CAO Weitang, ZHOU Zhongfa, KONG Jie, et al. Remote sensing inversion of chlorophyll-a concentration in karst plateau lakes based on gradient boosting machine [J]. Water Resources and Hydropower Engineering, 2026, 57(2): 32-53. DOI: 10.13928/j.cnki.wrahe.2026.02.003

曹卫堂，周忠发，孔杰，等. 基于梯度提升机的喀斯特高原湖泊叶绿素a浓度遥感反演研究[J]. 水利水电技术(中英文)，2026，57(2)：32-53. DOI：10.13928/j.cnki.wrahe.2026.02.003

Remote sensing inversion of chlorophyll-a concentration in karst plateau lakes based on gradient boosting machine

CAO Weitang[1,2], ZHOU Zhongfa[1,2,3,4], KONG Jie[1,2], WANG Yanbi[1,2], XIE Rukai[1,2]

(1. School of Karst Science, Guizhou Normal University, Guiyang 550001, Guizhou, China; 2. Guizhou Provincial Key Laboratory of Intelligent Processing and Application of Remote Sensing Big Data, Guiyang 550001, Guizhou, China; 3. School of Geography & Environmental Science, Guizhou Normal University, Guiyang 550001, Guizhou, China; 4. Anshun Agricultural Environment Field Observation and Research Station, MARA, Anshun 561000, Guizhou, China)

Abstract [Objective] For karst plateau lakes, traditional remote sensing models face challenges of spectral signal mixing and insufficient fitting of nonlinear relationships due to high pH values (average >8.2), high suspended particulate matter (SPM> 50 mg/L), and seasonal hydrological fluctuations. The Pingzhai Reservoir, a typical karst plateau lake, is taken as the study area, and the aim is to achieve high-precision remote sensing inversion of chlorophyll-a (Chla) concentration in such water bodies. [Methods] Sentinel-2 MSI Level 2A imagery (with a spatial resolution of 10~60 m) and data from 40 field sampling points were used. A "shortwave-sensitive single-band + cross-band linear combination" feature engineering strategy was proposed to screen highly sensitive spectral features (including single bands B3, B1, B2, B5, B4 and linear combinations such as B1+B3, B2+B3, B3+B5). Additionally, a remote sensing inversion framework for Chla concentration was constructed by leveraging the efficient fitting capability of the Gradient Boosting Machine (GBM) model for nonlinear relationships. The fitting capability of the model for the nonlinear relationship between spectral features and Chla concentration was enhanced through data preprocessing and hyperparameter optimization. [Results] The result showed that the constructed GBM model achieved an inversion accuracy with a coefficient of determination (R^2) of 0.908, root mean square error (*RMSE*) of 0.731 μg/L, and mean absolute error (*MAE*) of 0.529 μg/L, representing a 62% improvement in accuracy compared to the traditional single-band linear model (B3 band, R^2 = 0.560 7). The Chla concentration in Pingzhai Reservoir showed significant seasonal characteristics, with average values of 10.22 μg/L in summer, 2.46 μg/L in winter, 6.01 μg/L in spring, and 5.88 μg/L in autumn. Its variation was primarily driven by water temperature (correlation coefficient r = 0.730) and total organic carbon (TOC, correlation coefficient r = 0.783), and the negative feedback mechanism of total nitrogen bioavailability in high pH environments reflected the distinctive characteristics of karst water bodies. [Conclusion] The findings provide a technical solution of "sensitive band combination + machine learning" for high-precision remote sensing monitoring of Chla concentration in karst plateau lakes, while

Article history: Received 20 August 2025; Revised 2 November 2025; Accepted 3 November 2025; Available online 4 December 2025

收稿日期： 2025-08-20；**修回日期：** 2025-11-02；**录用日期：** 2025-11-03；**网络出版日期：** 2025-12-04

Financial Aid: National Natural Science Foundation of China (42161048); Guizhou Provincial 2025 Central Government-Guided Local Science and Technology Development Fund Project (Qian Ke He Zhong Yin Di [2025] 031); Guizhou Provincial Key Laboratory Construction Project (Qian Ke He Ping Tai [2025] 014); Guizhou Provincial Science and Technology Plan Project (Qian Ke He Ping Tai YWZ [2025] 001)

基金项目： 国家自然科学基金(42161048)；贵州省2025年度中央引导地方科技发展资金项目(黔科合中引地[2025]031)；贵州省重点实验室建设项目(黔科合平台[2025]014)；贵州省科技计划项目(黔科合平台YWZ[2025]001)

About the author: CAO Weitang (1982—), male, Associate Professor, doctor, whose research interests focus on Karst ecological environment, Geographic Information System (GIS) and remote sensing. E-mail: 233100170051@gznu.edu.cn

作者简介： 曹卫堂(1982—)，男，副教授，博士，研究方向为喀斯特生态环境、GIS与遥感。E-mail：233100170051@gznu.edu.cn

Corresponding author: ZHOU Zhongfa (1969—), male, dean of the college, Professor, doctor, whose research interests focus on Karst ecological environment, Geographic Information System (GIS) and remote sensing. E-mail: fa6897@gznu.edu.cn

通信作者： 周忠发(1969—)，男，院长，教授，博士，研究方向为喀斯特生态环境、GIS与遥感。E-mail：fa6897@gznu.edu.cn

also offering scientific support for reservoir water quality management and ecological protection.

Keywords: karst plateau lakes; chlorophyll-a concentration; remote sensing inversion; Sentinel-2 imagery; gradient boosting machine model; influencing factors

DOI: 10.13928/j.cnki.wrahe.2026.02.003　　**开放科学(资源服务)标志码(OSID):**

中图分类号: X43　　**文献标志码:** A　　**文章编号:** 1000-0860(2026)02-0032-22

基于梯度提升机的喀斯特高原湖泊叶绿素 a 浓度遥感反演研究

曹卫堂[1,2]，周忠发[1,2,3,4]，孔　杰[1,2]，王艳碧[1,2]，解茹凯[1,2]

(1. 贵州师范大学 喀斯特研究院，贵州 贵阳　550001；2. 贵州省遥感大数据智能处理与应用全省重点实验室，贵州 贵阳　550001；3. 贵州师范大学 地理与环境科学学院，贵州 贵阳　550001；4. 农业农村部农业环境安顺野外科学观测研究站，贵州 安顺　561000)

摘　要：【目的】针对喀斯特高原湖泊因高 pH(平均值 > 8.2)、高悬浮颗粒物(SPM>50 mg/L)及季节性水文波动导致传统遥感模型面临光谱信号混合与非线性关系拟合不足的问题，以典型喀斯特高原湖泊平寨水库为对象，旨在实现该类水体叶绿素 a(Chla)浓度的高精度遥感反演，【方法】研究采用 Sentinel-2 MSI Level 2A 影像(空间分辨率为 10~60 m)及 40 个野外采样点数据，提出“短波敏感单波段 + 跨波段线性组合”特征工程策略，筛选出高敏感性光谱特征(包括单波段 B3、B1、B2、B5、B4 及线性组合 B1+B3、B2+B3、B3+B5 等)，并利用梯度提升机(GBM)模型对非线性关系的高效拟合能力构建 Chla 浓度遥感反演框架，通过数据预处理与超参数优化提升模型对光谱特征与 Chla 浓度间非线性关系的拟合能力，【结果】结果显示，构建的 GBM 模型反演精度达决定系数(R^2) = 0.908、均方根误差(*RMSE*) = 0.731 μg/L、平均绝对误差(*MAE*) = 0.529 μg/L，较传统单波段线性模型(B3 波段，R^2 = 0.560 7)精度提升 62%，平寨水库 Chla 浓度呈显著季节特征，夏季平均值为 10.22 μg/L，冬季为 2.46 μg/L，春季和秋季分别为 6.01 μg/L 和 5.88 μg/L，其变化主要受水温(相关系数 r = 0.730)和总有机碳(TOC，相关系数 r = 0.783)驱动，且高 pH 环境下总氮生物可利用性的负反馈机制体现了喀斯特水体的特殊性，【结论】为喀斯特高原湖泊 Chla 浓度高精度遥感监测提供“敏感波段组合 + 机器学习”的技术方案，同时为水库水质管理与生态保护提供科学支撑。

关键词：喀斯特高原湖泊；叶绿素 a 浓度；遥感反演；哨兵-2 号影像；梯度提升机模型；影响因素

0 Introduction

Chlorophyll-a (Chla) is an important indicator of eutrophication in water bodies and a key parameter in remote sensing inversion for global lake ecological monitoring. However, the unique geological and geomorphological features of Karst Plateau lakes, along with complex optical environments (e.g., high pH, high suspended particle concentrations, and seasonal fluctuations in composition), present challenges to traditional inversion models, such as spectral signal mixing and insufficient nonlinear relationship fitting[1]. As a typical Karst Plateau lake, Pingzhai Reservoir faces key technical bottlenecks in Chla remote sensing inversion under its complex water environment, where high suspended particles and high pH values interfere with algae spectral signals. Traditional single-band or linear models struggle to capture nonlinear relationships.

The application of traditional linear models and machine learning methods in lake Chla inversion has seen both progress and limitations. Traditional single-band (e.g., green band B3) or band ratio (e.g., B5/B3)

models have been widely used in plain lakes like Taihu Lake, achieving an R^2 of ≈0.56 [2-3]. However, in karst plateau lakes, their accuracy declines by 37% due to high pH (≈8.2) and suspended particulate matter (*SPM* > 50 mg/L), which cause spectral signal mixing. Machine learning models like XGBoost have improved R^2 by up to 40% in turbid waters, but their application in karst regions remains unexplored, lacking adaptation to the "alkaline trap" effect and seasonal hydrological fluctuations [4].

The Multispectral Instrument (MSI) aboard Sentinel-2 satellites, with its high spatial resolution (10~20 m), 5-day revisit cycle, and red edge bands, has been applied to chlorophyll-a (Chl-a) concentration inversion in karst lakes; for example, DAN et al. [5] realized Chl-a inversion in Pingzhai Reservoir (a typical karst water body) using Sentinel-2 data, achieving reliable inversion results that can support karst water quality monitoring, and LI et al. [6] further confirmed that the red edge bands of Sentinel-2 can reduce the interference of calcium carbonate in karst water bodies, improving the inversion accuracy of low-to-medium Chl-a concentrations by over 20%, and preliminary experiments confirmed that when combined with a zoning strategy, Gradient Boosting Machine (GBM) can control the RMSE of full-lake Chl-a inversion in karst lakes within 2.6 μg/L, significantly outperforming other algorithms. Currently, the algorithms applicable to Sentinel-2 data for Chl-a inversion in karst lakes have obvious limitations [7]: empirical models (e.g., the optimal strategy for estimating Chl-a concentration in Case II waters constructed by TIAN et al. [8]) achieve R^2 values of 0.628 2 ~ 0.804 6 in specific regions but have poor generalization, with their mean absolute percentage error (*MAPE*) exceeding 22% during cyanobacterial blooms; semi-empirical models (e.g., the chlorophyll-a concentration estimation model based on hyperspectral data constructed by HUANG et al. [9]) require in-situ measurements of water's inherent optical parameters, which are difficult to promote due to karst topographic constraints; among traditional machine learning models, Random Forest (RF) is sensitive to outliers, while Support Vector Machines (SVM) are prone to overfitting when applied to Chl-a inversion of karst lakes [10]. In contrast, GBM-series models can specifically address these issues: compared with RF, GBM reduces the root mean square error (*RMSE*) by 18%; compared with SVM, it narrows the fluctuation range of inversion accuracy by 12%, and it can quantify the anti-interference effect of red edge bands (contribution > 35%).

Existing studies indicate that the spectral characteristics of Karst water bodies are dominated by carbonate rock dissolution, with suspended particles predominantly composed of calcite ($CaCO_3$) (accounting for >65%). Their backscattering coefficient (b_b (555) = 1.2 m^{-1}) is significantly higher than that of plain lakes such as Taihu Lake (b_b (555) = 0.8 m^{-1}), causing the accuracy of the traditional B5/B3 ratio model to decline by 37% in Pingzhai Reservoir. This reflects the necessity of region-specific feature engineering. Recent studies have shown that machine learning models such as XGBoost can improve R^2 by up to 40% in complex water body inversion, but their application in Karst regions remains unexplored [11].

