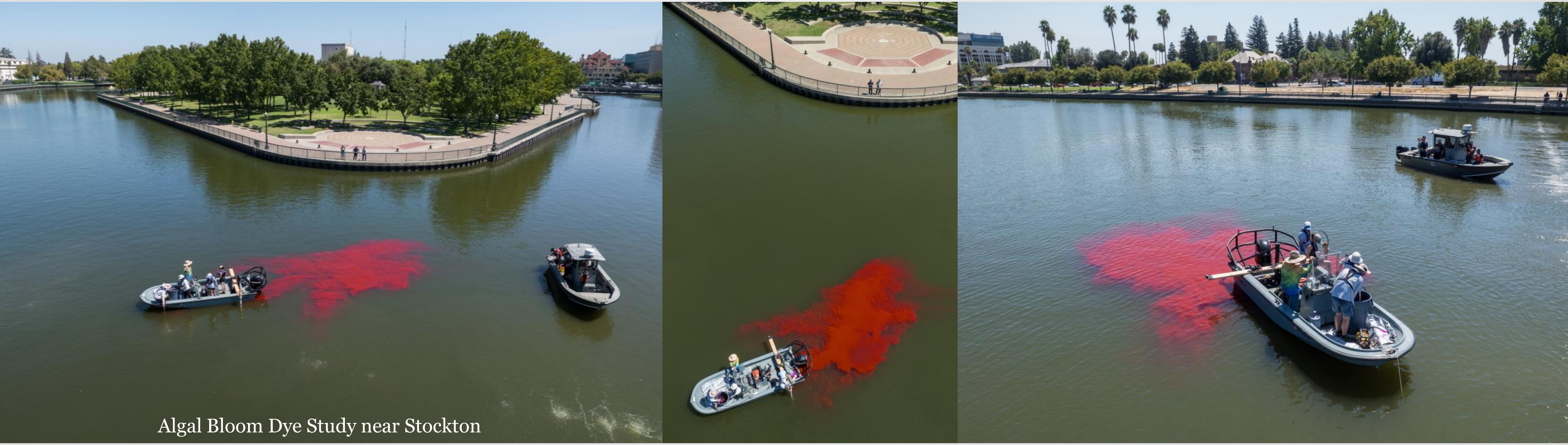
CALIFORNIA DEPARTMENT OF WATER RESOURCES

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

CWEMF Annual Meeting, May 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
- ❑ Data Preparation
- ❑ Model Development
- ❑ Initial Results and Observations
- ❑ Future directions

Saint Luis Lake (an example image)

Background

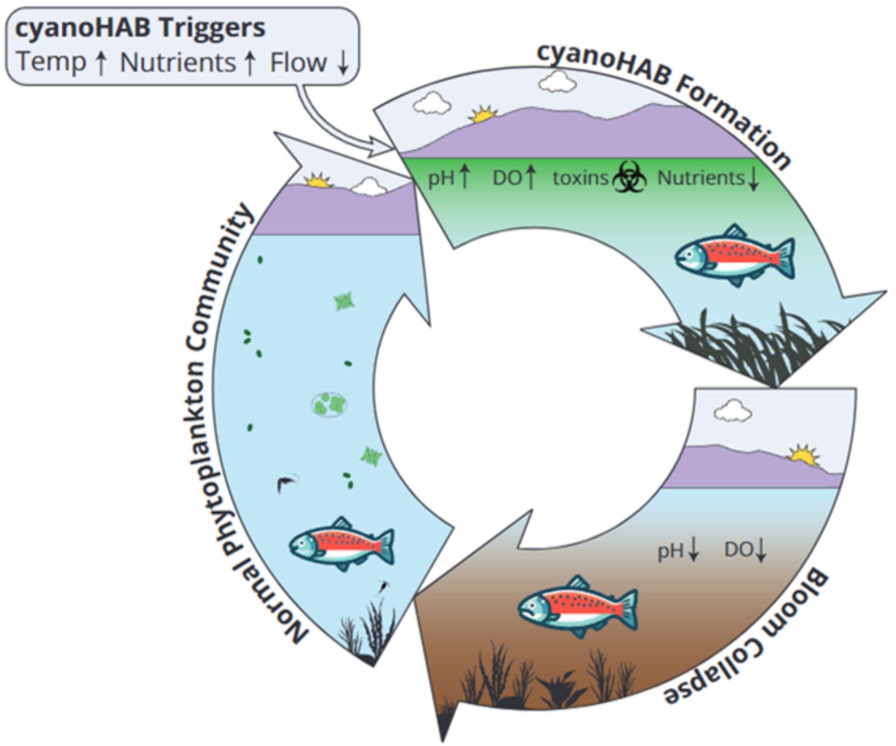
Photo Courtesy: California Water Quality Control Board

- 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

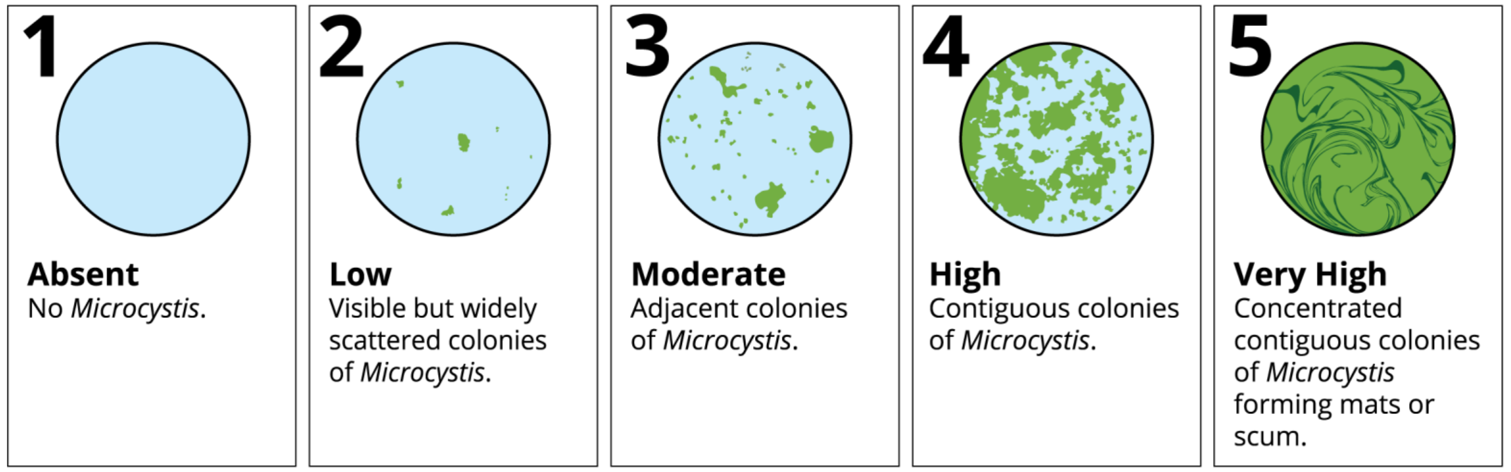
- Nutrient availability, low flow, and favorable water temperature influence the chance of Cyanobacteria bloom abundance in the Delta.



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

Source: Bouma–Gregson et al. 2024

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



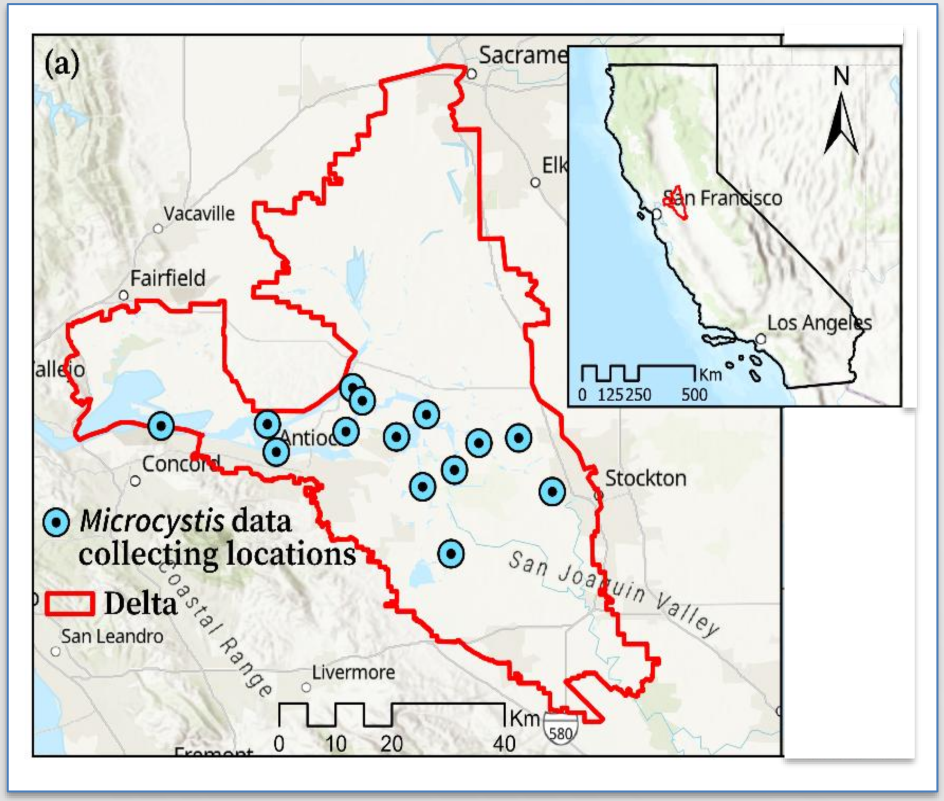
Visual Index

Source: Flynn et al. 2022

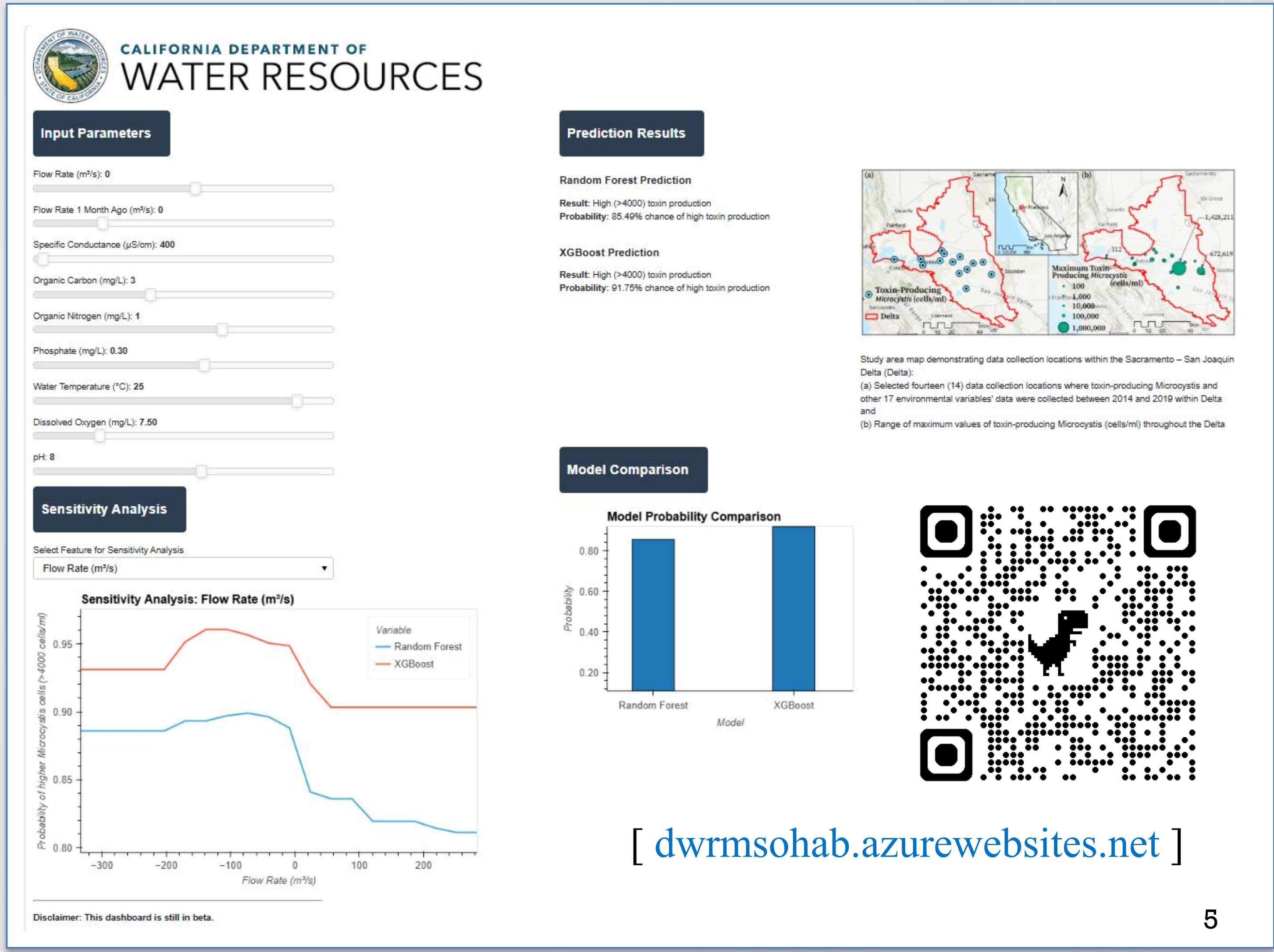
Background

- ❑ HAB modeling **Phase 1** predicted HAB risk using 5 developed ML models.

