AGU ADVANCING EARTH AND SPACE SCIENCES

Geophysical Research Letters

RESEARCH LETTER
10.1029/2026GL123513

Calibration of a Neural Network Ocean Closure for Improved Mean State and Variability

Pavel Perezhogin[1], Alistair Adcroft[2], and Laure Zanna[1]

[1]Courant Institute School of Mathematics, Computing, and Data Science, New York University, New York, NY, USA,
[2]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA

Key Points:

- Calibrated data-driven eddy parameterization reduces the mean state and variability errors by factors of 1.7–3.3 in coarse ocean models
- Ensemble Kalman Inversion effectively optimizes neural network parameters in the presence of chaotic ocean dynamics
- Efficient calibration is achievable with short simulations without integrating the ocean model to statistical equilibrium

Supporting Information:

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

Correspondence to:

P. Perezhogin,
pp2681@nyu.edu

Citation:

Perezhogin, P., Adcroft, A., & Zanna, L. (2026). Calibration of a neural network ocean closure for improved mean state and variability. *Geophysical Research Letters*, *53*, e2026GL123513. https://doi.org/10.1029/2026GL123513

Received 4 APR 2026
Accepted 12 JUN 2026

Abstract

Global ocean models exhibit biases in the mean state and variability, particularly at coarse resolution, where mesoscale eddies are unresolved. To address these biases, parameterization coefficients are typically tuned ad hoc. Here, we formulate parameter tuning as a calibration problem using Ensemble Kalman Inversion (EKI). We optimize parameters of a neural network parameterization of mesoscale eddies in two idealized ocean models at coarse resolution. The calibrated parameterization reduces errors in the time-averaged fluid interfaces and their variability by factors of 1.7–3.3 compared to the unparameterized model, depending on the metric and configuration. The EKI method is robust to noise in time-averaged statistics arising from chaotic ocean dynamics. Furthermore, we propose an efficient calibration protocol that bypasses integration to statistical equilibrium by carefully choosing an initial condition. These results demonstrate that systematic calibration can substantially improve coarse-resolution ocean simulations and provide a practical pathway for reducing biases in global ocean models.

Plain Language Summary

Ocean models used for climate prediction have persistent errors because they cannot capture small-scale swirling currents called eddies. Models often include mathematical corrections called parameterizations to approximate the effects of these missing eddies, but the adjustable settings in these corrections are usually chosen by hand through trial and error. We use a machine learning approach combined with an automatic tuning method to identify improved settings for an eddy parameterization in two simplified ocean simulations. Our tuning method reduces errors in both the average ocean state and its natural fluctuations by roughly half compared to an untuned model. Importantly, the method works well even when ocean statistics are noisy due to the chaotic nature of ocean currents, and it can be applied with relatively short simulations rather than waiting hundreds of years of simulations for the ocean model to fully adjust. These results offer a practical path toward reducing longstanding biases in the ocean models used for climate projections.

1. Introduction

Global ocean models exhibit substantial biases in the mean state and variability (Richter & Tokinaga, 2020; Wang et al., 2014). For example, representing variability of the Western Boundary Currents (WBC) is challenging across a wide range of horizontal resolutions of the ocean models, including non-eddy-resolving (Grooms et al., 2024), eddy-permitting (Juricke et al., 2020), and submesoscale-permitting (Uchida et al., 2022). The mean-state biases include errors at the air-sea interface and subsurface isopycnal structure (Adcroft et al., 2019; Griffies et al., 2015). These biases are often mitigated by incorporating various parameterizations (Andrejczuk et al., 2016; Chang et al., 2023; Grooms et al., 2024; Juricke et al., 2017, 2020). However, the adjustment of parameterization coefficients is frequently performed in an ad hoc manner. Here, we formalize this process as a calibration problem, see also Cooper and Zanna (2015); Cooper (2017).

Recently, multiple data-driven eddy parameterizations have been proposed to reduce biases in the ocean mean state and variability (Guillaumin & Zanna, 2021; Kamm et al., 2026; Perezhogin et al., 2025; C. Zhang et al., 2023; Zanna & Bolton, 2020). These parameterizations perform well at an eddy-permitting resolution (1/4°). However, their performance often degrades at a coarser resolution (1/2°), a particularly challenging resolution for testing eddy parameterizations (Jansen et al., 2019; Yankovsky et al., 2024). Developing skillful data-driven parameterizations for coarse ocean models is especially important, as these models are widely used in climate simulations (Grooms et al., 2024) and offer the greatest potential for bias reduction (Perezhogin et al., 2023). We build on recent works demonstrating that optimizing the parameterization parameters in online simulations can significantly improve the model fidelity at coarse resolution (Christopoulos et al., 2024; Frezat et al., 2022; Kochkov et al., 2021; Lee et al., 2025; Lopez-Gomez et al., 2022; Maddison, 2026; Ouala et al., 2024; Wagner et al., 2025; Yan et al., 2025). In contrast to physics-based parameterizations, machine-learned parameterizations often contain too many parameters for manual tuning, motivating the use of automatic calibration methods.

Calibration methods are designed to automatically adjust the free parameters of a parameterization by minimizing the mismatch between the output of the coarse-resolution ocean model and observations, that is, by minimizing a prescribed loss function. Calibration of ocean models is exceptionally expensive due to the long spin-up period, which can take hundreds of years (Mrozowska et al., 2025; Williamson et al., 2017), as well as the extended time window required to accurately estimate time-averaged statistics. In this work, we demonstrate the robustness of the Ensemble Kalman Inversion (M. A. Iglesias et al., 2013) calibration method to noise in temporal averages arising from the ocean's chaotic dynamics. Furthermore, we propose a simple method for calibrating fast physical processes without integrating the ocean model to statistical equilibrium, a longstanding challenge in climate modeling (DelSole & Tippett, 2024).

Our goal is to determine to what extent calibration of a neural-network eddy parameterization by Perezhogin et al. (2025) can improve the mean state and variability of an idealized GFDL MOM6 ocean model (Adcroft et al., 2019) at coarse resolution ($1/2°$). We constrain the neural network with physical equivariances to enhance generalization (Kashinath et al., 2021) and reduce the number of calibrated parameters. We consider two idealized ocean configurations, which serve different purposes. The simplest configuration is used to assess the convergence of the calibration algorithm and its robustness to the noise (O. R. Dunbar et al., 2021; O. R. Dunbar et al., 2022; Gjini et al., 2025; Howland et al., 2022). A more complex configuration is used to demonstrate the applicability of our calibration protocol to ocean models with long spin-up times.

2. Methods

In this section, we describe how improving the mean state and variability of the ocean can be framed as a calibration problem. We introduce idealized ocean models and eddy parameterization, followed by the choice of the calibration method, loss function, and a method to avoid long spin-up. A simplified workflow is illustrated in Figure 1.

2.1. Idealized Ocean Models

We consider two idealized configurations of the GFDL MOM6 ocean model: Double Gyre (DG, Perezhogin et al., 2024; C. Zhang et al., 2023; W. Zhang et al., 2025) and NeverWorld2 (NW2, Marques et al., 2022; Yankovsky et al., 2022, 2024). Both configurations represent adiabatic ocean dynamics and solve the stacked shallow water equations, with the circulation driven by prescribed wind stress. Our goal is to improve the time-averaged statistical properties of the ocean circulation in the parameterized simulation at a coarse horizontal resolution $1/2°$, and bring them closer to the statistics of the filtered and coarse-grained high-resolution simulation at resolution $1/32°$. All simulations use the biharmonic Smagorinsky eddy viscosity with coefficient $C_S = 0.06$ (Adcroft et al., 2019) to maintain numerical stability.

The DG configuration (Figure 1a) represents a midlatitude basin with two fluid layers. The imposed wind stress drives two counter-rotating gyres separated by an eastward jet, serving as an idealized model of western boundary current systems. The target ocean circulation is obtained by integrating the $1/32°$ model for 100 years from rest, discarding the first 10 years for spin-up. The coarse ($1/2°$) parameterized model is also evaluated in 100-year simulations. However, during calibration, the coarse model is integrated only for 20 years.

