JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2025MS004991

A Data-Driven Approach for Parameterizing Ocean Submesoscale Buoyancy Fluxes

Abigail Bodner[1], **Dhruv Balwada**[2], **and Laure Zanna**[1,3]

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

Key Points:

- A data-driven parameterization for mixed layer vertical buoyancy fluxes is developed using a Convolutional Neural Network (CNN)
- The CNN demonstrates high offline skill over a wide range of dynamical regimes and filter scales
- The large scale strain field, currently missing from oceanic submesoscale parameterizations, is identified as a significant input feature

Supporting Information:

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

Correspondence to:

A. Bodner,
abodner@mit.edu

Citation:

Bodner, A., Balwada, D., & Zanna, L. (2025). A data-driven approach for parameterizing ocean submesoscale buoyancy fluxes. *Journal of Advances in Modeling Earth Systems*, *17*, e2025MS004991. https://doi.org/10.1029/2025MS004991

Received 2 FEB 2025
Accepted 25 SEP 2025

Author Contributions:

Conceptualization: Abigail Bodner
Data curation: Abigail Bodner, Dhruv Balwada
Formal analysis: Abigail Bodner
Funding acquisition: Laure Zanna
Investigation: Abigail Bodner
Methodology: Abigail Bodner, Dhruv Balwada
Resources: Abigail Bodner
Software: Abigail Bodner, Dhruv Balwada
Supervision: Laure Zanna
Validation: Abigail Bodner, Dhruv Balwada, Laure Zanna

Abstract Parameterizations of $O(1-10)$km submesoscale mixed layer instabilities in General Circulation Models (GCMs) represent the effects of unresolved vertical buoyancy fluxes (VBF) in the ocean mixed layer. These submesoscale flows interact non-linearly with mesoscale and boundary layer turbulence, and it is challenging to account for all the relevant processes in physics-based parameterizations. In this work, we present a data-driven approach for the submesoscale parameterization, that relies on a Convolutional Neural Network (CNN) trained to predict mixed layer VBF as a function of relevant large-scale variables. The data used for training is given from 12 regions sampled from the global high-resolution MITgcm-LLC4320 simulation. When compared with the baseline of a submesoscale physics-based parameterization, the CNN demonstrates high offline skill across all regions, seasons, and filter scales tested in this study. During seasons when submesoscales are most active, which generally corresponds to winter and spring months, we find that the CNN prediction skill tends to be lower than in summer months. The CNN exhibits a dependency on the large scale strain field, a variable closely related to frontogenesis, which is currently missing from the submesoscale parameterizations in GCMs.

Plain Language Summary Upper ocean turbulence plays a key role in regulating the exchange of energy and heat between the ocean and atmosphere. Much of this turbulence occurs at submesoscales-spatial scales on the order of 1–10 km—which are too small to be directly resolved by climate models. Traditionally, these unresolved processes are estimated using parameterizations based on coarse model output. In this work, we take a different approach by using machine learning to predict the effects of submesoscale turbulence on the ocean system. Our method shows strong performance across a range of scales, locations, and seasons, offering a promising path toward reducing biases in climate models and improving their representation of upper ocean dynamics.

1. Introduction

General Circulation Models (GCMs) and future climate change projections are notoriously sensitive to parameterizations of unresolved phenomena near the ocean-atmosphere interface (IPCC, 2019; IPCC, 2021). Of particular importance is the ocean mixed layer, where turbulence modulates the transfer of properties—such as heat, momentum, and carbon—between the atmosphere and ocean interior (e.g., Bopp et al., 2015; Frankignoul & Hasselmann, 1977; Su et al., 2020). Turbulence in the ocean mixed layer spans a wide range of scales, from the $O(100)$km mesoscales, to $O(1-10)$km submesoscales, to $O(1-100)$m boundary layer turbulence, and all the way down to the molecular scales. A sensitive dynamical interplay between turbulence across all relevant scales sets the stratification in the upper ocean, yet models often struggle to capture these interactions, resulting in biases in upper ocean structure and heat exchange (Treguier et al., 2023). As opposed to mixing and homogenization dominated by boundary layer turbulence, submesoscale flows play a particularly important role in contributing to vertical transport in the ocean mixed layer primarily by shoaling, or restratifying, the mixed layer (Boccaletti et al., 2007; McWilliams, 2016; Taylor & Thompson, 2023).

The restratification effect is at leading order a result of instabilities formed along mixed layer fronts composed of sharp density gradients (Fox-Kemper et al., 2008; Gula et al., 2022). One of the primary submesoscale instabilities, known as mixed layer instabilities, produces vertical buoyancy fluxes (VBF) by slumping the fronts and restratifying the mixed layer. The effect of submesoscale restratification cannot be resolved in many GCMs, and is currently parameterized by the Mixed Layer Eddy parameterization (hereafter MLE, Fox-Kemper

Visualization: Abigail Bodner
Writing – original draft: Abigail Bodner
Writing – review & editing: Abigail Bodner, Dhruv Balwada, Laure Zanna

et al., 2011). Recent advances in submesoscale parameterization development propose new relationships between MLE and large scale properties of the mesoscale field (e.g., J. Zhang et al., 2023) and boundary layer turbulence (e.g., Bodner et al., 2023), yet these new approaches still struggle to capture the full range of complexity (Ajayi et al., 2021; Bachman et al., 2017; Callies & Ferrari, 2018; Lapeyre et al., 2006; Mahadevan et al., 2010).

Data-driven methods are emerging as powerful tools, with the ability to capture highly complex relationships between variables in turbulent flows. Advances in machine learning based parameterizations have yielded promising results for subgrid closures such as for ocean mesoscale momentum fluxes (Bolton & Zanna, 2019; Guillaumin & Zanna, 2021; Perezhogin et al., 2023; Zanna & Bolton, 2020), ocean boundary layer mixing (Sane et al., 2023; Souza et al., 2020), and atmospheric boundary layer mixing (e.g., Shamekh et al., 2023; Wang et al., 2022; Yuval & O'Gorman, 2020). Numerous examples for other machine learning applications exist both in the atmosphere and the ocean for inference of flow patterns and structures from data (e.g., Chattopadhyay et al., 2020; Dagon et al., 2022; Xiao et al., 2023; Zhu et al., 2023).

Here, we introduce a data-driven approach for parameterizing submesoscale-induced VBF in the ocean mixed layer. We train a Convolutional Neural Network (CNN) using high-resolution simulation data, with the goal of learning an improved functional relationship between mixed layer VBF and the large-scale variables that help set it. The data used to train and test the CNN is sampled from the MITgcm-LLC4320 ocean model (hereafter LLC4320, Menemenlis et al., 2021), which simulated the global ocean at a resolution of $1/48°$. The LLC4320 output has been widely studied for submesoscale applications, which cumulatively have demonstrated that submesoscale energetics and dynamics are captured relatively well down to its effective resolution (e.g., Gallmeier et al., 2023; Rocha et al., 2016; Su et al., 2018). In this paper, we describe the processing of the LLC4320 data and CNN architecture in Section 2. Results and sensitivity tests of the CNN prediction on unseen data are presented and compared with the baseline of the MLE parameterization in Section 3. In Section 4, we apply two complimentary methods to explain the relationship learned by the CNN and the mixed layer VBF. Discussion and concluding remarks are given in Section 5.

2. Data and Methods

2.1. Processing the LLC4320

The LLC4320 is a $1/48^o$ Massachusetts Institute of Technology general circulation model, named after its Latitude-Longitude polar Cap (LLC) grid with 4,320 points on each of the 13 tiles. The LLC4320 is initialized from the Estimating the Circulation and Climate of the Ocean, Phase II project, and is forced at the surface by atmospheric reanalysis, at 6 hourly temporal resolution. Model output from a total of 14 months is available at hourly frequency from September 2011 to November 2012 (Forget et al., 2015; Menemenlis et al., 2008, 2021). Before computing the CNN input and output variables, a common processing procedure was applied to the LLC4320 data, as described below.

Since the primary goal of our work is to parameterize the impact of submesoscale processes on mixed layer stratification, we focus on diagnosing subgrid VBF (Fox-Kemper et al., 2008). The large-scale fields, which may be resolved in a coarse-simulation, and subgrid impacts, which need to be parameterized, are defined with the help of filters. This is similar to approaches commonly used in the large-eddy simulation literature (Sagaut, 2005). First, to reduce the data volume, we averaged all LLC4320 variables over periods of 12 hr, effectively time-filtering the fastest motions. We note that these fast-varying motions, primarily composed of tides and internal waves, but also tied to atmospheric forcing, may impact the VBF at leading order (Balwada et al., 2018; Richards et al., 2021; Su et al., 2020; Uchida et al., 2019). However, the choice for the temporal filter was made with a focus on characteristic submesoscale variability and does not attempt to account for these high-frequency effects on the VBF. Next, a top-hat or coarsening spatial filter (denoted by $\overline{\cdot}$) was applied to decompose the simulation variables into large-scale and subgrid components. This allows us to define the subgrid VBF $\left(\overline{F^V}\right)$ as:

$$\overline{F^V} = \overline{wb} - \overline{w}\,\overline{b}, \tag{1}$$

where $\overline{w}$ and $\overline{b}$ correspond to the large-scale variables that can be resolved on a coarse-grid, and $\overline{wb}$ is the coarsened flux that is resolved in a high-resolution simulation. We used filter scales of $1°, 1/2°, 1/4°, 1/8°, 1/12°$, which are defined by the width of the coarsening box, applied by averaging over a fixed number of grid points in the original

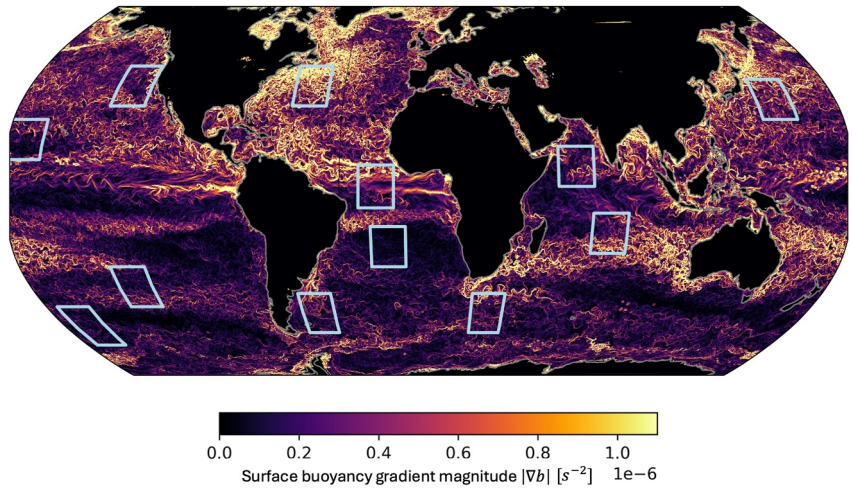

Figure 1. Snapshot of surface horizontal buoyancy gradient magnitude $|\nabla b|\ [s^{-2}]$ given by the global LLC4320 simulation coarsened to 1/4°. Buoyancy gradients are a key contributor to the characteristics of submesoscale flows, and their properties vary significantly between regions and season (Figure 2), motivating the choice of sampled regions used in this study (light blue boxes, exact coordinates listed in Table S1 in Supporting Information S1).

