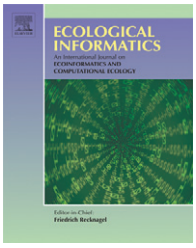
ECOLOGICAL INFORMATICS 2 (2007) 184–192

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/ecolinf



Prediction and elucidation of the population dynamics of *Microcystis* spp. in Lake Dianchi (China) by means of artificial neural networks

Hongbin Li[a], *Guoxiang Hou*[a,b,*], *Feng Dakui*[a], *Bangding Xiao*[b], *Lirong Song*[b], *Yongding Liu*[b]

[a]*Department of Ocean Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, PR China*
[b]*Institute of Hydrobiology, The Chinese Academy of Sciences, Wuhan 430072, PR China*

ARTICLE INFO

Keywords:
Algal dynamics
Algal bloom
Neural network
pH
Sensitivity analysis

ABSTRACT

Lake Dianchi is a shallow and turbid lake, located in Southwest China. Since 1985, Lake Dianchi has experienced severe cyanabacterial blooms (dominated by *Microcystis* spp.). In extreme cases, the algal cell densities have exceeded three billion cells per liter. To predict and elucidate the population dynamics of *Microcystis* spp. in Lake Dianchi, a neural network based model was developed. The correlation coefficient (R^2) between the predicted algal concentrations by the model and the observed values was 0.911. Sensitivity analysis was performed to clarify the algal dynamics to the changes of environmental factors. The results of a sensitivity analysis of the neural network model suggested that small increases in pH could cause significantly reduced algal abundance. Further investigations on raw data showed that the response of *Microcystis* spp. concentration to pH increase was dependent on algal biomass and pH level. When *Microcystis* spp. population and pH were moderate or low, the response of *Microcystis* spp. population would be more likely to be positive in Lake Dianchi; contrarily, *Microcystis* spp. population in Lake Dianchi would be more likely to show negative response to pH increase when *Microcystis* spp. population and pH were high. The paper concluded that the extremely high concentration of algal population and high pH could explain the distinctive response of *Microcystis* spp. population to +1 SD (standard deviation) pH increase in Lake Dianchi. And the paper also elucidated the algal dynamics to changes of other environmental factors. One SD increase of water temperature (WT) had strongest positive relationship with *Microcystis* spp. biomass. Chemical oxygen demand (COD) and total phosphorus (TP) had strong positive effect on *Microcystis* spp. abundance while total nitrogen (TN), biological oxygen demand in five days (BOD_5), and dissolved oxygen had only weak relationship with *Microcystis* spp. concentration. And transparency (Tr) had moderate positive relationship with *Microcystis* spp. concentration.

1. Introduction

Algal bloom, an explosive growth of phytoplankton, has become a chronic problem in many eutrophic freshwater lakes and reservoirs in China. With the procedure of urbanization and industrialization of China, explosion-like formations of algal blooms increasingly pollute fresh water ecosystems. They lead to enormous costs by affecting drinking water supply, aquaculture systems and tourism.3". Lake Dianchi is a representative, highly eutrophicated lake in Southwest China, which experienced severe cyanobacteria blooms (dominated by *Microcystis* spp.) and the algal biomass has exceeded three billion cells per liter

* *Corresponding author*. Institute of Hydrobiology, The Chinese Academy of Sciences, Wuhan 430072, PR China.
E-mail address: lrsong@ihb.ac.cn (G. Hou).

1574-9541/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.ecoinf.2007.03.007

in extreme cases. Further, Lake Dianchi is large (200 km^2) and 2.68 million inhabitants lived in the Dianchi basin. Therefore, improving the understanding of the dynamics of algal blooms and finally alleviating the damage of algal blooms has critical importance for the Dianchi basin. Due to the dominance of *Microcystis* spp., it was chosen as the output of our model in this study.

To alleviate the harmful impact, it is imperative to investigate the contribution of different environmental factors to the algal abundance, and to discover and systematize causal knowledge about the ecology of algae for better explanation, prediction and control of cyanobacteria blooms. Due to the diversity and connections of components governing the system's dynamics, aquatic ecosystems are very complex and possess nonlinear characteristics. Artificial neural networks are capable of modeling a complex nonlinear system. Many researchers have used feedforward neural networks to simulate the timing and magnitude of algal blooms and to forecast the cyanobacteria abundance (Yabunaka et al., 1997; Recknagel, 1997; Maier and Dandy, 2001; Wei et al., 2001; Hou et al., 2004), and recursive neural networks were applied also for the same purpose (Walter et al., 2001). Compared with other model approach, neural network models exhibited higher accuracy in the prediction of algal concentration, and artificial neural networks have become a popular and useful tool for modeling environmental systems (Maier and Dandy, 2001). In the present research, feedforward neural network was chosen for modeling algal concentration in Lake Dianchi.

Using the historical data of Lake Dianchi, the present research aimed at: (1) forecasting the abundance *Microcystis* spp. and determine the sensitivities of algal population to different environmental variables by means of feedforward neural network; (2) elucidating the relationships between algal abundance and environmental factors in Lake Dianchi.

2. Study site and data

Lake Dianchi (24°40′–25°03′ N, 102°37′–102°48′ E) is located in Kunming, Yunnan Province of China (see Fig. 1). It is a large plateau lake at an altitude of 1886.5 m with an area of 300 km^2, average depth of 4.7 m and a maximum depth of 11 m, and with 2.68 million residents in the Dianchi basin, and with the yearly burden of 216 million m^3 household waste water and 47.6 million m^3 industrial waste water.