The NW2 configuration represents an idealized Atlantic sector model with 15 fluid layers, featuring multiple circulation regimes, including a circumpolar current in an idealized Southern Ocean, midlatitude gyres, and equatorial flows. The high-resolution simulation ($1/32°$) was spun up in multiple stages in Marques et al. (2022). The coarse ocean model at $1/2°$ resolution is integrated for 5 years during the calibration stage and for 30,000 days from rest during evaluation of the calibrated parameterization. In all simulations, we analyze the final 800 days.

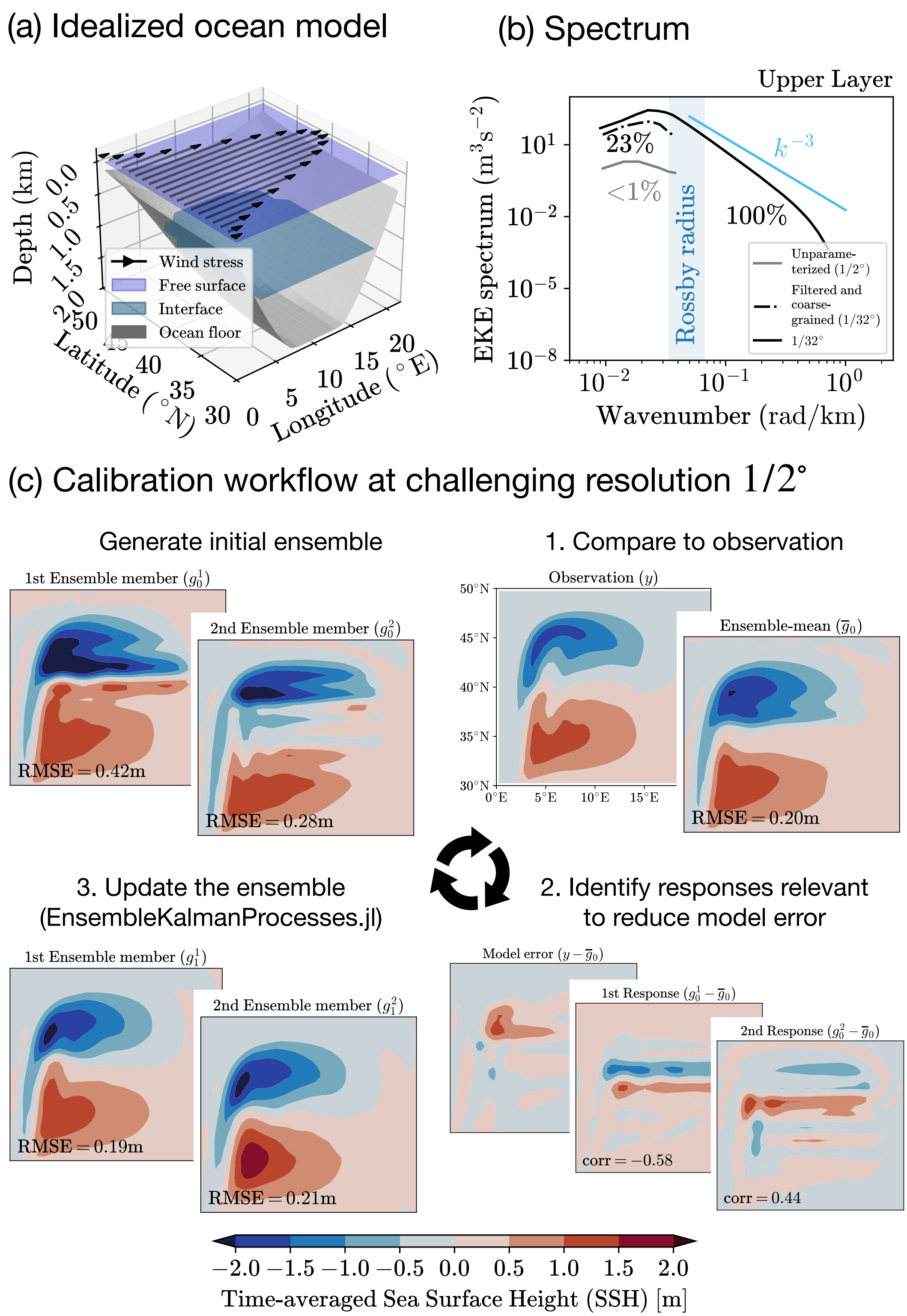

Figure 1. (a) Idealized wind-driven ocean model GFDL MOM6 in a double-gyre configuration. (b) The eddy kinetic energy spectrum as a function of isotropic horizontal wavenumber in the upper fluid layer and domain 5°E − 15°E × 35°N − 45°N. The percentages represent the integral over the spectrum relative to that of the high-resolution simulation. Panel (c) shows how the Ensemble Kalman Inversion calibration algorithm interacts with the coarse ocean model in order to update the free parameters of the parameterization such that the time-mean sea surface height is as close as possible to the filtered and coarse-grained high-resolution simulation ("observation").

2.2. Neural-Network Parameterization of Mesoscale Eddies

We use the recently developed data-driven parameterization of mesoscale eddies by Perezhogin et al. (2025), which modifies the horizontal momentum balance equation:

$$\partial_t \mathbf{u} = \cdots + \nabla \cdot \mathbf{T}, \tag{1}$$

where $\mathbf{T} \in \mathbb{R}^{2\times 2}$ is the horizontal stress tensor predicted by the parameterization, $\mathbf{u} = (u, v)$ is the vector of filtered horizontal velocities, and $\nabla = (\partial_x, \partial_y)$. This parameterization represents the inverse kinetic energy cascade across the grid scale ($\mathbf{u} \cdot (\nabla \cdot \mathbf{T}) > 0$ on average, Storer et al. (2023)) known as backscatter (Berner et al., 2009; Chasnov, 1991; Frederiksen & Davies, 1997; Jansen & Held, 2014; Juricke et al., 2019; Kraichnan, 1976).

The parameterized stress tensor is predicted as follows

$$\mathbf{T}(\mathbf{X}, \Delta) = \gamma \Delta^2 ||\mathbf{X}||_2^2 f_\phi(\mathbf{X}/||\mathbf{X}||_2), \tag{2}$$

where Δ is the coarse grid spacing, γ is a scaling coefficient, which is equal to 1 during training, f_ϕ is the Artificial Neural Network (ANN) with free parameters ϕ. The vector of input features, $\mathbf{X} \in \mathbb{R}^{27}$, consists of horizontal velocity gradients ($\sigma_S = \partial_y u + \partial_x v$, $\sigma_D = \partial_x u - \partial_y v$, $\omega = \partial_x v - \partial_y u$) evaluated on a 3×3 spatial stencil. Normalization of input features with $||\mathbf{X}||_2$ and output features with $\Delta^2 ||\mathbf{X}||_2^2$ introduces dimensional consistency and was shown to improve generalization to out-of-distribution data (Perezhogin et al., 2025).

The parameterization (Equation 2) encapsulates many hard constraints that preserve physical invariances, including Galilean invariance, dimensional consistency, and conservation of momentum and angular momentum. There were only two implemented soft constraints in Perezhogin et al. (2025)—rotational and reflectional invariances. This approach is reasonable when the parameterization is trained offline. However, online recalibration can violate invariances during optimization of the online loss function. Thus, in this work, we further constrain the parameterization (Equation 2) and implement rotational and reflectional invariances as hard constraints following Guan, Subel, et al. (2022); Pawar et al. (2023); Connolly et al. (2025) (see Text S1–S3 in Supporting Information S1). We will refer to the new parameterization as eANN (equivariant ANN).

We train the eANN on the global ocean data set at coarse resolutions in a range of $0.4°–1.5°$ using the same algorithm as in Perezhogin et al. (2025) and report similar offline performance (see Figure S1 in Supporting Information S1). We note that the set of resolutions covered in the training data set is well-suited to the online simulations we consider here ($1/2°$).

