Machine learning framework for red tide bloom severity classification in Charlotte Harbor, West Florida Shelf

Final Report for Task 9 of Water Quality Evaluation for Peace River Basin and Greater Charlotte Harbor Watershed in Southwest Florida, Florida Department of Environmental Protection (August 2023 – November 2025)

Author: Ahmed S. Elshall (aelshall@fgcu.edu)

Abstract: This study developed a machine learning framework to predict Karenia brevis red tide bloom severity along the southwest coast of Florida, specifically in Charlotte Harbor. We employed a Random Forest Classifier using weekly environmental data (1992–2024), including river discharge, nutrient loading (TN and TP), wind speed and direction, water temperature, salinity, and sea surface height anomalies. The model achieved high predictive performance, demonstrating a balanced accuracy of 0.887, precision of 0.90, and recall of 0.81 for bloom events. Karenia brevis cell counts, and their lagged values were the most significant predictors, followed by river discharge and nutrient inputs, confirming their critical role in bloom dynamics. The model exhibited slight overfitting, indicated by high training accuracy compared to validation accuracy, suggesting opportunities for further model optimization. Partial dependence and feature importance analyses highlighted nutrient loading and river discharge as primary drivers influencing bloom probability, consistent with ecological expectations. Operational forecasts with a 1-week horizon are recommended for practical management, with weekly updates and periodic retraining to maintain predictive reliability.

Final Report

1. Introduction

1.1 Overview: The overarching objective of this task is to develop a machine learning-based tool for predicting red tides in the Gulf of Mexico along the West Florida Shelf. The tool can be used to assess the role of river discharge and nitrogen and phosphorus loading on the probabilities of the maintenance and termination stages of red tides in the upper estuary and area of estuarine flux and serve as an early warning system. This report presents a machine learning framework to classify the severity levels of red tides, which are marine harmful algal blooms (HABs) caused by Karenia brevis along the southwest coast of Florida. We developed predictive models using key environmental drivers, including Peace River discharge, and nutrient concentrations including total nitrogen (TN) and total phosphorus (TP), wind speed and direction, water temperature and salinity, sea surface height (zos).

1.2 Deliverables: Co-PI Elshall used one-month summer salary for several tasks. First, co-PI Elshall attended several planning and progress meetings. Second, co-PI Elshall joined and lead two postdoc search committees, respectively. Thie first search committee was successful, and we hired the postdoc for watershed modeling. In the second search committee after interviewing 11 candidates, we failed to hire a postdoc. Three candidates accepted our offer, but they ended up not joining FGCU due to country of concern regulation, personal reasons, and visa issues, respectively. The budget for this task was reduced to one-month-summer salary, and accordingly the original scope was reduced from developing an early warning system for red tides, to providing preliminary results and proof-of concept as presented in this report. Additionally, this report discusses model reliability and provides suggestions for improving the workflow to enable this tool to be integrated into the dashboard and continue to run in real-time for monitoring environmental conditions in Charlotte Harbor that could potentially worsen red tide conditions. The main deliverables are six Jupyter notebooks covering river data (Peace River discharge, TN, and TP), atmospheric data (wind speed and direction), ocean data (Karenia brevis cell count, water temperature and salinity, and sea-surface height for ocean current detection), and machine learning framework to classify the bloom severity levels. These notebooks include Python code with data, metadata and figures. A second deliverable is this report documenting the method and describing the model limitations and suggestions for improvement. These deliverables including data, notebooks and report are published online as a Jupyter-book: https://aselshall.github.io/redtides

2. River data

2.1 Discharge

The Peace River significantly contributes freshwater, nitrogen (TN), and phosphorus (TP) loads to Charlotte Harbor, influencing coastal waters along southwest Florida where red tides frequently occur. Figure 1 shows key stations in the Peace River watershed near Charlotte Harbor. We utilized hydrological from the U.S. Geological Survey (USGS) monitoring station 02296750 near

Arcadia, which is strategically located near the river's mouth, to characterize freshwater inflow and nutrient dynamics relevant to Karenia brevis bloom events. Discharge from this station directly influences the nutrient budget of Charlotte Harbor, which in turn affects red tide development along the Gulf Coast. High discharge events at this station are closely linked to increased nutrient input (particularly nitrogen and phosphorus) into the coastal ecosystem, which can fuel Karenia brevis blooms. The station's long historical record and direct connection to coastal inflow make it critical for red tide prediction and monitoring. In a future study, other relevant stations could improve model performance or provide additional context for understanding nutrient dynamics. For example, PEACE RIVER RESERVOIR, which can influence downstream flow and nutrient concentrations through water retention and controlled releases. The station at PEACE RIVER AT NOCATEE FL is located downstream from the Arcadia station, closer to Charlotte Harbor. It can provide insights into how nutrient loading and discharge from Arcadia propagate downstream, influencing nutrient delivery into the estuary. In this study, the primary focus is Arcadia station, while the other stations can serve as supplementary sources for improving model resolution or understanding secondary drivers of bloom dynamics.

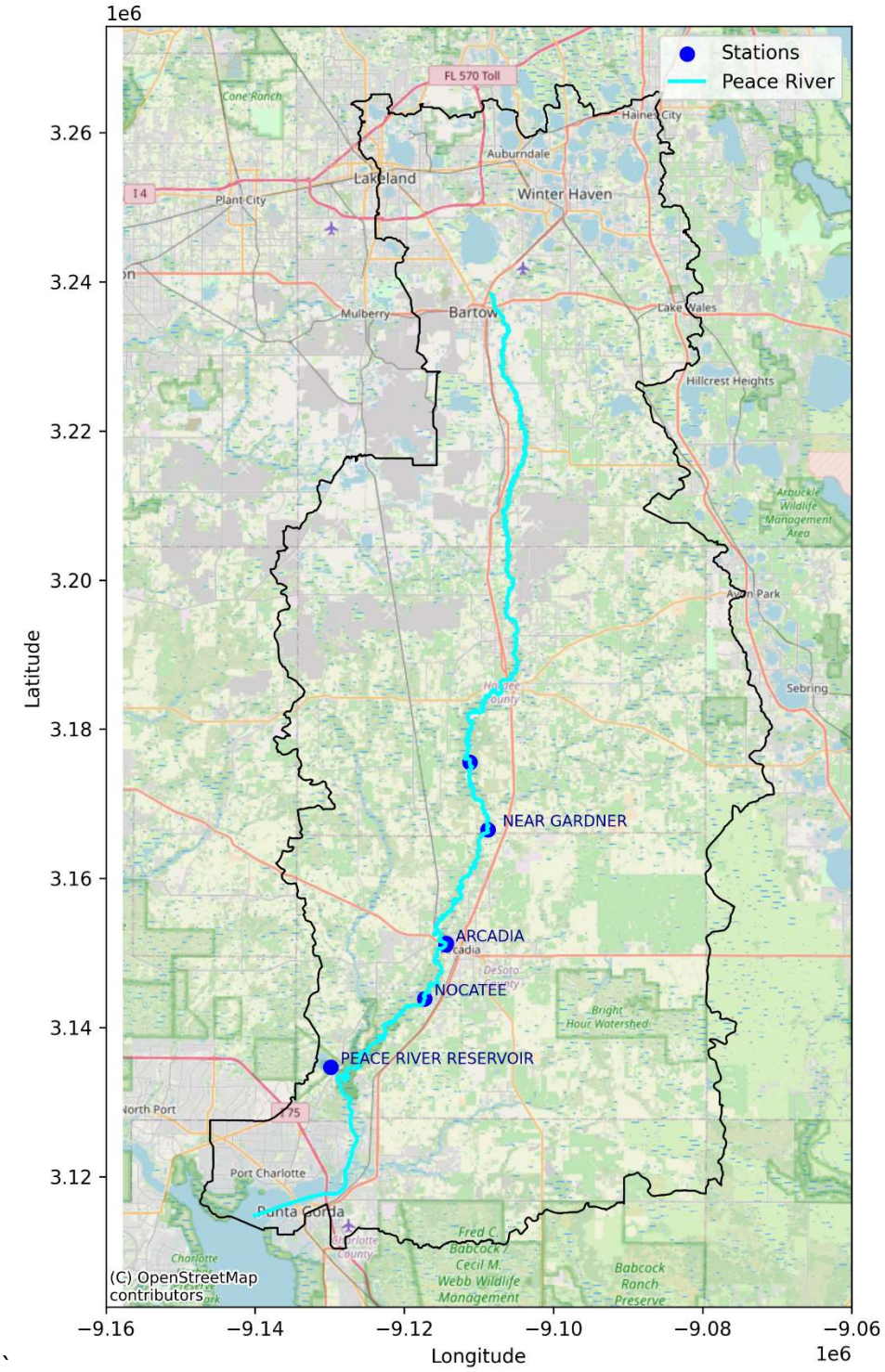

Figure 1 Peace River watershed showing key flow monitoring stations near Charlotte Harbor. The Peace River at Arcadia station (USGS 02296750) is the primary control point for freshwater discharge and nutrient loading into Charlotte Harbor. Additional stations at Nocatee, Peace River Reservoir, and Charlie Creek contribute to the broader understanding of hydrological and nutrient dynamics influencing red tide events.

The discharge data from USGS station 02296750 represents the daily streamflow measurements for the Peace River at Arcadia, Florida in cubic feet per second (cfs). Historical discharge records from this station extend back to 1912, providing extensive temporal coverage necessary for assessing long-term trends. As shown in Figure 2, discharge at Arcadia exhibits higher variability and larger peaks compared to Nocatee, indicating that Arcadia serves as a major inflow point for freshwater and nutrient loading into the Peace River system. This suggests that Arcadia's discharge dynamics are more influential in driving nutrient transport into Charlotte Harbor and potentially fueling red tide events. The collected data spans from 1987 to the present, capturing seasonal variations, storm events, and long-term hydrological trends. The station is part of the USGS National Water Information System (NWIS), which follows standardized methods for data collection and processing. However, the data may contain uncertainties due to factors such as instrument calibration errors, changes in channel geometry, flow estimation methods, and environmental factors like debris or vegetation influencing water flow. The data is marked with an approval code ('A'), indicating that it has been quality-checked and meets USGS data accuracy standards. However, for future study we need to account for measurement errors by learning about the error and distribution and missing data through using a physics-based model.