- ❑ This study used the 220 samples from 14 locations of the **Delta** region.



- ❑ An interactive dashboard was created.



[dwrmsohab.azurewebsites.net]

Goal and Objectives

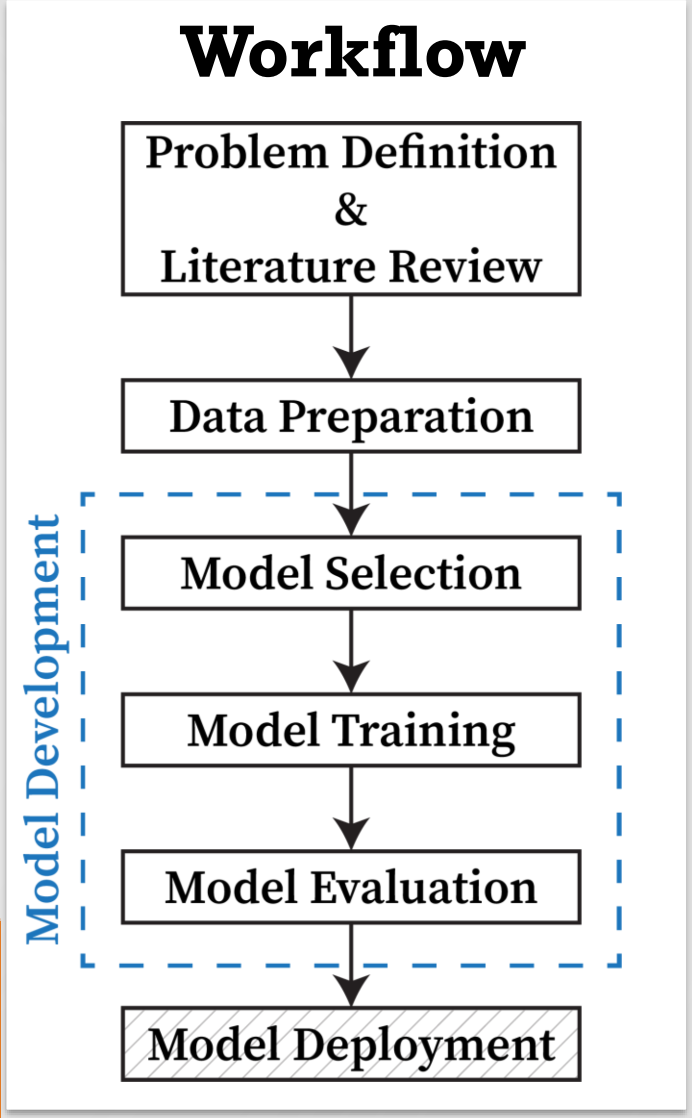
❑ Goal

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

❑ Study Objectives

- ✓ Predicting HAB status (**Visual Index)** using machine learning techniques in the Delta with **new dataset.**

- ✓ Creating **a dashboard** to predict HAB status **(Visual Index)** at user-defined locations.



Machine Learning Protocols

Study Data

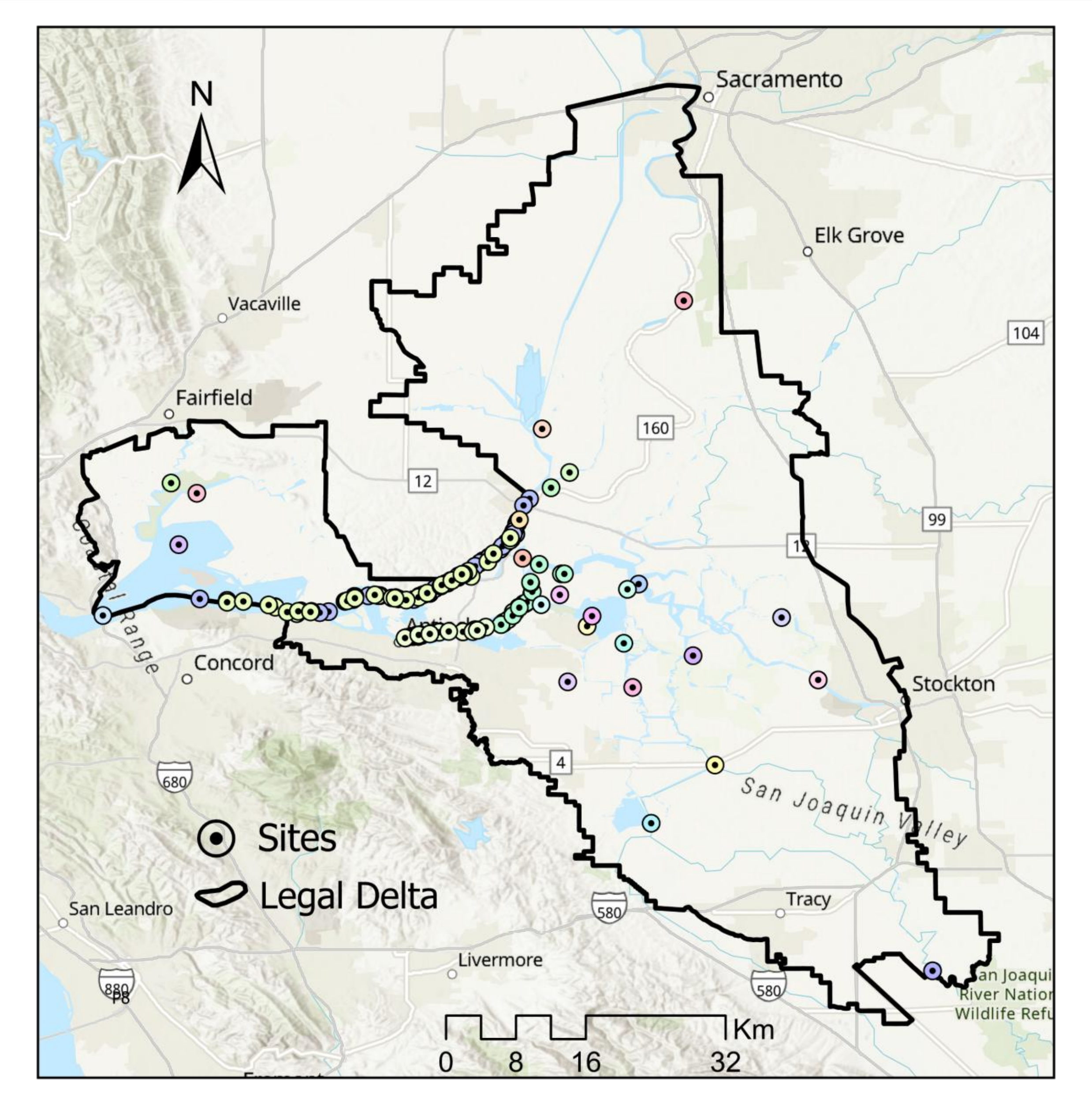
- 35 Sites (168 locations)

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

- Data Count: 1006 daily 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

Workflow

Data Preparation → ML Models Development → VI prediction (Absent, Low & High)

ML = Machine Learning
VI = Visual Index

Workflow

Data Preparation → ML Models Development → VI prediction (Absent, Low & High)

ML = Machine Learning
VI = Visual Index

Data: Input Data

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

Water quality factors

- Water temperature
- Conductivity

Physical processes

- Antecedent flow
- Antecedent velocity

(DSM2 Simulated)

Nutrient factors

- Dissolved ammonia
- Dissolved nitrate & nitrite
- Dissolved orthophosphate

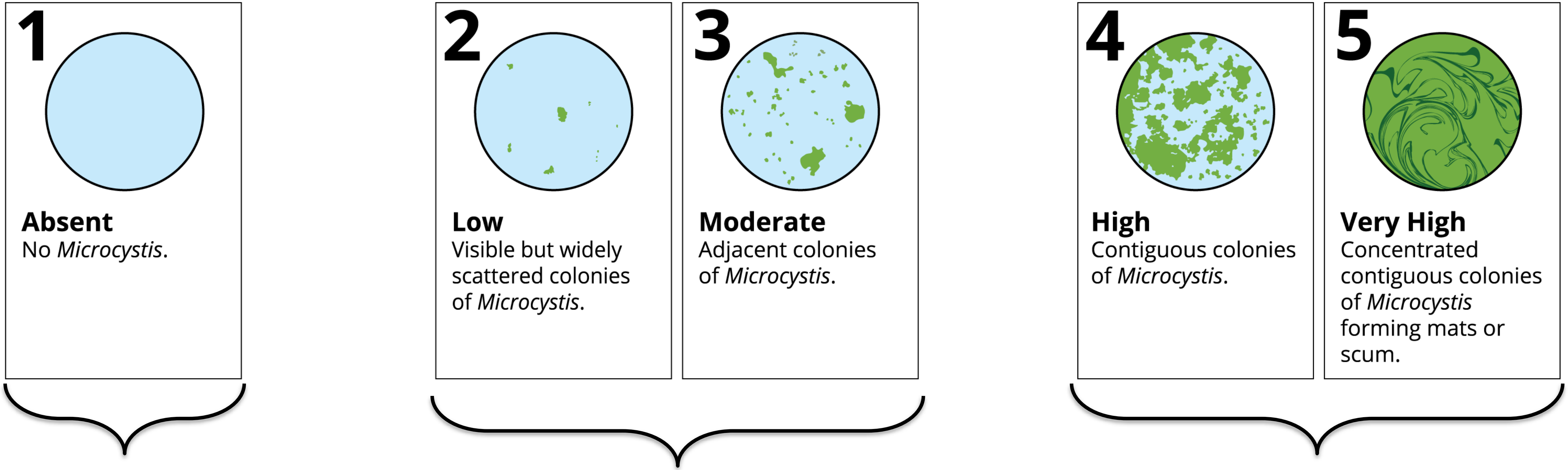
Light availability

- Turbidity

Data: Target Data

Target variable for ML based HAB modeling – Visual Index.

