Machine Learning-Based Harmful Algal Blooms (HABs) Modeling in the Sacramento-San Joaquin Delta

Gourab Saha, Peyman Namadi, Minxue He, Prabhjot Sandhu

Delta Modeling Section, Modeling Support Office, Department of Water Resources, Sacramento, CA 95814

Highlights

- **Environmental data** influencing ***Microcystis* abundance** at **16** Sacramento-San Joaquin Delta **(Delta) locations** has been considered for **Harmful Algal Blooms (HABs) modeling.**
- **Correlation** analysis showed **water temperature**, **pH**, **dissolved organic nitrogen**, and **phosphorus** had a **positive correlation**, and **antecedent flow, dissolved oxygen,** and **dissolved ammonia** had a **negative correlation** with ***Microcystis* (Average cells/ml)** .
- Initial data **analysis** inferred **water temperature** greater than **19 ⁰C, pH range** between **7.25** and **8.25, low flow, dissolved oxygen range of 7-9 mg/L,** and **increasing** availability of **organic nitrogen** and **phosphorus** contributed to higher ***Microcystis*** **abundance.**
- The **southern part** of **Delta** experienced several **HABs** episodes.
- Since ***Microcystis* (Avg. cells/ml)** data was **skewed**, the values were **log-transformed** for HABs modeling. The ***Microcystis* (Avg. cells/ml)** data was **divided** into **two** categories: (1) **Low** and (2) **High** (Caution), with **4,000 cell/ml** as the **threshold** per **California Voluntary Guidance**.
- **Five Machine Learning (ML) models**, including **logistic regression, decision tree, random forest, XGBoost**, and **artificial neural network,** were used to simulate *Microcystis* (Avg. cells/ml).
- The **70%** data was randomly selected for **models' training**. The remaining 30% of the data was used for **model evaluations**. Model **accuracy, precision, recall,** and **F1 score** were used for model evaluation.
- The **random forest (RF)** model predicted *Microcystis* (Avg. cells/ml) to be the **most accurate** among the selected five models with **0.88 accuracy, 0.89 precision, 0.99 recall,** and a **0.93 F1 score**.
- **RF model** accurately predicted **High *Microcystis* (Avg. cells/ml)** on **87%** of **occasions** and **Low *Microcystis* (Avg. cells/ml)** on **93%** of **occasions.**
- This **continuous modeling effort** will provide information to help **monitor** and **control** HABs, thereby **supporting** both the **ecological health** and the **long-term resilience** of the Bay-Delta ecosystem.

Background and Motivation

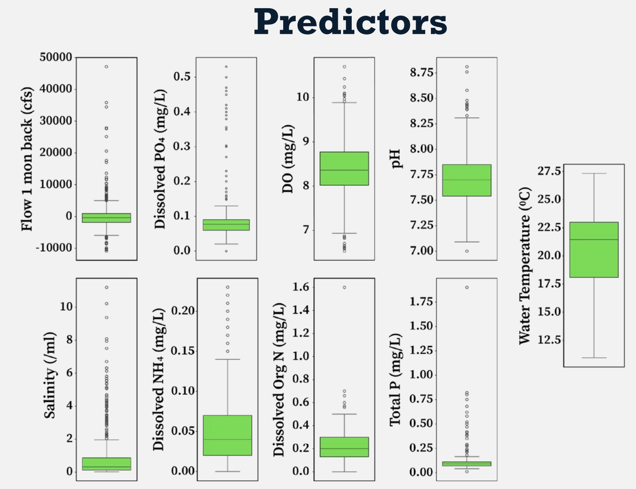
- **HABs** started becoming a critical concern for the **Delta** since **1999.**
- ***Microcystis***, a **Cyanobacteria** genus, is the primary reason behind **HABs.**
- **Prior** study **initiatives** closely **collected** relevant **data** influencing **HABs** abundance.
- This **study aims** to **statistically analyze** available HABs-related **environmental data** and **develop** an **HABs modeling framework** for the Delta.