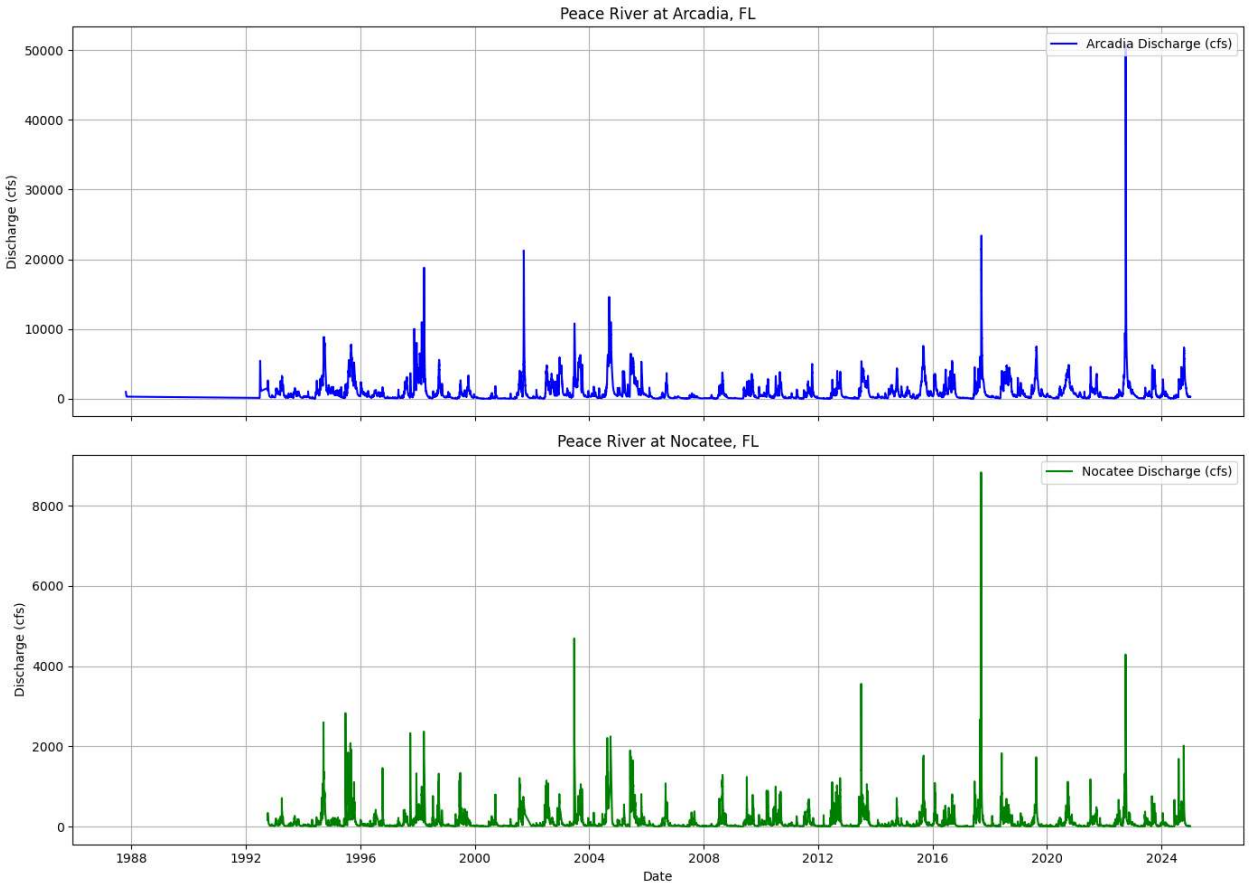

Figure 2. Discharge measurements (in cubic feet per second) at Peace River stations in Arcadia - USGS 02296750 (top) and Nocatee - USGS 2297100 (bottom) from 1987 to the present. Data reflect seasonal and episodic variations, with significant discharge peaks associated with storm events and wet season inflows.

2.1 Nutrients

TN and TP data were specifically sourced from the Water Atlas, integrating data from agencies 24 such as the Southwest Florida Water Management District (SWFWMD), Florida Department of Environmental Protection (FDEP), and local environmental organizations. Stations were selected based on proximity to the river's mouth and the tidal segments to ensure relevance to coastal nutrient dynamics. Selected segments included:

- Peace River (CHNEP) falling under the jurisdiction of the Coastal & Heartland National Estuary Partnership (CHNEP) and encompasses areas closer to the river's mouth near the Gulf
- Peace River Estuary (Upper Segment North, CHNEP) including the upper portion of the estuary where the Peace River transitions into tidal waters as it approaches the Gulf
- Tidal Peace River (CHNEP) covering the tidal sections of the Peace River within the CHNEP's monitoring area, reflecting the river's interaction with Gulf tides

By selecting these segments, we obtain TN and TP data relevant to the lower reaches of the Peace River near its confluence with the Gulf of Mexico as shown in Figure 3.

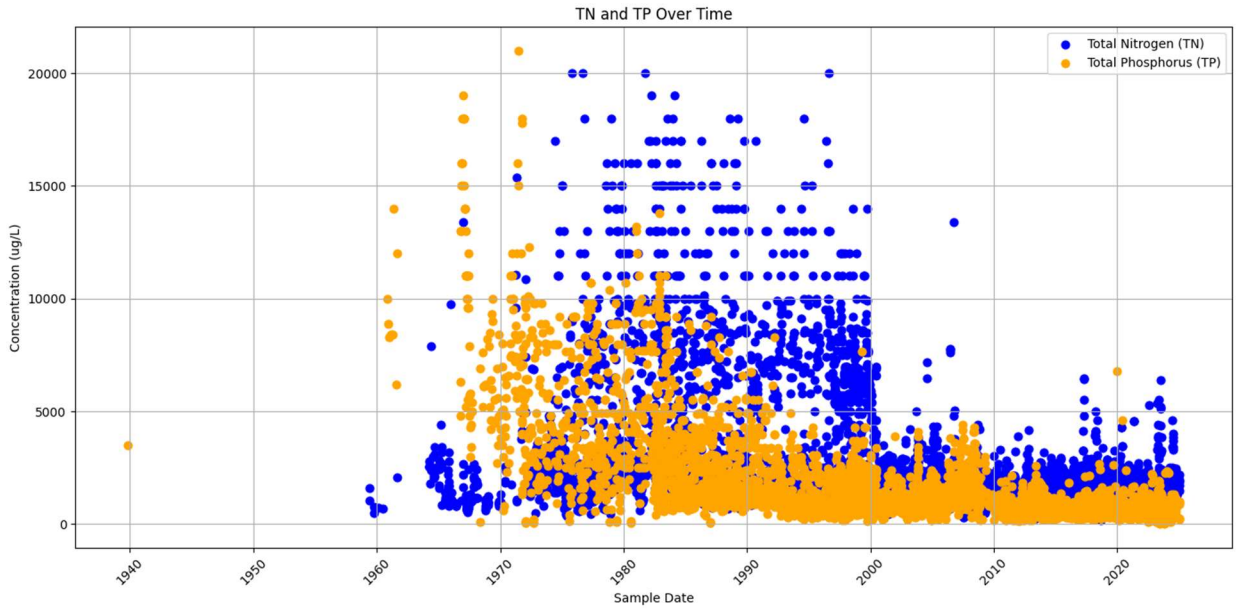

Figure 3. TN and TP data from Selected segments that are the Peace River (CHNEP), Peace River Estuary (Upper Segment North) (CHNEP), and Tidal Peace River (CHNEP) to provide coverage of freshwater-estuarine interactions

Given the identified 24 agencies and 377 stations, Table 1 shows stations with ≥30 TN and TP samples recorded between March 21, 1990, and November 5, 2024, at the lower reach showing a total of 1989 TN samples and 1397 samples. These are stations at the lower reach as shown in Figure 4.

Table 1. Stations are lower reach of Peace River with ≥30 TN and TP samples recorded between March 21, 1990, and November 5, 2024.

Data Source	Station Name	Station ID	# TN Samples	# TP Samples
USGS NWIS	PEACE RIVER AT ARCADIA FL	2296750	155	147
USGS NWIS	PEACE RIVER AT FT. OGDEN FL	2297330	161	156
USGS NWIS	PEACE RIVER AT RIVER MILE 14.82 NEAR FT. OGDEN FL	270318081593100	163	170
WIN 21FLGW	PEACE RIVER	3556	291	291
STORET FLPRMRWS	Peace River south of Desoto Marina	PR14	496	181
STORET FLPRMRWS	Peace River N of SR761	PR18	723	452

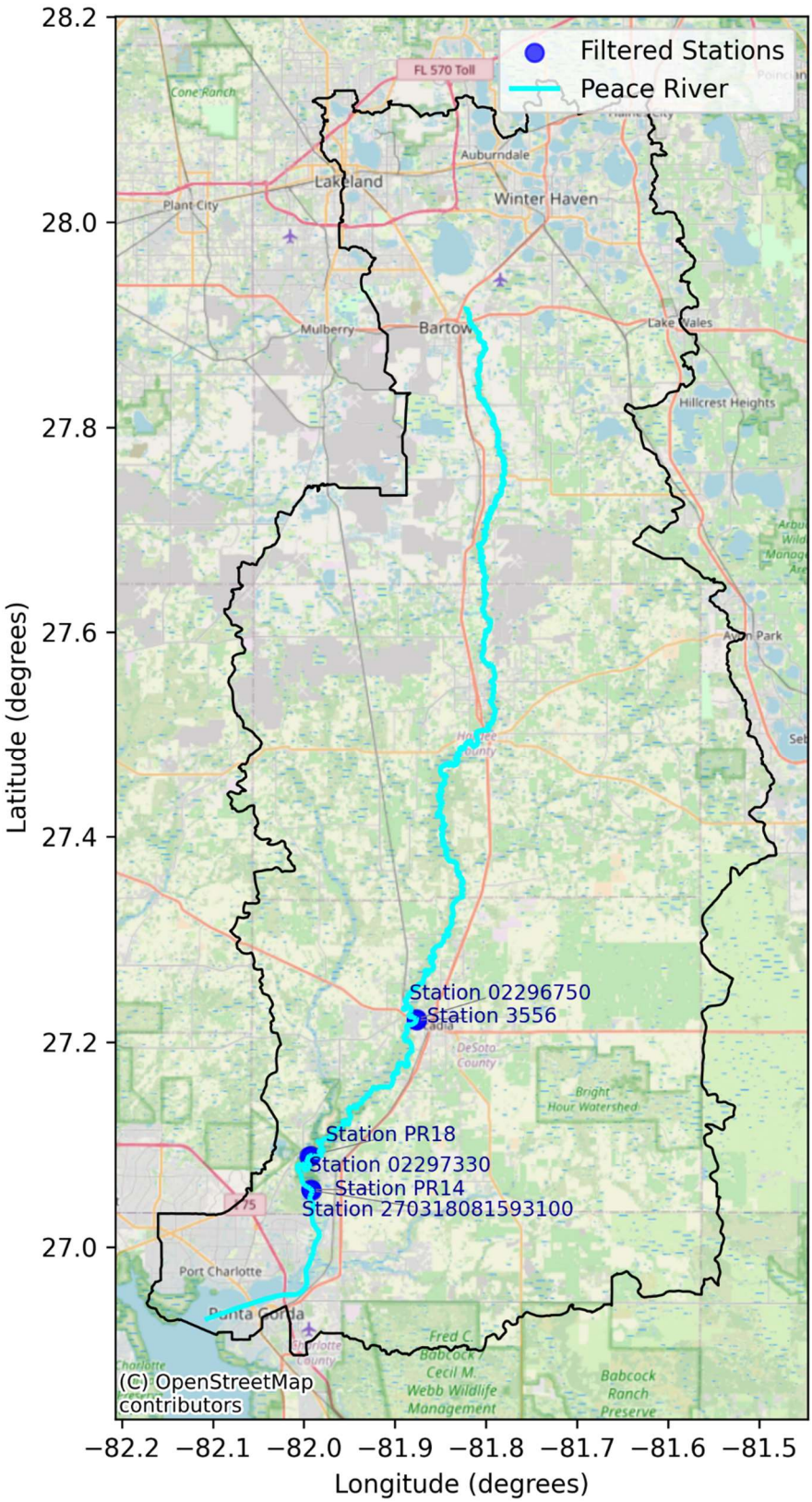

Figure 4. Peace River watershed showing key monitoring stations near Charlotte Harbor with ≥30 TN and TP samples recorded between March 21, 1990, and November 5, 2024.

The time-series data reveal distinct temporal patterns in TN and TP concentrations, with notable peaks in the late 1990s and early 2000s, highlighting periods of elevated nutrient loading that may have contributed to increased red tide severity. Also, the analysis revealed notable seasonal variations, with higher discharge and nutrient loads during the wet season (June–September). This period is critical as elevated freshwater flows enhance nutrient delivery to coastal waters, potentially exacerbating red tide conditions. The Arcadia station (USGS 2296750) data along with the TN and TP thus provide essential insights into the seasonal and long-term dynamics of freshwater and nutrient inputs influencing red tide events in the Gulf of Mexico.

Figure 5. Time-series data for Total Nitrogen (TN) and Total Phosphorus (TP) concentrations from key monitoring stations near Charlotte Harbor with ≥30 samples recorded between March 21, 1990, and November 5, 2024. The variability in TN and TP concentrations reflects seasonal and event-driven fluctuations, with higher peaks observed during specific periods, suggesting episodic nutrient loading that could influence red tide development.

3. Ocean Data

Oceanographic conditions, particularly the Loop Current, water temperature, and salinity, significantly influence the dynamics and persistence of Karenia brevis blooms along Florida's Gulf Coast. The Loop Current, a warm ocean current entering the Gulf of Mexico from the Caribbean Sea and exiting via the Florida Straits into the Atlantic Ocean, affects bloom dynamics by altering nutrient availability, water temperature, and coastal circulation patterns.