LLC4320 grid. As an example for the 1/4° spatial filter, averages are taken over 12 × 12 grid points of the 1/48° simulation grid.

For simplicity, we restrict our approach to depth-averaged mixed layer properties, and to this end, all 3D variables are averaged over the mixed layer depth (denoted by superscript z hereafter). Although recent work has suggested that submesoscale VBF can extend below the mixed layer (e.g., Siegelman et al., 2020), here we chose to remain close to the formulation of the MLE parameterization, composed of a depth-independent amplitude and a vertical structure function that determines the shape of the parameterization over the mixed layer (Equation S3 in Supporting Information S1). Here, the mixed layer depth, H_{ML}, is defined as the depth at which the potential density anomaly, ϱ, increased by 0.03 kg m^{-3} from its value at 10 m depth (de Boyer Montégut et al., 2004). ϱ is computed from the LLC4320 outputs of potential temperature and salinity fields, with reference pressure of 0 dbar and $\rho_0 = 1000$ kg m^{-3}.

To ensure that the training data represented a diverse set of dynamics, we included a mix of regions with strong and weak variability (e.g., Torres et al., 2018) as shown by blue boxes in Figure 1 (exact coordinates provided in Table S1 in Supporting Information S1). In the following section we describe how the LLC4320 data is processed for each of these regions to diagnose the VBF (CNN output) and a variety of inputs (Table 1) in preparation for the CNN training.

2.2. Input and Output Features

The subgrid quantity we are parameterizing using the CNN is the depth-averaged mixed layer VBF $\left(\overline{F^{V}}^{z}\right)$. This quantity is formally denoted as,

$$\mathbb{Y}_{wb} := \overline{F^{V}}^{z} = \overline{wb}^{z} - \overline{w}^{z}\overline{b}^{z}. \tag{2}$$

VBF in the mixed layer is largely a result of submesoscale flows (Boccaletti et al., 2007). This can be seen in the maximum cross-spectrum of w and b, analogous to maximum VBF, which is found to be predominantly in the submesoscale range and confined to the mixed layer (Figure 2a). However, variability across scales can differ between the different regions, and the filter scale choice (illustrated by the gray lines in Figure 2b) will impact the properties of the large scale CNN inputs and subgrid flux output. We have included all regions to gain a variety of dynamical regimes in our training data, but test the performance of the CNN over the different filter scales, and selected unseen regions and parts of the timeseries in Section 3.

Table 1
Features Used in the Convolutional Neural Network Method

	Variable	Symbol and definition
Input, $\mathbb{X}$	Buoyancy gradient magnitude	$\lvert\overline{\nabla b}^z\rvert = \sqrt{\left(\overline{b_x}^z\right)^2 + \left(\overline{b_y}^z\right)^2}$
	Coriolis parameter	$\overline{f}$
	Mixed layer depth	$\overline{H_{ML}}$
	Surface heat flux	$\overline{Q}$
	Surface wind stress magnitude	$\lvert\overline{\tau}\rvert = \sqrt{\overline{\tau_x}^2 + \overline{\tau_y}^2}$
	Boundary layer depth	$\overline{H_{BL}}$
	Strain magnitude	$\overline{\sigma}^z = \sqrt{(\overline{u_x}^z - \overline{v_y}^z)^2 + (\overline{v_x}^z + \overline{u_y}^z)^2}$
	Vertical vorticity	$\overline{\zeta}^z = \overline{v_x}^z - \overline{u_y}^z$
	Horizontal divergence	$\overline{\delta}^z = \overline{u_x}^z + \overline{v_y}^z$
Output, $\mathbb{Y}_{wb}$	Subgrid vertical buoyancy flux	$\mathbb{Y}_{wb} := \overline{wb}^z - \overline{w}^z\overline{b}^z$

Note. Overbar represents the top-hat spatial filter and superscript z represents a depth averaging over the mixed layer depth applied as part of the processing of the LLC4320 data (described in Section 2.2).

We choose to include input features that are correlated (Figure S4 in Supporting Information S1), or have known analytical relationships, with submesoscale VBF. We leverage the physical relevance demonstrated by variables that appear in the Fox-Kemper et al. (2011) and Bodner et al. (2023) versions of the MLE parameterization, as well as correlated large-scale velocity derivatives (e.g., Barkan et al., 2019; Balwada et al., 2021; J. Zhang et al., 2023). The input features (Table 1) consist of the depth-averaged horizontal buoyancy gradient magnitude, $\lvert\overline{\nabla b}^z\rvert$, where buoyancy is defined as $b = -g\varrho/\rho_0$, and $g = 9.81$ m/s^2 the gravity acceleration; Coriolis parameter, $\overline{f}$; mixed layer depth, $\overline{H_{ML}}$; surface heat flux, $\overline{Q}$; surface wind stress magnitude, $\lvert\overline{\tau}\rvert = \sqrt{\overline{\tau_x}^2 + \overline{\tau_y}^2}$; boundary layer depth, $\overline{H_{BL}}$, an output of the LLC4320 computed from the Richardson number criteria with the critical value of 0.3 (K-profile Parameterization, Large et al., 1994); depth-averaged strain magnitude, $\overline{\sigma}^z = \sqrt{(\overline{u_x}^z - \overline{v_y}^z)^2 + (\overline{v_x}^z + \overline{u_y}^z)^2}$; depth-averaged vertical vorticity, $\overline{\zeta}^z = \overline{v_x}^z - \overline{u_y}^z$; depth-averaged horizontal divergence, $\overline{\delta}^z = \overline{u_x}^z + \overline{v_y}^z$. Note that velocities (u, v, w) and wind stresses (τ_x, τ_y) are all interpolated to collocate with the tracer grid.

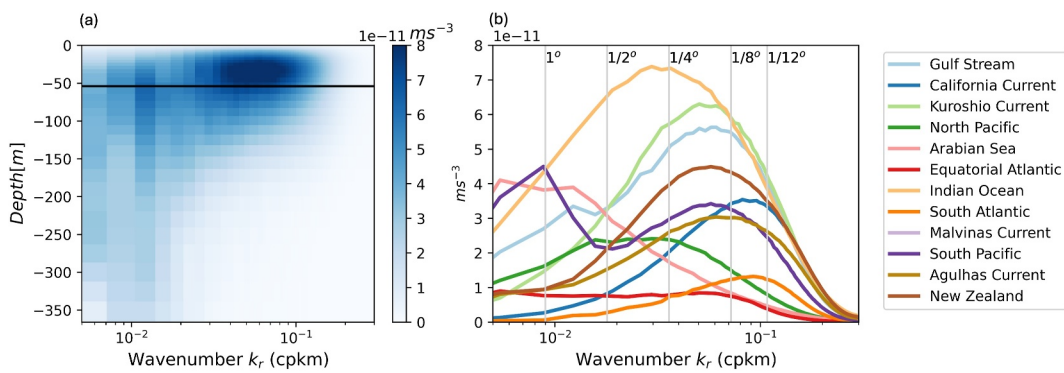

Figure 2. Isotropic cross-spectrum of w and b in variance-preserving form, averaged over the entire LLC4230 simulation duration (14 months): (a) Example of the depth varying cross-spectrum in the Gulf Stream region, illustrating that the vertical buoyancy fluxes is concentrated in the small scales and within the average mixed layer (black horizontal line). (b) Cross-spectrum averaged over the mixed layer depth for all regions. Vertical gray lines mark the filter scales used in this study with respect to the cross-spectrum variability.

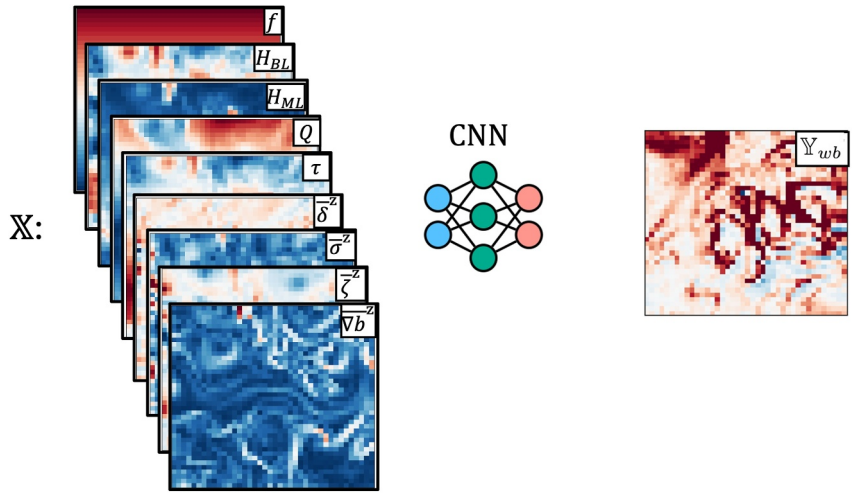

Figure 3. Schematic of the $1/4°$ Convolutional Neural Network (CNN) method with nine input features, $\mathbb{X}$, and one output, $\mathbb{Y}_{wb}$ (corresponding to Table 1). The CNN architecture is described in Section 2.3.

Formally, we define our nine input features as,

$$\mathbb{X} := \left(\left|\overline{\nabla b}^z\right|, \overline{f}, \overline{H_{ML}}, \overline{Q}, |\overline{\tau}|, \overline{H_{BL}}, \overline{\sigma}^z, \overline{\zeta}^z, \overline{\delta}^z\right), \tag{3}$$

and a single output as,

$$\mathbb{Y}_{wb} := \overline{wb}^z - \overline{w}^z \overline{b}^z \tag{4}$$

The CNN is trained to predict the subgrid fluxes as a function of the large scale variables, such that $S(\mathbb{X}) = \hat{\mathbb{Y}}_{wb}$, where S represents the CNN and $\hat{\mathbb{Y}}_{wb}$ its prediction. As explained in the following section, we train a separate CNN for each filter scale experiment by minimizing the Mean Squared Error (MSE) loss between $\mathbb{Y}_{wb}$ and $\hat{\mathbb{Y}}_{wb}$, where both $\mathbb{Y}_{wb}$, $\hat{\mathbb{Y}}_{wb}$ and the input, $\mathbb{X}$, are defined according to the filter scale choice. Figure 3 illustrates a schematic of the CNN inputs and output in the $1/4°$ filter scale experiment. We further discuss the sensitivity of the CNN to the choice of filter scale in Section 3.

