remote sensing

MDPI

Article

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

Highlights

What are the main findings?

- Hourly-scale eutrophication dynamics and driving factors are investigated in Lake Taihu, a large shallow lake.
- WRF-WOA-GeoSHAP model is proposed via coupling satellite remote sensing and machine learning.

What are the implications of the main findings?

- Explain the nonlinear relationships between spectral indices and chlorophyll-a concentrations in the Lake Taihu.
- Hourly-scale monitoring methods are developed to support eutrophication early warning in large shallow lake.

Abstract

Chlorophyll-a concentration is a core indicator for assessing lake eutrophication. Lake Taihu, situated in one of China's most urbanized regions, is highly susceptible to anthropogenic nutrient loading, highlighting the need for high-frequency monitoring and spatially resolved characterization of chlorophyll-a dynamics. Nevertheless, existing approaches have yet to simultaneously achieve hourly temporal resolution, lake-wide spatial coverage, and model interpretability in chlorophyll-a retrieval. To address these challenges, a novel WRF-WOA-GeoSHAP hybrid framework integrating wavelet denoising (W), random forest (RF), the whale optimization algorithm (WOA), and Geographic SHapley Additive exPlanations (GeoSHAP) was developed, with Lake Taihu serving as the study area. Spectral indices derived from Geostationary Ocean Color Imager-II (GOCI-II) remote-sensing reflectance products were paired with in situ chlorophyll-a measurements for quantitative retrieval. The GeoSHAP method was then utilized to characterize the spatial variations

Check for updates

Academic Editor: Teodosio Lacava

Received: 20 September 2026
Revised: 19 November 2026
Accepted: 23 November 2026
Published: 27 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

in spectral-index contributions to model predictions. Results showed that the WRF-WOA model achieved an R^2 of 0.775, an RMSE of 3.79 μg/L, and an MAE of 2.14 μg/L on the validation set, demonstrating its capability for hourly-scale chlorophyll-a retrieval across Lake Taihu and characterization of hourly variations in chlorophyll-a concentrations, while capturing nonlinear relationships between spectral indices and chlorophyll-a concentrations. GeoSHAP revealed spatially varying contributions of the six spectral indices, with FLHI exhibiting the highest overall contribution. Cross-sensor comparison with Sentinel-2 further demonstrated consistency in the major regional spatial gradients and chlorophyll-a accumulation patterns between the two sensors at the 250 m scale. These findings demonstrate the potential of the WRF-WOA-GeoSHAP framework for hourly-scale chlorophyll-a monitoring in large shallow lakes and provide technical support for eutrophication assessment and future early-warning research.

Keywords: lake eutrophication; hourly-scale monitoring; chlorophyll-a; hourly dynamics; satellite remote sensing; machine learning

1. Introduction

Lakes represent indispensable freshwater resource within the earth system, assuming a critical role in maintaining regional ecological balance and fostering global sustainable progress [1,2]. Under multiple stresses of global climate change and anthropogenic activities, particularly the discharges of nitrogen-phosphorus-rich waters, lake eutrophication intensified markedly in recent decades [3]. This intensification has manifested in the increasingly frequent cyanobacterial bloom events, posing severe threats to aquatic environmental security in China [3–5]. Freshwater eutrophication, characterized by recurrent cyanobacterial blooms events, has become a bottleneck restricting sustainable socioeconomic development. It is now widely recognized as a pervasive environmental challenge of global concern [6]. Chlorophyll-a concentration serves as a key indicator for assessing the water eutrophication [7]. Accordingly, the accurate acquisition of hourly-scale and high-precision spatiotemporal distribution patterns of chlorophyll-a concentration is particularly important for investigating the lake eutrophication.

Lake Taihu is the third largest freshwater lake in China, with a surface area of 2338 km^2. It is a vital water source occupying a critical position in the Yangtze River Delta region. However, accompanying rapid economic development, substantial inputs of industrial wastewater, municipal sewage, and agricultural fertilizers have been discharged into the lake via tributary inflows [8]. These pollutants caused increasingly severe eutrophication in the Lake Taihu [9]. Although the cyanobacteria presence in Lake Taihu can be traced back to the early 1960s, frequent cyanobacterial blooms emerged only in the late 1980s. In recent years, algal proliferation is accelerated substantially, with algal blooms occurring earlier, duration extending, and spatial coverage expanding, gradually establishing an 'algae-dominated habitat pattern' in the Lake Taihu [10]. Eutrophication pollution in Lake Taihu has become a major environmental problem in China, demanding effective monitoring and remediation [11]. Accurate simulation of cyanobacterial dynamics is therefore of paramount importance for bloom mitigation, environmental risk assessment, and the establishment of early warning systems.

The distribution of chlorophyll-a concentration is influenced by multiple factors, including natural lake topography, wind fields, and flow fields, and exhibits marked spatial heterogeneity [12]. Conventional monitoring approaches for chlorophyll-a concentration primarily relied on manual field sampling and automatic station observations [13]. Although these methods offered high measurement accuracy, they required substantial time

and financial resources, and their detection frequency and spatial coverage were inherently constrained by logistical and instrumental limitations [14,15]. Therefore, the hourly-scale spatiotemporal dynamics of lake water bodies could not be captured easily by aforementioned methods [16]. Over the past few years, satellite remote-sensing technology has made rapid progress, due to its inherent advantages of broad spatial coverage and high observational frequency [17]. These attributes established it as among the most prominent approaches for lake water quality assessment [18,19].

Multi-source remote-sensing technologies were applied for chlorophyll-a concentration estimation, and many algorithms have been developed based on bio-optical models, water surface reflectance and radiance (Table 1). These algorithms could be divided into four categories: (i) Empirical algorithms, including linear and multiple linear regression models, rely on statistical relationships between chlorophyll-a concentration and spectral reflectance. Although the calculation process is straightforward, their applicability is limited by specific water types and optical conditions [20]. (ii) Semi-analytical algorithms, such as three-band models and four-band models, combine radiative transfer theory with empirical relationships [21]. These algorithms effectively mitigate the non-algal component effects through the targeted band selection [22]. (iii) Analytical algorithms, encompassing spectral mixture analysis and algebraic methods, involve radiative transfer equations and optical properties. Albeit theoretically rigorous, they are computationally demanding and require accurate atmospheric correction. (iv) Machine learning algorithms, such as neural networks, support vector machines, and random forests, capture the complex nonlinear relationships between chlorophyll-a and spectral indices via extensive training datasets [23].

Despite the considerable progress made in machine-learning-based chlorophyll-a retrieval, several challenges remain. On the one hand, the optical properties of large inland shallow lakes are complex [24], and satellite-derived reflectance signals are susceptible to sensor noise and residual atmospheric effects, which can obscure spectral information relevant to chlorophyll-a retrieval. On the other hand, although GOCI-II provides hourly observations, reliable chlorophyll-a retrieval from these high-frequency observations remains challenging in optically complex and spatially heterogeneous lake waters [25,26]. Furthermore, the performance of machine-learning models is highly sensitive to hyperparameter configurations [27,28], while their limited interpretability may constrain confidence in retrieval results and their application in aquatic remote-sensing research [29,30].

Although SHapley Additive exPlanations (SHAP) has been widely used to quantify the contributions of input features to machine-learning predictions [31], global feature-importance summaries calculated across all samples may mask geographic variations in these contributions [32]. Local SHAP values explain individual predictions, but their spatial patterns can only be assessed when they are linked to the geographic locations of the corresponding samples [33]. This is particularly important for large shallow lakes, where chlorophyll-a concentrations and water optical properties exhibit pronounced spatial heterogeneity [34]. Meanwhile, hourly satellite observations provide the temporal sampling needed to retrieve short-term changes in chlorophyll-a distribution [35]; however, concentration maps alone cannot reveal which spectral indices contribute most strongly to model predictions in different lake regions [36]. Therefore, integrating hourly-scale retrieval with spatially explicit interpretation remains necessary. In this study, the absolute SHAP values for individual prediction samples were normalized to obtain relative feature contributions, linked to their corresponding geographic locations, and averaged across samples at the same pixel location. This spatially explicit SHAP-based procedure (GeoSHAP) was used to characterize the spatial heterogeneity of spectral-index contributions across Lake Taihu.

To address the above challenges, this study developed a WRF-WOA-GeoSHAP framework integrating wavelet denoising (W), random forest (RF), the whale optimization algo-

rithm (WOA), and a spatially explicit SHAP-based interpretation procedure (GeoSHAP). Lake Taihu was selected as a representative large shallow lake to investigate hourly-scale chlorophyll-a dynamics. The six spectral indices derived from Geostationary Ocean Color Imager-II (GOCI-II) remote-sensing reflectance products were paired with spatiotemporally matched in situ chlorophyll-a measurements to train the retrieval model. Wavelet denoising and WOA were employed to improve retrieval performance, while GeoSHAP was applied to the WRF-WOA predictions to quantify the location-specific mean relative contributions of individual spectral indices and characterize their spatial variations across Lake Taihu.

The main objectives of this study are to (i) develop an integrated satellite remote-sensing and machine-learning framework for hourly-scale chlorophyll-a retrieval, (ii) quantify the contributions of spectral indices to chlorophyll-a predictions and assess their spatial variability using GeoSHAP, and (iii) characterize the spatial distribution and hourly-scale variations in chlorophyll-a concentration and discuss their potential environmental controls in a large shallow lake. These findings are expected to provide methodological insights for eutrophication monitoring in Lake Taihu and technical support for remote-sensing-based water-quality retrieval in shallow-lake ecosystems.

Table 1. Summary of representative methods for satellite-based estimation of chlorophyll-a levels in lakes.

Study Area	Methodology	Temporal Resolution	Spatial Resolution	Spatial Interpretability	Satellite Data	References
Lake Balaton	RF	16 d	30 m	No	L4-9	[37]
Subalpine Lakes	BOMBER	5–16 d	10–30 m	No	S2A, L8	[38]
Wakasa Bay	SD	5–16 d	10–30 m	No	S2A, L8	[39]
Estonian Lakes	BR, MCI, C2RCC	5 d	10–60 m	No	S2, MSI	[40]
Hulun Lake	XGBoost	5 d	10 m	No	S2	[41]
Lake Taihu	BPNN	Continuous	High	No	proximal HSI	[42]
Lake Taihu	WRF-WOA-GeoSHAP	1 h	250 m	Yes	GOCI-II	This study

Abbreviations: BOMBER is the Bio-Optical model-based tool for estimating water quality and bottom properties from remote-sensing images, BPNN is the Back Propagation Neural Network, BR is the Band Ratio algorithm, C2RCC is the Case 2 Regional CoastColour processor, GOCI-II is the Geostationary Ocean Color Imager-II, L4-9 refers to Landsat 4-9, L8 refers to Landsat-8, MCI is the Maximum Chlorophyll Index, MSI is the MultiSpectral Instrument, proximal HSI is the Proximal Hyperspectral Imager, RF is the Random Forest model; S2/S2A refers to Sentinel-2/Sentinel-2A, SD is the Spectral Decomposition method, XGBoost is the Extreme Gradient Boosting model.

2. Materials and Methods

2.1. Study Area and Sampling Sites

Lake Taihu is the third largest freshwater lake in China, with a surface area of 2338 km^2 and an average water depth of 2 m, representing a typical large shallow lake. It is a vital water source occupying a critical position in the Yangtze River Delta region. This basin is characterized by a densely developed river network, rapid economic growth, and high population density [43]. However, sustained inputs of industrial wastewater, municipal sewage, and agricultural runoff via tributary inflows, driven by rapid industrialization and urbanization, have rendered it one of the most severely eutrophic lakes in China [44]. The water quality of Lake Taihu exhibits obvious spatial heterogeneity. The southeastern sector, densely colonized by submerged macrophytes, maintains relatively

favorable water quality but introduces optical interference into satellite remote-sensing signals. The macrophyte-dominated zone (approximately 15% of the total lake area) was therefore excluded from this study. To capture the spatial heterogeneity of water quality, hydrodynamic conditions, and algal distribution, six representative sampling sites were established across the principal ecological zones, including Meiliang station (S1), Wangyu station (S2), Lanshanzui station (S3), Wuguishan station (S4), Yangjiapu station (S5) and Changdou station (S6) (Figure 1).