We hypothesized that elevated water temperatures enhance the development and intensification of Karenia brevis blooms, while intermediate salinity levels create optimal conditions for their growth. Conversely, extreme salinity variations may inhibit bloom formation. To investigate these relationships, we analyzed long-term HAB datasets from the NOAA National Centers for Environmental Information (NCEI), focusing specifically on the Charlotte Harbor region. This dataset included detailed measurements of Karenia brevis cell counts, water temperature, and salinity, which were geographically segmented to facilitate regional analysis between Tampa Bay and Charlotte Harbor. Spatial analyses revealed correlations between higher bloom concentrations and shallower bathymetric features near the 300-meter contour, underscoring the role of water depth and Loop Current-driven circulation in bloom formation (Figure 6).

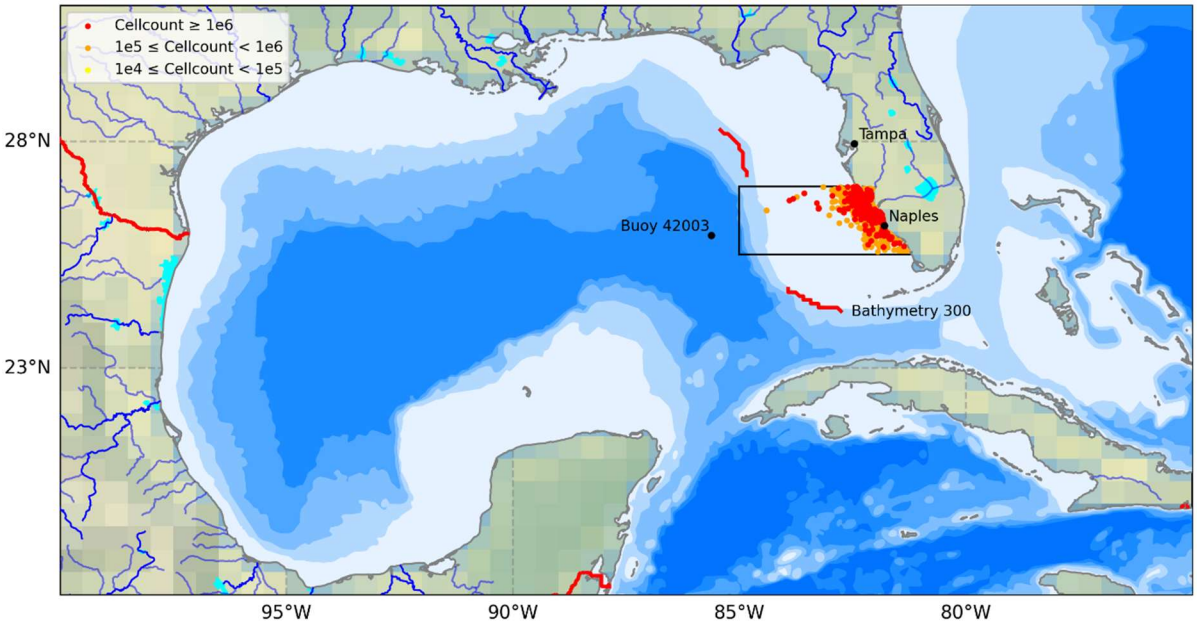

Figure 6. Gulf of Mexico bathymetry and Karenia brevis bloom severity near Charlotte Harbor. Red, orange, and yellow dots represent cell counts greater than or equal to 10^6, between 10^5 and 10^6, and between 10^4 and 10^5, respectively. Major cities (Tampa and Naples) and buoy 42003 are marked. Bathymetric contours appear in shades of blue, with deeper regions represented by darker shades. The 300-m bathymetry contour (two red segments) differentiates the north and south phases of the Loop Current.

Sea surface height anomalies further illustrate the influence of the Loop Current on regional oceanographic conditions. We hypothesized that elevated SSH anomalies associated with

intensified Loop Current conditions increase coastal nutrient upwelling, promoting favorable bloom conditions. Variations in Loop Current position and strength may also modulate coastal flow rates, temperature and salinity, directly affecting bloom intensity and distribution. Persistent high SEA SURFACE HEIGHT anomalies might indicate prolonged Loop Current activity linked to extended or intensified bloom events. Spatial SEA SURFACE HEIGHT anomaly patterns across the Gulf of Mexico (Figure 7) demonstrate the Loop Current's role in regional water elevation changes, with positive anomalies indicating higher sea levels and intensified Loop Current activity.

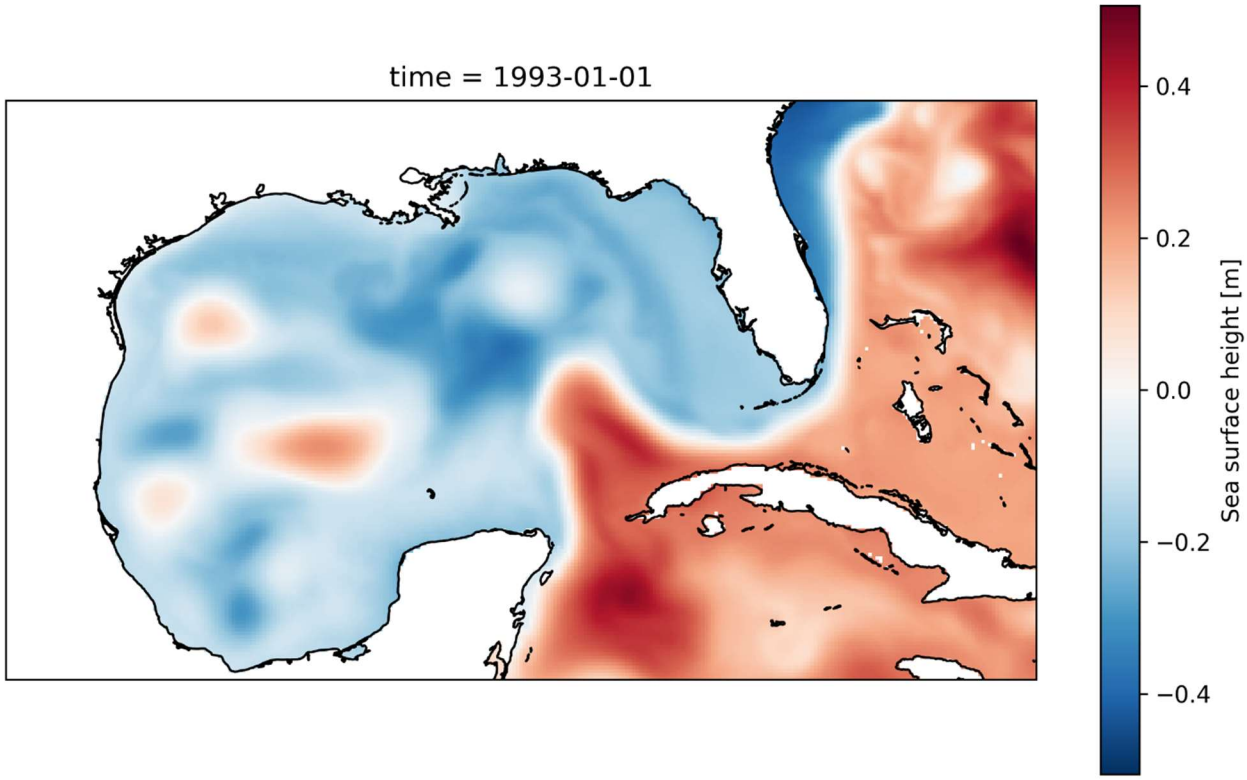

Figure 7. Spatial distribution of sea surface height anomalies in the Gulf of Mexico. Warm colors indicate positive anomalies (higher sea levels), while cooler colors show negative anomalies (lower sea levels).

We computed spatially averaged sea surface height above geoid (zos) across the Gulf from 1992 to 2024 to analyze temporal variability and trends. Results indicate clear cyclic variability, with an upward trend due to sea level rise (Figure 8). To isolate sea-level rise trend, we detrended the zos data using Locally Weighted Scatterplot Smoothing (LOWESS). Applying a smoothing parameter, we removed slow background trends while preserving shorter-term variations. The detrended data (Figure 9) clarify high-frequency zos variations attributable to the Loop Current. In future study, this simple approach of averaging sea surface height over the Gulf of Mexico can be readily replaced with more advanced analysis methods(Elshall et al., 2021, 2022b, 2022a; Maze et al., 2015; Weisberg et al., 2014) .These advanced analyses are needed to provide insights into the oceanographic conditions driving red tide dynamics and context to an integrated modeling approach for predicting and managing harmful algal bloom events.

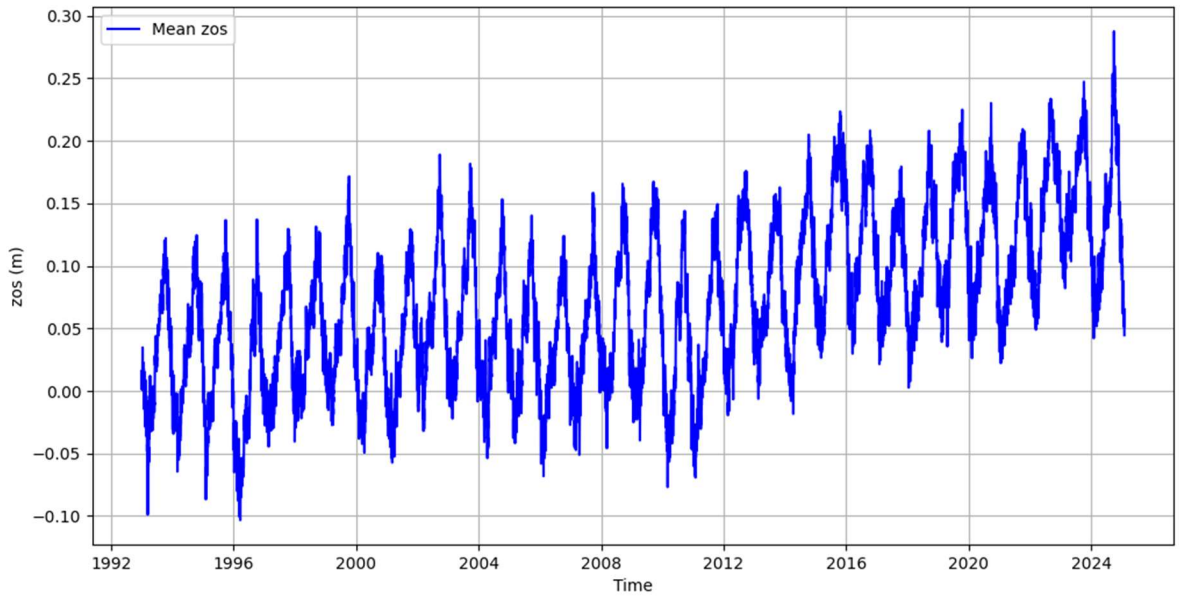

Figure 8. Time series of mean sea surface height anomaly (zos) in the Gulf of Mexico (1992–2024). The data reveal both seasonal and interannual variability, with an increasing long-term trend reflecting potential changes in Loop Current dynamics.

