Assessing Data-Driven Eddy-Parameterizations in an Atlantic Sector Model

David Kamm[1], **Pavel Perezhogin**[2], **Julie Deshayes**[1,3], **Etienne Meunier**[1,4], **Alexis Barge**[5]

[1]Sorbonne Université-CNRS-IRD-MNHN, LOCEAN Laboratory, Paris, France
[2]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
[3]Sorbonne Université, Université Paris Cité, CNRS, INRIA, LJLL, EPC ANGE, Paris, France
[4]INRIA, Paris, France
[5]Université Grenoble Alpes, CNRS, IGE, France

Key Points:

- We implement and test two data-driven eddy parameterizations in the NEMO Ocean model
- The equation-based formulation generalizes better to previously unseen model data than the neural network parameterization
- The equation-based parameterization improves the mean state similarly to an existing backscatter scheme

---

Corresponding author: David Kamm, `david.kamm@locean.ipsl.fr`

Abstract

Mesoscale eddies are the dominant reservoir of kinetic energy in the ocean and a key component of the Earth's climate system. Current state-of-the-art climate models can only partially resolve mesoscale processes due to the limited spatial resolution imposed by computational resources, making the accurate parameterization of unresolved kinetic energy transfers crucial. In recent years, a growing number of studies use machine learning tools to build data-driven eddy parameterizations. In this work, we implement two such parameterizations, Zanna and Bolton (2020) (ZB20) and Guillaumin and Zanna (2021) (GZ21), into the NEMO ocean model. We tune and evaluate them against a higher-resolution ground truth, and compare them to a baseline backscatter parameterization in an idealized Atlantic basin configuration. GZ21 is unaware of the model grid, resulting in a systematic bias in the predicted fluxes tied to the local grid spacing. It does not improve the coarse simulation concerning any metrics presented here. ZB20 accurately predicts subgrid fluxes and improves both local kinetic energy spectra and large-scale circulation. It yields an improved mean state comparable to that of the baseline parameterization, consistent with previous studies. We emphasize the importance of carefully designed training data, which takes into account all resolution-dependent model components, such as viscosity and model geometry. Our findings provide guidance for the future development of robust and generalizable data-driven eddy parameterizations.

Plain Language Summary

Ocean eddies are important for transporting energy and heat throughout the ocean. They are typically 10 to 100 km wide, too small to be fully captured by current climate models. In this work, we implement two methods using machine learning to represent the effect of unresolved eddies, Zanna and Bolton (2020) (ZB20) and Guillaumin and Zanna (2021) (GZ21). We test them in a computer model of the Atlantic Ocean and compare their performance to a non-machine learning approach and to a high-resolution reference. GZ21 performs poorly, mainly because it does not take into account the grid of the model. ZB20 performed better, improving the energy patterns and ocean circulation of the model similar to the non-machine learning approach. Our results highlight the importance of considering the model grid and other resolution-dependent model components when training machine learning models to represent ocean processes. Our findings can help to improve future developments to better represent ocean eddies.

1 Introduction

Simulating Earth's climate system is strongly constrained by computational capacity, carbon discipline, or both (Acosta et al., 2024). As a result, climate models most often optimize a trade-off between the model's complexity (the number of processes and climate components explicitly resolved), the spatial resolution of its components (horizontal and vertical), and the temporal duration of integrations (including preliminary experiments for equilibration and calibration of parameters). With the computational resources available today, current climate projections resolve the Oceans with a horizontal grid spacing between 10 km and 100 km (Hewitt et al., 2020). This range spans scales at which mesoscale eddies are not resolved, are partially resolved, or are explicitly resolved. They are the ocean's primary reservoir of kinetic energy (KE), and not accounting for them appropriately leads to biases in the energy cycle (Ferrari & Wunsch, 2009) and ultimately in the mean state (Juricke et al., 2020). At the largest scales, eddies extract available potential energy (APE) from the mean state, which then cascades downscale towards the deformation radius. There, baroclinic instabilities transform it into eddy kinetic energy (EKE). An inverse cascade transfers EKE back towards larger scales, thereby energizing the mean flow (Larichev & Held, 1995).

In very coarse ocean circulation models (approximately 1° horizontal resolution and coarser), even the largest eddies are unresolved, leading to excessive APE, nearly laminar flow, and erroneous tracer transports. Gent and Mcwilliams (1990) proposed a parameterization for tracer transport driven by baroclinic instabilities, extracting APE from the large-scale flow. As computational resources become more affordable, many global ocean and climate models enter the so-called "gray zone" of eddy-permitting horizontal resolution. Depending on the local stratification, latitude, and topography, different regions within such models resolve a different range of the eddy spectrum (Hallberg, 2013). The unresolved range still needs to be parameterized. At eddy-permitting resolution, it is the upscale transfer of EKE from the subgrid scales that is particularly missing, leading to an energetically too weak eddy field and mean circulation. Jansen and Held (2014) addressed this issue by introducing a prognostic equation for the subgrid KE, which informs a negative viscosity operator that transfers KE upscale. This kinetic energy backscatter approach was further developed (Jansen et al., 2015; Klöwer et al., 2018; Juricke et al., 2019; Jansen et al., 2019), yet it remains subject to ongoing research.

A growing number of studies use machine learning (hereafter ML) tools to build data-driven eddy parameterizations (Zanna & Bolton, 2020; Frezat et al., 2021; Guillaumin & Zanna, 2021; Guan et al., 2022; Balwada et al., 2025), some of which have already been implemented in state-of-the-art ocean models (Zhang et al., 2023; Perezhogin et al., 2024). The common approach is to use data from simulations with a high enough horizontal resolution to explicitly resolve eddies as "ground truth". The data are filtered and coarse-grained to diagnose the subgrid fluxes that are missing at coarse resolution. ML models are then trained to predict these fluxes from the coarse-resolution data and tested on unseen fields to evaluate their performance. A known issue with data-driven eddy parameterizations is that they generalize poorly to new flow regimes (Ross et al., 2023; Gultekin et al., 2024), highlighting the importance of an appropriate choice of training data (Yan et al., 2024). As with physics-based parameterizations, they need to be calibrated to the specific context in which they are employed. This can be challenging, especially for black-box parameterizations such as neural networks (hereafter NN), which have a vast number of degrees of freedom that are difficult to interpret in physical terms.

In this work, we implement two data-driven eddy closures, which target KE backscatter in the eddy-permitting regime: Zanna and Bolton (2020) (hereafter ZB20) and Guillaumin and Zanna (2021) (hereafter GZ21). They take different approaches to the problem. ZB20 trains a relevance vector machine algorithm to learn equations that best describe the subgrid fluxes from the coarse-grained velocity fields. GZ21 directly predicts the subgrid fluxes and their standard deviation using a NN, assuming a Gaussian probability distribution for the predicted fluxes. The scientific objective of this paper is to compare the two approaches with respect to local energy transfers and their effect on the large-scale ocean circulation. As a first step, we introduce the ocean model and the parameterizations to be tested (Section 2). We then explain the calibration of parameterizations for the configuration chosen in this study (section 3) prior to stating the results of the comparison for both short- and long-term metrics (section 4). Finally, we discuss all results (section 5) and conclude (section 6).

2 Ocean Model and Subgrid Parameterizations

We use the Nucleus for European Modelling of the Ocean (NEMO; Madec et al., 2023), a well-established modeling framework written in Fortran to simulate ocean dynamics. The DINO configuration (Diabatic Neverworld Ocean; Kamm et al., 2025) was developed for NEMO to serve as a testbed for mesoscale eddy parameterizations. It is of intermediate complexity, featuring an idealized Atlantic basin geometry, surface forcing, and a zonally periodic channel representing the Atlantic section of the Southern Ocean. As the name suggests, DINO extends the Neverworld2 configuration (Marques et al., 2022) by including diabatic processes and an idealized equation of state for temperature and salinity. This allows for the evaluation of eddy parameterizations with respect to their role in setting a cross-hemispheric circulation and stratification, in addition to their interaction with the ocean's energy cycle. It is thus realistic enough to simulate key metrics of the climate system, such as the Antarctic Circumpolar Current (ACC), Meridional Overturning Circulation (MOC), and meridional heat transport, but computationally inexpensive enough to serve as a hierarchy across a large range of model resolutions.

The eddy parameterizations implemented and tested in this study target the eddy-permitting regime, where the eddy field is partially resolved. As an example of this regime, we use DINO at 1/4°. The unparameterized reference run at this resolution is denoted as R4. At 1/16° horizontal resolution (denoted R16), we assume the model to be eddy-resolving and treat its solution as ground truth. For details about the configuration and the experimental design of R4 and R16, we refer to Kamm et al. (2025). All experiments are initialized after a 3000 year spin-up at 1° horizontal resolution to adjust the stratification to the surface forcing. After interpolation to the respective grids, we run R16 for 30 years and R4 for 600 years. We use years 20–30 of R16 as a higher-resolution reference for assessing short-term metrics, and years 20–30 of R4 for the corresponding unparameterized comparisons. Long-term metrics are assessed using years 550–600 of R4. In this section, we introduce the governing equations and the closure problem of the discretized model, along with the proposed parameterizations to address it.