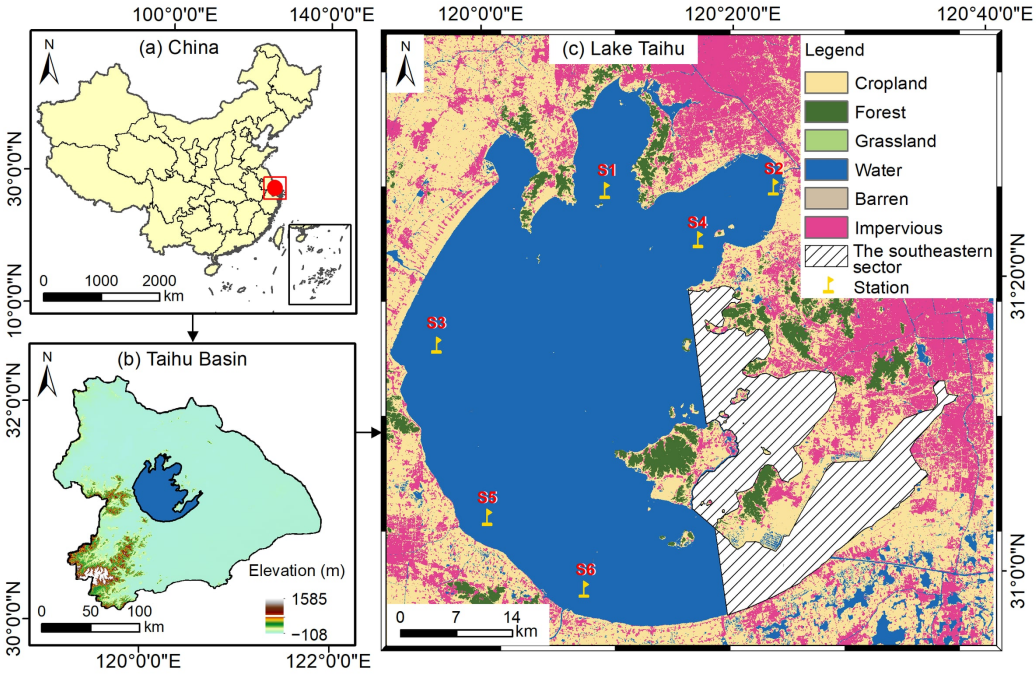

Figure 1. Location of sampling sites in the Lake Taihu. The red dot in (**a**) indicates the location of Lake Taihu in China, while the red rectangle outlines the area enlarged in (**b**).

2.2. Sample Collection and Measurement

Field sampling was conducted in Lake Taihu from January 2024 to December 2024. The geographic coordinates of sampling were recorded using a handheld GPS receiver (eTrex 32x; Garmin International, Inc., Olathe, KS, USA). At each sampling stations, water samples were collected at three depths, including the top layer, 0.5 m, and 1 m. The collected samples were combined in equal volumes, filtered through a 25 μm mesh (SEFAR PA; Sefar AG, Heiden, Switzerland), and a 100 mL aliquot was preserved for subsequent analysis [45]. This study determined chlorophyll-a concentration using the hot ethanol extraction method. Water samples (100–500 mL) were filtered onto GF/F glass microfiber filters (Whatman; Cytiva, Marlborough, MA, USA), which were subsequently immersed in 10 mL of 80 °C ethanol and extracted for 20 min. Spectrophotometric measurements of the extracts were performed using a UV–Vis spectrophotometer (UV-1800; Shimadzu Corporation, Kyoto, Japan) at 665 nm and 750 nm, with absorbance at 750 nm applied to correct for non-chlorophyll interference. To minimize the risk of contamination, all glassware and filtration apparatus were pre-soaked and cleaned with 10% hydrochloric acid solution prepared from analytical-grade hydrochloric acid (EMSURE®, Merck KGaA, Darmstadt, Germany).

2.3. Satellite Data and Processing

The Level-2 remote-sensing products of the Geostationary Ocean Color Imager-II (GOCI-II), provided by the National Ocean Satellite Center of South Korea, (https://www.

nosc.go.kr/eng/program/actionGociDownload.do; accessed on 11 April 2026), served as the primary satellite data source in this study. GOCI-II provides observations at a spatial resolution of 250 m and a temporal resolution of 1 h over the local observation area, enabling the characterization of sub-daily variations in inland waters. The official Level-2 remote-sensing reflectance products were directly used for subsequent spectral-index calculation and chlorophyll-a retrieval.

(i) Atmospheric correction

Atmospheric correction of the GOCI-II Level-2 products was performed through the operational GOCI-II Ground Segment processing system prior to product distribution [46]. The standard GOCI-II atmospheric correction algorithm is theoretically based on the previous GOCI atmospheric correction framework and was further improved for turbid waters using the additional 620 and 709 nm bands of GOCI-II [47,48]. The objective of atmospheric correction is to separate the water-leaving signal from atmospheric contributions in the top-of-atmosphere (TOA) observation. Ignoring contributions from sunglint, whitecaps, and bidirectional reflectance effects, the TOA reflectance at wavelength λ can be expressed as:

$$\rho_{\mathrm{TOA}}(\lambda) = \rho_{\mathrm{r}}(\lambda) + \rho_{\mathrm{am}}(\lambda) + t_{\mathrm{d}}^{\mathrm{v}}(\lambda)t_{\mathrm{d}}^{\mathrm{s}}(\lambda)\rho_{\mathrm{wn}}(\lambda) \quad (1)$$

where ρ_{r} is the Rayleigh-scattering reflectance, ρ_{am} represents aerosol multiple-scattering reflectance including aerosol–molecule interactions, $t_{\mathrm{d}}^{\mathrm{s}}(\lambda)$ and $t_{\mathrm{d}}^{\mathrm{v}}(\lambda)$ are the downward and upward diffuse atmospheric transmittances, respectively, and ρ_{wn} is the normalized water-leaving reflectance.

The Rayleigh contribution is calculated from the solar–sensor geometry and atmospheric conditions, whereas the aerosol contribution is estimated using the spectral relationships of aerosol multiple-scattering reflectance (SRAMS). The SRAMS relationship between two wavelengths can be represented as:

$$\rho_{am}^{M}(M_i, \lambda_2) = \sum_{n=1}^{D} c_n(M_i, \lambda_1, \lambda_2, \Theta)[\rho_{am}(\lambda_1)]^n \quad (2)$$

where M_i denotes the candidate aerosol model, D is the polynomial degree, c_n represents the coefficients stored in the look-up table, and Θ describes the solar–sensor viewing geometry. This relationship allows aerosol information derived from near-infrared wavelengths to be extrapolated to the visible spectral region.

For optically complex and turbid waters, the conventional assumption of negligible near-infrared water-leaving reflectance may not be valid. The GOCI-II atmospheric correction framework therefore incorporates additional spectral information at 620 and 709 nm, together with the 745 and 865 nm bands, to improve the iterative separation of water and aerosol reflectance. For example, the SR709 model estimates the normalized water-leaving reflectance at 745 nm from that at 709 nm as:

$$\rho_{wn}(745) = \sum_{n=0}^{3} l_n[\rho_{wn}(709)]^n \quad (3)$$

where l_n denotes the polynomial coefficients. Previous validation using simulated and in situ Rrs datasets showed that the GOCI-II turbid-water correction schemes improved the estimation of near-infrared water reflectance, particularly at 745 and 865 nm.

The final remote-sensing reflectance used in this study is related to the normalized water-leaving reflectance by:

$$R_{\mathrm{rs}}(\lambda) = \frac{\rho_{\mathrm{wn}}(\lambda)}{\pi} \quad (4)$$

Because atmospheric correction had already been implemented within the official GOCI-II Level-2 processing chain, no secondary atmospheric correction was conducted in

this study. The quality-screened official Level-2 products were directly adopted to preserve radiometric consistency and to avoid introducing additional uncertainty associated with different atmospheric-processing schemes.

(ii) Spatiotemporal matching

The GOCI-II Level-2 products were first subjected to quality screening, during which images severely affected by cloud contamination or failing to meet the satellite–in situ matching requirements were excluded. Considering the rapid short-term variation in cyanobacterial blooms in Lake Taihu, a strict temporal matching criterion of $\Delta t \leq 15$ min was adopted to minimize the temporal mismatch between satellite observations and in situ measurements. Although a wider temporal window could increase the number of satellites–in situ matchups, it could also introduce additional temporal representativeness errors and weaken the consistency required for hourly chlorophyll-a retrieval. The in situ observations were obtained from continuously operating monitoring stations and complementary manual field measurements and were spatially matched to the corresponding 250 m GOCI-II pixels. These observations were treated as representative of the mean chlorophyll-a concentration within the corresponding pixels. Under this point-to-pixel spatial matching assumption, the 15 min temporal criterion was retained to minimize additional temporal uncertainty and avoid compounding spatial and temporal mismatches. After quality control and spatiotemporal matching, approximately 300 GOCI-II scenes and valid satellite–in situ matchup samples were retained, providing an adequate sample size for subsequent machine-learning model training and validation.

2.4. Development of Remote-Sensing-Derived Indices

Based on GOCI-II remote-sensing imagery, this study calculated six representative spectral indices to estimate chlorophyll-a concentration. These indices effectively captured the optical characteristics of chlorophyll-a in aquatic environments and were well-suited for water quality monitoring [49,50]. Although GOCI-II provides 12 narrow spectral bands, only five native bands (B8–B12) were required for the six spectral indices used in this study. All reflectance values used in these calculations were obtained directly from the native GOCI-II bands at 660, 680, 709, 745, and 865 nm (Table 2), without spectral interpolation. AFAI, FLHI, RESI1, and RESI2 were calculated following formulations previously established for GOCI observations of Lake Taihu, whereas NDCI and TBI retained their classical mathematical structures, with the original wavelengths adapted to the nearest spectrally corresponding native GOCI-II bands.

Table 2. Characteristics of the GOCI-II spectral bands used in this study.

GOCI-II Band	Central Wavelength (nm)	Spatial Resolution	Temporal Resolution	Spectral Indices Used in This Study
B8	660	250 m	1 h	AFAI, FLHI, NDCI, TBI, $RESI_1$
B9	680	250 m	1 h	FLHI, $RESI_2$
B10	709	250 m	1 h	NDCI, TBI
B11	745	250 m	1 h	AFAI, FLHI, TBI, $RESI_1$, $RESI_2$
B12	865	250 m	1 h	AFAI

The specific calculation formula of alternative floating algae index (AFAI) was as follows:

$$\mathrm{AFAI} = R_{745} - [R_{660} + \frac{745 - 660}{865 - 660}(R_{865} - R_{660})] \quad (5)$$

where R_{660}, R_{745} and R_{865} denote the remote-sensing reflectance at the 660 nm red band, the 745 nm near-infrared band, and 865 nm long-wave near-infrared band, respectively. The AFAI quantifies the deviation of the 745 nm reflectance from the spectral baseline between 660 and 865 nm, thereby enhancing the spectral signal associated with surface-floating algae and exhibiting high sensitivity to cyanobacterial blooms [51].

The fluorescence line height index (FLHI) was calculated as follows:

$$\mathrm{FLHI} = R_{680} - [R_{660} + \frac{680 - 660}{745 - 660}(R_{745} - R_{660})] \quad (6)$$

where R_{680}, R_{660} and R_{745} denote the remote-sensing reflectance at the central wavelengths of 680 nm, 660 nm and 745 nm, respectively. The FLHI characterizes the elevation of the 680 nm reflectance above the spectral baseline between 660 and 745 nm, thereby enhancing the chlorophyll-a fluorescence signal while reducing background spectral effects [52].

The normalized difference chlorophyll index (NDCI) was defined as follows:

$$\mathrm{NDCI} = \frac{R_{709} - R_{660}}{R_{709} + R_{660}} \quad (7)$$

where R_{709} and R_{660} are the remote-sensing reflectance at the 709 nm and 660 nm bands. By exploiting the normalized spectral contrast between the red absorption and red-edge reflectance regions, NDCI reduces the influence of variations in overall reflectance magnitude and exhibits high sensitivity to chlorophyll-a variations [53].

The three-band index (TBI) was formulated as follows:

$$\mathrm{TBI} = (\frac{1}{R_{660}} - \frac{1}{R_{709}}) \times R_{745} \quad (8)$$

where R_{660}, R_{709} and R_{745} are the remote-sensing reflectance at 660 nm, 709 nm, and 745 nm, respectively. By combining the inverse-reflectance difference between the red and red-edge bands with the near-infrared reflectance, TBI enhances chlorophyll-a-related spectral variations while reducing interference from other optically active constituents, thereby improving its applicability to turbid shallow waters [54].

The red-edge spectral index 1 (RESI_1) was calculated as follows:

$$\mathrm{RESI}_1 = \frac{R_{745}}{R_{660}} \quad (9)$$

where R_{745} and R_{660} are the remote-sensing reflectance at the 745 nm and 660 nm bands. RESI_1 characterizes the spectral contrast between the red and near-infrared regions, providing an algae-sensitive spectral feature for subsequent chlorophyll-a retrieval [55].

The red-edge spectral index 2 (RESI_2) was expressed as follows:

$$\mathrm{RESI}_2 = \frac{R_{745}}{R_{680}} \quad (10)$$

where R_{745} and R_{680} denote the remote-sensing reflectance at the 745 nm and 680 nm bands. The RESI_2 characterizes the spectral contrast between the near-infrared and chlorophyll-a fluorescence regions, providing complementary information on algal biomass and chlorophyll-a variations [56].

