Remote sensing inversion of lake water quality parameters based on ensemble modelling

Li Xiaojuan¹, Huang Mutao^{1,*}, and Li Jianbao¹

¹ School of Hydropower Engineering and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract. In this paper, combined with water quality sampling data and Landsat8 satellite remote sensing image data, the inversion model of Chl-a and TN water quality parameter concentration was constructed based on machine learning algorithm. After the verification and evaluation of the inversion results of the test samples, Chl-a TN inversion model with high correlation between model test results and measured data was selected to participate in remote sensing inversion ensemble modelling of water quality parameters. Then, the ensemble remote sensing inversion model of water quality parameters was established based on entropy weight method and error analysis. By applying the idea of ensemble modelling to remote sensing inversion of water quality parameters, the advantages of different models can be integrated and the precision of water quality parameters inversion can be improved. Through the evaluation and comparative analysis of the model results, the entropy weight method can improve the inversion accuracy to some extent, but the improvement space is limited. In the verification of the two methods of ensemble modelling based on error analysis, compared with the optimal results of a single model, the determination coefficient (R²) of Chlorophyll a and TN concentration inversion results was increased from 0.9288 to 0.9313 and from 0.8339 to 0.8838, and the root mean square error was decreased from 14.2615 $\mu\text{g/L}$ to 10.4194 $\mu\text{g/L}$ and from 1.1002 mg/L to 0.8621 mg/L. At the same time, with the increase of the number of models involved in the set modelling, the inversion accuracy is higher.

1 INTRODUCTION

Water quality monitoring methods for inland lakes can be divided into conventional detection methods and remote sensing detection methods [1]. The conventional monitoring method is to collect and monitor water samples by setting up sampling points or stations manually. Remote sensing monitoring method is used for water quality parameter inversion and remote sensing monitoring through multiple band models and machine learning algorithms using spectral bands of remote sensing images combined with spectral characteristics of water quality elements. However, compared with conventional monitoring methods, remote sensing monitoring of water quality can not only obtain the temporal and spatial distribution of lake water quality parameters, but also has the advantages of fast response, low cost and wide monitoring range [2].

The rapid development of remote sensing technology and artificial intelligence algorithm provides us with many water quality parameter inversion models, such as empirical models, semi-analytical and semi-empirical models, analytical models and machine learning models. However, in the process of remote sensing monitoring of water quality in inland lakes, different inversion models show different applicability and significant along with the spatio-temporal variation of inland water bodies,

different inversion data sources and different inversion parameter objects [3-5]. It is a difficult problem in remote sensing inversion field that how to use different models effectively and reasonably, integrate the simulation results of different models, and get an optimal solution that is closer to the objective real value in the face of the diversity and regional limitations of the model and the complexity of the optical features of inland water. In the field of hydrological forecasting, the method of ensemble modelling is often used to synthesize the results of different models to get the optimal solution [6]. The method of set modelling is applied to remote sensing inversion to improve the accuracy of water quality parameter inversion on the basis of improving the stability of the model.

In this study, four currently popular and universal machine learning models of KNN, ANN, SVR and RF were used to construct the remote sensing inversion model collection of Chl-a, TP, TN concentration in the Donghu lake. The entropy weight method was used to build water quality parameter inversion ensemble model. And an ensemble modelling method of water quality parameter inversion based on model error analysis is proposed. The precision of water quality parameter inversion can be improved by integrating the inversion results of multiple models.

* Corresponding author: 362481958@qq.com

2 Study area

Donghu lake is located in the east of Wuchang, Wuhan city, Hubei province. Its geographical position is between $114^{\circ} 09' \sim 114^{\circ} 39' \text{ E}$ and $30^{\circ} 22' \sim 30^{\circ} 41' \text{ N}$. The total area of it is 32.5km².

Donghu lake is the second largest lake in Wuhan city. Its surrounding area is highly urbanized and industrialized, attracting more and more external population. With the increase of population density and the rapid development of industrial aquaculture and tourism, the eutrophication of Donghu lake is serious. According to the water quality test results from 2017 to 2018 released by Wuhan ecological environmental protection bureau, the eutrophication status of Donghu lake is generally stable in mild eutrophication state due to the environmental governance of Wuhan municipal government. From January to May of each year, the water body basically reaches the standard of medium nutrition, but in the second half of the year, the water body is in the state of mild eutrophication.

