FLORIDA DEPARTMENT OF ENVIRONMENTAL PROTECTION

Final Report

DEP Agreement No.:	**INV28**
Grantee Name:	Florida State University
Grantee Address:	874 Traditions Way, Third Floor, Tallahassee, FL 32306
Reporting Period:	07/31/2022– 04/30/2024
Project Number and Title:	Development of a Statewide Tool to Predict Harmful Algal Blooms in Freshwater Lakes

Project Background

Excess nutrient export from agricultural nonpoint source pollution, point discharges from wastewater treatment plants, and atmospheric deposition from air pollution cause eutrophication and associated harmful algal blooms in inland waters and freshwater systems. Florida’s lakes, rivers, and springs experience cyanobacteria blooms at the highest frequency in the nation. Along with endangering public health and wildlife, the blooms cost local economies hundreds of millions of dollars. For example, agricultural and urban development in Lake Okeechobee in Southeast Florida and the construction of the Central and South Florida Project for flood control have caused excessive nutrient inputs and cyanobacteria blooms.

Project Description & Location

This project aims to develop a statewide tool to analyze the relationship between Cyanobacteria concentrations (indicators of algal blooms), and watershed and water body independent variables such as land use, temperature, runoff, etc. Specifically, an online integrated map-viewer platform will be developed by the Florida State University that provides insights into the correlations between cyanobacteria and these variables. Advanced data analysis techniques will be used to develop the platform using existing water quality datasets and remote sensing data, so that decision-makers and water managers can select sustainable, innovative, and cost-effective watershed management strategies such as best management practices (BMPs) in regions experiencing harmful algal blooms. The tool will be applicable to different ranges of scales across the diverse climates of Florida.

Project Objectives and Tasks

The overarching goal of this project is to develop an online integrated map-viewer platform that analyzes the relationship between cyanobacteria concentrations and various watershed and water body independent variables, facilitating informed decision-making for sustainable watershed management strategies in regions experiencing harmful algal blooms. Specific tasks are to 1) develop a quality assurance (QA) manual for data evaluation and usability, including draft and final versions, 2) collect and analyze data on cyanobacterial blooms, including remote sensing data, existing datasets, and relevant factors affecting cyanobacteria concentrations, 3) develop and deliver a tool that establishes the relationship between cyanobacteria and watershed/water body independent variables, incorporating advanced numerical methods such as machine learning. deliver an online integrated map and a presentation showcasing the tool's application, and 4) prepare a comprehensive final report summarizing the project's results, including all tasks in the grant work plan. This report primarily focuses on the completion and findings of Task 2, which involved data collection and analysis related to cyanobacterial concentrations. The report provides insights into the collected data, its sources, and the calculated cyanobacteria concentrations presented through an interactive online map.

Task 1. Quality Assurance Project Plan (QAPP)

The QAPP was submitted by the Florida State University on November 16, 2022, and was approved by FDEP on December 08, 2022. No amendments have been made to the QAPP.

Task 2. Data Collection and Analysis

The online tool developed in Task 2 and the interim report were submitted by the Florida State University on June 6th, 2023. All data collected and calculated for this project have been deposited into the FDEP data repository.

2.1 Calculation of Cyanobacteria Concentrations

To effectively manage water quality, allocate monitoring resources, and aid managers in responding to cyanobacterial harmful algal blooms (CyanoHABs), it is crucial to have access to timely cyanoHAB data derived from satellites through web-based platforms. Addressing this need, we have developed an online map indicating the calculated cyanobacteria concentrations for the time period of 2002 to 2022 across Florida using satellite remote sensing data based on 7-day maximum value composites derived from different sensors: the European Space Agency Copernicus's Medium Resolution Imaging Spectrometer (MERIS; 2002-2012), Ocean and Land Colour Instrument (OLCI) on Sentinel-3A (2016-present), and OLCI on Sentinel-3B (2018-present). Remote sensing data have been obtained from Cyanobacteria Assessment Network (CyAN) project (https://oceancolor.gsfc.nasa.gov/projects/cyan/), which is a collaborative effort among the U.S. Environmental Protection Agency (EPA), the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and the United States Geological Survey (USGS). Remote sensing data obtained from the CyAN project include digital numbers (DN) that can be used to calculate cyanobacteria concentrations using Equation (1). The spatial resolution of the data is 300 meters, meaning each pixel represents an area of 300 square meters on the ground, and a 50-meter land mask. The temporal resolution depends on the sensor and date, with the best coverage since 2018 due to the utilization of sensors on two Sentinel-3 satellites. CyAN data are available in GeoTIFF format, with daily values (2022 to present), 7-day maximum values (2007 to present), and 14-day maximum values (2002-2008).

In our online tool, we have employed a 7-day maximum value approach for estimating cyanobacteria concentrations, driven by two key factors. Firstly, cyanobacteria blooms typically occur over the span of a few days to a week, making it necessary to consider a longer time period for accurate estimation. Secondly, the use of a 7-day composite minimizes the impacts of cloud cover and maximizes the frequency of available data based on a typical workweek to guide management decisions. Due to the unavailability of weekly data from CyAN for a certain period (i.e., 2002-2007), we generated weekly 7-day composite images by retaining the maximum value detected for each pixel from daily values within that specific timeframe using a raster calculator tool at ArcMap (V 10.8.1). Upon completion of creating 7-day maximum values for 2002 to the end of 2022 across Florida, rasters were processed using ArcMap. First, rasters were clipped to the Florida border, ensuring that the analysis focuses on the specific region of interest. Then, cyanobacteria concentrations were calculated using Equation 1.

DN Guide:

- 0 indicates below the threshold of CI detection limits (grey color)

- 1-253 are data.
- 254 is land (brown).
- 255 are no data (black--e.g., a cloudy pixel).
- To convert Digital Number (DN) to CI_cyano:

$$CI_{cyano}=10^{(DN*0.011714-4.1870866)} \qquad \text{Eq (1)}$$

That range is ~10,000 to 7,000,000 cells/ml. Each shapefile includes DNs, cyanobacteria concentrations, and whether cyanobacteria concentrations exceeded safety thresholds per pixel across Florida (see Table 1). Figure 1 indicates an example of cyanobacteria concentrations calculated in this study in Lake Okeechobee in January and February of 2022. To explore cyanobacteria concentrations in various lakes across Florida and for different time periods ranging from 2002 to 2022, please refer to the web application.

