GAUSSIAN MIXTURE MODELING DESCRIBES THE GEOGRAPHY OF THE SURFACE OCEAN CARBON BUDGET

Daniel C. Jones[1], Takamitsu Ito[2]

Abstract*—We use an unsupervised classification technique (i.e. Gaussian mixture modeling or GMM) to identify ocean regions with similar balances between processes that determine the surface budget of dissolved inorganic carbon. GMM objectively locates sub-populations in the distribution of carbon budget terms. We use a simple four-class description and find regimes that are broadly consistent with classical theoretical frameworks. Class 1 covers 24% of ocean surface area and corresponds to highly productive areas with strong vertical mixing, wind-driven open ocean upwelling, and absorption of atmospheric carbon dioxide. Class 2 covers 8% of ocean surface area and corresponds to regions of especially weak productivity. Class 3 covers 16% of ocean surface area and corresponds to wind-driven coastal and equatorial upwelling. Finally, class 4 covers the remaining 52% of ocean surface area and corresponds to the relatively unproductive subtropical gyres, which are typically characterized by downwelling and low surface nutrient concentrations. We argue that GMM may be a useful method for comparing biogeochemical regimes between climate models.

I. MOTIVATION

The global ocean is a critical part of Earth's climate system, in part because it absorbs atmospheric carbon dioxide from fossil fuel burning, cement production, and biomass burning, thereby slowing the rate of surface warming. At present, the ocean absorbs between 20 to 35 percent of anthropogenic CO_2 emissions [1], [2], [3]. The ocean's ability to transport carbon from the near-surface ocean into the deep interior, where it is out of contact with the atmosphere, is sometimes referred to as the ocean carbon pump. Broadly speaking, this pump consists of two components: the solubility pump and the biological pump. The solubility pump is a consequence of the global overturning circulation, whereby atmospheric carbon is more readily absorbed by cold, high latitude waters and subducted into the interior ocean via deep convection. The biological pump is a consequence of ocean ecology, by which carbon is transferred from a dissolved inorganic carbon (DIC) pool in the surface ocean to an organic carbon pool. Some fraction of this organic carbon is respired back to DIC throughout the water column, and a small fraction ultimately reaches the seabed. The net result is a vertical transfer of DIC away from the surface ocean into the deep interior, where it is out of contact with the atmosphere and unable to directly affect surface climate.

The processes that govern the surface carbon budget display considerable spatial variability. For example, air-sea gas exchange is highly nonuniform, due to spatial variability in mixed layer depths, near-surface winds, and carbonate chemistry parameters [4], [5]. Biological productivity varies based on the nutrient distribution and other ecological factors [6]. Physical transport, which controls the ocean solubility pump, is also spatially variable as the ocean's global overturning circulation is set by bathymetry, surface forcing, and internal dynamics, which all have their own spatial patterns. The rate of change of the surface carbon concentration is set by the residual of these processes. Our present understanding of the surface carbon budget relies on classical theoretical frameworks that describe balances between these processes. As a complement to existing expertise-driven approaches, it may be useful to develop a suite of alternative methods by which we can characterize the surface carbon budget. Unsupervised learning may offer such a possibility. In unsupervised learning, one applies a classification algorithm to an unlabeled dataset, and the algorithm attempts to identify sub-populations in the data distribution [7]. To the extent that such methods can be shown to be robust and objective, they could be useful for comparing

Corresponding author: D. Jones, dannes@bas.ac.uk [1]British Antarctic Survey, NERC, UKRI, Cambridge, UK [2]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA

biogeochemical regimes in different climate models, which are sometimes difficult to compare directly due to systematic biases [8].

In this note, we apply Gaussian mixture modeling, an unsupervised classification method, to the surface carbon budget derived from a numerical circulation and biogeochemistry model. We find that the surface carbon budget can be described using four different classes that roughly correspond to regimes found in classical theoretical frameworks. We briefly discuss the possibility of using unsupervised learning to compare climate models.

