

Supplementary Material
of
**Decoding Environmental Memory and Spatiotemporal
Heterogeneity of Algal Blooms: An Explainable Machine
Learning Framework with Ecological Thresholds**

Fei Li^a, Xin Sun^{a*1}, Yiqiang Wang^b, Weiqiao Wang^a, Ran Hao^a, Lihao Zhao^a

^aSchool of Environmental and Municipal Engineering, Xi'an University of Architecture and Technology, 13 Yanta Road, Xi'an 710055 China

^bShenzhen Water Group Co. Ltd., 1019 Shennan Road, Shenzhen 518031 China

¹ *Corresponding author: Xin Sun

School of Environmental and Municipal Engineering, Xi'an University of Architecture and Technology, Xi'an, 710055, PR China.

E-mail address: xinsunn@163.com

1 Feature selection

To balance information content against redundancy and multicollinearity, a three-stage feature-selection procedure was applied: mutual-information pre-screening, multicollinearity control and stepwise forward selection.

1.1 Mutual-information (MI) pre-screening

For each candidate predictor, the mutual information (MI) with the target variable was calculated to quantify non-linear dependence (Kraskov et al. 2004, Contreras-Reyes 2022, Jiang et al. 2022). Candidate features were then ranked in descending order of MI, and variables with higher information content were preferentially retained. For multiple lagged values and moving averages of the same physical variable, only the lag and MA with the highest MI were retained to reduce redundancy among highly similar variables. The corresponding statistics are reported in Table S1.

1.2 Multicollinearity control (VIF)

For the subset of variables retained after MI filtering, variance inflation factors (VIFs) were computed, and features with $VIF > 10$ were removed to control multicollinearity (Dormann et al. 2013, Vatcheva et al. 2016), thereby improving the stability and interpretability of subsequent models. Details of the elimination steps and the final retained sets are provided in Table S2.

1.3 Stepwise forward selection

On the basis of the reduced candidate set, stepwise forward selection was performed. Starting from a null model and using default hyperparameters, one candidate feature at a time was added and model performance on a validation set was evaluated. At each step, the feature that yielded the largest improvement in validation performance was retained. The procedure was terminated when adding further features no longer improved validation performance. This step ensured that an efficient and compact feature set was obtained without altering the underlying learning algorithms.

Table S1. Mutual information values for candidate predictors.

Feature	Mutual Information
WT	0.11976
WT_MA_15d	0.16541
WT_MA_10d	0.15916
WT_MA_7d	0.155714
WT_MA_5d	0.131705
WT_MA_3d	0.116991
WT_LV_3d	0.124396
WT_LV_10d	0.117198
WT_LV_7d	0.11282
WT_LV_5d	0.112345
WTLV_15d	0.099433

Feature	Mutual Information
WS2M	0.010648
WS2M_MA_7d	0.035362
WS2M_MA_10d	0.032724
WS2M_MA_15d	0.0267
WS2M_MA_5d	0.021828
WS2M_MA_3d	0.018394
WS2M_LV_15d	0.00613
WS2M_LV_3d	0.004125
WS2M_LV_5d	0.002502
WS2M_LV_7d	0
WS2M_LV_10d	0
WD2M	0.012258
WD2M_MA_15d	0.053972
WD2M_MA_7d	0.023424
WD2M_MA_10d	0.014118
WD2M_MA_5d	0.012123
WD2M_MA_3d	0.002814
WD2M_LV_15d	0.019698
WD2M_LV_3d	0.015169
WD2M_LV_10d	0.005756
WD2M_LV_7d	0.001448
WD2M_LV_5d	0.000162
TUR	0.159137

Feature	Mutual Information
TUR_MA_15d	0.173413
TUR_MA_10d	0.170854
TUR_MA_5d	0.167687
TUR_MA_7d	0.160112
TUR_MA_3d	0.156126
TUR_LV_3d	0.122039
TUR_LV_5d	0.115205
TUR_LV_10d	0.09989
TUR_LV_7d	0.097271
TUR_LV_15d	0.092447
TP	0.204086
TP_MA_15d	0.261636
TP_MA_10d	0.229782
TP_MA_7d	0.207209
TP_MA_5d	0.204959
TP_MA_3d	0.196615
TP_LV_5d	0.178401
TP_LV_3d	0.170414
TP_LV_10d	0.166753
TP_LV_7d	0.144278
TP_LV_15d	0.141625
SR	0.088843
SR_CUM_15d	0.093493

Feature	Mutual Information
SR_CUM_7d	0.089088
SR_CUM_3d	0.089002
SR_CUM_10d	0.084957
SR_CUM_5d	0.084056
SR_LV_5d	0.098166
SR_LV_lag_7d	0.091481
SR_LV_10d	0.088652
SR_LV_15d	0.085171
SR_LV_3d	0.084039
TN	0.149923
TN_MA_15d	0.181886
TN_MA_7d	0.177037
TN_MA_3d	0.171677
TN_MA_10d	0.169888
TN_MA_5d	0.169671
TN_LV_3d	0.147302
TN_LV_7d	0.128209
TN_LV_15d	0.127047
TN_LV_10d	0.124716
TN_LV_5d	0.116477
T2M	0.072182
T2M_MA_15d	0.139604
T2M_MA_10d	0.134058

Feature	Mutual Information
T2M_MA_7d	0.116755
T2M_MA_5d	0.098441
T2M_MA_3d	0.090759
T2M_LV_7d	0.067176
T2M_LV_10d	0.06707
T2M_LV_15d	0.066814
T2M_LV_5d	0.066313
T2M_LV_3d	0.058024
PS	0.055476
PS_MA_15d	0.122829
PS_MA_10d	0.090582
PS_MA_7d	0.090383
PS_MA_5d	0.070901
PS_MA_3d	0.051512
PS_LV_10d	0.057358
PS_LV_3d	0.05526
PS_LV_7d	0.054317
PS_LV_5d	0.04862
PS_LV_15d	0.035186
PRCP	0.001142
PRCP_CUM_15d	0.097359
PRCP_CUM_10d	0.072954
PRCP_CUM_7d	0.049766

Feature	Mutual Information
PRCP_CUM_5d	0.035207
PRCP_CUM_3d	0.005017
PRCP_LV_3d	0.007489
PRCP_LV_10d	0.007382
PRCP_LV_7d	0.003084
PRCP_LV_5d	0.003056
PRCP_LV_15d	0
pH	0.077691
pH_MA_15d	0.124306
pH_MA_10d	0.114697
pH_MA_5d	0.095012
pH_MA_3d	0.082099
pH_MA_7d	0.076152
pH_LV_7d	0.083942
pH_LV_5d	0.082454
pH_LV_15d	0.07451
pH_LV_3d	0.068846
pH_LV_10d	0.068809
N-P	0.138326
N-P_MA_15d	0.184927
N-P_MA_7d	0.154852
N-P_MA_10d	0.152914
N-P_MA_3d	0.149462

Feature	Mutual Information
N-P_MA_5d	0.138889
N-P_LV_3d	0.125829
N-P_LV_7d	0.115629
N-P_LV_5d	0.115453
N-P_LV_15d	0.102509
N-P_LV_10d	0.086576
NH3-N	0.034635
NH3-N_MA_15d	0.097491
NH3-N_MA_10d	0.071738
NH3-N_MA_7d	0.067802
NH3-N_MA_5d	0.058873
NH3-N_MA_3d	0.03687
NH3-N_LV_5d	0.024374
NH3-N_LV_7d	0.021643
NH3-N_LV_3d	0.020067
NH3-N_LV_10d	0.01857
NH3-N_LV_15d	0.007807
DO	0.082757
DO_MA_15d	0.108051
DO_MA_7d	0.095049
DO_MA_5d	0.094611
DO_MA_10d	0.092752
DO_MA_3d	0.074074

Feature	Mutual Information
DO_LV_3d	0.09685
DO_LV_10d	0.086583
DO_LV_5d	0.083133
DO_LV_7d	0.075876
DO_LV_15d	0.059176
CON	0.112804
CON_MA_15d	0.169601
CON_MA_10d	0.154025
CON_MA_7d	0.150379
CON_MA_3d	0.133721
CON_MA_5d	0.130344
CON_LV_15d	0.121137
CON_LV_7d	0.120467
CON_LV_5d	0.11048
CON_LV_10d	0.10721
CON_LV_3d	0.10061
PI	0.174527
PI_MA_15d	0.18821
PI_MA_5d	0.173016
PI_MA_10d	0.168803
PI_MA_3d	0.154946
PI_MA_7d	0.153646
PI_LV_5d	0.124246

Feature	Mutual Information
PI_LV_3d	0.123946
PI_LV_7d	0.12029
PI_LV_10d	0.112079
PI_LV_15d	0.098679

Table S2. Predictors removed during variance-inflation-factor screening.

