Evaluation of Machine Learning Techniques for Estimating Biogeochemical Properties on Seaglider Tracks in the Southern Ocean

Zachary Nachod

University of Washington, Seattle, WA

School of Oceanography

znachod@uw.edu

3/11/22

Abstract

The Southern Ocean is the largest oceanic carbon and heat sink on the planet with complex dynamics at a variety of scales. Reliable, accurate, and high resolution estimates of nitrate and carbonate system parameters (hereafter biogeochemical estimates) in the Southern Ocean would enable the analysis of mesoscale and submesoscale biogeochemical processes throughout the water column. This work explores the use of multiple methods, including several from the machine learning literature, for biogeochemical parameter estimation in the Southern Ocean. Training data for this work includes temperature, salinity, oxygen, and nitrate measurements from the 2019 R/V Thomas G. Thompson reoccupation of the I06S line and from Southern Ocean Carbon and Climate Observations and Modeling project (SOCCOM) floats deployed during this cruise. Four models for the estimation of nitrate were trained and validated for accuracy; these models included a random forest regression, a generalized additive model, a multiple linear regression, and a gradient boosted regression tree model. The random forest regression performed the best out of the four machine and statistical models on our nitrate test data with a median value for the absolute error of 0.09 μmol kg^{-1} and an interquartile range of 0.13 μmol kg^{-1} in the absolute error. Using this random forest model, we predicted the nitrate concentrations along the high resolution tracks of two Seagliders deployed on this cruise. We plan to later repeat this estimation process for pH along the Seaglider tracks as well. The nitrate and pH estimates from the random forest model can be used to improve our understanding of mesoscale and submesoscale processes related to carbon flux in this region of the Southern Ocean.

Plain Language Summary

In oceanography and other physical sciences, data science has gained increasing importance as more oceanographic instruments with enhanced resolution and accuracy are deployed in the world's oceans, resulting in a large increase in the number of observations. This research finds the statistical relationships (i.e., models) between quantities that have been directly measured and uses them to estimate nitrate on tracks of a remotely piloted ocean observing instrument called a Seaglider. These statistical relationships are found using data from research ships and autonomous ocean observing devices called Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) floats. The random forest model performed the best out of the four models tested, which included a generalized additive model, multiple linear regression, and a gradient boosted model. The random forest model was the best performing model as it had the least amount of error. Using the random forest model, we were able to estimate nitrate on two Seaglider tracks deployed on this cruise. These estimates along the Seaglider tracks allow us to look at processes that occur over shorter distances and periods of time than what SOCCOM floats and the ship can measure. This would enable the study of smaller processes that affect the planet's climate.

Introduction

Motivation

The planet is changing rapidly due to climate change. Climate change is a looming threat to society that can seem overwhelming to try to study and understand, let alone find a solution for. Individuals around the world are studying changes to environmental variables such as temperature, carbon dioxide content in the atmosphere, and even the nutritional distribution of

soil. There is so much data available that it has become next to impossible to manually sift through it to gain valuable insights. In oceanography and other physical sciences, data science has gained increasing importance as a way to handle the deluge of data from oceanographic instruments, which have increasingly enhanced resolution and accuracy.

Machine learning has revolutionized fields such as medicine, mathematics, and artificial intelligence (Ahmad 2019). While machine learning has already been used in oceanography for a variety of applications including species identification (Simmonds et al. 1996), wave modeling (James et al. 2018), and predicting climatology (Hsieh 2009), its use is still in its infancy. While previous research has shown that predicting dissolved oxygen with machine learning in the Southern Ocean is possible (Giglio et al. 2018), this paper explores the use of machine learning models in order to predict other biogeochemical variables in addition to oxygen in this basin.

Background

The global ocean has absorbed 24% of the world's total anthropogenic carbon (Friedlingstein et al. 2020) and 90% of the heat gained by the planet in the last few decades (Zanna et al. 2019). Out of the major oceans, the Southern Ocean plays a particularly significant role in global carbon and heat storage (Chen et al. 2019). The Southern Ocean completely encompasses the Antarctic, connecting all major oceans besides the Arctic Ocean, making it one of the most important oceanic regions. The Southern Ocean serves as a primary heat and carbon sink for the planet, accounting for 75% of global oceanic excess heat uptake and 40% of global oceanic excess carbon uptake (Chen et al. 2019). As more anthropogenic carbon enters the ocean, negative consequences such as ocean acidification will increase in magnitude. In addition, the likelihood of deoxygenation increases due to an increase in stratification (Turner et al. 1987). Stratification in the ocean increases due to an increase in global temperatures (Shepherd et al.

2019) from an increase in anthropogenic carbon in the atmosphere. Mesoscale biogeochemical changes from processes such as carbon flux can be modeled well (Hauck et al. 2013), however widescale direct observations of these impacts are not easily obtained.