Methods

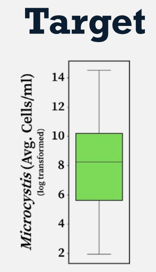
- Data Source: **CA DWR** and **UC Davis**
- Data Locations: **16 HABs** locations in **Delta**
- Data Collection Period: **2014 – 2019**
- Data Availability: **381 days**
- Predictor variables: **14**
- Target Variable: ***Microcystis* (Avg. cells/ml)**

Available ***Microcystis* (Avg. cells/ml)** data is **highly skewed**, given the variability of relevant environmental variables.

Predictors



Target



Workflow

Problem Definition & Literature Review

⬇

Data Analysis

⬇

Model Selection

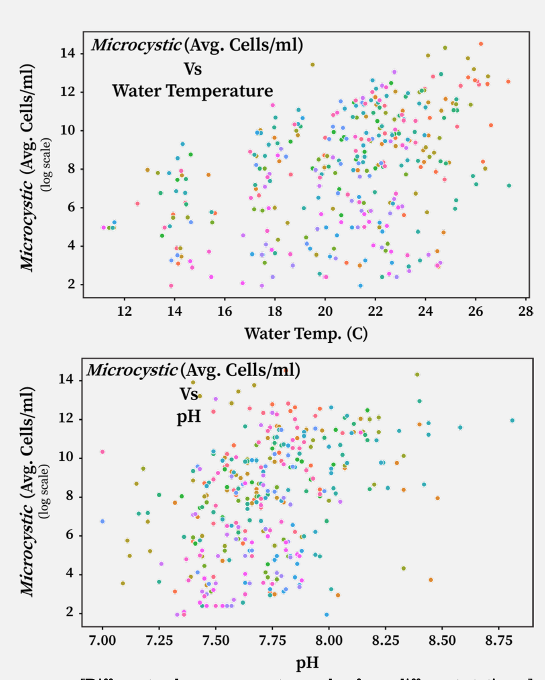
⬇

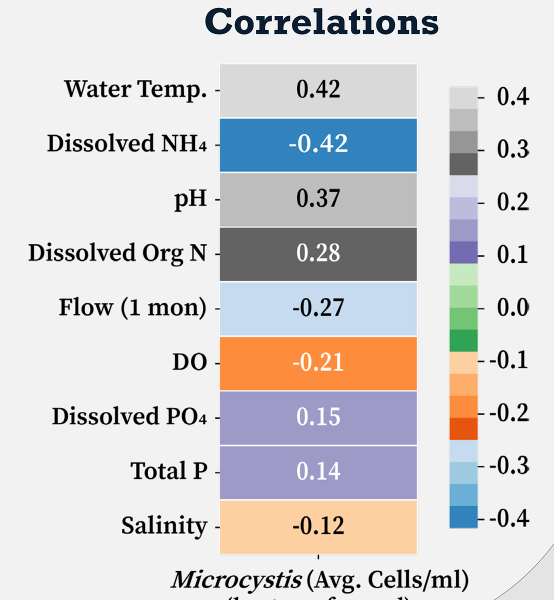
Model Development

⬇

Model Evaluation

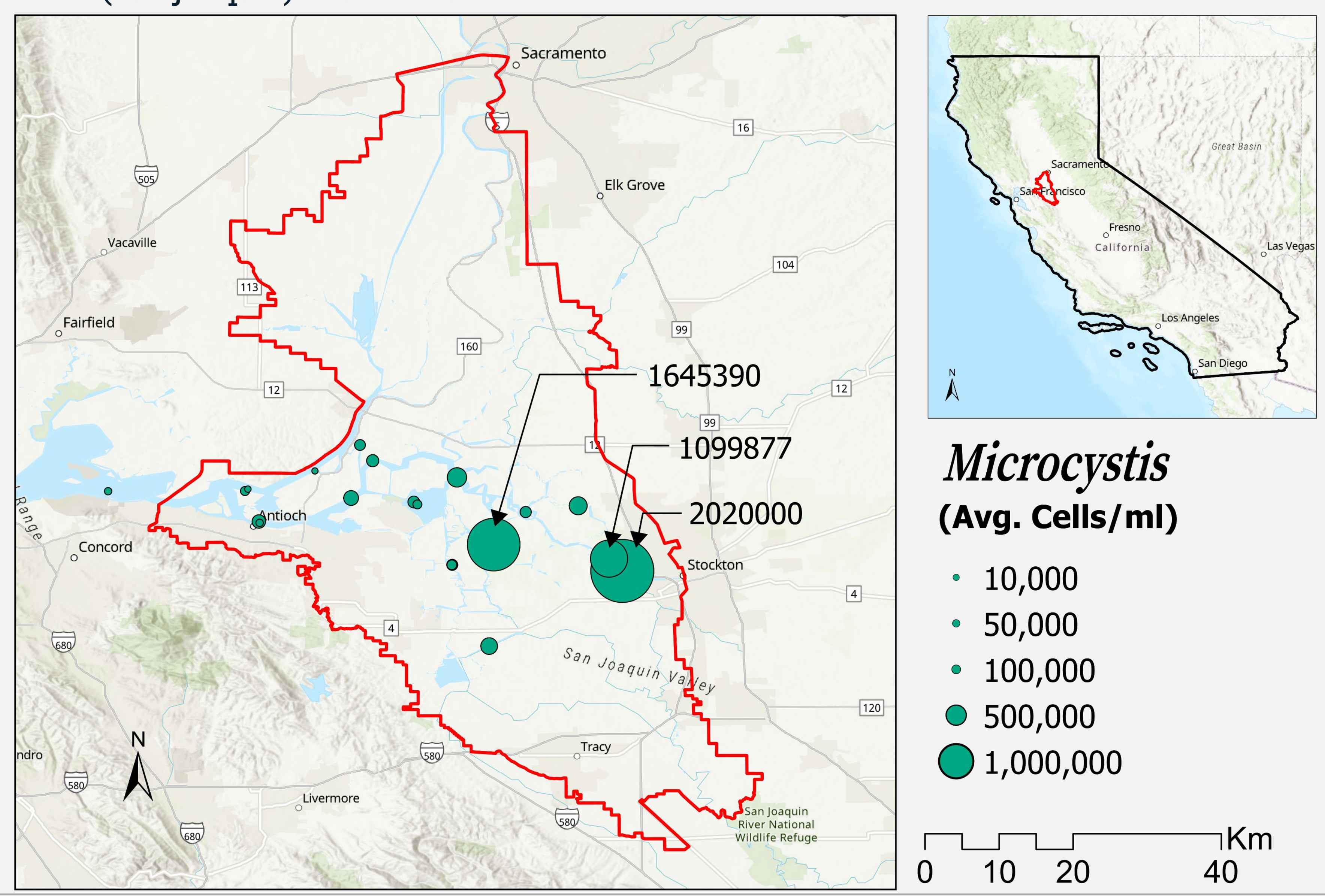
Water temperature, dissolved NH4, and pH had **higher correlation** with log transformed ***Microcystis* (Avg. cells/ml).**





Study Area

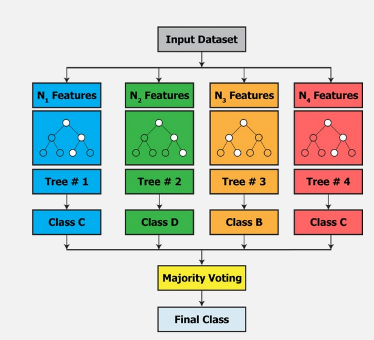
- Major Cyanobacterial *Microcystis* abundance events happened in the southern part (San Joaquin) of the Delta.



