Research on algae blooms forecasting based on the multivariate data driven method: a case study of the Chaohu Lake

Xiang bo¹, Zhao Yanchuang¹,², Wang Xinyuan¹, Zong Xin¹,²

¹Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, 100094 Beijing, China
²University of Chinese Academy of Sciences, 100049 Beijing, China

E-mail: xiangbo@radi.ac.cn

Abstract. The data driven is one of the main methods for forecasting algae bloom, which requires lots of continuous and accurate monitoring data. It is an effective way to increase sample data size by combining in-situ, and remote sensing data. The Chaohu Lake was taken as the case study. Based on water quality data (TLI), meteorological data (sunshine duration, temperature, wind speed, wind direction) and bloom grade data, provided respectively by remote sensing and in-situ monitoring, an artificial neural network was employed to build empirical data-driven models. The model accuracy was evaluated by algae bloom grade recognition rate and bloom trend recognition rate. The results showed that the bloom grade recognition rate of model driven by remote sensing data was better than others. Bloom trend recognition rate of model driven by in-situ data is higher than others. These results provide some insights for algae bloom forecasting.

1. Introduction
Algae bloom is one of the most widespread environmental problems of inland water in the world. It is important to develop methods for forecasting algae blooms, which is useful for water management and protection. There are mainly two kinds of methods for forecasting algae bloom: mechanical models and data driven models. The mechanical methods predict algae blooms by modeling the processes of the hydrological, hydrodynamic and algae growth. However, the mechanical methods, based on the processes of hydrological, hydrodynamic and algae growth, need a number of parameters, such as initial conditions and ecological variables. It is very complex to get all data, which limits the application of model driven methods on forecasting algae blooms.
The data driven approaches for forecasting algae blooms by constructing empirical-statistical models based on multi-source heterogeneous data, such as in-situ monitoring data, automatic monitoring data, remote sensing data. ANN (artificial neural network), Decision Tree, GA (genetic algorithm), SVM (support vector machine) are common methods to perform data driven processes. Wei et al. (2001) simulated the relationship between algae density and multi water quality factors, through ANN and predicted the outbreak of algae bloom. Lee et al. (2003) forecasted water quality factors and algae bloom of coastal waters in Hong Kong. Zeng et al. (2007) combined decision tree modeling and non-linear regression, to predict algae bloom of ‘six lakes’ in Beijing. Muttil et al. (2005) predicted algae bloom in Causeway Bay, Hong Kong by using GA. SVM can solve the problem of the empirical trial resulting from ANN. Lu et al. (2006) established the water safety warning evaluation model by SVM.

However, in order to ensure the accuracy and reliability of the data driven models, lots of continuous and accurate monitoring data are required. Datasets in-situ is often used to establish the data driven model. This type of data is usually intuitive and with high accuracy, but the collection of these data is higher time-consuming, laborious and costly. Besides, the information obtained by in-situ monitoring are usually at station scale. Remote sensing data can provide global detail information about eutrophication and algae blooms, quickly at different scales with low cost. Thus, it is an effective way to increase sample data size by combining in-situ data and remote sensing data. Here, we took Chaohu Lake as the case of study. Based on water quality data (TLI), meteorological data (sunshine duration, temperature, wind speed, wind direction) and bloom grade data, provided respectively by remote sensing and in-situ monitoring. An artificial neural network method was employed to build empirical data driven models. Then algae bloom grade recognition rate and bloom trend recognition rate were used to evaluate the model accuracy. We hope this study can provide some insights for algae bloom forecasting.

2. Study area

Figure 1. Location of Chaohu Lake

The Chaohu Lake is located between the Yangtze and the Huaihe rivers. It belongs to the north drainage of the Yangtze River downstream and is one of the five largest freshwater lakes in China. Fig. 1 shows the location of the Chaohu Lake. The study area is located in the humid subtropical monsoon climate zone with abundant rainfall. The annual average temperature is 15-16°C. The main wind
direction in winter is northeast and southeast in summer. The average annual precipitation is about 1100mm. The mean water depth is 2.69 m covering an area of 780 km$^2$. Phosphorus-rich strata are widely distributed in the north shore of the basin and phosphorus is input into the lake along runoff. It is not only the main drinking water source for Hefei and Chaohu cities, but also has the function of shipping, fishing, tourism, etc. the water quality of the Chaohu Lake is of the utmost importance.

