Research on Inversion Technology of Empirical Models for Water Chlorophyll Concentration Based on Sentinel-2 Images

Jiqiang Tan [1], Yilong Gong [1], Fuzhong Yin [2], Zhenfa Zhou [2], Song Bai [1], Hengfei Liu [1]

[1] Heilongjiang Geomatics Center of MNR, Harbin 150081, China

[2] Heilongjiang Administration of Surveying, Mapping and Geoinformation, Harbin 150081, China

Keywords: Multispectral Imagery, Chlorophyll Inversion, Remote Sensing Image Application

Abstract

To address the limitations of traditional fixed-point sampling for monitoring water chlorophyll concentration and improve the inversion accuracy of complex inland waters, this study took the Naoli River Nature Reserve as the research area and conducted research based on Sentinel-2 images and measured chlorophyll point data. First, the study performed combined calculations on multispectral bands to generate multiple derived bands, and selected "B4+B5+B6" as the optimal band combination through the coefficient of determination (R^2). Then, using this combination as input, various empirical models were constructed and evaluated using multiple indicators. The results showed that the univariate cubic function model had the highest R^2 and the smallest multiple error indicators, which was significantly better than other models, and successfully realized the spatial inversion of chlorophyll concentration in the study area. This study reveals the complex nonlinear relationship between chlorophyll and band combinations, provides a high-precision inversion technical scheme for water chlorophyll, offers data support for algal bloom early warning and water quality fluctuation tracking, and provides scientific references for the optimization of river basin management measures and ecological protection decisions.

1. Introduction

Water chlorophyll concentration is a core indicator for characterizing phytoplankton biomass (Zhou et al., 2023). Its content directly reflects the degree of water eutrophication (Gu et al., 2025), and it also serves as a key parameter for early warning of cyanobacterial bloom outbreaks (Zhang et al., 2025) and assessing the risk of water quality deterioration (Ma et al., 2013). Traditional monitoring relies on fixed-point sampling and laboratory analysis. Although it yields high-precision data, it has limitations such as limited coverage, low monitoring frequency, and difficulty in capturing spatial heterogeneity, which makes it unable to meet the needs of dynamic water quality management at the large watershed scale. With the development of remote sensing technology, multispectral satellite imagery has become a mainstream technical method for water chlorophyll concentration inversion, as it enables rapid acquisition of surface information over a large area.

The Naoli River water body has complex optical characteristics, and the long-term monitoring network in the protected area can provide synchronized in-situ chlorophyll data, offering a typical research scenario for the construction and validation of remote sensing inversion models. Based on the above context, this study takes the Naoli River as the research area, uses Sentinel-2 Level-2A imagery as the data source, and combines in-situ measured point data to construct a variety of empirical models. The coefficient of determination (R^2) is used as the evaluation index to determine the optimal inversion model, thereby realizing the spatial inversion of chlorophyll concentration in the study area. This study aims to provide a technical solution for high-precision inversion of chlorophyll concentration in complex inland water bodies, and at the same time offer scientific support for dynamic watershed water quality monitoring and ecological management.

2. Basic research data

2.1 Study area

The Naoli River is located in the hinterland of the Sanjiang Plain in Northeast China, with geographical coordinates ranging from 132°22′29″E to 134°13′45″E (east longitude) and 46°30′22″N to 47°24′32″N (north latitude). It has a drainage basin area of 23,988 square kilometers, and its main tributaries include first-level tributaries such as the Qixing River and the Hamatong River. The long-term monitoring network of the Naoli River Protected Area provides synchronized ground data (including chlorophyll concentration), which is used for the collaborative validation of "remote sensing imagery + in-situ measured data" to effectively calibrate model deviations. Its complex optical mixing characteristics provide a typical research scenario for the inversion of chlorophyll empirical models based on the red edge bands of Sentinel-2 satellites. The image of the test area is shown in Figure 1.