3 Ensemble model building of water quality parameters based on remote sensing inversion

3.1 Data

3.1.1 Measured water quality data

Measured water quality data include Chlorophyll a and TN. In this study, according to the outdoor water quality sampling scheme, three large-scale water quality sampling were complete by using water quality sampling equipment and laboratory water quality parameter concentration measuring instrument (46 sampling points in December 20, 2017; 49 sampling points in March 26, 2018; 43 sampling points in December 26, 2018). According to the Gebras criterion and the matching test of measured coordinate points, the outliers were eliminated, and a total of 94 valid sample points were obtained from three sampling.

3.1.2 Remote sensing satellite data

Landsat8 remote sensing images are mainly acquired by USGS download (<https://earthexplorer.usgs.gov>). According to the principle of time synchronization and the principle of remote sensing image availability, the acquisition time of remote sensing image is mainly determined according to the actual sampling time.

Landsat8 multi-spectral remote sensing image in December 20, 2017, March 26, October 26, 2018 were acquired, which is used as remote sensing modelling data of Chlorophyll a concentration inversion model.

3.2 Water quality parameter inversion models

In this study, a total of 8 remote sensing inversion models of Chlorophyll a and TN concentration in Donghu lake were constructed by using four machine learning algorithms of KNN, ANN, SVR and RF. Among them, there are 4 Chl-a inversion models and 4 TN concentration inversion models. Specific models and data are shown in table 1.

Table 1. Model and data

Data	Training sample	The measured values of Chl-a, TN concentration and Landsat8 band reflectivity on December 20, 2017 and October 26, 2018	71
	Test sample	The measured values of Chlorophyll a TN and Landsat8 band reflectivity on March 26, 2018	23
Models	Chl-a inversion model	Constructing inversion model using KNN, ANN, SVR and RF	4
	TN inversion model		4

The waveband reflectance of the pre-processed Landsat8 image was used as the model input, and Chl-a TN water quality parameter concentration was used as the model output for model training. The optimal parameters and model errors of different models are shown in table 2.

Table 2. Table of model optimal parameters

Water quality parameters	Bands of input	Arithmetic	The optimal parameters	R^2
Chl-a	B1-B7	KNN	K=8	0.8884
	B1-B7	ANN	The number of implicit nodes is 6	0.9288
	B1-B7	SVR	C=61, g=0.1	0.9079
	B1-B7	RF	N=700	0.8939
TN	B1-B7	KNN	K=9	0.7969
	B1-B7	ANN	The number of implicit nodes is 8	0.8202
	B1-B4	SVR	C=4.8, g=1	0.8339
	B1-B7	RF	N=600	0.802

According to the analysis of the determination coefficient, training error and test error of the inversion model of different water quality parameters, the fitting

degree and inversion accuracy of Chl-a inversion model and TN concentration inversion model are higher.

Measured values of water quality parameters on March 26, 2018 and the waveband reflectivity of the corresponding Landsat8 image were used to form 23 test samples. The model test was conducted according to the combination of input bands and the optimal parameters of each water quality parameter model in table 2. The correlation between the inversion test results of each model with different water quality parameters and the measured water quality parameters was compared. After testing and verification, Chlorophyll a TN inversion model can participate in ensemble modelling.

3.3 Ensemble model building

The key to constructing an ensemble model is how to establish the relationship between the ensemble model and each inversion model, that is, how to determine the weight of the model. It determines the ability of ensemble model to improve the stability of model and simulation accuracy [7]. In this paper, entropy weight method and error analysis method are used to integrate the inversion model to build the water quality parameter ensemble model.