II. Methods

Here we describe the ocean circulation and biogeochemistry model that we used to evaluate the surface carbon budget. We also describe the unsupervised learning method (i.e. Gaussian mixture modeling) that we applied to the surface carbon budget.

A. *Ocean biogeochemistry model*

To quantify the steady state global ocean surface carbon budget, we use a coarse resolution ocean circulation and biogeochemistry model [9]. The numerical model is an instance of MITgcm (http://mitgcm.org/, [10], [11]) with a simple biogeochemistry component [12]. The biogeochemistry package uses six tracers: DIC, alkalinity, PO_4, dissolved organic phosphorous, oxygen, and iron. The export of biological carbon out of the surface is calculated as a function of available light, PO_4, and iron. The model uses a fixed grid with a horizontal resolution of $2.8° \times 2.8°$ in latitude-longitude and 23 vertical levels with gradually increasing cell thickness, with relatively thin cells in the the rapidly-changing surface and thicker cells in the relatively quiescent interior. Unresolved transport is parameterized using an isopycnal thickness diffusion scheme with a uniform diffusivity of 1000 m^2/s [13]. We also impose along-isopycnal diffusion at the same rate [14], and mixed-layer processes are parameterized using the K-Profile Parameterization (KPP) scheme [15]. Vertical diffusivity is set to 0.3×10^{-4} m^2/s in the upper 2000 m and increases to 10^{-4} m^2/s in the interior ocean following an arctangent profile [16]. The Arctic is not included in this model, in part due to convergence issues with latitude-longitude grids near the poles. The model was spun up for 1000 years and then run for another 100 years for evaluation. We average the last 10 years of the simulation in order to construct the steady state budget.

We evaluate the steady state surface carbon budget in the top 185 m of the model domain. The budget can be expressed as follows:

$$\begin{aligned} 0 = &- \mathbf{u} \cdot \nabla_H C - -w \frac{\partial C}{\partial z} \\ &+ \nabla_H (\mathbf{K} \nabla_H C) + \frac{\partial}{\partial z}\left(K_z \frac{\partial C}{\partial z}\right) \\ &+ V_P (1 - f) K_H \Delta p\mathrm{CO}_2 \\ &+ \mathrm{FWF} + \mathrm{Bio}, \end{aligned} \qquad (1)$$

where C is the dissolved inorganic carbon concentration, $\mathbf{u}$ is the horizontal velocity vector, ∇_H is the horizontal component of the gradient operator, w is the vertical velocity, z is the depth coordinate, $\mathbf{K}$ is the diffusivity tensor, K_z is the vertical component of the diffusivity tensor, and $\Delta p\mathrm{CO}_2$ is the air-sea difference in partial pressure of CO_2. The terms on the RHS of equation (1) represent the processes of horizontal advection, vertical advection, horizontal diffusion, vertical diffusion, air-sea gas exchange, freshwater flux, and biological sources and sinks of DIC, respectively. In terms of model diagnostics, the unresolved, parameterized fluxes are contained in the diffusive terms of the budget.

B. *Gaussian mixture modeling*

Gaussian mixture modeling (GMM) attempts to represent the density of data in an abstract space as a linear combination of multi-dimensional Gaussian functions [17]. A GMM is “trained” by adjusting the means and covariances of the Gaussian functions. GMM has been applied to ocean temperature and salinity data in order to identify different “profile types” in different ocean regions [18], [19].

We follow the method of [20], wherein each term of the steady-state, two-dimensional barotropic vorticity budget equation is used as a feature for unsupervised classification; each term/feature represents a different physical process. The result of their classification analysis is a robust, algorithmically defined global geography of ocean dynamical regimes [20]. In our application, we use each term of the surface carbon budget in equation (1) as a feature for classification analysis; in doing so, we represent the distribution of data in a seven-dimensional abstract feature space. At every $2.8° \times 2.8°$ model grid cell, there is a value for each of the seven terms of equation (1). The vector of budget term values from a grid cell is used as a single seven-dimensional “observation” in the clustering analysis. We weight each grid cell by its ocean surface area. We do *not* standardize the budget term values beforehand, as we want the terms to retain their relative magnitudes. This should not affect the GMM fitting procedure, as the covariances of the Gaussian functions are generally allowed to scale as needed to fit the data.