To probe the cyanobacteria blooms and seek for an appropriate control solution, a pilot experiment area (6 km^2) was curtained off by waterproof enclosures from other area of Lake Dianchi in July 2000. And some blooms control plans were conducted in the pilot experiment area after that time. First, the planktivorous fishes (Silver carp, *Hypophthalmichthys molitrix* Cuvier et Valencienes and big-head carp, *Aristichthys nobilis* Richardson) were cultured with the population of 75 g of fish per cubic meter of water. Second, submerged aquatic vegetation was cultivated with the average density of 4.5 submerged higher plants per square meter. However, the control plan has not shown significant effects on water quality up till now. The reason may be that more time is needed for the effects to reveal, or that some more understanding about the algal abundance in Lake Dianchi is still needed.

From September 2000 to December 2002, sampling was undertaken once each month at ten sites in the experiment area measuring eight environmental factors such as total nitrogen (TN, mg/L), total phosphorus (TP, mg/L), chemical oxygen demand (COD, mg/L), biological oxygen demand in five days (BOD_5, mg/L), dissolved oxygen (DO, mg/L), pH, transparency (Tr, cm), and water temperature (WT, °C). The Chlorophyll a concentration and *Microcystis* spp. concentration were also measured. The measurements were conducted according to Jin and Tu (1990). And some statistics that describe the measured data were listed in Table 1.

3. Methods

3.1. The network structure

In order to model the relationship between the eight environment factors and the concentrations of *Microcystis* spp., a three-layer feedforward neural network was programmed (Fig. 2)

Fig. 1 – Lake Dianchi in Yunnan Province, Southwest China.

Table 1 – Some statistics on the measured data in Lake Dianchi (2000.9–2002.12)

Variable	Mean	Standard deviation	Min	Max
TN(mg/L)	3.07	1.48	0.47	10.95
TP(mg/L)	0.30	0.13	0.052	0.83
COD(mg/L)	15.40	5.93	5.00	53.50
BOD_5(mg/L)	9.69	6.07	0.00	27.50
DO(mg/L)	7.14	1.72	2.00	12.32
pH	8.88	0.56	7.40	10.20
Tr(cm)	29.88	12.87	0.00	80.00
WT(°C)	16.81	4.42	10.00	26.00
Chl *a*(mg/L)	0.17	0.11	0.02	0.60
Microcystis (Cells/L)	$2.53^{*}10^{8}$	$3.42^{*}10^{8}$	$1.65^{*}10^{6}$	$3.24^{*}10^{9}$

Note: the statistics were computed from 280 data samples (ten sampling sites for 28 months).

using the neural network toolbox in Matlab 7.0. In the neural network, environmental factors were treated as the network inputs and algal biomass as the network output respectively. And we chose the historical data of environmental factors in two lagged months as the network inputs after comparing the results with one or three month lagged inputs. With one month lagged inputs, the trained network showed much lower prediction ability than that with two month lagged inputs (the R^2 between the predictions and the targets less than 0.80); with three month lagged inputs, the prediction ability of the trained network showed no improvement to that of the trained network with two month lagged inputs. Stated clearly, the values of TN (mg/L), TP (mg/L), COD (mg/L), BOD_5 (mg/L), DO (mg/L), pH, Tr (cm) and WT (°C) in the $(n-2)$th month, the $(n-1)$th month were included in the network inputs (as shown in Fig. 2).

The input layer of the neural network comprised 16 neurons corresponding with the two month lagged history of the eight environmental factors, while the output is the *Microcystis* spp. concentration. Each neuron is connected to all neurons of adjacent layer. Neurons receive and send signals through these connections. Signals are transmitted only in one direction, from input layer to output layer through hidden layer. Connections are given a weight that modulated the intensity of the signal they transmit (Fig. 3).

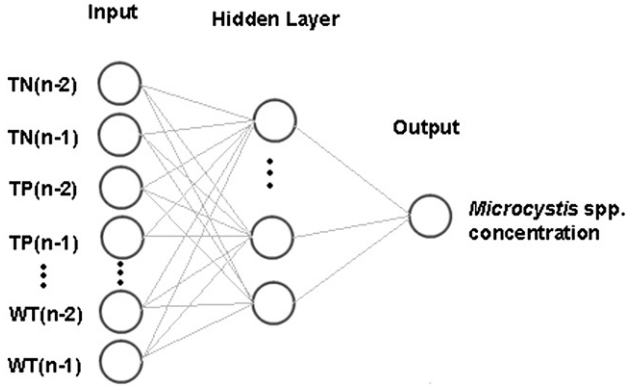


Fig. 2 – Neural network structure for predicting *Microcystis* spp. concentration of *n*th month with two month lagged inputs.

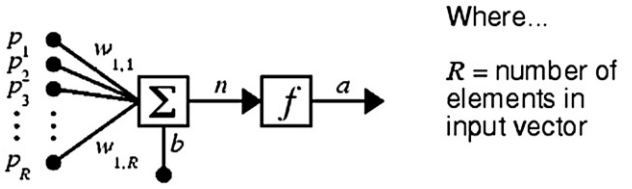


Fig. 3 – A neuron mapping a *R*-element vector *p* to a scalar *a* (*a*=*f* (*wp*+b).