2.3. CNN Architecture and Training

Each experiment, designed with a given filter scale, is trained and tested independently, but all CNNs shared a common architecture. We use a CNN architecture for regression inspired by applications for mesoscale eddy parameterizations (Bolton & Zanna, 2019; Guillaumin & Zanna, 2021; Perezhogin et al., 2023). A hyperparameter sweep over the number of hidden layers, kernel size, learning rate, and weight decay, was used to find the best performing CNN. The CNN is trained over 100 epochs while minimizing the MSE loss (shown in Figure S1 in Supporting Information S1). The hyperparameters were tuned against the $1/4°$ filter scale experiment, and remained fixed throughout all other experiments. Results presented here are based on a CNN with a kernel size of 5×5 in the first layer, followed by 7 hidden convolutional layers with kernel size of 3×3, a learning rate of 2×10^{-4}, and weight decay of 0.02. The total number of learnable parameters is approximately 300,000.

Prior to applying the CNN, all variables listed in Table 1 are normalized by a single mean and standard deviation computed over all regions (example shown in Figure S2 in Supporting Information S1). To train the CNN, we randomly select 80% of the $\sim 10,000$ samples given from all regions combined. The remaining 20% is left unseen by the CNN, and is used to test the CNN prediction. Results are compared in the following section with the target LLC4320 data and the Bodner et al. (2023) version of the MLE parameterization, which is used here as a baseline (referred to as the "MLE parameterization" hereafter). In Sections 3.1 and 3.2, we examine other split choices between the train and test data sets to include subsets of the sampled regions or timeseries, respectively.

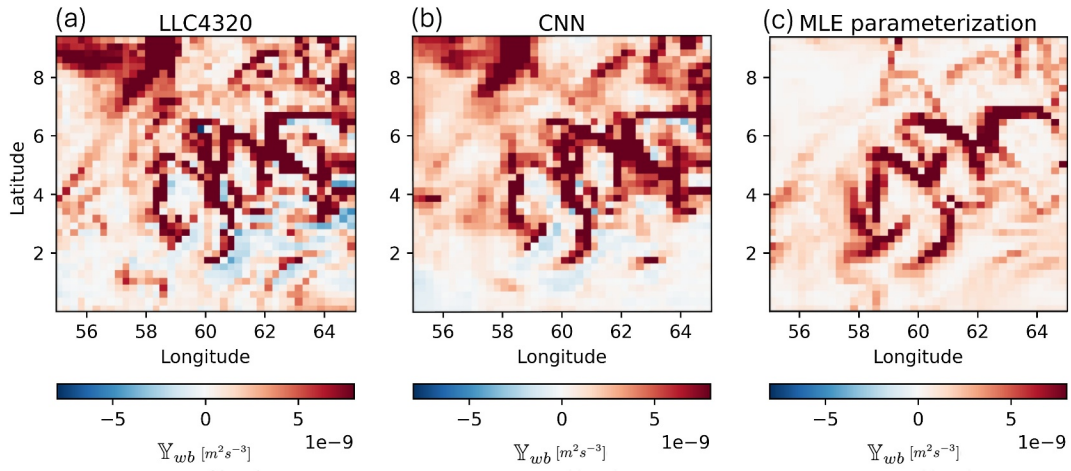

Figure 4. Snapshot taken from the Arabian Sea region of the depth-averaged subgrid vertical buoyancy fluxes $\mathbb{Y}_{wb}\ [\mathrm{m^2 s^{-3}}]$ given by (a) the LLC4320, (b) Convolutional Neural Network prediction in physical space, and (c) the Bodner et al. (2023) version of the Mixed Layer Eddy parameterization. A $1/4°$ filter scale is applied here. An example of (a) with filter scales of $1/12°, 1/8°, 1/2°, 1°$ is shown in Figure S3 in Supporting Information S1.

3. CNN Prediction of Subgrid Submesoscale Fluxes

Once trained, the CNN has learned a functional mapping, $S(\mathbb{X})$, between the input features, $\mathbb{X}$, and subgrid mixed layer VBF, $\mathbb{Y}_{wb}$. In this section, we examine the extent to which the CNN can make skillful predictions on data that was not included in the training process. For this purpose, we compare the CNN prediction with the target LLC4320 data held out from training, and test whether the CNN improves on the baseline given by the Bodner et al. (2023) version of the MLE parameterization.

Illustrated by an example from the $1/4°$ filter scale experiment (Figure 4), the CNN is able to capture much of the fine-scale structure and sign of the subgrid VBF. The majority of the fluxes are positive, which is the bulk restratification effect inferred by the MLE parameterization. The negative fluxes exhibited in the LLC4320, and captured by the CNN but not the MLE parameterization, are another indication that the CNN is able to capture finer signatures of submesoscale VBF (Torres et al., 2025). This can be further seen in the joint histogram of the VBF given by the target LLC4320 and those predicted by the CNN and MLE parameterizations (Figure 5). The joint histograms are computed over the entire unseen test data set, which provides a comparison over several orders of magnitude of the VBF. In the case of positive fluxes, the CNN prediction remains close to the LLC4320 VBF, as can be seen by the alignment along the one-to-one gray line, an improvement on the MLE parameterization which deviates from the LLC4320 VBF. For the negative fluxes, the one-to-one alignment is less pronounced, likely due to the significantly smaller number of negative samples seen by the CNN (less than 5% of the total samples). However, the ability of the CNN to predict of negative fluxes is still an improvement on the MLE parameterization, which does not include negative fluxes by construction.

To test whether the CNN also has skill in predicting bulk effects, such as is inferred by the MLE parameterization, the CNN predictions on unseen test data are averaged over each month to form a seasonal cycle, and is compared with the equivalent for the MLE parameterization and LLC4320 target data. We find that in all regions, the CNN prediction captures the seasonality and bulk effects of the LLC4320 data, and outperforms the MLE parameterization, particularly where fluxes appear to be strongest during the winter and spring months where the LLC4320 VBF can be as large as three times the MLE prediction (Figure 6 for the $1/4°$ filter scale experiment and more quantitatively in the analysis described below).

Prediction skill of the CNN and MLE parameterization for all filter scales are quantified in terms of R^2 values relative to the LLC4320 target data. The R^2 value provides a useful metric for the average skill by incorporating the MSE of the parameterization prediction compared with the target data (calculation is described in Equation S1 in Supporting Information S1). We find that the CNN R^2 remain at a value of at least 0.2 higher than that of the MLE in all filter scale experiments and in all regions (Figure 7). As the magnitude of $\mathbb{Y}_{wb}$ varies spatially

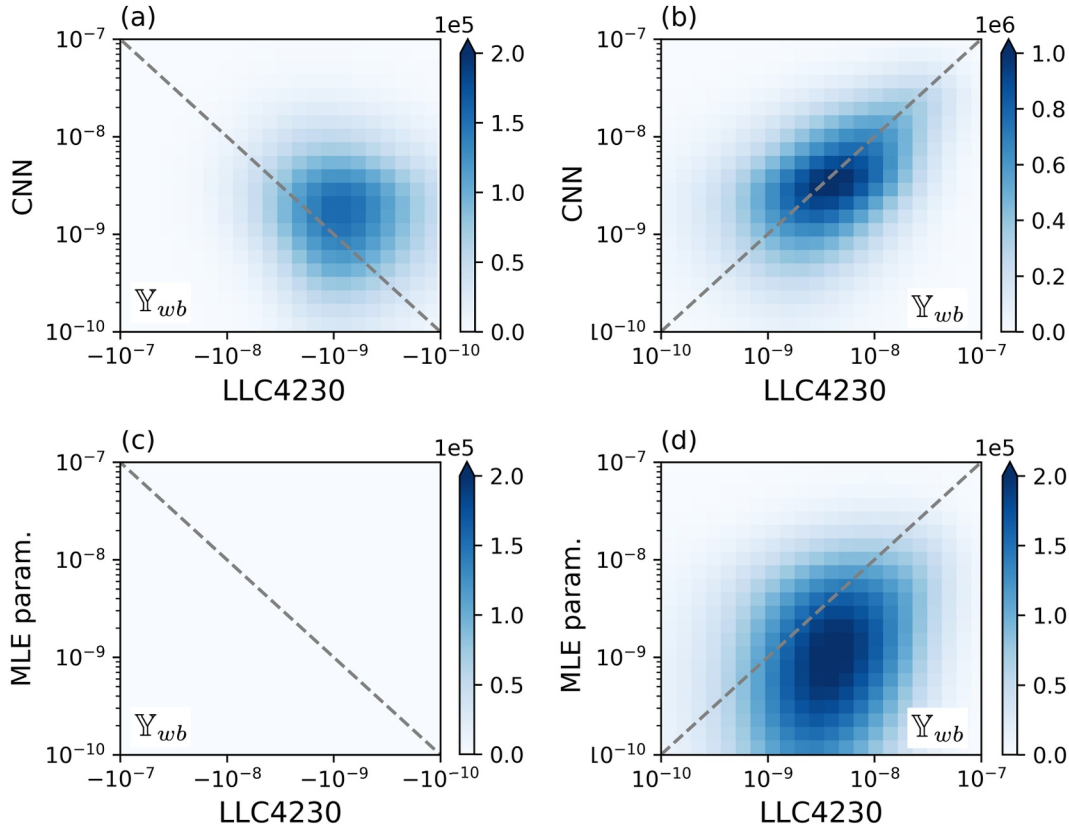

Figure 5. Joint histogram of the 1/4° $\mathbb{Y}_{wb}$ $[\mathrm{m^2 s^{-3}}]$: (a, b) Convolutional Neural Network (CNN) prediction and LLC4320 data, and (c, d) Mixed Layer Eddy parameterization and LLC4320 data. Panels (a, c) correspond to negative fluxes and (b, d) to positive fluxes. The CNN predictions remain close to the target LLC4320 in both positive and negative values of $\mathbb{Y}_{wb}$. Note that the colorbar in (b) is an order of magnitude larger than the others due to the high concentration along the diagonal.

(Figure 2), this impacts the predicted output of the CNN in the different regions. The largest filter scales tend to have skill nearing an R^2 values of 1, which then decreases as the filter scale becomes smaller. In the large filter scale experiments, the fields tend to be smoother, as much of the subgrid spatial variability is averaged out, thus presenting an easier learning problem for the CNN. The CNN prediction skill is found to be especially sensitive in the small filter scale experiments, where it performs well in some regions, e.g. in the California Current where all R^2 values are above 0.8, but less so in others, e.g. in the Indian Ocean region where the R^2 values in the small filter scale experiments are below 0.3. In the regions that exhibit large sensitivity to the filter scale such as the Kuroshio Current, Indian Ocean, and South Pacific, the skill of the MLE parameterization drops significantly as well.

It is worth noting that the CNN prediction skill can be sensitive to the initialization weights even when training identical experiment configurations (e.g., Otness et al., 2023). However, we find that for each given filter scale experiment, the sensitivity to the initialization weights is smaller than an R^2 value of 0.1 (Figure S5 in Supporting Information S1), indicating that the large range of R^2 values displayed in Figure 9 reflects the sensitivity due to regional variability rather than properties of the CNN.