Feature	VIF
SR_LV_5d	236.08
WT_MA_15d	78.89
WT	52
SR_CUM_15d	28.17
PS_MA_15d	24.04
TP_MA_15d	17.6
CON_MA_15d	13.85
TN_MA_15d	12.26
WT_LV_3d	11.54

Table S2b. Final predictor sets retained for the XGBoost models used in explainability and transition analyses.

Scheme	Data subset	Number of predictors	Retained predictors
W1	Global	11	TP, COD, TUR, TOA_SW_DWN, CON, TN, WT, N-P, pH, DO, NH3-N
W1	Stable–East	10	TUR, COD, TN, TOA_SW_DWN, DO, pH, TP, WT, NH3-N, CON
W1	Stable–West	10	WT, NH3-N, pH, TOA_SW_DWN, CON, TN, COD, N-P, TUR, TP
W1	Outbreak–East	9	TN, NH3-N, CON, TUR, TOA_SW_DWN, TP, COD, N-P, DO
W1	Outbreak–West	9	NH3-N, pH, CON, TOA_SW_DWN, COD, WT, TUR, N-P, TP
W1	Decline–East	10	TUR, pH, CON, TP, TOA_SW_DWN, DO, TN, NH3-N, T2M, COD
W1	Decline–West	9	NH3-N, TUR, CON, TOA_SW_DWN, TN, DO, WT, pH, N-P
W2	Global	11	TP, PI, NH3-N_MA_15d, TOA_SW_DWN, TUR_MA_15d, pH_MA_15d, N-P_MA_15d, PI_MA_15d, CON, DO_MA_15d, CON_LV_15d
W2	Stable–East	17	TUR, TUR_LV_3d, pH_MA_15d, PI_MA_15d, TOA_SW_DWN, TN, DO_MA_15d, PI, CON_LV_15d, N-P_MA_15d, NH3-N_MA_15d, PRECTOTCORR_CUM_15d, TP, CON, TUR_MA_15d, DO, NH3-N_LV_5d
W2	Stable–West	12	NH3-N_MA_15d, NH3-N, N-P_MA_15d, TOA_SW_DWN, pH_MA_15d, PI_MA_15d, TUR_MA_15d, PRECTOTCORR_CUM_15d, CON_LV_15d, DO_MA_15d, WD2M_MA_15d, CON
W2	Outbreak–East	12	pH_MA_15d, NH3-N_MA_15d, TUR_MA_15d, TOA_SW_DWN, PI_MA_15d, N-P_MA_15d, TP, TN, CON_LV_15d, CON, DO_MA_15d, WD2M_MA_15d
W2	Outbreak–West	13	NH3-N_MA_15d, NH3-N, CON, pH_MA_15d, TOA_SW_DWN, N-P_MA_15d, PI_MA_15d, TUR_MA_15d, PRECTOTCORR_CUM_15d, PI, CON_LV_15d, DO_MA_15d, WD2M_MA_15d
W2	Decline–East	12	NH3-N_MA_15d, pH_LV_7d, N-P_MA_15d, DO_MA_15d, TUR, CON, pH_MA_15d, TOA_SW_DWN, TUR_MA_15d, PI_MA_15d, TP, TN
W2	Decline–West	17	NH3-N_MA_15d, NH3-N, N-P_MA_15d, CON_LV_15d, TUR_MA_15d, DO_MA_15d, PI_MA_15d, TOA_SW_DWN, PI, CON, TN, pH_MA_15d, N-P, PRECTOTCORR_CUM_15d, WD2M_MA_15d, NH3-N_LV_5d, DO_LV_3d

Feature selection was performed separately for each data subset and predictor scheme. The table reports the XGBoost feature sets because XGBoost-W2 was used for the integrated-importance and SHAP analyses

2 Machine-learning model development

2.1 Overall Experimental Design

Several representative supervised-learning algorithms were considered for algal-density modelling, including Random Forest (RF) (Biau and Scornet 2016), k-Nearest Neighbours (KNN) (Zhang and Zhou 2007), Extreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016) and Light Gradient Boosting Machine (LightGBM) (Hajihosseini et al. 2023). RF aggregates multiple decision trees with random sampling of features and samples, thereby reducing variance and improving generalisation for high-dimensional, non-linear data (Biau and Scornet 2016). XGBoost and LightGBM are efficient implementations of gradient-boosting decision trees (GBDT) with strong non-linear fitting capacity and automatic feature selection, and they perform particularly well for sparse features and complex feature interactions (Chen and Guestrin 2016, Hajihosseini et al. 2023). KNN, as a distance-based lazy-learning algorithm, provides a structurally distinct baseline for comparison with tree-based methods (Zhang and Zhou 2007). These algorithms have been widely applied in chlorophyll-a and harmful algal bloom prediction and have demonstrated high predictive accuracy and robustness in lakes and reservoirs (Hajihosseini et al. 2023, Tian et al. 2024, Nguyen et al. 2025, Song et al. 2025, Sun et al. 2025), making them suitable candidates for algal-density modelling in this study.

On the basis of the preprocessed data and constructed features, a unified modelling workflow was adopted. First, the 8,259 processed samples were randomly divided into a training set and a test set at a ratio of 4:1; the test set remained completely independent during model training and hyperparameter tuning and was used only for final performance evaluation (Tao et al. 2024). Five-fold cross-validation was then carried out within the training set to assess generalisation and guide hyperparameter selection. Specifically, the training set was partitioned into five approximately equal folds; in each round, one fold was used as the validation set and the remaining four folds were used for model training. This procedure was repeated five times so that each fold served once as the validation set, and the average of the five validation results was taken as the performance estimate for a given hyperparameter combination (Kohavi 1995, Arlot and Celisse 2010). This strategy maximises the use of limited samples, reduces the randomness of a single split and provides a consistent evaluation basis across algorithms.

Hyperparameters were tuned using grid search combined with five-fold cross-validation. For each algorithm, discrete candidate sets were defined for key hyperparameters, such as the number of trees, maximum depth and minimum samples per split for RF; learning rate, number of base learners and maximum number of leaves for XGBoost and LightGBM; and the number of neighbours and distance metric for KNN. Grid search

exhaustively evaluated all combinations in the predefined hyperparameter grid. For each combination, full five-fold cross-validation was conducted on the training set, and performance metrics such as R^2 and mean squared error (MSE) were computed. The combination with the best average performance was selected as the optimal configuration for that algorithm (Bergstra and Bengio 2012, Hajihosseini et al. 2023). After determination of the optimal hyperparameters, the final model was retrained using the entire training set and evaluated on the independent test set to assess predictive accuracy and stability.

This modelling and tuning procedure was applied to the global dataset as well as to the eastern and western subsets for the stable, outbreak and decline periods, under both feature schemes without lagged variables (W1) and with lagged variables (W2), thereby ensuring consistent and reproducible model construction across spatiotemporal scenarios and feature configurations.

2.2 Hyperparameter Optimization Methodology by Model

2.2.1 Random Forest (RF)

RF aggregates multiple decision trees to reduce variance and improve generalization. We optimized six hyperparameters to balance performance and computational cost.

Search Space:

n_estimators: [50, 100, 200, 300]. This controls the scale of the forest;

larger values improve performance but increase cost.

