Supplementary Materials for

Geographically Interpretable Data-Driven Modeling for Hourly-Scale Monitoring of Chlorophyll-a Dynamics in Large Shallow Lake Using Satellite Remote Sensing

Jianqi Zhao [1], Yuan Xu [2,3], Zhengxian Zhang [1,2,4,5,*], Yi Liu [1], Chenhui Wu [6,7], Wei Tang [3], Wenhao Ding [3] and Haoyang Qiang [1]

1 Co-Innovation Center of Sustainable Forestry in Southern China, Institute of Ecological Civilization Construction and Forestry Development, Nanjing Forestry University, Nanjing 210037, China; jianqi_zhao@njfu.edu.cn (J.Z.); yiliu@njfu.edu.cn (Y.L.); qianghaoyang2@njfu.edu.cn (H.Q.)

2 Key Laboratory of Taihu Basin Water Resources Management, Ministry of Water Resources, Nanjing 210029, China; xuyuan@scimee.com

3 National Key Laboratory of Water Disaster Prevention, Nanjing Hydraulic Research Institute, Nanjing 210029, China; sltjsb@slt.gxzf.gov.cn (W.T.); whding@niglas.ac.cn (W.D.)

4 Key Laboratory of Integrated Regulation and Resources Development of Shallow Lakes of Ministry of Education, Hohai University, Nanjing 210098, China

5 Key Laboratory for Water Ecology Management and Protection in River Source Areas, Ministry of Water Resources, Qinghai University, Xining 810016, China

6 Key Laboratory of Flood&Drought Disaster Defense, The Ministry of Water Resources, Nanjing 210098, China; wuchh@zuwe.edu.cn

7 School of Hydraulic Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China

* Correspondence: zhangzx@whu.edu.cn

1. Hybrid modeling framework

Conventional black-box prediction models exhibit several limitations in chlorophyll-a concentration retrieval, notably limited accuracy and poor generalizability. To address these shortcomings, this study developed a novel WRF-WOA-GeoSHAP hybrid model integrating wavelet denoising (W), random forest (RF), the whale

optimization algorithm (WOA), and GeoSHAP, with Lake Taihu serving as the study area. Spectral indices derived from GOCI-II remote-sensing reflectance products were paired with in situ chlorophyll-a measurements to develop the retrieval model. The six selected spectral indices collectively capture distinct optical characteristics of the water column, and GeoSHAP was used to characterize the spatial heterogeneity of their contributions to chlorophyll-a retrieval. The proposed WRF-WOA-GeoSHAP model can provide technical support for hourly-scale monitoring of chlorophyll-a dynamics and eutrophication in large shallow lakes. The detailed workflow is described as follows.

