Kristina Confesor
04/25/2023
OEAS 805
Project Report: Preliminary Classification of *K. brevis* HAB conditions on the WFS

Abstract

K. brevis is a designated HAB forming species of dinoflagellates at the WFS due to their brevotoxin production, however, the causes and location for where blooms form is still unknown. The purpose of this capstone project is to create a preliminary classifier model that can predict bloom versus no bloom conditions in this region. Hydrogeographic and nutrient data from FWRI and CMEMS were combined for model training and testing. A supervised MLP Classifier model was created to predict bloom versus no bloom conditions on an April 2019 WFS cruise. No bloom conditions were found in the cruise data, however differences between model and in situ PO_4^{3-} data may have affected the prediction results despite the model's accuracy scores. Future studies will use different combinations of nutrients as well as add other parameters in order to better characterize the *K. brevis* blooms that occur in this region with the eventual goal of accurate forecasting.

Introduction

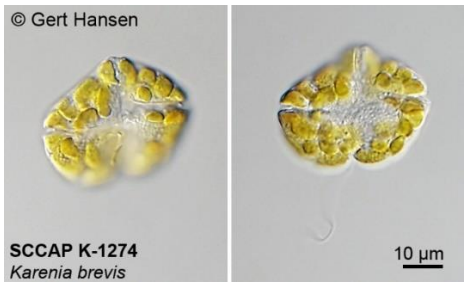

Figure 1: *K. brevis* morphology (Hansen et al., 2000).

Harmful algal blooms (HABs) have the capacity to negatively impact humans as well as surrounding ecosystems around the world (Ho & Michalak, 2015). Monitoring and predicting these HAB events can improve the ability to mitigate their effects. *Karenia brevis* (previously identified as *Gymnodinium breve* and *Ptychodiscus brevis* (Lin et al., 1981) is an example of a HAB forming toxic dinoflagellate that contributes to red tide bloom events on the West Florida Shelf (WFS) (Figure 1). The WFS itself is composed of mixed carbonate and siliciclastic sediments at nearshore isobaths (Harrison et al., 2003) that transition to being carbonate-dominated towards the southmost part (Hine et al., 2008). Shellfish harvested and consumed in *K. brevis* infested waters can result in potential oceanic and human mortalities due to the brevotoxin *K. brevis* produces (Kirkpatrick et al., 2004). Documented cases of brevotoxins aerosolized and ingested via breathing (Fleming et al., 2005) have also occurred from the wave actions of *K. brevis* HABs around coastal areas. The lethal toxicity of the produced brevotoxins as well as the documented prolific growth of *K. brevis* results in the species as a designated prominent HAB at the WFS.

Creating models to predict or classify *K. brevis* HAB events is crucial in understanding what causes them, where they are more likely to occur, and the severity of which these blooms occur at. Currently, there is no single hypothesis explaining the occurrences of *K. brevis* HABs along the WFS. It is known that nutrients play a significant role during bloom formation, where NO_3^- is utilized at initiation phases of a bloom, while NH_4^+ is consumed during maintenance and stationary phases (Bronk et al., 2014). Other nutrient impacts remain to be elucidated. This capstone project utilizes supervised machine learning to classify potential *K. brevis* bloom occurrences on the WFS in a April 2019 cruise (Confesor et al., 2022) based on an assortment of nutrients (PO_4^{3-}, NO_3^-, and Si). Cell counts will be used as a measure of bloom formation, with the target variable specifically as *K. brevis* cells counted per liter. A classifier model is necessary as *K. brevis* HAB events have changed over the past decade, and as stated before, we do not have a clear hypothesis as to what causes these blooms to occur yet.

