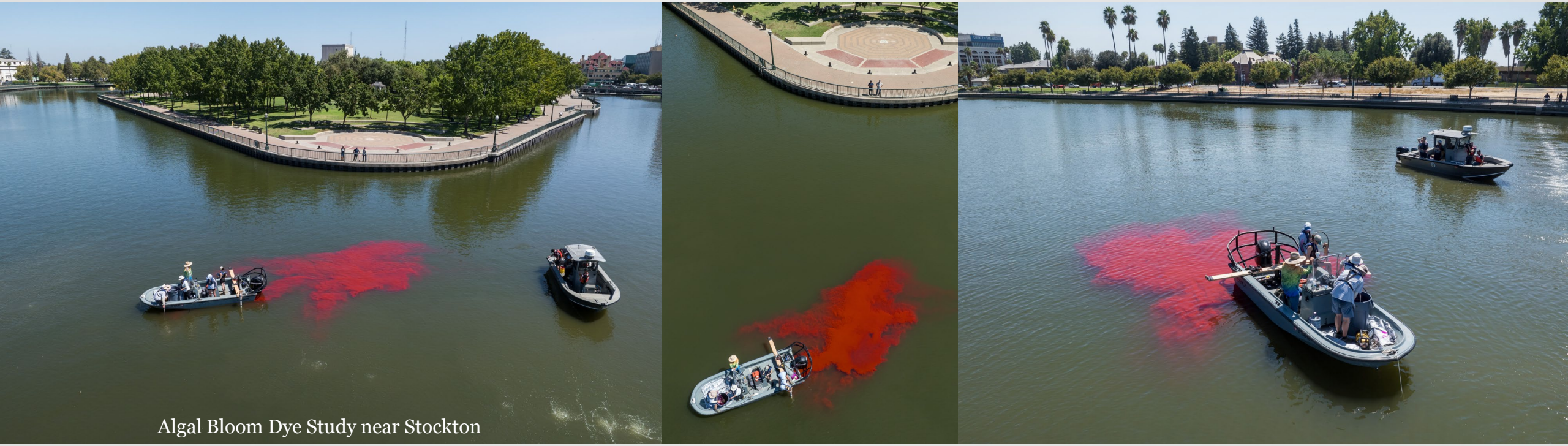
CALIFORNIA DEPARTMENT OF WATER RESOURCES

Machine Learning-Based Harmful Algal Blooms (HABs) Modeling in the Delta (Phase 2)

Data Science PWT, November 13, 2025



Algal Bloom Dye Study near Stockton

CALIFORNIA DEPARTMENT OF WATER RESOURCES

Gourab Saha, Ph.D., P.E.
Modeling Support Office, DWR

Overview

- ❑ Background
- ❑ Goal and Objectives
- ❑ Study Phases
- ❑ ML Models and Evaluation Metrics
- ❑ Study Workflow
  - ✓ Data Preparation
  - ✓ Model Development
  - ✓ Results Analysis
  - ✓ Initial Observations
- ❑ Future Directions and Final Thoughts

San Luis Lake (an example image)

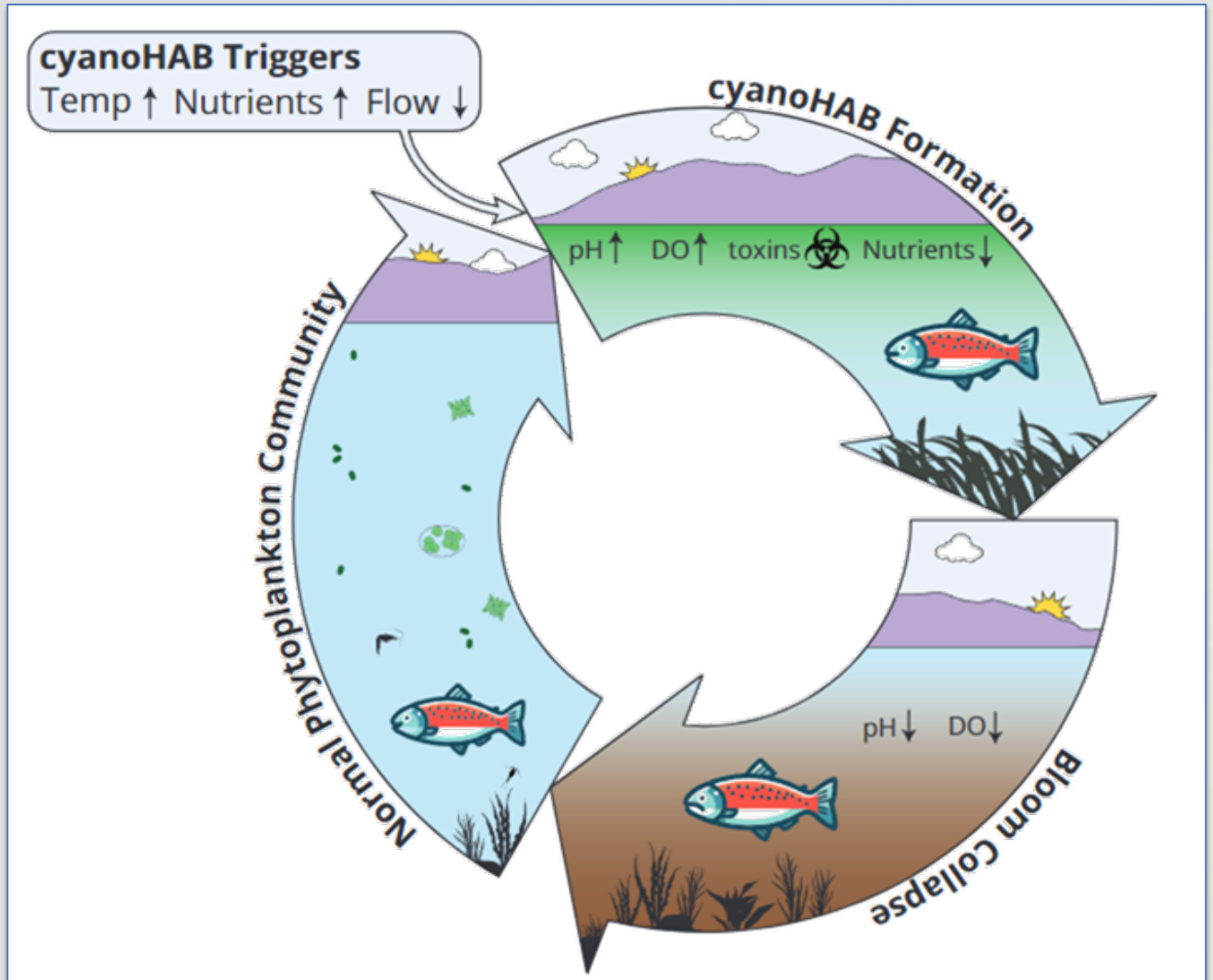
Background

Photo Courtesy: California Water Quality Control Board

- Growth and existence duration of Cyanobacteria increases with changing climatic condition in the Sacramento – San Joaquin Delta (Delta).

- Cyanobacteria produced toxins affect fish, pets, wildlife, and humans, especially fishermen and recreational swimmers (may cause liver cancer and neurological damage).

Background: CyanoHAB Cycle



Conceptual Model of factors hypothesized to trigger harmful algal blooms in the Delta

Source: Bouma–Gregson et al. 2024

- Nutrient availability, low flow, and favorable water temperature influence the chance of Cyanobacteria bloom abundance in the Delta.

Background: CA Recreational Guidelines (Voluntary)

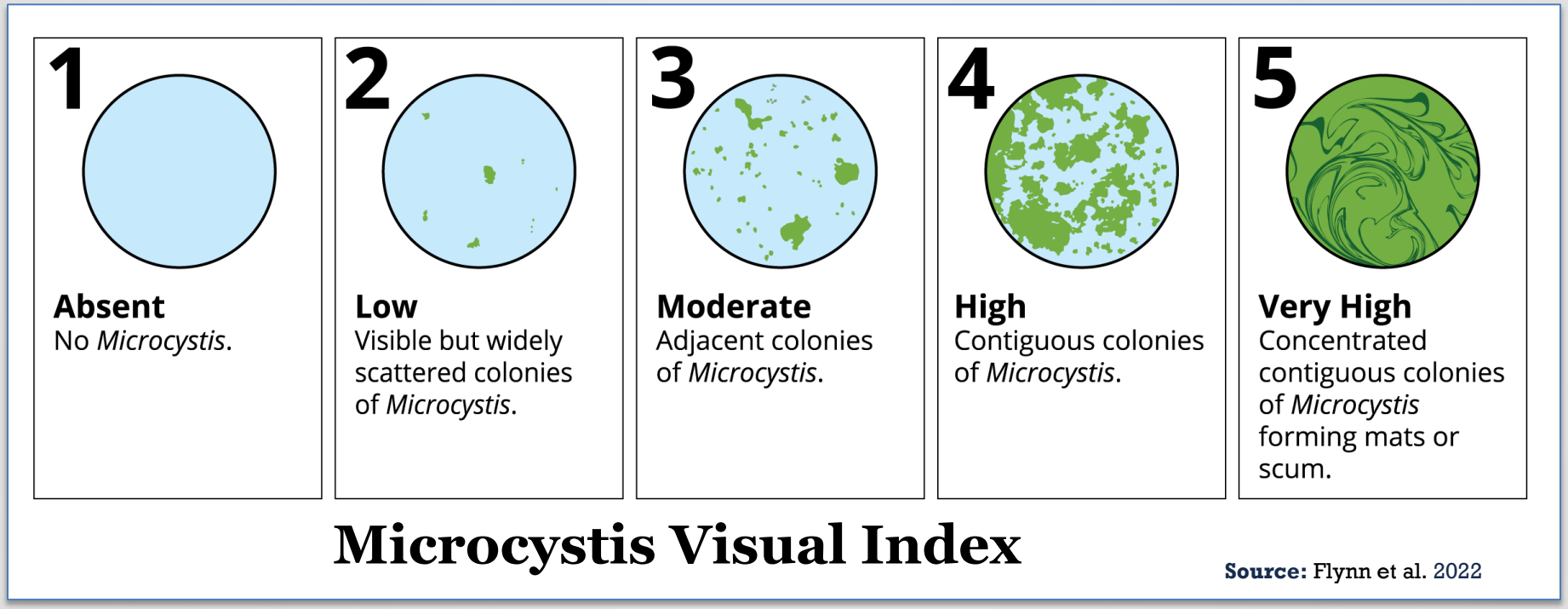
- ✓ The CA recreation advisories indicate toxigenic Cyanobacterial Abundance (***Microcystis***) **4,000 cells/ml** as **trigger for caution.**

Trigger Levels For Human and Animal Health				
Criteria*	No Advisory [a]	Caution (TIER 1)	Warning (TIER 2)	Danger (TIER 3)
Total Microcystins [b]	< 0.8 µg/L	0.8 µg/L	6 µg/L	20 µg/L
Anatoxin-a	Non-detect [c]	Detected [c]	20 µg/L	90 µg/L
Cylindrospermopsin	< 1 µg/L	1 µg/L	4 µg/L	17 µg/L
Cell Density of potential toxin producers	< 4,000 cells/mL	4,000 cells/mL	_____	_____

Source: https://mywaterquality.ca.gov/habs/resources/habs_response.html#advisory_signs_guidance

Background: Microcystis Visual Index

❑ A visual scale, named **Microcystis Visual Index (MVI)**, was developed based on photographic and visual observations by Environmental Monitoring Program to monitor surface cyanobacteria colonies (Flynn et al. 2022).