3.3.1 Remote sensing inversion ensemble modelling of water quality parameters based on entropy weight method

The relative error of each model is used as the evaluation index to calculate the information entropy of the model, determine the weight of the model, and construct the ensemble model. The main calculation process is

(1) The calculation formula of relative error weight of each model is

$$e_i(k) = \frac{\hat{y}_i(k) - y(k)}{y(k)} \quad (1)$$

$$P_i(k) = \frac{|e_i(k)|}{\sum_{k=1}^n |e_i(k)|} \quad (2)$$

(2) The calculation formula of entropy value of each model is

$$H_i = -\frac{1}{\ln n} \sum_{k=1}^n P_i(k) \ln P_i(k) \quad (3)$$

(3) The weight calculation formula of the model is

$$D_i = 1 - H_i \quad (4)$$

$$W_i = \frac{1}{N-1} \left(1 - \frac{D_i}{\sum_{i=1}^N D_i} \right) \quad (5)$$

(4) Ensemble model

$$y = W_1 \times \hat{y}_1(k) + W_2 \times \hat{y}_2(k) + W_3 \times \hat{y}_3(k) + \dots + W_N \times \hat{y}_N(k) \quad (6)$$

Where i is the model number; $y(k)$ is the measured data of water quality parameters of the k th sample; $\hat{y}_i(k)$ is the inversion value of water quality parameters of the k th sample in the simulation results of the i th model; n is the number of training samples; N is the number of models participating in the set modelling.

3.3.2 Remote sensing inversion ensemble modelling of water quality parameters based on error analysis

It is possible to avoid using uniform weights for the model results using the weights determined by the model error analysis in the application of set modelling. The inversion precision and error distribution of the same model are different in different concentration gradients or in different water bodies. Therefore, the weight of the model should be adjusted according to the concentration classification of water quality parameters or the difference of water body area, so as to obtain the optimal solution of water quality parameter inversion.

3.3.2.1 Model error analysis

(1) Model error classification based on concentration gradient of water quality parameters

In this paper, the concentration gradient is divided according to the frequency distribution of training samples with different concentration ranges. The concentration of Chl-a was divided into 8 groups, and the concentration gradient was 0-10 μ g/L, 10-20 μ g/L, 20-30 μ g/L, 30-40 μ g/L, 40-70 μ g/L, 70-100 μ g/L, 100-110 μ g/L and more than 110 μ g/L. TN concentration was divided into 8 groups, and the concentration gradient was 0-0.5mg/L, 0.5-0.7mg/L, 0.7-0.9mg/L, 0.9-1.1mg/L, 1.1-3mg/L, 3-6mg/L and greater than 6mg/L.

On the basis of the concentration gradient partition, the sample data of each gradient was counted. And the RMSE of the inversion results of each model set on the concentration gradient was calculated as the error information of the concentration gradient of the model.

(2) Model error classification based on the location of Donghu sub-lake

According to the sampling location of the training sample, the samples were divided into 5 categories such as Guozheng lake, Tanglinghu lake, Miaohu lake, Yujiahu lake and Houhu lake. On the basis of the location division of this region, the RMSE of the inversion results of each model set on different sub-lakes was calculated and counted as the error information of the location of the model in this sub-lake region.

3.3.2.2 Ensemble modelling process

According to ensemble modeling method based on error analysis and the results of model set error analysis, the modeling process is

(1) Based on the sample data set and remote sensing inversion theory, a remote sensing inversion model set of water quality parameters was constructed. In 3.2, Chl-a and TN concentration inversion model set based on four machine learning algorithms was constructed.

(2) The water quality parameter concentration data of the modelled training samples were graded according to the concentration range and sub-lake region. Error of samples in different concentration gradients or different sub-lake was counted. The RMSE of each model involved in the corresponding partition were calculated. Then the weight distribution of the model can be distributed according to it.

(3) The concentration of water quality parameters at each sample point is calculated using the ensemble model. The result of the ensemble model is calculated using the weight distribution result obtained in the second step, which is the optimal inversion result of water quality parameter concentration. The expression of set modelling optimal inversion result and model weight is

$$X_a = \sum_{i=1}^n K_i x_i \quad (7)$$

$$K_i = \frac{R_i^{-1}}{\sum_{i=1}^n R_i^{-1}} \quad (8)$$

Where R_i is the model error of the i th model. The expression method of model error is root mean square error, relative error, and comprehensive error. In order to reduce the complexity of collection modelling, this paper uses RMSE to describe the error information of different models. K_i is weights.

