Analysis of cyanobacteria bloom in the Waihai part of Dianchi Lake, China

Hu Sheng a, Hui Liu a,b, Cuiyu Wang a, Huaicheng Guo a,⁎, Yong Liu a, Yonghui Yang a

a College of Environmental Sciences and Engineering, Peking University, Beijing 100871, China
b Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong, China

A R T I C L E   I N F O

Article history:
Received 15 November 2011
Received in revised form 17 March 2012
Accepted 20 March 2012
Available online xxxx

Keywords:
Missing data
Multiple imputation
Logistic regression
Cyanobacteria bloom

A B S T R A C T

Blue-green algae (BGA) bloom is a typical phenomenon in eutrophied lakes. However, up to now, no environmental mechanism has been commonly accepted. Systematic and complete data sets of BGA blooms and environmental factors without any missing data are rare, which seriously affected previous studies. In this study, a bootstrapping based multiple imputation algorithm (EMB) was first applied to reconstruct a complete data set from the available data set with missing data, hence forming a basis for quantitatively relating BGA bloom to contributing factors. Then, the probability of BGA bloom outbreak was simulated using a binomial (or binary) logistic regression model, which is an effective tool for recognizing key contributing factors. The results suggest that 1) the outbreak frequency or probability of BGA bloom tends to first increase and then decrease with a turning point between June and September each year; 2) air temperature, relative humidity, and precipitation were significant positive factors correlated with outbreak frequency, whereas wind speed and the number of sunshine hours were negative factors; 3) water temperature had a strong positive effect on the probability of BGA bloom outbreak, whereas other water quality factors, such as concentrations of organics and nutrients, were not so significant. However, water quality factors, such as NO3−, SD, pH, NH4−, COD and DO, still need to be concerned, which had a potential to aggravate the outbreak of BGA bloom in Dianchi Lake, if they were out of control.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Eutrophication in lakes or reservoirs can lead to blue-green algae (BGA) blooms, which are detrimental to aquatic ecosystems and often pose potentially serious health threats to people with access to the waters (Heisler et al., 2008; Maestrini and Granelli, 1991; Pitois et al., 2001). When a BGA bloom breaks out, cyanobacteria will reproduce at a rapid rate and multiply quickly in a short time, turning the lake or reservoir green locally or even globally. In China, the public and government are passionately concerned with the relative events, and a number of experts and scholars are very interested in forecasting when and at what probability and how long it will take for the BGA bloom to break out, with the help of remote sensing technology which provides a fast and large-scale monitoring. These researches usually pay attention to the surface bloom phenomenon quantified by whether or not a bloom breaks out or the area of the bloom, which is easily recognized by eyes and remote sensing images. The quantifying method is more directly accepted by the public and government, other than biomass commonly quantified by chlorophyll-a (Chl-a). Generally, the concentration of Chl-a represents the degree of a BGA bloom (Hou et al., 2004), but it is unclear what Chl-a level represents in the emergence of a BGA bloom, especially a surface BGA bloom, which is the concern of most citizens. On the other hand, data interpreted by remote sensing imaging have been used to indicate the status of BGA blooms (Xie et al., 2010). In our study, we prefer to use the data interpreted by remote sensing imaging, so as to build up the relationship between key influencing environmental factors (including nutrient concentration and weather conditions) and the eco-environmental hazards, which indicated whether a BGA bloom breaks out in a given day (dichotomous variable).

Many ecological risk assessment (ERA) methods, such as linear regression, artificial neural networks, Bayesian networks, aquatic system models, etc. (Arhonditsis et al., 2003; Borsuk et al., 2004; Hamilton and Al, 2008; Lek, 1999; Wong et al., 2007), have supported these kinds of researches under uncertainty, which is mostly expressed by the probability distribution of the ecological effects. Two types of uncertainty are always taken into consideration, i.e., epistemic uncertainty and aleatory uncertainty (Chen, in press; Chen et al., 2011; Reckhow, 1994). When analyzing the data sets in real world, we will inevitably encounter the problem of missing data, which is a big challenge to the unbiasedness, accuracy, robustness of the results, and it even has an influence up on the analyzation process. This problem results in another kind of uncertainty, derived from imperfect data sets, which is often easy to be ignored. In China, data imperfection is common, because historically monitored data is lacking in observations in most research area, and remote sensing technology tends to be interrupted by clouds.

