



Forecasting performance of algae blooms based on artificial neural networks and automatic observation system

Mi Eun Kim^a, Tae Seok Shon^b, Kyung Sok Min^c, Hyun Suk Shin^{a,*}

^aDepartment of Civil Engineering, Pusan National University, Busan, 609-735, Korea
Tel. +82 51 510 3288; Fax: +82 51 517 3287; email: hsshin@pusan.ac.kr

^bBrain Korea 21 Division for Ubiquitous-Applied Construction of Port Logistics Infrastructures, Pusan National University, Busan, 609-735, Korea

^cDepartment of Environmental Engineering, Kyungbook National University, Daegu, 702-701, Korea

Received 14 October 2011; Accepted 24 December 2011

ABSTRACT

In recent, South Korea has faced on water quality management problems in reservoir and river because of increasing water temperature (T_w) and rainfall frequency caused by climate change. For these reasons, South Korea has set up automatic water quality monitoring system for preventing early algae blooms in five representative watershed. Also, Government makes a greater effort for preparing remedy with numerical models which handle water quality accidents quality in advance by predicting variation of water quality factors on account of change of weather conditions and source of pollutants in the future. Many countries have conducted various studies on forecasting water quality by artificial neural network which has a number of advantages, as opposed to the traditional models based on methods like data driven self-adaptive methods, generalization through learning the data presented, universal functional approximators, and nonlinear for forecasting. Daecheong reservoir located in Geum river has suitable environment for algae blooms because it has lots of contaminants that are flowed by rainfall in Daejeon and Chungcheong area. This study selected Daecheong reservoir in the Geum river and used the data of the automatic water quality observation system. By using back propagation algorithm (BPNN) of feed forward neural networks, a model has been built to forecast the algae blooms over short periods (1, 3 and 7 d). In this model, input parameters considered the hydrologic and water quality factors as following: inflow, outflow, average areal precipitation, air temperature (T_a), T_w , dissolved oxygen (DO), total organic carbon (TOC), pH, chlorophyll-a (chl-a), total nitrogen (TN), and total phosphorous (TP) in Daecheong reservoir. However, the chl-a data of automatic water quality observation system has some missing data caused by defect and maintenance in the system. Through carrying out correlation analysis, interpolation has been implemented by neural network with BPNN. Correlation analysis has been implemented to analyze lag time and components that sensitively responded to chl-a by referring the interpolated data and water quality and hydrologic factors in all. Based on the results of the data, after selecting input parameters for algae bloom prediction model, each case has been verified along with making various models. As a result of this research, the short term algae bloom prediction models showed minor errors in the prediction of the 1 d and the 3 d. Components of water quality such as T_w , pH, and TOC showed the closest correlation with chl-a and the models have been built with them. Therefore, the models will be very effective to control the water quality of Daecheong reservoir in South Korea by predicting a day to seven days.

Keywords: Artificial neural network; Forecasting; Algae bloom; Chlorophyll-a

*Corresponding author.

1. Introduction

In recent, the indiscriminate industrialization and two urban concentration have undergone difficulty of the water quality management since 1970s. Especially, in the case of reservoir, water quality management problems to conserve ecological system have raised owing to characteristics such as little variation of flow velocity and the obvious stratified phenomenon during summer. In these days, there are not only getting more leisure activities around the river and reservoir due to increasing concerns for recreation activities but also the reservoir is used to serve many purposes which are to supply water in urban and industry. Therefore, various problems related to water quality management have been generated in reservoir. These affect greatly the occurrence of water pollution on the main reason, which occur on account of the key limiting nutrient brought into the reservoir. The inflow of the nutrient to phytoplankton growth such as nitrogen or iron makes the ecological system more complicated known as eutrophication. Eutrophication is the movement of a body of water's trophic status in the direction of increasing plant biomass, by the addition of artificial or natural substances, such as nitrates and phosphates to an aquatic system. The phenomenon caused by a variety of problems can occur where eutrophic conditions interfere with drinking water treatment. That is, provided that algae which is a type of plant with no stems or leaves that grows in water or on the surfaces is enhanced by nutrient, dissolved oxygen (DO) concentration on aquatic ecosystem of standard decreases steadily. After that pH and biochemical oxygen demand (BOD) concentration increased contrastively. These kinds of relation of water quality and hydrologic factors cause destruction in ecosystem balance. This reflects having immediate connection between algae growth and water quality factors.