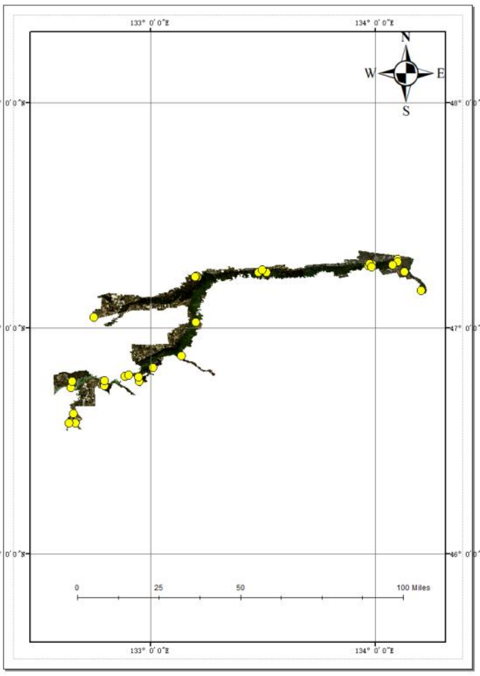

Fig 1. Naoli River Experimental Area Images.

2.2 Sentinel-2 imagery

Sentinel-2 imagery offers significant technical advantages in water chlorophyll inversion: the MultiSpectral Instrument (MSI) it carries includes three red edge bands (B5-B7, 705-783 nm), which accurately capture the characteristic absorption peaks of chlorophyll in the red edge region—this is a core advantage not possessed by series such as Landsat and Modis. Its high spatial resolution of 10 meters can clearly reflect the spatial heterogeneity of chlorophyll in small water units, avoiding inversion deviations caused by mixed pixels and adapting to the complex water conditions of the Naoli River. Additionally, its short revisit period of 5 days enables dynamic monitoring of chlorophyll concentration.

In this study, Sentinel-2 Level-2A products were selected as the data source. This level of products has completed preprocessing procedures including radiometric calibration, atmospheric correction (including the removal of influences such as aerosols and water vapor), and geometric correction. They can be directly used for water chlorophyll concentration inversion, effectively avoiding the technical costs of independently conducting preprocessing work such as radiometric calibration and atmospheric correction in the study, while ensuring data reliability.

3. Research Methods

To invert water chlorophyll concentration, the following steps are required: first, acquire multispectral imagery and perform preprocessing such as radiometric correction and atmospheric correction on it. Then, generate chlorophyll inversion data by combining the multispectral imagery with in-situ measured data. After conducting band combination on the chlorophyll inversion data, apply multiple empirical models for model inversion. Evaluate these empirical models using the coefficient of determination as the evaluation index to select the optimal empirical model, and finally invert the chlorophyll concentration in the study area. The overall workflow is shown in Figure 2.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-1113-2026 |

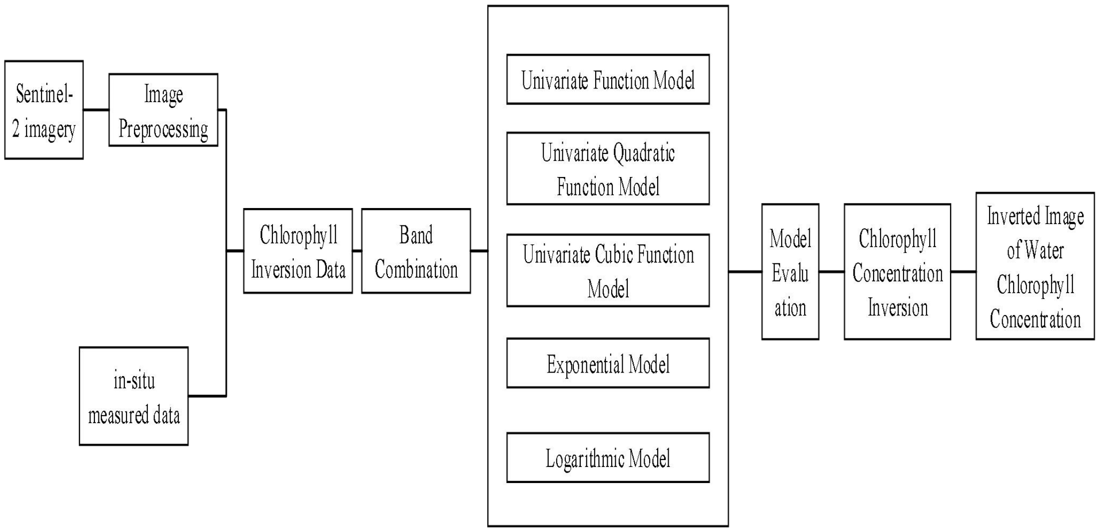

Fig 2. Flow Chart for Water Chlorophyll Concentration Inversion.