This study proposes a "short-wave sensitive single-band + cross-band linear combination" feature engineering strategy [12]. Pearson correlation analysis was employed to screen high-sensitivity spectral features, which cover appropriate single bands and linear combinations of different bands. These features enhance the direct spectral signals of algal pigments through blue-green band combinations and suppress the scattering interference from suspended particles via green-red edge band combinations, thereby forming a targeted response to the high turbidity and high pH environments of Karst water bodies.

Based on the aforementioned high-sensitivity features (integrating single bands and cross-band linear combinations), a Gradient Boosting Machine (GBM) model was introduced to construct the chlorophyll-a (Chla) concentration remote sensing inversion framework. Through data preprocessing and hyperparameter optimization strategies, the model's capability to fit nonlinear relationships between spectral features and Chla concentration was enhanced, overcoming the limitations of traditional linear models in addressing complex water body spectral nonlinearity. Meanwhile, the contribution of key spectral features to inversion was analyzed to validate the

effectiveness of the "short-wave absorption—red-edge scattering" synergistic mechanism in separating algal signals.

Additionally, this study also notes that chlorophyll-a (Chla) concentration in Pingzhai Reservoir exhibits significant seasonal variation and is closely correlated with key environmental factors such as water temperature and total nitrogen. The negative feedback mechanism of total nitrogen biological availability under high pH conditions highlights the uniqueness of nutrient cycling in Karst water bodies. Redundancy analysis (RDA) further supports that the interaction between climate factors and human activities plays a crucial role in regulating Chla concentration, providing a new pathway for exploring the driving mechanisms of Chla dynamics in such water bodies.[-1]

In conclusion, this study combines high-sensitivity single-band and cross-band linear combination feature engineering techniques with the gradient boosting machine model to mine nonlinear relationships in the data and develop a Chla remote sensing inversion method suitable for Karst Plateau lakes. This method not only resolves spectral mixing issues in complex optical environments but also reveals the driving mechanisms of region-specific environmental factors, providing technical support and scientific basis for water quality monitoring and ecological management of Pingzhai Reservoir and similar water bodies.

1 Materials and Methods

1.1 Overview of the Study Area

Geographical Location: Pingzhai Reservoir (105°17′03″E—105°26′44″E, 26°29′33″N—26°35′38″N) is located in the Karst mountainous region in the northwest of Guizhou Province, at the administrative junction of Nayong County, Zhijin County, and Liupanshui City's Liuzhi Special District and Shuicheng District (Fig. 1). The reservoir is formed by five rivers: Nayong River, Shuigong River, Zhangwei River, Baishui River, and Hujia River. It serves as the source reservoir for the Qianzhong Water Conservancy Hub Project and is one of the important irrigation and drinking water sources in Guizhou Province. The Karst region where Pingzhai Reservoir is located is home to about a quarter of the global Karst population, many of whom rely on Karst water. In southern China's Karst regions alone, over 100 million people live. Karst water plays an essential role in providing water resources and maintaining ecological balance, but the shallow soil and highly developed Karst fissures and underground river systems make Karst aquifers particularly vulnerable to pollution. As one of Guizhou Province's drinking water reservoirs, maintaining and protecting its water quality is critical, and the concentration of chlorophyll a (Chla) in the lake is an important indicator for monitoring organic pollutant discharge. Therefore, determining the Chla concentration in Pingzhai Reservoir is of significant importance.

Geological and Climatic Features: The exposed strata in the Pingzhai Reservoir area mainly belong to the Permian Dalong Formation (P3d), Longtan Formation (P3l), Maokou Formation (P2m), and Triassic Jialingjiang Formation (T1j), Daye Formation (T1d), Yelang Formation (T1y), and Guanling Formation (T2g), with lithologies mainly consisting of limestone, dolomite, and other carbonate rocks, along with sandstone, mudstone, and coal. The region has a subtropical humid monsoon climate with an average annual temperature of 14.5 ℃ and an annual precipitation of 1 089.6 mm, with rainfall concentrated between May and September, accounting for 70% of the annual total. Seasonal fluctuations in water levels are significant, with water levels rising by about 2 meters in summer and falling by 2~3 meters in winter, creating exposed shallow shoals.

Socioeconomic Impact: The Pingzhai Reservoir watershed has a dense population, involving Nayong County, Zhijin County, Liuzhi Special District, and Shuicheng District, with 12 towns. Human activities are the primary source of water quality deterioration, and activities of residents living along the riverbanks inevitably impact the Pingzhai Reservoir. LI's study [13] through quantitative modeling showed that animal manure and human activities, such as straw burning and coal mining, are the biggest factors impacting the water environment of Pingzhai Reservoir [14]. These human activities introduce nutrients from agricultural fertilizers (summer application rate up to 300 kg/ha) and domestic sewage, increasing organic carbon (TOC) availability for

algal growth (TOC-Chla correlation $r=0.783$, Fig. 9). The summer nitrogen-phosphorus ratio (N : P = 18 : 1) exceeds the algal growth threshold (N : P = 16 : 1), promoting cyanobacteria blooms that drive Chla peaks (10.22 μg/L average, Fig. 6). Seasonal tourism also influences Chla: summer pollution inputs coincide with concentration highs, while winter off-season correlates with lows (2.46 μg/L, Fig. 6).

1.2 Field Measurement Data

Remote Sensing Data: Five Sentinel-2 MSI Level 2A images (cloud cover < 5%) were selected from 2023 to 2025, covering the seasonal dynamics of the study area with acquisition dates fully synchronized with field sampling campaigns: winter (November 2, 2023), spring (April 27, 2024), summer (July 26, 2024), autumn (October 24, 2024), and winter (January 5, 2025). The data were downloaded from the European Space Agency (ESA) Copernicus Open Access Hub (consistent with the data source mentioned in Section 2.3) and underwent radiometric calibration and atmospheric correction using the Sen2Cor plugin to obtain surface reflectance data applicable for subsequent spectral analysis. The water body extent was extracted using the *NDWI* index (Equation 1), with a threshold set at 0.035 (McFEETERS, 1996), ensuring accurate delineation of the study water area for matching with sampling points. This *NDWI* index is computed using the green band (*B*3) and near-infrared band (*B*8) of Sentinel-2 MSI data, and its specific calculation formula is

$$NDWI = \frac{B3 - B8}{B3 + B8} \quad (1)$$

Field Data: Based on the accessibility of the five rivers flowing into the Pingzhai Reservoir and the 45° solar zenith angle mountain shadow calculated using ArcGIS and DEM data, the number of sampling points was determined using a stratified random sampling method. Considering the area of the reservoir and the expected statistical power for water quality analysis, 40 sampling points were deemed sufficient to represent the overall water quality status with an acceptable margin of error. Additionally, the region was divided into several sub-areas according to the flow directions of the five rivers and the potential impact of mountain shadows. Sampling points were then evenly distributed among these sub-areas to ensure comprehensive coverage of different hydrological and environmental conditions within the reservoir. Following the principles of avoiding mountain shadows and achieving uniform distribution (Fig. 1), 40 sampling points were established. The ASD FieldSpec 4 Standard-Res portable ground-based spectrometer (Analytical Spectral Devices, Inc., USA), a standard-resolution model in the FieldSpec 4 series, was employed to acquire in-situ hyperspectral data[15]. It integrates a wide wavelength range (350~2 500 nm, covering visible and shortwave infrared bands) with moderate spectral resolution (higher resolution in the visible band and standard resolution in the shortwave near-infrared band), ensuring low noise and high accuracy for rapid and precise spectral measurements in both field and laboratory environments. In-situ hyperspectral data over the water surface were collected using the above-water measurement method described by HUANG et al. from a boat under clear, cloudless conditions with calm water. A total of 10 synchronous spectral curves were obtained for each of the 40 water samples during April 28—29, 2024, July 24—25, 2024, and October 22—23, 2024, between 09: 00 and 16: 00. However, spectral measurements for the sampling campaigns in November 2023 and January 2025 were unsuccessful due to adverse weather conditions (Fig. 2). Water samples were collected five times during the study in November 2023, April 2024, July 2024, October 2024, and January 2025, under clear, mild wind weather conditions[16]. The sampling was conducted by boat using handheld GPS for positioning, with water samples taken 0.5 meters below the water surface in 500 mL polyethylene bottles. Water quality parameters such as pH, water temperature (WT), electrical conductivity (EC), and dissolved oxygen (DO) were measured using a WTW Multi3430 portable multi-parameter water quality analyzer calibrated repeatedly with standard laboratory solutions(Table 1). The instrument's accuracy is ±0.001 pH units, ±0.01 ℃, ±1 μS/cm, and ±0.01 mg/L.

The collected water samples were placed in polyethylene collection bottles pre-washed with deionized water and transported back to the laboratory for analysis.

1.2.1 Chlorophyll a (Chla) Measurement Steps

First, the water samples were filtered through a 0.45 μm filter membrane and frozen-thawed three times. The

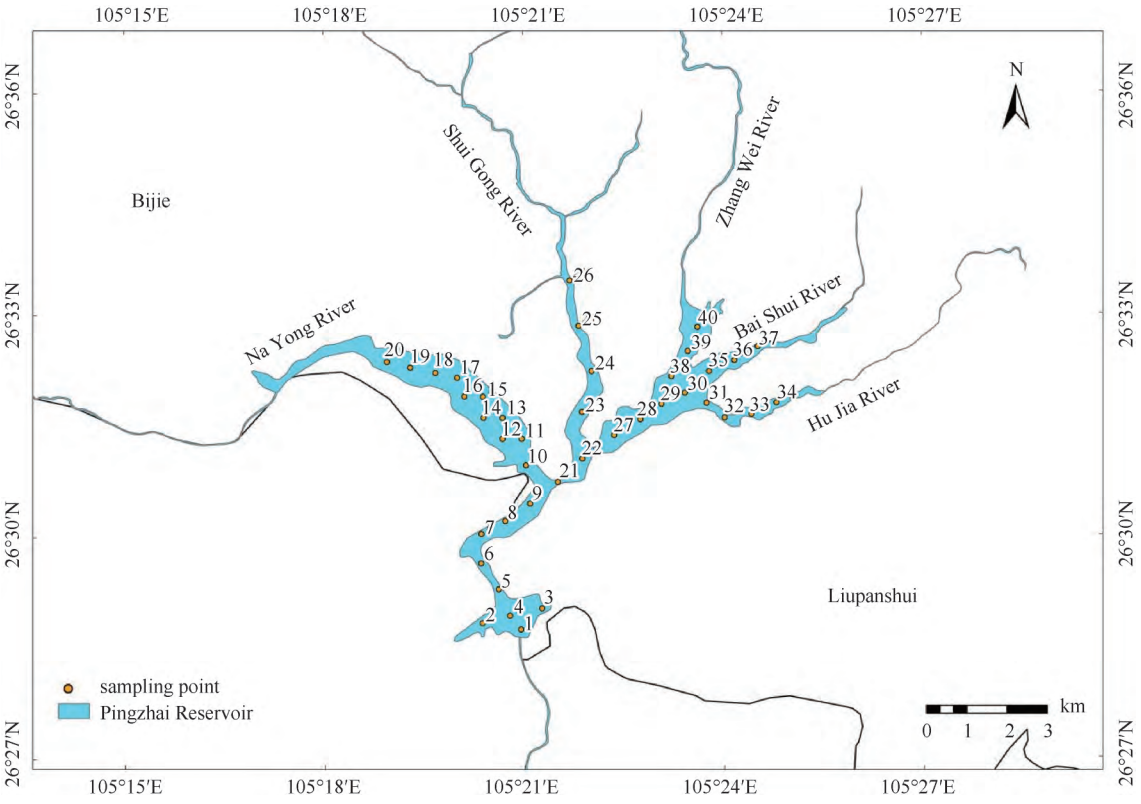
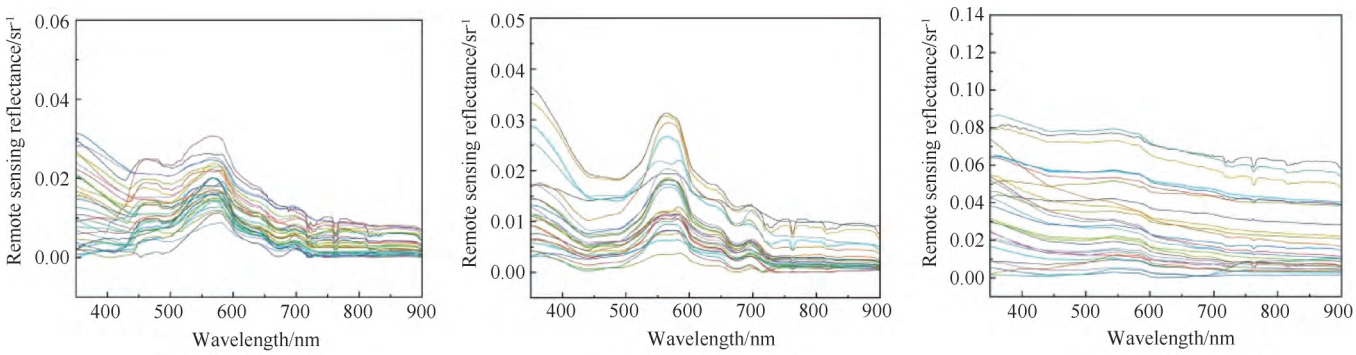