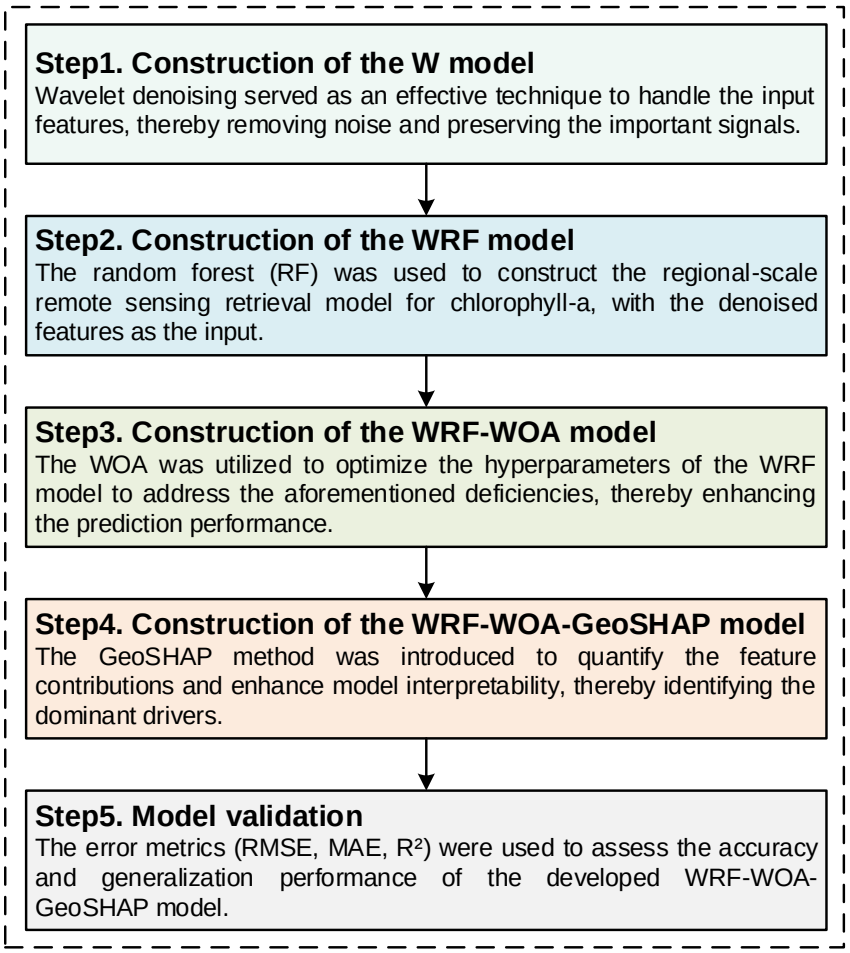

Fig. S1. Modelling steps for the WRF-WOA-GeoSHAP hybrid model.

(i) Step1. Construction of the W model

In chlorophyll-a retrieval, wavelet denoising was used to reduce noise in the input features [1–3]. In this study, the six spectral indices were processed independently, whereas the measured chlorophyll-a concentrations were retained unchanged as the response variables. To avoid mixing station-specific temporal patterns, data from the six

monitoring stations were processed separately. Within each station, observations were first arranged in ascending chronological order, and the time intervals between adjacent observations were calculated to identify temporal discontinuities. The original sequences were then divided into approximately continuous segments based on the observed sampling intervals. Each eligible segment was denoised independently, and the reconstructed segments were subsequently recombined. To ensure the applicability of the conventional discrete wavelet transform, only segments with sufficiently regular sampling intervals and adequate sample lengths for three-level decomposition were retained for subsequent model construction. Furthermore, the training and validation subsets were denoised independently to prevent information leakage between model training and validation.

A three-level discrete wavelet transform (DWT) was applied to each eligible spectral-index sequence using the Daubechies 3 (db3) wavelet [4]. To address boundary effects caused by the finite length of the input sequences, symmetric extension was adopted at both ends of each segment [5]. The wavelet decomposition was expressed as follows.

$$DWT_{\psi,sym} f[(i)] = \left[D_1, D_2, \ldots, D_j, A \right] \tag{1}$$

where $f(i)$ is the original feature sequence, DWT is the discrete wavelet transform, D_j denotes the detail coefficients at decomposition level j, A denotes the approximation coefficients, ψ is the wavelet basis function. In this study, the db3 wavelet was employed with the decomposition level fixed at j=3.

Following wavelet decomposition, thresholding was applied to the detail coefficients to suppress noise-related fluctuations while retaining the dominant signal characteristics. The universal threshold method (VisuShrink) was adopted, in which the threshold was adaptively determined according to the estimated noise level and the number of observations in each input sequence [6]. Specifically, the noise standard deviation was estimated using the median absolute deviation of the finest-scale detail coefficients.

$$\sigma = \frac{median(|D_1(k) - median(D_1)|)}{0.6745} \tag{2}$$

where σ represents the estimated noise standard deviation, and $D_1(k)$ denotes the detail coefficients at the finest decomposition level. The constant 0.6745 is the normalization factor used to obtain a robust estimate of the standard deviation under the assumption of

Gaussian noise.

The universal threshold was subsequently calculated as:

$$\lambda = \sigma\sqrt{2\ln N} \tag{3}$$

where λ is the adaptive threshold and N represents the number of observations in the corresponding input sequence. Accordingly, the threshold was independently estimated for each combination of spectral index, monitoring station, and continuous temporal segment, rather than applying a fixed threshold across all input sequences.

In this study, hard thresholding was applied to the detail coefficients.

$$\tilde{D}_j(k) = \begin{cases} D_j(k), & |D_j(k)| \geq \lambda \\ 0\ , & |D_j(k)| < \lambda \end{cases} \tag{4}$$

where $D_j(k)$ is the original wavelet detail coefficient, $D\tilde{j}$ denotes the coefficient after thresholding, and λ is the adaptively estimated universal threshold. The approximation coefficients A were retained unchanged to preserve the dominant information contained in the original spectral-index sequences.