2.1 Momentum equations, coarse-graining and the closure problem

We consider the momentum balance for the zonal (u) and meridional (v) components of the velocity vector field $\mathbf{u} = \mathbf{u}_h + \mathbf{k}\, w$ for an incompressible fluid on a rotating sphere under the Boussinesq and hydrostatic approximation. The subscript $[\dots]_h$ denotes the vector component in the horizontal plane, $\mathbf{k}$ is the unit vector in the vertical direction, and w is the amplitude of velocity in the vertical direction. These equations are discretized on grids with different horizontal resolutions for the experiments presented in this study. To account for this in the notation, we follow Guillaumin and Zanna (2021), denoting fields discretized at coarse and fine resolutions with the superscripts $\downarrow$ and $\uparrow$, respectively. Hence, the momentum equation at coarse resolution (R4) reads

$$\partial_t\, \mathbf{u}_h^{\downarrow} + \left[\left(\nabla \times \mathbf{u}^{\downarrow} \right) \times \mathbf{u}^{\downarrow} + \frac{1}{2} \nabla\, {\mathbf{u}^{\downarrow}}^2 \right]_h + f\, \mathbf{k} \times \mathbf{u}_h^{\downarrow} + \frac{1}{\rho_0}\, \nabla_h p^{\downarrow} = \mathbf{F}^{\mathbf{u}^{\downarrow}} + \mathbf{D}^{\mathbf{u}^{\downarrow}}, \qquad (1)$$

where p is the pressure, f is the Coriolis frequency, ρ_0 is a reference density, $\mathbf{F}^{\mathbf{u}}$ represents wind forcing at the surface ($z = 0$) and friction at the bottom ($z = D$). The viscous terms are collected in $\mathbf{D}^{\mathbf{u}}$ and will be discussed in detail later. Similarly, the momentum equation at high resolution (R16) reads

$$\partial_t\, \mathbf{u}_h^{\uparrow} + \left[\left(\nabla \times \mathbf{u}^{\uparrow} \right) \times \mathbf{u}^{\uparrow} + \frac{1}{2} \nabla\, {\mathbf{u}^{\uparrow}}^2 \right]_h + f\, \mathbf{k} \times \mathbf{u}_h^{\uparrow} + \frac{1}{\rho_0}\, \nabla_h p^{\uparrow} = \mathbf{F}^{\mathbf{u}^{\uparrow}} + \mathbf{D}^{\mathbf{u}^{\uparrow}}, \qquad (2)$$

We consider Equation 2 to describe the time evolution of the velocity field $\mathbf{u}^{\uparrow}$ at sufficiently high horizontal resolution to explicitly resolve mesoscale eddies. We treat it as our target ground truth and neglect processes that are not resolved at R16, namely the submesoscale.

To define the closure problem, we introduce a coarse-graining operator $\overline{(\cdot)}$, which maps fine-grid variables from R16 onto the coarser grid of R4. The coarse-grained R16 variables are denoted C16 in the following. We perform the coarse-graining by first applying a Gaussian spatial filter to suppress small-scale variability, followed by a spatial average over the coarse grid cell. To a first approximation, such an operator commutes with spatial and temporal derivatives and acts only in the lateral direction. However, Aluie (2019) demonstrates that this assumption is not strictly valid for vector fields discretized on a sphere. For simplicity, we adopt it nonetheless. By applying the coarse-graining operator to each term of Equation 2, we can formulate the coarse-grained momentum equation as

$$\partial_t \overline{\mathbf{u}^\uparrow}_h + \left[\left(\nabla \times \overline{\mathbf{u}^\uparrow}\right) \times \overline{\mathbf{u}^\uparrow} + \frac{1}{2}\nabla \overline{\mathbf{u}^\uparrow}^2\right]_h + f\,\mathbf{k} \times \overline{\mathbf{u}^\uparrow}_h + \frac{1}{\rho_0}\,\nabla_h \overline{p^\uparrow} = \mathbf{F}^{\overline{\mathbf{u}^\uparrow}} + \mathbf{D}^{\overline{\mathbf{u}^\uparrow}} + \mathcal{S} - \mathcal{D}. \quad (3)$$

It takes essentially the same form as Equation 1, with additional subgrid eddy momentum fluxes $\mathcal{S}$ and $\mathcal{D}$ arising from the nonlinearity of the advection and viscosity terms. These terms close the coarse-grained momentum equation with respect to the high-resolution solution and are given by

$$\mathcal{S} = \left[\left(\nabla \times \overline{\mathbf{u}^\uparrow}\right) \times \overline{\mathbf{u}^\uparrow} + \frac{1}{2}\nabla \overline{\mathbf{u}^\uparrow}^2\right]_h - \overline{\left[\left(\nabla \times \mathbf{u}^\uparrow\right) \times \mathbf{u}^\uparrow + \frac{1}{2}\nabla\, {\mathbf{u}^\uparrow}^2\right]_h} \quad (4)$$

$$\mathcal{D} = \mathbf{D}^{\overline{\mathbf{u}^\uparrow}} - \overline{\mathbf{D}^{\mathbf{u}^\uparrow}}. \quad (5)$$

They reflect the momentum fluxes across the filter scale of the coarse-graining operator. The idea of subgrid closures is that by including them in the coarse model, its solution $\mathbf{u}^\downarrow$ approaches the coarse-grained ground truth $\overline{\mathbf{u}^\uparrow}$. The closure problem is that the fluxes depend on the unknown fine-scale velocity field $\mathbf{u}^\uparrow$ and need to be parameterized for coarse-resolution model simulations.

The viscosity terms collected in $\mathbf{D}^{\mathbf{u}}$ are themselves parameterized. They are devoted to the dissipation of KE cascading toward the grid scale. This is not strictly motivated by physical principles but rather serves to ensure the numerical stability of the model (Madec et al., 2023). For vertical viscosity, we use the turbulent kinetic energy closure (Gaspar et al., 1990), denoted as $\mathbf{D}^{\mathbf{u}}_{\text{TKE}}$. Since it does not explicitly depend on the horizontal resolution of the grid, the associated subgrid fluxes are close to zero, i.e., $\mathbf{D}^{\overline{\mathbf{u}^\uparrow}}_{\text{TKE}} \approx \overline{\mathbf{D}^{\mathbf{u}^\uparrow}_{\text{TKE}}}$. We chose the *biharmonic Smagorinsky* model for horizontal viscosity (Smagorinsky, 1963; Griffies & Hallberg, 2000), denoted as $\mathbf{D}^{\mathbf{u}}_{\text{SMA}}$, with the default settings provided in NEMO (see Madec et al., 2023, for details). It is both grid- and flow-dependent and highly scale-selective, dissipating only near the smallest resolved scales. Consequently, coarse-graining the Smagorinsky term removes most of its associated dissipation, leaving $\mathbf{D}^{\overline{\mathbf{u}^\uparrow}}_{\text{SMA}} \gg \overline{\mathbf{D}^{\mathbf{u}^\uparrow}_{\text{SMA}}}$ and thus $\mathcal{D} \approx \mathbf{D}^{\overline{\mathbf{u}^\uparrow}}_{\text{SMA}}$. Substituting into the right-hand side of Equation 3 leaves us with

$$\partial_t \overline{\mathbf{u}^\uparrow}_h + ... = \mathbf{F}^{\overline{\mathbf{u}^\uparrow}} + \mathbf{D}^{\overline{\mathbf{u}^\uparrow}}_{\text{TKE}} + \mathcal{S}, \quad (6)$$

rendering Equation 3 horizontally inviscid. This implies that our target fields $\overline{\mathbf{u}^\uparrow_h}$ are not a stable solution of the coarse-discretized model (Equation 1). For the sake of numerical stability, we therefore accept the excessive dissipation introduced by the viscous term and set $\mathcal{D} = 0$ in a-posteriori simulation rollouts. Introducing a subgrid parameterization $\tilde{\mathcal{S}}$ in the coarse-discretized model (Equation 1) to account for the fluxes of KE from unresolved processes then reads

$$\partial_t \mathbf{u}^\downarrow_h + ... = \mathbf{F}^{\mathbf{u}^\downarrow} + \mathbf{D}^{\mathbf{u}^\downarrow}_{\text{TKE}} + \mathbf{D}^{\mathbf{u}^\downarrow}_{\text{SMA}} + \tilde{\mathcal{S}}, \quad (7)$$

When comparing Equation 6 and Equation 7, it becomes clear that to assess parameterizations with respect to the coarse-grained ground truth, the combined parameterized subgrid fluxes of $\mathbf{D}^{\mathbf{u}^\downarrow}_{\text{SMA}}$ and $\tilde{\mathcal{S}}$ need to be considered to approximate the diagnosed subgrid fluxes $\mathcal{S}$. The KE transfers due to the diagnosed fluxes on the coarse-grained velocities are then given by

$$\mathcal{T} = \overline{\mathbf{u}^\uparrow} \cdot \mathcal{S} \quad (8)$$

while those due to the parameterized fluxes on the coarse velocities are given by

$$\tilde{\mathcal{T}} = \mathbf{u}^\downarrow \cdot \left[\mathbf{D}^{\mathbf{u}^\downarrow}_{\text{SMA}} + \tilde{\mathcal{S}}\right] \quad (9)$$

In the eddy-permitting regime, $\mathcal{T}$ typically represents the net upscale kinetic energy transfer from unresolved subgrid eddies to the resolved eddy field. However, the transfer can be either upscale or downscale locally, depending on the instantaneous velocity field. The goal of eddy parameterizations in this regime is to predict these fluxes accurately from coarse-resolution data. The data-driven approaches tested here aim to address this by learning to infer the fluxes using diagnosed fields from high-resolution simulations as training data. In the following sections, we introduce the methodology and implementation, alongside a physics-based kinetic energy backscatter parameterization for reference.

2.2 Kinetic energy backscatter parameterization (KEB20)

We employ the kinetic energy backscatter parameterization as implemented and tested in NEMO in Perezhogin (2020). It aims to represent the inverse cascade of KE from the subgrid-scale, which is missing at eddy-permitting resolution. The kinetic energy dissipated by the biharmonic viscosity term is diagnosed and re-injected through a negative Laplacian viscosity operator.

$$\tilde{\mathcal{S}}_{KEB} = \nabla_h \left(\nu \, \nabla_h \, \mathbf{u}_h^{\downarrow} \right) \tag{10}$$

The backscatter viscosity coefficient $\nu(t, z, y, x) \leq 0$ is computed locally from the subgrid kinetic energy $e(t, z, y, x)$ as

$$\nu = -c_{\text{back}} \Delta^{\downarrow} \sqrt{\max(e, 0)}, \tag{11}$$

where $c_{\text{back}} = 0.4\sqrt{2}$, following Perezhogin (2020), and $\Delta^{\downarrow}$ denotes the respective grid spacing at coarse resolution. The subgrid kinetic energy follows from an additional prognostic equation, as first proposed in Jansen and Held (2014) and further developed in Jansen et al. (2015), Jansen et al. (2019), and Juricke et al. (2019). Here, it is defined as

$$\partial_t \, e = \gamma_{\text{KEB}} \, \dot{E}_{\text{diss}} + \dot{E}_{\text{back}} + \nu_e \, \nabla_h^2 \, e, \tag{12}$$

where $\dot{E}_{\text{diss}} \geq 0$ is the rate of energy dissipation by the biharmonic Smagorinsky operator, $\dot{E}_{\text{back}} \leq 0$ is the rate of energy backscatter by the negative Laplacian viscosity operator Equation 10, and the last term corresponds to the diffusion of subgrid kinetic energy with a diffusivity $\nu_e = 300 \, \text{m}^2 \, \text{s}^{-1}$. $\gamma_{KEB} \approx 1$ is a non-dimensional coefficient that controls the strength of the backscatter parameterization (Klöwer et al., 2018; Juricke et al., 2019).