As submesoscale seasonality greatly impacts the variability of VBF (demonstrated in Figure 6), we examine the skill (in terms of R^2 values) of the CNN and MLE parameterization averaged only over winter and summer months (Figure 8). During winter, when mixed layer VBF tend to be stronger, the skill of both the CNN and MLE parameterization in the smaller filter scale experiments drops compared with its equivalent in summer. This is particularly true for regions where fluxes are very strong during winter, such as in the Kuroshio Current, where the 1/8° and 1/12° filter scale experiments shows no skill (negative R^2) during winter compared with an R^2 value above 0.8 in summer. Similarly, the Gulf Stream, Agulhas Current, Malvinas Current, and the Southern Ocean near New Zealand all exhibit a drop in skill in both the MLE and CNN with small filter scales. In other regions

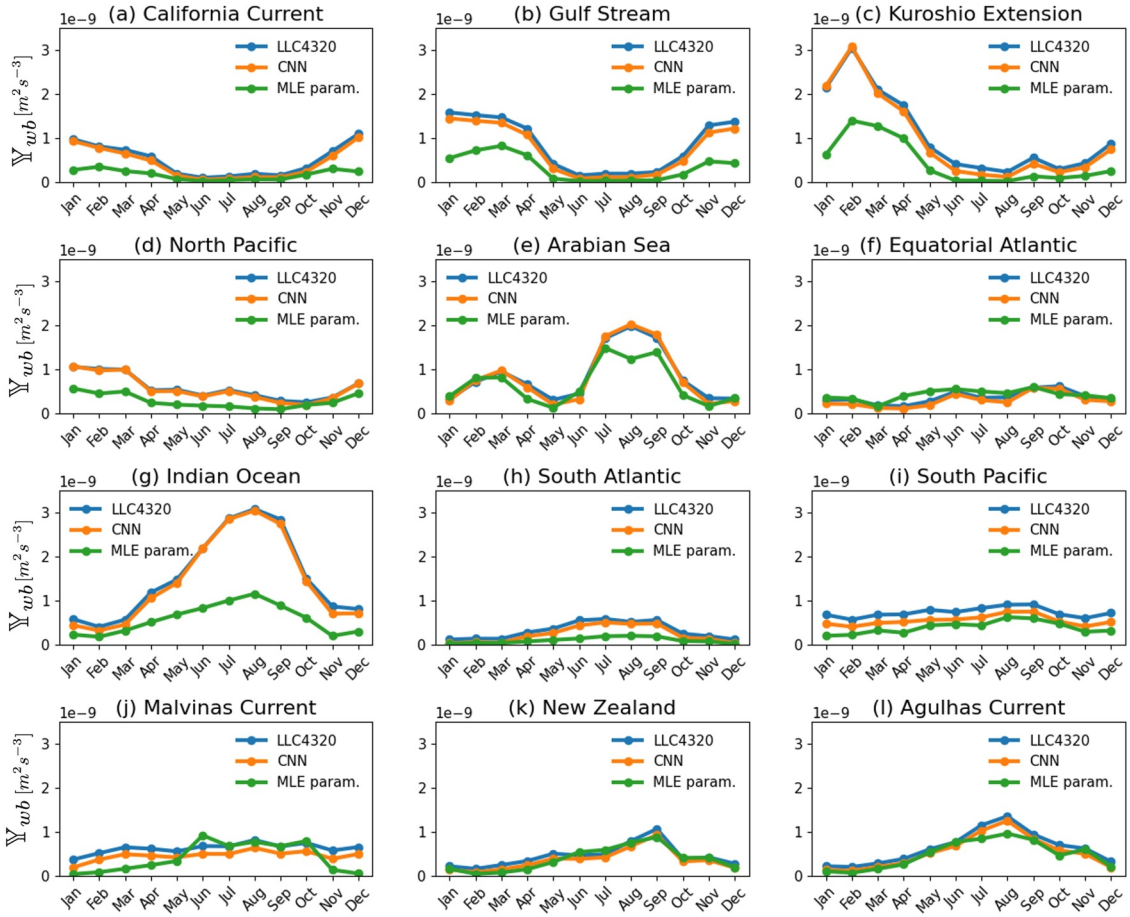

Figure 6. Area-weighted spatial average of $\mathbb{Y}_{wb}\left[\mathrm{m^2 s^{-3}}\right]$ decomposed by region in the $1/4°$ filter scale experiment. In each panel, Convolutional Neural Network (CNN) predictions of $\mathbb{Y}_{wb}$ on unseen test data (orange) are averaged over each month and compared with the LLC4320 target data (blue) and the Mixed Layer Eddy (MLE) parameterization (green). In all regions, the CNN prediction stays close to the LLC4320 target data and surpasses estimates from the MLE parameterization.

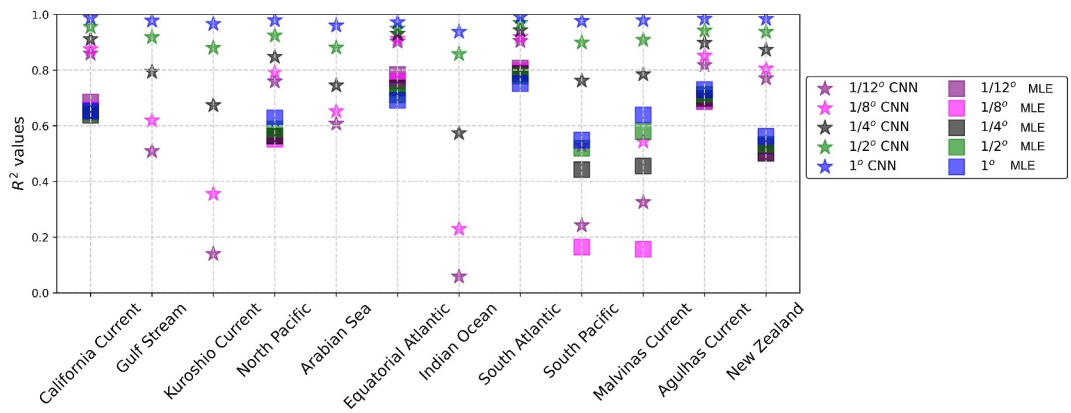

Figure 7. R^2 values of Convolutional Neural Network (CNN) prediction on unseen data (stars) and the Mixed Layer Eddy (MLE) parameterization estimates (squares) decomposed by regions. Colors represent the different filter scale experiments. Note that R^2 is a point-wise estimate and not an averaged quantity as in Figure 6. Negative R^2 values, exhibited in the MLE parameterization (i.e., Gulf Stream, Kuroshio Current, Arabian Sea, and Indian Ocean in all filter scales, and South Pacific and Malvinas Current in the $1/12°$ filter scale), are removed from this figure. The CNN skill exceeds that of the MLE parameterization in all regions and for all filter scales.

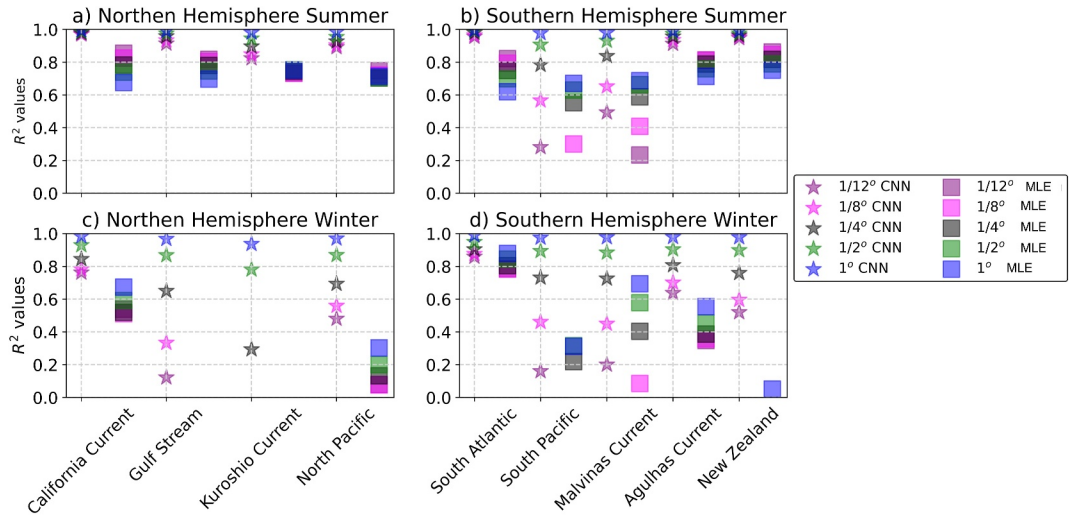

Figure 8. Same as Figure 7, R^2 values of Convolutional Neural Network (CNN) prediction on unseen data (stars) and the Mixed Layer Eddy parameterization estimates (squares) decomposed by regions. Colors represent the different filter scale experiments. Here we include an average over winter (summer) months: January, February, March, and summer (winter) months: July, August, September for regions in the Northern (Southern) Hemisphere. The CNN skill is generally higher in summer (weak vertical buoyancy fluxes [VBF]) compared with winter (strong VBF). Note that we have not included equatorial regions here as the submesoscale equatorial seasonality is less trivial.

with less of a pronounced seasonal cycle, such as in the South Pacific or South Atlantic regions, where the CNN skill is roughly the same for summer and winter, and differences appear within the 0.1 range that can be explained by the CNN initialization sensitivity. The R^2 values of the MLE parameterization still drops in both regions during winter in all filter scale experiments. One reason for this decrease in skill could be related to the scale of mixed layer instabilities captured by the filter scale. As submesoscale mixed layer instabilities tend to be larger in winter due to the deep winter mixed layer depths (Dong et al., 2020), the small filter scale experiments may not be fully capturing the implied VBF. The seasonal dependence manifested by skill in the CNN and MLE parameterization could thus be attributed to both the scale of instabilities and their energetic properties, where the small filter scale experiments especially struggle to predict the strongest fluxes, generally exhibited during winter and spring months (e.g., Callies et al., 2015; Johnson et al., 2016).

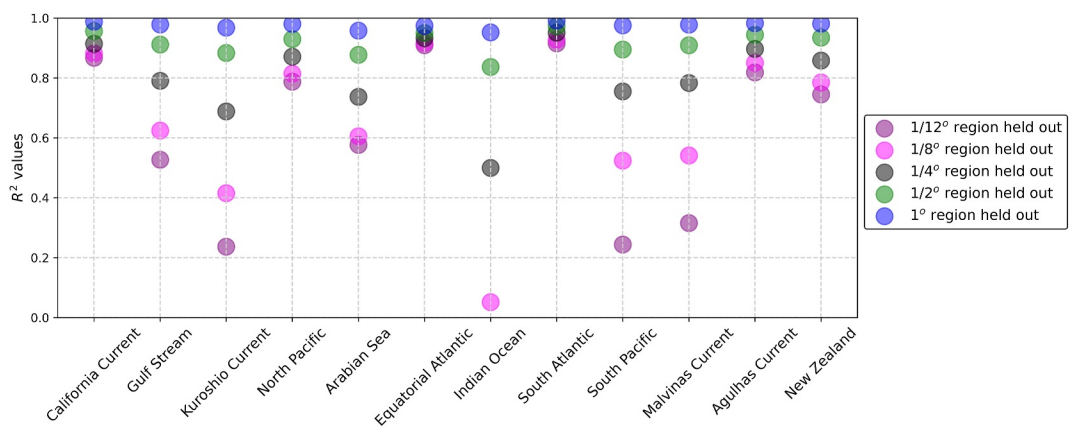

Figure 9. R^2 values of Convolutional Neural Network (CNN) prediction on regions held out during training. Colors represent the different filter scale experiments. Results are consistent with Figure 7 which include all regions for training, suggesting that the CNN is able to generalize onto unseen regions.