In particular, total phosphorus (TP) total nitrogen (TN) as key factors for evaluating eutrophication are represented in reservoir and river, which are closely related to chlorophyll-a (chl-a) being essential factor to raise eutrophication before water quality accidents are seriously caused. All things considered, accidents related to water quality are able to prevent in advance, providing that effective forecasting models are established.

As the beginning of Streeter-Phelps model, various models have been developed to simulate the phenomenon of water quality's change quantitatively. The developed models like QUAL2E, WASP, and artificial neural network (ANN) have been acknowledged and used frequently. Especially, ANN model has been used at home and abroad widely for purpose of supervision through water quality prediction. Such water quality prediction model has been built by using ANN and widely used on the research such as prediction of time series'

change and evaluation of adaptability. A research on the algae blooms forecast using ANN has been widely researched from various countries since 1990. Generally, the researches usually study on water quality components and prediction for 1 d to 5 d, or even a month by using back propagation algorithm (BPNN).

Friedrich et al. [1] used water quality data that had been measured for 12 y as input data of ANN which was orthophosphate, nitrate, secchi depth (SD), water depth, DO, water temperature (Tw), and chl-a. The models were to predict algal species. Also, ANN model for predicting species abundance and succession of blue-green algae was developed, which was to control of harmful algae blooms with water quality factors such as SD, DO, solar radiation, Tw, chl-a, nitrate, and orthophosphate [2]. Karul et al. [3] had considered various water quality factors to analyze and predict chl-a in the reservoir of river, a neural network prediction model was built by a variety of water quality factors (phosphorus, nitrogen, alkalinity, suspended solids, pH, Tw, DO, electrical conductivity) as in Levenberg-Marquardt (tangent-sigmoid) structure. This was to model non-linear behavior in eutrophication process. Hugh et al. [4] used genetic ANN model for dynamic predictions of algal abundance with the key driving variables which were phosphorus, nitrogen, underwater light, and Tw. The models were to predict over 30-d. On the other hand, Dogan et al. [5] had built the BOD prediction model by applying BPNN of neural network with chemical oxygen demand, temperature, DO, water flow, chl-a, and nutrients for 2 y among the water factors. The model was established because BOD was an important parameter for usage conditions of surface water. Kuo et al. [6] used neural network to inquire the relation with the key factors of eutrophication that influence a number of water quality indicators such as DO, TP, chl-a, and secchi disk depth in a reservoir. Jeong et al. [7] developed a temporal autoregressive recurrent neural network model that could predict time series changes as a month and a day on phytoplankton dynamics in river. The model was based on BPNN of neural network as inputs which were only chl-a in a month of lag time. Palani et al. [8] forecasted the chl-a a week ahead using general regression neural networks. The forecast model was based on DO, temperature, and chl-a with lag time of two weeks as input data, which was to show the eutrophication dynamics with respect to time and space. Singh et al. [9] described ANN models for computing the DO and BOD levels with measured water quality data for 10 s. The models were established by pH, total solids, chloride, phosphate, 5-d BOD, DO, nitrate nitrogen, and chemical oxygen demand as input data.

According to the previous researches, it was found that there were many prediction researches for each

components of water quality such as BOD and DO, but there are rarely many prediction researches for checking concentration of chl-a by considering water quality and hydrologic factors as well as the models predicted chiefly chl-a by a week or a month. The concentration of chl-a is a main factor that causes eutrophication. If short term forecast about this element is implemented, water quality management will be much more efficient and it is able to prevent the water pollution beforehand.

There is a current pending issue under the conditions of water quality in South Korea. As it is, water quality management in Daecheong reservoir is needed because of the sudden increase of algae blooms in Geum river.

The purpose of this study is to develop chl-a forecast models by a day considering the most important factors though correlation analysis among the water quality and hydrologic factors using back propagation neural network for effective water quality management in reservoir of South Korea.

The purposes of this research are to develop the chl-a forecasting model based on ANNs and to evaluate its performance for real management. In doing this, the factors affecting algae blooms including climatic, hydrologic, and water quality data have been fully and statistically considered. These models could be used as a forecasting and warning tool for preventing eutrophication problems in real field.