The total number of classes N is a free parameter in GMM. Although one can use statistical tests to estimate the value of N with the highest likelihood relative to an overfitting "penalty" term (e.g. Bayesian Information Criterion or BIC), one can also use N as a description of the complexity of the statistical model. A simple statistical model with small N will likely be easier to interpret than a complex statistical model with large N. In this way, a range of GMM models with different N constitutes a model hierarchy, and one may be able to learn about the system under investigation based on how it changes as one adds or removes sources of complexity [21].

We use the Scikit-learn machine learning package in Python to carry out the classification analysis [22]. We use the expectation-maximization algorithm to determine the means and covariances of the Gaussians that have the highest probability of correctly representing the steady-state budget data as a linear combination of multi-dimensional Gaussian functions. We use the "full" covariance type to allow the Gaussians to change their orientations and covariances in any way that increases the overall probability of the distribution. We use every ocean grid cell from the numerical model, which is 4447 data points (or "observations") in total. Once the GMM has been trained, we use it to assign a class label to each grid cell. Specifically, GMM assigns to each grid cell a probability distribution across all of the classes, and it assigns each grid cell to the class with the maximum posterior probability.

III. RESULTS

In our implementation of GMM, each class broadly represents a different distribution of balances in the terms of equation (1). Specifically, each of the four seven-dimensional Gaussians can be described by a set of means (a seven-dimensional vector) and covariances (a tensor) across the different processes. For simplicity, we only show the means of the classes (Figure 1). We see that there are classes with less biological productivity (e.g. class 2), and classes with more biological productivity (e.g. class 1), corresponding to a large export of DIC from the surface waters. Note that these mean values do not necessarily represent every observation in a given class; they are means of the Gaussian functions used to represent the data. The surface area occupied by each class is shown in Figure 2.

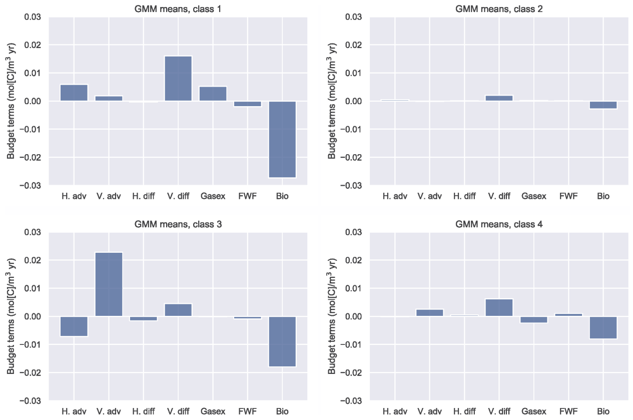

Fig. 1. Each class features a different balance distribution across the terms of the surface carbon budget. Here we plot the means of the four multi-dimensional Gaussian functions used to statistically model the data density. The terms are horizontal advection (H. Adv.), vertical advection (V. Adv.), horizontal diffusion (H. Diff.), vertical diffusion (V. Diff.), air-sea gas exchange (Gasex.), freshwater flux (FWF), and biological sources and sinks of carbon (Bio).

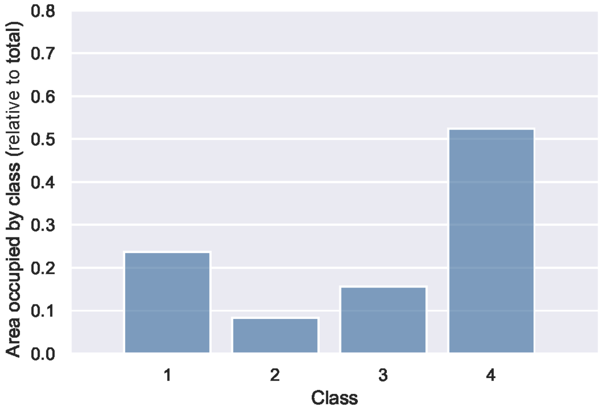

Fig. 2. The surface area occupied by each class, relative to the total ocean surface area.

