Earth and Space Science

RESEARCH ARTICLE
10.1029/2022EA002618

Identifying Ocean Submesoscale Activity From Vertical Density Profiles Using Machine Learning

Leyu Yao[1], **John R. Taylor**[1], **Dani C. Jones**[2,3], **and Scott D. Bachman**[4]

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK, [2]Cooperative Institute for Great Lakes Research (CIGLR), University of Michigan, Ann Arbor, MI, USA, [3]British Antarctic Survey, NERC, UKRI, Cambridge, UK, [4]National Center for Atmospheric Research, Boulder, CO, USA

Key Points:

- A machine learning algorithm is used to identify the signature of submesoscale eddies in the ocean
- The method is applied to two model data set and shown to be a good predictor of submesoscale activity
- The advantage of the method is that it can be applied to vertical density profiles without needing information about horizontal variation

Correspondence to:

J. R. Taylor,
j.r.taylor@damtp.cam.ac.uk

Citation:

Yao, L., Taylor, J. R., Jones, D. C., & Bachman, S. D. (2025). Identifying ocean submesoscale activity from vertical density profiles using machine learning. *Earth and Space Science*, *12*, e2022EA002618. https://doi.org/10.1029/2022EA002618

Received 24 FEB 2023
Accepted 2 JAN 2025

Author Contributions:

Conceptualization: Leyu Yao, John R. Taylor
Data curation: Scott D. Bachman
Formal analysis: John R. Taylor
Investigation: Leyu Yao
Methodology: Leyu Yao
Supervision: John R. Taylor, Dani C. Jones
Visualization: Leyu Yao
Writing – original draft: Leyu Yao, John R. Taylor
Writing – review & editing: Leyu Yao, John R. Taylor, Dani C. Jones, Scott D. Bachman

Abstract Submesoscale eddies are important features in the upper ocean where they mediate air-sea exchanges, convey heat and tracer fluxes into ocean interior, and enhance biological production. However, due to their small size (0.1–10 km) and short lifetime (hours to days), directly observing submesoscales in the field generally requires targeted high resolution surveys. Submesoscales increase the vertical density stratification of the upper ocean and qualitatively modify the vertical density profile. In this paper, we propose an unsupervised machine learning algorithm to identify submesoscale activity using vertical density profiles. The algorithm, based on the profile classification model (PCM) approach, is trained and tested on two model-based data sets with vastly different resolutions. One data set is extracted from a large-eddy simulation (LES) in a 4 km by 4 km domain and the other from a regional model for a sector in the Southern Ocean. We show that the adapted PCM can identify regions with high submesoscale activity, as characterized by the vorticity field (i.e., where surface vertical vorticity ζ is similar to Coriolis frequency f and Rossby number $Ro = \zeta/f \sim \mathcal{O}(1)$), using solely the vertical density profiles, without any additional information on the velocity, the profile location, or horizontal density gradients. The results of this paper show that the adapted PCM can be applied to data sets from different sources and provides a method to study submesoscale eddies using global data sets (e.g., CTD profiles collected from ships, gliders, and Argo floats).

Plain Language Summary In this paper we describe a new method based on Machine Learning techniques for identifying the tell-tale signatures of submesoscale (1–10 km) eddies from individual density profiles. Our method is based on the hypothesis that submesoscale eddies alter the shape of the buoyancy profile within the surface mixed layer. We start by re-scaling (normalizing) buoyancy and depth within the mixed layer so that our algorithm classifies each profile based only on the shape of the normalized buoyancy profile in the mixed layer. We then use the profile classification method and Gaussian mixture model to classify buoyancy profiles from two very different model data sets: a high resolution LES that resolves submesoscales and the largest 3D turbulence, and a regional model of the Scotia Sea that resolves mesoscale and submesoscale eddies. In both cases, our algorithm shows significant skill in detecting the presence of submesoscale eddies from individual buoyancy profiles.

1. Introduction

Submesoscales are dynamical features with horizontal scales between 0.1 and 10 km, vertical scales of 0.01–1 km, and life cycles spanning hours to days (McWilliams, 2016). Often found on the periphery of mesoscale eddies and other regions with strong horizontal density gradients, submesoscales arise from a variety of instabilities in the mixed layer fueled by energy from fronts, including baroclinic instability (Boccaletti et al., 2007), symmetric instability (SI) (Taylor & Ferrari, 2009), and centrifugal instability (Gula et al., 2016). Submesoscales have a crucial influence on the stratification of the upper ocean, especially the depth and properties of the ocean surface mixed layer. The mixed layer is characterized by homogeneous water properties in the vertical direction due to vigorous turbulent mixing. The mixed layer transfers particles, momentum, and energy between the atmosphere and ocean interior, and the varying thickness of the mixed layer controls the amount of heat content and energy that can interact with the atmosphere. Many processes can affect the mixed layer depth (MLD) both temporally and spatially, such as surface wind stress, buoyancy forcing, and internal waves (de Boyer Montégut et al., 2004; McCreary et al., 2001; Mellor & Durbin, 1975; Panassa et al., 2018). In particular, coherent submesoscale eddies arising through mixed layer baroclinic instability can lead to a persistent stable density stratification in the surface mixed layer, resulting in a shallower mixed layer and an altered density profile within the eddy (Mahadevan et al., 2010; McWilliams, 2016; Taylor, 2016; Taylor & Ferrari, 2010). Submesoscale eddies also play an important role in supplying nutrients to phytoplankton in sunlit layers, as well as enhancing biological production and bio-diversity in the ocean (Lévy et al., 2018; Mahadevan & Tandon, 2006; Whitt et al., 2019). It was revealed in Omand et al. (2015) through numerical simulations and observations that during the spring transition, surface water that is rich in particulate organic carbon (POC) often descends along the periphery of submesoscale eddies, which accounts for as much as half of the spring-time export of POC from the highly productive subpolar oceans but is unresolved in global carbon cycle models. Further, in a recent study, Taylor et al. (2020) used high resolution large-eddy simulations (LES) to study the influence of submesoscale eddies on the concentration and export of sinking particles from the ocean mixed layer. It was found that the development of submesoscale eddies causes the re-stratification of the mixed layer and reduces the rate of vertical mixing, enhancing the rate of export from gravitational settling (Taylor et al., 2020).

Despite the critical roles that submesoscales play in the ocean ecosystem and the ocean carbon cycle, many unanswered questions remain, particularly about their distribution and influence on a global scale. Since submesoscales are small and relatively short-lived, directly observing them typically requires targeted high-resolution surveys. Global data sets typically lack the resolution in space and time to capture submesoscales. In particular, Argo and BioGeoChemical-Argo (BGC-Argo) floats provide global coverage of hydrographic and biogeochemical properties, but the profiles collected by Argo and BGC-Argo floats do not have sufficient spatial or temporal coverage to capture submesoscales (Wong et al., 2020). Global models are just beginning to partially resolve the submesoscale range (Rocha et al., 2016) and submesoscales are still unresolved in climate models (Dong et al., 2020). For example, Garabato et al. (2022) compares the submesoscale-resolving observations obtained from nine bottom-anchored moorings deployed in the northeast Atlantic Ocean from September 2012 to September 2013 as part of the Ocean Surface Mixing–Ocean Submesoscale Interaction Study (OSMOSIS) project (Buckingham et al., 2016; Damerell et al., 2016; Thompson et al., 2016) with the output of the LLC4320 simulation (Menemenlis et al., 2021; Torres et al., 2018), one of the most realistic high-resolution submesoscale-permitting global ocean models. They concluded that current submesoscale-permitting models may substantially understate the downscale kinetic energy transfer from the mesoscale to submesoscale. Thus, it would be extremely useful if we could determine whether an individual vertical profile belongs to a region with high submesoscale activity. This would permit global surveys of submesoscale eddies and their influence on ocean physics and biogeochemistry using existing global data sets.

To achieve this goal, we adapt an unsupervised machine learning algorithm proposed by Maze et al. (2017) and used to analyze data from ocean floats in recent years (Houghton & Wilson, 2020; Jones et al., 2019; Maze et al., 2017; Rosso et al., 2020) to classify ocean vertical density profiles in two existing model-based data sets from Taylor et al. (2020) and Bachman et al. (2017). Our algorithm is able to classify density profiles according to the presence or absence of submesoscale activity without any information on the geographic locations of the profiles. Although here we only apply the method to model data, this study provides an important step toward using sparse global data sets such as Argo and BGC-Argo to study submesoscale eddies.

The fundamental insight behind the proposed method is that while submesoscales have been known to effectively re-stratify the upper ocean (Gula et al., 2022), they also qualitatively change the shape of the vertical density profile in the upper ocean. This is illustrated in Figure 1 which shows six buoyancy profiles inside and outside of a submesoscale eddy (three profiles each) from a very high resolution LES reported in Taylor et al. (2020). Details of the model will be discussed in the next section. As can be seen in Figure 1, the increase in stratification from about 200 to 400 m depth is less abrupt inside the eddy and more abrupt outside the eddy, while the change in buoyancy inside and outside the eddy is quite similar over the upper 400 m. In addition, stable stratification reaches a shallower depth inside the eddy. By "learning" the qualitative differences between these density profiles, a machine learning algorithm should be able to identify the signature of submesoscale activity in vertical density profiles.

Below, Section 2 introduces the model-based data sets used in this paper, Section 3 demonstrates the unsupervised machine learning algorithm for training the statistical models, and Section 4 presents the results of applying the statistical models with different sets of parameters to the data sets and the interpretation of the results. Section 5 gives the conclusion of this paper and directions for future work.