`max_depth`: [10, 15, 20]. Explores medium to high complexity trees to capture features without overfitting.

`min_samples_split`: [2, 5, 10]. Pre-pruning parameter to control node splitting.

`min_samples_leaf`: [1, 2, 4]. Pre-pruning parameter to set the minimum samples required at a leaf node.

`n_bins`: [128, 256]. Balances GPU efficiency with the precision of split points.

`split_criterion`: ['mse', 'poisson']. Tests adaptability to different response variable distributions (continuous vs. count-based).

2.2.2 k-Nearest Neighbours (KNN)

KNN serves as a distance-based baseline algorithm. We optimized the neighbor count and weight function.

Search Space:

`n_neighbors`: [3, 5, 7, 9, 11, 13, 15]. This determines model smoothness; smaller k values are sensitive to local noise, while larger k values provide smoother predictions but may miss local features.

`weights`: ['uniform', 'distance']. Compares equal weighting versus distance-based weighting (closer neighbors contribute more).

`algorithm`: Fixed to 'brute' for consistent GPU performance.

2.2.3 XGBoost

XGBoost provides strong non-linear fitting capacity. We performed sequential tuning on tree structure and regularization parameters.

Search Space:

max_depth: [3, 5, 7, 9]. Balances model expressiveness and overfitting risk.

eta (Learning Rate): [0.01, 0.05, 0.1, 0.2]. Controls the contribution of each tree; lower rates require more trees but yield robust models.

gamma: [0, 0.1, 0.3]. Specifies the minimum loss reduction required to make a split (regularization).

subsample & colsample_bytree: [0.7, 0.8, 0.9]. Row and column sampling ratios to introduce randomness and improve generalization.

num_boost_round: [50, 100, 200, 300]. Optimized alongside early stopping.

2.2.4 LightGBM

LightGBM is optimized for efficiency and handling complex feature interactions. Parameters were tuned to accommodate its leaf-wise growth strategy.

Search Space:

num_leaves: [15, 31, 63, 127]. Key complexity control; leaf-wise growth requires more leaves than depth-wise methods.

learning_rate: [0.01, 0.05, 0.1, 0.2]. balances training speed and accuracy.

feature_fraction & bagging_fraction: [0.7, 0.8, 0.9]. Increases diversity and suppresses overfitting via subsampling.

min_data_in_leaf: [5, 10, 20, 50]. Critical regularization to prevent leaves with too few samples.

2.3 Optimization Results

The optimal hyperparameter configurations were determined for the global dataset and specific spatiotemporal subsets (Stable, Outbreak, and Decline periods; East and West banks) under two feature schemes: W1 (without lagged variables) and W2 (with lagged variables). The corresponding statistics are reported in Table S3.

Table S3a Optimal hyperparameters for RF.

Scheme	Data subset	Algorithm	Optimal hyperparameters
W1	Global	RF	n_estimators=300; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=poisson
W1	Stable–East	RF	n_estimators=300; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=mse
W1	Stable–West	RF	n_estimators=300; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=poisson
W1	Outbreak–East	RF	n_estimators=100; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=poisson
W1	Outbreak–West	RF	n_estimators=300; max_depth=20; min_samples_split=5; min_samples_leaf=1; n_bins=256; split_criterion=poisson
W1	Decline–East	RF	n_estimators=100; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=mse
W1	Decline–West	RF	n_estimators=100; max_depth=15; min_samples_split=2; min_samples_leaf=1; n_bins=256; split_criterion=mse
W2	Global	RF	n_estimators=200; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=poisson
W2	Stable–East	RF	n_estimators=300; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=mse
W2	Stable–West	RF	n_estimators=300; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=poisson
W2	Outbreak–East	RF	n_estimators=100; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=256; split_criterion=poisson
W2	Outbreak–West	RF	n_estimators=300; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=mse
W2	Decline–East	RF	n_estimators=100; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=mse
W2	Decline–West	RF	n_estimators=300; max_depth=20; min_samples_split=2; min_samples_leaf=1; n_bins=128; split_criterion=mse

Table S3b Optimal hyperparameters for KNN.

Scheme	Data subset	Algorithm	Optimal hyperparameters
W1	Global	KNN	n_neighbors=5; weights=uniform
W1	Stable - East	KNN	n_neighbors=3; weights=uniform
W1	Stable - West	KNN	n_neighbors=5; weights=uniform
W1	Outbreak - East	KNN	n_neighbors=3; weights=uniform
W1	Outbreak - West	KNN	n_neighbors=5; weights=uniform
W1	Decline - East	KNN	n_neighbors=3; weights=uniform
W1	Decline - West	KNN	n_neighbors=3; weights=uniform
W2	Global	KNN	n_neighbors=3; weights=uniform
W2	Stable - East	KNN	n_neighbors=3; weights=uniform
W2	Stable - West	KNN	n_neighbors=3; weights=uniform
W2	Outbreak - East	KNN	n_neighbors=3; weights=uniform
W2	Outbreak - West	KNN	n_neighbors=3; weights=uniform
W2	Decline - East	KNN	n_neighbors=3; weights=uniform
W2	Decline - West	KNN	n_neighbors=3; weights=uniform

Table S3c Optimal hyperparameters for XGBoost.

Scheme	Data subset	Algorithm	Optimal hyperparameters
W1	Global	XGBoost	max_depth=9; eta=0.1; gamma=0; subsample=0.9; colsample_bytree=0.9; num_boost_round=300
W1	Stable - East	XGBoost	max_depth=7; eta=0.1; gamma=0; subsample=0.8; colsample_bytree=0.7; num_boost_round=100
W1	Stable - West	XGBoost	max_depth=7; eta=0.1; gamma=0; subsample=0.7; colsample_bytree=0.9; num_boost_round=300
W1	Outbreak - East	XGBoost	max_depth=7; eta=0.1; gamma=0; subsample=0.9; colsample_bytree=0.8; num_boost_round=300
W1	Outbreak - West	XGBoost	max_depth=7; eta=0.05; gamma=0; subsample=0.7; colsample_bytree=0.9; num_boost_round=300
W1	Decline - East	XGBoost	max_depth=9; eta=0.05; gamma=0; subsample=0.8; colsample_bytree=0.8; num_boost_round=100
W1	Decline - West	XGBoost	max_depth=7; eta=0.1; gamma=0; subsample=0.8; colsample_bytree=0.9; num_boost_round=300
W2	Global	XGBoost	max_depth=9; eta=0.1; gamma=0; subsample=0.9; colsample_bytree=0.8; num_boost_round=300
W2	Stable - East	XGBoost	max_depth=7; eta=0.1; gamma=0; subsample=0.8; colsample_bytree=0.8; num_boost_round=100
W2	Stable - West	XGBoost	max_depth=9; eta=0.05; gamma=0; subsample=0.8; colsample_bytree=0.9; num_boost_round=300
W2	Outbreak - East	XGBoost	max_depth=9; eta=0.1; gamma=0; subsample=0.8; colsample_bytree=0.8; num_boost_round=100
W2	Outbreak - West	XGBoost	max_depth=7; eta=0.1; gamma=0; subsample=0.8; colsample_bytree=0.7; num_boost_round=200
W2	Decline - East	XGBoost	max_depth=9; eta=0.1; gamma=0; subsample=0.8; colsample_bytree=0.8; num_boost_round=100
W2	Decline - West	XGBoost	max_depth=9; eta=0.1; gamma=0; subsample=0.8; colsample_bytree=0.9; num_boost_round=100

Table S3d Optimal hyperparameters for LightGBM.