Here we describe the GMM classes and the classical theoretical frameworks to which they approximately correspond. Despite the fact that GMM was not given any information about the latitude-longitude locations of the grid cells, it is still able to identify spatially coherent regimes in the surface carbon budget (Figure 3). Along with the distribution across processes (Figure 1), the spatial distribution of the labels helps in our attempt to interpret the classes. Class 1 corresponds to the highly productive open ocean, which is dominated by wintertime convection and mixing, seen in the term balance as vertical diffusion (Figure 1). Class

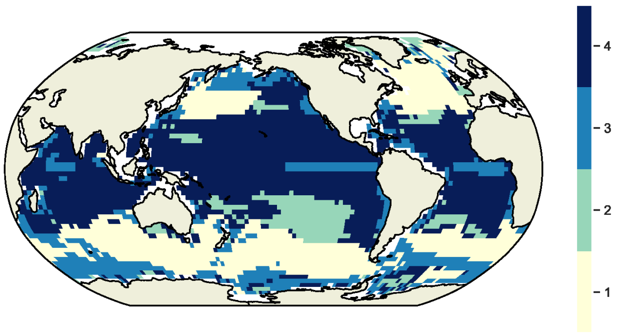

Fig. 3. GMM produces spatially coherent regimes for the surface carbon budget, despite the fact that it is not given any information about the location of the grid cells. Here we show the labels assigned by GMM to each grid cell. Regions with no data are masked out in white.

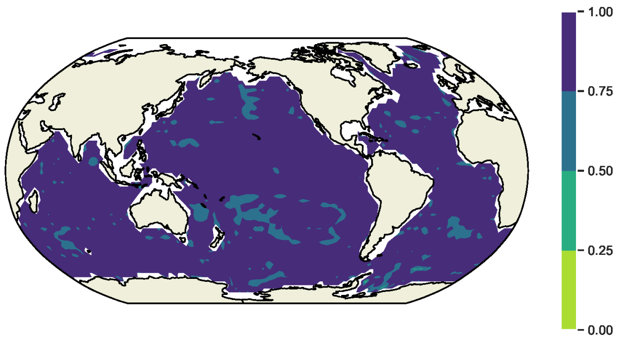

Fig. 4. Posterior probabilities can be used to identify transition regions between classes or where simple characterization of the class type is difficult. Regions with no data are masked out in white.

2 corresponds to relatively isolated patches of low productivity. Class 3 features wind-driven upwelling, as evidenced by the larger values of the advective terms. Both coastal and open ocean wind-driven upwelling can bring nutrients to the surface, where they can enable primary productivity and encourage the export of carbon out of the surface layer. This class is possibly the high nutrient low chlorophyll regime, which features relatively shallow mixed layer depths and (typically) iron limited productivity [23]. Finally, class 4 corresponds to the relatively unproductive gyres and tropics. The subtropical gyres are typically characterized by low local values of productivity, although they may still contribute significantly to the global carbon budget due to their large size [24]. Low productivity in the subtropical gyres is typically explained as a result of large-scale downwelling due to the wind-driven convergence of surface waters, which prevents productivity-enabling nutrients from reaching the surface waters [25]. Overall, the carbon distribution appears to reflect the nutrient budget, i.e. biological productivity is high in regions where upwelling can supply nutrients to the surface and low in regions where downwelling suppresses surface nutrient availability.

For each seven-dimensional observation of the steady state budget terms at a grid cell, GMM calculates a probability distribution across all four Gaussians. It labels each grid cell based on the Gaussian with the highest posterior probability. The maximum posterior probability is a measure of confidence in GMM's assignment of a grid cell to a class and can be used to characterize boundaries between classes. In this application, we find that the maximum posterior probability values are high ($\geq 90\%$) in the tropics and subtropics, with somewhat lower values between classes in the Southern Ocean. This may simply reflect the complex spatial structure of the classes in the Southern Ocean, which features numerous transition regions.