3.4 Result verification and evaluation

3.4.1 Verification

According to the three modelling methods, the inversion results of 11 ensemble models constructed by each set modelling method are analysed statistically and compared with the inversion statistical results of a single model. As shown in table 3, table 4 and table 5. By analyzing the comparative results, the following conclusions can be drawn:

Table 3.1 Statistical table of entropy weight method Chl-a ensemble modeling results

Model	Statistical parameter	RMSE	R^2
Single model	maximum	18.0955	0.9288
	minimum	14.2615	0.8884
	average	16.4906	0.9047
Double model	maximum	14.5469	0.9236
	minimum	10.8558	0.8686
	average	12.4060	0.9027
Three model	maximum	12.7484	0.9233
	minimum	11.0002	0.8995
	average	11.9140	0.9113
Four model	—	11.6716	0.9154

(1) The remote sensing inversion ensemble model of water quality parameters based on entropy weight method has limited improvement in the inversion accuracy of Chl-a and TN; (2) The inversion effect of ensemble modelling is better than that of single model. (3) The precision of water quality parameter inversion increases with the increase of the number of models involved in ensemble modelling, and the model becomes more stable. (4) The ensemble modelling based on error classification of sub-lake performs well, and the method is proved to be reasonable and effective.

Table 3.2 Statistical table of entropy weight method TN ensemble modeling results

Model	Statistical parameter	RMSE	R^2
Single model	maximum	1.4926	0.8339
	minimum	1.1002	0.7969
	average	1.3223	0.8132
Double model	maximum	1.0314	0.8303
	minimum	0.8445	0.8057
	average	0.9367	0.8196
Three model	maximum	0.9367	0.8821
	minimum	0.8588	0.8532
	average	0.8908	0.8590
Four model	—	0.8621	0.8838

Table 4.1 Statistical table of concentration gradient error classification Chl-a ensemble modeling results

Model	Statistical parameter	RMSE	R^2
Single model	maximum	18.0955	0.9288
	minimum	14.2615	0.8884
	average	16.4906	0.9047
Double model	maximum	14.1367	0.9327
	minimum	10.4417	0.9049
	average	11.3869	0.9168
Three model	maximum	11.7015	0.9343
	minimum	10.0799	0.9166
	average	10.7324	0.9273
Four model	—	10.4194	0.9313

Table 4.2 Statistical table of concentration gradient error classification TN ensemble modeling results

Model	Statistical parameter	RMSE	R^2
Single model	maximum	1.4926	0.8339
	minimum	1.1002	0.7969
	average	1.3223	0.8132
Double model	maximum	1.0512	0.8427
	minimum	0.8462	0.8248
	average	0.9469	0.8177
Three model	maximum	0.9538	0.8525
	minimum	0.8606	0.8121
	average	0.9075	0.8416
Four model	—	1.4926	0.8339

Table 5.1 Statistical table of error classification Chi-a ensemble modeling results in sub-lake area

Model	Statistical parameter	RMSE	R^2
Single model	maximum	18.0955	0.9288
	minimum	14.2615	0.8884
	average	16.4906	0.9047
Double model	maximum	14.3526	0.9321
	minimum	10.2281	0.9015
	average	11.4317	0.9162
Three model	maximum	11.0887	0.9351
	minimum	10.1446	0.9231
	average	10.5151	0.9299
Four model	—	18.0955	0.9288

Table 5.2 Statistical table of error classification TN ensemble modeling results in sub-lake area

Model	Statistical parameter	RMSE	R^2
Single model	maximum	1.4926	0.8339

	minimum	1.1002	0.7969
	average	1.3223	0.8132
	maximum	1.0512	0.8427
Double model	minimum	0.8462	0.8248
	average	0.9469	0.8177
	maximum	0.9538	0.8525
Three model	minimum	0.8606	0.8121
	average	0.9075	0.8416
	maximum	0.9075	0.8416
Four model	—	1.4926	0.8339

3.4.2 Evaluation

The optimal values of the inversion results of the single model and the ensemble model were combined with the measured concentration of water quality parameters in the test samples to make scatter distribution figure (As shown in figure 1, 2 and 3) for the analysis and evaluation of the set modelling results.

