

A Data-Driven Approach for Parameterizing Submesoscale Vertical Buoyancy Fluxes in the Ocean Mixed Layer

Abigail Bodner¹, Dhruv Balwada², Laure Zanna³

¹Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

²Lamont-Doherty Earth Observatory, Columbia University, New York, NY, USA

³Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

Key Points:

- We improve the parameterization of submesoscale vertical buoyancy fluxes using a Convolutional Neural Network (CNN).
- The CNN demonstrates high offline skill compared with the physics-based submesoscale parameterization over a wide range of dynamical regimes, seasonality, and resolutions.
- We identify strong dependency on the large scale strain field, which is currently missing from submesoscale parameterizations implemented in General Circulation Models.

Abstract

The parameterizations of submesoscale ($< 10\text{km}$) ocean surface flows are critical in capturing the subgrid effects of vertical fluxes in the ocean mixed layer, yet they struggle to infer the full-complexity of these fluxes in relation to the large scale variables that help set them. In this work, we present a data-driven approach for the submesoscale parameterization, utilizing information from the high-resolution submesoscale-permitting MITgcm-LLC4320 simulation (LLC4320). The new parameterization is given by a Convolutional Neural Network (CNN) trained to infer the subgrid mixed layer vertical buoyancy fluxes as a function of relevant large scale variables. In contrast to previous physics-based approaches, such as the Mixed Layer Eddy (MLE) parameterization, here the CNN infers vertical fluxes that are directly computed from the LLC4320 data, where the submesoscales are resolved down to a resolution of approximately 2km. The CNN has significantly high skill compared with the MLE parameterization, which we demonstrate over a wide range of dynamical regimes and resolutions. We find that the improved skill can be attributed to learned physical relationships between submesoscale fluxes and the large scale strain field, currently missing from submesoscale parameterizations in General Circulation Models.

Plain Language Summary

This work provides a data-driven method for inferring small scale buoyancy fluxes in the upper ocean given from high resolution simulation output. Our method provides high skill compared with the estimates currently used in climate models, suggesting that information from the large scale strain field is key to reducing model biases.

1 Introduction

General Circulation Models (GCMs) and future climate change projections are notoriously sensitive to parameterizations of unresolved phenomena at the ocean-atmosphere interface (IPCC, 2019, 2021). Of particular importance is the ocean mixed layer, where multiscale turbulence modulates the transfer of properties – such as heat, momentum, and carbon – between the atmosphere and ocean interior (e.g., Frankignoul & Hasselmann, 1977; Bopp et al., 2015; Su et al., 2020; Fox-Kemper et al., 2022). Ocean surface submesoscales flows typically appear on the time scale of hours to days and $O(1)\text{km}$ in spatial scale (e.g., McWilliams, 2016). Submesoscale flows are sandwiched between mesoscale $O(100)\text{km}$ flows and boundary layer turbulence $O(1)\text{m}$, and play an important role in vertical transport in the ocean mixed layer. As opposed to mixing and homogenization of mixed layer buoyancy dominated by boundary layer turbulence, submesoscale flows tend to exhibit a positive vertical buoyancy flux and thus have a net restratification effect (Boccaletti et al., 2007; Thomas et al., 2008; Mahadevan, 2016; Johnson et al., 2016; McWilliams, 2016; Balwada et al., 2021; Taylor & Thompson, 2023). Such positive buoyancy fluxes are at leading order a result of mixed layer instabilities formed along submesoscale fronts, with dynamics attributed to multiscale interactions between the mesoscale flow field and boundary layer turbulence (Fox-Kemper et al., 2008; McWilliams, 2016; Bachman & Klocker, 2020; Gula et al., 2022; Bodner et al., 2023). It is thus a sensitive interplay between mesoscale flows, submesoscale flows, boundary layer turbulence, and surface forcing that set the stratification in the ocean mixed layer, for which GCM parameterizations struggle to capture in its entirety.

The effects of submesoscale vertical buoyancy fluxes are inferred in many GCMs by the Mixed Layer Eddy (MLE) parameterization (Fox-Kemper et al., 2011; Calvert et al., 2020; Bodner et al., 2023), which represents the positive vertical buoyancy flux produced by mixed layer instabilities formed along submesoscale fronts (Fox-Kemper et al., 2008). The MLE parameterization is cast in the form of a streamfunction (e.g., Gent & Mcwilliams, 1990; Griffies, 1998; Ferrari et al., 2008), which represents the net restrat-

ification effect prescribed by submesoscale eddy fluxes. Frontal sharpening, known as frontogenesis, helps set the buoyancy gradient from which the mixed layer instabilities are formed. Frontogenesis also contributes vertical buoyancy flux at scales typically smaller than MLE (Hoskins & Bretherton, 1972; Lapeyre et al., 2006; Shakespeare & Taylor, 2013; Bodner et al., 2020; Gula et al., 2022; Srinivasan et al., 2023). The implementation of the MLE parameterization in GCMs (Fox-Kemper et al., 2011) involves a rescaling factor to enhance the subgrid buoyancy gradient, which tends to be suppressed by the smoother GCM (e.g., Stanley et al., 2020). The frontal width rescaling factor is taken to be proportional to the mixed layer deformation radius (Fox-Kemper et al., 2011; Calvert et al., 2020), yet this scaling is found to not satisfactorily hold when frontogenesis or strong surface forcing are present (e.g., Lapeyre et al., 2006; Mahadevan et al., 2010; Callies & Ferrari, 2018). Bodner et al. (2023) had recently modified the MLE parameterization with a new scaling for frontal width set based on turbulence-induced frontogenesis, linking the MLE parameterization with the Energetics-based Planetary Boundary Layer scheme (ePBL; Reichl & Hallberg, 2018).

Recent developments in high resolution numerical modeling have presented a new class of submesoscale permitting simulations with kilometer scale resolution (e.g., Rocha, Chereskin, et al., 2016; Schubert et al., 2019; Ajayi et al., 2021). In a comprehensive comparison across eight different submesoscale permitting simulations, Uchida et al. (2022) find that the MLE parameterization struggles to capture the full range of complexity and magnitude given by submesoscale vertical buoyancy fluxes, even when submesoscale fronts are resolved and the rescaling factor is not necessary. Studies have suggested that such missing physics can be attributed to the relevance of the large scale flow in setting submesoscale vertical buoyancy fluxes, namely, properties of strain, divergence, and vorticity (e.g., Lapeyre et al., 2006; Capet et al., 2008; McWilliams, 2017; Barkan et al., 2019; Balwada et al., 2021). J. Zhang et al. (2023) propose a parameterization for submesoscale vertical buoyancy flux enhanced by the effects of the large scale strain field, yet does not include a relationship with boundary layer turbulence. Furthermore, the interaction of submesoscale fronts with surface forcing, and in particular in regards to surface cooling or wind orientation, can induce an Ekman Buoyancy Flux, which emerges as a central component to setting mixed layer stratification, serving as another source for frontogenesis, and setting Potential Vorticity conditions for Symmetric Instability (Thomas, 2005; Thomas et al., 2013; Bachman et al., 2017; Wenegrat & Thomas, 2020). Thus, even the most recent advancements in submesoscale parameterizations present incomplete physical relationships between submesoscale vertical buoyancy fluxes and the large scale variables that help set it.

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 parameterizations have yielded promising results in improving physics for subgrid closures. The seminal work by Bolton and Zanna (2019) demonstrated that Neural Networks (NNs) are able to learn subgrid ocean mesoscale momentum fluxes, while remaining physically-constrained. Zanna and Bolton (2020) and Guillaumin and Zanna (2021) respectively trained a NN to learn the underlying equation and statistics of subgrid mesoscale fluxes as given from data. An example for ocean boundary layer turbulence closures can be found in Souza et al. (2020), whom apply a Bayesian approach to estimate empirical parameters in the K-Profile Parameterization (Large et al., 1994) from an ensemble of large eddy simulations. A complimentary approach has been taken in Sane et al. (2023), where empirical parameters from ePBL are replaced with a NN constrained by observations. In the atmosphere, several comparable approaches have trained NNs to infer vertical buoyancy fluxes in the atmospheric boundary layer (e.g., Yuval & O’Gorman, 2020; Wang et al., 2022; Shamekh et al., 2023). Numerous examples for applications exist both in the atmosphere and the ocean for use of machine learning 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).