Model Selection and Development

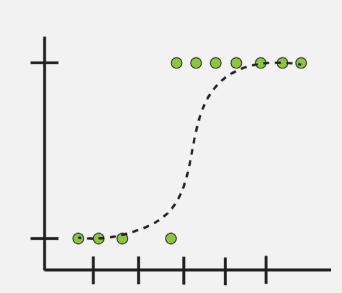
ML model development

- This study treated **HABs modeling** as s **classification problem** (using **4,000 cells/ml** as the threshold).
- **Five** popular **ML models** are selected to solve **classification problems**.
- Based on the correlation matrix and initial data analysis, **water temperature, pH, DO, dissolved NH_4, dissolved organic N, dissolved PO_4, total P, antecedent flow, and salinity** were selected as predictors for **ML model development**.
- **70% Training** and **30% Testing** Data
- Data were **selected randomly** for modeling.
- **Hyperparameters** were **tuned** using the **grid search method.**

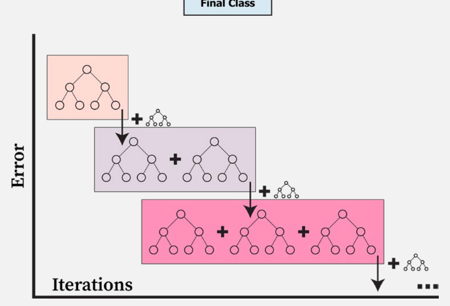


Random Forest Model: An ensemble of multiple decision trees that improves accuracy and reduces overfitting. It is predicted by averaging the outputs of all decision trees (for regression) or by majority voting (for classification).