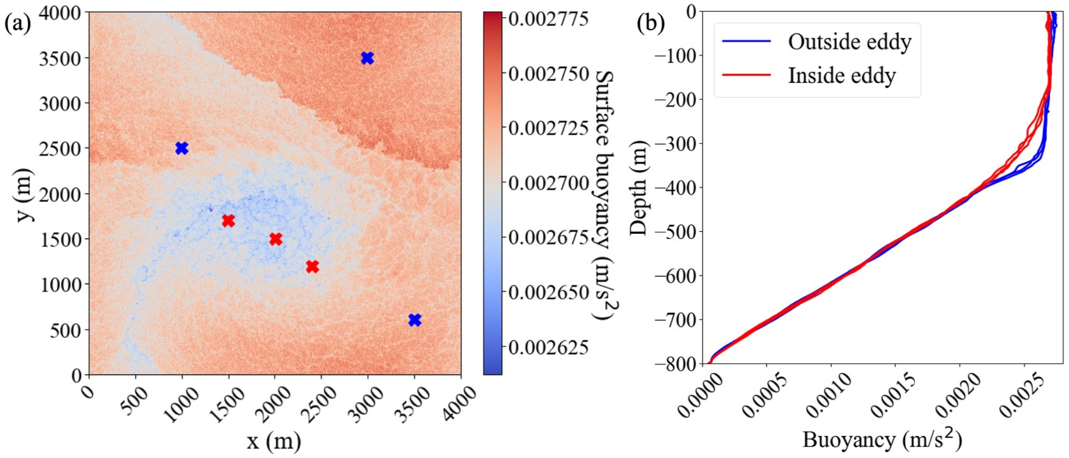

Figure 1. (a) Surface buoyancy with locations of sample profiles inside submesoscale eddy (red) and outside submesoscale eddy (blue), and (b) sample profiles from inside and outside submesoscale eddy.

2. Data

Here, we apply our new method to two model-based data sets. The first data set is extracted from the high-resolution, idealized domain LES discussed in Taylor et al. (2020). The second data set is part of the series of offline-nested numerical simulations with realistic bathymetry, specifically of the Drake Passage and Scotia Sea regions of the Southern Ocean (SO) (Bachman et al., 2017). We will refer to these two data sets as the "LES data set" and the "SO data set," respectively. Here, we briefly describe each model. More details can be found in the corresponding papers.

The first main difference between the two models is their domain size and resolution. The defining characteristic of an LES is that it explicitly resolves the largest three-dimensional turbulent motions. This has the key advantage that the influence of submesoscale dynamics on boundary layer turbulence is explicitly captured by the model, although the computational expense required to achieve this resolution places a significant constraint on the grid spacing and hence the model domain. The LES model in Taylor et al. (2020) simulates an idealized patch of open ocean with a domain size of 4 km in both horizontal directions and a depth of 800 m. Each horizontal direction is discretized into 1,024 grid points, so that the horizontal resolution is 3.9 m in both directions. There are 257 uniformly-sized grid cells in the vertical direction, giving a vertical resolution of about 3.1 m. The simulation has a background horizontal buoyancy gradient and is forced with a constant surface cooling by applying a constant negative buoyancy flux at the sea surface. The wind stress is set to zero in the LES and the simulation starts with a weakly stratified layer of 300 m depth on top of a strongly stratified pycnocline of 500 m thickness. The computational domain can be seen as an idealized representation of a small patch of open ocean but without any larger-scale variability. A submesoscale eddy develops about 3 days after initialization of the LES. Further details about the model setup and results can be found in Taylor et al. (2020). Here, we extract data from the LES from three-dimensional fields saved at six time steps ranging from about 5 to 6 days after the initialization of the LES, all of which are after the submesoscale eddy has developed.

The second model that we use was described in Bachman et al. (2017) and used a series of offline-nested simulations with the MIT General Circulation Model (MITgcm) to study submesoscale activity in the Drake Passage and Scotia Sea regions of the Southern Ocean. Simulations of horizontal resolutions 1/12°, 1/24°, 1/48°, 1/96°, and 1/192° were run in the study. The lowest resolution model (1/12°) was initialized from an ocean state estimate from 23 April 2015 provided by the Copernicus Marine Environment Monitoring Service Global Ocean 1/12° Physics Analysis (CMEMS), which is produced by Mercator Ocean (http://marine.copernicus.eu), to simulate conditions found during the Surface Mixed Layer at Submesoscales (SMILES) research cruise which took place at this time. The open boundary conditions were one-way nested, updated once per day, and relaxed to the CMEMS state estimate of the subsequent day over a 2° wide sponge region on all edges of the domain. The model was forced with wind stress and net surface heat flux sampled daily from the European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric analysis, which were interpolated from 1/4° to the appropriate resolution. The data we use in this paper come from the simulation with horizontal resolution of 1/192° (roughly 500 m) and covering the area from 60°S to 48°S and 80°W to 40°W. The resolution was sufficient to capture the larger end of the submesoscale range. The vertical spacing of the grid points is 5 m over the top 100 m, and the grid spacing increases by a factor of 1.1 for each level below, up to a maximum spacing of 50 m. The depth contains 125 levels, with a maximum depth of 4,600 m. We use density profiles calculated from temperature and salinity using the GSW Oceanographic Toolbox (McDougall & Barker, 2011) and extracted from the 1/192° model on May 13, which is sufficiently long after the model was initialized with an eddying state estimate on April 23 for submesoscale eddies to develop, and is one of a handful of dates when the full three-dimensional fields were saved from the simulations reported in Bachman et al. (2017).

It is important to note that since the LES is forced with a constant net surface heat flux and the SO model is forced with daily snapshots of wind stress and surface heat forcing from ECMWF, neither of the models that we use includes a diurnal cycle. We are therefore not able to test whether our method is able to distinguish between re-stratification associated with the diurnal cycle from submesoscale eddies. This will likely be an important consideration when applying our method to ocean observations, but we leave this test for future work.

Small-scale turbulence is represented very differently in the two models. The LES model solves the non-hydrostatic and Boussinesq equations with sufficient resolution to capture the largest and most energetic turbulent overturning motions, capturing dynamical interactions between boundary layer turbulence and submesoscales (Taylor et al., 2020). The MITgcm in Bachman et al. (2017) solves the Boussinesq hydrostatic equations and uses the K-Profile Parametrization (KPP) to represent the vertical mixing of momentum and tracers in the surface boundary layer (Bachman et al., 2017). KPP models mixing in the surface boundary layer and the ocean interior using empirical flux curves, surface forcing, and the model shear and stratification (Large et al., 1994). We expect that the differences in the representations of small-scale turbulence and mixing between the two data sets will be reflected to some degree in the vertical structure of the density profiles. Applying our method to both models allows us to test its robustness to different representations of turbulent mixing.

3. Methods

In this section, we describe the unsupervised machine learning algorithm which is based on the profile classification model (PCM) proposed in Maze et al. (2017). We first describe our pre-processing of the model output, which is the key adaptation of the PCM to identify submesoscale activity (Section 3.1). We then introduce the PCM and several key parameters for training the PCMs (Section 3.2). The results of the experiments and discussions are shown (Section 4).

3.1. Pre-Processing of Model Output

The central hypothesis of our method is that submesoscale eddies lead to characteristic changes in the *shape* of the density profile within the mixed layer which can be used to diagnose the presence of submesoscale activity. In addition to their influence on the MLD (defined by a fixed change in density from the ocean surface, see below), submesoscale eddies can lead to a weak but persistent stable density stratification within the mixed layer (Taylor, 2016; Taylor & Ferrari, 2010; Whitt & Taylor, 2017). Other processes that re-stratify the upper ocean lead to qualitatively different density profiles. For example, solar heating can generate a step-like density profile where a warm, shallow mixing layer becomes isolated from the remnant mixed layer (Brainerd & Gregg, 1995).

In order to isolate changes in the density profile with respect to the MLD, we adapt the original algorithm for PCM (Maze et al., 2017) by first re-scaling each density profile based on MLD, effectively adopting a new dimensionless vertical coordinate with values between −1 and 0, where −1 indicates the base of the mixed layer and 0 indicates the surface. Next, we subtract the vertical mean density from each profile before applying the PCM. Without these steps, the PCM would be dominated by changes in the depth and density of the mixed layer, which vary geographically and in time and are not necessarily indicative of submesoscale activity. For example, while submesoscale eddies are known to reduce the MLD, shallow mixed layers also develop in response to seasonal and diurnal heat fluxes.

Various criteria have been used to define the MLD (e.g., de Boyer Montégut et al., 2004). Since submesoscale eddies develop by extracting potential energy, thereby increasing the density stratification in the upper ocean, we use a fixed density change with respect to the surface value to diagnose the MLD. Here, we express the density change in terms of buoyancy, $\Delta b \equiv -g\Delta\rho/\rho_0$, where g is the gravitational acceleration and ρ_0 is a reference density and we use $\Delta b = 3.924 \times 10^{-4}\mathrm{m/s}^2$ which corresponds to a temperature change of about 0.2°C as suggested by de Boyer Montégut et al. (2004). Other frequently used difference-based criteria for the MLD include $\Delta T = 0.8$°C and $\Delta\sigma_t = 0.125\sigma_t$, where the latter looks for the depth at which the in-situ density has increased by $\Delta\sigma_t$ from the surface value and $\sigma_t = \rho - 1000\ \mathrm{kg/m}^3$ (Kara et al., 2000). The study by de Boyer Montégut et al. (2004) compared multiple choices for the difference-based criteria and showed that temperature thresholds of 0.5°C and 0.8°C yield similar results to $\Delta T = 0.2$°C during periods of mixed layer deepening, but the larger values of ΔT are not sensitive enough to capture the mixed layer re-stratification in the spring and result in delayed shoaling of mixed layer. Note that as will be shown below, some of the important differences between the density profiles in submesoscale active and inactive regions occur at the base of the mixed layer, so using a temperature or density threshold that includes the mixed layer base is important for separating submesosale active and inactive regions. There also exist criteria based on the vertical gradient of density profiles. However, the experimental study by Brainerd and Gregg (1995) suggests that difference-based criterion for MLD is more stable than gradient-based criterion, which requires sharp gradient-resolved profiles. As global observational data sets typically have much lower vertical resolution compared to model-based data, the more robust difference-based criterion for MLD suits our purpose for future application better.

The LES model uses buoyancy as a dynamical variable and hence applying the criterion based on buoyancy is straightforward. For the SO model, the potential density is first calculated from the temperature and salinity profiles using the GSW Oceanographic Toolbox (McDougall & Barker, 2011). Then the potential density is scaled by the gravitational acceleration and a constant reference density, $\rho_0 = 1020\ \mathrm{kg/m}^3$, to give the buoyancy profiles.

For each vertical buoyancy profile, we find the depth where the buoyancy decreases by Δb from the surface buoyancy value and retain only the portion of the profile above this depth. Then, to ensure that all the resulting profiles have the same number of entries, we linearly interpolate each profile with values above the mixed layer base onto 100 vertical levels equally spaced between the mixed layer base and the ocean surface.