2.5. 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 (Figure 2).

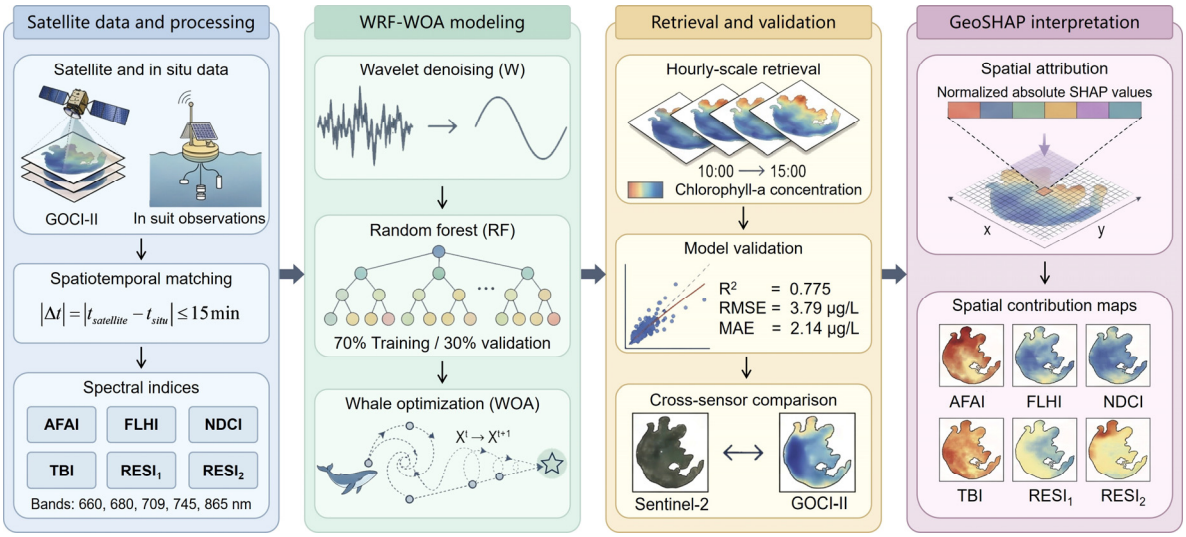

Figure 2. The hybrid framework for tracking the hourly dynamics of chlorophyll-a concentration.

(i) Step1. Construction of the W model

In chlorophyll-a retrieval, wavelet denoising was used to reduce noise in the input features [57–59]. 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 [60]. To address boundary effects caused by the finite length of the input sequences, symmetric extension was adopted at both ends of each segment [61]. The wavelet decomposition was expressed as follows:

$$\mathrm{DWT}_{\psi,\mathrm{sym}} f[(i)] = \left[D_1, D_2, \ldots, D_j, A\right] \quad (11)$$

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 [62]. Specifically, the noise standard deviation was estimated using the median absolute deviation of the finest-scale detail coefficients.

$$\sigma = \frac{\text{median}(|D_1(k) - \text{median}(D_1)|)}{0.6745} \tag{12}$$

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:

$$\lambda = \sigma\sqrt{2\ln N} \tag{13}$$

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.

$$\tilde{D}_j(k) = \begin{cases} D_j(k), & \left|D_j(k)\right| \geq \lambda \\ 0\,, & \left|D_j(k)\right| < \lambda \end{cases} \tag{14}$$

where $D_j(k)$ is the original wavelet detail coefficient, $\tilde{D}_j$ 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.

$$\tilde{f}(i) = \text{IDWT}_{\psi,\text{sym}}\left[\tilde{D}_1, \tilde{D}_2, \ldots, \tilde{D}_j, A\right] \tag{15}$$

where $\tilde{f}(i)$ 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 [61].

$$J_{\max} = \left\lfloor \log_2\left(\frac{N}{L_\psi - 1}\right) \right\rfloor \tag{16}$$

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, $L_\psi = 6$. 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 [63]. It has been widely applied in land surface monitoring and remote-sensing

parameter retrieval [64]. 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.

$$\tilde{f}(I) = [\tilde{f}(1), \tilde{f}(2), \ldots, \tilde{f}(i)] \tag{17}$$

where $\tilde{f}(I)$ and $\tilde{f}(i)$ 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.

$$\tilde{y}(I_m) = \int_T \tau(\tilde{f}(I_m)) dP(\tau/\theta) \tag{18}$$

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, $\tau(\tilde{f}(I_m))$ 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 [65]. The prediction performance of the RF model depends strongly on its hyperparameter settings [28]. To address this issue, WOA was employed to optimize the hyperparameters of the RF model within the WRF framework [66]. 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:

$$L_\theta = \frac{1}{5}\sum_{k=1}^{5} \mathrm{RMSE}_k(\theta) \tag{19}$$

$$\begin{cases} \tilde{y}^*(I_m) = \int_T \tau(\tilde{f}(I_m)) \mathrm{d}P(\tau/\theta^*) \\ \theta^* = \mathrm{argmin}_{\boldsymbol{\theta}\in\Omega} L(\theta) \end{cases} \tag{20}$$

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 3. 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 [67]. 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 [68,69]. 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:

$$\tilde{y}^*(I_m) = \phi_0 + \sum_{i=1}^{I} \phi[\tilde{y}^*{}_i(I_m)] \tag{21}$$

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:

$$\phi[\tilde{y}_i^{\,*}(I_m)] = \sum_{S \subseteq \tilde{f}(I) \setminus \{\tilde{f}(i)\}} \frac{|S|!(|I| - |S| - 1)!}{|I|!} \left[u[\tilde{f}(S \cup \{i\})] - u[\tilde{f}(S)] \right] \tag{22}$$

where I denotes the set of all input features, S is a subset of features excluding feature i, $u(\tilde{f}(S))$ is the model prediction based on the feature subset S, whereas $u[\tilde{f}(S \cup \{i\})]$ 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:

$$C_{(m,i)} = \frac{\left|\phi[\tilde{y}_i^{\,*}(I_m)]\right|}{\sum_{i=1}^{I} \left|\phi[\tilde{y}_i^{\,*}(I_m)]\right|} \tag{23}$$

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:

$$\overline{C}_{(m,i)} = \frac{1}{N_m} \sum_{k=1}^{N_m} C_{(m,i,k)} \tag{24}$$

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. $\overline{C}_{(m,i)}$ 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.

2.6. Model Validation

(i) Chlorophyll-a-based trophic status classification

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 [70]. The eutrophic category was further divided into light eutrophic and severe eutrophic levels (Table 4) [71]. These concentration thresholds were applied to the retrieved chlorophyll-a fields to quantify the areal proportions and hourly variations in 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 4. Chlorophyll-a-based trophic status classification in Lake Taihu.

Trophic Status		Chlorophyll-a Concentration (ug/L)
Oligotrophic		≤1.6
Mesotrophic		(1.6, 10]
Eutrophic	Light eutrophic	(10, 26]
	Severe eutrophic	>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 are provided below.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{C}_i - C_i\right)^2} \tag{25}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{C}_i - C_i\right| \tag{26}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{C}_i - C_i\right)^2}{\sum_{i=1}^{n}\left(C_i - \overline{C}\right)^2} \tag{27}$$

where $\hat{C}_i$ is the predicted value for the i-th sample, C_i is the corresponding in situ measured value, $\overline{C}$ 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.

3. Results

3.1. Statistical Characteristics of In Situ Chlorophyll-a

The chlorophyll-a concentration dataset constructed in this study had a mean value of 17 μg/L and a coefficient of variation of 82.2%, indicating a high degree of data dispersion. The chlorophyll-a concentration exhibited a right-skewed distribution, with most

samples falling within the low-concentration range and only a small number extended toward higher concentrations (Figure 3). Previous studies have shown that chlorophyll-a concentration in natural waters typically follows a lognormal or skewed distribution. The distribution observed in this study was consistent with the Pearson type III distribution curve. Additionally, several outliers were identified in the dataset, likely corresponding to periods of cyanobacterial blooms or localized algal scum events. Overall, the chlorophyll-a concentration was generally low, with few occurrences of eutrophic levels, suggesting that the water condition of Lake Taihu remained fairly favorable and stable during the study period.

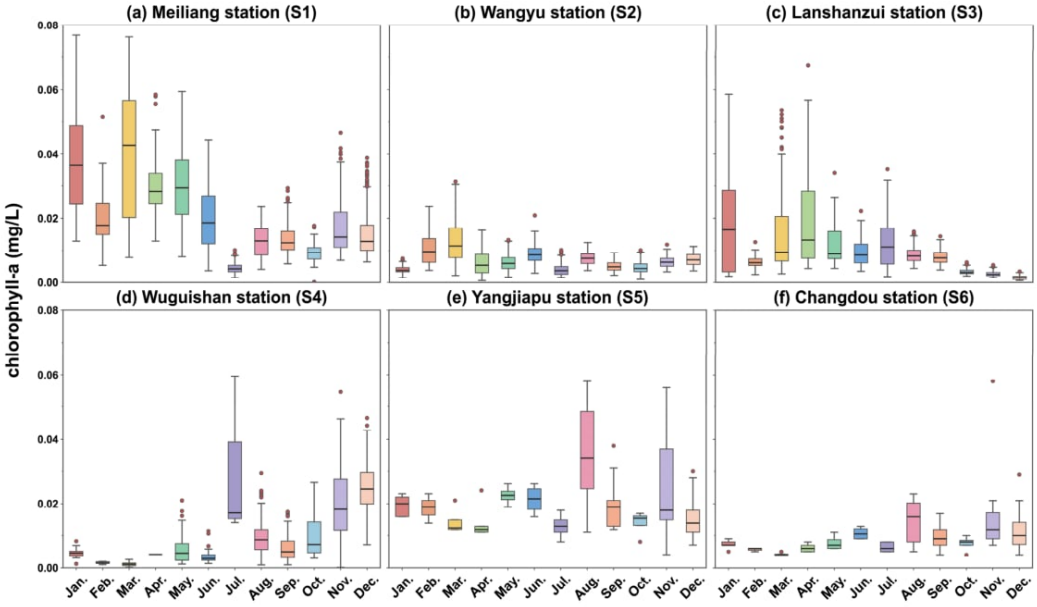

Figure 3. Statistical distribution of the chlorophyll-a concentration dataset in 2024. **The red dots in the boxplots indicate outliers.**

3.2. Spatiotemporal Distribution of Remote-Sensing-Derived Indices

Considering both sampling accessibility and the availability of cloud-free remote-sensing imagery, two representative dates within the cyanobacterial bloom season (May–September) were selected for hourly-scale analysis. As shown in Figure 4, the spatial distributions of the six remote-sensing-derived indices on 17 May and 25 August 2024 exhibited pronounced spatial heterogeneity across Lake Taihu. High-value areas were consistently concentrated in Meiliang Bay and the northwestern bay region, where relatively high chlorophyll-a-related spectral responses were observed. In addition, elevated index values were distributed along parts of the western littoral zone and extended toward the central lake, whereas relatively lower values generally occurred in the eastern and southeastern regions. Overall, the six indices collectively revealed a clear spatial pattern characterized by localized high-value accumulation in the northwestern and Meiliang Bay regions and marked spatial gradients from nearshore zones toward the lake center, reflecting the strong spatial heterogeneity of chlorophyll-a-related optical conditions in Lake Taihu.