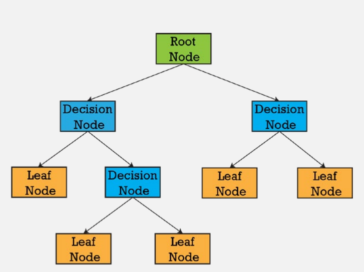
ML model used



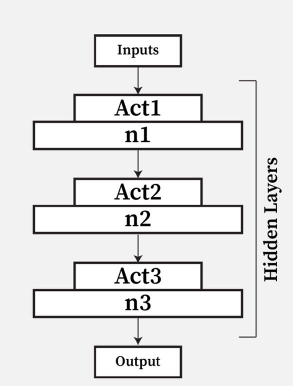
Logistic Regression Model: A simple model used for binary classification predicts probabilities by applying a logistic function to the input features. It predicts the likelihood of an event based on linear relationships between the features and the outcome.



XGBoost Model: A highly efficient boosting algorithm that builds models sequentially, where each new model corrects the errors of the previous one. It uses gradient descent to optimize predictions, making it robust for large, complex datasets.



Decision Tree Model: This model makes predictions by splitting data into branches based on feature values, forming a tree-like structure. It classifies by moving down the tree to a decision point (leaf) based on feature splits.

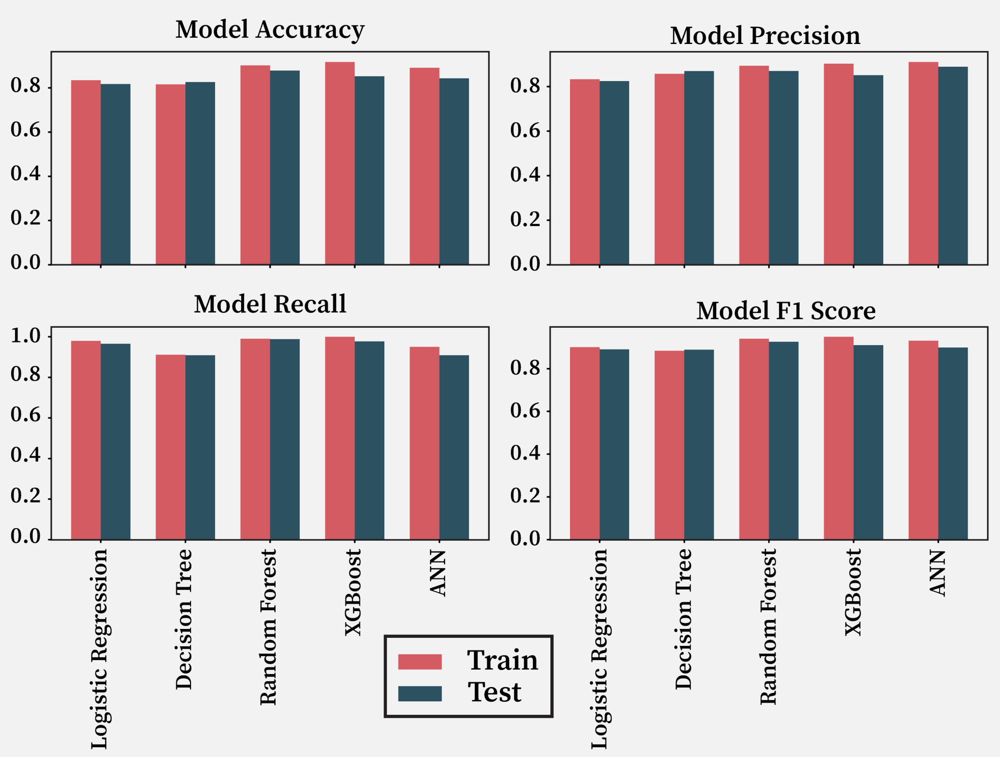


Artificial Neural Network (ANN) Model: ANNs consist of layers of interconnected nodes (neurons) that process data. They predict by learning complex patterns and relationships between inputs and outputs through multiple layers and backpropagation.