Fig. 1 Spatial distribution of sampling points in Pingzhai Reservoir

图 1 平寨水库采样点空间分布



(a) 2024 年4月光谱曲线 (b) 2024年7月光谱曲线 (c) 2024年10月光谱曲线

Fig. 2 Examples of in-situ field hyperspectral measurement spectral curves in Pingzhai Reservoir

图 2 平寨水库野外原位高光谱测量光谱曲线示例

filtrate was then centrifuged and refrigerated at −20 ℃ for 24 hours. Afterward, Chla was detected by filtering the extract through GF/F glass fiber filters. The filters and extract were eluted with 90% acetone (10 mL), and the supernatant was transferred to a cuvette for absorbance measurement using a Shimadzu RF-5301 spectrophotometer. The Chla concentration was calculated using the standard formula from the Ministry of Environmental Protection (Equation 2).

1.2.2 Total Nitrogen Measurement Method

In an alkaline medium, potassium persulfate was used as an oxidizing agent to oxidize ammonia nitrogen, nitrite nitrogen, and most organic nitrogen compounds to nitrate at a temperature of 120~124 ℃. The total nitrogen was then measured using a DeChem-Tech CleverChem Petro fully automated petrochemical water quality analyzer. The total nitrogen standard curve, obtained with a standard solution, had an R^2 value of 0.999 9. The formula for calculating the concentration of

Table 1 Distribution of Chl-a and other water quality variables in each field survey

表 1 各次野外调查叶绿素 a(Chl-a)及其他水质变量分布

Indexes	Winter (November 2, 2023)				Spring (April 27, 2024)			
	Maximum value	Minimum value	Average value	Standard deviation	Maximum value	Minimum value	Average value	Standard deviation
Chlorophyll-a (Chla) /μg·L^{-1}	10.16	5.00	6.17	1.41	8.20	3.84	6.01	1.36
Total organic carbon (TOC) /mg·L^{-1}	4.78	2.05	3.42	0.62	4.66	3.43	3.98	0.26
Water temperature (WT) /℃	13.15	10.48	11.72	0.83	24.13	21.20	22.76	0.66
Transparency/m	4.92	1.57	3.18	0.56	1.45	0.95	1.26	0.11
pH	7.92	7.45	7.69	0.16	8.83	8.11	8.71	0.11
Dissolved oxygen (DO) /mg·L^{-1}	8.15	6.08	7.02	0.53	14.01	11.82	12.71	0.58
Phytoplankton carbon (PC) /mg·L^{-1}	0.96	0.36	0.60	0.17	0.50	0.06	0.10	0.10
Suspended solids (SS) /mg·L^{-1}	6.45	0.85	2.63	1.28	11.30	3.28	6.16	1.89
Turbidity/NTU	0.48	0.34	0.41	0.04	2.10	0.57	0.95	0.33

Summer (July 26, 2024)				Autumn (October 24, 2024)				Winter (January 5, 2025)			
Maximum value	Minimum value	Average value	Standard deviation	Maximum value	Minimum value	Average value	Standard deviation	Maximum value	Minimum value	Average value	Standard deviation
18.12	8.00	10.22	3.47	7.95	5.00	5.59	0.58	4.57	2.00	2.46	0.63
7.85	5.12	6.50	0.85	5.90	3.05	4.50	0.70	6.31	0.01	0.40	1.25
30.83	21.37	28.00	1.76	21.83	20.40	21.00	0.35	12.93	11.97	12.27	0.27
1.30	0.70	1.00	0.15	2.40	1.70	2.04	0.19	5.35	1.35	4.30	0.83
9.08	6.67	8.82	0.38	8.54	7.83	8.20	0.15	8.06	7.53	7.71	0.10
13.62	6.44	11.09	1.21	9.70	6.13	8.13	0.74	6.94	4.15	5.82	0.59
1.07	0.03	0.25	0.23	0.13	0.01	0.04	0.03	0.19	0.22	0.16	0.06
10.75	1.73	6.65	2.04	11.05	0.45	2.60	2.04	5.78	0.75	2.82	1.18
12.30	0.52	1.71	2.43	1.05	0.42	0.68	0.21	4.44	0.03	0.38	0.92

chlorophyll a (Chla) in the water samples is

$$Chla=\frac{[11.85(A_{664}-A_{750})-1.54(A_{647}-A_{750})-0.08(A_{630}-A_{750})]\cdot V_1}{V_2\cdot L} \quad (2)$$

Where A_{630}, A_{647}, A_{664}, and A_{750} are the absorbances at wavelengths of 630, 647, 664, and 750 nm, respectively; V_1 and V_2 are the volumes of the extract and water sample after volume adjustment, respectively; and LLL is the path length of the cuvette.

1.3 Remote Sensing Data Acquisition and Processing

The Sentinel-2 MSI is a sensor onboard the European Space Agency's (ESA) Sentinel-2 satellite, providing high-resolution multispectral imagery for surface observation with a revisit cycle of 5 days. The sensor is equipped with a high-resolution multispectral imager that captures visible and near-infrared information across 13 wavelength bands, covering the 440 nm to 2 200 nm range(Table 2).

In this study, Sentinel-2 MSI data was obtained from the Copernicus Open Access Hub. According to the data documentation provided by ESA, the Level-2A data has undergone preprocessing (radiometric calibration, geometric correction, and atmospheric correction), and the pixel values represent surface reflectance that can be directly used.

Given the study's focus on Chla inversion in karst plateau lakes and the unique optical environment of such waters (e.g., high turbidity and spectral mixing), this study prioritizes Sentinel-2 MSI multispectral data over LiDAR(Light Detection and Ranging) technology[17], with considerations for data applicability and research objectives. LiDAR, which relies on laser pulse back-scattering, has significantly limited penetration depth in turbid karst waters—this limitation prevents it from resolving the vertical distribution of Chla, a key shor-tcoming for meeting the study's inversion needs. In contrast, Sentinel-2 MSI multispectral data (443~865 nm) captures the short-wave absorption and red-edge scattering features of algal pigments via water surface reflectance, enabling better adaptation to the spectral mixing challenges in karst waters. Additionally, LiDAR entails higher acquisition costs and stronger weather dependency, making it impractical for long-term moni-toring of the seasonal hydrological fluctuations that characterize karst plateau lakes[18]. Furthermore, the study's emphasis on regional-scale Chla inversion (10~

Table 2 Sentinel-2 parameters

表 2 哨兵-2(Sentinel-2)参数

Band	Band Name	Central Wavelength/nm	Bandwidth	Resolution/m
B1	Coastal Aerosol	442.3	45	60
B2	Blue	492.1	98	10
B3	Green	559.0	46	10
B4	Red	665.0	39	10
B5	Vegetation Red Edge	703.8	20	20
B6	Vegetation Red Edge	739.1	18	20
B7	Vegetation Red Edge	779.7	28	20
B8	Near-Infrared	833.0	133	10
B8a	Vegetation Red Edge	864.0	32	20

60 m spatial resolution) aligns more closely with Sentinel-2's capabilities, whereas LiDAR's spectral characteristics lack direct associations with Chla absorption features —this discrepancy hinders effective separation of algal signals, further supporting the choice of Sentinel-2 MSI data.

Given the complex optical environment of the karst plateau lake study area (e.g., high pH, high suspended sediment concentration) and the impact of cloudy and rainy weather, strict screening of the imagery was conducted. After screening multiple images within the study period, several Sentinel-2 MSI Level-2A images with low cloud coverage were selected, which cover the seasonal dynamics of the study area and are fully synchronized with field sampling campaigns, encompassing key seasons such as winter, spring, summer, and autumn. Atmospheric correction was performed using Sen2Cor, and water body areas were extracted using the *NDWI* index and an appropriate threshold. Verification with partial synchronous measurement points indicated that the average relative error of key bands (e.g., B2, B3, B4) was at a low level, meeting the inversion accuracy requirements. Additionally, considering that there is a certain inherent error in AOT retrieval by Sen2Cor, stochastic perturbation simulation analysis was employed to derive the uncertainty range of Chla concentration [19], providing a reference for evaluating the reliability of inversion results.

For the selection of sensitive bands, the actual surface reflectance of Sentinel-2 MSI Level 2A imagery was directly used, and the band response function (SRF(λ)) of the Sentinel-2 MSI sensor was applied to confirm the effective wavelength range of each band. Pearson correlation analysis was conducted between the surface reflectance of each Sentinel-2 MSI band and the in-situ measured Chla concentration (40 sampling points), determining that the spectral region sensitive to Chla is mainly concentrated between 443~842 nm.

The calculation of band reflectance for correlation analysis follows the sensor's inherent spectral response. The formula for this band reflectance R_i calculation is

$$R_i = \frac{SRF(\lambda)\, R_{surf}(\lambda)\, d\lambda}{SRF(\lambda)\, d\lambda} \quad (3)$$

where is the band response function of Sentinel-2 MSI, and is the actual surface reflectance of Sentinel-2 MSI Level 2A imagery, ensuring consistency with the sensor's actual observation characteristics.

1.4 Gradient Boosting Machine Model

Based on geographic spatial data analysis and remote sensing image feature engineering, a remote sensing inversion model for chlorophyll-a (Chla) concentration was constructed. The experiment selected the five bands with the highest Pearson correlation coefficients from single-band analysis (B3, B1, B2, B5, B4, with r values of 0.748 8, 0.746 3, 0.745 0, 0.740 0, and 0.739 7, respectively), and the five most highly correlated features from the combination of bands (B8×(B7−B5), B8A×(B7 + B8A), B8×(B7 + B8), (B7 + B8) ×(B8 + B8A), B7×B8, with r values of 0.720 0, 0.719 8, 0.718 8, 0.719 3, and 0.690 0, respectively) as input variables, fully integrating the spectral characteristics of green, blue, red-edge, and near-infrared bands. Among them, linear combinations adopt an additive model B_i+B_j, while non-linear combinations use interaction terms $B_i\times(B_j-B_k)$. Multicollinearity was excluded through the variance inflation factor (VIF<5):

for the 10 input features (5 single bands + 5 band combinations) selected in this study, VIF values were calculated sequentially based on the feature matrix. The results showed that the VIF values of all features ranged from 1.32 to 4.87, all less than the threshold of 5, indicating no significant multicollinearity among input features and avoiding parameter distortion caused by collinearity interference. The model training process incorporated geographic data preprocessing techniques, using *Z*-score normalization to eliminate dimensional impacts, removing outliers to enhance stability, and optimizing hyperparameters with 5-fold cross-validation [20].