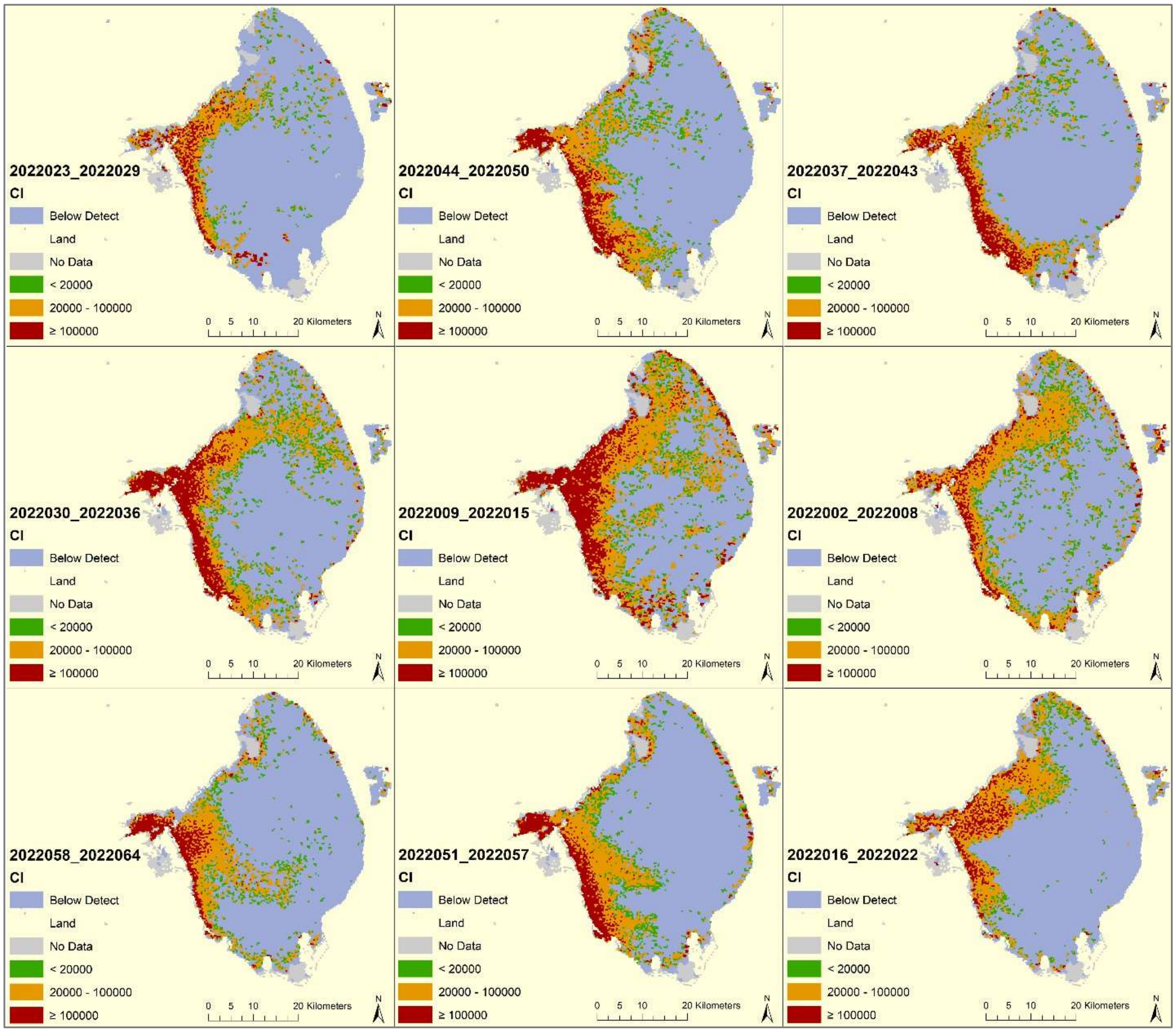


Figure 1. Visualization of cyanobacteria concentrations (cells/ml, 7-day maximum) in Lake Okeechobee during January and February of 2022, with color-coded safety thresholds.

2.2 Safety Thresholds for Cyanobacteria Concentrations

While satellite observations are unable to detect toxins (Stumpf et al., 2016), they can accurately measure the abundance of cyanoHABs (cyanobacterial harmful algal blooms) (Kutser, 2009). Considering the limitations in toxin monitoring (Clark et al., 2017), cyanoHAB abundance assessed through the CyAN app may be more suitable for evaluating risks in Florida. Cell counts and microcystin concentrations are commonly utilized as indicators of potential health hazards, with various states adopting customized thresholds based on local data (Graham et al., 2009). For instance, Oklahoma and Massachusetts have developed specific guidelines to determine safe levels. In Oklahoma, a warning is issued to lake users if cell counts surpass 100,000 cells mL^{-1} or microcystin concentrations exceed 20 μg L^{-1}. On the other hand, Massachusetts has established guidelines stating that an advisory against water contact should be issued when cell counts go beyond 70,000 cells mL^{-1} or microcystin concentrations exceed 14 μg L^{-1}.

The World Health Organization (WHO) provides estimates of microcystin concentrations corresponding to the cell abundance at each guideline level. The U.S. Environmental Protection Agency (EPA) has also established a drinking water health advisory for cyanobacteria microcystin toxin (U.S. EPA, 2015). The WHO employs a three-level guideline approach, utilizing chlorophyll-a, a widely present photosynthetic pigment, and cyanobacterial cell abundance (cells mL^{-1}) to determine associated risks and issue warnings or closures. These values take into consideration potential exposures through various recreational activities such as contact with water, ingestion, and inhalation. These parameters are used in our online Tool as indicators to estimate the potential health risks associated with engaging in recreational activities in environments where cyanobacteria are present (see Table 1).

Table 1. Thresholds used in our CyanoHABs Florida online tool to evaluate the risks of cyanobacteria based on WHO's Guidelines for Safe Practice in Managing Recreational Waters.

Guidance level or situation	How guidance level derived	Health risks	Typical actions
Relatively low probability of adverse health effects **20000 cyanobacterial cells/ml** *or* 10 ug chlorophyll-a/liter with dominance of cyanobacteria	From human bathing epidemiological study	Short-term adverse health outcomes, e.g., skin irritations, gastrointestinal illness	Post on-site risk advisory signs Inform relevant authorities
Moderate probability of adverse health effects **100 000 cyanobacterial**	From provisional drinking-water guideline value for	Potential for long-term illness with some cyanobacterial species	Watch for scums or conditions conducive to scums and further

Guidance level or situation	How guidance level derived	Health risks	Typical actions
cells/ml *or* 50 ug chlorophyll-a/liter with dominance, of cyanobacteria	microcystin-LR and data concerning other cyanotoxins	health outcomes, e.g., skin irritations, gastrointestinal illness	investigate hazard Post on-site risk advisory signs Inform relevant authorities
High probability of adverse health effects Cyanobacterial scum formation in areas where whole-body contact and/or risk of ingestion/aspiration occur.	Inference from oral animal lethal poisoning. Actual human illness case histories	Potential for acute poisoning Potential for long-term illness with cyanobacterial species Short-term adverse activities health outcomes, e.g., skin irritations, gastrointestinal illness	Immediate action to control contact with scums; possible prohibition of swimming and other water contact activities Public health follow-up investigation Inform public and relevant authorities

2.3 Development of an Online Application for Cyanobacteria Concentrations

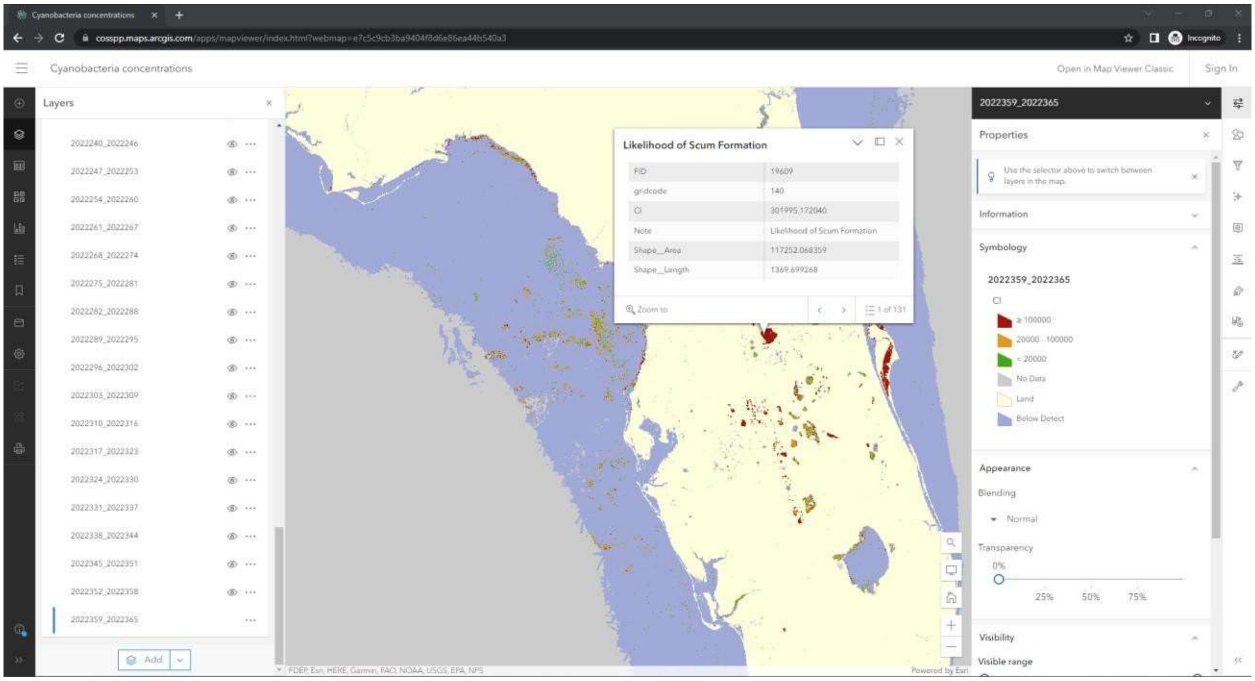
We have developed an online map alongside an online map package that includes shapefiles representing weekly cyanobacteria concentrations across Florida. This map also includes the evaluation of cyanobacteria-related risks based on WHO safety thresholds. In addition to cyanobacteria concentrations and risk assessment, the developed web application includes several features to enhance user experience and functionality:

1. Dropdown Menu for Selecting Years: The application includes a dropdown menu that allows users to select different years for cyanobacteria concentration data. This feature facilitates temporal analysis and comparison of cyanobacteria blooms over multiple years.
2. Zooming In and Out: Users can zoom in and out of the map to view cyanobacteria concentrations at different levels of detail. This feature allows for a closer examination of specific areas or a broader view of the entire map.
3. Search for Location: The application provides a search function that allows users to enter a specific location or address. This enables them to quickly navigate to their desired area of interest without manually searching for it on the map.
4. Inserting Coordinates: Users can input specific coordinates to navigate directly to a particular location on the map.
5. Basemap Selection: The web application offers a variety of basemap options for users to choose from. They can select different basemaps such as satellite imagery, street maps, topographic maps, or custom basemaps, depending on their preference or the specific context of analysis.
6. Layer List: The application provides a layer list that displays the available data layers and allows users to toggle their visibility on or off. This feature enables users to customize the displayed information based on their specific interests or analysis requirements.