The ensemble model based on entropy weight method is simple to determine the weights, but the improvement of model inversion accuracy is limited. The ensemble modelling method based on concentration gradient error classification performs best in improving the inversion accuracy of the model. However, a unified standard cannot be formed due to the great influence of human subjective factors on the division of concentration gradient. The ensemble modelling method based on different sub-lake error classification has the same improvement effect as that of concentration gradient error classification in inversion accuracy. And the weight of the model is determined mainly according to the error distribution in the spatial position, which is easy to form a unified standard. However, this modelling method is only applicable to inland urban lakes with obvious difference in water quality distribution characteristics of sub-lakes like Wuhan Donghu lake. The precision effect is unknown for the other lakes. Therefore, in this paper, the application of the Donghu model is mainly based on the concentration gradient error classification of water quality parameters. For the inversion results of remote sensing images in different periods, a uniform concentration gradient classification standard is adopted.

Figure 1.1 Chl-a ensemble modeling test sample inversion results based on entropy weight method

Figure 1.2 TN ensemble modeling test sample inversion results based on entropy weight method

Figure 2.1 Chl-a ensemble modeling test sample inversion results based on concentration gradient error classification

Figure 2.2 TN ensemble modeling test sample inversion results based on concentration gradient error classification

Figure 3.1 Chl-a ensemble modeling test sample inversion results based on different sub-lake error classification

Figure 3.2 TN ensemble modelling test sample inversion results based on different sub-lake error classification

4 Conclusion

In this paper, combined with large-scale water quality sampling data and Landsat8 remote sensing image data, four machine learning algorithms of KNN, ANN, SVR and RF were used to construct 8 Chl-a, TN concentration inversion models. There are four inversion models for Chl-a and TN concentration parameters.

The classical entropy weight method of ensemble modelling was introduced to construct the ensemble model with weighted processing method for different models to determine the validity and feasibility of ensemble modelling. On this basis, the inversion results of Chlorophyll a and TN concentration were analyzed on the different concentration gradient and sub-lake area. This paper proposes a weight processing method based on error classification in concentration gradient or different sub-lake. Empirical comparison shows that entropy weight method can improve the inversion accuracy to some extent, but the improvement space is limited. However, in the verification of the two methods of ensemble modelling based on error analysis, with the increase of the number of models, the inversion accuracy keeps increasing. Compared with the optimal results of a single model, the determination coefficient (R^2) of Chlorophyll a and TN concentration inversion results was increased from 0.9288 to 0.9313 and from 0.8339 to 0.8838, and the root mean square error was decreased from 14.2615 μ /L to 10.4194 μ /L and from 1.1002mg/L to 0.8621mg/L, taking the ensemble model based on the classification of concentration gradient error as an example. It can be seen that applying the idea of ensemble modelling to remote sensing inversion of water quality parameters can overcome the limitations and shortcomings of single model inversion, integrate the information of different models, make multiple models complement each other in water quality inversion, improve the precision of water quality parameters inversion, and obtain the optimal value of inversion results.

Funding

This research was funded by the National Natural Science Foundation of China (No. 51579108).

References

1. Haidong, Z. *Retrieval of Chlorophyll a in inland lake based on MERIS data-a case study on Poyang Lake*. Master's Thesis. Central China Normal University (2016)
2. Hui, Z. *Retrieval and Analysis Water Quality Parameters in Poyang Lake Based on Multi-source Remote Sensing Data*. Doctoral Thesis. Nanchang University (2011)
3. Xiaobi, Z. *Regression of Chlorophyll content based on hyperspectral remote sensing data in Chaohu Lake*. Master's Thesis. Anhui University of Architecture (2012)
4. Yunmei L. etc. *Three Wavelengths Factor Model of Chlorophyll Concentrations Inversion Based on Different Spectral Dominating Factors for Inland Lake* Environmental sciences **31**(12):2873-2881. (2010)
5. Fucheng, L. and Xia L. *Lianyungang, China. Chlorophyll Concentration Estimation Model Based on Canopy Spectra of Spartina Anglica* IITA Association **4**(2010)
6. Zhe, Y. Denghua, Y. Zhiyong, Y. Jun, Y. *Ensemble model and its application in runoff simulation and forecast* J Hydraul Eng-Asce **45**(03):351-359 (2014)
7. Menezes, L. M. D. Bunn D W. Taylor J W. *Review of guidelines for the use of combined forecasts*. Eur J Oper Res **120**(1):190-204 (2000)