2.3. Ensemble Kalman Inversion

The goal of calibration is to adjust the tunable parameters of the parameterization so that the statistics of the coarse ocean model are as close as possible to those of the filtered, coarse-grained high-resolution data. This problem can be framed as the minimization of the following loss function (Gjini et al., 2025):

$$\mathcal{L}(\theta) = ||R^{-1/2}(y - \mathcal{G}(\theta))||_2^2. \tag{3}$$

Here, $\theta \in \mathbb{R}^{n_p}$ is a vector of length n_p representing parameters to be calibrated, $y \in \mathbb{R}^{n_o}$ is a vector of observations of length n_o (i.e., in our case, statistics of filtered and coarsened high-resolution simulation), $\mathcal{G}(\theta)$ is a forward model evaluation (i.e., statistics of the coarse ocean simulation, performed at a given set of parameters θ), and R is the covariance matrix of the observational error.

We solve the optimization problem (Equation 3) using a gradient-free optimization method—Ensemble Kalman Inversion (M. A. Iglesias et al., 2013) implemented in the software package `EnsembleKalmanProcesses.jl` (O. Dunbar et al., 2022). A particular method of Kalman inversion we use among the methods implemented in the package is the Ensemble Transform Kalman Inversion (ETKI, Huang et al., 2022). This choice is made for two reasons: the possibility to explore the parameter space (the ensemble size is not tightly fixed to the number of parameters) and scalability with respect to the observational dimension (n_o).

The ETKI method is initialized by sampling an ensemble of n_e parameter vectors, denoted by $\theta_0^1, \ldots, \theta_0^{n_e}$. The forward model is evaluated at these parameter vectors, giving $g_0^1 = \mathcal{G}(\theta_0^1), \ldots, g_0^{n_e} = \mathcal{G}(\theta_0^{n_e})$. Here, the superscript indexes the ensemble members, while the subscript denotes the iteration number, where 0 corresponds to the initial ensemble. The ensemble is then updated over multiple iterations, as described below (see Figure 1c for illustration).

At each iteration, the ensemble-mean parameter vector is updated as follows (Gjini et al., 2025):

$$\overline{\theta}_{j+1} = \overline{\theta}_j + \delta t \Theta_j \left(I + \delta t G_j^T R^{-1} G_j\right)^{-1} G_j^T R^{-1} \left(y - \overline{g}_j\right), \tag{4}$$

where $\delta t > 0$ is the scheduler step (M. Iglesias & Yang, 2021), j is the iteration number, I is the identity matrix, $\overline{\theta}_j = 1/n_e \sum_i \theta_j^i$ is the ensemble-mean parameter vector, $\overline{g}_j = 1/n_e \sum_i g_j^i$ is the ensemble-mean forward model evaluation, Θ_j and G_j are normalized perturbation matrices:

$$\Theta_j = \frac{1}{\sqrt{n_e - 1}} \left(\theta_j^1 - \overline{\theta}_j, \ldots, \theta_j^{n_e} - \overline{\theta}_j\right) \in \mathbb{R}^{n_p \times n_e}, \tag{5}$$

$$G_j = \frac{1}{\sqrt{n_e - 1}} \left(g_j^1 - \overline{g}_j, \ldots, g_j^{n_e} - \overline{g}_j\right) \in \mathbb{R}^{n_o \times n_e}. \tag{6}$$

The mechanism of the update Equation 4 is as follows. We first compare the ensemble-mean prediction $\overline{g}_j$ with the observational vector y (panel 1 in Figure 1c). Next, we identify the individual responses that project onto the model error, that is, $G_j^T R^{-1} \left(y - \overline{g}_j\right) \neq 0$ (see panel 2 in Figure 1c). Finally, we update the ensemble-mean parameter vector $\overline{\theta}_j$ along a direction that, on average, reduces the model error. The described mechanism is similar to gradient descent using the ensemble-based approximation of the gradient and preconditioning (Chada et al., 2020; Vernon et al., 2025).

The update of the ensemble mean (Equation 4) is followed by the update of the perturbations:

$$\Theta_{j+1} = \Theta_j \left(I + \delta t G_j^T R^{-1} G_j\right)^{-1/2}, \tag{7}$$

which shrinks the ensemble to the consensus.

2.4. Loss Function and Parameters

Our goal is to improve both the mean and variability of the interfaces between fluid layers, as these are often used to assess the impact of eddy parameterizations (Balwada et al., 2025; Grooms et al., 2024; Juricke et al., 2020; Yankovsky et al., 2024). Fluid interfaces, unlike other spatial fields, are associated with large-scale horizontal circulation patterns (i.e., the streamfunction; Vallis (2017)), which are well resolved on the coarse grid.

The observational vector y consists of the time-averaged interfaces (denoted by $\overline{\eta}^t$) and their temporal standard deviation (STD, denoted by $\sqrt{\overline{\eta'^2}^t}$):

$$y = \begin{bmatrix} \overline{\eta}^t \\ \sqrt{\overline{\eta'^2}^t} \end{bmatrix}, \tag{8}$$

where η is the flattened 3D array of the normalized vertical coordinate of the fluid interfaces (for normalization, see Text S4 in Supporting Information S1). The output from filtered and coarse-grained high-resolution simulation (vector y) and the output of the coarse ocean model (vectors g_j^i) are processed in the same manner.

We specify the simplest observational noise model ($R = I$), where I is the identity matrix, so as not to alter the inner product and the loss function, which already contains normalized variables:

$$\mathcal{L}(\theta) = \| y - \overline{g}_j \|_2^2. \tag{9}$$

The loss function (Equation 9) represents a multi-objective optimization problem, in which we intentionally chose equal weights for the time-averaged and the standard deviations of the interfaces. That way, we assign equal contributions for the potential energy bias in representation of the mean state and eddies, which are proportional to $(\overline{\eta}^t)^2$ and $\overline{\eta'^2}^t$, respectively. The energy-based l_2 norm is a popular choice in the analysis of model errors (Tuppi et al., 2023) and optimal disturbances (Zasko et al., 2023). Unlike Yankovsky et al. (2024) and Pudig et al. (2026), we exclude domain-integrated metrics from the calibration protocol as they provide little information about the spatial patterns of the mean state and variability. Additionally, domain-integrated metrics may disrupt the optimization loss dynamics because the calibration algorithm struggles to strike a balance between errors in spatial fields and scalar fields.

We consider the weights and biases of the eANN in the last layer, as well as a coefficient in front of the parameterization γ, as a vector of tunable parameters θ of size $n_p = \dim(\theta) = 14$. Restricting calibration to the deepest layers of a neural network is a common practice (Pahlavan et al., 2024). Furthermore, the first layer is responsible for feature extraction (Guan, Chattopadhyay, et al., 2022), and calibrating it can significantly degrade the eANN's generalization ability. We note that the number of parameters was considerably reduced by using rotational and reflection equivariances as hard constraints (see Text S2 and Table S1 in Supporting Information S1). For constructing an initial ensemble, we perturb the offline trained values of parameter vector θ by 25% of their magnitude (see Table S2 in Supporting Information S1). Note that manually tuning 14 parameters via grid search is challenging, as it would require approximately 10^{14} simulations.

Once the calibration problem is defined, the ETKI update rule (Equation 4) depends on two parameters: the ensemble size (n_e) and scheduler step (δt). We use $n_e = 100$ in the DG configuration, following the recommendation of O. Dunbar et al. (2025) for our parameter dimension. In the NW2 configuration, the ensemble size is doubled because coarse ocean simulations blow up for a bigger percentage of the ensemble members. The scheduler step controls the tradeoff: larger values accelerate the convergence, but the ensemble spread collapses faster. We choose fixed values of δt manually ($\delta t = 2$ in DG and $\delta t = 0.025$ in NW2) to ensure monotonic decrease of the loss function (Equation 9) while avoiding excessively rapid ensemble collapse. The sensitivity to δt and the random seed is shown in Figure S2 in Supporting Information S1.

2.5. Calibrating an Ocean Model Without Integrating to Statistical Equilibrium

Calibration in the NW2 configuration is computationally demanding, as a single simulated day requires 400 times as many CPU-core hours as in the DG configuration. Moreover, the spin-up time from rest is approximately 100 years in NW2, compared with approximately 5 years in DG. Consequently, we developed a method that reduces the computational cost of calibration in the NW2 configuration by using short simulations (5 years):

- We initialize the coarse ocean model with the filtered and coarse-grained snapshot from the spun-up high-resolution simulation,
- We allow the coarse ocean model to adjust and partially forget the initial eddy field during a short spin-up (1,000 days), and compute the loss function by averaging the statistics over an additional 800 days.