7. Bookmark: Users can create bookmarks or save specific locations of interest within the application. This feature allows for easy navigation and quick access to frequently visited areas or important points on the map.
8. Print: The application includes a printing functionality, allowing users to generate printable maps or reports of their selected areas and data layers.
9. Query: Users can perform queries or spatial analysis on the data, allowing them to extract specific information or insights based on their criteria or spatial relationships.

Figure 2 indicates an overview of the web application alongside these features. These features collectively enhance the usability and functionality of the web application, providing users with intuitive tools for exploring and analyzing cyanobacteria concentrations across Florida. In addition to web applications, a map package is available online for users to download and open in desktop GIS software such as ArcMap. A map package is a file format that bundles together all the necessary data, including shapefiles, associated with the project. By providing a map package for download, users can access all the shapefiles and other data layers included in the package for further analysis, exploration, and customized visualizations using desktop GIS software. The web application and map package can be accessed using the links below:

1. Online map: https://cosspp.maps.arcgis.com/home/item.html?id=e7c5c9cb3ba9404f8d6e86ea44b540a3
2. Map package: https://cosspp.maps.arcgis.com/home/item.html?id=a9b4380188be4384b3c586700b0e876b



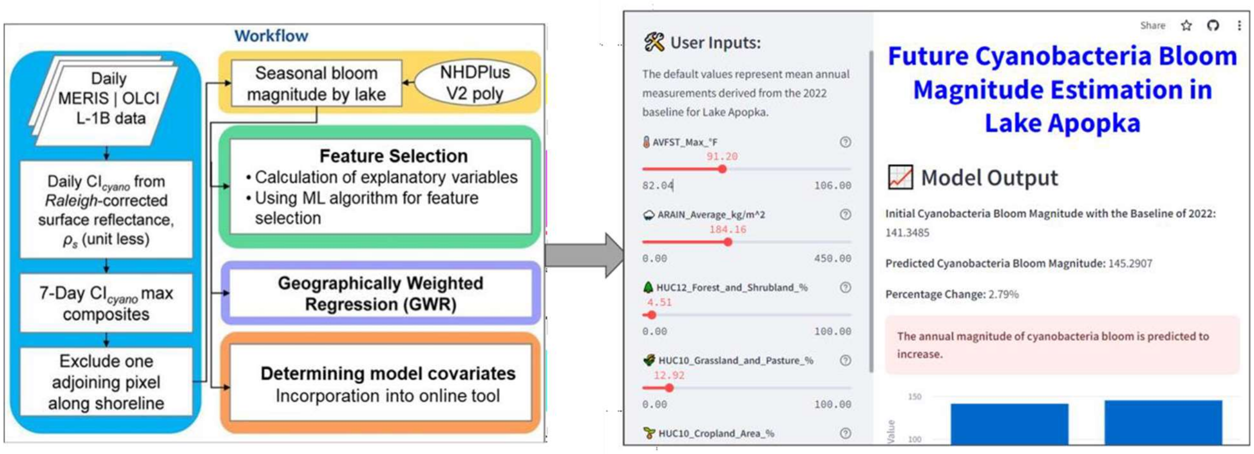
Likelihood of Scum Formation

FID	19609
gridcode	140
CI	301995.172040
Note	Likelihood of Scum Formation
Shape__Area	117252.068359
Shape__Length	1369.699268

Figure 2. depicts an overview of the web application, showcasing cyanobacteria concentrations (measured in cells/ml, with 7-day maximum values) in Florida's lakes spanning the period from 2002 to 2022. The application allows users to interact with the map by clicking on specific locations. Upon selection, attribute tables are displayed, providing detailed information on concentrations as well as safety statuses determined by the World Health Organization (WHO) safety thresholds.

Task 3. Tool Development and Verification of Success

In this task, we established a comprehensive statewide tool to examine the correlation between Cyanobacteria concentrations (indicative of algal blooms) and a range of independent variables including land use, temperature, rainfall, etc. Subsequently, an integrated online map-viewer platform was created to enhance understanding of the connections between cyanobacteria and these diverse variables within watersheds and water bodies across the state. Detailed information regarding data and methodologies is outlined below. Additionally, to enhance clarity, a conceptual workflow illustrating the data flow and analysis methods is presented in Figure 3


Workflow
Daily MERIS | OLCI L-1B data
Daily CI_{cyano} from *Raleigh*-corrected surface reflectance, ρ_s (unit less)
7-Day CI_{cyano} max composites
Exclude one adjoining pixel along shoreline
Seasonal bloom magnitude by lake
NHDPlus V2 poly
Feature Selection
• Calculation of explanatory variables
• Using ML algorithm for feature selection
Geographically Weighted Regression (GWR)
Determining model covariates
Incorporation into online tool
Share
Value
150
100

User Inputs:

The default values represent mean annual measurements derived from the 2022 baseline for Lake Apopka.

AVFST_Max_°F: 91.20 (82.04 – 106.00)

ARAIN_Average_kg/m^2: 184.16 (0.00 – 450.00)

HUC12_Forest_and_Shrubland_%: 4.51 (0.00 – 100.00)

HUC10_Grassland_and_Pasture_%: 12.92 (0.00 – 100.00)

HUC10_Cropland_Area_%

Future Cyanobacteria Bloom Magnitude Estimation in Lake Apopka

Model Output

Initial Cyanobacteria Bloom Magnitude with the Baseline of 2022: 141.3485

Predicted Cyanobacteria Bloom Magnitude: 145.2907

Percentage Change: 2.79%

The annual magnitude of cyanobacteria bloom is predicted to increase.

Figure 3. Satellite data processing and analysis workflow for processing and analyzing satellite data, emphasizing the key methods and steps involved in investigating the relationship between watershed variables and changes in cyanobacteria bloom magnitude in Florida lakes from 2002 to 2022. Additionally, it showcases the development of an inline tool for predicting cyanobacteria in freshwater lakes based on established relationships.

3.1 Seasonal Bloom Magnitude

The Cyanobacteria bloom magnitude aims to capture two fundamental aspects of algal blooms: the quantity of biomass and the duration of the bloom. While metrics such as frequency and spatial extent offer insights into the temporal and spatial characteristics of the bloom within a lake, they do not specifically address the seasonal/annual intensity. To address this, a spatial-temporal mean is employed to capture the quantity and duration of the entire lake's cyanobacteria biomass over the course of a year. Consequently, we estimated the bloom magnitude as the spatiotemporal mean cyanobacteria biomass within a lake throughout the year using the following equation (Mishra et al., 2023; Schaeffer et al, 2024):

$$\textit{Mean bloom magnitude} = \frac{a_p}{A_{lake}} \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T} \sum_{t=1}^{T} \sum_{p=1}^{P} CI_{cyano,p,t,m}$$

Where the indices *P* and *T* denote the number of valid pixels in a lake or water body and the number of composite (time) sequences in each month (e.g., four in a month), respectively. Here, *M* represents the number of months in a season or the annual study period, *ap* is the area of a pixel, and A_{lake} is the area of the lake extracted from the National Hydrography Dataset Plus version 2.0 (NHDPlusV2) lake vector layer. Utilizing only the valid pixel area for calculating the spatial mean could introduce bias to the estimates. An excess of invalid pixels during high-concentration bloom events might lead to underestimation, while more invalid pixels over periods of bloom absence or non-detect pixels may result in overestimation of the bloom magnitude. To address this, we incorporated the lake area into the equation for estimating bloom magnitude. Henceforth, for brevity, we refer to the spatiotemporal mean cyanobacteria bloom magnitude as simply "bloom magnitude."