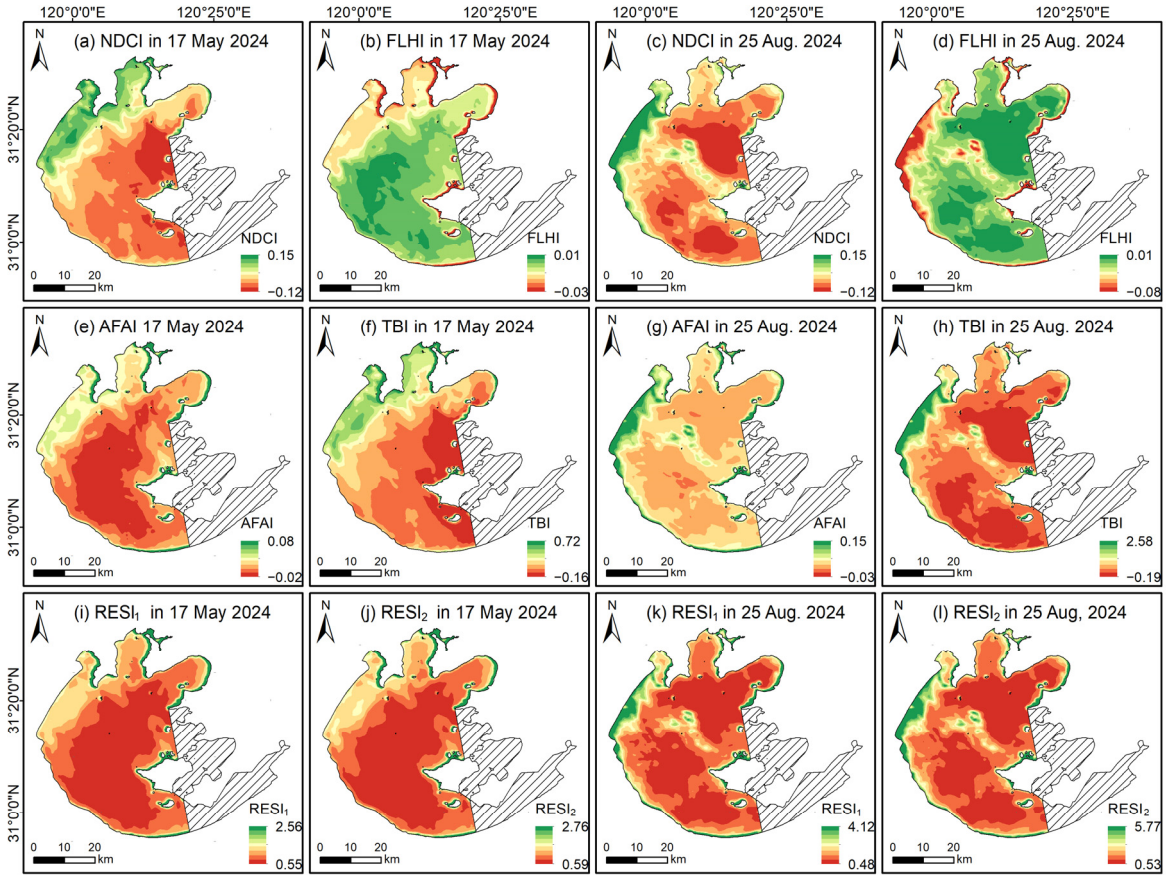

Figure 4. Spatial distribution of remote-sensing indices. NDCI, FLHI, AFAI, TBI, $RESI_1$ and $RESI_2$ denote the normalized difference chlorophyll index, the fluorescence line height index, alternative floating algae index, the three-Band index, the red-edge spectral index 1, and the red-edge spectral index 2, respectively.

3.3. Model Training and Ablation Analysis

This study adopted a stratified random sampling strategy, with 70% of all valid samples used to construct the training set and the remaining 30% reserved as an independent validation set. During model training, the WOA optimized the RF hyperparameters using five-fold cross-validation conducted within the training set, with the mean RMSE across folds serving as the optimization objective. The 30% validation set was not involved in model fitting, feature selection, or hyperparameter optimization, and was retained as a held-out dataset for evaluating and comparing the generalization performance of the predefined model configurations. Based on this procedure, the optimized model was used for chlorophyll-a retrieval across Lake Taihu.

To further investigate the individual contributions of wavelet denoising and WOA to chlorophyll-a retrieval, four model configurations, including RF, RF-WOA, WRF, and WRF-WOA, were established for ablation analysis (Figure 5). The RF and WRF models were used as baseline configurations without WOA optimization, whereas RF-WOA and WRF-WOA incorporated WOA-based hyperparameter optimization. All four models adopted the same training and validation data partitioning strategy. For the WOA-based models, identical hyperparameter search ranges and optimization settings were applied to ensure comparable optimization conditions. The detailed wavelet configurations, hyperparameter search ranges, WOA optimization settings, and the corresponding optimal hyperparameter configurations are provided in Supplementary Table S3.

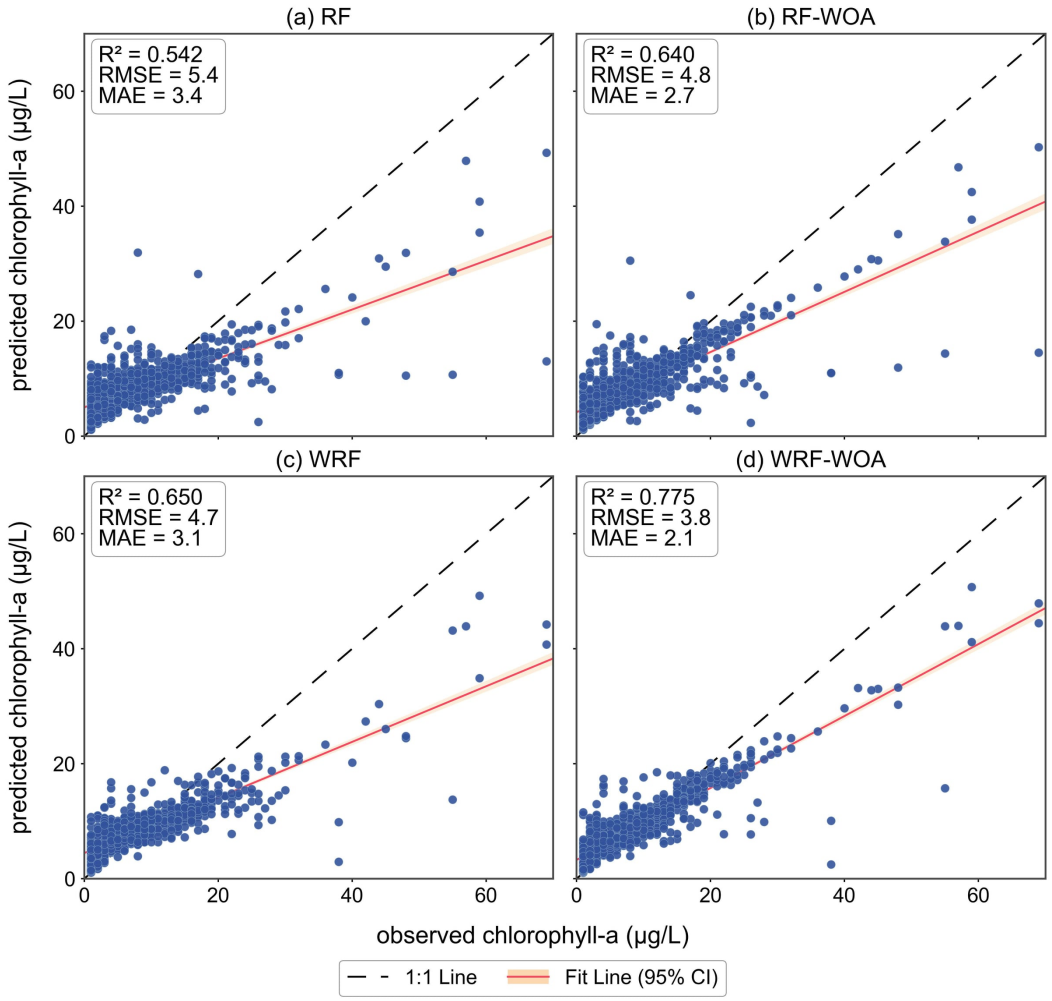

Figure 5. Ablation analysis of different model configurations for chlorophyll-a retrieval in Lake Taihu. RF is the random forest; WOA is the whale optimization algorithm; WRF is the wavelet-denoised random forest.

As shown in Figure 5, the conventional RF model achieved an R^2 of 0.542, with RMSE and MAE values of 5.4 and 3.4 μg/L, respectively. After introducing wavelet denoising, the WRF model achieved an improved R^2 of 0.650, while RMSE and MAE decreased to 4.7 and 3.1 μg/L, respectively. Similarly, the RF-WOA model achieved an R^2 of 0.640, with RMSE and MAE values of 4.8 and 2.7 μg/L, indicating the positive contribution of WOA-based hyperparameter optimization. The WRF-WOA model exhibited the best overall performance, achieving an R^2 of 0.775, an RMSE of 3.8 μg/L, and an MAE of 2.1 μg/L. Compared with the conventional RF model, the RMSE and MAE of WRF-WOA decreased by 29.63% and 38.24%, respectively. These results demonstrated that wavelet denoising and WOA optimization individually improved retrieval performance, while their combination yielded the most accurate chlorophyll-a estimates.

3.4. Spatiotemporal Retrieval of Chlorophyll-a Concentration

To evaluate the spatial and temporal applicability of the WRF-WOA model, complementary validation experiments were conducted at the station and seasonal levels (Figure 6). For spatial assessment, separate models were developed using observations from each of the six monitoring stations to examine modeling stability under different local conditions. For temporal assessment, the 2024 dataset was divided into a cyanobacterial bloom period (May–September) and a non-bloom period (the remaining months), and

separate models were established for each period. Furthermore, observations from all six stations throughout 2024 were integrated into an overall dataset, which was partitioned using the stratified random sampling strategy described in Section 3.3. The same 70%/30% training–validation strategy and WOA optimization procedures were applied to each experiment. The resulting overall WRF-WOA model was subsequently used for lake-wide hourly chlorophyll-a retrieval.

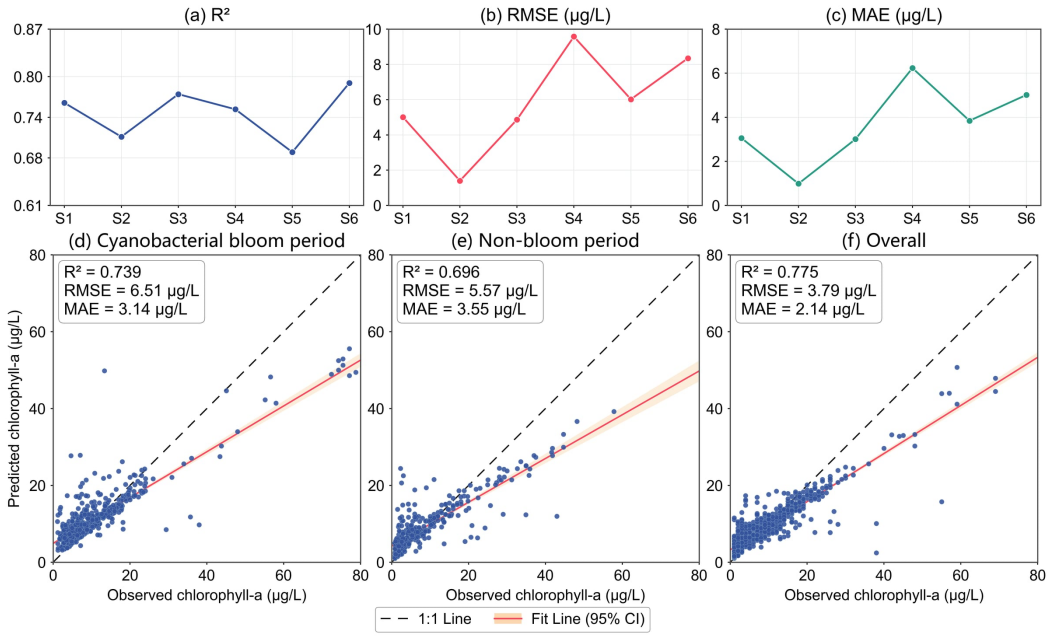

Figure 6. Spatiotemporal validation of the WRF-WOA model in Lake Taihu.

The station-specific models achieved R^2 values ranging from 0.69 to 0.79, indicating stable predictive performance among the six monitoring stations (Figure 6a–c). For temporal assessment, the bloom-period and non-bloom-period models achieved R^2 values of 0.739 and 0.696, with corresponding RMSE values of 6.51 and 5.57 μg/L and MAE values of 3.14 and 3.55 μg/L, respectively (Figure 6d,e). When all samples were integrated, the overall model achieved an R^2 of 0.775, an RMSE of 3.79 μg/L, and an MAE of 2.14 μg/L (Figure 6f). These results indicate that the WRF-WOA framework maintained satisfactory predictive performance across the evaluated station-specific and seasonal conditions.

The hourly-scale retrieval results from 10:00 to 15:00 were further examined for two representative dates during the cyanobacterial bloom season (May–September), namely 17 May and 25 August 2024 (Figure 7). The retrieved chlorophyll-a fields exhibited pronounced spatial heterogeneity and coherent hour-to-hour variations across Lake Taihu. High-concentration areas were primarily distributed in Meiliang Bay and the northwestern part of the lake, with locally elevated concentrations also occurring along the western littoral zone, whereas relatively low concentrations generally occurred in the central and eastern open-water regions. Across successive hourly observations, the locations, spatial extents, and boundaries of high-concentration patches varied noticeably, indicating localized expansion, contraction, and spatial redistribution within the lake.