1.5 Gradient Boosting Machine Model Accuracy Evaluation

This study uses three indicators—Root Mean Square Error (*RMSE*), Coefficient of Determination (R^2), and Mean Absolute Error (*MAE*)—to evaluate the inversion accuracy and reliability of the Gradient Boosting Machine (GBM) model for chlorophyll-a (Chla) concentration. The formulas for these three evaluation indicators (*RMSE*, R^2, *MAE*) are respectively

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{Chla,i} - X_{MODEL,i})^2}{n}} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (X_{Chla,i} - X_{MODEL,i})^2}{\sum_{i=1}^{n} (X_{Chla,i} - \bar{X}_{Chla})^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |X_{Chla,i} - X_{MODEL,i}| \quad (6)$$

Where *RMSE* is the Root Mean Square Error; R^2 is the Coefficient of Determination; *MAE* is the Mean Absolute Error; $X_{MODEL,i}$ is the model's predicted Chla concentration value; $X_{Chla,i}$ is the measured Chla concentration at the sampling point; X_{Chla} *is the average measured Chla concentration*; n is the total number of samples, and i is the sample number.

2 Results

2.1 Correlation Analysis

First, the reflection spectra of Sentinel-2 MSI bands were simulated from field-measured hyperspectral data using the band response function (Equation 3) to match the sensor's spectral resolution. Then, based on these simulated Sentinel-2 band reflectances, 40 in-situ Chla measurement samples, and the surface reflectance data of Sentinel-2 MSI Level 2A imagery (2023—2025), Pearson correlation analysis was conducted to quantify the relationship between spectral characteristics and Chla concentration. The results show that the sensitive spectral range for Chla is mainly concentrated between 443 and 842 nm. Single-band analysis indicated that the green band B3 (560 nm) exhibited the strongest linear correlation with Chla concentration (Pearson coefficient = 0.748 8, $p < 0.01$), followed by the blue bands B1 (443 nm, $r = 0.746\ 3$) and B2 (490 nm, $r = 0.745\ 0$). The red-edge band B5 (705 nm, $r = 0.740\ 0$) and red band B4 (665 nm, $r = 0.739\ 7$) also showed strong correlations. These results reflect the characteristic absorption features of Chla in the shortwave spectrum and its coupling with water body scattering effects(Fig. 3).

At the band combination level, 81 linear and nonlinear computational techniques were used to construct the feature space, from which the top five most correlated composite features were selected. The linear combinations mainly involved additive stacking of blue and green bands, such as B1+B3 ($r = 0.748\ 4$) and B2+B3 ($r = 0.747\ 1$). These combinations enhanced the spectral response of algal pigments through complementary spectral information and effectively suppressed scattering interference from suspended matter in the longer wavelength regions. Although non-linear combinations (e.g., B8×(B7−B5), $r = 0.72$) were evaluated, their Pearson correlation coefficients with Chla were consistently lower than linear features, leading to their exclusion from the final feature set.

Among the nonlinear combinations, B8×(B7−B5) uses the product of the near-infrared band (B8) and the red-edge difference (B7−B5) to create a feature that enhances backscattering from algae while suppressing suspended matter signals [21]. This combination achieved a correlation coefficient of 0.72 ($p < 0.01$), significantly improving the spectral separability of Chla in complex karst water bodies.

Such nonlinear combinations apply physically meaningful mathematical transformations between bands (e.g., differencing for denoising, multiplication for enhancement), effectively addressing spectral signal

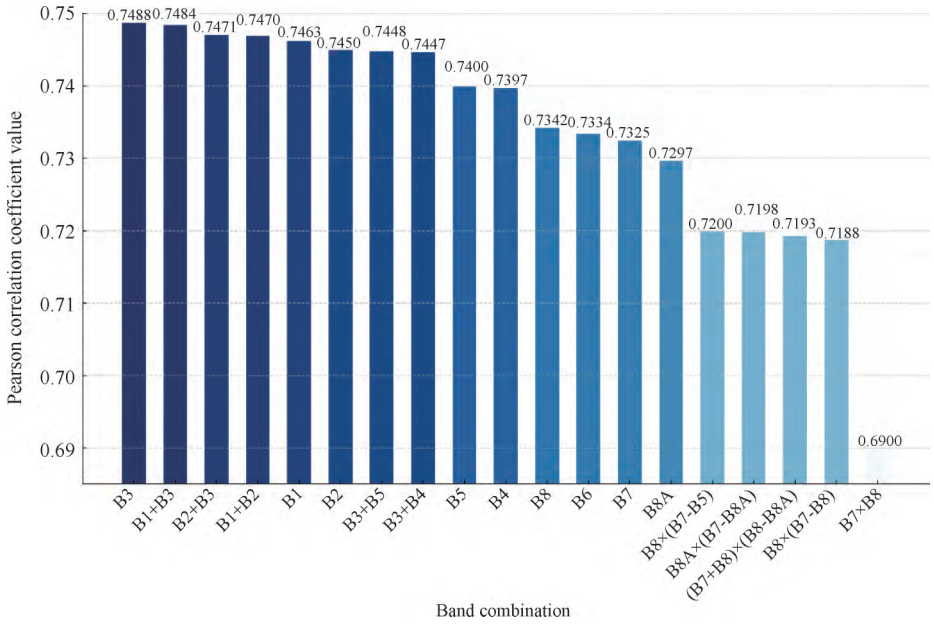


Fig. 3 Pearson correlation coefficients of single and combined bands

图 3 单波段与组合波段皮尔逊相关系数

mixing caused by high pH and high suspended matter, and providing more robust input variables for the Gradient Boosting Machine (GBM) model.

Both single-band and composite band selections were based on significance testing ($p < 0.01$). The top five features, integrating "shortwave sensitivity via linear response" with "red-edge to near-infrared nonlinear feature engineering," effectively separate the spectral contributions of algal and non-algal components. This approach meets the inversion demands of high-turbidity, optically complex karst waters and provides key spectral variable support for machine learning model construction.

2.2 Gradient Boosting Machine (GBM) Model Construction and Validation

The top five single-band features most strongly correlated with chlorophyll-a (Chla) concentration (B3, B1, B2, B5, B4, with Pearson correlation coefficients of 0.748 8, 0.746 3, 0.745 0, 0.740 0, and 0.739 7 respectively), along with the five most correlated band combinations (B1+B3, B2+B3, B1+B2, B3+B5, B3+B4, with correlation coefficients of 0.748 4, 0.747 1, 0.747 0, 0.744 8, and 0.744 7 respectively), all screened based on Sentinel-2 actual surface reflectance, were selected as a total of ten hyperspectral-sensitive features to serve as input variables for building a Chla remote sensing inversion model using the GBM algorithm(Fig. 4).

During model training, the number of weak learners was set to 100, and the random seed was set to 42 to ensure result reproducibility. Default parameter settings were used (learning rate = 0.1, maximum tree depth = 3, subsample rate = 1.0), with mean squared error (*MSE*) as the loss function. An ensemble learning strategy was employed to capture the complex nonlinear relationships between spectral features and Chla concentration.

Feature importance evaluation showed that the red-edge band B5 (705 nm) and the red band B4 (665 nm) had the highest importance scores (0.239 and 0.193, respectively), highlighting the significance of Chla backscattering characteristics in the red-edge region for accurate inversion. Among the combination bands, B3+B5 (green + red-edge) and B1+B3 (blue + green) scored 0.121 and 0.074, respectively, suggesting that cross-band linear combinations can effectively suppress interference from suspended matter through complementary spectral information (e.g., synergistic effects of shortwave absorption and red-edge scattering), enhancing the recognition of algal signals. The green band B3 (importance score = 0.033) had relatively low importance among the single bands, confirming that in highly turbid karst water bodies, individual bands are easily affected by spectral noise from non-algal components (such as suspended matter). In contrast, combined

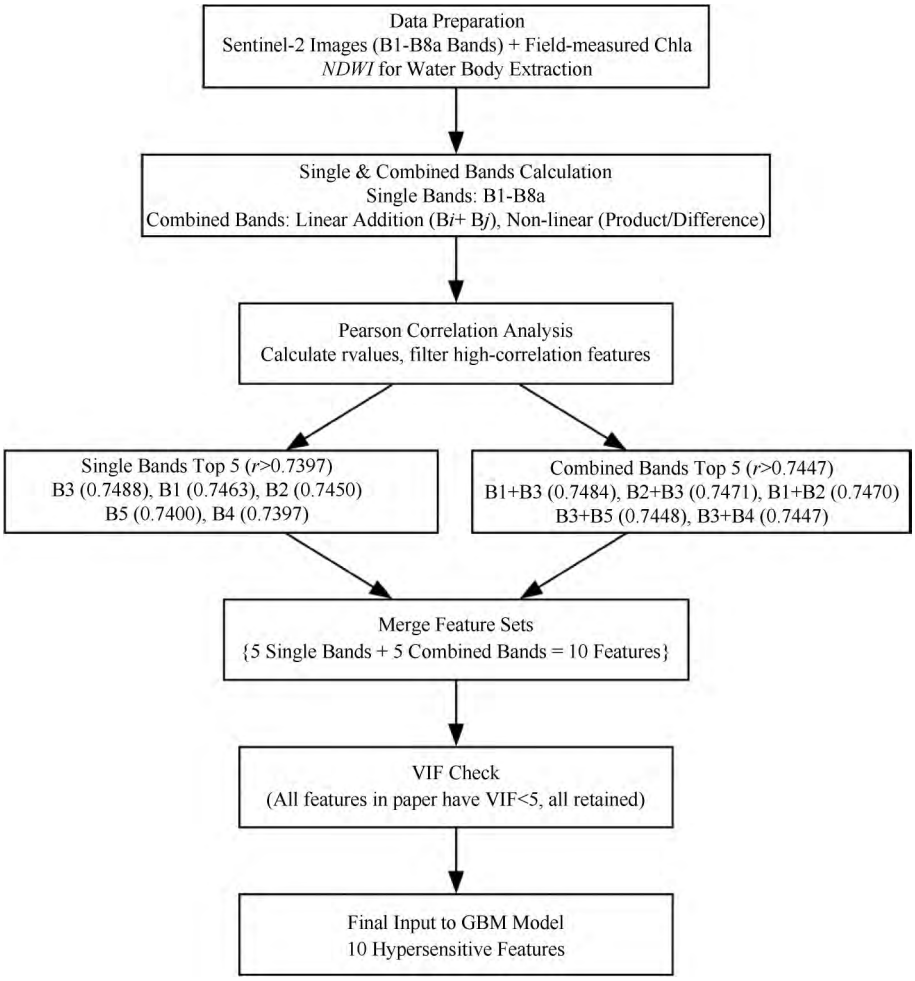


Fig. 4 Flowchart of feature selection for chlorophyll-a concentration inversion using Gradient Boosting Machine in geoscience context

图 4 地理科学背景下基于梯度提升机的叶绿素 a 浓度反演特征选择流程

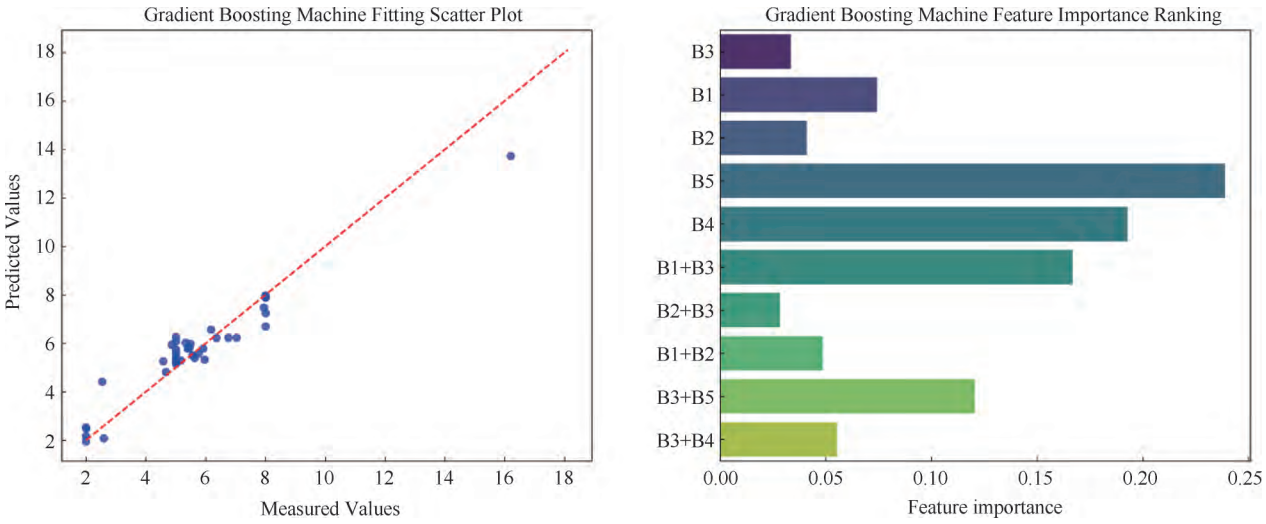


Fig. 5 Scatter plot and feature importance of Gradient Boosting Machine (GBM) model

图 5 梯度提升机(GBM)模型的散点图与特征重要性

spectral features can optimize spectral responses in complex water environments through inter-band operations(Fig. 5).