Therefore, an unbiased and efficient method is urgently needed to overcome the data limitations and to derive a reliable relationship between the BGA blooms and the main influencing factors. Missing data are a frequent complication in any real-world study, and thus there are
conventional methods to tackle missing data (Allison, 2001; Enders, 2010; Little and Rubin, 2002), such as 1) deletion, including listwise deletion and pairwise deletion (Acock, 2005; Schafer and Graham, 2002); and 2) imputation, including single imputation such as arithmetic mean imputation (Wilks, 1932), regression imputation (Buck, 1960), and hot-deck imputation (e.g., Ford, 1983), and multiple imputation such as imputation-posterior (IP) (Tanner and Wong, 1987), expectation-maximization (EM) (Dempster et al., 1977), EM with sampling (EMs), EM with importance resampling (EMis), and EM with bootstrapping (EMB) (e.g., Honaker and King, 2010; King et al., 2001). Notably, multiple imputation was first proposed by Rubin (1976) to impute “m” values for each missing item and create “m” completed data sets, which reflect the uncertainty levels of the missing values filled with different imputations. The EMB algorithm, as an unbiased, efficient, and fast method of multiple imputation, is more robust for distributional and small sample problems.

In this study, we attempt to generate an analysis framework to serve the ubiquitous problem of missing data in ERA simply by application of the EMB algorithm to impute the missing data and conduct an uncertainty analysis, together with binomial logistic regression (BLR) modeling to identify the response relationship between BGA blooms and the influencing environmental factors (such as water quality factors and weather condition factors), to explore the characteristics and laws for outbreak of BGA blooms in the Waihai part of Dianchi Lake, China.

2. Materials and methods

2.1. Site description

Dianchi Lake Basin is located in Kunming, Yunnan Province, in the middle of the Yungui Plateau in southwestern China. The lake basin has an area of 2920 km² and was divided into 16 sub-basins for our research (Fig. 1 left). The lake basin is in a northern sub-tropical moist monsoon climate, with average annual precipitation of approximately 1004 mm, average temperature of approximately 14.5 °C, 2470 sunshine hours per year, relative humidity of 74%, and mean wind speed of about 2.5 m/s. Dianchi Lake is southwest of Kunming City in the basin and has a lake area of 309 km² when the water level is 1887.4 m. The lake is 40.4 km long from south to north and 7 km wide from east to west. The lakeshore is 163.2 km long, and the mean water depth is 4.4 m. Because the lake is downstream of the city, large amount of municipal sewage, industrial wastewater, and high nonpoint loads of nutrients, such as nitrogen and phosphorus, are discharged into the lake, particularly in the rainy season. After the establishment of the Haigeng Dam in 1996, Dianchi Lake was separated into two parts, i.e., Caohai in the north and Waihai in the south (Fig. 1 right). The north part of Waihai became a dead space with low water flow and high pollutant concentration (Pan et al., 2004). As a result, the lake water is heavily polluted and suffers from severe eutrophication, especially in the north part of Waihai, where BGA blooms are experienced almost annually since the 1990s.

In this study, we attempt to explain the relationship between outbreak frequency (or probability) of cyanobacterial bloom each month and environmental factors, such as weather conditions and water quality factors during the research period from 1998 to 2009. However, The BGA bloom status (whether or not the bloom outbreak in a given day) was acquired by interpreting remote sensing images, but these data were only obtained daily from April to October between 2004 and 2008 and were easily disturbed by clouds. Therefore, the BGA bloom status data were dichotomous (noted as 0–1 variable, 1 means the BGA bloom outbreak, 0 means the opposite) with quite a few missing values. Weather factors were acquired from the Kunming weather station of the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/). We chose 17 weather condition factors: mean pressure (MP), daily highest pressure (DHP), daily lowest pressure (DLP), mean temperature (MT), daily highest temperature (DHT), daily lowest temperature (DLT),

![Fig. 1. The Dianchi Lake Basin and monitoring sites on Dianchi Lake.](http://example.com/fig1.png)

Please cite this article as: Sheng, H., et al., Analysis of cyanobacteria bloom in the Waihai part of Dianchi Lake, China, Ecological Informatics (2012), doi:10.1016/j.ecoinf.2012.03.007
mean relative humidity (MRH), lowest relative humidity (LRH), precipitation (20–20 h) (PP), mean wind speed (MWS), maximum wind speed (10-minute mean wind speed) (MAXWS), direction of maximum wind speed (DMAXWS), extreme wind speed (EWS), direction of extreme wind speed (DEWS), number of sunshine hours (SH), global radiation (GR), and net total radiation (NTR). In addition, there are eight conventional monitoring sites at Waihai Dianchi Lake (Fig. 1 right). The water quality at the Huizhong monitoring site is the poorest. This site is north of Waihai, which is a heated area for BGA blooms; thus it is conjectured that the reason is that the Huizhong monitoring site has the highest nutrient conditions which promote the growth and reproduction of blue-green algae, resulting easily in an outbreak of BGA bloom when the weather condition is suitable. But what nutrient conditions (or water quality factors) are more critical for BGA bloom? To solve this problem, we obtained 12 water quality factors from the Huizhong monitoring site with the monitoring frequency once a month, including Chl-a, total phosphorus (TP), total nitrogen (TN), ammonia nitrogen (NH4–N), nitrate nitrogen (NO3–N), permanganate index (CODMn), chemical oxygen demand (COD), biochemical oxygen demand (BOD), dissolved oxygen (DO), Secchi disk depth (SD), pH, and water temperature (TEMP).

