A Scalable Predictive Tool to Identify Vulnerable Estuarine Areas to Harmful Algae Blooms across the Panhandle

Task 5: Final Report

DEP Award Number: INV29

Prepared by:

Dr. Ebrahim Ahmadisharaf

Dr. Ming Ye

Dr. Xiuming Sun

Sumon Hossain Rabby

Date: May 2024

Table of Contents

List of Figures

List of Tables

INV29

Executive Summary

This project aimed to develop a tool for detecting and forecasting harmful algae blooms (HABs) in the Florida panhandle estuarine systems. Empirical understanding of the drivers behind HABs were developed through comprehensive analyses of potential drivers utilizing machine learning (ML) algorithms, literature review, and survey questionnaires of the experts. These assisted us in developing empirical relationships between HABs and their drivers across four selected estuarine systems—Apalachicola, St. Joseph, St. Andrews and Pensacola-Perdido—across the panhandle. The selection was based on the availability of data in these systems in terms of chlorophyll-*a* and the drivers.

Chlorophyll-*a* and cyanobacteria were initially selected as the HAB indicators. However, cyanobacteria data were very sparse; this hindered us from developing an efficient model for this indicator. Subsequently, all model development and analyses were done using chlorophyll-*a* as the HAB indicators. Our model was forced with a variety of water quality, hydrologic and meteorologic inputs. Since an intention was to develop models that are widely applicable in estuarine systems, we examined several model versions and embedded the ones with minimum input data requirements in the final model. The final model requires air temperature, salinity, pH and nutrients (only for one estuarine system).

The project employed ML algorithms, including Random Forest and eXtreme Gradient Boosting (XGBoost), to develop the predictive HAB model. The model was tuned using optimization techniques like bootstrapping and iterative learning to best predict HABs. We validated the model against historical chlorophyll-*a* data in the estuarine systems. The model performance showed satisfactory performance in terms of fit metrics (e.g., R^2 between 0.50 and 0.61) in our validation phase. We developed four system specific models, one for each estuarine system. Our model evaluations found that the models are not transferable across the systems. That is, the validated model for a given system does not perform well in prediction of HABs in another system.

The developed models were embodied in a web-based tool that allows the user to predict HABs in any of the four estuarine systems and identify the vulnerability against HABs under different environmental scenarios such as warmer temperatures and shifting salinity regimes. Nine scenarios, which differed in terms of air temperature, salinity and pH, were evaluated via the tool alongside the historical conditions. Two HAB characteristics were studied: frequency and severity (based on chlorophyll-*a* concentration). Evaluating six hypothetical scenarios showed that pH increases and warmer temperatures increase the frequency and severity of HABs. The frequency of occurrence can increase by up to 90% and the maximum chlorophyll-*a* concentration can exceed 50 μg/L multiple times. Among the four estuarine systems, Pensacola-Perdido was predicted to be the most vulnerable one, while St. Joseph showed the lowest level of vulnerability to HABs.

The predictive models and web-based tool can assist in planning for water pollution control strategies and proactive mitigation of HABs in estuarine systems. Future work should focus on expanding data collection efforts (e.g., using remotely sensed data) to improve the performance of predictive models and expanding the geographic domain of the tool to estuarine systems beyond the panhandle.

1 Project Basic Information

1.1 Project location

Four estuarine systems across the Florida Panhandle—Apalachicola, St. Joseph, St. Andrews and Pensacola-Perdido—were the geographic focus of this project.

1.2 Project background

There have been occurrences of harmful algae blooms (HABs) and the frequency of these events may increase in the future due to environmental change such as warmer temperatures, land cover change (and subsequent nonpoint source pollution of nutrients) and pollution exported by septic tanks. Florida's Blue-Green Algae Task Force (2019) recommends developing technologies that can detect and forecast HABs to enable proactive responses. It is, therefore, necessary to assess the vulnerability of estuarine systems to HABs under current and plausible future conditions. Such vulnerability assessments require analyses of historical observations in algae blooms and their applicable drivers, predictive frameworks, and developing plausible environmental scenarios.

1.3 Project description

Florida State University (FSU) identified key HAB drivers through joint assessments of potential drivers (e.g., nutrients, septic tanks, hydrology, and climate) versus existing HAB observations using machine learning (ML) models, literature reviews, online surveys of regional experts, as well as geospatial and time series analyses. Empirical relationships were developed between HABs and their drivers at selected estuarine areas across the Panhandle using various ML algorithms (e.g., Random Forest). These relationships were used in models that predict HABs based on chlorophyll-a concentration. Quantitative metrics (e.g., R^2) were used to measure the performance of the models. A web-based tool was subsequently implemented based on the developed models. The tool was then applied to identify vulnerable areas to HABs across the panhandle under a diverse range of environmental scenarios such as warmer temperatures and shifting salinity regimes. The methods and results were documented in this final report.

1.4 Original project timeline vs. actual completion timeline

The tasks were completed by the corresponding end date and all deliverables were submitted by the designated due date. It is notable that DEP granted a change order on March 7, 2024 to extend the due dates of Task 5. A summary of the planning and completion dates are shown in Table 1. DEP have already approved all deliverables for Tasks 1-4.

Table 1. Original project timeline vs. actual completion timeline.

Task/ Deliverable No.	Task or Deliverable Title	Task Start Date	Original Task End Date	Actual Task End Date
1	Quality Assurance Manual			
1a	Draft Quality Assurance Manual	7/01/2021	8/22/2022	8/22/2022
1b	Final Quality Assurance Manual	7/01/2021	10/21/2022	11/21/2022
2	Data Collection/Processing, Literature Review, Surveys and Preliminary Analyses			

2a	Interim report	7/01/2021	5/10/2024	1/21/2023
2b	Final report	7/01/2021	5/10/2024	3/31/2023
3	Data-driven Modeling and Tool Development			
3a	Interim report	7/01/2021	5/10/2024	6/30/2023
3b	Final report	7/01/2021	5/10/2024	10/31/2023
4	Verification, Finalization, and Application of the Tool for Vulnerability Assessments	7/01/2021	5/10/2024	2/28/2024
5	Final Report			
5a	Draft Final Report	7/01/2021	3/10/2024	3/18/2024
5b	Final Report	7/01/2021	5/10/2024	6/10/2024

1.5 Project financial summary vs. projected costs

Cost reimbursable grant funding must not exceed the category totals for the project as indicated below. Projected costs during the project period are summarized in Table 2.

Table 2. Project financial summary vs. actual project costs.

Category totals	**Original funding**	**Projected costs***
Salaries	$225,810	$246,979
Fringe	$107,072	$86,634
Miscellaneous/Other Expenses	$800	$200
Overhead/Indirect Cost (10%)	28,916	$28,785
Total:	**$362,598**	**$362,598**

Task No.	**Budget category**	**Original funding**	**Projected costs***
1	Salaries	$12,345	$5,931
	Fringe	$3,561	$1,234
	Overhead/Indirect Cost (10%)	$1,591	$716
	Total for Task:	**$17,497**	**$7,881**
2	Salaries	$23,637	$16,012
	Fringe	$9,029	$7,736
	Miscellaneous/Other Expenses	$200	$200
	Overhead/Indirect Cost (10%)	$2,953	$2,032
	Total for Task:	**$35,819**	**$25,980**
3	Salaries	$88,452	$114,165
	Fringe	$42,725	$35,098
	Miscellaneous/Other Expenses	$400	$0
	Overhead/Indirect Cost (10%)	$11,314	$13,082
	Total for Task:	**$142,891**	**$162,345**
4	Salaries	$72,589	$81,559
	Fringe	$43,426	$34,656
	Miscellaneous/Other Expenses	$200	$0
	Overhead/Indirect Cost (10%)	$9,347	$9,347

	Total for Task:	**$125,562**	**$125,562**
5	Salaries	$28,787	$29,312
	Fringe	$8,331	$7,910
	Overhead/Indirect Cost (10%)	$3,711	$3,608
	Total for Task:	**$40,829**	**$40,830**
	Total for All Tasks:	**$362,598**	**$362,598**

* A change order was submitted for this re-budget and was accepted by DEP.

1.6 Anticipated benefits

Based on the proposed framework, FSU developed data, models, a web-based tool and analyses with the following benefits:

1) A homogenized database of a HAB indicator (chlorophyll-*a* concentration) and pertinent HAB drivers in each of the four estuarine systems.
2) Identified driving factors of HABs in each of the four estuarine systems of the panhandle.
3) Statistical relationships between HAB indicator (chlorophyll-*a* concentration) and environmental features in each estuarine system.
4) Developed ML models to predict chlorophyll-*a* concentrations in the four estuarine systems of the panhandle.
5) A web-based tool that allows the user to predict chlorophyll-*a* concentrations in the four estuarine systems of the panhandle.
6) Assessed the vulnerability of each estuarine system under hypothetical scenarios using the tool.

2 Summary of Data Collection

The growth of algae requires a set of specific environmental conditions that are location specific. We collected information from literature review and conducted expert surveys to acquire relevant information about the important features in the estuarine systems. Based on this information, we collected necessary data to develop and validate the ML models.

First, we collected all the available data for the panhandle area, including Apalachicola Bay, Apalachee-St. Marks, St. Andrews, St. Jospeh, Choctawhatchee Bay and Pensacola and Perdido Bays. The data included water quality, hydrologic, meteorologic, land cover, land surface imperviousness, and locations of septic tanks from different sources. All the collected data were stored in a common folder on SharePoint backed up by Florida State University (FSU) servers. A summary of the data is shown in Table 3. We also acquired historical data of cyanobacteria cell counts from Florida Fish and Wildlife Conservation Commission (FWC). The cyanobacteria cell counts are the most direct and informative data to represent occurrence of HABs. Sixty-two percent of the experts in our online survey questionnaire (conducted under Task 2) suggested that cell counts (cells/L) of algal species can be used as the HAB indicator. The dataset of cell counts included 777 events that took place along the panhandle area at different locations (Figure 1) from 1982 to early 2022 (Figure 2). Table 4 summarizes the collected data.

Among all the estuarine systems, Apalachicola Bay had the most extensive water quality data (e.g., chlorophyll-*a* concentration) across the panhandle as part of continuous monitoring by Apalachicola National Estuarine Research Reserve (ANERR). Data used in our tool development are shown in Tables 5 and 6. Additional data were collected from other estuarine systems to validate our models (Table 7 - Table 10).

Table 3. Parameters and metadata availability for the data collected from different sources.

Parameter(s)		**Brief description**	**Metadata availability**	**Usability in our study**	**Data provider/ source**
Water Quality	Water_Quality_Nutrients_Apalachicola_Bay	Water quality and nutrient data obtained for different stations are stored in different sub-folders named by the station name. The original data are in .csv format. Refer to Table 2 for more details about these data, their period or records and frequency of data. The locations of the stations are detailed in Table 3.	All the metadata provided by the primary data producer are stored in the sub-folders under folder "Metadata_WQ_Nutrients_Apalachicola_Bay". There are several metadata files (in .docx format), all of which are available in the aforementioned folder. The "readme.rtf" file located in the folder is also provided by the primary data producer. Additionally, we included a list called "sampling_stations_ANERR.xlsx" that presents the details of the monitoring stations.	The primary data producer shared only the data passed their standard data checks. We processed part of these data that have better temporal coverage and have conducted our preliminary analysis on the processed data.	Apalachicola National Estuarine Research Reserve
	Cyanobacteria_Entire_Panhandle	A .csv file was obtained that contains all available cyanobacteria cell counts, dates of the records, species of the cyanobacteria, coordinates of the observations and the names of the counties across the coastal-estuarine systems of the entire panhandle. We generated a .kmz file employing the coordinates which is also available under the folder. Refer to Table 3 for more details.	No metadata obtained.	Due to the sparsity of spatial and temporal distribution of the observations, these datasets cannot be used for model development.	Florida Fish and Wildlife Conservation Committee

	Water_Quality_Entire_Panhandle_IWR_Runs63	All available water quality data that includes physical, chemical, biological parameters and pesticides are stored under the "All Available WQ Data Extracted" folder. We also extracted the relevant water quality parameters by excluding the pesticides data and stored them under "Relevant Parameters". All the parameters have unique MasterCodes and the unique dataset (in .csv format) for each parameter is named by the MasterCode. The .csv file contains the station IDs, dates of sampling	Details of the MasterCodes that represent the parameters including the units of measurements are listed in "Parameter MasterCodes List.csv". The details of the stations including the waterbody ID (WBID) and locations for which IWR records the data are presented in "IWR_Stations_WBID_List_Panhandle.csv". No other metadata were obtained.	These datasets will be used further for validation of the proposed model once it is developed in Task 3-4.	Impaired Water Rules (IWR) Runs 63
Meteorologic	Apalachicola East Bay Meteorology (EB_met)	All available meteorologic data obtained from the primary data producer are stored as .csv files. The data are available for only one station at Apalachicola Bay. The location of the station is shown in Table 3 and the details of the data are presented in Table 2.	Similar to the water quality and nutrients data for Apalachicola Bay, there were several metadata files (.docx) provided by the primary data producer, all of which are stored under "Metadata_Meteorology_Apalachicola_Bay".	We processed and included these data in our preliminary analyses.	Apalachicola National Estuarine Research Reserve
Septic Tanks	N/A	The locations of septic tanks for the entire panhandle are available in .shp format.	No metadata obtained.	The data will be used in our further analyses	Florida Department of Health
Hydrologic	Discharge Gage Height (Water Level)	Data available for 17 selected USGS stream gauges are stored.	The locations and details of the stations are available in "17_selected_USGSStreamGage_locations.xlsx" file.	Data for one station near Apalachicola Bay were included in our preliminary analyses.	USGS

Imperviousness_and_Land_Cover	The folder consists of a ArcGIS project file (.aprx) as well as a geodatabase file named "HAB.gdb" that contains the land cover and imperviousness for different years extracted from the national land cover dataset website and clipped for the entire panhandle area. For more information, we refer to Table 4.	No metadata was obtained.	The data will be used in our further analyses.	USGS
Processed_Data_3stations_Apalachicola_Bay	We processed the data for the three stations in Apalachicola Bay and compiled the water quality and nutrient datasets with the meteorologic and hydrologic datasets obtained for nearby stations (one for meteorology by ANERR and one for hydrology by USGS). For additional details, refer to section 1 in Task 2 report. Processed and compiled datasets for each station are available as .csv formats.	N/A	Our main tool development will be based on these datasets.	N/A

Table 4. Cyanobacteria cell-count data collected for the estuaries across the panhandle.