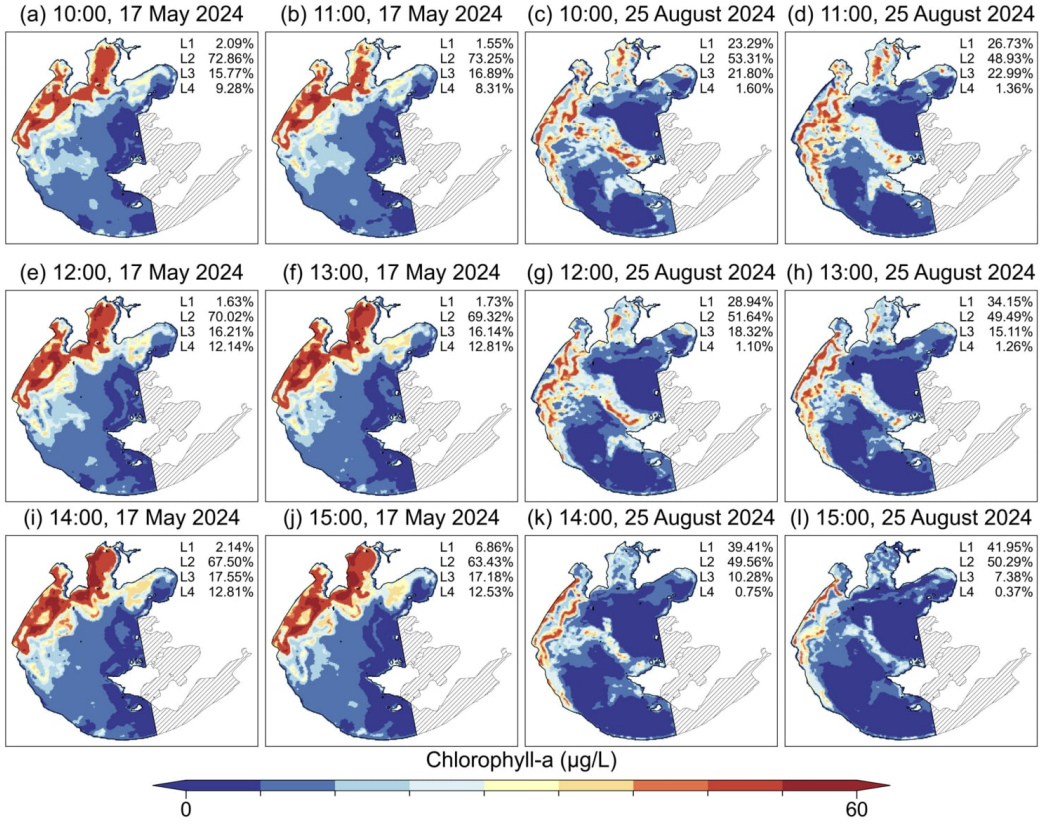

Figure 7. Hourly-scale retrieval of chlorophyll-a concentration in Lake Taihu from 10:00 to 15:00 on two representative dates during the cyanobacterial bloom season (17 May and 25 August 2024), based on the WRF-WOA model. L1: oligotrophic; L2: mesotrophic; L3: light eutrophic; and L4: severe eutrophic. Percentages indicate the proportion of valid lake pixels in each trophic category at the corresponding hour.

To quantify these variations, the chlorophyll-a concentrations were classified into four trophic levels, and their areal proportions were calculated for each hour (Figure 7). On 17 May, mesotrophic conditions dominated the lake, although their proportion decreased from 72.86% at 10:00 to 63.43% at 15:00. The proportion of severe eutrophic areas increased from 9.28% to 12.53%, reaching a maximum of 12.81% at 13:00–14:00. In contrast, on 25 August, the proportion of oligotrophic areas increased from 23.29% to 41.95%, while light eutrophic areas decreased from 21.80% to 7.38%. The combined proportion of light and severe eutrophic areas decreased from 23.40% to 7.75%. These differences highlight contrasting hourly changes in chlorophyll-a distribution between the two dates, demonstrating the capability of the WRF-WOA model to characterize hourly-scale chlorophyll-a dynamics in Lake Taihu.

3.5. Interpretability of Remote-Sensing-Derived Indices

The WRF-WOA-GeoSHAP model was employed to quantify the contribution of each input feature to chlorophyll-a concentration, thereby revealing the spatially varying contributions of different remote-sensing-derived indices to model predictions. As shown in Figure 8, FLHI exhibited the highest overall contribution, indicating the importance of fluorescence-related spectral information in chlorophyll-a retrieval. FLHI reflected the fluorescence signal associated with phytoplankton photosynthesis and was closely related to algal biomass. NDCI showed notable contributions in regions with relatively low chlorophyll-a concentrations, suggesting its relevance to chlorophyll-a retrieval under lower-concentration conditions. $RESI_1$, which incorporates red and near-infrared wavelengths,

exhibited sensitivity to the optical characteristics of algae and showed high contributions in the central lake area. TBI and AFAI also provided complementary spectral information for chlorophyll-a retrieval under complex optical conditions.

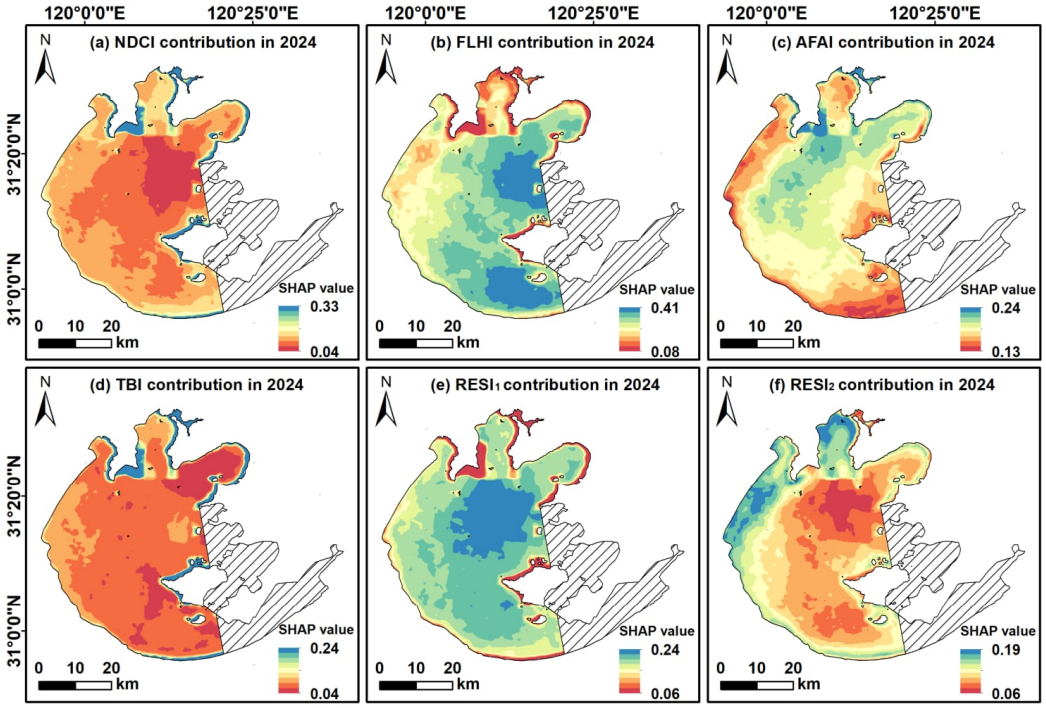

Figure 8. Feature importance of remote-sensing-derived indices. NDCI, FLHI, AFAI, TBI, $RESI_1$ and $RESI_2$ denote the normalized difference chlorophyll index, the fluorescence line height index, alternative floating algae index, the three-Band index, the red-edge spectral index 1, and the red-edge spectral index 2, respectively. For each prediction sample, absolute SHAP values were normalized by their sum and linked to geographic locations, followed by averaging across samples within each pixel. Accordingly, the values labeled as "SHAP value" represent spatially aggregated normalized absolute SHAP contributions rather than raw signed SHAP values, with larger values indicating greater relative contributions to model predictions.

In summary, all six remote-sensing-derived indices selected in this study exhibited contributions to chlorophyll-a retrieval, with each index capturing different aspects of the optical characteristics associated with chlorophyll-a. This finding indicates that reliable chlorophyll-a retrieval in large shallow lakes depends on the combined information provided by multiple spectral indices rather than any single index. An individual remote-sensing-derived index provides limited information and cannot fully represent the optical complexity of inland waters. The spatial variation in feature contributions further indicates that the relative importance of spectral indices changes across different lake regions, reflecting their differing sensitivities to local optical conditions and algal biomass.

4. Discussion

4.1. Applicability of GOCI-II Level-2 Products for Chlorophyll-a Retrieval

The direct use of operational GOCI-II Level-2 remote-sensing reflectance (Rrs) products provides a reasonable basis for subsequent chlorophyll-a retrieval. The standard GOCI-II atmospheric correction algorithm was specifically developed for the sensor and incorporates additional spectral information to improve the separation of atmospheric and water-leaving signals, particularly over optically complex and turbid waters. Initial validation of the operational algorithm showed acceptable agreement between satellite-derived and in situ

Rrs, supporting the use of the standard Level-2 products as input for downstream ocean-color applications.

Previous GOCI-II studies have also directly adopted operational Level-2 Rrs products for algal-bloom detection and other water-color applications. Jing et al. used standard GOCI-II Level-2 data and directly compared the retrieved Rrs with field measurements before applying multiple spectral algorithms for harmful algal bloom detection; relatively strong agreement was reported at several visible bands relevant to bloom identification [72]. Similarly, Zhang et al. directly employed GOCI-II Level-2 products, including fully atmospherically corrected Rrs, for algal-bloom detection in optically complex coastal waters [73]. More recently, Zhao et al. validated GOCI-II Level-2 Rrs against in situ observations and subsequently used these products for bloom identification and classification, further demonstrating their applicability to quantitative water-color analysis [74]. Collectively, these previous validations and applications indicate that the operational GOCI-II Level-2 Rrs products can serve as standardized and physically corrected inputs for subsequent spectral-index calculation and chlorophyll-a retrieval. Accordingly, the present study directly adopted the quality-screened Level-2 Rrs products to maintain consistency with the operational GOCI-II processing framework.

4.2. Spatiotemporal Dynamics and Drivers of Chlorophyll-a

Hourly-scale chlorophyll-a concentration fields were generated using the WRF-WOA-GeoSHAP hybrid model for two dates during the cyanobacterial bloom season (17 May and 25 August 2024). The retrieval results revealed pronounced spatial heterogeneity across Lake Taihu, with relatively high chlorophyll-a concentrations occurring predominantly in Meiliang Bay and the northwestern nearshore region, while low concentrations were observed in the central and eastern open-water regions (Figure 7). In addition to this regional pattern, the spatial extent, boundaries, and local distribution of high-concentration patches varied between successive hourly observations, indicating short-term redistribution of algal biomass within the lake.

The accumulation of chlorophyll-a in the northwestern region and Meiliang Bay can be attributed to the combined effects of external nutrient loading and hydrodynamic transport. The northwestern Lake Taihu receives substantial nutrient inputs from densely urbanized and industrialized catchments and major tributary inflows, providing favorable conditions for algal growth and accumulation [44]. Meanwhile, wind-driven circulation during the cyanobacterial bloom season can redistribute surface algal biomass toward the northern and northwestern littoral zones. The semi-enclosed morphology of Meiliang Bay further restricts water exchange and favors local algal retention and accumulation [75]. These processes provide a plausible mechanistic explanation for both the spatial hotspots and the hour-to-hour redistribution observed in the retrieval results.

At a broader temporal scale, chlorophyll-a concentration also exhibited clear seasonal variability throughout 2024 [76]. Water temperature is one of the most direct environmental factors affecting algal growth in lakes [77,78]. Seasonal increases in water temperature, particularly during the warm season, together with favorable nutrient availability, create suitable conditions for cyanobacterial proliferation and bloom development [79]. Therefore, the observed chlorophyll-a dynamics in Lake Taihu reflect the combined influence of seasonal environmental forcing, external nutrient supply, and short-term hydrodynamic redistribution [80].

4.3. Spatial Interpretability of Spectral-Index Contributions

Conventional SHAP analysis provides a quantitative interpretation of the contributions of individual features to model predictions. However, global feature importance summa-

rizes contributions across all samples, which may obscure spatial differences in feature contributions. Although local SHAP values can explain individual predictions, they do not directly reveal the geographical distribution of feature contributions without explicit spatial association. In contrast, GeoSHAP links the geographic coordinates of prediction samples to their corresponding SHAP values and aggregates the normalized absolute contributions at pixel locations. This approach extends conventional feature attribution from sample-level interpretation to spatially explicit analysis, allowing the identification of regional differences in spectral-index contributions. Therefore, GeoSHAP provides spatial information for understanding the spatial variation in feature importance within machine-learning models, particularly in large shallow lakes characterized by spatial heterogeneity.