In this work, we introduce a data-driven approach for parameterizing submesoscale-vertical buoyancy fluxes. The main goal is to train a Convolutional Neural Network (CNN) on realistic simulated data to learn an improved functional relationships between submesoscale vertical buoyancy fluxes and the large scale physics that help set it. The data used to train and test the CNN is sampled from the global ocean MITgcm-LLC4320 (hereafter LLC4320, Menemenlis et al., 2021) simulated at a high resolution of $1/48^\circ$, developed ahead of the Surface Water and Ocean Topography (SWOT) altimeter mission. 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., Rocha, Chereskin, et al., 2016; Rocha, Gille, et al., 2016; Su et al., 2018; Torres et al., 2018; Dong et al., 2020; Uchida et al., 2022; Gallmeier et al., 2023). We describe the LLC4320 data, preprocessing, and choices of CNN inputs and output in section 2. Results of the CNN prediction on unseen data are presented and compared with the MLE parameterization in section 3, together with sensitivity and generalization tests for robustness. In section 4, we apply two complimentary methods to explain the relationship learned by the CNN between the physical variables used as inputs and the submesoscale vertical buoyancy flux. Discussion and concluding remarks are given in section 5.

2 Data and methods

2.1 Designing the learning problem

We choose to include input features that are correlated, or have known analytical relationships, with submesoscale vertical buoyancy fluxes. We thus leverage the physical relevance demonstrated by the MLE parameterization, and extract variables which appear in the Fox-Kemper et al. (2011) and Bodner et al. (2023) formulations. We also include correlated large scale velocity derivatives as demonstrated in 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, $|\overline{\nabla b}^z|$, where overbar represents the filtering operator described in section 2.2, and superscript z represents a depth averaging operator over the mixed layer depth; Coriolis parameter, \overline{f} ; mixed layer depth, H_{ML} ; surface heat flux, \overline{Q}^* ; surface wind stress magnitude, $|\tau| = \sqrt{\overline{\tau_x}^2 + \overline{\tau_y}^2}$; boundary layer depth, H_{BL} ; 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 we exclude the Brunt-Väisälä frequency, $\overline{N^2}$, (Fox-Kemper et al., 2008) as its average quantity in the mixed layer contains identical information as H_{ML} , and thus would not contribute to the CNN training.

Formally, we define our 9 input features as,

$$\mathbf{X} = (|\overline{\nabla b}^z|, \overline{f}, H_{ML}, \overline{Q}^*, |\tau|, H_{BL}, \overline{\sigma}^z, \overline{\zeta}^z, \overline{\delta}^z), \quad (1)$$

and a single output as,

$$\mathbf{Y} = \overline{w}^z \overline{b}^z - \overline{w} \overline{b}^z \quad (2)$$

For compactness, we will frequently refer to the output as $\overline{w' b'}^z := \overline{w}^z \overline{b}^z - \overline{w} \overline{b}^z$, where $(\cdot)'$ represents the submesoscale. The CNN provides the subgrid fluxes as a function, S , of the large scale variables, such that $S(\mathbf{X}) \rightarrow \mathbf{Y}$. Figure 1 illustrates a schematic of the CNN with 9 input features and 1 output. The processing involved in computing all inputs and output from the LLC4320 is described in more detail below.

2.2 Input and output features

The LLC4320 is a $1/48^\circ$ Massachusetts Institute of Technology general circulation model (MITgcm), named after its Latitude-Longitude polar Cap (LLC) grid with 4320

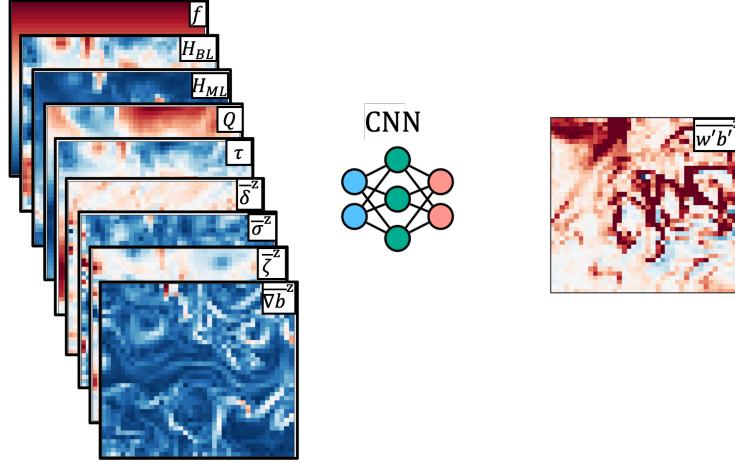


Figure 1. Schematic of CNN method with 9 input features and one output feature (table 1) at a $1/4^\circ$ resolution.

CNN inputs, \mathbf{X}

Depth-averaged buoyancy gradient magnitude	$ \bar{\nabla}b^z $
Coriolis parameter	\bar{f}
Mixed layer depth	\bar{H}_{ML}
Surface heat flux	\bar{Q}^*
Surface wind stress magnitude	$ \bar{\tau} = \sqrt{\bar{\tau}_x^2 + \bar{\tau}_y^2}$
Boundary layer depth	\bar{H}_{BL}
Depth-averaged strain magnitude	$\bar{\sigma}^z = \sqrt{(\bar{u}_x^z - \bar{v}_y^z)^2 + (\bar{v}_x^z + \bar{u}_y^z)^2}$
Depth-averaged vertical vorticity	$\bar{\zeta}^z = \bar{v}_x^z - \bar{u}_y^z$
Depth-averaged horizontal divergence	$\bar{\delta}^z = \bar{u}_x^z + \bar{v}_y^z$

CNN Output, \mathbf{Y}

$$\text{Depth-averaged subgrid vertical buoyancy flux} \quad \bar{w}'\bar{b}'^z := \bar{w}^z\bar{b}^z - \bar{w}\bar{b}^z$$

Table 1. Input and output features used in the CNN method. Overbar represents the filtering operator described in 2.2, and superscript z represents a depth averaging operator over the mixed layer depth.

points on each of the 13 tiles. The LLC4320 is initialized from the Estimating the Circulation and Climate of the Ocean (ECCO), Phase II project, and is forced at the surface by atmospheric reanalysis, at 6 hourly temporal resolution. The LLC4320 simulation also includes a synthetic surface pressure field to mimic tidal forcing present in the real ocean. A total of 14 months are available at hourly frequency from September 2011 to November 2012 (Forget et al., 2015; Menemenlis et al., 2008, 2021).

Figure 2 illustrates a snapshot of the global surface horizontal buoyancy gradient from LLC4320 coarsened to $1/4^\circ$, with boxes highlighting the sampled regions used in this study. Table 2 lists the coordinates of each of the 12 sampled regions of approximately $15^\circ \times 15^\circ$. Buoyancy gradients are a key variable in the MLE parameterization, which helps motivate the choices of regions selected for training. We include a mix of regions with strong variability and others more quiescent (e.g., Torres et al., 2018) in effort to train the CNN on a range of dynamical regimes influenced by seasonality and local flow properties.