For accurate remote sensing data, it is essential to ensure a sufficient spatial resolution. In our case, we require a minimum of three pixels per lake, with each pixel measuring 300 by 300. By applying this criterion, we identified 134 lakes in Florida that meet the requirement of having at least three pixels and are of a size suitable for utilizing remote sensing data for cyanobacteria concentrations. Figure 4 depicts the geographical distribution of these lakes, and detailed information, including the name and characteristics of each lake, can be found in the data directory.

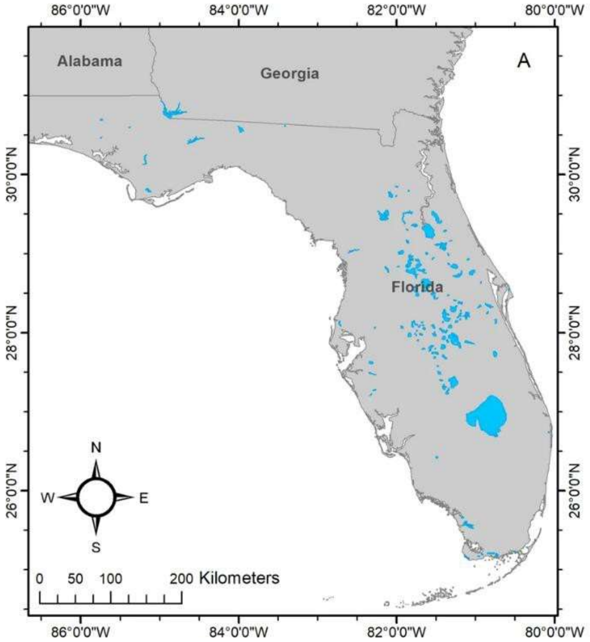


Figure 4: Geographical distribution of 134 lakes in Florida meeting the criteria for accurate remote sensing data analysis. Each lake, represented on the map, possesses a minimum of three pixels, each measuring 300 by 300, making them suitable for the assessment of cyanobacteria concentrations using remote sensing data.

3.2 Meteorological Variables

We utilized monthly climate data to explore the correlation between observed differences in bloom magnitude and various climate variables. The monthly climate data were obtained from the North American Land Data Assimilation System Phase 2 (NLDAS-2) (https://ldas.gsfc.nasa.gov/nldas). NLDAS integrates a large quantity of observation-based and model reanalysis data to drive offline (not coupled to the atmosphere) land-surface models (LSMs) and executes at 1/8th-degree grid spacing over central North America, enabled by the Land Information System (LIS). NLDAS forcing drives four land-surface models: NASA's Mosaic, NOAA's Noah, the NWS Office of Hydrological Development's (OHD) SAC, and Princeton's implementation of VIC. Obtained climate data included average surface skin temperature (AVFST, °F), liquid precipitation (rainfall, ARAIN, kg/m^2), subsurface runoff (baseflow, BGRUN, kg/m^2), and surface runoff (non-infiltrating, SSRUN, kg/m^2).

To enhance our analysis, we derived additional features from the monthly climate data by calculating the statistical mean, minimum, and maximum of a climate variable over specific time periods (e.g., a year). Examples include determining the maximum annual temperature (°F) or computing the cumulative precipitation (kg/m^2) by summing precipitation over January to December. The aggregation of climate variables was tailored to lacustrine cyanobacterial algal bloom phenology in lakes (Coffer et al., 2020). Notably, the final selection of climate variables was not predetermined but instead driven by a data-driven approach utilizing the Random Forest model to identify variable importance.

Furthermore, we obtained the U.S. Climate Extreme Index (CEI) dataset for each year by climate region from the National Climate Data Center (NCDC) website https://www.ncei.noaa.gov/access/monitoring/cei/graph). The CEI quantifies observed changes in climate within each region, summarizing a comprehensive set of multidimensional climate variables in nine climate regions defined by the National Center for Environmental Information (Karl and Koss, 1984). This dataset was employed during the observation period to establish correlations between the simplified and summarized state of climate and occurrences of cyanobacterial harmful algal blooms (cyanoHAB) in FL lakes.

3.3 Land Use and Land Cover (LULC) data

We obtained annual Land Use and Land Cover (LULC) data for the observation period from the United States Department of Agriculture (USDA) National Agricultural Statistical Service (NASS)(https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php). For each lake, we extracted the relevant LULC data within hydrological units at two hierarchical levels that enclose the lake. The Hydrologic Unit Code (HUC) is a hierarchical land area classification system established by the United States Geological Survey (USGS) based on surface hydrologic features in a standardized geographical framework. The United States is partitioned into successively smaller hydrologic units, classified into regions (HUC-2), subregions (HUC-4), basins (HUC-6), sub-basins (HUC-8), watersheds (HUC-10), and sub-watersheds (HUC-12). In this study, we utilized HUC -10, and -12 to consider LULC and physical factors surrounding a lake at watershed to sub-watershed scale, and their impact on bloom magnitude.

Annual acreage information for relevant LULC types, including cropland area, wetland, grassland and pasture, forest and shrubland, and developed area, was extracted two HU boundaries with HU codes ten and twelve (HUC10, HUC12). Pixel counts from the cropland Data Layers (CDL) were converted to acreage by LULC type. Additionally, we calculated the fraction of acreage for each LULC class in each HU, considering the area of the corresponding HU. We incorporated HUCs at two different scales enclosing a lake (HUC10 and HUC12) and allowed the Random Forest feature selection (described below) to determine the relationship between the spatial scale of LULC variables and their impact on bloom magnitude. In our dataset, there are 84 HUC-10 and 165 HUC-12 units enclosing our targeted lakes. Figure 5 displays these HUC-10 and HUC-12 units enclosing Lake Okeechobee.

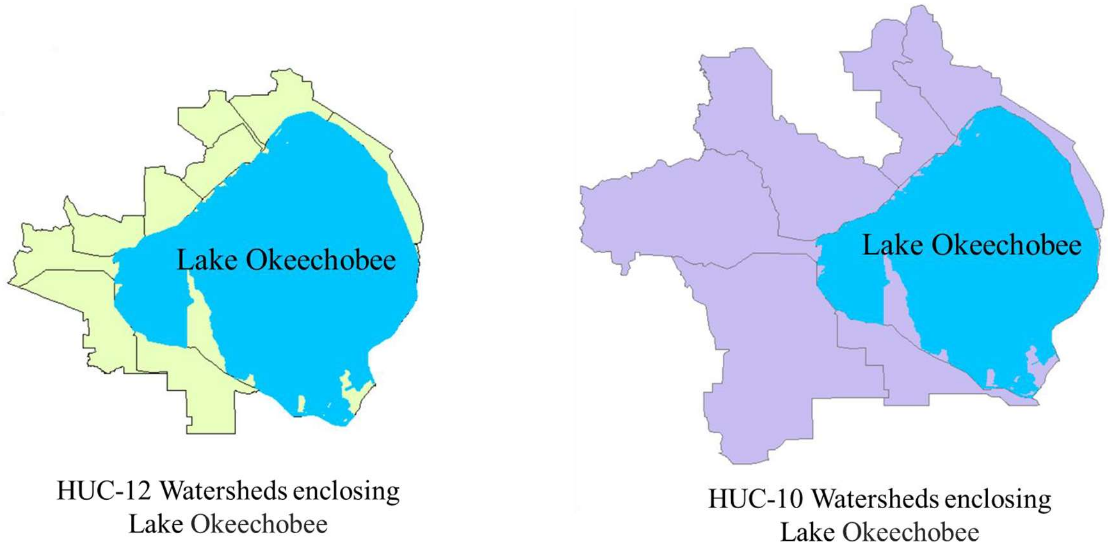


Figure 5. Visualization of Hydrologic Unit Code 10 (HUC-10) and Hydrologic Unit Code 12 (HUC-12) boundaries enclosing Lake Okeechobee.