2. Materials and methods

2.1. Study area

Daecheong reservoir is one of the major water supplies in Korea. The length of the dam is 495 m, the volume is 1,234,000 m³, and the height is 72 m. Its basin area is 3204 km² excluding Yongdam reservoir basin. The dam is used as multi-purposes as water supply and industrial water. Daecheong reservoir stores 1,649 m² of water supply and 1,300 m² of the water uses for living and industrial water. Also, it is greatly important in a broad sense because the area offers people around the reservoir outdoor spaces for leisure activities. This basin's precipitation converges in June to August, and the maximum rainfall distributes in July to August. And the dry season, which is January to April and November to December, has only 15% of the whole precipitation for a year. This is the typical form of Korea. For this circumstance, possibility for inflow of contamination is higher compared to other countries. Accordingly, automatic water quality observation system has been installed to supervise the change of the water quality. However, as of 2011, this year alone there have been several occurrences of algae blooms due to decreased water level in the area. Especially, algae blooms caution is consistently

alerted in every summer since 2001. This is the main reason of summer eutrophication, which is caused from the long period of water's stagnation and its contaminant. In light of this, the Korean government has demonstrated standards of the algae bloom forecast by selecting the water quality factors such as Tw, chl-a, and TP in the area. This research selected the basin as the suitable site to establish the effective model of algae bloom forecast over short periods (Fig. 1).

2.2. Data description

The data used on this study has been collected from automatic water quality observation system in Daecheong reservoir. This has various water quality variables Tw, pH, DO, total organic carbon (TOC), TN, TP, chl-a from 2009 to 2010 and hydrological variables (inflow quantity of Daecheong reservoir, outflow quantity, average areal precipitation, air temperature (Ta)) in Fig. 2. Both water quality and hydrologic data are to analyze what are the most effective factors related to the growth of algae. The period of the data's usage is all the same, but chl-a of the data collected from automatic water quality observation system has missing values over some periods unlike other water quality factors.

2.3. Theory of back propagation neural network

ANN with BPNN is considerably useful owing to its broad applicability related to many problems such as principal prediction and modeling on the various



Fig. 1. Study area on Geum river basin in South Korea.