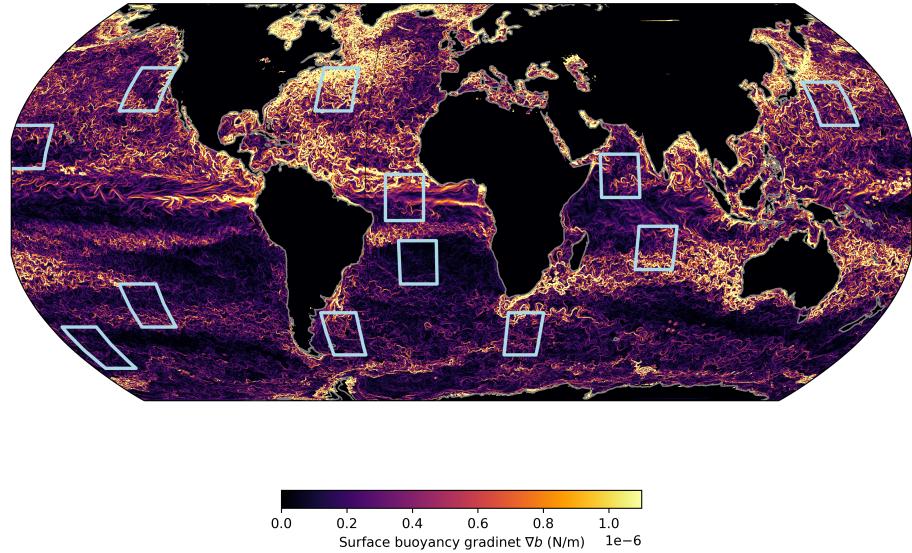


Figure 2. Snapshot of global surface horizontal buoyancy gradient ∇b (N/m) given by the LLC4320 simulation coarsened to $1/4^\circ$. Light blue boxes mark the regions from table 2.

From each of the regions listed in table 2, the LLC4320 output is preprocessed in preparation of the CNN training, in which we compute the CNN input and output features listed in table 1.

In order to filter the synthetic fast-varying wave field from the LLC4320 output, the temporal resolution of each variable is converted from hourly into an average over 12 hour window periods. The potential density anomaly σ_0 is computed from the potential temperature and salinity fields, with reference pressure of 0 dbar and $\rho_0 = 1000 \text{ kg m}^{-3}$. The mixed layer depth, H_{ML} , is defined as the depth at which σ_0 increased by 0.03 kg m^{-3} from its value at 10m depth (de Boyer Montégut et al., 2004). The buoyancy field is then defined as $b = -g\sigma_0/\rho_0$, where $g = 9.81 \text{ m/s}^2$ is the gravity acceleration. Velocities and wind stresses are interpolated to commute with the tracer grid. The boundary layer depth H_{BL} is already an LLC4320 output computed from the K-profile Parameterization criteria for which the Richardson number exceeds the critical value of 0.3 (KPP, Large et al., 1994).

We restrict our method to learning the depth-averaged submesoscale vertical buoyancy fluxes. This approach remains close to the MLE parameterization, which was developed as a bulk formula for the depth-averaged submesoscale vertical buoyancy fluxes (Fox-Kemper et al., 2011; Bodner et al., 2023), and a vertical structure function determines the shape of the parameterization at depth (eq. (S4) in the supplementary material). To this end, all 3D variables are averaged over the mixed layer depth, which we denote by superscript z . Horizontal filtering is applied to all input and output variables to bring the high-resolution, fine-scale LLC4320 data, to the low-resolution large-scale, such that it is comparable to a GCM resolution. We have tested our method over res-

Region	Latitudinal Range	Longitudinal Range
1. Gulf Stream	(30,45)	(-60,-45)
2. South Atlantic	(-30,-15)	(-25,-10)
3. Equator Atlantic	(-8,8)	(-30,-15)
4. Malvinas Current	(-55,-40)	(-60,-45)
5. California Current	(30,45)	(-140,-125)
6. South Pacific	(-45,-30)	(-140,-125)
7. Kuroshio Extension	(25,40)	(145,160)
8. North Pacific	(10,25)	(-180,-165)
9. Southern Ocean, New Zealand	(-60,-45)	(-175,-160)
10. Agulhas Current	(-55,-40)	(20,35)
11. Indian Ocean	(-25,-10)	(70,85)
12. Arabian Sea	(0,15)	(55,70)

Table 2. Coordinate range of sampled regions from the LLC4320 corresponding to the blue boxes in figure 2.

solutions of $1^\circ, 1/2^\circ, 1/4^\circ, 1/8^\circ, 1/12^\circ$. To achieve this, we include both spatial and temporal filtering. In space, we apply a top-hat filter. As an example for the $1/4^\circ$ resolution case, the top-hat filter is an average over 12 grid points of the original LLC4320 grid (Loose et al., 2022). An additional 7-day temporal filter is applied such that all large-scale variables are assumed to vary on timescales larger than a week, e.g., larger than the submesoscale. Note that the 7-day temporal filter is larger than the inertial period with the added intention of reducing noise from tides and internal waves that may have otherwise contaminated the sampled data (Torres et al., 2018; Jones et al., 2023). To review, for any variable ξ , the filtering procedure is defined as both in horizontal space $\langle \xi \rangle_h$ and in time $\widetilde{\langle \xi \rangle}_t$. For compactness, we denote the combination of both averaging operators as $\bar{\xi} := \widetilde{\langle \xi \rangle}_h$, which is applied to all input and output features. If ξ is a three-dimensional variable, we also include the superscript $\bar{\xi}^z$ which corresponds to a vertical average over the mixed layer depth.

Thus, the sub-grid process we are parameterizing is defined by a combined temporal and spatial filter, which is the fast-varying (within the 12 hour to 7 day range) and small-scale (below the spatial filter scale) mixed layer buoyancy flux. This quantity is designed to be the CNN output, formally written as the total turbulent stress tensor, $w' b'^z := \bar{w}^z \bar{b}^z - \bar{w} b^z$. The CNN output of mixed-layer-averaged vertical buoyancy fluxes is thus akin to an estimate of submesoscale fluxes. We demonstrate this in the variance-preserving co-spectrum of w and b in the mixed layer, shown in figure 3. It can be seen that the maximum covariance is predominantly in the submesoscale range. Special cases indicate that more complex dynamics influence mixed layer buoyancy fluxes, either by strong mesoscale activity, such as in the Arabian Sea, or by turbulent activity, such as in the Equatorial Atlantic. Naturally, the filter scale (illustrated by the grey lines in figure 3b) will impact the properties of the learned subgrid flux, which we explore in section 3. We have included all regions to gain a variety of dynamical regimes in our training data. In sections 3.1 and 3.2 we test the ability of the CNN to generalize over different seasons and regions, respectively.

2.3 CNN architecture and training

All features listed in table 1 are normalized by a global mean and standard deviation computed over all regions. To train the CNN, we randomly select 80% of the, approximately, 10,000 samples given from all regions combined, after temporal and spatial filtering is applied. The remaining 20% is left unseen by the CNN and is used only

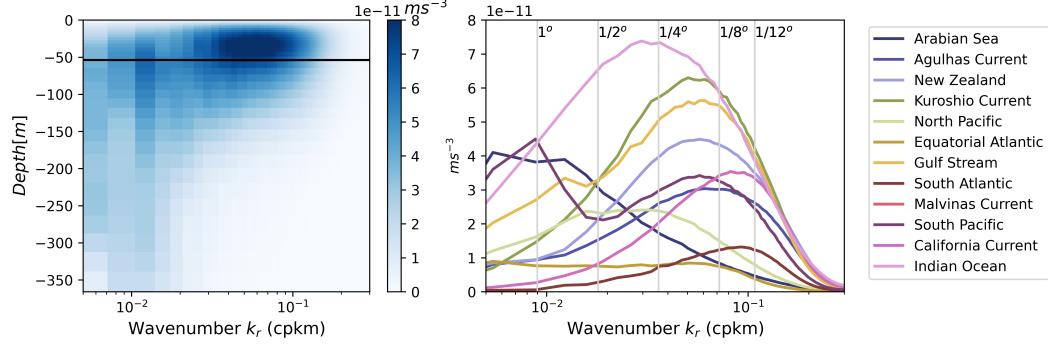


Figure 3. Isotropic cross-spectrum of w and b in variance-preserving form, averaged over the entire simulation duration (14 months). Example of the depth varying cross-spectrum in the Gulf Stream region (left), where the black horizontal line is the average mixed layer depth. On the right is the same but across all regions and includes the depth average over the mixed layer depth. Vertical grey lines mark the filter scales used in this study with respect to the cross-spectrum variability.

to test the prediction of the trained CNN and compare with the target LLC4320 data and MLE parameterization in section 3 below.