After threshold processing, the inverse discrete wavelet transform (IDWT) was applied to reconstruct the denoised spectral-index sequences.

$$\tilde{f}(i) = IDWT_{\psi,sym}\left[\tilde{D}_1, \tilde{D}_2, \ldots, \tilde{D}_j, A\right] \tag{5}$$

where $f(i)$ is the feature sequence after thresholding, $IDWT$ is the inverse discrete wavelet transform. To determine whether a three-level decomposition was feasible, the maximum useful decomposition level was calculated based on the input sequence length and the filter length of the selected wavelet [5].

$$J_{\max} = \left\lfloor \log_2\left(\frac{N}{L_\psi - 1}\right) \right\rfloor \tag{6}$$

where J_{max} represents the maximum useful decomposition level, N is the number of observations in each segment, and L_ψ is the filter length of the selected wavelet. For the db3 wavelet, L_ψ=6. Therefore, only segments satisfying $J_{max}\geq 3$ were retained for subsequent analysis. The denoised sequences reconstructed from the eligible segments were then combined to form the input feature dataset for the WRF model, while the corresponding in situ chlorophyll-a measurements were retained as the original response

variables.

(ii) Step2. Construction of the WRF model

The random forest (RF) is an effective machine-learning technique for regression tasks [7]. It has been widely applied in land surface monitoring and remote sensing parameter retrieval [8]. Accordingly, RF was used in this study to construct the regional-scale chlorophyll-a retrieval model for Lake Taihu, using the denoised spectral indices obtained from the W model as inputs. The corresponding formulation of the WRF model is given as follows.

The feature set obtained from the W model was used as the input to the WRF model.

$$\tilde{f}(I) = [\tilde{f}(1), \tilde{f}(2), \ldots, \tilde{f}(i)] \tag{7}$$

where $f(I)$ and $f(i)$ represent the denoised feature set and the reconstructed feature sequence after wavelet denoising, respectively, and i is the number of input features.

The resulting feature set was used to train the RF model for chlorophyll-a retrieval.

$$\tilde{y}(I_m) = \int_T \tau(\tilde{f}(I_m)) dP(\tau / \theta) \tag{8}$$

where m denotes a pixel in the Lake Taihu, $\tilde{y}(I_m)$ denotes the predicted chlorophyll-a concentration at pixel m, T denotes the number of decision trees, $\tau(f(I_m))$ denotes the prediction made by an individual decision tree, θ represents the hyperparameters of the WRF model, and $P(\tau/\theta)$ denotes the probability distribution of the decision-tree predictions given θ.

(iii) Step3. Construction of the WRF-WOA model

The whale optimization algorithm (WOA) is a heuristic optimization algorithm inspired by the bubble-net foraging behavior of humpback whales [9]. The prediction performance of the RF model depends strongly on its hyperparameter settings [10]. To address this issue, WOA was employed to optimize the hyperparameters of the RF model within the WRF framework [11]. The search ranges of the RF hyperparameters are listed in Table 3. The mean RMSE across the five folds of cross-validation performed on the training set was adopted as the objective function. The WOA was run with a population size of 30 whales and a maximum of 100 iterations. The detailed optimization procedure is provided in the Supplementary Materials. The corresponding formulation of the WRF-WOA model is given as follows.