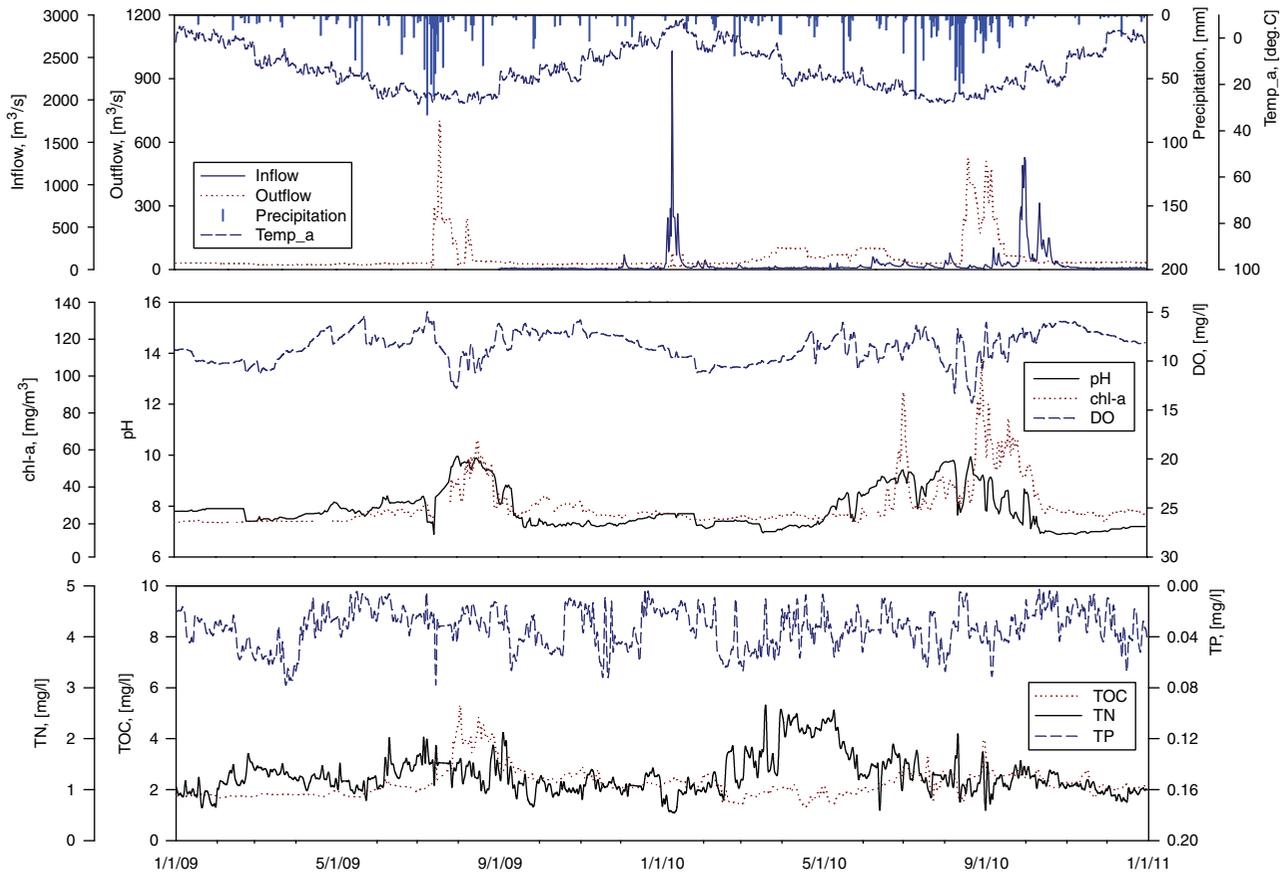


Fig. 2. Time series of observed data used in this study at Geum river basin.

purposes. According to a supervised learning method in the process, this neural network requires a set of training data in order to learn the relationships among several data and testing data to validate.

The architecture of BPNN is made of three nodes as input, hidden, and output node. The node to node like input-hidden and hidden-output is connected by weights and bias.

In building an BPNN, one must select an appropriate activation function, determine the number of hidden nodes, and estimate the corresponding parameters by using an approximate computational scheme. The objective is to find the reasonable BPNN that will give an approximation to true output within a specified error. The use of approximating functions is needed for superposition of sigmoid and hyperbolic tangent as following:

$$\text{Sigmoid function: } \phi(v) = \frac{1}{1 + \exp(-v)}$$

$$\text{Hyperbolic tangent: } \phi(v) = \frac{1 - \exp(-v)}{1 + \exp(-v)}$$

An description of BPNN system considering the hyperbolic tangent activation function can be derived as:

$$\tilde{y}_t^{(k)}(x) = \gamma_k + \sum_{j=1}^h a_{jk} \tanh \left[\beta_j + \sum_{i=1}^n [\omega_j^{(i)} x_t^{(i)}] \right]$$

where, the coefficients as γ , α , β , and ω are parameters of the ANN model. The coefficient γ_k is associated with the output node k , the coefficient α_{jk} is associated with the hidden node j and output node k , the coefficient β_j is associated with the hidden node j only, and the coefficient $\omega_j^{(i)}$ is associated with the input (i) and the hidden node j .

Each output node receives data through each weighted value of all the hidden nodes. Each node prints out the result value by changing added up data that has been used with non-linear function for producing values. This converges as the output, which is approached by ANN using BPNN in the research.

2.4. Methods

We develop the BPNN for daily chl-a forecasting model, called Algae Blooms-Real Time Forecast-Neural

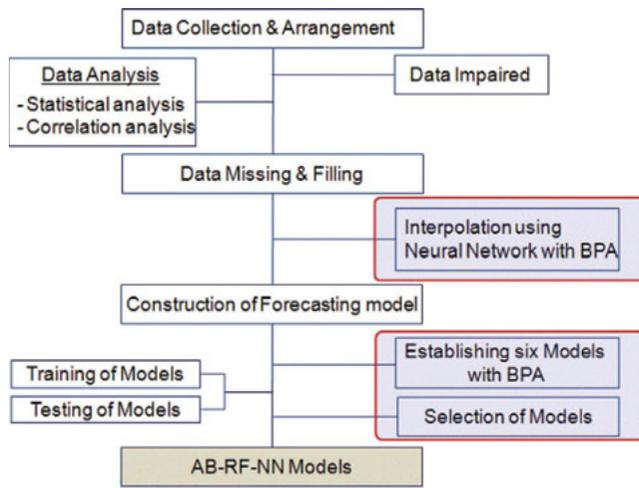


Fig. 3. Flow chart for constructing AB-RF-NN models.

Networks (AB-RF-NN). Fig. 3 generally shows process establishing AB-RF-NN model which is flow chart on this study.

First of all, the analysis of data was implemented for basic statistical and correlation analysis between elements of hydrology and water quality which directly affect algae blooms. In case of correlation analysis, the research performed serial correlation and cross correlation analysis to grasp the relationship among the elements. The theory related to correlation analysis is following as:

To check the seasonal change and cyclical repeatability of the time series data, serial correlation analysis (r_k) can derive the following Eq. (1):

$$r_k = \frac{\sqrt{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (1)$$

\bar{x} means the average, and n means the number of the analyzed data.