3.1 Band Combination

The band combination technique based on Sentinel-2 multispectral imagery is one of the core techniques for improving the inversion accuracy of parameters such as water chlorophyll. In the inversion process, the reflectance information of these 13 bands and the spectral information of a single band can be obtained simultaneously at a single in-situ measurement point. Through the band combination strategy, the original 13 bands can be expanded into a large number of combined bands, which provide the model with more abundant spectral feature dimensions. This enables the model to accurately capture the nonlinear correlation between chlorophyll and spectral reflectance, screen out the optimal band combination, and significantly reduce inversion errors. The band combination process is divided into two steps: "acquiring band factors" and "recombining bands".

Single-band factor extraction is conducted by directly retaining the reflectance data of the 13 original bands as independent factors, which serves to preserve the original spectral characteristics of the imagery. Exhaustive double-band combination is implemented based on the permutation and combination principle, where the 13 original bands are paired two by two to generate all possible double-band combination factors. These double-band factors can initially reflect the synergistic or inhibitory relationships between different bands, thereby providing a targeted feature basis for the subsequent band recombination process.

Band recombination is carried out through three specific forms to construct diverse spectral features: band addition recombination involves the superposition calculation of reflectances from two or more band factors with the formula of "Band A + Band B"; band subtraction recombination highlights the spectral differences between target parameters and interfering factors by calculating the reflectance difference of two band factors in the form of "Band A - Band B"; normalized recombination adopts the formula of "(Band A - Band B)/(Band A + Band B)", which can effectively eliminate the interference of external factors such as light intensity, observation angle and residual atmospheric effects on spectral reflectance data.

3.2 Empirical Models

Empirical models are inversion models constructed based on the statistical correlation between in-situ measured data and remote sensing features. They do not require a clear physical mechanism and only establish variable mapping relationships by fitting data patterns. Due to their advantages of simple operation and efficient calculation, they are widely used in water chlorophyll inversion. Combining the band characteristics of Sentinel-2 imagery and in-situ chlorophyll measured data, this study selects the following five types of typical empirical models for inversion.

Univariate function models are applicable to scenarios with a significant single correlation between spectral features and chlorophyll concentration. Their prominent advantage is the small number of parameters, which ensures high interpretability and enables direct inference of chlorophyll concentration through reflectance variations of the corresponding band combination. However, these models only utilize information from a single band combination, making it difficult to avoid spectral interference caused by other substances in the water body.

Univariate quadratic function models capture the nonlinear correlation between chlorophyll concentration and spectral features by introducing a quadratic term. They are well-suited for water bodies with low content of interfering substances, where the relationship between chlorophyll concentration and spectral reflectance exhibits obvious nonlinear characteristics that cannot be effectively fitted by linear models.

Univariate cubic function models can fit more complex nonlinear variation trends with the introduction of a cubic term, compared with low-order univariate function models. Their advantage lies in the ability to comprehensively and accurately capture the complex correlation between spectral information and chlorophyll concentration, making them particularly applicable to water bodies with a slightly higher level of interfering substances. Nevertheless, these models are prone to overfitting issues and have specific requirements for the sample size to ensure the reliability of fitting results.

Exponential models take a single band combination factor as the independent variable, and their core lies in reflecting the correlation between spectral reflectance and chlorophyll concentration through the nonlinear characteristics of exponential functions. They are adapted to the nonlinear response characteristics between chlorophyll and spectral features, and can effectively alleviate the spectral saturation problem that often occurs in scenarios with high chlorophyll concentration.

