when resolution is changed.

	ACC transport [Sv]
1/32°	235.3
Control(1/3°)	242.7
Yankovsky24(1/3°)	237.4
ZB20-Reynolds(1/3°)	230.2
ANN(1/3°)	241.6
Control(1/4°)	245.1
Yankovsky24(1/4°)	229.9
ZB20-Reynolds(1/4°)	225.4
ANN(1/4°)	236.9
Control(1/6°)	243.3
Yankovsky24(1/6°)	230.4
ZB20-Reynolds(1/6°)	219.5
ANN(1/6°)	228.8

Table S3. Online results in idealized configuration NeverWorld2. The ACC transport through the Drake Passage at 0°E averaged over 800 days.

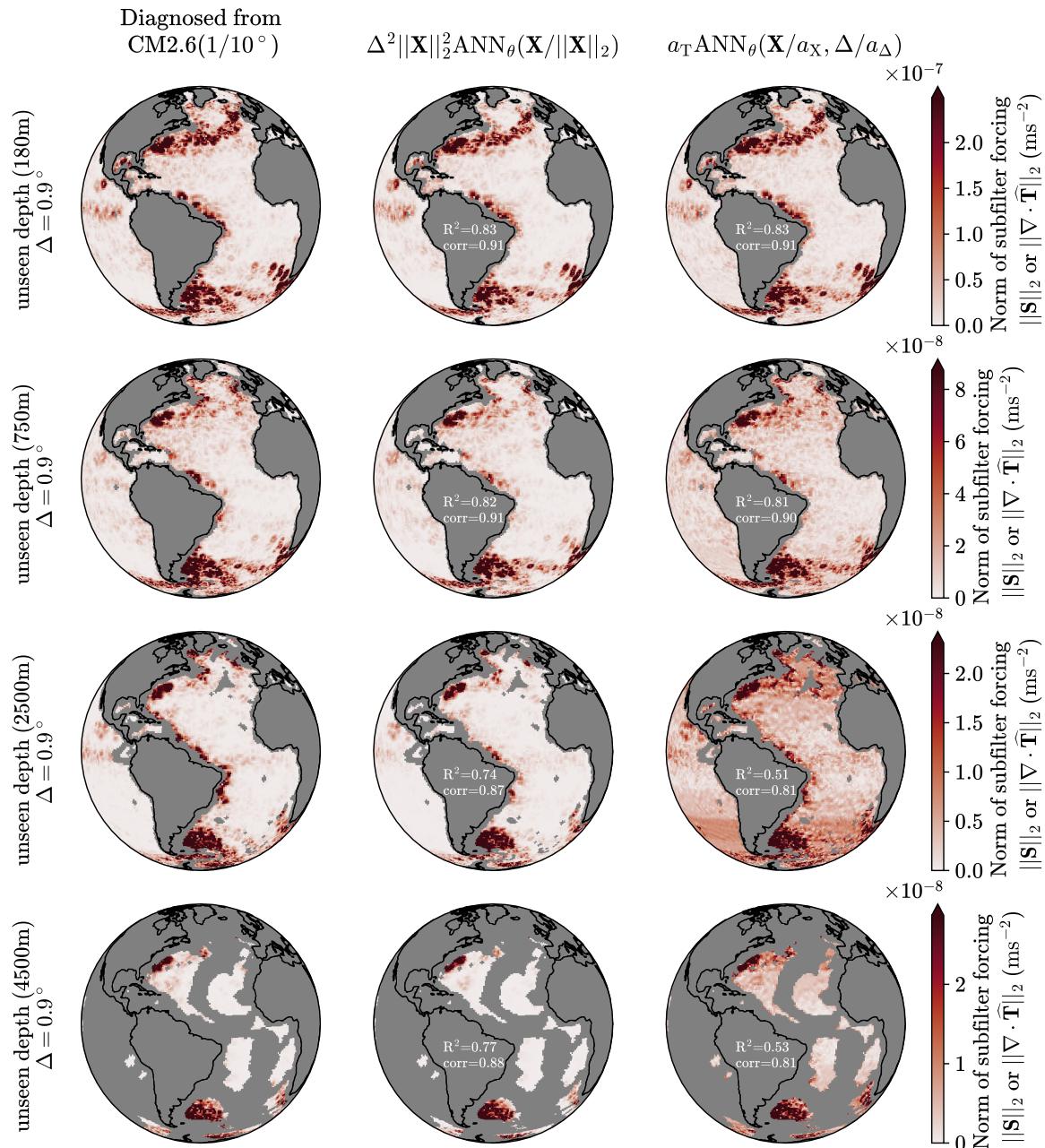


Figure S1. Extension of Figure 1 in the main text with generalization of two ANN parameterizations to multiple unseen depths, but seen resolution used for training (0.9°).