An elementary neuron with *R* inputs is shown in Fig. 2. The input vector is weighted with an appropriate weight vector ***w***. The sum of the weighted inputs and the bias forms the input to the transfer function *f*. Neurons may use any differentiable transfer function *f* to generate their output. In our study, the output layer used linear transfer function, and tan-sigmoid transfer function was used for other network layers.

3.2. *The network training*

One difficulty in the application of an artificial neural network lies in determining the number of hidden layer nodes and analyzing their influence on the network output. Up till now, no assured methods were found for determining the number of hidden nodes. In our study, an empirical formula was employed to calculate the number of hidden nodes (Ma et al., 2002; Xiong et al., 2002).

The suggested (Ma et al., 2002) formula is

$$n_y = \sqrt{n_x * n_z} \quad (1)$$

where n_y is the number of hidden layer nodes, n_x the number of input layer nodes, and n_z the number of output layer nodes. According to experience, the appropriate number of hidden nodes should be a little bigger than the calculated n_y. And for good performance of the trained network we tried networks with different number of hidden nodes around the number computed from the empirical formula.

Training algorithm selection is the second difficulty in the ANN modeling. The back-propagation algorithm was created by generalizing the Widrow–Hoff learning rule (Xu and Wang, 2002). And there are many variations of the back-propagation (BP) algorithm such as back-propagation with momentum, Levenberg–Marquardt algorithm, BFGS Quasi-Newton algorithm, resilient backpropagation, scaled conjugate gradient algorithm, conjugate gradient with Powell/Beale restarts, Fletcher–Powell conjugate gradient, Polak–Ribiére conjugate gradient, one-step secant algorithm, variable learning rate backpropagation (MathWorks Inc., 2004). But it is very difficult to know which training algorithm will be the best one for a given problem (MathWorks Inc., 2004). So we tried different training algorithms in the network training.

Overfitting often occurs during neural network training (Tzafestas et al., 1996). The error on the training set is driven to a very small value, but when new data is presented to the network the error is large. The network has memorized the training examples, but it has not learned to generalize to new situations. Another word, the trained network has poor generalization ability. So improving generalization ability of networks is another difficulty we have to face when using

feedforward neural networks. Cross-validation (Amari et al., 1997; Rivals and Personnaz, 1999) and regularization (Girosi et al., 1995; Chen and Hagan, 1999) are the two approaches used widely for improving generalization of feedforward neural network.

In the cross-validation technique the available sample data is divided into three subsets, the training subset, the validation subset, and the testing subset. The training set is used for computing the gradient and updating the network weights. The validation set is used to monitor the generalization error. The error on the validation set is monitored during the training process. When the network begins to overfit the data, the error on the validation set will typically begin to rise. When the validation error increases for a specified number of iterations, the training is stopped, and the weights at the minimum of the validation error are returned. The test subset is used for verifying the network performance, i.e., the test set error is used to compare different networks.

The regularization approach involves modifying the performance function. While the typical performance function for training feedforward neural networks is the mean sum of squares of the network errors (mse), the regularization approach modifies the mse performance function by adding a term that consists of the mean of the sum of squares of the network weights and we call the modified performance function msereg. Using msereg performance function in the network training causes the network to have smaller weights, and this will force the network response to be smoother and less likely to overfit. Eqs. (2) and (3) are the mse and msereg performance functions respectively.

$$\mathrm{mse} = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)^2 \tag{2}$$

where t_i denotes the target for the ith input, and a_i represent the output of the network for the ith input.

$$\mathrm{msereg} = \lambda \cdot mse + (1-\lambda)\mathrm{msw} \tag{3}$$

where λ is the performance ratio, and msw represents the mean squared weights, i.e.,

$$\mathrm{msw} = \sum_{j=1}^{n} w_j^2.$$

And from our experience, the different combination of hidden nodes number, training algorithms, and generalization improving techniques has a complex effect on the generalization ability of neural networks. Because of this, we tried all the combinations of numbers of hidden nodes, from 4 to 11, the eleven training algorithm aforementioned, and the two improving techniques for finding a trained neural network with good generalization ability.

Because the components of input data have different orders of magnitude, they were standardized before network training, so that they had means of zero and standard deviations of one. The standardizing conversion used the relationship

$$X_s = (X_o - \overline{X})/\sigma_x \tag{4}$$

where X_s denotes the standardized variable, X_o is the original variable, and $\overline{X}$ and σ_x represent the mean and standard deviation of the original variable respectively.

3.3. The data set (the input–output examples)

For a specific sampling site, all the measured factors had historical values of 28 months (from September 2000 to December 2002). According the network structure, one input–output example must have one value of *Microcystis* spp. concentration for a specific month as the output, and have two month lagged values of eight environmental variables as the inputs. Because the first two month, September and October 2000, had no corresponding two month lagged values of environmental variables, only 26 input–output examples could be constructed from the historical data for one specific sampling site. Hence, the whole data set consisted of 260 input–output examples, which constructed from 28 month historical data of the ten sampling sites.

Because neural network models usually have poor performance for extrapolating, the 52 examples from site No. 4 and site No. 7, where the algal blooms showed moderate magnitude instead of very high or very low magnitude, were reserved as the testing dataset for model testing. The other 208 examples were used as the training dataset when using regularization, while 52 randomly selected examples were used as the validation subset and the remaining 156 examples as the training subset when using cross-validation (or early stopping).