Data

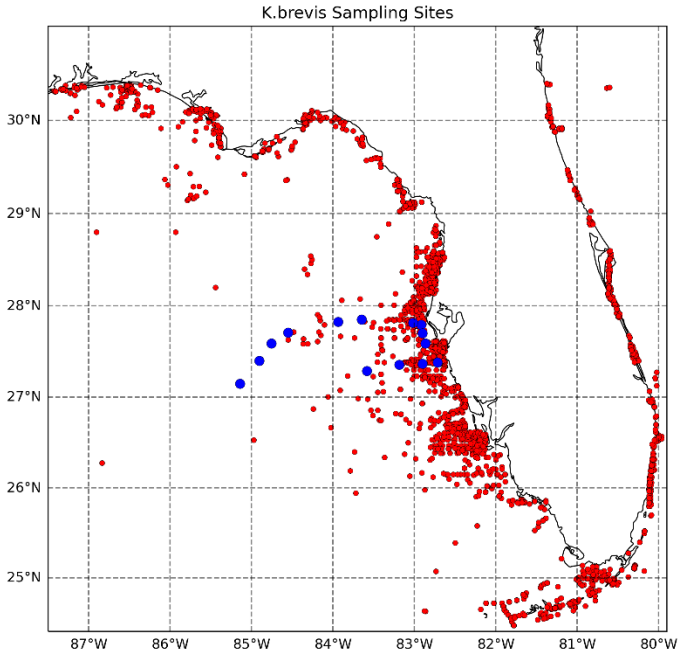

Figure 2. WFS Sampling sites from 2015-2020. Red dots indicate FWRI and CMEMS combined data used in the MLP model construction, while blue dots indicate in situ samples taken from the April 2019 WFS cruise.

The GitHub repository for this capstone project is available to the public here (Confesor, 2023) with full data access descriptions, notebooks, and package requirements. The target variable consists of *K. brevis* cell counts/L taken by the Fish and Wildlife Research Institute (FWRI) over 2015-2023 around the coasts of Florida. This dataset contains datetime, latitude, longitude, temperature, salinity, and cell count data for each sample taken. Currently, it has over 60,000 samples that span from 2015 through 2020 with some salinity and temperature data missing, as well as no nutrient data available to the public. Instead, other salinity, temperature, PO_4^{3-}, NO_3^-, and Si variables were accessed from the GLORYS12V1 (0.083° × 0.083° spatial resolution) as well as the Global Ocean Biogeochemistry Hindcast (0.25° × 0.25° spatial resolution) datasets (Lien et al., 2021) to replace missing values in the FWRI dataset. These Copernicus Marine Environment Monitoring Service (CMEMS) datasets have been taken since 1993 and cover the time and geographic range that the *K. brevis* cell counts were taken up until 2020 (Lien et al., 2021). The datasets are in the form of a NetCDF (.nc) file, where variables will have to be extracted on the basis of the FWRI parameters. The Multilayer Perceptron (MLP) Classifier from the *Scikit-learn* package (Pedregosa et al., 2011) was used for supervised model training and testing, where *K. brevis* bloom conditions were predicted in a 2019 WFS cruise (Confesor et al., 2022) (Figure 2).

Methods

The capstone project was divided into three distinct steps: data cleaning and merging, exploratory data analysis, and model construction (Figure 3). Each step has a separate Jupyter Notebook that can be accessed via the GitHub repository (Confesor, 2023).