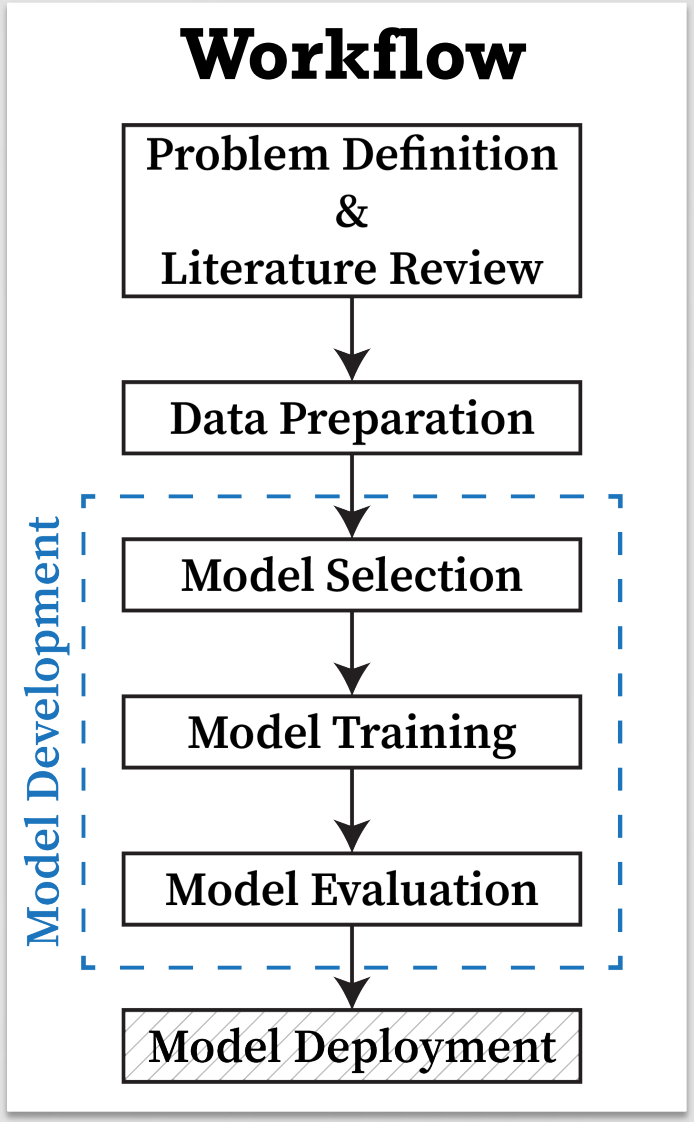
Goal and Objectives

❑ Goal

- ✓ Develop a **machine learning (ML)** based **HABs** modeling **tool** for **Delta**.

❑ Study Objectives

- ✓ To predict HAB using ML techniques in the Delta.
- ✓ To create **a dashboard** to predict HAB at user-defined locations.



Machine Learning Protocols

Study Phases

Phase 1

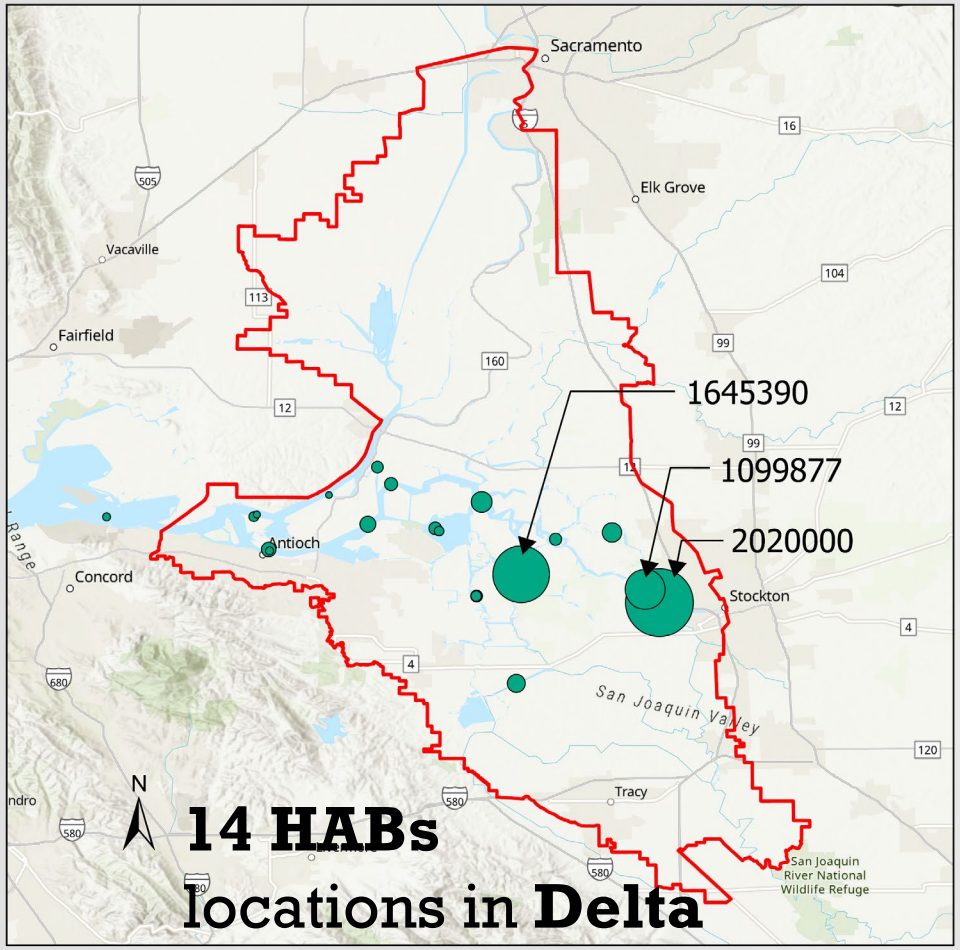
- Predict HAB Risk (based on toxigenic Microcystis cell count)
- Sample size: 220 samples

Phase 2

- Predict HAB Status (based on Microcystis Visual Index)
- Sample size: 1006 samples

Phase 1

Phase 1: Study Data



Microcystis
(Avg. Cells/ml)

- 10,000
- 50,000
- 100,000
- 500,000
- 1,000,000

0 10 20 40 Km

- ✓ Data Period: **2014 – 2019**
- ✓ Data Availability: samples of **220 days**

[Most of these samples were collected during Summer and Fall.]

Environmental factors

- ❖ Water Temperature
- ❖ Salinity
- ❖ pH
- ❖ DO

Nutrient factors

- ❖ Dissolved Ammonia
- ❖ Dissolved Org. Nitrogen
- ❖ Dissolved Phosphate
- ❖ Total Phosphorus

Physical Processes

- ❖ Antecedent Flow (daily average flow, a month back before Microcystis data collection date)

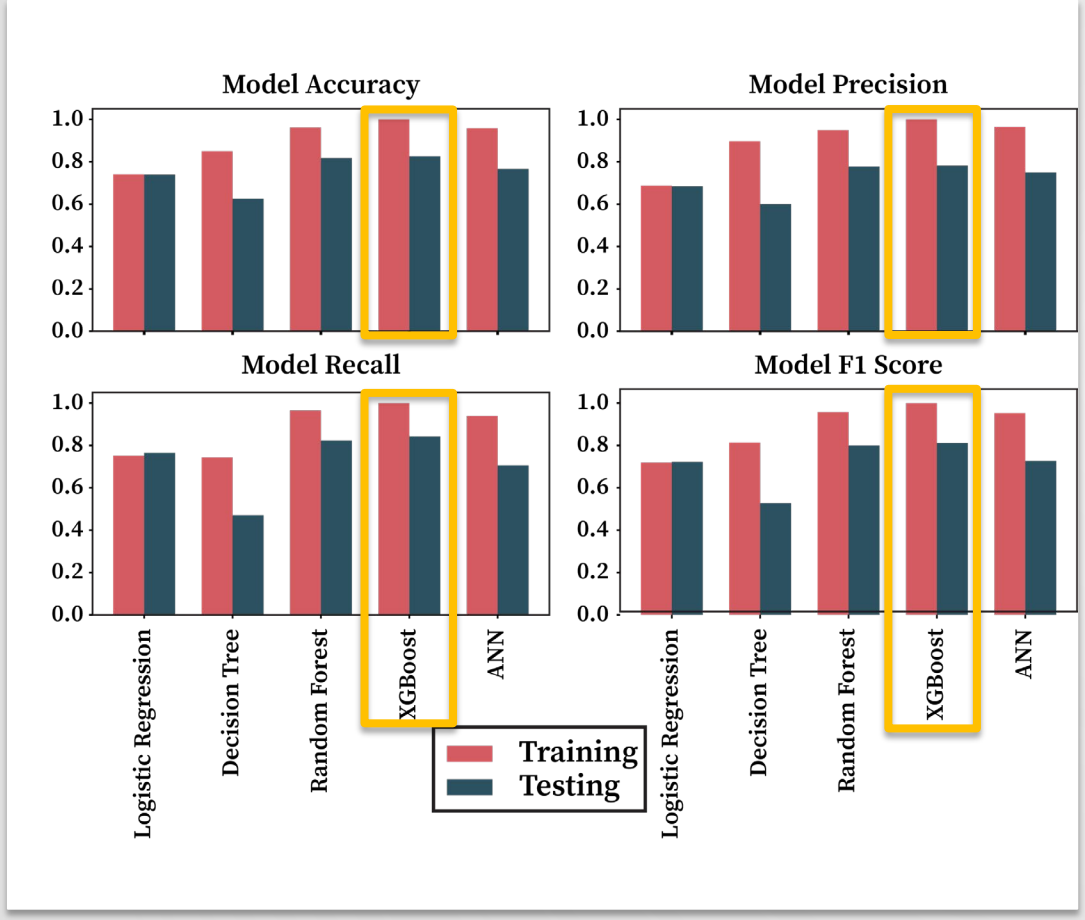
Data Sources

Data collected under **Cyanobacteria and water quality data collection for the upper San Francisco Estuary** project (DWR-UCD Dataset).

