Limnology (2015) 16:179–191
DOI 10.1007/s10201-015-0454-7

RESEARCH PAPER

Combination of artificial neural network and clustering techniques for predicting phytoplankton biomass of Lake Poyang, China

Jiacong Huang[1] · Junfeng Gao[1] · Yinjun Zhang[2]

Received: 8 January 2015 / Accepted: 15 April 2015 / Published online: 9 May 2015

Abstract A single artificial neural network (ANN) model is inadequate for predicting phytoplankton biomass in a large lake due to its high spatial heterogeneity. In this study, ANN was combined with a clustering technique to simulate phytoplankton biomass in a large lake (Lake Poyang) using a 7-year dataset. Two ANN models (named ANN_Downstream and ANN_Upstream) were developed for the downstream and upstream areas based on the k-means clustering results of 17 sampling sites at Lake Poyang, China. They performed better than ANN_Poyang (an ANN model for the whole lake), indicating the success of the clustering technique in improving ANN models for predicting phytoplankton biomass in different sub-regions of the large lake. A sensitivity analysis based on ANN_Downstream and ANN_Upstream showed that phytoplankton dynamics responded differently to environmental variables in different sub-regions of Lake Poyang. This case study demonstrated the good performance of ANN models in describing phytoplankton dynamics, and the potential of coupling ANN with a clustering technique to describe the spatial heterogeneity of natural ecosystems.

Keywords Chlorophyll *a* · Artificial neural network · Clustering · Lake Poyang · Sensitivity analysis

Handling Editor: Toshiya Katano.

✉ Junfeng Gao
gaojunf@niglas.ac.cn

[1] Key Laboratory of Watershed Geographic Sciences, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, 73 East Beijing Road, Nanjing 210008, China

[2] China National Environmental Monitoring Centre, 8(B) Dayangfang Beiyuan Road, Chaoyang District, Beijing 100012, China

Introduction

Lake eutrophication is one of the most serious water quality problems around the world. Phytoplankton biomass (represented by chlorophyll *a*) is an important indicator for evaluating the status of lake eutrophication, and has thus been intensively investigated by lake researchers and managers for decades (Harmel et al. 2015; Huang et al. 2015b; Mooij et al. 2010). Modeling chlorophyll *a* (Chl *a*) helps us to understand the influence of various environmental variables on phytoplankton dynamics, and thus support lake managers in taking measures to combat lake eutrophication (Huang et al. 2012b, 2015a). It is acceptable to assume that Chl *a* is homogeneous in small lakes (Mooij et al. 2007). However, in large lakes (e.g., Lake Erie and Lake Taihu), Chl *a* dynamics are highly heterogeneous (Huang et al. 2012a; Zhang et al. 2008). The ability to describe the spatial heterogeneity of Chl *a* in these large lakes is thus needed.

Simple regression models have been widely used to predict phytoplankton biomass. However, their model fits are generally unsatisfactory due to the non-linear relationship between Chl *a* and other environmental variables (Huszar et al. 2006; Phillips et al. 2008). Mechanistic models describe biological and physical processes of phytoplankton explicitly (Jørgensen 2010; Robson 2014). However, their applications are constrained by the requirement for adequate understanding of phytoplankton dynamics (Huang et al. 2014). As an alternative to regression and mechanistic models, artificial neural network (ANN) models were demonstrated to have high potential in modeling highly non-linear and complex natural systems (Panda et al. 2010; Rao and Alvarruiz 2007; Singh et al. 2009). They have thus been increasingly used to forecast water quality in complex aquatic ecosystems (Jeong et al.

2008; Wu et al. 2014a). Their methodological issues (e.g., determination of network architecture and choice of performance criteria) have been thoroughly discussed, and well supported our ANN modeling practice (Maier et al. 2010; Wu et al. 2014b). However, the ANN technique has so far scarcely been used in describing spatial differences of phytoplankton dynamics in large lakes (Jørgensen 2008). This is mainly due to the limited ability of ANN in describing spatial processes related to phytoplankton dynamics.

The limitation of ANN in spatial modeling may be overcome by coupling with clustering techniques (e.g., k-means method, self-organizing map and hierarchical clustering). These techniques have been widely used in partitioning entities with similar characteristics into a group (Chang et al. 2011; Dietrich et al. 2013; Monem and Hashemy 2011). Among these techniques, the k-means method has been widely used for clustering purposes in environmental sciences (Areerachakul and Sanguansintukul 2010; Ay and Kisi 2014; Elangasinghe et al. 2014). The k-means method implements clustering by computing the distance between entities and cluster centers (Celebi et al. 2013). Ease of implementation, high efficiency and rich case studies were the main reasons for its popularity (Jain, 2010). Inspired by the successful clustering case studies using the k-means method, we argued that it has the potential to partition a large lake with high spatial heterogeneity into several sub-regions, and an ANN model could thus be developed for each sub-region. These ANN models for the sub-regions of the lake could then be compared to investigate the spatial difference of phytoplankton dynamics in a lake.

The phytoplankton dynamics of Lake Poyang, the largest freshwater lake in China, have become increasingly important in recent years due to the small-scale algal bloom events which occurred in 2007 and 2013 (Wu et al. 2014d; Yi and Chen 2015). To combat the severe droughts and floods in Lake Poyang (Ye et al. 2013), lake managers have planned to build a dam (named Poyang Dam) in the downstream area of Poyang. Economic, ecological and environmental issues regarding this project have been debated by researchers and stakeholders (Yi and He 2012). Investigating the change of phytoplankton dynamics due to the dam construction project would be helpful in assessing the potential impacts on lake eutrophication.

The main objectives of this paper are to predict phytoplankton dynamics in different areas of Lake Poyang, and to quantify the impacts of environmental variables on phytoplankton dynamics. ANN and k-means clustering techniques were used for these purposes. Based on a 7-year dataset, three ANN models were developed to simulate Chl *a* dynamics in different areas of Lake Poyang. Spatial and temporal characteristics of Chl *a* dynamics in Lake Poyang were analyzed based on the simulation results. A sensitivity analysis was carried out to quantify the influence of water quality, hydrodynamic and meteorological conditions on the Chl *a* dynamics in both the downstream and upstream areas of Lake Poyang. The impacts of Poyang Dam on phytoplankton dynamics were analyzed.

Materials and methods

Study area

Lake Poyang, located in the northern Jiangxi Province, China, has a surface area of 3283 km^2 and a mean depth of 5.1 m. It has five inflows (Ganjiang, Fuhe, Xinjiang, Raohe and Xiushui rivers) and one outflow connecting with the Yangtze River (Fig. 1). However, the water in the Yangtze River may flow into Lake Poyang to reduce its flood discharge during heavy rainfall periods. The mean hydraulic retention time is as short as 10 days due to the large discharges of the connected rivers (Wang and Dou 1998). The water level changes significantly within a year, resulting in large variation of lake area (Feng et al. 2011).