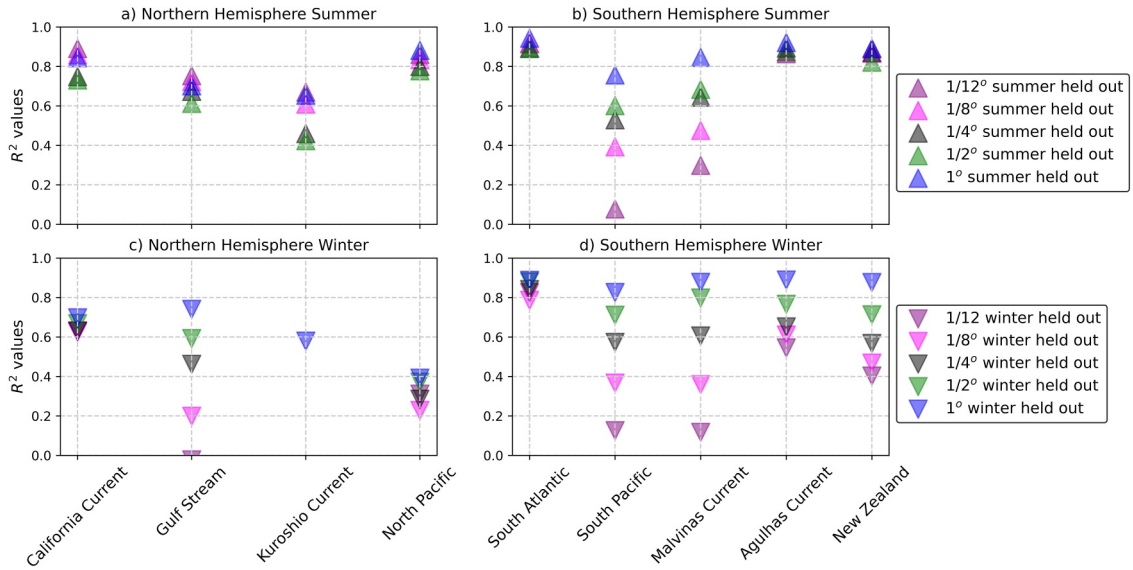

Figure 10. R^2 values of Convolutional Neural Network (CNN) prediction on seasons held out during training: summer (upward triangle), winter (downward triangle). Colors represent the different filter scale experiments. Same as Figure 8, we average over winter (summer) months: January, February, March, and summer (winter) months: July, August, September for regions in the Northern (Southern) Hemisphere. The CNN skill in the Northern Hemisphere regions, in particular near strong ocean boundary current, drops compared with its equivalent in Figure 8.

To better understand the dependency of our method on the training data, and in particular on regional and seasonal variability, in the following subsections we perform two sensitivity tests by holding out parts of the training data, retraining the CNN, and examining CNN prediction skill on unseen regions or selected fractions of the timeseries.

3.1. Holding out Regions From Training

We test the ability of the CNN to make predictions on regions that are not included in the training data. We thus generate 12 new data sets that correspond to removing one region at a time from the training data set. We retrain the CNN in 12 different experiments, and make predictions on a different unseen region each time. The R^2 values of the CNN on the unseen regional data (Figure 9) remain consistent with those found on the full training set (Figure 7) across filter scales and over all regions. This suggests that the training data covers a wide enough range of dynamical regimes that enables generalization of the CNN on regions not included in training, an especially important result given that a fairly small number of regions were included in training compared with the full ocean.

3.2. Holding Out Seasonality From Training

We perform two experiments in which we hold out winter and summer months from the training data, to examine the ability of the CNN to make predictions on unseen seasonal variability. We thus create two new training and test data sets to better understand the overall sensitivity of our method to submesoscale seasonality:

- *Winter held out* refers to training data which excludes from the time series the months of January, February, March from all regions in the Northern Hemisphere, and July, August, September from regions in the Southern Hemisphere. Note again that we have removed equatorial regions from the analysis entirely. The remainder of the time series—for example spring, summer, fall—is used to train the CNN, and predictions are made on the unseen winter data.
- *Summer held out* is same as the above, where we now exclude July, August, September from the Northern Hemisphere and January, February, March from the Southern Hemisphere. Equatorial regions are once again excluded.

In these experiments, we find that results differ between regions in the Northern and Southern Hemispheres. In the Northern Hemisphere regions, the R^2 values in the "Summer held out" experiments decrease by a margin larger than 0.1 compared with the CNN trained on the full timeseries (Figure 10a compared with Figure 8). We find the

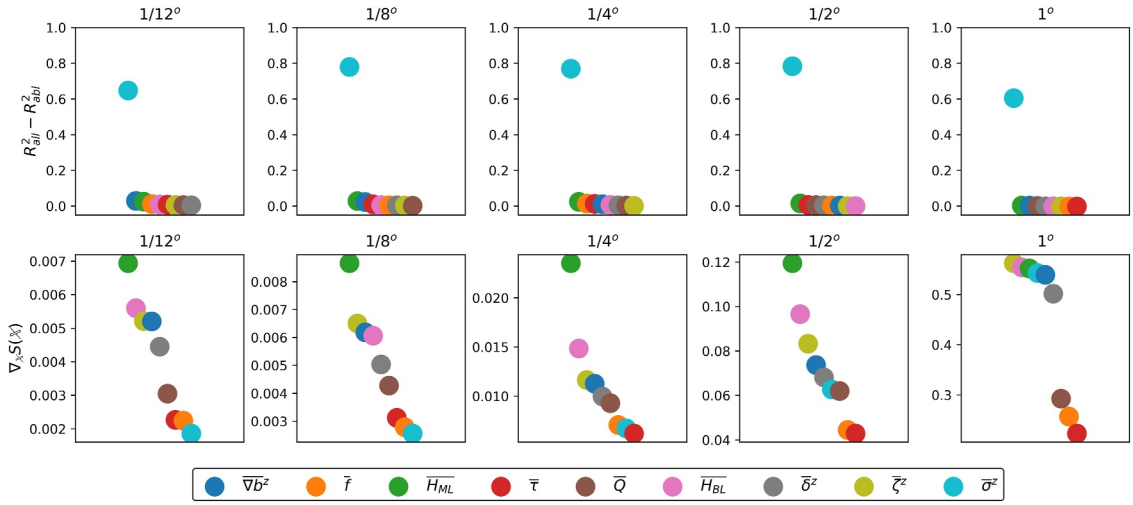

Figure 11. Explainability methods: Top panels are the relative R^2 values, $R^2_{all} - R^2_{abl}$, between the Convolutional Neural Network containing all input features and results from the ablation experiment, where one input feature is removed at a time. Bottom panels are the Jacobian $\nabla_{\mathbb{X}} S(\mathbb{X})$ of the output with respect to individual inputs (in normalized units). In both methods, a high score indicates sensitivity to input features. Columns represent the filter scale experiments. Note that the entire unseen data set (including all regions and seasons) was used here.

largest decreases in particular in the small filter scale experiments in regions affected by strong ocean boundary currents (i.e., the Gulf Stream and Kuroshio Current). Contrarily, the Southern Hemisphere regions are found to be consistent with the predictions for experiments trained on the full timeseries (Figures 8 and 10b). The CNN predictions in the "Winter held out" experiments result in lower R^2 values in the Northern Hemisphere (Figure 10c) compared with predictions trained on the full timeseries. In the Kuroshio Current, for example, all filter scales smaller than 1^o result in negative R^2 values, indicating that there is no skill in the CNN prediction in these cases. In the Southern Hemisphere regions (Figure 10d), the "Winter held out" experiments also display a decrease in R^2 values but it is still within the margin that can be explained by sensitivity to the initialization weights (Figure S5 in Supporting Information S1). These results reinforce the need for the distribution of the data used to train the CNN to include the strongest seasonal fluxes, and particularly from regions where submesoscales are most active, such as near ocean boundary currents.

Results from both the seasonal and regional sensitivity experiments indicate that the learned relationships between the input features and $\mathbb{Y}_{wb}$ can extend over a variety of dynamical regimes, especially in the large filter scale experiments. In the following section we delve deeper in attempt to interpret these relationships.

4. Local and Non-Local Feature Importance

We have shown that the CNN improves on the MLE parameterization, but an important remaining question is why? What relationships are learned between the input variables and $\mathbb{Y}_{wb}$ that lead to better predictions by the CNN? With such complex and nonlinear relationships, it is difficult to decipher which input feature is most important and for what reason. Many methods exist that help explain and interpret the dependency of CNN outputs to its inputs (e.g., Ribeiro et al., 2016; Selvaraju et al., 2017; Van den Broeck et al., 2022; Zeiler & Fergus, 2014). Here, we have chosen two complimentary methods that help gain insight on the learned relationships and the importance of individual inputs to $\mathbb{Y}_{wb}$.

4.1. Impact of Input Feature on CNN Prediction Skill

To test the dependency of the CNN prediction on certain input features, we perform a set of ablation experiments, where we remove one input feature at a time, retrain the CNN, and examine the resulting prediction skill in terms of the relative R^2 value. This relative R^2 value is taken as the difference between R^2_{all}, resulting from the experiment with all input features included, and R^2_{abl}, resulting from the ablation experiment for each input. A high $R^2_{all} - R^2_{abl}$ value indicates that the skill has dropped in a particular ablation experiment, meaning that the CNN strongly depends on said input feature (top panels in Figure 11). Notably, strain demonstrates the strongest

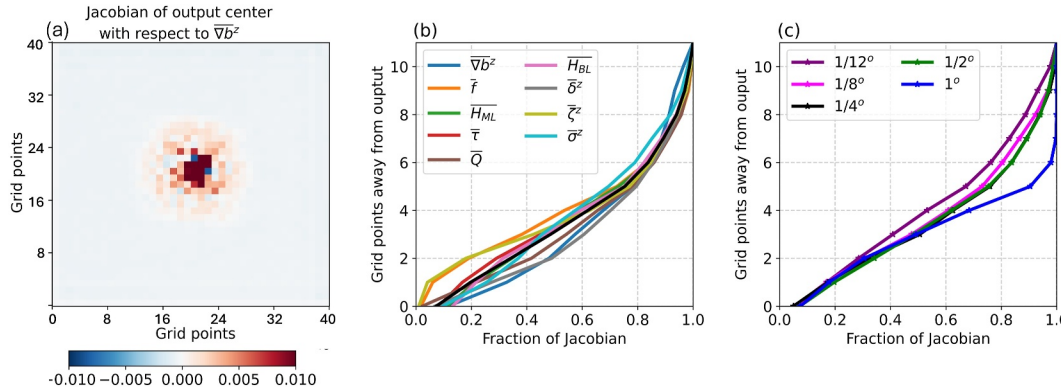

Figure 12. (a) Example of the Jacobian at the center of the input domain with respect to the buoyancy gradient field in a 1/4° filter scale experiment. (b) Radial grid point number compared with fraction of $\nabla_{\mathbb{X}} S(\mathbb{X})$ for all input features in the 1/4° filter scale experiment. Black line is the mean over all inputs. (c) Average fraction of Jacobian for all filter scales experiments. We find that seven grid points captures 90% of the Jacobian fraction, which corresponds to the number of grid points required to capture sensitivity between the input map and a single output grid cell. In all panels the Jacobian is computed in normalized units.

dependency of the CNN consistently across all filter scales. We include a more detailed view of the distributions in Figure S6 in Supporting Information S1 to complement the R^2 values in the ablation experiment, where the contribution of strain is found to be important for getting the overall magnitude of the fluxes right, in particular those of weaker magnitude. Interestingly, there appears to be very little sensitivity to the removal of any other input feature, including those used by the MLE parameterization. We discuss this further in the following section.