The spatial contribution patterns obtained in this study highlighted the non-uniform roles of different spectral indices across Lake Taihu. Although FLHI exhibited the highest overall contribution, the contributions of individual indices varied spatially, indicating that their relative importance depended on local optical conditions and chlorophyll-a distributions. This spatial heterogeneity suggests that the spectral information associated with chlorophyll-a is not represented uniformly across the lake. The complementary contributions of the six spectral indices therefore provide a more comprehensive representation of the spatially heterogeneous optical characteristics of Lake Taihu than any single index alone. Compared with conventional global SHAP summaries, GeoSHAP further identified the spatial locations where different spectral indices contributed more strongly, providing a more geographically explicit understanding of the spectral information used by the model.

4.4. Cross-Sensor Consistency of Chlorophyll-a Retrievals

To further evaluate the robustness of the WRF-WOA-GeoSHAP model, cross-sensor validation was conducted using Sentinel-2 imagery acquired on 17 May and 25 August 2024. Chlorophyll-a concentration was retrieved from the Sentinel-2 imagery using the four-band model and calibrated against in situ observations [81]. Because the native spatial resolution of Sentinel-2 (10 m) differs substantially from that of GOCI-II (250 m), the Sentinel-2 retrievals were additionally resampled to 250 m to provide a scale-consistent comparison with GOCI-II. Both the original 10 m Sentinel-2 results and the resampled 250 m products are shown in Figure 9.

The two sensors exhibited broadly consistent regional spatial patterns on both dates, with relatively high chlorophyll-a concentrations mainly distributed in Meiliang Bay and the northwestern nearshore region, while lower concentrations generally occurred in the central and eastern open-water regions. Local differences remained in the magnitude, spatial extent, and boundaries of high-concentration patches, which may be associated with differences in sensor characteristics, spectral configuration, spatial resolution, and retrieval uncertainty. After resampling Sentinel-2 to the same 250 m spatial resolution as GOCI-II, the major spatial gradients and regional accumulation patterns remained comparable, indicating that the GOCI-II retrievals were capable of resolving the dominant lake-scale spatial heterogeneity of chlorophyll-a concentration. Overall, the cross-sensor comparison demonstrated that the WRF-WOA-GeoSHAP model could reliably reproduce the major spatial distribution characteristics of chlorophyll-a at the 250 m scale, supporting the applicability of GOCI-II for hourly-scale monitoring of chlorophyll-a dynamics in Lake Taihu.

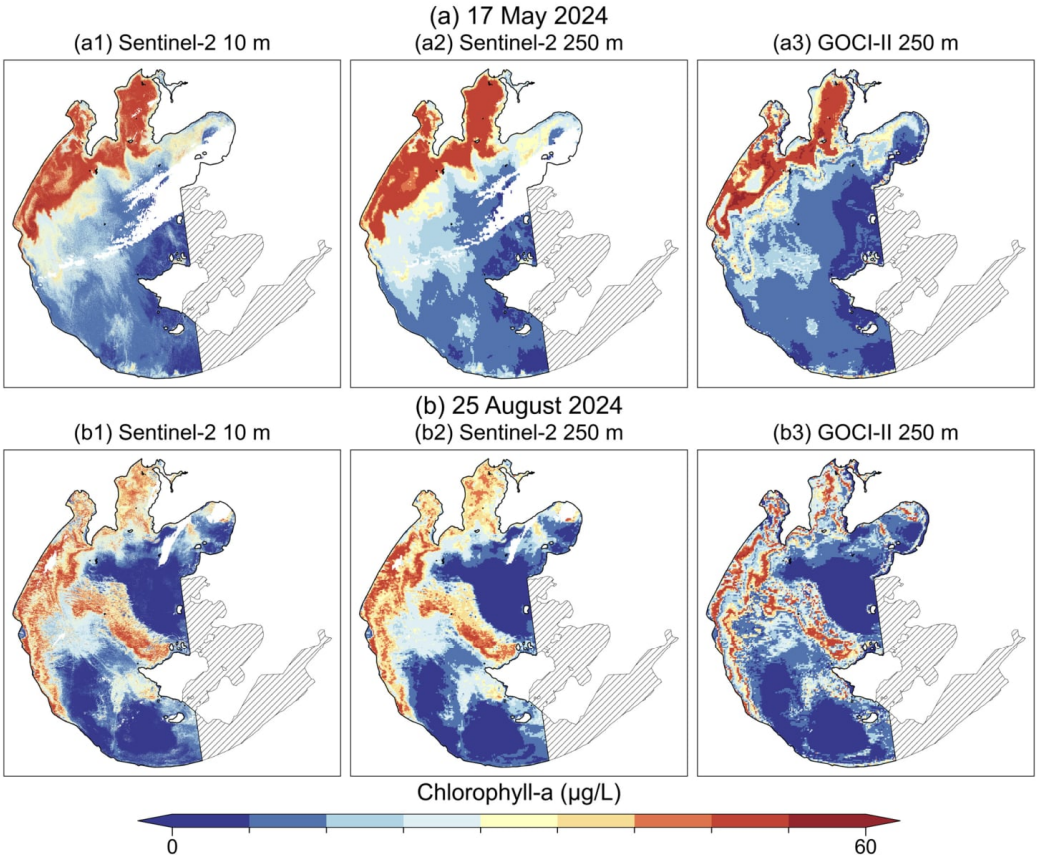

Figure 9. Cross-sensor comparison of chlorophyll-a retrievals from Sentinel-2 and GOCI-II in Lake Taihu on 17 May and 25 August 2024. Sentinel-2 retrievals are shown at the original 10 m resolution and after resampling to 250 m for comparison with the 250 m GOCI-II retrievals.

4.5. Methodological Advantages

To evaluate the retrieval performance of the proposed WRF-WOA model, twelve satellite-based chlorophyll-a retrieval models were compared, including a conventional statistical model, a gray prediction model, neural networks, and tree-based machine learning algorithms (Figure 10). The MLR and GM(1,N) models exhibited relatively poor performance, with R^2 values of 0.04 and −0.40, respectively. Similarly, the neural network models, including MLP, CNN, and LSTM, showed limited predictive performance, with negative R^2 values of −6.42, −0.31, and −0.16, respectively. In contrast, the tree-based machine learning models achieved better retrieval accuracy. Among them, GBDT, LightGBM, and XGBoost yielded R^2 values of 0.35, 0.45, and 0.42, respectively, while the conventional RF model achieved a higher R^2 of 0.54, with RMSE and MAE values of 5.41 and 3.43 μg/L, respectively. These results indicate that tree-based ensemble learning methods achieved better retrieval performance with the dataset used in this study.

These results suggest that wavelet denoising and WOA optimization provide complementary improvements to chlorophyll-a retrieval. Wavelet denoising decomposes the input spectral-index sequences into components at different scales and applies thresholding to suppress noise-related fluctuations while retaining the dominant signal characteristics. This preprocessing may improve the representation of informative spectral features and reduce the influence of high-frequency fluctuations on subsequent model training. Meanwhile, WOA optimizes the hyperparameter configuration of RF, thereby improving its predictive performance [27,65]. By improving input-feature quality and model configuration, respectively, these two strategies contribute to the overall retrieval performance of the WRF-WOA

model. Nevertheless, some underestimation remained in the high-concentration range, indicating that further improvements are needed under high chlorophyll-a concentrations.

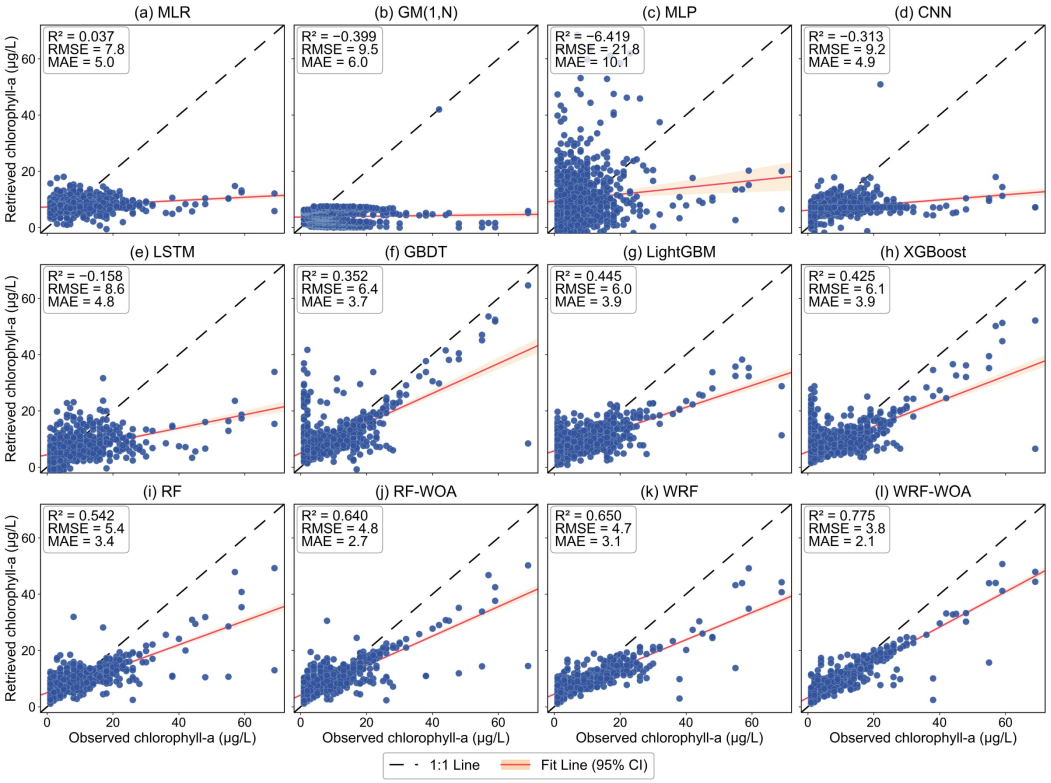

Figure 10. Comparison of model fitting accuracy. MLR is the multiple linear regression; GM(1,N) is the multivariable gray model; MLP is the multilayer perceptron; CNN is the convolutional neural network; LSTM is the long short-term memory; GBDT is the gradient boosting decision tree; LightGBM is the 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.

4.6. Limitations and Future Research

Similar to many studies, this study has limitations, and its findings should be interpreted with caution. Although GOCI-II Level-2 remote-sensing reflectance products were used after operational atmospheric correction, atmospheric correction methods were not compared. Future studies should evaluate alternative correction schemes and quantify their effects on chlorophyll-a retrieval. In addition, the 1 h observation interval of GOCI-II limits the framework's ability to capture sub-hourly variations. Multi-source remote-sensing and spatiotemporal fusion may help achieve sub-hourly monitoring and improve temporal continuity. Meanwhile, the limited number of sampling sites and the 250 m spatial resolution of GOCI-II constrain fine-scale spatial characterization. The present validation framework evaluated model performance across different monitoring stations and seasonal conditions. Future studies will further extend this assessment using independent observations from previously unobserved stations and dates to provide additional evidence of the model's spatial and temporal transferability.

5. Conclusions

Chlorophyll-a concentration is a key indicator of lake eutrophication. However, simultaneously achieving hourly temporal resolution, lake-wide spatial coverage, and model interpretability in chlorophyll-a retrieval for large shallow lakes remains challenging. To address these challenges, a novel WRF-WOA-GeoSHAP hybrid model integrating wavelet

denoising (W), random forest (RF), the whale optimization algorithm (WOA), and Geographic SHapley Additive exPlanations (GeoSHAP) was developed, with Lake Taihu serving as the study area. Spectral indices derived from Geostationary Ocean Color Imager-II (GOCI-II) remote-sensing reflectance products were paired with in situ chlorophyll-a measurements for quantitative retrieval. The GeoSHAP method was then utilized to characterize the spatial variations in spectral-index contributions to model predictions. The main conclusions are summarized as follows:

(1) The WRF-WOA-GeoSHAP model achieved satisfactory retrieval accuracy and robustness, effectively retrieving lake-wide chlorophyll-a concentrations at hourly intervals and characterizing their temporal variations, while capturing nonlinear relationships between spectral indices and chlorophyll-a concentrations in Lake Taihu. The framework provides technical support for hourly-scale chlorophyll-a monitoring under valid GOCI-II observation conditions.

(2) The retrieved chlorophyll-a fields exhibited pronounced spatial heterogeneity, with relatively high concentrations mainly distributed in Meiliang Bay and the northwestern nearshore region, together with locally elevated concentrations along parts of the western littoral zone. Across successive hourly observations, the spatial extent, boundaries, and local distribution of high-concentration patches varied noticeably, indicating short-term changes in the spatial distribution of chlorophyll-a within the lake.

(3) Cross-sensor comparison with Sentinel-2 showed that, after resampling Sentinel-2 data to the native 250 m resolution of GOCI-II, the two sensors exhibited consistent major regional spatial gradients and chlorophyll-a accumulation patterns. Although local differences remained in concentration magnitude and patch boundaries, the results indicated that the 250 m GOCI-II retrievals could characterize dominant lake-scale spatial heterogeneity and support hourly-scale chlorophyll-a monitoring in Lake Taihu.