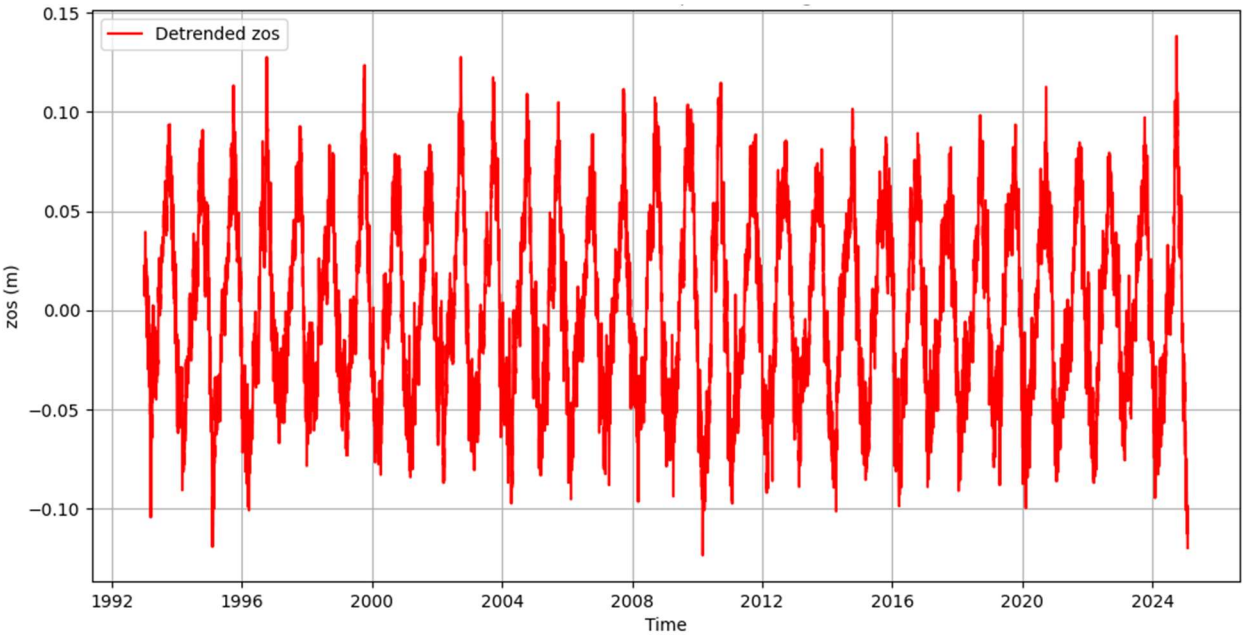

Figure 9. Detrended sea surface height anomaly using LOWESS smoothing, highlighting high-frequency variations associated with Loop Current activity.

4. Atmospheric Data

Atmospheric conditions, particularly wind speed and direction, influence the dynamics of coastal processes and marine harmful algal blooms (HABs). We hypothesized that wind speed and

direction substantially affect the formation, transport, and persistence of red tides. Prevailing onshore winds may facilitate the transport of offshore Karenia brevis blooms toward coastal waters, increasing bloom intensity near shorelines. Conversely, offshore winds may disperse blooms and reduce their coastal impact.

To characterize these conditions, we obtained historical wind data from the NOAA National Data Buoy Center (NDBC), specifically utilizing data from buoy station 42003 located near 300-m Bathymetry (Figure 8). The selected station provides long-term hourly measurements of wind speed and direction, along with air and water temperature, spanning from 1990 to 2022. The dataset was systematically downloaded, merged, and cleaned to address discrepancies. Wind data were analyzed to illustrate temporal patterns and distributions. The method involved resampling wind and environmental data to a daily scale by calculating the mean, median, and mode of wind direction and the maximum of other parameters (wind speed, air temperature, and water temperature). The wind direction was converted into sine and cosine components, which were weighted by wind speed to compute a circular mean using vector averaging. The median direction was computed using circular statistics, while the mode was determined by identifying the most frequently occurring wind direction for each day. Finally, wind speed, air temperature, and water temperature were aggregated using their daily maximum values.

Wind speed data exhibit clear seasonal and episodic patterns, with higher wind speeds observed during storm events and the winter months. Extreme wind events were defined using the 99th percentile of the wind speed distribution, which corresponds to 13.1 m/s (Figure 10). This threshold was selected because wind speeds exceeding this value represent high-energy atmospheric conditions that could enhance bloom transport and mixing. Storm events, when available, were overlaid to assess alignment with extreme wind patterns. Gaps in the wind speed record reflect periods of instrument failure or data collection interruptions, which are common in long-term buoy datasets.

Wind speed shows a distinct seasonal cycle (Figure 11), with higher values during winter and lower values in summer, contrasting with air and water temperature patterns, which peak in late summer. The inverse relationship suggests that calmer atmospheric conditions during warmer months may contribute to bloom persistence nearshore. The wind rose (Figure 12) indicates that easterly and northeasterly winds are most common, with moderate to strong winds predominantly from the east. This suggests that wind-driven transport of blooms toward the shore is likely, reinforcing the hypothesis that onshore winds facilitate bloom concentration along the coast. The consistency of wind patterns aligns with known coastal dynamics, where sustained onshore winds can enhance nutrient delivery and bloom persistence through water column mixing and coastal upwelling.

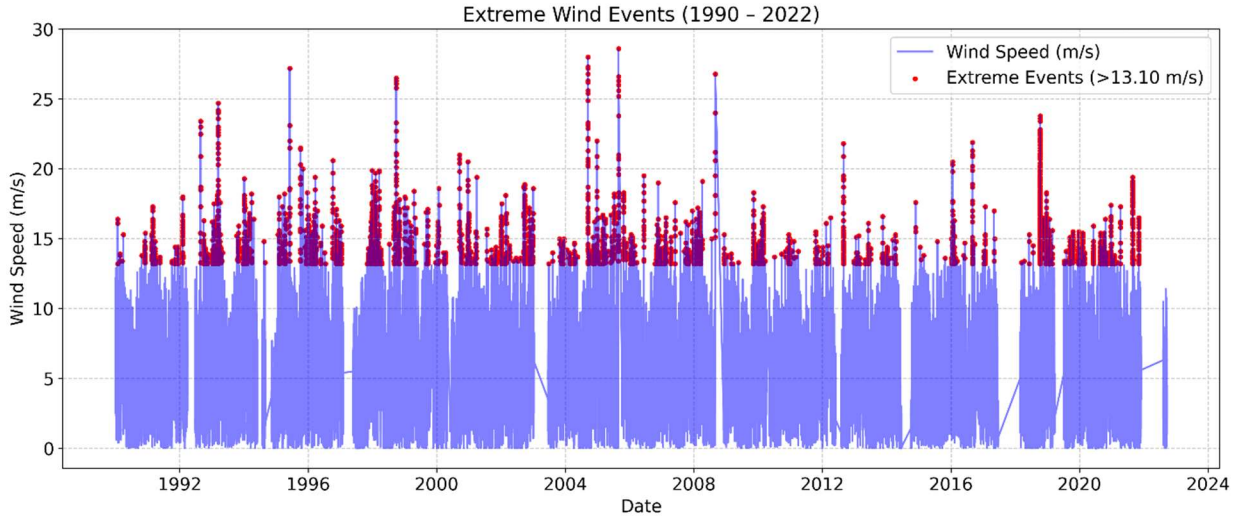

Figure 10. Time-series of wind speed (1990–2022) from NOAA station 42003. Red dots indicate extreme wind events (>13.1 m/s). The data reflect seasonal and interannual variability, with higher wind speeds observed during storm events and specific seasons

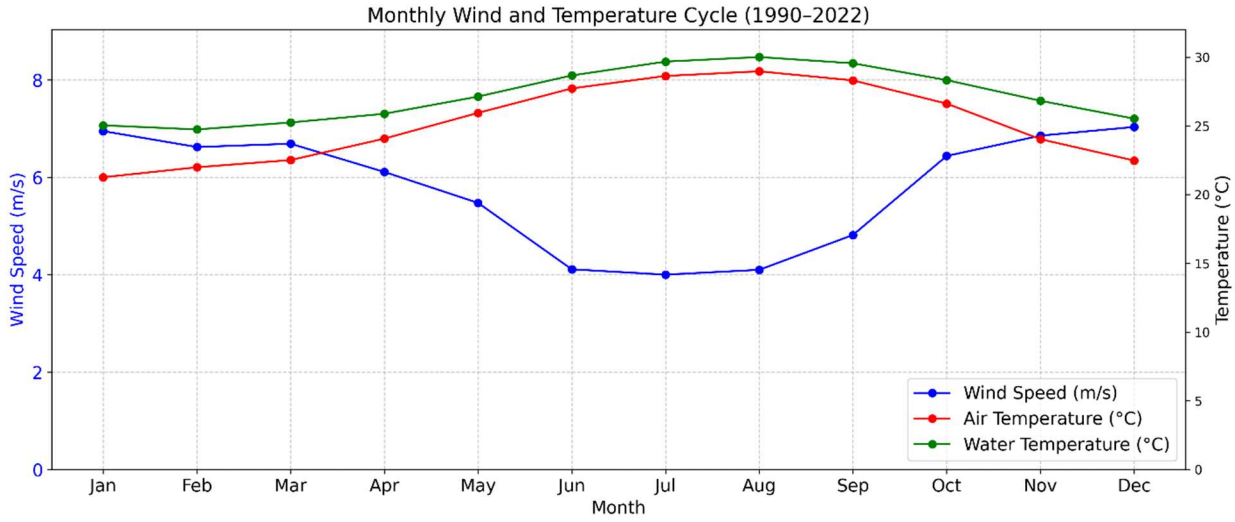

Figure 11. Monthly average wind speed, air temperature, and water temperature from NOAA station 42003 (1990–2022). Wind speed (blue) is shown on the left axis, while air and water temperature (red and green, respectively) are shown on the right axis.

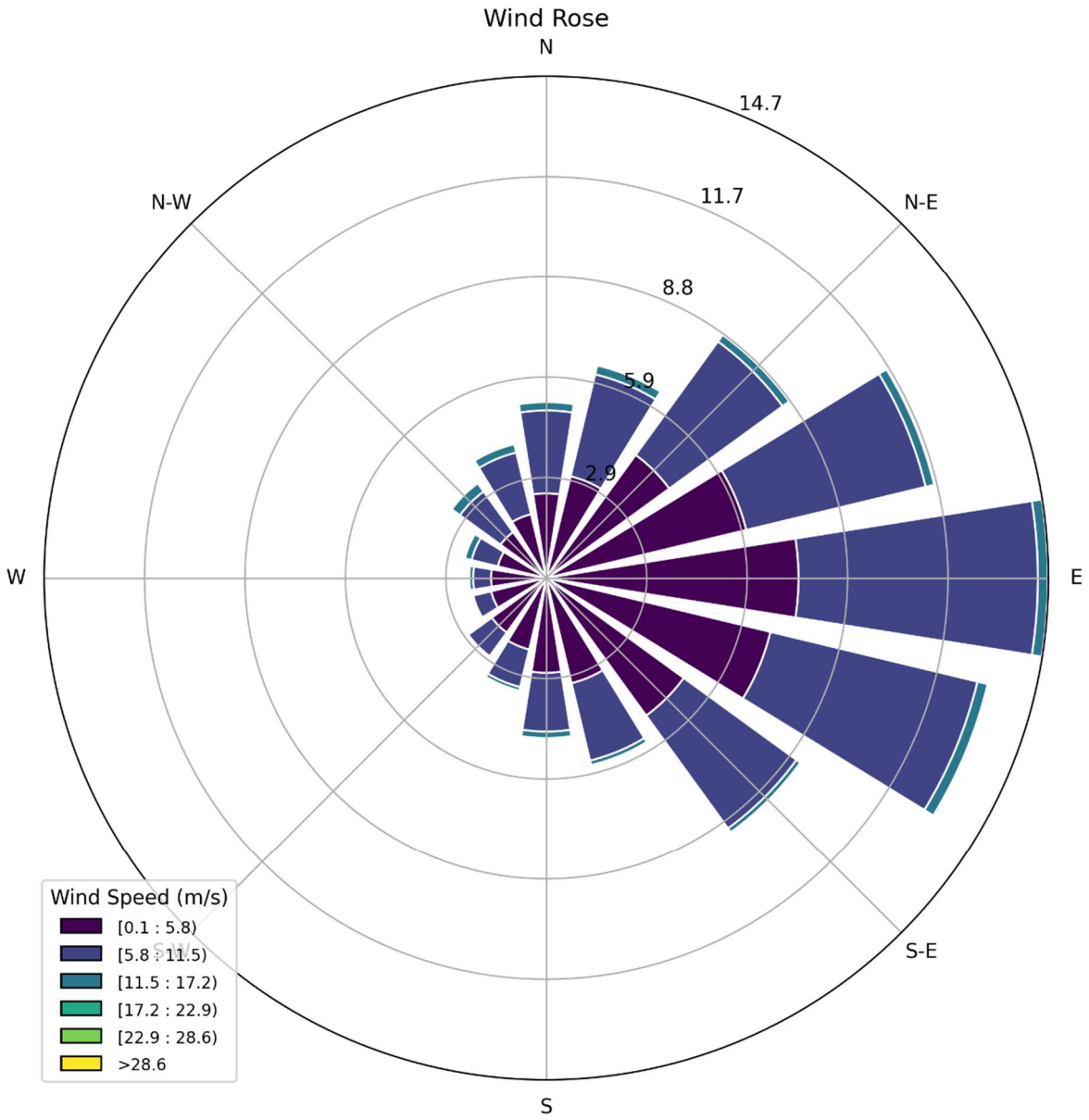

Figure 12. Wind rose diagram of wind speed and direction from NOAA station 42003 (1990–2022). Colors represent wind speed ranges, and the length of the bars represents the frequency of wind from each direction.