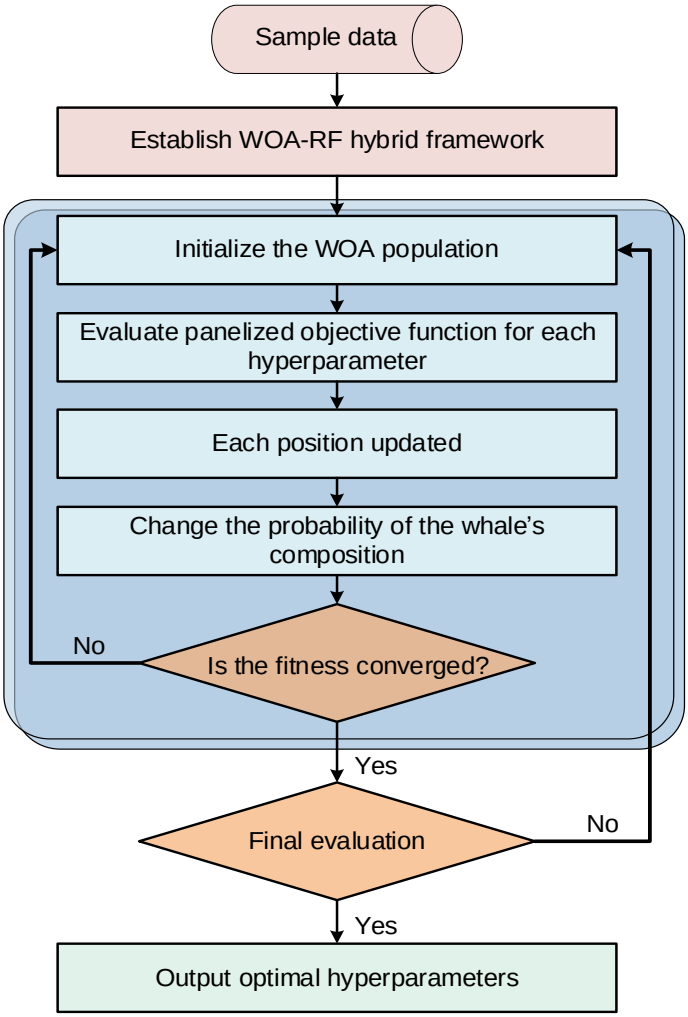

Fig. S2. The calculation process of the optimizing WRF hyperparameters based on the whale optimization algorithm (WRF-WOA) [9,12].

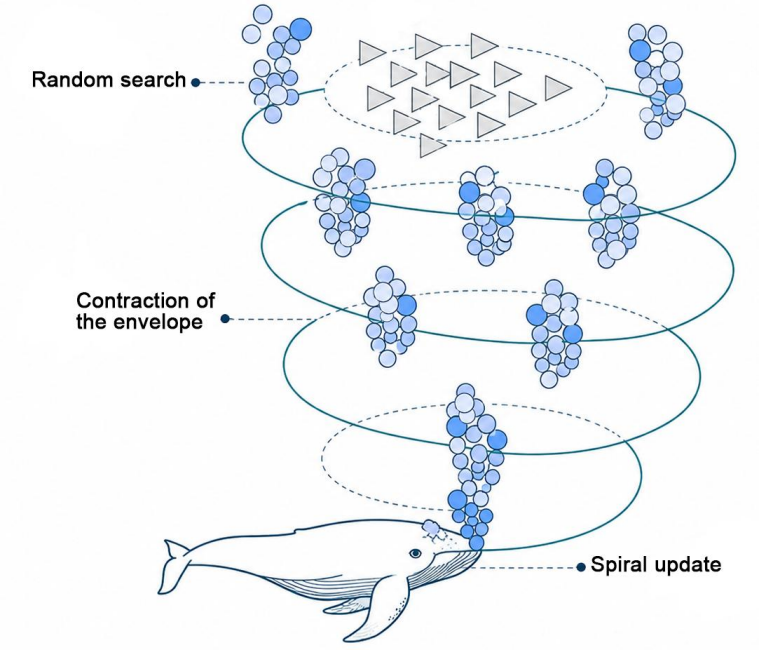

Fig. S3. The illustration of the whale optimization algorithm (WOA).

(i) The WOA searched for the optimal solution through iterative position updates, and the position update formulas were given as follows.

$$\begin{cases} D = \left| C \cdot X^*(t) - X(t) \right| \\ X(t+1) = X^*(t) - A \cdot D \\ A = 2a \cdot r - a \\ C = 2 \cdot r \end{cases} \tag{9}$$

where t is the current iteration number, D is the spatial distance between the current individual whale and the optimal solution, $X^*(t)$ is the current optimal position vector, $X(t)$ is the individual position vector of a whale, $X(t+1)$ is the optimal position vector after interation, a is a vector that gradually shrinks the whale into the encircling range, the value range of a is [0,2] and the r is a random vector within the interval [0,1], A and C are coefficient vectors.