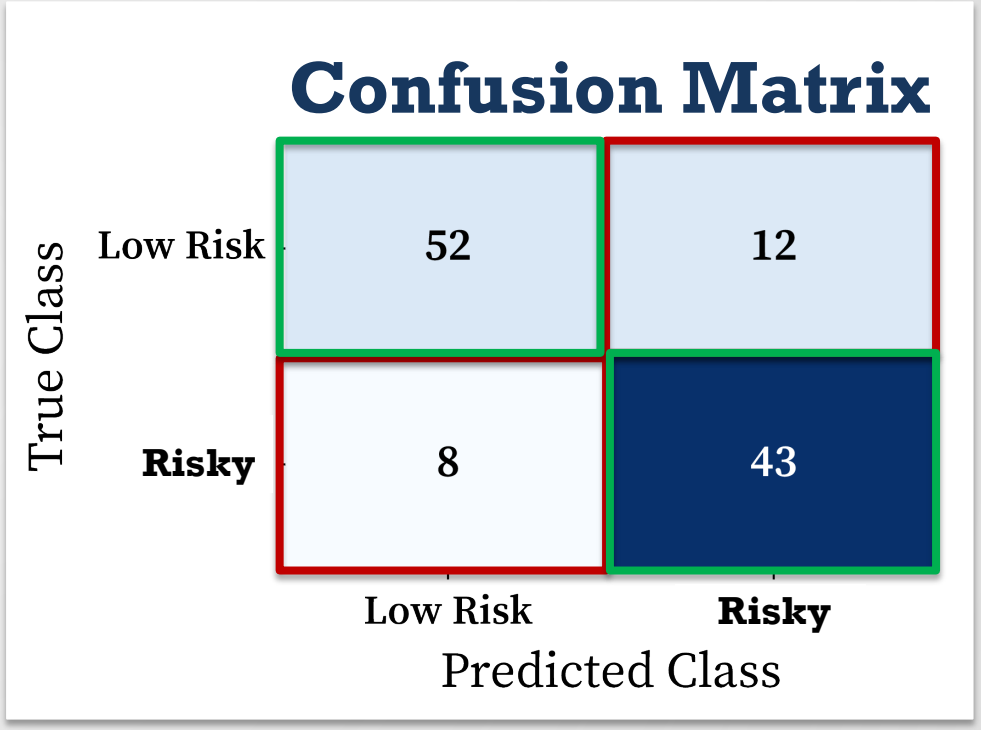
(Source: https://portal.edirepository.org/nis/mapbrowse?packageid=edi.1076.1)

Phase 1: Model Results

- **XGBoost** model **most accurately** predicted **Risk** of ***Microcystis*** (Avg. cells/ml) among five considered ML models.



- XGBoost model correctly predicts the Low Risk and Risky (caution) categories 81% and 84% of occasions, respectively, on the testing dataset.



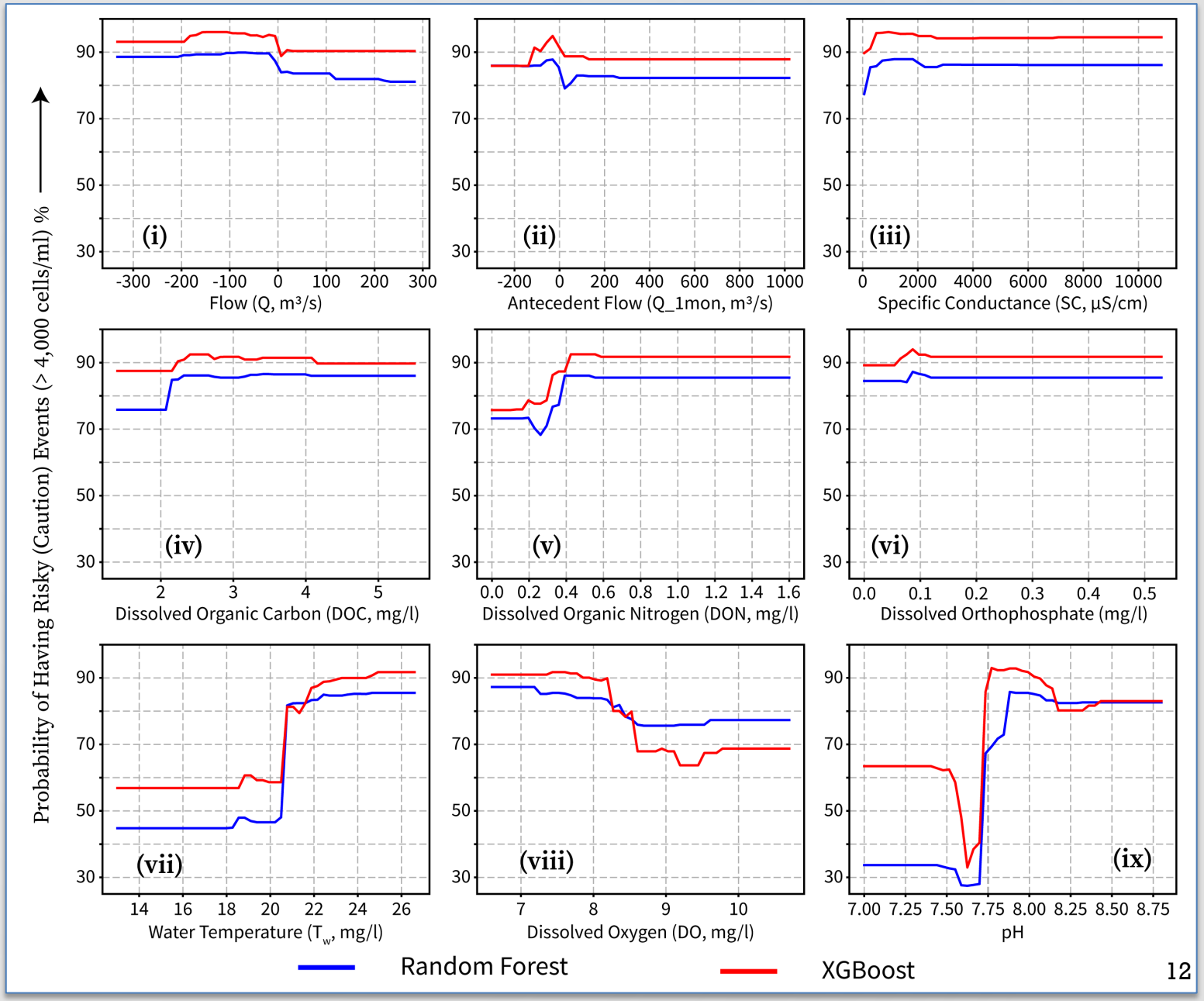
- 64 Testing Samples for **True Low-Risk** Class
- 51 Testing Samples for **True Risky** Class

Phase 1: ML Model-Based Sensitivity

- Identified each environmental variable's sensitivity by keeping all other input variables fixed.

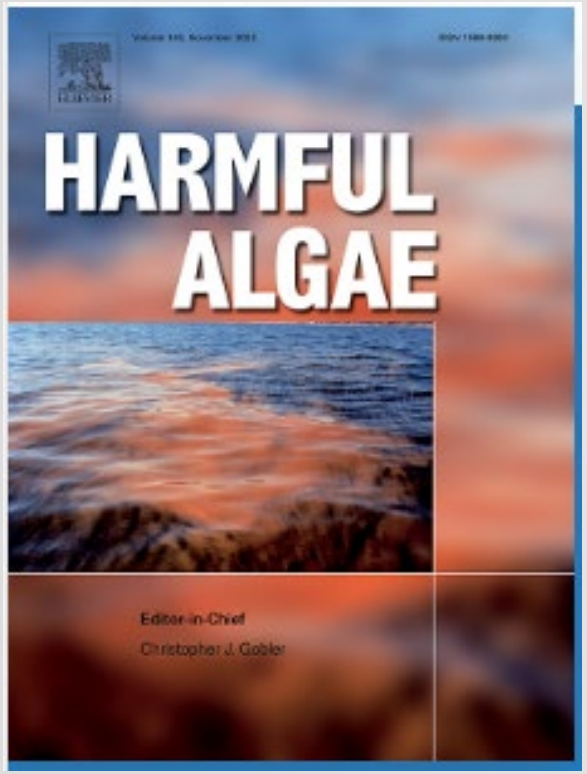
- Identified Water Temperature, pH, and Dissolved Oxygen as highly sensitive.

- pH and Dissolved Oxygen are consequence of the bloom.