5. Combined weekly data

Weekly resampling was performed by computing weekly averages of key variables including wind direction, wind speed, air temperature, water temperature, salinity, and Karenia brevis cell counts. Water temperature and wind direction and speed were interpolated using seasonal averages, using the weekly historical mean data for the corresponding week across all available years. Missing values for salinity, Peace River discharge, TN and TP were filled using linear

interpolation. Lastly, missing Karenia brevis cell count values were filled with zeros as sampling generally does not occur for no bloom periods. The finalized interpolated dataset was then constrained to the period from January 1, 1993, to December 31, 2023, and saved for subsequent analysis.

The presented weekly data in Figure 13 offers comprehensive insights into the temporal variability of factors influencing Karenia brevis blooms. Blooms exhibit sporadic peaks over the period analyzed, indicating episodic rather than continuous occurrences. The detrended sea surface height anomaly (zos) highlights high-frequency variations linked to the Loop Current, potentially influencing bloom formation through nutrient availability and coastal circulation. Salinity and water temperature display clear seasonal patterns, emphasizing their role in bloom initiation and persistence, with intermediate salinity and elevated temperatures often correlating with bloom peaks. Wind data illustrate variable directions and intensities, which could influence bloom transport and dispersion. River discharge, along with nutrient concentrations (TN and TP) from Peace River, shows significant variability, highlighting episodic freshwater inflow events that may alter coastal nutrient dynamics and contribute to bloom conditions. Together, these data facilitate an integrated analysis of environmental controls governing harmful algal bloom dynamics using machine learning.

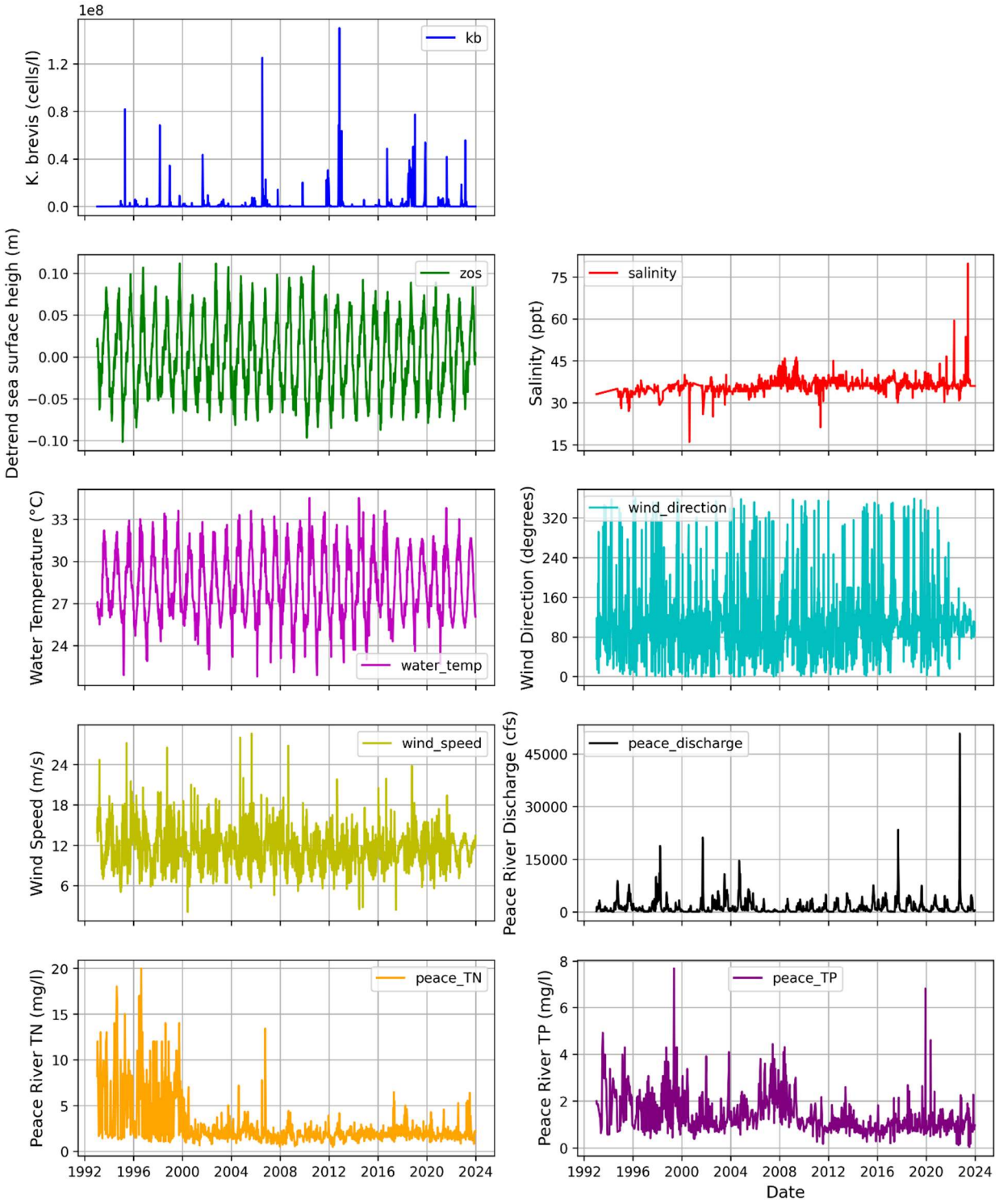

Figure 13. Weekly time series of environmental and oceanographic variables related to Karenia brevis bloom dynamics in the Charlotte Harbor region.

6. Machine learning model

6.1 Method

The development of an accurate early-warning system for Karenia brevis blooms using machine learning (ML) framework involves multiple environmental drivers. These include river discharge, nutrient concentrations (TN and TP) atmospheric conditions (wind speed and direction), water temperature, salinity, and detrended sea surface height anomalies (zos). The following details outline the methodological process comprehensively.

The weekly interpolated environmental dataset (1992–2024) was preprocessed to ensure data quality and robustness for classification modeling. A robust scaling approach (RobustScaler) was applied to reduce the influence of outliers, particularly in river discharge and nutrient inputs, enhancing the model's resistance to anomalous events(Medina et al., 2022). RobustScaler scales data by removing the median and dividing by the interquartile range, making it resistant to outliers. To prevent data leakage, features were scaled after splitting into training and test sets to ensure unbiased model evaluation. Data leakage occurs when information from the test set is used during training, leading to overly optimistic model performance and poor generalization to unseen data. Feature engineering captured temporal patterns in Karenia brevis bloom dynamics. Lag features included 1- and 2-week lags of Karenia brevis counts, zos anomalies, salinity, water temperature, wind speed, river discharge, TN, and TP to account for temporal autocorrelation (Medina et al., 2024). A 4-week rolling average for Peace River discharge was used to reflect cumulative hydrological influences on bloom development.

A Random Forest Classifier from Scikit-Learn was selected for its ability to handle complex interactions and class imbalance. Class imbalance occurs when the number of samples in one class significantly exceeds the other in a classification task, which can bias the model toward predicting the majority class and reduce performance for the minority class. The model used 100 estimators meaning it constructed 100 decision trees, applied balanced class weights to adjust for class imbalance by giving higher weight to the minority class, and ensured reproducibility with a fixed random state 42. The binary classification task identified weekly bloom presence based on Karenia brevis cell counts ≥ 100,000 cells/L. Training data spanned from 1993 to 2018, while testing used data from 2019 onward to maintain temporal independence. We compare a 1-week and 2-week forecast. A 1-week forecast predicts bloom events one week ahead, capturing short-term persistence and immediate driver effects. A 2-week forecast predicts bloom status two weeks out, requiring longer lag features and potentially sacrificing some accuracy due to changing bloom dynamics, but providing more lead time for management. If the accuracy drop is small, a 2-week forecast may be preferred; otherwise, a 1-week forecast with weekly updates might be optimal. The model will update weekly using real-time data, with periodic retraining (e.g., annually) to adjust for long-term changes, ensuring timely and accurate predictions.

6.2 Results

Figure 14 compares the model's predicted bloom status (red, dashed) with the actual bloom status (blue, solid) over the test period (2019–2024). The alignment between the predicted and actual values indicates the model’s ability to capture bloom events accurately over time. Some

mismatches suggest that the model may occasionally overpredict or underpredict bloom events, but overall, it captures the bloom patterns effectively.

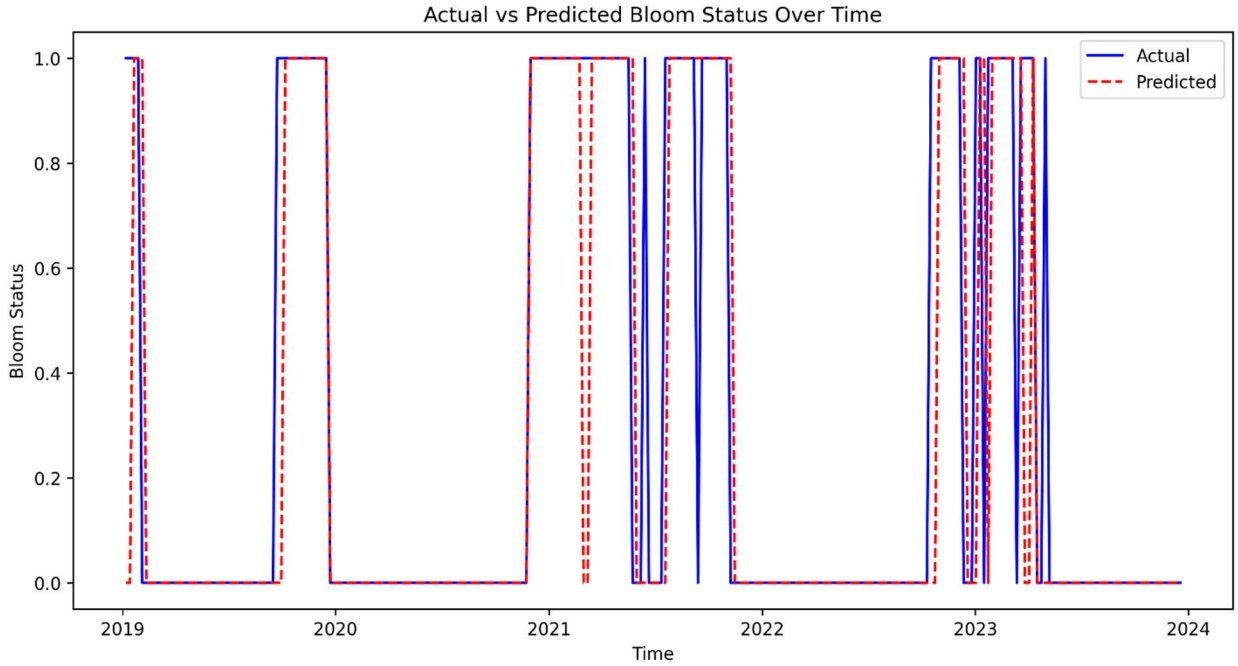

Figure 14. actual vs predicted bloom status over time.