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



Absent (VI = 1)

Low (VI = 2)

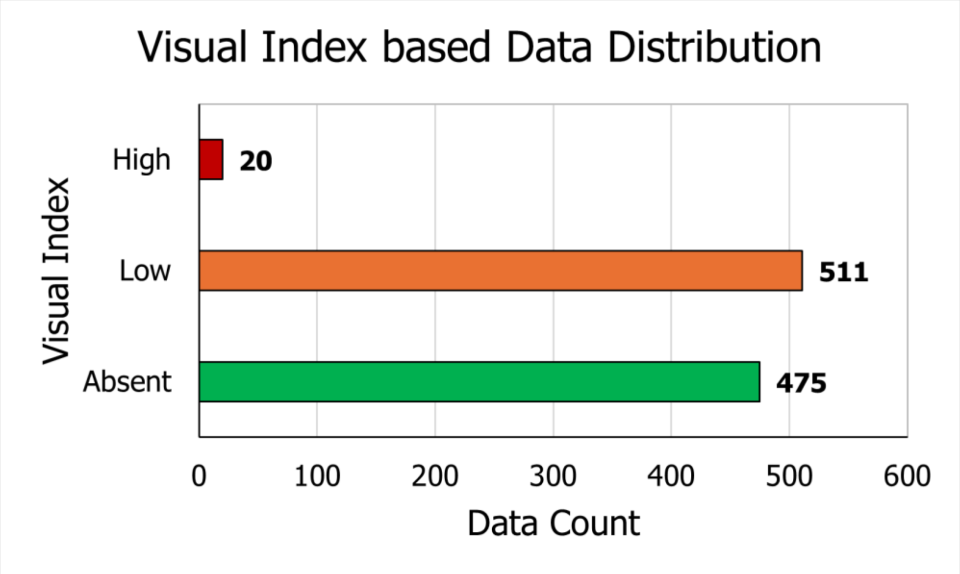
High (VI = 3)

VI = Visual Index

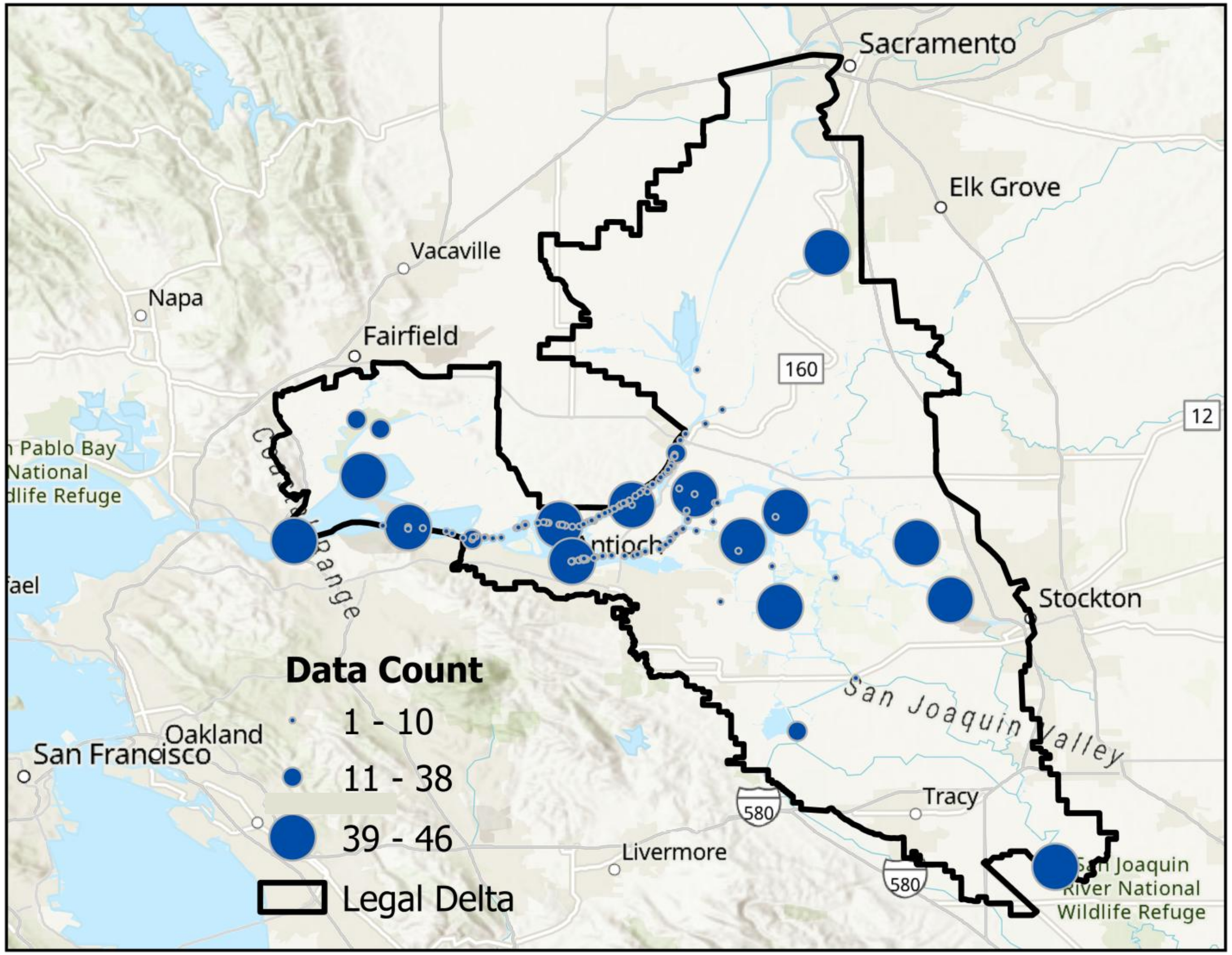
Source: https://doi.org/10.15447/sfews.2024v22iss1art2

Data Distribution

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



Number of Data per Site



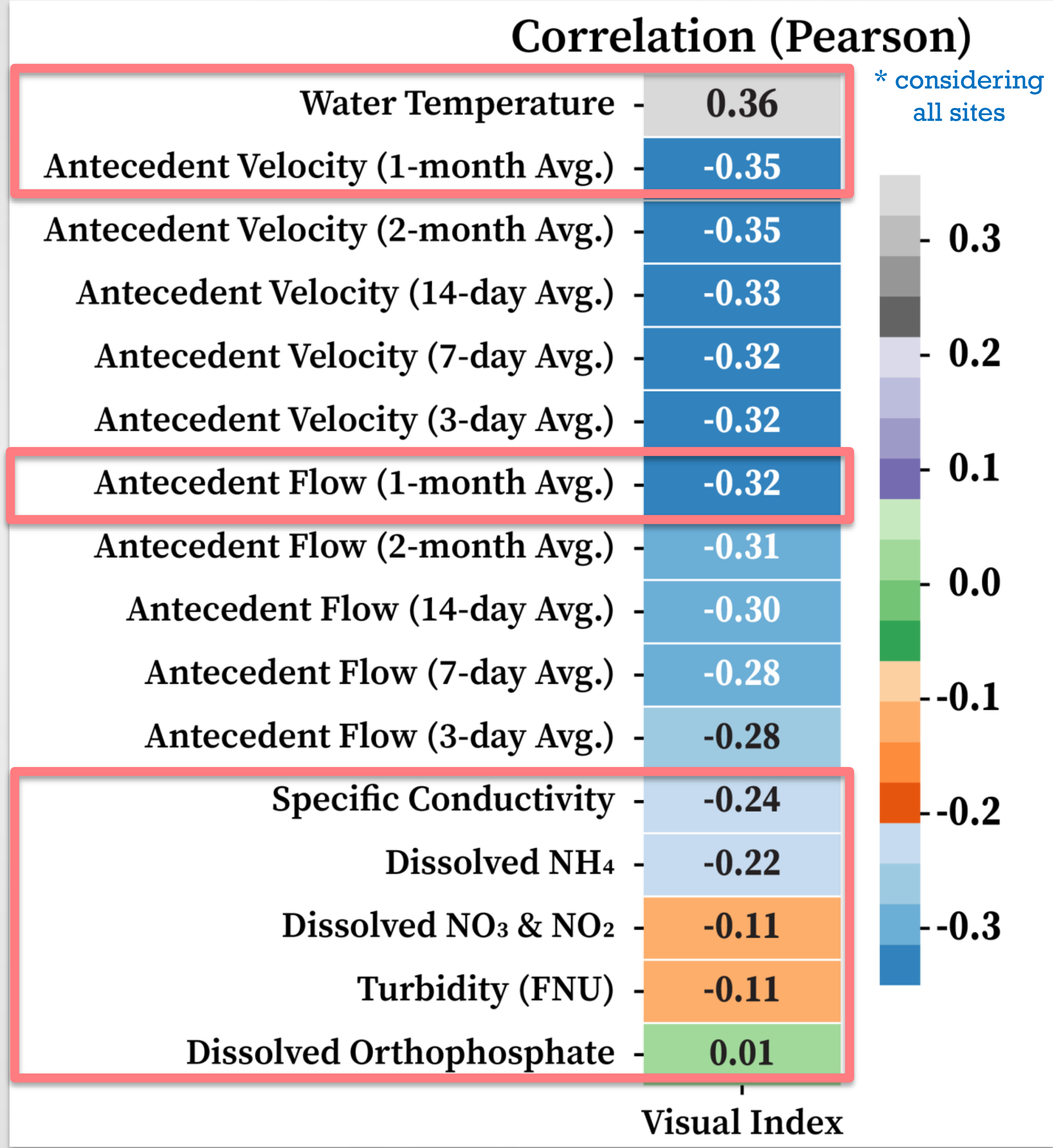
Data Analysis

- Visual Index correlations used to select initial machine learning inputs.

- Positive correlation: If the variable value goes up (e.g., water temperature up), then high probability of having more HABs.

- Negative correlation: If the variable value goes up (e.g., velocity up), then 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.

Workflow

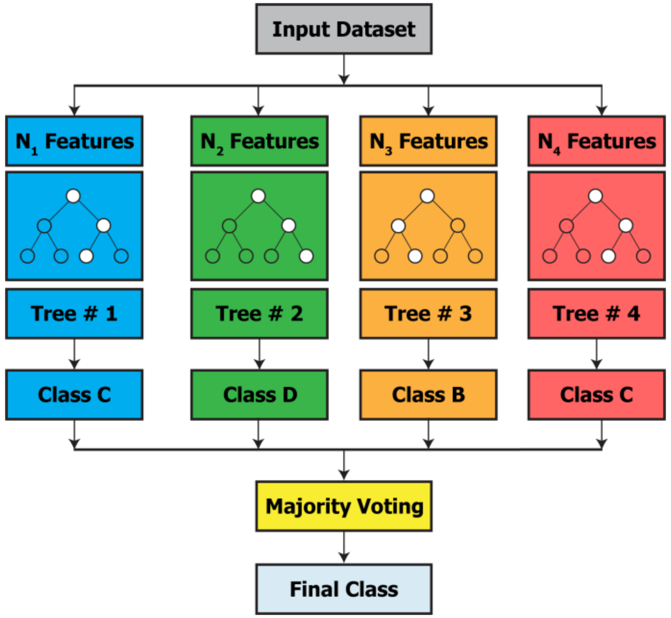

Data Preparation
ML Models Development
VI prediction (Absent, Low & High)

Model Selection

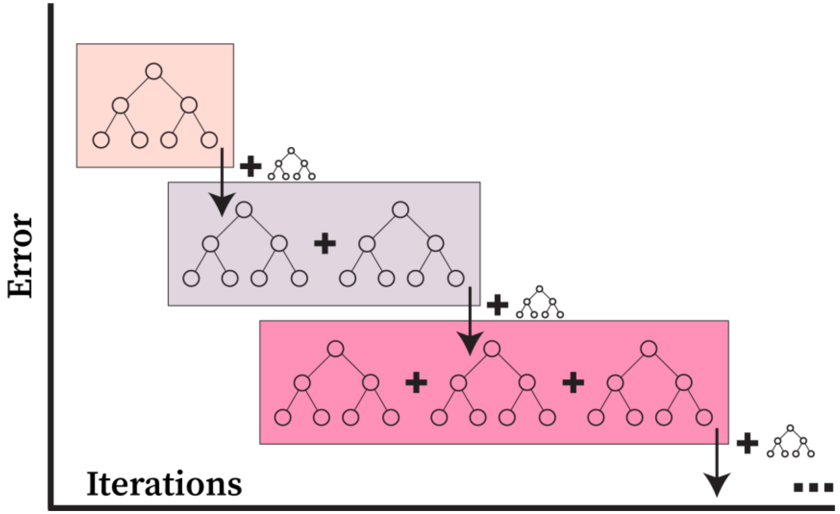
- ❑ Developed 3 types of HABs Machine Learning models
  - ✓ Random Forest (RF)
  - ✓ XGBoost
  - ✓ Artificial Neural Network (ANN)