IV. Discussion

In general, using budget terms as inputs to an unsupervised learning algorithm allows us to interpret clustering results in terms of balances between processes, representing an important link between data-derived results and the process-based physical and biogeochemical understanding that underpins much of modern oceanography. This approach may offer a viable bridge between machine learning methods and more traditional approaches.

We use four classes in this example implementation of GMM for ease of interpretation. Based on the BIC score, we could improve the overall likelihood of the GMM by increasing the number of classes to somewhere between 14-19 (see appendix). Although this would enhance the ability of the GMM to statistically describe the data density, it could decrease our ability to understand the results in terms of existing conceptual frameworks. The tradeoff between accuracy of representation and interpretability is a familiar contrast in ocean modeling. One strategy for dealing with this contrast is to use a hierarchical approach, in which we try to learn about a system by comparing models with different levels of complexity [21]. In terms of GMM, this would amount to changing the maximum number of classes and comparing results.

GMM as applied to budget terms may be a useful method for comparing different climate models, for example the ensemble members of the Climate Model Intercomparison Project [8]. These models often features biases with respect to each other, but they display similar physical and biogeochemical regimes characterized by balances between processes. Unsupervised learning may offer a set of methods for objectively identifying these regimes in different models; the properties of the objective regimes could be compared, as opposed to comparing different geographical regions, which are often chosen using crude and somewhat arbitrary latitude-longitude boxes.

One limitation of this study is the relatively coarse resolution of the model; an application of GMM to a high-resolution biogeochemical state estimate like B-SOSE would be a welcome extension to this study [26]. We have also not thoroughly explored the many alternative unsupervised classification methods avalable, including DBSCAN and variational Bayesian approaches.

APPENDIX

Here we present additional information about the GMM classification results. BIC tends to increase as the likelihood of the statistical model increases with the total number of classes N, but that tendency is offset by a penalty term which discourages overfitting. Usually, one would choose the value of N with the minimum value of BIC, if the goal is to create a detailed statistical description of the dataset. The BIC mean score reaches a minimum at 19 classes, although the error suggests that the minimum could be between 14-19 (Figure 5(a)). In order to examine the distinctiveness of the clusters, we use two complementary dimensionality reduction techniques. First, we use principal component analysis (PCA) to project the data onto three PC axes that together explain 91% of the variance (52% PC1, 29% PC2, and 10% PC3). Projections of the principal components into 2D space show that class 1, which corresponds to the highly productive open ocean, is reasonably distinct from the others (Figure 5(b-d)). Class 2 is tightly clustered around the origin, which is consistent with the low values of the flux terms that characterize this class. Next, although classes 3 and 4 have some overlap around the origin, they do have distinct structures in PC space, with class 3 showing a larger spread along the PC1 axis.

For an alternative view on the distinctiveness of the classes, we employ t-SNE, a technique for exploring structures in high-dimensional data [27]. The t-SNE technique creates two-dimensional "maps" from high-dimensional data using non-linear transformations. It has a tunable parameter called "perplexity" which roughly corresponds to the attention paid to local versus global aspects of the data in feature space (see https://distill.pub/2016/misread-tsne/ for details). As we increase perplexity, we see class 1 emerge as a distinct feature. Classes 2-4 have some considerable regions of overlap, but classes 3 and 4 have some distinct lobes above and below the class 1 cluster.