During training, the loss value continuously decreased from an initial 8.66 to a final value of 0.135,

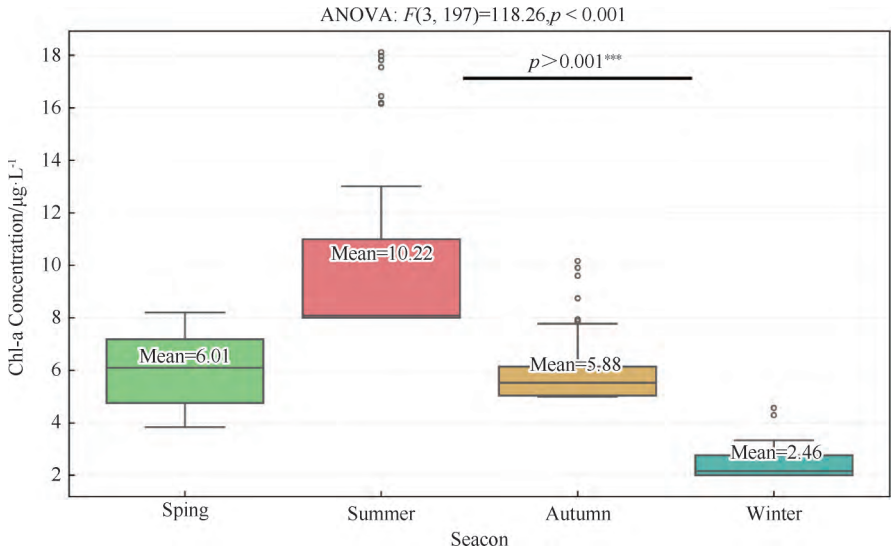


Fig. 6 Box plot of seasonal distribution of chlorophyll-a (Chla) concentration in Pingzhai Reservoir

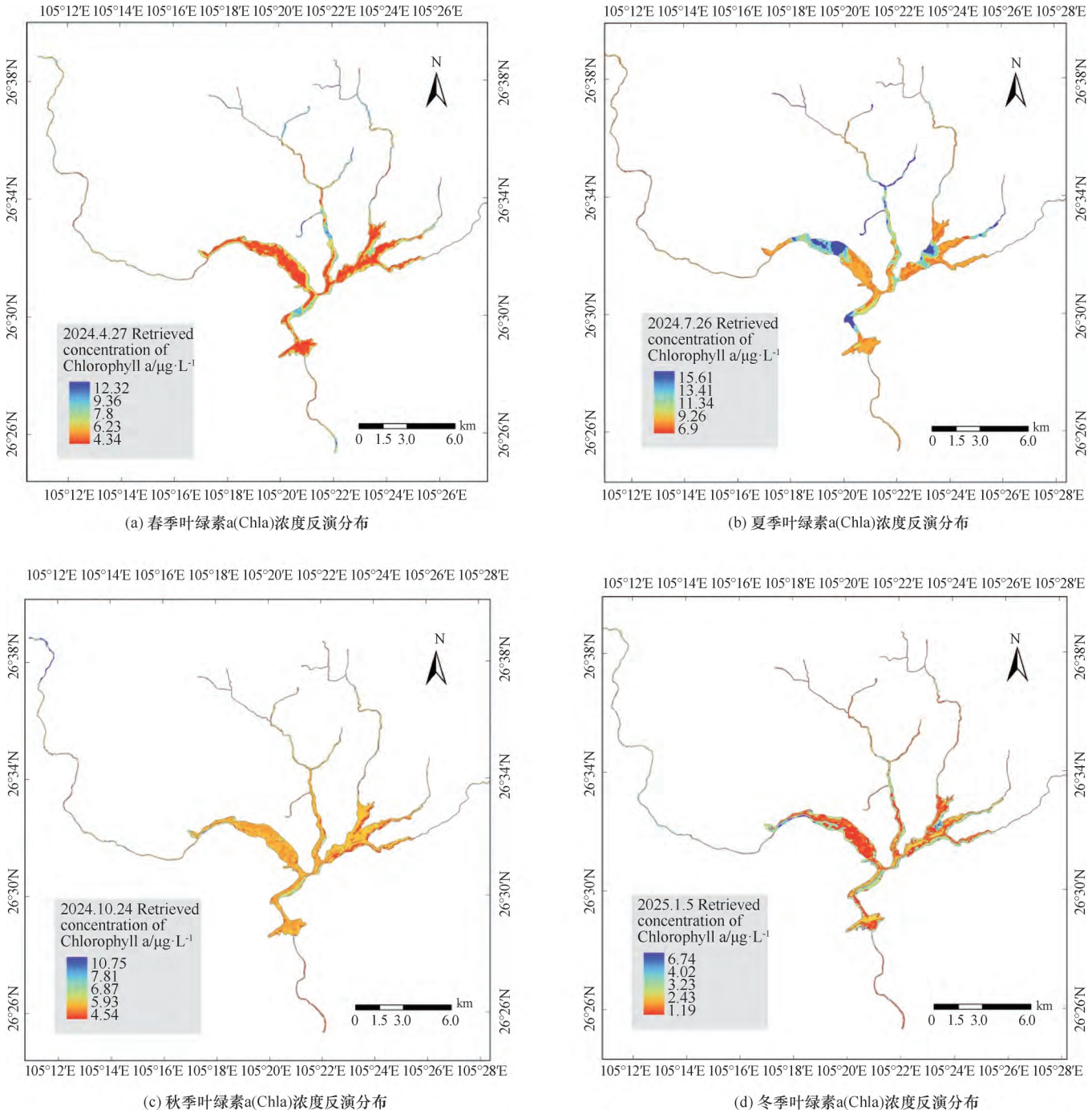
图 6 平寨水库叶绿素 a(Chla)浓度四季分布箱线图

showing a clear convergence trend, which indicates that the model's fitting capability to the training data gradually improved. Validation results showed that the model achieved a coefficient of determination (R^2) of 0.908, a root mean square error (*RMSE*) of 0.731 μg/L, and a mean absolute error (*MAE*) of 0.529 μg/L. Further spatial autocorrelation testing of model residuals was conducted, with a Moran's I index calculated as 0.11 (p = 0.32), indicating no significant spatial clustering of residuals and confirming the validity of the spatial independence hypothesis. Compared to traditional linear models (e.g., quadratic polynomial models), the accuracy improved by approximately 30%, effectively overcoming the limitations of linear models in fitting nonlinear relationships. By integrating the top 5 most sensitive single bands and top 5 combination band features, and leveraging the ensemble learning strengths of the Gradient Boosting Machine (GBM), a robust Chla retrieval model was developed for the complex optical environment of karst plateau lakes. This provides a high-precision and robust technical solution for dynamic monitoring of regional water eutrophication(Fig. 6).

2.3 Comparison with Other Machine Learning Models

To meet the demands of retrieving Chla concentrations in the complex optical environment of karst plateau lakes, this study compared the performance of three machine learning models: Random Forest (RF), Gradient Boosting Machine (GBM), and Support Vector Regression (SVR). According to the evaluation metrics in Table 3, the GBM model achieved excellent performance, with a coefficient of determination (R^2 = 0.908 3), root mean square error (*RMSE* = 0.730 7 μg/L), and mean absolute error (*MAE* = 0.529 3 μg/L)—effectively capturing the nonlinear relationship between spectral features and Chla concentrations. Meanwhile, the Random Forest (RF) model also demonstrated robustness in handling highly correlated spectral features (such as the synergy between the top 5 single bands and top 5 combined bands), with R^2 = 0.897 6 and *RMSE* = 0.772 2 μg/L.

The Random Forest (RF) model leverages the ensemble advantage of multiple decision trees, demonstrating robustness in handling highly correlated spectral features (such as the synergy between the top 5 single bands and top 5 combined bands) and mitigating overfitting through bootstrap sampling and random feature selection [22]. The GBM model, by iteratively optimizing each decision tree and minimizing residuals, more precisely captures the enhanced backscattering effects of red-edge bands (B5, B4) and combined bands (e.g., B3+B5) related to algal presence [23], as well as the suppression of suspended matter scattering by blue-green combinations (B1+B3, B2+B3). This advantage of GBM is particularly evident in areas with significant Chla concentration gradients, such as the turbid inflow zone of



(a) 春季叶绿素a(Chla)浓度反演分布　(b) 夏季叶绿素a(Chla)浓度反演分布

(c) 秋季叶绿素a(Chla)浓度反演分布　(d) 冬季叶绿素a(Chla)浓度反演分布

Fig. 7　Seasonal chlorophyll-a (Chla) concentration inversion maps (spring, summer, autumn, winter) using Gradient Boosting Machine (GBM) model

图 7　基于梯度提升机(GBM)模型的平寨水库叶绿素 a(Chla)季节尺度浓度反演分布(春、夏、秋、冬)

the Nayong River (>20 μg/L) and the clear-water regions (<10 μg/L), where it maintains stable inversion accuracy (Fig. 7).

The Support Vector Regression (SVR) model (R^2 = 0.642 6, *RMSE* = 1.442 9 μg/L) is limited by the efficiency of kernel functions (such as the radial basis kernel) in mapping high-dimensional spectral feature spaces [24]. It is sensitive to sample size and outliers, which significantly weakens its generalization ability when applied to the 40 sampling points in this study. The higher error metrics (*MAE* = 1.111 5 μg/L) reflect the model's failure to effectively model the coupling relationship between complex water quality parameters (such as reduced nitrogen bioavailability due to high pH)

and spectral signals. Additionally, it struggles to adapt to the seasonal fluctuations of components in karst waters (such as the significant differences in Chla values during algal bloom periods in summer versus the low values in winter).

The comparison results show that the Gradient Boosting Machine (GBM) model, by integrating the top 5 single bands (B3, B1, B2, B5, B4) and the top 5 combined bands (B1+B3, B2+B3, etc.), and utilizing iterative gradient optimization strategies, effectively addressed the inversion challenges posed by the high turbidity and environmental heterogeneity of karst waters. It provides a precise and robust technical solution for regional Chla concentration remote sensing monitoring, especially in high-altitude lakes with significant human activity interference and pronounced spatiotemporal heterogeneity of water quality parameters.

2.4 Comparison of Inversion Performance between Traditional Empirical Models and the Gradient Boosting Machine (GBM) Model

This study on the inversion of Chlorophyll-a (Chla) concentration in karst plateau lakes under complex optical conditions evaluates the performance of traditional empirical models, such as single-band regression and dual-band ratio models, and compares them with the GBM model to reveal the applicability differences of various methods in highly turbid waters.