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 hyper-parameter 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 Mean Squared Error (MSE) loss (shown in figure S2). Results presented here are based on a CNN with a kernel size of 5X5 in the first layer, 7 hidden convolutional layers with kernel size of 3x3, a learning rate of 2×10^{-4} , and weight decay of 0.02, which gives an MSE of 0.23 in normalized units. The total number of learnable parameters is approximately 300,000.

3 CNN prediction of subgrid submesoscale fluxes

Once training is complete, the relationship between the CNN inputs and output is optimized over the learnable parameters. In other words, the CNN has learned a functional form, $S(\mathbf{X})$, between the input features, \mathbf{X} , and submesoscale vertical buoyancy fluxes, \mathbf{Y} . In this section, we examine the extent to which the relationships learned by the CNN generalize onto data that was not included in the training process. For this purpose, we evaluate the skill of the CNN compared with the target LLC4320 data as well as the Bodner et al. (2023) version of the MLE parameterization. In subsections 3.1 and 3.2 we test whether our results are sensitive to the choice of training on a subset of the timeseries or sampled regions, respectively.

Figure 4 displays a single sample of the $1/4^\circ$ submesoscale vertical buoyancy fluxes, $\overline{w'b'}^z$, as given by the LLC4320 data (4a), CNN (4b), and MLE parameterization (4c). The sample is from the Arabian Sea region during the month of September. For illustration purposes, the LLC4320 with filter scales corresponding to $1/12^\circ, 1/8^\circ, 1/2^\circ, 1^\circ$, is shown in figure S3. The majority of the fluxes appear to be positive, which is the bulk restratification effect inferred by the MLE parameterization. However, the CNN captures much more of the fine-scale structure as well as the sign, including intermittent negative fluxes.

To examine the statistics beyond a single sample, we compute the joint histogram of the $1/4^\circ$ LLC4320 submesoscale fluxes and those given by the CNN. The joint histogram is computed over the entire unseen test dataset, which contains 20% of random samples over all regions, and provides a metric over several orders of magnitude. The LLC4320-CNN joint histogram is compared with that of the LLC4320 and the MLE parameterization in figure 5. In the case of positive fluxes, the CNN prediction remains close to the LLC4320 data, as can be seen by the alignment along the one-to-one grey line, a significant improvement on the MLE parameterization. For the negative fluxes the one-to-one alignment is less pronounced, likely due to the significantly smaller number of negative samples, however the CNN prediction is an improvement on the MLE parameterization which does not infer negative fluxes by construction.

We examine the ability of the CNN in the $1/4^\circ$ resolution experiment to capture seasonality in figure 6, which shows the seasonal cycle of the spatially-weighted-average of $\overline{w'b^z}$ decomposed by region. In each panel, the LLC4320 target data is compared with the MLE parameterization and the CNN predictions on unseen data. Once again, in virtually all regions the CNN prediction outperforms the MLE parameterization, particularly where fluxes appear to be strongest during the winter and spring months.

The CNN and MLE parameterization skill in terms of R^2 values decomposed by region and across all resolution experiments are shown in figure 7. The CNN prediction skill remains above that of the MLE in all filter scale experiments and in all regions. The lowest resolution tends to have the highest skill and it decreases as resolution increases. Note that the MLE parameterization is a bulk formula for wb^z , which is not as sensitive to the different resolutions due to the frontal rescaling factor. As a reminder, figure 3 illustrates how different filter scales define the amount of variability captured in the data. The variability of $\overline{wb^z}$ varies both in scale and by location, which impacts the learned output of the CNN in the different regions. The CNN prediction skill is found to be especially sensitive in the high-resolution experiments, where it performs well in some regions (e.g. the New Zealand region), but not in others (e.g. the Indian Ocean region). In the low resolution experiments, the fields tend to be smoother, as much of the variability is averaged out, thus presenting an easier learning problem for the CNN.

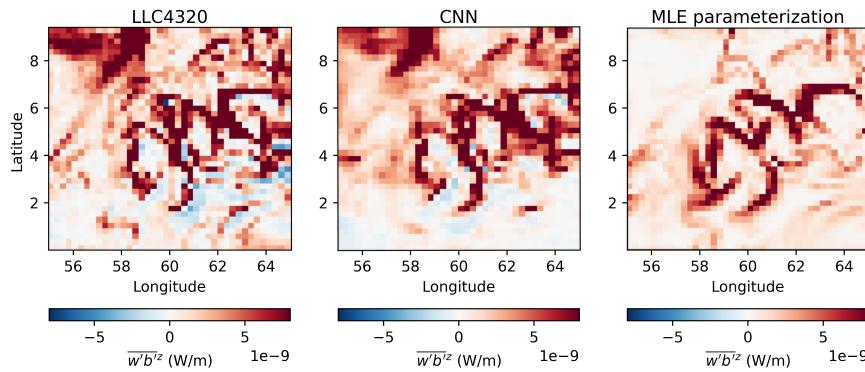


Figure 4. $1/4^\circ$ snapshot taken from the Arabian Sea during the month of September of the depth-averaged LLC4320 subgrid vertical buoyancy flux [W/m^2] (a), CNN prediction for same in physical space (b), and the MLE parameterization (c).

To better understand the dependency of our method on the training data, in the following subsections we perform two sensitivity tests by holding out parts of the train-

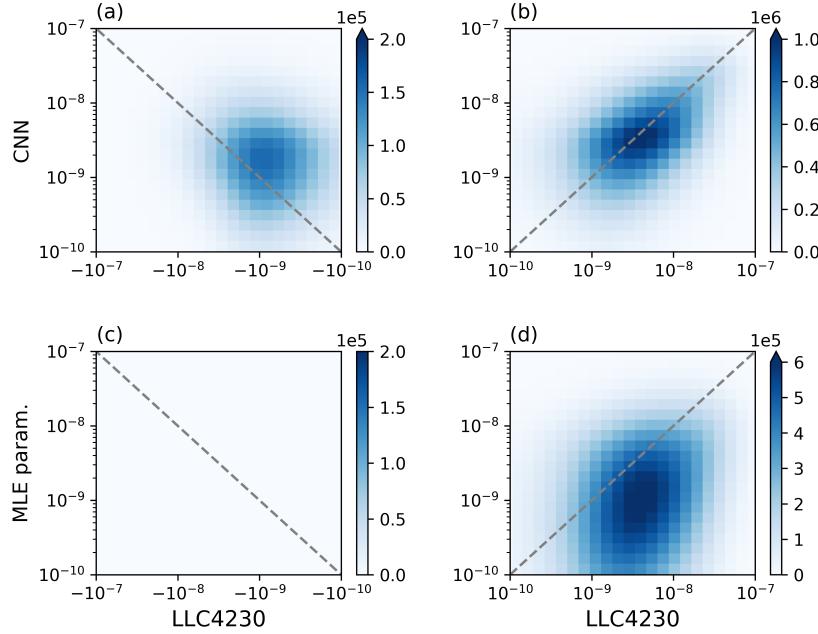


Figure 5. Joint histogram between the $1/4^\circ$ CNN prediction and LLC4230 data (a) and (b) and between the MLE parameterization and LLC4230 data (c) and (d). Left panels (a, c) are negative fluxes and (b, d) are positive.

ing data, retraining the CNN, and examining CNN prediction skill on the unseen regional or seasonal data.

3.1 Holding out seasonality from training

Submesoscale fluxes tend to be strongest during months of winter and spring (Callies et al., 2015; Johnson et al., 2016, e.g.). The strong seasonality is also observable in the climatology presented in figure 6 with respect to the associated hemisphere. 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 generalize on unseen seasonal variability, with skill given in terms of the R^2 value. We thus create two new training and test datasets 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 that we have removed equatorial regions from the analysis entirely as the submesoscale equatorial seasonality is less trivial. The remainder of the time series—e.g. 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.