(ii) The WOA simulated the hunting behavior of whales to search for the optimal solution. The shrinking encircling mechanism and the spiral position update formula were given as follows.

$$X(t+1) = \begin{cases} X^*(t) - A \cdot D, p < 0.5 \\ D' \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t), p \geq 0.5 \end{cases} \tag{10}$$

$$D' = \left| X^*(t) - X(t) \right| \tag{11}$$

where b is a constant that defines the shape of the logarithmic spiral, l is a random number within the interval [-1, 1], p is a random probability parameter within the interval [0, 1], D' is the spatial distance between the current individual whale and the optimal solution, the shrinking encircling is achieved by reducing the value of a, and the range of A is [$-a$, a].

(iii) The WOA conducted a global search to find the optimal solution. The corresponding equation was as follows.

$$\begin{cases} D = \left| C \cdot X_{rand} - X \right| \\ X(t+1) = X_{rand} - A \cdot D \end{cases} \tag{12}$$

where X_{rand} is a random position vector selected from the current population.

Through the above optimization procedure, the optimal set of hyperparameters minimizing the objective function was obtained. Accordingly, the WOA was used in this

study to optimize the hyperparameters of the WRF model. The WRF-WOA model corresponding formulas was expressed as follows.

$$L_\theta = \frac{1}{5}\sum_{k=1}^{5} RMSE_k(\theta) \tag{13}$$

$$\begin{cases} \theta^* = \arg\min_{\theta\in\Omega} L(\theta) \\ \tilde{y}^*(I_m) = \int_T \tau(\tilde{f}(I_m))dP(\tau/\theta^*) \end{cases} \tag{14}$$

where θ^* denotes the optimized hyperparameter set of the WRF model, Ω is the search space of the hyperparameters, $L(\theta)$ is the objective function (mean RMSE) evaluated for a given hyperparameter set θ, and $\tilde{y}^*(I_m)$ denotes the chlorophyll-a concentration predicted at pixel m using the optimized model.

Table S1. Search ranges of the RF hyperparameters considered in WOA optimization.

Parameter	Description	Search range
$n_{estimators}$	Number of trees in the forest	100-2000
max_{depth}	Maximum depth of the tree	5-50
$min_{samples,split}$	Minimum samples required to split an internal node	2-20
$min_{samples,leaf}$	Minimum samples required to be at a leaf node	1-10
$max_{features}$	Feature-selection strategy at each split	sqrt, log2, None

Note: The $max_{features}$ parameter was treated as a categorical hyperparameter with three candidate feature-subsampling strategies: the square-root rule (sqrt), the logarithmic rule (log2), and the all-feature strategy (None).

(iv) Step4. Construction of the WRF-WOA-GeoSHAP model

Machine-learning models are often difficult to interpret because their internal decision-making processes are not readily accessible, which may limit confidence in their predictions and their application in aquatic remote sensing [13]. To examine the contributions of input features to chlorophyll-a predictions and improve model interpretability, this study developed an explainable machine-learning framework integrating WRF-WOA with GeoSHAP. Based on cooperative game theory, SHAP decomposes model predictions into baseline values and feature-specific contributions, thereby quantifying the impact of each input feature on the predicted outcome [14]. In

this study, the absolute SHAP values were normalized, linked to the geographic coordinates of the prediction samples, and averaged across samples at the same location to characterize the spatial heterogeneity of feature contributions. The corresponding calculation formulas are expressed as follows.

The model prediction was decomposed into a baseline value and feature-specific contributions, as follows.

$$\tilde{y}^*(I_m) = \phi_0 + \sum_{i=1}^{I} \phi[\tilde{y}_i^*(I_m)] \tag{15}$$

where ϕ_0 denotes the baseline value, and $\phi[\tilde{y}_i^*(I_m)]$ is the contribution of the feature i to the predicted chlorophyll-a concentration at pixel m.

The SHAP value of feature i was calculated as follows.

$$\phi[\tilde{y}_i^*(I_m)] = \sum_{S \subseteq \tilde{f}(I) \setminus \{\tilde{f}(i)\}} \frac{|S|!(|I|-|S|-1)!}{|I|!} \left[u[\tilde{f}(S \cup \{i\})] - u[\tilde{f}(S)] \right] \tag{16}$$

where I denotes the set of all input features, S is a subset of features excluding feature i, u(f(S)) is the model prediction based on the feature subset S, whereas u[f(S∪{i})] represents the model prediction after feature i is included. The difference between these two predictions represents the marginal contribution of feature i.