Just as the MLD varies across different regions and seasons, the vertical-mean buoyancy in the mixed layer varies as well. In order to eliminate the influence of vertical-mean buoyancy on the algorithm while maintaining the shapes of the profiles, we subtract the vertical mean from each buoyancy profile.

3.2. PCM

The PCM is an unsupervised machine learning algorithm first proposed in Maze et al. (2017) for classifying Argo temperature profiles in the North Atlantic Ocean. Jones et al. (2019) used this method to analyze Argo temperature profiles in the Southern Ocean, and more recently, Rosso et al. (2020) applied PCM to Argo temperature and salinity data, and Houghton and Wilson (2020) used this approach to relate properties in the Pacific Ocean to the El Niño index. PCM uses a Gaussian Mixture Model (GMM), which decomposes the probability density function (PDF) of a data set into a weighted sum of Gaussian distributions (Maze et al., 2017). We use the Python "pyXpcm" package (Maze, 2018) to perform the PCM on the re-scaled buoyancy profiles. Below, we give a brief overview of the GMM and PCM algorithms. More details on the methods can be found in Maze et al. (2017).

GMM fits data with a linear combination of multi-dimensional Gaussian distributions with unknown means and unknown variances. Suppose we have a set $\mathbf{X}$ of N profiles, each with D entries, and suppose that we want to group the profiles into K classes. GMM uses an expectation-maximization algorithm (described in Maze et al. (2017)) and attempts to find the optimal parameters (weight $\boldsymbol{\lambda}$, mean $\boldsymbol{\mu}$, and covariance $\boldsymbol{\Sigma}$) to represent the PDF, $p(\mathbf{X})$, as a weighted sum of Gaussian distributions as follows:

$$p(\mathbf{X}) = \sum_{k=1}^{K} \lambda_k \mathcal{N}(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

where the multi-dimensional Gaussian distribution $\mathcal{N}(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ has the form

$$\mathcal{N}(\mathbf{X};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=\frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_k|}}\exp\left(-\frac{1}{2}(x-\boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(x-\boldsymbol{\mu}_k)\right). \tag{2}$$

The Gaussian distribution in Equation 2 has D dimensions, which correspond to the number of vertical levels in each vertical profile. As the number of dimensions grows, the number of entries needed to be calculated in each covariance matrix $\boldsymbol{\Sigma}_k$ grows exponentially. Hence, it is sometimes necessary to reduce the dimensionality of the data sets before using GMM, and a principal component analysis (PCA) is used for this purpose. Following the criterion for selecting the number of principal components in Maze et al. (2017), Jones et al. (2019), and Rosso et al. (2020), we retain the minimum number of principal components that accounts for at least 95% of the variance.

Another parameter in the PCM algorithm is the number of profiles to use in the training step. For the LES data set, we train the PCM with $N_{train} = 120$ profiles and $N_{train} = 1200$ profiles to explore the effect of the sparsity of the sampled points. To produce a statistically robust test, we randomly select 10 different sets of N_{train} profiles from the LES data set for each $N_{train} = 120$ and $N_{train} = 1200$. For each set of training profiles, we randomly select the N_{train} profiles from three-dimensional model output saved at six different times ranging from 5.1 to 6.1 days after initialization (i.e., $N_{train}/6$ profiles are randomly selected from each time step and then combined to form a training set). We then train PCMs with $N_{train} = 120$ and $N_{train} = 1200$ on each of the 10 training sets.

For the SO data set, we started by selecting 10,000 density profiles from randomly selected locations. In order to exclude the possible influence of the bottom in shallow water, we exclude all points with bottom depth less than 500 m from our analysis. In addition, there is a sponge forcing region in the simulations that is north of 50°S, so any points in this region are discarded as well. After applying the exclusion criteria mentioned above, we are left with 6,397 vertical profiles.

The last key parameter for training the PCM is the number of classes, K. Since we are primarily interested in determining whether a vertical buoyancy profile belongs to a submesoscale eddy or not, we start with $K = 2$ for both the LES data set and the SO data set. Based on the results presented in Taylor et al. (2020), the region inside the submesoscale eddy has a shallower MLD, compared to outside the eddy. Thus, we calculate the average MLD for each of the $K = 2$ classes, and assign the class with the shallower average MLD as "Eddy" and the other class as "Non-eddy." Although the SO data set contains both mesoscale and submesoscale eddies, these labels refer just to submesoscale eddies which have a stronger influence on the mixed layer stratification. We also experiment with $K > 2$ on the LES data set (Section 4.1.2).

3.3. Testing and Evaluation of PCM Performance

We test the PCM performance qualitatively by visualizing the classification results and the mean profiles for the two classes. We also evaluate the quantitative skill by assigning "truth labels" to each profile using a threshold applied to the filtered vertical vorticity as described in detail below. It is worth noting that the PCMs are trained on and applied to buoyancy profiles and we evaluate the skill of the PCMs, either qualitatively or quantitatively, based on the velocity field which is not shown to the PCMs, allowing us to independently test the ability of the PCM to identify regions affected by submesoscale activity. Also note that our definition of the "truth labels" is somewhat arbitrary and other choices are possible.

For a statistical model such as the PCM that assigns data into one of two categories, each classification result falls into one of four outcomes: true positive, false positive, true negative, and false negative. Here, "positive" and "negative" stands for the two classes, and we assign these to "eddy" and "non-eddy," respectively, while "true" and "false" indicates whether the classification agrees with the truth label. These outcomes are commonly used to calculate skill scores, such as precision, recall, and F-1 scores, for classification models and to evaluate them quantitatively. The *precision score* is the ratio of number of "true positives" outcomes to the total number of positive classifications (i.e., the sum of "true positive" and "false positive" outcomes). The *recall score* is the ratio of "true positive" outcomes to the number of cases that are actually positive (i.e., the sum of "true positive" and "false negative" outcomes). In other words, the precision measures the fraction of profiles classified as "Eddy" by the PCM that are actually in the "Eddy" category, and the recall measures what fraction of profiles actually in the "Eddy" category are classified as "Eddy" by the PCM. The F-1 score, which is the weighted average of precision and recall, is a more balanced measure of the skill of the statistical model, especially for data sets with uneven distributions between positive and negative cases (Pedregosa et al., 2011).

4. Results

4.1. LES Data Set

After pre-processing the LES data set, we apply a PCA to the re-scaled and normalized buoyancy profiles. The first 7 principal components account for 64.3%, 17.0%, 5.3%, 3.9%, 1.9%, 1.7% and, 1.1% of variance respectively, adding up to 95.2% of total explained variance. Thus, we choose $N_{pc} = 7$ when training PCMs on the LES data set.

4.1.1. PCM With $K = 2$

In order to calculate the scores, we need to generate "ground truth" labels for the data sets. For the LES data set, we use the vertical vorticity $(\bar{\zeta})$ evaluated at the sea surface to generate truth labels. In order to remove the influence of small-scale turbulence, we calculate the vorticity from smoothed velocity fields using a moving average with 100 × 100 points (covering an area of about 390 m × 390 m). As submesoscale eddies are characterized by Rossby number $Ro = \zeta/f \sim \mathcal{O}(1)$, where ζ is the surface vertical vorticity and here $f = 1.28 \times 10^{-4}\text{s}^{-1}$ is the Coriolis frequency, the submesoscale eddy can be equivalently characterized as having $\zeta \sim f \approx 10^{-4}$ s^{-1} (Thomas et al., 2008). Thus, truth labels for the LES data set are generated by assigning points with $\bar{\zeta} \geq 10^{-4}$ s^{-1} as "Eddy" and points with $\bar{\zeta} < 10^{-4}$ s^{-1} as "Non-eddy."

To test the performance of the PCM, we randomly select $N_{test} = 10000$ profiles from each of the six output files from the LES and apply the PCM to this testing data set. For each test, the skill scores of the PCM, namely the precision, recall, and F-1 scores are calculated using the metrics in the *scikit-learn* package and recorded (Pedregosa et al., 2011). The means and standard deviations of the skill scores from the tests of PCMs with $N_{train} = 120$ and $N_{train} = 1200$ on the LES data set are calculated. Since profiles from a single time step of the LES data set might be correlated, we test the homogeneity of profiles in the LES data set by randomly selecting 100 ensembles of n profiles from the LES data set, where $n = 10, 20, \ldots$, and calculating the mean and standard deviation of Euclidean distance between the ensemble mean and the overall mean (Figure 2). The Euclidean distance converges to 0 after about $n = 10^4$, which implies that little new information is added by including more profiles beyond this point. However, the Euclidean distance decreases from $n = 10^2$ to $n = 10^3$, implying that there is new information added in this range, and justifying our choice of the two values of N_{train}.

Here, we first present an example of applying the classification algorithms to $N_{test} = 10000$ randomly selected testing profiles from a single time step, about 5.5 days after the initialization of the LES. Unless otherwise specified, we use the PCM trained with $N_{train} = 120$ and $N_{pc} = 7$. Figure 3 shows the shaded contour plots of surface speed (V_s), smoothed surface vertical vorticity $(\bar{\zeta})$, and MLD at this time step. In Figure 4, panel (a) shows the "truth" labels of Eddy and Non-eddy generated as described in Section 3.3 with contour of the classification results in red, while panel (b) shows the classification from the PCM trained with $N_{train} = 120$ and $N_{pc} = 7$ and applied to the test profiles with contour of the "truth" labels in red. The truth or prediction labels colored orange are points belonging to the submesoscale eddy and those colored blue are points outside the submesoscale eddy.

The submesoscale eddy in the domain does not have a clearly defined boundary, so the truth labels assigned for the profiles are not absolute (Figures 3 and 4). Which profiles are assigned to Eddy and Non-eddy depends on the vorticity threshold and the smoothing operation, and different parameter choices will result in different "truth" labels. The subjectivity in this process does affect the skill scores of the PCM. However, if we compare the generated truth labels and the classification from the PCM visually, they agree well, especially around the central part of the submesoscale eddy. The main disagreements are around the periphery of the eddy and how far the eddy extends. These plots demonstrate that the PCM indeed is able to identify whether a vertical buoyancy profile belongs to a submesoscale eddy or not, at least for the LES data set, without the information on the locations of the profiles in the domain.