2.3 Zanna-Bolton Parameterization (ZB20)

Zanna and Bolton (2020) write the subgrid eddy fluxes as the divergence of the stress tensor $\mathbf{T}$

$$\mathcal{S} = \nabla_h \cdot \mathbf{T}. \tag{13}$$

The authors used an equation discovery algorithm to approximate the components of the stress tensor in terms of the coarse velocity fields. As basis functions, they chose relative vorticity $\zeta = \partial_x v^{\downarrow} - \partial_y u^{\downarrow}$, horizontal divergence $\delta = \partial_x u^{\downarrow} + \partial_y v^{\downarrow}$, shearing deformation $\varepsilon_s = \partial_y u^{\downarrow} + \partial_x v^{\downarrow}$, and stretching deformation $\varepsilon_n = \partial_x u^{\downarrow} - \partial_y v^{\downarrow}$. The diagnosed subgrid fluxes $\mathcal{S}$ from the same coarse-graining approach as presented here served as the ground truth. The authors used high-resolution data generated from an idealized double-gyre configuration as training data, similar to the configuration used in Cooper and Zanna (2015). They arrived at the following equation to approximate the stress tensor:

$$\tilde{\mathbf{T}} = \kappa \left[\begin{pmatrix} -\zeta \, \varepsilon_s & \zeta \, \varepsilon_n \\ \zeta \, \varepsilon_n & \zeta \, \varepsilon_s \end{pmatrix} + \frac{1}{2} \left(\zeta^2 + \varepsilon_s^2 + \varepsilon_n^2 \right) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] \tag{14}$$

The discovered equation resembles a special case of the nonlinear gradient model, derived in earlier LES studies through a Taylor-series expansion of the subgrid fluxes (e.g., Leonard, 1975; Meneveau & Katz, 2000). We refer to Jakhar et al. (2024) for a detailed discussion on the similarities between the equation discovery and Taylor-series methods. When tested *a posteriori* in an idealized eddy-permitting ocean model, Perezhogin et al. (2024) found that the ZB20 parameterization underestimates backscatter and addressed this issue by applying spatial filtering schemes to suppress noise close to the grid scale and enhance the KE backscatter towards larger scales. The ZB20 parameterization evaluated here corresponds to "ZB-smooth" in their paper, where a Gaussian-weighted spatial low-pass filter is convoluted N times with each component of the stress tensor:

$$\tilde{\mathcal{S}}_{ZB} = \nabla_h \cdot \left(\mathcal{G}^N * \tilde{\mathbf{T}}\right), \tag{15}$$

where $\mathcal{G}$ is defined on a 3×3 horizontal stencil, given by

$$\mathcal{G} = \frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}. \tag{16}$$

They found $N = 4$ to be optimal in the eddy-permitting regime; therefore, we will adopt this value. In Zanna and Bolton (2020), the value for κ is approximated as the averaged regression coefficients from the equation discovery algorithm. Following Perezhogin et al. (2024) and previous nonlinear gradient model studies (Meneveau & Katz, 2000; Chen et al., 2003), we associate the coefficient with the local grid scale $\Delta^{\downarrow}$:

$$\kappa = -\gamma_{ZB} {\Delta^{\downarrow}}^2 \tag{17}$$

This leaves us with a non-dimensional tuning coefficient $\gamma_{ZB} \geq 0$ to control the magnitude of the momentum fluxes, similar to γ_{KEB} in the previous section.

2.4 Guillaumin-Zanna parameterization (GZ21)

Guillaumin and Zanna (2021) takes a stochastic approach to predict both the mean and standard deviation of a Gaussian distributed subgrid momentum forcing. GZ21 is a *convolutional neural network* (CNN) trained on high-resolution surface data from the CM2.6 coupled climate model (Delworth et al., 2012). The coarse-graining procedure for diagnosing the ground-truth subgrid fluxes is similar to the one presented here, but it maps from 0.1° to 0.4° horizontal resolution, targeting slightly coarser models than the test configuration used in this study. The CNN uses the coarse-resolution horizontal velocity field as input to infer the mean and standard deviation of a Gaussian distribution representing the instantaneous subgrid fluxes. The subgrid forcing entering the right-hand side of the momentum equations is then sampled as

$$\tilde{\mathcal{S}}_{GZ} = \gamma_{GZ} \left[\mathbf{CNN}^{(\mathrm{mean})}(\mathbf{u}_h^{\downarrow}) + \mathcal{W}\,\mathbf{CNN}^{(\mathrm{std})}(\mathbf{u}_h^{\downarrow})\right], \tag{18}$$

where $\mathcal{W}$ is a spatial random field sampled from uncorrelated standard normal distributions (i.e., white noise), and $\gamma_{GZ} \geq 0$ is again a non-dimensional tuning coefficient. The calibration procedure for the tuning coefficients is described in Section 3.

In the original publication, the parameterization is tested a posteriori in a barotropic double gyre configuration (Klöwer et al., 2018). In a later study (Zhang et al., 2023), the parameterization was introduced into the Modular Ocean Model version 6 (MOM6 Adcroft et al., 2019) and tested in a baroclinic two-layer model. They found that the model lacks generalization across different depths and horizontal resolutions, as it was trained only on surface data and at a single coarse resolution.

The CNN consists of eight convolutional layers, with approximately a quarter of a million weights learned during its training phase using the Python package PyTorch. To couple the CNN from Python to the NEMO model, we employ the *EOPHIS* package (Barge, Le Sommer, & IGE-MEOM, 2024; Barge, Meunier, & Kamm, 2024). It enables the integration of arbitrary Python code into NEMO via its native coupler, OASIS (Craig et al., 2017). At each time step, the velocity fields are sent to parallel Python processes, where the subgrid forcing is inferred according to Equation 18, sent back to the NEMO processes, and finally applied to update the right-hand side of the time-stepping scheme.

3 Tuning strategy

The parameterizations presented above include parameters that remain to be chosen (namely γ_{KEB}, γ_{ZB}, and γ_{GZ}), a phase referred to as ”tuning” in climate modeling (Hourdin et al., 2017). They were developed for a specific ocean model and tested in a specific configuration. The chosen parameters need to be revisited, as we are using another ocean model in yet another configuration. We illustrate this by highlighting the flaws in using the initially chosen parameters for each parameterization after implementation, as part of an initial evaluation of the domain-integrated KE (see Figure 1), prior to explaining our tuning approach.

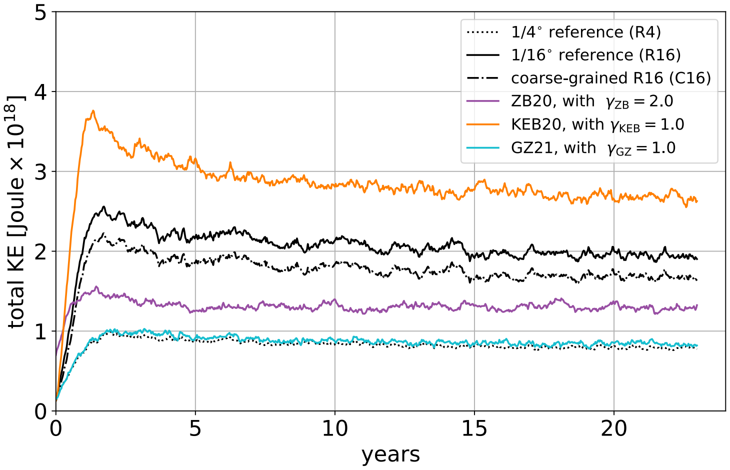

Figure 1. Kinetic energy integrated over the entire domain for R4 (black dotted), R16 (black solid), C16 (black dash-dotted) and the untuned ZB20 (purple), GZ21 (blue) and KEB20 (orange) experiment.

As shown by Kamm et al. (2025), the subgrid fluxes resolved in C16 lead to an upscale transfer of KE from the subgrid to resolved scales. This results in more than twice the total KE when compared to R4: a target we want to reach by parameterizing these fluxes. Using $\gamma_{\mathrm{KEB}} = 1.0$, KEB20 overestimates the integrated backscatter by almost a factor of two, while GZ21 with $\gamma_{\mathrm{GZ}} = 1.0$ barely energizes the flow at all (Figure 1). ZB20 with $\gamma_{\mathrm{ZB}} = 2.0$ is close to the target, which can be explained by the fact that ZB20, as implemented here following Perezhogin et al. (2024), was tuned in NW2, a very similar configuration to DINO. For a meaningful comparison of the parameterizations, it is essential to calibrate them in a common context and with respect to common evaluation metrics.

The most straightforward approach is to directly compare the diagnosed subgrid momentum fluxes $\mathcal{S}$ (Equation 4) to the instantaneously predicted fluxes from each parameterization, namely $\tilde{\mathcal{S}}_{KEB}$ (Equation 10), $\tilde{\mathcal{S}}_{ZB}$ (Equation 15), and $\tilde{\mathcal{S}}_{GZ}$ (Equation 18). This a priori approach is intuitive for data-driven parameterizations, as they were trained on instantaneous fluxes similarly. However, it also comes with similar challenges: it implicitly assumes that the coarse-grained high-resolution solution of Equation 6, $\overline{\mathbf{u}^{\uparrow}}$, behaves as the coarse-discretized low-resolution fields $\mathbf{u}^{\downarrow}$ of Equation 1. As discussed in

Section 2.1, this is not strictly the case, partly due to the additional viscous dissipation from the Smagorinsky term required at coarse resolution.