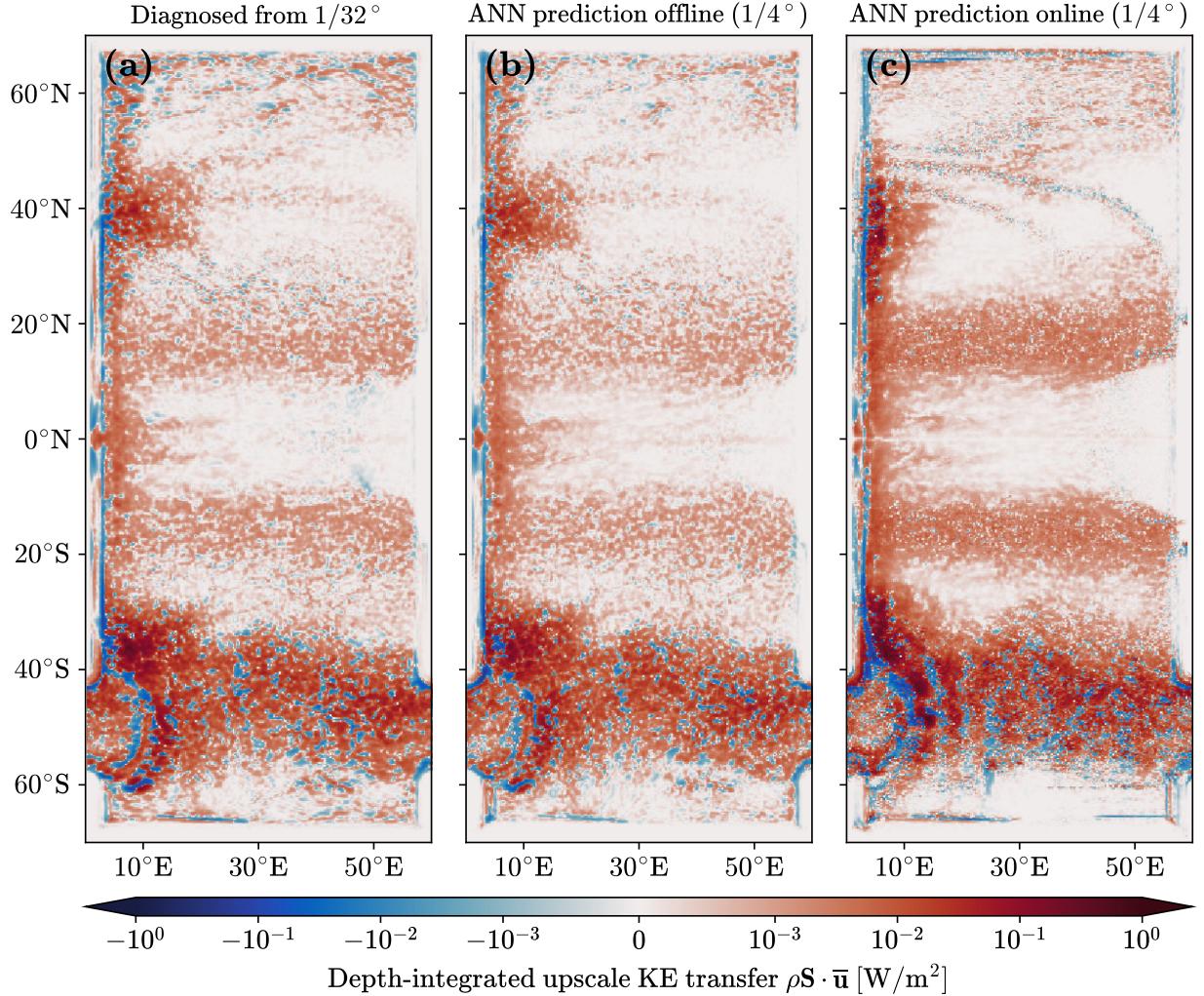


Figure S2. Upscale KE transfer (positive numbers correspond to backscatter) averaged over 800 days and integrated over depth in idealized NeverWorld2 configuration. (a) Diagnosed from high-resolution ($1/32^\circ$) simulation by filtering and coarsegraining, (b) and (c) predicted by the ANN offline and online, respectively, at coarse resolution $1/4^\circ$. The ANN was trained on global ocean data and thus generalizes well to a new configuration as seen in the accurate prediction of KE transfer offline. Prediction offline means that filtered and coarsegrained snapshots of the high-resolution model were given as inputs to the ANN. Slight degradation