(4) GeoSHAP analysis revealed spatial variations in the relative contributions of multiple spectral indices to chlorophyll-a retrieval, with FLHI exhibiting the highest overall contribution. Spatial differences in feature contributions highlighted the varying importance of spectral indices across lake regions, providing a geographically explicit interpretation of the retrieval results.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/rs18234091/s1, Figure S1: Modeling steps for the WRF-WOA-GeoSHAP hybrid model; Figure S2: The calculation process of the optimizing WRF hyperparameters based on the whale optimization algorithm (WRF-WOA); Figure S3: The illustration of the whale optimization algorithm (WOA); Table S1: Search ranges of the RF hyperparameters considered in WOA optimization; Table S2: Chlorophyll-a-based trophic status classification in Lake Taihu; Table S3: The parameter settings of WRF-WOA-GeoSHAP hybrid model; Table S4: Comparison of retrieval accuracy among the evaluated models. The references cited in the Supplementary Materials are included in the main reference list [28,57–70].

Author Contributions: Conceptualization, Y.X. and Z.Z.; methodology, Z.Z.; software, J.Z.; validation, Z.Z.; formal analysis, J.Z. and Z.Z.; investigation, W.T.; resources, Z.Z.; data curation, W.D. and H.Q.; writing—original draft preparation, J.Z.; writing—review and editing, Z.Z. and Y.L.; visualization, J.Z.; supervision, C.W. and Y.L.; project administration, W.D.; funding acquisition, Y.X. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (Grant Nos. 42401136 and 42501393); the Guangxi Key Research and Development Program (Grant No. AB25069137); the Basic Research Program of Jiangsu (Grant No. BK20251072); the Hebei Provincial Water Conservancy Science and Technology Program Project (Grant No. 2026-19); the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China (Grant Nos. 24KJB170011

and 25KJB170003); the Philosophy and Social Sciences Research Program for Higher Education Institutions in Jiangsu Province (Grant No. 2025SJYB0351); the International (Regional) Cooperation and Exchange Program of the National Natural Science Foundation of China (Grant No. W2523017); the Open Foundation of Key Laboratory of Taihu Basin Water Resources Management, Ministry of Water Resources; the Open Fund of the Key Laboratory of Integrated Regulation and Resources Development of Shallow Lakes of Ministry of Education (Grant No. B240203001); the Open Research Fund Program of Key Laboratory for Water Ecology Management and Protection in River Source Areas, Ministry of Water Resources (Grant No. 2025slbsy01); the Open Funds of the Key Laboratory of Mountain Hazards and Engineering Resilience, Chinese Academy of Sciences (KLMHER-K44); the Open Research Fund of Key Laboratory of Flood&Drought Disaster Defense, the Ministry of Water Resources (Grant No. KF202605); the College Students' Innovation Training Projects of Nanjing Forestry University (Grant No. 202610298001Z); and the College Students' Innovation Training Projects of Jiangsu Second Normal University (Grant No. X2025144360099).

Data Availability Statement: The datasets used or generated during this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the Korea Ocean Satellite Center (KOSC), Korea Institute of Ocean Science and Technology (KIOST), for providing the GOCI-II satellite data, and the European Space Agency (ESA) for providing the Sentinel-2 satellite data.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AFAI	Alternative floating algae index
DWT	A discrete wavelet transform
FLHI	Fluorescence line height index
GeoSHAP	Geographic SHapley Additive exPlanations
GOCI-II	Geostationary Ocean Color Imager-II
IDWT	A inverse discrete wavelet transform
MLR	Conventional multiple linear regression
NDCI	Normalized difference chlorophyll index
$RESI_1$	The red-edge spectral index 1
$RESI_2$	The red-edge spectral index 2
RF	Random forest model
TBI	The three-Band index
W	Wavelet denoising
WOA	Whale optimization algorithm
WRF-WOA-GeoSHAP	The hybrid approach for tracking the hourly dynamics of chlorophyll-a concentration

References

1. Downing, J.A.; Prairie, Y.T.; Cole, J.J.; Duarte, C.M.; Tranvik, L.J.; Striegl, R.G.; McDowell, W.H.; Kortelainen, P.; Caraco, N.F.; Melack, J.M.; et al. The Global Abundance and Size Distribution of Lakes, Ponds, and Impoundments. *Limnol. Oceanogr.* **2006**, *51*, 2388–2397. [CrossRef]
2. Tranvik, L.J.; Downing, J.A.; Cotner, J.B.; Loiselle, S.A.; Striegl, R.G.; Ballatore, T.J.; Dillon, P.; Finlay, K.; Fortino, K.; Knoll, L.B.; et al. Lakes and Reservoirs as Regulators of Carbon Cycling and Climate. *Limnol. Oceanogr.* **2009**, *54*, 2298–2314. [CrossRef]
3. Ho, J.C.; Michalak, A.M.; Pahlevan, N. Widespread Global Increase in Intense Lake Phytoplankton Blooms since the 1980s. *Nature* **2019**, *574*, 667–670. [CrossRef] [PubMed]
4. Zhang, Z.; Li, Y.; Wang, X.; Li, H.; Zheng, F.; Liao, Y.; Tang, N.; Chen, G.; Yang, C. Assessment of river health based on a novel multidimensional similarity cloud model in the Lhasa River, Qinghai-Tibet Plateau. *J. Hydrol.* **2021**, *603*, 127100. [CrossRef]
5. Zhang, Z.; Wang, X.; Li, Y.; Liu, Y.; Xu, Y.; Li, J.; Ding, W.; Li, H.; Yang, H. River Ecosystem Health Assessment in the Qinghai-Tibet Plateau: A Novel Hybrid Method Based on Artificial Intelligence and Multi-Source Data Fusion. *Expert Syst. Appl.* **2024**, *251*, 124078. [CrossRef]
6. Dillon, P.J.; Rigler, F.H. The Phosphorus-Chlorophyll Relationship in Lakes1,2. *Limnol. Oceanogr.* **1974**, *19*, 767–773. [CrossRef]

7. Carlson, R.E. A Trophic State Index for Lakes1. *Limnol. Oceanogr.* **1977**, *22*, 361–369. [CrossRef]
8. Zhang, J.; Huang, J.; Ji, Y.; Liu, L.; Huang, Z.; Luo, S.; Gao, J. Altering Water Flow Pathway to Enhance Nutrient Retention in Lowland Areas. *Water Res.* **2025**, *287*, 124429. [CrossRef] [PubMed]
9. Qin, B.; Xu, P.; Wu, Q.; Luo, L.; Zhang, Y. Environmental Issues of Lake Taihu, China. *Hydrobiologia* **2007**, *581*, 3–14. [CrossRef]
10. Yan, X.; Xia, Y.; Ti, C.; Shan, J.; Wu, Y.; Yan, X. Thirty Years of Experience in Water Pollution Control in Taihu Lake: A Review. *Sci. Total Environ.* **2024**, *914*, 169821. [CrossRef] [PubMed]
11. Lv, Y.; Zhang, M.; Kong, M.; Zhang, L.; Yin, H. Source Identification and Phosphorus Transformation of Suspended Particulate Matter in the Eutrophic Bays of a Large Lake (Lake Taihu, China). *Water Res.* **2026**, *289*, 124835. [CrossRef] [PubMed]
12. Sagan, V.; Peterson, K.T.; Maimaitijiang, M.; Sidike, P.; Sloan, J.; Greeling, B.A.; Maalouf, S.; Adams, C. Monitoring Inland Water Quality Using Remote Sensing: Potential and Limitations of Spectral Indices, Bio-Optical Simulations, Machine Learning, and Cloud Computing. *Earth Sci. Rev.* **2020**, *205*, 103187. [CrossRef]
13. Kallio, K.; Malve, O.; Siivola, E.; Kervinen, M.; Koponen, S.; Lepistö, A.; Lindfors, A.; Laine, M. Spatiotemporal Analysis of Lake Chlorophyll-a with Combined in Situ and Satellite Data. *Environ. Monit. Assess.* **2023**, *195*, 465. [CrossRef] [PubMed]
14. Cao, H.; Han, L.; Li, L. A Deep Learning Method for Cyanobacterial Harmful Algae Blooms Prediction in Taihu Lake, China. *Harmful Algae* **2022**, *113*, 102189. [CrossRef] [PubMed]
15. Wang, S.; Zhang, X.; Chen, N.; Wang, W. Classifying Diurnal Changes of Cyanobacterial Blooms in Lake Taihu to Identify Hot Patterns, Seasons and Hotspots Based on Hourly GOCI Observations. *J. Environ. Manag.* **2022**, *310*, 114782. [CrossRef] [PubMed]
16. Li, J.; Tian, L.; Song, Q.; Sun, Z.; Yu, H.; Xing, Q. Temporal Variation of Chlorophyll-a Concentrations in Highly Dynamic Waters from Unattended Sensors and Remote Sensing Observations. *Sensors* **2018**, *18*, 2699. [CrossRef] [PubMed]
17. Xue, K.; Ma, R.; Hu, M.; Wang, M.; Cheng, Y.; Jiang, L.; Huang, Z. Geostationary Satellite Observations Reveal Diurnal Dynamics of Particulate Organic Carbon in Optically Complex Lakes. *J. Hydrol.* **2026**, *672*, 135353. [CrossRef]
18. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep Learning in Environmental Remote Sensing: Achievements and Challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [CrossRef]
19. Tao, Y.; Yang, M.; Ren, J.; Zhu, H.; Liang, Z.; Li, J.; Cui, H. Towards Precision Limnology: An Explainable AI Framework Decoding Spatiotemporal Algal Dynamics in Chinese Major Lakes. *J. Hazard. Mater.* **2026**, *501*, 140879. [CrossRef] [PubMed]
20. Ramgirkar, A.; Rao, V.; Talhar, J.; Mishra, T.K.; Narayanan, S.J.; Satapathy, S.M.; Perumal, B. Mining of Heterogeneous Time Series Information for Predicting Chlorophyll Accumulation in Oceans. *Sustain. Comput. Inform. Syst.* **2024**, *42*, 100980. [CrossRef]
21. Gitelson, A.A.; Dall'Olmo, G.; Moses, W.; Rundquist, D.C.; Barrow, T.; Fisher, T.R.; Gurlin, D.; Holz, J. A Simple Semi-Analytical Model for Remote Estimation of Chlorophyll-a in Turbid Waters: Validation. *Remote Sens. Environ.* **2008**, *112*, 3582–3593. [CrossRef]
22. Le, C.; Li, Y.; Zha, Y.; Sun, D.; Huang, C.; Lu, H. A Four-Band Semi-Analytical Model for Estimating Chlorophyll a in Highly Turbid Lakes: The Case of Taihu Lake, China. *Remote Sens. Environ.* **2009**, *113*, 1175–1182. [CrossRef]
23. Du, Y.; Li, J.; Zhang, B.; Yan, K.; Zhao, H.; Wang, C.; Mu, Y.; Zhang, F.; Wang, S.; Wang, M. Monitoring Cyanobacterial Blooms in China's Large Lakes Based on MODIS from Both Terra and Aqua Satellites with a Novel Automatic Approach. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103830. [CrossRef]
24. Sun, D.; Li, Y.; Wang, Q.; Le, C.; Lv, H.; Huang, C.; Gong, S. Specific Inherent Optical Quantities of Complex Turbid Inland Waters, from the Perspective of Water Classification. *Photochem. Photobiol. Sci.* **2012**, *11*, 1299–1312. [CrossRef] [PubMed]
25. Shi, K.; Zhang, Y.; Qin, B.; Zhou, B. Remote Sensing of Cyanobacterial Blooms in Inland Waters: Present Knowledge and Future Challenges. *Sci. Bull.* **2019**, *64*, 1540–1556. [CrossRef] [PubMed]
26. Su, H.; Lu, X.; Chen, Z.; Zhang, H.; Lu, W.; Wu, W. Estimating Coastal Chlorophyll-a Concentration from Time-Series OLCI Data Based on Machine Learning. *Remote Sens.* **2021**, *13*, 576. [CrossRef]
27. Probst, P.; Wright, M.N.; Boulesteix, A.-L. Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
28. 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. [CrossRef]
29. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef] [PubMed]
30. Xi, L.; Qi, Z.; Feng, Y.; Cao, X.; Zou, J.; Han, J. An Explainable Machine Learning Approach (SHAP) to Assessing Desertification Risk and Its Drivers in the Ring-Tarim Basin, 1990–2020. *Environ. Impact Assess. Rev.* **2026**, *118*, 108309. [CrossRef]
31. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.
32. Hu, Y.; Wu, C.; Meadows, M.E.; Feng, M. Pixel Level Spatial Variability Modeling Using SHAP Reveals the Relative Importance of Factors Influencing LST. *Environ. Monit. Assess.* **2023**, *195*, 407. [CrossRef] [PubMed]
33. Waldock, C.; Wegscheider, B.; Josi, D.; Calegari, B.B.; Brodersen, J.; Jardim de Queiroz, L.; Seehausen, O. Deconstructing the Geography of Human Impacts on Species' Natural Distribution. *Nat. Commun.* **2024**, *15*, 8852. [CrossRef] [PubMed]