3.4. Model validation and neural network based sensitivity analysis approach

Model validation was based on R^2 values between the observed and predicted concentrations of *Microcystis* spp. in the testing dataset. The network with highest R^2 value in all trained networks was selected as the best-predicting neural network. For visual comparison, the observed algal concentrations and predicted concentrations by the best-predicting network were plotted in the same figure. Using the trained network with best performance, one type of sensitivity analysis, 'Most Influencing Parameter' sensitivity analysis was implemented.

Many researches have applied sensitivity analysis approaches to determine the impact of input variables on output (Goh, 1995; Lek et al., 1996; Siginer, 1997; Dimopoulos et al., 1999; Jeong et al., 2001). In our study, the sensitivity analysis approach used is similar to the one described in Zar (1984) and Jeong et al. (2001). To compute the sensitivity of algal biomass to one variable (the explanatory variable), two simulations were made. In the first simulation, the trained network was fed with original input vectors and the output values represented the predictions with non-disturbed inputs. In the second simulation, the two components of the explanatory variable in the input vectors were disturbed by +1 SD (standard deviation) and the network outputs represented predictions with disturbed inputs by +1 SD of the explanatory variable. Subtracting the non-disturbed predictions from the disturbed outputs, the sensitivity of algal biomass to +1 SD increase of the explanatory variable on every data point in the whole data set was obtained. And the

Table 2 – The mean R^2 values for different combinations of training algorithms and improving generalization approaches

Network training algorithm	Generalization improving approach	
	Cross-validation	Regularization
BPM	0.53	0.55
LM	0.68	0.60
BFG	0.64	0.83
RBP	0.57	0.71
SCG	0.62	0.83
CGB	0.67	0.78
CGF	0.63	0.80
CGP	0.59	0.75
OSS	0.63	0.76
VLR	0.69	0.62

Abbreviation: BPM, back-propagation with momentum; LM, Levenberg–Marquardt algorithm; BFG, BFGS Quasi-Newton algorithm; RBP, resilient backpropagation; SCG, scaled conjugate gradient algorithm; CGB, conjugate gradient with Powell/Beale restarts; CGF, Fletcher–Powell conjugate gradient; CGP, Polak–Ribiére conjugate gradient; OSS, one-step secant algorithm; VLR, variable learning rate backpropagation.

sensitivity of mean algal biomass to the variable could be calculated by averaging the sensitivity on all data points in the whole data set. Here, the SDs of all variables (see Table 1) was computed on the whole dataset.

Furthermore, to explain the critical influence of pH changes on algal bloom formation, the scatter plot approach was applied to investigate with detail the sensitivity of *Microcystis* spp. biomass to pH changes at all the data points in the whole data set. The ecological dynamics often shows nonlinearity. The response of algal biomass to pH increase at one specific data points not only depended on the magnitude of pH increase, but also depended on the values of different variables at this specific data points. If a significant pattern existed between algal biomass responses to pH increase and the values of one specific variable, the scatter plot of algal responses to pH increase versus the values of that variable could visualize the pattern. Therefore, scatter plot is a very useful tool for detecting significant patterns between sensitivities and a certain variable.

4. Results and discussion

4.1. *The trained neural network and its validation*

The performance of every trained network was evaluated through correlation coefficients (R^2) between its predictions and the observed algal concentrations in the testing dataset. Then, the mean R^2 values for different combinations of training algorithms and two approaches for generalization improving were computed (see Table 2). BFGS Quasi-Newton algorithm or scaled conjugate gradient algorithm combined with regularization approach showed superiority for the given dataset in this study (Table 2). The neural network with best performance was found when trying the combination of 8 hidden nodes, BFGS Quasi-Newton algorithm and regularization approach with the performance ratio of 0.5. The trained network was used to predict *Microcystis* spp. concentrations in site No. 4 and site No. 7, and the predicted results were depicted in Fig. 4.

Fig. 4 showed that the predictions were almost consistent with the measured biomass of *Microcystis* spp. The timing of all significant peak of *Microcystis* spp. in both sites was well recognized, even though there were some significant errors (i.e., June 2001 and Sep. 2002 in Site No. 4, and June 2001 in Site No. 7).

4.2. *Results of sensitivity analysis*

Fig. 5 plotted the mean responses of *Microcystis* spp. population to +1 SD increase of different variables. And Fig. 6 was a whisker plot which represented the responses of *Microcystis* spp. biomass to +1 SD increase of different variables with some more detail. From the Fig. 5 and Fig. 6, +1 SD increase of water temperature (WT) had strongest positive relationship with *Microcystis* spp. biomass. Chemical oxygen demand (COD) and total phosphorus (TP) had strong positive effect on algal abundance while total nitrogen (TN), BOD_5, and dissolved oxygen (DO) had only weak relationship with *Microcystis* spp. concentration. And transparency (Tr) had moderate positive relationship with *Microcystis* spp. concentration. At last, only pH increase, it had to be mentioned that, had the strong negative influence on the *Microcystis* spp. concentration.