HABs = Harmful Algal Blooms

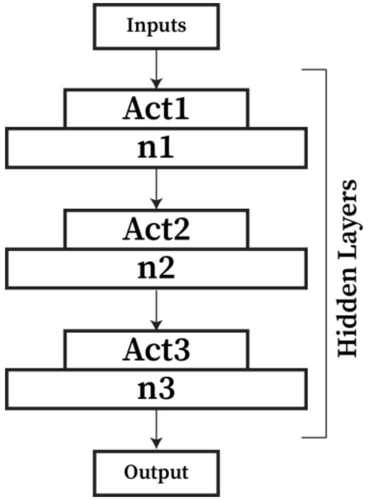
Random Forest



XGBoost

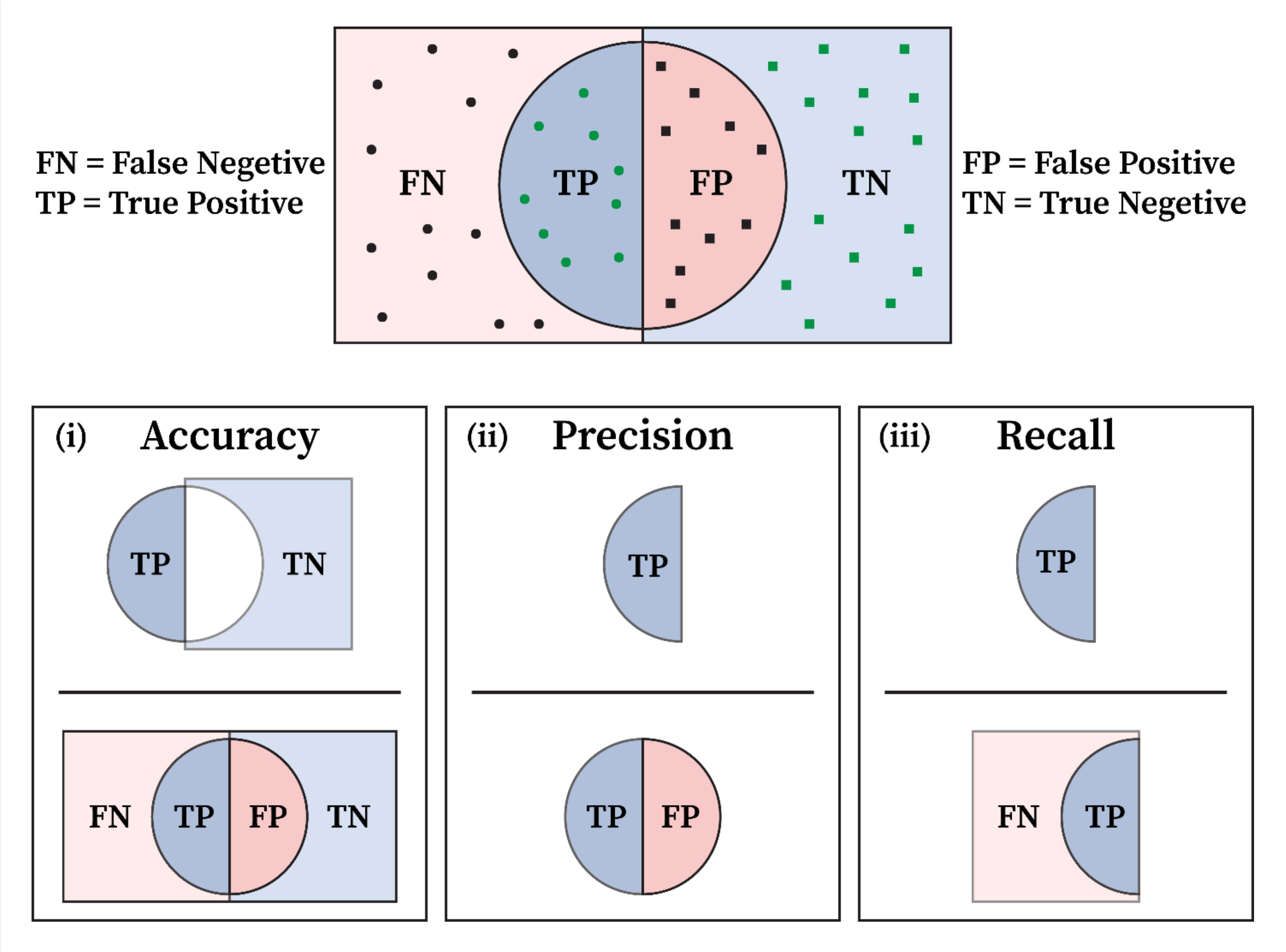


ANN



python™

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.

Model Development

Inputs

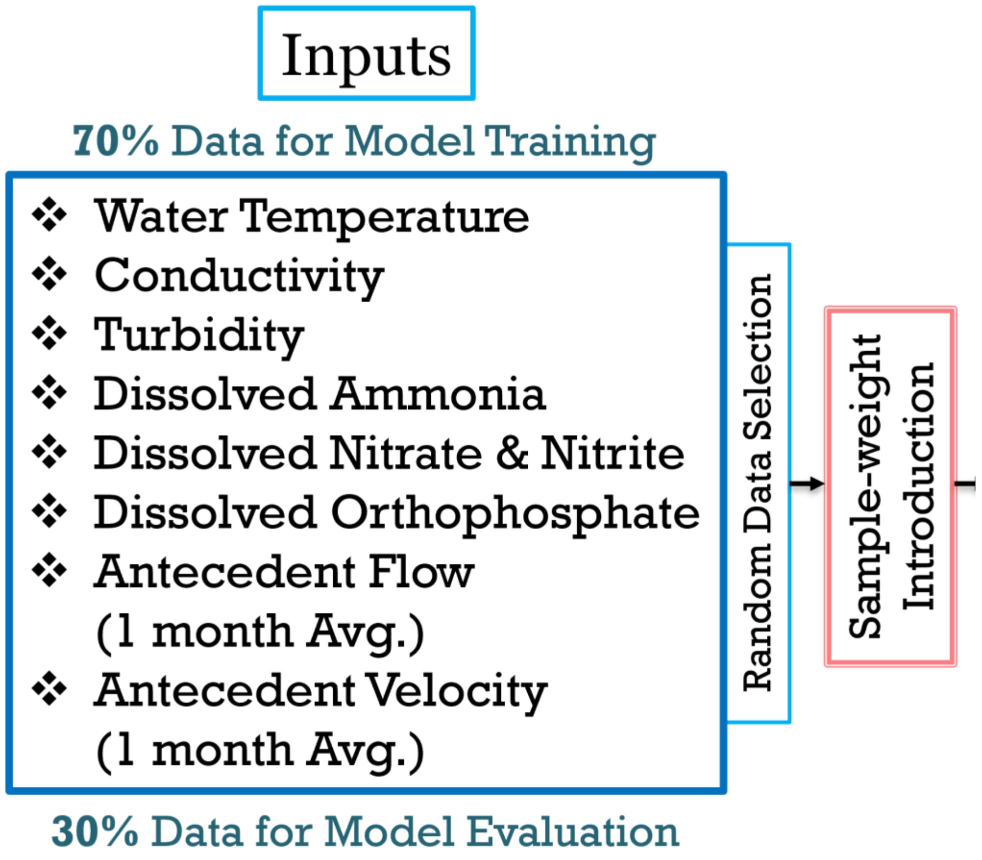
70% Data for Model Training

- Water Temperature
- Conductivity
- Turbidity
- Dissolved Ammonia
- Dissolved Nitrate & Nitrite
- Dissolved Orthophosphate
- Antecedent Flow (1 month Avg.)
- Antecedent Velocity (1 month Avg.)