of prediction online is related to the difference in magnitude of small-scale velocity gradients and large-scale circulation patterns in the coarse ocean model.

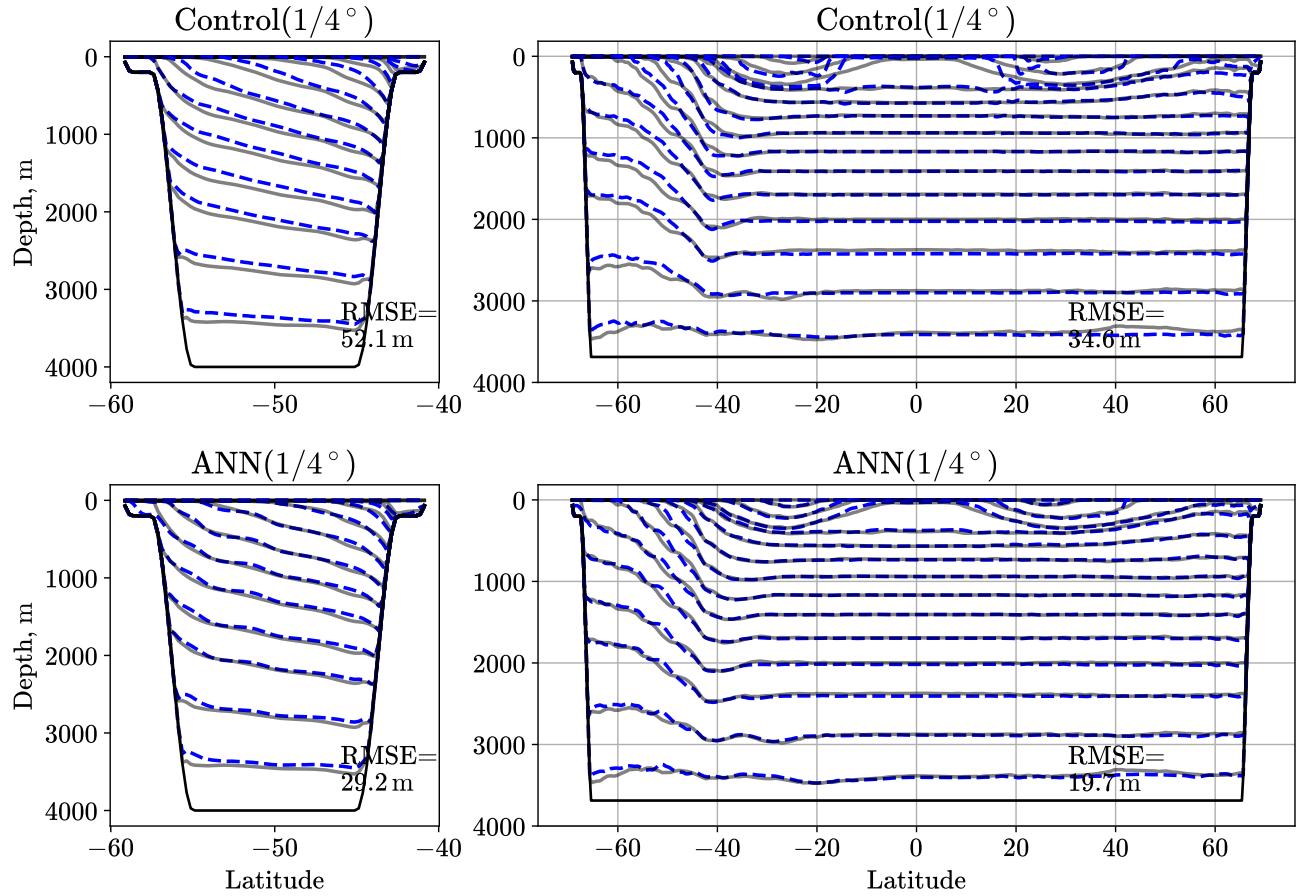


Figure S3. Online results in idealized configuration NeverWorld2. The 1000-days averaged isopycnal interfaces in the meridional transect of Drake Passage (Longitude 0°E, left column) and at Longitude 45°E. The blue dashed lines show the position of interfaces in the coarse-resolution ($1/4^\circ$) experiment, and the gray lines show the interfaces of the high-resolution model $1/32^\circ$. The root mean squared errors (RMSE) between coarse and high-resolution models are provided.