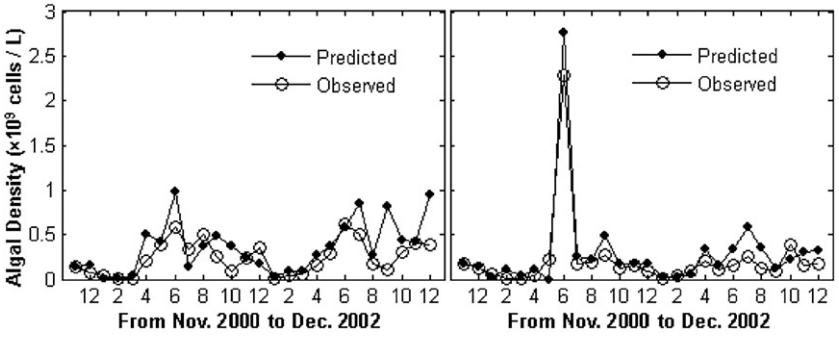


Fig. 4 – One-month-ahead predictions of *Microcystis* spp. concentration in Lake Dianchi (Left panel: site No. 4; Right panel: site No. 7).

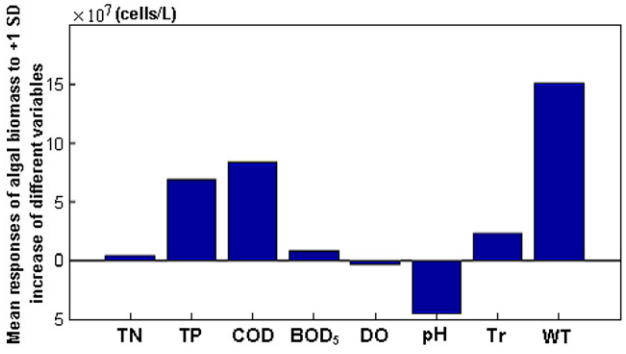


Fig. 5 – Mean responses of *Microcystis* spp. biomass to +1 SD (standard deviation) increase of different variables (Abbreviation: TN, total nitrogen; TP, total phosphorus; COD, chemical oxygen demand; BOD5, biological oxygen demand in five days; DO, dissolved oxygen; Tr, transparency; WT, water temperature).

From Fig. 6, we could find some outliers. Why these outliers? For some input patterns with high values of a certain variable, +1 SD increase of that variable might push such input patterns out of the that variable's range in the training dataset. Then the trained was very likely to produce poor outputs which led to those outliers in Fig. 6 because of poor extrapolation ability of the network.

4.3. *Relationships between algal abundance and environmental factors*

The results of sensitivity analysis in Fig. 5 revealed that *Microcystis* spp. had strong negative response to +1 pH increase in Lake Dianchi. This finding conflicts with the related hypothesis in some researches on hypereutrophic Lake Kasumigaura (Japan). By analyzing the data in hypereutrophic Lake Kasumigaura from 1981 to 1998, Yabunaka et al. (1997) concluded that pH had strong positive relation with chl *a* concentration.

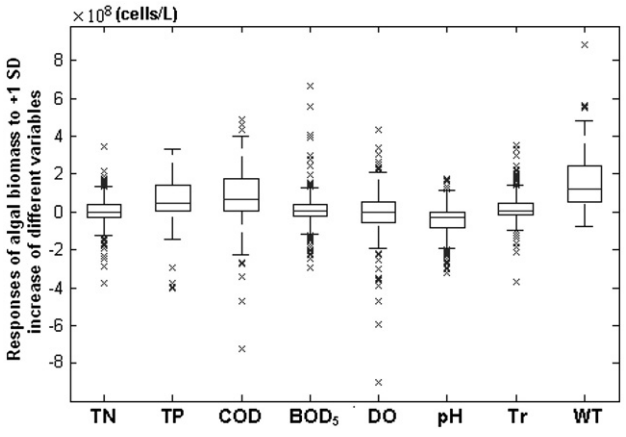


Fig. 6 – Responses of *Microcystis* spp. biomass to +1 SD (standard deviation) increase of different variables (Abbreviation: TN, total nitrogen; TP, total phosphorus; COD, chemical oxygen demand; BOD5, biological oxygen demand in five days; DO, dissolved oxygen; Tr, transparency; WT, water temperature).

Table 3 – Chlorophyll a, *Microcystis* spp. and pH in Lake Dianchi (2000.9–2002.12) and Lake Kasumigaura (1984–1993)

	Lake Dianchi			Lake Kasumigaura		
	Mean	Min	Max	Mean	Min	Max
Chlorophyll a (mg/L)	0.165	0.02	0.602	0.073	0.00069	0.28
Microcystis spp. ($^*10^6$ cells/L)	253	1.65	3236	38.64	0.001	644.12
pH	8.88	7.4	10.2	8.75	7.12	10.13

The statistics of Lake Kasumigaura were collected from Recknagel (2005).

The increase in photosynthesis led by the increase in phytoplankton cells, they stated that, decreases the number of carbonic acid ions and pH increases. Wei et al. (2001) also found that pH had strong positive influence on *Microcystis* growth after analyzing the data in Lake Kasumigaura from 1982 to 1996.

Why algal concentration in Lake Dianchi showed distinctive response to minor change of pH from that in Lake Kasumigaura? To answer the question, let's probe the algal biomass and the pH in both lakes first.

Recknagel et al. (2006) reported some limnological properties of Lake Kasumigaura (1984–1993). The statistics of chlorophyll a, *Microcystis* spp. and pH in Lake Dianchi and in Lake Kasumigaura were collected and presented in Table 3. From that table, it could be learned that chl *a* concentration in Lake Dianchi was twice more than that in Lake Kasumigaura, and *Microcystis* spp. in Lake Dianchi was six times more than that in Lake Kasumigaura and pH in Lake Dianchi was a little higher also.