Because SHAP values can be positive or negative, their absolute values were used to quantify the magnitude of each feature's contribution. These values were then normalized to represent the relative contribution of each feature to chlorophyll-a retrieval, enabling comparison of feature contributions across pixels. The normalized contribution weight of feature i at pixel m was calculated as follows.

$$C_{(m,i)} = \frac{\left|\phi[\tilde{y}_i^*(I_m)]\right|}{\sum_{i=1}^{I} \left|\phi[\tilde{y}_i^*(I_m)]\right|} \tag{17}$$

where $C(m,i)$ denotes the contribution weight of feature i at pixel point m in the Lake Taihu.

To further characterize the spatial heterogeneity of feature contributions, the normalized contribution weights of prediction samples associated with the same pixel location were averaged. The mean contribution weight of feature i at pixel m was calculated as follows.

$$\bar{C}_{(m,i)} = \frac{1}{N_m}\sum_{k=1}^{N_m} C_{(m,i,k)} \quad (18)$$

where $C_{(m,i,k)}$ represents the normalized contribution weight of feature i at pixel m for the k-th prediction sample, and Nm denotes the total number of valid prediction samples associated with pixel m. $\bar{C}_{(m,i)}$ represents the mean contribution weight of feature i at the corresponding geographic location.

Mapping these mean contribution weights across Lake Taihu revealed the spatial distribution of spectral-index contributions, allowing the dominant spectral indices and their spatial variations among lake regions to be identified.

2. Model validation

Chlorophyll-a concentration is an important indicator for evaluating the trophic status of lakes and reservoirs. To quantitatively characterize the spatiotemporal variations in eutrophication across Lake Taihu, this study used chlorophyll-a concentration as the basis for trophic-status classification. According to the concentration thresholds of 1.6, 10, and 26 μg/L, three primary trophic categories were distinguished, including oligotrophic, mesotrophic, and eutrophic conditions [15]. The eutrophic category was further divided into light eutrophic and severe eutrophic levels (Table S2) [16]. These concentration thresholds were applied to the retrieved chlorophyll-a fields to quantify the areal proportions and hourly variations of different trophic levels, providing a consistent basis for subsequent analysis of eutrophication dynamics in Lake Taihu. This classification represents a chlorophyll-a-based assessment rather than a comprehensive evaluation involving multiple water-quality indicators.

Table S2. Chlorophyll-a-based trophic status classification in Lake Taihu.

Trophic status		Chlorophyll-a concentration (μg/L)
Oligotrophic (L1)		≤ 1.6
Mesotrophic (L2)		(1.6,10]
Eutrophic	Light eutrophic (L3)	(10,26]
	Severe eutrophic (L4)	> 26

Note: The upper bounds of the mesotrophic and light eutrophic intervals are included, whereas their lower bounds are excluded.

(ii) Performance evaluation of modeling

The accuracy of model simulation results was evaluated via three error indicators, including the root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2). Their calculation formulas were provided below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{C}_i - C_i\right)^2} \tag{19}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{C}_i - C_i\right| \tag{20}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{C}_i - C_i\right)^2}{\sum_{i=1}^{n}\left(C_i - \bar{C}\right)^2} \tag{21}$$

where $\hat{C}_i$ is the predicted value for the *i-th* sample, C_i is the corresponding in-situ measured value, $\bar{C}$ is the average value of all measured values, n is the total amount of samples used for evaluation. *RMSE*, *MAE*, and R^2 are used to assess the accuracy and generalization performance of the developed WRF-WOA-GeoSHAP model.

In this study, all calculations were performed using the Python version 3.14

Table. S3. The parameter settings of WRF-WOA-GeoSHAP hybrid model.

Parameter	Description	Value
wavelet	Wavelet basis function type	Daubechies 3
level	Decomposition level	3
mode	Thresholding mode	Hard
$n_{estimators}$	Number of trees in the forest	1432
max_{depth}	Maximum depth of the tree	31
$min_{samples,split}$	Minimum samples required to split an internal node	2
$min_{samples,leaf}$	Minimum samples required to be at a leaf node	1
$max_{features}$	Feature-selection strategy at each split	sqrt
n_{whales}	Population size (number of whales)	30
$n_{iterations}$	Maximum number of iterations	100
a_{decay}	Convergence factor decay coefficient	2
b	Spiral constant defining logarithmic spiral	1

Table. S4. Comparison of retrieval accuracy among the evaluated models.