Data	**Description**	**Source**	**Notes**
Cyanobacteria cell counts	• Period of record: September 1982 - April 2022 • Sampling frequency: Irregular (777 records for the entire panhandle) • Spatial coverage: Bays, estuaries and nearshore coastal areas of the panhandle • Contains abundance (cells/L), taxa, county name, location (latitude, longitude), sampling date and sampling depth (m)	Florida Fish and Wildlife Conservation (FWC)	No systematic monitoring and no time series were found. A total of 777 observations were found, out of which 85 records did not contain abundance (cells/L).
All other water quality parameters	• All available data extracted and stored parameter-wise for all the stations in entire panhandle. • Parameter includes nutrients, physical water quality parameters, metals, microbiological (E. coli) and different biotoxins. • Sampling period and frequency are very sparse as well as vary for different parameters and different stations.	Impaired Water Rules (Run 63)	Not regular time series for any parameters at a single station.
Hydrologic (discharge and gauge height)	• Period of record: 2002 - 2021 • Frequency/temporal resolution: Daily • Spatial coverage: 16 stream gauges selected based on their locations near to the downstream, that record the stream flows coming into the major bays and estuaries from the upstream rivers.	USGS	Regularly monitored time series
Land cover	Available for years 2004, 2006, 2008, 2011, 2013, 2016 and 2019	National Land Cover Dataset – USGS (2019 release)	N/A

Land surface imperviousness	Available for years 2001, 2006, 2008, 2011, 2013 and 2019	National Land Cover Dataset – USGS (2019 release)	N/A
Septic tanks	Spatial distribution of the septic tanks	Florida Department of Health	N/A

Table 5 Data collected for the Apalachicola Bay.

Data	**Frequency of sampling or temporal resolution**	**Number of locations / stations**	**Period of records obtained**	**Source**
Physical water quality parameters: pH, turbidity, dissolved oxygen, water temperature, salinity, specific conductivity and depth of water samples	30 minutes	5	January 2002-December 2021 for three stations—Cat Point (CP), Dry Bar (DB) and East Bay (EB). Only the EB station has data for both surface (EB_s) and bottom (EB_b) waters samples. January 2017 - December 2021 for the two other stations—Pilot's Cove (PC) and Little St. Marks (LM).	Apalachicola National Estuarine Research Reserve
Nutrients: Phosphate, nitrite + nitrate, ammonium and chlorophyll-a	Once a month with irregular sampling intervals	10	April 2002 - December 2021 for 10 stations. Only one station (EB) has data for both surface and bottom waters samples.	Apalachicola National Estuarine Research Reserve

Meteorologic variables: Air temperature, total precipitation, wind speed, maximum wind speed, wind direction, standard deviation in wind direction and total photosynthetically active radiation	15 minutes	One meteorologic station at East Bay (EB_met).	January 2002 - December 2021	Apalachicola National Estuarine Research Reserve
Hydrologic variable: Discharge and water level	Daily	One nearby stream gauge (Apalachicola River near Sumatra, AR)	January 2002 - December 2021	USGS

Table 6. A list of the nutrient, physical water quality parameters, meteorologic and hydrologic monitoring stations in the Apalachicola Bay.

Station name	Station ID	Station code	Latitude (˚)	Longitude (˚)	Parameters
Pilot's Cove	PC	apapcnut	29.61	-85.02	Nutrients, physical water quality, and chlorophyll-*a*
East Bay-Surface	Es	apaesnut	29.79	-84.88	Nutrients, physical water quality, and chlorophyll-*a*
East Bay-Bottom	Eb	apaebnut	29.79	-84.88	Nutrients, physical water quality, and chlorophyll-*a*
East Bay-Bridge	Eg	apaegnut	29.73	-84.95	Nutrients, physical water quality, and chlorophyll-*a*
Nick Hole	NH	apanhnut	29.65	-84.93	Nutrients, physical water quality, and chlorophyll-*a*
Cat Point	CP	apacpnut	29.70	-84.88	Nutrients, physical water quality, and chlorophyll-*a*
West Pass	WP	apawpnut	29.64	-85.09	Nutrients, physical water quality, and chlorophyll-*a*
Dry Bar	DB	apadbnut	29.67	-85.06	Nutrients, physical water quality, and chlorophyll-*a*
Mid Bay	MB	apambnut	29.67	-84.99	Nutrients, physical water quality, and chlorophyll-*a*
Apalachicola River	AR	aparvnut	29.78	-85.04	Nutrients, physical water quality, and chlorophyll-*a*
Sike's Cut	SC	apascnut	29.79	-84.88	Nutrients, physical water quality, and chlorophyll-*a*
East Bay-Meteorological Stations (NOAA)	Eb_Met	apaebmet	29.77	-84.88	Meteorological variables

Table 7. Statistical description of the dataset used for optimal machine learning model for Apalachicola Bay (Data source: ANERR & NOAA).

Parameter	Count	Mean	Std.	Min	25%	50%	75%	Max
Salinity (ppt)	875	16.6	11.5	0.0	5.8	16.6	27.1	38.0
Turbidity (NTU)	875	10.7	9.8	0.0	5.2	7.8	12.4	101.6
DO (mg/l)	875	8.3	9.2	3.0	6.2	7.3	8.5	104.4
pH	875	8.0	0.3	6.3	7.8	8.0	8.2	8.5
Air temperature (°C)	875	21.7	6.8	5.0	16.0	21.4	28.2	44.7
Daily maximum air temperature (°C)	875	24.7	7.2	7.9	19.4	25.5	30.4	60.7
Daily maximum air temperature 1 day ago (°C)	875	23.6	9.1	-40.0	19.1	24.4	30.2	40.9
Daily maximum air temperature 2 days ago (°C)	875	23.4	9.2	-40.0	19.0	24.9	30.2	43.2
Daily maximum air temperature 3 days ago (°C)	875	23.7	8.7	-41.8	20.2	24.4	29.9	38.7
Daily maximum air temperature 4 days ago (°C)	875	24.0	6.5	-44.4	19.5	24.2	29.4	38.2
Daily maximum air temperature 5 days ago (°C)	875	24.1	6.6	-44.8	19.6	25.4	29.0	50.2
Daily maximum air temperature 6 days ago (°C)	875	24.0	7.4	-44.6	19.9	24.8	29.0	57.8
Daily maximum air temperature 7 days ago (°C)	875	23.8	7.0	-37.6	19.4	24.5	29.2	51.0
Chlorophyll-*a* (ug/L)	875	6.7	5.2	0.4	2.9	5.4	8.8	32.2

Table 8. Statistical description of the dataset used for optimal machine learning model for St. Joseph Bay (Data source: WIN & NOAA).

Parameter	Count	Mean	Std.	Min	25%	50%	75%	Max
Salinity (ppt)	111	28.5	2.5	17.9	27.9	29.1	29.7	31.8
Turbidity (NTU)	111	1.4	1.3	0.3	0.7	1.0	1.5	7.8
DO (mg/l)	111	7.6	1.1	4.7	6.7	7.6	8.3	10.3
pH	111	8.1	0.2	7.1	8.0	8.1	8.1	9.2
Air temperature (°C)	111	21.0	6.3	10.4	15.7	21.3	27.3	32.3
Daily maximum air temperature (°C)	111	22.6	12.8	-17.8	17.2	26.7	31.7	37.2
Daily maximum air temperature 1 day ago (°C)	111	18.5	16.8	-17.8	15.0	25.0	29.2	35.6
Daily maximum air temperature 2 days ago (°C)	111	22.9	12.3	-17.8	19.4	27.2	30.3	35.0
Daily maximum air temperature 3 days ago (°C)	111	23.2	12.1	-17.8	21.7	26.1	30.6	36.1
Daily maximum air temperature 4 days ago (°C)	111	23.0	13.1	-17.8	22.2	26.7	30.3	35.0

Daily maximum air temperature 5 days ago (°C)	111	21.2	14.1	-17.8	21.1	25.0	28.9	35.0
Daily maximum air temperature 6 days ago (°C)	111	22.8	12.2	-17.8	21.7	25.0	30.3	35.6
Daily maximum air temperature 7 days ago (°C)	111	21.7	14.2	-17.8	20.6	24.4	31.1	33.3
Chlorophyll-*a* (ug/L)	111	4.1	2.7	0.8	2.2	3.2	5.3	13.0

Table 9. Statistical description of the dataset used for optimal machine learning model for St. Andrews Bay (Data source: EPA STORET & NOAA).

Parameter	**Count**	**Mean**	**Std.**	**Min**	**25%**	**50%**	**75%**	**Max**
Salinity (ppt)	110	18.6	12.3	0.0	6.1	22.5	29.5	35.0
Turbidity (NTU)	110	5.4	7.0	0.4	2.1	3.6	5.7	47.0
DO (mg/l)	110	5.6	2.1	0.7	4.4	5.6	7.2	9.8
pH	110	7.3	0.7	4.8	7.0	7.5	7.8	8.3
Temperature (°C)	110	24.3	5.9	11.2	19.6	25.3	29.2	36.1
Daily maximum air temperature (°C)	110	28.1	6.1	15.0	24.4	30.6	32.8	36.7
Daily maximum air temperature 1 day ago (°C)	110	27.8	6.0	11.1	24.2	30.6	31.7	35.6
Daily maximum air temperature 2 days ago (°C)	110	28.3	5.7	15.0	24.4	30.8	32.8	37.8
Daily maximum air temperature 3 days ago (°C)	110	27.5	6.9	8.3	22.8	30.6	32.8	35.6
Daily maximum air temperature 4 days ago (°C)	110	27.2	8.5	7.8	20.6	31.1	33.9	36.1
Daily maximum air temperature 5 days ago (°C)	110	26.6	8.8	5.0	20.6	30.8	32.8	35.0
Daily maximum air temperature 6 days ago (°C)	110	27.4	7.3	5.0	24.4	30.6	32.2	35.6
Daily maximum air temperature 7 days ago (°C)	110	28.4	6.8	13.3	25.0	30.6	32.6	38.9
Chlorophyll-*a* (ug/L)	110	6.0	11.6	0.7	1.1	3.2	5.3	96.0

Table 10 Statistical description of the dataset used for optimal machine learning model for Pensacola-Perdido Bay (Data source: STORET & NOAA)

Parameter	**Count**	**Mean**	**Std.**	**Min**	**25%**	**50%**	**75%**	**Max**
Salinity (ppt)	273	7.7	8.8	0.0	0.1	2.9	15.8	29.0
Turbidity (NTU)	273	7.2	23.2	0.9	3.0	4.0	6.0	350.0
DO (mg/l)	273	6.6	2.9	-0.9	4.9	6.9	8.5	18.4
pH	273	6.7	0.9	3.5	6.3	6.7	7.2	8.5
Air temperature (°C)	273	22.5	6.1	8.7	16.9	23.5	27.9	33.6
Daily maximum air temperature (°C)	273	26.0	7.0	9.4	20.6	29.4	31.1	36.7
Daily maximum air temperature 1 day ago (°C)	273	25.9	6.6	7.2	20.6	28.9	31.1	37.8
Daily maximum air temperature 2 days ago (°C)	273	25.2	7.0	9.4	17.8	28.3	30.6	37.8
Daily maximum air temperature 3 days ago (°C)	273	25.9	5.9	10.0	21.7	27.8	30.6	37.2
Daily maximum air temperature 4 days ago (°C)	273	25.8	6.2	7.8	21.7	26.7	31.1	35.6
Daily maximum air temperature 5 days ago (°C)	273	26.5	6.0	10.6	24.4	28.9	31.1	35.6
Daily maximum air temperature 6 days ago (°C)	273	25.9	5.7	9.4	21.7	27.2	30.6	35.0
Daily maximum air temperature 7 days ago (°C)	273	26.3	5.6	11.1	22.2	28.3	31.1	36.7
Chlorophyll-*a* (ug/L)	273	10.1	19.0	0.6	5.0	5.0	8.5	194.0

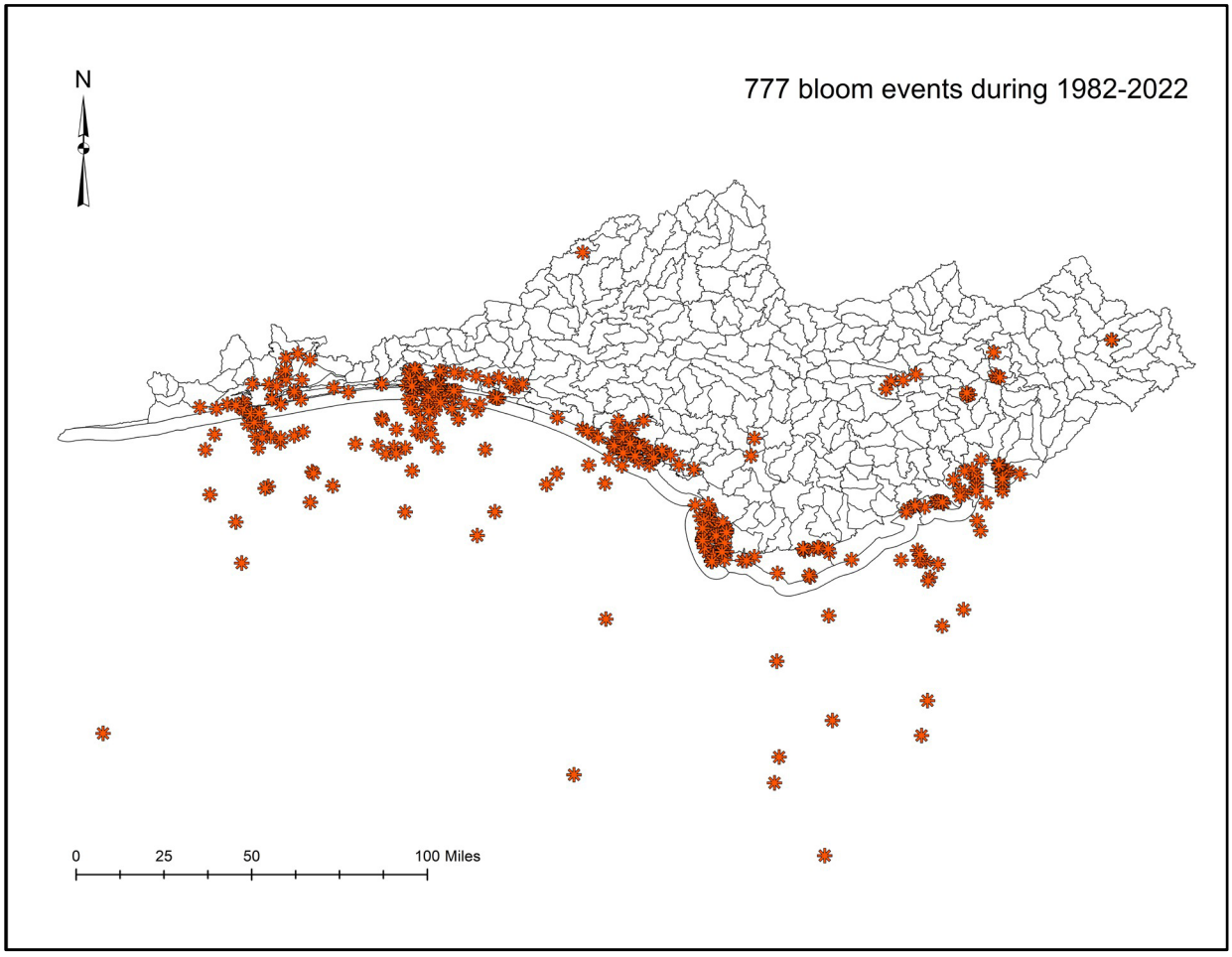

Figure 1. Spatial distribution of HAB events across the panhandle area from 1982 to 2022.