CALIFORNIA DEPARTMENT OF WATER RESOURCES

Phase 1: Manuscript under Review

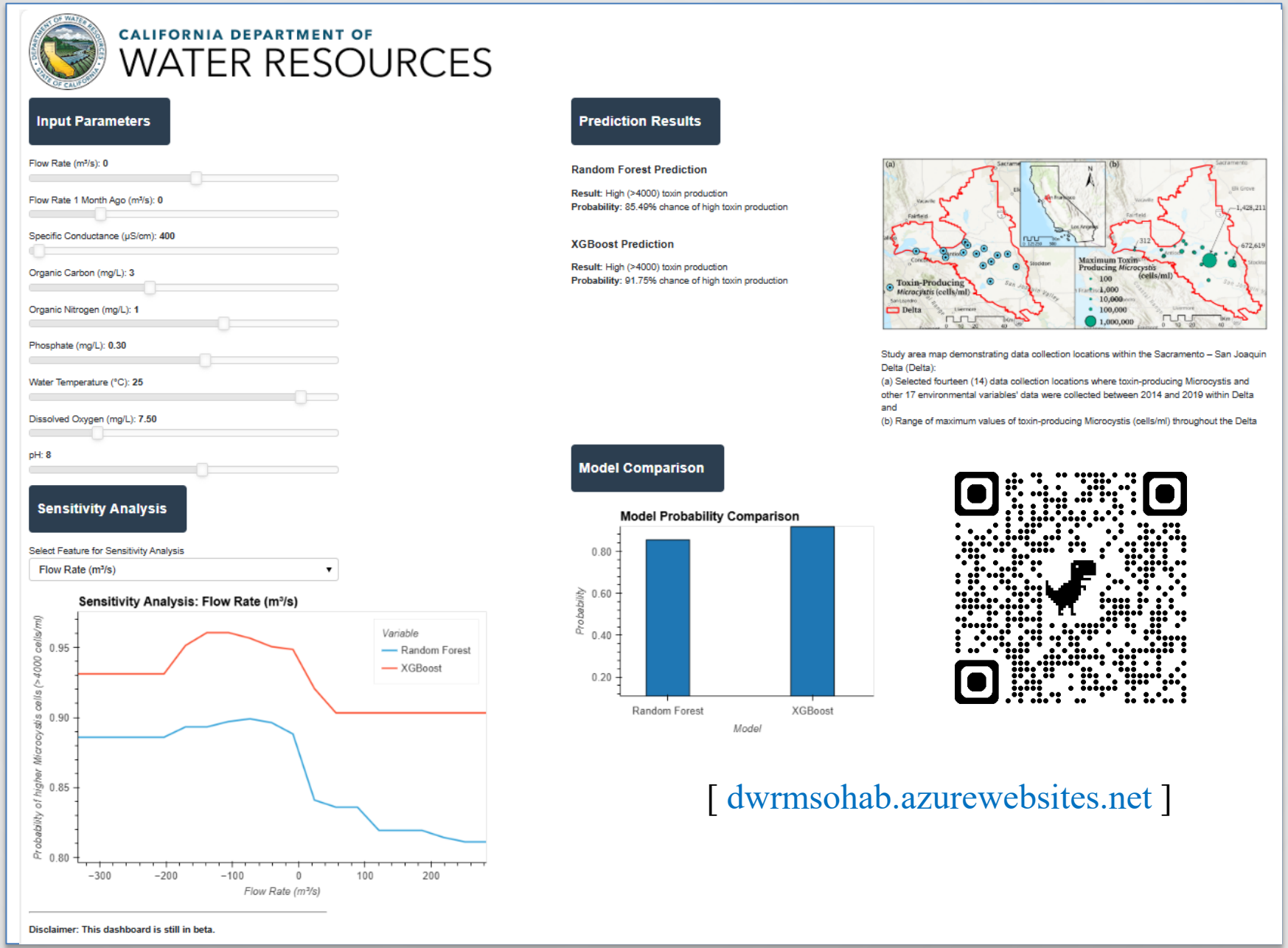


- ❑ We have submitted a manuscript from the HAB ML modeling (Phase 1) work.
- ❑ Currently, the manuscript is under review.

Phase 1: Dashboard

- An interactive dashboard was created.

My colleague, Peyman Namadi, will provide a demo.

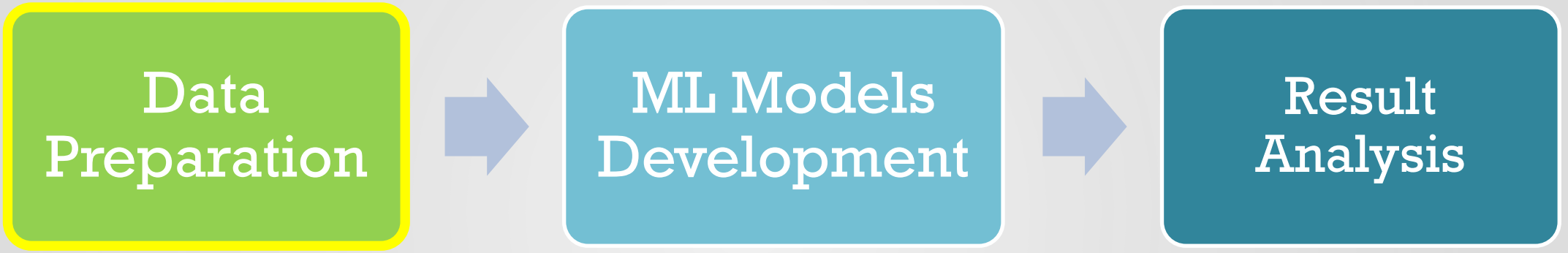


[dwrmsohab.azurewebsites.net]

Phase 2

Study Workflow

ML = Machine Learning



ML = Machine Learning

Phase 2: Study Data

- 35 Sites (168 locations)

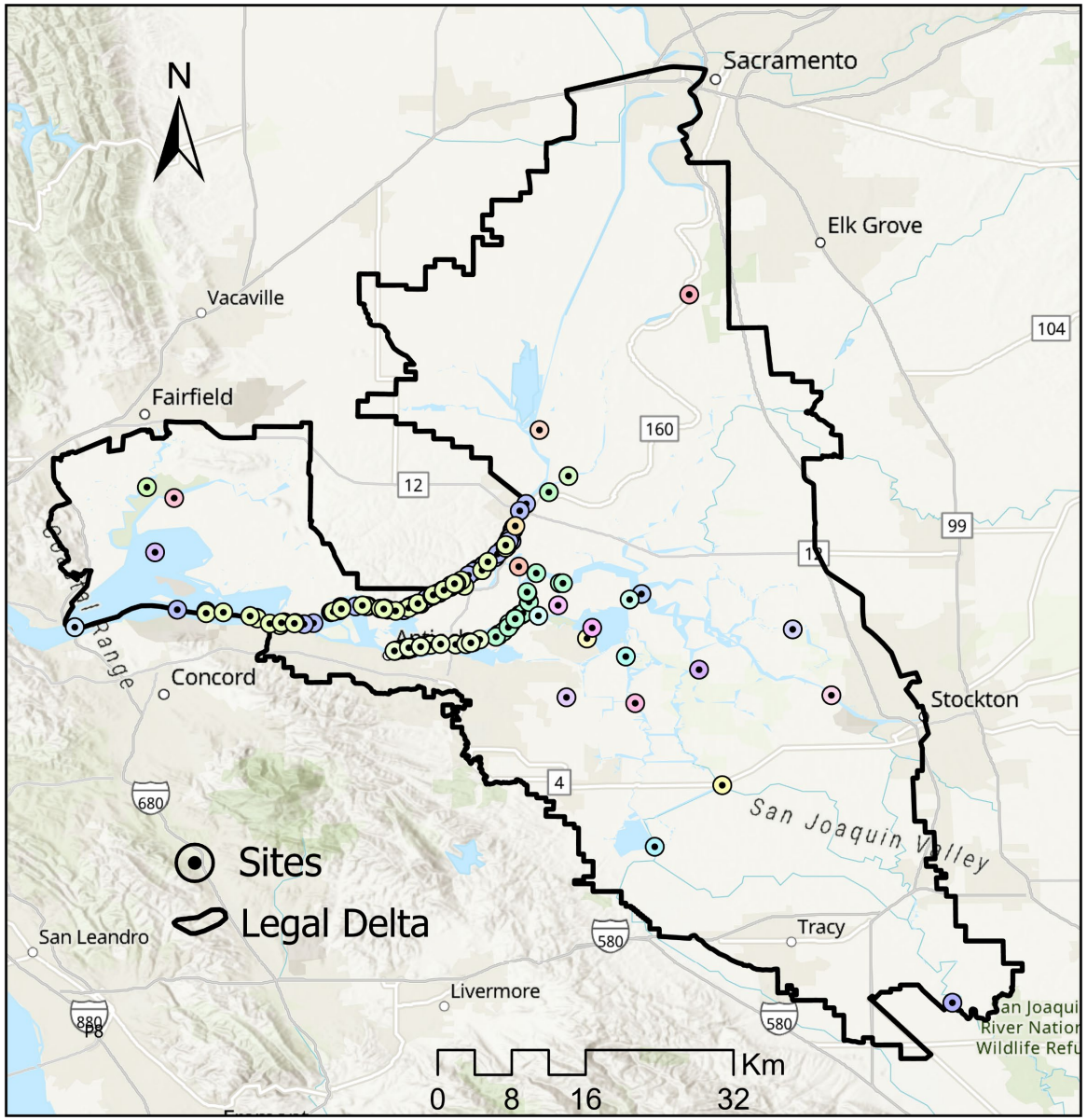
  (4 floating sites' data was collected from 137 locations. Remaining 31 sites' data was collected from their designated locations.)

- Data Count: 1,006 samples

  (Data collected during summer and fall has been used.)

- Data Period: 2014 - 2022

Note: Data collected from colleagues from Division of Integrated Science and Engineering



Input Data

✓ Data available for 8 environmental variables that influence **HABs**.