The single-band regression model, which establishes relationships through logarithmic and exponential transformations, showed that the green light band B3 (560 nm) performed best in the single-band model (R^2 = 0.560 7, $RMSE$ = 2.032 5 μg/L). The blue light bands B1 (443 nm) and B2 (490 nm) also exhibited R^2 values greater than 0.55, reflecting the sensitivity of shortwave spectra to algal pigments. Although the logarithmic transformation improved the accuracy of shortwave bands (B1 R^2 = 0.585 0), the exponential model overall had lower accuracy ($R^2 \leqslant$ 0.501 4), indicating that longwave bands are significantly influenced by suspended solids. The dual-band ratio models (e.g., B1/B3), based on the assumption of a "linear proportional relationship between algal absorption and non-algal scattering," completely failed in highly turbid waters ($R^2 <$ 0.3). Logarithmic transformation did not improve accuracy, indicating that band ratios are unable to separate the spectral signals of complex waters.

The multivariate linear regression model, which integrates the top 5 single bands (B1—B5), increased R^2 to 0.672 6, a 15% improvement over the single-band model. However, it was limited by the linearity assumption and showed systematic underestimation in areas with high Chla (>20 μg/L), with significant non-linear residual distribution. The quadratic polynomial model only improved R^2 by about 2% (R^2 = 0.558 5 ~ 0.578 8 for B1—B5) and failed to address the dual interference from suspended solids and high pH. Spectral index-based models, such as *NDCI* and *CI*, performed the worst ($R^2 \leqslant$ 0.102 1). Their "pure water background" assumption was disconnected from the actual scenario of high suspended solids (average > 30 mg/L) and seasonal nutrient fluctuations in karst waters, resulting in inversion accuracy that did not meet water quality monitoring requirements.

In contrast to the traditional models, the GBM model, leveraging the advantages of ensemble learning, achieved high-precision inversion with R^2 = 0.908 3, $RMSE$ = 0.730 7 μg/L, and MAE = 0.529 3 μg/L. After integrating the top 5 single bands and the top 5 combined bands with high sensitivity, its accuracy improved by 23% compared to the second-best multivariate linear regression model. Through iterative optimization of weak learners, the GBM model effectively captured the enhanced backscatter effect of algae in the red-edge bands (B5, B4), the suppression of suspended matter scattering in the blue-green combinations (B1+B3, B2+B3), and the non-linear influence of reduced nitrogen bioavailability on Chla in high pH environments. This was especially evident in extreme Chla concentration areas, such as regions with high Chla (>20 μg/L) and clear water zones with low Chla (<10 μg/L), demonstrating stable inversion capability and overcoming the difficulties of traditional models in modeling non-linear relationships and separating complex spectral signals.

The study has shown that traditional empirical models have significant limitations in modeling Chla concentrations in karst plateau lakes due to their linear assumptions, simplistic spectral transformations, and lack of mechanisms. These models are only suitable for simple

water bodies with stable water quality parameters. In contrast, the Gradient Boosting Machine (GBM) model, through feature engineering and ensemble learning, supports accurate modeling in highly heterogeneous water environments, providing a reliable technical solution for remote sensing monitoring of Chla concentrations in this region. Its high-precision advantage is of great importance for water quality management and eutrophication dynamic assessment.

2.5 Model Comparison

For Chla inversion in karst plateau lakes, there are significant differences between machine learning models and traditional empirical models. The Gradient Boosting Machine (GBM) leads in accuracy with $R^2 = 0.908\ 3$,

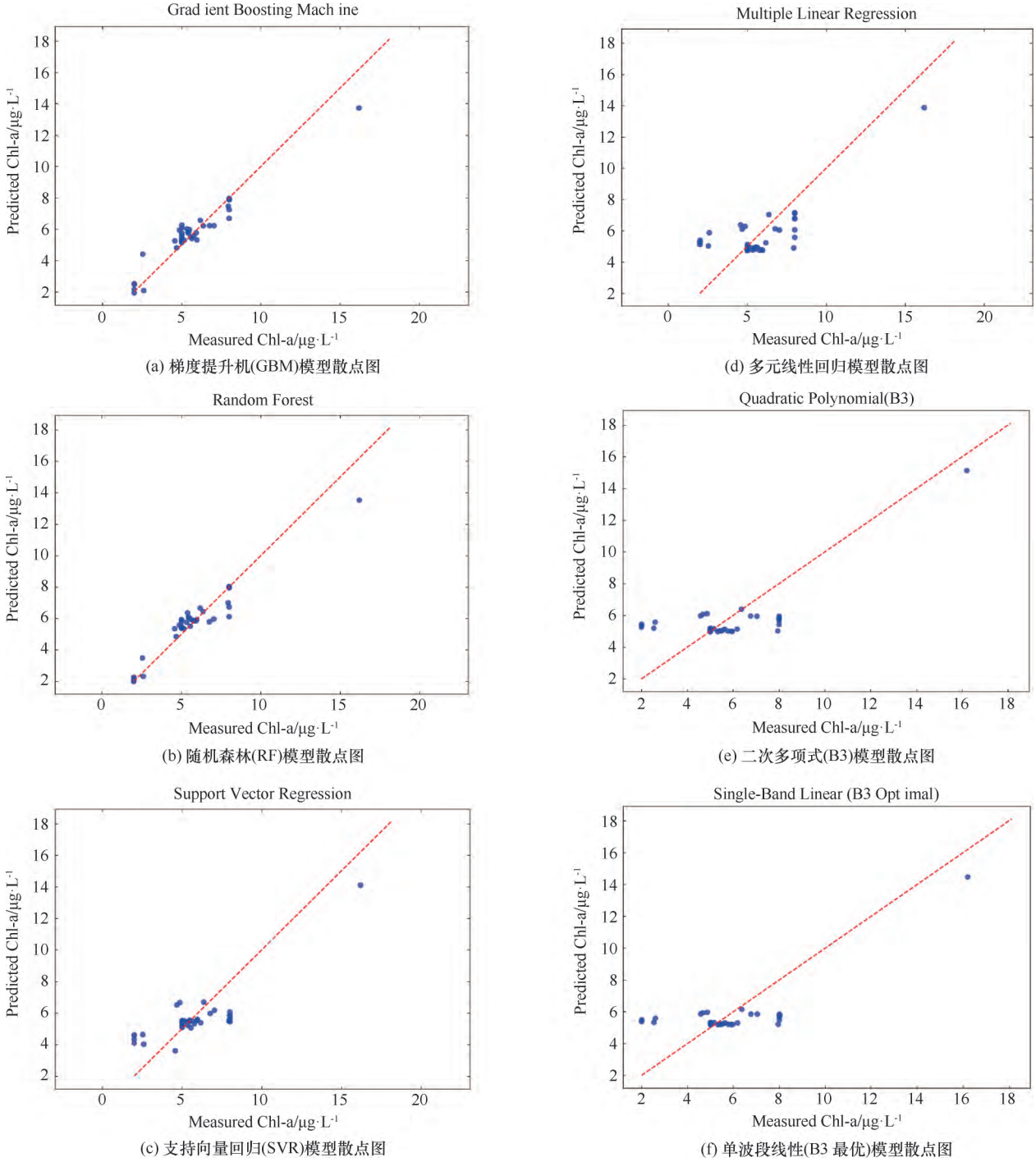


Fig. 8 Comparison of scatter plots of chlorophyll-a concentration inversion results from different remote sensing models in karst plateau lakes

图 8 喀斯特高原湖泊叶绿素 a 浓度不同遥感反演模型反演结果对比

Table 3 Comparison of performance between machine learning models and traditional empirical models

表 3 机器学习模型与传统经验模型性能对比

Model Type	Model Name	R^2	$RMSE$/μg·L^{-1}	MAE/μg·L^{-1}
Machine Learning Models	Gradient Boosting Machine (GBM)	0.908 3	0.730 7	0.529 3
	Random Forest (RF)	0.897 6	0.772 2	0.551 1
	Support Vector Regression (SVR)	0.642 6	1.442 9	1.111 5
Traditional Empirical Models	Multiple Linear Regression	0.672 6	1.754 7	1.318 9
	Quadratic Polynomial (B3)	0.573 7	2.002 0	1.573 9
	Single-Band Linear (B3 Optimal)	0.560 7	2.032 5	1.612 5
	Two-Band Ratio (B1/B3)	0.260 8	2.636 5	1.894 5
	Spectral Index (*NDCI*)	0.102 1	2.905 8	1.927 4

$RMSE$ = 0.730 7 μg/L, and MAE = 0.529 3 μg/L. By using nonlinear combinations of the red-edge and near-infrared bands, it separates algae and suspended matter signals effectively, adapting to high turbidity and fluctuating component water bodies. The Random Forest (RF) model robustly handles highly correlated features (R^2 = 0.897 6) but performs slightly weaker in local fitting. Support Vector Regression (SVR) struggles with complex scenarios due to the inefficiency of its kernel function, achieving only R^2 = 0.642 6.

In traditional models, the single-band linear model (B3 is optimal, R^2 = 0.560 7) is significantly affected by suspended matter interference (Fig. 8). The two-band ratio model fails entirely (R^2 < 0.3). Multivariate linear regression (R^2 = 0.672 6) underestimates high-value areas due to linear assumptions, while spectral index models (*NDCI*, *CI*) are largely unusable due to the "pure water background" assumption (R^2 ≤ 0.102 1, Table 3).

3 Discussion

3.1 Physical Mechanisms of the GBM Model

The GBM model's superior performance stems from its ability to capture the "short-wave absorption-red edge scattering" synergistic mechanism, which addresses the spectral mixing challenge in karst waters. The red-edge band B5 (705 nm) and green-red edge combination B3+B5 (feature importance scores: 0.239 and 0.121, respectively) leverage Chla's strong absorption in blue-green wavelengths (443 ~ 560 nm) and enhanced backscattering at red-edge regions. This mechanism effectively separates algal signals from high suspended particulate matter (*SPM* > 50 mg/L, predominantly $CaCO_3$), whose backscattering coefficient (b_b(555) = 1.2 m^{-1}) exceeds that of plain lakes (e.g., Taihu Lake, 0.8 m^{-1}). In contrast, traditional linear models fail due to the "alkaline trap" effect (high pH ≥ 8.2), which weakens the linear correlation between single bands (e.g., B3, R^2 = 0.560 7) and Chla. The GBM's ensemble learning strategy thus overcomes nonlinear spectral interference, as validated by its 62% accuracy improvement over single-band models.

3.2 Implications for Water Quality Monitoring

This study provides a practical "sensitive band combination + machine learning" framework for karst lake eutrophication management. The model's high accuracy (R^2 = 0.908, $RMSE$ = 0.731 μg/L) enables seasonal tracking of Chla hotspots, such as the Nayong River inlet (summer average: 10.22 μg/L), where agricultural fertilizer inputs (300 kg/ha) drive algal blooms [25]. The negative correlation between Chla and total nitrogen (r = −0.58) under high pH highlights the unique nutrient cycling in karst waters, suggesting that controlling organic carbon inputs (TOC-Chla r = 0.783) may be more effective than nitrogen reduction. Operational applications include: (1) establishing riparian buffer zones to intercept nutrient loads; (2) implementing seasonal fertilizer restrictions (≤200 kg/ha in summer); and (3) integrating Sentinel-2 data into early warning systems (trigger alerts at Chla >8 μg/L). These strategies address the dual challenges of spectral mixing and hydrological fluctuations in karst environments [26].

3.3 Model Transferability and Limitations

While the GBM model shows promise for karst lakes, its transferability requires regional adaptation to optical properties. The model is optimized for karst-specific conditions (high $CaCO_3$ backscattering, alkaline pH), differing from plain lakes (e.g., Taihu Lake) and non-karst waters. Validation in global karst systems

(e. g., Lake Bafa, Turkey; Tonle Sap Lake, Cambodia) is needed to establish regional correction factors. Current limitations include: (1) reliance on surface reflectance, limiting vertical Chla profile inversion in turbid waters; (2) sensitivity to seasonal hydrological fluctuations (e. g., winter low-water periods); and (3) data scarcity in extremely high-*SPM* scenarios (*SPM* >100 mg/L)[27]. Future work should integrate UAV hyperspectral imagery (0.5 ~ 2 m resolution) and LiDAR to resolve vertical dynamics[28], and couple the model with hydrological-social frameworks to simulate climate-anthropogenic impacts[29] (e. g., temperature rise, land use change).