3.4 Estimation of Total Nitrogen and Phosphorus

In this study, we estimated Total Nitrogen (TN) and Total Phosphorus (TP) for each HUC-10 and HUC-12 watershed enclosing the lakes using the EPA approach, which involves assessing the relationship between nutrient levels and the proportion of land dedicated to specific uses (https://www.epa.gov/sites/default/files/2015-11/documents/2008_04_18_nps_watershed_handbook_ch08.pdf). This method considers the impact of land use on nutrient runoff and aids in predicting TN and TP concentrations in water bodies based on the surrounding land cover composition. Such estimation is widely utilized in watershed management and environmental studies to comprehend the influence of human activities on nutrient levels in aquatic systems.

Since we have already included the percentages of various land uses as explanatory variables in predicting cyanobacteria concentrations, and also estimated TN and TP based on land uses, we refrain from using TN and TP directly for predicting cyanobacteria. This approach is adopted to prevent redundancy, address collinearity concerns, and avoid potential overfitting. By concentrating on the most relevant variables, we aim for accurate predictions in our model. Nonetheless, in the final online tool, we present baseline TN and TP values. Additionally, the tool allows users to estimate TN and TP based on their input for different land use types. This feature enhances the tool's versatility, providing users with the ability to estimate TN and TP according to varying land use scenarios.

3.5 Feature Selection with Random Forest Model

We have compiled an extensive dataset encompassing 300 physical and climate variables for each of the 134 lakes, as outlined in Table 2. Recognizing the intricacies of this dataset, we utilized a Random Forest (RF) regression model for feature selection. RF identifies the most important features by aggregating insights from multiple decision trees. It assesses how each feature contributes to prediction accuracy, assigning higher importance scores to those consistently enhancing the model's performance. This ensemble approach enables robust identification of key variables influencing outcomes, such as cyanobacteria concentrations in our study. In addition, RF models have proven effective in discerning significant variables, even in datasets with a high number of features. Furthermore, our previous study demonstrated the effectiveness of the RF algorithm in predicting HAB in Florida (Yan et al., 2024), highlighting its suitability for our current analysis. Thus, we have utilized the RF algorithm to identify the most significant features impacting cyanobacteria concentrations in our study.

Table2. Details of the physical and climate variables utilized as input for the algorithm aimed at identifying the most significant features impacting cyanobacteria concentrations in our study.

Variable Type	Variable Name
Dependent	Annual Cyanobacteria bloom magnitude
Explanatory	Minimum annual temperature
Explanatory	Maximum annual temperature
Explanatory	Average annual temperature
Explanatory	Minimum annual rainfall
Explanatory	Maximum annual rainfall
Explanatory	Average annual rainfall
Explanatory	Sum of annual rainfall
Explanatory	Palmer Drought Severity Index (PDSI)
Explanatory	Percentage of cropland area in HUC-10 watershed enclosing each lake
Explanatory	Percentage of wetland area in HUC-10 watershed enclosing each lake

Explanatory	Percentage of grassland and pasture in HUC-10 watershed enclosing each lake
Explanatory	Percentage of forest and shrubland in HUC-10 watershed enclosing each lake
Explanatory	Percentage of developed area in HUC-10 watershed enclosing each lake
Explanatory	Percentage of cropland area in HUC-12 watershed enclosing each lake
Explanatory	Percentage of wetland area in HUC-12 watershed enclosing each lake
Explanatory	Percentage of grassland and pasture in HUC-12 watershed enclosing each lake
Explanatory	Percentage of forest and shrubland in HUC-12 watershed enclosing each lake
Explanatory	Percentage of developed area in HUC-12 watershed enclosing each lake

Utilizing feature ranks and importance scores, we identified 6 key LULC and climate features crucial for modeling bloom magnitude, as detailed in Table 3.

Selected LULC features

- *All_crops_acr_pct_hu10*: is the percentage of the total acreage of all croplands in the HUC 10, representing the agricultural activity in the hydrologic unit surrounding a lake under study. Therefore, that would serve as a proxy of nutrient loading to a lake in the form of excess nutrients transferred from surrounding agricultural land to the lake through surface runoff.

- *Forest_shrub_acr_pct_hu12:* is the percent area of the HU with code 12 surrounding a lake covered by forest and shrubland. Lakes in hydrologic units with higher forest and shrubland cover would be expected to be in pristine condition with less anthropogenic disturbance.

- *Grassland_pasture_acr_pct_hu10*: is the percent area of the HU with code ten surrounding a lake covered by grassland and pasture. Grasslands and pastures can act as sources by working as a nonpoint source of excessive fertilizer. It can also serve as a sink by absorbing nutrients from the surface runoff by taking the role of cover crops.

- *Developed_acr_pct_hu12:* is the percent area of the HU with code 12 surrounding a lake covered by developed areas. Developed areas can act as nutrient sources, contributing to higher nutrient levels in a lake and influencing bloom conditions.

Selected climate features

- T_{max}: represents the maximum monthly temperature recorded from January to December.

- Average monthly precipitation is the mean precipitation over the months of January to December.

Table 3: Key Land Use and Climate Features Identified for Modeling Bloom Magnitude through Feature Ranks and Importance Scores.

Selected features	**Description**

AVFST_Max	Maximum of air **temperature** observed over a year.
ARAIN_Average	The average annual **precipitation**.
HUC12_TN	The average total **nitrogen** concentrations of the HU with code 12 surrounding a lake.
HUC10_TP	Average total **phosphorus** concentrations of the HU with code 10 surrounding a lake.
HUC10_ % cropland area	Percentage of the total acreage of all croplands in the HUC 10, representing the **agricultural activity** in the hydrologic unit surrounding a lake under study.
HUC12_ %developed area	Percent area of the HU with code 12 surrounding a lake covered by developed area, representing **urban areas**.

3.6 Geographically Weighted Regression (GWR)

In this study, we employed Geographically Weighted Regression (GWR), a spatial statistical method designed for modeling spatially heterogeneous processes. GWR allows for varying relationships between a response variable and a set of covariates across geographic space (Fotheringham et al., 2001). GWR extends ordinary least-square (OLS) regression. Using a spatial weight matrix allows models to vary over space, addressing the non-stationary effect of independent variables on the response variable (Fotheringham et al., 2001).

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon \quad \text{(S1)}$$

Where y_i is the dependent variable at lake year i; β_{i0} refers to the regression intercept; β_{ik} refers to the independent parameter; X_{ik} is the value of the k^{th} regression parameter; ε_i refers to the model residuals at lake year location *i*.

$$\underset{i}{\beta} = (X^T \underset{i}{W} X)^{-1} X^T W \underset{i}{y} \quad \text{(S2)}$$

$$w_{ij} = [-\frac{1}{2}(\frac{d_{ij}}{b})^2] \quad \text{(S3)}$$

where d_{ij} is the Euclidian distance between observation point *j* and regression point *i* with planar coordinates, and b is the kernel bandwidth.

GWR has been recognized as a superior approach (Kang et al., 2023) compared to classical linear regression, especially when the effects of independent variables exhibit spatial variability. Unlike classical linear regression, which assumes that data comes from an independent and identically distributed population of random variables and does not consider the geographical location of variables, GWR explicitly incorporates spatial information into the regression model. This enables the detection of spatial variation in the relationship among variables.

In this study, GWR was applied to model localized physical and anthropogenic factors surrounding lakes, as outlined in Table 3, and their association with bloom magnitude. The primary component of GWR is the spatial weight matrix, wherein closer observations are assigned larger weights defined by spatial kernel functions such as a Gaussian function (Brunsdon et al., 2002). In this study, we utilized an adaptive kernel as the kernel type and employed a bandwidth method to determine the bandwidth parameter. The adaptive kernel, combined with the specified bandwidth method, allowed for a customized spatial weighting scheme tailored to the unique characteristics of each lake, enhancing the precision of the GWR models in capturing the spatially varying relationships within the study area.

The six independent variables were scaled to a range of zero to one before training the GWR regression models. This scaling facilitates the comparison of model coefficient maps and the relative effects of independent variables based on the magnitude or size of the coefficients. It is important to note that variable selection for GWR was performed 'globally' using a Random Forest model, not 'locally.' This approach aimed to capture local relations without training over-fitted GWR models, which can occur with local variable selection. Additionally, meaningful variables with broader significance across the FL were chosen to draw meaningful conclusions in a statewide-wide study. Consequently, localized regression models are calibrated by data from surrounding locations. **GWR produces n sets of model coefficients and model R^2 (local R^2), corresponding to the number of lakes (134)**, allowing for visualization through descriptive statistics or surface maps. GWR statistics, including an overall R^2 of 0.5222, are summarized, and additional metrics can be referenced in Table 4.