The confusion matrix (Figure 15) summarizes model predictions against actual observed conditions, providing clarity on the type and frequency of prediction errors. Specifically, out of 179 actual no-bloom cases, 172 were correctly identified (true negatives, TN), and 7 were incorrectly labeled as blooms (false positives, FP). Conversely, among 80 actual bloom instances, 65 were accurately predicted (true positives, TP), whereas 15 bloom instances were missed by the model (false negatives, FN).

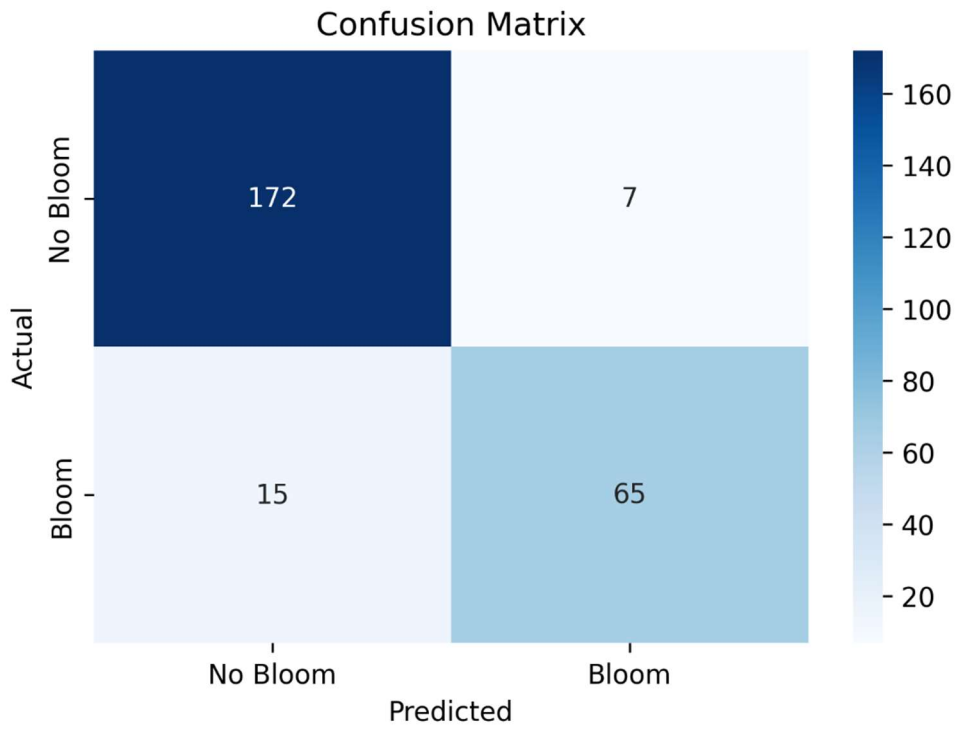

Figure 15. Confusion matrix shows the model's classification performance, where the model correctly predicted 172 non-bloom cases and 65 bloom cases, with 7 false positives and 15 false negatives. The high values along the diagonal reflect strong predictive accuracy.

Key metrics derived from the confusion matrix are accuracy, precision, recall, F1 score, and support. Accuracy measures the overall percentage of correct predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{65}{65 + 172 + 7 + 15} = 0.92 \ \ (92\%)$$

The precision matric measures how many of the predicted positive cases (blooms) were correct:

$$Precision = \frac{TP}{TP + FP} = \frac{65}{65 + 7} = 0.9 \ \ (90\%)$$

High precision means fewer false positives. That is the model does not incorrectly predict blooms very often. Recall (sensitivity or true positive rate) measures how many of the actual positive cases (blooms) were correctly identified:

$$Recall = \frac{TP}{TP + FN} = \frac{65}{65 + 15} = 0.81 \ \ (81\%)$$

High recall means the model catches most of the bloom events. F1-Score is the harmonic mean of precision and recall, which balances the trade-off between precision and recall

$$F1 = 2\frac{Precision \times Recall}{Precision + \ Recall} = 2\frac{0.9 \times 0.81}{09 + 0.81} = 0.86 \ \ (86\%)$$

A high F1-score means the model achieves a good balance between precision and recall. Support is the number of actual samples in each class that are No Bloom (179 total instances) and Bloom (80 total instances).

Table 2. Summary of key metrics

Class	Precision	Recall	F1-score	Support
No Bloom	0.92	0.96	0.94	179
Bloom	0.9	0.81	0.86	80
Accuracy			0.92	259
Macro Avg	0.91	0.89	0.9	259
Weighted Avg	0.91	0.92	0.91	259

Table 2 shows a summary of key metrics. The bloom class precision of 0.90 indicates that when the model predicts a bloom, there is a 90% likelihood it is correct. For the no-bloom class, precision is higher (0.92), indicating slightly greater reliability in predicting non-bloom conditions. The recall (also known as sensitivity) measures the model's ability to detect actual bloom events, with the bloom class recall at 0.81 indicating the model successfully identified 81% of all actual bloom occurrences. The recall for no-bloom cases is notably higher at 0.96, meaning it correctly identified 96% of all no-bloom instances. The F1-score provides a balanced measure combining precision and recall, especially useful when class imbalance exists. An F1-score of 0.86 for blooms reflects good balance between precision and recall, though highlighting room for improvement in recall. Finally, support reflects the number of observations for each class within the test dataset (179 no-bloom and 80 bloom cases), showing the imbalance which the model effectively managed by using balanced class weights. Overall, the balanced accuracy (0.887), combined with the confusion matrix, precision, recall, and F1-score metrics, demonstrates robust performance suitable for practical use in operational forecasting. Improvements could focus on reducing false negatives to further enhance bloom detection accuracy.

The feature importance plot (Figure 16) ranks the relative contribution of each input feature to predicting Karenia brevis bloom severity. Feature importance reflects how much each feature reduces classification error, calculated from improvements in the Gini Index across all trees. Karenia brevis cell counts (kb) are the most influential feature, highlighting the strong autocorrelation of bloom persistence. Lagged bloom counts (kb_prev1, kb_prev2) rank next, confirming that past bloom status predicts future blooms. Peace River discharge (discharge_4w_avg) and nutrient loading (TN, TP) are also key drivers, reinforcing the role of nutrient-rich runoff in fueling blooms (Medina et al., 2022). Sea surface height anomalies (zos) have moderate importance, supporting the hypothesis that Loop Current variability influences nutrient upwelling. However, this importance can improve through improved feature engineering. Wind speed, direction, and salinity are less important, suggesting they have secondary effects on bloom transport rather than initiation. This analysis confirms that bloom persistence and nutrient inputs are the strongest short-term drivers, guiding future forecasting and mitigation strategies.

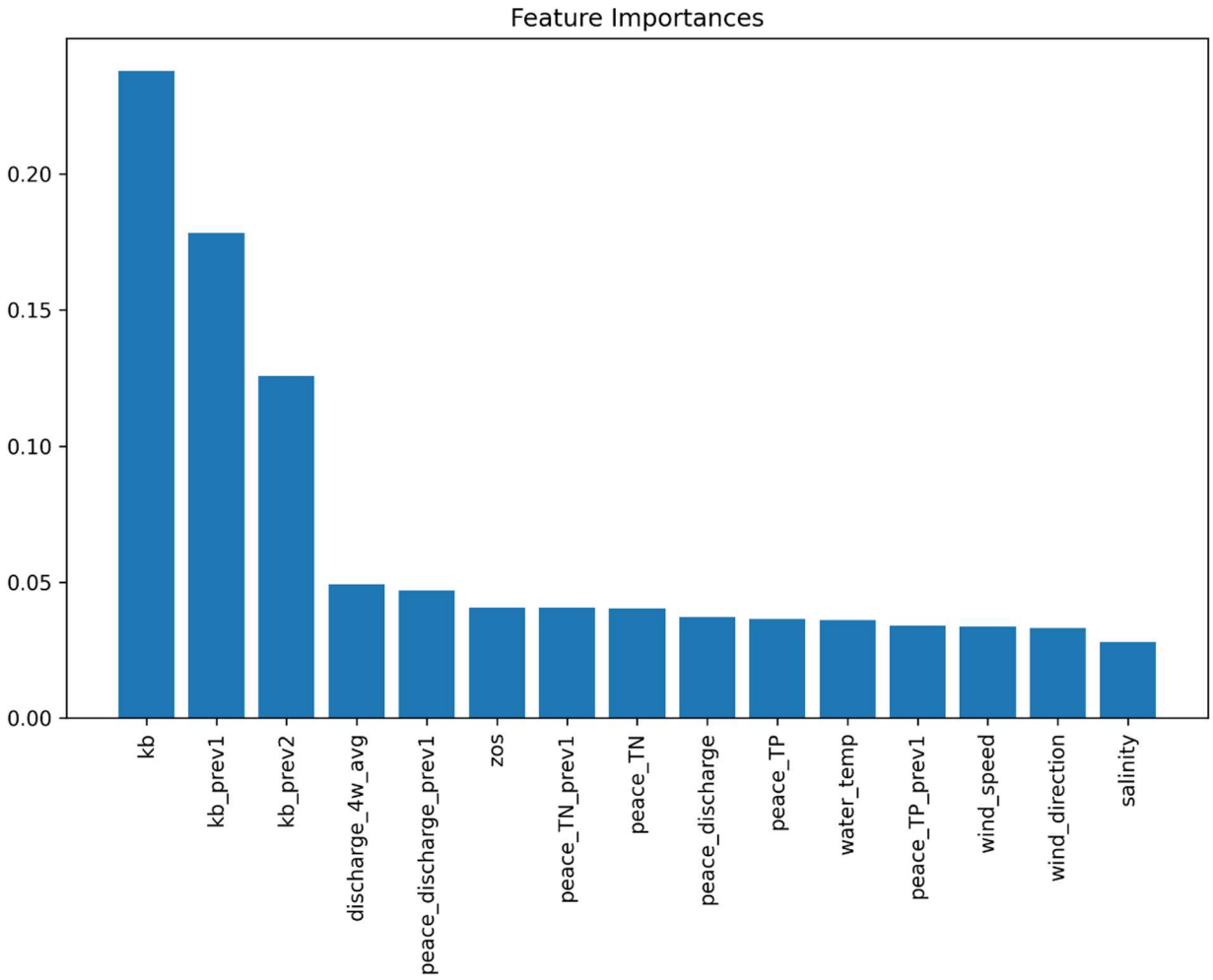

Figure 16. Feature importance ranks the input variables based on their contribution to the model's predictive accuracy. Karenia brevis cell counts (current and lagged) are the most influential, followed by Peace River discharge and nutrient loading, highlighting the key drivers of bloom dynamics.

The learning curve (Figure 17) shows how model accuracy changes as the training set size increases. The x-axis represents the training set size, while the y-axis shows accuracy. The training score (blue) remains high (~1.0), indicating that the model fits the training data very well, suggesting potential overfitting. The validation score (orange) starts low but increases with more data, eventually leveling off around 0.85, indicating that the model is generalizing well but may have reached a saturation point where additional data will not significantly improve performance. The gap between the training and validation scores confirms overfitting, which could be reduced by simplifying the model (e.g., reducing tree depth) or increasing regularization. The stable validation score suggests the model has learned most of the useful patterns, but fine-

tuning could further improve generalization. Also, trying models like Logistic Regression, Decision Trees, or Support Vector Machines (SVM) might help reduce overfitting.