Chlorophyll *a* in Lake Poyang was low (6.9 µg/l in 2011). However, high spatial differences existed, with higher chlorophyll *a* (9.7 µg/l in 2011) in the south than in the north (3.7 µg/l in 2011). Chlorophyll *a* peak (34.4 µg/l) was observed in eastern Lake Poyang by Wu et al. (2014c). Algal bloom events were occasionally observed in Lake Poyang, e.g., in October 2007 (Yi and Chen 2015) and May 2013 (Wu et al. 2014d). Phytoplankton growth in Lake Poyang was found to be limited by light, because nutrient (nitrogen and phosphorus) levels were high for

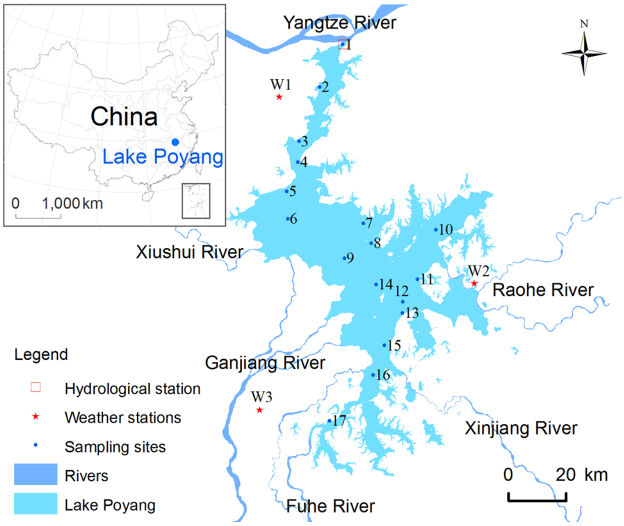

Fig. 1 Locations of on Lake Poyang, sampling sites and hydrological and weather stations

phytoplankton growth due to the high loading of inflow rivers (Lu et al. 2012).

Data

Meteorological, hydrological and water quality data during 2005–2011 in Lake Poyang are presented with their basic statistical descriptions in Table 1. Eight variables were selected for model development based on the following facts:

1. Phytoplankton biomass in a lake is mostly represented by Chl *a* (Wu et al. 2011). The biological processes of phytoplankton are widely recognized to be affected by water temperature, nitrogen, phosphorus and light conditions (Fragoso Jr et al. 2008; Hamilton and Schladow 1997; Jørgensen and Bendoricchio 2001).
2. Although precipitation was scarcely included in mechanistic models for phytoplankton dynamics, it has the potential to dilute Chl *a*, especially during the heavy rainfall period. Precipitation was thus included for ANN model development.
3. Hydrological data were also included in these ANN models, given that hydrological conditions are an important factor that dominates the dynamics of the Lake Poyang ecosystem (Shankman et al. 2006).

The water quality data included four variables, i.e., the chlorophyll *a* concentration (Chl *a*, μg/l), water temperature (WT, °C), total phosphorus (TP, mg/l) and nitrogen (TN, mg/l). The Chl *a*, WT, TP and TN were obtained at 4 sampling sites (sites 2, 7, 11 and 12 in Fig. 1) by the Ministry of Environmental Protection of the People's Republic of China. These samples were collected seasonally between 2005 and 2008, and were obtained monthly between 2009 and 2011. In 2011, monthly sampling was carried out at another 13 sampling sites (Fig. 1). A total of 344 samples were collected during this 7-year period. The meteorological and hydrological conditions for these 344 samples were obtained from the following two daily datasets on the sampling day.

The daily meteorological data, i.e., precipitation (Pr, mm) and sunshine hours (PAR, h), were obtained from three weather stations of the China Meteorological Administration (W1–3 in Fig. 1). The meteorological conditions for each sampling site (sites 1–17 in Fig. 1) were represented by the data from its nearest weather station.

The daily hydrological data included discharge (Q, m^3/s) and water level (WL, m) of the outflow. These data were recorded at the hydrological station of site 1 (Fig. 1) by the Bureau of Hydrology, Changjiang Water Resources Commission. Although the Chl *a* dynamics were significantly affected by the five inflows (see "Study area") of Lake Poyang, their hydrological conditions were not included in the ANN model, because the discharge and water level of these five inflows were positively correlated with those of the outflow measured at site 1. For example, the measured data in 2010 showed a high correlation coefficient (0.83) between the outflow and largest inflow (Ganjiang) discharges.

Clustering approach

A clustering approach was used to cluster 17 sampling sites (Fig. 1) into two clusters by minimizing the distance of sampling sites in a cluster. Although many clustering techniques (e.g., k-means method, self-organizing map and hierarchical clustering) exist (Chang et al. 2011; Dietrich et al. 2013; Monem and Hashemy 2011), the k-means method was most widely used in similar classification cases to this study (Jain 2010), and has good performance in these case studies (e.g., Areerachakul and Sanguansintukul 2010; Ay and Kisi 2014). It was thus selected to cluster sampling sites in this study.

Environmental variables at a sampling site may change significantly in a year. Thus, yearly average data of TN,

Table 1 Data collected during 2005–2011 in Lake Poyang for developing ANN models

Item	Variable	Unit	Temporal resolution	Mean	Range
Water quality data	Chl *a*	μg/l	Monthly to seasonally	6.3	[0.02, 39.3]
	WT	°C	Monthly to seasonally	18.4	[2.0, 37.0]
	TP	mg/l	Monthly to seasonally	0.07	[0.01, 1.5]
	TN	mg/l	Monthly to seasonally	1.02	[0.19, 4.21]
Meteorological data	Pr	mm	Daily	3.7	[0, 53.6]
	PAR	h	Daily	4.5	[0, 12.5]
Hydrological data	Q	m^3/s	Daily	3601	[−3040, 19,100]
	WL	m	Daily	11.4	[7.5, 19.8]

Chl a chlorophyll *a* concentration, *WT* water temperature, *TP* total phosphorus, *TN* total nitrogen, *Pr* precipitation, *PAR* sunshine hours, *Q* discharge (minus value represents shows that the water from Yangtze River was has flowed into Lake Poyang), *WL* water level

TP, Chl *a*, Pr, PAR, WL and WT were used to represent the environmental characteristics of a sampling site. Euclidean metric was used to compute the distance between sampling sites and cluster centers. The main steps in implementing k-means clustering were given as follows (Jain 2010):

- Select an initial partition with two clusters, and repeat the following two steps until cluster membership stabilizes.
- Generate a new partition by assigning each pattern to its closest cluster center.
- Calculate new cluster centers.

Further details about the k-means clustering technique can be found in Jain (2010) and Liao (2005). The k-means clustering process was implemented in Matlab software (Table 4 in Appendix).

Artificial neural network models

The artificial neural network technique was used in this study based on two hypotheses: (a) the ANN technique could be used in spatial modeling of Chl *a* based on clustering of the sampling sites in Lake Poyang, and (b) phytoplankton biomass (represented by Chl *a*) could be predicted by the water quality and hydrological and meteorological conditions in Lake Poyang.

Based on clustering results using the k-means method, three ANN models (named ANN_Poyang, ANN_Downstream and ANN_Upstream) were developed to simulate Chl *a* in Lake Poyang. The ANN_Downstream and ANN_Upstream models were developed using field data from the sites in the downstream and upstream areas (Fig. 2), respectively. These two models aimed to simulate Chl *a* in the upstream and downstream areas of Lake Poyang, respectively. Their simulation results were compared with those of ANN_Poyang, which used all the field data (344 samples) to simulate Chl *a* of the whole lake. The comparison between ANN_Poyang and other two models (ANN_Downstream and ANN_Upstream) aimed to reveal the contribution of the clustering technique to improve the performance of ANN models. These ANN models were developed in Matlab software (Table 5 in Appendix). Further information on ANN implementation (data processing, model architecture, performance evaluation and model validation) is described below.