Verification of the success: Figure 6 showcases the relationship between the predicted and the observed annual cyanobloom magnitudes for 134 lakes in Florida during the study period, offering insights into the model’s performance. With an R^2 value of 0.522, the model demonstrates a strong predictive capability, particularly significant in the context of cyanobacterial bloom prediction. This level of accuracy is noteworthy, as predicting cyanobacterial blooms is inherently challenging due to their complex nature and the multitude of influencing environmental factors.

The substantial predictive power is further highlighted by a Residual Squares value of 0.3236, indicating that the model is not only capable of capturing the general trends in cyanobacterial bloom occurrences but also minimizes the error margin in its predictions. Such precision in prediction is crucial for effective monitoring and management of water bodies, potentially allowing for timely intervention and mitigation strategies against bloom events.

The results signify a considerable achievement in predictive modeling of cyanobacteria, yet they also suggest an opportunity for ongoing improvement. By continuously refining the model, particularly focusing on reducing residual errors and enhancing its explanatory power, we can aim to achieve even greater accuracy in predicting and managing cyanobacterial blooms.

Table 4. Geographically Weighted Regression (GWR) statistics, showcasing the overall R^2 value and additional relevant metrics for the modeling of spatially varying relationships between

dependent and explanatory variables in the study.

OBJECTID	VARNAME	VARIABLE	DEFINITION
1	Neighbors	25	
2	ResidualSquares	0.3236	
3	EffectiveNumber	23.5546	
4	Sigma	0.0942	
5	AICc	-69.5149	
6	**R^2**	**0.5222**	
7	Dependent Field	0	Norm_CyAN_
8	Explanatory Field	1	HUC10_crop
9	Explanatory Field	2	HUC10_gras
10	Explanatory Field	3	HUC12_deve
11	Explanatory Field	4	HUC12_fore
12	Explanatory Field	5	ARAIN_Aver
13	Explanatory Field	6	AVFST_Max_

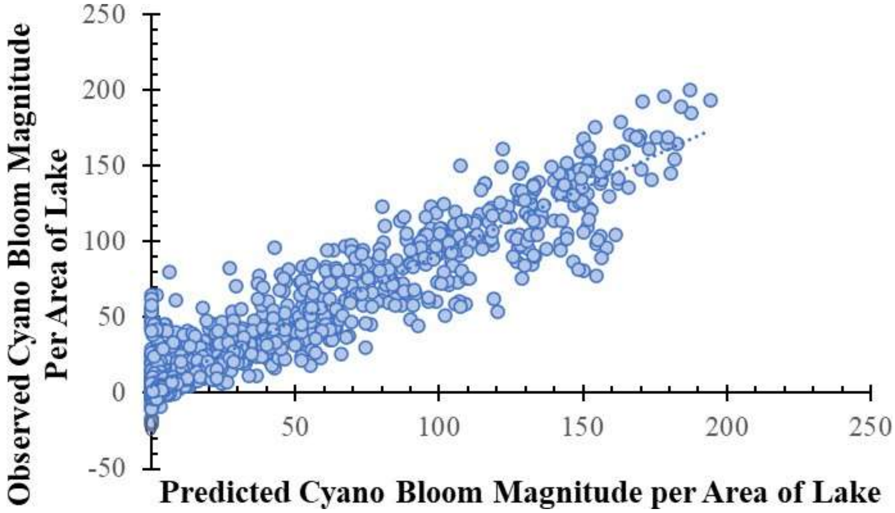


Figure 6. Scatter Plot comparing Predicted Annual Cyanobloom Magnitude to Observed Annual Cyanobloom Magnitude for 134 lakes in Florida within the study period. This visual representation provides insights into the accuracy and effectiveness of the predictive model in capturing cyanobloom occurrences.

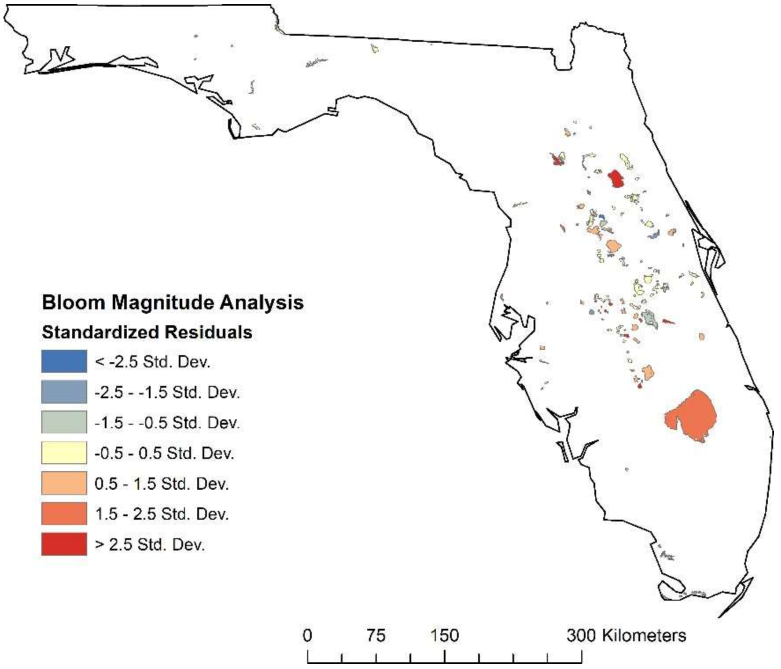
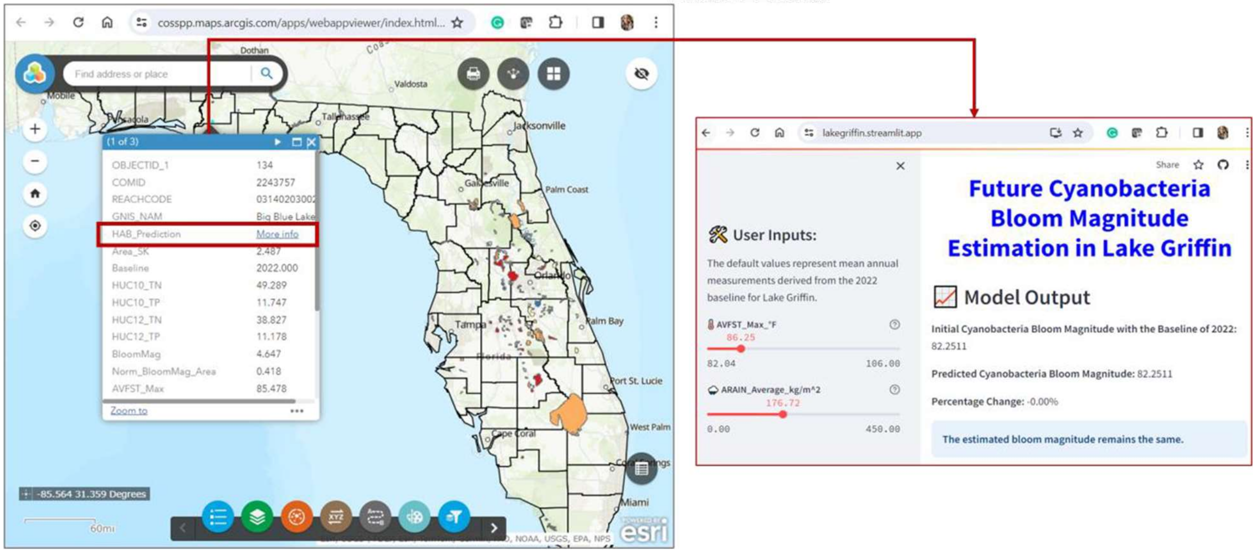


Figure 7. Spatial Distribution of Standardized Residuals (StdResid) from the Geographically Weighted Regression (GWR) Model. Standardized residuals highlight areas of significant deviations between observed and predicted values, aiding in the assessment of model performance and identification of spatial patterns in model errors.