Cross correlation analysis used to associate water quality and hydrologic elements can derive the following Eq. (2):

Table 1

Structure of AB-RF-NN models with selected water quality variables in the study

Output	Model	Factors of input data for neural network						
		If	Tw	pH	DO	TOC	TP	Chl-a
Chl-a	1	-	-	-	-	-	-	-
	2	-	-	X	X	-	-	X
	3	-	-	X	X	X	-	X
	4	X	-	X	X	-	-	X

$$r_{xy}(k) = \frac{c_{xy}(k)}{S_x S_y} \quad (2)$$

C_{xy} means covariance of x and y , S_x and S_y means standard deviation of x , y .

By using the methods mentioned above, chl-a with some missing values among the water quality factors could be supplemented by ANN. After that, the result performed the validation over peak periods in chl-a on accuracy of produced values over the missing data.

In this research, the AB-RF-NN models for algae bloom forecast with neural network over the short terms have been built by inflow (If), Tw, pH, DO, TOC, TP, and chl-a as shown in Table 1. But one disadvantage of the AB-RF-NN models is rarely to set peak value to train. For this reason, a set of training data was used for 2010, a set of testing data was collected for 2009. For calibration/validation activity on the models, performance functions between the observed values and the calculated output are used as following; RMSE (Root Mean Square Error), and R^2 (Correlation Coefficient).

3. Results and discussions

First, the result of statistical analysis showed similar trends as two groups which are hydrologic factors group and water quality factors group in Table 2.

Table 2
Result of basic statistical analysis for variables in concern

	If	Of	Precip	Ta	Tw	pH	DO	TOC	TN	TP	Chl-a
Max	2573.3	700	78.40	28.12	31.11	9.96	14.26	5.32	2.58	0.08	89.30
Min	0.0	14.70	0.0	-8.02	3.41	6.90	4.99	1.32	0.54	0.0	0.15
Average	60.58	55.82	3.00	12.98	16.94	7.86	8.54	2.24	1.28	0.03	9.53
Stdev	159.64	73.03	8.76	9.50	8.53	0.76	1.55	0.56	0.38	0.02	11.98
Skew	8.46	4.27	4.63	-0.26	-0.09	1.05	0.41	1.93	1.20	0.38	2.68
Kurt	101.04	21.61	25.01	-1.11	-1.45	0.19	-0.18	5.23	1.54	-0.38	8.33

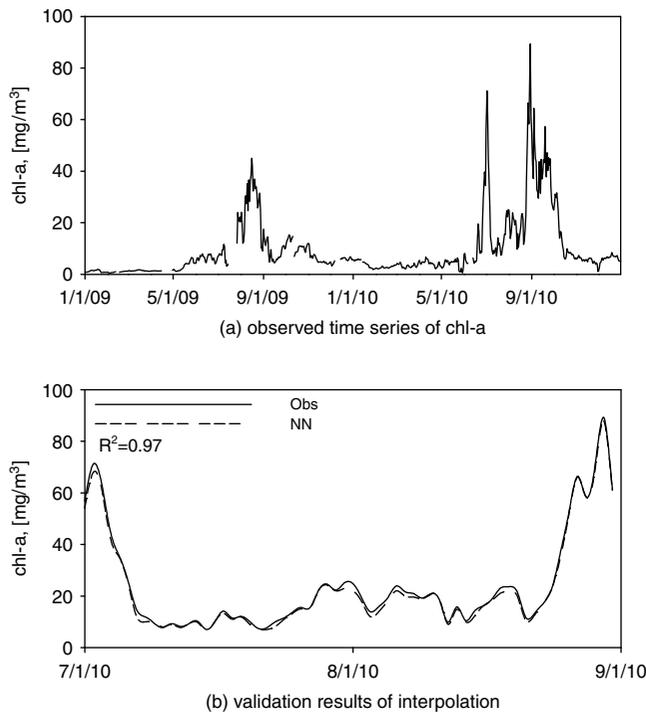


Fig. 4. Results of interpolating missing data of chl-a using AB-RF-NN model.