Data processing

The variables, resulting in a multicollinearity problem of the ANN models, were firstly detected using multicollinearity analysis (Demirhan 2014). These variables were then excluded for ANN model development. To ensure that all input variables received equal attention during the training process, the min–max normalization method was used to rescale all the variables to the range of 0.1–0.9. The dataset for the ANN models was randomly divided into three sub-sets (training, testing and validation datasets). The training dataset was used to estimate the unknown connection weights in the training algorithm. The testing dataset was used to prevent the ANN model from overfitting during the training process. The validation dataset was used to evaluate the generalization ability of the trained ANN model (Maier et al. 2010). The data division proportion for training, testing and validation datasets was 15:2:3 (Fig. 2). This data partition proportion has been widely used in previous studies (see the review by Maier and Dandy (2000)).

Model architecture

All these three ANN models included an input layer, a hidden layer and an output layer (Fig. 2). This structure has been widely used due to its flexibility and good performance (Hornik et al. 1989; Maier and Dandy 2000; Scardi and Harding Jr 1999). The input layer consisted of seven input variables with a linear transfer function, and the output layer included a single node of Chl *a* with a sigmoid transfer function. The hidden layer linked the input and output variables by its nodes. The node number of the hidden layer normally varied from 2 to 20, and was determined by a trial-and-error approach in this study.

A Levenberg–Marquardt backpropagation algorithm was used to train these three ANN models. The weights (w_{n+1}) in the Levenberg–Marquardt backpropagation algorithm were adjusted during training using the following equation (Parisi et al. 1996):

$$w_{n+1} = w_n - \eta_n d_n, \tag{1}$$

where w_n is the weight vector of the network, η_n is the step size, d_n is a vector defining the direction of descent, and n is the iteration number. The model training process was stopped when the errors for the testing set began to rise, or the training error (mean squared error) was lower than the goal value (0.0001), or a maximum of 1000 iterations was reached. This early stopping rule prevented the ANN models from overfitting.

Performance evaluation

The simulation results were compared with measurements at the 4 sampling sites (sites 2, 7, 11 and 12 in Fig. 1), because long-term measurement data were available at these 4 sampling sites (see Section "Data"). Six widely used indicators were used to compare the model fits of these three ANN models. The mathematical equations and the value ranges of these indicators are given in Table 2.

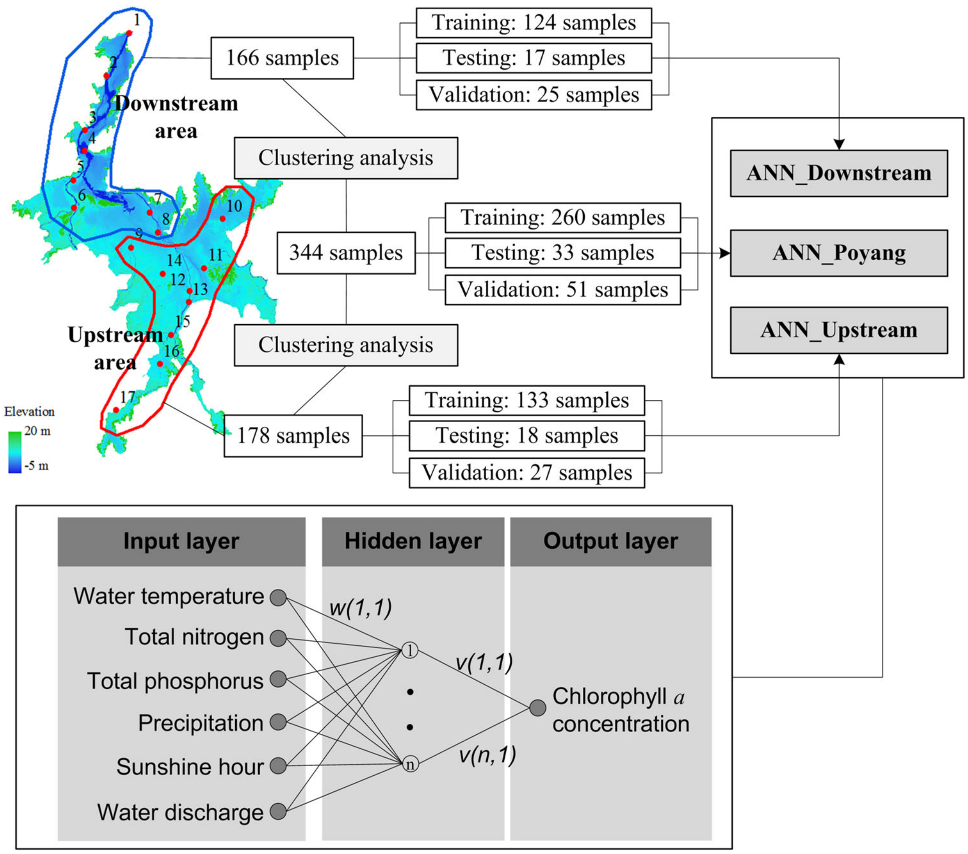

Fig. 2 Neural network structures of ANN_Poyang, ANN_Downstream and ANN_Upstream

Table 2 Indicators to evaluate model fits

Indicator	Description	Unit	Equation	Range
MAE	Mean absolute error	µg l^{-1}	$\mathrm{MA\%E} = \sum_{i=1}^{n} \left\lvert C_i - \hat{C}_i \right\rvert / n$	$[0, +\infty)$
MA%E	Mean absolute percent error	%	$\mathrm{MA\%E} = \sum_{i=1}^{n} \frac{\lvert C_i - \hat{C}_i \rvert}{\hat{C}_i} / n$	$[0, +\infty)$
RMSE	Root mean square error	µg l^{-1}	$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} (C_i - \hat{C}_i)^2 / n}$	$[0, +\infty)$
r^2	Coefficient of determination		$r_2 = \left(\frac{\sum_{i=1}^{n} (\hat{C}_i - \overline{\hat{C}})(C_i - \overline{C})}{\sqrt{\sum_{i=1}^{n} (\hat{C}_i - \overline{\hat{C}})^2 \sqrt{\sum_{i=1}^{n} (C_i - \overline{C})^2}}} \right)^2$	[0.0, 1.0]
d	Index of agreement		$d = 1 \frac{\sum_{i=1}^{n} (\hat{C}_i - C_i)^2}{\sum_{i=1}^{n} \left(\left\lvert C_i - \overline{\hat{C}} \right\rvert + \left\lvert \hat{C}_1 - \overline{\hat{C}} \right\rvert \right)^2}$	[0.0, 1.0]
NSE	Nash–Sutcliffe efficiency		$\mathrm{NSE} = 1 \frac{\sum_{i=1}^{n} (C_i - \hat{C}_i)^2}{\sum_{i=1}^{n} (C_i - \overline{\hat{C}})^2}$	$(-\infty, 1]$

C_i and $\hat{C}_i$ the simulation and measured chlorophyll *a* in Lake Poyang, $\overline{C}$ and $\overline{\hat{C}}$ the mean values of simulation and observed chlorophyll *a* in Lake Poyang $\left(\overline{C} = \sum_{i=1}^{n} C_i, \cdots \overline{\hat{C}} = \sum_{i=1}^{n} \hat{C}_i \right)$, *n* the number of days when both simulated and observed results are available during the modeling period

Mean absolute error (MAE) and root mean square error (RMSE) describe the model deviations with the same units as Chl *a* (i.e., µg l^{-1}). Mean absolute percent error (MA%E) is a dimensionless measure ranging from 0 to infinity. The MA%E value of 0 indicates that the simulation Chl *a* exactly agrees with the measured Chl *a*. Three correlation-based measures were used, including the coefficient of determination (r^2), index of agreement (*d*), and Nash–

Sutcliffe efficiency coefficient (NSE). Higher values of these three indicators imply better model fit. More details about these model fit indicators can be found in previous studies (e.g., Harmel and Smith 2007; Krause et al. 2005; Nash and Sutcliffe 1970).