Logarithmic models break through the sensitivity limitation of linear models by amplifying weak spectral differences, making them one of the important models for chlorophyll inversion in low-concentration water bodies. Their main advantage is targeting the nonlinear correlation of low-concentration chlorophyll and amplifying subtle spectral variations to improve inversion sensitivity. A notable disadvantage is that the independent variable of logarithmic models must be a positive value, which requires truncation processing for negative values generated during spectral calculation to ensure the validity of model operation.

3.3 Model Evaluation Indicators

Combining the technical requirements of chlorophyll inversion and data characteristics, this study constructs a comprehensive evaluation system including the Coefficient of Determination (R^2), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Relative Mean Absolute Error (RMAE). These four types of indicators complement each other from four dimensions: goodness of fit, error magnitude, robustness, and applicability. Together, they support a comprehensive evaluation of model performance and provide a scientific basis for selecting the optimal inversion model.

In Equations 1 to 4, x_i represents the true value of the i-th sample, y_i represents the predicted value of the i-th sample, and n represents the total number of evaluation data samples.

The Coefficient of Determination (R^2) is a core indicator for measuring model fitting goodness in regression analysis, with its value generally ranging from 0 to 1. A value closer to 1 signifies stronger explanatory power and better fitting performance of the model, while a negative value indicates model invalidity. It is widely applied to evaluate the prediction fitting effects of linear, nonlinear and machine learning regression models, and its calculation formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)^2} \tag{1}$$

Mean Squared Error (MSE) quantifies the overall magnitude of a model's inversion error by calculating the average of the squared differences between predicted values and true values. It is particularly sensitive to outliers (such as extreme values of chlorophyll concentration occurring in inversion) and can amplify the impact of larger errors. Thus, it is suitable for scenarios where emphasis is placed on squared error loss. A smaller MSE value indicates a smaller overall inversion error of the model and higher prediction accuracy. Its calculation formula is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{2}$$

Mean Absolute Error (MAE) reflects the average level of a model's inversion error by calculating the average of the absolute differences between predicted values and true values. Compared with MSE, MAE is less affected by outliers, resulting in more robust outcomes. It can intuitively show the average deviation between the model's predicted values and the actual chlorophyll concentration. A smaller MAE value indicates a smaller average deviation of the model's inversion results and stronger prediction stability for most samples. Its calculation formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - y_i\right| \tag{3}$$

Relative Mean Absolute Error (RMAE) introduces the mean of true values for normalization based on MAE. This eliminates the impact of dimensional differences in chlorophyll concentration, facilitating the comparison of model accuracy across different study areas and concentration gradients. Meanwhile, it can more objectively reflect the relative proportion of errors to true values, avoiding deviations in error evaluation caused by the magnitude of absolute concentrations. RMAE is usually presented as a percentage; a smaller value indicates a lower relative proportion of the model's inversion error and stronger adaptability to chlorophyll at different concentrations. Its calculation formula is as follows:

$$RMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|x_i - y_i\right|} \tag{4}$$

4. Experimental Results and Analysis

A part of the Naoli River basin was selected as the study area for water chlorophyll concentration inversion. The inversion was conducted in the middle of May 2020. Combined with the chlorophyll concentration in the study area derived from the model inversion, an analysis was carried out based on the experimental results.

4.1 Experimental Data

Sentinel-2 remote sensing images covering the study area and measured data from 30 sampling points were selected for the period from May 1, 2020 to May 10, 2020. Based on the spatial heterogeneity of water bodies in the Naoli River Basin, this study adopted a grid sampling method to deploy 30 sampling points, covering the main stream and major tributaries of the river to ensure the spatial representativeness of the sampling distribution. All water samples were collected at 0.5 m below the water surface to avoid the interference of photodegradation in surface water on chlorophyll concentration. Water turbidity was measured synchronously during sampling, with a range of 5-30 NTU. By analyzing the correlation between turbidity and chlorophyll concentration, the interference mechanism of turbidity on inversion results was clarified. Subsequently, the impact of turbidity was reduced by optimizing the band combination selection scheme, ensuring data reliability and model inversion accuracy. According to the spatial relationship between the measured sampling points and the raster images, chlorophyll concentration inversion data were obtained. Among the bands of Sentinel-2 imagery, Band 1 (Coastal Aerosol), Band 9 (Water Vapor), and Band 10 (Cirrus Detection) were not used in the inversion. Example data are shown in Table 1.