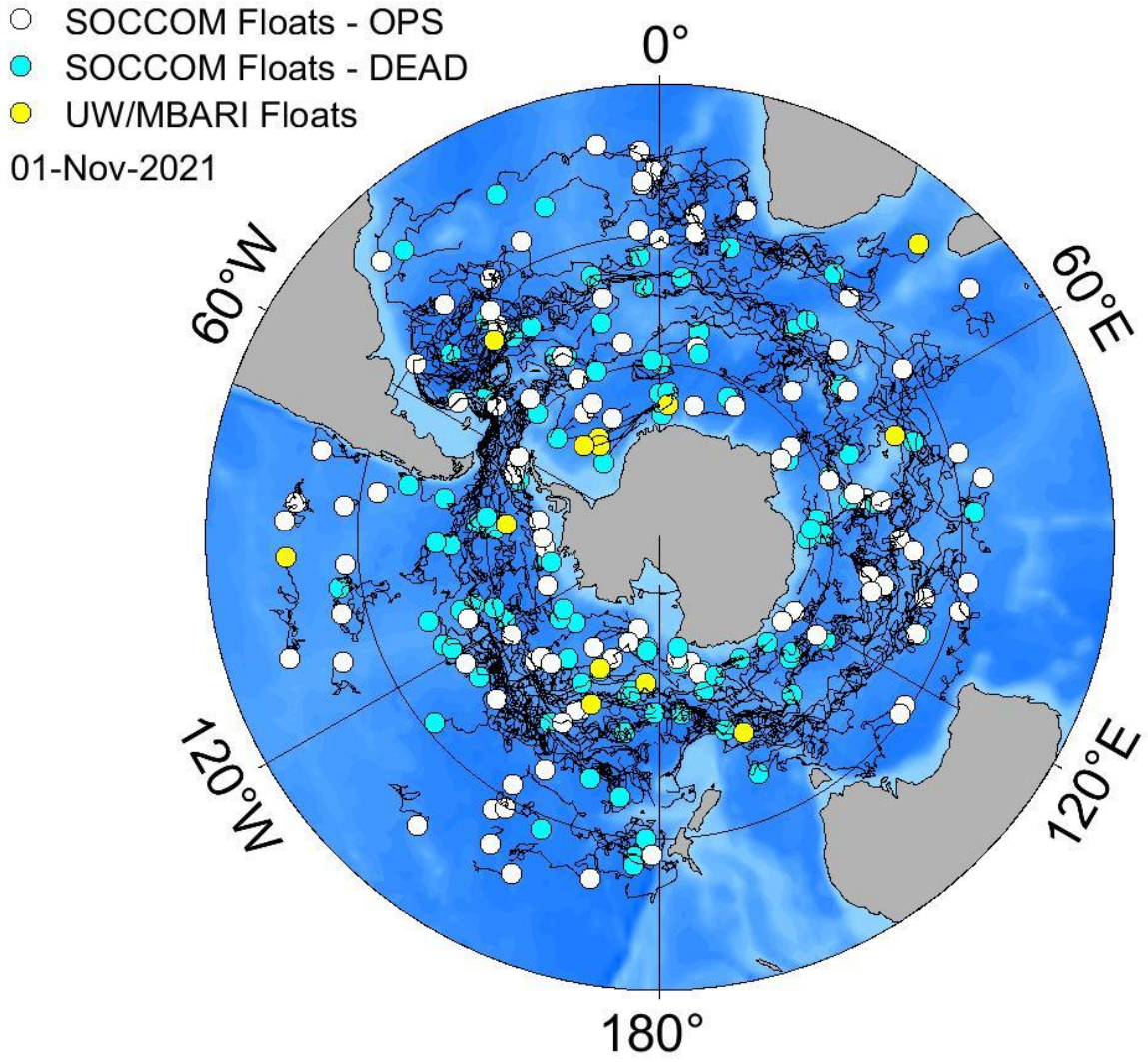

Figure 1. Map of active SOCCOM floats. Yellow and white are operational floats while blue are non-operational floats. Source: (The Trustees of Princeton University 2020)

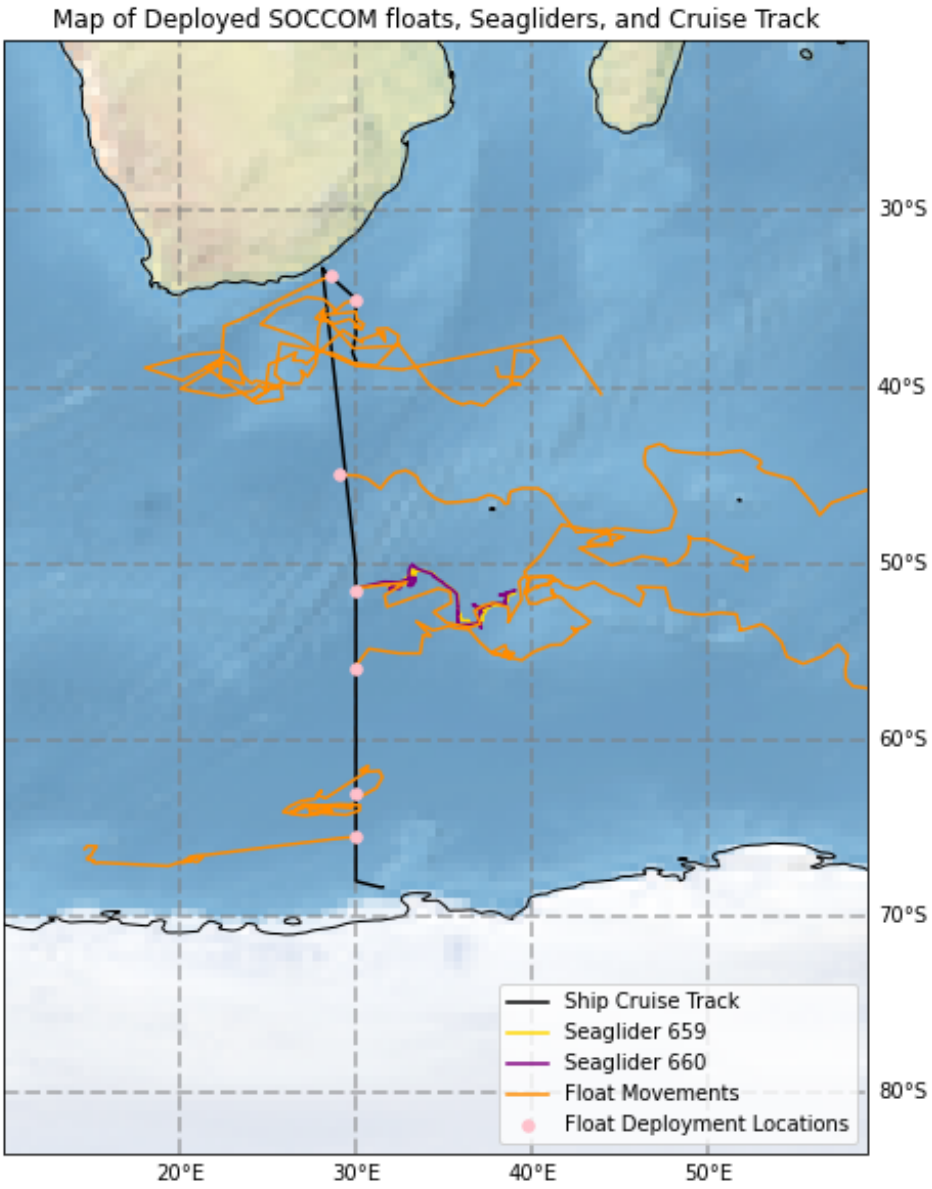

Figure 2. Map of the area of interest in the Southern Ocean showing the ship track, Seaglider tracks, float deployments, and float motion.

Another reason to study the Southern Ocean in this project is that there are already many ocean observing instruments in place. In the region of the Southern Ocean we are studying there were three major sources of relevant oceanographic data collected from 2019 to 2022 (Figures 1 and 2). The first two are Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) floats (Russell et al. 2014) and shipboard data from nearby hydrographic

expeditions which both collect conductivity-temperature-depth (CTD) and biogeochemical data such as oxygen, nitrate, and pH. SOCCOM floats collect data from the surface to between 1400 to 2000 meters depth whereas the cruises collect data to the seafloor. These floats collect relatively low resolution spatial (2–5 meter resolution) and temporal (5–10 day resolution) data, impeding the study of mesoscale and submesoscale processes in the ocean like carbon flux. The third type of instrument used to collect data are Seagliders which collect data at a much higher temporal and horizontal resolution from the surface to 1000 meters depth. This enables submesoscale processes to be observed though typically fewer properties are measured than in the float and ship datasets. What if we could estimate high resolution biogeochemical properties (e.g., nitrate) from the properties that are measured by the Seaglider? Previously, oxygen at 150 meters was predicted using temperature, salinity, location, and time information from Argo floats in the Southern Ocean using the random forest model (Giglio et al. 2018). Building on their approach, various models were trained and tested to predict nitrate at multiple depths along the tracks of two Seagliders of interest in the Southern Ocean.