4 Analysis of the Correlation Between Chlorophyll-a (Chla) and Water Quality and Environmental Parameters

As a key indicator of phytoplankton biomass, the concentration of chlorophyll-a (Chla) is closely related to the organic carbon content, physicochemical properties, and particulate characteristics of water bodies. Based on Pearson correlation analysis, the study discusses eight parameters: total organic carbon (TOC), water temperature (WT), transparency, pH, dissolved oxygen (DO), phytoplankton carbon (PC), suspended solids (SS), and turbidity (Fig. 9). As shown in Figure 9, positive correlations (promoting Chla accumulation) are shown in blue; negative correlations inhibiting Chla accumulation) are shown in orang. Allparameters are calculated based on data fromj 40 field sampling points.

4.1 Relationship Between Total Organic Carbon (TOC) and Chla

TOC is significantly positively correlated with Chla (Pearson coefficient 0.783), indicating a coupling effect between the organic carbon cycle in the water body and phytoplankton biomass[30]. TOC is a comprehensive indicator of dissolved and particulate organic carbon, sourced from both organic products and remnants generated by phytoplankton metabolism, as well as exogenous inputs of terrestrial organic matter. High TOC concentrations provide abundant carbon sources for algal growth, promoting the synthesis of photosynthetic pigments. Conversely, during algal proliferation, inorganic carbon is fixed through photosynthesis and converted into organic carbon, further increasing TOC levels. This bidirectional positive feedback relationship is particularly evident in eutrophic waters, reflecting the intrinsic connection between the carbon biogeochemical cycle and primary productivity.

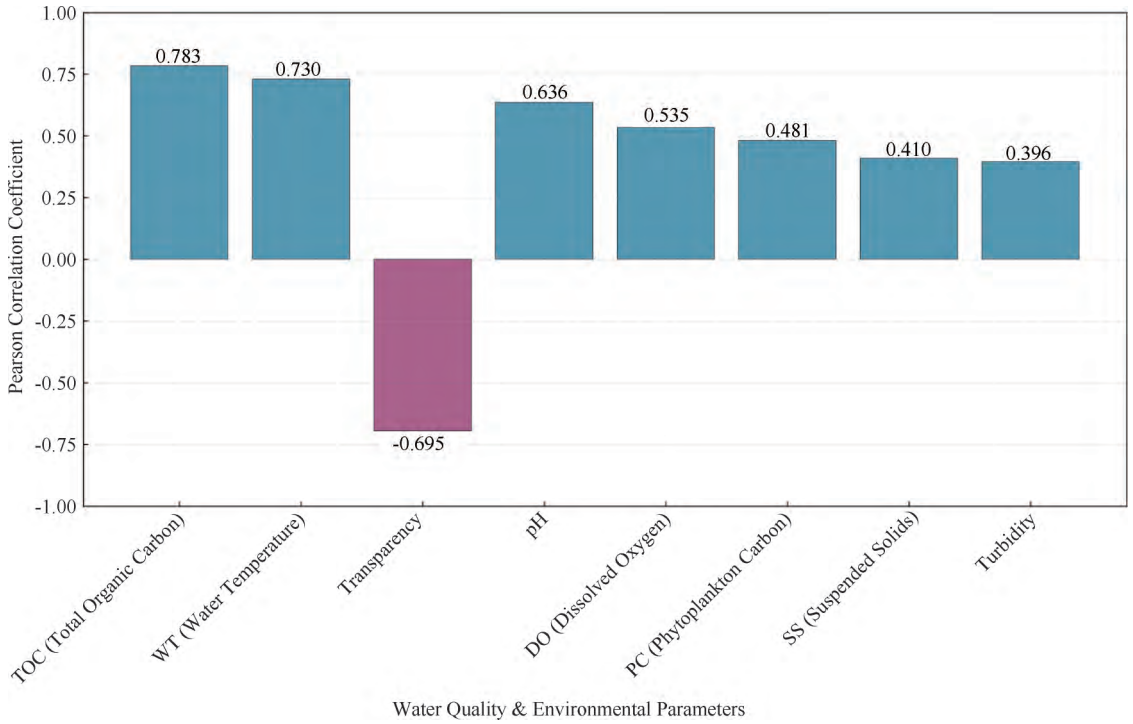


Fig. 9 Pearson correlation coefficients between chlorophyll-a (Chla) and water quality & environmental parameters

图 9 叶绿素 a(Chla)与水质及环境参数的皮尔逊相关系数

4.2 Relationship Between Water Temperature (WT) and Chla

Water temperature (WT) is significantly positively correlated with Chla (0.730), confirming the regulatory effect of temperature on algal physiological activities. Water temperature indirectly regulates Chla accumulation by influencing enzyme activity, cell division rate, and the stratification structure of the water body. Optimal temperatures (20~30 ℃) enhance the efficiency of algal photosynthesis, accelerating chlorophyll synthesis and cell proliferation. Meanwhile, as the temperature rises, water stratification intensifies, weakening the exchange of nutrients between the warm upper layer and the bottom layer, which leads to explosive growth of surface algae under nutrient-rich conditions. In karst lakes, the summer peak in water temperature coincides with high Chla concentration areas, indicating the key role of temperature as a driving factor in phytoplankton community dynamics.

4.3 Relationship Between Transparency and Chla

Transparency is significantly negatively correlated with Chla (−0.695), reflecting the impact of algal biomass on the optical properties of water. As Chla concentration increases, phytoplankton proliferates in large numbers, and the absorption and scattering of light by cell particles intensify, resulting in a decrease in light penetration depth in the water. This negative correlation aligns with the ecological characteristics of "turbid water-type" lakes, where algae, as the primary suspended particles, cause transparency to decrease exponentially with increasing Chla concentration. Additionally, fine particulate suspended matter produced by carbonate weathering in karst waters may interfere with transparency, but algal biomass dominates on a seasonal scale, making it the core factor influencing optical transparency.

4.4 Relationship Between pH and Chla

pH is significantly positively correlated with Chla (0.635), primarily influenced by the dual effects of algal photosynthesis and the carbonate system balance. Phytoplankton consume CO_2 through photosynthesis, leading to the decomposition of HCO_3^- into CO_2 to replenish the carbon source, which subsequently raises the pH of the water. Conversely, a high pH environment (typically > 8.0) is more suitable for the growth of alkaline-tolerant algae, such as blue-green algae, forming a positive feedback mechanism of "pH increase-algal growth-further pH increase." In karst regions, carbonate rock dissolution makes the water pH alkaline (7.5 ~ 8.5), providing a favorable environment for blue-green algae outbreaks, which strengthens the correlation between pH and Chla and reflects the regulatory role of regional geological background in aquatic ecological processes.

4.5 Relationship Between Dissolved Oxygen (DO) and Chla

DO is significantly positively correlated with Chla (0.534), with the key mechanism being the contribution of algal photosynthesis to dissolved oxygen. In surface waters, high Chla concentration zones correspond to areas with dense phytoplankton, and the oxygen released during photosynthesis is the main source of DO. Especially under sufficient light conditions, algal metabolic activities can cause DO concentrations to reach supersaturated levels. Although phytoplankton respiration consumes oxygen at night, the dominant role of photosynthesis during the day results in a positive correlation between DO and Chla. Furthermore, the differences in algal biomass mediated by Chla influence the oxygen balance in the water, with high Chla waters typically having a stronger oxygen production capacity and maintaining higher DO levels, reflecting the direct connection between primary productivity and DO dynamics.

4.6 Relationship Between Phytoplankton Carbon (PC) and Chla

PC is positively correlated with Chla (0.481), as both are quantitative indicators of phytoplankton biomass, mainly reflecting the coordinated changes in cellular carbon reserves and photosynthetic pigments. As a key component of chloroplasts, Chla content is linearly correlated with cellular carbon content (each milligram of Chla corresponds to approximately 50 ~ 100 mg of carbon), especially in communities dominated by a single type of algae, where this correlation is more pronounced. Although the carbon/chlorophyll ratio differs among algae species (e.g., diatoms have a lower ratio than blue-green algae), overall, PC and Chla levels together reflect the consistency of biomass accumulation during phyto-

plankton growth, providing a theoretical basis for inferring phytoplankton carbon reserves from Chla.

4.7 Relationship Between Suspended Solids (SS) and Chla

Suspended solids (SS) are positively correlated with Chla (0.410), reflecting the contribution of algae to organic suspended solids. SS consists of inorganic particles (e.g., clay, carbonate debris) and organic particles (e.g., algae, biological remnants). In eutrophic water environments, organic particles produced by algal proliferation make up the primary component of SS, especially during the summer algal bloom period when the positive correlation between Chla and SS is significantly enhanced [31]. Additionally, inorganic suspended solids may adsorb nutrients, promoting algal growth, forming an indirect positive feedback of "increased suspended solids-enhanced nutrient availability-algal proliferation," which further strengthens the relationship between the two. However, when inorganic particles from external inputs, such as sediments, dominate, the correlation may weaken, reflecting the spatial heterogeneity of particulate sources.

4.8 Relationship Between Turbidity and Chla

Turbidity is positively correlated with Chla (0.396), reflecting the impact of algal cells on the light scattering ability of water. As a comprehensive indicator of water turbidity, turbidity increases with the concentration of suspended particles, especially particles larger than 2 μm. When Chla concentration increases, phytoplankton cells, as highly efficient scattering particles, directly cause an increase in turbidity. Conversely, waters with low Chla concentrations have fewer particles and thus lower turbidity. Although turbidity is affected by non-algal particles (e.g., sediment, humus), the positive correlation with Chla still suggests that in karst lakes, where algae are the primary suspended particles, phytoplankton biomass is an important driver of turbidity changes. This correlation is particularly significant in seasons without large-scale external sediment input [32].

The correlation between chlorophyll-a and various parameters indicates that phytoplankton growth is regulated by a combination of factors, including carbon source supply (TOC), temperature conditions (WT), optical environment (transparency, turbidity), chemical stoichiometry (pH, DO), and particulate characteristics (SS, PC). Significantly positively correlated parameters (TOC, WT, pH, DO, PC, SS, turbidity) collectively promote Chla accumulation through mechanisms such as material cycling, physiological adaptation, and environmental stress. In contrast, transparency is negatively correlated with Chla due to the shading effect of algae. These relationships not only reflect the intrinsic interconnections of the aquatic environmental system but also provide scientific evidence for the selection of Chla-sensitive parameters in remote sensing inversion of water quality in karst lakes. This has significant implications for regional eutrophication monitoring and ecological model development [33].

5 Conclusion

This study focuses on the Pingzhai Reservoir, a typical lake in the karst plateau, and conducts a remote sensing inversion study on chlorophyll-a (Chla) concentration using Sentinel-2 MSI remote sensing imagery and the Gradient Boosting Machine (GBM) model. Through correlation analysis between measured water sample data and remote sensing bands, the electromagnetic response characteristics of Chla concentration in karst plateau lakes are identified. Among the single bands, the green light band B3 (560 nm) shows the strongest correlation with Chla concentration (r = 0.748 8). The red-edge and near-infrared bands, through nonlinear combinations (e.g., B8 × (B7−B5), r = 0.72), effectively enhance the algal spectral signal and suppress the interference of high suspended matter, providing an optimized basis for model development.

The model constructed using the 10 highly sensitive features (5 single bands + 5 combined bands) shows excellent performance, with an R^2 value of 0.908 3, *RMSE* value of 0.730 7 μg/L, and *MAE* value of 0.529 3 μg/L. This is a 62% improvement in accuracy compared to the traditional linear model (B3, R^2 = 0.560 7). The model fully demonstrates the advantage of capturing the complex nonlinear relationship between Chla concentration and remote sensing features. The red-edge band B5 (feature importance 0.239) and the combined band B3+B5 (feature importance 0.121) make significant contributions to the model, confirming that

Chla concentration is influenced by the "short-wave absorption-red-edge scattering" mechanism. This provides theoretical support for the selection of bands and model construction for Chla remote sensing inversion in karst plateau lakes.