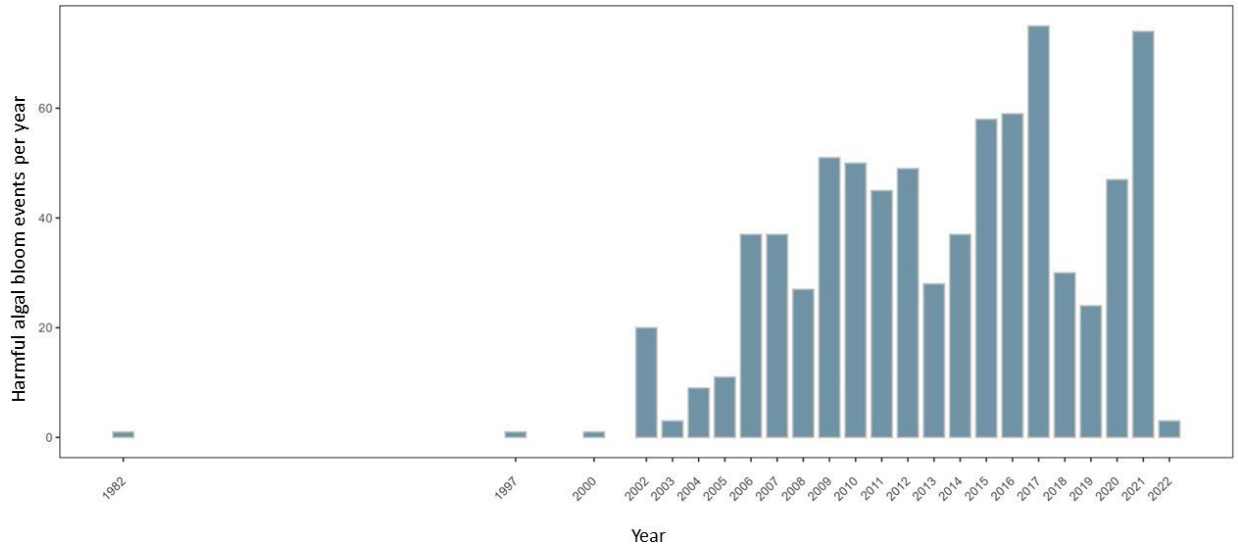

Figure 2. Historical HAB events took place across the panhandle between 1982 and 2022. No data

was reported for the period of 1982-1997.

3 Methods for Tool Development

3.1 Model approach description

In our project, we applied time series analyses, geospatial analyses and ML algorithms. Previous research (e.g., Xie et al., 2012; Zhu et al., 2020) has proved their feasibility and capability in analyzing and predicting HABs and more broadly water quality. The ML algorithms are available through open-source packages that users can access by R or python programming languages. The ML models can handle large and noisy ecological datasets, adapt to dynamic data, detect complex patterns, and continuous learning (Pichler et al., 2023). Another important advantage of ML models is the possibility to implicitly model complex nonlinear relationships between predictors and target variables without relying on explicit mechanistic knowledge (Cruz et al., 2021; Yu et al., 2021). Two main regression algorithms were used in our project: Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). The best model with optimal selected features was used to assess the transferability of the model to the other estuarine systems.

By applying RF, we bootstrapped samples of D_j of size N from training set to train individual trees. Bootstrapping is an iterative process, in which, in each iteration the samples can be replaced using random choices of bootstrapping indices utilizing numpy library in python. For example, three (size N here) datapoints x, y and z can be selected randomly as the following combinations: xyz, xxy, xyy, xzz, yyz, yzz, and xzz. In each of these combinations, sample size is three (D_j). We trained the individual tree with each of the bootstrapped samples and calculated their mean prediction. The sklearn library function for the random forest regression in Python 3 was used in our predictions. The bootstrapping technique potentially improves the model generalizability in predictions of unseen data (out-of-sample).

The regular library function in Python sklearn module was used to apply XGboost algorithm. No bootstrapping was done for the XGboost model. The model grows in the number of trees (weak/base learners) in a sequential manner iteratively by analyzing the errors in the previous model and correcting the errors by adding more trees. After training the final models, each model was iterated 100 times, and the average performance was recorded to compare their performances.

3.2 Target variables

3.2.1 Cyanobacteria biomass as target variable

Firstly, we used the historical HAB observations of cyanobacteria cell counts as individual events and developed an ML model. For each taxon, there is a range of their morphological characteristics (e.g., the range of their length, width, and height). This dimension information can be used to calculate the biovolume of the taxon. However, in the HAB dataset provided to us by FWC, the dimension information was missing; therefore, we used average values from the literature to estimate the biovolume of the taxa. For the taxon that was not identified to a species level, an average of the identified taxon was used. Equation 1 shows how to calculate the biomass:

$$\text{Biomass (mg/L)} = \text{Cell counts (cells/L)} * \text{Biovolume } (\mu m^3/\text{cell}) * \text{Cell density } (\text{mg}/\mu m^3) \quad \text{Eq. 1}$$

Since the biomass distributed within a wide range, we used log transformation to normalize the target data. This model can predict cell counts and analyze the potential drivers of HABs. However, due to the high sparsity of the HAB data, we could not develop estuarine-specific HAB models for

each singe bay-estuarine area (i.e., one prediction model for Apalachicola and another for Pensacola). That said, we developed one ML model for the entire panhandle estuaries based on the entire historical cell count dataset (i.e., all events in Table 4). Details of the model results were documented in section 3.4.1. In addition, due to the limitation of data availability of cell counts, this model was not used for further validation and transferability to other estuarine systems.

3.2.2 Chlorophyll-*a* concentration as target variable

Due to the limited size of the cell-count data, we developed our HAB prediction tool by using chlorophyll-*a* as the target variable, which was deemed acceptable based on the literature review and experts' opinion in our survey questionnaire (Task 2).

Preliminarily, we planned to develop the models for the panhandle area, including Apalachicola Bay, Apalachee-St. Marks, St. Andrews, St. Jospeh, Choctawhatchee Bay and Pensacola and Perdido Bays. However, the data of Choctawhatchee Bay and Apalachee-St. Marks did not consist of chlorophyll-*a* concentration, which means we could not have target variable for this bay-estuary system. Therefore, we developed the ML models for the other four estuarine systems. Apalachicola Bay had the most extensive water quality data (e.g., chlorophyll-*a*) among the systems as part of continuous monitoring by Apalachicola National Estuarine Research Reserve (ANERR). Therefore, and for brevity reasons, we provide detailed description of the model development for Apalachicola Bay as the representative system. Model validation results are provided for each system in detail in Section 3.5.

3.3 Feature selection

3.3.1 Feature selection for Cyanobacteria biomass models

Based on the empirical experience, we classified the factors that may influence cyanobacteria biomass into four main categories of land use/land cover (LULC), meteorologic, hydrologic, and septic tank density.

The historical LULC data was extracted from the national land cover dataset (NLCD). We used this national data over the Florida land use map since the Florida classification over the historical years (1982-2022) is not consistent. The NLCD data had 15 classes of open water, open space (developed land), low intensity developed land, medium intensity developed land, high intensity developed land, barren land, deciduous forest, evergreen forest, mixed Forest, shrub scrub, herbaceous, hay pasture, cultivated crops, woody wetlands, emergent herbaceous wetlands. In addition to these LULC data, we extracted land surface imperviousness data (in %) as it explains pertinent hydrologic processes and can influence on the cyanobacteria biomass. The LULC and imperviousness data was assigned to the upstream hydrologic unit code 12 (HUC12) watershed of each of the historical bloom events (Figure 3).

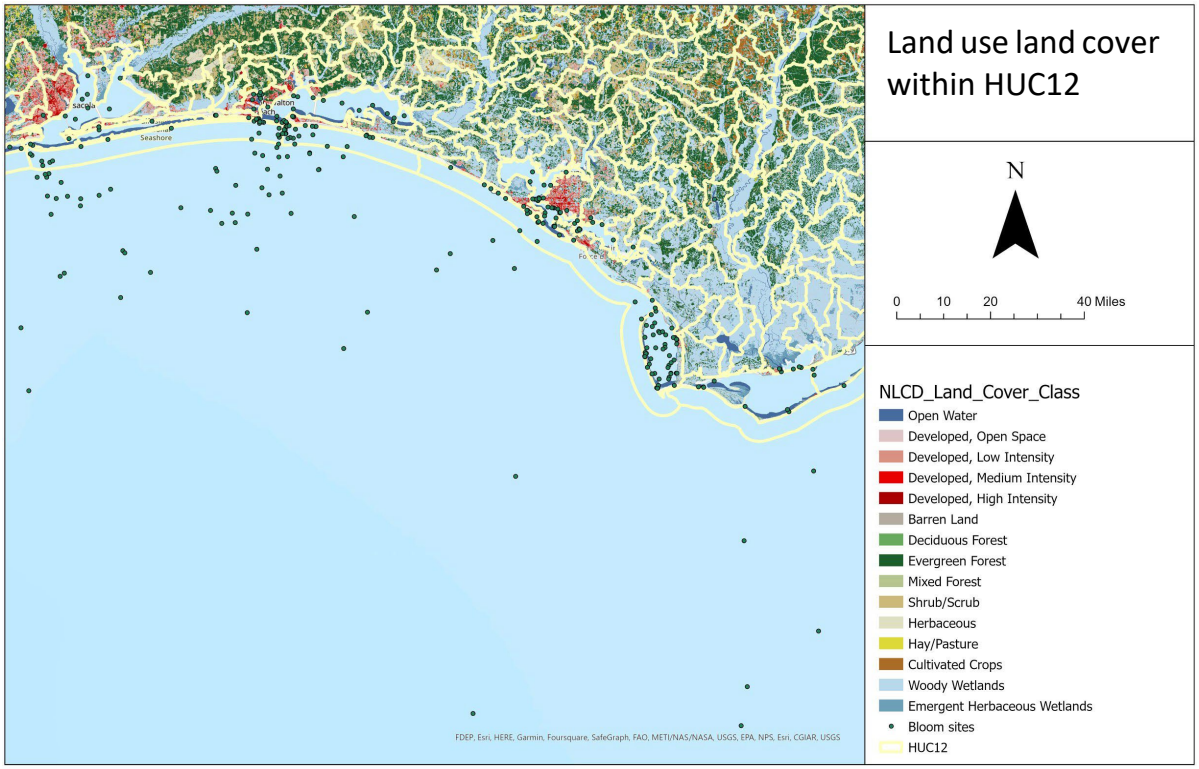

Figure 3. Land use/land cover distribution across HUC12 watersheds of the Florida panhandle.

Meteorological data were obtained for the day and location (latitude and longitude) that each historical bloom event occurred. Data included daily precipitation (mm), maximum temperature (Celsius degree), minimum temperature (Celsius degree), daylight duration (seconds/day) and solar radiation (W/m^2).

We acquired observed streamflow and water level recorded from upstream stream gauges as the hydrologic predictors. For consistency with the other LULC and meteorologic drivers, we used the stream gauges from United States Geological Survey (USGS) in the upstream HUC12 watersheds, where bloom events occurred. However, there were only a few stream gauges that fall inside the HUC12 watersheds; many bloom events did not have a stream gauge in the upstream HUC12 watershed. Therefore, we used the stream gauges within the upstream HUC8 watersheds, in which the historical bloom events occurred. If there were more than one gauge in the HUC8, the closest gauge upstream was chosen to obtain the hydrologic data (Figure 4). Since the stream gauges are in the upstream watersheds, there might be lag effects of the hydrologic factors. Thus, we included data from previous days (1-7, 14 or 30 days) to consider the lag effects.