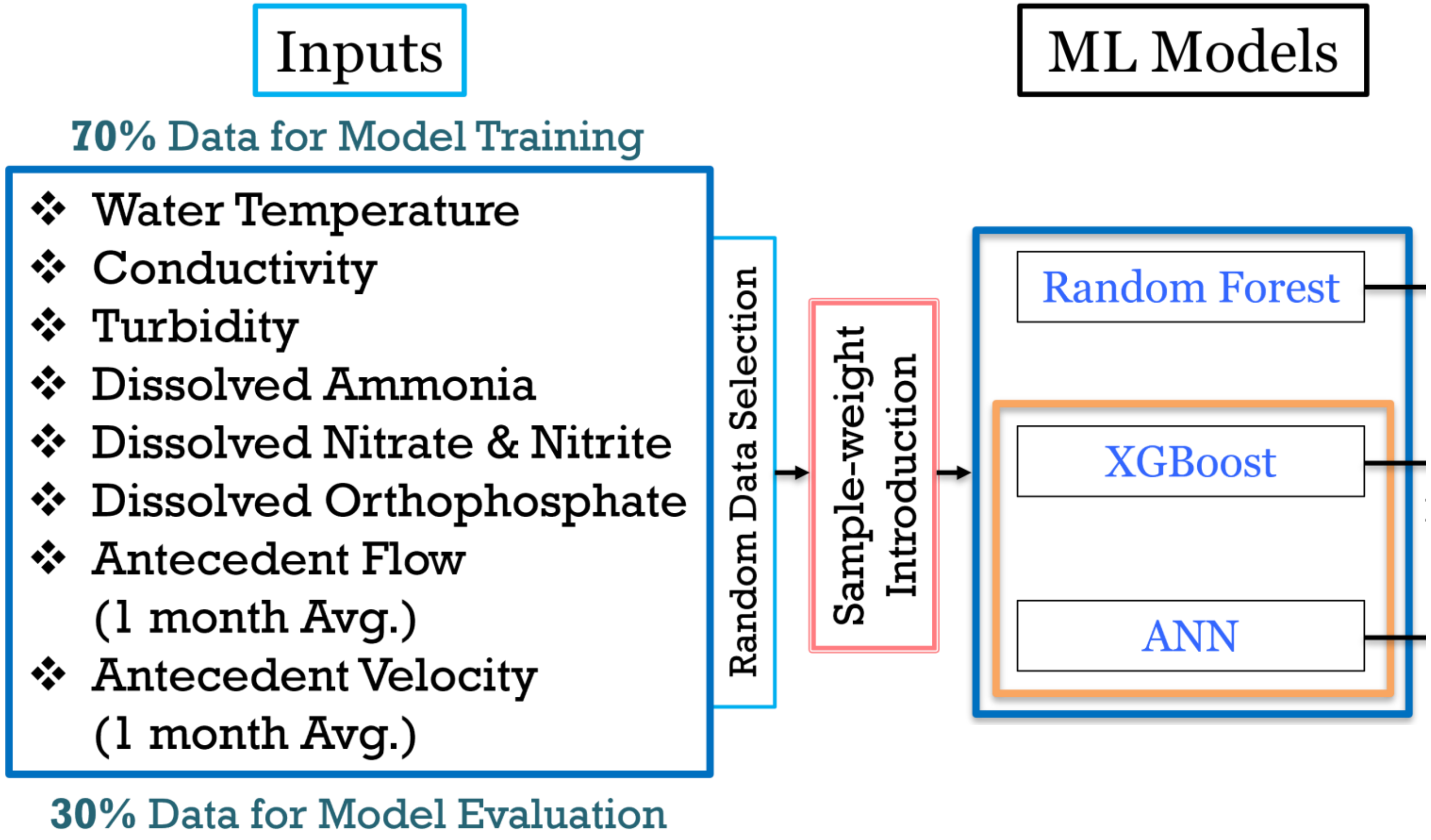
Random Data Selection

30% Data for Model Evaluation

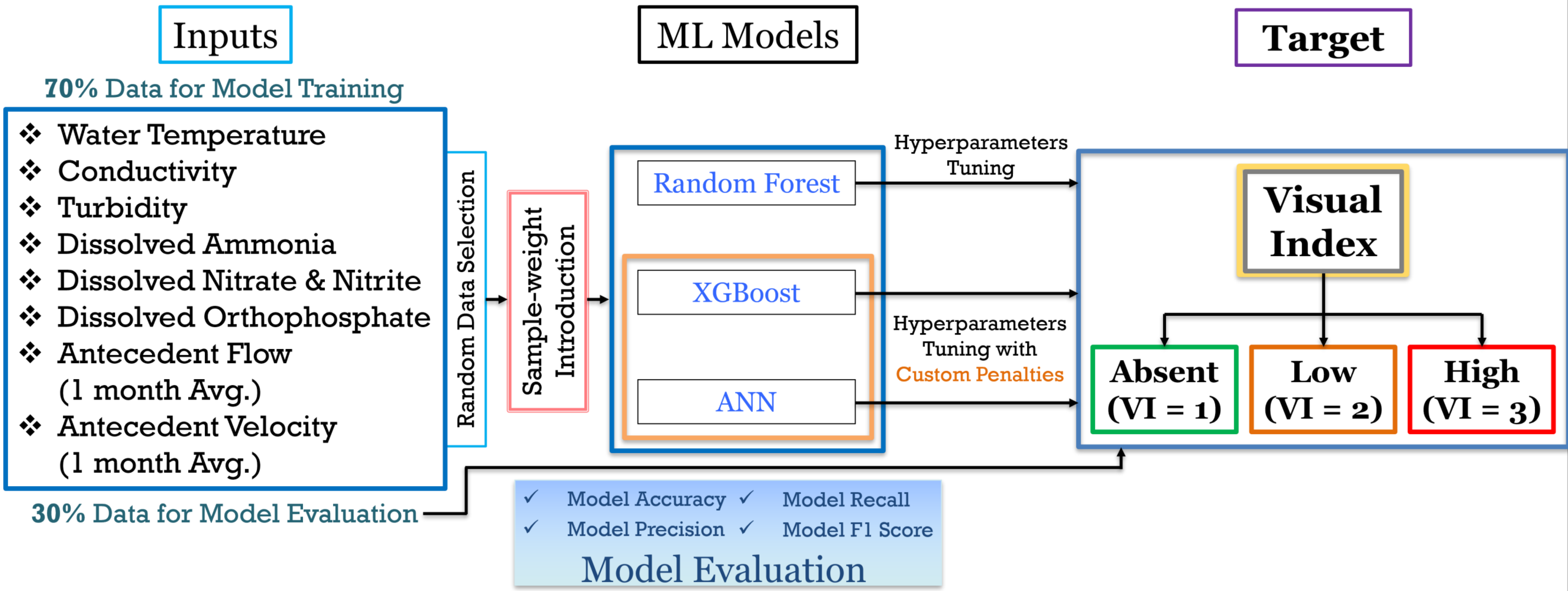
Model Development



Model Development



Model Development



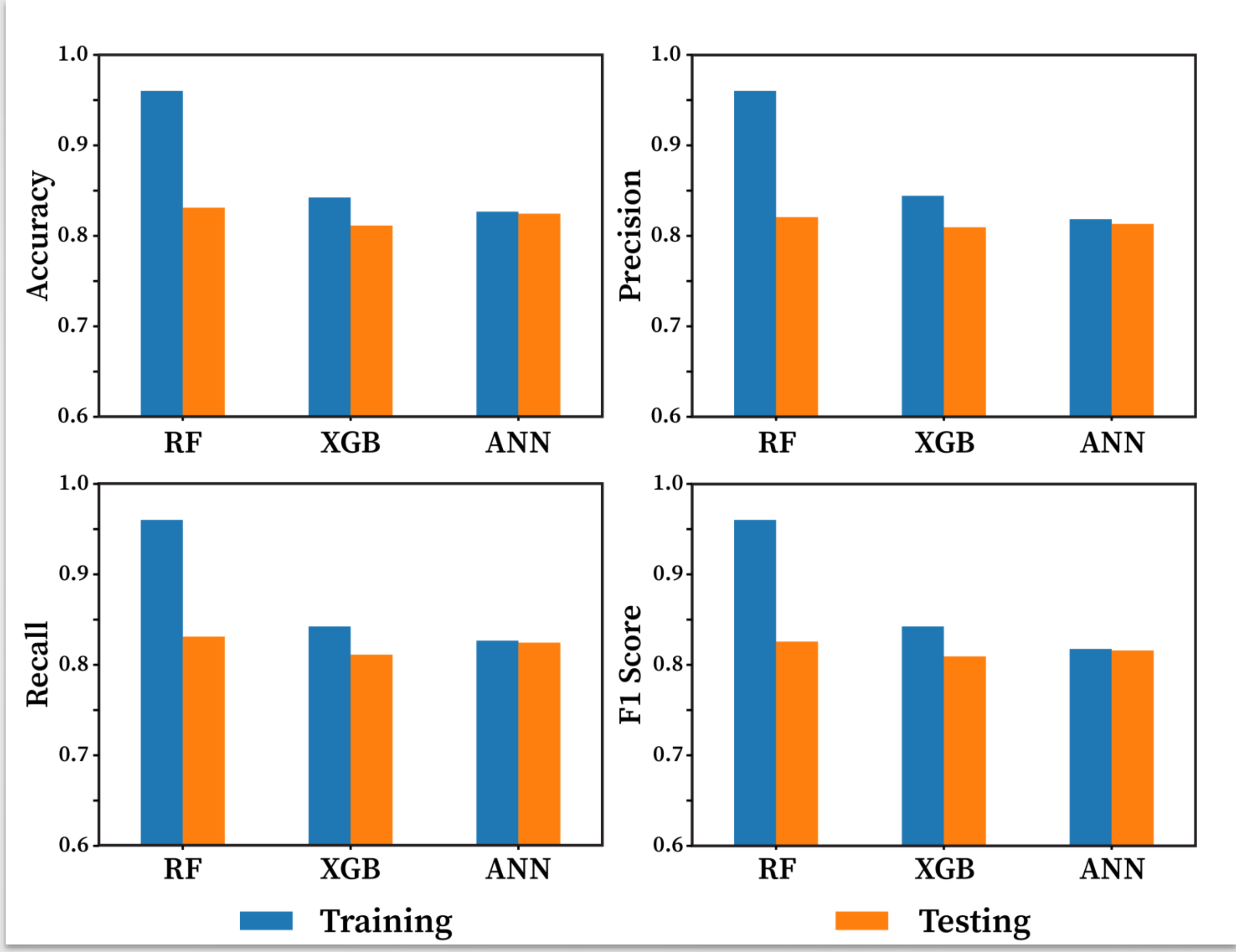
Workflow

Data Preparation → ML Models Development → VI prediction (Absent, Low & High)

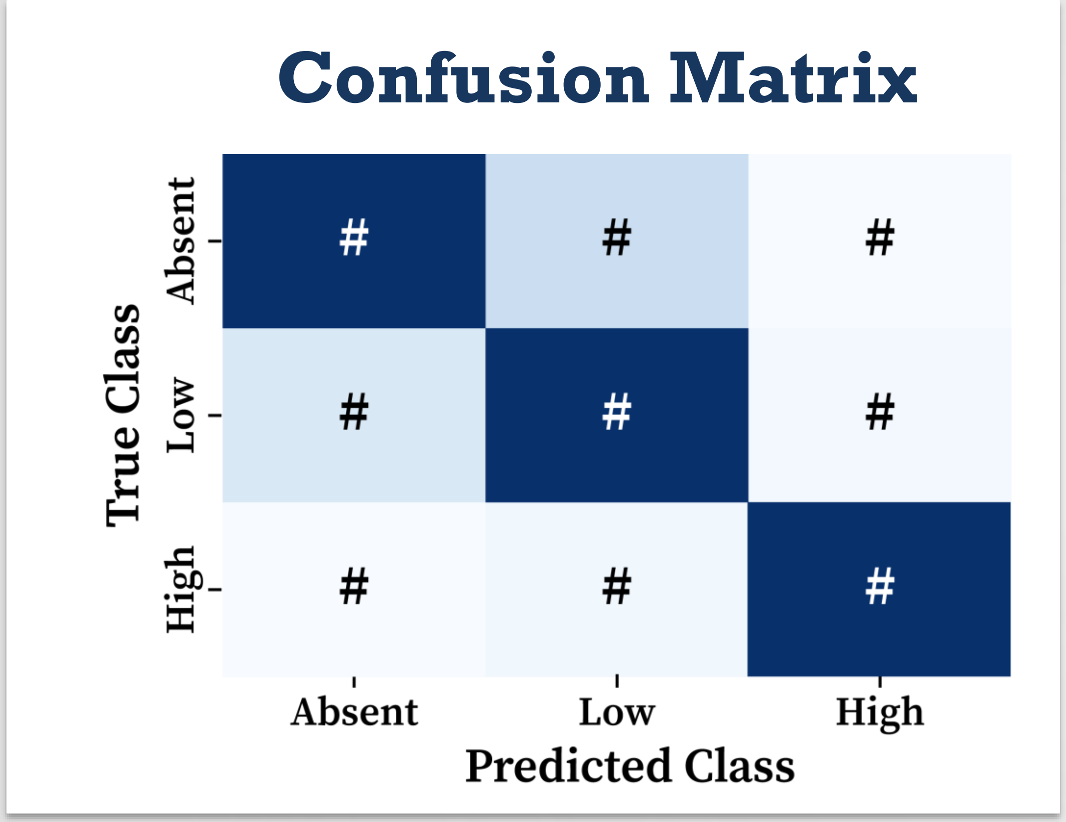
Initial ML Models' Results

- Three (3) ML models (Random Forest, XGBoost, ANN) were developed on eight (8) input variables to predict ***three VI classes*** (Absent, Low, High).

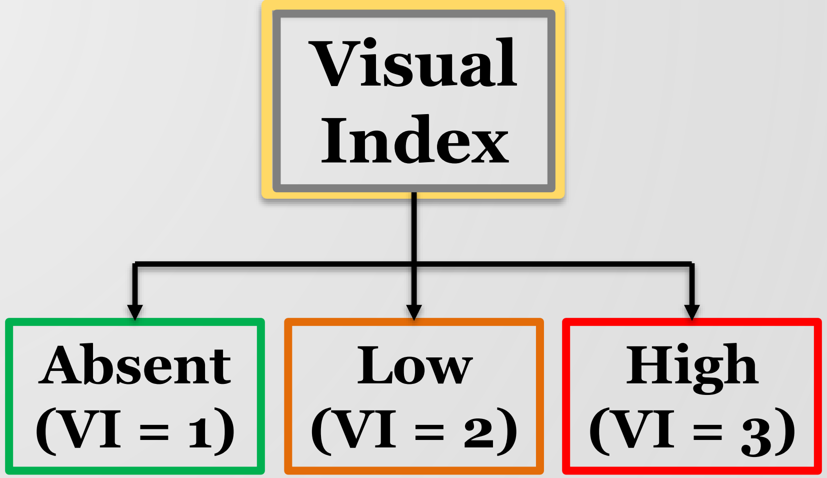
- All three ML models, including Random Forest, XGBoost, and ANN, demonstrated similar Visual Index prediction performance with test accuracy 0.83, 0.81, and 0.82, respectively.



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.



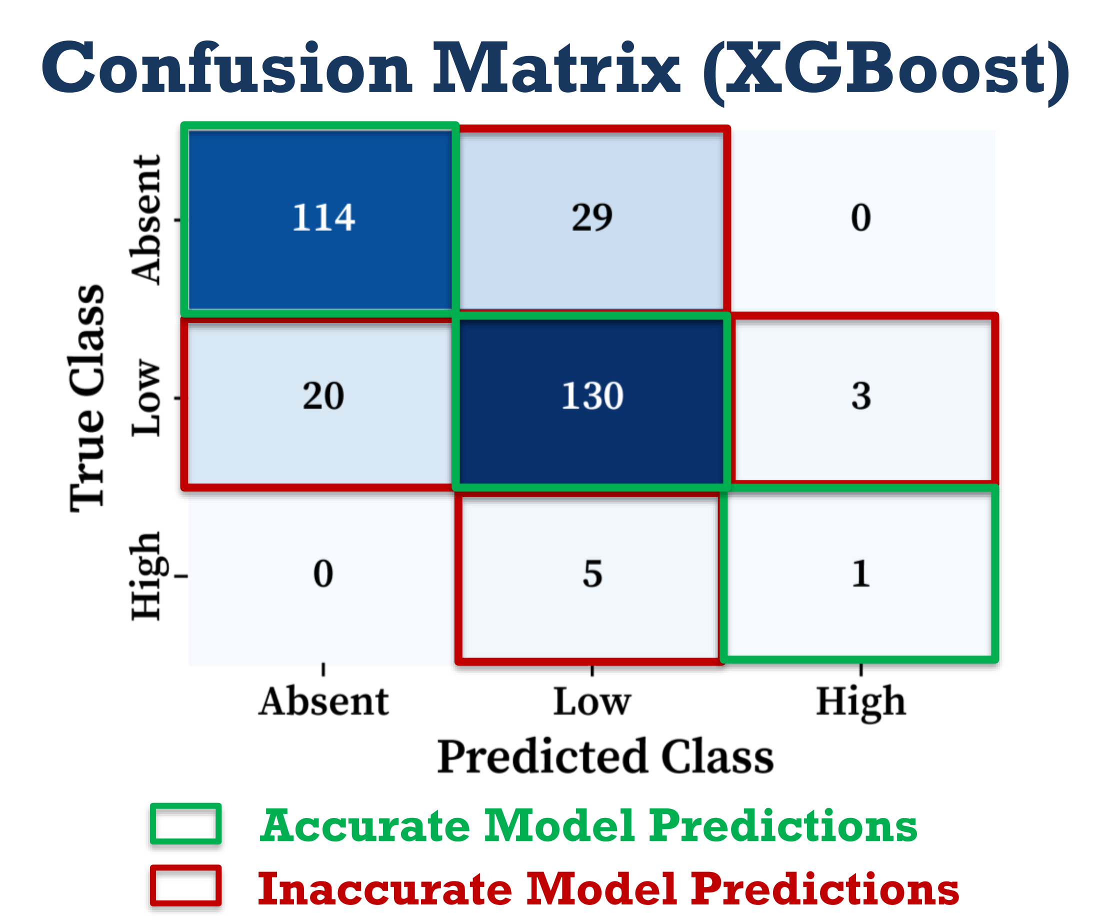
True Class means the **Observed Values.**

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

XGBoost Model Results (As an example)

- XGBoost model correctly predicts the Absent VI 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 VI category correctly.
- XGBoost model predicts High VI category on 1 out of 5 occasions (19%).