Here, the time-averaging interval (800 days) controls the amount of noise in the evaluation of the loss function (Equation 9). We found that considerably reducing the time-averaging interval degrades the performance of the calibrated parameterization (Figure S5 in Supporting Information S1). On the contrary, reducing the spin-up time (1,000 days) to zero does not degrade performance (Figure S5 in Supporting Information S1). This likely happened because, contrary to other studies on ocean model calibration, we use the exact initial condition, which allows us to evaluate the loss function without waiting for the ocean model to adjust. We suggest that the simulation length and its split into spin-up and time-averaging intervals are hyperparameters that should be optimized in future applications.

3. Results

We consider recalibration of the eddy parameterization in the coarse ocean model at a challenging resolution of $1/2°$, which resolves less than 25% of the eddy kinetic energy in the DG configuration in the filtered and coarse-

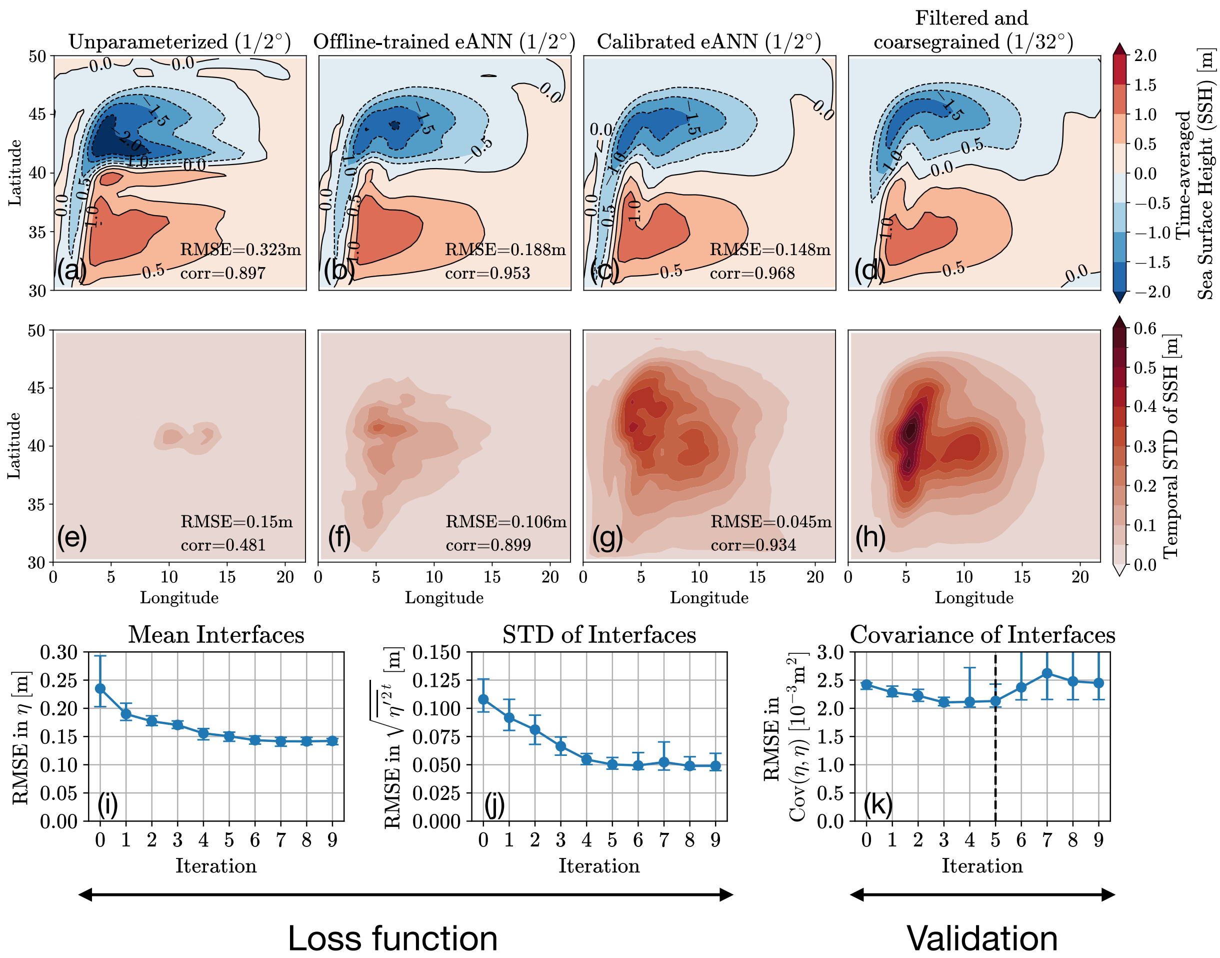

Figure 2. Calibration of the eddy parameterization (eANN) in Double Gyre configuration. The upper row shows time-averaged sea surface height (SSH), and the second row shows the temporal standard deviation of SSH. On these panels, all simulations are 100 years long and results are averaged over 90 years. We show coarse ocean models (1/2°) with: (a), (e) no parameterization, (b), (f) parameterization trained offline with manually adjusted scaling coefficient ($\gamma = 1.4$), (c), (g) calibrated parameterization. Panels (d),(h) show the filtered and coarse-grained high-resolution simulation (1/32°). The lowest row shows the convergence of the calibration process, as assessed by two metrics included in the loss function (i), (j) and one metric excluded from the loss function (k). Blue markers show the ensemble median, and error bars show the 25% and 75% quantiles.

grained high-resolution model, see Figure 1b. First, we evaluate the efficiency of the calibration algorithm in a simple DG configuration and demonstrate its robustness to noise. Second, we apply the calibration algorithm to a substantially more expensive NW2 configuration and show that long-time statistics can be improved using only short simulations for calibration.

3.1. Simple Ocean Configuration

Our goal is to show that calibrating a sufficiently expressive parameterization, such as a neural network, can significantly improve the mean state and variability of the ocean model. For this purpose, we run the calibration algorithm described in Section 2.3 for 10 iterations. At each iteration, this algorithm runs an ensemble of coarse online simulations in the DG configuration described in Section 2.1, updates the weights and biases of the parameterization described in Section 2.2, attempting to minimize the loss function described in Section 2.4.

The coarse unparameterized model has a strong bias in the time-averaged sea surface height (SSH) with Root Mean Squared Error (RMSE) equal to 0.323 m, and too low variability (RMSE in SSH STD is 0.15 m), see Figures 2a and 2e.

We first isolate the effect of tuning only the scaling coefficient γ in front of the offline-trained eANN parameterization, while keeping the neural network's weights and biases unchanged. The optimal coefficient $\gamma = 1.4$ is

10.1029/2026GL123513

determined as described in Figure S3 in Supporting Information S1. The offline-trained eANN with tuned coefficients reduces biases in the mean state and variability by 41.8% and 29.3%, respectively, compared to the unparameterized model; see Figures 2b and 2f.

Second, we show the effect of jointly calibrating the scaling coefficient and the last layer of the neural network. The calibrated eANN parameterization reduces biases in the mean state and variability by 21.3% and 57.5%, respectively, compared to the offline-trained parameterization with optimal scaling coefficient (Figures 2c and 2g). The spatial pattern of the SSH STD is accurately reproduced in the parameterized model, both near the boundary (Longitude = 5) and in the boundary current extension (Longitude = 10). This is especially important for future applications of our approach, as reproducing the variability in WBC extension is a challenging problem in global ocean models (Juricke et al., 2020; Uchida et al., 2022).

The statistics of the simulations used for calibration are affected by noise arising from a relatively short time-averaging interval (10 years) compared to the evaluation runs (90 years); see Figure S4 in Supporting Information S1. Nevertheless, the calibration method is robust to noise, as evidenced by the monotonic decrease of the loss function (Figures 2i and 2j).