The upper panel in figure 8 shows the results of the two seasonality experiments. R^2 values of the CNN prediction on the unseen summer are found to be relatively skillful across resolutions and over the different regions included in the seasonality experi-

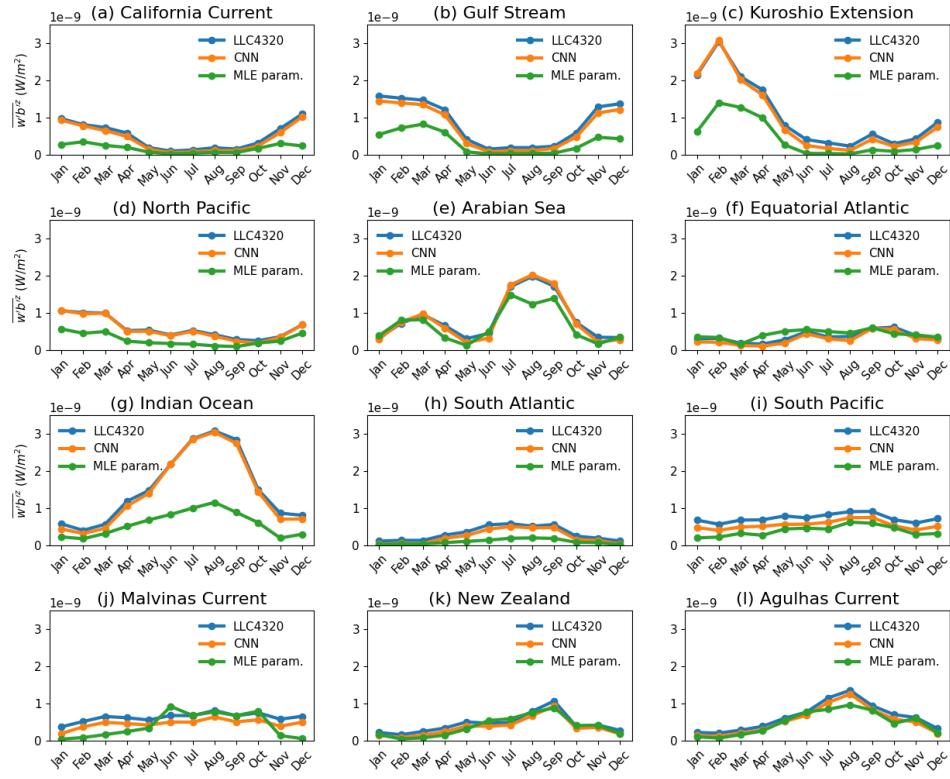


Figure 6. Monthly averages of $\overline{w'b'^z}$ as given per region by the LLC4320 data (blue), the CNN prediction on unseen data (orange), and the MLE parameterization (green).

ment. Contrarily, CNN predictions on winter months do not yield as skillful results, with consistently equal or lower R^2 value. Although not entirely surprising, these results suggest that the CNN is able to generalize and make skillful predictions when the strongest fluxes, generally exhibited during winter and spring months, are included in the training data.

3.2 Holding out regions from training

We perform another set of experiments to test the ability of the CNN to generalize on regions that are not included in the training data. We thus generate 12 new datasets that correspond to removing one region at a time from the training dataset. We retrain the CNN in 12 different experiments, and make predictions on a different unseen region each time.

The lower panel in figure 8 illustrates that R^2 values of the CNN on the unseen regional data remain consistent with those found on the full training set (figure 7) across resolutions 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. Thus, the learned relation-

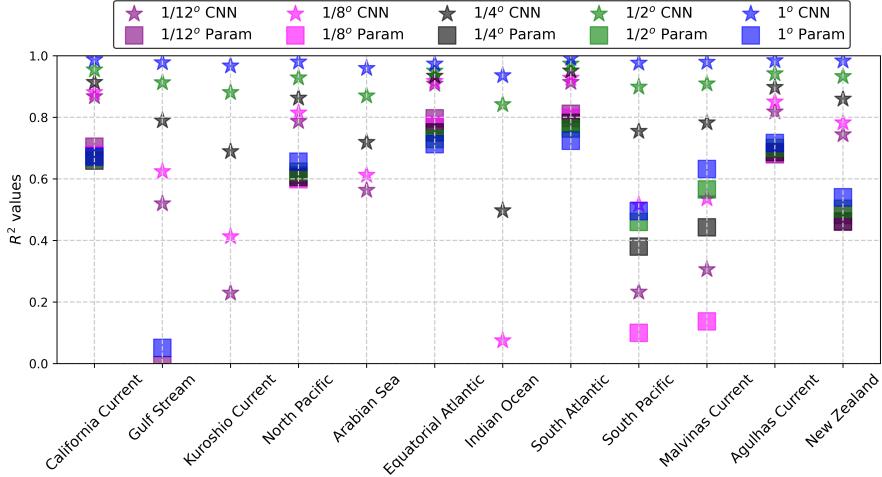


Figure 7. R^2 values of CNN prediction (stars) and the MLE parameterization estimates (squares) over all regions. Colors represent the different resolution experiments. Note that R^2 a point-wise estimate and not an average quantity as in figure 6.

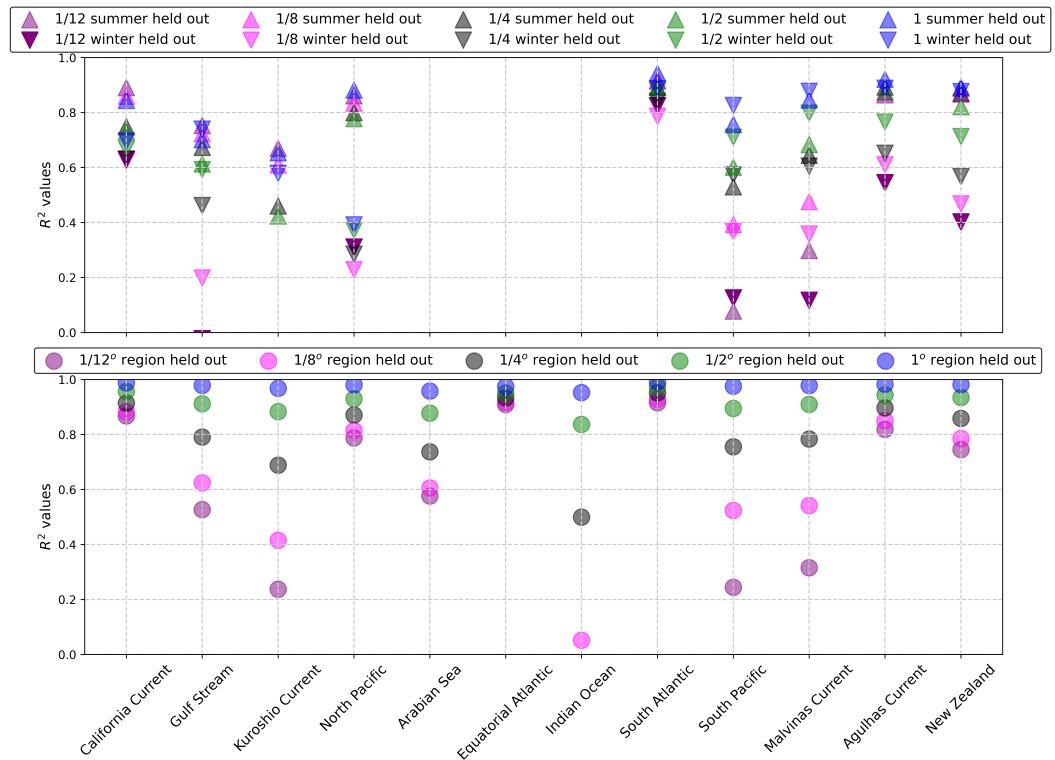


Figure 8. R^2 values of seasonality and location sensitivity experiments. CNN prediction on unseen data: summer (upward triangle), winter (downward triangle), and each region (circles). Colors represent the different resolution experiments.

ships between the input features and $\overline{w'b^z}$ by the CNN are realizable across dynamical regimes, especially in the low resolution experiments.