Figure 5a shows the re-scaled buoyancy profiles above the mixed layer base classified by the PCM trained with $N_{train} = 120$ and $N_{pc} = 7$. The average of the profiles assigned to each class are shown in solid lines with filled

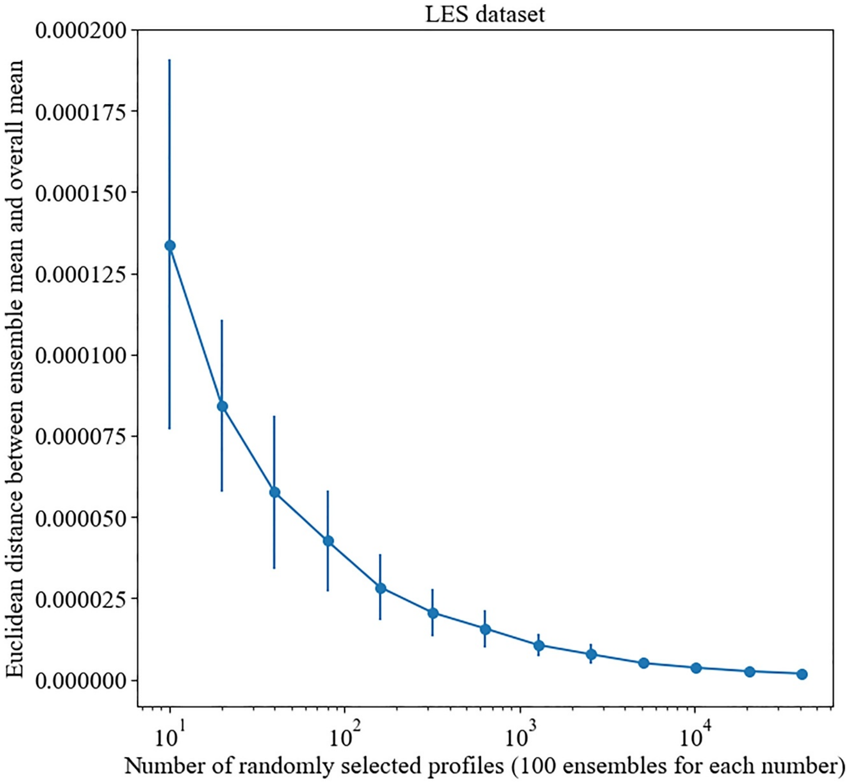

Figure 2. Mean and standard deviation of the Euclidean distance between the ensemble mean profile and overall mean profile for *n* randomly selected profiles from the large-eddy simulation data set. Hundred ensembles are used for each value of *n*.

shading indicating one standard deviation range about the mean. By comparing profiles classified as Eddy and Non-eddy, we can clearly see that there are qualitative differences between profiles in each class. In particular, the profiles in the Eddy class tend to have smoother transitions across the mixed layer base. The reasons for this are not immediately clear. One explanation might be that the eddy confines convection to a relatively shallow layer (similar to Taylor and Ferrari (2010)) and the submesoscale eddy re-stratifies the mixed layer below this depth.

Figure 6 illustrates the Gaussian Mixture Model (GMM) applied in principal component space for the LES data set. The coordinates of each point are the first two coefficients in the principal component expansion. Note that we apply the GMM to $N_{pc} = 7$ principal components, and hence this figure can be viewed as a 2-dimensional projection of the 7-dimensional principal component space. The solid, dashed, and dotted elliptical contours

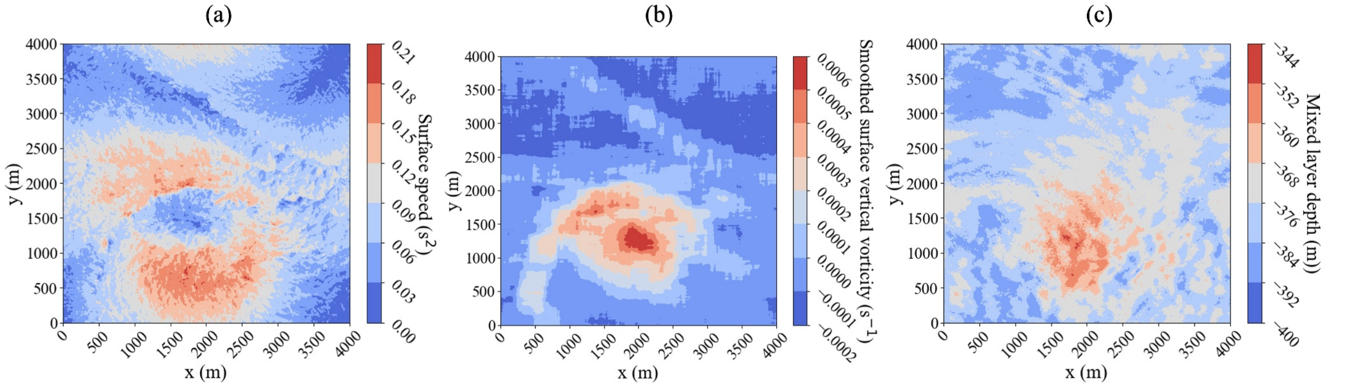

Figure 3. Shaded contour plots of (a) surface speed, (b) smoothed surface vertical vorticity $\bar{\zeta}$, and (c) mixed layer depth for a time step about 5.5 days after the initialization of the large-eddy simulation.

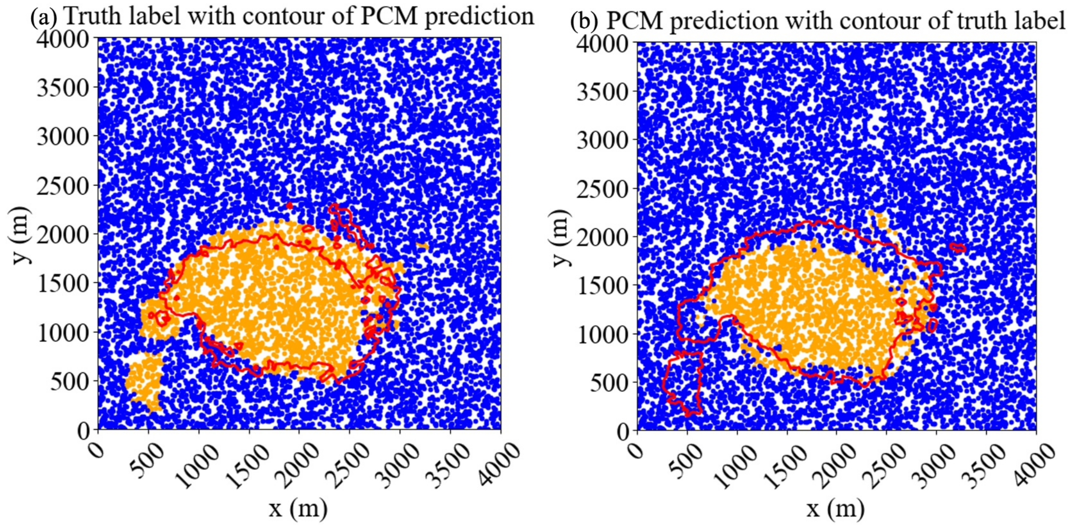

Figure 4. (a) Truth labels with contour of profile classification model (PCM) predictions and (b) PCM predictions with contour of truth labels for a time step about 5.5 days after the initialization of the large-eddy simulation. Truth labels or predictions colored orange correspond to Eddy and those colored blue correspond to Non-eddy. Contours are in red.

indicate the one, two, and three standard deviation confidence ellipses of the Gaussian distributions fitted by the GMM and projected onto the first two principal component space. The Gaussian distributions and the individual points are colored according to the PCM classification, where blue represents "Non-eddy" and orange represents "Eddy." The two classes occupy quite distinct regions in the principal component space and the Gaussian distributions capture their spread well (Figure 6), bearing in mind that only two of the seven dimensions of principal component space are shown.

In addition to the qualitative comparison, we can evaluate the skill of the PCM quantitatively using the truth labels defined above. As described in Section 3.2, we fix $K = 2$, $N_{pc} = 7$, and vary the PCM parameter $N_{train} \in \{120, 1200\}$. PCMs with each pair of parameters are trained 10 times with 10 different sets of training profiles and then tested on six sets of $N_{test} = 10000$ profiles from the six time steps. Precision and recall scores

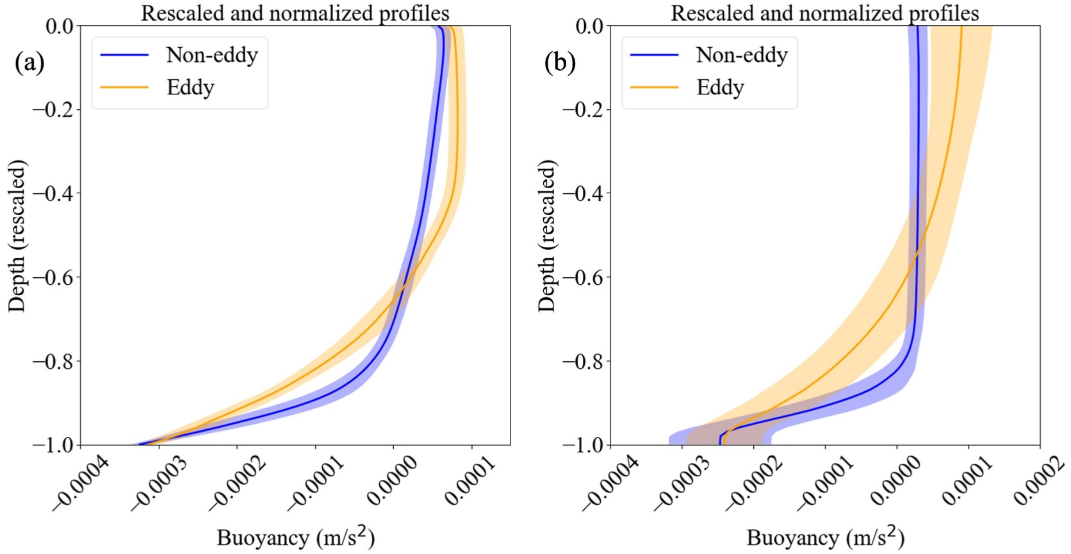

Figure 5. Mean re-scaled and normalized buoyancy profiles classified as Eddy and Non-eddy by PCMs trained with and applied to (a) the large-eddy simulation data set and (b) the SO data set, respectively (as discussed in Section 4.1.1 and Section 4.2, respectively), with shaded one standard deviation ranges.