B2	B3	B4	B5	B6	B7	B8	B8A	B11	B12	chlorophyll concentration（μg/L）
0.0542	0.0619	0.0258	0.0454	0.0673	0.0508	0.1784	0.1271	0.039	0.0358	2.15
0.041	0.089	0.0151	0.0774	0.0618	0.0854	0.1323	0.1894	0.0314	0.0438	4.32
0.0454	0.0473	0.0286	0.0584	0.0418	0.0655	0.1574	0.1006	0.0334	0.0202	0.92
0.065	0.0621	0.021	0.0554	0.0881	0.0612	0.2068	0.1316	0.0498	0.0314	5.87
0.0348	0.0782	0.0177	0.0641	0.0462	0.0715	0.1718	0.1638	0.027	0.0374	1.76
0.0542	0.0619	0.0258	0.0454	0.0673	0.0508	0.1784	0.1271	0.039	0.0358	2.15
...	...	...	...	...	...	...	...	...	...	...

Tab 1. Example Data of Chlorophyll Concentration Inversion

4.2 Chlorophyll Concentration Inversion

Band combination operations were performed on the chlorophyll concentration inversion data, resulting in a total of 10,120 combined bands. These combined bands were evaluated based on the in-situ measured data and the Coefficient of Determination (R^2). Finally, the top 10 combined bands with the highest correlation were obtained, as shown in Figure 3.

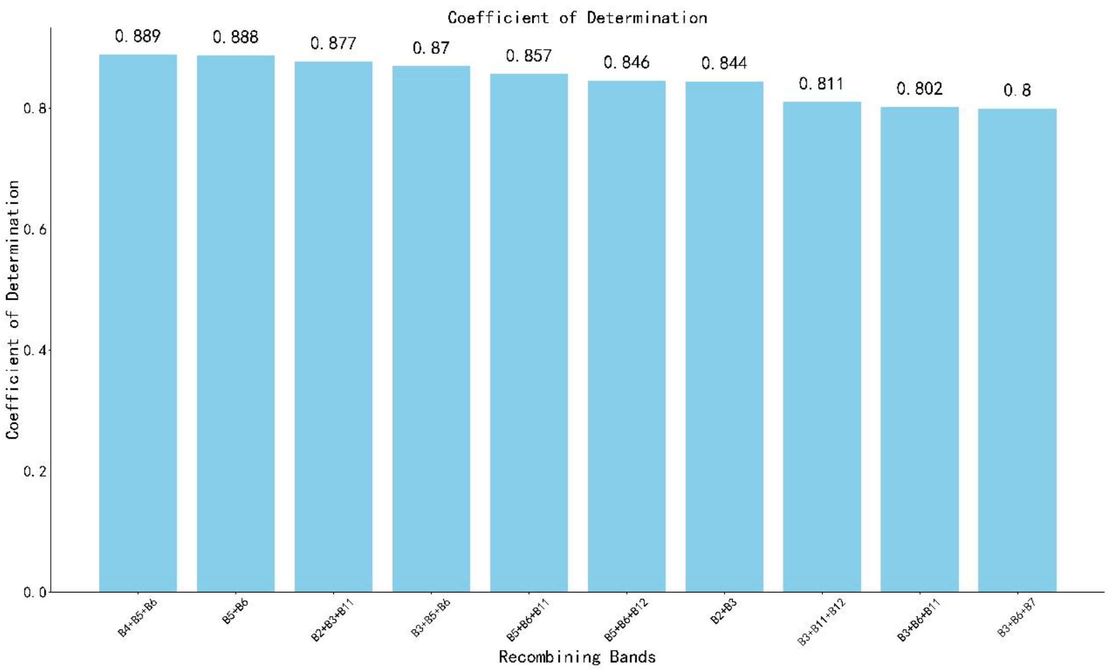

Fig 3. Screening of Correlations between Band Combinations and Determination Coefficients

The band combination with the highest correlation is B4 + B5 + B6, and the three bands correspond to the Red band, Red Edge 1 band, and Red Edge 2 band respectively. Using B4 + B5 + B6 as the basic data for the inversion model, combined with five empirical models including the unary function, quadratic function, cubic function, logarithmic function, and exponential function, the R^2 of each model was calculated, as shown in Table 2.