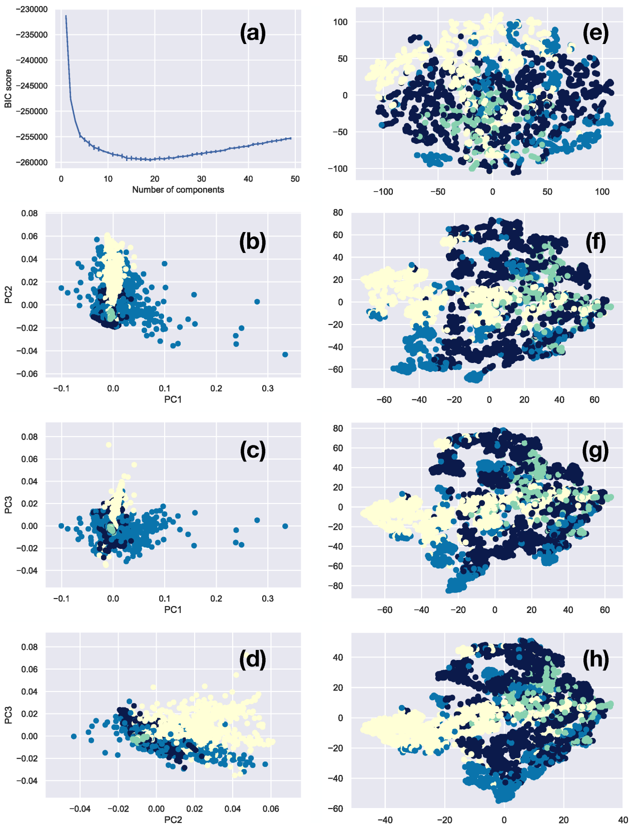

Fig. 5. Additional clustering diagnostics. (a) Mean and standard deviation of BIC scores from 25 independent instances of GMM for each value of N. (b-d) Reduced dimensionality view using a three-component PCA, viewed as three different projections onto 2D space. (e-h) Reduced dimensionality view using t-SNE for perplexity values of 5, 30, 50, and 100, respectively. The axes are the arbitrary t-SNE dimensions. Color values correspond to those in Figure 3.

ACKNOWLEDGMENTS

DJ is supported by the Natural Environment Research Council, UKRI (grant NE/N018028/1). TI is supported by the US National Science Foundation (OCE1737188, OPP-1744755).

REFERENCES

[1] C. Sabine, R. Feely, N. Gruber, R. Key, K. Lee, J. L. Bullister, R. Wanninkhof, C. S. Wong, D. W. Wallace, B. Tilbrook, F. J. Millero, T.-H. Peng, A. Kozyr, T. Ono, and A. F. Rios, "The oceanic sink for anthropogenic CO2," *Science*, vol. 305, no. 367, 2004.

[2] R. Houghton, "Balancing the global carbon budget," *Annual Review of Earth and Planetary Sciences*, vol. 35, no. 1, pp. 313–347, 2007.

[3] S. Khatiwala, F. Primeau, and T. Hall, “Reconstruction of the history of anthropogenic CO2 concentrations in the ocean,” *Nature*, vol. 462, pp. 346 EP –, 2009.

[4] T. Takahashi, S. Sutherland, R. Wanninkhof, C. Sweeney, R. A. Feely, D. W. Chipman, B. Hales, G. Friederich, F. Chavez, C. Sabine, A. Watson, D. C. E. Bakker, U. Schuster, N. Metlz, H. Yoshikawa-Inoue, M. Ishii, T. Midorikawa, Y. Nojiri, A. Kortzinger, T. Steinhoff, M. Hoppema, J. Olafsson, T. S. Arnarson, B. Tilbrook, T. Johannessen, A. Olsen, R. Bellerby, C. S. Wong, B. Delille, N. R. Bates, and H. J. W. d. Barr, “Climatological mean and decadal change in surface ocean pCO2, and net sea-air CO2 flux over the global oceans,” *Deep Sea Research II*, vol. 56, pp. 554–577, 2009.

[5] D. C. Jones, T. Ito, Y. Takano, and W.-C. Hsu, “Spatial and seasonal variability of the air-sea equilibration timescale of carbon dioxide,” *Global Biogeochemical Cycles*, vol. 28, no. 11, pp. 1163–1178, 2014.