Model validation

A cross-validation approach was used by partitioning the dataset into the training, testing and validation datasets. In the ANN_Poyang model, the full dataset included 344 samples. The data of the first 51 samples were used for validation. Then, the data of sequence samples (52–102, 103–153, 154–204, 205–255, 256–306, and 307–344) were used in turn for validation. This cross-validation strategy ensured that the data of each sample have been used for validation, and improved the stability of the ANN model developed with a limited amount of data. The Chl *a* output of ANN_Poyang was the mean Chl *a* output of seven ANN models in Fig. 3. This cross-validation strategy was also used for both ANN_Upstream and ANN_Downstream. Further details about this cross-validation method can be found in previous studies (Burden et al. 1997; Huang et al. 2014; Stone 1974).

Sensitivity analysis

One of the sensitivity analysis methods (mentioned as 'Perturb' method in Gevrey et al. 2003) was used to quantify the contributions of the environmental variables on the phytoplankton dynamics in different areas of Lake Poyang. In the sensitivity analysis, each input variable at each sampling site was changed by 10 %, while all other variables were kept fixed. All the input variables in ANN_Downstream and ANN_Upstream were tested. Thus, 14 simulations were simulated and compared with the base simulation (without changing any input variable) for each ANN model. The sensitivity value of each testing variable (*Sa*) was calculated by the relative change of the simulation Chl *a* due to its 10 % change:

$$\mathrm{Sa} = \frac{\left|C_i^{\mathrm{Test}} - C_i^{\mathrm{Base}}\right|}{C_i^{\mathrm{Base}} 10\ \%}, \tag{2}$$

where C_i^{Test} and C_i^{Base} are the simulation Chl *a* of sample *i* from the test simulation and base simulation, respectively. A variable with a higher sensitivity value implies that it is more important for phytoplankton dynamics.

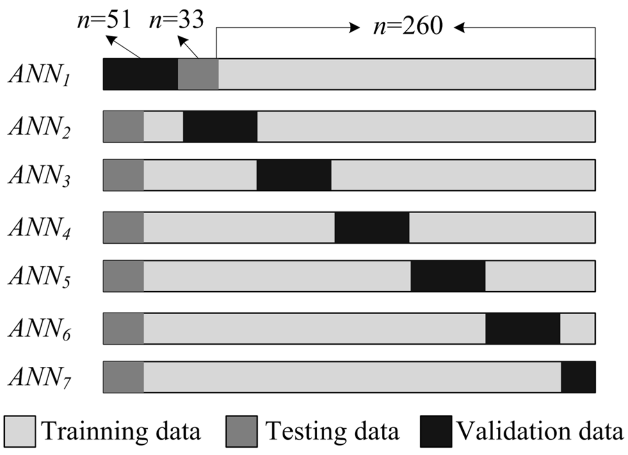

Fig. 3 Procedure of the cross-validation used in ANN_Poyang (redrawn from Huang et al. 2014). *n* represents the sampling number, ANN$_i$ (i = 1, …, 7) represents an ANN model

Results

ANN model development

Some results from the ANN model development can be listed as follows:

- *K-means clustering* The k-means clustering process resulted in two groups (sites 1–8 and 9–17 in Fig. 2). Sites 1–8 are located in the downstream area of Lake Poyang, and sites 9–17 are located in upstream area (Fig. 2).
- *Multicollinearity analysis* The multicollinearity analysis results showed that the variable of WL has a condition index of 17.3, implying a multicollinearity problem. This multicollinearity problem was overcome by excluding the variable of WL. There was no severe multicollinearity problem in the models using six variables without WL (Fig. 2).
- *Node number for hidden layer* Eight nodes for the hidden layer were found to be the optimal selection for these three ANN models using a trial-and-error approach (Sect. "Model architecture").
- *Model training* During the ANN model training process, none of the ANN models reached the goal and maximum iterations.

Chlorophyll *a* dynamics

Six model fit indicators were used to evaluate the model performances of ANN_Poyang, ANN_Downstream and ANN_Upstream (Table 3). The ANN_Downstream model well predicted the Chl *a* in the downstream area of Lake Poyang, with relatively high values of r^2, *d* and NSE. Its model fit was significantly better than that of ANN_Poyang. The model fit of ANN_Upstream was slightly better than that of ANN_Poyang.

Table 3 Model fits of ANN_Poyang, ANN_Downstream and ANN_Upstream

Area	Model	*n*	RMSE (μg l^{-1})	MAE (μg l^{-1})	MA%E (%)	r^2	*d*	NSE
Downstream	ANN_Poyang	166 (Fig. 2)	2.70	2.38	0.747	0.121	0.47	−0.355
	ANN_Downstream	166 (Fig. 2)	1.50	0.98	0.268	0.610	0.827	0.584
Upstream	ANN_Poyang	178 (Fig. 2)	5.14	3.33	4.712	0.396	0.623	0.298
	ANN_Upstream	178 (Fig. 2)	4.16	2.77	4.762	0.589	0.796	0.540

n sample number, *RMSE* root mean square error, *MAE* mean absolute error, *MA%E* mean absolute percent error, r^2 coefficient of determination, *d* index of agreement, *NSE* Nash–Sutcliffe efficiency