Fig. 5 described results of AB-RF-NN models established in these results. As seen in Fig. 5, when all of the factors were considered in model 1, the result was successful that correlation coefficient was over 0.9 except 7 d. But, in case of model 3, compared to model 4, the result was not good that R^2 was around 0.6. In this result, the forecast of chl-a was highly sensitive reaction to TOC factor. The chl-a concentration had significant correlation with TOC among the water quality factors. Although, the inflow factor was not sensitive to estimate algae growth as known in model 2 and 4, the results showed correlation coefficient in over 0.8 ranges. It is clear that the factor is important to manage water quality management in reservoir. Overall, the factors affecting the eutrophication were not only inflow, Tw, TOC, and TP, but also pH and DO with lag time of 1–3 d had an important effect on that before chl-a occurred at $t+1$. The result in Table 5 was shown architecture of AB-RF-NN models for forecasting chl-a by training which exactly described what used as input factors with lag time and architecture (I-H-O: the number of input layer-hidden layer-output layer).

As a result, the model 1 and 4 selected for 1, 3 and 7 d ahead were better than other models as in Fig. 6. The results of AB-RF-NN models presented values of

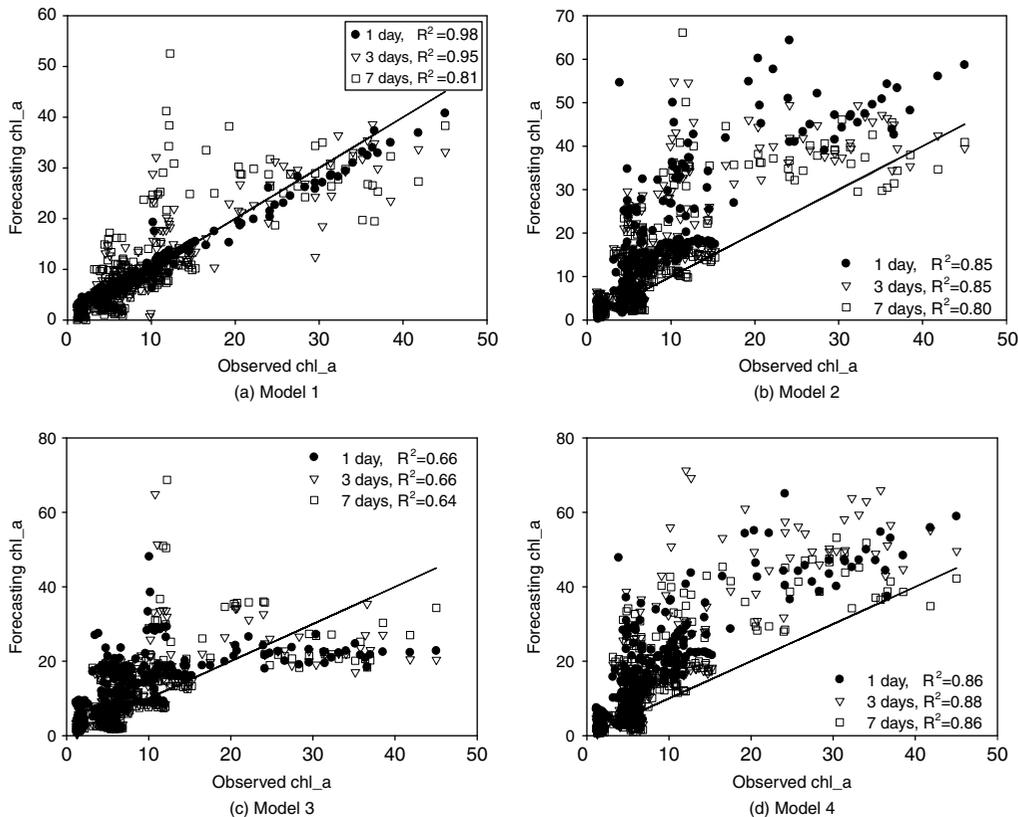


Fig. 5. Results of chl-a forecasting by AB-RF-NN models: Scatter plots between observed and forecasted data.