3. Materials and methods

3.1 Data source
Water quality data from 12 monitoring sites (Figure 1) from 2013 to 2015 were provided by Chaohu Authority. Chlorophyll a (Chl–a), total phosphorus (TP), total nitrogen (TN), transparency (SD), chemical oxygen demand (COD), and algae density were included. The monitoring frequency is once a day from May to October each year and once a week for the rest of the months. The meteorological data were obtained from the National Meteorological Information Center, China Meteorological data network. The accuracy of sunshine duration, average temperature, and wind speed are 0.1h, 0.1 °C and 0.1m/s respectively. The monitoring frequency is once a day. The algae bloom coverage data between 2013 and 2015 were obtained from Chaohu cyanobacteria bloom inversion product datasets, which were provided by the National Science and Technology Platform - National Earth system data sharing platform - Lake - Watershed Science Data Center. MODIS surface reflectance product (MOD09) was obtained from the National Aeronautics and Space Administration, (NASA) website.

3.2 Data preprocessing

3.2.1 Grade of algae bloom feature. In agreement with the actual monitoring of Chaohu, the method grading algae bloom feature in this study was according to that employed by the China Environmental Monitoring Center and the Anhui Environmental Monitoring Center Station. As shown in Table 1, the algae feature in the Chaohu Lake was divided into five grades. The algae density threshold is employed by in-situ data and the algae coverage threshold are employed by the remote sensing data. In order to determine the algae bloom coverage for each site, the averaged value of 3*3 grid, fixing the station as center, were calculated.

| Algae bloom feature | Algae density ($10^4$ /L) | Algae coverage (%) |
|---------------------|--------------------------|--------------------|
| No bloom            | <200                     | ≤5                 |
| Sporadic bloom      | ≥200                     | ≤10                |
| Local bloom         | ≥500                     | ≤40                |
| Regional bloom      | ≥1000                    | ≤60                |
| Global bloom        | ≥10000                   | >60                |

3.2.2 Trophic Level Index (TLI). The Trophic Level Index is a general indicator to evaluate the eutrophication status of inland water. It is also one of the most important quality factors for algae
bloom prediction model inputs. TLI was calculated based on in-situ data according to the following formulas (Wang et al., 2002; Jin et al., 1990).

$$TLI(\sum) = \sum_{j=1}^{n} W_j \times TLI(j) \quad (1); \quad \sum_{j=1}^{n} W_j = \sum_{j=1}^{n} r_{ij}^2$$

$TLI(j)$ is the $j$th composite indicator with the corresponding weight $W_j$. The $r_{ij}$ value is the correlation coefficient for the relationship between the reference chlorophyll concentration and each indicator (Table 2). The TLI of all the observation points was calculated from 5 components, including chemical oxygen demand (COD), total phosphorus (TP), total nitrogen (TN), chlorophyll a ($Chl-a$), and Secchi depth (SD). Formulas for the TLI of each component are given below:

$$TLI(chl-a) = 10(2.5 + 1.086 \ln(chl-a)) \quad (2)$$
$$TLI(TP) = 10(9.436 + 1.624 \ln(TP)) \quad (3)$$
$$TLI(TN) = 10(5.453 + 1.694 \ln(TN)) \quad (4)$$
$$TLI(SD) = 10(5.118 + 1.94 \ln(SD)) \quad (5)$$
$$TLI(COD) = 10(0.109 + 2.661 \ln(COD)) \quad (6)$$

Equations 2 to 6 are empirical regression equations based on a survey of eutrophication levels from more than 20 lakes in China (Jin et al., 1990).

| $r_{ij}$ | $r_{ij}^2$ |
|---------|-----------|
| 1       | 1         |
| 0.84    | 0.7056    |
| 0.82    | 0.6724    |
| -0.83   | 0.6889    |
| 0.83    | 0.6889    |

Table 2. Correlation between $Chl-a$, and other substances influencing TLI

Please refer to our previous research work about the procedure of TLI calculation based on remote sensing data (Xiang et al., 2015), which will not be described in detail here.

3.3 Multivariate data driven method

Based on water quality data (TLI), meteorological data (sunshine duration, temperature, wind speed, wind direction) and bloom grade data; an artificial neural network was employed to build an empirical model. Meteorological data and bloom grade data were used as the input and output layer respectively. Neural networks are a machine learning algorithm that imitates brain processes and was originally developed to solve non-linear problems like over-fitting, pattern recognition, clustering, and time-series prediction (Mitchell, 1997). They have been successfully applied in environmental sciences (Sattari et al., 2012; Krasnopolsky and Chevallier, 2003; Krasnopolsky and Schiller, 2003; Song et al., 2014a). In this study, the neural network was designed as a feed-forward type and divided into three layers (input layer, hidden layer, and output layer). Back propagation algorithm was used to determine the parameters in the network. In-situ data, remote sensing data and their combination between 2013 and 2014 were selected to perform the driven process. In order to prevent over-fitting, the number of nodes in hidden layer was adjusted based on the results of verification.
3.4 Evaluation of model accuracy

In-situ data, remote sensing data, and their combination in 2015 were selected to evaluate the model accuracy respectively. Bloom grade recognition rate and bloom trend recognition rate were used to assess the predictive effect of different types of data driven models. To calculate bloom trend recognition rate, assuming that it is correct when the difference between forecast bloom grade and the true grade is limited in one grade.