Apparent, the various data sets collected for our research are with different temporal scopes and scales, even with some annoying missingness, which is a major obstacle for revealing the relationship between key factors and outbreak of BGA bloom by statistical analysis. Then, in our research, we imputed the missing data by EMB algorithm which is firstly introduced in the political sciences, and built up a group of binomial logistic regression models to find the probable key factors determining the outbreak of BGA bloom.

2.2. Research methods

2.2.1. EMB algorithm for missing data imputation

The EMB algorithm is a bootstrap-based multiple imputation algorithm that imputes "m" values for each missing cell in a data matrix and creates "m" "completed" datasets (Honaker et al., 2010). It is an unbiased and efficient method with two basic assumptions: one is that the complete data set (both observed and unobserved) is multivariate and normal; the other is that the data are missing at random (MAR). If we denote the (n×k) data set as D (with observed part Dobs and unobserved part Dmis), then the first assumption is written as

\[ D \sim N_p(\mu, \Sigma). \]  

MAR is one type of mechanisms for missing data (others are 'missing completely at random,' MCAR, and 'missing not at random,' MNAR) and requires that the missing pattern depends on the observed data Dobs, not the unobserved data Dmis. If we introduce the matrix M to indicate whether or not a cell is a missing data, with cells \( m_{ij} = 1 \) if \( d_{ij} \in D_{mis} \) and \( m_{ij} = 0 \) otherwise, then the second assumption can be defined as

\[ p(D|M) = p(D|M^{obs}). \]  

The EMB algorithm has two steps: 1) bootstrap the data to simulate estimation uncertainty, and 2) run the EM algorithm to identify the posterior mode for the bootstrapped data. More specifically, we supposed the complete-data parameters \( \theta = (\mu, \Sigma) \). First, we bootstrapped the missing data "m" times, and for each initiated \( \theta \) with a prior distribution, drew \( D_{mis} \) from the distribution as follows:

\[ D_{mis} \sim p(D_{mis}|D^{obs}, \mu, \Sigma). \]

Then, we imputed \( D_{mis} \) into the data sets to form the "completed" data sets, and thus determined the maximum likelihood estimator of \( \theta \), which is the priori for the next iteration until it is convergent. The likelihood function is given as

\[ \theta = (\mu, \Sigma) \sim p(D^{obs}, D_{mis}). \]

The assumption above is not so strict because the condition of multivariate normal is hardly satisfied, such as the 0–1 variable that exists in the data set. A previous study indicated that this assumption works well even when discrete or non-normal variables are included (Schafer and Olsen, 1998). That is, with some transformations, nearly all kinds of variables can be imputed. In addition, \( m = 5 \) for the number of imputed datasets is adequate unless the rate of missing data is very high. The EMB algorithm was implemented by the Amelia II program in R.

2.2.2. BLR classification models

BLR is a form of multiple regression that is applied when the dependent variable is dichotomous. It is a probabilistic non-linear regression and used for predicting the probability of occurrence of a 0–1 event by fitting data to a logit function, so that it recognizes the impact of predictor variables that may be either numerical or categorical (Hosmer and Lemeshow, 2000). The basic assumption is that there exists a continuous variable \( y_i \) in the real number field to determine whether \( y_i \) occurred or not; that is,

\[ y_i = \begin{cases} 1, & y_i^* > c \\ 0, & y_i^* \leq c \end{cases} \]

where \( c \) is a threshold (generally 0). When \( y_i \) equals 1, the event happens; otherwise the event does not occur.