The calibration algorithm schedules an ensemble of 100 ocean simulations at every iteration. Thus, over 10 iterations, we have 1,000 simulations with different parameter vectors. Here, we describe how we chose one simulation shown in Figures 2c and 2g. Our parameterization is strongly constrained by physics and thus cannot compensate for all numerical model errors. As a result, we face not only parametric uncertainty but also model-form uncertainty (structural error) (Prévost et al., 2025; Shin & Howland, 2026; Williamson et al., 2017). In such a setting, an optimal set of parameter values depends considerably on the choice of the loss function, which determines which biases to prioritize compensating. In this work, we chose a simple loss function (see Section 2.4) after extensive experimentation. Due to structural errors, minimizing the loss function may cause other physical metrics to deteriorate. Here, we show that the RMSE in the covariance matrix of interfaces is improving over the first 3 iterations, then remains on a plateau until the 5th iteration, and after that starts growing (Figure 2(k)). Consequently, a further small improvement of the loss function is possible, but at the expense of generating strong unphysical modes of variability. This phenomenon in model tuning is known as overfitting (Williamson et al., 2017). We use the described validation metric to implement early stopping. That is, we consider the parameterization from the 5th iteration to be the best-performing one. Furthermore, we average the parameter vector over an ensemble to select a parameterization shown in panels (c,g) in Figure 2. We note that the need for validation here may stem from the optimization being underconstrained: we consider only one simple ocean configuration. Thus, seeking a single parameter set suitable for multiple flow regimes (Lopez-Gomez et al., 2022; Wagner et al., 2025) might eliminate the need for a validation procedure, as we show in the next section.

3.2. Ocean Configuration Featuring Multiple Flow Regimes

We now consider calibrating the eddy parameterization in the NW2 configuration, which features multiple flow regimes. We apply the calibration algorithm developed in the DG configuration with minimal adjustments. These include model initialization protocol and simulation length (described in Section 2.5), adjusting the scheduler step (δt) and ensemble size (n_e) (Section 2.4), and reducing the number of iterations to 4.

Below, we show the evaluation of coarse ocean models, which are integrated to the statistical equilibrium for 30,000 days from a state of rest and use the last 800 days to compute statistics.

The unparameterized coarse ocean model displays strong biases in the mean ocean state, see the position of fluid interfaces in the idealized Southern Ocean (60°S–40°S), and too low variability of the SSH (Figures 3a and 3b). Offline-trained parameterization allows only for a slight reduction in the mean state bias (RMSE in averaged interfaces drops by 13.4%), leaving variability effectively unchanged (RMSE in SSH STD reduces by 4.4%), see Figure 3c. The parameterization calibrated in the DG configuration is more skillful in reducing the model biases: the mean state was improved by 41.9%, and the variability was improved by 20.9% (Figure 3d). In these two parameterized models, we tuned the scaling coefficient (γ) to minimize the error on the presented metrics; see Figures S6 and S7 in Supporting Information S1. That way, we evaluate the effectiveness of the spatial pattern predicted by the parameterization, leaving the coefficient as a tunable parameter that often does not generalize across configurations and must be tuned manually (Balwada et al., 2025; Kamm et al., 2026; Perezhogin et al., 2024).

10.1029/2026GL123513

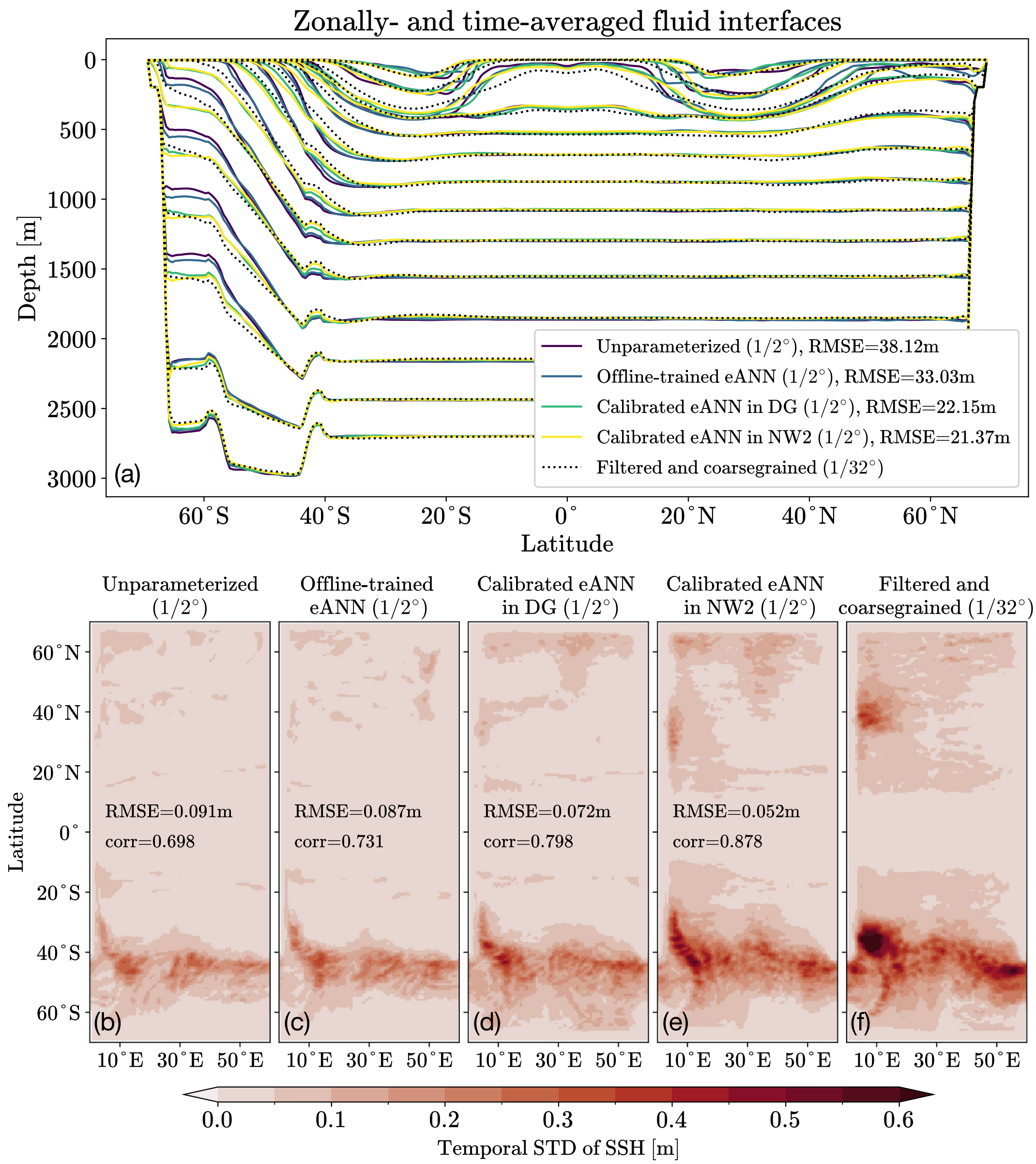

Figure 3. Evaluation of calibrated parameterizations in 30,000-day simulations in configuration NeverWorld2. (a) Zonally- and time-averaged vertical coordinate of internal fluid interfaces. Lower row shows temporal standard deviation of sea surface height for simulations with: (b) unparameterized model, (c) parameterization trained offline with manually adjusted scaling coefficient ($\gamma = 0.5$), (d) parameterization calibrated in Double Gyre with manually adjusted $\gamma = 0.6$, (e) parameterization calibrated in short 5-year runs in NeverWorld2 configuration, (f) filtered and coarse-grained high-resolution model.

We now evaluate the eddy parameterization calibrated in the NW2 configuration in short simulations. This parameterization does not require further adjustment of the scaling coefficient γ and can be effectively applied in long 30,000-day simulations. The ocean model with this parameterization achieves a similar error in the mean state compared to the parameterization calibrated in a simpler configuration (Figure 3a). However, it improves the error and the pattern correlation of the induced variability by 27.8% and 10%, respectively (Figure 3e). This demonstrates that the calibration algorithm effectively learns the spatial patterns specific to the NW2 configuration.

One of the major advantages of using the calibrated parameterizations is the improvement of the deep Southern Ocean stratification (Figure 3(a)). Some challenges remain. The calibrated parameterization does not sufficiently enhance variability in the western boundary current extension (30°N–50°N × 0°E−20°E) (Figure 3e) and has limited improvement in the isopycnal structure in the upper 500 m of tropical and Southern Ocean (Figure 3a). This occurs because the deep Southern Ocean dominates the loss function, thereby reducing the weight assigned to other dynamically important regions. The presence of multiple dynamical regimes therefore renders the optimization problem effectively overconstrained.