Instead, we tune each parameterization a posteriori through repeated simulation rollouts. We use the domain-integrated KE to optimize the scalar coefficients γ_{KEB}, γ_{ZB}, and γ_{KEB} for a minimal root-mean-square error (RMSE) over the last four years of 23-year-long tuning rollouts. We found the optimal values with respect to domain-integrated KE to be $\gamma_{\text{KEB}} = 0.87$, $\gamma_{\text{ZB}} = 3.0$, and $\gamma_{\text{GZ}} = 6.0$. These experiments were then continued for a total simulation length of 30 years for comparison (see Figure 2).

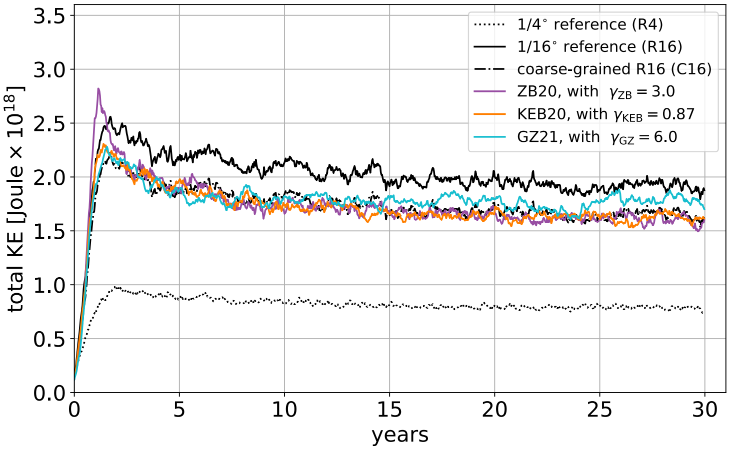

Figure 2. Kinetic energy integrated over the entire domain for R4 (black dotted), R16 (black solid), C16 (black dash-dotted) and the tuned ZB20 (purple), GZ21 (blue) and KEB20 (orange) experiment.

4 Results

Since the domain-integrated KE was used for tuning, it cannot serve as an evaluation metric for comparison. Additionally, it does not reveal whether the flow is energized in the correct location and at the appropriate scales. In the following, we focus on the spatial and spectral patterns of kinetic energy and its associated transfers. Then, we compare their impact on long-term metrics of the ocean circulation, which are crucial for the climate system.

4.1 Short-term metrics

We decompose the total domain-integrated KE into meridional and vertical profiles (panels **(a)** and **(b)** of Figure 3, respectively). The high-resolution reference experiment (R16, black solid line) exhibits consistently higher KE across all latitudes and depths compared to the low-resolution reference experiment (R4, black dotted line). These differences are most pronounced in regions where the flow is most energetic, namely, vertically towards the surface (panel **(b)**) and meridionally in the Southern Ocean, the Equatorial region, and close to the Gulf Stream separation (panel **(a)**). After coarse-graining the high-resolution reference experiment, most of this additional kinetic energy is retained (C16, black dashed line). This demonstrates how energy is not only introduced at the expanded spectrum of smaller, resolved scales but also transferred to larger scales captured at coarse resolution throughout the domain.

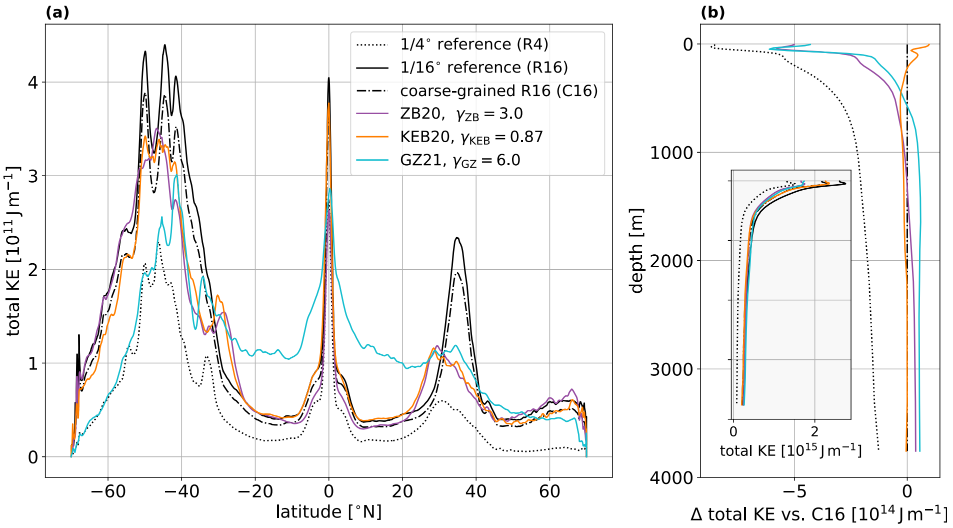

Figure 3. Kinetic energy integrated over depth and longitude (panel **(a)**) and latitude and longitude (panel **(b)**). The profiles correspond to averages over the last ten years of the R4 (black dotted), R16 (black solid), C16 (black dash-dotted), and the ZB20 (purple), GZ21 (blue), and KEB20 (orange) experiments. Panel **(b)** depicts the respective difference to the C16 ground truth for better visibility, while the inlet shows the overall shape of the kinetic energy profiles, using the same color coding as panel (a).

The meridional KE profiles of KEB20 (orange) and ZB20 (purple) are broadly similar. They perform well in the Southern Ocean region, with slightly less KE in the northern part of the periodic channel. ZB20 barely improves the integrated KE on the equator, while KEB20 closely follows the target. Both parameterizations succeed in representing KE in the tropical bands but fail to capture the eddy-rich Gulf Stream separation region. They inject too little KE and exhibit a southward shift of the maximum KE, as does the unparameterized reference experiment (R4). GZ21 (blue) injects too much KE at low latitudes and too little at high latitudes. This is likely due to the non-uniform grid size of DINO, which is roughly one-third the size at the poles compared to that at the equator. Even though GZ21 was trained on non-uniform data, it is agnostic to the local grid size and cannot infer spatial derivatives of the flow field (see Equation 18). This presumably leads to poor generalization across different resolutions and dynamical regimes.

The depth profiles of domain integrated KE reveal a surface-intensified flow for all experiments (inlet in panel **(b)** of Figure 3). Again, we find that overall, R16 is more energetic than R4 and that, after coarse-graining, most of the additional KE is retained. KEB20 reproduces the depth profile of the target ground truth well, with slightly higher KE close to the surface. Both ZB20 and GZ21 underestimate KE close to the surface and overestimate it at depth. This bias is more pronounced for GZ21, which was trained only on surface data.

The results presented so far reveal only the net effect of resolved and parameterized subgrid fluxes on mean KE: an upscale flux of momentum, leading to a net increase in KE across the domain. Snapshots of surface KE tendency due to the subgrid terms show that the associated KE transfers across the coarse grid scale can be locally upscale or downscale, fluctuating across a wide range of spatial scales (see Figure 4).

As already suggested by the meridional profiles of mean KE (panel **(a)** of Figure 3), the most intense instantaneous fluxes are indeed found in the Southern Ocean, around

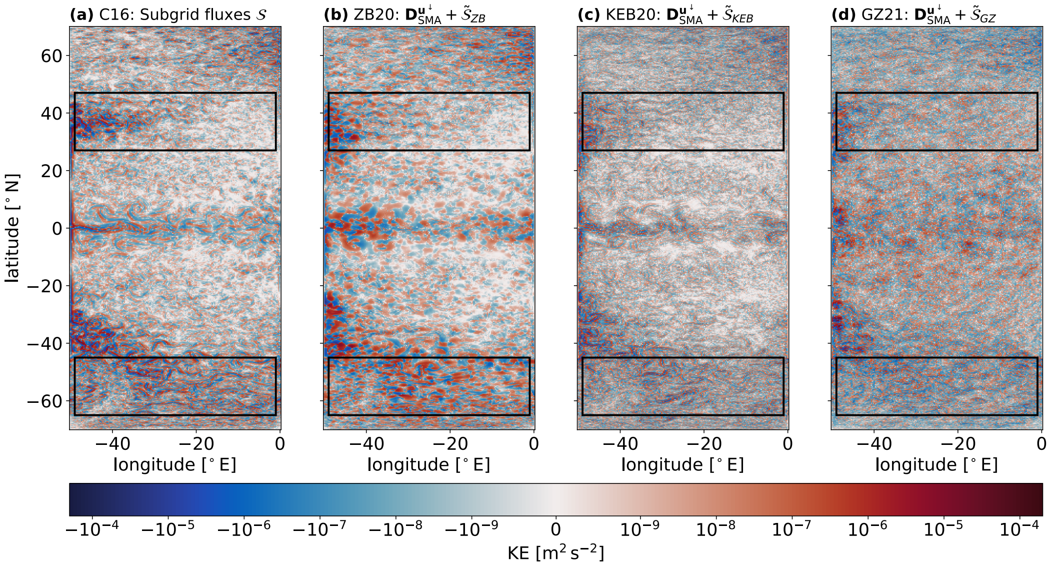

Figure 4. Snapshots of surface kinetic energy tendency due to **(a)** the diagnosed subgrid fluxes $\mathcal{S}$ and the combined parameterized fluxes $\mathbf{D}^{\mathbf{u}^{\downarrow}}_{\mathrm{SMA}} + \tilde{\mathcal{S}}$ of **(b)** ZB20, **(c)** KEB20 and **(d)** GZ21, as defined in Equation 8 and Equation 9. Black boxes indicate the Southern Ocean and the Gulf Stream region.

the Equator, and downstream of the Gulf Stream separation for all experiments. In the case of GZ21, we also find a clear latitudinal pattern, with negative KE fluxes dominating at high latitudes, while positive KE fluxes dominate at low latitudes. A direct comparison of the instantaneous fields is not straightforward. Instead, we analyze the signal after projecting it into spectral space. We focus on surface fields in the two regions indicated by the black boxes in Figure 4: the Southern Ocean (SO) and the Gulf Stream region (GS). The fields are first interpolated to equidistant Cartesian coordinates using the *xemf* Python package (Zhuang et al., 2023). Then, we use the *xrft* Python package (Uchida et al., 2023) to compute the isotropic power density spectrum of KE and the isotropic transfer spectrum from the diagnosed and parameterized KE fluxes (see Figure 5).