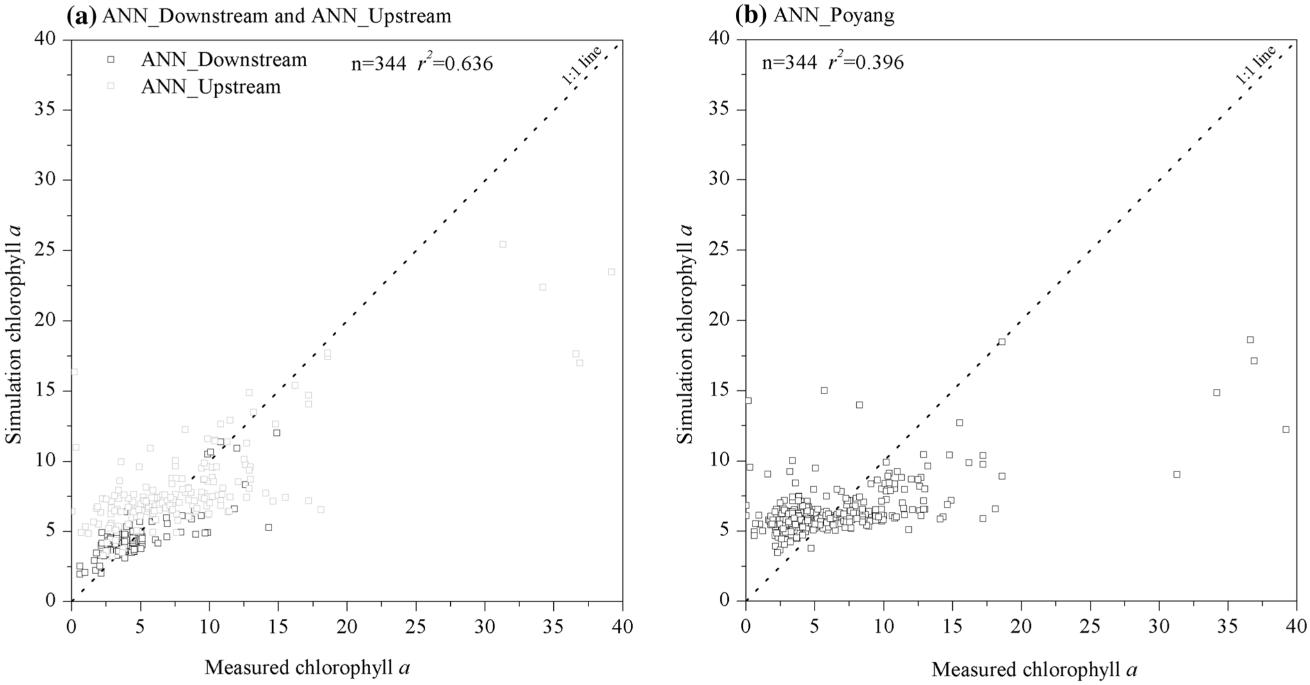

Fig. 4 Measured and simulation chlorophyll *a* of ANN_Downstream, ANN_Upstream and ANN_Poyang models in Lake Poyang

Figure 4 shows the simulation results of ANN_Downstream, ANN_Upstream and ANN_Poyang. It can be clearly seen that the Chl *a* dynamics in the downstream area was better simulated than that in the upstream area. Several Chl *a* peaks in the upstream area were underestimated.

The simulation results of ANN_Downstream and ANN_Upstream were compared with the simulation results of ANN_Poyang at four sampling sites (sites 2, 7, 11 and 12) with relatively long-term sampling data (Fig. 5). Site 2 was located in the downstream area of Lake Poyang. Its Chl *a* dynamics between 2005 and 2011 were predicted well by ANN_Downstream. However, ANN_Poyang showed clear overestimation of Chl *a* between 2005 and 2007. The model fit of ANN_Upstream at site 7 was better than that of ANN_Poyang. The Chl *a* dynamics at sites 11 and 12 were not adequately predicted by either ANN_Poyang or ANN_Upstream.

If the Poyang Dam was constructed in the downstream area, the environmental conditions in the downstream area would be more similar to those in the upstream area. To predict phytoplankton dynamics of the downstream area in the case of building a dam, simulation Chl *a* at site 2 from ANN_Upstream was compared with the simulation results from ANN_Downstream. The simulation Chl *a* from ANN_Upstream showed about 84 % increase compared with that from ANN_Downstream (Fig. 6).

Sensitivity analysis

Sensitivity values of the input variables in ANN_Downstream and ANN_Upstream were compared (Fig. 7). These values represented the sensitivity of Chl *a* dynamics to environmental variables in the downstream and upstream areas. It was found that Chl *a* dynamics were not sensitive to sunshine hours (PAR) and precipitation (Pr) in both the

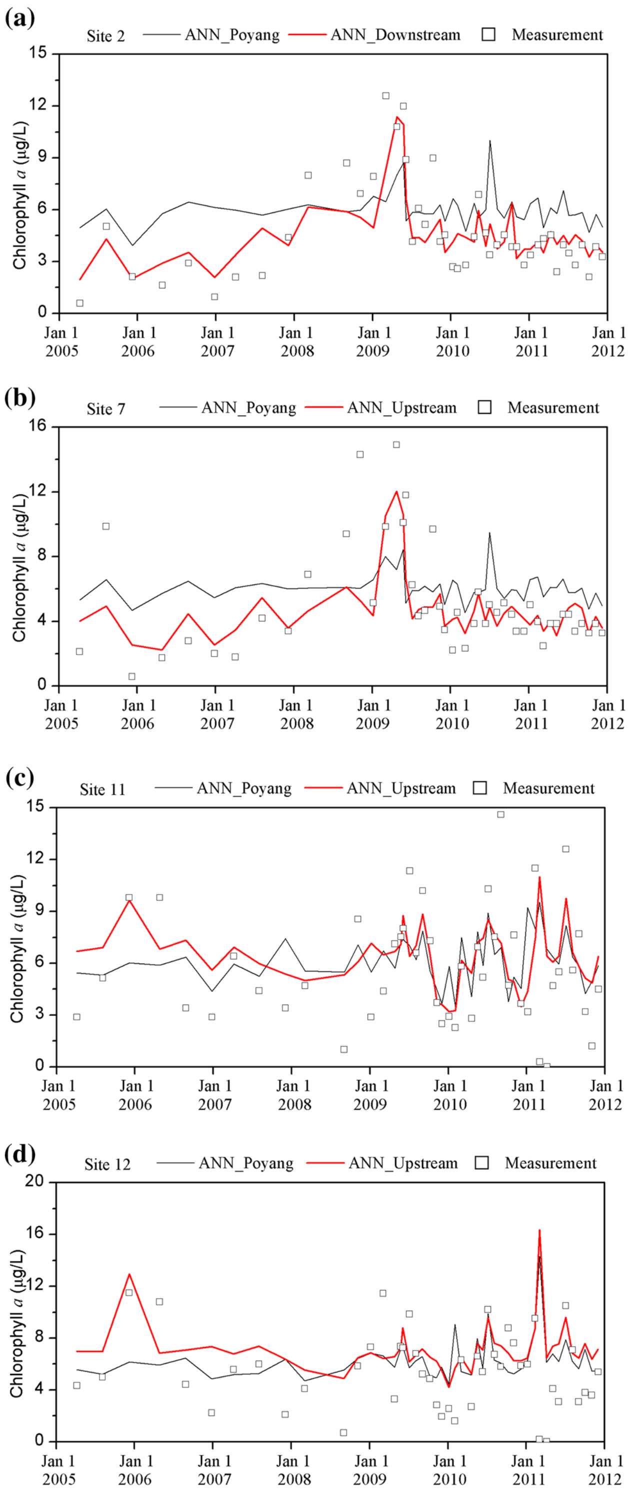

Fig. 5 Measured and simulation chlorophyll *a* of ANN_Poyang, ANN_Downstream and ANN_Upstream models at **a** site 2, **b** site 7, **c** site 11 and **d** site 12

upstream and downstream areas (sensitivity value less than 0.15). However, Chl *a* dynamics responded differently to environmental factors in different areas of Lake Poyang.