4. Discussion

In this study, we explored the effectiveness of a calibration approach in improving the mean state and variability of an idealized ocean model by adjusting parameters of the data-driven mesoscale eddy parameterization of Perezhogin et al. (2025).

We found that the EKI can significantly improve the errors in the mean state and variability at a challenging resolution of $1/2°$. The improvement amounts to factors of $2.2 - 3.3$ in the DG configuration and a factor of 1.7 in the NW2 configuration compared to the unparameterized model. Further, the improvement amounts to factors of $1.3 - 2.3$ in the DG configuration and $1.5 - 1.7$ in the NW2 configuration, compared to the offline-trained parameterization with the optimally tuned coefficient. The parameterization calibrated in a simple configuration (DG) can effectively reduce model biases in an unseen configuration (NW2) after manually adjusting the scaling coefficient. This demonstrates that the spatial patterns learned by the calibrated parameterization generalize well across ocean configurations. Nevertheless, we show that calibrating in the NW2 configuration directly further reduces the variability error by a factor of 1.4 compared to the parameterization calibrated in a simpler configuration.

The calibration algorithm modifies the offline-trained parameterization as follows (Figure S8 in Supporting Information S1). The predicted along-gradient fluxes exhibit markedly different spatial patterns, with greater skewness toward stronger upgradient fluxes (backscatter). Further, the cross-gradient fluxes were attenuated by $30 - 40\%$ without changing the spatial pattern, and the isotropic stress remained almost unchanged. A similar effect cannot be achieved by tuning a single coefficient in front of the offline-trained parameterization. The considerable change in the spatial pattern of the along-gradient fluxes can be explained as follows. First, the stronger upgradient fluxes compensate for the excessive numerical dissipation (biharmonic Smagorinsky) that was not observed during training. Second, the along-gradient fluxes, both up- and downgradient, produce strong responses that project onto the ocean model biases (Fox-Kemper & Menemenlis, 2008; Griffies & Hallberg, 2000; Jansen & Held, 2014).

We demonstrated that the EKI method is robust to noise in time-averaged statistics arising from the ocean's chaotic dynamics. Furthermore, the calibration can be performed efficiently in 5-year simulations without integrating to statistical equilibrium (approximately 100 years) if an accurate estimate of the initial ocean state is provided.

The advantages of the EKI algorithm include convergence in a few iterations and parallelism across ensemble members. The computational cost of the calibration in the DG configuration is only 1,000 CPU-core hours, with each simulation requiring 1 CPU core for 1 hr, and 1,000 simulations performed. The small computational cost allowed us to repeat the calibration 50–100 times while experimenting with the calibration protocol, including the loss function, noise model, parameters to be calibrated, and their prior distributions. Once the calibration problem is set up, tuning of the EKI parameters (scheduler step and ensemble size) is straightforward. The computational cost of the calibration in the NW2 configuration is much higher—50,000 CPU-core hours. This highlights the importance of assessing the calibration protocol in a simple, idealized configuration.

We found that efficient calibration can be achieved by perturbing a subset of the neural network's parameters by roughly 25% on average (Table S2 in Supporting Information S1). This emphasizes that the offline training plays a significant role in the calibration process, as it determines the weights and biases of the first layer of the neural network and provides initialization for calibrating the last layer.

We found that calibration in a configuration with only one flow regime tends to be underconstrained, whereas calibration in a configuration with multiple flow regimes tends to be overconstrained. In the former case, a careful validation protocol is essential to prevent overfitting (Williamson et al., 2017). In the latter case, improvements from calibration are largely concentrated in the ocean's most energetic region. Achieving improvements across multiple flow regimes ultimately requires introducing additional parameters into the calibration procedure, which can be done in various ways (Abernathey & Marshall, 2013; Hallberg, 2013; Liu et al., 2012; Prévost et al., 2025; Tuppi et al., 2023). However, this direction must be pursued with caution, as it may increase the risk of overfitting and reduce generalization capabilities (Maddison, 2026).

Our calibration protocol, which bypasses the integration to statistical equilibrium, represents a compromise between optimizing weather forecasting skill (Frezat et al., 2022; Kochkov et al., 2021, 2024; Maddison, 2026; Ouala et al., 2024) and optimizing the climate metrics in statistical equilibrium (Williamson et al., 2017; Mrozowska et al., 2025; O. R. Dunbar et al., 2021). The method can be extended to improve the mean state and variability of realistic ocean models using modern data-assimilation systems (Delworth et al., 2020) or reanalysis products (Lellouche et al., 2021), which provide estimates of the ocean state. We emphasize that short simulations (a few years) are not suitable for inferring parameters governing processes that evolve over millennial timescales, such as changes in background vertical diffusivity. However, the proposed approach can provide an initial estimate of parameters controlling processes that act on interannual timescales, particularly those influencing the air–sea interface.

Our calibration framework provides a systematic approach for improving parameterizations and reducing biases in global ocean models built on the MOM6 dynamical core, including those used by GFDL (Adcroft et al., 2019) and NCAR (Grooms et al., 2024).

Conflict of Interest

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

Availability Statement

The calibration script, eANN weights, and plots are available at Perezhogin (2026a). The simulation data and the offline analysis can be found at Perezhogin (2026b). Ensemble Kalman Inversion library is available at O. Dunbar et al. (2025).

Acknowledgments
This project is supported by Schmidt Sciences, as part of the M^{2}LInES project. This research was also supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. We would like to acknowledge high-performance computing support from the Derecho system, project code UOSC0036, provided by the NSF National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation. We are grateful to Dhruv Balwada and Fabrizio Falasca for their valuable comments and suggestions. We also thank the editor, Dr. Gregory Johnson, as well as two anonymous reviewers for their constructive comments that helped improve the quality and presentation of this paper.

References

Abernathey, R. P., & Marshall, J. (2013). Global surface eddy diffusivities derived from satellite altimetry. *Journal of Geophysical Research: Oceans*, *118*(2), 901–916. https://doi.org/10.1002/jgrc.20066

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., et al. (2019). The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3167–3211. https://doi.org/10.1029/2019MS001726

Andrejczuk, M., Cooper, F., Juricke, S., Palmer, T., Weisheimer, A., & Zanna, L. (2016). Oceanic stochastic parameterizations in a seasonal forecast system. *Monthly Weather Review*, *144*(5), 1867–1875. https://doi.org/10.1175/MWR-D-15-0245.1

Balwada, D., Perezhogin, P., Adcroft, A., & Zanna, L. (2025). Design and implementation of a data-driven parameterization for mesoscale thickness fluxes. *Authorea Preprints*. https://doi.org/10.22541/essoar.174835313.30541637/v1

Berner, J., Shutts, G., Leutbecher, M., & Palmer, T. (2009). A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ecmwf ensemble prediction system. *Journal of the Atmospheric Sciences*, *66*(3), 603–626. https://doi.org/10.1175/2008JAS2677.1

Chada, N. K., Chen, Y., & Sanz-Alonso, D. (2020). Iterative ensemble kalman methods: A unified perspective with some new variants. *arXiv preprint arXiv:2010.13299*. https://doi.org/10.48550/arXiv.2010.13299

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. https://doi.org/10.1029/2023GL105757

Chasnov, J. R. (1991). Simulation of the Kolmogorov inertial subrange using an improved subgrid model. *Physics of Fluids A: Fluid Dynamics*, *3*(1), 188–200. https://doi.org/10.1063/1.857878

Christopoulos, C., Lopez-Gomez, I., Beucler, T., Cohen, Y., Kawczynski, C., Dunbar, O. R., & Schneider, T. (2024). Online learning of entrainment closures in a hybrid machine learning parameterization. *Journal of Advances in Modeling Earth Systems*, *16*(11), e2024MS004485. https://doi.org/10.1029/2024MS004485

Connolly, A., Cheng, Y., Walters, R., Wang, R., Yu, R., & Gentine, P. (2025). Deep learning turbulence closures generalize best with physics-based methods. *Authorea Preprints*. https://doi.org/10.22541/essoar.173869578.80400701/v1