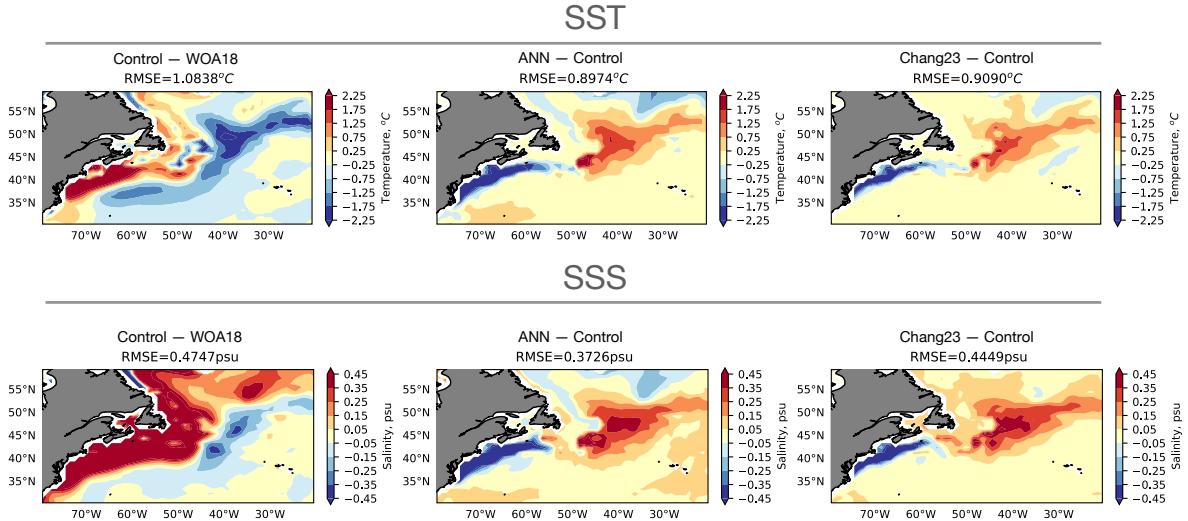


Figure S4. Online results in the global ocean-ice model OM4 (Adcroft et al., 2019), North Atlantic region. Comparison of the ANN parameterization to a baseline parameterization tested in Chang et al. (2023). We consider biases in sea surface temperature (SST), sea surface salinity (SSS). Model output is averaged over years 1981-2007. The observational data for SST and SSS is given by the World Ocean Atlas 2018 (WOA18, Locarnini et al. (2018)). Root mean square errors (RMSEs) between simulations and observations are provided.