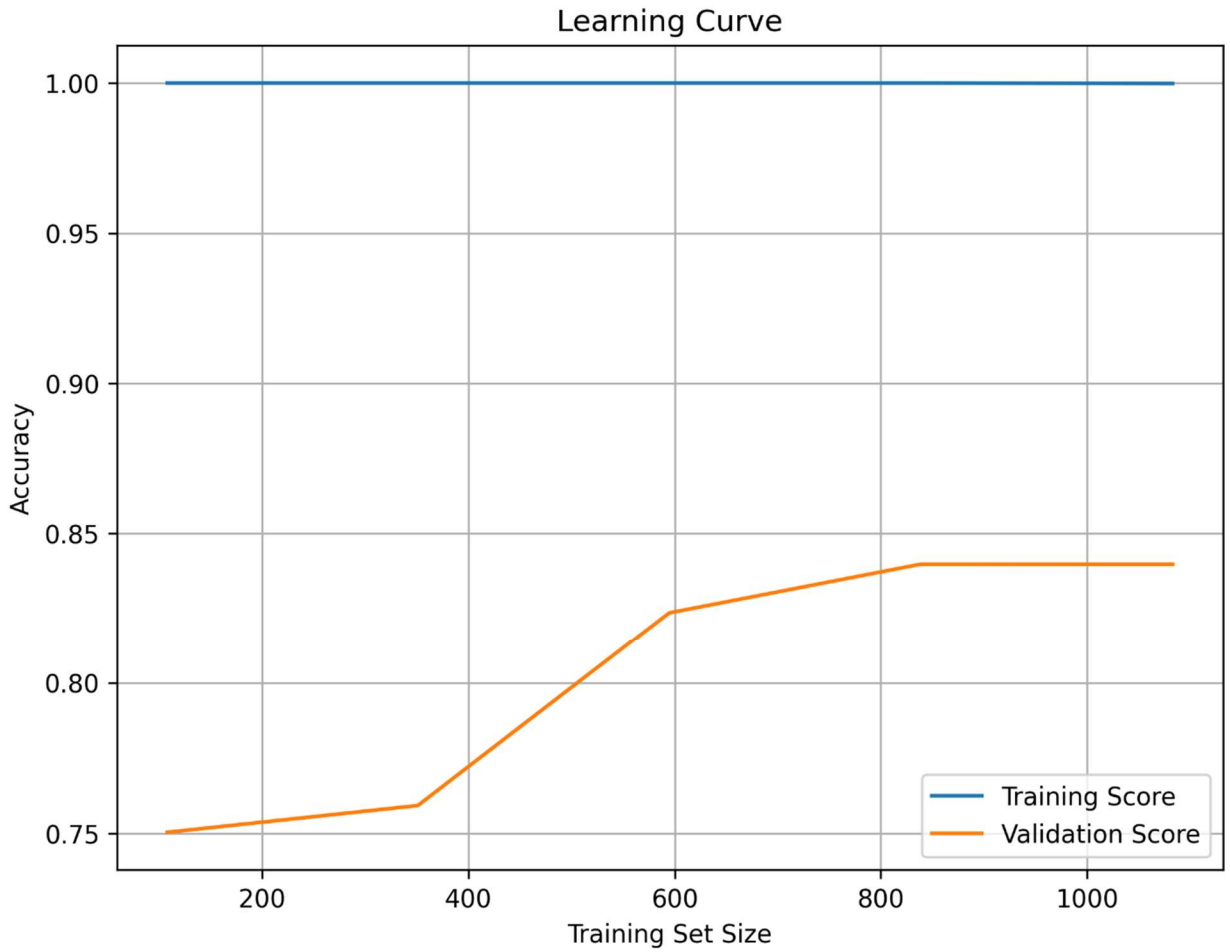

Figure 17. Learning curve shows how model accuracy changes with increasing training data size. The high training score (~1.0) and the gap with the validation score (~0.85) suggest mild overfitting, but the validation score's plateau indicates that the model has learned most of the meaningful patterns in the data.

The pair plot (Figure 18) shows the relationships and distributions between Karenia brevis (kb), Peace River discharge, Peace River total nitrogen (TN), and Peace River total phosphorus (TP). Peace River discharge has a right-skewed distribution with several high outliers, which suggests episodic high-discharge events. Peace River TN and Peace River TP both show clustered patterns, indicating that nutrient concentrations are relatively stable most of the time but spike under certain conditions. There is no clear linear correlation between bloom intensity (kb) and river discharge or nutrient levels, but higher nutrient loads and river discharge events may correspond to increased bloom intensity, as suggested by the scatter clusters.

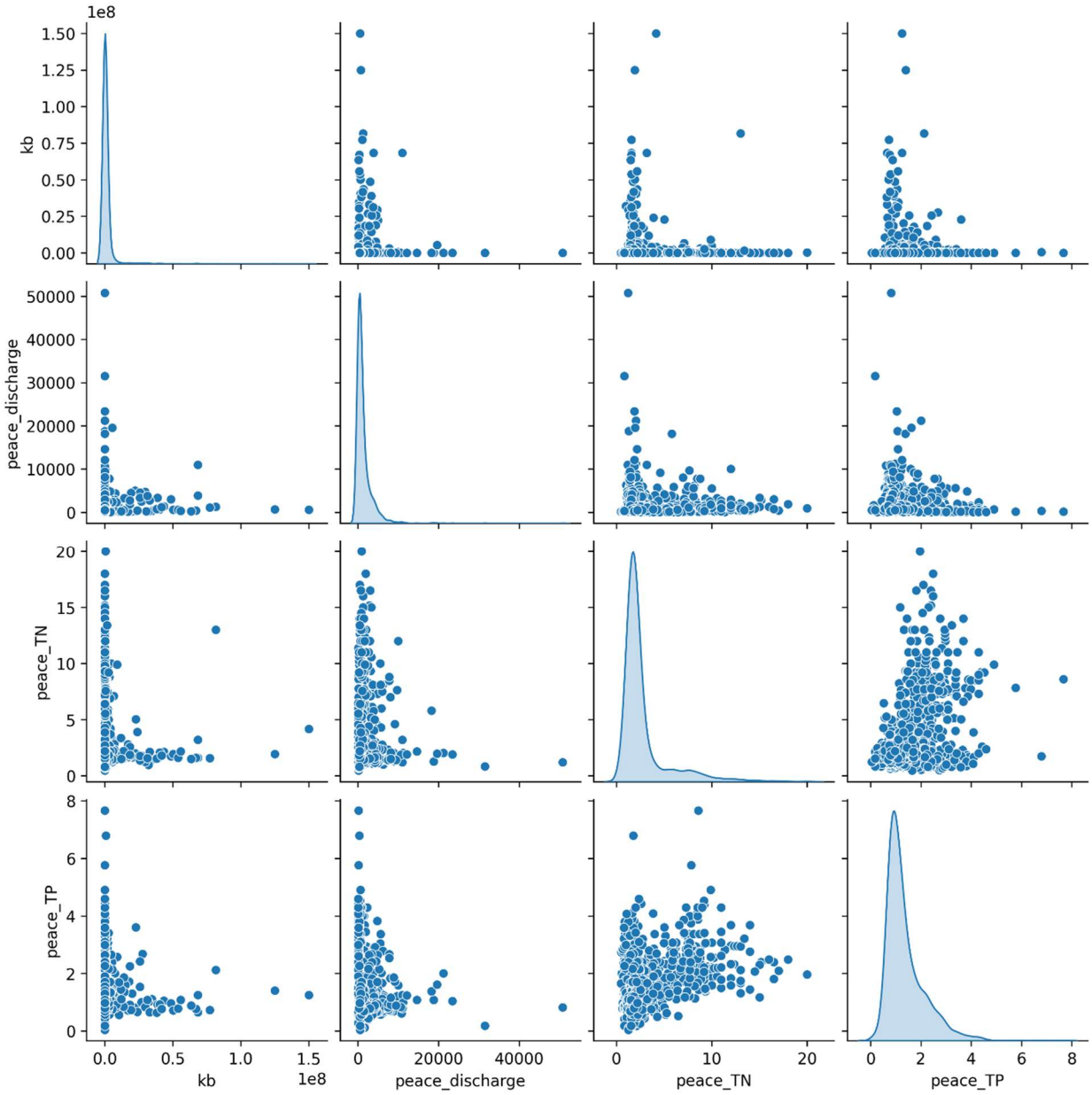

Figure 18. Pair plot of environmental variables visualizes relationships between key environmental variables, including Karenia brevis cell counts, Peace River discharge, total nitrogen (TN), and total phosphorus (TP). The scatter plots reveal nonlinear patterns and clustering, suggesting complex interactions influencing bloom formation.

Partial dependence plots (Figure 19) illustrate the effect of Peace River discharge, TN, and TP on Karenia brevis bloom predictions while holding other variables constant. The y-axis represents the predicted bloom probability, and the x-axis shows the range of feature values. For Peace River discharge, the model predicts higher bloom probability as discharge increases from low to moderate levels, but the relationship weakens at higher discharge levels, suggesting saturation

or other limiting factors. For Peace River TN and Peace River TP, the model shows increased bloom probability with rising nutrient concentrations up to 2.5–5 units, after which the effect plateaus, indicating diminishing returns from higher nutrient levels. The clustering of black ticks at the bottom reflects the actual data distribution, highlighting where the model has more training data. These patterns confirm that nutrient loading and river discharge are key bloom drivers, consistent with ecological understanding.

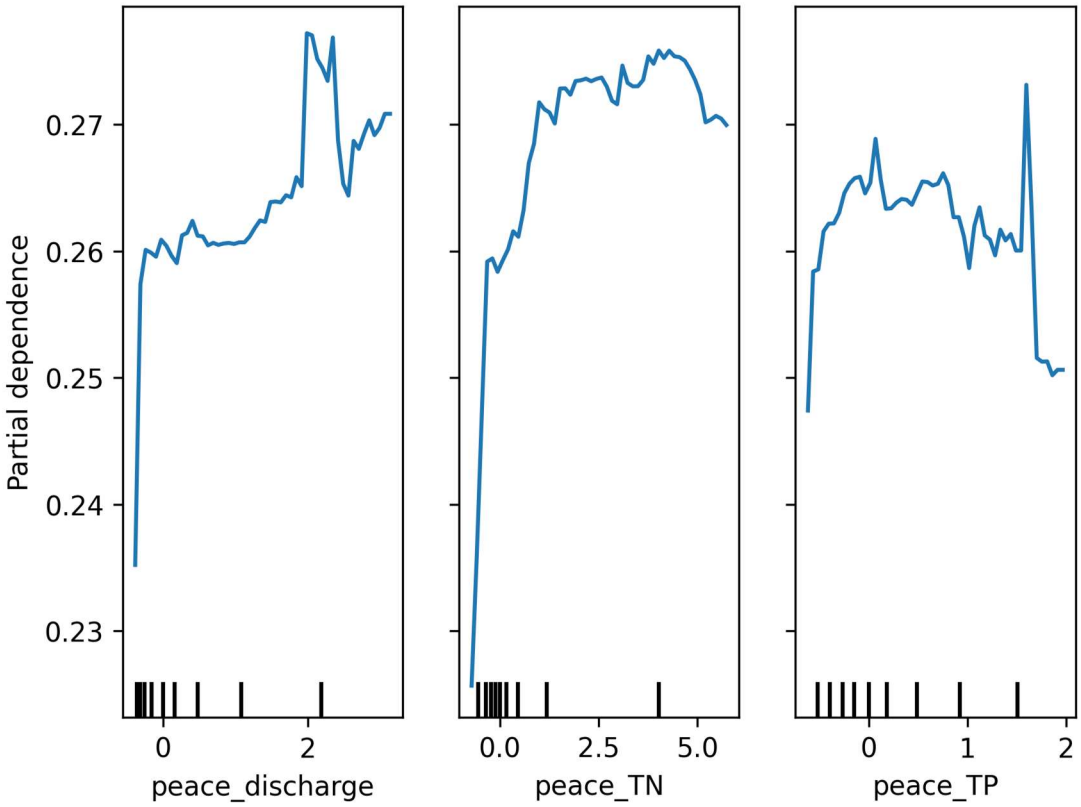

Figure 19. Partial dependence plots show the marginal effect of Peace River discharge, TN, and TP on the predicted bloom probability. Increased nutrient loading and discharge initially raise bloom likelihood, but the effect levels off at higher values, indicating potential saturation or other limiting factors.