Scheme	Data subset	Algorithm	Optimal hyperparameters
W1	Global	LightGBM	num_leaves=127; learning_rate=0.1; feature_fraction=0.9; bagging_fraction=0.8; min_data_in_leaf=5; num_boost_round=300
W1	Stable - East	LightGBM	num_leaves=63; learning_rate=0.2; feature_fraction=0.7; bagging_fraction=0.8; min_data_in_leaf=10; num_boost_round=100
W1	Stable - West	LightGBM	num_leaves=63; learning_rate=0.2; feature_fraction=0.8; bagging_fraction=0.8; min_data_in_leaf=20; num_boost_round=300
W1	Outbreak - East	LightGBM	num_leaves=31; learning_rate=0.2; feature_fraction=0.9; bagging_fraction=0.8; min_data_in_leaf=5; num_boost_round=200
W1	Outbreak - West	LightGBM	num_leaves=31; learning_rate=0.2; feature_fraction=0.8; bagging_fraction=0.8; min_data_in_leaf=10; num_boost_round=200
W1	Decline - East	LightGBM	num_leaves=31; learning_rate=0.2; feature_fraction=0.9; bagging_fraction=0.8; min_data_in_leaf=5; num_boost_round=100
W1	Decline - West	LightGBM	num_leaves=31; learning_rate=0.2; feature_fraction=0.8; bagging_fraction=0.8; min_data_in_leaf=5; num_boost_round=300
W2	Global	LightGBM	num_leaves=127; learning_rate=0.2; feature_fraction=0.9; bagging_fraction=0.8; min_data_in_leaf=20; num_boost_round=300

Scheme	Data subset	Algorithm	Optimal hyperparameters
W2	Stable - East	LightGBM	num_leaves=31; learning_rate=0.2; feature_fraction=0.9; bagging_fraction=0.8; min_data_in_leaf=10; num_boost_round=100
W2	Stable - West	LightGBM	num_leaves=63; learning_rate=0.2; feature_fraction=0.7; bagging_fraction=0.8; min_data_in_leaf=10; num_boost_round=200
W2	Outbreak - East	LightGBM	num_leaves=31; learning_rate=0.2; feature_fraction=0.9; bagging_fraction=0.8; min_data_in_leaf=20; num_boost_round=50
W2	Outbreak - West	LightGBM	num_leaves=31; learning_rate=0.1; feature_fraction=0.8; bagging_fraction=0.8; min_data_in_leaf=10; num_boost_round=300
W2	Decline - East	LightGBM	num_leaves=15; learning_rate=0.2; feature_fraction=0.7; bagging_fraction=0.8; min_data_in_leaf=5; num_boost_round=100
W2	Decline - West	LightGBM	num_leaves=31; learning_rate=0.1; feature_fraction=0.9; bagging_fraction=0.8; min_data_in_leaf=5; num_boost_round=100

Table S4 Independent-test performance for all algorithms, predictor schemes, and data subsets.

Scheme	Data subset	Algorithm	R²	MSE ($\times 10^{12}$)	RMSE ($\times 10^6$)	MAE ($\times 10^6$)
W1	Global	RF	0.7008	33.285	5.769	3.785
W1	Global	KNN	0.4270	63.738	7.984	5.111
W1	Global	XGBoost	0.7505	27.755	5.268	3.387
W1	Global	LightGBM	0.7540	27.360	5.231	3.322
W1	Stable–East	RF	0.7486	22.105	4.702	2.527
W1	Stable–East	KNN	0.5113	42.976	6.556	3.474
W1	Stable–East	XGBoost	0.7682	20.383	4.515	2.456
W1	Stable–East	LightGBM	0.7734	19.927	4.464	2.452
W1	Stable–West	RF	0.6928	43.434	6.590	4.092

Scheme	Data subset	Algorithm	R ²	MSE ($\times 10^{12}$)	RMSE ($\times 10^6$)	MAE ($\times 10^6$)
W1	Stable–West	KNN	0.5478	63.927	7.995	5.037
W1	Stable–West	XGBoost	0.7371	37.168	6.097	3.813
W1	Stable–West	LightGBM	0.7626	33.562	5.793	3.705
W1	Outbreak–East	RF	0.6900	25.621	5.062	3.336
W1	Outbreak–East	KNN	0.6604	28.068	5.298	2.981
W1	Outbreak–East	XGBoost	0.7855	17.727	4.210	2.742
W1	Outbreak–East	LightGBM	0.7909	17.288	4.158	2.787
W1	Outbreak–West	RF	0.6562	29.848	5.463	4.082
W1	Outbreak–West	KNN	0.3305	58.116	7.623	5.638
W1	Outbreak–West	XGBoost	0.7139	24.836	4.984	3.621

Scheme	Data subset	Algorithm	R ²	MSE ($\times 10^{12}$)	RMSE ($\times 10^6$)	MAE ($\times 10^6$)
W1	Outbreak–West	LightGBM	0.7231	24.036	4.903	3.587
W1	Decline–East	RF	0.7463	29.596	5.440	3.394
W1	Decline–East	KNN	0.5387	53.809	7.335	4.056
W1	Decline–East	XGBoost	0.7712	26.690	5.166	3.115
W1	Decline–East	LightGBM	0.7768	26.041	5.103	3.048
W1	Decline–West	RF	0.7309	16.285	4.035	2.970
W1	Decline–West	KNN	0.6023	24.067	4.906	3.343
W1	Decline–West	XGBoost	0.7705	13.890	3.727	2.627
W1	Decline–West	LightGBM	0.7759	13.563	3.683	2.700
W2	Global	RF	0.8179	20.251	4.500	2.788

Scheme	Data subset	Algorithm	R ²	MSE ($\times 10^{12}$)	RMSE ($\times 10^6$)	MAE ($\times 10^6$)
W2	Global	KNN	0.7991	22.343	4.727	2.520
W2	Global	XGBoost	0.8490	16.802	4.099	2.539
W2	Global	LightGBM	0.8431	17.450	4.177	2.631
W2	Stable–East	RF	0.7902	18.446	4.295	2.233
W2	Stable–East	KNN	0.6934	26.965	5.193	2.435
W2	Stable–East	XGBoost	0.8153	16.238	4.030	2.074
W2	Stable–East	LightGBM	0.8216	15.688	3.961	2.077
W2	Stable–West	RF	0.8796	17.022	4.126	2.494
W2	Stable–West	KNN	0.8379	22.917	4.787	2.227
W2	Stable–West	XGBoost	0.9007	14.034	3.746	2.063

Scheme	Data subset	Algorithm	R ²	MSE ($\times 10^{12}$)	RMSE ($\times 10^6$)	MAE ($\times 10^6$)
W2	Stable–West	LightGBM	0.8899	15.566	3.945	2.391
W2	Outbreak–East	RF	0.8142	15.360	3.919	2.453
W2	Outbreak–East	KNN	0.8809	9.846	3.138	1.693
W2	Outbreak–East	XGBoost	0.8452	12.794	3.577	2.267
W2	Outbreak–East	LightGBM	0.8397	13.247	3.640	2.434
W2	Outbreak–West	RF	0.7830	18.837	4.340	3.332
W2	Outbreak–West	KNN	0.7877	18.426	4.293	3.130
W2	Outbreak–West	XGBoost	0.8213	15.514	3.939	2.901
W2	Outbreak–West	LightGBM	0.8266	15.050	3.879	2.865
W2	Decline–East	RF	0.8453	18.043	4.248	2.719

Scheme	Data subset	Algorithm	R ²	MSE ($\times 10^{12}$)	RMSE ($\times 10^6$)	MAE ($\times 10^6$)
W2	Decline–East	KNN	0.8044	22.822	4.777	2.399
W2	Decline–East	XGBoost	0.8800	13.998	3.741	2.241
W2	Decline–East	LightGBM	0.8836	13.581	3.685	2.082
W2	Decline–West	RF	0.7995	12.136	3.484	2.421
W2	Decline–West	KNN	0.7584	14.623	3.824	2.417
W2	Decline–West	XGBoost	0.8266	10.495	3.240	2.094
W2	Decline–West	LightGBM	0.8191	10.948	3.309	2.204

MSE is expressed in 10^{12} (cells/L)²; RMSE and MAE are expressed in 10^6 cells/L. W1 and W2 used matched independent test sets within each data subset.

3 Explainability tools

To provide mechanistic interpretation for the predictive models, structural, perturbation-based and game-theoretic perspectives were combined to build an explainability framework and analyse key drivers and response patterns of algal density. In recent years, explainable artificial intelligence (XAI) methods such as feature importance, accumulated local effects and SHAP have been widely applied in environmental and water-quality studies and increasingly used to identify algal-bloom drivers and threshold effects, thereby enhancing the usability and credibility of models in water-environment management.