Cooper, F. C. (2017). Optimisation of an idealised primitive equation ocean model using stochastic parameterization. *Ocean Modelling*, *113*, 187–200. https://doi.org/10.1016/j.ocemod.2016.12.010

Cooper, F. C., & Zanna, L. (2015). Optimisation of an idealised ocean model, stochastic parameterisation of sub-grid eddies. *Ocean Modelling*, *88*, 38–53. https://doi.org/10.1016/j.ocemod.2014.12.014

DelSole, T., & Tippett, M. K. (2024). Tuning Earth system models without integrating to statistical equilibrium. *Journal of Advances in Modeling Earth Systems*, *16*(12), e2024MS004230. https://doi.org/10.1029/2024MS004230

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J.-H., Dunne, K. A., et al. (2020). SPEAR: The next generation GFDL modeling system for seasonal to multidecadal prediction and projection. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001895. https://doi.org/10.1029/2019MS001895

Dunbar, O., Constantinou, N. C., Lopez-Gomez, I., Iñigo, A. G., Bolewski, J., Bach, E. E., et al. (2025). CliMA/EnsembleKalmanProcesses.jl: v2.5.0 [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.16616326

Dunbar, O., Lopez-Gomez, I., Garbuno-Iñigo, A., Huang, D. Z., Bach, E., Wu, J.-l., et al. (2022a). Ensemblekalmanprocesses. jl: Derivative-free ensemble-based model calibration. *Journal of Open Source Software*, *7*(80), 4869. https://doi.org/10.21105/joss.04869

Dunbar, O. R., Duncan, A. B., Stuart, A. M., & Wolfram, M.-T. (2022b). Ensemble inference methods for models with noisy and expensive likelihoods. *SIAM Journal on Applied Dynamical Systems*, *21*(2), 1539–1572. https://doi.org/10.1137/21M1410853

Dunbar, O. R., Garbuno-Inigo, A., Schneider, T., & Stuart, A. M. (2021). Calibration and uncertainty quantification of convective parameters in an idealized GCM. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2020MS002454. https://doi.org/10.1029/2020MS002454

Fox-Kemper, B., & Menemenlis, D. (2008). Can large eddy simulation techniques improve mesoscale rich ocean models? *Washington DC American Geophysical Union Geophysical Monograph Series*, *177*, 319–337. https://doi.org/10.1029/177GM19

Frederiksen, J. S., & Davies, A. G. (1997). Eddy viscosity and stochastic backscatter parameterizations on the sphere for atmospheric circulation models. *Journal of the Atmospheric Sciences*, *54*(20), 2475–2492. https://doi.org/10.1175/1520-0469(1997)054⟨2475:EVASBP⟩2.0.CO;2

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. https://doi.org/10.1029/2022MS003124

Gjini, R., Morzfeld, M., Dunbar, O. R., & Schneider, T. (2025). The ensemble kalman inversion race. *arXiv preprint arXiv:2511.15853*. https://doi.org/10.48550/arXiv.2511.15853

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. https://doi.org/10.1175/1520-0493(2000)128⟨2935:BFWASL⟩2.0.CO;2

Griffies, S. M., Winton, M., Anderson, W. G., Benson, R., Delworth, T. L., Dufour, C. O., et al. (2015). Impacts on ocean heat from transient mesoscale eddies in a hierarchy of climate models. *Journal of Climate*, *28*(3), 952–977. https://doi.org/10.1175/JCLI-D-14-00353.1

Grooms, I., Agarwal, N., Marques, G. M., Pegion, P., & Yassin, H. (2024). The stochastic GM+ E closure: A framework for coupling stochastic backscatter with the gent and McWilliams parameterization. *Authorea Preprints*. https://doi.org/10.22541/essoar.172118408.85625257/v1

Guan, Y., Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2022a). 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. https://doi.org/10.1016/j.jcp.2022.111090

Guan, Y., Subel, A., Chattopadhyay, A., & Hassanzadeh, P. (2022b). Learning physics-constrained subgrid-scale closures in the small-data regime for stable and accurate LES. *Physica D: Nonlinear Phenomena*, *443*, 133568. https://doi.org/10.1016/j.physd.2022.133568

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534. https://doi.org/10.1029/2021MS002534

Hallberg, R. (2013). Using a resolution function to regulate parameterizations of Oceanic mesoscale eddy effects. *Ocean Modelling*, *72*, 92–103. https://doi.org/10.1016/j.ocemod.2013.08.007

Howland, M. F., Dunbar, O. R., & Schneider, T. (2022). Parameter uncertainty quantification in an idealized GCM with a seasonal cycle. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002735. https://doi.org/10.1029/2021MS002735

Huang, D. Z., Huang, J., Reich, S., & Stuart, A. M. (2022). Efficient derivative-free Bayesian inference for large-scale inverse problems. *Inverse Problems*, *38*(12), 125006. https://doi.org/10.1088/1361-6420/ac99fa

Iglesias, M., & Yang, Y. (2021). Adaptive regularisation for ensemble kalman inversion. *Inverse Problems*, *37*(2), 025008. https://doi.org/10.1088/1361-6420/abd29b

Iglesias, M. A., Law, K. J., & Stuart, A. M. (2013). Ensemble kalman methods for inverse problems. *Inverse Problems*, *29*(4), 045001. https://doi.org/10.1088/0266-5611/29/4/045001

Jansen, M. F., Adcroft, A., Khani, S., & Kong, H. (2019). Toward an energetically consistent, resolution aware parameterization of ocean mesoscale eddies. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2844–2860. https://doi.org/10.1029/2019MS001750

Jansen, M. F., & Held, I. M. (2014). Parameterizing subgrid-scale eddy effects using energetically consistent backscatter. *Ocean Modelling*, *80*, 36–48. https://doi.org/10.1016/j.ocemod.2014.06.002

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. https://doi.org/10.1029/2019MS001855

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. https://doi.org/10.1016/j.ocemod.2019.03.009

Juricke, S., Palmer, T. N., & Zanna, L. (2017). Stochastic subgrid-scale ocean mixing: Impacts on low-frequency variability. *Journal of Climate*, *30*(13), 4997–5019. https://doi.org/10.1175/JCLI-D-16-0539.1

Kamm, D., Deshayes, J., Perezhogin, P., Meunier, E., & Barge, A. (2026). Assessing data-driven eddy-parameterizations in an Atlantic sector model. *Authorea Preprints*. https://doi.org/10.22541/essoar.177100611.18240844/v1

Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., et al. (2021). Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200093. https://doi.org/10.1098/rsta.2020.0093

Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., & Hoyer, S. (2021). Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, *118*(21), e2101784118. https://doi.org/10.1073/pnas.2101784118

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066. https://doi.org/10.1038/s41586-024-07744-y

Kraichnan, R. H. (1976). Eddy viscosity in two and three dimensions. *Journal of the Atmospheric Sciences*, *33*(8), 1521–1536. https://doi.org/10.1175/1520-0469(1976)033⟨1521:EVITAT⟩2.0.CO;2

Lee, X. K., Ramadhan, A., Souza, A., Wagner, G. L., Silvestri, S., Marshall, J., & Ferrari, R. (2025). Nori: An ml-augmented ocean boundary layer parameterization. *arXiv preprint arXiv:2512.04452*. https://doi.org/10.48550/arXiv.2512.04452

Lellouche, J.-M., Eric, G., Romain, B.-B., Gilles, G., Angélique, M., Marie, D., & others (2021). The copernicus global 1/12 Oceanic and sea ice GLORYS12 reanalysis. *Frontiers in Earth Science*, *9*, 698876. https://doi.org/10.3389/feart.2021.698876

Liu, C., Köhl, A., & Stammer, D. (2012). Adjoint-based estimation of eddy-induced tracer mixing parameters in the global ocean. *Journal of Physical Oceanography*, *42*(7), 1186–1206. https://doi.org/10.1175/JPO-D-11-0162.1

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2022MS003105. https://doi.org/10.1029/2022MS003105

Maddison, J. R. (2026). Online learning in idealized ocean gyres. *Journal of Advances in Modeling Earth Systems*, *18*(2), e2024MS004883. https://doi.org/10.1029/2024MS004883