Table 5
Architecture of AB-RF-NN models for forecasting chl-a by training

Model	Output	Inputs	Structure
1	$t + 1$	(inflow, Tw, TOC, TP, chl-a) _{t, t-1'} (pH, DO) _t	12-20-1
	$t + 3$	(inflow, Tw, TOC, TP, chl-a) _{t, t-1'} (pH, DO) _t	12-20-1
	$t + 7$	(inflow, Tw, TOC, chl-a) _{t, t-1, t-2'} (TP) _{t, t-1'} (pH, DO) _t	16-24-1
2	$t + 1$	(inflow, Tw) _{t, t-1'} (TOC, TP) _t	6-10-1
	$t + 3$	(inflow, Tw) _{t'} (TOC, TP) _t	4-10-1
	$t + 7$	(inflow, Tw) _{t'} (TOC, TP) _t	4-10-1
3	$t + 1$	(inflow, Tw, TP) _t	3-8-1
	$t + 3$	(inflow, Tw) _{t, t-1, t-2'} (TP) _t	7-16-1
	$t + 7$	(inflow, Tw) _{t, t-1, t-2'} (TP) _t	7-20-1
4	$t + 1$	(Tw, TOC, TP) _{t, t-1}	6-8-1
	$t + 3$	(Tw, TOC) _{t, t-1'} (TP) _{t, t-1, t-2}	7-14-1
	$t + 7$	(Tw, TOC) _{t, t-1'} (TP) _{t, t-1, t-2}	7-14-1

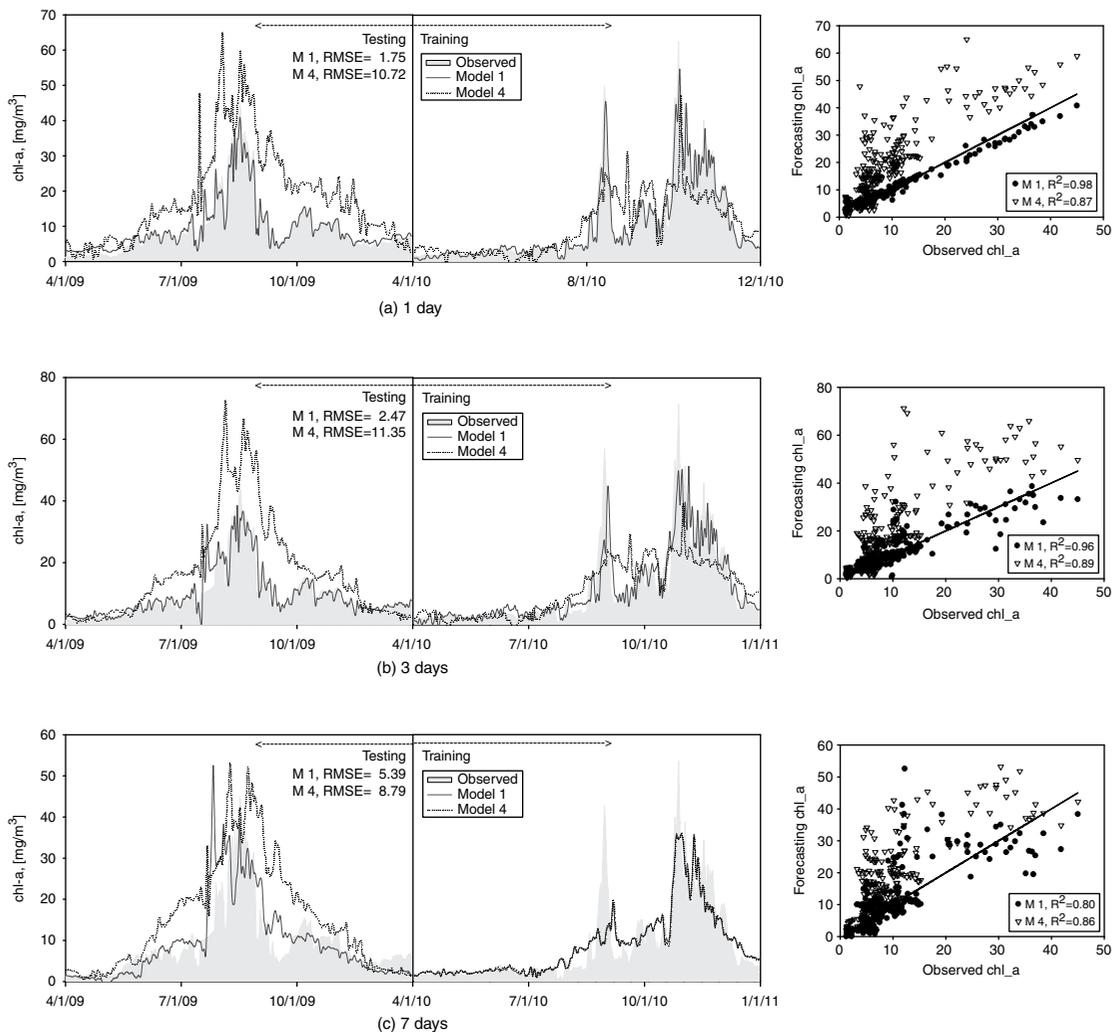


Fig. 6. Results from selected AB-RF-NN models with high performance based on R² and RMSE criteria.

training in 2010 and testing in 2009 and the scatter plot graph to the testing results by a day. From the results, 1 d and 3 d ahead had excellent correlation coefficient in the range of over 0.8 to 0.9, the case of 7 d ahead showed decent result to correlation coefficient of over 0.8.