3.1 Model-based (structural) feature importance

Based on XGBoost and other tree-based models, model-based feature importance was first calculated. Mean decrease in impurity (MDI) was used by accumulating the reduction in the loss function attributable to each feature across all splitting nodes, weighted by the number of samples, and then normalising the results to obtain the relative contribution of each feature to the reduction in overall prediction error (Louppe et al. 2013). This index is directly derived from the training process and provides an efficient global ranking of feature importance. However, MDI may be biased when features are correlated or have different value ranges (Strobl et al. 2007). In this study, it was therefore mainly used as a baseline for comparison with perturbation-based and game-theoretic importance

measures rather than for stand-alone mechanistic interpretation.

3.2 Perturbation-based feature importance (permutation importance)

To quantify the practical impact of each environmental factor on model performance, permutation importance was further applied. For a given feature, its values in the test set were randomly permuted while all other features were kept unchanged, and the R^2 and MSE of the model on the perturbed data were recalculated. The decrease in performance relative to the original data was taken as the importance of that feature (Fisher et al. 2019). Compared with structural indices, permutation importance directly reflects the deterioration in model performance when a given predictor is disrupted and can partially alleviate the tendency of tree-based models to favour high-frequency or wide-range features. Because environmental predictors in algal-bloom studies are often correlated, the resulting importance scores were interpreted with attention to potential interactions rather than over-emphasising any single feature.

3.3 Game-theoretic SHAP decomposition and response curves

Building on the global importance measures, the SHAP (SHapley Additive exPlanations) method was introduced to decompose model outputs from a game-theoretic perspective. SHAP treats the model prediction as a cooperative payoff generated by all features and computes a Shapley value for each feature at each sample, thereby quantifying its positive or negative contribution at the local level and summarising its

global importance and dominant effect direction across the dataset (Lundberg and Lee 2017). In environmental and algal-bloom studies, SHAP has been widely used to identify the relative importance of water temperature, nutrients, light and other factors and to characterise their spatiotemporal patterns, providing a quantitative basis for management decisions (Lee et al. 2022, Chen et al. 2025, Gao et al. 2025, Ma et al. 2025).

Taken together, structural feature importance provides an efficient overview of model sensitivity, permutation importance quantifies the contribution of each feature to predictive performance and SHAP-based analyses reveal non-linearity, interactions and threshold behaviour at the local scale. The combination of these tools enables comparison of the relative importance of drivers across models and scenarios and supports detailed interpretation of how lagged versus instantaneous features operate at different time scales, thereby providing more actionable mechanistic insights for designing algal-bloom control strategies.

3.4 segmented-SHAP analysis

For each leading predictor, continuous piecewise-linear models with zero to three breakpoints were compared using BIC. A segmented model was considered supported when its BIC improved by at least 2 relative to a linear model. Each fitted segment was required to contain at least 30 observations and at least 15% of the available data.

Uncertainty was assessed using non-parametric bootstrap resampling. Single-breakpoint analyses used 300 replicates, whereas multi-breakpoint analyses used 100 replicates to estimate the recovery of the selected breakpoint number and ordered confidence intervals. Supported values are reported as model-derived transition ranges rather than universal regulatory thresholds.

Table S5 Integrated importance rankings of the three leading predictors in each XGBoost-W2 model.

Model	Rank	Predictor	Integrated score
Global	1	TOA_SW_DWN	0.893
Global	2	TP	0.882
Global	3	N-P_MA_15d	0.781
Stable–East	1	TUR	0.900
Stable–East	2	DO	0.688
Stable–East	3	DO_MA_15d	0.580
Stable–West	1	TOA_SW_DWN	1.000
Stable–West	2	pH_MA_15d	0.535
Stable–West	3	TUR_MA_15d	0.491
Outbreak–East	1	TP	0.960
Outbreak–East	2	TN	0.715
Outbreak–East	3	pH_MA_15d	0.573

Outbreak–West	1	pH_MA_15d	0.962
Outbreak–West	2	PI	0.914
Outbreak–West	3	PRECTOTCORR_CUM_15d	0.724
Decline–East	1	TP	1.000
Decline–East	2	TUR	0.645
Decline–East	3	N-P_MA_15d	0.449
Decline–West	1	N-P_MA_15d	1.000
Decline–West	2	TUR_MA_15d	0.647
Decline–West	3	CON_LV_15d	0.428

Table S6 Segmented-SHAP transition estimates and uncertainty assessment for the leading predictors.

Model	Rank	Predictor	Reported transition(s)	95% CI	Model evidence	Bootstrap support	Reporting status
Global	1	TOA_SW_DWN	9.87	9.755–10.01	$\Delta\text{BIC}=205.33$	1.00	Single breakpoint retained
Global	2	TP	0.05185	0.05133–0.05296	$\Delta\text{BIC}=1529.23$	1.00	Single breakpoint retained
Global	3	N-P_MA_15d	24.85	24.15–25.28	$\Delta\text{BIC}=1149.92$	1.00	Single breakpoint retained
Stable–East	1	TUR	39.34	34.47–43.16	$\Delta\text{BIC}=84.56$	1.00	Single breakpoint retained
Stable–East	2	DO	10.16	9.981–10.33	$\Delta\text{BIC}=132.49$	1.00	Single breakpoint retained
Stable–East	3	DO_MA_15d	11.1	10.7–11.74	$\Delta\text{BIC}=72.41$	1.00	Single breakpoint retained
Stable–West	1	TOA_SW_DWN	9.11	8.902–9.53	$\Delta\text{BIC}=27.08$	1.00	Single breakpoint retained
Stable–West	2	pH_MA_15d	7.781	7.752–8.495	$\Delta\text{BIC}=6.30$	0.86	Single breakpoint retained
Stable–West	3	TUR_MA_15d	19.68	16.72–21.58	$\Delta\text{BIC}=48.30$	1.00	Single breakpoint retained
Outbreak–East	1	TP	0.02529	0.02045–0.05388	$\Delta\text{BIC}=49.03$	1.00	Single breakpoint retained
Outbreak–East	2	TN	0.6784	0.645–0.866	$\Delta\text{BIC}=44.00$	1.00	Single breakpoint retained
Outbreak–East	3	pH_MA_15d	8.766	8.359–8.801	$\Delta\text{BIC}=44.97$	1.00	Single breakpoint retained
Outbreak–West	1	pH_MA_15d	7.822	7.757–8.601	$\Delta\text{BIC}=20.16$	0.99	Single breakpoint retained

Model	Rank	Predictor	Reported transition(s)	95% CI	Model evidence	Bootstrap support	Reporting status
Outbreak–West	2	PI	4.566	4.396–4.765	Δ BIC=246.51	1.00	Single breakpoint retained
Outbreak–West	3	PRETOTCORR_CUM_15d	24.07; 68.854002	Broad and overlapping	BIC-selected 2-breakpoint model	same-k=0.79	Exploratory transition ranges
Decline–East	1	TP	0.07024	0.06574–0.07295	Δ BIC=168.92	1.00	Single breakpoint retained
Decline–East	2	TUR	51.561686; 73.452563	See ordered bootstrap intervals in Results/Table 2	BIC-selected 2-breakpoint model	same-k=0.99	Multi-breakpoint result retained
Decline–East	3	N-P_MA_15d	24.6	23.77–25.66	Δ BIC=103.21	1.00	Single breakpoint retained
Decline–West	1	N-P_MA_15d	27	25.36–28.72	Δ BIC=87.96	1.00	Single breakpoint retained
Decline–West	2	TUR_MA_15d	44.066409; 51.803833	See ordered bootstrap intervals in Results/Table 2	BIC-selected 2-breakpoint model	same-k=0.80	Multi-breakpoint result retained
Decline–West	3	CON_LV_15d	319.6	312.8–356	Δ BIC=34.34	0.98	Single breakpoint retained

Single-breakpoint models were retained for standardized reporting when BIC and bootstrap support were adequate. Multi-breakpoint models were retained for decline-stage turbidity because the single-breakpoint fits were weak. The western-outbreak precipitation result is exploratory.