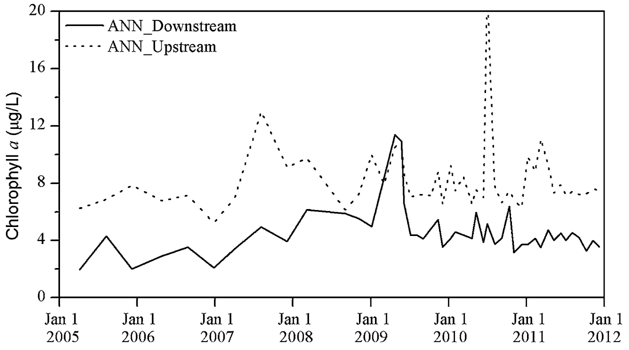

Fig. 6 Simulation chlorophyll *a* at site 2 using ANN_Downstream and ANN_Upstream

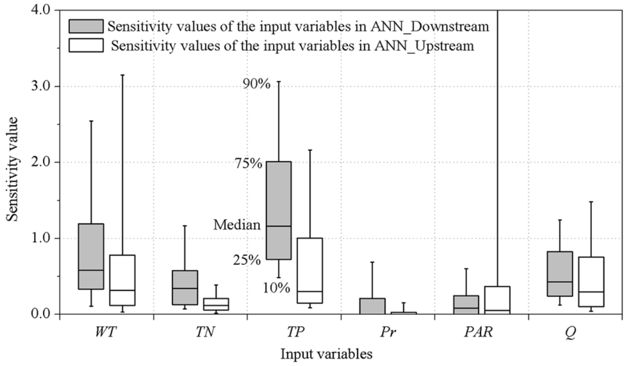

Fig. 7 Sensitivity values of the input variables in ANN_Downstream and ANN_Upstream. *WT* water temperature, *TN* total nitrogen, *TP* total phosphorus, *Pr* precipitation, *PAR* sunshine hours, *Q* discharge

In the downstream area, Chl *a* dynamics were sensitive to water temperature (WT), discharge (Q) and total phosphorus (TP), and were dominated by total phosphorus (TP) with its high sensitivity value at 1.25. In the upstream area, water temperature (WT) was the most sensitive variable with a sensitivity value of 0.32. The sensitivity values of several input variables (e.g., WT) had wide ranges.

Discussion

Performance of the ANN models

The model fit of ANN_Downstream (Table 3) was acceptable compared with previous case studies on Chl *a* predictions in two reviews (Arhonditsis and Brett 2004; Phillips et al. 2008). This acceptable model fit of ANN_Downstream was mainly attributed to the reasonable clustering of the sampling sites and the high potential of the ANN technique in simulating phytoplankton dynamics. It also demonstrated the high potential of the ANN technique in modeling aquatic systems with limited prior knowledge. It is particularly useful in case studies where the process mechanism has not been adequately studied and good model fits are preferred. However, the model fits of these three ANN models (Table 3) were not as good as the model fits of other water quality models using the ANN technique (Panda et al. 2010; Singh et al. 2009). This is because the phytoplankton dynamics in such a large lake, influenced by intensive human activity, are very complex and challenging to predict (Huang et al. 2014).

The sample amount was not large compared with previous case studies using ANN models (see the review by Maier and Dandy 2000; Maier et al. 2010). However, the cross-validation method was used to alleviate data limitation, and resulted in seven ANN models for each ANN model. Their model fits did not vary considerably, implying the reliability of these ANN models. Some methodological issues (e.g., data pre-processing, selection of input variables, determination of network architecture and choice of performance criteria) were determined based on previous studies (Bowden et al. 2005; Maier et al. 2010).

Contribution of clustering to ANN models

Based on the clustering results of 17 sampling sites, the Chl *a* dynamics in the upstream and downstream areas were simulated using ANN_Downstream and ANN_Upstream, respectively. The better performance of these two ANN models than ANN_Poyang (Table 3) indicated that the partition of the sampling sites was reasonable and necessary for such a large lake as Lake Poyang. The study demonstrated a successful clustering of spatial sites in a lake using the k-means clustering method, and was a primary step for the ANN technique in describing spatial heterogeneity. Modelers from other disciplines are encouraged to couple ANN with clustering techniques for modeling other target variables (e.g., precipitation and flow discharge) with high spatial heterogeneity. Many other clustering methods, such as self-organizing map, hierarchical clustering and the subtractive clustering method used in previous studies, could also be used for spatial clustering of a lake (Doan et al. 2005).

The relatively low model fit of ANN_Upstream implies that spatial differences of the Chl *a* dynamics at sites 9–17 (Fig. 1) existed. Further partition of these sampling sites would be helpful to improve the model fit. However, this was not carried out in this study due to the limited amount of field data. This shortage could be overcome with an increasing amount of measured data.

Influences of environmental conditions on chlorophyll *a* dynamics

Phytoplankton growth is widely recognized to be limited by temperature, nutrients and photosynthetically active radiation in lakes (Arhonditsis and Brett 2005; Fragoso et al. 2008). However, algal blooms have so far rarely been found in Lake Poyang, even with a relatively high nutrient concentration for phytoplankton growth. The sensitivity analysis results (Fig. 7) showed that hydrological conditions were important factors affecting phytoplankton growth in the downstream area of Lake Poyang. This conclusion was in agreement with previous studies on other lake ecosystems (Rangel et al. 2012; Souza Cardoso and Motta Marques 2009). From this perspective, the inclusion of hydrodynamic conditions is necessary for a proper prediction of phytoplankton biomass in Lake Poyang.

Nitrogen and phosphorus are major nutrients that limit phytoplankton growth in freshwater lakes. The sensitivity analysis results (Fig. 7) showed that phytoplankton growth in Lake Poyang was more limited by phosphorus than nitrogen. This result was reasonable and in accordance with previous studies based on TN/TP ratios (Smith 1982). It is generally recognized that phytoplankton growth is limited by phosphorus with a high TN/TP ratio (higher than 17 in Smith 1982), and was limited by nitrogen with a low TN/TP (lower than 10 in Smith 1982). The measured data showed that 60.2 % of the 344 samples had TN/TP ratios higher than 17, indicating phosphorus limitation of phytoplankton growth. Only 14.8 % of these samples had TN/TP ratios lower than 10, indicating nitrogen limitation of phytoplankton growth.

Although the nutrient conditions were not significantly heterogeneous in Lake Poyang, the sensitivity analysis results (Fig. 7) indicated that phytoplankton dynamics respond differently to total phosphorus between the downstream and upstream areas. Phytoplankton dynamics were sensitive to total phosphorus in the upstream area, while they were less sensitive to total phosphorus in the downstream area. This result was attributed to the dominance of hydrological conditions for phytoplankton dynamics in the downstream area.

The high sensitivity values of water temperature in ANN_Downstream and ANN_Upstream indicated that temperature was a key limiting factor for phytoplankton growth in both the downstream and upstream areas of Lake Poyang. This was consistent with previous studies on the relationship between water temperature and phytoplankton growth (Davis et al. 2009). The sensitivity analysis results showed that Chl *a* dynamics were slightly affected by precipitation, indicating a slight decrease of Chl *a* due to dilution. This was different from the lakes with severe algal blooms (e.g., Lake Taihu in China), where precipitation affected Chl *a* significantly, especially during the heavy rainfall period (Huang et al. 2014).