As the pH of freshwater is determined by its CO_2 budget (Stumm and Morgan, 1970) alkaline conditions are likely for a hypereutrophic lake such as Lake Dianchi because of limited

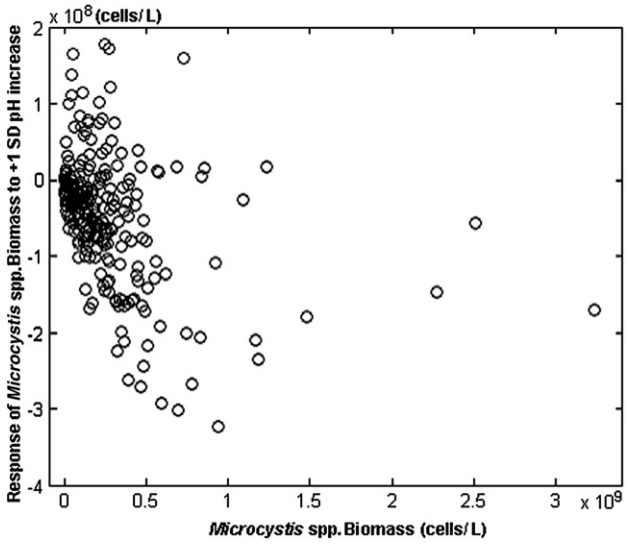


Fig. 7 – Response of *Microcystis* spp. biomass to +1 SD pH increase vs. the corresponding one-month-lagged biomass of *Microcystis* spp. for the 260 examples in the whole data set.

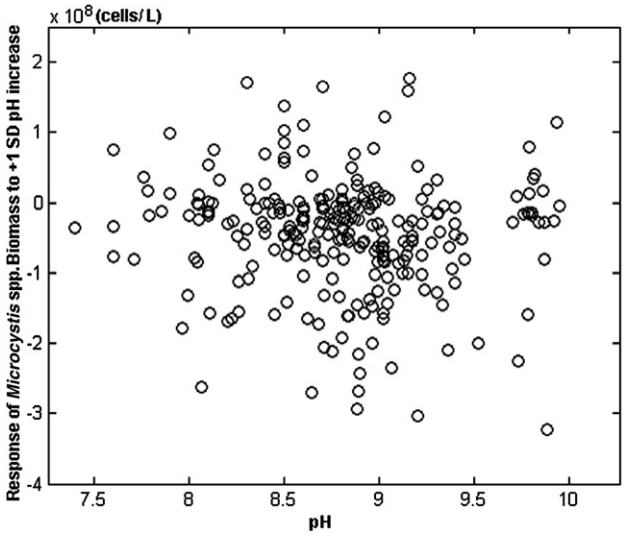


Fig. 8 – Response of *Microcystis* spp. biomass to +1 SD pH increase vs. the corresponding one-month-lagged pH for the 260 examples in the whole data set.

availability of dissolved CO_2 led by the photosynthesis of highly abundant algal populations. For this reason, high pH means limited supply of CO_2 for the photosynthesis of algal population. Nevertheless extremely high abundance of algal populations in Lake Dianchi has much more demand of CO_2 for its photosynthesis. Therefore, an increase of pH in Lake Dianchi with extremely high concentrated algal population and high pH means a cut in the very limited CO_2 supply, and would be likely to inhibit the algal growth. To sum up, extremely high abundance of algal population and high pH, especially the high algal population relative to that in Lake Kasugaura, could be a potential explanation for the distinctive phenomenon in Lake Dianchi, a strong negative response of chl *a* to a minor pH increase. However, this is only a hypothesis now. To prove the truth of this hypothesis, some algal culture experiments in controlled conditions are still needed.

Furthermore, we thought that the higher the abundance of *Microcystis* spp. and the higher the pH, an increase of pH would be more likely to have a negative (or strong negative) influence on *Microcystis* spp. growth. To search for the evidences for the hypothesis, two scatter graphs were depicted (Figs. 7 and 8).

In Fig. 7, there were 260 points which corresponding with 260 data points in the whole dataset. One point, represented by a circle in the plot, showed the response of *Microcystis* spp. to +1 SD pH increase versus the one month lagged *Microcystis* spp. biomass for a certain example in the whole data set. And Fig. 8 scattered the *Microcystis* spp. response to pH increase versus the one month lagged pH for the 260 examples in the whole dataset. Both the two scatter graphs revealed a top-left to right-down trend, and provide clear evidences for the hypothesis aforementioned.

Because this hypothesis about the link between *Microcystis* spp. concentration and pH is so different from some literature findings, we dived into the raw data for more evidences. To investigate the responses of *Microcystis* spp. to pH increase from the raw data, we first constructed the pH-*Microcystis*-relationship dataset, each data point in which consisted of the *Microcystis* spp. biomass of the *n*th month (MB(*n*)), the pH of *n*th month (pH(*n*))and the *Microcystis* spp. biomass of the (*n*+1)th month (MB(*n*+1))at one identical sampling site. Because of one-month-lagged structure, there were 270 data points in this

Table 4 – Relationship between *Microcystis* spp. biomass of (*n*+1)th month and one-month-lagged pH, one-month-lagged *Microcystis* spp. biomass in Lake Dianchi (2000.9–2002.12)