Table S7. East–west water-quality contrasts evaluated with Welch's t-test, Benjamini–Hochberg correction, and Hedges' g.

Variable	West n	East n	West mean	East mean	Welch t	Adjusted p	Hedges g
pH	4284	3975	8.11	8.375	-29.210	<0.001	-0.638
WT	4284	3975	18.86	19.69	-4.280	<0.001	-0.094
DO	4284	3975	9.749	9.304	10.369	<0.001	0.227
PI	4284	3975	3.987	2.769	57.700	<0.001	1.261
NH3-N	4284	3975	0.02546	0.02547	-0.156	0.883	-0.003
TP	4284	3975	0.06938	0.03867	48.762	<0.001	1.062
TN	4284	3975	1.221	0.9332	22.896	<0.001	0.502
CON	4284	3975	406.9	407.1	-0.147	0.883	-0.003
TUR	4284	3975	46.83	40.24	9.075	<0.001	0.200

Positive Hedges' g indicates a higher mean on the western functional zone; negative values indicate a higher mean on the eastern shore.

Fig. S1. Integrated feature-importance profile for the global XGBoost-W2 model. Predictors are ordered by the integrated score; bars show normalized model gain, permutation importance, and mean absolute SHAP importance, and diamonds denote the integrated score.

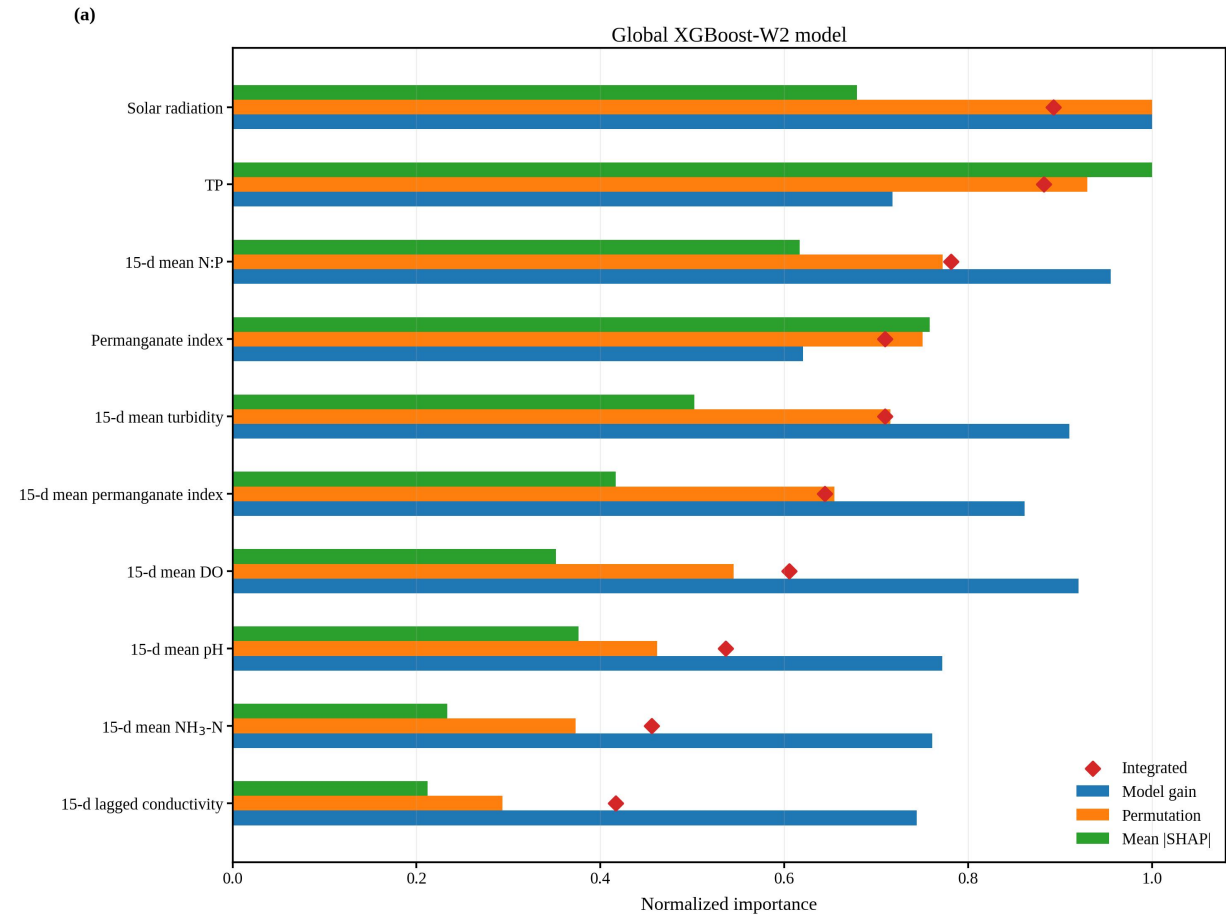


Fig. S2. SHAP dependence plots for the five highest-ranked predictors in each eastern functional-zone stage model. Panels (a)–(e), (f)–(j), and (k)–(o) represent stable, outbreak, and decline stages, respectively. Solid lines are continuous piecewise fits, dashed vertical lines mark model-derived transitions, shaded bands show bootstrap 95% confidence intervals, and horizontal lines indicate zero SHAP contribution.

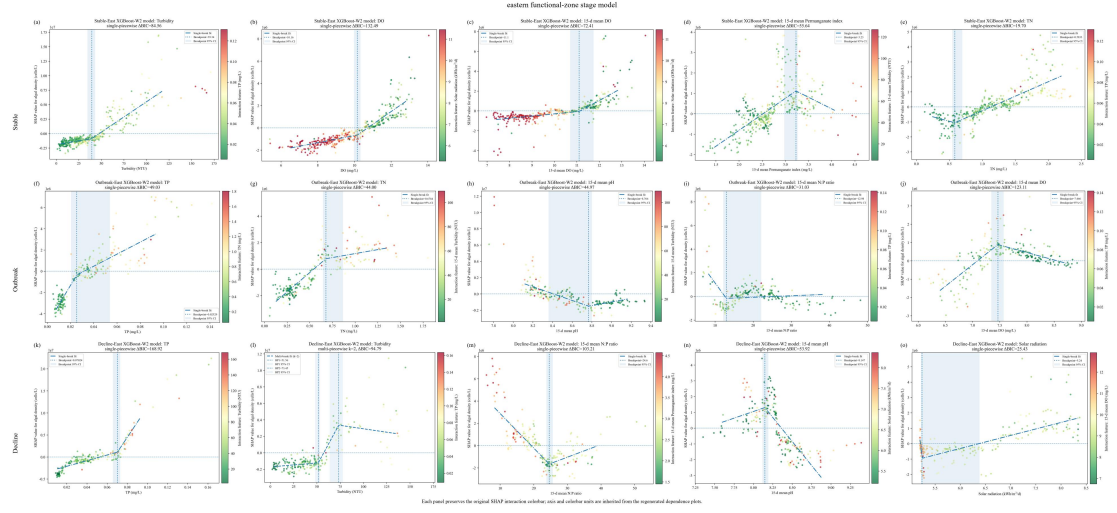


Fig. S3. SHAP dependence plots for the five highest-ranked predictors in each western functional-zone stage model. Panels (a)–(e), (f)–(j), and (k)–(o) represent stable, outbreak, and decline stages, respectively. The outbreak-stage precipitation ranges are exploratory, and decline-stage turbidity uses the BIC-selected two-transition model. Other plot elements follow Fig. S2.