Marques, G. M., Loose, N., Yankovsky, E., Steinberg, J. M., Chang, C.-Y., Bhamidipati, N., et al. (2022). NeverWorld2: An idealized model hierarchy to investigate ocean mesoscale eddies across resolutions. *Geoscientific Model Development*, *15*(17), 6567–6579. https://doi.org/10.5194/gmd-15-6567-2022

Mrozowska, M. A., Avery, J., Stoustrup, A., Nuterman, R., Johnsen, C.-J., Thormann, A., & Jochum, M. (2025). Bayesian optimization with GPU acceleration for ocean models. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(3), e2024JH000517. https://doi.org/10.1029/2024JH000517

Ouala, S., Chapron, B., Collard, F., Gaultier, L., & Fablet, R. (2024). Online calibration of deep learning sub-models for hybrid numerical modeling systems. *Communications Physics*, *7*(1), 402. https://doi.org/10.1038/s42005-024-01880-7

Pahlavan, H. A., Hassanzadeh, P., & Alexander, M. J. (2024). Explainable offline-online training of neural networks for parameterizations: A 1D gravity wave-QBO testbed in the small-data regime. *Geophysical Research Letters*, *51*(2), e2023GL106324. https://doi.org/10.1029/2023GL106324

Pawar, S., San, O., Rasheed, A., & Vedula, P. (2023). Frame invariant neural network closures for kraichnan turbulence. *Physica A: Statistical Mechanics and its Applications*, *609*, 128327. https://doi.org/10.1016/j.physa.2022.128327

Perezhogin, P. (2026a). m2lines/calibration-MOM6-v1.2 [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.20262174

Perezhogin, P. (2026b). Simulation data for calibration in MOM6 [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.20261601

Perezhogin, P., Adcroft, A., & Zanna, L. (2025). Generalizable neural-network parameterization of mesoscale eddies in idealized and global ocean models. *Geophysical Research Letters*, *52*(19), e2025GL117046. https://doi.org/10.1029/2025GL117046

Perezhogin, P., Zanna, L., & Fernandez-Granda, C. (2023). Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003681. https://doi.org/10.1029/2023MS003681

Perezhogin, P., Zhang, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2024). A stable implementation of a data-driven scale-aware mesoscale parameterization. *Journal of Advances in Modeling Earth Systems*, *16*(10), e2023MS004104. https://doi.org/10.1029/2023MS004104

Prévost, L. M., Regayre, L. A., Johnson, J. S., McNeall, D., Milton, S., & Carslaw, K. S. (2025). Detection of structural deficiencies in a global aerosol model to explain limits in parametric uncertainty reduction. *EGUsphere*, *2025*, 1–65. https://doi.org/10.5194/egusphere-2025-4795

Pudig, M., Zhang, W., Smith, K. S., & Zanna, L. (2026). Parameterizing isopycnal mixing via kinetic energy backscatter in an eddy-permitting ocean model. *Journal of Advances in Modeling Earth Systems*, *18*(3), e2025MS005497. https://doi.org/10.1029/2025MS005497

Richter, I., & Tokinaga, H. (2020). An overview of the performance of CMIP6 models in the tropical Atlantic: Mean state, variability, and remote impacts. *Climate Dynamics*, *55*(9), 2579–2601. https://doi.org/10.1007/s00382-020-05409-w

Shin, E. Y., & Howland, M. F. (2026). Accelerated Bayesian calibration and uncertainty quantification of RANS turbulence model parameters for stratified atmospheric boundary layer flows. *Boundary-Layer Meteorology*, *192*(1), 3. https://doi.org/10.1007/s10546-025-00945-6

Storer, B. A., Buzzicotti, M., Khatri, H., Griffies, S. M., & Aluie, H. (2023). Global Cascade of kinetic energy in the ocean and the atmospheric imprint. *Science Advances*, *9*(51), eadi7420. https://doi.org/10.1126/sciadv.adi7420

Tuppi, L., Ekblom, M., Ollinaho, P., & Järvinen, H. (2023). Simultaneous optimization of 20 key parameters of the integrated forecasting system of ECMWF using OpenIFS. Part I: Effect on deterministic forecasts. *Monthly Weather Review*, *151*(6), 1325–1346. https://doi.org/10.1175/MWR-D-22-0209.1

Uchida, T., Le Sommer, J., Stern, C., Abernathey, R. P., Holdgraf, C., Albert, A., et al. (2022). Cloud-based framework for inter-comparing submesoscale-permitting realistic ocean models. *Geoscientific Model Development*, *15*(14), 5829–5856. https://doi.org/10.5194/gmd-15-5829-2022

Vallis, G. K. (2017). *Atmospheric and Oceanic fluid dynamics*. Cambridge University Press. https://doi.org/10.1017/9781107588417

Vernon, S., Bach, E., & Dunbar, O. R. (2025). Nesterov acceleration for ensemble kalman inversion and variants. *Journal of Computational Physics*, *535*, 114063. https://doi.org/10.1016/j.jcp.2025.114063

Wagner, G. L., Hillier, A., Constantinou, N. C., Silvestri, S., Souza, A., Burns, K. J., et al. (2025). Formulation and calibration of catke, a one-equation parameterization for microscale ocean mixing. *Journal of Advances in Modeling Earth Systems*, *17*(4), e2024MS004522. https://doi.org/10.1029/2024MS004522

Wang, C., Zhang, L., Lee, S.-K., Wu, L., & Mechoso, C. R. (2014). A global perspective on CMIP5 climate model biases. *Nature Climate Change*, *4*(3), 201–205. https://doi.org/10.1038/nclimate2118

Williamson, D. B., Blaker, A. T., & Sinha, B. (2017). Tuning without over-tuning: Parametric uncertainty quantification for the nemo ocean model. *Geoscientific Model Development*, *10*(4), 1789–1816. https://doi.org/10.5194/gmd-10-1789-2017

Yan, F. E., Frezat, H., Le Sommer, J., Mak, J., & Otness, K. (2025). Adjoint-based online learning of two-layer quasi-geostrophic baroclinic turbulence. *Journal of Advances in Modeling Earth Systems*, *17*(7), e2024MS004857. https://doi.org/10.1029/2024MS004857

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. https://doi.org/10.1029/2023MS004093

Yankovsky, E., Zanna, L., & Smith, K. S. (2022). Influences of mesoscale ocean eddies on flow vertical structure in a resolution-based model hierarchy. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003203. https://doi.org/10.1029/2022MS003203

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

Zasko, G. V., Glazunov, A. V., Mortikov, E. V., Nechepurenko, Y. M., & Perezhogin, P. A. (2023). Optimal energy growth in stably stratified turbulent couette flow. *Boundary-Layer Meteorology*, *187*(1–2), 395–421. https://doi.org/10.1007/s10546-022-00744-3

Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2023). Implementation and evaluation of a machine learned mesoscale eddy parameterization into a numerical ocean circulation model. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003697. https://doi.org/10.1029/2023MS003697

Zhang, W., Kuo, Y.-H., Silvestri, S., Adcroft, A., Hallberg, R. W., & Griffies, S. M. (2025). A WENO finite-volume scheme for the evolution of potential vorticity in isopycnal ocean models. *Authorea Preprints*. https://doi.org/10.22541/essoar.175380391.18723979/v1

References From the Supporting Information

Bachman, S. D. (2024). An eigenvalue-based framework for constraining anisotropic eddy viscosity. *Journal of Advances in Modeling Earth Systems*, *16*(8), e2024MS004375. https://doi.org/10.1029/2024ms004375

Cesa, G., Lang, L., & Weiler, M. (2022). A program to build E(N)-Equivariant steerable CNNs. In *International conference on learning representations (iclr)*.

10.1029/2026GL123513

Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2023). Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. *Journal of Advances in Modeling Earth Systems*, *15*(1), e2022MS003258. https://doi.org/10.1029/2022MS003258

Smith, R. D., & McWilliams, J. C. (2003). Anisotropic horizontal viscosity for ocean models. *Ocean Modelling*, *5*(2), 129–156. https://doi.org/10.1016/s1463-5003(02)00016-1

Weiler, M., & Cesa, G. (2019). General E(2)-Equivariant steerable CNNs. In *Conference on neural information processing systems (neurips)*.