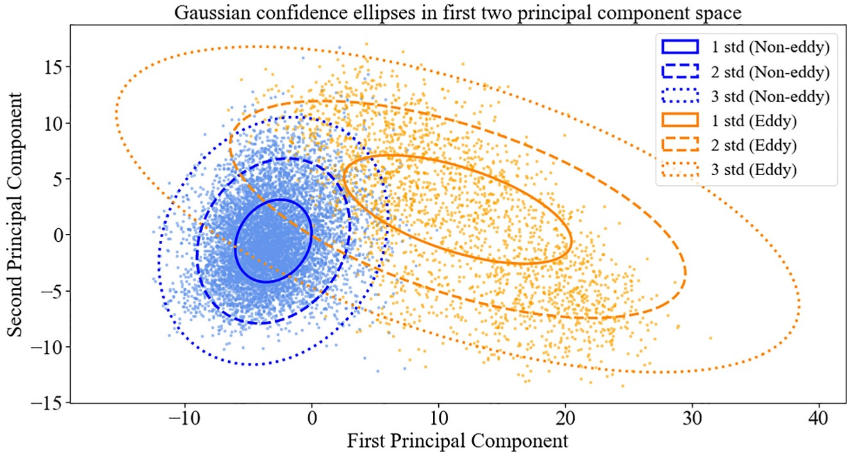

Figure 6. Confidence ellipses of the GMM for the two classes in the abstract principal component space, where the points represent the first two principal components of 10,000 vertical buoyancy profile sampled for testing the profile classification model and are colored according to their classes (Blue corresponds to "Non-eddy" and Orange corresponds to "Eddy"). The solid, dashed, and dotted elliptical contours correspond to one, two, and three standard deviation confidence range, respectively.

are calculated for each PCM and statistics of the scores are calculated over PCMs with the same set of parameters. Tables 1 and 2 show the means and standard deviations of the precision, recall, and F-1 scores for PCMs trained on 120 and 1,200 profiles, respectively.

In the truth labels of the LES data set, around 80% of the points are classified as Non-eddy and around 20% are classified as Eddy. Thus, if an algorithm classifies every point as Non-eddy, it should achieve a weighted recall score of 0.8. The PCMs trained with 120 and 1,200 profiles both achieve weighted recall scores higher than 0.8, which shows that PCM can successfully classify buoyancy profiles in our data set.

4.1.2. PCM With $K > 2$

Although the main focus of this paper is to classify whether a vertical buoyancy profile belongs to a region with high submesoscale activity or not, requiring only two classes in the classification, we also experiment with grouping profiles into multiple classes using PCM on the LES data set. To find the most suitable number of classes for the LES data set, we calculate the Bayesian information criterion (BIC) score for PCMs trained with $N_{train} = 1200$, $N_{pc} = 7$, and varying the number of classes, K. The BIC score rewards high likelihood but penalizes overfitting. The BIC score for PCM is calculated as follows

$$BIC = -2llh \times N_{train} + \ln(N_{train}) \times \left[N_{pc} \times K + K \frac{N_{pc}(N_{pc}+1)}{2} + K - 1 \right], \tag{3}$$

where llh stands for the log-likelihood of the trained PCM. The first term effectively rewards any increases in the statistical likelihood of the PCM, while the second term penalizes overfitting. Below, Figure 7 shows the BIC score for PCMs trained on the LES data set with $N_{train} = 1200$, $N_{pc} = 7$ and number of classes $K \in [2, 10]$.

Table 1
Precision, Recall, and F-1 Scores for Profile Classification Model Trained on Large-Eddy Simulation Data Set With $K = 2$, $N_{pc} = 7$, and $N_{train} = 120$

	Precision	Recall	F-1 score
Mean	0.913	0.888	0.894
Standard deviation	0.016	0.050	0.042

We use the "elbow method" to select the optimal value for K, which is a heuristic method for choosing the number of clusters for unsupervised clustering models, such as k-means, by finding the inflection point on the BIC curve and selecting the corresponding number of clusters (Dangeti, 2017). This method works because the inflection point on the BIC curve indicates when the model fits the data relatively well but without too much overfitting. In our case, the inflection point on the BIC curve is at $K = 5$. Here, we show an example of results from applying PCM trained with $K = 5$, $N_{train} = 1200$, and $N_{pc} = 7$ to the testing set of profiles from the time step about 5.5 days

Table 2
Precision, Recall, and F-1 Scores for Profile Classification Model Trained on Large-Eddy Simulation Data Set With $K = 2$, $N_{pc} = 7$, and $N_{train} = 1200$

	Precision	Recall	F-1 score
Mean	0.897	0.851	0.863
Standard deviation	0.003	0.010	0.008

after the LES initialization. The classification result (Figure 8a) and the corresponding mean re-scaled and normalized profiles with one standard deviation range for each class (Figures 8b and 8c) are shown.

It is worth noting that with $K = 5$, PCM has the ability to further distinguish regions outside the central part of the submesoscale eddy. It may be possible to identify regions with certain characteristics near a submesoscale eddy, such as a submesoscale front, by selecting classes in PCM predictions using relevant criteria. The individual profiles in those regions can then be isolated and explored in more detail. By comparing the labels produced by the PCM in Figure 8 with Figure 3, we see that Class 1 is the central part of the submesoscale eddy, Class 3 and 4 are around the edge of the eddy, while Class 2 and 5 are farthest from the eddy. Class 5 corresponds to a submesoscale front where the horizontal density gradient and the vertical vorticity are large (Figure 3b).

4.2. SO Data Set

In this section, we apply the PCM described above to the simulation of a sector of the Southern Ocean described in Section 2. This simulation covers a much larger area than the LES data set while still permitting submesoscale eddies. Figure 9 shows the surface speed (V_s, top panel) and the surface vertical vorticity (ζ, bottom panel) from the Bachman et al. (2017) model. Due to its large domain size, this data set contains many submesoscale eddies, visible as small-scale features in the surface vorticity. However, the areas in the model domain with the highest surface speed and the most negative surface vertical vorticity are, in fact, mesoscale eddies and meanders rather than submesoscale eddies. Enhanced submesoscale activity can be found along the peripheries of some mesoscale eddies in the southeast areas of the domain, and along the west and south coasts of South America. Due to the extremely energetic mesoscale eddy field, this region presents a difficult test for diagnosing submesoscale activity. For example, since the mesoscale eddies in this region are associated with large Rossby number ($Ro \equiv |\zeta|/|f| \sim 1$), the local vertical vorticity alone is insufficient to distinguish between mesoscale and submesoscale eddies.

In order to diagnose submesoscale activity and generate truth labels for the SO model, we calculate the high-pass filtered vorticity magnitude, $\hat{\zeta}$, by taking the moving average of the absolute value of the surface vertical vorticity ($|\zeta|$) and subtracting from it the absolute value of the moving average of surface vertical vorticity (ζ). That is,

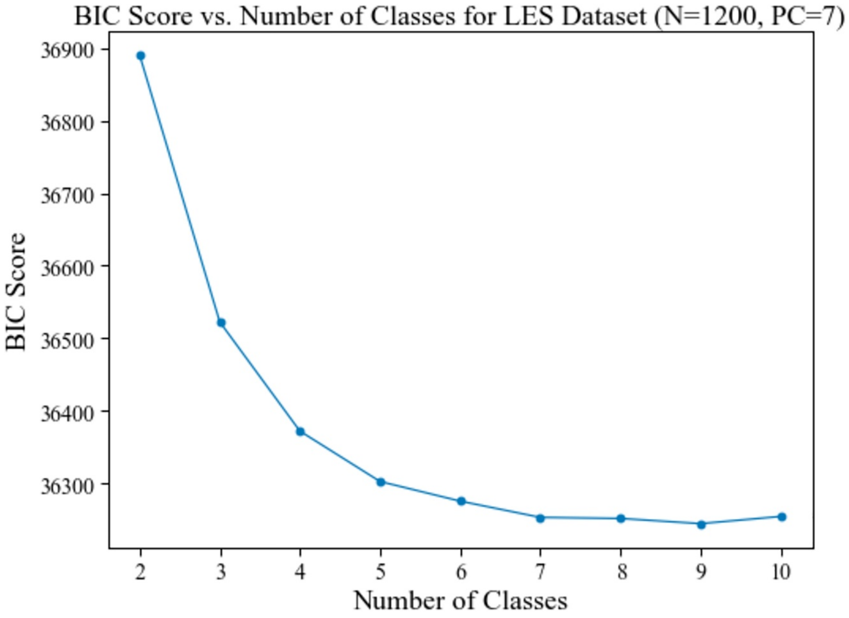

Figure 7. Bayesian information criterion score versus number of classes (K) for PCMs trained with $N_{train} = 1200$ and $N_{pc} = 7$ on the large-eddy simulation data set.

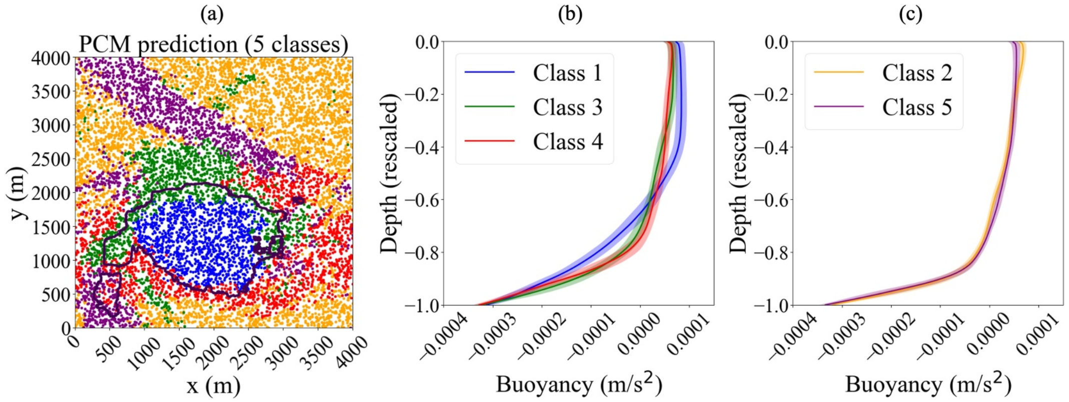

Figure 8. (a) Classification using profile classification model trained with $K = 5$, $N_{train} = 1200$, and $N_{pc} = 7$, with $\bar{\zeta} = 10^{-4}$ s^{-1} contour in black, and mean re-scaled and normalized profiles with one standard deviation range for (b) Classes 1, 3, 4, and (c) Classes 2, 5.