The coarse R4 experiment has less KE than the high-resolution R16 experiment across all scales for both the SO and GS regions (panels **(a)** and **(b)** of Figure 5; note that, overall, the GS region is approximately one order of magnitude less energetic than the SO). The coarse-graining operator only affects higher wavenumbers, drastically reducing the KE towards the coarse grid scale. At larger scales, the increased KE is retained. This increased KE in the coarse-grained high-resolution simulation, compared to the low-resolution simulation, demonstrates the inverse cascade, which transfers KE from the subfilter scale all the way to the largest scales. Turbulence theory predicts a $k^{-\frac{5}{3}}$ slope for the KE spectrum, where energy cascades from small to large scales (Kraichnan, 1967) (see the gray dashed line in Figure 5). We find that the high-resolution spectrum follows the predicted slope between scales of $40-200$ km in the SO and $50-120$ km in the GS region. We diagnose the transfer spectrum with isotropic cross-spectra between the coarse-grained velocity field $\overline{\mathbf{u}_h^{\uparrow}}$ and the subgrid fluxes $\mathcal{S}$ (see panels **(c)** and **(d)** of Figure 5). On average, the subgrid fluxes extract KE at scales smaller than approximately 70 km in the SO and 80 km in the GS region, while injecting KE at larger scales. This aligns with the range of scales predicted by the $k^{-\frac{5}{3}}$ power law fitted to the R16 spectrum. Consistent with the power density spectra, the transfers in the GS region are ap-

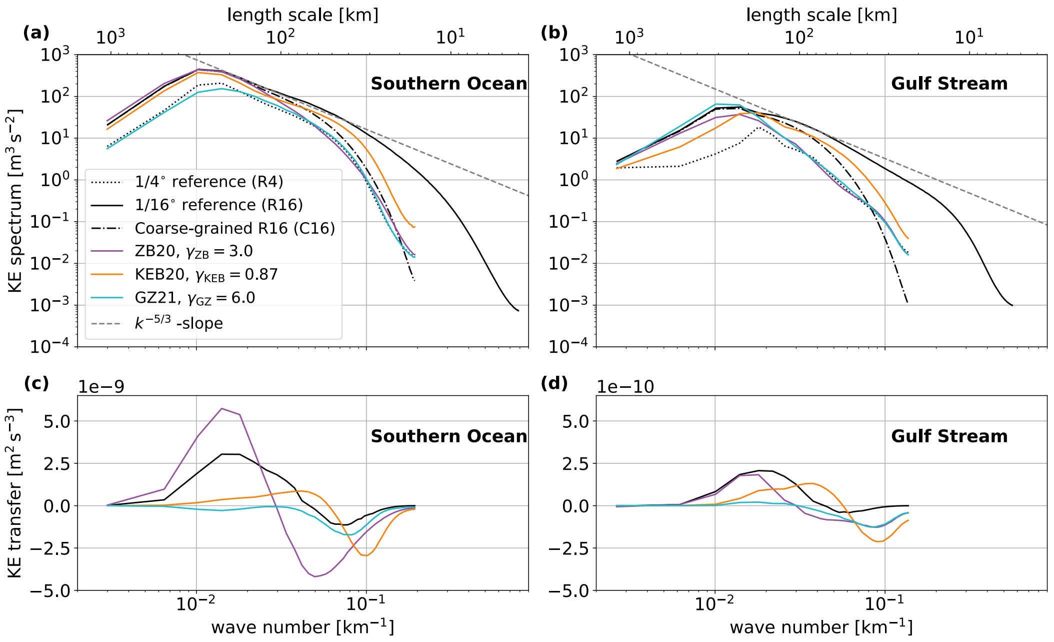

Figure 5. The two upper panels show the power spectrum of eddy-kinetic-energy for all experiments in the **(a)** Southern Ocean and the **(b)** Gulf-Stream region. The gray dashed line represents a $k^{-\frac{5}{3}}$ power law. Panels **(c)** and **(d)** show the corresponding kinetic energy transfers due to the diagnosed (black solid line) and parameterized subgrid fluxes. They include the dissipative fluxes of the biharmonic Smagorinsky viscosity. All spectra are derived from velocity fields within the black boxes in Figure 4, interpolated onto an equidistant Cartesian grid. We use the Python library *xrft* (Uchida et al., 2023) and apply linear detrending along with a Hann window for smoothing. The spectra are averaged over the years 20 to 30 of each simulation.

proximately one order of magnitude smaller than those in the SO. It is these momentum fluxes that are approximated by parameterizations such as KEB20, ZB20, and GZ21. When comparing their KE transfers to the ground truth, we must also take the dissipative fluxes of the Smagorinsky scheme into account (see section 2.1).

Overall, we find that the combined KE fluxes from parameterizations, including contributions from the Smagorinsky scheme (Equation 9), are more dissipative than those diagnosed from the coarse-grained ground truth (Equation 8). This demonstrates the additional dissipation necessary for stable simulations at coarse resolution, as discussed in Section 2.1. The positive KE fluxes of the KEB20 experiment are reduced and shifted towards smaller scales in both the GS region and the SO. The resulting power spectrum is less energetic for larger scales and more energetic for smaller scales compared to the C16 target spectrum. In the GS region, these differences are more pronounced. ZB20, on the other hand, has larger upscale and downscale fluxes that are shifted towards larger scales in the SO. They remain shifted towards larger scales but are much reduced in the GS region. The KE spectra in both regions suggest that ZB20 hardly affects scales below 90 km, as they closely follow the coarse R4 experiment. This is due to the Gaussian kernel $\mathcal{G}$ (Equation 15) filtering the smaller scales of the subgrid forcing $\mathcal{S}_{ZB}$. GZ21 and Smagorinsky combined have a net negative KE transfer across all scales in the SO. In the GS region, the fluxes are dissipative towards smaller scales and only marginally positive for larger scales. The KE spectra hardly change in the SO compared to the R4 reference experiment, suggesting that there is little contribution from the parameterization in that region. Scales above approximately 200km agree well with the ground truth spectrum of the GS region.

To further illustrate the different performances of the parameterizations in the SO and GS regimes, we need to go beyond the instantaneous transfers of their predicted subgrid fluxes and towards the evolution of the circulation pattern over time. Even after the relatively short integration period of 30 years, we identify substantial differences in the large-scale circulation, notably the meridional heat transport (see Figure 6). The total meridional heat transport (panel **(a)** of Figure 6) has a stronger southward component at higher resolution, with enhanced southward and decreased northward transport. GZ21 only has a marginal but contrary effect on the total heat transport. Both the ZB20 and KEB20 experiments agree well with the ground truth experiment. Following previous studies (e.g., Jayne & Marotzke, 2002), we diagnose the time mean eddy contribution of the meridional heat transport as

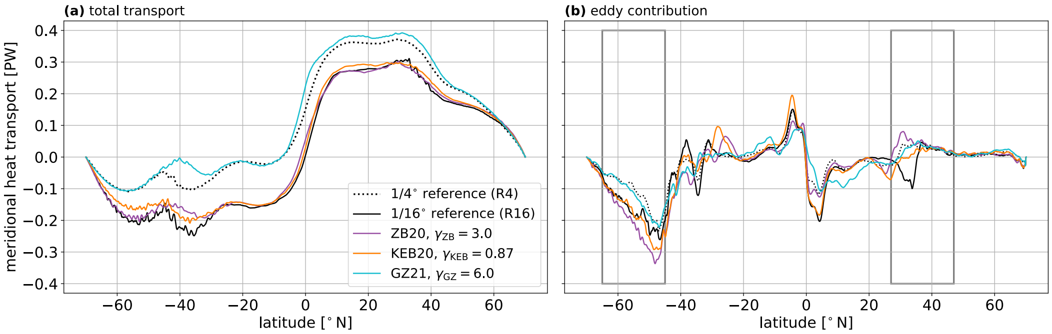

Figure 6. Zonally and vertically averaged meridional heat transport for all experiments. Panel **(a)** shows the total transport averaged over years 20 to 30 of each experiment. Panel **(b)** only shows the eddy contribution to the meridional heat transport. The boxes mark the latitudes of the Southern Ocean region (left) and Gulf Stream separation region (right).

$$\langle v' \, \theta' \rangle = \langle v \, \theta \rangle - \langle v \rangle \, \langle \theta \rangle \,, \tag{19}$$

where we use angle brackets to denote a time average over the ten years of data collection and θ is the potential temperature. The largest differences in the eddy contribution in R4 versus R16 arise in the SO and GS (marked by boxes in panel **(b)** of Figure 6). In the SO, R16 has an enhanced southward eddy heat transport. This is well captured by ZB20 and KEB20, with slightly overestimated eddy transports. As for the total heat transport, GZ21 does not reproduce this behavior.