3.6 Development of an Online Open-Source Tool

We have integrated regression equations for each lake into our online web app tool, enabling users to manipulate variables such as % land uses (cropland, developed areas, forest, and grassland) along with meteorological variables (maximum temperature and average rainfall) to predict bloom magnitude. The tool is developed using ArcGIS Online and Python, and the Python code can be found in Appendix B. Metrics values from 2022 serve as the baseline, allowing the tool to predict bloom magnitude and the % increase or decrease compared to 2022. We checked the accuracy of the model by predicting observed HAB data for 134 lakes. The results are presented in Figure 6. No expertise in GIS or Python is required. The tool features functionalities such as search, and filtering based on counties and bloom magnitude. It is open source and easy to use. Additionally, a map package is provided for download, enabling more in-depth analysis in GIS desktop applications.

Our tool comprises two components: a spatial tool indicating bloom magnitude for 2022, color-coded based on quantile classification; and a second part where users can click on each lake. This action redirects them to the online interface for inputting values and predicting cyanobacteria bloom magnitude. These two components are shown in Figure 8.



(1 of 3)

OBJECTID_1	134
COMID	2243757
REACHCODE	03140203002
GNIS_NAM	Big Blue Lake
HAB_Prediction	More info
Area_SK	2.487
Baseline	2022.000
HUC10_TN	49.289
HUC10_TP	11.747
HUC12_TN	38.827
HUC12_TP	11.178
BloomMag	4.647
Norm_BloomMag_Area	0.418
AVFST_Max	85.478

Zoom to

User Inputs:

The default values represent mean annual measurements derived from the 2022 baseline for Lake Griffin.

AVFST_Max_°F: 86.25 (82.04 – 106.00)

ARAIN_Average_kg/m^2: 176.72 (0.00 – 450.00)

Future Cyanobacteria Bloom Magnitude Estimation in Lake Griffin

Model Output

Initial Cyanobacteria Bloom Magnitude with the Baseline of 2022: 82.2511

Predicted Cyanobacteria Bloom Magnitude: 82.2511

Percentage Change: -0.00%

The estimated bloom magnitude remains the same.

Figure 8. Spatial Visualization of Cyano bloom Magnitude in 2022 - Color-coded representation based on quantile classification, providing a comprehensive overview of bloom intensity across lakes alongside a predictive component, allowing estimation of cyanobacteria bloom magnitude through consideration of land use, land cover, and meteorological variables.

- Access our developed online tool at https://cosspp.maps.arcgis.com/apps/webappviewer/index.html?id=65fbe71c7c6940a09d09049df2f378f3.
- For further in-depth analysis, download the map package for GIS desktop use from https://cosspp.maps.arcgis.com/home/item.html?id=1764ecfee36149c886408fb34e4b5391

Project Timeline and Budget Summary:

Task/ Deliverable No.	Task or Deliverable Title	Task Start Date	Task End Date
1	Quality Assurance Manual		
1a	Draft Quality Assurance Manual	Upon Execution	9/30/2022
1b	Quality Assurance Manual	Upon Execution	11/30/2022
2	Data Collection and Analysis	Upon Execution	
2a	Interim Report	Upon Execution	06/30/2023
3	Tool Development and Verification of Success	Upon Execution	
3a	Interim Report	Upon Execution	01/31/2024
4	Final Report	Upon Execution	

4a	Draft Final Report	Upon Execution	2/29/2024
4b	Final Report	Upon Execution	06/10/2024

BUDGET DETAIL BY TASK:

Budget Category	Budget Amount
Total for Task 1:	**$50,785**
Total for Task 2:	**$118,501**
Total for Task 3:	**$118,501**
Total for Task 4:	**$50,787**
Total	**$338,575**

*No adjustments were made to the overall budget for each task; it simply involved reallocating funds among tasks due to modifications in personnel costs and their associated fringe benefits. Additionally, we extended the deadline for the final task to accommodate the revisions requested by the FDEP grant manager.

Project Schedule vs. Actual Completion

The project has adhered closely to the original timeline and scope as detailed in the Grant Work Plan. This alignment is a confirmation to the effective planning, management, and execution of the project tasks by the team. Despite the complex nature of developing a statewide predictive tool for harmful algal blooms, the project has met its milestones on time and has successfully navigated the challenges inherent in such an ambitious endeavor.

Schedule Adherence

From the outset, the project was structured around a detailed timeline that included key milestones such as the development of the Quality Assurance Project Plan (QAPP), data collection and analysis, model development and testing. Each phase was completed as planned, with no significant deviations from the scheduled dates.

Unexpected Site Conditions and Adjustments

Given the statewide scope of the project, potential for unexpected site conditions and the need for adjustments was anticipated. However, through diligent planning and flexible management strategies, the project team was able to preempt and mitigate these risks effectively. There were no reported unexpected site conditions that necessitated significant adjustments to the project plan.

Significant Delays and Corrections

Remarkably, the project encountered no significant unexpected delays or corrections. This achievement is particularly notable given the project's reliance on field data collection and analysis, which are often susceptible to delays due to weather conditions, equipment malfunctions, or data quality issues.

Deviations from the Original Project Plan

The project's adherence to the original project plan, without any significant deviations, underscores the robustness of the initial project design and the effectiveness of the project management approach. This fidelity to the planned process ensured that the project objectives were met within the established timeframe and budget, ultimately contributing to the project's success.

Discussion on the Anticipated Benefits Realization

Given the project's on-time completion and adherence to the planned activities, it's critical to assess whether the anticipated benefits, particularly regarding Best Management Practices (BMP) and their expected removal efficiency, have been or will likely be achieved. The project aimed at developing a predictive tool for HABs in freshwater lakes across Florida, with several key anticipated benefits:

- **Enhanced Predictive Capabilities**: The tool integrates advanced data analysis and machine learning algorithms to predict HAB occurrences. This capability allows for proactive rather than reactive measures in managing water quality, directly contributing to the efficacy of BMPs aimed at nutrient reduction and bloom prevention.
- **Informed Decision-Making**: By providing detailed analysis and visualization of the relationships between cyanobacteria concentrations and various environmental variables, the tool empowers water managers and decision-makers. This informed decision-making facilitates the selection and implementation of BMPs tailored to specific conditions and challenges of different water bodies.
- **Increased Efficiency of BMPs**: The predictive tool's insights into nutrient loading and its impact on HAB occurrences help refine the application of BMPs. By understanding the variables most significantly associated with bloom events, management practices can be optimized for maximum nutrient removal efficiency.
- **Cost-Effectiveness**: The ability to predict and thereby prevent HABs can significantly reduce the costs associated with bloom management, such as treatment, cleanup, and economic losses related to recreational and commercial water use disruptions.

Considering these points, the project's successful execution suggests that the anticipated benefits are well on their way to being realized. The development and implementation of the predictive tool marks a significant advancement in the management of freshwater resources in Florida. By providing a means to better understand and predict HAB occurrences, the tool directly supports the enhanced effectiveness and efficiency of BMPs.

While it's too early to quantify the exact improvement in BMP removal efficiency, the project's outcomes align with the goals of improving water quality management and reducing the incidence and impact of HABs. Continued monitoring and analysis will be essential to fully assess the long-term benefits and the tool's contribution to BMP performance improvements.