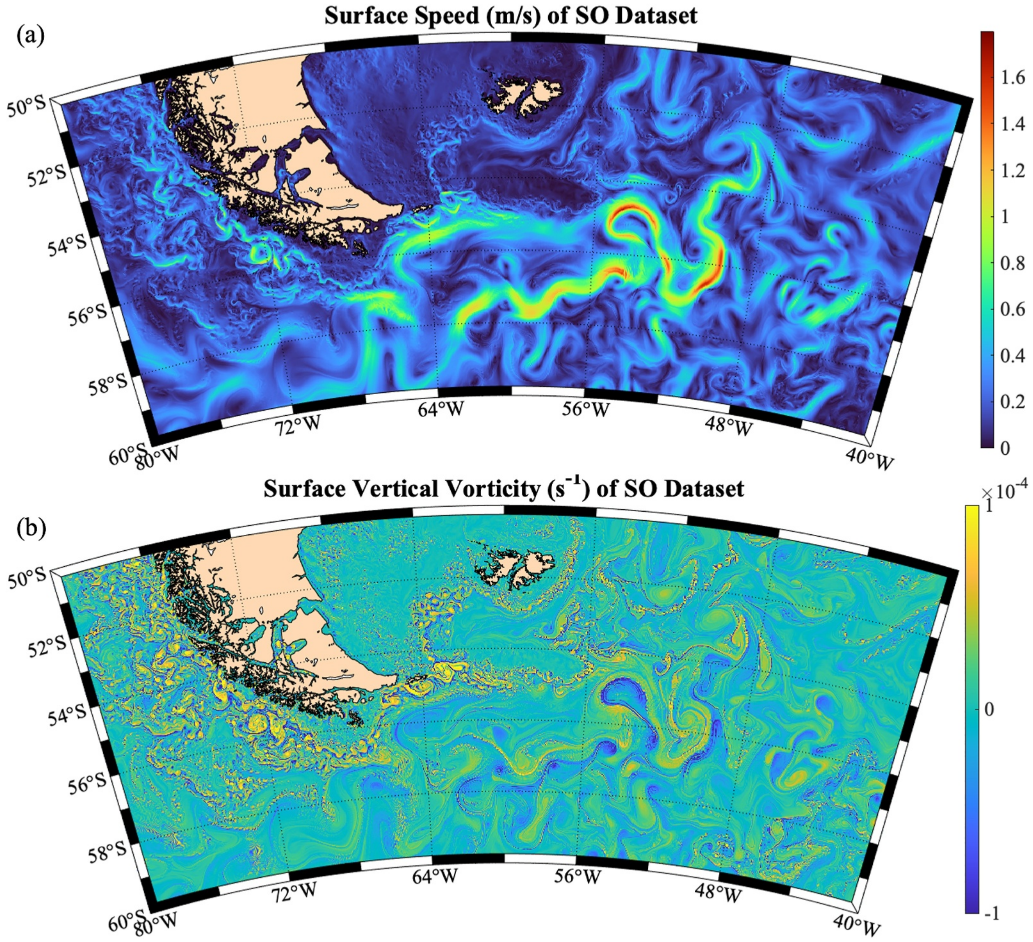

Figure 9. Snapshot of (a) Surface speed and (b) surface vertical vorticity for the SO data set on May 13 in the simulation.

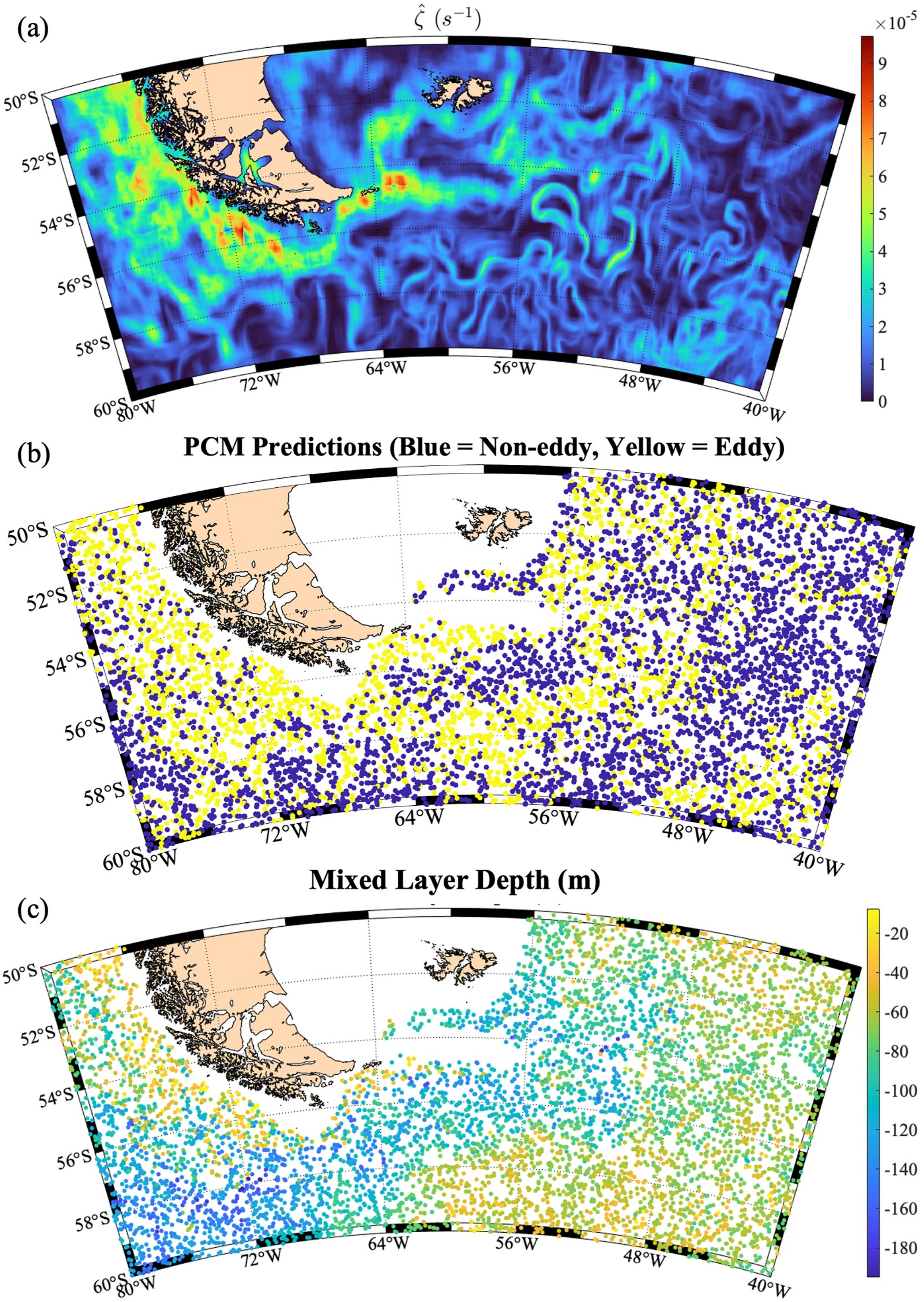

Figure 10. (a) $\hat{\zeta}$, (b) profile classification model predictions, and (c) mixed layer depth for the SO data set.

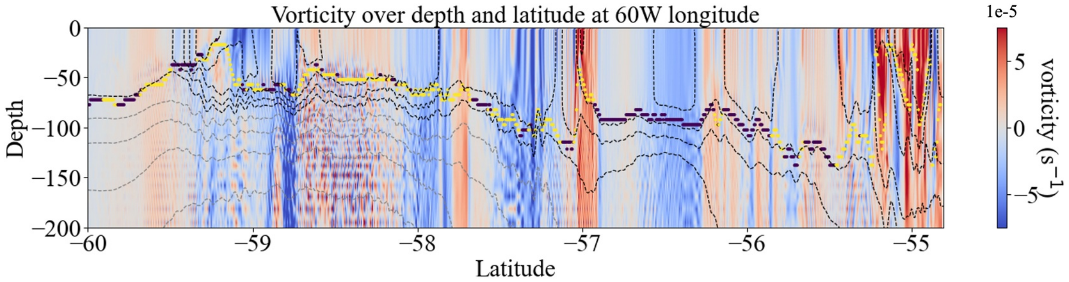

Figure 11. North-South slice of SO model domain at 60°W showing the vertical vorticity in color, buoyancy contours with 0.0005 m/s^2 spacing in black dashed lines on top, buoyancy contours with 0.001 m/s^2 spacing in gray at bottom, and scattered points with color corresponding to profile classification model classification (Eddy in yellow, Non-eddy in dark purple) at depth corresponding to mixed layer depth.

$$\hat{\zeta} = \overline{|\zeta|} - |\overline{\zeta}|, \tag{4}$$

where $\overline{\cdot}$ stands for the operation of taking the moving average over a window of 100 × 100 grid points, or about 50 × 50 km. As shown in the top panel of Figure 10, $\hat{\zeta}$ highlights regions with enhanced submesoscale activity.

As described in Section 2 and 3, we re-scale and normalize the buoyancy profiles before applying the PCM algorithm. The first four principal components of the pre-processed SO data set account for 60.7%, 27.1%, 6.6%, 3.1% of variance respectively, summing up to 97.5% of total variance. We train the PCM with parameters $N_{pc} = 4$ and $K = 2$ using all 6,397 profiles in the SO data set. We then apply the trained PCM to the same profiles and separate them into two classes.

Figure 10b shows the classification results. The locations of profiles classified as "Eddy" (shown as yellow points in Figure 10b) generally coincide with regions of enhanced submesoscale activity as diagnosed by $\hat{\zeta}$ (Figure 10a). Some regions with high submesoscale activity also coincide with shallow mixed layers (Figure 10c), notably the region west of southern Chile. However, enhanced submesoscale activity and submesoscale eddy classification are also seen in the high-speed ACC jets where the MLD is not shallow (Figure 10c). It is also worth noting that during the pre-processing of the SO data set, we discard profiles at locations with bottom depths shallower than 500 m from our analysis, so coastal regions near South America and over the continental shelf are excluded. Although there can be high submesoscale activity in the coastal regions (Figure 10a), we exclude these regions from our analysis to avoid situations where the MLD is comparable to the bottom depth.