[6] C. R. McClain, S. R. Signorini, and J. R. Christian, “Subtropical gyre variability observed by ocean-color satellites,” *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 51, no. 1, pp. 281 – 301, 2004. Views of Ocean Processes from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) Mission: Volume 1.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[8] A. Anav, P. Friedlingstein, M. Kidston, L. Bopp, P. Ciais, P. Cox, C. Jones, M. Jung, R. Myneni, and Z. Zhu, “Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models,” *Journal of Climate*, vol. 26, no. 18, pp. 6801–6843, 2013.

[9] Y. Takano, T. Ito, and C. Deutsch, “Projected centennial oxygen trends and their attribution to distinct ocean climate forcings,” *Global Biogeochemical Cycles*, vol. 32, no. 9, pp. 1329–1349, 2018.

[10] J. Marshall, A. Adcroft, C. Hill, and L. Perelman, “A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers,” *Journal of Geophysical Research*, vol. 102, pp. 5753–5766, 1997.

[11] J. Marshall, C. Hill, L. Perelman, and A. Adcroft, “Hydrostatic, quasi-hydrostatic, and nonhydrostatic ocean modeling,” *Journal of Geophysical Research*, vol. 102, pp. 5733–5752, 1997.

[12] S. Dutkiewicz, M. J. Follows, and J. G. Bragg, “Modeling the coupling of ocean ecology and biogeochemistry,” *Global Biogeochemical Cycles*, vol. 23, no. 4, p. GB4017, 2009.

[13] P. Gent and J. Mcwilliams, “Isopycnal mixing in ocean circulation models,” *Journal of Physical Oceanography*, vol. 20, pp. 150–155, 1990.

[14] M. Redi, “Oceanic isopycnal mixing by coordinate rotation,” *Journal of Physical Oceanography*, vol. 12, pp. 1154–1158, 1982.

[15] W. Large, J. Mcwilliams, and S. Doney, “Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization,” *Reviews of Geophysics*, vol. 32, no. 4, pp. 363–403, 1994.

[16] K. Bryan and L. J. Lewis, “A water mass model of the world ocean,” *Journal of Geophysical Research: Oceans*, vol. 84, no. C5, pp. 2503–2517, 1979.

[17] D. Reynolds, *Gaussian Mixture Models*, pp. 659–663. Boston, MA: Springer US, 2009.

[18] G. Maze, H. Mercier, R. Fablet, P. Tandeo, M. L. Radcenco, P. Lenca, C. Feucher, and C. Le Goff, “Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean,” *Progress in Oceanography*, vol. 151, pp. 275–292, 2017.

[19] D. C. Jones, H. J. Holt, A. J. S. Meijers, and E. Shuckburgh, “Unsupervised Clustering of Southern Ocean Argo Float Temperature Profiles,” *Journal of Geophysical Research - Oceans*, vol. 40, no. 2, pp. 1556–13, 2019.

[20] M. Sonnewald, C. Wunsch, and P. Heimbach, “Unsupervised Learning Reveals Geography of Global Ocean Dynamical Regions,” *Earth and Space Science*, vol. 6, p. 784 794, 2019.

[21] I. Held, “The gap between simulation and understanding in climate modeling,” *Bulletin of the American Meteorological Society*, vol. 86, no. 11, pp. 1609–1614, 2005.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] J. Pitchford and J. Brindley, “Iron limitation, grazing pressure and oceanic high nutrient-low chlorophyll (HNLC) regions,” *Journal of Plankton Research*, vol. 21, pp. 525–547, 03 1999.

[24] W. J. Jenkins and S. C. Doney, “The subtropical nutrient spiral,” *Global Biogeochemical Cycles*, vol. 17, no. 4, 2003.

[25] R. G. Williams and M. J. Follows, *Ocean Dynamics and the Carbon Cycle: Principles and Mechanisms*. Cambridge University Press, 2011.

[26] A. Verdy and M. R. Mazloff, “A data assimilating model for estimating Southern Ocean biogeochemistry,” *Journal of Geophysical Research - Oceans*, vol. 122, no. 9, pp. 6968–6988, 2017.

[27] L. v. d. Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.