In consequence of constructing various ANN considering hydrologic, water quality, and lag time, the constructed neural network models by case, which were combined inflow, T_w , TOC, pH, DO, TP, and chl-a with lag time, had high degree of correlation among the a variety factors, the degree of observational errors kept low until 1 d and 3 d. Furthermore, as to the forecast of 7 d, if the models are trained using long term data because there was the limitation due to the limited data set from observation system, much better results than present rate of errors will be expected. The lack of fit between the observed and estimated data indicates that new patterns must be incorporated into the model, and thus the model should be recalibrated and revalidated as more data collected. Even though the available data size is small, reasonably good results were obtained for the water quality forecast.

4. Conclusions

This study demonstrated the application of back propagation neural network to the task of modeling and forecast of algae blooms. All of the lakes or rivers had different characteristics including a variety of water quality and hydrologic factors, but the established models widely could utilize, provided that the models use the non-linear relation in characteristics of neural network. This study showed that non-linear relationships in the eutrophication phenomenon could be modeled reasonably well. For chl-a was sensitive factor about the behavior of eutrophication in reservoir, along with the chl-a concentrations, ANN models can also be used to estimate functions of environmental parameters. Especially, AB-RF-NN models can be used as algae bloom estimators though correlation analysis and training by neural network considering various water quality and hydrologic factors.

The study where longer training data might be available is required to make more precise results regarding

superiority in ecological system. In this sense, it makes water quality management easier in reservoir and river. On top of that, mechanism between water quality and hydrologic factors could be found for eutrophication. Furthermore, it could be tried to extract knowledge of characteristics of each river by applying to various rivers in South Korea with validated neural network in specific freshwater systems.

Finally, this study can be applied directly to maintain reasonable water quality in the reservoir and to prevent the water quality in future accidents.

Acknowledgements

This research was supported by the Korea Ministry of Environment as “The Eco-Innovation Project: Non-point source pollution control research group”.

References

- [1] R. Friedrich, F. Mark, H. Pia and K.I. Yabunaka, Artificial neural network approach for modeling and prediction of algal blooms, *J. Ecol. Model.*, 96 (1997) 11–28.
- [2] R. Friedrich, ANNA-Artificial neural network model for predicting species abundance and succession of blue-green algae, *J. Hydrobiologia*, 349 (1997) 47–57.
- [3] C. Karul, S. Soyupak, A.F. Cilesiz, N. Akbay and E. Germen, Case studies on the use of neural networks in eutrophication modeling, *J. Ecol. Model.*, 134 (2000) 145–152.
- [4] W. Hugh and R. Friedrich, Towards a generic artificial neural network model for dynamic predictions of algal abundance in freshwater lakes, *J. Ecol. Model.*, 146 (2001) 69–84.
- [5] E. Dogan, R. Köklü and B. Şengörür, Estimation of biological oxygen demand using artificial neural network, *International Earthquake Symposium KOCAELI*, 2007.
- [6] J.T. Kuo, M.H. Hsieh, W.S. Lung and N. She, Using artificial neural network for reservoir eutrophication prediction, *J. Ecol. Model.*, 200 (2007) 171–177.
- [7] K.S. Jeong, D.K. Kim, J.M. Jung, M.C. Kim and G.J. Joo, Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshment phytoplankton dynamics, *J. Ecol. Model.*, 211 (2008) 292–300.
- [8] S. Palani, S.Y. Liong and P. Tkalic, An ANN application for water quality forecasting, *J. Mar. Pollut. Bull.*, 56 (2008) 1586–1597.
- [9] K.P. Singh, A. Basant, A. Malik and G. Jain, Artificial neural network modeling of the river water quality—a case study, *J. Ecol. Model.*, 220 (2009) 888–895.
- [10] R.V. Thomann and J.A. Mueller, *Principles of surface water quality modeling and control*, Prentice-Hall (1987), ISBN 89-425-0063-3.