After detaching from the western boundary, the Gulf Stream representation in R16 propagates almost entirely zonally eastward and leads to eddy-induced stirring against the meridional temperature gradient. This produces a net southward eddy heat transport. Some eddies are advected northward, causing a net northward eddy transport north of the separation point. In the R4 experiment, this signal is shifted southward with reduced eddy-induced southward transport and increased northward transport. While ZB20 and KEB20 increase the respective northward and southward transports, none of the parameterizations improve the latitude of the Gulf Stream separation with respect to R16. We visualize these differences in the Gulf Stream pathway with geostrophic currents derived from sea surface height fields of the western part of the GS region (see Figure 7). In the R16 experiment, we observe an intense jet in the narrow western boundary layer, which detaches around $36°N$ and splits into several branches to the north, east, and northeast. For all coarse experiments, with and without parameterization, the Gulf Stream

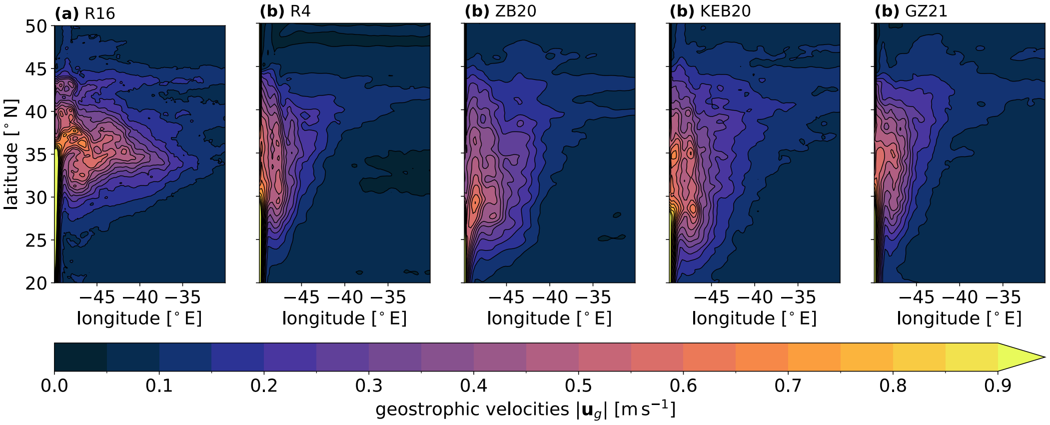

Figure 7. Geostrophic velocities averaged over years 20 to 30 for the **(a)** R16, **(b)** R4, **(c)** ZB20, **(d)** KEB20 and **(e)** GZ21 experiments. The meridional extent corresponds to the Gulf Stream separation region defined by the upper black box in Figure 4.

is much weaker, broader, and does not clearly separate from the coast, while the largest northward velocities do not reach as far north as in R16. The sensitivity of Gulf Stream separation to horizontal resolution, as well as the impact of topography, has already been discussed at length (e.g., Chassignet & Xu, 2017). This mechanism is not accounted for in the closure problem formulated here and is therefore not addressed by the subgrid fluxes predicted by ZB20, KEB20, or GZ21.

4.2 Long-term metrics

We identified differences between two data-driven eddy parameterizations and a more traditional process-based backscatter parameterization concerning their predicted momentum fluxes. In order to assess these with respect to a high-resolution ground truth, we are limited by computational cost to 30 years of simulation rollouts; this duration is too short a timescale for long-adjusting climate metrics. As a compromise, we afford 600-year rollouts of the low-resolution experiments, namely R4, ZB20, and KEB20. We exclude GZ21 from this comparison due to its large biases in KE transfers, KE profiles, and meridional heat transport, which we observe within the first 30 years. We analyze the last 50 years of each 600-year-long experiment to compare the mean state with respect to the slowly adjusting stratification and the associated large-scale ocean circulation.

Starting with the MOC in density space, we find that all experiments exhibit an interhemispheric clockwise overturning cell above a denser anti-clockwise bottom cell (Figure 8). The upper cell is associated with the model equivalent of North Atlantic Deep Water (NADW), forming at the surface, close to the northern boundary. There, it sinks, flows along a nearly adiabatic pathway crossing almost the entire basin, and resurfaces in the Southern Ocean channel. The lower cell is associated with the model equivalent of Antarctic Bottom Water (AABW). This is the densest water mass, which forms close to the southern boundary and flows northward along the bottom topography. It slowly loses buoyancy through bottom mixing and returns southward, resurfacing in the southern part of the periodic channel.

Based on the circulation pattern, we roughly define the model representation of AABW and NADW as water with densities $\sigma_2 > 36\,\mathrm{kg\,m^{-3}}$ and $36\,\mathrm{kg\,m^{-3}} \geq \sigma_2 > 35\,\mathrm{kg\,m^{-3}}$, respectively. This definition does not represent the observed water mass properties of

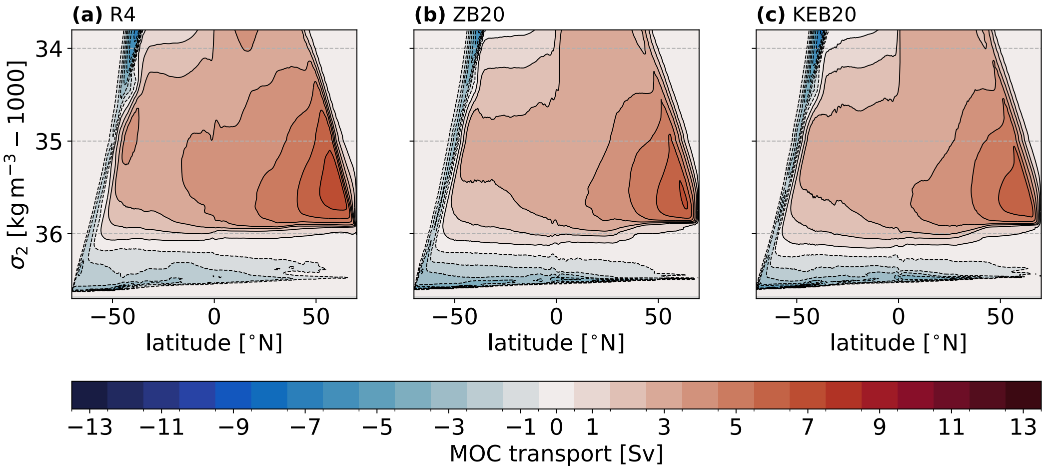

Figure 8. Meridional overturning circulation strength in density space for the **(a)** unparameterized reference experiment, **(b)** Zanna-Bolton 2020 and **c)** Kinetic-Energy-Backscatter 2020 parameterization.

AABW and NADW but serves to characterize the idealized mean stratification and circulation, following the model of Vallis (2017).

The overall overturning structure is not affected by the parameterizations; however, both ZB20 and KEB20 weaken the upper cell and strengthen the lower cell of the meridional overturning circulation compared to the reference, with KEB20 having a stronger effect than ZB20 (Figure 8 and panel **(d)** of Figure 9). As we cannot afford to run the ground truth experiment for the same adjustment time, we cannot comment on which is more realistic. These changes in the large-scale circulation are not a direct consequence of the KE transfers discussed in the previous section, but arise indirectly through slow changes in stratification caused by the intensified eddy field. Indeed, more energetic eddies lead to enhanced lateral stirring of tracers, ultimately flattening the isopycnal surfaces. We find this effect to be most pronounced in the Southern Ocean channel, where the isopycnals are steeper and backscatter is stronger (as shown for the $\sigma_2 = 27\,\mathrm{kg\,m^{-3}}$ and $\sigma_2 = 26\,\mathrm{kg\,m^{-3}}$ isopycnals in panel **(a)** of Figure 9). This leads to substantial changes in stratification: the volume of AABW increases at the expense of NADW for both KEB20 and ZB20 (see panel **(c)** of Figure 9). Changes in the volumes of AABW and NADW are ultimately reflected in the overturning circulation strength: more AABW is associated with a stronger lower cell, and less NADW with a weaker upper cell (panel **d)** of Figure 9). The changes in stratification also impact the ACC through thermal wind balance: the steeper the slope of the isopycnals in the channel, the stronger the current. Consequently, we diagnose the strongest transport in the unparameterized reference experiment (see panel **(b)** of Figure 9). ZB20 exhibits the flattest isopycnals overall and, accordingly, the weakest ACC transport. The presented response patterns of the ACC, MOC, and stratification are directly relevant for the application of backscatter parameterizations in global ocean models (e.g., Juricke et al., 2020; Grooms et al., 2025; Perezhogin et al., 2025).

5 Discussion

The assessment of two data-driven eddy closures performed in this study raises several issues regarding their a priori training methodology and generalizability to differ-

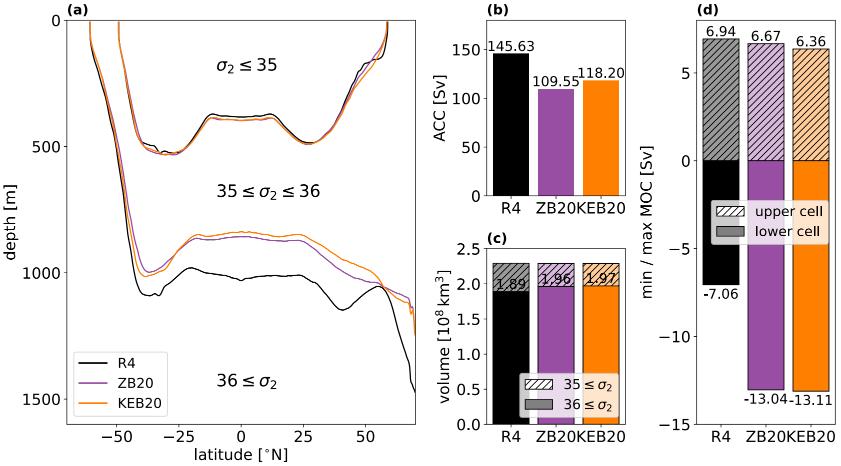

Figure 9. Comparison of (a) zonal mean stratification, **(b)** ACC strength, **(c)** total volumes of NADW (hatched) and AABW (non-hatched) and **(d)** the maximum overturning strength of the lower (non-hatched) and upper (hatched) cell for the R4, ZB20 and KEB20 experiment.

ent ocean models and dynamical regimes. In this section, we will discuss these findings and propose possible avenues for addressing them in future studies.