Spatiotemporal distribution analysis shows that the Chla concentration in Pingzhai Reservoir is jointly driven by natural factors and human activities: the peak Chla concentration in summer (average 10.22 μg/L) coincides with increased water temperature ($r = 0.730$) and agricultural non-point source pollution input. From June to August, the fertilizer application rate in the basin reaches 300 kg/ha, and the nitrogen-phosphorus ratio (N : P = 18 : 1) significantly exceeds the optimal threshold for algal growth (N : P = 16 : 1), promoting explosive proliferation of nitrogen-tolerant algae such as cyanobacteria. Low winter temperatures (<10 ℃) inhibit algal metabolic activities, combined with the off-season of tourism (reduced tourist volume and decreased domestic pollution sources), collectively leading to the trough of Chla concentration (average 2.46 μg/L). Reduced bioavailability of total nitrogen in high pH environments (8.5~9.0) confirms the 'alkaline trap' effect in karst water bodies, recommending that management departments prioritize control of organic carbon inputs. Redundancy analysis (RDA) shows that water temperature and total nitrogen jointly explain 68.3% of the Chla concentration variation, providing a new method for analyzing the relationship between climate factors and human activities.

The GBM model developed herein ($R^2 = 0.908$) enables high-precision seasonal monitoring of Chla concentrations in karst lakes, facilitating targeted surveillance of inflow zones (e.g., Nayong River inlet) during summer bloom periods when average Chla reaches 10.22 μg/L. Integrating Sentinel-2 data with in-situ sampling, this framework supports adaptive management strategies. Based on the identified negative correlation ($r = -0.58$) between Chla and total nitrogen under high pH conditions—highlighting the "alkaline trap" effect—specific mitigation measures include establishing riparian buffer zones to intercept nutrient loads and regulating summer fertilizer application rates to ≤200 kg/ha within the watershed. Leveraging the model's 10~60 m spatial resolution, real-time Chla inversions can be integrated into early warning systems to trigger alerts when surface concentrations exceed 8 μg/L. Seasonal dynamics (summer peaks, winter lows) further inform adaptive policies, such as restricting lakeshore tourism during high-Chla months (June-August) and prioritizing wetland restoration in exposed shoals during winter low-water periods, thereby providing actionable technical solutions for karst reservoir water quality management and ecological protection.

This study, through innovative feature engineering of highly sensitive single bands and cross-band linear combinations, combined with the efficient capture of nonlinear relationships by the Gradient Boosting Model, has developed a remote sensing inversion method for Chla concentration in karst plateau lakes. This method not only addresses the issue of spectral mixing in complex optical environments but also reveals the driving mechanisms of region-specific environmental factors. It provides technical support and scientific basis for water quality monitoring and ecological management of Pingzhai Reservoir and similar water bodies.

Although this study has achieved certain results in Chla remote sensing inversion for karst plateau lakes, due to the complex terrain of the study area, uneven distribution of sampling points caused by local ecological conditions, and insufficient matching of actual field data with remote sensing imagery in terms of temporal resolution, the model's generalizability under extremely complex water quality conditions still requires further verification. Future research could combine remote sensing data with higher spatial and temporal resolution, deepen the application of multi-source heterogeneous data fusion, and strengthen the study of the mechanism behind the relationship between karst-specific hydrogeological conditions and Chla spectral responses to improve the accuracy and adaptability of the Chla remote sensing inversion model, providing stronger technical support for the precise monitoring and scientific management of the water ecological environment in karst plateau lakes.

Future research is proposed to focus on three specific avenues: First, transnational validation of the GBM model across global karst lakes (e.g., Lake Bafa in Turkey and Tonle Sap Lake in Cambodia) using

consistent Sentinel-2 processing protocols, aiming to establish regional adaptation strategies for Chla inversion. Second, multi-source data fusion integrating UAV hyperspectral imagery (0.5 ~ 2 m resolution) with Sentinel-2 and in-situ sensor data via a Kalman filter framework, which will enhance spatial-temporal resolution for micro-scale gradient analysis. Third, mechanistic studies on karst hydrogeological effects, including laboratory simulations of carbonate dissolution impacts on Chla spectra and field profiling of vertical Chla-turbidity relationships. Additionally, developing a coupled hydrological-social model to simulate climate-anthropogenic impacts (e.g., temperature rise, land use change) on Chla dynamics will support adaptive management frameworks for karst water bodies. These directions aim to bridge current gaps in model generalizability, data integration, and process-based understanding.

参考文献(References):

[1] XIAO X F, ZHANG S L, GUO W, et al. Environmental pollution characteristics of surface water and groundwater in southwest China and its research prospects [J]. Earth and Environment, 2023, 51(5): 564-573.

[2] SUN D Y, ZHOU X Y, LI Y M, et al. Hyperspectral remote sensing of chlorophyll-a concentration in Taihu Lake based on optical classification [J]. Environmental Science, 2013, 34(8): 3002-3009.

[3] LI Y Q, LAI Y Q, ZHANG J, et al. Review on remote sensing inversion methods of Chlorophyll a in Taihu Lake [J]. IOP Conference Series: Earth and Environmental Science, 2020, 467(1): 012135.

[4] MUSGROVE M, OPSAHL S P, MAHLER B J, et al. Source, variability, and transformation of nitrate in a regional karst aquifer: Edwards aquifer, central Texas [J]. Science of the Total Environment, 2016, 568: 457-469.

[5] DAN Y S, ZHOU Z F, LI S H, et al. Retrieval of chlorophyll-a concentration in pingzhai reservoir based on sentinel-2 [J]. Environmental Engineering, 2020, 38(3): 180-185.

[6] LI Y L. Study on Remote Sensing Inversion of Chlorophyll-a Concentration in Karst Plateau Lakes [D]. Guiyang: Guizhou Normal University, 2022.

[7] ZHENG P, XU X F, XU X H, et al. Multi-year same-phase "satellite-ground" collaborative remote sensing inversion of chlorophyll-a concentration in Poyang Lake [J]. Journal of Yangtze River Scientific Research Institute, 2024, 41(7): 48-56.

[8] TIAN S, ZHU X T, ZHANG H, et al. An optimal strategy for estimating chlorophyll-a concentration in case II waters to support sustainable development goals [J]. Advances in Space Research, 2025, 75(10): 7195-7211.

[9] HUANG Q H, HE Z H, LIANG H, et al. Estimation of chlorophyll-a concentration in baihua lake water based on hyspectral data [J]. Environmental Science & Technology, 2019, 42(1): 134-141.

[10] WASEHUN E T, BENI L H, DI VITTORIO C A, et al. Comparative analysis of Sentinel-2 and PlanetScope imagery for chlorophyll-a prediction using machine learning models [J]. Ecological Informatics, 2025, 85: 102988.

[11] PAN W B, YU F, LI J L, et al. Quantification of chlorophyll-*a* in inland waters by remote sensing algorithm based on modified equivalent spectra of Sentinel-2 [J]. Ecological Informatics, 2025, 87: 103061.

[12] HAN W L, ZHAO Q C. Retrieving chlorophyll-a concentrations in Baiyangdian Lake from sentinel-2 data using Kolmogorov-Arnold networks [J]. Water, 2025, 17(15): 2346.

[13] LI Z K, YANG X K, RAN L S, et al. Assessing the impacts of cascade reservoirs on Pearl River environmental status using machine learning and satellite-derived chlorophyll-a concentrations [J]. Journal of Environmental Management, 2025, 382: 125406.

[14] ZHANG Y R, ZHOU Z F, ZHANG H T, et al. Quantifying the impact of human activities on water quality based on spatialization of social data: A case study of the Pingzhai Reservoir Basin [J]. Water Supply, 2020, 20(2): 688-699.

[15] ZHENG W, LU X, LI Y, et al. Hyperspectral identification of chlorophyll fluorescence parameters of suaeda salsa in coastal wetlands [J]. Remote Sensing, 2021, 13(11): 2066.

[16] ASWATHY T S, ACHU A L, FRANCIS S, et al. Assessment of water quality in a tropical ramsar wetland of southern India in the wake of COVID-19 [J]. Remote Sensing Applications: Society and Environment, 2021, 23: 100604.

[17] WANG J, LI W, ZHAO W Q, et al. Remote sensing inversion of CODMn in deep karst plateau lakes and reservoirs based on Sentinel-2 data [J]. Remote Sensing Technology and Application, 2024, 39(1): 98-109.

[18] YAN F L, LI Y Z, FAN X T, et al. Estimating chlorophyll-a concentrations in optically shallow waters using Gaofen-1 wide-field-of-view (GF-1 WFV) datasets from Lake Taihu, China [J]. Remote Sensing, 2025, 17(7): 1299.

[19] SAI J V, VASANTH K, NARAYANAN M, et al. Comparison of chlorophyll-a retrieval from aqua MODIS using SeaDAS and ArcGIS of Visakhapatnam coast at bay of Bengal [J]. Asian Journal of Current Research, 2025, 10(2): 166-176.

[20] GUPTA A. Assessment of bio-optical parameters retrieval in the coastal waters using local to global optimization [J]. Journal of Earth System Science, 2024, 133(4): 228.

[21] WANG X M, DENG Y, TUO Y C, et al. Study on the temporal and spatial distribution of chlorophyll a in Erhai Lake based on multispectral data from environmental satellites [J]. Ecological Informatics, 2021, 61: 101201.

[22] GU W B, LIANG J, YANG L, et al. Retrieval of chlorophyll-a concentration in Nanyi Lake using the AutoGluon framework [J]. Water, 2025, 17(15): 2190.

[23] JEONG B, LEE S M, HEO J, et al. Deep learning-based retrieval of chlorophyll-a in lakes using sentinel-1 and sentinel-2 satellite imagery [J]. Water, 2025, 17(11): 1718.

[24] ZHU W D, LIU S, LUAN K F, et al. Research on the inversion of chlorophyll-a concentration in the Hong Kong coastal area based on convolutional neural networks [J]. Journal of Marine Science and Engineering, 2024, 12(7): 1119.

[25] KE X L, LIU M, DENG X Z. Ecological zoning of lake nutrients based on remote sensing inversion parameters and double-constraint spatial clustering algorithm [J]. Progress in Geography, 2012, 31(3): 315-323.

[26] LI G Y. Improved water security for China's efforts to build itself into a stronger country and rejuvenate the Chinese nation on all fronts by pursuing Chinese modernization [J]. Water Resources Development Research, 2024, 24(3): 1-3.

[27] CHEN Y T, MA B D, LI X X, et al. Dustfall effect on hyperspectral inversion of chlorophyll content: A laboratory experiment [J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2018, IV-3: 65-69.

[28] WU C Y, XIE M, LIN L, et al. 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 [J]. Water, 2025, 17(6): 855.

[29] BYRNE A, LOMEO D, OWOKO W, et al. LAQUA: A LAndsat water QUality retrieval tool for east African lakes [J]. Remote Sensing, 2024, 16(16): 2903.

[30] XIE R K, ZHOU Z F, KONG J, et al. Remote sensing inversion study of total organic carbon concentration in Karst Plateau Lakes-Taking Pingzhai reservoir as an example [J]. Geocarto International, 2024, 39(1): 2343006.

[31] ZHANG L, ZHANG C, MA C F, et al. Machine learning-based retrieval of chlorophyll-a and total suspended matter from HY-3A CZI: Model development, validation, and application [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2025, 227: 613-631.

[32] KONG J, ZHOU Z F, XIE R K, et al. Tracing the source and behaviour of sulphate in karst reservoirs, using stable isotopes and Bayesian isotopic-mixing models [J]. Science of the Total Environment, 2025, 958: 177994.

[33] SALAH M, SALEM S I, UTSUMI N, et al. 3LATNet: Attention based deep learning model for global Chlorophyll-a retrieval from GCOM-C satellite [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2025, 220: 490-508.

(责任编辑　王海锋)