Implications for lake management

This study investigated the response of Chl *a* dynamics to environmental variables, and identified the important factors which dominated Chl *a* dynamics in both the downstream and upstream areas of Lake Poyang. Different water management strategies could thus be proposed for different areas of Lake Poyang. Phytoplankton dynamics were most sensitive to water temperature in the downstream area; however, they were most sensitive to total phosphorus in the upstream area. This indicated that reduction of external phosphorus loading could be used as a measure to prevent eutrophication in the upstream area of Lake Poyang .

The impacts of Poyang Dam construction on phytoplankton dynamics were investigated based on the simulation results in Fig. 6. The simulation Chl *a* at site 2 by ANN_Upstream was significantly higher (an increase of 83.8 %) than that by ANN_Downstream (Fig. 6), indicating that the Chl *a* concentration would increase in the case of building the Poyang Dam. The conclusion was in consensus with previous studies in other large reservoir dams in the world, such as the Three Gorge Dam, the Danube River Dam and the Aswan Dam (Humborg et al. 1997; White 1988; Zeng et al. 2006). From this perspective, adequate measures should be taken to alleviate or avoid algal blooms if the Poyang Dam is built.

Conclusions

Seventeen sampling sites at Lake Poyang were clustered into two groups using the k-means method, and an ANN model was developed for each group to simulate phytoplankton biomass. These two ANN models performed better than another ANN model that simulated phytoplankton biomass in the whole lake, implying the necessity of partitioning a large lake into two (or more) regions and developing an ANN model for each sub-region. Sensitivity analysis based on these two ANN models showed that phytoplankton dynamics responded differently to environmental variables in different sub-regions. This study demonstrated the high performance of ANN models and the value of coupling a clustering technique with an ANN model for environmental modeling. Supported by the increasing amount of spatial data, a combination of ANN and clustering techniques will be useful in investigating the spatial heterogeneity of complex natural ecosystems.

Acknowledgments The project was financially supported by National Basic Research Program of China (No. 2012CB417006). The authors would like to thank the Ministry of Environmental Protection of the People's Republic of China, Changjiang Water Resources Commission and the National Meteorological Information Center for providing the measured data for model development. We would like to thank the editor and reviewers for their extensive review, which significantly improved the manuscript.

Appendix

See below Appendix Tables 4 and 5.

Table 4 MATLAB code to implement clustering

```
% Read data from Excel file.
Data = xlsread('Data')
X = zscore(Data);

% K-Means clustering.
[idx,ctrs,sumD,D] = kmeans(X, 2, 'Replicates',5);
plot(X(idx==1,1),X(idx==1,2),'ro','MarkerSize',6)
hold on
plot(X(idx==2,1),X(idx==2,2),'bo','MarkerSize',6)
```

Table 5 MATLAB code to develop ANN model

```
% Read input and output data from Excel file.
Ins = xlsread('Data_Ins');
Outs = xlsread('Data_Outs');

% Read the disorder number from 1-344.
ind = xlsread('Data_Order');
trainInd = ind(85:344);
valInd = ind(1:51);
testInd = ind(52:84);

% Construct training, validation and testing sampling.
[trainP,valP,testP] = divideind(Ins,trainInd,valInd,testInd);
[trainT,valT,testT] = divideind(Outs,trainInd,valInd,testInd);
trainSamples = [];
trainSamples.P = trainP;
trainSamples.T = trainT;
validateSamples = [];
validateSamples.P = valP;
validateSamples.T = valT;
testSamples = [];
testSamples.P = testP;
testSamples.T = testT;

% Model development.
net=newff(minmax(Ins),[8,1],{'tansig','purelin'},'trainlm');
net.trainParam.epochs = 1000;
net.trainParam.show = 100;
net.trainParam.lr = 0.1;
net.trainParam.goal =0.0001;
net.trainParam.mc=0.9;
net = init(net);

% [net,tr]=train(net,InTrain,MeaTrain);
[net,tr] =
train(net,trainSamples.P,trainSamples.T,[],[],validateSamples,testSamples);

% Model simulation using the training and validation datasets.
SimTrain = sim(net,trainSamples.P);
SimVal = sim(net,validateSamples.P);

% Save the training plot, network and the simulation reslults.
saveas(gcf, 'Train.jpg', 'jpg');
save('ANN.mat', 'net');
sims = [SimTrain; SimTrain];
Arr2Txt(sims', 'Results.txt');
```

References

Areerachakul S, Sanguansintukul S (2010) Clustering analysis of water quality for canals in Bangkok, Thailand. In: Taniar D, Gervasi O, Murgante B, Pardede E, Apduhan B (eds) Computational science and its applications. Springer, Heidelberg, pp 215–227

Arhonditsis GB, Brett MT (2004) Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Mar Ecol Prog Ser 271:13–26

Arhonditsis GB, Brett MT (2005) Eutrophication model for Lake Washington (USA): part I. Model description and sensitivity analysis. Ecol Model 187:140–178

Ay M, Kisi O (2014) Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. J Hydrol 511:279–289

Bowden GJ, Dandy GC, Maier HR (2005) Input determination for neural network models in water resources applications. Part 1–background and methodology. J Hydrol 301:75–92

Burden FR, Brereton RG, Walsh PT (1997) Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy. Analyst 122:1015–1022

Celebi ME, Kingravi HA, Vela PA (2013) A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst Appl 40:200–210

Chang K, Gao J, Wu W, Yuan Y (2011) Water quality comprehensive evaluation method for large water distribution network based on clustering analysis. J Hydroinform 13:390–400

Davis TW, Berry DL, Boyer GL, Gobler CJ (2009) The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of *Microcystis* during cyanobacteria blooms. Harmful Algae 8:715–725

Demirhan H (2014) The problem of multicollinearity in horizontal solar radiation estimation models and a new model for Turkey. Energy Convers Manag 84:334–345

Dietrich JP, Popp A, Lotze-Campen H (2013) Reducing the loss of information and gaining accuracy with clustering methods in a global land-use model. Ecol Model 263:233–243

Doan C, Liong S, Karunasinghe D (2005) Derivation of effective and efficient data set with subtractive clustering method and genetic algorithm. J Hydroinform 7:219–233

Elangasinghe MA, Singhal N, Dirks KN, Salmond JA, Samarasinghe S (2014) Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. Atmos Environ 94:106–116

Feng L, Hu C, Chen X, Li R, Tian L, Murch B (2011) MODIS observations of the bottom topography and its inter-annual variability of Poyang Lake. Remote Sens Environ 115:2729–2741

Fragoso CR Jr, Marques DMLM, Collischonn W, Tucci CEM, Van Nes EH (2008) Modelling spatial heterogeneity of phytoplankton in Lake Mangueira, a large shallow subtropical lake in South Brazil. Ecol Model 219:125–137

Gevrey M, Dimopoulos I, Lek S (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Model 160:249–264

Hamilton DP, Schladow SG (1997) Prediction of water quality in lakes and reservoirs. Part I model description. Ecol Model 96:91–110

Harmel RD, Smith PK (2007) Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. J Hydrol 337:326–336

Harmel RD, Smith PK, Migliaccio KW, Chaubey I, Douglas-Mankin KR, Benham B, Shukla S, Muñoz-Carpena R, Robson BJ (2015) Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: a review and recommendations. Environ Model Softw 57:40–51

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2:359–366

Huang J, Gao J, Hörmann G (2012a) Hydrodynamic-phytoplankton model for short-term forecasts of phytoplankton in Lake Taihu, China. Limnologica 42:7–18