The precision-recall (PR) curve evaluates the model's performance across different classification thresholds, showing how precision changes with recall. High recall means the model captures more true blooms but may increase false positives, lowering precision, while high precision means fewer false positives but possibly missing some blooms. The model maintains high precision (~ 0.9) even at moderately high recall (~ 0.7), indicating a strong balance between capturing true blooms and avoiding false positives. The sharp drop in precision at very high recall (>0.8) suggests the model struggles to maintain precision when trying to capture all blooms. This balance reflects the model's strong performance, as indicated by the high F1-score and balanced accuracy. Adjusting the decision threshold can prioritize higher recall (capturing more blooms) or

higher precision (minimizing false alarms) based on operational needs, with the F1-score or PR area under curve guiding the optimal threshold.

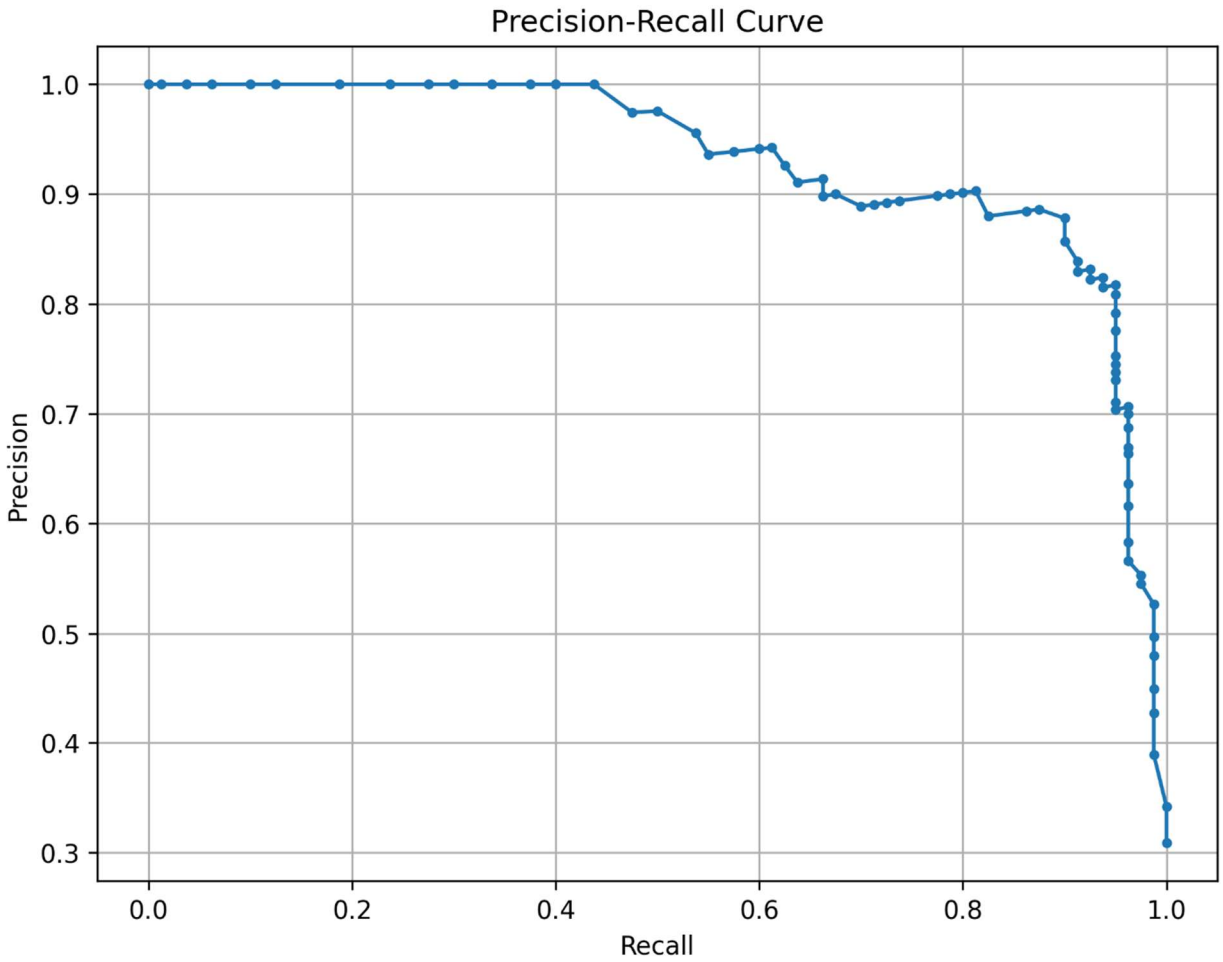

Figure 20. Precision-recall curve illustrates the trade-off between precision and recall at different classification thresholds. The model maintains high precision (~0.9) at moderate recall values (~0.7), indicating that it effectively balances detecting true blooms while minimizing false positives.

6.3 Discussion

This framework incorporates insights from prior studies as well as novel features. We include river discharge and nutrient load features, recognizing their critical role in coastal bloom intensification (Medina et al. 2022STE Nitrogen). We use lagged and current bloom state as features, an approach validated by Medina et al. (2024) to significantly improve predictive accuracy. Future work should test multiple algorithms (tree ensembles, SVM, LSTM), to address findings on other studies (Elshall et al., 2020; Li et al., 2021) that different methods have trade-offs in recall and precision. In particular, the use of LSTM addresses the proposal's note that sequential models can outperform static models for HAB prediction. Even though we used a basic ML model, our preprocessing (robust scaling, etc.) and augmentation (oversampling) steps to

reduce input data uncertainty and handle outliers resulted in higher accuracy than previous studies. Also, the focus on balanced accuracy and recall highlights the need for reliable bloom detection (even at the cost of a few false alarms) in an early warning context, which is emphasized in the literature.

With respect to designing a forecast system, we first need an offline phase to maximize performance leveraging all historical data, extensive feature engineering, and robust models like XGBoost or LSTM. Once the best model and feature set are identified (with balanced accuracy as a key metric), the model can be streamlined for deployment – using the most important drivers, automating data ingestion, and running forecasts on a weekly schedule. This ensures we achieve high predictive skill in the research setting and maintain feasibility for real-time harmful algal bloom forecasting along Florida's southwest coast. For deployment, restricting real-time available data and planning periodic retraining would be needed for operational forecast tools (Medina et al. 2024).

7. Conclusions

This study developed a predictive machine learning model for forecasting Karenia brevis (K. brevis) bloom occurrences along the West Florida Shelf, utilizing a comprehensive environmental dataset spanning from 1992 to 2024. The dataset included key environmental drivers such as K. brevis cell counts, river discharge, nutrient concentrations (total nitrogen and total phosphorus), atmospheric conditions (wind speed and direction), water temperature, salinity, and sea surface height anomalies (zos). Data preprocessing involved robust scaling to minimize outlier impacts, linear and seasonal interpolation methods to handle missing values, and temporal segmentation to avoid data leakage and ensure rigorous model evaluation.

This study successfully demonstrated that machine learning, specifically a Random Forest Classifier, can reliably predict Karenia brevis (K. brevis) bloom occurrences along the West Florida Shelf with high predictive performance (balanced accuracy of 0.887). The model correctly predicted 92% of the test samples (accuracy). It had a high precision of 90%, meaning it rarely issued false bloom alerts. The recall of 81% indicates that the model detected most bloom events but still missed 15 cases (false negatives). The F1-score of 86% reflects that the model maintains a good balance between precision and recall. Temporal autocorrelation of bloom presence was identified as the primary predictor, followed by river discharge and nutrient loadings (total nitrogen and phosphorus). Partial dependence analysis confirmed the positive relationship between river discharge, nutrient concentrations, and bloom likelihood, highlighting critical environmental thresholds for bloom formation. It is recommended that the developed model be implemented as a decision-support tool integrated into real-time environmental dashboards for coastal management in Charlotte Harbor. The model should utilize a 1-week forecast horizon, updated weekly, to balance prediction accuracy with operational practicality. Given evidence of moderate overfitting, hyperparameter tuning or regularization methods such as adjusting tree depth and number of estimators should be explored to further improve generalization and reliability.

Future work should focus on integrating advanced sequential modeling techniques, such as Long Short-Term Memory (LSTM) networks, to better capture complex temporal dependencies in bloom dynamics. Additional efforts should also be dedicated to enhancing input data quality through reanalysis, assimilation techniques, and data augmentation. Exploring Bayesian approaches, such as Temporal Fusion Transformers (TFT), could provide uncertainty quantification beneficial for risk-based decision-making. Additionally, model interpretability can be enhanced using SHapley Additive exPlanations (SHAP) to reveal clear contributions of individual predictors. Lastly, continuous validation against emerging data and periodic model recalibration are crucial for adapting to changing environmental conditions and ensuring sustained predictive accuracy.

Data availability statement
Data and codes that support this study are accessible from: https://aselshall.github.io/redtides

Funding statement
Final Report for Task 9 of Water Quality Evaluation for Peace River Basin and Greater Charlotte Harbor Watershed in Southwest Florida, Florida Department of Environmental Protection (August 2023 – November 2025)

Competing interests
None

AI assistance statement
Data analysis and plotting was conducted using standard Python packages including pandas and matplotlib with code generation assistance from GTP-4o and GPT-4o-mini via Jupyter AI. AI assistance from GPT-4o was utilized for providing review comments, refining text for succinctness and clarity, and providing suggestions for restructuring paragraphs to improve logic flow. AI contributions were verified and contextualized to ensure accuracy and relevance.

Ethical standards
The research meets all ethical guidelines, including adherence to the legal requirements of the study country and Florida Gulf Coast University.

Author contributions
Conceptualization: A.E., Methodology: A.E, Data curation: A.E, Data visualization: A.E. Writing original draft: A.E.

Acknowledgements
Mewcha Amha Gebremedhin developed the Jupyter-book that hosts the data. Matthew Duus assisted with data collection and discussion.

References

Elshall, A., Ye, M., Kranz, S., Harrington, J., Yang, X., Wan, Y., & Maltrud, M. (2021, December 16). Machine learning for red tide prediction in the Gulf of Mexico along the West Florida Shelf [poster]. https://doi.org/10.1002/essoar.10509597.1

Elshall, A., Ye, M., Kranz, S. A., Harrington, J., Yang, X., Wan, Y., & Maltrud, M. (2022a). Application-specific optimal model weighting of global climate models: A red tide example. *Climate Services*, *28*, 100334. https://doi.org/10.1016/j.cliser.2022.100334

Elshall, A., Ye, M., Kranz, S. A., Harrington, J., Yang, X., Wan, Y., & Maltrud, M. (2022b). Earth system models for regional environmental management of red tide: Prospects and limitations of current generation models and next generation development. *Environmental Earth Sciences*, *81*(9), 256. https://doi.org/10.1007/s12665-022-10343-7

Maze, G., Olascoaga, M. J., & Brand, L. (2015). Historical analysis of environmental conditions during Florida Red Tide. *Harmful Algae*, *50*, 1–7. https://doi.org/10.1016/j.hal.2015.10.003

Medina, M., Kaplan, D., Milbrandt, E. C., Tomasko, D., Huffaker, R., & Angelini, C. (2022). Nitrogen-enriched discharges from a highly managed watershed intensify red tide (*Karenia brevis*) blooms in southwest Florida. *Science of The Total Environment*, *827*, 154149. https://doi.org/10.1016/j.scitotenv.2022.154149

Medina, M., Julian, P., Chin, N., & Davis, S. E. (2024). An early-warning forecast model for red tide (*Karenia brevis*) blooms on the southwest coast of Florida. *Harmful Algae*, *139*, 102729. https://doi.org/10.1016/j.hal.2024.102729

Weisberg, R. H., Zheng, L., Liu, Y., Lembke, C., Lenes, J. M., & Walsh, J. J. (2014). Why no red tide was observed on the West Florida Continental Shelf in 2010. *Harmful Algae*, *38*, 119–126. https://doi.org/10.1016/j.hal.2014.04.010