Offline metrics in CM2.6

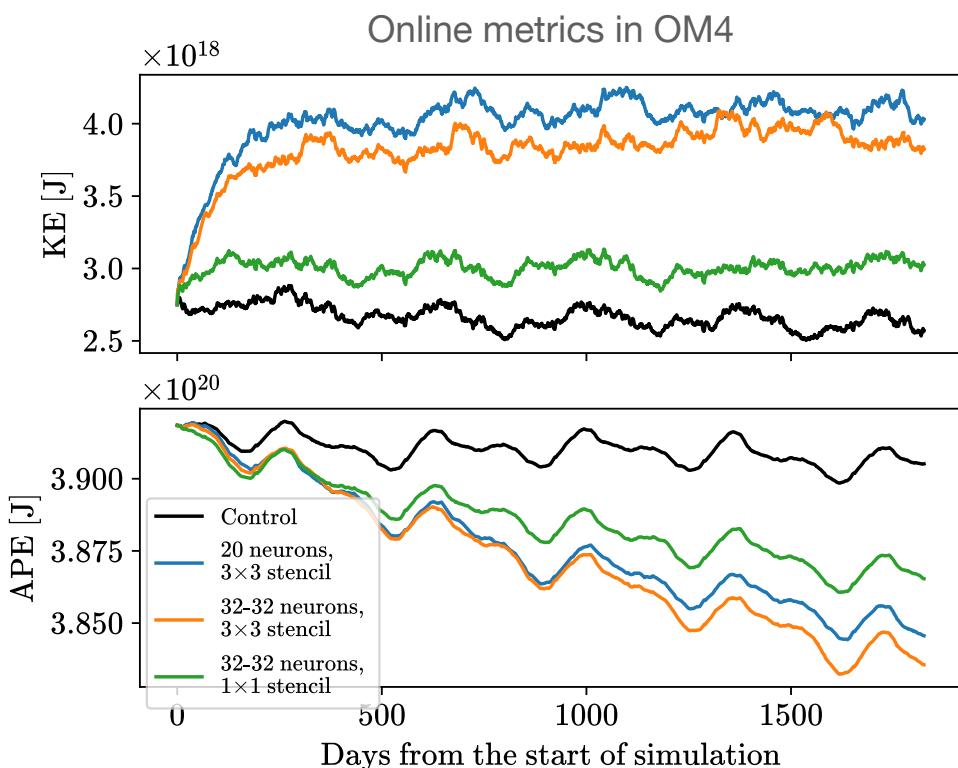
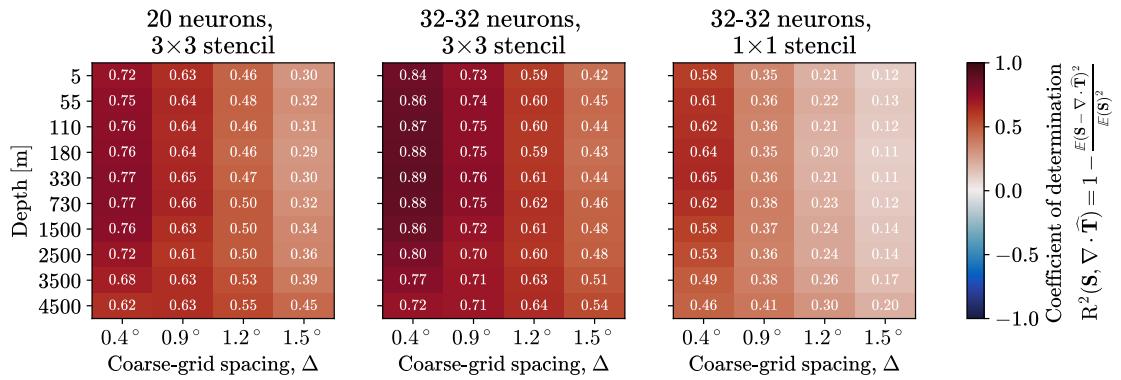


Figure S5. Upper row: Offline performance for the ANN parameterization with dimensional scaling on CM2.6 data. Three versions of the parameterization differ in the number of neurons used or the size of the spatial stencil. In particular, the left column is the ANN used in online simulations, and the center column is the ANN used in offline analysis but retrained using all available data. Lower row: Kinetic energy (KE) and available potential energy (APE) in short 5-year OM4 parameterized simulations for these three versions of the ANN parameterization.