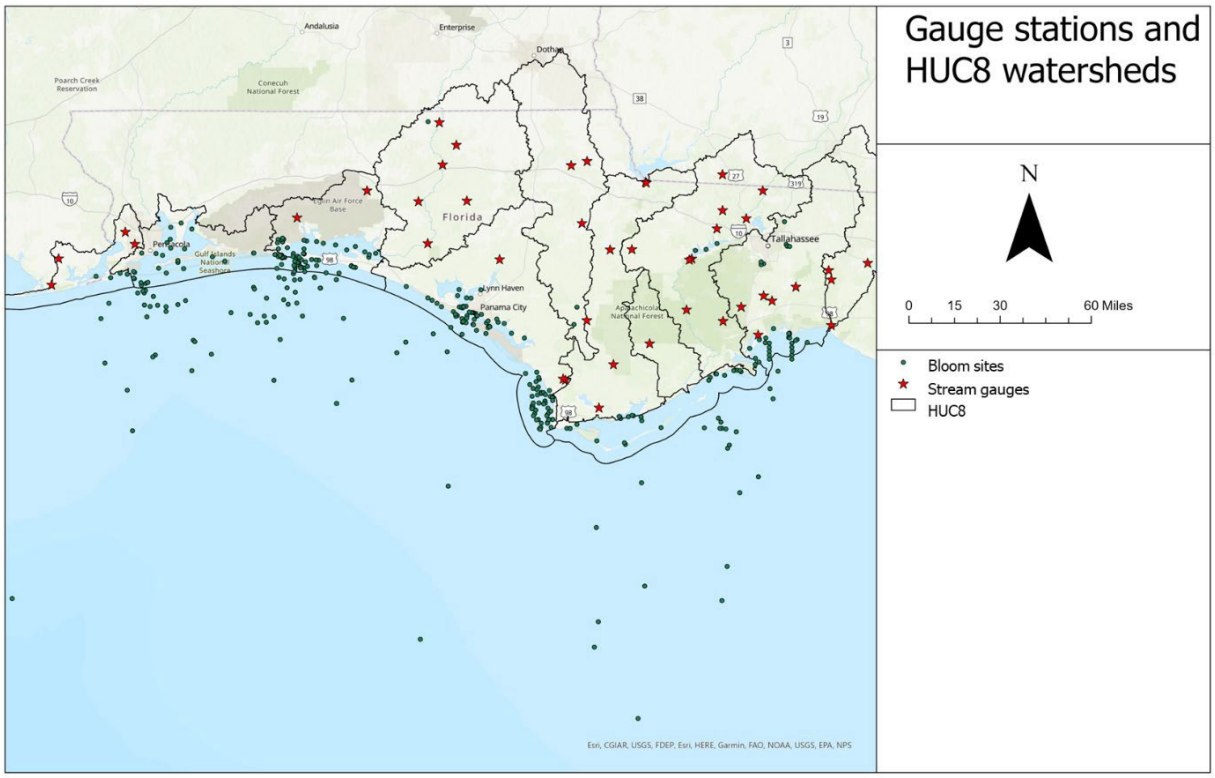

Figure 4. Stream gauges in the upstream HUC8 watersheds of the historical bloom events.

Septic tank density was considered as another predictor. The data was acquired from the Florida Department of Health as part of Task 2. The number of septic tanks for the upstream HUC12 watershed, consistent with LULC and meteorologic data, was calculated. We then divided this number by the upstream watershed drainage area to estimate the septic tank density.

Before we ran the model, statistical analyses were conducted on the target and predictor data to ensure an efficient selection of the predictors. This was done because high correlations among the predictors would increase the risk of overfitting in the model. The Pearson's linear correlation coefficient was used to detect any linear relationships among the predictors. From our analyses on the LULC data, different classes of 'Developed' land use had high linear correlations (> 0.80) (Figure 5); therefore, they were merged as one class of 'Developed land'. Similar analyses were conducted for land surface imperviousness (Figure 6) and hydrologic variables (Figure 7).

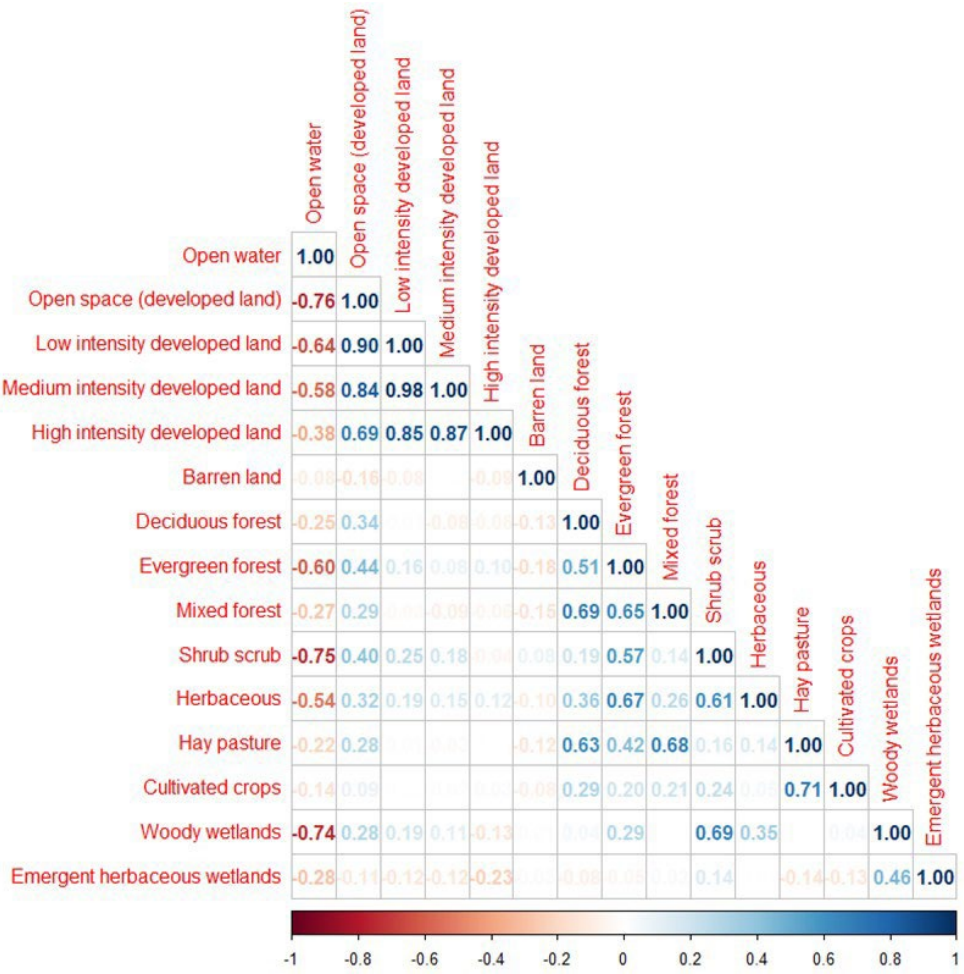

Figure 5. Pearson's correlation coefficients between the land use/land cover (LULC) variables.

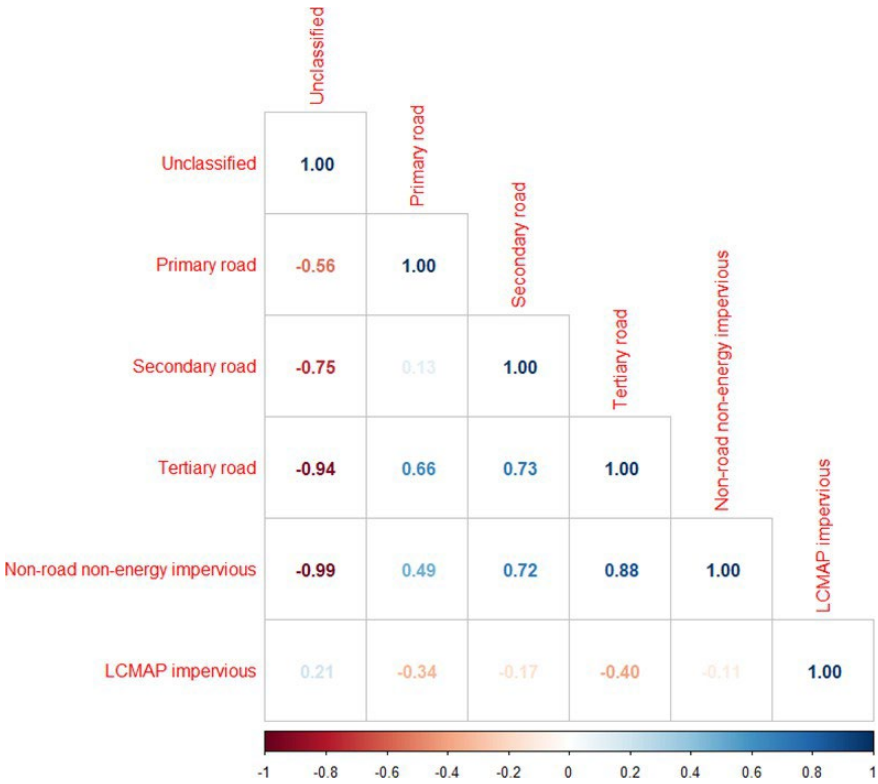

Figure 6. Pearson's correlation coefficients between the land surface imperviousness variables.

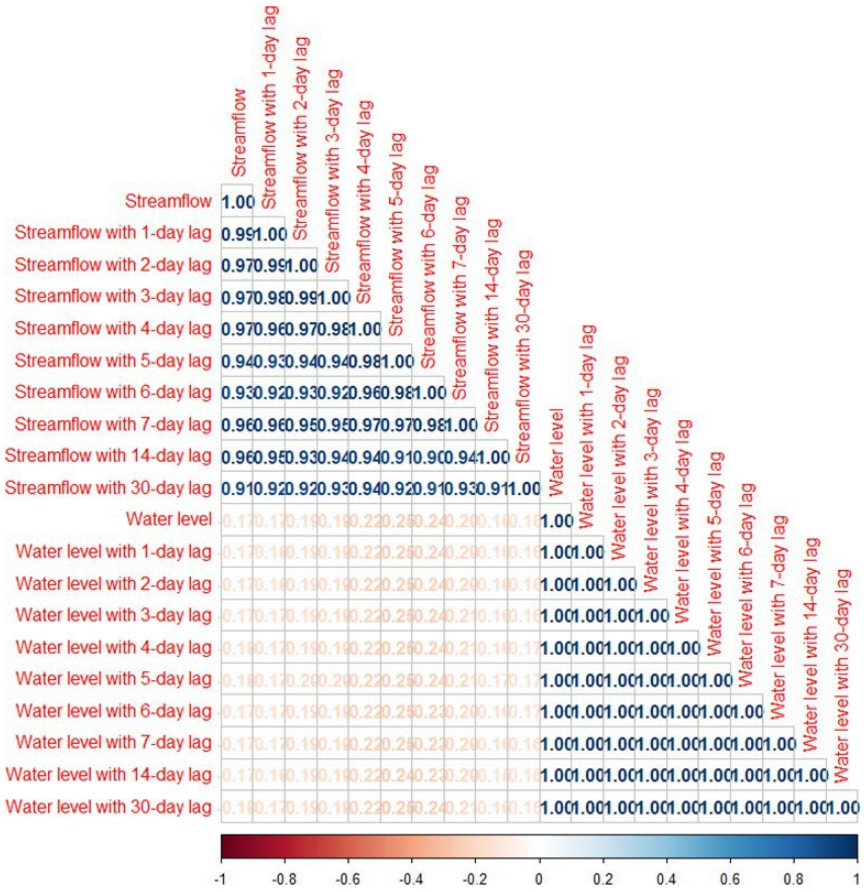

Figure 7. Pearson's correlation coefficients between the hydrologic variables.

3.3.2 Feature selection for chlorophyll-*a* concentration models

As mentioned before, Apalachicola Bay had the most extensive water quality data (e.g., chlorophyll-a) among the systems and for brevity reasons, we provide detailed description of the model development here for Apalachicola Bay as the representative system. The transferability of the models will be shown in the following section 3.5. We conducted a set of exploratory analyses on the final datasets we used to develop the ML models, including the distribution analysis that provided with an idea of the data quality (Figure 8). We observed that no water quality parameters follow a normal distribution, which adds greater challenge in choosing predictive models. We also analyzed the mutual correlationships between the water quality parameters and meteorologic variables (Figure 9) as well as used empirical knowledge to identify the redundant features. A primary investigation with Shapely Additive Explanation (SHAP) used with a non-optimized random forest model enabled us to identify the features that are least important in predicting chlorophyll-*a* concentrations (Figure 10).

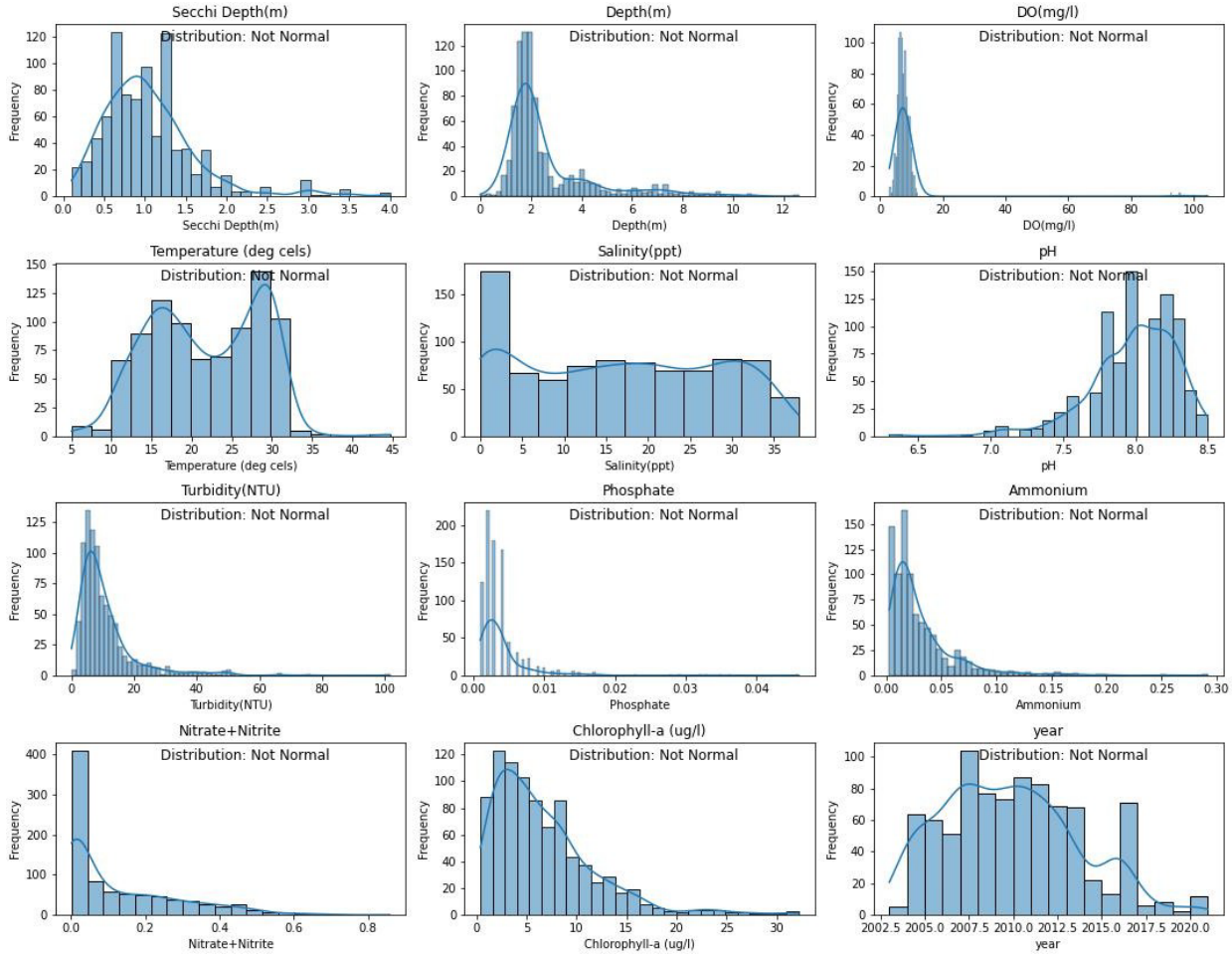

Figure 8. Distribution of water quality data in the Apalachicola Bay.