4 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 physical variables used as inputs and $w'b^z$ 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 why. Many methods exist that help explain and interpret the dependency of CNN outputs to its inputs. Here, we have chosen two complimentary methods that help gain insight on the learned local and non-local relationships and the importance of individual inputs to $w'b^z$.

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 $1 - R^2$ value. If removing a certain input feature results in a very low R^2 value, or high $1 - R^2$ value, it indicates that the CNN strongly depends on said input feature. The top panel in figure 9 displays the $1 - R^2$ value given from the ablation experiments for each input features, and compared across all resolutions.

Notably, removing strain, $\bar{\sigma}^z$, as an input feature results in high $1 - R^2$ values, demonstrating the largest reduction in skill consistently across all resolutions. Interestingly, there appears to be very little sensitivity to the removal of any other input feature, including those used by the MLE parameterization. These results strongly suggest that the primary reason the CNN predictions surpass those of the MLE parameterization, are due to the newly-captured non-local relationship between submesoscale vertical buoyancy fluxes and the large scale strain field. Note that the relevance of strain to submesoscale fluxes is not a new result (e.g., Balwada et al., 2021; J. Zhang et al., 2023), but these findings emphasize the relevance of strain to improving submesoscale $w'b^z$ parameterizations.

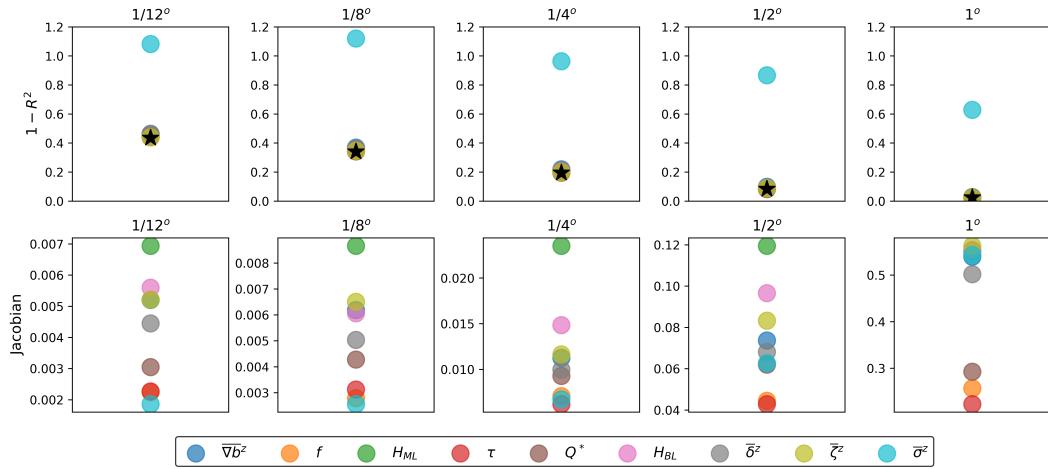


Figure 9. Top panels shows $1 - R^2$ values of ablation experiments where we remove one input feature at a time over all resolutions. Note a score of 1 indicates skill dropped and feature is very important. Bottom panels are the same but for the Jacobian $\nabla \mathbf{x} S(\mathbf{X})$ of the output with respect to individual inputs.

4.2 Sensitivity of output to relative to input features

We next apply a complimentary method to the ablation experiment above. We compute the Jacobian, defined as the gradients of the CNN output with respect to the input features, $\nabla_{\mathbf{X}} S(\mathbf{X})$. Note that unlike the ablation experiment, where we examined the CNN skill on the full output domain, the Jacobian is a local metric, considering only the sensitivity of a single output point to a single point in the input feature map. The Jacobian is especially useful to evaluate the point-wise sensitivity of the output to each input feature by taking gradients along the CNN weights (e.g., Ross et al., 2023). We compute the Jacobian over the entire unseen test dataset, 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 values given by the ablation experiments in figure 9, where for the Jacobian, a high score indicates that the CNN output, $w'b^z$, is sensitive to local changes in a certain input. We find that the highest-ranked input feature, for which $w'b^z$ is most sensitive to, is the mixed layer depth, H_{ML} , which is generally a 1D, local, physical property, determined by surface forcing. The sensitivity to mixed layer depth is followed by sensitivity to boundary layer depth, H_{BL} , the buoyancy gradient and vorticity. Note that $w'b^z$ does not appear to be sensitive to local-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° resolution experiment that the Jacobian exhibits sensitivity of $w'b^z$ to vorticity, divergence, and strain, indicating that these fields are significantly non-local unless the largest filter is applied.

To further understand the relevance of locality, in figure 10, we examine the Jacobian of the output center point with respect to the full domain of each input feature. Figure 10a shows an example for the buoyancy gradient input feature, 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. Figure 10b is an example for the $1/4^\circ$ resolution case, where on average, we find that 7 grid points away from the center is sufficient for capturing 90% of the Jacobian fraction, e.g. the 90% of sensitivity between the output and input features. This relatively local receptive field is also found to be consistent across resolutions (figure 10c), despite the varying importance of input features found previously.

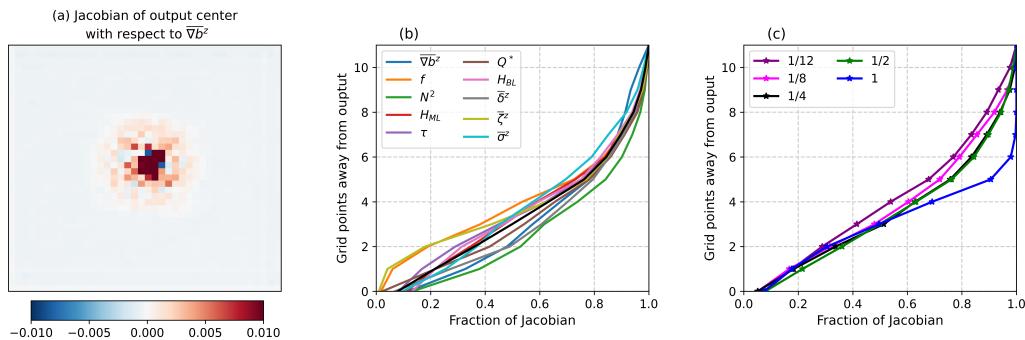


Figure 10. (a) Example of Jacobian of center output point with respect to $1/4^\circ$ buoyancy gradient input field. (b) Radial grid point number compared with fraction of $\nabla_{\mathbf{X}} S(\mathbf{X})$. Black is the mean over all inputs. (c) Average fraction of Jacobian for all resolutions experiments. Note that the Jacobian is computed in normalized units.

5 Discussion and Conclusions

The parameterization for submesoscale vertical buoyancy fluxes 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 this work, we introduce an improved parameterization based on a data-driven approach, where a CNN is trained to learn submesoscale vertical buoyancy fluxes, $\overline{w'b'}$, given by large scale variables that help set it. The subgrid flux, $\overline{w'b'}$, is inferred by the CNN as a function of 9 large-scale input features with known relevance to submesoscale vertical buoyancy flux: $\overline{\nabla b}^z$, 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 resolution experiments of $1/12^\circ$, $1/8^\circ$, $1/4^\circ$, $1/2^\circ$, 1° and compare with the Bodner et al. (2023) formulation of the MLE parameterization. We consistently find that the CNN predictions improves on the MLE parameterization, with higher R^2 values across all regions, seasons, and resolutions tested in this study. We additionally perform several sensitivity experiments, where we test the CNN's ability to generalize on regional or temporal data held out during training. It is found that the CNN, in particular in the low resolution experiments, is able to make skillful predictions on unseen data as long as it is trained on seasons where submesoscales are most active, which generally corresponds to winter and spring months. We have also shown that the CNN is able to generalize on most regions that were held out during training, with dependency on the dominant scales of variability in $\overline{w'b'}$.