Scope and Limitations

Machine learning models heavily depend on the quality and quantity of their training data. To predict biogeochemical data in the Southern Ocean there are only so many sources of data. The biggest source of data, SOCCOM floats, only have a population of about 100 deployed floats (Figure 1) and most of these floats are not relevant to the location of the Seaglider tracks (Figure 2) and cannot be used. In total, we use data from 543 SOCCOM float profiles and 55 shipboard profiles. Seaglider data in the Southern Ocean are even more limited as these instruments are expensive to deploy, need a pilot, and the low frequency of Southern Ocean research cruises makes it so there are few tracks historically available. Therefore, the success of

the machine learning models could be limited by the amount of data available for training. The small amount of data also emphasizes the importance of making sure the data is accurate and organized through data cleaning and data quality checking. Over time more data will become available through existing and future SOCCOM floats as well as from other float arrays through the Global Ocean Biogeochemistry Array (GO-BGC) program.

Methods

Data Sources

The training dataset consists of the SOCCOM float data and nearby shipboard hydrographic data. The shipboard data is from the R/V Thompson when it reoccupied the I06S line between South Africa and Antarctica along 30°E in 2019. Both of these datasets are publicly available and updated regularly through Princeton's SOCCOM data access portal (https://soccom.princeton.edu/content/data-access) and Scripps' CLIVAR and Carbon Hydrographic Data Office (CCHDO) cruise data portal (https://cchdo.ucsd.edu/cruise/325020190403), respectively. The SOCCOM float dataset is composed of seven floats each making millions of measurements over their lifespan (Figure 2). These seven floats were all deployed from this cruise and therefore overlap with the ship track and Seaglider data temporally and spatially. The files specifically used for the SOCCOM floats are the SPROF (synthetic profiles) data files in the Argo Global Data Assembly Centers (GDAC) in the netCDF file format. Other cruise data in the Southern Ocean are few and far between due to the undesirable conditions of this basin for sailing.

Testing Models

To test each model's accuracy, the overall dataset is split into 3 different subsets of data, the training, validation, and test datasets, consisting of 80%, 10%, and 10% of the full dataset, respectively. The training dataset is then used to train the models with all available parameters including the parameters that are intended to be estimated. After the models have been fitted the training subset of data (without the nitrate data) is fed through the model. Since the model has seen these exact points before as these were the data points the model was trained on, the error between the predicted and measured points should be fairly small. Each model uses all of the predictor training variables (Table 1) to make their predictions.

Table 1. Variables from the shipboard hydrographic and SOCCOM float data used to train and test the various models.

Variable Name	**Units**	**Variable Type**
Position (latitude, longitude)	Degrees	Predictor
Month of year (as two variables using sine and cosine, following Sharp et al. 2021)	N/A	Predictor
Pressure	Decibars	Predictor
Conservative Temperature	°C	Predictor
Absolute Salinity	g kg^{-1}	Predictor
Dissolved Oxygen	µmol kg^{-1}	Predictor
Nitrate	µmol kg^{-1}	Predictand

The validation dataset is then fed through each model again using all of the variables except nitrate. Since nitrate measurements exist in the training and validation datasets, the absolute error can be calculated by subtracting the estimated data from the measured data and taking the absolute value. This gives an array of absolute error values the size of the dataset fed

through the models for each datapoint. The distribution of the absolute error for each model can be summarized through the median value and interquartile range for each model's absolute error. Different configurations of the models were tested in this fashion to try and find the most accurate version of each type of model. The lower the absolute error values, the better the model is at estimating nitrate. After finding an acceptable iteration of each model, the test subset of data is finally fed through the models. The test data has not been touched or looked at before this point, so the test data provides an unbiased error value for each model to show how well it can make predictions using data that was not used to build the model.

Training Models

Multiple types of statistical/machine learning models were explored for estimating the biogeochemical variables of interest (nitrate) including: a generalized additive model (GAM), a random forest model (Breiman 2001), a gradient boosted regression tree model (XGBoost (Chen et al. 2016)), and a multiple linear regression (MLR). Multiple linear regression uses multiple predictor variables to generate a linear line of best fit no matter the true linearity of the predictor variables. MLR is historically the primary method of data estimation used by scientists. Thus, we used it as a baseline to see how much our other models improved over classical methods of data estimation. The GAM model is a generalized linear model that's predictor variables can utilize smooth functions which is a function that has continuous derivatives up to a desired order over an established domain. These smooth functions allow for the GAM model to use nonlinear predictor variables, an improvement over the MLR model where nonlinear variables are present. The random forest model utilizes multiple randomly generated decision trees which are trained on different bootstrapped subsamples of the training dataset to produce a model with reduced variance. Each tree produces a class prediction and the class with the most votes from all the

trees becomes the final prediction of the model. The gradient boosted regression tree model is similar to the random forest model where multiple decision trees are generated. However, the gradient boosted model creates sequential weak decision trees that improve overtime that result in a final stronger decision tree used for the final prediction of the model.

Training the models involves giving each model a series of data points that provide the model with a reference to the patterns and behaviors of the dataset it is trying to predict. This includes having preliminary data of both variables we want to estimate and the variables we think can be used for this estimation. The models selected all have different settings and characteristics to tweak to find the best way to predict biogeochemical properties from the float and shipboard data.

Various positional, temporal, and biogeochemical variables were included in the training dataset for the model (Table 1). The position variables are latitude and longitude measured in decimal degrees. The temporal variables included two variables for the month of the year derived from the equations: $mn_{\sin} = \sin(2\pi \times mn/12)$ and $mn_{\cos} = \cos(2\pi \times mn/12)$ where mn is the month (Sharp et al. 2021). Both of these variables cycle from -1 to 1 but since they are offset from one another they allow the month information to be represented in a cyclic way from year to year in the model. The standard CTD variables include pressure (decibars), salinity (g kg^{-1}), and temperature (°C). The rest of the biogeochemical data was measured by either a Niskin bottle on the ship or by various sensors on the SOCCOM floats.

MATLAB was used to read in the shipboard data, apply the quality control (QC) flags, and create a .csv file of the remaining data to be read into Python. Python was used to train the various models as well as to read in the SOCCOM float data and apply the appropriate QC flags. The Pandas, Xarray, and NumPy packages were used to create and organize the full training

dataset for the models (see Appendix A for URL links to these packages). Creating the full dataset involved combining both the shipboard and float data into one dataset with the same variables and units. To perform conversions between oceanographic variables, the Python implementation of the Gibbs SeaWater (GSW) Oceanographic Toolbox of the International Thermodynamic Equation Of Seawater – 2010 (TEOS-10) was used.