4.2. Sensitivity of Output Relative to Input Features

We next apply a complimentary method to the ablation experiment above. The Jacobian of the CNN prediction, $S(\mathbb{X})$, is computed with respect to the input features by taking gradients along the CNN weights, $\nabla_{\mathbb{X}} S(\mathbb{X})$. The Jacobian is an especially useful metric to evaluate the point-wise sensitivity of the output to each input feature (e.g., Ross et al., 2023). Note that unlike the ablation experiment, where we examined the R^2 value on the full output domain, the Jacobian considers only the sensitivity of a single output grid cell to a single grid cell in the input feature map. Here, we compute the Jacobian over the entire unseen test data set, and examine its *average* values for each input feature, thus providing a metric for how sensitive, on average, the CNN output is to each input feature. We contrast the Jacobian with the $R^2_{all} - R^2_{abl}$ values given by the ablation experiments (Figure 11), where for the Jacobian, a high score indicates that the CNN prediction, $S(\mathbb{X})$, is sensitive to point-wise changes in a certain input. We find that the highest-ranked input feature, for which $S(\mathbb{X})$ is most sensitive to, is the mixed layer depth, H_{ML}, which is generally a one-dimensional, local property. The sensitivity to mixed layer depth is followed by sensitivity to boundary layer depth, the buoyancy gradient, and vorticity. Note that $S(\mathbb{X})$ does not appear to be sensitive to point-wise changes in surface heat flux, surface wind stress, or Coriolis, which is likely due to these fields being smoother in the LLC4320 at the scales relevant for the Jacobian. Despite strain being the most important feature in the previous section, it is only in the 1° filter scale experiment that the Jacobian exhibits sensitivity of $S(\mathbb{X})$ to vorticity, divergence, and strain, indicating that the impact of these fields is most apparent at the large scale. Specifically, since strain is coming from the large-scale mesoscale field, it is only in the 1° filter scale experiment, when the signature of strain is contained in a single grid cell, that is detectable by the Jacobian method.

4.3. Receptive Field of the CNN

To further understand the relevance of locality, we follow the analysis in Ross et al. (2023) for examining the Jacobian of the output center point with respect to the full domain of each input feature. An example for the buoyancy gradient input feature is shown in Figure 12a, where the shaded area illustrates the CNN's receptive field needed to predict a single output point. Averaging over that halo, we examine the fraction of Jacobian over the number of grid points, which can be thought of as the percentage of sensitivity for each input feature that is being captured by the CNN (Figures 12b and 12c). We find that 7 grid points away from the center is sufficient for

capturing 90% of the Jacobian fraction, that is, 90% of sensitivity between the output and input features. This relatively local receptive field is found to be consistent across filter scales despite the varying importance of input features found previously.

This result complements the analysis in Gultekin et al. (2024), which found that the Guillaumin and Zanna (2021) CNN skill saturates at a stencil of seven grid points for coarse-graining factors of 4, 8, 12, 16. As discussed in Section 2.3, the CNN architecture choice used here is motivated by that used in Guillaumin and Zanna (2021), however, the physical phenomena we are parameterizing is different, and so are the input and output features. An investigation of the significance of a seven grid point stencil and its dependence on the CNN architecture in both cases is left for a future study.

5. Discussion and Conclusions

The parameterization for submesoscale VBF plays a key role in setting stratification in the ocean mixed layer, and as such contributes to the exchange between the ocean and atmosphere systems in GCMs. In contrast to previous physics-based approaches, here we develop a new data-driven parameterization, where a CNN is trained to predict mixed layer VBF. The subgrid flux, $\mathbb{Y}_{wb}$, is inferred by the CNN as a function of 9 large-scale input features with known relevance to submesoscale VBF: $\overline{\nabla b}^z$, f, H_{ML}, N^2, Q, τ, H_{BL}, $\overline{\sigma}^z$, $\overline{\delta}^z$, $\overline{\zeta}^z$ (see Table 1). The data used for training is given from 12 regions sampled from the global high-resolution LLC4320 simulation output. The CNN is trained over a random selection of 80% of all data, while the remaining 20% is unseen by the CNN and is used for testing. We perform five filter scale experiments of $1/12°, 1/8°, 1/4°, 1/2°$, l° and compare with a baseline given by the Bodner et al. (2023) formulation of the MLE parameterization.

We consistently find that the CNN predictions improve on the MLE parameterization, with higher R^2 values across all regions, seasons, and filter scales tested in this study. We additionally perform several sensitivity experiments, where we test the CNN's ability to make predictions on regional or temporal data held out during training. It is found that the CNN, in particular in the larger filter scale experiments, is able to make skillful predictions on unseen data, but does poorly during when submesoscales are most active, generally corresponding to winter months and near ocean boundary currents.

The significant improvement on the MLE parameterization indicates that the CNN has learned new meaningful relationships between the input features and $\mathbb{Y}_{wb}$, such that it is able to make skillful predictions over widely different dynamical regimes. We applied two complimentary explainability methods which enable a closer look at the relationships between the CNN output and input features. We find that the CNN exhibits strong dependency on the local relationship between $\mathbb{Y}_{wb}$ and the mixed layer depth, a 1D property driven by surface forcing, and strong non-local dependency on the large scale strain field.

An important limitation of these methods is in detecting the importance of features that are strongly correlated with other fields, such that removing one feature may not lead to differences in the CNN prediction skill. In particular, fields associated with surface forcing, for example, mixed layer, boundary layer, surface heat flux, and wind stress are expected to be correlated (as is also found to be true in the LLC4320 data, Figure S4 in Supporting Information S1). As such, it is not necessarily that strain is dominant over the other variables but rather that of all the input features used in the CNN there is no other feature that compensates for its contribution. As an example for ∇b and H_{MLD} which rank high in the Jacobian method, there are other input features that are correlated with them, such as H_{BLD} (Figure S4 in Supporting Information S1). For this reason the CNN still has the needed information to make skillful predictions even when these input features are removed. Nonetheless, results from the ablation experiments suggest that the primary reason the CNN predictions surpass those of the MLE parameterization are due to the newly-captured non-local relationship between $\mathbb{Y}_{wb}$ and the large scale strain field, on which the CNN is strongly dependent (Figure S6 in Supporting Information S1). Note that strain is known for its role in constraining submesoscale fluxes and contributing to frontal intensification (e.g., Balwada et al., 2021; Shcherbina et al., 2013; Sinha et al., 2023), where the confluence or diafluence of the flow impacts conditions for baroclinic instability (Spall, 1997) or contribute to frontogenesis, where an ageostrophic secondary circulation intensifies the front (e.g., Hoskins & Bretherton, 1972; Shakespeare & Taylor, 2013) and helps set the frontal length scale, a key scaling factor in the MLE parameterization (Bodner et al., 2023; Calvert et al., 2020; Fox-Kemper et al., 2011). These findings emphasize the relevance of strain to improving submesoscale VBF parameterizations, such as recently proposed by J. Zhang et al. (2023).

We have demonstrated that the CNN improves on the MLE parameterization in an offline setting. However, it is important to note that during training, the CNN minimizes the MSE loss—a point-wise metric closely related to the R^2 value (as shown in Equation S1 in Supporting Information S1). The MLE parameterization, on the other hand, is designed to represent the bulk effects of submesoscale VBF, which has been proven to be an effective estimate compared with various submesoscale resolving simulations and observations (e.g., Capet et al., 2008; Mensa et al., 2013; Richards et al., 2021; Uchida et al., 2022). A next important step is to explore the implications of better captured point-wise mixed layer VBF in a GCM and compare online with the bulk effect of the MLE parameterization, such as in a recent attempt by Zhou et al. (2024) in a Regional Ocean Modeling System (ROMS). We have designed our method to correspond with the existing implementation of the MLE parameterization in GCMs, where the theoretical expression for $\mathbb{Y}_{wb}$ in Equation S2 in Supporting Information S1 can simply be replaced with the CNN. A relatively small receptive field of 7 grid points is found to be sufficient at capturing relationships between the input features and $\mathbb{Y}_{wb}$, which suggests that a smaller network may aid future implementation efforts in GCMs (C. Zhang et al., 2023). A decomposition may be preferred to distinguish the bulk restratification effect with the intermittent negative fluxes, and will allow a more natural relationship with VBF already estimated in boundary layer turbulence parameterizations (Large et al., 1994; Reichl & Hallberg, 2018; Sane et al., 2023). The exact formulation, implementation, and evaluation of impact on climate variables is left for future work.

It should be noted that the method presented here is limited by the data and processing choices used for training the CNN. First, the resolution of the LLC4320 does not capture the relevant scale of mixed layer instabilities globally (Dong et al., 2020) which impacts the VBF signature and the CNN skill. The promising skill found in the large filter scale experiments, with minimal seasonal sensitivity, suggests the method is most robust when the output entirely contains the mixed layer VBF (Figure 2). Second, the nonlinear interactions between submesoscale and boundary layer turbulence are not resolved in the LLC4320, where boundary layer turbulence is entirely parameterized (Large et al., 1994; Taylor & Thompson, 2023). These interactions are intertwined with the definition of mixed layer depth, which we hinge on for averaging purposes, yet do not consider its vertical structure. Third, the presence of inertia gravity waves, or other high frequency effects, such as inertial oscillations, tides, winds, and diurnal cycles, may alter the VBF at first order and cannot easily be disentangled from the submesoscales within the 12-hr averaging performed in the processing stage (Balwada et al., 2018; Delpech et al., 2024; Richards et al., 2021; Su et al., 2020). Lastly, our method is designed only for mixed layer VBF, and is likely missing contributions of submesoscale VBF extending below the mixed layer (Siegelman et al., 2020).

