The Analysis and Prediction of Harmful Algal Blooms (K.brevis) through Machine Learning Modules

Injun Cho[1], Shahin Alam[1], Ming Ye[1]

1. Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, 32304

Introduction

- Harmful Algal blooms (HAB), such as the red tide occurrences, cause ecological and societal harm throughout Florida.
- Machine learning (ML) is utilized to predict potential blooms through real-time abiotic data.
- Not much analysis on the Sarasota region, constantly affected by the phenomena.
- Using Elshall at the Florida Gulf Coast University's machine learning model, different abiotic data will be collected to understand and predict red tide occurrences in the area.
- The machine learning model, primarily designed for the fort Charlotte region, will be evaluated on how applicable it is to other watersheds.
- Data from Manatee river of Sarasota will be used to evaluate the ML module's effectiveness.

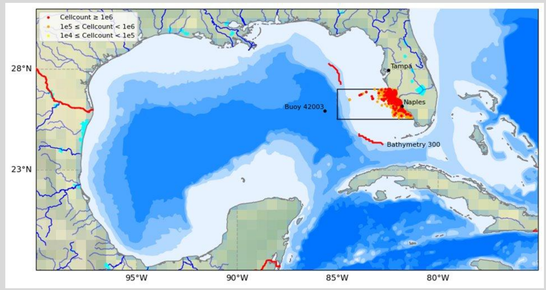


Figure 1: Spatial distribution of red tide intensity showing nearshore amplification and the influence of shelf dynamics and local environmental drivers.

Study Area & Data

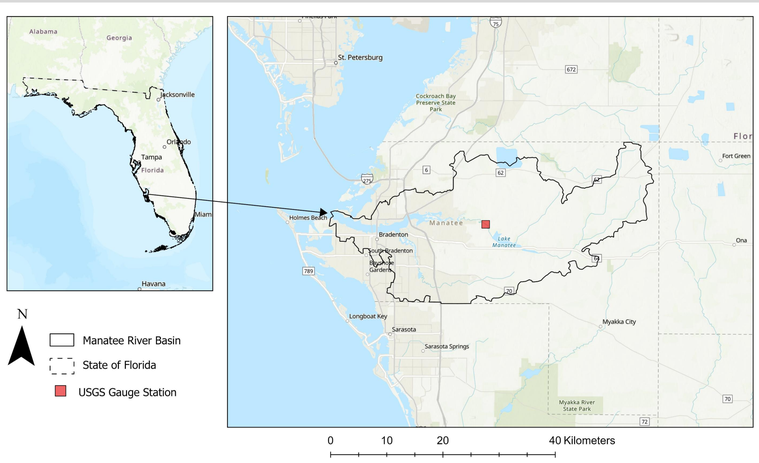


Figure 2. *Location of the Manatee River Basin in Southwest Florida, showing watershed boundaries and the USGS*

The Manatee river basin, forming on the northern border of Sarasota, flows west towards a semi-closed estuarine system on Sarasota Bay. Draining over 900 km^2 across the majority of Manatee county, the river serves as an important source of phosphorous and nitrogen discharge into the ocean from the surrounding area. Especially, seasonal rainfall influences such nutrient discharge and nutrient loading.

Data have been collected from the following sources:

- Time Range: available data from 1994 to 2024
- Wind Data: National buoy data center (ID: 42013)
- HAB Data: NCEI HAB dataset
- ZOS Data: Climate Data Store Sea level dataset
- Manatee river discharge: USGS Water Atlas data

Methods

Procedures

- Data was collected on the desired study region
- ML framework modified to fit the data used for Sarasota and the Manatee river basin
- Cleaned and formatted datasets to match Elshall's ML framework
- Trained model using 1994–2023 historical data
- Generated predictions using data starting from 2023

Data Analysis

- Compared predicted vs. observed 2025 red tide events
- Calculated prediction accuracy (%)
- Evaluated classification performance metrics

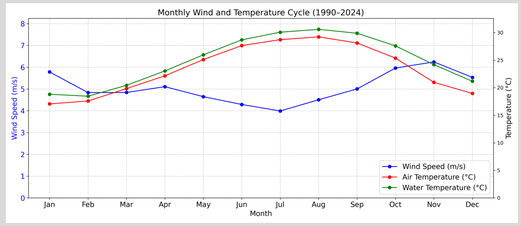


Figure 3. *Compiled wind & temperature data of the Manatee river & Sarasota region through Python*

Machine Learning Models

Objective

Predicting the K. brevis occurrences one week ahead using environmental and physical predictors. The classification framework distinguishes the bloom conditions based on weekly aggregated data.