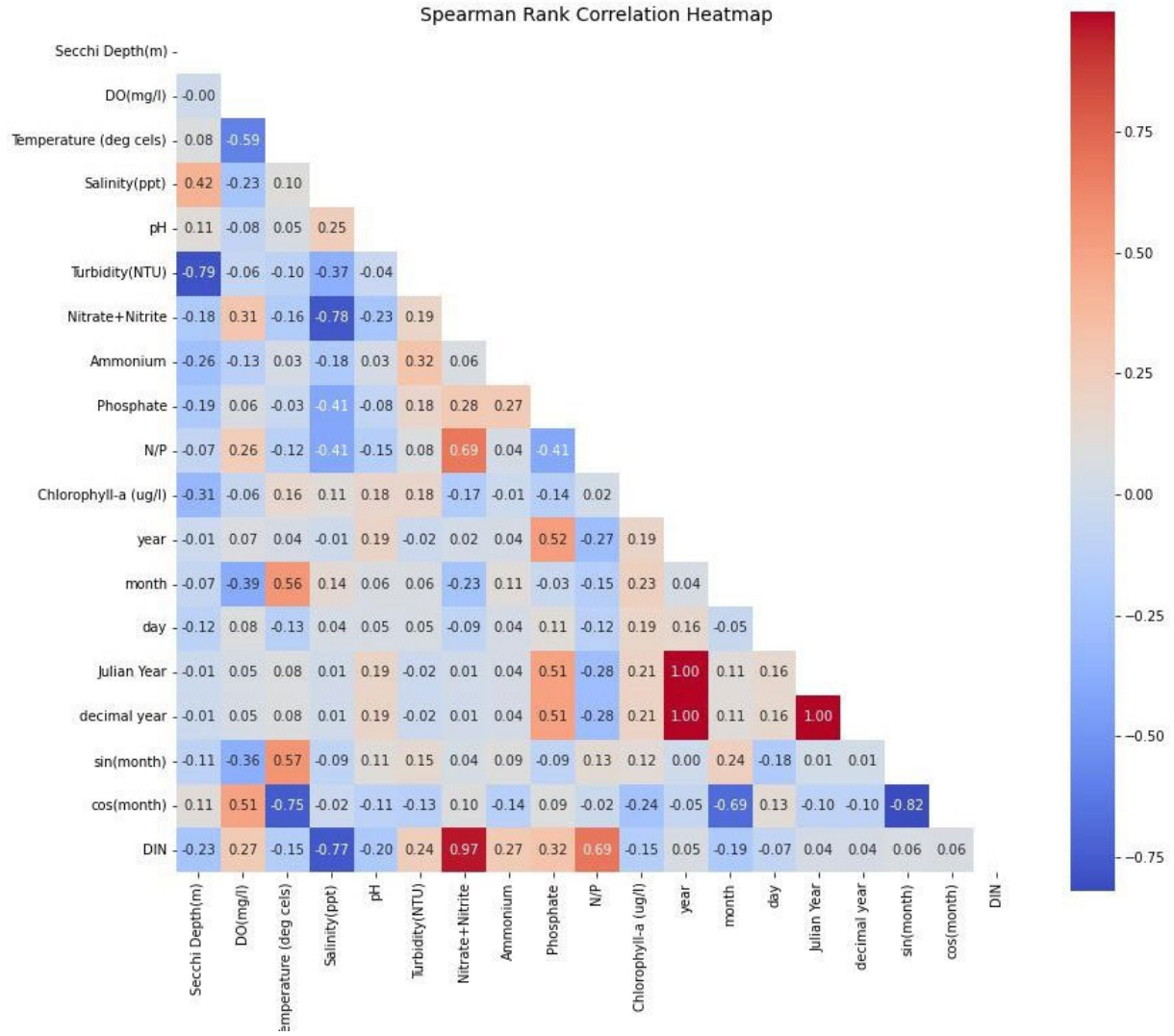

Figure 9. Pearson's correlation coefficients between the water quality parameters and time variables in the Apalachicola Bay.

Figure 10. Shapely Additive Explanation (SHAP) feature importance on a Random Forest regression's prediction of chlorophyll-*a* concentration in the Apalachicola Bay.

All the redundant features were excluded to reduce the dimensionality of the dataset and to lessen the ML model over-complicacy. In summary, meterologic variables, except for the lags of maximum air temperature showed less importance in predicting chlorophyll-*a* concentrations. Given that some of the water quality or meteorologic data may not be available in some cases, three scenarios, which represent three distinct models in predicting chlorophyll-*a* concentration (target variable), were prposed: (1) Only using water quality parameters; (2) only using meteorologic variables; and (3) using both water quality parameters and meteorologic variables. The following equations summarize the inputs in these models, all of which predict chlorophyll-*a* (chl-*a*).

$$Chl\text{-}a = ff(XX1, XX2, XX3, \ldots\ldots\ldots\ldots)$$

Model 1:

$XX, XX, XX, \ldots$ = SSSSSSSShii dSSeeeeh, SSSSSSiiSSiieeSS, TTTTTTTTiiTTiieeSS, TTSSTTeeSSTTSSeeTTTTSS, DDDD, NNiieeTTSSeeSS+NNiieeTTiieeSS, *Phosphate, Nitrogen/phosphorus, and Julian year*

Model 2:

XX1, XX2, XX3, … = Daily maximum, minimum and mean of 'Air temperature, Relative Humidity, Barometric pressure, Wind speed, Wind direction, Standard deviation in wind direction, Total photosynthetically active radiation, and Total precipitation' as well as all these features' lags up to previous seven days.

Model 3:

XX1, *XX2*, *XX3*, … = *SSSSSSSShii* dSSeeeeh, *SSSSSSiiSSiieeSS*, *TTTTTTTTiiTTiieeSS*, Water tSSTTeeSSTTSSeeTTTTSS, *DDDD*, *NNiieeTTiieeSS+NNiieeTTSSeeSS*, Phosphate, Nitrogen/phosphorus, Julian Year, Maximum air temperature, and its lags up to previous seven days.

3.4 Model performance

3.4.1 ML models using cyanobacteria biomass as target variable

We used random forest as the ML algorithm to estimate the cyanobacteria biomass. For the technical implementation, *randomForest* package in R was utilized. The random forest model was run initially with 47 predictors to identify the hydrologic factors with a higher importance. The results showed that streamflow with a 7-day lag and water level with a 3-day lag showed the highest importance among the hydrologic variables. These two were selected as the hydrologic variables in the final model implementation. A total of 26 variables were used as predictors in the second run of the random forest model. The whole dataset (529 data points) was split into train (70%) and test (30%) randomly. The random forest could successfully predict the biomass of cyanobacteria with an R^2 of approximately 0.38 and root mean squared error (RMSE) at around 1.18 (mg/L; Table 11). For the test dataset, the model can predict the target variable with a R^2 of approximately 0.2 and RMSE around 2.45 (mg/L). Due to the uncertainty in estimating the biomass and the fact that cell growth of cyanobacteria in water is easily influenced by changes in the water (nutrients, water temperature, wind etc.), the performance of our model is relatively satisfactory. The model performance can improve with the growth of data amount.

Table 11. Model performance of the ML model for the entire Florida panhandle using cyanobacteria biomass as the target variable.

Model	**Model performance indicator**	
	R^2	**RMSE**
Train data	0.38	1.18
Test data	0.2	2.45

RMSE: root mean squared error.

3.4.2 ML models using chlorophyll-a concentration as target variable

Three optimization algorithms were used to tune the hyperparameters and find the optimal ML models: Bayesian Optimization, Grid Search and Randomized Search. The best performed model will be used to other bay systems to test the transferability of our model.

Among the three models tested, Model 2 with a set of sole meteorological parameters performed not well enough to be considered for further analysis. Model 1 and model 3 showed satisfactory results. For model 1 (using only water quality parameters as the predictors), the XGboost model

tuned with Bayesian Optimization algorithm was found to result in the highest accuracy (highest R^2 and lowest RMSE) among all (Figure 11). The R^2 was 0.64 and RMSE was 3.05 (ug/l). Other evaluation metrics are shown in Table 12.

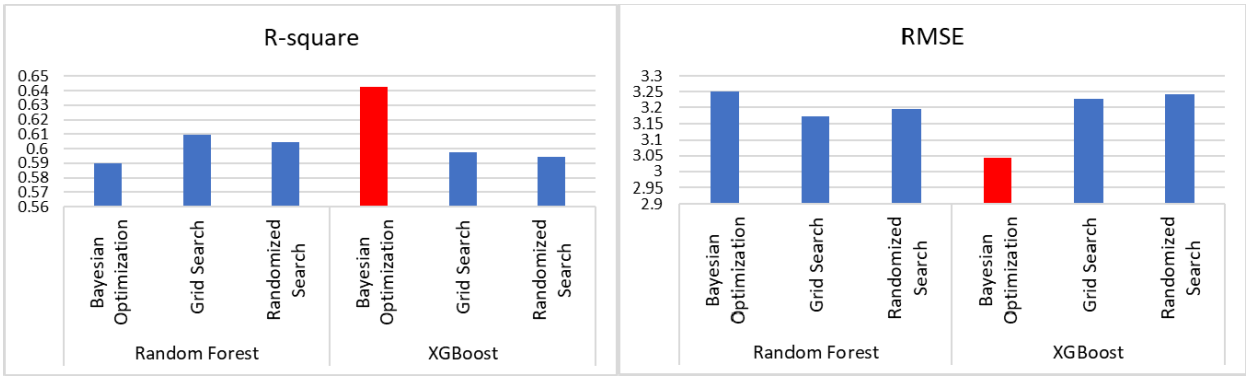

Figure 11. Performance comparisons for Model 1 (using only water quality parameters) in the Apalachicola Bay.

Table 12. Performance evaluation metrics for Model 1 (using only water quality parameters) in the Apalachicola Bay. Blue text specifies the best model.

ML algorithm	Optimization algorithm	R^2	RMSE (ug/l)	MAE (ug/l)	PBIAS (%)
Random Forest	**Bayesian Optimization**	0.59	3.25	2.26	-39.72
	Grid Search	0.61	3.17	2.26	-40.67
	Randomized Search	0.60	3.20	2.28	-39.48
XGBoost	Bayesian Optimization	0.64	3.04	2.26	-35.15
	Grid Search	0.60	3.23	2.34	-33.76
	Randomized Search	0.59	3.24	1.46	-39.45

For Model 3, the inclusion of the lags of maximum air temperature as additional features with water quality parameters resulted in the optimization algorithms obtaining the less complex models performing better (Figure 12).

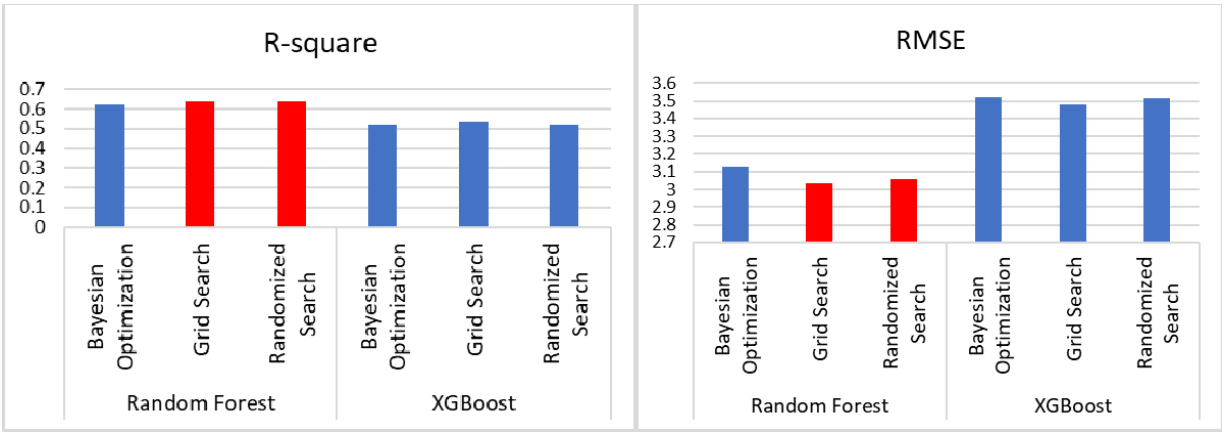

Figure 12. Performance comparisons for Model 3 (using both water quality and meteorologic variables) in the Apalachicola Bay.

Both *grid search* and *randomized search* algorithms indicate that the Random Forest regression (with number of trees = 1) or in other words Decision Tree Regression results in the best accuracy among all with R^2 of around 0.64. The other evaluation metrics and their comparisons are shown in Table 13.

Table 13. Performance evaluation metrics for Model 3 (using both water quality and meteorologic variables) in the Apalachicola Bay. Blue text specifies the best model.

ML algorithm	Optimization algorithm	R^2	RMSE (ug/l)	MAE (ug/l)	PBIAS (%)
Random Forest	**Bayesian Optimization**	0.62	3.13	2.21	-38.79
	Grid Search	0.64	3.04	2.19	-37.47
	Randomized Search	0.64	3.06	2.22	-37.09
XGBoost	**Bayesian Optimization**	0.52	3.52	2.45	-36.17
	Grid Search	0.53	3.48	2.40	-35.97
	Randomized Search	0.52	3.51	2.44	-37.71

3.5 Transferability of the model to other estuarine systems

The optimal model (section 3.4) was used to predict chlorophyll-*a* concentration in other estuarine systems. However, due to the limited data availability, there was not sufficient environmental feature data from all the other systems. We had to adjust the optimal model regarding the environmental parameters. Even though, the model could not perform well to predict the other systems' chlorophyll-*a* with the goodness of fitting (R^2) negative (Figure 13). This could be because that all the bay-estuary systems have different environmental and social characteristics,

which leads the most important environmental features differ from each other. Another import reason could be that the machine learning model relies on the large data set to learn the patterns, the Apalachicola Bay system has the largest data set (875 data points), which domains the performance of the model and could not be validated in other systems which has only around 100 data points. In addition, we tried to integrate 80% data sets from each of system to develop one model and then use it to predict the rest 20% of each system. However, this turned out also a poor R^2, which indicates that the systems are quite site specific and the transferability among the systems is inefficient.