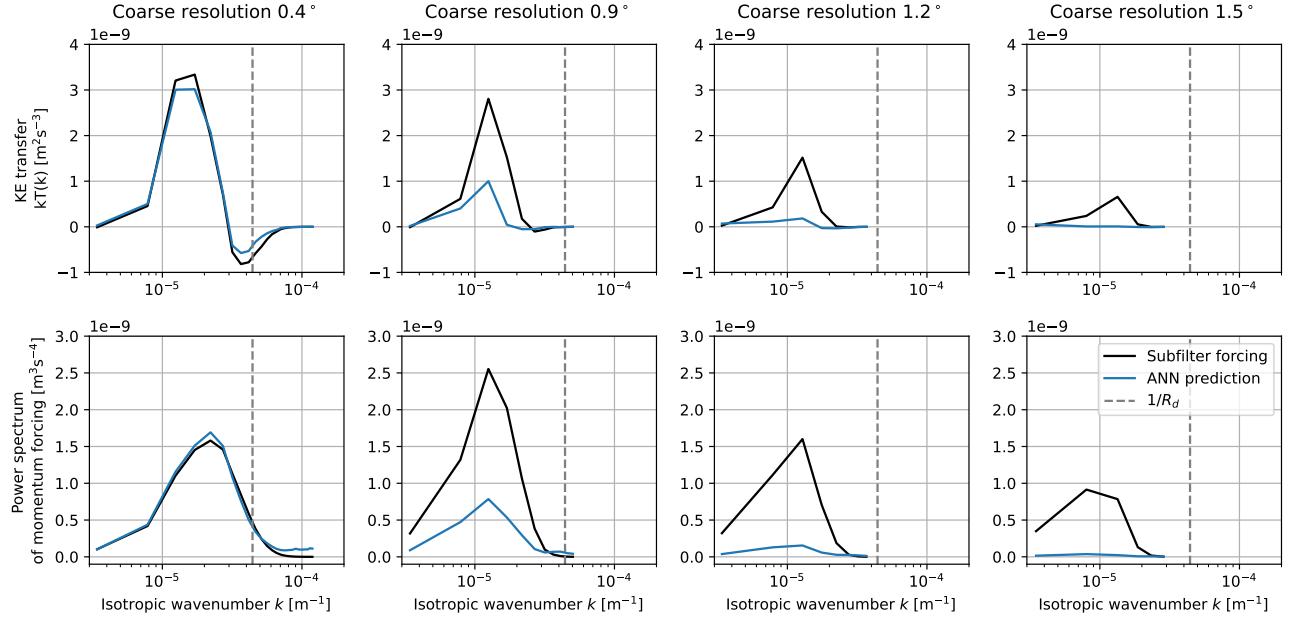


Figure S6. (Upper row) Offline kinetic energy (KE) transfer spectrum, where $T(k) = 2\pi k \text{Re}(\mathcal{F}(\mathbf{u})^* \mathcal{F}(\mathbf{S}))$, and \mathcal{F} is the 2D Fourier transform, Re is the real part, and $*$ is the complex conjugate. (Lower row) power spectrum of subfilter forcing $2\pi k \mathcal{F}(\mathbf{S})^* \mathcal{F}(\mathbf{S})$. Spectra are computed in the North Atlantic region $(25 - 45^\circ\text{N}) \times (60 - 40^\circ\text{W})$ and at depth 5m. $R_d = 22.6\text{km}$ is the Rossby deformation radius in this region. Results are shown for an ANN used in online simulations.

We can identify two effects of the coarsening of the resolution on the diagnosed and predicted eddy fluxes. First, the diagnosed interscale energy transfer vanishes once the Rossby deformation radius becomes unresolved. This can be explained by the blocking of the inverse energy cascade on the scales much larger than the forcing scale (deformation radius). Second, the ANN parameterization predicts even smaller kinetic energy transfer at these coarse resolutions ($\approx 1^\circ$). It is a subject of future studies whether we should attempt to achieve more accurate predictions at these resolutions with improved architecture of the ANN or consider alternative parameterization approaches, such as parameterizing buoyancy fluxes instead, Balwada et al. (2025).

References

- Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., ... others (2019). The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, 11(10), 3167–3211. doi: <https://doi.org/10.1029/2019MS001726>
- Agdestein, S. D., & Sanderse, B. (2025). Discretize first, filter next: Learning divergence-consistent closure models for large-eddy simulation. *Journal of Computational Physics*, 522, 113577. doi: <https://doi.org/10.1016/j.jcp.2024.113577>
- Balwada, D., Perezhigin, P., Adcroft, A., & Zanna, L. (2025). Design and implementation of a data-driven parameterization for mesoscale thickness fluxes. *Authorea Preprints*.
- Chang, C.-Y., Adcroft, A., Zanna, L., Hallberg, R., & Griffies, S. M. (2023). Remote versus local impacts of energy backscatter on the North Atlantic SST biases in a global ocean model. *Geophysical Research Letters*, 50(21), e2023GL105757. doi: <https://doi.org/10.1029/2023GL105757>
- Griffies, S. M., Winton, M., Anderson, W. G., Benson, R., Delworth, T. L., Dufour, C. O., ... others (2015). Impacts on ocean heat from transient mesoscale eddies in a hierarchy of climate models. *Journal of Climate*, 28(3), 952–977. doi: <https://doi.org/10.1175/JCLI-D-14-00353.1>
- Grooms, I., Loose, N., Abernathey, R., Steinberg, J., Bachman, S. D., Marques, G. M., ... Yankovsky, E. (2021). Diffusion-based smoothers for spatial filtering of gridded geophysical data. *Journal of Advances in Modeling Earth Systems*, 13(9), e2021MS002552. doi: <https://doi.org/10.1029/2021MS002552>

- Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, 13(9), e2021MS002534. doi: <https://doi.org/10.1029/2021MS002534>
- Gultekin, C., Subel, A., Zhang, C., Leibovich, M., Perezhigin, P., Adcroft, A., ... Zanna, L. (2024). An analysis of deep learning parameterizations for ocean subgrid eddy forcing. *arXiv preprint arXiv:2411.06604*. doi: <https://doi.org/10.48550/arXiv.2411.06604>
- Leith, C. (1996). Stochastic models of chaotic systems. *Physica D: Nonlinear Phenomena*, 98(2-4), 481–491. doi: [https://doi.org/10.1016/0167-2789\(96\)00107-8](https://doi.org/10.1016/0167-2789(96)00107-8)
- Locarnini, M., Mishonov, A., Baranova, O., Boyer, T., Zweng, M., Garcia, H., ... others (2018). World ocean atlas 2018, volume 1: Temperature. *NOAA Atlas NESDIS*.
- Loose, N., Abernathey, R., Grooms, I., Busecke, J., Guillaumin, A., Yankovsky, E., ... others (2022). GCM-filters: A Python package for diffusion-based spatial filtering of gridded data. *Journal of Open Source Software*, 7(70).
- Marques, G. M., Loose, N., Yankovsky, E., Steinberg, J. M., Chang, C.-Y., Bhamidipati, N., ... others (2022). NeverWorld2: An idealized model hierarchy to investigate ocean mesoscale eddies across resolutions. *Geoscientific Model Development*, 15(17), 6567–6579. doi: <https://doi.org/10.5194/gmd-15-6567-2022>
- Perezhigin, P., Balakrishna, A., & Agrawal, R. (2024). Large eddy simulation of ocean mesoscale eddies. In *Proceedings of the Summer Program 2024, Center for Turbulence Research, Stanford University* (p. 507-516). Retrieved from <https://arxiv.org/abs/2501.05357>
- Perezhigin, P., & Glazunov, A. (2023). Subgrid parameterizations of ocean mesoscale eddies based on Germano decomposition. *Journal of Advances in Modeling Earth Systems*, 15(10).

doi: <https://doi.org/10.1029/2023ms003771>

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. doi: <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. doi: <https://doi.org/10.1016/j.cma.2022.115457>

Sane, A., Reichl, B. G., Adcroft, A., & Zanna, L. (2023). Parameterizing vertical mixing coefficients in the ocean surface boundary layer using neural networks. *Journal of Advances in Modeling Earth Systems*, 15(10), e2023MS003890. doi: <https://doi.org/10.1029/2023MS003890>

Smagorinsky, J. (1963). General circulation experiments with the primitive equations: I. the basic experiment. *Monthly weather review*, 91(3), 99–164. doi: [https://doi.org/10.1175/1520-0493\(1963\)091%3C0099:GCEWTP%3E2.3.CO;2](https://doi.org/10.1175/1520-0493(1963)091%3C0099:GCEWTP%3E2.3.CO;2)

Srinivasan, K., Chekroun, M. D., & McWilliams, J. C. (2024). Turbulence closure with small, local neural networks: Forced two-dimensional and β -plane flows. *Journal of Advances in Modeling Earth Systems*, 16(4), e2023MS003795. doi: <https://doi.org/10.1029/2023MS003795>

Yankovsky, E., Bachman, S., Smith, K. S., & Zanna, L. (2024). Vertical structure and energetic constraints for a backscatter parameterization of ocean mesoscale eddies. *Journal of Advances in Modeling Earth Systems*, 16(7), e2023MS004093. doi: <https://doi.org/10.1029/2023MS004093>

Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17), e2020GL088376. doi: <https://doi.org/10.1029/2020GL088376>

Zhang, C., Perezhigin, P., Adcroft, A., & Zanna, L. (2024). Addressing out-of-sample issues in multi-layer convolutional neural-network parameterization of mesoscale eddies applied near coastlines. *arXiv preprint arXiv:2411.01138*. doi: <https://arxiv.org/abs/2411.01138>