Model Name	Model Formula	R^2	MSE	MAE	RMAE
Univariate Function Model	144.23x -18.05	0.866	0.686	0.646	0.828

Model Name	Model Formula	R²	MSE	MAE	RMAE
Univariate Quadratic Function Model	$1431.35x^{2}$-299.73x+15.76	0.916	0.432	0.605	0.657
Univariate Cubic Function Model	$-54588.13x^3+27103.97x^2-4288.54x+220.41$	0.923	0.389	0.520	0.623
Logarithmic Model	21.9ln(x)+45.42	0.835	0.846	0.679	0.920
Exponential Model	$0.55e^{17.25x}-4.31$	0.909	0.466	0.625	0.682

Tab 2. Inverse formula and evaluation index comparison

After comparing multiple empirical models, the Univariate Cubic Function Model achieved the highest R^2, while its error terms (MSE, MAE, and RMAE) were all the smallest. This indicates that this empirical model exhibits the best inversion performance in terms of data.

4.3 Analysis of Experimental Results

This study conducted inversion research using Sentinel-2 remote sensing images covering the Naoli River study area from May 1 to May 10, 2020 as the data source, combined with in-situ measured chlorophyll concentration data from 30 sampling points. The experiment formed a closed-loop verification from three dimensions: data preprocessing, band combination selection, and model comparison, and the results demonstrated clear scientific validity and practical applicability.

To fully explore spectral information for screening the optimal band combination, the study performed full permutation and combination operations (including single-band, dual-band, and triple-band combinations) on 10 valid bands (excluding B1, B9, and B10, which are non-water-sensitive bands), generating a total of 10,120 band combinations (including addition, subtraction, normalization, and other recombination forms). Using the coefficient of determination (R^2) as the core evaluation index combined with in-situ measured data, the top 10 combinations with the highest correlation were selected (as shown in Figure 3). Among them, the B4+B5+B6 combination achieved an R^2 value of 0.897, which was significantly higher than other combinations (the R^2 range of the Top 10 combinations was 0.821-0.897). Its superiority stems from the synergistic and complementary effects of the three bands: B4 (red band, 665 nm) is a strong absorption band for chlorophyll, and its reflectance shows a significant negative correlation with chlorophyll concentration (in-situ data verification indicated that when the chlorophyll concentration increased from 0.92 μg/L to 5.87 μg/L, the B4 reflectance decreased from 0.0286 to 0.021); B5 (red edge 1 band, 705 nm) and B6 (red edge 2 band, 740 nm) can sensitively capture the "red edge shift" phenomenon. The combination of the three not only retains the absorption characteristics but also enhances the sensitivity to concentration gradient differences, effectively avoiding the defect that a single band is susceptible to interference from water turbidity and external light.

To clarify the performance differences among different empirical models, the study constructed a four-dimensional evaluation system consisting of "R^2 + Mean Squared Error (MSE) + Mean Absolute Error (MAE) + Relative Mean Absolute Error (RMAE)", and conducted comprehensive verification on five types of empirical models with the B4+B5+B6 combination as the input variable (as shown in Table 2). The univariate cubic function model performed the best, with an R^2 of 0.923 and the smallest values of MSE (0.389), MAE (0.520), and RMAE (0.623), demonstrating excellent goodness of fit and error control ability. The univariate quadratic function model (R^2=0.916) and the exponential function model (R^2=0.909) ranked second, with slightly higher errors in fitting high-concentration samples. The univariate function model (R^2=0.866) and the logarithmic function model (R^2=0.835) showed relatively weak performance; the latter required truncation of some data due to the positivity requirement of the input variable, leading to large inversion deviations for low-concentration chlorophyll. This result confirms that there is a complex nonlinear relationship between chlorophyll concentration and the band combination, and the multi-order term of the cubic function can accurately capture this correlation, which is consistent with the physical mechanism of chlorophyll absorption-scattering in water bodies.