Model	R^2	RMSE	MAE
MLR	0.04	7.84	5.00
GM(1,N)	-0.40	9.45	6.01
MLP	-6.42	21.77	10.08
CNN	-0.31	9.16	4.94
LSTM	-0.16	8.60	4.83
GBDT	0.35	6.43	3.71
LightGBM	0.45	5.95	3.89
XGBoost	0.42	6.06	3.88
RF	0.54	5.41	3.43
RF-WOA	0.64	4.79	2.72
WRF	0.65	4.73	3.12
WRF-WOA	0.77	3.79	2.14

Note: MLR is the multiple linear regression; GM(1,N) is the multivariable grey model; MLP is the multilayer perceptron; CNN is the convolutional neural network; LSTM is the long short-term memory; GBDT is the gradient boosting decision tree; LightGBM is the light gradient boosting machine; XGBoost is the extreme gradient boosting; RF is the random forest; WOA is the whale optimization algorithm; WRF is the wavelet-denoised random forest.

References

1. Taswell, C. The What, How, and Why of Wavelet Shrinkage Denoising. *Comput. Sci. Eng.* **2000**, *2*, 12–19, doi:10.1109/5992.841791.
2. Md Yusof, M.R.; Ariffin, A.K.; Ismail, M. Wavelet Denoising as Signal Feature-Dependent Kernel Convolution. *J. Phys. Conf. Ser.* **2019**, *1185*, 012026, doi:10.1088/1742-6596/1185/1/012026.
3. Halidou, A.; Mohamadou, Y.; Ari, A.A.A.; Zacko, E.J.G. Review of Wavelet Denoising Algorithms. *Multimed. Tools Appl.* **2023**, *82*, 41539–41569, doi:10.1007/s11042-023-15127-0.
4. Mallat, S.G. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693, doi:10.1109/34.192463.
5. Lee, G.; Gommers, R.; Waselewski, F.; Wohlfahrt, K.; O'Leary, A. PyWavelets: A Python Package for Wavelet Analysis. *J. Open Source Softw.* **2019**, *4*, 1237, doi:10.21105/joss.01237.
6. Donoho, D.L.; Johnstone, I.M. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika* **1994**, *81*, 425–455, doi:10.2307/2337118.
7. Liaw, A.; Wiener, M. Classification and Regression by randomForest.; 2007.
8. Belgiu, M.; Drăguţ, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31, doi:10.1016/j.isprsjprs.2016.01.011.

9. Mirjalili, S.; Lewis, A. The Whale Optimization Algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67, doi:10.1016/j.advengsoft.2016.01.008.
10. Bilolikar, D.K.; More, A.; Gong, A.; Janssen, J. How to Out-Perform Default Random Forest Regression: Choosing Hyperparameters for Applications in Large-Sample Hydrology. *Hydrol. Res.* **2025**, *57*, 61–77, doi:10.2166/nh.2025.075.
11. Fang, K.; Wu, F.; Gao, X.; Li, Z. Multi-Source Data Fusion and Heuristic-Optimized Machine Learning for Large-Scale River Water Quality Parameters Monitoring. *Remote Sens.* **2026**, *18*, 320, doi:10.3390/rs18020320.
12. Rana, N.; Latiff, M.S.A.; Abdulhamid, S.M.; Chiroma, H. Whale Optimization Algorithm: A Systematic Review of Contemporary Applications, Modifications and Developments. *Neural Comput. Appl.* **2020**, *32*, 16245–16277, doi:10.1007/s00521-020-04849-z.
13. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115, doi:10.1016/j.inffus.2019.12.012.
14. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67, doi:10.1038/s42256-019-0138-9.
15. Zheng, B.-H.; Xu, Q.-J.; Zhu, Y.-Z. [Primary Study on Enacting the Lake Nutrient Control Standard]. *Huan Jing Ke Xue Huanjing Kexue* **2009**, *30*, 2497–2501.
16. Xu, H.; Paerl, H.W.; Qin, B.; Zhu, G.; Gaoa, G. Nitrogen and Phosphorus Inputs Control Phytoplankton Growth in Eutrophic Lake Taihu, China. *Limnol. Oceanogr.* **2010**, *55*, 420–432, doi:10.4319/lo.2010.55.1.0420.