Assume that dependent variable \( y_i^* \) and predictor variable \( x_i \) have a linear relationship as follows:

\[ y_i^* = \alpha + \beta x_i + e_i. \]

So that

\[ P(y_i = 1|x_i) = P(y_i^* > 0) = \frac{e^{-\alpha - \beta x_i}}{1 + e^{-\alpha - \beta x_i}} = F(\alpha + \beta x_i) \]

where \( F \) is the distribution function of a random variable \( -e_i \), which is symmetrical. When \( F \) is logistically distributed, it can be written as

\[ \log(\frac{p_i}{1-p_i}) = \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i. \]

with an odds ratio of

\[ \text{odds} = \frac{p_i}{1-p_i}. \]

A maximum likelihood estimation is generally used to estimate the parameters \( \alpha \) and \( \beta \). For each sample

\[ P(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}. \]

When the samples are isolated, the likelihood function is

\[ L(\alpha, \beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \]

and the log-likelihood function is

\[ \ln L(\alpha, \beta) = \sum_{i=1}^{n} y_i \ln p_i^{y_i}(1-p_i)^{1-y_i} - \sum_{i=1}^{n} (y_i (\alpha + \beta x_i) - \ln(1 + e^{\alpha + \beta x_i})). \]
Then, we solve for the parameters $\alpha$ and $\beta$ by letting partial derivatives of log-likelihood function be zero. The Newton–Raphson iterative algorithm is always used for the solution.

2.2.3. Analytical process

We can generate an analytical process for BGA blooms with missing data in three basic steps (Fig. 2): 1) impute “m” values for each missing cell and create “m” data sets using the EMB algorithm. 2) Use BLR modeling to identify the relationship between the BGA bloom status and environmental factors. 3) Combine the results from each analysis to compute a multiple imputation estimate and standard error. This method is as follows (Honaker et al., 2010):

$$q_j = \frac{1}{m} \sum_{j=1}^{m} q_j$$  \hspace{1cm} (13)

$$\text{SE}(q_j)^2 = \frac{1}{m} \sum_{j=1}^{m} \text{SE}(q_j)^2 + S_q^2(1 + 1/m)$$  \hspace{1cm} (14)

where $q_j$ ($j=1, \ldots, m$) is the average of the $m$ separate estimates, $\text{SE}(q_j)^2$ is the estimated variance (squared standard error) of $q_j$ from the data set $j$, and $S_q^2$ is the sample variance across “m” point estimates:

$$S_q^2 = \frac{1}{m-1} \sum_{j=1}^{m} (q_j - q)^2.$$  \hspace{1cm} (15)

3. Results

3.1. Missing data imputation

As mentioned in Section 2.1, the weather conditions and water quality factors have different temporal precisions (the former is once a day and the latter is once a month). Our main purpose was to identify the relationship between water quality factors and the status of BGA blooms. However, the BGA bloom status data were collected daily, which did not match the monthly collection of water quality data; in addition, some bloom status data were missing. Thus, we had to impute the missing data. We imputed the weather condition and

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Results of missing data imputation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types</td>
<td>Factors</td>
</tr>
<tr>
<td>Weather condition factors</td>
<td>MP</td>
</tr>
<tr>
<td></td>
<td>DHP</td>
</tr>
<tr>
<td></td>
<td>DLP</td>
</tr>
<tr>
<td></td>
<td>MT</td>
</tr>
<tr>
<td></td>
<td>DHT</td>
</tr>
<tr>
<td></td>
<td>DLT</td>
</tr>
<tr>
<td></td>
<td>MBH</td>
</tr>
<tr>
<td></td>
<td>LRH</td>
</tr>
<tr>
<td></td>
<td>PP</td>
</tr>
<tr>
<td></td>
<td>MWS</td>
</tr>
<tr>
<td></td>
<td>MAXWS</td>
</tr>
<tr>
<td></td>
<td>EWS</td>
</tr>
<tr>
<td></td>
<td>SH</td>
</tr>
<tr>
<td></td>
<td>GR</td>
</tr>
<tr>
<td></td>
<td>NTR</td>
</tr>
<tr>
<td>Water quality factors</td>
<td>Chl-a</td>
</tr>
<tr>
<td></td>
<td>TP</td>
</tr>
<tr>
<td></td>
<td>TN</td>
</tr>
<tr>
<td></td>
<td>NH₄-N</td>
</tr>
<tr>
<td></td>
<td>NO₃-N</td>
</tr>
<tr>
<td></td>
<td>COD</td>
</tr>
<tr>
<td></td>
<td>COD</td>
</tr>
<tr>
<td></td>
<td>BOD</td>
</tr>
<tr>
<td></td>
<td>DO</td>
</tr>
<tr>
<td></td>
<td>SD</td>
</tr>
<tr>
<td></td>
<td>pH</td>
</tr>
</tbody>
</table>

Var., variance; prop., proportion; coef., coefficient (The bold emphasis in the table indicated the variables with a great mass of data missing).
BGA bloom status data by creating five data sets. Then, we matched the imputed data sets with the water quality data and imputed five data sets for each. Namely, 25 water quality data sets were created.