Water quality factors

- Water temperature
- Specific conductivity

Physical processes

- Antecedent flow
- Antecedent velocity

} (DSM2 Simulated)

Nutrient factors

- Dissolved ammonia
- Dissolved nitrate & nitrite
- Dissolved orthophosphate

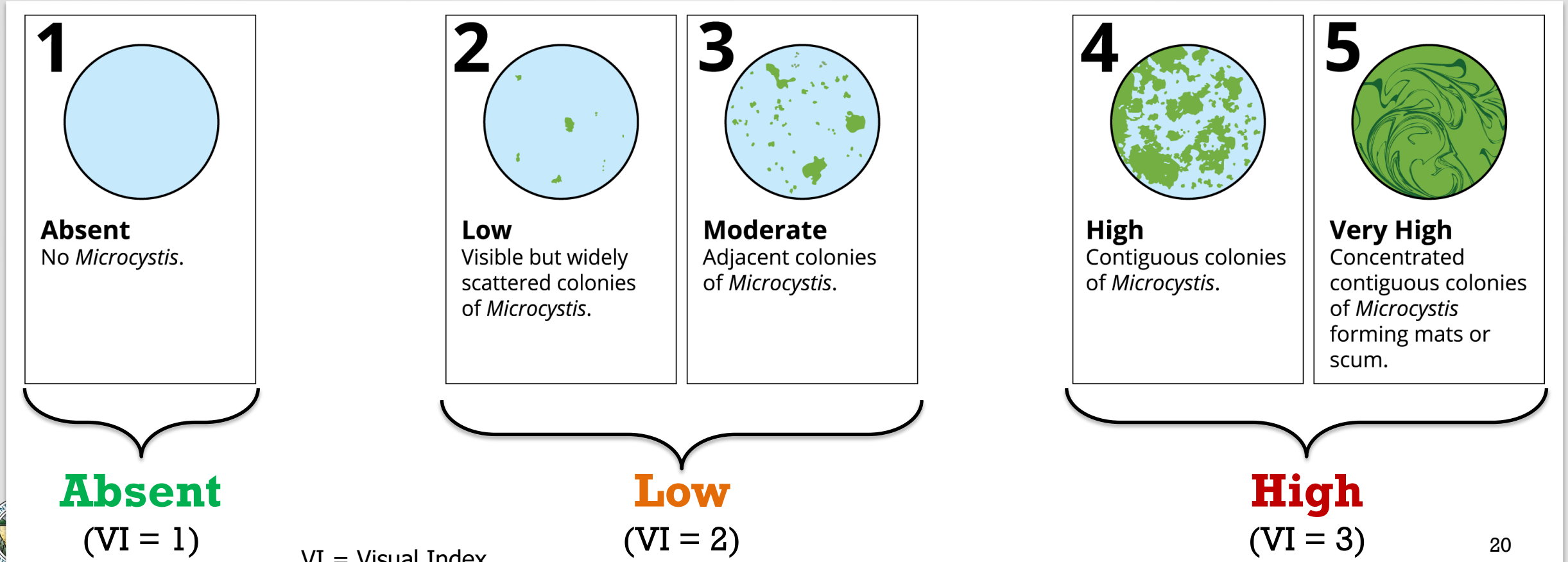
Light availability

- Turbidity

Phase 2: Target Data

Target variable for ML based HAB modeling – Microcystis Visual Index.

- 5 categories of **Microcystis Visual Index scale** converted to 3 categories scale recommended in the Bouma-Gregson et al. (2024).

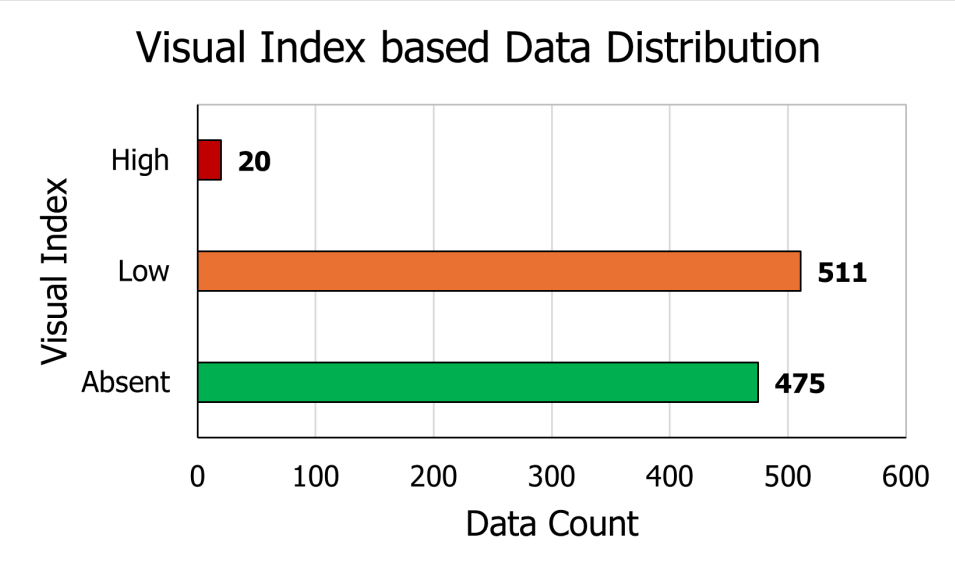


VI = Visual Index

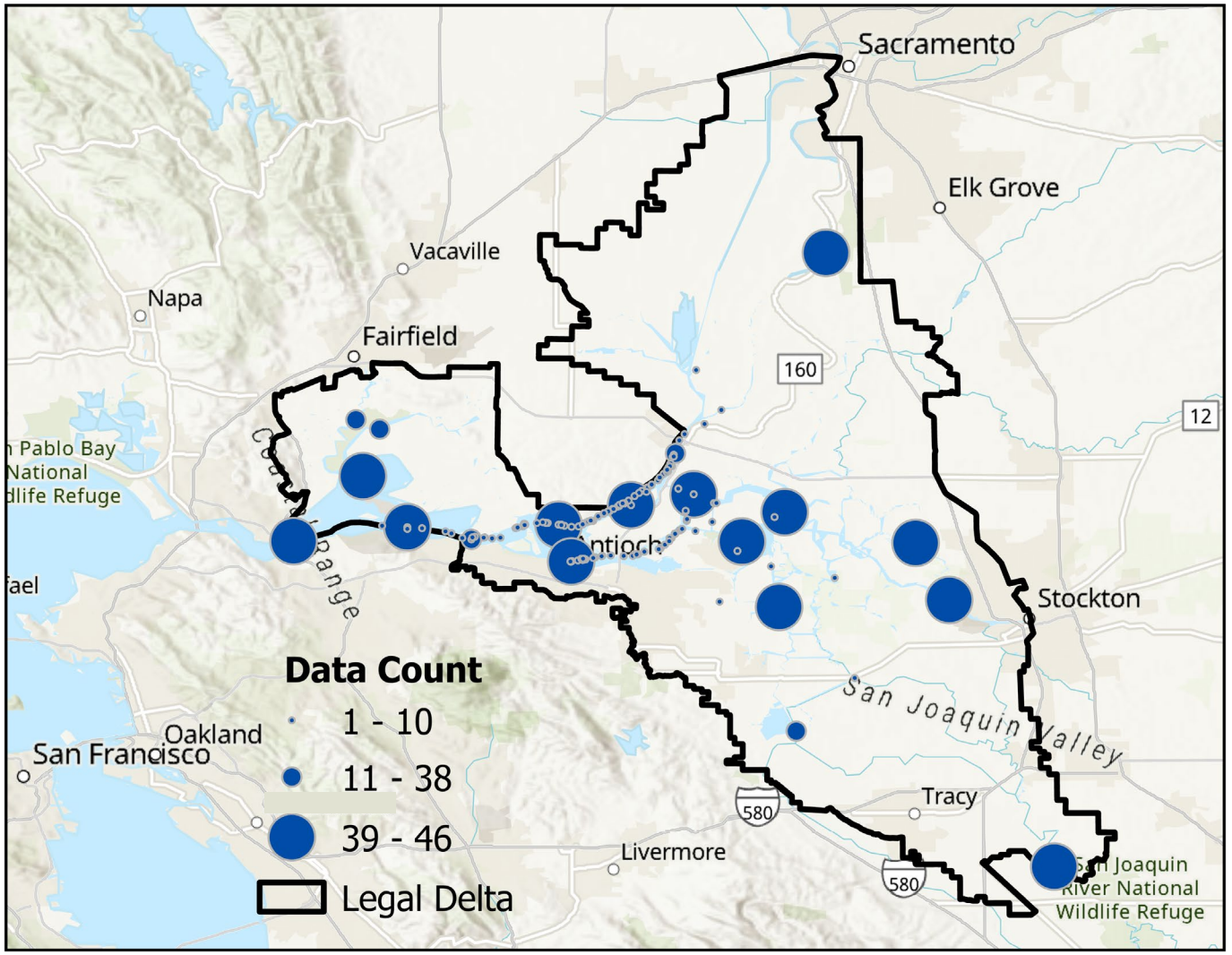
Source: https://doi.org/10.15447/sfews.2024v22iss1art2

Target Data Distribution

- Imbalance exists in Visual Index category-based data distribution.