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



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

Initial Observations

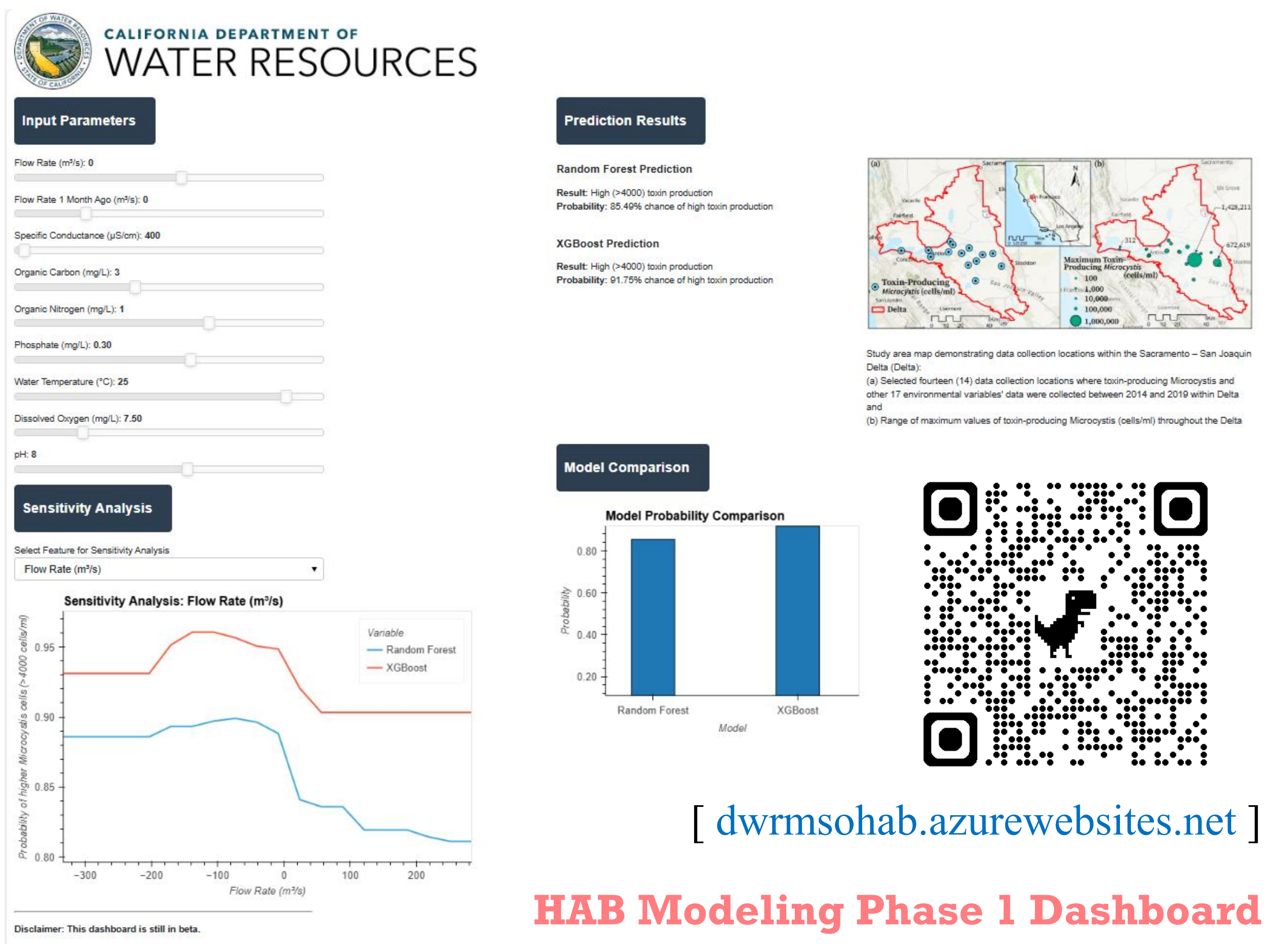
- All three ML models predicted **Visual Index (VI)** with test accuracy 0.83 (Random Forest), 0.81 (XGBoost), and 0.82 (ANN).
- As an example, XGBoost model accurately predicted Absent VI on 80% (114 out of 143) of occasions and Low VI on 85% (130 out of 153 of occasions).
- All three models struggled to predict High **Visual Index**. Investigation is on going to accurately identify the potential reasons.

Future Directions

- Investigate the dataset and ML models more to understand the reasons behind models' under-prediction of High **Visual Index** category.

- Develop an interactive dashboard that enables users to instantly simulate **Visual Index** under user-defined environmental conditions.

- Document and deploy final models (source code), datasets, and dashboard on GitHub/Microsoft Azure and make them publicly available.



[dwrmsohab.azurewebsites.net]

HAB Modeling Phase 1 Dashboard

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**.

DELTA DASH

CALIFORNIA DEPARTMENT OF WATER RESOURCES

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
- Zhenlin.Zhang@water.ca.gov
- Kevin.He@water.ca.gov

Thank You!

CALIFORNIA DEPARTMENT OF WATER RESOURCES

Extra Slides

Data Distribution

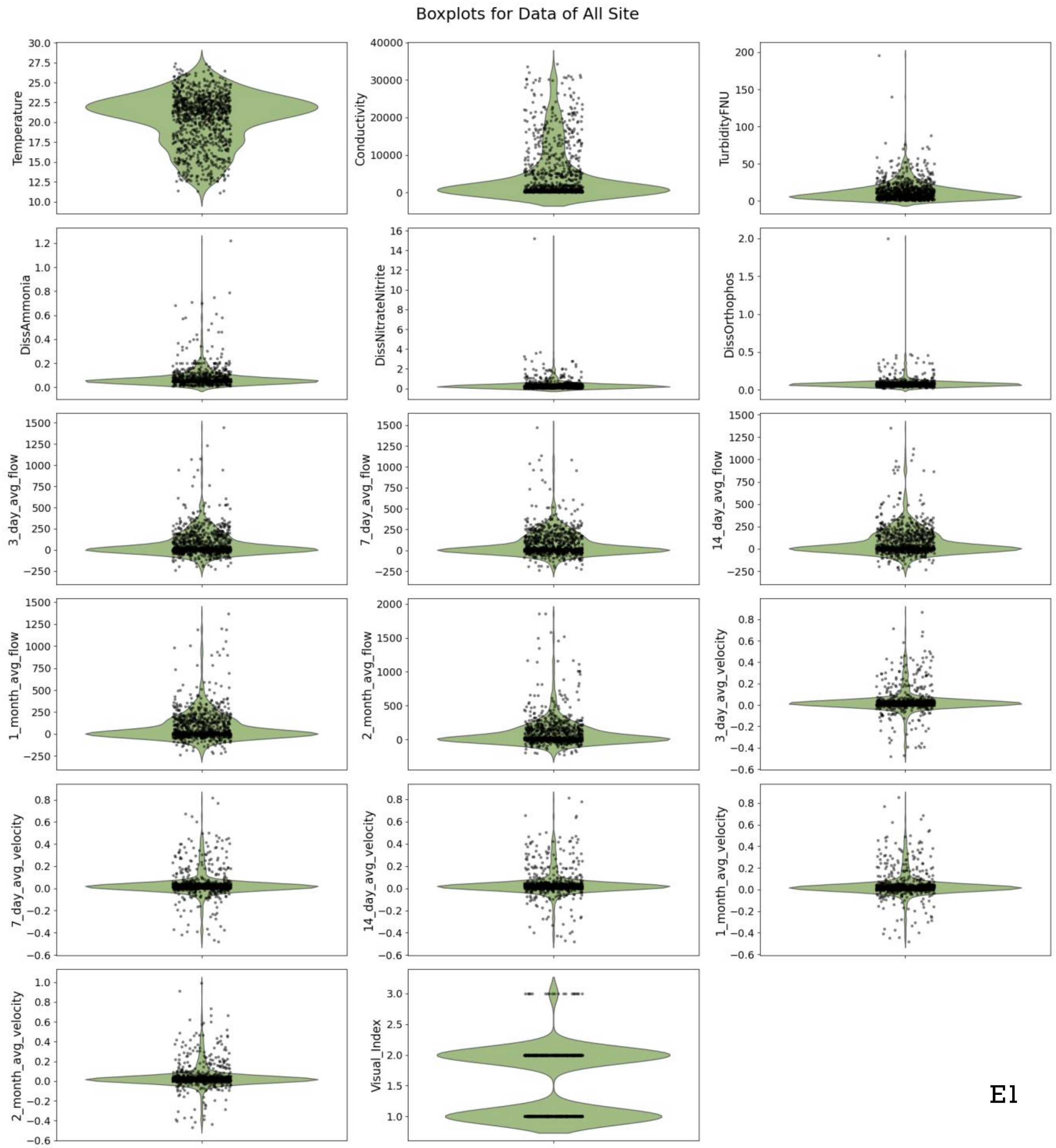
- ❑ Violin plots for all sites' data combined.

❑ Antecedent Flow

- ❖ 3-days avg. flow
- ❖ 7-days avg. flow
- ❖ 14-days avg. flow
- ❖ 1-month avg. flow
- ❖ 2-months avg. flow

❑ Antecedent Velocity

- ❖ 3-days avg. velocity
- ❖ 7-days avg. velocity
- ❖ 14-days avg. velocity
- ❖ 1-month avg. velocity
- ❖ 2-months avg. velocity