To create the random forest and multiple linear regression models, the scikit-learn package was used. For the generalized additive model, the pyGAM package was utilized. For the gradient boosted model, the XGBoost package was used. The creation of the generated figures was assisted by the Matplotlib, Seaborn, and Dataframe_image packages. The calculation of the error statistics was done using the SciPy and scikit-learn packages.

The generalized additive model allows for each predictor variable going into the model to be fit with either a linear or spline function. The default of the GAM model is to model every variable with a spline function. However, some of the predictor variables have more linear relationships to the variables we want to predict, therefore it is better to incorporate both types of functions to increase performance. We use eight predictor variables (latitude, longitude, mn_{sin}, mn_{cos}, pressure, temperature, salinity, and dissolved oxygen) to make a model to estimate nitrate, which means there are 256 different combinations of these function types. Through testing each combination iteratively and obtaining the error statistics of each model iteration, the most optimized GAM model was found. The best function type for each predictor variable is spline except for the longitude variable where the best function was linear.

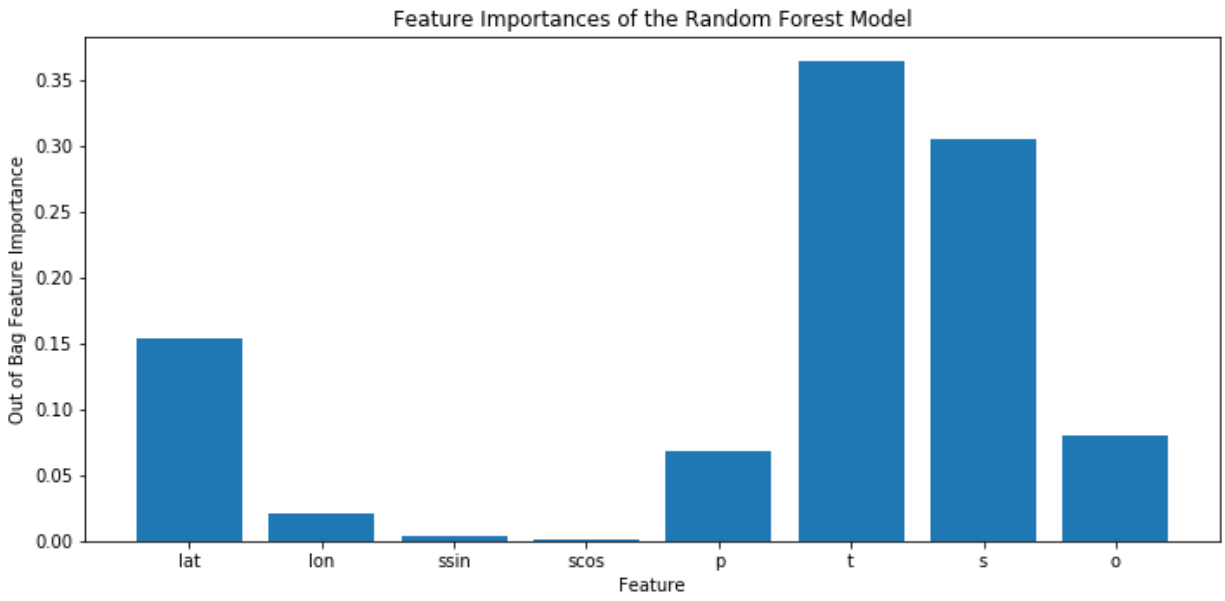

Figure 3. Out-of-bag feature importances of the random forest model. The importances are quantified on a scale of 0 to 1, such that the sum of all 8 feature importances is equal to 1. The higher the individual feature importance, the more predictive ability that variable has in the model.

The random forest model was optimized in multiple ways by adjusting parameters fed to the model. The first optimization made was setting the number of trees generated to 1000. Generating more trees would be redundant and the error change would be negligible, while less than 1000 trees would not be enough to create an accurate model. The maximum number of features was set to ⅓ so that the nearest integer to N/3 (where N = 8, the number of predictor variables/"features") number of features are considered at each tree split. This value for this parameter is usually optimal for random forest regression (Breiman 2001). The random forest model also provides metrics of relative feature importance once the model is created. These metrics show what features are the most important in predicting the response variable. When we initially ran the random forest model the longitude and month variables had the least amount of importance (Figure 3). These variables were then removed one at a time in the next iteration of the model to see if the error became smaller from their absence. We found that the model with all training variables still performed the best, while the model that removed the longitude and kept

the month variables performed the worst. However, the model that kept longitude but removed the month variables was very close in error to the model that kept all training variables.

Predicting on the Seaglider Tracks

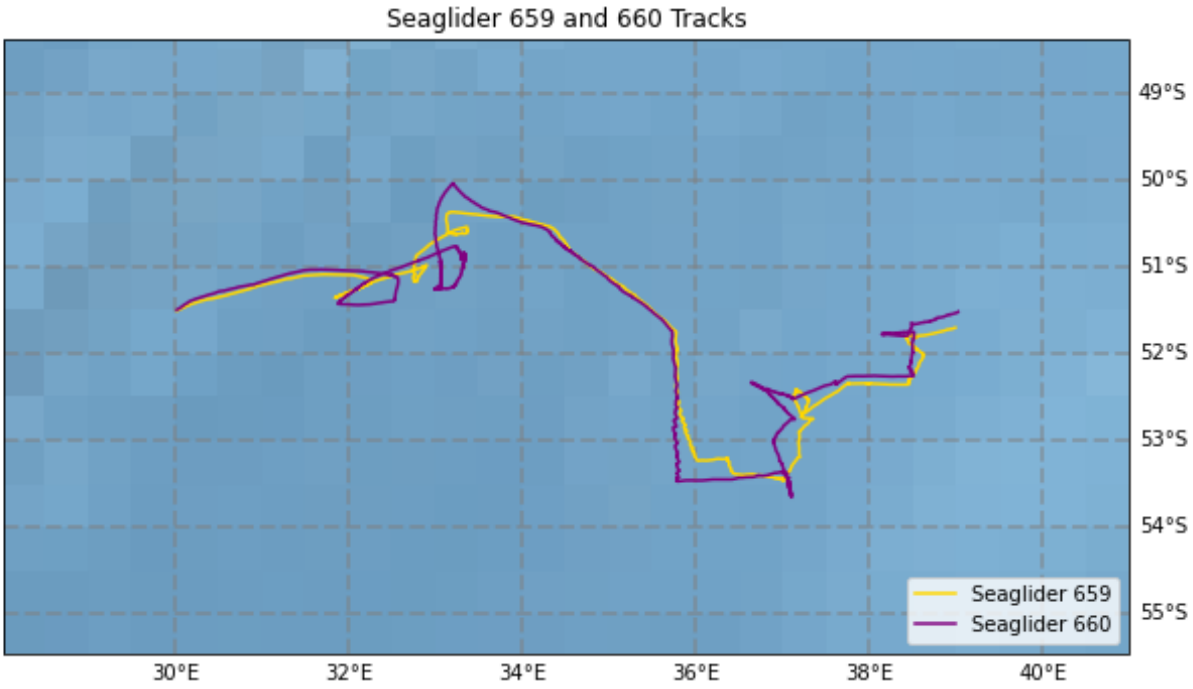

Figure 4. Zoomed in map of Seaglider tracks deployed off of the R/V Thompson.