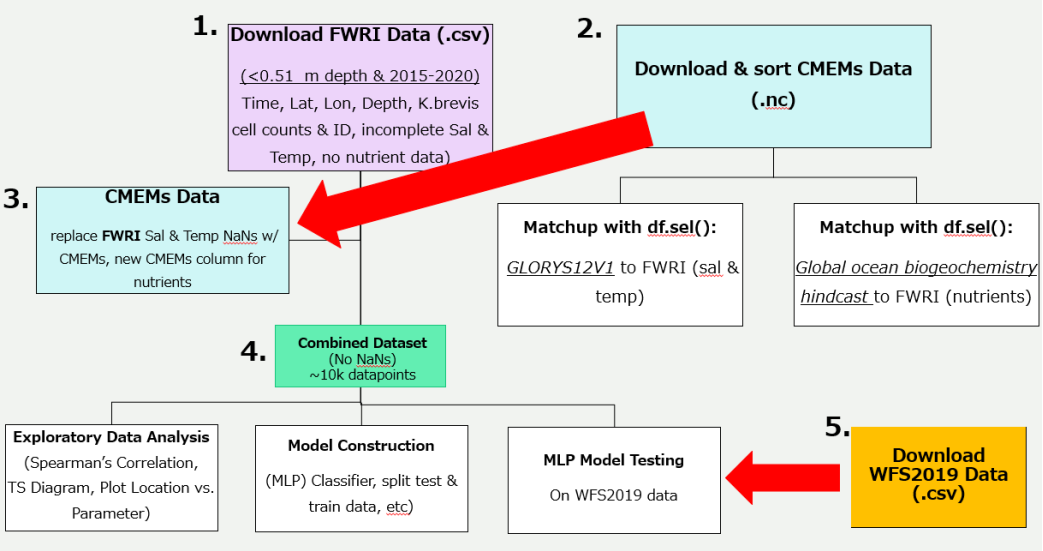

Figure 3. Capstone Project Pipeline. Data was collected from databases, cleaned, and merged, then exploratory data analysis occurred prior to model construction and testing.

Data Cleaning and Merging:

Datasets were first obtained from FWRI, CMEMS, and a 2019 WFS cruise. FWRI data was indexed to include only data points above 0.51 m depth between 2015-2020, and then designated an ID of "Bloom" or "No Bloom". A bloom for the purposes of this project was defined as cell counts equal to or above 50 per Liter ($\geq$ 50 cells/L). The resulting FWRI time, latitude, longitude, and depth parameters were matched up with CMEMs data. Temperature and salinity CMEMS data replaced any missing FWRI data, while new columns containing PO_4^{3-}, NO_3^-, and Si CMEMS variables were created. The resulting 10,000 datapoints were merged into a new .csv file and used for exploratory data analysis and model construction (the functions for these steps can be found in the Confesor_Functions.py file in the GitHub repository (Confesor, 2023)).

Exploratory Data Analysis:

	***K. brevis* (cells/L)**	**Depth (m)**	**Temperature (C)**	**Salinity**	**NO_3^- (um)**	**PO_4^{3-} (um)**	**Si (um)**
std.dev	1.606814e+06	0.141290	5.028454	6.053556	1.644527	0.003074	4.926889
mean	1.299599e+05	0.405495	24.817921	31.900248	1.248983	0.002111	10.179175
median	0.000000e+00	0.500000	25.600000	34.363842	0.762936	0.001221	9.892280
max	7.729556e+07	0.500000	39.100000	47.290000	11.818180	0.074222	24.072092
min	0.000000e+00	0.100000	5.700000	0.150000	0.004355	0.000027	2.476244

Table 1. Standard deviation, mean, median, maximum, and minimum values of parameters.

Basic statistical analysis was first done on the merged and cleaned dataset (Table 1). All parameters had some variance except for PO_4^{3-}, where values were around 0.00211 um and had a small standard deviation (0.00307). *K. brevis* cell counts had the largest standard deviation (1.06×10^6), where counts ranged from 0- 7.73×10^7 cells/L.