Table 1 lists the results of the missing data imputation. It illustrates that only some of the factors had missing data (the sample variance was 0, or the missing proportion was 0.00%) and factors with many missing data (e.g., EWS with 50.7% missing data and NO$_3$–N with 72.92% missing data) made small variance contributions (variance proportions: EWS = 0.01% and NO$_3$–N = 2.31%). Thus, the coefficient of variation almost entirely originated from the observed data, not the missing data. In other words, imputation did not incorporate more information into the dataset, which could have resulted in biases. Additionally, the variance proportion is the proportion of sample variance of total variance, which represents additional information.

3.2. Analysis of weather condition factors

The frequency of BGA bloom outbreak is represented by the proportion of outbreak days in 1 month. Therefore, we calculated the frequency of BGA bloom outbreak for each month from 1998 to 2009 using the imputation results for weather condition data and BGA bloom status data to create five such data sets. The mean, maximum, and minimum of the frequency of BGA bloom outbreaks each month during 1 year were also calculated and displayed in Fig. 3 to analyze the trends. In Fig. 3, the line with black points represents the mean, and the gray region formed by the minimum and the maximum is the scope of the mean. The frequency of BGA bloom outbreaks from January to December first increased and then decreased. The peak point occurred between June and September. The scope of each point seemed narrower from April to October than in other months, which may be because the observed data ranged from April to October, whereas data for the other months were missing and imputed, which could have caused greater variance.

To identify the relationship between weather condition factors and the frequency of BGA bloom outbreaks, we obtained the means of each weather condition factor by month and calculated the correlation coefficient between these factors and the frequency of BGA bloom outbreak. Three methods are available for calculating the correlation coefficient: (a) for all data, (b) for data classified by year, and (c) for data classified by month. Method (a) was used to provide...
an overview of the correlation, and methods (b) and (c) were used to determine if the correlation was mainly influenced by variation during the month or during the year. Specifically, when calculating the correlation coefficient for each year [method (b)], the correlation was calculated by month and the distribution of correlation coefficients could be obtained for each year, as with method (c). A clear result is displayed in Fig. 4.

Fig. 4(a) shows the general correlation between weather condition factors and frequency of BGA bloom outbreak. Air temperature (MT, DHT, and DLT), relative humidity (MRH and LRH), and precipitation (PP) had significant positive correlations with the frequency of BGA bloom outbreak. In contrast, wind speed (MWS, MAXWS, and EWS) and the number of sunshine hours (SH) had significant negative correlations with the frequency of BGA bloom outbreak. It is easy to understand why the correlation with air temperature was positive and that for wind speed was negative. BGA may be more active at a higher air temperature within the appropriate range and wind speed can stir the lake water to disturb the BGA on the surface. However, it is difficult to explain why relative humidity and precipitation had a positive influence, whereas the number of sunshine hours had a negative influence. It is likely that other factors with a high correlation exist. For example, Dianchi Basin is in a northern sub-tropical moist monsoon climate, in which the hot season is rainy and precipitation is thus correlated with temperature.

Fig. 4(b) and (c) indicate that the correlations between weather condition factors and frequency of BGA bloom outbreak are mainly influenced by variations during the month, not during the year. The bold lines in the boxplot for each variable in (b) indicate the median and had the same correlations as shown in (a). The length of the whisker, which represents variation during the year, was short for most variables, suggesting that differences between years are not so clear. Unexpectedly, the medians of some variables in (c) showed entirely different correlations than in (a) and the whisker length was much longer than in (b) for all variables, which indicates that monthly variation was significant.

3.3. Analysis of water quality factors

There were 25 imputations for water quality factors and BGA bloom status with 144 samples of each (one sample per month from 1998 to 2009). However, the status of the BGA bloom in each imputation was not balanced because the sample size of the “positive” (i.e., BGA bloom outbreak with a label “1”) was not equal to that of the “negative” (i.e., no

![Fig. 4. The values and distributions of correlation coefficients for the weather condition factors.](image-url)
BGA bloom outbreak with a label “0”). The “positive” was nearly twice the “negative.” Because this imbalance would result in bias when building the BLR models, a balanced sample size was necessary and additional samples were created as follows: 1) randomly draw samples from the “positive” with the same number of samples from the “negative” without replacement n times (in this study, n=100); 2) for each imputation, 100 new samples were generated, which means that a total of 2500 samples were available for model building. Thus, 2500 BLR models were built to predict the probability of BGA bloom outbreak. The distribution of the regression coefficients for each variable by BLR is displayed in Fig. 5 and the univariate statistics for the regression coefficients are shown in Table 2.