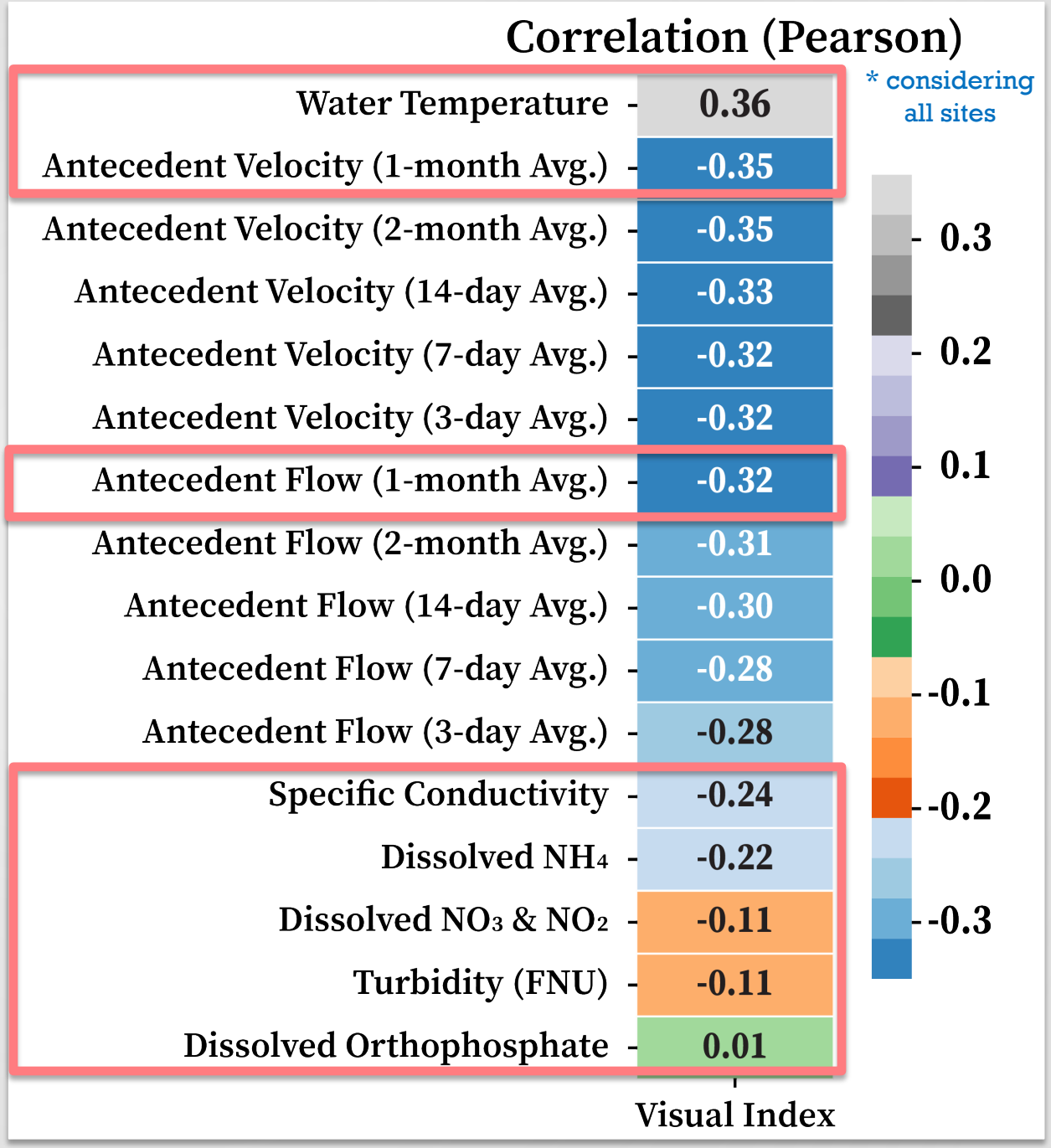
Number of Data per Site



CALIFORNIA DEPARTMENT OF WATER RESOURCES

Data Analysis

- Positive correlation: increasing water temperature, high probability of having more HABs.
- Negative correlation: high velocity, less probability of having more HABs.
- Antecedent Flow (1-month Avg.) and Velocity (1-month Avg.) had high correlations and selected for machine learning models development.

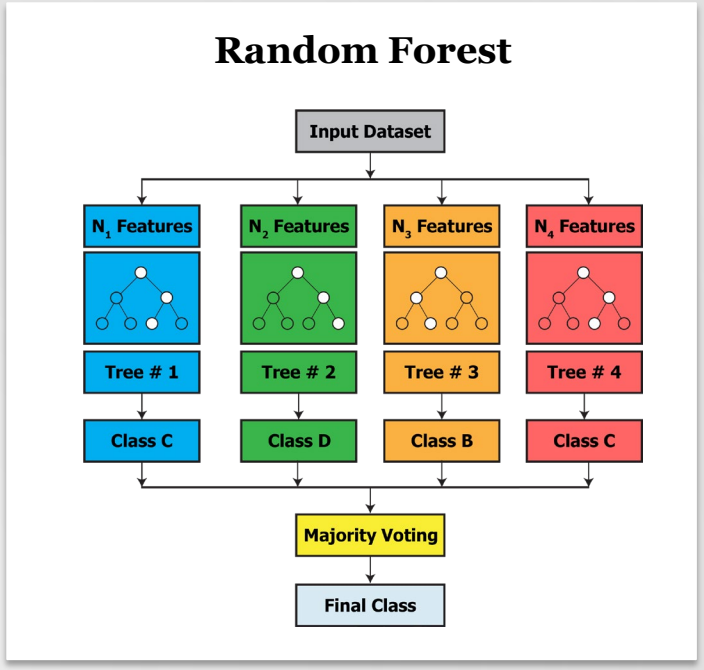


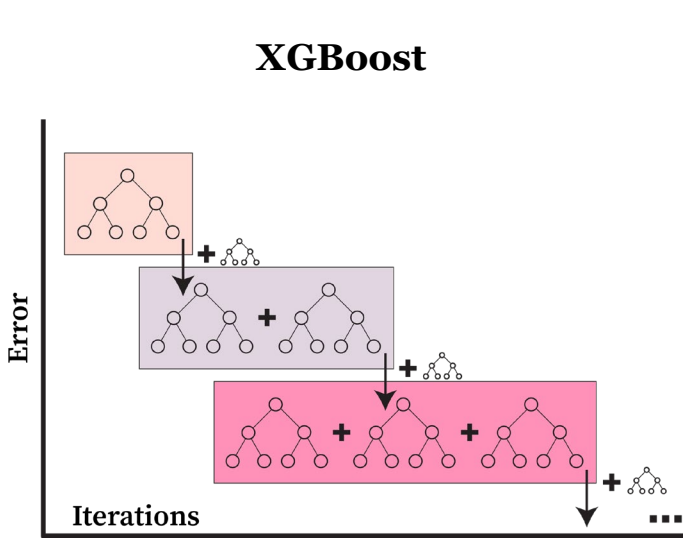
*Antecedent Flow and Velocity are DSM2 simulated values.

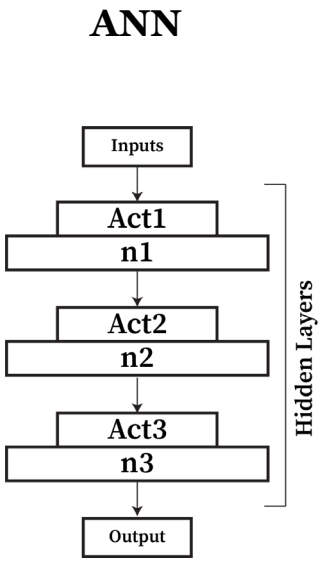
ML = Machine Learning

Model Selection

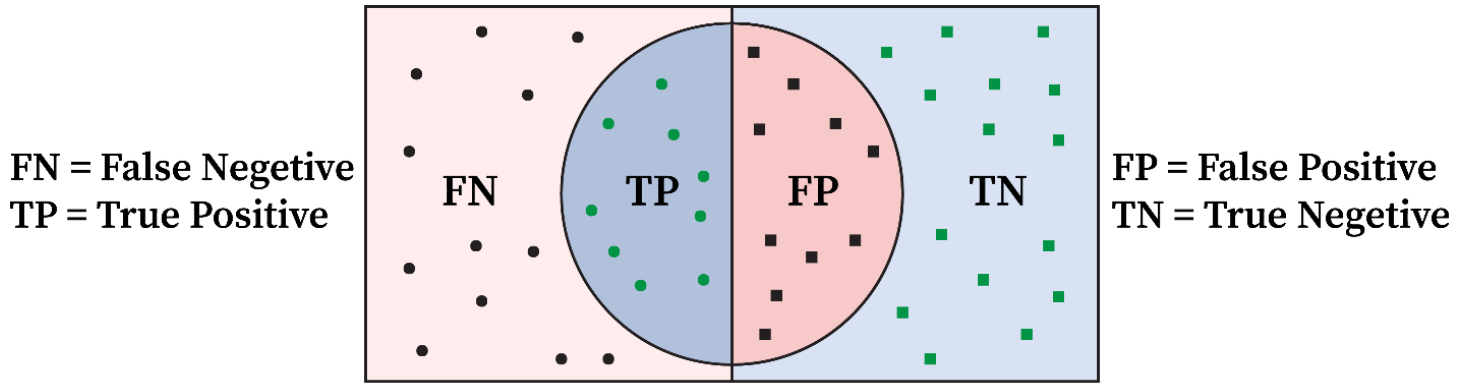
- ❑ Selected ML models for HABs model development
  - ✓ Random Forest (RF)
  - ✓ XGBoost
  - ✓ Artificial Neural Network (ANN)