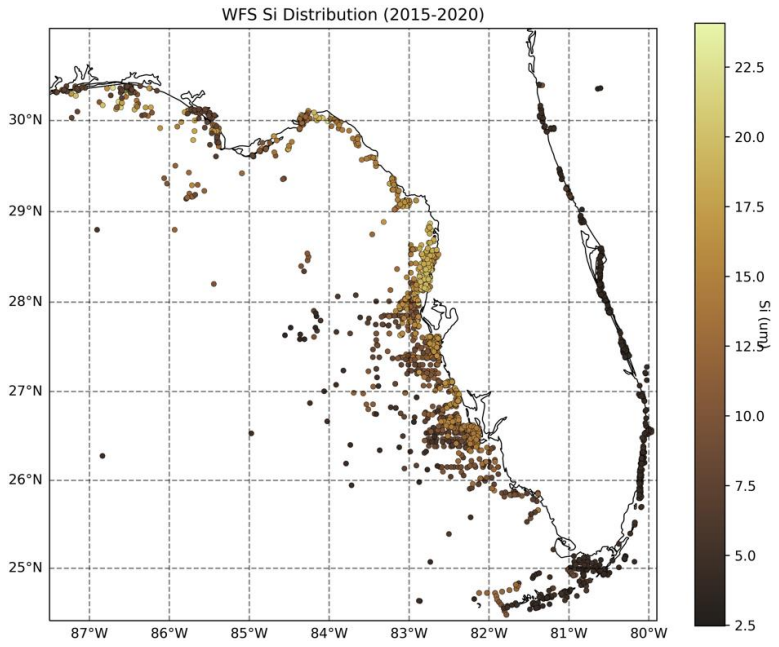

Figure 4. Distribution of merged and cleaned Si datapoints along the WFS.

Parameters were plotted via sample location as well. The most significant qualitative trend was found in the Si plot (Figure 4). Si was most abundant inshore and north of the WFS, and gradually declined southward. This follows prior literature where values are most abundant at northern and nearshore isobaths of the WFS (Harrison et al., 2003)

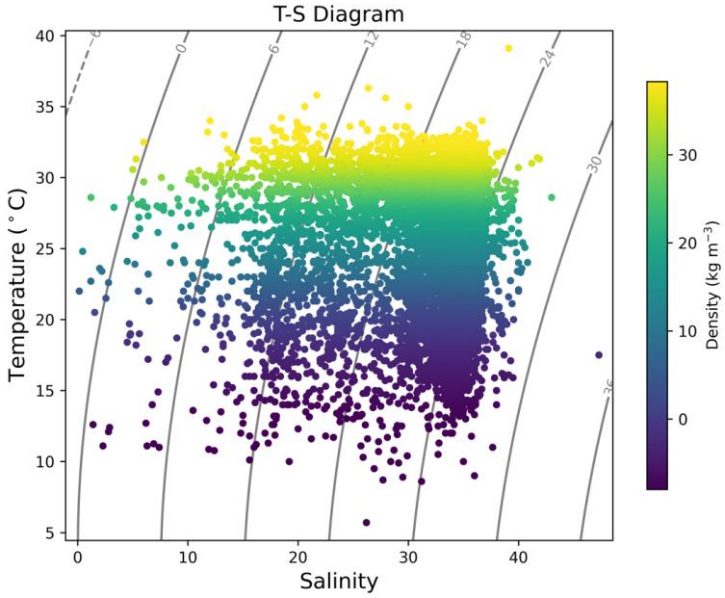

Figure 5. T-S Diagram of cleaned and merged datapoints.

A temperature-salinity (T-S) diagram was also created (Figure 5). Density was lower at colder temperatures (close to 0 kg*m^{-3}), while higher at hotter temperatures (above 30 kg*m^{-3}). Salinity did not have as much of an impact on density, where different values could be found across different salinity values, indicating that temperature has a more significant impact on density values on this particular dataset.

Test	Rho	P-value
K. brevis vs Salinity	0.067445	1.262214e-11
K. brevis vs Temperature	-0.061056	8.786673e-10
K. brevis vs NO_3^-	0.128968	1.381598e-38
K. brevis vs PO_4^{3-}	0.021748	2.913160e-02
K. brevis vs Si	0.167928	1.492788e-64

Table 2. Spearman's correlations between parameters of interest.

Lastly, nonparametric testing was done on the parameters to test for potential correlations and significance (Table 2). Spearman's correlations were done to look at the relationship between *K. brevis* cell counts and both the hydrogeographic and nutrient data. All p-values were significant, indicating potential significant relationships and trends that could be used to classify *K. brevis* HAB conditions in this region.