Evaluation Metrics:

- **Model Accuracy**: How often the model's predictions are correct overall.
- **Model Precision**: Of the predictions made for a specific category, how many were actually correct.
- **Model Recall**: Out of all true cases in a category, how many were correctly predicted.
- **Model F1 Score**: A balance between precision and recall showing overall model performance.

Initial Results



- The higher the values of all performance metrics (accuracy, precision, recall, and F1 score), the better.

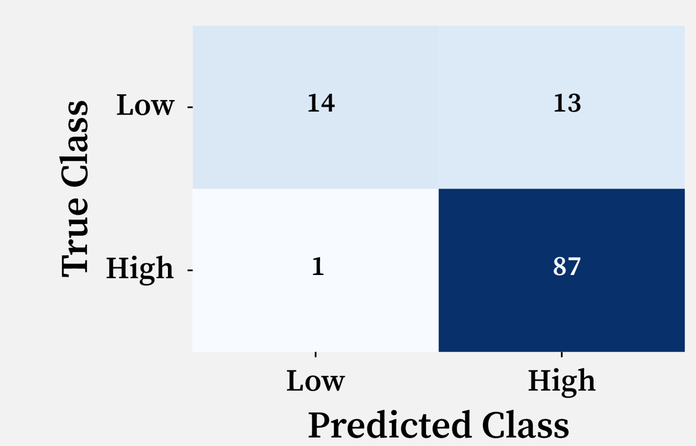
- **RF model most accurately predicted** *Microcystis* (Avg. cells/ml) among five considered ML models.
- **RF model** demonstrated the **highest accuracy** (0.88), **highest recall** (0.99), and **highest F1 score** (0.93) on the testing dataset, making the model appropriate **in classifying** the *Microcystis* (Avg. cells/ml) responsible for HABs.

- **Low** category: ***Microcystis* (Avg. cells/ml) < 4,000***
- **High** category: ***Microcystis* (Avg. cells/ml) > 4,000***

* California Voluntary Guidance

- **RF model** correctly predicts the **Low** category **93%** of the time (14 out of 15 times) on the testing dataset.
- On **87%** of occasions, the model predicts the **High** category correctly.
- The **RF model** predicts **Low** category on one occasion, though the **True** class is **High** (which implies under-prediction and undesirable).
- The model predicts **High** on **13** occasions (**13%**), but the **True** class is **Low**.

Confusion Matrix



Future Direction

- **Expand** the **dataset** to **enhance** the **generalization capability** of the proposed **ML models.**
- **Develop** an **interactive dashboard** that enables users to instantly **simulate *Microcystis*** under **user-defined environmental conditions.**

Acknowledgements

We want to express gratitude to our colleagues **Silvia Angles**, **Rosemary Hartman**, **Daphne Gille**, and **Shaun Philippart** for their valuable insights on HABs modeling and the dataset used in this study. The views expressed in this poster are those of the authors and not of their employer.

References

1. Hartman R, Rasmussen N, Bosworth D, Berg M, Ateljevich E, Flynn T, Wolf B, Pennington T, Khanna S. 2022. *Temporary Urgency Change Petition of 2021 and Emergency Drought Salinity Barrier: Impact on Harmful Algal Blooms and Aquatic Weeds in the Delta.* Sacramento (CA): California Department of Water Resources. October 2022. 188 pp. + appendix.
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*
3. Environmental Protection Agency, US. (2019). *Recommended Human Health Recreational Ambient Water Quality Criteria or Swimming Advisories for Microcystins and Cylindrospermopsin (EPA Document Number: 822-R-19-001)*. https://www.epa.gov/sites/default/files/2019-05/documents/hh-rec-criteria-habs-document-2019.pdf

Contact Us

Gourab Saha
Water Resources Engineer
Delta Modeling Section, DWR
Email: gourab.saha@water.ca.gov

Peyman Namadi
Water Resources Engineer
Delta Modeling Section, DWR
Email: peyman.hosseinzadehnamadi@water.ca.gov