After the models are tested and their errors are analyzed, one of the models will be selected to predict data on the two Seaglider tracks (Figure 4). The estimated nitrate will then be used to investigate mesoscale and submesoscale biogeochemical processes. Once this work is replicated for pH data, the pH data can be used with alkalinity to calculate carbon flux on the track. An idea of the possible error in the Seaglider track estimations will be known as the model error and variance would be quantified at this time.

Results

Each of the four models successfully predicted nitrate concentrations from the training, validation, and test subsets of data. From calculating the mean, the standard deviation, the

absolute median, and the interquartile range, the random forest model shows the least amount of error in estimating nitrate followed closely by the gradient boosted model. On the other hand, multiple linear regression had the most error out of the four models and the generalized additive model showed intermediate error in nitrate estimation.

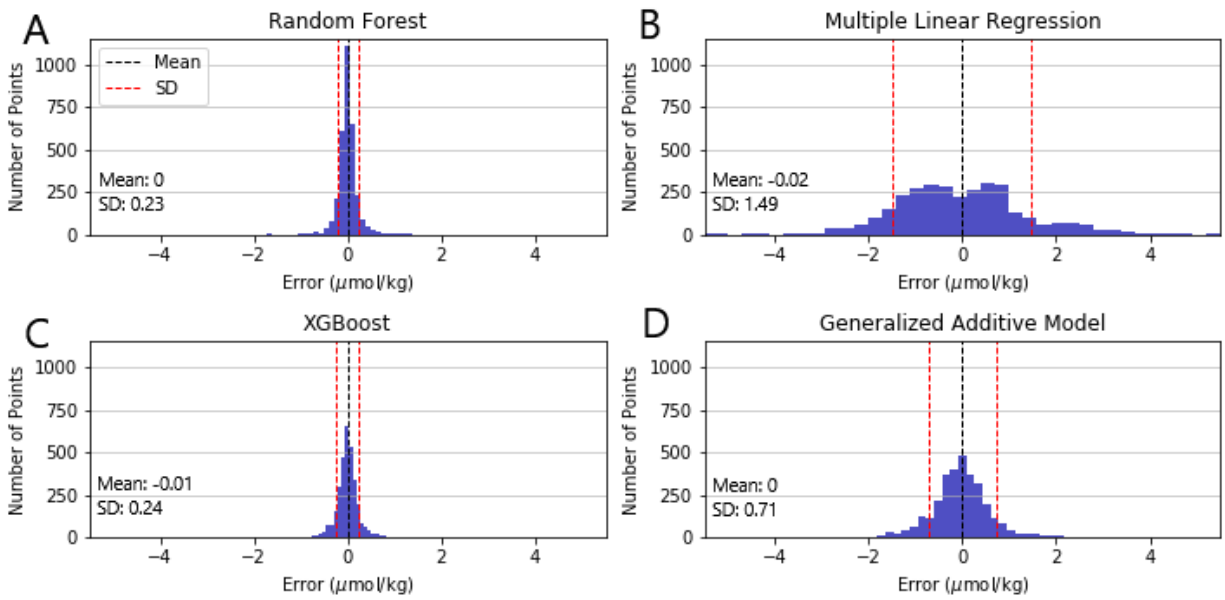

Figure 5. Histograms of the error between measured and predicted nitrate values for the validation dataset. The red dotted lines represent the standard deviation while the black dotted line represents the mean.

Histograms of the error distributions for the nitrate predictions (Figure 5) from each of the models gives insight into the accuracy of each model. Each of the four models have a mean centered close to zero. However, their standard deviation ranges vary widely. The random forest model (Figure 5a) has an absolute error range of -3.7 to 1.8 µmol $kg^{-1}$with a standard deviation range of 0.23. The multiple linear regression model (Figure 5b) has an absolute error range of -8.63 to 6.4 µmol $kg^{-1}$with a standard deviation range of 1.49. The gradient boosted model (Figure 5c) has an absolute error range of -2.23 to 1.79 with a standard deviation of 0.24. The generalized additive model (Figure 5d) has an absolute error range of -3.38 to 5.28 µmol $kg^{-1}$with a standard deviation range of 0.71.

Table 2. Median and interquartile range of the absolute error for the predictions using the four models. Statistics for the training and validation data subsets are shown.

Model	**Training Median**	**Training IQR**	**Validation Median**	**Validation IQR**	**Test Median**	**Test IQR**
Random Forest	0.03	0.05	0.08	0.12	0.09	0.13
MLR	0.87	1.73	0.86	1.70	0.84	1.68
GAM	0.33	0.66	0.33	0.66	0.33	0.68
XGBoost	0.09	0.13	0.11	0.16	0.11	0.16

We also calculated the median and interquartile range of the absolute error for each model in order to summarize their predictive skill. The absolute median depicts the center of a skewed distribution, the closer to zero the value is the more accurate the model is. The interquartile range (IQR) of the absolute median gives the range where the middle 50% of the data lies. The smaller the IQR, the smaller or narrower the distribution of error is. If the median is close to zero and the IQR is small, the more accurate the model is. The random forest model is at least an order of magnitude smaller than the MLR model and the GAM model in both the median absolute error and its IQR for the training, validation, and test data (Table 2). XGBoost is not far behind the random forest model with a test median difference of 0.02 and an IQR difference of 0.03. Similar relationships between the performance of the four models in the test data when compared to the validation data solidifies the trends seen in the error data (Figure 5; Table 2) proving that the results are unbiased and replicable. Finally, we assess the performance of the different models by looking at the absolute errors for the test data in pressure-date space to see if there is any obvious structure to the error fields (Figures 6c and 7c). Any such structure would indicate that our model is potentially missing something that could be included to improve the prediction.