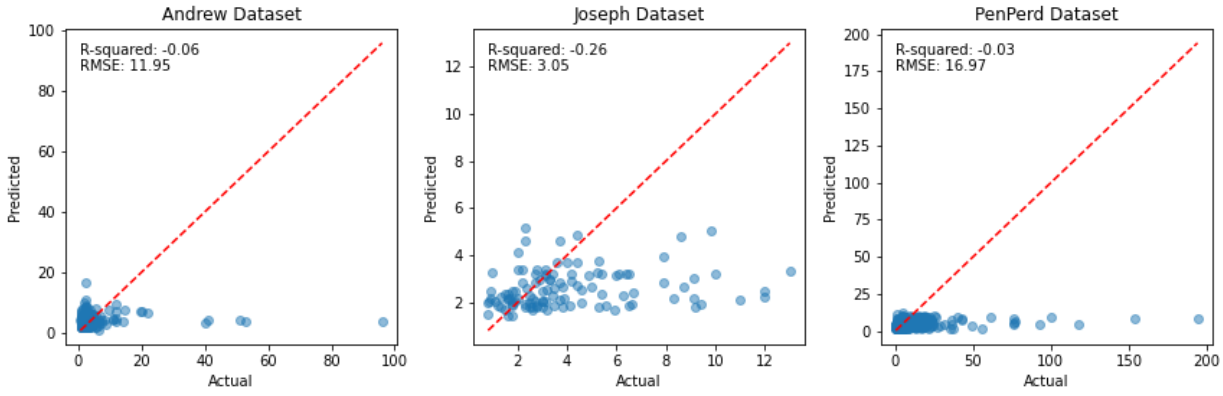

Figure 13. Performance of applying optimal ML model from Aplachicola system to other estuarine systems of St. Andrews, St. Joseph, and Pensacola-Paerdido.

3.6 Model optimization: hyperparameter tuning

To solve the problem of transferability, we developed site-specific models for each of the estuarine systems: Apalachicola, St. Joseph, St. Andrews, and Pensacola-Perdido. Although the models were site-specific, we followed the same framework and input features to implement the models for consistency reasons. Considering the overall data availability for all the systems, Physical WQ Parameters: Salinity (ppt), Turbidity (NTU), DO (mg/L), Water Temperature (deg cels) and pH were used as the predictors for the model development. The systems' physical and chemical parameters differ from each other, and the distribution of data is summarized in Figure 14. To improve the performance of the models, RandomizedSearchCV technique was applied. This technique includes 5-fold cross validation on 80% training samples, 50 iterations, negative RMSE scoring, and search space with number of trees from 1 to 300, maximum depth of each tree at 1 to 7, and learning rat at 0.0001 to 0.011. Furthermore, linear residual correction was applied for model improvement. The models were significantly improved after the linear residual correction (Figure 15). The higher quantiles of chlorophyll-*a* concentrations are crucial of indicating the HAB severity, that indicates the importance of improving the predicting ability of the models.

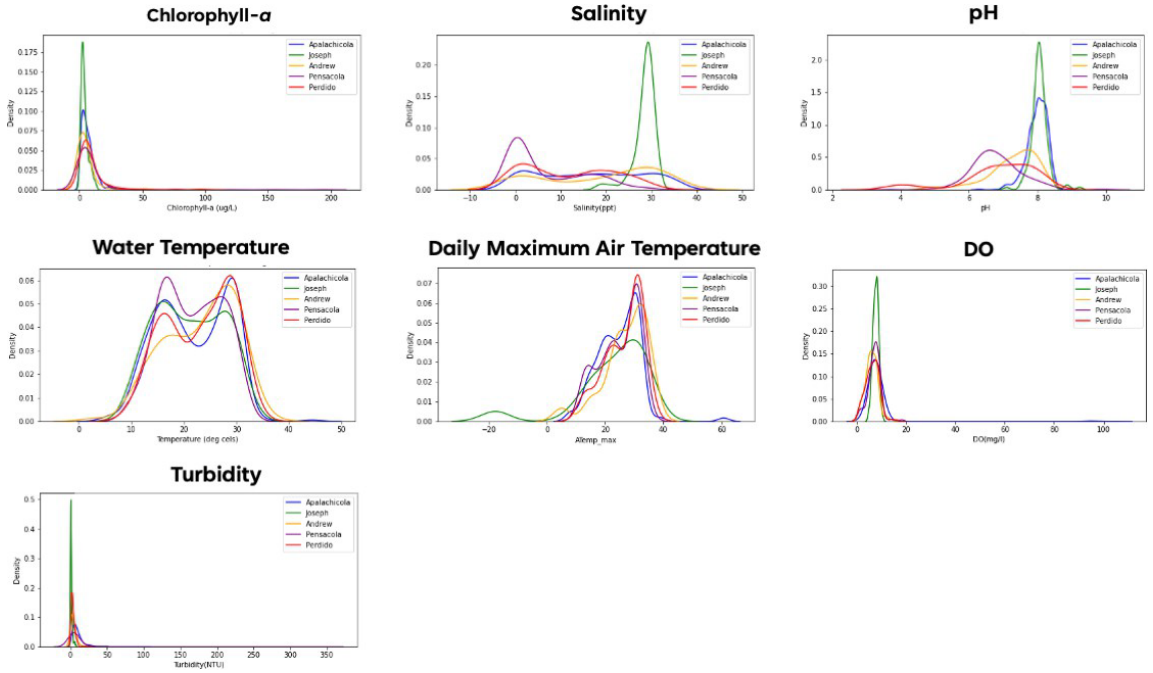

Figure 14. Data distributions of input features and target variable among the estuarine systems: Apalachicola, St. Joseph, St. Andrews, Pensacola, and Perdido.

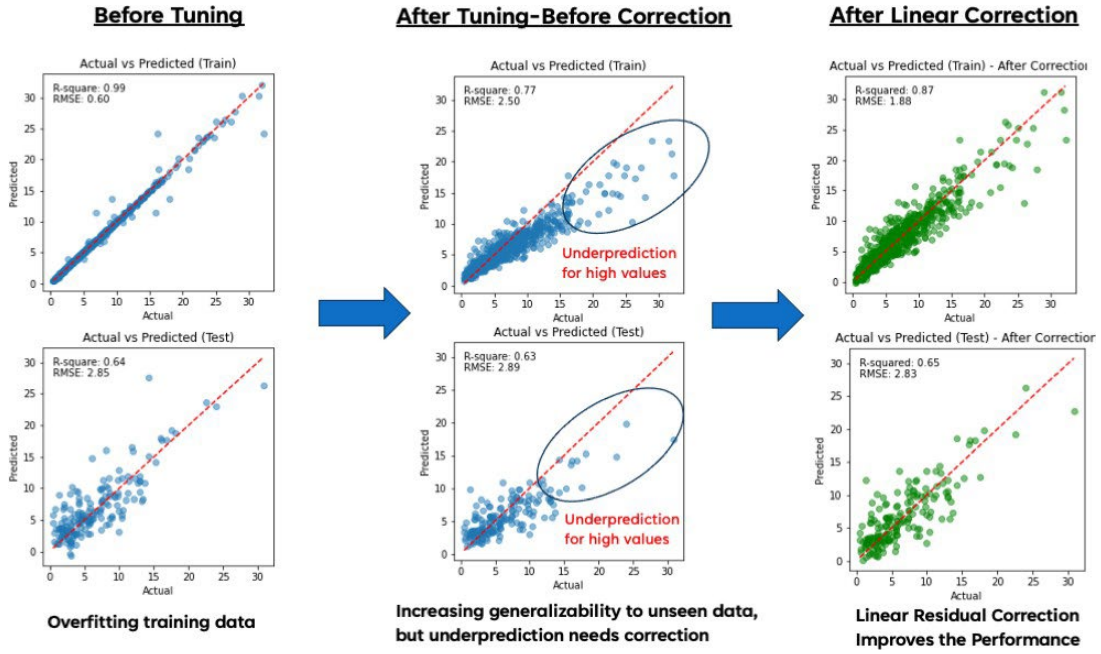

Figure 15. Performance of the finalized model for the Aplachichola Bay before and after applying the linear residual correction.

After optimzing the model for Apalachicola, the same modeling frame was applied to other systmems by using the same framworks, details are summarized in Table 14. The site-specific models showed satisfactory performance in predicting the chlorophyll-*a* concentration based on the chosen predictors (Figure 16). The models showed very good fitting with training datasets (R^2 ranged from 0.69 to 0.97), and with test datasets (R^2 ranged from 0.50 to 0.61). Chlorophyll-*a* as an indicator for HABs can be impacted by complex ecological systems. This is similar to related studies such as Kim et al. (2022) who reported R^2 for test data from 0.2 to 0.7 by applying ML models in predicting chlorophyll-*a* concentrations. Considering we did not use nutrients as predictors (while previous studies did), our model performance was judged to be satisfactory. These models were then used to develop web-based tool for these estuarine systems.

Table 14. Summary of the estuarine system specific models for predicting HABs.

Bay-Estuary system	**Regression model**	**Test data size**	**Input features**
Apalachicola	XGBoost	20%	Physical water quality parameters: Salinity (ppt), DO (mg/l), Turbidity (NTU), pH, Water temperature (deg cels) Climatic: Maximum air temperature and its lags up to previous seven days
St. Joseph	Random Forest	20%	Physical water quality parameters: Salinity (ppt), DO (mg/l), Turbidity (NTU), pH Climatic: Maximum air temperature and its lags up to previous seven days
St. Andrews	XGBoost	20%	
Pensacola-Perdido	Random Forest	30%	

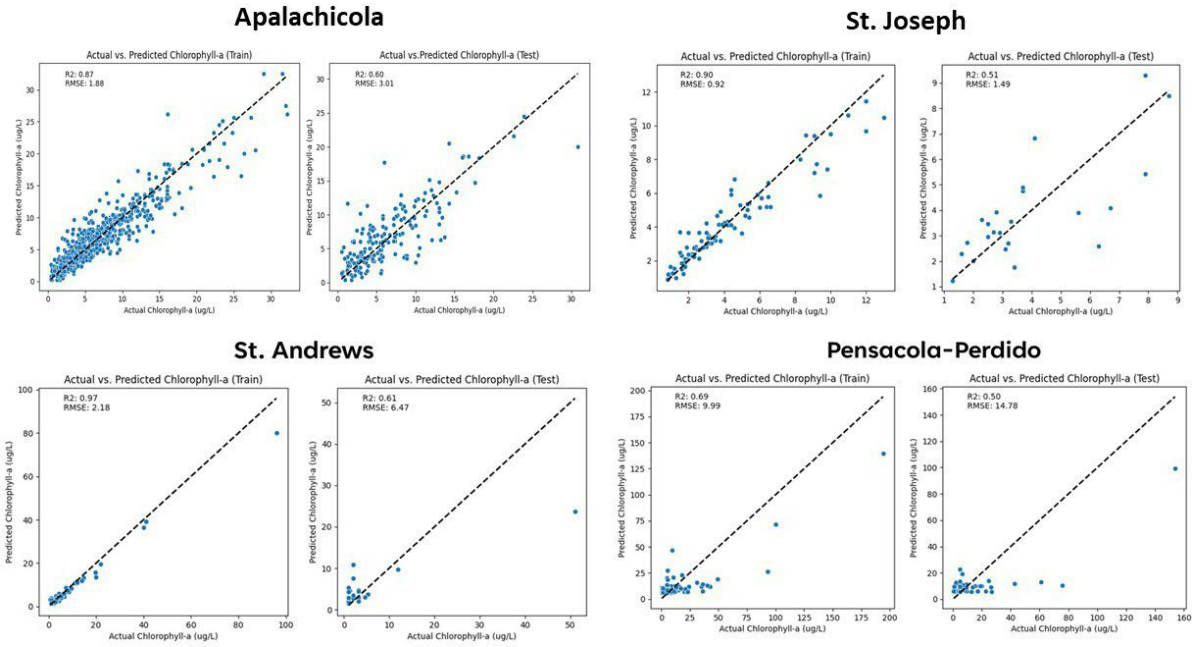

Figure 16. Performance metrics of the final models for both training and testing datasets in the four study systems—Apalachicola, St. Joseph, St Andrews, and Pensacola-Perdido—in terms of R^2 and RMSE.

3.7 Development of a web-based tool for chlorophyll-a prediction

A web-based user interface was developed based on the best ML models for the four estuarine systems using *streamlit (https://fdephab-tool.streamlit.app/)* and can be used to evaluate the vulnerability of each system under different hypothetical (what-if) scenarios (details shown in the next session). There are two main functions for this tool: HAB prediction and vulnerability assessment. For each of the four estuarine system, the user can use the 'Prediction' function to see the performance of the embedded model and to use their own data to predict the chlorophyll-*a* concentration. Further, under the function 'Vulnerability', the user can adjust three environmental features—salinity, air temperature, and pH—to observe vulnerable locations to HABs. The user can choose to show maps of HABs in terms of frequency of occurrence, maximum chlorophyll-*a* concentration, median chlorophyll-*a* concentration, and boxplots of chlorophyll-*a* concentration at different locations (each referring to a monitoring station; Figure 17). The user also has the option to download the maps and .csv files of the predicted chlorophyll-*a* values.