Two abnormal phenomena were observed during the inversion, with specific analyses as follows:

R^2 of empirical models < R^2 of band combinations: Theoretically, the R^2 of empirical models should be higher than that of band combinations, but the opposite was true for the univariate function model and logarithmic model in this study. The reason is that during inversion, the data were divided into a training set (80%) and a test set (20%): empirical models were fitted based on the training set and evaluated using the test set, while band combinations directly calculated R^2 using the full dataset, leading to this abnormal phenomenon. The stability of the models can be improved by increasing the sample size and enriching band data.

Negative chlorophyll concentration values in inversion: When using the fitted models to invert the water body of the Naoli River and its surrounding areas, negative values or extremely high values appeared in some regions. The cause is that the images include not only water bodies but also rice planting areas and bare soil areas. However, the models fitted based on in-situ measured points of water bodies show poor inversion performance for non-water areas, while their inversion performance for chlorophyll concentration in water bodies is consistent with the spectral characteristics of the images.

5. Conclusion

This study took a partial area of the Naoli River as the research area, focusing on the demand for water chlorophyll concentration inversion. Based on Sentinel-2 data and in-situ measured data from 30 sampling points, a complete research framework of "band combination selection - empirical model comparison - inversion accuracy verification" was constructed to systematically conduct water chlorophyll concentration inversion research. The main conclusions and prospects are as follows:

The study performed combination operations on 10 valid bands, generating 10,120 derived bands, and finally selected B4 + B5 + B6 as the optimal band combination. On this basis, a comparison of five types of empirical models (univariate function, univariate quadratic function, univariate cubic function, logarithmic function, and exponential function) showed that the univariate cubic function model (R^2 = 0.923) could most accurately characterize the complex nonlinear correlation between chlorophyll concentration and the band combination. Its goodness of fit was significantly better than that of other models. Finally, spatial inversion of chlorophyll concentration in the Naoli River study area was realized based on this model, providing an efficient remote sensing technical scheme for regional dynamic water quality monitoring and cyanobacterial bloom early warning, and effectively making up for the limitations of traditional fixed-point sampling, such as limited coverage and poor timeliness.

Future research can be optimized from two aspects: first, expand the size of in-situ measured samples and cover different seasons to improve the model’s adaptability to chlorophyll concentration gradients and environmental interference; second, combine the advantage of Sentinel-2’s short revisit cycle to carry out temporal dynamic analysis of chlorophyll concentration, further explore the variation trend of chlorophyll concentration, and provide more comprehensive scientific support for water quality protection and ecological management.

References

Gu, Y.Z., Deng, R.R., Li, J.Y., Guo, Y., Hua, Z.Q., Kuang, Z.Y., Huang, D., 2025. Chlorophyll-a concentration inversion and eutrophication assessment in the Beijing River Basin. Bulletin of Surveying and Mapping, (06), 24-29+72.

Ma, T.F., 2013. Effect of cyanobacteria bloom on the variation trend of water quality parameters. Guangdong Water Resources and Hydropower, (S1), 54-57.

Zhang, Y., Xiong, J., Tang, M.Z., Zhang, X., 2025. A preliminary study on water bloom early warning and monitoring based on satellite remote sensing chorophyll products in typical lakes and reservoirs in Hubei province. Environmental Monitoring in China, 41(04), 267-278.

Zhou, Q., Chen, G.J., Li, R., Chen, X.L., Zhang, T., Han, Q.H., Wang, L., 2023. A comparative study on spatio-temporal variations and environmental drives of lake-water chlorophyll-a and diatom communities in Lake Yilong. Journal of Lake Sciences, 35(06), 1891-1904.