An extension of the work presented here may include incorporating data from other submesoscale permitting simulations with different boundary layer closures, high-frequency motions, or mixed layer vertical structures to better understand the sensitivities to these choices (Gultekin et al., 2024; Uchida et al., 2022). Likewise, examining the sensitivity of the CNN method to other filtering choices may clarify the role high frequency motions have on the large-scale VBF signature required for a submesoscale parameterization (Delpech et al., 2024; Jones et al., 2023; Richards et al., 2021; Su et al., 2020). An investigation of causal links (e.g., Camps-Valls et al., 2023) or an equation discovery approach (e.g., Zanna & Bolton, 2020) may enable a closer comparison with J. Zhang et al. (2023) and the significance of strain over different regions and seasons, and provide additional insights on the relationship between input features and submesoscale VBF. For example, the dependence of the CNN on strain may implicitly suggest a frontal length scale that combines several input features (or their gradients) into a single rescaling factor with regional or seasonal variability (such as proposed by Bodner et al., 2023). These future investigations may help determine the minimal number of input features and guide optimal CNN architectures or further MLE parameterization developments under the various scenarios. Beyond the modeling framework, the utility of our work can also be made amenable to observational data. In particular, the Surface Water and Ocean Topography (SWOT) altimeter mission is starting to provide measurements of sea surface height at a submesoscale permitting resolution (Archer et al., 2025; Morrow et al., 2019). Complimentary to other recent data-driven inference approaches (e.g., Bolton & Zanna, 2019; Qiu et al., 2016; Souza et al., 2025), the method presented here provides an opportunity to leverage surface fields derived from SWOT and other satellite products to infer subsurface VBF and gain new insights of upper ocean dynamics.

Conflict of Interest

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

Data Availability Statement

Data from the LLC4320 simulation can be accessed using the `llcreader` Python package (Abernathey, 2019). Variables from the LLC4320 output in the regions used in this study are stored on the LEAP-Catalog https://catalog.leap.columbia.edu/feedstock/highresolution-ocean-simulation-llc4320-12hourly-averaged-3d-regions. Code used to process the LLC4320, train the CNN, and generate the figures in this manuscript can be found at https://doi.org/10.5281/zenodo.15620970 (Bodner et al., 2025). Diagnostics incorporate open source Python packages: `xhistogram`, xhistogram.readthedocs.io; `fastjmd95` (Abernathey, 2020), `xmitgcm` (Abernathey et al., 2021).

Acknowledgments

This research received support through Schmidt Sciences. AB was supported by a grant from the Simons Foundation: award number 855143, Bodner. We thank members of the M^2LInES project for support and constructive feedback during the formulation of ideas, in particular, Pavel Perezhogin, Chris Pedersen, Ryan Abernathey, Carlos Fernandez-Granda, and Fabrizio Falasca. The authors would also like to thank the Pangeo Project for providing open-source code which enabled timely analysis for working with the LLC4320 data. This research was also supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

Abernathey, R. (2019). Petabytes of ocean data, part I: Nasa ecco data portal [Software]. *xmitgcm.readthedocs.io*. https://medium.com/pangeo/petabytes-of-ocean-data-part-1-nasa-ecco-data-portal-81e3c5e077be

Abernathey, R. (2020). fastjmd95: Numba implementation of jackett & mcdougall (1995) ocean equation of state [Software]. https://zenodo.org/records/4498376

Abernathey, R., Busecke, J., Smith, T., Banihirwe, A., Fernandes, F., Bourbeau, J., et al. (2021). Xgcm: General circulation model postprocessing with xarray [Software]. *Zenodo*. https://zenodo.org/records/4421428

Ajayi, A., Le Sommer, J., Chassignet, E. P., Molines, J.-M., Xu, X., Albert, A., & Dewar, W. (2021). Diagnosing cross-scale kinetic energy exchanges from two submesoscale permitting ocean models. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2019MS001923. https://doi.org/10.1029/2019ms001923

Archer, M., Wang, J., Klein, P., Dibarboure, G., & Fu, L.-L. (2025). Wide-swath satellite altimetry unveils global submesoscale ocean dynamics. *Nature*, *640*(8059), 691–696. https://doi.org/10.1038/s41586-025-08722-8

Bachman, S. D., Fox-Kemper, B., Taylor, J. R., & Thomas, L. N. (2017). Parameterization of frontal symmetric instabilities. I: Theory for resolved fronts. *Ocean Modelling*, *109*, 72–95. https://doi.org/10.1016/j.ocemod.2016.12.003

Balwada, D., Smith, K. S., & Abernathey, R. (2018). Submesoscale vertical velocities enhance tracer subduction in an idealized Antarctic circumpolar current. *Geophysical Research Letters*, *45*(18), 9790–9802. https://doi.org/10.1029/2018gl079244

Balwada, D., Xiao, Q., Smith, S., Abernathey, R., & Gray, A. R. (2021). Vertical fluxes conditioned on vorticity and strain reveal submesoscale ventilation. *Journal of Physical Oceanography*, *51*(9), 2883–2901. https://doi.org/10.1175/jpo-d-21-0016.1

Barkan, R., Molemaker, M. J., Srinivasan, K., McWilliams, J. C., & D'Asaro, E. A. (2019). The role of horizontal divergence in submesoscale frontogenesis. *Journal of Physical Oceanography*, *49*(6), 1593–1618. https://doi.org/10.1175/jpo-d-18-0162.1

Boccaletti, G., Ferrari, R., & Fox-Kemper, B. (2007). Mixed layer instabilities and restratification. *Journal of Physical Oceanography*, *37*(9), 2228–2250. https://doi.org/10.1175/jpo3101.1

Bodner, A., Balwada, D., & Zanna, L. (2025). Submeso_param_net: Code repository for a data-driven submesoscale parameterization [Software]. *Zenodo*. https://zenodo.org/records/15620971

Bodner, A., Fox-Kemper, B., Johnson, L., Van Roekel, L. P., McWilliams, J. C., Sullivan, P. P., et al. (2023). Modifying the mixed layer eddy parameterization to include frontogenesis arrest by boundary layer turbulence. *Journal of Physical Oceanography*, *53*(1), 323–339. https://doi.org/10.1175/jpo-d-21-0297.1

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399. https://doi.org/10.1029/2018ms001472

Bopp, L., Lévy, M., Resplandy, L., & Sallée, J.-B. (2015). Pathways of anthropogenic carbon subduction in the global ocean. *Geophysical Research Letters*, *42*(15), 6416–6423. https://doi.org/10.1002/2015gl065073

Callies, J., & Ferrari, R. (2018). Baroclinic instability in the presence of convection. *Journal of Physical Oceanography*, *48*(1), 45–60. https://doi.org/10.1175/jpo-d-17-0028.1

Callies, J., Ferrari, R., Klymak, J. M., & Gula, J. (2015). Seasonality in submesoscale turbulence. *Nature Communications*, *6*(1), 6862. https://doi.org/10.1038/ncomms7862

Calvert, D., Nurser, G., Bell, M. J., & Fox-Kemper, B. (2020). The impact of a parameterisation of submesoscale mixed layer eddies on mixed layer depths in the nemo ocean model. *Ocean Modelling*, *154*, 101678. https://doi.org/10.1016/j.ocemod.2020.101678

Camps-Valls, G., Gerhardus, A., Ninad, U., Varando, G., Martius, G., Balaguer-Ballester, E., et al. (2023). Discovering causal relations and equations from data. *Physics Reports*, *1044*, 1–68. https://doi.org/10.1016/j.physrep.2023.10.005

Capet, X., Campos, E., & Paiva, A. (2008). Submesoscale activity over the Argentinian shelf. *Geophysical Research Letters*, *35*(15), L15605. https://doi.org/10.1029/2008gl034736

Chattopadhyay, A., Hassanzadeh, P., & Pasha, S. (2020). Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports*, *10*(1), 1317. https://doi.org/10.1038/s41598-020-57897-9

Dagon, K., Truesdale, J., Biard, J. C., Kunkel, K. E., Meehl, G. A., & Molina, M. J. (2022). Machine learning-based detection of weather fronts and associated extreme precipitation in historical and future climates. *Journal of Geophysical Research: Atmospheres*, *127*(21), e2022JD037038. https://doi.org/10.1029/2022jd037038

de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., & Iudicone, D. (2004). Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. *Journal of Geophysical Research*, *109*(C12), C12003. https://doi.org/10.1029/2004jc002378

Delpech, A., Barkan, R., Srinivasan, K., McWilliams, J. C., Arbic, B. K., Siyanbola, O. Q., & Buijsman, M. C. (2024). Eddy–internal wave interactions and their contribution to cross-scale energy fluxes: A case study in the California current. *Journal of Physical Oceanography*, *54*(3), 741–754. https://doi.org/10.1175/jpo-d-23-0181.1

Dong, J., Fox-Kemper, B., Zhang, H., & Dong, C. (2020). The scale of submesoscale baroclinic instability globally. *Journal of Physical Oceanography*, *50*(9), 2649–2667. https://doi.org/10.1175/jpo-d-20-0043.1

Forget, G., Campin, J.-M., Heimbach, P., Hill, C., Ponte, R., & Wunsch, C. (2015). Ecco version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development*, *8*(10), 3071–3104. https://doi.org/10.5194/gmd-8-3071-2015

Fox-Kemper, B., Danabasoglu, G., Ferrari, R., Griffies, S., Hallberg, R., Holland, M., et al. (2011). Parameterization of mixed layer eddies. III: Implementation and impact in global ocean climate simulations. *Ocean Modelling*, *39*(1–2), 61–78. https://doi.org/10.1016/j.ocemod.2010.09.002

Fox-Kemper, B., Ferrari, R., & Hallberg, R. (2008). Parameterization of mixed layer eddies. Part I: Theory and diagnosis. *Journal of Physical Oceanography*, *38*(6), 1145–1165. https://doi.org/10.1175/2007jpo3792.1

Frankignoul, C., & Hasselmann, K. (1977). Stochastic climate models, part ii application to sea-surface temperature anomalies and thermocline variability. *Tellus*, *29*(4), 289–305. https://doi.org/10.3402/tellusa.v29i4.11362

Gallmeier, K. M., Prochaska, J. X., Cornillon, P., Menemenlis, D., & Kelm, M. (2023). An evaluation of the llc4320 global ocean simulation based on the submesoscale structure of modeled sea surface temperature fields. *Geoscientific Model Development Discussions*, *16*(23), 1–42. https://doi.org/10.5194/gmd-16-7143-2023

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

Gula, J., Taylor, J., Shcherbina, A., & Mahadevan, A. (2022). Submesoscale processes and mixing. In *Ocean mixing* (pp. 181–214). Elsevier.