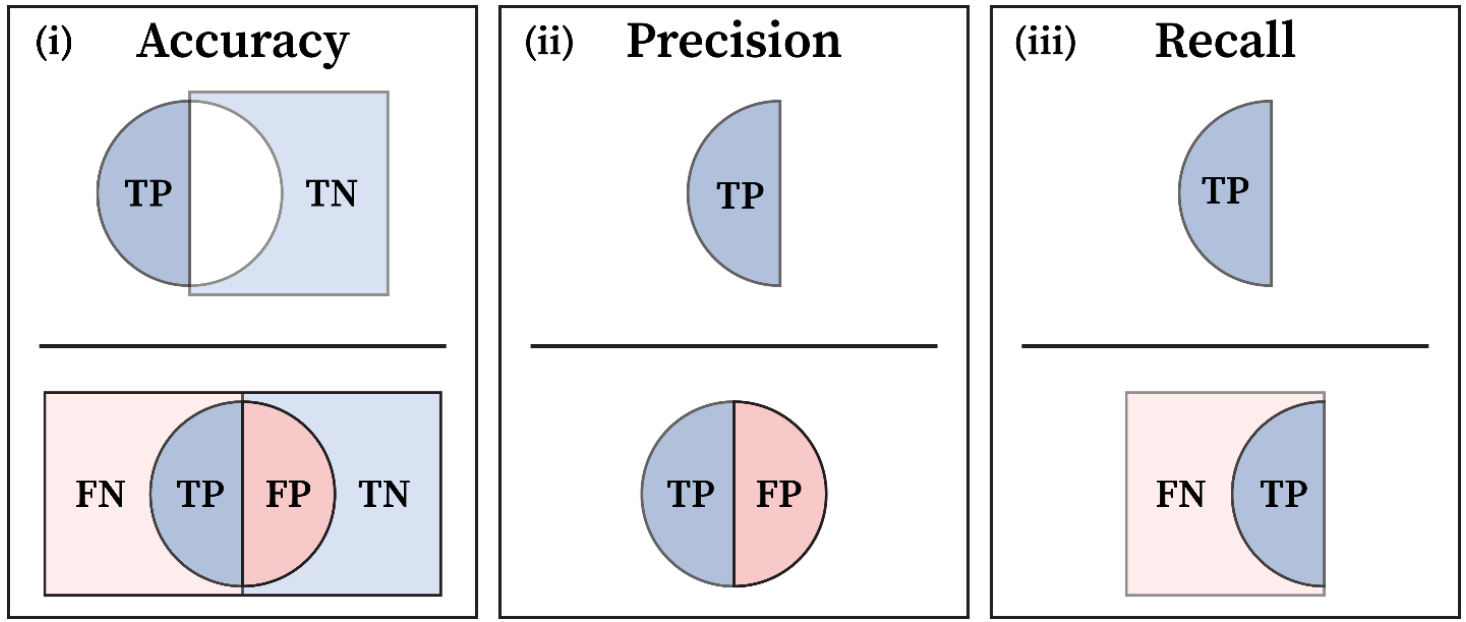






Model Performance Evaluation Metrics



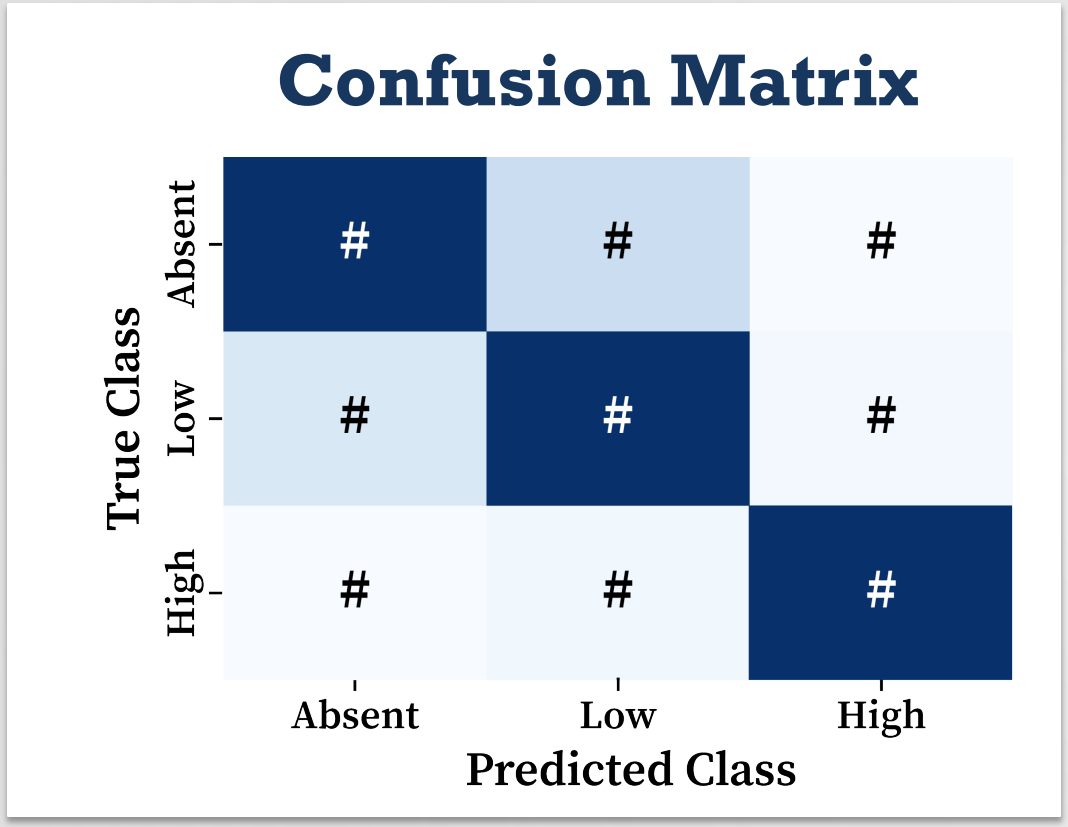


- ✓ **Model F1 Score**: A balance between precision and recall showing overall model performance.

  [**No figure for Model F1 Score*]

- ✓ **Model Recall**: Out of all true cases in a category, how many were correctly predicted.
- ✓ **Model Precision**: Of the predictions made for a specific category, how many were actually correct.
- ✓ **Model Accuracy**: How often the model's predictions are correct overall.

Confusion Matrix Definition

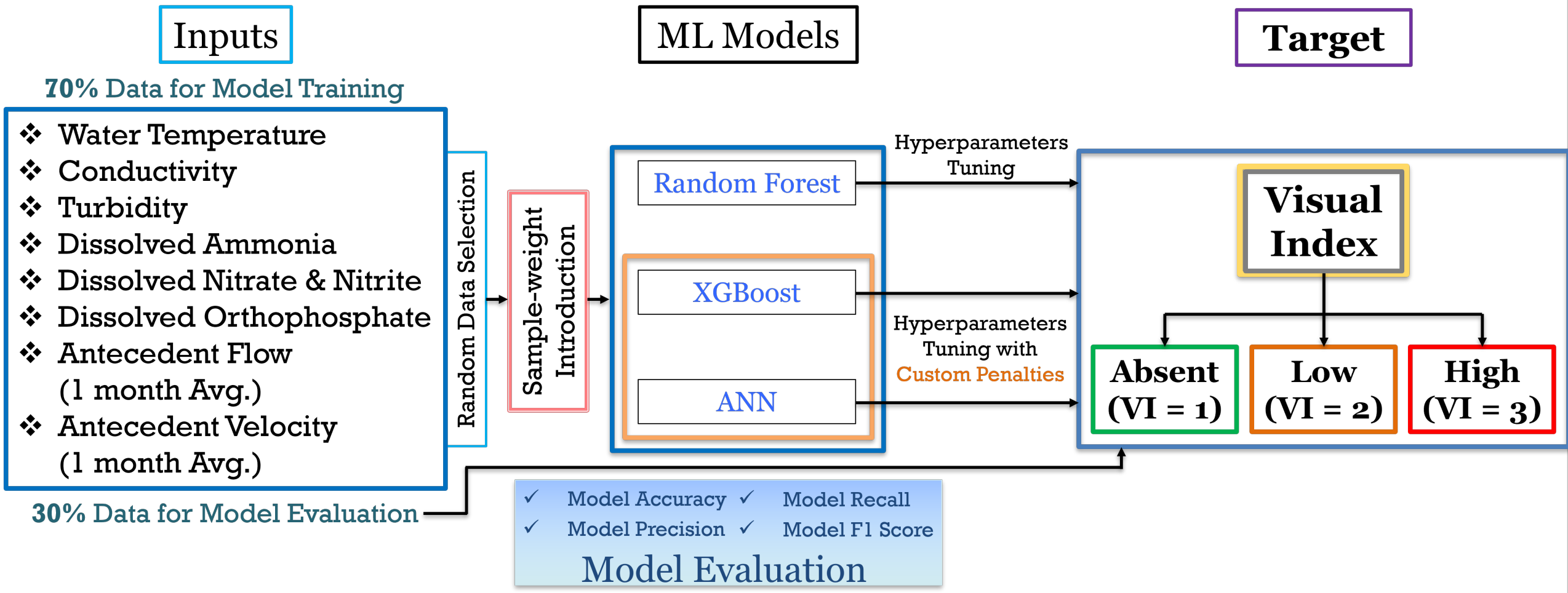


A confusion matrix is a **table** that compares predicted and actual values for a dataset to **evaluate the performance** of a classification model in machine learning.

True Class means the **Observed Values**.

Predicted Class means the **ML Model Predicted Values**.