The significant improvement on the MLE parameterization indicates that the CNN has learned meaningful relationships between the input features and $\overline{w'b'}$ that are likely not captured by the MLE parameterization, and are able to generalize over widely different dynamical regimes. Thus, we applied two complimentary explainability methods which enable a closer look at the relationships between the CNN output, $\overline{w'b'}$, and the CNN input features. The point-wise dependency is computed by the Jacobian of the output with respect to inputs along the CNN weights. We find that a strong dependency between $\overline{w'b'}$ and the mixed layer depth emerges. In contrast, from a set of ablation experiments, we find that including strain as an input feature significantly improves skill in terms of R^2 values. To summarize, we find that the CNN exhibits strong dependency on the local relationship between $\overline{w'b'}$ and the mixed layer depth, a 1D property driven by surface forcing, and non-locally on the large scale strain field, a variable currently missing from the MLE parameterization in GCMs. An interesting application here would be to test whether this result holds in other submesoscale permitting simulations, such as those compared in Uchida et al. (2022). Other parameterizations such as J. Zhang et al. (2023) have suggested a theoretical formulation that includes a relationship between submesoscale vertical buoyancy flux with the strain field. An equation discovery approach (e.g., Zanna & Bolton, 2020), may enable a closer comparison with J. Zhang et al. (2023), and whether a strong relationship between strain and submesoscale fluxes as well as a new shape function emerge in a similar fashion.

We have shown that the CNN improves on the MLE parameterization in an offline setting. A next natural step is to explore the implications of better captured $\overline{w'b'}$ in a GCM and compare with the MLE parameterization online. We have designed our method to correspond with the existing implementation of the MLE parameterization in GCMs, where $\overline{w'b'}$ in (S3) 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 $\overline{w'b'}$, 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 vertical buoyancy fluxes already estimated in boundary layer turbulence parameterizations (Large et al., 1994; Reichl & Hallberg, 2018). The

exact formulation, implementation, and evaluation of impact on climate variables is left for future work.

Acknowledgements

AB was supported by a grant from the Simons Foundation: award number 855143. AB, DB and LZ received M²LInES research funding through the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. We thank members of the M²LInES 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.

References

- 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.
- 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.
- Bachman, S. D., & Klocker, A. (2020). Interaction of jets and submesoscale dynamics leads to rapid ocean ventilation. *Journal of Physical Oceanography*, 50(10), 2873–2883.
- 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.
- 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.
- Boccaletti, G., Ferrari, R., & Fox-Kemper, B. (2007). Mixed layer instabilities and restratification. *Journal of Physical Oceanography*, 37(9), 2228–2250.
- Bodner, A. S., Fox-Kemper, B., Johnson, L., Van Roekel, L. P., McWilliams, J. C., Sullivan, P. P., ... Dong, J. (2023). Modifying the mixed layer eddy parameterization to include frontogenesis arrest by boundary layer turbulence. *Journal of Physical Oceanography*, 53(1), 323–339.
- Bodner, A. S., Fox-Kemper, B., Van Roekel, L. P., McWilliams, J. C., & Sullivan, P. P. (2020). A perturbation approach to understanding the effects of turbulence on frontogenesis. *Journal of Fluid Mechanics*, 883.
- 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.
- 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.
- Callies, J., & Ferrari, R. (2018). Baroclinic instability in the presence of convection. *Journal of Physical Oceanography*, 48(1), 45–60.
- Callies, J., Ferrari, R., Klymak, J. M., & Gula, J. (2015). Seasonality in submesoscale turbulence. *Nature communications*, 6(1), 6862.
- 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.
- Capet, X., McWilliams, J. C., Molemaker, M. J., & Shchepetkin, A. (2008). Mesoscale to submesoscale transition in the california current system. part ii: Frontal processes. *Journal of Physical Oceanography*, 38(1), 44–64.
- 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.
- 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.
- 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: Oceans*, 109(C12).
- Dong, J., Fox-Kemper, B., Zhang, H., & Dong, C. (2020). The seasonality of submesoscale energy production, content, and cascade. *Geophysical Research Letters*,

- 47(6), e2020GL087388.
- Ferrari, R., McWilliams, J. C., Canuto, V. M., & Dubovikov, M. (2008). Parameterization of eddy fluxes near oceanic boundaries. *Journal of Climate*, 21(12), 2770–2789.
- 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.
- Fox-Kemper, B., Danabasoglu, G., Ferrari, R., Griffies, S., Hallberg, R., Holland, M., ... Samuels, B. (2011). Parameterization of mixed layer eddies. iii: Implementation and impact in global ocean climate simulations. *Ocean Modelling*, 39(1-2), 61–78.
- 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.
- Fox-Kemper, B., Johnson, L., & Qiao, F. (2022). Ocean near-surface layers. In *Ocean mixing* (pp. 65–94). Elsevier.
- Frankignoul, C., & Hasselmann, K. (1977). Stochastic climate models, part ii application to sea-surface temperature anomalies and thermocline variability. *Tellus*, 29(4), 289–305.
- 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*, 1–42.
- Gent, P. R., & Mcwilliams, J. C. (1990). Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, 20(1), 150–155.
- Griffies, S. M. (1998). The gent–mcwilliams skew flux. *Journal of Physical Oceanography*, 28(5), 831–841.
- Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, 13(9), e2021MS002534.
- Gula, J., Taylor, J., Shcherbina, A., & Mahadevan, A. (2022). Submesoscale processes and mixing. In *Ocean mixing* (pp. 181–214). Elsevier.
- Hoskins, B. J., & Bretherton, F. P. (1972). Atmospheric frontogenesis models: Mathematical formulation and solution. *Journal of the atmospheric sciences*, 29(1), 11–37.
- IPCC. (2019). *Special report on the ocean and cryosphere in a changing climate* [Book]. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: 10.1017/9781009157964
- IPCC. (2021). *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (Vol. In Press) [Book]. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: 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.
- 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.
- Lapeyre, G., Klein, P., & Hua, B. L. (2006). Oceanic restratification forced by surface frontogenesis. *Journal of Physical Oceanography*, 36(8), 1577–1590.
- 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.
- Loose, N., Abernathey, R., Grooms, I., Busecke, J., Guillaumin, A., Yankovsky, E.,

- ... others (2022). Gcm-filters: A python package for diffusion-based spatial filtering of gridded data. *Journal of Open Source Software*, 7(70).
- Mahadevan, A. (2016). The impact of submesoscale physics on primary productivity of plankton. *Annual review of marine science*, 8, 161–184.
- Mahadevan, A., Tandon, A., & Ferrari, R. (2010). Rapid changes in mixed layer stratification driven by submesoscale instabilities and winds. *Journal of Geophysical Research: Oceans*, 115(C3).
- McWilliams, J. C. (2016). Submesoscale currents in the ocean. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472(2189), 20160117.
- McWilliams, J. C. (2017). Submesoscale surface fronts and filaments: Secondary circulation, buoyancy flux, and frontogenesis. *Journal of Fluid Mechanics*, 823, 391–432.
- Menemenlis, D., Campin, J.-M., Heimbach, P., Hill, C., Lee, T., Nguyen, A., ... Zhang, H. (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 2km grid oceanographic version 1.0. Ver.
- 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*.
- Reichl, B. G., & Hallberg, R. (2018). A simplified energetics based planetary boundary layer (epbl) approach for ocean climate simulations. *Ocean Modelling*, 132, 112–129.
- 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.
- Rocha, C. B., Gille, S. T., Chereskin, T. K., & Menemenlis, D. (2016). Seasonality of submesoscale dynamics in the kuroshio extension. *Geophysical Research Letters*, 43(21), 11–304.
- 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.
- 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*.
- Schubert, R., Schwarzkopf, F. U., Baschek, B., & Biastoch, A. (2019). Submesoscale impacts on mesoscale agulhas dynamics. *Journal of Advances in Modeling Earth Systems*, 11(8), 2745–2767.
- 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.
- 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.
- Souza, A. N., Wagner, G., Ramadhan, A., Allen, B., Churavy, V., Schloss, J., ... others (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.
- Srinivasan, K., Barkan, R., & McWilliams, J. C. (2023). A forward energy flux at submesoscales driven by frontogenesis. *Journal of Physical Oceanography*, 53(1), 287–305.
- Stanley, Z., Grooms, I., Kleiber, W., Bachman, S., Castruccio, F., & Adcroft, A. (2020). Parameterizing the impact of unresolved temperature variability on the