Fig. 5 illustrates the variation in the regression coefficients. All coefficients varied from negative to positive except TEMP, which was nearly positive at all times. We inferred that TEMP had a strong positive influence on the probability of BGA bloom outbreak but that

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chl-a</td>
<td>-118.465</td>
<td>-5.604</td>
<td>-0.793</td>
<td>-0.925</td>
<td>4.934</td>
<td>25.202</td>
<td>82.205</td>
<td>-9.80</td>
</tr>
<tr>
<td>TP</td>
<td>-16.513</td>
<td>-2.499</td>
<td>-0.375</td>
<td>-0.010</td>
<td>2.554</td>
<td>16.801</td>
<td>14.290</td>
<td>-362.79</td>
</tr>
<tr>
<td>TN</td>
<td>-6.542</td>
<td>-0.055</td>
<td>0.203</td>
<td>0.215</td>
<td>0.482</td>
<td>2.272</td>
<td>0.211</td>
<td>2.14</td>
</tr>
<tr>
<td>NH₄-N</td>
<td>-2.770</td>
<td>-0.735</td>
<td>-0.303</td>
<td>0.314</td>
<td>0.415</td>
<td>35.248</td>
<td>4.808</td>
<td>6.99</td>
</tr>
<tr>
<td>NO₃-N</td>
<td>-129.365</td>
<td>-5.531</td>
<td>-3.195</td>
<td>-4.484</td>
<td>-1.209</td>
<td>4.650</td>
<td>49.445</td>
<td>-1.57</td>
</tr>
<tr>
<td>COD_Mn</td>
<td>-0.300</td>
<td>0.008</td>
<td>0.072</td>
<td>0.096</td>
<td>0.148</td>
<td>2.965</td>
<td>0.023</td>
<td>1.85</td>
</tr>
<tr>
<td>COD</td>
<td>-0.083</td>
<td>-0.022</td>
<td>-0.005</td>
<td>0.004</td>
<td>0.022</td>
<td>0.320</td>
<td>0.001</td>
<td>10.50</td>
</tr>
<tr>
<td>BOD</td>
<td>-0.076</td>
<td>-0.127</td>
<td>0.048</td>
<td>-0.015</td>
<td>0.153</td>
<td>0.704</td>
<td>0.055</td>
<td>-16.13</td>
</tr>
<tr>
<td>DO</td>
<td>-0.526</td>
<td>-0.129</td>
<td>-0.013</td>
<td>0.009</td>
<td>0.122</td>
<td>2.443</td>
<td>0.043</td>
<td>24.09</td>
</tr>
<tr>
<td>SD</td>
<td>-10.720</td>
<td>-3.614</td>
<td>-2.088</td>
<td>-2.253</td>
<td>-0.620</td>
<td>2.887</td>
<td>4.611</td>
<td>-0.95</td>
</tr>
<tr>
<td>pH</td>
<td>-12.954</td>
<td>-1.863</td>
<td>-1.209</td>
<td>-1.335</td>
<td>-0.631</td>
<td>1.474</td>
<td>1.540</td>
<td>-0.93</td>
</tr>
<tr>
<td>TEMP</td>
<td>-0.005</td>
<td>0.228</td>
<td>0.279</td>
<td>0.282</td>
<td>0.332</td>
<td>1.217</td>
<td>0.006</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Table 2

Univariate statistics of the regression coefficients for water quality factors.
other variables had weak influences. In addition, the peak of each histogram was near zero except for SD and TEMP. SD had a negative effect because transparency will decrease when a BGA bloom occurs.

In Table 3, the minimum (Min.), first quartile (1st Qu.), median (Median), mean (Mean), third quartile (3rd Qu.), maximum (Max.), variance (Variance), and coefficient of variation (Var. coef.) provide more details about the closeness of the mean to zero and the coefficients of variation. The results indicate that Chl-a, TP, COD, BOD, and DO had means very close to zero but standard variances were large enough to make the coefficients of variation nearly greater than 10; thus the influence of these factors on the probability of BGA bloom outbreak were very weak. TN, NO$_3$–N, COD$_{Mn}$, SD, and pH had similar features in that the first quartile, median, mean, and third quartile were simultaneously positive or negative. This suggests that all these factors will have a relative steady influence on the probability of BGA bloom outbreak. The median and mean of NH$_4$–N had an opposite sign, and the coefficient of variation was relatively large, indicating that the influence of NH$_4$–N can possibly be ignored. TEMP was obviously strong enough to influence the probability of BGA bloom outbreak, but it was still unusual to find the minimum as a negative value, even if it was close to zero. To conclude, TEMP had a strong positive effect; TN, NO$_3$–N, COD$_{Mn}$, SD, and pH had relatively steady effects; TN and COD$_{Mn}$ had positive effects; and the other factors had negative effects. Thus, the influences of Chl-a, TP, NH$_4$–N, COD, BOD, and DO can be ignored.