Classifier

A random forest classifier was utilized for this model. It follows a decision tree based on aggregated data such as sea surface height, nutrient input, temperature, and wind speed.

Evaluation

Model performance was quantified using a simple balanced accuracy of success rates. Datasets will be split into training and testing for this purpose.

Results and Discussion

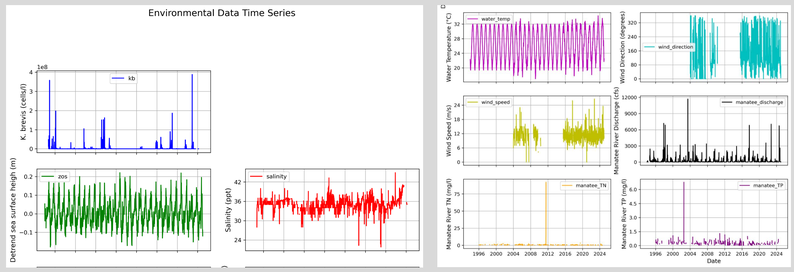
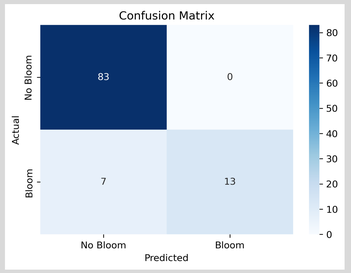


Figure 4. *Compiled data of all the environmental factors relevant to the decision making during 1994-2024.*

Results and Discussion

- Environmental datasets differed between the Sarasota coastal region and the Manatee River basin in both temporal coverage and variable availability.
- For example, wind data were limited in duration and derived from an offshore buoy (NDBC 42013),
- The machine learning model achieved an accuracy of 85% when applied to the Manatee River basin dataset (compared to the 90% accuracy of the original model).
- Given the geographic proximity and broadly similar conditions between the two regions, this result suggests that the model retains predictive capability when applied to nearby environments.



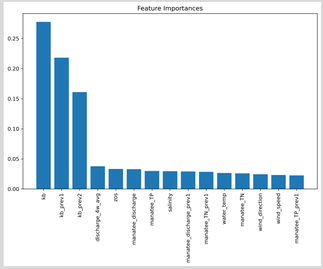


Figure 5,6. *Visual representation of confusion matrix (left) and the feature importance (right)*

Conclusion and Future Work

- Results demonstrated predictive capability for harmful algal bloom occurrence in the Sarasota region.
- Given the close geographic proximity and similar coastal-oceanographic conditions of the Manatee River basin, model transferability is plausible; however, regional differences in data availability and spatial resolution introduce uncertainty.
- Future studies should be conducted to apply different datasets from different environments.

- Additionally, future work should incorporate higher-resolution coastal wind and circulation datasets to better represent nearshore transport processes.
- Current model does not account for aerosolized brevetoxin dispersion, which drives respiratory health impacts.

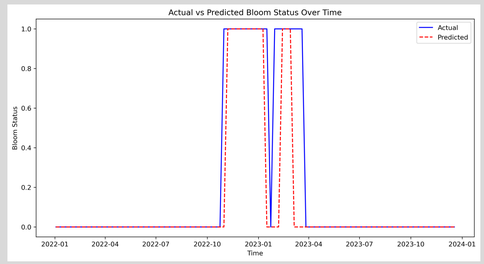


Figure 7. Comparison of observed and machine-learning-predicted harmful algal bloom (HAB) status in the Sarasota region from 2022–2023. The model successfully captures the timing of bloom events with minor deviations in onset and duration.

Acknowledgements and References

I would like to express my sincere gratitude to everyone who contributed to and supported this research. In particular, I thank Dr. Ye and Mr. Shahin Alam, for their guidance and assistance throughout the project. I also extend my appreciation to the members of the UROP team and the UROP program leaders for their support and mentorship during the research process.