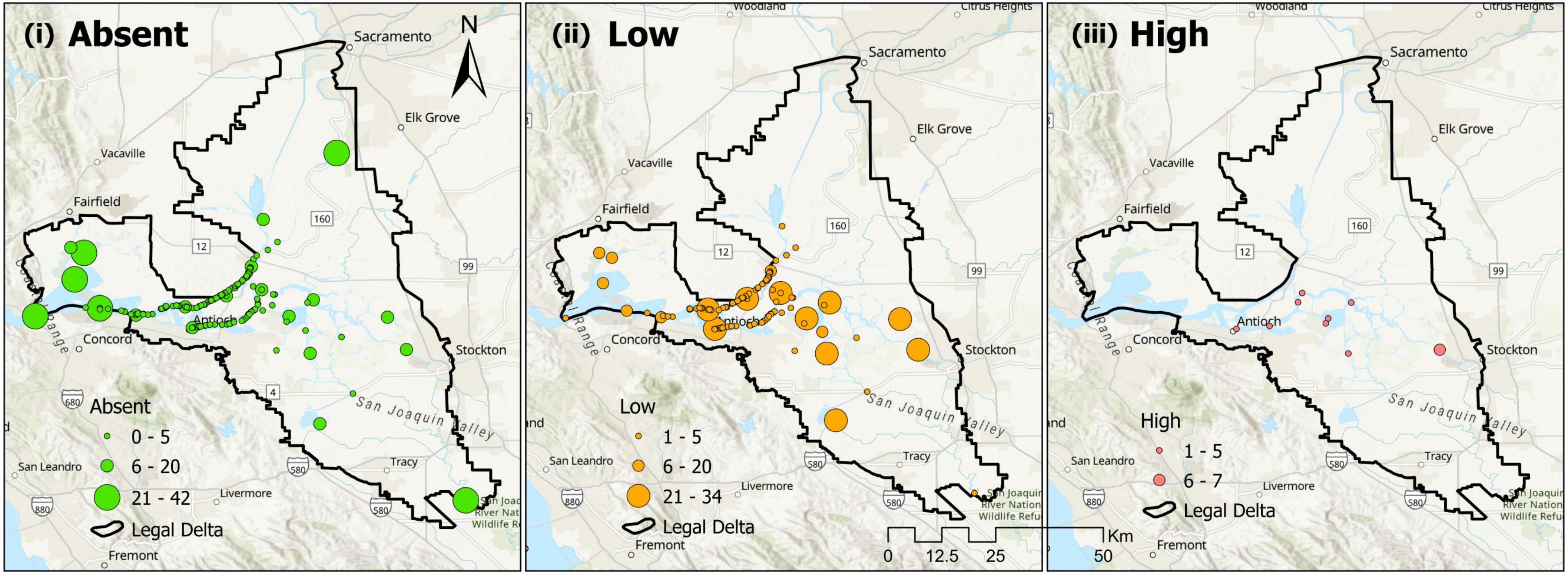


CALIFORNIA DEPARTMENT OF WATER RESOURCES

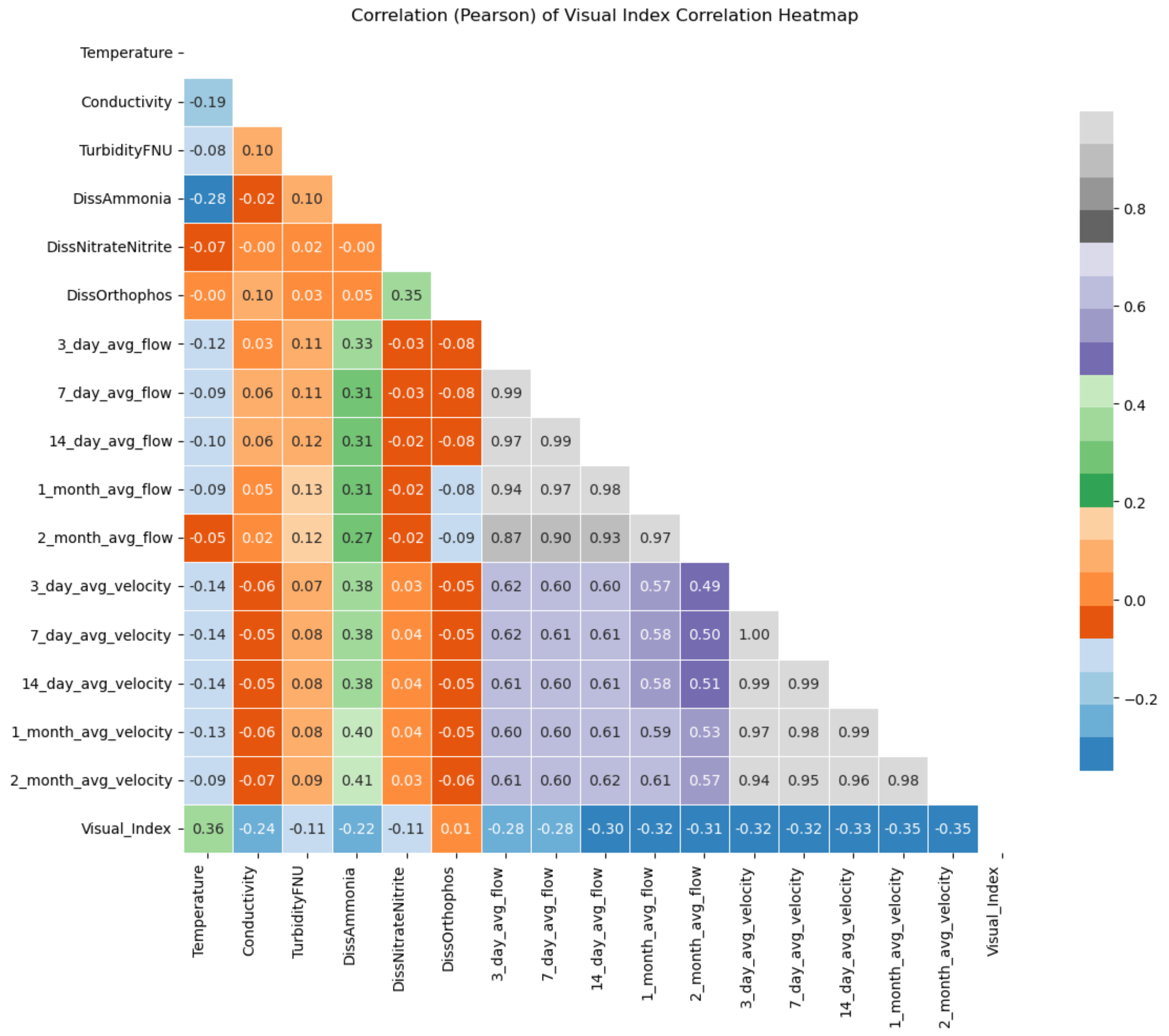
E1

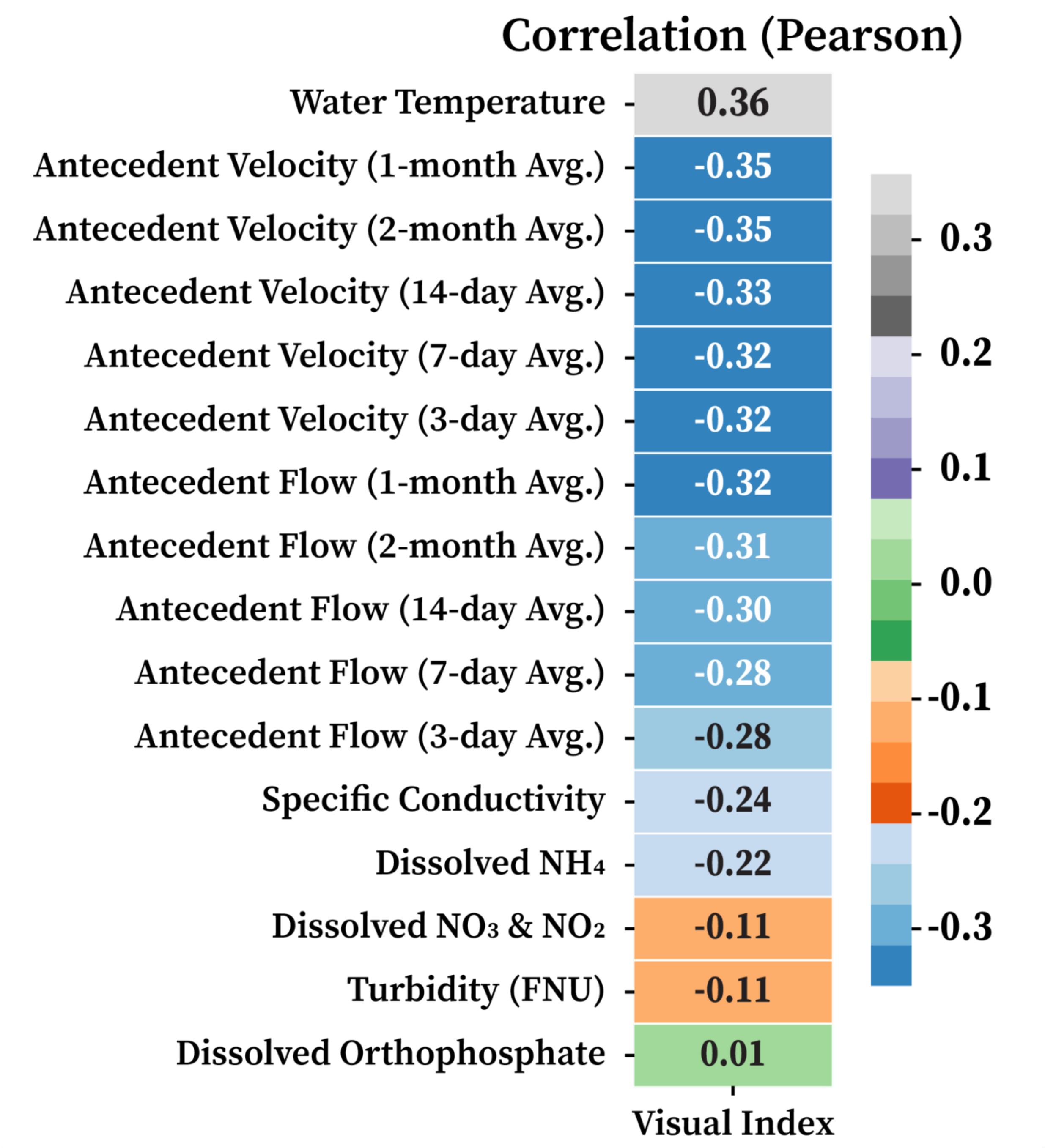
Data Distribution

- **Higher number** of **Low** and **High** **Visual Index data per site** are available in the **Central Delta**.



Data Analysis





- ❑ Correlations considering all sites

Code Snippets (Sample Weights)

HAB	Sample Count	Sample Weight
Absent	475	0.705964912
Low	511	0.656229615
High	20	16.7666666

Code Snippets (Hyperparameters)

Random Forest

```python
# Define the parameter space for Random Forest
rf_params = {
    'n_estimators': [10, 50, 100, 200, 300, 400, 500, 700, 800, 1000],
    'max_depth': [None, 5, 10, 15, 20, 25, 30, 40, 50, 80, 100],
    'min_samples_split': [2, 4, 6, 8, 10, 12, 16, 20, 24, 36, 50],
    'min_samples_leaf': [1, 2, 3, 4, 5, 8, 10, 12, 16, 20],
    'max_features': [None, 'sqrt', 'log2', 0.1, 0.2, 0.5, 0.8],
    'max_samples': [None, 0.1, 0.2, 0.5, 0.8],
    'criterion': ['gini', 'entropy', 'log_loss']
}
```

XGBoost

```python
# Define hyperparameter search space
xgb_params = {
    'max_depth': [2, 4, 6, 8, 10, 12, 15],            # Extended depth range
    'learning_rate': [0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5],  # Added smaller learning rates
    'subsample': [0.5, 0.6, 0.7, 0.8, 0.9, 1.0],      # Expanded subsample range
    'colsample_bytree': [0.5, 0.6, 0.7, 0.8, 0.9, 1.0],  # Adjusted feature sampling
    'gamma': [0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 1], # Regularization control
    'min_child_weight': [1, 3, 5, 7, 10, 15, 20],     # Minimum sum of instance weight
    'reg_alpha': [0, 0.01, 0.1, 0.5, 1, 5, 10],       # L1 regularization
    'reg_lambda': [0, 0.01, 0.1, 0.5, 1, 5, 10],      # L2 regularization
    'base_score': [0.5, 0.6, 0.7, 0.8],               # Initial prediction score
    'booster': ['gbtree', 'dart'],        # Boosting method
    'tree_method': ['exact', 'approx', 'hist'],  # Computational methods
    'num_class': [3],                                 # Ensure multi-class compatibility
    'objective': ['multi:softprob'],                  # Multi-class probability output
    'seed': [42]                                      # Random seed for reproducibility
}
```

ANN

```python
# Hyperparameter grid for manual search
ann_params = {
    'neurons': [16, 32, 64, 128],          # Number of neurons
    'hidden_layers': [2, 3, 4],            # Number of hidden layers
    'dropout_rate': [0.05, 0.1, 0.15, 0.2, 0.25, 0.3],  # Dropout rates
    'activation': ['relu', 'tanh', 'elu', 'sigmoid'],    # Activation functions
    'epochs': [200, 300, 500],                  # Epochs
    'batch_size': [16, 32, 64]             # Batch sizes
}
```

Code Snippets (Custom Loss Function)

```python
def xgb_custom_penalty_function(preds, dtrain):
    labels = dtrain.get_label().astype(int)
    sample_weights = dtrain.get_weight()
    num_class = len(np.unique(labels))

    penalty_matrix = np.array([
        [0.0, 0.7, 1.0],   # Class 0 true
        [0.8, 0.0, 1.0],   # Class 1 true
        [1.2, 1.7, 0.0]    # Class 2 true
    ], dtype=np.float32)

    # Softmax prediction
    preds = preds.reshape(-1, num_class)
    preds -= np.max(preds, axis=1, keepdims=True)  # numerical stability
    exp_preds = np.exp(preds)
    probs = exp_preds / np.sum(exp_preds, axis=1, keepdims=True)

    # Create one-hot encoded labels
    onehot_labels = np.eye(num_class)[labels]

    # Get per-class penalties for each sample
    penalties = penalty_matrix[labels]

    # Adjust gradients with penalties
    grad = (probs - onehot_labels) * penalties
    grad *= sample_weights[:, None]

    # Hessian: simplified but safe
    hess = probs * (1.0 - probs)

    # Boost Hessian by penalty influence (approximate second-order)
    hess *= sample_weights[:, None]
    hess = np.clip(hess, 1e-6, 10.0)

    return grad, hess
```

An example Penalty matrix

```
[0.0, 0.7, 1.0],
[0.8, 0.0, 1.0],
[1.2, 1.7, 0.0]
```

Custom Penalty function for XGBoost

Code Snippets (ANN model architecture)

```
# Model definition with custom penalty loss
def build_ann_model(neurons=64, dropout_rate=0.3, activation='relu', hidden_layers=2):
    model = Sequential()
    model.add(Dense(neurons, activation=activation, input_shape=(X_train.shape[1],)))
    model.add(Dropout(dropout_rate))
    # Add hidden layers
    for _ in range(hidden_layers - 1):
        model.add(Dense(neurons, activation=activation))
        model.add(Dropout(dropout_rate))
    # Output layer
    model.add(Dense(3, activation='softmax'))  # Multi-class classification
    model.compile(optimizer='adam', loss=custom_penalty_loss, metrics=['accuracy'])
    return model
```