MB(n)	pH(*n*)	NP	NI	ND	pH(*n*)			MB(*n*) (*10^8 cells/L)			MB(*n*+1) (*10^8 cells/L)		
					Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Low	Low	41	33	8	8.34	7.60	8.70	0.37	0.02	1.06	1.04	0.02	4.77
	Moderate	30	22	8	8.85	8.71	9.01	0.38	0.04	1.05	1.06	0.05	4.50
	High	18	6	12	9.58	9.13	10.0	0.79	0.52	1.08	0.86	0.06	3.26
Moderate	Low	27	21	6	8.38	7.40	8.70	1.66	1.14	2.37	2.89	0.83	6.94
	Moderate	23	13	10	8.91	8.76	9.04	1.75	1.15	2.39	1.75	0.23	3.54
	High	39	14	25	9.57	9.05	10.2	1.58	1.11	2.37	1.94	0.02	11.9
High	Low	27	9	18	8.27	7.60	8.68	4.15	2.50	14.8	4.18	0.64	14.8
	Moderate	35	10	25	8.91	8.71	9.04	4.52	2.43	10.9	2.92	0.11	9.39
	High	24	11	13	9.41	9.06	9.96	5.31	2.48	12.3	4.42	0.05	11.7

Abbreviation: MB, *Microcystis* spp. biomass; MB(*n*), *Microcystis* spp. biomass of nth month; NP, number of points in a specific subset (For example, the subset with low MB and low pH has 41 data points); NI, number of MB increases in a specific subset from the *n*th month to the next month at an identical sampling site; ND, ND, number of MB decreases in a specific subset from the *n*th month to the next month at an identical sampling site.

Note: 1) High MB means MB>2.394*108 cells/L, moderate MB means MB>1.082*10^8 cells/L and MB≤2.394*10^8 cells/L, low MB means MB<=1.082*10^8 cells/L. 2) High pH means pH>9.04, moderate pH means pH>8.7 and pH≤9.04, low pH means pH<8.7. 3) To investigate the responses of *Microcystis* spp. to pH increase from the raw data, we first constructed the pH-*Microcystis*-relationship dataset, each data point in which consisted of MB(*n*), pH(*n*) and MB(*n*+1) at one identical sampling site. Because of one-month-lagged structure, there were 270 data points in this dataset first, and 264 data points after deleting six points where MB(*n*) or MB(*n*+1) was extremely high, i.e., high than 2*10^9 cells/L. This dataset were grouped to 9 subset, such as the low MB(*n*) and low pH(*n*) subset with 41 data points, the low MB(*n*) and moderate pH(*n*) subset with 30 data points,…, and the high MB(*n*) and high pH(*n*) subset with 24 data points. 4) All the five subsets with high MB(*n*) or high pH(*n*) showed more decreases of *Microcystis* spp. biomass from the nth month to the next month than increases of *Microcystis* spp. biomass. This supported the hypothesis that high pH and high algal population led to negative response of *Microcystis* spp. concentration to pH increase.

dataset first, and 264 data points after deleting six points where *Microcystis* spp. biomass of the *n*th month or the (*n*+1)th month was extremely high, i.e., higher than $2^{*}10^{9}$ cells/L. Then this dataset were grouped to 9 subset, such as the low MB(*n*) and low pH(*n*) subset with 41 data points, the low MB(*n*) and moderate pH(*n*) subset with 30 data points,…, and the high MB (*n*) and high pH(*n*) subset with 24 data points (Table 4). Last, we computed the statistical properties for these nine subsets, and computed the number of increases of *Microcystis* spp. biomass from *n*th month to the next month, and the number of decreases of *Microcystis* spp. biomass from *n*th month to the next month for the nine subsets respectively. Table 4 represented the results. In Table 4, all the five subsets with high MB (*n*) or high pH(*n*) showed more decreases of *Microcystis* spp. biomass from the *n*th month to the next month than increases of *Microcystis* spp. biomass. Therefore, the raw data could gave direct support for the hypothesis that high pH and high algal population led to negative response of *Microcystis* spp. concentration to pH increase. Furthermore, it could be noted in Table 4 that algal population were more likely to show positive response to pH increase when algal population and pH were low or moderate. Therefore the response of *Microcystis* spp. biomass was dependent on algal biomass and pH. When *Microcystis* spp. population and pH were moderate or low, the response of *Microcystis* spp. population would be more likely to be positive in Lake Dianchi. This is consistent with previous findings (Yabunaka et al., 1997; Recknagel, 1997). However, *Microcystis* spp. population in Lake Dianchi would be more likely to show negative response to pH increase when *Microcystis* spp. population and pH were high, contrary to previous findings. In addition, Figs. 7 and 8 also gave some evidences for the two different responses of algal population to pH changes.

Reynolds (1984) and Shapiro (1990) presented that dominance of blue-green algae had positive response to high water temperature, nutrient availability. And Hou et al. (2004) also concluded that water temperature, TP, TN, COD had positive effect on algal abundance. The results of sensitivity analysis in Fig. 5 showed strong positive response of *Microcystis* spp. to WT, TP and COD increase, which confirmed the literature findings aforementioned.

TN showed weak relationship with *Microcystis* spp. in Fig. 5. An abundance of nutrient supply for the algal proliferation in the highly eutrophicated Lake Dianchi, we think, was the reason for that.