Appendices

Appendix A: Data Directory

TASK 3

1. Cyanobacteria concentrations
2. Meteorological Variables:
3. Land use land cover
4. Regression Equations

<u>Appendix B: Code for Developing Online Tool</u>

Appendix B contains the code for developing an online tool. The provided code, exemplified using Lake Apopka as a sample, has been crafted to accommodate all 134 lakes.

```
import streamlit as st

import pandas as pd

import numpy as np


# Display the title with blue color and centered text

title_markdown = "<h1 style='color: blue; text-align: center;'>Future Cyanobacteria Bloom Magnitude Estimation in Lake Apopka</h1>"

st.markdown(title_markdown, unsafe_allow_html=True)


# Initial values according to the baseline of 2022 for Lake Apopka

initial_values = {

    'Norm_CyAN': 141.348473,

    'AVFST_Max': 88.106,

    'ARAIN_Average': 184.16,

    'HUC12_forest_and_shrubland_4': 4.511227711,

    'HUC10_grassland_and_pasture_3': 12.92275406,

    'HUC10_cropland_area_1': 3.332721843,

    'HUC12_developed_area_5': 27.27513883

}


# Coefficients for Lake Apopka
```

```
coefficients = {
    'intercept': 2.707261329,
    'AVFST_Max': 0.057913537,
    'ARAIN_Average': -0.112755289,
    'HUC12_forest_and_shrubland_4': -0.363781812,
    'HUC10_grassland_and_pasture_3': -5.514615137,
    'HUC10_cropland_area_1': -2.882319595,
    'HUC12_developed_area_5': -2.836429498
}


# Equations variables
b1, c1, d1, e1, f1, g1 = 82.04, 163.72, 0, 0, 0, 0.052616068
b2, c2, d2, e2, f2, g2 = 90.86, 223.83, 80.3992991, 81.38497115, 86.75640259, 79.36556518


# Sidebar for user inputs with icons
st.sidebar.markdown("<h2 style='font-size: 24px;'>     User Inputs:</h2>",
unsafe_allow_html=True)
st.sidebar.write("The default values represent mean annual measurements derived from the
2022 baseline for Lake Apopka.")


# Slider variables:
b3, c3, d3, e3, f3, g3 = 82.04, 0.00, 0.00000000, 0.000000000, 0, 0.000000000
b4, c4, d4, e4, f4, g4 = 106.00, 450.00, 100.00, 100.00, 100.00, 100.00


# User Input in the sidebar with colorful labels
AVFST_Max_user = st.sidebar.slider("**     AVFST_Max_°F**", b3, b4,
initial_values['AVFST_Max'], step=0.1, key="avfst_max", help="Adjust the annual
maximum air temperature.")
```

```
ARAIN_Average_user = st.sidebar.slider("**      ARAIN_Average_kg/m^2**", c3, c4,
initial_values['ARAIN_Average'], step=0.1, key="arain_average", help="Adjust the annual
average rainfall.")

HUC12_forest_and_shrubland_4_user = st.sidebar.slider("**
HUC12_Forest_and_Shrubland_%**", d3, d4,
initial_values['HUC12_forest_and_shrubland_4'], step=0.1, key="huc12_forest_shrubland",
help="Modify the percentage of forest and shrubland within the HUC12 watershed enclosing
the lake.")

HUC10_grassland_and_pasture_3_user = st.sidebar.slider("**
HUC10_Grassland_and_Pasture_%**", e3, e4,
initial_values['HUC10_grassland_and_pasture_3'], step=0.1,
key="huc10_grassland_pasture", help="Modify the percentage of grassland and pasture
within the HUC10 watershed enclosing the lake.")

HUC10_cropland_area_user = st.sidebar.slider("**      HUC10_Cropland_Area_%**",
float(f3), float(f4), initial_values['HUC10_cropland_area_1'], step=0.1,
key="huc10_cropland", help="Modify the percentage of cropland within the HUC10
watershed enclosing the lake.")

HUC12_developed_area_5_user = st.sidebar.slider("**•      HUC12_Developed_Area_%**",
float(g3), float(g4), initial_values['HUC12_developed_area_5'], step=0.1,
key="huc12_developed", help="Modify the percentage of developed area within the HUC12
watershed enclosing the lake.")


# Calculate Predicted Magnitude

Y = coefficients['intercept'] + \

   coefficients['AVFST_Max'] * (AVFST_Max_user - b1) / (b2 - b1) + \

   coefficients['ARAIN_Average'] * (ARAIN_Average_user - c1) / (c2 - c1) + \

   coefficients['HUC12_forest_and_shrubland_4'] * (HUC12_forest_and_shrubland_4_user -
d1) / (d2 - d1) + \

   coefficients['HUC10_grassland_and_pasture_3'] *
(HUC10_grassland_and_pasture_3_user - e1) / (e2 - e1) + \

   coefficients['HUC10_cropland_area_1'] * (HUC10_cropland_area_user - f1) / (f2 - f1) + \

   coefficients['HUC12_developed_area_5'] * (HUC12_developed_area_5_user - g1) / (g2 -
g1)
```

```
final_bloom_magnitude = Y * 194.0458755
percentage_change = (final_bloom_magnitude - initial_values['Norm_CyAN']) /
initial_values['Norm_CyAN'] * 100


# Main content to display the output with an icon
st.header("    Model Output")


# Bar chart data
chart_data = pd.DataFrame({
    'Magnitude Type': ['Initial Bloom Magnitude', 'Predicted Bloom Magnitude'],
    'Magnitude Value': [initial_values['Norm_CyAN'], final_bloom_magnitude]
})


# Display the final result with bold text
st.write(f"**Initial Cyanobacteria Bloom Magnitude with the Baseline of 2022:**
{initial_values['Norm_CyAN']:.4f}")
st.write(f"**Predicted Cyanobacteria Bloom Magnitude:** {final_bloom_magnitude:.4f}")


# Display the percentage change with bold text
st.write(f"**Percentage Change:** {percentage_change:.2f}%")


# Display a message based on the change with color and bold text
threshold = 0.001


if abs(percentage_change) < threshold:
    st.info("**The estimated bloom magnitude remains the same.**")
elif percentage_change > 0:
    st.error("**The annual magnitude of cyanobacteria bloom is predicted to increase.**")
```

```
else:
    st.success("**The annual magnitude of cyanobacteria bloom is predicted to decrease.**")

# Bar chart
chart_data = pd.DataFrame({
    'Magnitude Type': ['Initial Bloom Magnitude', 'Predicted Bloom Magnitude'],
    'Magnitude Value': [initial_values['Norm_CyAN'], final_bloom_magnitude]
})

# Display the bar chart
st.bar_chart(chart_data, x='Magnitude Type', y='Magnitude Value')
```

References:

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (2002). Geographically weighted summary statistics—a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, 26(6), 501-524.

Clark, J. M., Schaeffer, B. A., Darling, J. A., Urquhart, E. A., Johnston, J. M., Ignatius, A. R., ... & Stumpf, R. P. (2017). Satellite monitoring of cyanobacterial harmful algal bloom frequency in recreational waters and drinking water sources. *Ecological indicators*, *80*, 84-95.

Coffer, M. M., Schaeffer, B. A., Darling, J. A., Urquhart, E. A., & Salls, W. B. (2020). Quantifying national and regional cyanobacterial occurrence in US lakes using satellite remote sensing. *Ecological indicators*, 111, 105976.

EPA, US. (2015). Drinking Water Health Advisory for the Cyanobacterial Microcystin Toxins. *US Environmental Protection Agency Office of Water, Health and Ecological Criteria Division Washington, DC*. https://www.epa.gov/cyanohabs/world-health-organization-who-1999-guideline-values-cyanobacteria-freshwater

Fotheringham, A. S., Charlton, M. E., & Brunsdon, C. (2001). Spatial variations in school performance: a local analysis using geographically weighted regression. *Geographical and environmental Modellin*g, 5(1), 43-66.

Graham, J. L., Loftin, K. A., & Kamman, N. (2009). Monitoring recreational freshwaters. *Lakelines*, *29*, 18-24.

Kang, T., Wang, H., He, Z., Liu, Z., Ren, Y., & Zhao, P. (2023). The effects of urban land use on energy-related CO2 emissions in China. *Science of The Total Environment*, 870, 161873.

Karl, T., & Koss, W. J. (1984). Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983.

Mishra, S., Stumpf, R. P., Schaeffer, B. A., & Werdell, P. J. (2023). Recent changes in cyanobacteria algal bloom magnitude in large lakes across the contiguous United States. *Science of The Total Environment*, 897, 165253.

Schaeffer, B. A., Reynolds, N., Ferriby, H., Salls, W., Smith, D., Johnston, J. M., & Myer, M. (2024). Forecasting freshwater cyanobacterial harmful algal blooms for Sentinel-3 satellite resolved US lakes and reservoirs. *Journal of Environmental Management*, 349, 119518.

Shang, W., Jin, S., He, Y., Zhang, Y., & Li, J. (2021). Spatial–Temporal Variations of Total Nitrogen

Stumpf, R. P., Davis, T. W., Wynne, T. T., Graham, J. L., Loftin, K. A., Johengen, T. H., ... & Burtner, A. (2016). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful algae*, *54*, 160-173.

Yan, Z., Kamanmalek, S., & Alamdari, N. (2024). Predicting coastal harmful algal blooms using integrated data-driven analysis of environmental factors. *Science of The Total Environment*, 912, 169253.