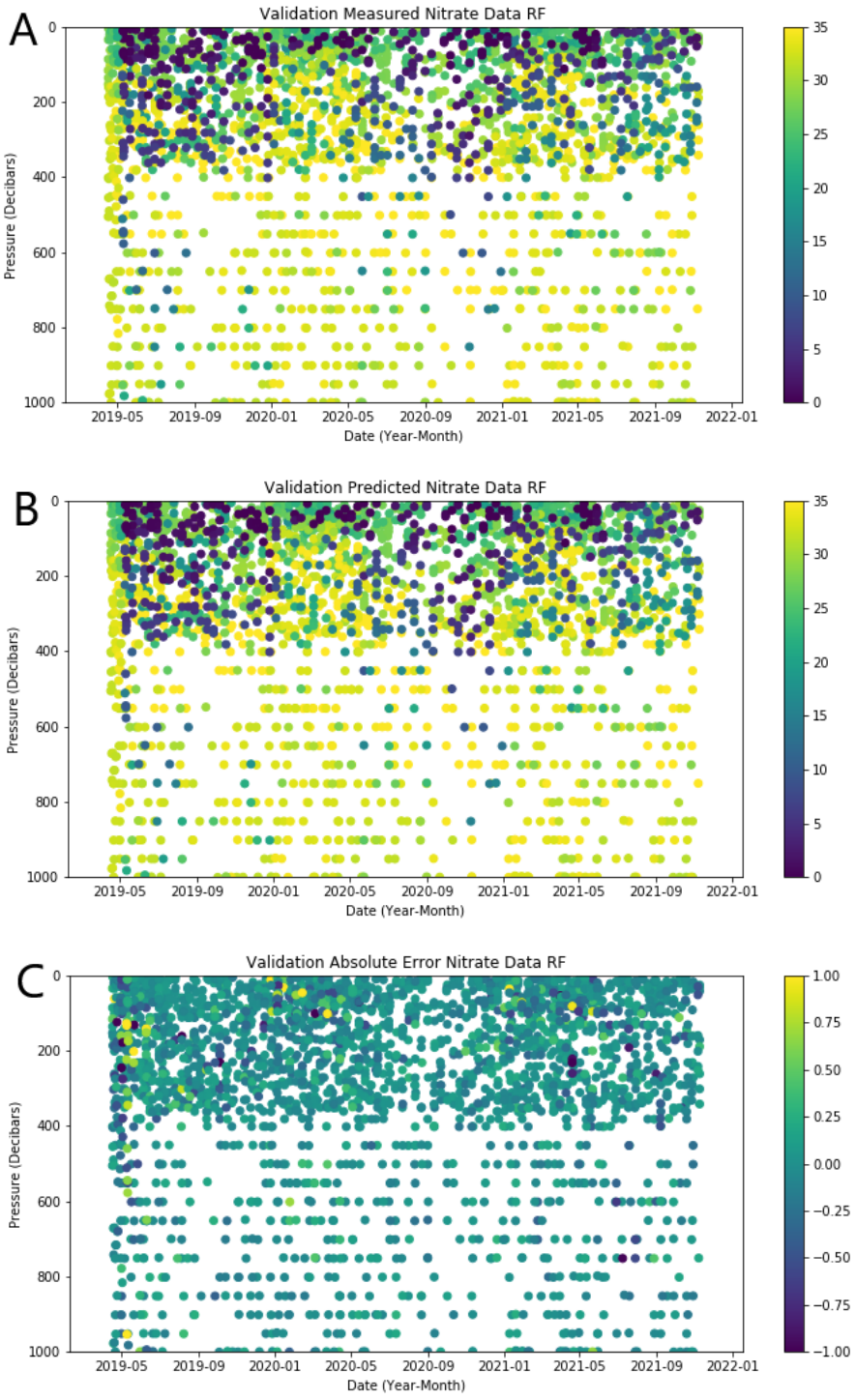

Figure 6. Scatter plots of A) measured (by ship hydrography and SOCCOM floats) and B) predicted (using the random forest) nitrate values for the validation dataset (10% of total dataset) and C) the absolute difference between these values. Nitrate is measured in µmol kg^{-1}. The increased frequency of points over the span of 0 to 1000 dbar on the left of each graph (before 2019-05) is the data from the ship. The data in the rest of the figure (i.e., later times) are from SOCCOM floats, which have a higher sampling rate in the upper 400 meters of the water column than at depth.