Using the 2500 BLR models, the probability of BGA bloom outbreak can be calculated for the 2500 data sets of 144 data each. As illustrated above, monthly variability was more significant than yearly variability, and so the mean probability and frequency for each month were calculated. The results are shown in Fig. 6.

As shown in Fig. 6(a), the probability and frequency were very similar between April and October and within the range of observed

![Fig. 6. Comparison between outbreak frequency and prediction probability.](image-url)
BGA bloom status data. From January to March and from November to December, the frequency was slightly higher than the probability. Overall, the trends were the same.

Fig. 6(b) and (c) provide more details for further analysis. The heights of each box reveal the robust nature of the median, which is a bold black line in the box. The median trends in both (b) and (c) were the same as that of the mean in (a). From June to September, the height of each box was significantly lower than in the other months, which indicates that these 4 months have a steady high probability of BGA bloom outbreak.

4. Discussion

Here, we highlight some of the previously described results to further discuss the key factors for a BGA bloom outbreak: (1) the frequency of BGA bloom outbreak varied from January to December with an initial increase and then a decrease between June and September, whereas the differences between the years 1998 to 2009 were minor. (2) Air temperature, relative humidity, and precipitation were significantly and positively correlated with the frequency of BGA bloom outbreak, whereas wind speed and the number of sunshine hours had a significant negative correlation. (3) TEMP had a strong positive effect on the probability of BGA bloom outbreak, while TN, NO₃⁻N, COD₈₅₀, SD, and pH had relative steady effects, in detail, TN and COD₈₅₀ were positive and the other factors were negative. The effects of Chl-a, TP, NH₄⁻N, COD, BOD, and DO could be ignored. (4) Both the predicted probability and the frequency of BGA bloom outbreak followed the same trends with a good fit from April to October. There was a steady high probability of BGA bloom outbreak between June and September.

So far, it seems that the results listed above are somewhat incredible because they, to a certain extent, contradict the popular hypotheses in the literatures (Elser et al., 2007; Jonge and Elliott, 2002; Lathrop et al., 1998; Ryther, 1971). Generally, nitrogen (N) and phosphorus (P) are considered as the most important nutrient elements for the aquatic plant growth, which are needed for protein synthesis, and DNA, RNA, energy transfer, respectively. Thus, a number of reports suggest to reduce the nitrogen and phosphorus inputs from terrestrial run-off, riverine input, organic pollution from sewage outfall and/or mariculture activities, atmospheric deposition (dry and wet) (Hodgkiss, 1998), to control eutrophication and BGA bloom occurrence in the lake or reservoir, simultaneously or single-handedly, depending on which one is the limiting factor or both (Conley et al., 2009). Furthermore, research suggests that, to reduce the inputs of both P and N, dissolved oxygen
concentrations (DO), transparency (SD), and other water quality conditions should be significantly improved (Wulff et al., 2007), which infers that water quality is also important for the outbreak of BGA bloom. Why did these factors not turn out to be substantial in our study? On the contrary, was water temperature, which was rarely concerned by the researchers, prominent?

To answer this question, we reconsidered the water quality factors, particularly the organics (such as COD\textsubscript{Mn}, COD, and BOD) and nutrients (such as TN, TP, NH\textsubscript{4}–N, and NO\textsubscript{3}–N), to determine more directly if these factors really have nothing to do with the probability of BGA bloom outbreak in Dianchi Lake with a further mathematical experiment (sensitivity analysis) of each variable in the BLR. The process was as follows.

First, we designed an interval and an initial value for each variable because classical one-at-a-time sensitivity analysis experiments change one variable at a time in a given interval with the other variables unchanged to determine the fate of the dependent variable. Table 3 lists the univariate statistics for the water quality factors, which were useful to create these intervals and initial values. The mean of each variable was a good choice for the initial value, whereas the intervals were either first and third quartiles or minimums and maximums. Because the minimum and maximum values were too extreme, we conservatively chose first quartiles and third quartiles as intervals.

Then, we conducted a sensitivity analysis for the 2500 BLR models with each interval divided into 100 points to calculate the probability of BGA bloom outbreak for each point. In the same way, we calculated the basic statistics such as minimum (Min.), first quartile (1st Qu.), median (Median), mean (Mean), third quartile (3rd Qu.), and maximum (Max.).