Multilayer Perceptron (MLP) Classifier:

	Precision	Recall	f1-score	support
0	0.89	0.94	0.91	1608
1	0.68	0.53	0.59	405
Accuracy			0.86	2013
Macro Avg.	0.79	0.73	0.75	2013
Weighted Avg.	0.85	0.86	0.85	2013

Table 3. MLP Classification Report (hidden_layer_sizes=(300,300,300), random_state=1, max_iter=5000), where 0 = No Bloom conditions, while 1= Bloom conditions.

Supervised machine learning was used during model creation. The dataset was split into training and testing sets, and then the MLP Classifier itself was looped through layer and neuron combinations to find the combination with the best accuracy & f1-scores. The combination that yielded the best accuracy was 3 layers of 300 neurons at 5000 iterations with a random state of 1 (Table 3).The model then was used to predict *K. brevis bloom* conditions on WFS2019 cruise data, and also cross-validated with the Kfold python package (Pedregosa et al., 2011).

Discussion

Out[8]:

	0	1
0	1.0	0.000000e+00
1	1.0	0.000000e+00
2	1.0	2.748071e-192
3	1.0	6.685027e-258
4	1.0	3.722210e-290
5	1.0	1.878796e-201
6	1.0	4.134108e-246
7	1.0	2.229301e-174
8	1.0	5.310508e-236
9	1.0	0.000000e+00
10	1.0	2.675700e-255
11	1.0	0.000000e+00
12	1.0	0.000000e+00
13	1.0	0.000000e+00

Figure 7. Prediction results for WFS 2019 data. The 0 indicates No Bloom conditions, while 1 indicates Bloom conditions.

Resulting classification predictions on the April WFS 2019 data indicate that there were no bloom conditions classified on any of the samples collected (Figure 7). However, it is important to note that the

PO_4^{3-} values collected in situ are much higher than the values used for the model by 3 orders of magnitude (Confesor et al., 2022), indicating that this particular nutrient may not be a good predictor for bloom conditions. Even though there are significant correlations between cell counts and each of the parameters (Table 2), some parameters do a better job of predicting bloom versus no bloom conditions.

From the MLP classification report (Figure 6), the precision is higher in the no bloom classifications compared to the bloom classifications, indicating that this particular model is more accurate in predicting when there are no bloom conditions, but not as much for bloom conditions. The differing PO_4^{3-} values between model and predictor data may account for discrepancies as a result. Removing this nutrient from future models may improve the accuracy scores.

The root mean squared error (RMSE) values for 5 splits from the Kfold cross-validation were all around 0.40 (Confesor, 2023), indicating that the model is still accurate in predicting bloom versus no bloom conditions. A RMSE closer to 0 indicates better accuracy. Other parameter combinations need to be tested in order to elucidate which model best classifies bloom versus no bloom conditions.

Conclusions

Much work still remains to predict *K. brevis* HABs at the WFS. This capstone project is a step in accurate forecasting of *K. brevis* bloom events. A combination of nutrients and hydrogeographic data are good parameters to include in a HAB classifier model, where some nutrients are better predictors of HAB conditions compared to others. Future work to add to this specific model include adding PCA analysis to the EDA pipeline, determining the accuracy and/or standard deviation of CMEMS matchups to FWRI data, creating different models with different parameters (removing PO_4^{3-} and adding *Trichodesmium* related variables such as N_2-fixation rates), constraining geographical boundaries to only the WFS area and not the entire Floridian peninsula, more data to predict *K. brevis* HAB conditions on, and adding seasonal categories to the model as *K. brevis* only occurs during certain months of the year. This model is still in its preliminary stages but has potential for a broader impact in the WFS area and in turn protecting the wildlife in this region.