A vertical North-South slice at 60°W is shown in Figure 11 to demonstrate finer structures in the vertical profiles. Color shading show the vertical vorticity, black contours show isopycnals (with constant buoyancy). The signature of submesoscale eddies are detectable as small regions of surface-intensified vorticity. The mixed layer base is denoted by dots which are colored based on the classification of the corresponding profile. The PCM classification for Eddy (yellow points) highlights the regions with intensified surface vorticity quite well. However, quite a few profiles between 58°W and 59°W are classified as Eddy but no significant vorticity feature can be seen near the ocean surface, so they are in fact "false positives" in the classification. This region of false positives contains high levels of topography-generated internal waves, which are visible through the small-scale vorticity fluctuations in the pycnocline whose amplitude increases downwards (these waves are discussed in Bachman et al. (2017)). We speculate that enhanced mixing associated with the energetic internal wavefield could smooth the change in stratification at the base of the mixed layer, thereby making the shape of the density profiles similar to those corresponding to submesoscale eddies. It might be possible to further refine the classification method to reduce the number of false positive results (e.g., by filtering out internal waves), but we leave this for future work.

Figure 5b shows the mean re-scaled and normalized buoyancy profiles for both classes and their corresponding one standard deviation range. The profiles classified as Non-eddy are nearly homogeneous in the upper 80% of the mixed layer. Profiles classified as Eddy, which fall in regions with high submesoscale activity, are more stratified throughout the mixed layer. Similar to the results from the LES data set, the mixed layer base is less distinct in profiles that are classified as Eddy.

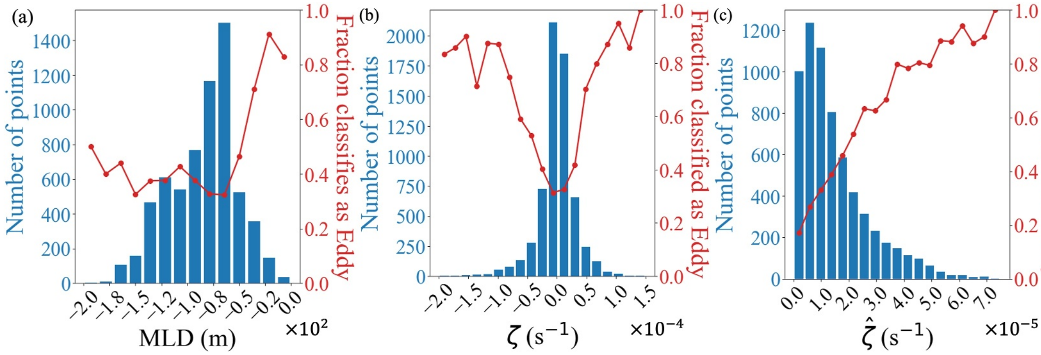

Figure 12. Histograms (blue) of (a) mixed layer depth, (b) ζ, and (c) $\hat{\zeta}$ from the SO data set. Red lines indicate the fraction of the points within each histogram bin that were classified as Eddy.

There are a few notable differences between the classified profiles in the LES data set and the SO data set. For example, the profiles classified as Non-eddy in the LES data set are more stratified in the mixed layer than in the SO data set. This may be caused by the fundamentally different way that turbulence is represented in the two models. With its high resolution, the LES explicitly resolves the largest scales of 3D turbulence, while the SO model uses the KPP to parameterize 3D turbulence. Another cause may be the differences in the physical processes that are included in the two models. For example, the LES model does not include wind forcing, which is a major driver of upper ocean turbulence in the Southern Ocean. The LES is forced by cooling the surface, which also removes potential vorticity. Hence SI might be active in the LES. SI is known to lead to a density profile characterized by a well-mixed "convective layer" overlying a weakly stratified layer with active SI (Taylor & Ferrari, 2010). The weak stratification visible in the non-eddy classes of the LES might be due to SI. The resolution of the SO model is likely too coarse to fully capture SI (Bachman & Taylor, 2014), and this might explain why the profiles in the non-eddy case have a very small buoyancy gradients in the mixed layer.

Figure 12 shows histograms of MLD, ζ, and $\hat{\zeta}$ for all profiles in blue. Red lines indicate the fraction of points within each histogram bin that are classified as Eddy. Note that most profiles with a MLD shallower than ∼45 m are classified as Eddy (Figure 12a). The profiles with mixed layer shallower than 45 m mainly come from the region off the west coast of South America, which coincides with one of the regions with highest submesoscale activity. However, if we were to classify profiles according to their MLD (e.g., assign profiles with mixed layer shallower than 45 m as Eddy and the rest Non-eddy), the vast majority of regions in the open ocean with high submesoscale activity would be misclassified as Non-eddy. This again emphasizes the benefit of using the shape of the density profiles for the classification, rather than using only the MLD. It can also be seen that the majority of points with high local vorticity, $|\zeta| > 5 \times 10^{-5}\ \mathrm{s}^{-1}$, are classified as Eddy (Figure 12b). There is a slight bias where more profiles with cyclonic vorticity ($\zeta > 0$) are classified as Eddy compared to the same magnitude of anticyclonic vorticity ($\zeta < 0$). This is consistent with the well known cyclonic bias associated with submesoscale eddies (McWilliams, 2016). Most points with high $\hat{\zeta}$ are classified as Eddy (Figure 12c), demonstrating the skill of the method when applied to the SO data set.

More quantitatively, we assign "truth" labels for the SO data set and calculate the skill scores for the PCM as we have done for the LES data set. We generate the truth labels by assigning profiles with $\hat{\zeta} \geq 2 \times 10^{-5}\ \mathrm{s}^{-1}$ as Eddy and $\hat{\zeta} < 2 \times 10^{-5}\ \mathrm{s}^{-1}$ as Non-eddy. Note that there is some inherent ambiguity in the choice of the truth label and for this data set. The precision, recall, and F-1 scores for the Eddy and Non-eddy classes are shown in Table 3. The precision score for the Non-eddy class is quite high, which means that of all the profiles classified as Non-eddy, 86% of them have Non-eddy truth labels. The recall scores for Eddy and Non-eddy are similarly moderate, indicating that of all profiles with Eddy and Non-eddy truth labels, 67% and 69% are correctly classified, respectively. However, the precision score for the Eddy class is lower, with only 43% of profiles classified as Eddy having the correct

Table 3
Precision, Recall, and F-1 Scores for Profile Classification Model Trained on SO Data Set With $K = 2$, $N_{pc} = 4$, and $N_{train} = 6397$

	Precision	Recall	F-1 score	Support
Eddy	0.43	0.67	0.52	1,621
Non-eddy	0.86	0.69	0.77	4,776
Weighted average	0.75	0.69	0.71	6,397

Note. Support indicates the total number of profiles with the Eddy and Non-eddy truth labels, respectively.

truth labels. This may be partially due to the bias in the SO data set itself, where the profiles are not evenly distributed over the two classes (1,621 with Eddy truth label and 4,776 with Non-eddy truth label as shown as "Support" in Table 3). As can be seen in Figure 12c, the profiles are distributed heavily over low $\hat{\zeta}$ values and there is a thin tail over higher $\hat{\zeta}$ values. With such large number of profiles with low $\hat{\zeta}$ and Non-eddy truth labels, even if only a small fraction of these profiles are misclassified as Eddy, they can outnumber the correctly classified Eddy profiles, resulting in the low precision score.

Unlike for the LES data set, where the whole model domain resolves only one submesoscale eddy and the threshold in $\bar{\zeta}$ sets a quite clear boundary of the eddy, the high-pass filtered vorticity magnitude $\hat{\zeta}$ for the SO data set can only highlight regions with intensified submesoscale activity and cannot resolve or distinguish individual submesoscale eddies in the field. The coarser resolution and the inherent ambiguity associated with distinguishing submesoscales from the extremely energetic mesoscale eddy field introduces uncertainties to the truth labels, resulting in lower skill scores for the SO data set is lower than those for the LES data set. By examining the histogram of $\hat{\zeta}$ and the fraction of Eddy profiles in Figure 12c, we can see that although the vast majority of profiles with very high and very low $\hat{\zeta}$ are correctly classified as Eddy and Non-eddy (87% and 76% for $\hat{\zeta} > 5 \times 10^{-5}$ s^{-1} and $\hat{\zeta} < 10^{-5}$ s^{-1} respectively), there are a large number of profiles in between, where the truth label may not correctly reflect the presence or absence of submesoscale eddies. Moreover, as shown in Figure 11 and explained above, processes like bottom topography-induced internal waves can cause false positives to occur and affect the PCM performance as well.

5. Conclusion

In this paper, we used a modified PCM to identify submesoscale eddies using vertical density profiles. We adapted the PCM by first normalizing the density (equivalently buoyancy) profiles within the mixed layer. Specifically, we introduced a new pre-processing step: the projection of density values onto a dimensionless vertical coordinate scaled by the MLD, wherein −1 corresponds to the base of the mixed layer and 0 corresponds to the surface. This step allows the PCM to identify structures within the mixed layer, as opposed to structures over a selected depth range. The adapted PCM was then applied to two existing model-based data sets from Taylor et al. (2020) and Bachman et al. (2017).

In Section 4, we applied the adapted PCM to a high resolution (~4 m horizontal grid spacing) LES and a 500 m resolution model of a sector of the Southern Ocean (SO). The PCM is able to identify whether a vertical buoyancy profile is within a submesoscale eddy in the LES data set and, statistically, whether it is in a region with high submesoscale activity for the SO data set without any additional information (e.g., velocity, location, or time). The method works for both data sets, despite the fact that the LES and SO models have very different resolutions and capture different physical processes, which demonstrates the robustness of the method.

Aside from the promising results obtained from applying the adapted PCM to the two model-based data sets, we need to keep in mind that these two data sets are sampled from somewhat idealized simulations. That is, there are additional processes that are not represented in the models, including a diurnal cycle and Langmuir circulation, which might influence the PCM classification and performance. Both a thin diurnal warm layer driven by the solar insolation and Langmuir turbulence might change the vertical profile of the mixed layer, and the influence of these processes should be considered when applying the adapted PCM to more realistic model-based data sets (e.g., Dauhajre & McWilliams, 2018; Menemenlis et al., 2021; Torres et al., 2018) or real-world observations.