5. Conclusion

This study showed that neural networks were capable of not only predicting the algal abundance successfully but also elucidating the driving factors for the algal proliferation by means of sensitivity analysis. While much of the findings of sensitivity analysis corresponded well with existing theory on the dynamics of algal population, pH was found to have strong negative relationship with *Microcystis* spp. concentration in Lake Dianchi. Compared with Lake Kasumigaura, we think the severely eutrophication and algal abundance could explain the distinctive relationship between *Microcystis* spp. concentration and pH. Furthermore, we found that the response of *Microcystis* spp. population was dependent on *Microcystis* spp. concentration and pH level after analyzing an in-depth investigation on the raw dataset. When *Microcystis* spp. population and pH were moderate or low, the response of *Microcystis* spp. population would be more likely to be positive in Lake Dianchi; *Microcystis* spp. population in Lake Dianchi would be more likely to show negative response to pH increase when *Microcystis* spp. population and pH were high. However, this hypothesis still needs more confirmations. Further comparative studies in lakes with diverse eutrophication, and algal culture experiments in controlled conditions might support or discard the hypothesis.

Acknowledgements

The authors would like to thank the two anonymous reviewers for their helpful, in-depth comments and suggestions. This research was supported by the National Basic Research Program of China 2002CB412300 and the National Natural Science Foundation of China (No. 50209003), the Chinese Academy of Sciences Project (KSCX2-1-10).

REFERENCES

Amari, S.-i., Murata, N.K.-R., Finke, M., Yang, H.H., 1997. Asymptotic statistical theory of overtraining and cross-validation. IEEE Transactions on Neural Networks 8 (5), 985–996.

Chen, D.D., Hagan, M.T., 1999. Optimal use of regularization and cross validation in neural network modeling. Proceedings of the 1999 International Joint Conference on Neural Networks, vol. 2, pp. 1275–1280.

Dimopoulos, Y., Chronopoulos, J., Chronopoulou, S.A., Lek, S., 1999. Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city. Ecological Modelling 120, 157–165.

Girosi, F., Jones, M., Poggio, T., 1995. Regularization theory and neural networks architectures. Neural Computation 7, 219–269.

Goh, A.T.C., 1995. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering 9, 143–151.

Hou, G.X., Song, L.R., Liu, J.T., Xiao, B.D., Liu, Y.D., 2004. Modeling of cyanobacterial blooms in hypereutrophic Lake Dianchi, China. Journal of Freshwater Ecology 19 (4), 623–629.

Jeong, K.S., Joo, G.J., Kim, H.W., Ha, K., Recknagel, F., 2001. Prediction and elucidation of algal dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. Ecological Modelling 146, 115–129.

Jin, X., Tu, Q., 1990. Methods for Research of Eutrophicated Lakes, 2nd edition. Meteorological Press, Beijing.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagner, S., 1996. Application of neural networks to modeling non-linear relationships in ecology. Ecological Modelling 90, 39–52.

Ma, X.X., He, X.J., Zhao, D.Q., Wang, X.Y., 2002. Influence of B-P networks hidden layer on water quality evaluation result. International Journal Hydroelectric Energy 20, 16–18.

Maier, H.R., Dandy, G.C., 2001. Neural Network Based Modelling of Environmental Variables: A Systematic Approach. Mathematical and Computer Modelling 33, 669–682.

MathWorks Inc., 2004. Neural Network Toolbox: User's Guide (Matlab 7.0). The Natick, MA.

Recknagel, F., 1997. ANNA — Artificial neural network model for predicting species abundance and succession of blue-green algae. Hydrobiologia 349, 47–57.

Recknagel, F., Kim, B., Takamura, N., Welk, A., 2006. Unravelling and forecasting algal population dynamics in two lakes

different in morphometry and eutrophication by neural and evolutionary computation. Ecological Informatics 2, 133–151.
Reynolds, C.S., 1984. The Ecology of Freshwater Phytoplankton. Cambridge University Press, Cambridge, p. 384.
Rivals, I., Personnaz, L., 1999. On cross validation for model selection. Neural Computation 11, 863–870.
Shapiro, J., 1990. Current beliefs regarding dominance of by blue-greens: the case for the importance of CO_2 and pH. Verhandlungen der Internationalen Vereinigung fur Limnologie 24, 38–54.
Siginer, I., 1997. Some artificial neural network applications to greenhouse environmental control. Computers and Electronics in Agriculture 18, 167–186.
Stumm, W., Morgan, J.J., 1970. Aquatic Chemistry. Wiley, New York.
Tzafestas, S.G., Dalianis, P.J., Anthopoulos, G., 1996. On the overtraining phenomenon of backpropagation neural networks. Mathematics and Computers in Simulation 40, 507–521.
Walter, M., Recknagel, F., Carpenter, C., Bormans, M., 2001. Predicting eutrophication effects in the Burrinjuck Reservoir (Australia) by means of the deterministic model SALMO and the recurrent neural network model ANNA. Ecological Modelling 146 (1–3), 97–114.
Wei, B., Sugiura, N., Maekawa, T., 2001. Use of artificial neural network in the prediction of algal blooms. Water Research 35 (8), 2022–2028.
Xiong, L.H., Guo, S.L., Wang, Y., 2002. Study and application of artificial neural network in real time flood forecasting. International Journal Hydroelectric Energy 20, 28–31.
Xu, S.G., Wang, J., 2002. Fuzzy optimal decision for structure of feedforward neural networks and its application in runoff forecast. International Journal of Hydroelectric Energy 20, 35–37.
Yabunaka, K., Hosomi, M., Murakami, A., 1997. Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. Water Science and Technology 36 (5), 89–97.
Zar, J.H., 1984. Biostatistical Analysis, 2nd edition. Prentice-Hall, NJ, p. 718.