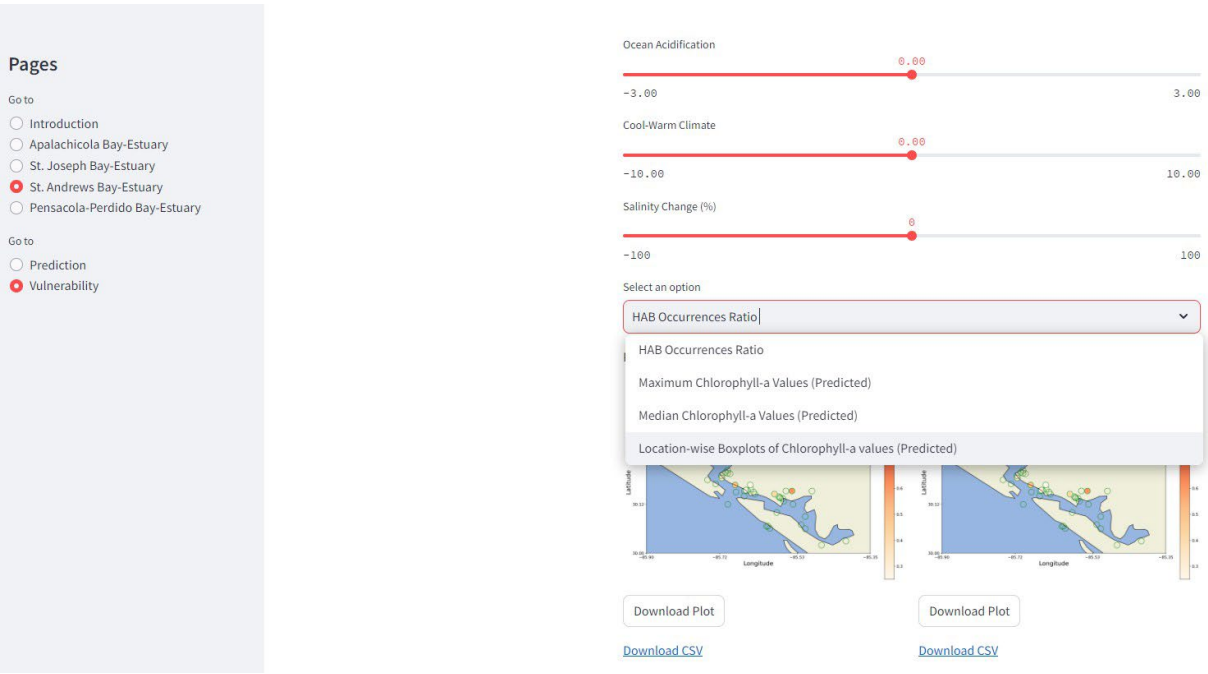

Figure 17. A screenshot of the web-based tool for the four estuarine systems.

3.8 Hypothetical scenarios and the underlying assumptions

We defined different what-if scenarios based on three inputs of the HAB model: air temperature, salinity and pH. The rationale for how each feature can affect HABs is provided in Table 15. The four estuarine systems were assessed via our tool (ML models) in terms of HAB vulnerability under each scenario.

Table 15. Description of the what-if scenarios evaluated in this project.

Scenario name	**Scenario description**	
S0: Business-as-Usual	Historical	Based on the historical observations
S1: Alkalification	pH increases	Algal biomass requires CO_2 for their photosynthesis. Further, they produce organic carbon-products after their die offs. These complex dynamics make chlorophyll-*a* concentration very relevant to the dissolved CO_2 in water. Given the ocean water hypothesized to experience fluctuations in the dissolved CO_2 level in the coming future, pH level will change, which implies possible changes in chlorophyll-*a* concentration or ecological status.
S2: Acidification	pH decreases	
S3: Warmer climate	Max air temperature increases	

S4: Cooler climate	Max air temperature decreases	Air Temperatures dictates the water temperature, the overall water quality dynamics, and hence the chlorophyll-*a* concentration. Changes in climate in the coming future can pose threats to the ecological status.
S5: Salinity increase	Salinity increases	Freshwater flows from the river and tidal seawater from the ocean mixes in estuaries. Anomalies in the incoming freshwater flow (e.g., streamflow droughts) may change the salinity level. Hydro-climatic extremes such as storm surges and hurricanes may change the mixing rate and hence the salinity level of the bays. The observed relevance of salinity level in dictating chlorophyll-*a* concentration in our study provides opportunities to see how the ecological status (HABs) in the estuaries of the Panhandle may experience changes under varying salinity regimes.
S6: Salinity decrease	Salinity decreases	

4 Documentation of Results

Based on the what-if scenarios proposed in Section 3.7, we assessed the HAB frequency (Table 16, Figure 18) and predicted chlorophyll-*a* concentration for each monitoring site (Figure 19-Figure 22) for each of the scenarios at extreme cases.

Table 16. HAB frequency ratio (%) in the four estuarine systems under various what-if scenarios.

Scenario description	Scenario ID	Apalachicola	St. Joseph	St. Andrews	Pensacola-Perdido
Historical	S0	13.94	2.70	13.64	23.55
pH increased by 3	S1	14.40	0.90	15.45	93.48
pH decreased by 3	S2	12.46	2.70	35.45	22.83
Max air temperature increase by 50˚ F	S3	28.34	4.50	18.18	85.14
Max air temperature decrease by 50˚ F	S4	7.54	4.50	4.55	16.30
Salinity increase by 100%	S5	6.74	2.70	10.00	40.94
Salinity decrease by 100%	S6	5.14	1.80	28.18	17.03

Figure 18. HAB frequency ratio (%) at extreme cases for the four estuarine systems under multiple hypothetical scenarios of change in salinity, air temperature or pH.

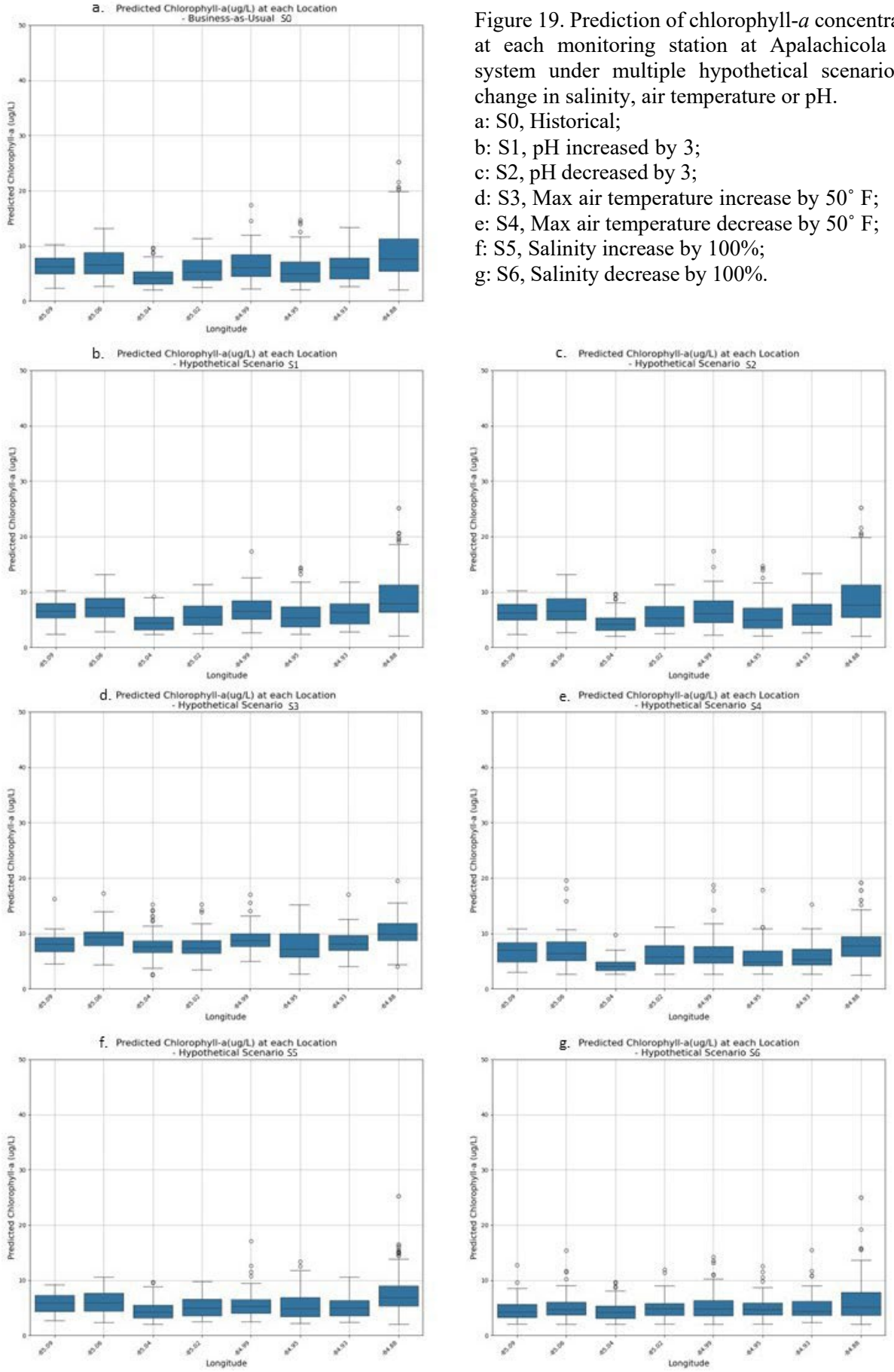

Figure 19. Prediction of chlorophyll-*a* concentration at each monitoring station at Apalachicola Bay system under multiple hypothetical scenarios of change in salinity, air temperature or pH.
a: S0, Historical;
b: S1, pH increased by 3;
c: S2, pH decreased by 3;
d: S3, Max air temperature increase by 50˚ F;
e: S4, Max air temperature decrease by 50˚ F;
f: S5, Salinity increase by 100%;
g: S6, Salinity decrease by 100%.

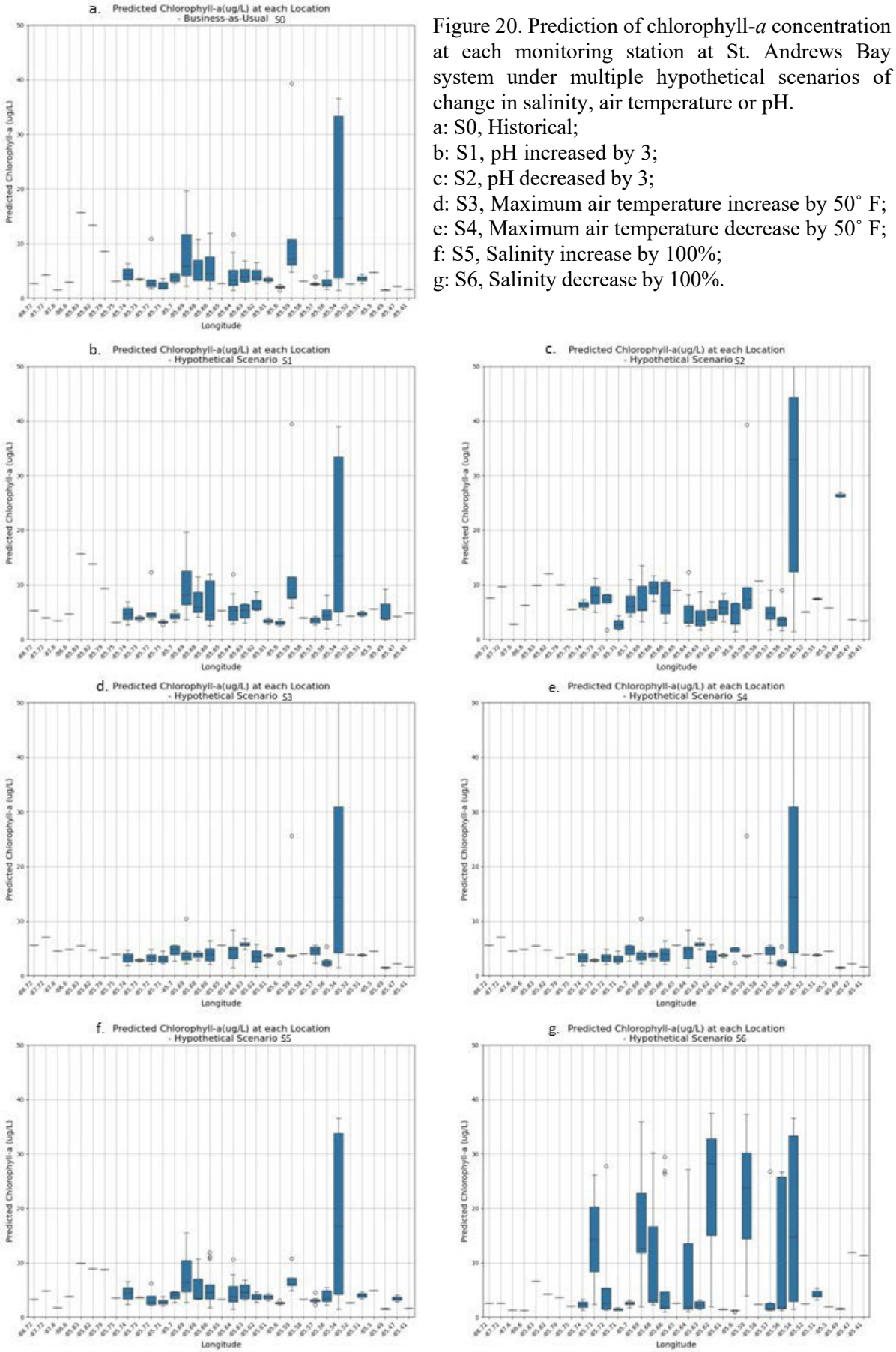

Figure 20. Prediction of chlorophyll-*a* concentration at each monitoring station at St. Andrews Bay system under multiple hypothetical scenarios of change in salinity, air temperature or pH.
a: S0, Historical;
b: S1, pH increased by 3;
c: S2, pH decreased by 3;
d: S3, Maximum air temperature increase by 50˚ F;
e: S4, Maximum air temperature decrease by 50˚ F;
f: S5, Salinity increase by 100%;
g: S6, Salinity decrease by 100%.