In addition, as we mentioned when discussing the classification results of the SO data set (Section 4.2), we excluded profiles sampled from locations with bottom depth shallower than 500 m from our analysis. In shallow waters over the shelf the surface and bottom mixed layers can interact, leading to qualitatively different density profiles compared to the open ocean. However, the adapted PCM may still work in coastal or shallow regions in the ocean, particularly if it is trained exclusively using vertical profiles sampled from those regions.

Future work could test our method using high resolution (submesoscale-permitting) global models or global observational data sets such as Argo floats. The ability to identify and study submesoscale eddies using global data sets could allow systematic study of the influence of submesoscale eddies on biogeochemistry and to identify possible responses of submesoscale activity to the changing climate.

Data Availability Statement

Python and MATLAB code for training models, analyzing data, and generating figures, along with the underlying data from the LES and MITgcm models used is available for download at https://doi.org/10.17863/CAM.86279.

Acknowledgments

DJ is supported by a UK Research and Innovation Future Leaders Fellowship (Grant MR/T020822/1). The data and code used to complete this manuscript can be downloaded via the University of Cambridge repository (http://doi.org/10.17863/CAM.86279) (Yao & Taylor, 2024).

References

Bachman, S. D., & Taylor, J. R. (2014). Modelling of partially-resolved oceanic symmetric instability. *Ocean Modelling*, *82*, 15–27. https://doi.org/10.1016/j.ocemod.2014.07.006

Bachman, S. D., Taylor, J. R., Adams, K. A., & Hosegood, P. (2017). Mesoscale and submesoscale effects on the mixed layer depth in the Southern Ocean. *Journal of Physical Oceanography*, *47*(9), 2173–2188. https://doi.org/10.1175/jpo-d-17-0034.1

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

Brainerd, K. E., & Gregg, M. C. (1995). Surface mixed and mixing layer depths. *Deep Sea Research Part I: Oceanographic Research Papers*, *42*(9), 1521–1543. https://doi.org/10.1016/0967-0637(95)00068-H

Buckingham, C. E., Naveira Garabato, A. C., Thompson, A. F., Brannigan, L., Lazar, A., Marshall, D. P., et al. (2016). Seasonality of submesoscale flows in the ocean surface boundary layer. *Geophysical Research Letters*, *43*(5), 2118–2126. https://doi.org/10.1002/2016GL068009

Damerell, G. M., Heywood, K. J., Thompson, A. F., Binetti, U., & Kaiser, J. (2016). The vertical structure of upper ocean variability at the porcupine abyssal plain during 2012–2013. *Journal of Geophysical Research: Oceans*, *121*(5), 3075–3089. https://doi.org/10.1002/2015JC011423

Dangeti, P. (2017). *Statistics for machine learning build supervised, unsupervised, and reinforcement learning models using both Python and R*. Packt Publishing.

Dauhajre, D., & McWilliams, J. (2018). Diurnal evolution of submesoscale front and filament circulations. *Journal of Physical Oceanography*, *48*(10), 2343–2361. https://doi.org/10.1175/JPO-D-18-0143.1

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

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

Garabato, A. C. N., Yu, X., Callies, J., Barkan, R., Polzin, K. L., Frajka-Williams, E. E., et al. (2022). Kinetic energy transfers between mesoscale and submesoscale motions in the open ocean's upper layers. *Journal of Physical Oceanography*, *52*(1), 75–97. https://doi.org/10.1175/JPO-D-21-0099.1

Gula, J., Molemaker, M., & McWilliams, J. (2016). Topographic generation of submesoscale centrifugal instability and energy dissipation. *Nature Communications*, *7*(1), 12811. https://doi.org/10.1038/ncomms12811

Gula, J., Taylor, J., Shcherbina, A., & Mahadevan, A. (2022). Chapter 8 - Submesoscale processes and mixing. In *Ocean mixing*. Elsevier.

Houghton, I. A., & Wilson, J. D. (2020). El Niño detection via unsupervised clustering of Argo temperature profiles. *Journal of Geophysical Research: Oceans*, *125*(9), e2019JC015947. https://doi.org/10.1029/2019JC015947

Jones, D. C., Holt, H. J., Meijers, A. J. S., & Shuckburgh, E. (2019). Unsupervised clustering of Southern Ocean Argo float temperature profiles. *Journal of Geophysical Research: Oceans*, *124*(1), 390–402. https://doi.org/10.1029/2018JC014629

Kara, A., Rochford, P., & Hurlburt, H. (2000). An optimal definition for ocean mixed layer depth. *Journal of Geophysical Research*, *105*(C7), 16803–16821. https://doi.org/10.1029/2000JC900072

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

Lévy, M., Franks, P. J. S., & Smith, K. S. (2018). The role of submesoscale currents in structuring marine ecosystems. *Nature Communications*, *9*(1), 4758. https://doi.org/10.1038/s41467-018-07059-3

Mahadevan, A., & Tandon, A. (2006). An analysis of mechanisms for submesoscale vertical motion at ocean fronts. *Ocean Modelling*, *14*(3), 241–256. https://doi.org/10.1016/j.ocemod.2006.05.006

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

Maze, G. (2018). pyXpcm: Ocean Profile Classification Model. https://doi.org/10.5281/zenodo.3906236

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Lopez Radcenco, M., Lenca, P., et al. (2017). Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean. *Progress in Oceanography*, *151*, 275–292. https://doi.org/10.1016/j.pocean.2016.12.008

McCreary, J. P., Jr., Kohler, K. E., Hood, R. R., Smith, S., Kindle, J., Fischer, A. S., & Weller, R. A. (2001). Influences of diurnal and intraseasonal forcing on mixed-layer and biological variability in the central Arabian Sea. *Journal of Geophysical Research*, *106*(C4), 7139–7155. https://doi.org/10.1029/2000JC900156

McDougall, T., & Barker, P. (2011). Getting started with TEOS-10 and the Gibbs Seawater (GSW) oceanographic toolbox. *SCOR/IAPSO WG*, *127*(532), 1–28.

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

Mellor, G. L., & Durbin, P. A. (1975). The structure and dynamics of the ocean surface mixed layer. *Journal of Physical Oceanography*, *5*(4), 718–728. https://doi.org/10.1175/1520-0485(1975)005⟨0718:TSADOT⟩2.0.CO;2

Menemenlis, D., Hill, C., Henze, C. E., Wang, J., & Fenty, I. (2021). *Southern Ocean Pre-SWOT Level-4 Hourly MITgcm LLC4320 Native Grid 2km Oceanographic Dataset Version 1.0*. NASA Physical Oceanography Distributed Active Archive Center. https://doi.org/10.5067/PRESW-ASJ10

Omand, M. M., D'Asaro, E. A., Lee, C. M., Perry, M. J., Briggs, N., Cetinić, I., & Mahadevan, A. (2015). Eddy-driven subduction exports particulate organic carbon from the spring bloom. *Science*, *348*(6231), 222–225. https://doi.org/10.1126/science.1260062

Panassa, E., Völker, C., Wolf-Gladrow, D., & Hauck, J. (2018). Drivers of interannual variability of summer mixed layer depth in the Southern Ocean between 2002 and 2011. *Journal of Geophysical Research: Oceans*, *123*(8), 5077–5090. https://doi.org/10.1029/2018JC013901

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

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

Rosso, I., Mazloff, M. R., Talley, L. D., Purkey, S. G., Freeman, N. M., & Maze, G. (2020). Water mass and biogeochemical variability in the Kerguelen sector of the Southern Ocean: A machine learning approach for a mixing hot spot. *Journal of Geophysical Research: Oceans*, *125*(3), e2019JC015877. https://doi.org/10.1029/2019JC015877

Taylor, J. R. (2016). Turbulent mixing, restratification, and phytoplankton growth at a submesoscale eddy. *Geophysical Research Letters*, *43*(11), 5784–5792. https://doi.org/10.1002/2016gl069106

Taylor, J. R., & Ferrari, R. (2009). On the equilibration of a symmetrically unstable front via a secondary shear instability. *Journal of Fluid Mechanics*, *622*, 103–113. https://doi.org/10.1017/S0022112008005272

Taylor, J. R., & Ferrari, R. (2010). Buoyancy and wind-driven convection at mixed layer density fronts. *Journal of Physical Oceanography*, *40*(6), 1222–1242. https://doi.org/10.1175/2010jpo4365.1

Taylor, J. R., Smith, K. M., & Vreugdenhil, C. A. (2020). The influence of submesoscales and vertical mixing on the export of sinking tracers in large-eddy simulations. *Journal of Physical Oceanography*, *50*(5), 1319–1339. https://doi.org/10.1175/jpo-d-19-0267.1

Thomas, L., Tandon, A., & Mahadevan, A. (2008). Submesoscale processes and dynamics. *Ocean Modeling in an Eddying Regime*, *177*, 17–38. https://doi.org/10.1029/177GM04

Thompson, A. F., Lazar, A., Buckingham, C., Garabato, A. C. N., Damerell, G. M., & Heywood, K. J. (2016). Open-ocean submesoscale motions: A full seasonal cycle of mixed layer instabilities from gliders. *Journal of Physical Oceanography*, *46*(4), 1285–1307. https://doi.org/10.1175/JPO-D-15-0170.1

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

Whitt, D. B., Lévy, M., & Taylor, J. R. (2019). Submesoscales enhance storm-driven vertical mixing of nutrients: Insights from a biogeochemical large eddy simulation. *Journal of Geophysical Research: Oceans*, *124*(11), 8140–8165. https://doi.org/10.1029/2019JC015370

Whitt, D. B., & Taylor, J. R. (2017). Energetic submesoscales maintain strong mixed layer stratification during an autumn storm. *Journal of Physical Oceanography*, *47*(10), 2419–2427. https://doi.org/10.1175/jpo-d-17-0130.1

Wong, A. P. S., Wijffels, S. E., Riser, S. C., Pouliquen, S., Hosoda, S., Roemmich, D., et al. (2020). Argo data 1999–2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats. *Frontiers in Marine Science*, *7*, 700. https://doi.org/10.3389/fmars.2020.00700

Yao, L., & Taylor, J. (2024). *Supporting data for identifying ocean submeoscale activity from vertical density profiles using machine learning*. University of Cambridge Data Repository. [data and sourcecode]. https://doi.org/10.17863/CAM.86279