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

Hoskins, B. J., & Bretherton, F. P. (1972). Atmospheric frontogenesis models: Mathematical formulation and solution. *Journal of the Atmospheric Sciences*, *29*(1), 11–37. https://doi.org/10.1175/1520-0469(1972)029<0011:afmmfa>2.0.co;2

IPCC. (2019). *Special report on the ocean and cryosphere in a changing climate*. Cambridge University Press. https://doi.org/10.1017/9781009157964

IPCC. (2021). Climate change 2021: The physical science basis. In *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. https://doi.org/10.1017/9781009157896

Johnson, L., Lee, C. M., & D'Asaro, E. A. (2016). Global estimates of lateral springtime restratification. *Journal of Physical Oceanography*, *46*(5), 1555–1573. https://doi.org/10.1175/jpo-d-15-0163.1

Jones, C. S., Xiao, Q., Abernathey, R. P., & Smith, K. S. (2023). Using lagrangian filtering to remove waves from the ocean surface velocity field. *Journal of Advances in Modeling Earth Systems*, *15*(4), e2022MS003220. https://doi.org/10.1029/2022ms003220

Lapeyre, G., Klein, P., & Hua, B. L. (2006). Oceanic restratification forced by surface frontogenesis. *Journal of Physical Oceanography*, *36*(8), 1577–1590. https://doi.org/10.1175/jpo2923.1

Large, W. G., McWilliams, J. C., & Doney, S. C. (1994). Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization. *Reviews of Geophysics*, *32*(4), 363–403. https://doi.org/10.1029/94rg01872

Mahadevan, A., Tandon, A., & Ferrari, R. (2010). Rapid changes in mixed layer stratification driven by submesoscale instabilities and winds. *Journal of Geophysical Research*, *115*(C3), C03017. https://doi.org/10.1029/2008jc005203

McWilliams, J. C. (2016). Submesoscale currents in the ocean. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *472*(2189), 20160117. https://doi.org/10.1098/rspa.2016.0117

Menemenlis, D., Campin, J.-M., Heimbach, P., Hill, C., Lee, T., Nguyen, A., et al. (2008). ECCO2: High resolution global ocean and sea ice data synthesis. *Mercator Ocean Quarterly Newsletter*, *31*(October), 13–21.

Menemenlis, D., Hill, C., Henze, C., Wang, J., & Fenty, I. (2021). Pre-SWOT level-4 hourly MITgcm LLC4320 native 2 km grid oceanographic version 1.0. Version.

Mensa, J. A., Garraffo, Z., Griffa, A., Özgökmen, T. M., Haza, A., & Veneziani, M. (2013). Seasonality of the submesoscale dynamics in the gulf stream region. *Ocean Dynamics*, *63*(8), 923–941. https://doi.org/10.1007/s10236-013-0633-1

Morrow, R., Fu, L.-L., Ardhuin, F., Benkiran, M., Chapron, B., Cosme, E., et al. (2019). Global observations of fine-scale ocean surface topography with the surface water and ocean topography (SWOT) mission. *Frontiers in Marine Science*, *6*, 232. https://doi.org/10.3389/fmars.2019.00232

Otness, K., Zanna, L., & Bruna, J. (2023). Data-driven multiscale modeling of subgrid parameterizations in climate models. *arXiv preprint arXiv:2303.17496*.

Perezhogin, P., Zanna, L., & Fernandez-Granda, C. (2023). Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model. *arXiv preprint arXiv:2302.07984*, *15*(10), e2023MS003681. https://doi.org/10.1029/2023ms003681

Qiu, B., Chen, S., Klein, P., Ubelmann, C., Fu, L.-L., & Sasaki, H. (2016). Reconstructability of three-dimensional upper-ocean circulation from SWOT sea surface height measurements. *Journal of Physical Oceanography*, *46*(3), 947–963. https://doi.org/10.1175/jpo-d-15-0188.1

Reichl, B. G., & Hallberg, R. (2018). A simplified energetics based planetary boundary layer (EPBL) approach for ocean climate simulations. *Ocean Modelling*, *132*, 112–129. https://doi.org/10.1016/j.ocemod.2018.10.004

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Richards, K., Whitt, D., Brett, G., Bryan, F., Feloy, K., & Long, M. (2021). The impact of climate change on ocean submesoscale activity. *Journal of Geophysical Research: Oceans*, *126*(5), e2020JC016750. https://doi.org/10.1029/2020jc016750

Rocha, C. B., Chereskin, T. K., Gille, S. T., & Menemenlis, D. (2016). Mesoscale to submesoscale wavenumber spectra in Drake Passage. *Journal of Physical Oceanography*, *46*(2), 601–620. https://doi.org/10.1175/jpo-d-15-0087.1

Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2023). Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. *Journal of Advances in Modeling Earth Systems*, *15*(1), e2022MS003258. https://doi.org/10.1029/2022ms003258

Sagaut, P. (2005). *Large eddy simulation for incompressible flows: An introduction*. Springer Science & Business Media.

Sane, A., Reichl, B. G., Adcroft, A., & Zanna, L. (2023). Parameterizing vertical mixing coefficients in the ocean surface boundary layer using neural networks. *arXiv preprint arXiv:2306.09045*, *15*(10), e2023MS003890. https://doi.org/10.1029/2023ms003890

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Shakespeare, C. J., & Taylor, J. R. (2013). A generalized mathematical model of geostrophic adjustment and frontogenesis: Uniform potential vorticity. *Journal of Fluid Mechanics*, *736*, 366–413. https://doi.org/10.1017/jfm.2013.526

Shamekh, S., Lamb, K. D., Huang, Y., & Gentine, P. (2023). Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences*, *120*(20), e2216158120. https://doi.org/10.1073/pnas.2216158120

Shcherbina, A. Y., D'Asaro, E. A., Lee, C. M., Klymak, J. M., Molemaker, M. J., & McWilliams, J. C. (2013). Statistics of vertical vorticity, divergence, and strain in a developed submesoscale turbulence field. *Geophysical Research Letters*, *40*(17), 4706–4711. https://doi.org/10.1002/grl.50919

Siegelman, L., Klein, P., Rivière, P., Thompson, A. F., Torres, H. S., Flexas, M., & Menemenlis, D. (2020). Enhanced upward heat transport at deep submesoscale ocean fronts. *Nature Geoscience*, *13*(1), 50–55. https://doi.org/10.1038/s41561-019-0489-1

Sinha, A., Callies, J., & Menemenlis, D. (2023). Do submesoscales affect the large-scale structure of the upper ocean? *Journal of Physical Oceanography*, *53*(4), 1025–1040. https://doi.org/10.1175/jpo-d-22-0129.1

Souza, A. N., Silvestri, S., Deck, K., Bischoff, T., Flierl, G., & Ferrari, R. (2025). Surface to seafloor: A generative machine learning framework for decoding the ocean interior state. *arXiv preprint arXiv:2504.15308*.

Souza, A. N., Wagner, G., Ramadhan, A., Allen, B., Churavy, V., Schloss, J., et al. (2020). Uncertainty quantification of ocean parameterizations: Application to the k-profile-parameterization for penetrative convection. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002108. https://doi.org/10.1029/2020ms002108

Spall, M. A. (1997). Baroclinic jets in confluent flow. *Journal of Physical Oceanography*, *27*(6), 1054–1071. https://doi.org/10.1175/1520-0485(1997)027<1054:bjicf>2.0.co;2

Su, Z., Torres, H., Klein, P., Thompson, A. F., Siegelman, L., Wang, J., et al. (2020). High-frequency submesoscale motions enhance the upward vertical heat transport in the global ocean. *Journal of Geophysical Research: Oceans*, *125*(9), e2020JC016544. https://doi.org/10.1029/2020jc016544

Su, Z., Wang, J., Klein, P., Thompson, A. F., & Menemenlis, D. (2018). Ocean submesoscales as a key component of the global heat budget. *Nature Communications*, *9*(1), 775. https://doi.org/10.1038/s41467-018-02983-w

Taylor, J. R., & Thompson, A. F. (2023). Submesoscale dynamics in the upper ocean. *Annual Review of Fluid Mechanics*, *55*(1), 103–127. https://doi.org/10.1146/annurev-fluid-031422-095147

Torres, H. S., Klein, P., Menemenlis, D., Qiu, B., Su, Z., Wang, J., et al. (2018). Partitioning ocean motions into balanced motions and internal gravity waves: A modeling study in anticipation of future space missions. *Journal of Geophysical Research: Oceans*, *123*(11), 8084–8105. https://doi.org/10.1029/2018jc014438

Torres, H. S., Wineteer, A., Rodriguez, E., Klein, P., Thompson, A. F., Perkovic-Martin, D., et al. (2025). Submesoscale eddy contribution to ocean vertical heat flux diagnosed from airborne observations. *Geophysical Research Letters*, *52*(2), e2024GL112278. https://doi.org/10.1029/2024gl112278

Treguier, A.-M., de Boyer Montégut, C., Bozec, A., Chassignet, E. P., Fox-Kemper, B., Hogg, A. M., et al. (2023). *The mixed layer depth in the ocean model intercomparison project (OMIP): Impact of resolving mesoscale eddies*. EGUsphere.

Uchida, T., Balwada, D., Abernathey, R., McKinley, G., Smith, S., & Levy, M. (2019). The contribution of submesoscale over mesoscale eddy iron transport in the open southern ocean. *Journal of Advances in Modeling Earth Systems*, *11*(12), 3934–3958. https://doi.org/10.1029/2019ms001805

Uchida, T., Le Sommer, J., Stern, C., Abernathey, R., Holdgraf, C., Albert, A., et al. (2022). Cloud-based framework for inter-comparing submesoscale permitting realistic ocean models. *Geoscientific Model Development Discussions*, *15*(14), 1–32. https://doi.org/10.5194/gmd-15-5829-2022

Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the tractability of shap explanations. *Journal of Artificial Intelligence Research*, *74*, 851–886. https://doi.org/10.1613/jair.1.13283

Wang, P., Yuval, J., & O'Gorman, P. A. (2022). Non-local parameterization of atmospheric subgrid processes with neural networks. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS002984. https://doi.org/10.1029/2022ms002984

Xiao, Q., Balwada, D., Jones, C. S., Herrero-González, M., Smith, K. S., & Abernathey, R. (2023). Reconstruction of surface kinematics from sea surface height using neural networks. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003709. https://doi.org/10.1029/2023ms003709

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. https://doi.org/10.1038/s41467-020-17142-3

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

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6–12, 2014, proceedings, part I 13* (pp. 818–833).

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. *arXiv preprint arXiv:2303.00962*, *15*(10), e2023MS003697. https://doi.org/10.1029/2023ms003697

Zhang, J., Zhang, Z., & Qiu, B. (2023). Parameterizing submesoscale vertical buoyancy flux by simultaneously considering baroclinic instability and strain-induced frontogenesis. *Geophysical Research Letters*, *50*(8), e2022GL102292. https://doi.org/10.1029/2022gl102292

Zhou, S., Dong, J., Xu, F., Jing, Z., & Dong, C. (2024). A neural network-based submesoscale vertical heat flux parameterization and its implementation in regional ocean modeling system (roms). *arXiv preprint arXiv:2403.05028*.

Zhu, R., Li, Y., Chen, Z., Du, T., Zhang, Y., Li, Z., et al. (2023). Deep learning improves reconstruction of ocean vertical velocity. *Geophysical Research Letters*, *50*(19), e2023GL104889. https://doi.org/10.1029/2023gl104889