We note that the coarse-graining approach employed in both studies fails to yield a numerically stable model at runtime. This is expressed by the difference between the governing equations of the coarse grained high resolution (Equation 6) and the low resolution (Equation 1) experiment, where the latter requires a viscous dissipation term $\mathbf{D}^{\mathbf{u}\downarrow}_{\mathrm{SMA}}$ to suppress gradients near the grid scale. The inconsistency between the coarse-grained target model and the coarse model used for simulation rollouts is a known issue, causing poor a posteriori (online) performance of closures that were trained a priori (offline) (Sanderse et al., 2024). A promising approach to this problem is online learning, where the parameterization is trained during runtime, adjusting the model trajectory step by step. This method implicitly informs the ML model about the coarse model, which has been found to lead to more stable online simulations (see e.g., Rasp, 2020; Frezat et al., 2022). However, this is not straightforward for general circulation models like NEMO, as it requires a differentiable solver or a differentiable adjoint model (Frezat et al., 2024). An alternative approach for this specific case could be to inform the ML model about the additional dissipative fluxes explicitly through the Smagorinsky scheme, since we can diagnose them a priori from the coarse-grained data (Mana & Zanna, 2014). One approach could be to follow the considerations from more traditional backscatter closures, such as KEB20, to constrain the domain-integrated upscale KE transfer of the predicted fluxes $\tilde{\mathcal{S}}$ by the domain-integrated downscale fluxes of the viscosity operator during training. While there have been efforts for the combined treatment of the dissipative and backscatter fluxes (e.g., Perezhogin & Glazunov, 2023), none of these approaches eliminates the need for additional numerical dissipation in coarse-resolution ocean models and the associated risk of error compensation. This remains an open question for the ocean modeling community.

As ZB20 and GZ21 were already trained offline prior to this evaluation study, we employed an online tuning strategy targeting the total integrated KE, thereby allowing for a controlled compensation of the numerically imposed dissipation and backscatter during the evaluation. The tuning procedure exposed an explicit dependency of the subgrid fluxes on the size of the grid cell, which was not detected during the training phase. For ZB20, the grid dependency is accounted for through the discretization of the discovered equation and the association of its free parameter to the local grid scale (see Equation 17) following the theoretical arguments of previous studies. A key advantage of the ZB20 approach thus lies in its interpretability, in contrast to the black-box deep learning approach of GZ21, which is agnostic to the model grid. ZB20 outperforms GZ21 across all metrics shown in this study, suggesting that information about the grid is essential for the generalizability of ML-based eddy parameterizations (see Perezhogin et al., 2025, for a detailed discussion).

We did not tune each vertical layer independently, nor with horizontally varying coefficients, as suggested by Zhang et al. (2023). This decision was motivated by two factors: First, this would involve too many simulation rollouts for the a posteriori approach chosen here. Second, it could lead to overfitting: in a first attempt for GZ21, we found that the optimal vertical profile of $\gamma_{\mathrm{GZ}}(z)$ would invert the sign of the predicted net KE flux for a few layers at the bottom of the mixed layer. While this demonstrates the difficulty for GZ21 to generalize from surface training data to the baroclinic flow of the model, injecting energy where the model predicts extraction and vice versa suggests overfitting rather than capturing a physical mechanism. Similarly, dependencies on the grid, latitude, or flow should be detected in the training phase of the data-driven approaches.

The KE and KE transfer spectra (see Figure 5) show that ZB20 is more effective in energizing the large-scale than the baseline backscatter parameterization KEB20. The scales at which energy is injected in KEB20 arise from the negative Laplacian operator, which is not strictly derived from first principles. These differences in the instantaneous KE transfers only translate to minor differences in the slowly adjusting mean state. Both parameterizations energize the eddy field, leading to enhanced tracer stirring and a flattening of isopycnal slopes, especially in the Southern Ocean. This modifies the water mass properties at depth, slowly changing the large scale circulation towards a weaker ACC and upper MOC cell and a stronger bottom MOC cell. The described mechanism is consistent with the model study by Marshall et al. (2017) and the backscatter parameterization experiments performed in a global ocean model by Juricke et al. (2020). We find that ZB20 is slightly more effective in flattening the isopycnals compared to KEB, resulting in a weaker ACC. KEB20, on the other hand, produces more AABW at the expense of NADW, favoring a slightly stronger deep cell and a slightly weaker upper cell of the MOC.

Ultimately, we cannot evaluate these long-term metrics quantitatively with respect to the high-resolution ground truth, as we cannot afford 600 years of the R16 experiment. Nevertheless, the spectra, KE profiles, and integrated meridional heat transport of the first 30 years indicate that ZB20 and KEB20 both perform well in the Southern Ocean, but not in the Gulf Stream region. We attribute this to a misrepresentation of the bathymetry and the western boundary layer at coarse resolution. Previous studies in global ocean models found that backscatter improves the Gulf Stream separation, mainly by countering the excessive numerical dissipation of the western boundary current at coarse resolution (Juricke et al., 2020; Chang et al., 2023). We cannot reproduce these results in the idealized model used in this study. At coarse resolution, the Gulf Stream is overall less energetic, and none of the tested backscatter parameterizations lead to a clear improvement in the separation or path of the current. The DINO configuration has a very idealized bathymetry, with an exponential slope at the boundary and vertical walls above 2000 m (see Appendix A of Kamm et al., 2025, for details). Western boundary currents can, therefore, only be represented in a simplified way. In this context, the representation of bathymetry affected by horizontal resolution may play a more significant role than KE pathways in controlling the separation and course of the Gulf Stream. This is not addressed by any of the parameterizations presented here and does not emerge naturally from the coarse-graining method of ZB20 and GZ21. There are other approaches to this problem, such as terrain-following representations of the bathymetry or volume penalization informed by the high-resolution bathymetry (Debreu et al., 2022), but these are beyond the scope of this study.

6 Conclusion

We have implemented and tested two data-driven eddy parameterizations, Zanna and Bolton (2020) (ZB20) and Guillaumin and Zanna (2021) (GZ21), in the NEMO ocean model, using a configuration of intermediate complexity (Kamm et al., 2025). Our work is pioneering, as it is among the first to integrate a neural network into the NEMO model. Both parameterizations were trained a priori on subgrid fluxes diagnosed from coarse-grained, high-resolution data generated by different ocean models than the one in which they were tested. We emphasize that this approach neglects the enhanced viscous fluxes necessary for numerical stability in coarse simulation rollouts. We compensate for this by tuning the combined fluxes of each parameterization and the viscosity scheme a posteriori to match the domain-integrated kinetic energy of the ground truth.

GZ21 does not generalize well to the previously unseen model data. It was trained only on surface data and is agnostic to the spatial grid. The closure, therefore, tends to underestimate the KE backscatter at high latitudes and near the surface while overestimating it at low latitudes and at depth. It does not improve the KE spectra or the large-scale circulation patterns compared to the unparameterized control run. ZB20 generalizes better, as it is inherently grid-aware through the discretization of the discovered equation and the dimensional scaling introduced by Perezhogin et al. (2024). The KE spectra resemble the ground truth well, especially in the Southern Ocean. ZB20 exhibits KE spectra closer to the R16 experiment when compared to the KEB20 baseline experiment, since it energizes the large scale more efficiently. Both ZB20 and KEB20 improve the basin-scale circulation through the energized eddy field. In the Southern Ocean, the enhanced tracer stirring flattens the isopycnals, leading to a weaker ACC, a stronger deep cell of the MOC, and a weaker upper cell of the MOC, which is in line with previous studies (Marshall et al., 2017; Juricke et al., 2020).

We cannot assess whether the data-driven ZB20 parameterization outperforms the baseline in terms of its effect on the mean state, as we cannot afford to run the ground truth to thermodynamic equilibrium. Based on the 30 years available, we nevertheless identify substantial differences in the mean flow pattern in the Gulf Stream region. All parameterizations evaluated here produce a Gulf Stream that is too weak, overly broad, and separates too far south. We attribute this to the misrepresentation of the idealized bathymetry at coarse resolution, which is not addressed by the backscatter parameterizations presented in this study.

The study offers insight into the impact of the training strategy for data-driven eddy parameterizations, particularly regarding their generalizability across different modeling environments. As guidance for future developments, we propose to *i)* inform data-driven parameterizations with the grid to allow for the inference of spatial derivatives and *ii)* consider all resolution-dependent model components when formulating the target ground truth. For *ii)*, we highlight the excessive numerical dissipation introduced by the viscosity scheme at coarse resolution, which leads to a discrepancy between the a priori and a posteriori behavior of the data-driven parameterizations. Finally, we advocate that the ultimate aim of the data-driven approach should be to develop physically interpretable momentum closures, as this facilitates generalizability to other modeling frameworks.

Data Availability Statement

The DINO configuration is available via Kamm et al. (2025), and all experiments presented here can be reproduced using Kamm (2026b). The processed simulation data and the code to generate all figures in this manuscript are available via Kamm (2026a).

Conflict of Interest declaration

The authors declare there are no conflicts of interest for this manuscript.

Acknowledgments

This project is supported by Schmidt Sciences, LLC. This work was granted access to the HPC resources of IDRIS under the allocation 2023-A0140107451 and 2024-A0160107451 made by GENCI and has been supported by the European Union's Horizon Europe research and innovation program (grant no. 101081193) as a contribution to the OptimESM project. We would like to thank all members of M2LInES, especially Laure Zanna, Alistair Adcroft, and Dhruv Balwada, for their insightful comments and discussions.

References

Acosta, M. C., Palomas, S., Paronuzzi Ticco, S. V., Utrera, G., Biercamp, J., Bretonniere, P.-A., … Balaji, V. (2024). The computational and energy cost of simulation and storage for climate science: lessons from cmip6. *Geoscientific Model Development*, *17*(8), 3081–3098. Retrieved from `https://gmd.copernicus.org/articles/17/3081/2024/` doi: 10.5194/gmd-17-3081-2024

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., … Zhang, R. (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. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001726` doi: https://doi.org/10.1029/2019MS001726

Aluie, H. (2019). Convolutions on the sphere: Commutation with differential operators. *GEM-International Journal on Geomathematics*, *10*(1), 9.

Balwada, D., Perezhogin, P., Adcroft, A., & Zanna, L. (2025). Design and implementation of a data-driven parameterization for mesoscale thickness fluxes. *Authorea Preprints*.

Barge, A., Le Sommer, J., & IGE-MEOM. (2024, December). *Eophis.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.14288831` doi: 10.5281/zenodo.14288831

Barge, A., Meunier, E., & Kamm, D. (2024, September). *morays-community/nemo-dino.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.13851911` doi: 10.5281/zenodo.13851911

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.