34. Cheng, C.; Wei, Y.; Lv, G.; Xu, N. Remote Sensing Estimation of Chlorophyll-a Concentration in Taihu Lake Considering Spatial and Temporal Variations. *Environ. Monit. Assess.* **2019**, *191*, 84. [CrossRef] [PubMed]
35. Zhao, M.; Li, H.; Li, H.; Zhang, X.; Ding, X.; Gong, F. Assessment of GOCI-II Satellite Remote Sensing Products in Lake Taihu. *Front. Mar. Sci.* **2024**, *11*, 1488219. [CrossRef]
36. Woo Kim, Y.; Kim, T.; Shin, J.; Lee, D.-S.; Park, Y.-S.; Kim, Y.; Cha, Y. Validity Evaluation of a Machine-Learning Model for Chlorophyll a Retrieval Using Sentinel-2 from Inland and Coastal Waters. *Ecol. Indic.* **2022**, *137*, 108737. [CrossRef]
37. Li, H.; Somogyi, B.; Chen, X.; Luo, Z.; Blix, K.; Wu, S.; Duan, Z.; Tóth, V.R. Leveraging Landsat and Google Earth Engine for Long-Term Chlorophyll-a Monitoring: A Case Study of Lake Balaton's Water Quality. *Ecol. Inform.* **2025**, *90*, 103245. [CrossRef]
38. Bresciani, M.; Cazzaniga, I.; Austoni, M.; Sforzi, T.; Buzzi, F.; Morabito, G.; Giardino, C. Mapping Phytoplankton Blooms in Deep Subalpine Lakes from Sentinel-2A and Landsat-8. *Hydrobiologia* **2018**, *824*, 197–214. [CrossRef]
39. Yadav, S.; Yamashiki, Y.; Susaki, J.; Yamashita, Y.; Ishikawa, K. Chlorophyll Estimation of Lake Water and Coastal Water Using Landsat-8 and Sentinel-2a Satellite. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *XLII-3/W7*, 77–82. [CrossRef]
40. Ansper, A.; Alikas, K. Retrieval of Chlorophyll a from Sentinel-2 MSI Data for the European Union Water Framework Directive Reporting Purposes. *Remote Sens.* **2018**, *11*, 64. [CrossRef]
41. Zhang, S.; A, Y.; Wang, L.; Wang, Y.; Zhang, X.; Zhu, Y.; Ma, G. Monitoring of Low Chl-a Concentration in Hulun Lake Based on Fusion of Remote Sensing Satellite and Ground Observation Data. *Remote Sens.* **2024**, *16*, 1811. [CrossRef]
42. Sun, X.; Zhang, Y.; Shi, K.; Zhang, Y.; Li, N.; Wang, W.; Huang, X.; Qin, B. Monitoring Water Quality Using Proximal Remote Sensing Technology. *Sci. Total Environ.* **2022**, *803*, 149805. [CrossRef] [PubMed]
43. Deng, X.; Xu, Y.; Han, L.; Song, S.; Yang, L.; Li, G.; Wang, Y. Impacts of Urbanization on River Systems in the Taihu Region, China. *Water* **2015**, *7*, 1340–1358. [CrossRef]
44. Qin, B.; Zhu, G.; Gao, G.; Zhang, Y.; Li, W.; Paerl, H.W.; Carmichael, W.W. A Drinking Water Crisis in Lake Taihu, China: Linkage to Climatic Variability and Lake Management. *Environ. Manag.* **2010**, *45*, 105–112. [CrossRef] [PubMed]
45. Zhang, Y.; Zhang, Y.; Shi, K.; Zhou, Y.; Li, N. Remote Sensing Estimation of Water Clarity for Various Lakes in China. *Water Res.* **2021**, *192*, 116844. [CrossRef] [PubMed]
46. Ahn, J.-H.; Kim, K.-S.; Lee, E.-K.; Bae, S.-J.; Lee, K.-S.; Moon, J.-E.; Han, T.-H.; Park, Y.-J. Introduction of GOCI-II Atmospheric Correction Algorithm and Its Initial Validations. *Korean J. Remote Sens.* **2021**, *37*, 1259–1268. [CrossRef]
47. Gordon, H.R.; Wang, M. Retrieval of Water-Leaving Radiance and Aerosol Optical Thickness over the Oceans with SeaWiFS: A Preliminary Algorithm. *Appl. Opt.* **1994**, *33*, 443. [CrossRef] [PubMed]
48. Ahn, J.-H.; Park, Y.-J.; Kim, W.; Lee, B. Simple Aerosol Correction Technique Based on the Spectral Relationships of the Aerosol Multiple-Scattering Reflectances for Atmospheric Correction over the Oceans. *Opt. Express* **2016**, *24*, 29659. [CrossRef] [PubMed]
49. Hu, C.; Lee, Z.; Ma, R.; Yu, K.; Li, D.; Shang, S. Moderate Resolution Imaging Spectroradiometer (MODIS) Observations of Cyanobacteria Blooms in Taihu Lake, China. *J. Geophys. Res. Oceans* **2010**, *115*, 2009JC005511. [CrossRef]
50. Le, C.; Hu, C.; English, D.; Cannizzaro, J.; Chen, Z.; Feng, L.; Boler, R.; Kovach, C. Towards a Long-Term Chlorophyll-a Data Record in a Turbid Estuary Using MODIS Observations. *Prog. Oceanogr.* **2013**, *109*, 90–103. [CrossRef]
51. Wang, M.; Jiang, L. Recovery of Pixels with Extremely Turbid Waters and Intensive Floating Algae from False Cloud Masking in Satellite Ocean Color Remote Sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *137*, 104408. [CrossRef]
52. Zhao, M.; Bai, Y.; Li, H.; He, X.; Gong, F.; Li, T. Fluorescence Line Height Extraction Algorithm for the Geostationary Ocean Color Imager. *Remote Sens.* **2022**, *14*, 2511. [CrossRef]
53. Mishra, S.; Mishra, D.R. Normalized Difference Chlorophyll Index: A Novel Model for Remote Estimation of Chlorophyll-*a* Concentration in Turbid Productive Waters. *Remote Sens. Environ.* **2012**, *117*, 394–406. [CrossRef]
54. Gurlin, D.; Gitelson, A.A.; Moses, W.J. Remote Estimation of Chl-*a* Concentration in Turbid Productive Waters—Return to a Simple Two-Band NIR-Red Model? *Remote Sens. Environ.* **2011**, *115*, 3479–3490. [CrossRef]
55. Son, Y.B.; Min, J.-E.; Ryu, J.-H. Detecting Massive Green Algae (*Ulva prolifera*) Blooms in the Yellow Sea and East China Sea Using Geostationary Ocean Color Imager (GOCI) Data. *Ocean Sci. J.* **2012**, *47*, 359–375. [CrossRef]
56. Huang, C.; Shi, K.; Yang, H.; Li, Y.; Zhu, A.; Sun, D.; Xu, L.; Zou, J.; Chen, X. Satellite Observation of Hourly Dynamic Characteristics of Algae with Geostationary Ocean Color Imager (GOCI) Data in Lake Taihu. *Remote Sens. Environ.* **2015**, *159*, 278–287. [CrossRef]
57. Taswell, C. The What, How, and Why of Wavelet Shrinkage Denoising. *Comput. Sci. Eng.* **2000**, *2*, 12–19. [CrossRef]
58. Md Yusof, M.R.; Ariffin, A.K.; Ismail, M. Wavelet Denoising as Signal Feature-Dependent Kernel Convolution. *J. Phys. Conf. Ser.* **2019**, *1185*, 012026. [CrossRef]
59. Halidou, A.; Mohamadou, Y.; Ari, A.A.A.; Zacko, E.J.G. Review of Wavelet Denoising Algorithms. *Multimed. Tools Appl.* **2023**, *82*, 41539–41569. [CrossRef]
60. Mallat, S.G. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. [CrossRef]
61. 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. [CrossRef]

62. Donoho, D.L.; Johnstone, I.M. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika* **1994**, *81*, 425–455. [CrossRef]
63. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
64. 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. [CrossRef]
65. Mirjalili, S.; Lewis, A. The Whale Optimization Algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [CrossRef]
66. 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. [CrossRef]
67. 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. [CrossRef]
68. 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. [CrossRef] [PubMed]
69. Li, Z. GeoShapley: A Game Theory Approach to Measuring Spatial Effects in Machine Learning Models. *Ann. Am. Assoc. Geogr.* **2024**, *114*, 1365–1385. [CrossRef]
70. Zheng, B.-H.; Xu, Q.-J.; Zhu, Y.-Z. [Primary Study on Enacting the Lake Nutrient Control Standard]. *Huan Jing KE Xue* **2009**, *30*, 2497–2501. [PubMed]
71. 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. [CrossRef]
72. Jing, Y.; Feng, C.; Chen, T.; Zhu, Y.; Li, C.; Tao, B.; Song, Q. Use of GOCI-II Images for Detection of Harmful Algal Blooms in the East China Sea. *Geosci. Lett.* **2024**, *11*, 2. [CrossRef]
73. Zhang, C.; Tao, B.; Li, Y.; Ai, L.; Zhu, Y.; Liang, L.; Huang, H.; Li, C. Evaluation of Rayleigh-Corrected Reflectance on Remote Detection of Algal Blooms in Optically Complex Coasts of East China Sea. *Remote Sens.* **2024**, *16*, 2304. [CrossRef]
74. Zhao, M.; Li, H.; Zhang, X.; Ding, X.; Gong, F. A Novel Method for Distinguishing Different Algal Blooms in the East China Sea via Geostationary Ocean Color Satellite Observations. *Front. Mar. Sci.* **2025**, *12*, 1549711. [CrossRef]
75. Wu, T.; Qin, B.; Zhu, G.; Luo, L.; Ding, Y.; Bian, G. Dynamics of Cyanobacterial Bloom Formation during Short-Term Hydrodynamic Fluctuation in a Large Shallow, Eutrophic, and Wind-Exposed Lake Taihu, China. *Environ. Sci. Pollut. Res.* **2013**, *20*, 8546–8556. [CrossRef] [PubMed]
76. Liu, L.; Dong, Y.; Kong, M.; Zhou, J.; Zhao, H.; Wang, Y.; Zhang, M.; Wang, Z. Towards the Comprehensive Water Quality Control in Lake Taihu: Correlating Chlorphyll a and Water Quality Parameters with Generalized Additive Model. *Sci. Total Environ.* **2020**, *705*, 135993. [CrossRef] [PubMed]
77. Peng, J.; Chen, J.; Liu, S.; Liu, T.; Deng, F.; Fan, Y.; De Maeyer, P. Dynamics of the Risk of Algal Blooms Induced by Surface Water Temperature in an Alpine Eutrophic Lake under Climate Warming: Insights from Lake Dianchi. *J. Hydrol.* **2024**, *643*, 131949. [CrossRef]
78. Yang, Y.; Huang, C.; Li, Y.; Li, J.; Lyu, H.; Dong, X.; Zhu, Y.; Wang, G.; Li, J.; Wang, X. Multi-Scale Analysis of Thermal Forcing of Heatwaves on Cyanobacterial Blooms in Lake Taihu: Mechanisms, Spatial Heterogeneity, and Future Projections. *Environ. Res.* **2026**, *306*, 125185. [CrossRef] [PubMed]
79. Luo, X.; Li, N.; Zhang, Y.; An, S.; Zhang, Y.; Shi, K.; Zhu, G.; Qin, B. Summer Heatwaves Increased Surface Dissolved Oxygen through Algal Photosynthetic Compensation in a Eutrophic Lake. *Ecol. Indic.* **2026**, *185*, 114763. [CrossRef]
80. Paerl, H.W.; Xu, H.; McCarthy, M.J.; Zhu, G.; Qin, B.; Li, Y.; Gardner, W.S. Controlling Harmful Cyanobacterial Blooms in a Hyper-Eutrophic Lake (Lake Taihu, China): The Need for a Dual Nutrient (N & P) Management Strategy. *Water Res.* **2011**, *45*, 1973–1983. [CrossRef] [PubMed]
81. Wu, C.; Xie, M.; Lin, L.; He, S.; Luo, C.; Dong, H. A Study on Spatiotemporal Downscaling Methods for Chlorophyll-a Concentration in Taihu Lake Based on Remote Sensing Data from Sentinel-2 MSI and COMS-1 GOCI. *Water* **2025**, *17*, 855. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.