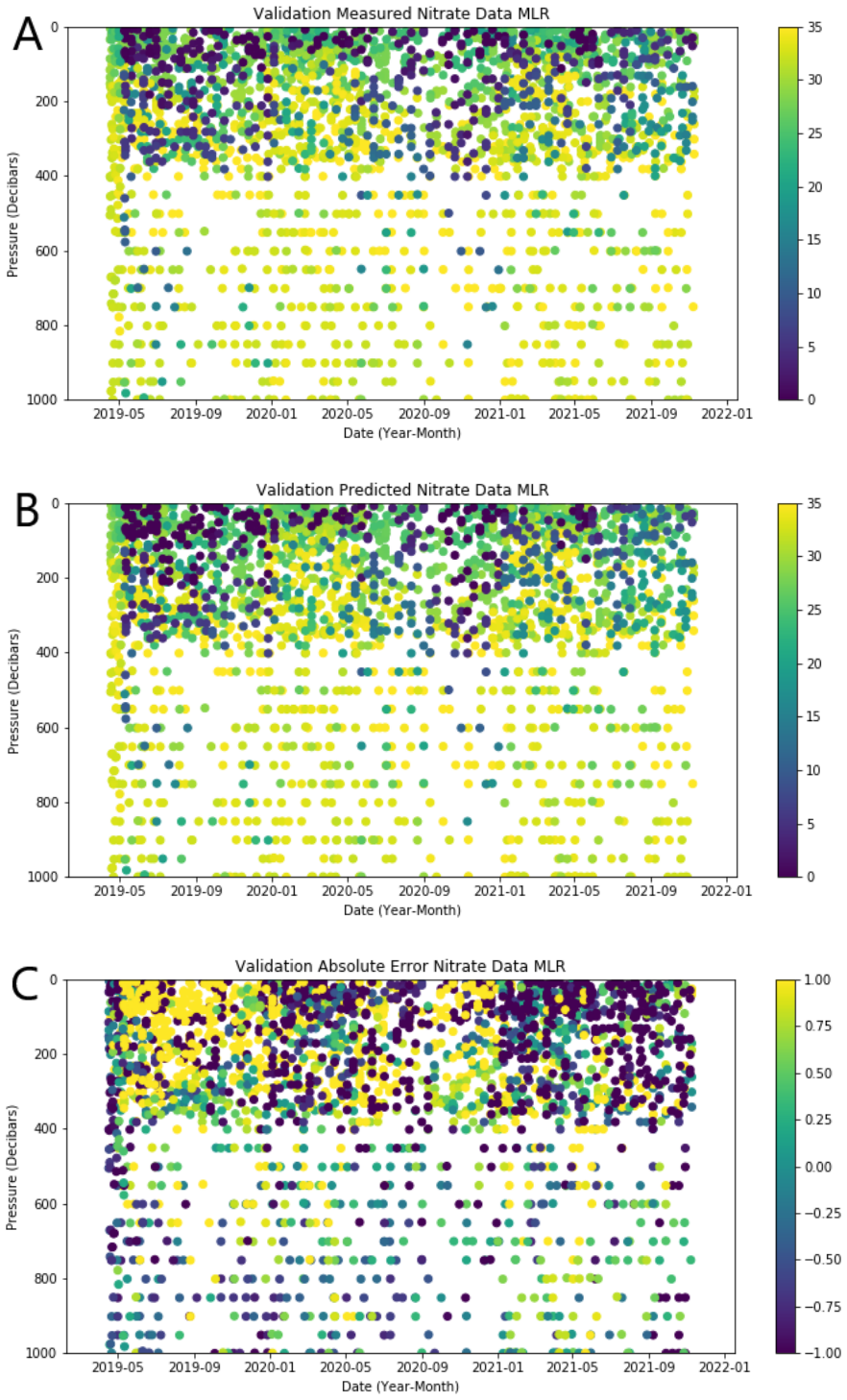

Figure 7. Scatter plots of A) measured (by ship hydrography and SOCCOM floats) and B) predicted (using multiple linear regression) nitrate values for the validation dataset (10% of total dataset) and C) the absolute difference between these values. Nitrate is measured in μmol kg^{-1}. The increased frequency of points over the span of 0 to 1000 dbar on the left of each graph (before 2019-05) is the data from the ship. The data in the rest of the figure (i.e., later times) are from SOCCOM floats, which have a higher sampling rate in the upper 400 meters of the water column than at depth.

Discussion

The random forest model performed the best on the validation dataset out of the four models with the smallest standard deviation in its error histogram and the smallest IQR in the absolute error (Figure 5; Table 3). The MLR model performed the worst, with the widest range of error values on the validation data. The MLR also had a high median absolute error and interquartile range in the absolute error for the training data which indicates that the model may have been poorly trained due to the nonlinear nature of the predictor variables (Table 3).

Inspecting the measured and predicted nitrate data for the random forest model throughout the water column (Figure 6c) confirms that the absolute error between the two sets of values is generally fairly small. The small nature of the absolute error is further supported by the small absolute median and IQR values (Table 3). These considerations were factored into selecting the random forest model to be the most accurate model for the use in estimating nitrate along the Seaglider tracks (discussed below). Additionally, upon looking at the measured and predicted nitrate data for the MLR throughout the water column, there are obvious differences between the two sets of values (Figure 7). There are noticeable patterns to the differences between the measured and predicted values which means there is some clear variability the MLR model is not accounting for. Above ~400 dbar, this variability appears to resemble a seasonal cycle.

The validation and test subset of data points both pointed to the random forest model being the best model. The model was run with the test data and found the absolute error median to be 0.09 and the test IQR to be 0.13, both values being close to their validation subset of data counterparts. This error is the uncertainty of the nitrate concentration estimation on the Seaglider

estimates. The smaller the error is, the less uncertainty there is in the estimations which makes the produced data more accurate and reliable.

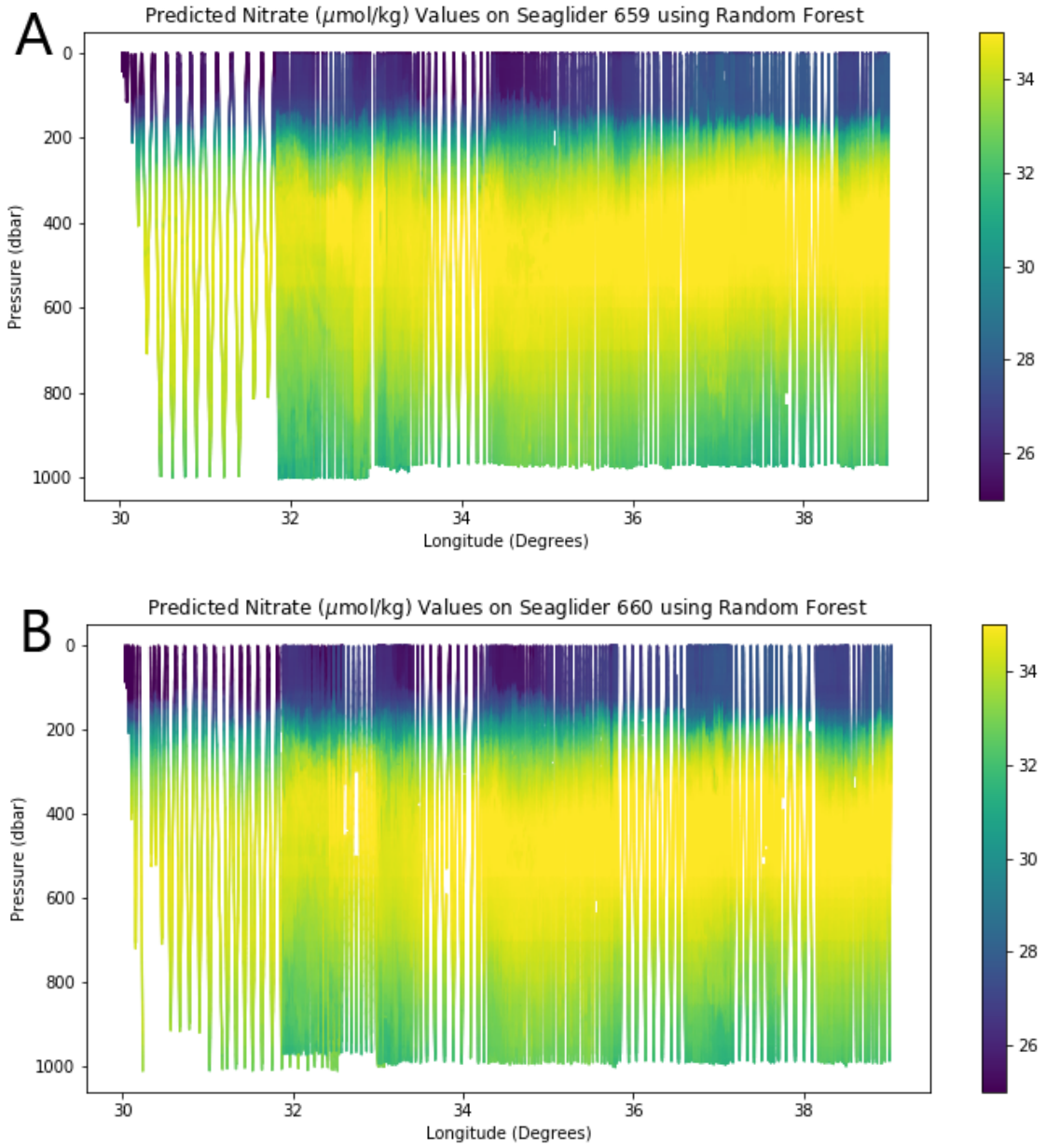

Figure 8. Predicted nitrate values (μmol kg^{-1}) on A) Seaglider 659 and B) Seaglider 660's track using the random forest model.