4 Results and analysis

Fig. 2 shows the accuracy of model driven by different types of data (in-situ data, remote sensing data as well as their combination). The blue, red and green solid lines, are respectively the bloom grade recognition rate of remote sensing data (RS) driven model, the in-suit data drive model, and their combination data driven model. The blue dotted, red dashed and green dotted lines are respectively corresponding to blooms trend recognition rate.

As can be seen in Fig. 2, there are totally 9 stations whose bloom grade recognition rate of remote sensing data (RS) driven model is not lower than that of in-situ data. 5 stations are in the western half lake and the others are in the eastern half lake. There are only 2 stations whose bloom grade recognition rate of collection data driven model is higher than that of the single dataset. 5 stations are in the western half lake and the others are in the eastern half lake. The in-situ data have more effect on the blooming trend recognition rate than the remote sensing data and the collection data. The bloom trend recognition rate of model driven by collection data fluctuates between that of the single data.

Table 3. Statistical results for accuracy of models driven by different types of data

| Type of data | Bloom grade recognition rate (%) | Standard deviation (%) | Bloom trend recognition rate (%) | Standard deviation (%) |
|--------------|----------------------------------|------------------------|----------------------------------|------------------------|
| In-situ      | 47.26                            | 21.56                  | 71.67                            | 10.68                  |
| RS           | 51.64                            | 20.34                  | 64.73                            | 17.06                  |
| I&R          | 49.05                            | 19.60                  | 68.69                            | 9.10                   |

The statistical results for accuracy of models driven by different types of data based on 12 stations are shown in Table 3. For bloom grade recognition rate, the accuracy of model driven by remote sensing
data is better than others. According to bloom trend recognition rate, the accuracy of model driven by in-situ data is higher than others. The reasons are that the model driven by remote sensing data, forecast algae bloom grade by algae coverage and the model driven by in-situ data, forecast algae bloom grade by algae density.

Algae coverage reflects the extent of the lake surface covered by algae blooms, which is a visible trait; while algae density reflects the degree of enrichment of algae in the lake, which is a latent trait. Visible traits characterize the current situation of algae bloom and are more suitable to forecast bloom grade. Latent traits characterize the possibility of algae bloom and are more suitable to forecast bloom trends. The algae bloom grade in the western half lake is more than that in the eastern half lake; the difference between visible and latent features is more than that in the eastern half lake. Thus, the difference of effect of data type on model accuracy is more significant in the western half lake.

5. Conclusion
The data driven is one of the main methods to forecast algae bloom. The approach requires lots of continuous and accurate monitoring data. It is an effective way to increase sample data size by combining in-situ and remote sensing data. The Chaohu Lake was the case of study. Based on water quality data (TLI), meteorological data (sunshine duration, temperature, wind speed, wind direction) and bloom grade data, provided respectively by remote sensing and in-situ monitoring. An artificial neural network was employed to build empirical data driven models. The model accuracy was evaluated by algae bloom grade recognition rate and bloom trend recognition rate. The results showed that the bloom grade recognition rate of model driven by remote sensing data was better than others. Bloom trend recognition rate of model driven by in-situ data is higher than others. These results provide some insights for algae bloom forecasting.

6. References
[1] Wei B, Sugiura N and Maekawa T 2001 Water Res. 35(8) 2022–2028.
[2] Lee JHW, Huang Y, Dickman M, et al. 2003 Ecol. Model. 159(2-3): 179–201.
[3] Zeng Y, Yang Z and Liu J 2007 Adv. Water Sci. 18(1): 79-85.
[4] Muttil N and Lee JHW 2005 Ecol. Model. 189(3-4): 363–376.
[5] Lu M, Zhang Z and Feng B 2006 Comput. Eng. 32 (15): 44-45.
[6] Wang M, Liu X and Zhang J 2002 Environ. Monit. China 18(5): 47–49.
[7] Jin X, Tu Q, Zhang Z, Jiang X, Wang Y, Zhu X, Shu J, Xu N, Huang C and Xu R 1990 Specifications for Lake Eutrophication Surve. 2nd Edition. Beijing: China Environment Science Press: 291.
[8] Xiang B, Song J, Wang X and Zhen J 2015 Water SA 5: 753-761.