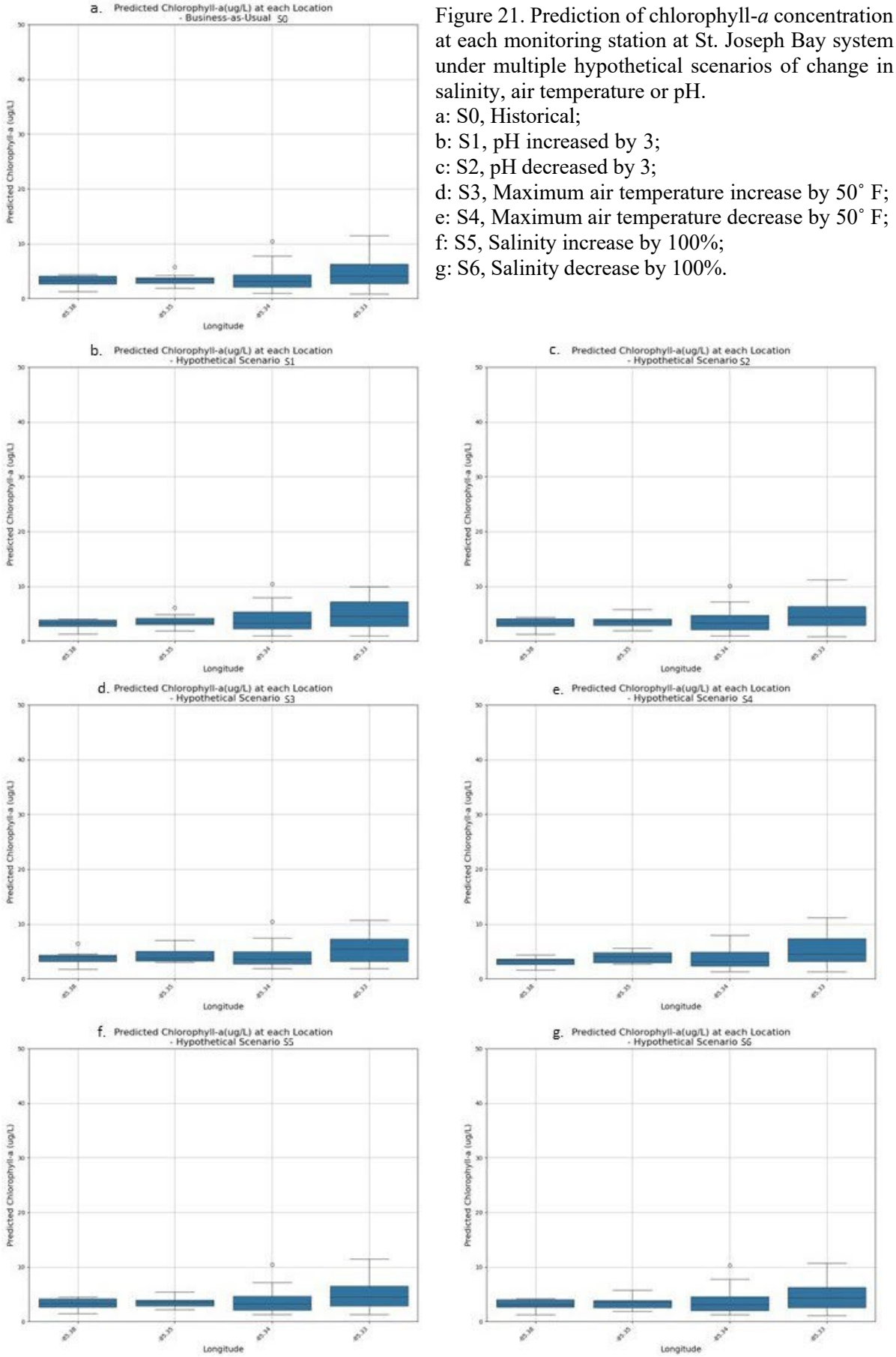

Figure 21. Prediction of chlorophyll-*a* concentration at each monitoring station at St. Joseph Bay system under multiple hypothetical scenarios of change in salinity, air temperature or pH.
a: S0, Historical;
b: S1, pH increased by 3;
c: S2, pH decreased by 3;
d: S3, Maximum air temperature increase by 50˚ F;
e: S4, Maximum air temperature decrease by 50˚ F;
f: S5, Salinity increase by 100%;
g: S6, Salinity decrease by 100%.

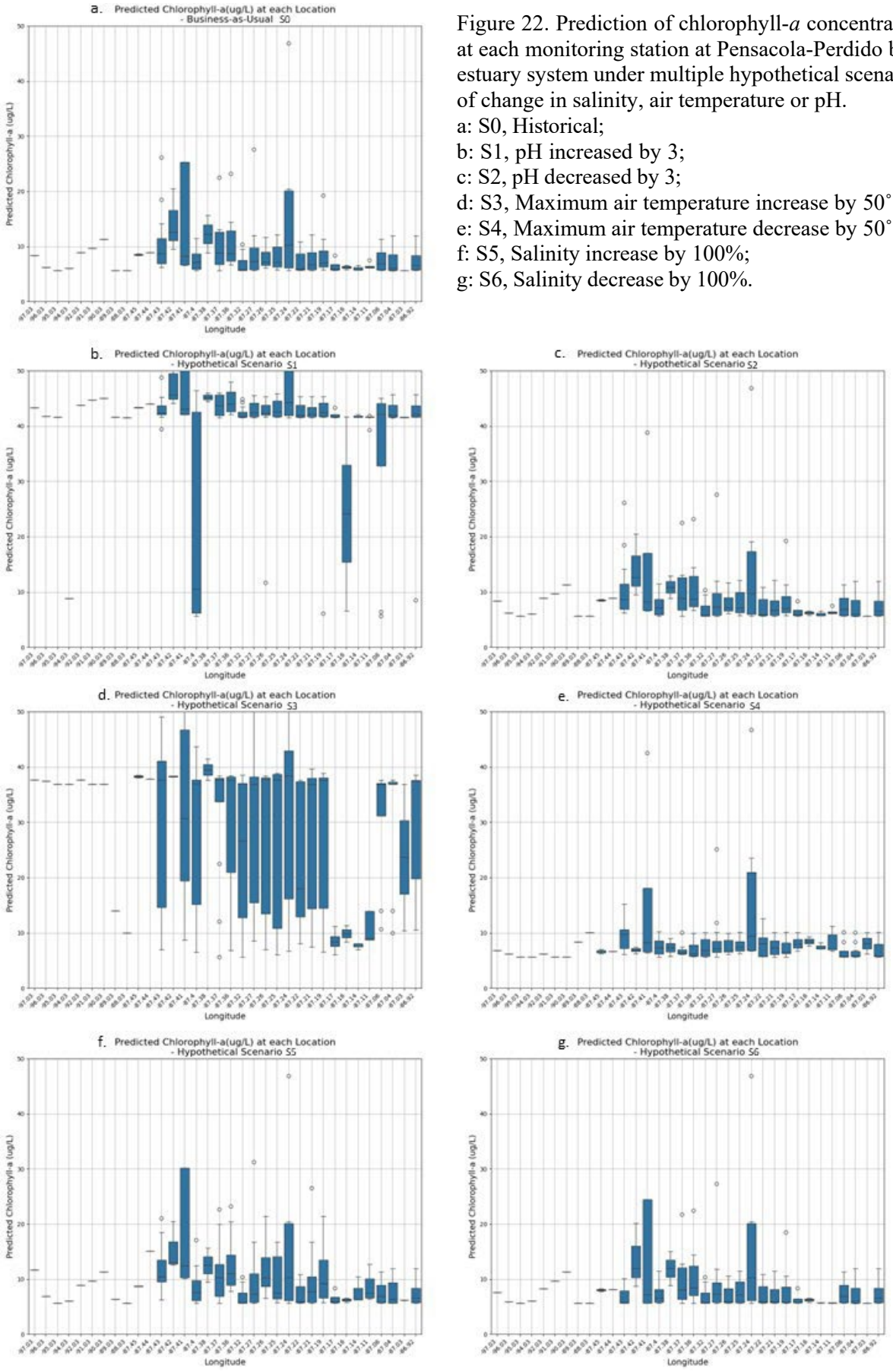

Figure 22. Prediction of chlorophyll-*a* concentration at each monitoring station at Pensacola-Perdido bay-estuary system under multiple hypothetical scenarios of change in salinity, air temperature or pH.
a: S0, Historical;
b: S1, pH increased by 3;
c: S2, pH decreased by 3;
d: S3, Maximum air temperature increase by 50˚ F;
e: S4, Maximum air temperature decrease by 50˚ F;
f: S5, Salinity increase by 100%;
g: S6, Salinity decrease by 100%.

5 Discussion of the Results

We found that ML models can satisfactorily predict the chlorophyll-*a* concentration based on small number of water quality parameters and air temperature. Among the ML algorithms we examined, XGBoot was the best in performance. We improved the performance of standard ML model in predicting high chlorophyll-*a* concentrations, which are vital for HAB occurrence, by applying the linear residual correction approach. The developed models worked satisfactorily only for the estuarine system they were developed for. That is, these models are not transferable across different estuarine systems. These models allow for HAB prediction and vulnerability assessments in estuarine systems of the Florida panhandle.

As mentioned before, the four estuarine systems had different characteristics in terms of water quality dynamics. As in different ranges of these parameters, there could be different cyanobacteria species composition that can lead to different range of chlorophyll-*a* concentration. In turn, when environmental conditions change (e.g., warmer temperatures), different species react differently to the changes. Therefore, the chlorophyll-*a* concentration can change differently in different estuarine systems. Of the scenarios evaluated, the case of pH increases by 3 and maximum temperature increases by 50˚ F, showed the most influential effect on the HABs frequency. Such extreme environmental events, including significant rises in maximum temperatures, could substantially elevate HABs risks, emphasizing the urgent need for effective monitoring and predictive tools in water quality management. The Apalachicola system was most sensitive to the increase of daily maximum temperature. The Pensacola-Perdido system was expected the worst experience both in terms of HAB frequency and concentration when either pH or daily maximum temperature increased. On the other hand, increases in salinity regimes influenced HAB frequency but did not much influence the HAB severity (chlorophyll-*a* concentrations). Compared to Pensacola-Perdido system, the St. Andrews system showed opposite relationships; decreases of pH and decrease of salinity led to more frequent and severe HABs. The difference between the results of different estuarine systems can be due to the difference of species composition that favors various optimal environmental conditions. From the historical observations, the chlorophyll-*a* concentration from St. Joseph system was mostly under 10 µg/L; this can be the reason that even under extreme changes we evaluated (e.g., 50˚ F increase in maximum air temperature), not much influence was predicted. In summary, among the four estuarine systems, Pensacola-Perdido was predicted to be the most vulnerable one, while St. Joseph showed the lowest level of vulnerability to HABs.

Our assessments of the HAB characteristics—frequency and severity—under multiple hypothetical what-if scenarios provide a baseline for detecting potential hotspots of HABs in the future. In turn, these analyses can guide planning and development of adaptive management strategies that consider the likelihood of more frequent and extreme HABs.

6 Fulfilment of the Anticipated Benefits

Through the work documented before, FU has successfully completed the activities proposed and realized the anticipated benefits listed at the beginning of this report.

1) A unified database of a HAB indicator (chlorophyll-*a* concentration) and pertinent HAB drivers in each of the four estuarine systems were prepared. This database can be used to develop similar models and analyze relationships between chlorophyll-*a* and pertinent drivers.
2) Using the database, we identified driving factors of HABs in each of the four estuarine systems of the panhandle.
3) Statistical relationships were conducted between chlorophyll-*a* and multiple environmental features (water quality parameters and meteorologic variables) in each estuarine system.
4) ML models were to predict chlorophyll-*a* concentrations in the four estuarine systems of the panhandle. The models are efficient in predicting high chlorophyll-*a* concentrations that are vital for HAB detection.
5) A web-based tool that allows the user to predict chlorophyll-*a* concentrations in the four estuarine systems of the panhandle was developed. This tool is easy-to-use and public domain. Therefore, we expect that the tool can be utilized easily in the future for HAB prediction and vulnerability assessments.
6) We assessed the vulnerability of each estuarine system under hypothetical scenarios—differed in terms of pH, salinity, and air temperature—using the web-based tool.

7 Limitations and Recommendations for Future Work

The incorporation of ML models in HAB research provides several advantages, primarily through enhanced prediction capabilities and the analyses of complex datasets. While the results showed satisfactory performance by the ML models in predicting HABs, improvements can be made by introducing additional data. Especially for the systems of St. Joseph, St. Andrews and Pensacola-Perdido, that had <200 chlorophyll-*a* data points. With a larger data set, our models may be improved. However, we also acknowledge that grab sampling for monitoring is expensive and time-consuming. Satellites imagery and remote sensing technologies provide an alternative to investigate in HABs. This approach is not only cost-efficient, but also enables the collection of data over extensive years. Furthermore, the capacity for rapid processing of satellite data facilitates the near real-time observation of algal bloom developments, playing a critical role in the early detection of HABs. Additionally, satellites provide a rich historical dataset, allowing for the analysis of long-term trends in HAB occurrences, thereby offering insights into environmental changes that may influence their frequency and intensity.

Our geographic domain was the Florida panhandle. The developed models and web-based tool can be evaluated for other estuarine systems, such as the other part of the Florida Gulf. Through additional analyses and model development, HAB predictions and vulnerability assessment can be done for the entire Florida Gulf. Such efforts would identify what system needs priority attention for HAB prevention in the future.

8 References

1. Cruz, R. C., Reis Costa, P., Vinga, S., Krippahl, L., Lopes, M. B. (2021) A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. *Journal of Marine Science and Engineering*, *9*(3), 283.
2. Florida Blue-Green Algae Task Force (2019) Consensus Document #1. Available at: https://floridadep.gov/sites/default/files/Final%20Consensus%20%231_0.pdf.
3. Kim, K. M., & Ahn, J. H. (2022). Machine learning predictions of chlorophyll-a in the Han river basin, Korea. *Journal of Environmental Management*, *318*, 115636.
4. Pichler, M., Hartig, F. (2023) Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, *14*(4), 994-1016.
5. Xie, Z., Lou, L., Ung, W.K., Mok, K.M. (2012) Freshwater Algal Bloom Prediction by Support Vector Machine in Macau Storage Reservoirs. *Mathematical Problems in Engineering* 1-12:397473.
6. Yu, P., Gao, R., Zhang, D., Liu, Z. P. (2021) Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecological Indicators*, *123*, 107334.
7. Zhang, Y., Lin, H., Chen, C., Chen, L., Zhang, B., Gitelson, A.A. (2011) Estimation of chlorophyll-a concentration in estuarine waters: case study of the Pearl River estuary, South China Sea. *Environmental Research Letters* 6:024016.
8. Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., Ye, L. (2022) A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health* 1(2):107-116.