References

Bronk, D. A., Killberg-Thoreson, L., Sipler, R. E., Mulholland, M. R., Roberts, Q. N., Bernhardt, P. W., Garrett, M., O'Neil, J. M., & Heil, C. A. (2014). Nitrogen uptake and regeneration (ammonium regeneration, nitrification and photoproduction) in waters of the West Florida Shelf prone to blooms of Karenia brevis. *Harmful Algae*, *38*, 50-62. https://doi.org/https://doi.org/10.1016/j.hal.2014.04.007

Confesor, K. A. (2023). *Preliminary Classification of Karenia Brevis HAB conditions on the West Florida Shelf (WFS)* https://doi.org/https://github.com/kconf001/MLP_Classifier_KbrevisHABs

Confesor, K. A., Selden, C. R., Powell, K. E., Donahue, L. A., Mellett, T., Caprara, S., Knapp, A. N., Buck, K. N., & Chappell, P. D. (2022). Defining the Realized Niche of the Two Major Clades of Trichodesmium: A Study on the West Florida Shelf [Original Research]. *Frontiers in Marine Science*, *9*. https://doi.org/10.3389/fmars.2022.821655

Fleming, E., Backer, C., & Baden, G. (2005). Overview of Aerosolized Florida Red Tide Toxins: Exposures and Effects. *Environmental Health Perspectives*, *113*(5), 618-620. https://doi.org/10.1289/ehp.7501

Hansen, N., Larsen, J., & Moestrup, Ø. (2000). Phylogeny of some of the major genera of dinoflagellates based on ultrastructure and partial LSU rDNA sequence data, including the erection of three new genera of unarmoured dinoflagellates. *Phycologia*, *39*, 302-317.

Harrison, S. E., Locker, S. D., Hine, A. C., Edwards, J. H., Naar, D. F., Twichell, D. C., & Mallinson, D. J. (2003). Sediment-starved sand ridges on a mixed carbonate/siliciclastic inner shelf off west-central Florida. *Marine Geology*, *200*(1), 171-194. https://doi.org/https://doi.org/10.1016/S0025-3227(03)00182-8

Hine, A. C., Halley, R. B., Locker, S. D., Jarrett, B. D., Jaap, W. C., Mallinson, D. J., Ciembronowicz, K. T., Ogden, N. B., Donahue, B. T., & Naar, D. F. (2008). Coral Reefs, Present and Past, on the West Florida Shelf and Platform Margin. In B. M. Riegl & R. E. Dodge (Eds.), *Coral Reefs of the USA* (pp. 127-173). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6847-8_4

Ho, J. C., & Michalak, A. M. (2015). Challenges in tracking harmful algal blooms: A synthesis of evidence from Lake Erie. *Journal of Great Lakes Research*, *41*(2), 317-325. https://doi.org/https://doi.org/10.1016/j.jglr.2015.01.001

Kirkpatrick, B., Fleming, L. E., Squicciarini, D., Backer, L. C., Clark, R., Abraham, W., Benson, J., Cheng, Y. S., Johnson, D., Pierce, R., Zaias, J., Bossart, G. D., & Baden, D. G. (2004). Literature Review of Florida Red Tide: Implications for Human Health Effects. *Harmful Algae*, *3*(2), 99-115. https://doi.org/10.1016/j.hal.2003.08.005

Lien, V. S., Nilsen, J. E., Perivoliotis, L., Sotiropoulou, M., Denaxa, D., & Ehrhart, S. (2021). BioGeoChemical product provided by the Copernicus Marine Service. *EGU General Assembly* https://doi.org/https://doi.org/10.5194/egusphere-egu21-5625

Lin, Y.-Y., Risk, M., Ray, S. M., Van Engen, D., Clardy, J., Golik, J., James, J. C., & Nakanishi, K. (1981). Isolation and structure of brevetoxin B from the "red tide" dinoflagellate Ptychodiscus brevis (Gymnodinium breve). *Journal of the American Chemical Society*, *103*(22), 6773-6775. https://doi.org/10.1021/ja00412a053

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, *12*(null), 2825–2830.