Phase 2: Model Development

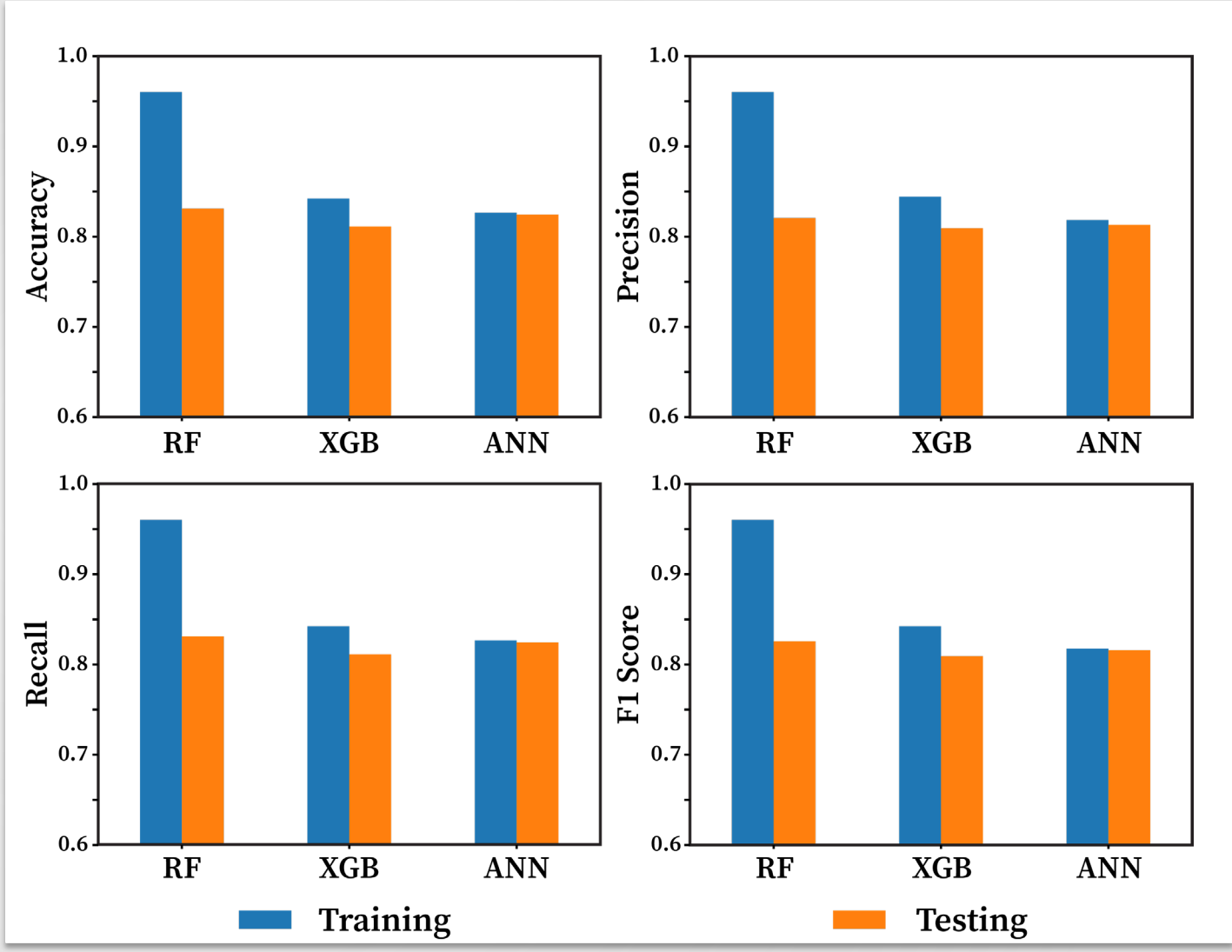


Workflow

ML = Machine Learning

Phase 2: Model Results

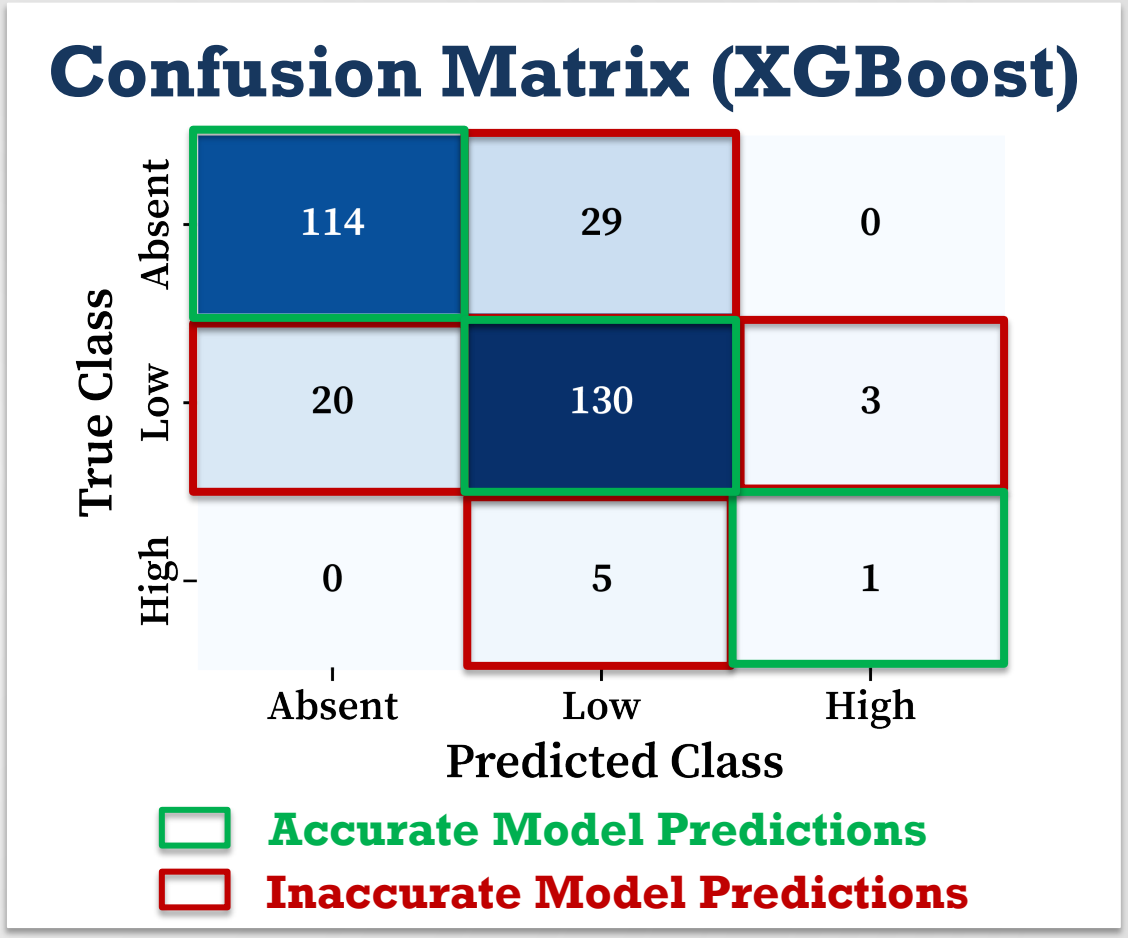
- All three ML models, including Random Forest, XGBoost, and ANN, demonstrated similar Visual Index prediction performance with test accuracy 0.83, 0.81, and 0.82, respectively.



Phase 2: XGBoost Results (An example)

- XGBoost model correctly predicts the Absent HAB category 80% of the time (114 out of 143 times) on the testing dataset.
- On 85% (130 out of 153 times) of occasions, the model predicts the Low HAB category correctly.

Most of High HAB category prediction by the selected Machine Learning models was inaccurate, which require further investigation.



- 143 Testing Samples for **True Absent HAB** Class
- 153 Testing Samples for **True Low HAB** Class
- 6 Testing Samples for **True High HAB** Class

Phase 2: Observations

- All three ML models predicted **Visual Index (VI)** with test accuracy 0.83 (Random Forest), 0.81 (XGBoost), and 0.82 (ANN).

- As an example, XGBoost model accurately predicted Absent HAB on 80% (114 out of 143) of occasions and Low HAB on 85% (130 out of 153 of occasions).

- All three models struggled to predict High **HAB**. These models can differentiate between 'Absent HAB' and 'HAB (Low + High)'. However, the models cannot differentiate between 'Low HAB' and 'High HAB' conditions.

Future Directions

- Investigate more about the **Microcystis Visual Index** data, agencies collect the data, and data collection methods.
- Feature engineering to improve the input features selection for Visual Index prediction.
- Develop an interactive dashboard that enables users to instantly simulate **Microcystis Visual Index** under user-defined environmental conditions.

Final Thoughts

The study demonstrates the opportunity to extend machine learning-based Harmful Algal Bloom modeling throughout the Delta and the State.

- ❑ HAB modeling (Phase 2) project is still on going and we are open to new suggestions.
- ❑ The developed machine learning model will be shared at **#DeltaDash.**
- ❑ Make interested parties aware of our modeling efforts and future data requirements.
- ❑ Create a symbiotic relationship among agencies to **monitor** and **restrict harmful algal blooms**.

CALIFORNIA DEPARTMENT OF
WATER RESOURCES

DELTA DASH

Acknowledgements

- **Modeling Support Office, DWR**
- **Ellen Preece, Rosemary Hartman, Shaun Philippart, Silvia Angles**, and **Daphne Gille, DWR**
- **Leslie Palencia, MWQI**
- **Keith Bouma-Gregson, USGS**

References

1. Flynn, T., Lehman, P., Lesmeister, S., and Waller, S. 2022. A Visual Scale for Microcystis Bloom Severity. *https://doi.org/10.6084/m9.figshare.19239882.v1*

2. Bouma–Gregson K, Bosworth D H, Flynn T M, Maguire A, Rinde J, and Hartman R. 2024. Delta Blue(green)s: The Effect of Drought and Drought-Management Actions on Microcystis in the Sacramento–San Joaquin Delta. San Francisco Estuary and Watershed Science. *https://doi.org/10.15447/sfews.2024v22iss1art2*

- Gourab.Saha@water.ca.gov
- Peyman.Hosseinzadehnamadi@water.ca.gov
- Kevin.He@water.ca.gov

Thank You!

CALIFORNIA DEPARTMENT OF WATER RESOURCES

Phase 1: Dashboard

[dwrmsohab.azurewebsites.net]

CALIFORNIA DEPARTMENT OF
WATER RESOURCES

Input Parameters

Flow Rate (m³/s): **0**

Flow Rate 1 Month Ago (m³/s): **0**

Specific Conductance (µS/cm): **400**

Organic Carbon (mg/L): **3**

Organic Nitrogen (mg/L): **1**

Phosphate (mg/L): **0.30**

Water Temperature (°C): **25**

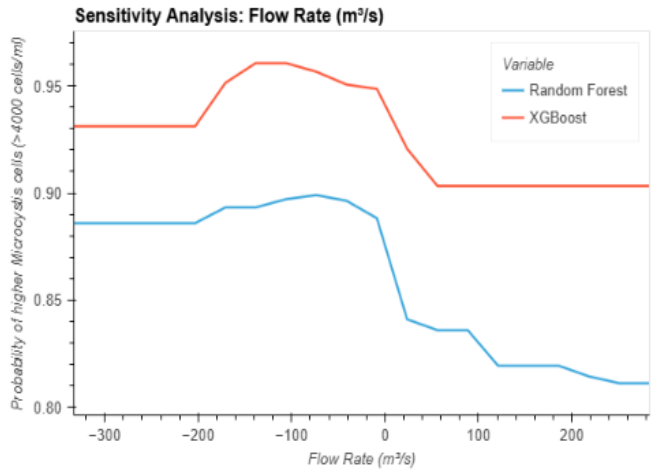
Dissolved Oxygen (mg/L): **7.50**

pH: **8**

Sensitivity Analysis

Select Feature for Sensitivity Analysis

Flow Rate (m³/s)



Disclaimer: This dashboard is still in beta.

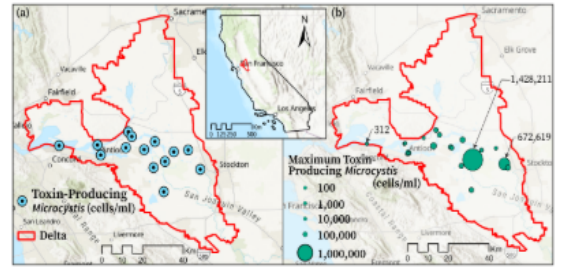
Prediction Results

Random Forest Prediction

Result: High (>4000) toxin production
Probability: 85.49% chance of high toxin production

XGBoost Prediction

Result: High (>4000) toxin production
Probability: 91.75% chance of high toxin production



Study area map demonstrating data collection locations within the Sacramento – San Joaquin Delta (Delta):
(a) Selected fourteen (14) data collection locations where toxin-producing Microcystis and other 17 environmental variables' data were collected between 2014 and 2019 within Delta and
(b) Range of maximum values of toxin-producing Microcystis (cells/ml) throughout the Delta

Model Comparison