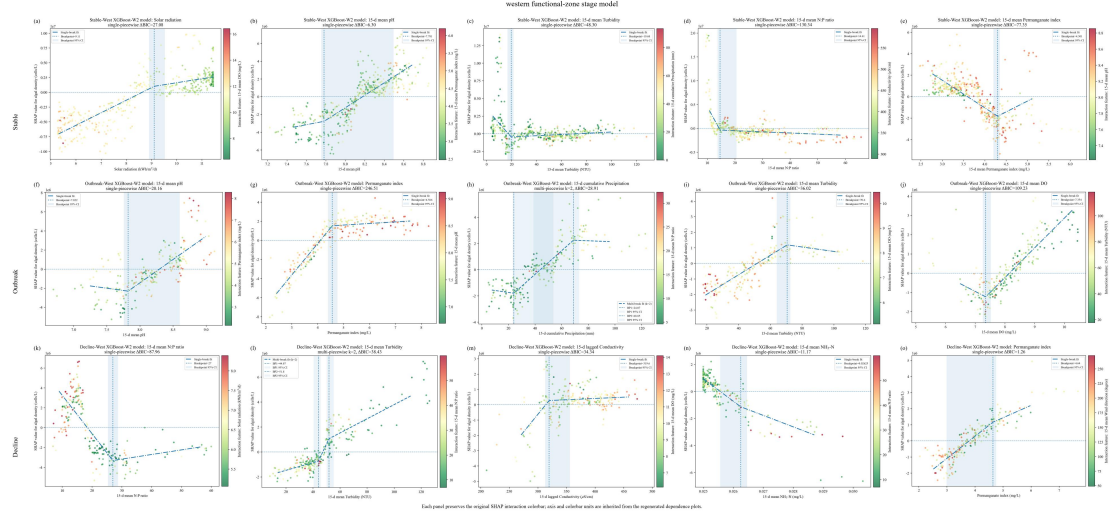


Fig. S4. SHAP dependence plots for the five highest-ranked predictors in the global XGBoost-W2 model. Plot elements follow Fig. S2. The fitted transitions describe changes in model response within the sampled domain and are not universal ecological or regulatory thresholds.

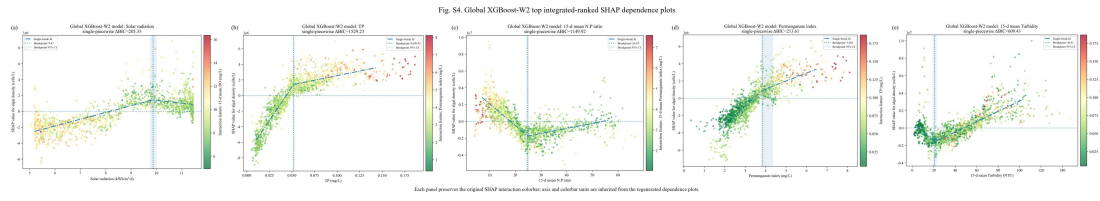


Fig. S5. PCA representation of the six monitoring stations after clustering with the final water-quality dataset. Marker annotations show the assigned zone and bootstrap membership confidence.

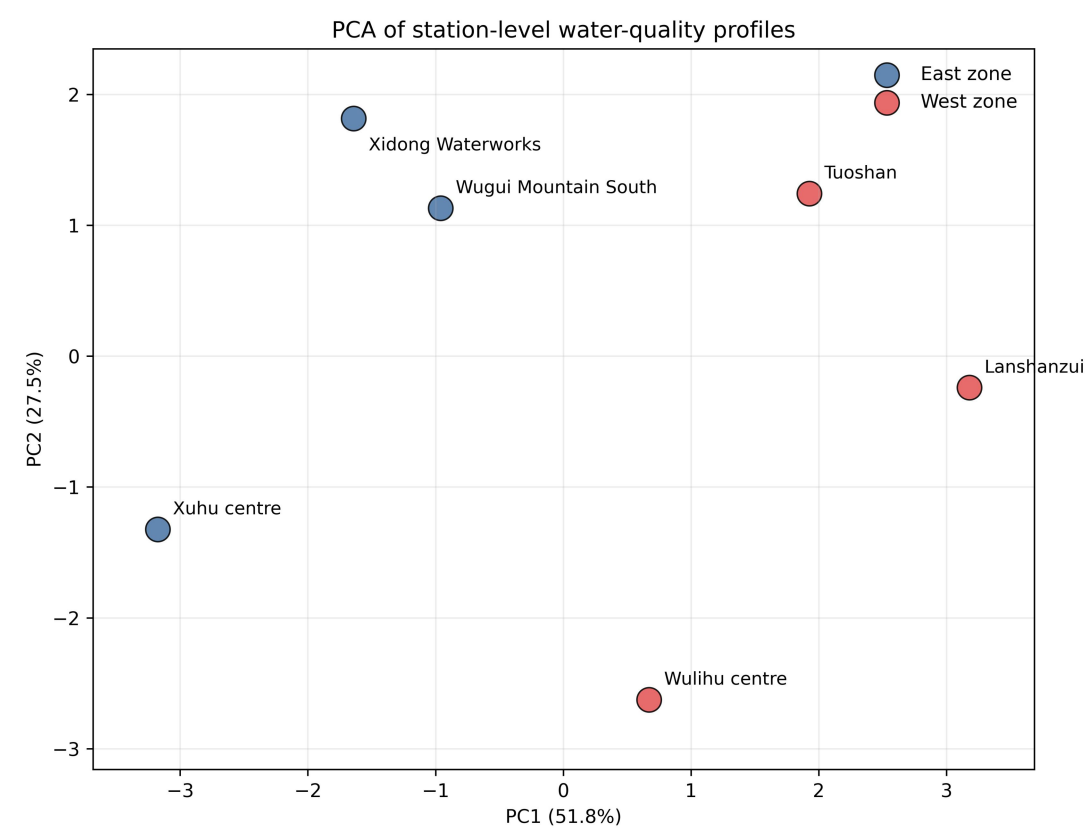


Fig. S7. Independent-test residuals versus predicted algal density for the global and six zone–stage XGBoost-W2 models. The dashed horizontal line marks zero residual. Increasing residual spread at higher predicted density indicates greater uncertainty for extreme bloom observations.