Finally, we added the statistics for each variable to Fig. 7. From Fig. 7, it is obvious that the probability of BGA bloom occurrence influenced by the mean of each variable (noted by the bold dotted dashed red line) can be divided into three categories: Chl-a, TP, NH\textsubscript{4}–N, COD, BOD, and DO have no influence; TEMP, TN and COD\textsubscript{Mn} have positive influences; and NO\textsubscript{3}–N, SD, and pH have negative influences. These results are exactly the same as the previous analysis. To find something different, the median (noted by the solid line), the first and third quartiles (noted by the dashed lines) of each variable were also provided in Fig. 7, indicating that: (1) reduction of the concentration of NH\textsubscript{4}–N, COD, and DO will

![Fig. 8. Matrix relation diagram for water temperature and air temperature.](image-url)
have a potential to lower the probability of BGA bloom outbreak, because a median is more robust than a mean, and the median was descending whereas the mean was horizontal. Besides, the median of TN and COD$_{mn}$ were more horizontal than the mean, which indicated that these two factors may have little correlation with the outbreak probability in Dianchi Lake. (2) A greater uncertainty will be created if we extrapolate too far from the initial points, because the interval between the first and third quartiles was narrow when near the initial points but wide when far away. (3) TEMP changes had a significant impact on the probability of BGA bloom outbreak, by contrast, the changes of other water quality factors did minor at the given intervals.

Since water temperature is important for the probability of BGA bloom outbreak and it has a close relationship with air temperature (Jun et al., 2010), we plotted a matrix relationship diagram for water temperature and air temperature in Fig. 8. In the lower panel, a scatter plot with a smooth curve is displayed to illustrate that water temperature and air temperature are highly correlated. In the diagonal panel, a histogram is plotted to show that all temperatures are two-tailed. In the upper panel, we added BGA bloom status information to the scatter plot, showing that higher temperatures mean a higher outbreak frequency.

From the above results, it is not difficult to find out that water quality factors, such as NO$_3$–N, SD, pH, NH$_4$–N, COD and DO, were also important factors for the outbreak of BGA bloom, even though they were not so significant as water temperature, which was direct but not substantial, highly correlated with air temperature. We conjectured that nutrients in Dianchi Lake were available in the right amount for cyanobacteria's growth and reproduction, and the weather conditions (such as wind speed and the number of sunshine hours, etc.) were comfortable. When water became warmer, cyanobacteria would reproduce rapidly, then BGA bloom would probably turn out (Paerl et al., 2008). So the probability of BGA bloom displayed increasing from February to August and decreasingly from August to the next February each year (Fig. 6), which had the same trend with Chl-a in Dianchi Lake (Sheng et al., 2012).

Therefore, in the present situation, with sufficient nutrients and favorable weather conditions, water temperature was the key factor which directly controlled the probability of BGA bloom outbreak in Dianchi Lake. Nevertheless, when conditions change, water quality factors will not be ignored, especially NO$_3$–N, SD, pH, NH$_4$–N, COD and DO, which still have a potential to control the outbreak of BGA bloom in Dianchi Lake.

5. Conclusion

In summary, through the EMB algorithm to create a full data set of water quality, weather condition and BGA bloom status without prominently introducing additional information, together with a group of BLR models to build up the relationship between water quality and BGA bloom status, our study suggests that the water quality conditions are appropriate to support an annual BGA bloom outbreak at a suitable water temperature, which is greatly influenced by air temperature. Furthermore, changing a single water quality factor had little effect on reducing the probability of BGA bloom outbreak in the present situation, but water quality factors, such as NO$_3$–N, SD, pH, NH$_4$–N, COD and DO should also be concerned in that they may have a potential to worsen the state. Still, we should be prudent to our research results, after all, they were based on incomplete information with many uncertainties, and our methods or models were not so sophisticated and precise. Besides, the BGA bloom status was not really an informative variable, in which outbreak intensity information was neglected. Maybe BGA bloom area is a good choice for further research, which is also imprecise when the lake is heterogenous and not well-mixed, such as Dianchi Lake. Whatever, a further research should be conducted to determine how to address the multiple factors involved and how to restore the environment to inhibit BGA bloom outbreaks.

Acknowledgments

This study was supported by the National Major Science and Technology Programs (No. 2008ZX07102-001). Weather condition data were provided by the China Meteorological Data Sharing Service System.

References

Pan, J., Li, W., Chen, K., 2004. A study on the environmental effect in the zone of restoration of aquatic plants covered at the northeast of Dianchi Lake: I. the effect on the controlling of alga blooming. Hupo Kexue 16 (2), 141–148.


Please cite this article as: Sheng, H., et al., Analysis of cyanobacteria bloom in the Waihai part of Dianchi Lake, China, Ecological Informatics (2012), doi:10.1016/j.ecoinf.2012.03.007