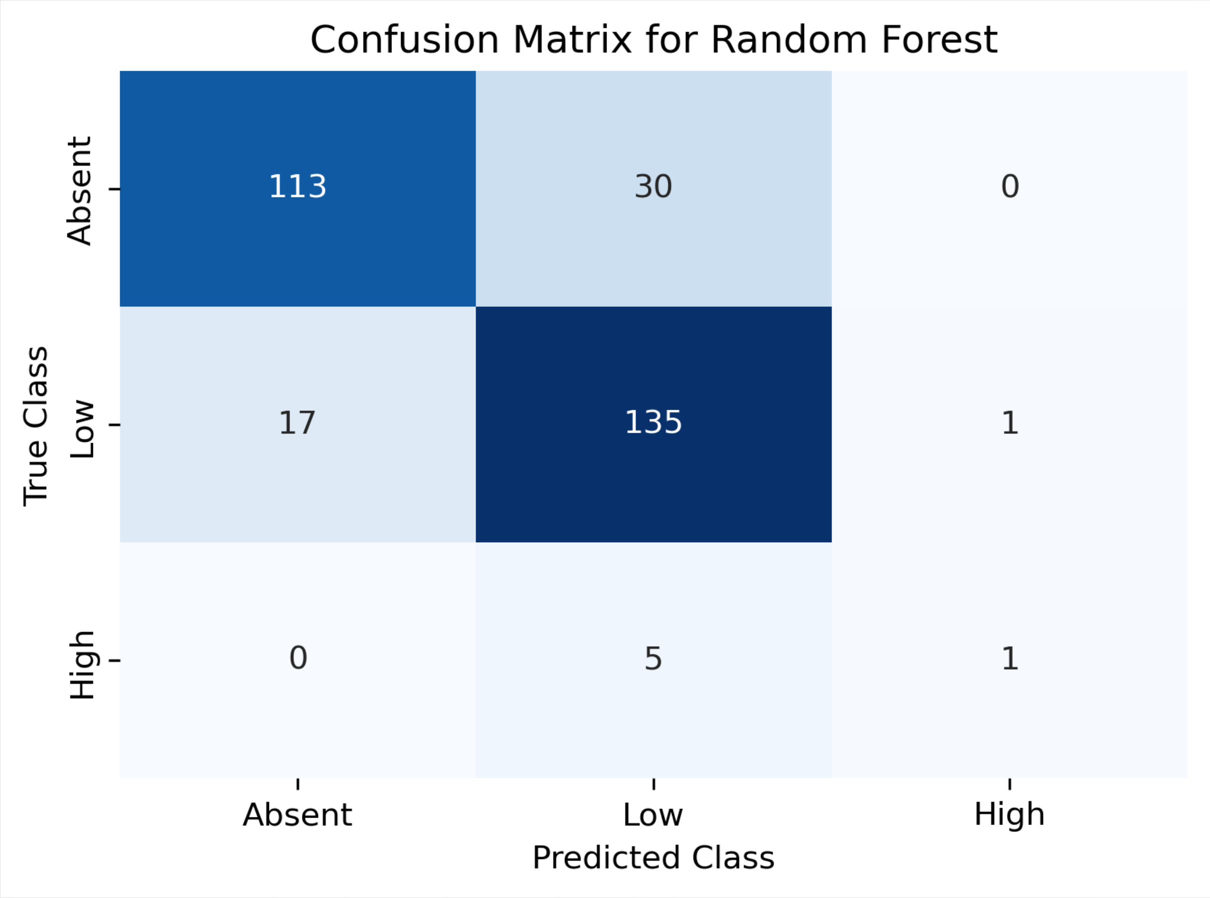
Initial Observations

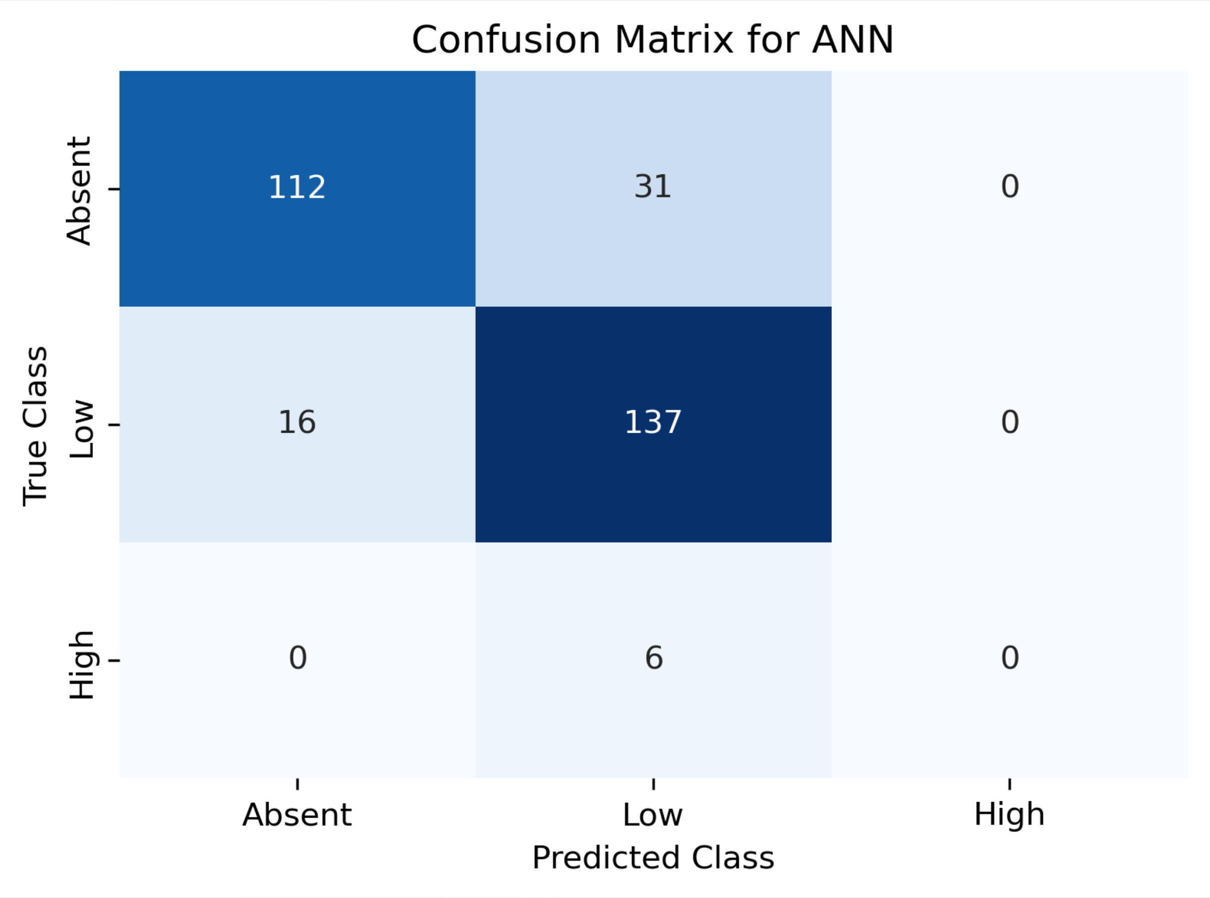
- Visual Index correlation analysis showed
  - positive correlation: water temperature
  - negative correlation: antecedent flow and velocity (last 1 month avg.), conductivity, dissolved ammonia, dissolved nitrate & nitrite, and turbidity.

- Dropped environmental variables to develop machine learning model:
  - 3-days, 7-days, 14-days, and 2-months average antecedent flows
  - 3-days, 7-days, 14-days, and 2-months average antecedent velocities

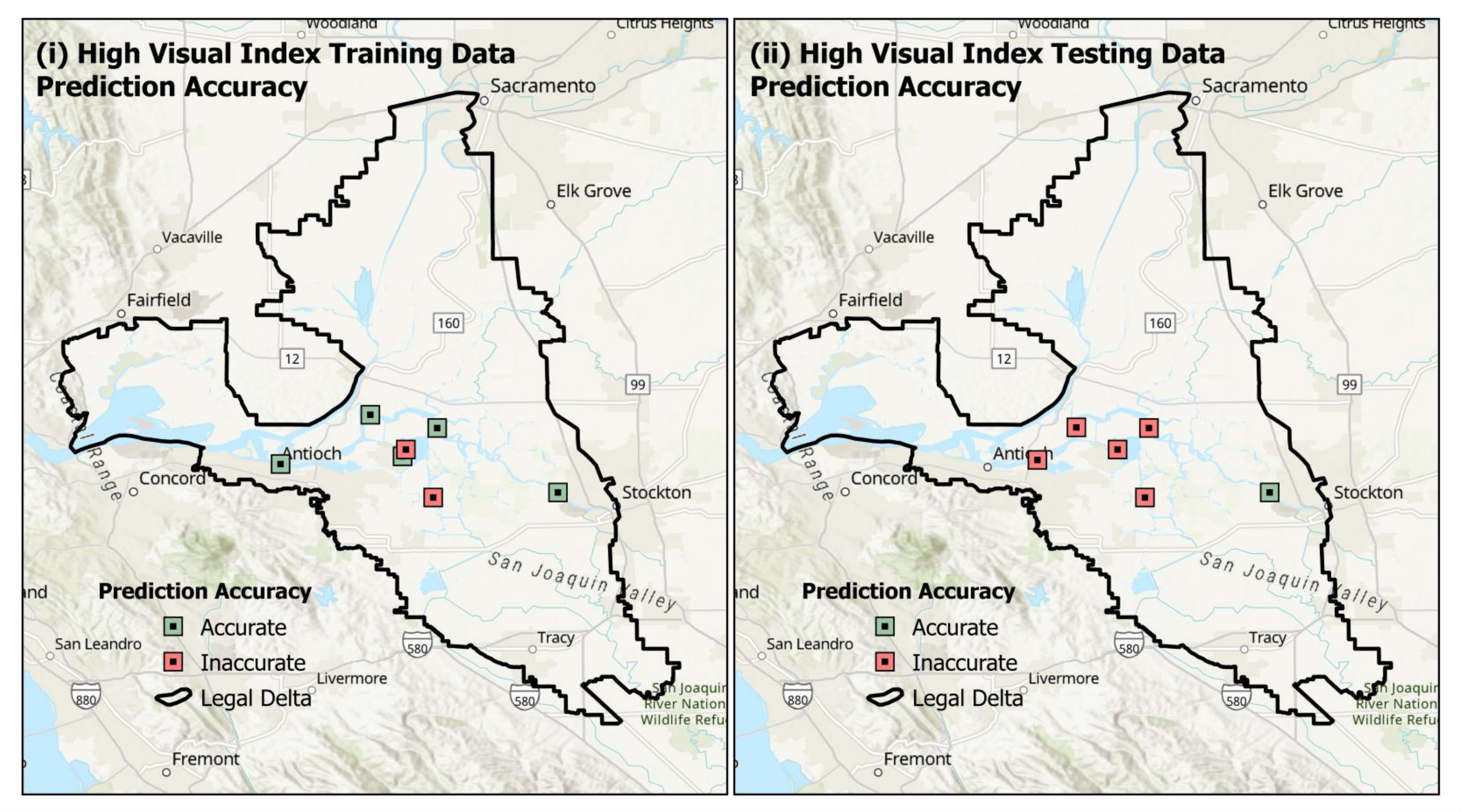
  - Dropped because of high correlation with 1-month average antecedent flow and velocity.

Confusion Matrices





Visual Index 3 Prediction Accuracy



Station	Station_1	Latitude	Longitude	Date	Temperature	Conductivity	TurbidityFNU	DissAmmonia	DissNitrateNitrite	DissOrthophos	1_month_avg_flow	1_month_avg_velocity	Visual_Index
EZ6-SJR	EZ6-SJR_10	38.02783	-121.73738	7/15/2021	21.65	3322	4.5	0.05	0.065	0.08	43.0094	0.0177	3
EZ2-SJR	EZ2-SJR_20	38.07761	-121.67781	7/15/2021	22.36	2102	4.8	0.05	0.111	0.081	12.9058	0.0122	3
D26	D26	38.07664	-121.5669	8/9/2016	22.76	196	8	0.1	0.19	0.05	-49.8753	-0.0084	3
D28A	D28A	37.97048	-121.573	7/11/2016	23.09	377	10	0.01	0.06	0.04	-70.0151	-0.0586	3
P8	P8	37.97817	-121.3823	8/11/2020	26.19	388	6.5	0.05	0.96	0.251	8.7877	0.0081	3
D19	D19	38.04376	-121.6148	9/6/2016	21.1	1026	7	0.04	0.17	0.06	0.7196	0.0048	3