Huang J, Gao J, Hörmann G, Mooij WM (2012b) Integrating three lake models into a Phytoplankton Prediction System for Lake Taihu (Taihu PPS) with Python. J Hydroinform 14:523–534

Huang J, Gao J, Mooij WM, Hörmann G, Fohrer N (2014) A comparison of three approaches to predict phytoplankton biomass in Gonghu Bay of Lake Taihu. J Environ Inform 24:39–51

Huang J, Gao J, Xu Y, Liu J (2015a) Towards better environmental software for spatio-temporal ecological models: lessons from developing an intelligent system supporting phytoplankton prediction in lakes. Ecol Inform 25:49–56

Huang J, Gao J, Zhang Y, Xu Y (2015b) Modeling impacts of water transfers on alleviation of phytoplankton aggregation in Lake Taihu. J Hydroinform 17:149–162

Humborg C, Ittekkot V, Cociasu A, Bv Bodungen (1997) Effect of Danube River dam on Black Sea biogeochemistry and ecosystem structure. Nature 386:385–388

Huszar V, Caraco N, Roland F, Cole J (2006) Nutrient–chlorophyll relationships in tropical–subtropical lakes: do temperate models fit? Biogeochemistry 79:239–250

Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31:651–666

Jeong KS, Kim DK, Jung JM, Kim MC, Joo GJ (2008) Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshwater phytoplankton dynamics. Ecol Model 211:292–300

Jørgensen SE (2008) Overview of the model types available for development of ecological models. Ecol Model 215:3–9

Jørgensen SE (2010) A review of recent developments in lake modelling. Ecol Model 221:689–692

Jørgensen SE, Bendoricchio G (2001) Fundamentals of ecological modelling. Elsevier, Amsterdam

Krause P, Boyle D, Bäse F (2005) Comparison of different efficiency criteria for hydrological model assessment. Adv Geosci 5:89–97

Liao TW (2005) Clustering of time series data—a survey. Pattern Recogn 38:1857–1874

Lu M, Zeng D, Liao Y, Tong B (2012) Distribution and characterization of organochlorine pesticides and polycyclic aromatic hydrocarbons in surface sediment from Poyang Lake, China. Sci Total Environ 433:491–497

Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ Modell Softw 15:101–124

Maier HR, Jain A, Dandy GC, Sudheer KP (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ Modell Softw 25:891–909

Monem M, Hashemy S (2011) Extracting physical homogeneous regions out of irrigation networks using fuzzy clustering method: a case study for the Ghazvin canal irrigation network. J Hydroinform 13:652–660

Mooij W, Janse J, De Senerpont Domis L, Hülsmann S, Ibelings B (2007) Predicting the effect of climate change on temperate shallow lakes with the ecosystem model PCLake. Hydrobiologia 584:443–454

Mooij WM, Trolle D, Jeppesen E, Arhonditsis G, Belolipetsky PV, Chitamwebwa DBR, Degermendzhy AG, DeAngelis DL, De Senerpont Domis LN, Downing AS, Elliott JA, Fragoso CR Jr, Gaedke U, Genova SN, Gulati RD, Håkanson L, Hamilton DP, Hipsey, ‘tHoen J, Hülsmann S, Los FH, Makler-Pick V, Petzoldt T, Prokopkin IG, Rinke K, Schep SA, Tominaga K, Van Dam AA, Van Nes EH, Wells SA, Janse JH (2010) Challenges and opportunities for integrating lake ecosystem modelling approaches. Aquat Ecol 44:633–667

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I. A discussion of principles. J Hydrol 10:282–290

Panda RK, Pramanik N, Bala B (2010) Simulation of river stage using artificial neural network and MIKE 11 hydrodynamic model. Comput Geosci 36:735–745

Parisi R, Di Claudio ED, Orlandi G, Rao BD (1996) A generalized learning paradigm exploiting the structure of feed forward neural networks. IEEE Trans Neural Netw 7:1450–1460

Phillips G, Pietiläinen OP, Carvalho L, Solimini A, Lyche Solheim A, Cardoso AC (2008) Chlorophyll–nutrient relationships of different lake types using a large European dataset. Aquat Ecol 42:213–226

Rangel L, Silva LS, Rosa P, Roland F, Huszar VM (2012) Phytoplankton biomass is mainly controlled by hydrology and phosphorus concentrations in tropical hydroelectric reservoirs. Hydrobiologia 693:13–28

Rao Z, Alvarruiz F (2007) Use of an artificial neural network to capture the domain knowledge of a conventional hydraulic simulation model. J Hydroinform 9:15–24

Robson BJ (2014) When do aquatic systems models provide useful predictions, what is changing, and what is next? Environ Model Softw 61:287–296

Scardi M, HardingJr LW (1999) Developing an empirical model of phytoplankton primary production: a neural network case study. Ecol Model 120:213–223

Shankman D, Keim BD, Song J (2006) Flood frequency in China’s Poyang Lake region: trends and teleconnections. Int J Clim 26:1255–1266

Singh KP, Basant A, Malik A, Jain G (2009) Artificial neural network modeling of the river water quality: a case study. Ecol Model 220:888–895

Smith VH (1982) The nitrogen and phosphorus dependence of algal biomass in lakes: an empirical and theoretical analysis. Limnol Oceanogr 27:1101–1112

Souza Cardoso L, Motta Marques D (2009) Hydrodynamics-driven plankton community in a shallow lake. Aquat Ecol 43:73–84

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc 36:111–147

Wang S, Dou H (1998) Record of Chinese lakes (in Chinese). Science Press, Beijing

White GF (1988) The environmental effects of the high dam at Aswan. Environ Sci Policy Sustain Develop 30:4–40

Wu N, Schmalz B, Fohrer N (2011) Distribution of phytoplankton in a German lowland river in relation to environmental factors. J Plankton Res 33:807–820

Wu N, Huang J, Schmalz B, Fohrer N (2014a) Modeling daily chlorophyll a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches. Limnology 15:47–56

Wu W, Dandy GC, Maier HR (2014b) Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environ Modell Softw 54:108–127

Wu Z, He H, Cai Y, Zhang L, Chen Y (2014c) Spatial distribution of chlorophyll a and its relationship with the environment during summer in Lake Poyang: a Yangtze-connected lake. Hydrobiologia 732:61–70

Wu Z, Wang J, Xia Y, Chen Y (2014d) A record of bloom in Lake Poyang: *Hydrodictyon reticulatum* (in Chinese). J Lake Sci 26:481–484

Ye X, Zhang Q, Liu J, Li X, Xu C (2013) Distinguishing the relative impacts of climate change and human activities on variation of streamflow in the Poyang Lake catchment, China. J Hydrol 494:83–95

Yi R, Chen Y (2015) http://news.sciencenet.cn/html/showsbnews1.aspx?id=192119. Accessed on April 15, 2015

Yi R, He D (2012) May Poyang Dam change River-Lake relationship? (in Chinese). China Three Gorges 5:47–49

Zeng H, Song L, Yu Z, Chen H (2006) Distribution of phytoplankton in the Three-Gorge Reservoir during rainy and dry seasons. Sci Total Environ 367:999–1009

Zhang H, Culver DA, Boegman L (2008) A two-dimensional ecological model of Lake Erie: application to estimate dreissenid impacts on large lake plankton populations. Ecol Model 214:219–241