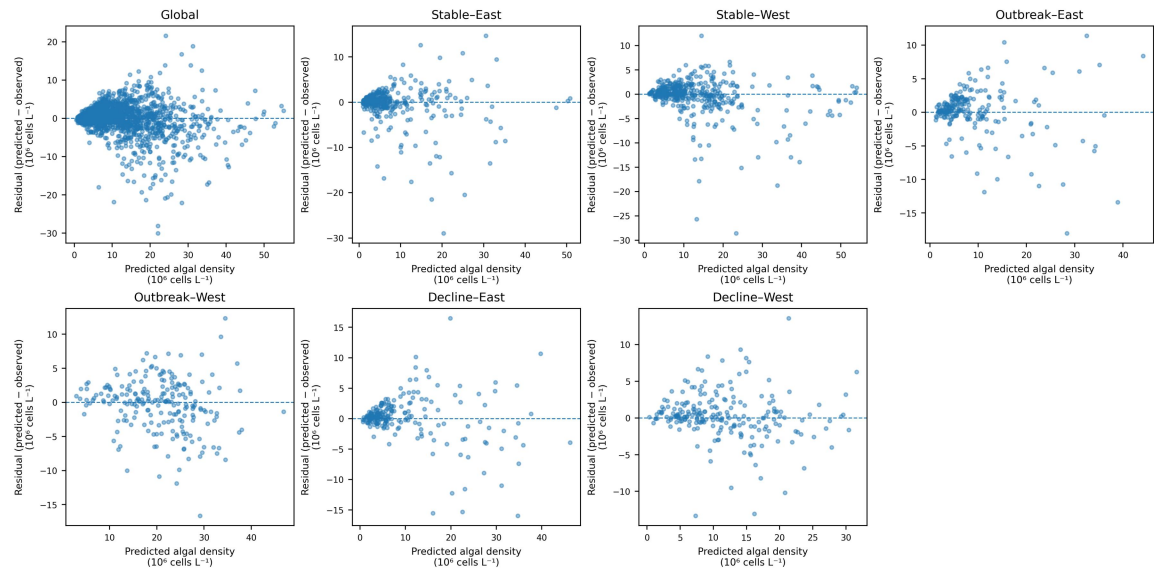
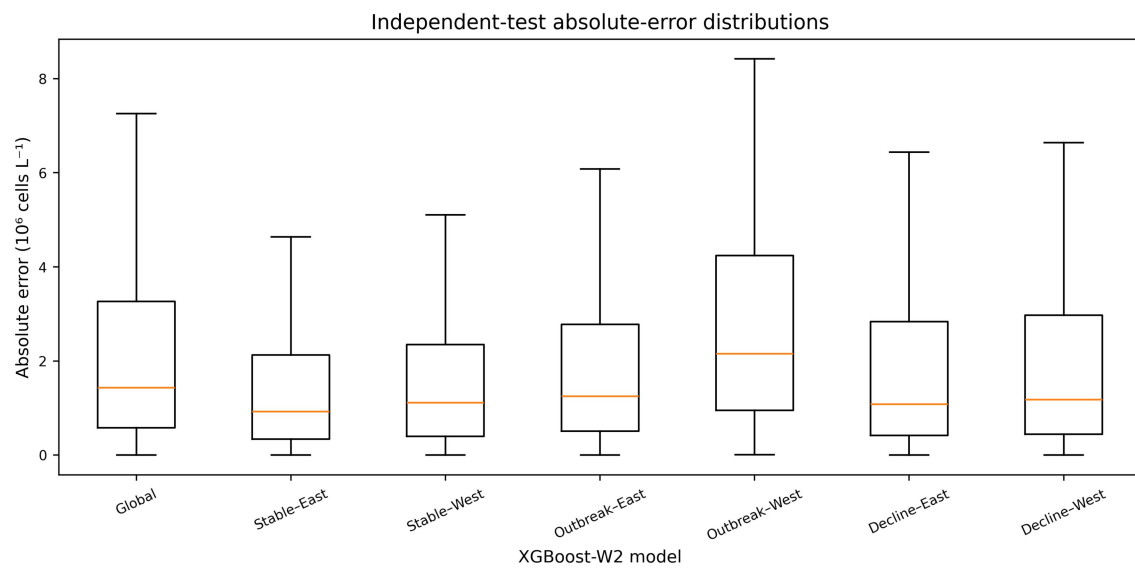


Fig. S8. Independent-test absolute-error distributions for the global and six zone–stage XGBoost-W2 models. Boxes show the interquartile range and central lines show the median; extreme outliers are omitted from the display to preserve readability.



References

- Arlot, S. and Celisse, A. 2010 A survey of cross-validation procedures for model selection. *4 %J Statistics Surveys*(none), 40-79, 40
- Bergstra, J. and Bengio, Y. 2012 Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13, 281-305
- Biau, G. and Scornet, E. 2016 A random forest guided tour. *Test* 25(2), 197-227.<https://doi.org/10.1007/s11749-016-0481-7>
- Chen, S., Huang, J., Huang, J., Wang, P., Sun, C., Zhang, Z. and Jiang, S. 2025 Explainable deep learning identifies patterns and drivers of freshwater harmful algal blooms. *Environmental Science and Ecotechnology* 23.<https://doi.org/10.1016/j.esec.2024.100522>
- Chen, T. and Guestrin, C. 2016 XGBoost: A Scalable Tree Boosting System. *Arxiv*.<https://doi.org/arXiv:1603.02754>
- Contreras-Reyes, J.E. 2022 Mutual information matrix based on Renyi entropy and application. *Nonlinear Dynamics* 110(1), 623-633.<https://doi.org/10.1007/s11071-022-07665-3>
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Garcia Marquez, J.R., Gruber, B., Lafourcade, B., Leitao, P.J., Muenkemueller, T., McClean, C., Osborne, P.E., Reineking, B., Schroeder, B., Skidmore, A.K., Zurell, D. and Lautenbach, S. 2013 Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36(1), 27-46.<https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Fisher, A., Rudin, C. and Dominici, F. 2019 All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* 20(177), 1-81
- Gao, L., Shangguan, Y., Sun, Z., Shen, Q. and Zhou, L. 2025 A novel algal bloom risk assessment framework by integrating environmental factors based on explainable machine learning. *Ecological Informatics* 87.<https://doi.org/10.1016/j.ecoinf.2025.103098>
- Hajihosseini, M., Maghsoudi, A. and Ghezelbash, R. 2023 A Novel Scheme for Mapping of MVT-Type Pb-Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm. *Natural Resources Research* 32(6), 2417-2438.<https://doi.org/10.1007/s11053-023-10249-6>
- Jiang, P., Son, K., Mudunuru, M.K. and Chen, X. 2022 Using Mutual Information for Global Sensitivity Analysis on Watershed Modeling. *Water Resources Research* 58(10).<https://doi.org/10.1029/2022wr032932>
- Kohavi, R. 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection, pp. 1137–1143, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada.
- Kraskov, A., Stogbauer, H. and Grassberger, P. 2004 Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics* 69(6 Pt 2), 066138-066138.<https://doi.org/10.1103/PhysRevE.69.066138>
- Lee, D., Kim, M., Lee, B., Chae, S., Kwon, S. and Kang, S. 2022 Integrated explainable deep learning prediction of harmful algal blooms. *Technological Forecasting and Social*

Change 185, 122046.<https://doi.org/10.1016/j.techfore.2022.122046>

Loupe, G., Wehenkel, L., Sutera, A. and Geurts, P. 2013 Understanding variable importances in forests of randomized trees, pp. 431–439, Curran Associates Inc., Lake Tahoe, Nevada.

Lundberg, S. and Lee, S.-I. 2017 A Unified Approach to Interpreting Model Predictions. Arxiv.<https://doi.org/10.1016/j.techfore.2022.122046>

Ma, Y., Wang, J., Huo, S., Wang, D., Wang, Y., Li, J., Chen, J. and Feng, L. 2025 Explainable machine learning reveals climate warming increases risk of algal blooms in lakes and reservoirs. *Water Research* 287.<https://doi.org/10.1016/j.watres.2025.124460>

Nguyen, H.Q., Arias, M.E., Zhang, Q., Tarabih, O.M., Armstrong, C., Sun, D. and Philips, E.J. 2025 A comparative analysis of data-driven modeling approaches to forecast cyanobacteria algal blooms in eutrophic lake discharge canals. *Journal of Environmental Management* 380.<https://doi.org/10.1016/j.jenvman.2025.124834>

Song, Y., Shen, C. and Hong, Y. 2025 Comparing the performance of 10 machine learning models in predicting Chlorophyll a in western Lake Erie. *Journal of Environmental Management* 380.<https://doi.org/10.1016/j.jenvman.2025.125007>

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. 2007 Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics* 8, 25-25.<https://doi.org/10.1186/1471-2105-8-25>

Sun, G., Zhu, W., Qian, X., Wei, C., Xie, P., Shi, Y., Cao, X. and He, Y. 2025 Machine Learning Models for Chlorophyll-a Forecasting in a Freshwater Lake: Case Study of Lake Taihu. *Water* 17(8).<https://doi.org/10.3390/w17081219>

Tao, Y., Ren, J., Zhu, H., Li, J. and Cui, H. 2024 Exploring spatiotemporal patterns of algal cell density in lake Dianchi with explainable machine learning. *Environmental Pollution* 356.<https://doi.org/10.1016/j.envpol.2024.124395>

Tian, J., Wang, G., Xiang, D., Huang, S. and Li, W. 2024 Causality Analysis and Prediction of Riverine Algal Blooms by Combining Empirical Dynamic Modeling and Machine Learning Techniques. *Water Resources Research* 60(5).<https://doi.org/10.1029/2023wr036334>

Vatcheva, K.P., Lee, M., McCormick, J.B. and Rahbar, M.H. 2016 Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale)* 6(2).<https://doi.org/10.4172/2161-1165.1000227>

Zhang, M.-L. and Zhou, Z.-H. 2007 ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7), 2038-2048.<https://doi.org/10.1016/j.patcog.2006.12.019>