- large-scale density field: Part 1. theory. *Journal of Advances in Modeling Earth Systems*, 12(12), e2020MS002185.
- Su, Z., Torres, H., Klein, P., Thompson, A. F., Siegelman, L., Wang, J., ... Hill, C. (2020). High-frequency submesoscale motions enhance the upward vertical heat transport in the global ocean. *Journal of Geophysical Research: Oceans*, 125(9), e2020JC016544.
- 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.
- Taylor, J. R., & Thompson, A. F. (2023). Submesoscale dynamics in the upper ocean. *Annual Review of Fluid Mechanics*, 55, 103–127.
- Thomas, L. N. (2005). Destruction of potential vorticity by winds. *Journal of physical oceanography*, 35(12), 2457–2466.
- Thomas, L. N., Tandon, A., & Mahadevan, A. (2008). Submesoscale processes and dynamics. *Ocean modeling in an Eddying Regime*, 177, 17–38.
- Thomas, L. N., Taylor, J. R., Ferrari, R., & Joyce, T. M. (2013). Symmetric instability in the gulf stream. *Deep Sea Research Part II: Topical Studies in Oceanography*, 91, 96–110.
- Torres, H. S., Klein, P., Menemenlis, D., Qiu, B., Su, Z., Wang, J., ... Fu, L.-L. (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.
- Uchida, T., Le Sommer, J., Stern, C., Abernathey, R., Holdgraf, C., Albert, A., ... others (2022). Cloud-based framework for inter-comparing submesoscale permitting realistic ocean models. *Geoscientific Model Development Discussions*, 1–32.
- 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.
- Wenegrat, J. O., & Thomas, L. N. (2020). Centrifugal and symmetric instability during ekman adjustment of the bottom boundary layer. *Journal of Physical Oceanography*, 50(6), 1793–1812.
- 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.
- 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.
- Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17), e2020GL088376.
- Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., & Zanna, L. (2023). Implementation and evaluation of a machine learned mesoscale eddy parameterization into a numerical ocean circulation model. *arXiv preprint arXiv:2303.00962*.
- 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.
- Zhu, R., Li, Y., Chen, Z., Du, T., Zhang, Y., Li, Z., ... Wu, L. (2023). Deep learning improves reconstruction of ocean vertical velocity. *Geophysical Research Letters*, 50(19), e2023GL104889.

Supplementary Material

The formula for Ψ_{MLE} in Fox-Kemper et al. (2008) is provided by a scaling for,

$$\overline{w' b'}^z \propto \frac{H_{ML} |\nabla_H b|^z}{|f|}, \quad (\text{S3})$$

where H_{ML} is the mixed layer depth, f is the Coriolis parameter, w is vertical velocity, b is buoyancy, and $\nabla_H b$ is the horizontal buoyancy gradient. We follow the notation in Fox-Kemper et al. (2008), where the horizontal spatial resolution of the GCM is denoted (\cdot) and $(\cdot)'$ is the unresolved subgrid variable. Superscript z represents a vertical average over the mixed layer depth. The scaling for submesoscale vertical buoyancy flux represents the bulk extraction of potential energy by MLEs within the mixed layer. A shape function $\mu(z)$ estimates the depth, z , at which the MLE fluxes are activated,

$$\mu(z) = \max \left(0, \left[1 - \left(\frac{2z}{H_{ML}} + 1 \right)^2 \right] \left[1 + \frac{5}{21} \left(\frac{2z}{H_{ML}} + 1 \right)^2 \right] \right) \quad (\text{S4})$$

where $\mu(z)$ is set to vanish at the surface and below the mixed layer H_{ML} .

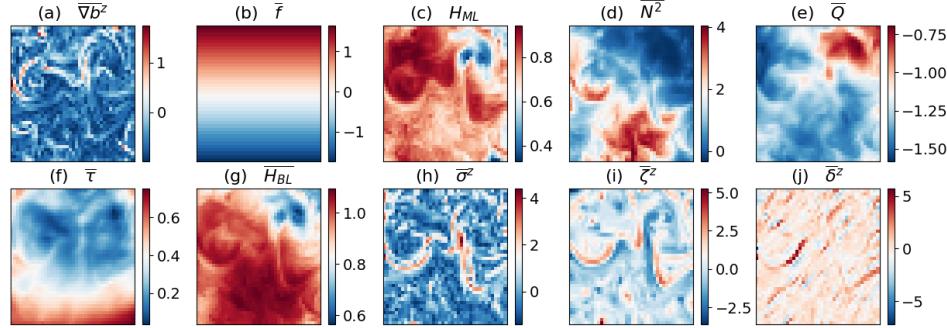


Figure S1. Snapshots of the $1/4^\circ$ normalized inputs.

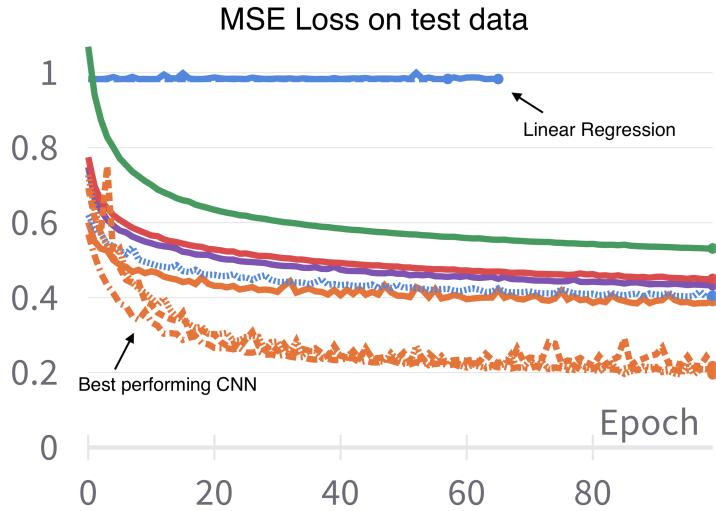


Figure S2. Example of loss function during the hyper-parameter sweep of the CNN. Solid blue line is the case of a simple linear regression, which is not sufficient to reduce the MSE loss.

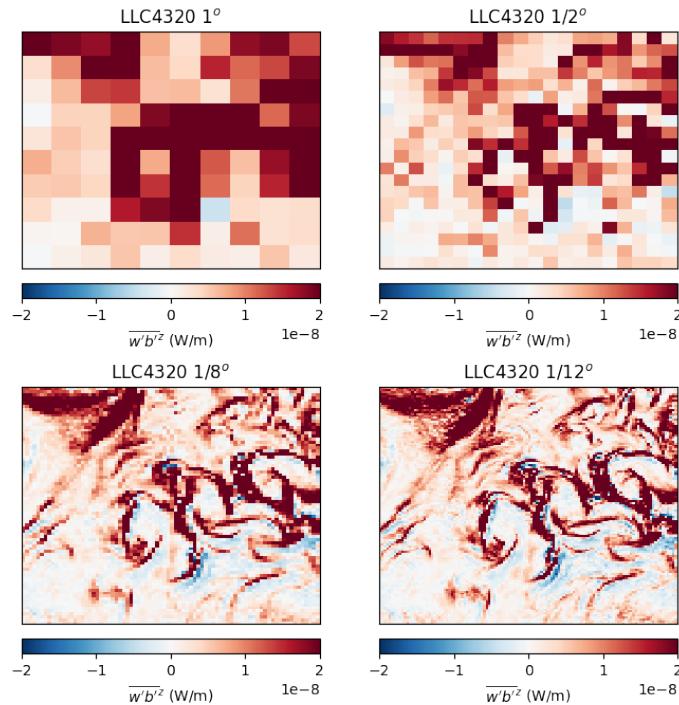


Figure S3. Same as figure 4a for resolutions of $1^\circ, 1/2^\circ, 1/8^\circ, 1/12^\circ$