We estimated nitrate along the two Seaglider tracks (Figure 8) using the best performing model (the random forest model) on the latitude, longitude, month, pressure, temperature, salinity, and oxygen data from these two Seagliders. The nitrate estimates generally agree with

what is known about the region and is plausible data for this region of the Southern Ocean . The general trend between the latitudes 50°S and 53°S for nitrate concentrations are lower concentrations of 25 - 30 μmol kg^{-1} at the surface from 0 to 200 meters. A peak of ~34 μmol kg^{-1} of nitrate occurs at ~400 meters depth with a slight decrease to 32-33 μmol kg^{-1} down past 1000 meters depth. (Sigman et al. 2000). The estimated data seems to follow the observed behavior described above closely, following the signals that were noted by previous observations. These nitrate estimates (Figure 8) will be analyzed in further detail to investigate mesoscale and submesoscale biogeochemical processes throughout the water column in this region. A similar procedure to the one discussed here for nitrate will be undertaken for pH in the future. Estimates of pH along the Seaglider tracks can be used to calculate the carbon flux at each point and then be analyzed to gain better understanding of mesoscale and submesoscale processes involving carbon flux.

Conclusion

The continuous application of novel machine learning and other data science techniques in oceanography is important for the future development of the field. New methods in data estimation and prediction enable research that would initially not be possible. The estimation of nitrate data using the random forest model proves that complex methods of machine learning can be applied to many other fields outside of theory. While the random forest model performed the best, some of the other models tested were nearly as accurate. This work utilized three other ways to estimate data that performed better than the historically used multiple linear regression. The use of any three of the improved models would provide better results in future research that needs to estimate biogeochemical variables. The availability of so many potential models ensures

that the growth of similar methods of data estimation can be continuously developed into the future.

Additionally, accurate data estimation allows us to get more out of every observation we take. Oceanography is an inherently expensive field to begin research in as the costs of sensors and deployment of sensors is expensive due to the use of research vessels. The design of sensors is also expensive due to the volatile nature of the ocean where more precautions need to be taken to successfully get the measurements needed for research. If more data can be estimated accurately from existing data from sensor networks such as SOCCOM floats, then that will free up resources to better optimize our observing networks. This helps make oceanography more accessible by potentially reducing the cost of research.

Expansion

Once a working model for estimating biogeochemical parameters in the Southern Ocean is created, it can be replicated for every other major oceanographic body of water tailoring to each region's specific qualities. Every ocean is different, so the model for the Southern Ocean may vary wildly from the model in the Pacific Ocean. Once biogeochemical data is able to be predicted in every major oceanographic body of water if given the latitude, longitude, depth, temperature, and salinity we would expand the dataset to train models to include more types of oceanographic instruments and types of data such as cabled observatories like the Regional Cabled Array off the coast of Washington and Station ALOHA to ORCA buoys. At the end of the day a machine learning model is only useful if the data it can predict is needed so such work would need to be guided by broader research questions of interest.

Acknowledgements

I would like to thank Alison Gray, Paige Lavin, and Dhruv Balwada for advising me on this project as well as serving as mentors in the physical oceanography field. I would also like to thank the 2021-2022 School of Oceanography Senior Thesis instructors for their time and for teaching me invaluable skills in scientific research and writing. A special thank you to the R/V Thompson crew and scientists who contribute to the collection of data across countless voyages. Finally, I would like to thank my fellow peers in the School of Oceanography and in the OCEAN 444/445 class for their help in peer reviews and providing feedback.

References

Ahmad, H. 2019. Machine learning applications in oceanography. Aquatic Research 2: 161-169.

Breiman, L. 2001. Random Forests. Machine Learning **45:** 5-32.

Chen, H., A. K. Morrison, C. O. Dufour, and J. L. Sarmiento. 2019. Deciphering Patterns and Drivers of Heat and Carbon Storage in the Southern Ocean. Geophysical Research Letters **46:** 3359-3367.

Chen, T. and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Friedlingstein, P., et al. 2020. Global Carbon Budget 2020. Earth System Science Data **12:** 3269-3340.

Giglio, D., V. Lyubchich, and M. R. Mazloff. 2018. Estimating Oxygen in the Southern Ocean Using Argo Temperature and Salinity. Journal of Geophysical Research: Oceans **123:** 4280-4297.

Hauck, J., C. Völker, T. Wang, M. Hoppema, M. Losch, and D. A. Wolf-Gladrow. 2013. Seasonally Different Carbon Flux Changes in the Southern Ocean in Response to the Southern Annular Mode. Global Biogeochemical Cycles **27:** 1236-1245.

Hsieh, W. W. 2009. Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels. Cambridge University Press.

James, S. C., Y. Zhang, and F. O'Donncha. 2018. A Machine Learning Framework to Forecast Wave Conditions. Coastal Engineering **137:** 1-10.

Russell, J. and others 2014. The Southern Ocean Carbon and Climate Observations and Modeling Program (SOCCOM). Ocean Carbon Biogeochem. Newsletter **7:** 1-5.

Sharp, J. D., Fassbender, A. J., Carter, B. R., Lavin, P. D., & Sutton, A. J. 2021. A monthly surface pCO 2 product for the California Current Large Marine Ecosystem. Earth System Science Data Discussions, 1-48.

Shepherd, J. G., P. G. Brewer, A. Oschlies, and A. J. Watson. 2017. Ocean Ventilation and Deoxygenation in a Warming World: Introduction and Overview. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **375:** 20170240.

Sigman, D. M., Altabet, M. A., McCorkle, D. C., Francois, R., & Fischer, G. 2000. The δ15N of nitrate in the Southern Ocean: Nitrogen cycling and circulation in the ocean interior. Journal of Geophysical Research: Oceans **105(C8)**:19599-19614.

Simmonds, J. E., F. Armstrong, and P. J. Copland. 1996. Species Identification Using Wideband Backscatter with Neural Network and Discriminant Analysis. ICES Journal of Marine Science **53:** 189-195.

Turner, R. E., W. W. Schroeder, and W. J. Wiseman. 1987. The Role of Stratification in the Deoxygenation of Mobile Bay and Adjacent Shelf Bottom Waters. Estuaries **10:** 13-19.

Zanna, L., S. Khatiwala, J. M. Gregory, J. Ison, and P. Heimbach. 2019. Global Reconstruction of Historical Ocean Heat Storage and Transport. Proceedings of the National Academy of Sciences **116:** 1126-1131.

Appendix A

Table 3. URLs to Python packages used in project.

Package Name	**Documentation URL**
Pandas	https://pandas.pydata.org
Xarray	https://docs.xarray.dev/en/stable/
Numpy	https://numpy.org
GSW	https://teos-10.github.io/GSW-Python/
scikit-learn	https://scikit-learn.org/stable/
pyGAM	https://pygam.readthedocs.io/en/latest/
XGBoost	https://xgboost.readthedocs.io/en/stable/index.html
SciPy	https://scipy.org
Matplotlib	https://matplotlib.org
Dataframe_image	https://pypi.org/project/dataframe-image/
Seaborn	https://seaborn.pydata.org