Chassignet, E. P., & Xu, X. (2017). Impact of horizontal resolution (1/12° to 1/50°) on gulf stream separation, penetration, and variability. *Journal of Physical Oceanography*, *47*(8), 1999 - 2021. Retrieved from `https://journals.ametsoc.org/view/journals/phoc/47/8/jpo-d-17-0031.1.xml` doi: 10.1175/JPO-D-17-0031.1

Chen, S., Ecke, R. E., Eyink, G. L., Wang, X., & Xiao, Z. (2003). Physical mechanism of the two-dimensional enstrophy cascade. *Physical review letters*, *91*(21), 214501.

Cooper, F. C., & Zanna, L. (2015). Optimisation of an idealised ocean model, stochastic parameterisation of sub-grid eddies. *Ocean Modelling*, *88*, 38–53.

Craig, A., Valcke, S., & Coquart, L. (2017). Development and performance of a new version of the oasis coupler, oasis3-mct_3. 0. *Geoscientific Model Development*, *10*(9), 3297–3308.

Debreu, L., Kevlahan, N.-R., & Marchesiello, P. (2022). Improved gulf stream separation through brinkman penalization. *Ocean Modelling*, *179*, 102121.

Delworth, T. L., Rosati, A., Anderson, W., Adcroft, A. J., Balaji, V., Benson, R., … Zhang, R. (2012). Simulated climate and climate change in the gfdl cm2.5 high-resolution coupled climate model. *Journal of Climate*, *25*(8), 2755 - 2781. Retrieved from `https://journals.ametsoc.org/view/journals/clim/25/8/jcli-d-11-00316.1.xml` doi: 10.1175/JCLI-D-11-00316.1

Ferrari, R., & Wunsch, C. (2009). Ocean circulation kinetic energy: Reservoirs, sources, and sinks. *Annual Review of Fluid Mechanics*, *41*(1), 253–282.

Frezat, H., Balarac, G., Le Sommer, J., Fablet, R., & Lguensat, R. (2021, Feb). Physical invariance in neural networks for subgrid-scale scalar flux modeling. *Phys. Rev. Fluids*, *6*, 024607. Retrieved from `https://link.aps.org/doi/10.1103/PhysRevFluids.6.024607` doi: 10.1103/PhysRevFluids.6.024607

Frezat, H., Fablet, R., Balarac, G., & Sommer, J. L. (2024). *Gradient-free online learning of subgrid-scale dynamics with neural emulators.* Retrieved from `https://arxiv.org/abs/2310.19385`

Frezat, H., Le Sommer, J., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003124.

Gaspar, P., Grégoris, Y., & Lefevre, J.-M. (1990). A simple eddy kinetic energy model for simulations of the oceanic vertical mixing: Tests at station papa and long-term upper ocean study site. *Journal of Geophysical Research: Oceans*, *95*(C9), 16179–16193.

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

Griffies, S. M., & Hallberg, R. W. (2000). Biharmonic friction with a smagorinsky-like viscosity for use in large-scale eddy-permitting ocean models. *Monthly Weather Review*, *128*(8), 2935–2946.

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.

Guan, Y., Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2022). Stable a posteriori les of 2d turbulence using convolutional neural networks: Backscattering analysis and generalization to higher re via transfer learning. *Journal of Computational Physics*, *458*, 111090.

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534.

Gultekin, C., Subel, A., Zhang, C., Leibovich, M., Perezhogin, P., Adcroft, A., … Zanna, L. (2024). An analysis of deep learning parameterizations for ocean subgrid eddy forcing. *arXiv preprint arXiv:2411.06604*.

Hallberg, R. (2013, December). Using a resolution function to regulate parameterizations of oceanic mesoscale eddy effects. *Ocean Modelling*, *72*, 92–103. Retrieved 2023-11-09, from `https://linkinghub.elsevier.com/retrieve/pii/S1463500313001601` doi: 10.1016/j.ocemod.2013.08.007

Hewitt, H. T., Roberts, M., Mathiot, P., Biastoch, A., Blockley, E., Chassignet, E. P., … Zhang, Q. (2020). Resolving and parameterising the ocean mesoscale in earth system models. *Current Climate Change Reports*, *6*, 137–152.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q.,

. . . Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589 - 602. Retrieved from `https://journals.ametsoc.org/view/journals/bams/98/3/bams-d-15-00135.1.xml` doi: 10.1175/BAMS-D-15-00135.1

Jakhar, K., Guan, Y., Mojgani, R., Chattopadhyay, A., & Hassanzadeh, P. (2024). Learning closed-form equations for subgrid-scale closures from high-fidelity data: Promises and challenges. *Journal of Advances in Modeling Earth Systems*, *16*(7), e2023MS003874.

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.

Jansen, M. F., & Held, I. M. (2014). Parameterizing subgrid-scale eddy effects using energetically consistent backscatter. *Ocean Modelling*, *80*, 36–48.

Jansen, M. F., Held, I. M., Adcroft, A., & Hallberg, R. (2015). Energy budget-based backscatter in an eddy permitting primitive equation model. *Ocean Modelling*, *94*, 15–26.

Jayne, S. R., & Marotzke, J. (2002). The oceanic eddy heat transport. *Journal of Physical Oceanography*, *32*(12), 3328–3345.

Juricke, S., Danilov, S., Koldunov, N., Oliver, M., & Sidorenko, D. (2020). Ocean kinetic energy backscatter parametrization on unstructured grids: Impact on global eddy-permitting simulations. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001855.

Juricke, S., Danilov, S., Kutsenko, A., & Oliver, M. (2019). Ocean kinetic energy backscatter parametrizations on unstructured grids: Impact on mesoscale turbulence in a channel. *Ocean Modelling*, *138*, 51–67.

Kamm, D. (2026a, January). *Processed data and notebook for data-driven eddy parameterisation paper.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.18405109` doi: 10.5281/zenodo.18405109

Kamm, D. (2026b, January). *vopikamm/dino-parameterizations-2026: v0.1.2 – included restart files.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.18413986` doi: 10.5281/zenodo.18413986

Kamm, D., Deshayes, J., & Madec, G. (2025). Dino: A diabatic model of pole-to-pole ocean dynamics to assess subgrid parameterizations across horizontal scales. *Geoscientific Model Development*, *39*(1), 125–142.

Klöwer, M., Jansen, M. F., Claus, M., Greatbatch, R. J., & Thomsen, S. (2018). Energy budget-based backscatter in a shallow water model of a double gyre basin. *Ocean Modelling*, *132*, 1–11.

Kraichnan, R. H. (1967). Inertial ranges in two-dimensional turbulence. *Physics of fluids*, *10*(7), 1417.

Larichev, V. D., & Held, I. M. (1995). Eddy amplitudes and fluxes in a homogeneous model of fully developed baroclinic instability. *Journal of physical oceanography*, *25*(10), 2285–2297.

Leonard, A. (1975). Energy cascade in large-eddy simulations of turbulent fluid flows. In *Advances in geophysics* (Vol. 18, pp. 237–248). Elsevier.

Madec, G., Bell, M., Blaker, A., Bricaud, C., Bruciaferri, D., Castrillo, M., . . . Wilson, C. (2023, July). NEMO Ocean Engine Reference Manual. *Zenodo*. Retrieved 2023-12-19, from `https://zenodo.org/records/8167700` (Publisher: Zenodo) doi: 10.5281/zenodo.8167700

Mana, P. P., & Zanna, L. (2014). Toward a stochastic parameterization of ocean mesoscale eddies. *Ocean Modelling*, *79*, 1–20.

Marques, G. M., Loose, N., Yankovsky, E., Steinberg, J. M., Chang, C.-Y., Bhamidipati, N., . . . Zanna, L. (2022, September). NeverWorld2: an idealized model hierarchy to investigate ocean mesoscale eddies across resolutions. *Geoscientific Model Development*, *15*(17), 6567–6579. Retrieved 2022-12-13, from `https://gmd.copernicus.org/articles/15/6567/2022/` (Publisher: Copernicus GmbH) doi: 10.5194/gmd-15-6567-2022

Marshall, J., Scott, J. R., Romanou, A., Kelley, M., & Leboissetier, A. (2017). The dependence of the ocean's moc on mesoscale eddy diffusivities: A model study. *Ocean Modelling*, *111*, 1–8.

Meneveau, C., & Katz, J. (2000). Scale-invariance and turbulence models for large-eddy simulation. *Annual Review of Fluid Mechanics*, *32*(1), 1–32.

Perezhogin, P. (2020). Testing of kinetic energy backscatter parameterizations in the nemo ocean model. *Russian Journal of Numerical Analysis and Mathematical Modelling*, *35*(2), 69–82.

Perezhogin, P., Adcroft, A., & Zanna, L. (2025). Generalizable neural-network parameterization of mesoscale eddies in idealized and global ocean models. *arXiv preprint arXiv:2505.08900*.

Perezhogin, P., & Glazunov, A. (2023). Subgrid parameterizations of ocean mesoscale eddies based on germano decomposition. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003771.

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.

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and lorenz 96 case study (v1.0). *Geoscientific Model Development*, *13*(5), 2185–2196. Retrieved from `https://gmd.copernicus.org/articles/13/2185/2020/` doi: 10.5194/gmd-13-2185-2020

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).

Sanderse, B., Stinis, P., Maulik, R., & Ahmed, S. E. (2024). Scientific machine learning for closure models in multiscale problems: A review. *arXiv preprint arXiv:2403.02913*.

Smagorinsky, J. (1963). General circulation experiments with the primitive equations: I. the basic experiment. *Monthly weather review*, *91*(3), 99–164.

Uchida, T., Rokem, A., Soler, S., Nicholas, T., Abernathey, R., frederic nouguier, . . . Moon, Z. (2023, February). *xgcm/xrft: v1.0.1.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7621857` doi: 10.5281/zenodo.7621857

Vallis, G. K. (2017). Atmospheric and oceanic fluid dynamics. In (p. 853-859). Cambridge University Press.

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.

Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, *47*(17), e2020GL088376.

Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2023). Implementation and evaluation of a machine learned mesoscale eddy parameterization into a numerical ocean circulation model. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003697.

Zhuang, J., raphael dussin, Huard, D., Bourgault, P., Banihirwe, A., Raynaud, S., . . . Li, X. (2023, September). *pangeo-data/xesmf: v0.8.2.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.8356796` doi: 10.5281/zenodo.8356796