Int. J. Environ. Res., 6(1):95-108, Winter 2012
ISSN: 1735-6865

Machine Learning for Predictive Management: Short and Long term Prediction of Phytoplankton Biomass using Genetic Algorithm Based Recurrent Neural Networks

Kim, D. K.[1], Jeong, K. S.[2], McKay, R. I. B.[1], Chon, T. S.[2] and Joo, G. J.[1*]

[1] School of Computer Science and Engineering, Seoul National University, Seoul,
151-721, South Korea

[2] Department of Biological Science, Pusan National University, Busan, 609-735, South Korea

Received 11 Jan. 2011; Revised 2 July 2011; Accepted 19 July 2011

ABSTRACT: In the regulated Nakdong River, algal proliferations are annually observed in some seasons, with cyanobacteria (*Microcystis aeruginosa*) appearing in summer and diatom blooms (*Stephanodiscus hantzschii*) in winter. This study aims to develop two ecological models forecasting future chlorophyll *a* at two time-steps (one-week and one-year forecasts), using recurrent neural networks tuned by genetic algorithm (GA-RNN). A moving average (MA) method pre-processes the data for both short- and long-term forecasting to evaluate the effect of noise downscaling on model predictability and to estimate its usefulness and trend prediction for management purposes. Twenty-five physicochemical and biological components (e.g. water temperature, DO, pH, dams discharge, river flow, rainfall, zooplankton abundance, nutrient concentration, etc. from 1994 to 2006) are used as input variables to predict chlorophyll *a*. GA-RNN models show a satisfactory level of performance for both predictions. Using genetic operations in the network training enables us to avoid numerous trial-and-error model constructions. MA-smoothed data improves the predictivity of models by removing residuals in the data prediction and enhancing the trend of time-series patterns. The results demonstrate efficient development of ecological models through selecting appropriate network structures. Data pre-processing with MA helps in forecasting long-term seasonality and trend of chlorophyll *a*, an important outcome for decision makers because it provides more reaction time to establish and control management strategies.

Key words: Genetic algorithm, Nakdong River, Biomass, Management, Sensitivity analysis, Time-series prediction

INTRODUCTION

In ecological management, predictions are used to permit timely responses to likely changes. Different responses require varying timescales, so predictions are needed on a range of time scales. These issues are particularly important in environmental management, where relatively low predictability of ecosystems often combines with a paucity of data. In this paper, we consider one particular case, management of water quality in a South Korean river system, the lower Nakdong. The insights gained may be of far wider application. Effective management strategies require prediction on a range of scales, from days to decades. We consider here only two timescales: one-week-ahead prediction (sufficient for short-term management responses such as dam releases) and one-year-ahead prediction, suitable for medium-term management responses such as medium-scale construction works. We argue that these different time horizons require different modeling strategies, demonstrating this through a case study.

Strategy development in riverine water quality management requires consideration of a range of factors, including rainfall, river flow and physicochemical characteristics, together with their interaction with biological entities (Jeong *et al.*, 2006a). An ecological model of a river system is one way to understand these relationships, and so predict the future pattern of change in some property of interest. However in ecosystems modeling, it can be difficult to control model complexity and design good model components. Even though the continuing development of monitoring systems enables researchers to collect a wide range of data, the large number of input variables leads to an excess of

[*] Corresponding author E-mail: lter.ecoinfo@gmail.com

complexity, making it difficult to discover an efficient model structure (Jørgensen, 1992). Determining an appropriate structure, providing acceptable accuracy while minimizing the number of variables and parameters, is an important focus in ecological modeling and management strategy development.

The eutrophication of freshwater systems is an internationally important issue, seen worldwide, resulting from the interaction of a wide variety of factors (Wetzel, 1983). Cyanobacterial blooms are one of the world's commonest – and most costly – water pollution and eutrophication phenomena, occurring in both lotic and lentic ecosystems (Moss, 1998). Anthropogenic impacts such as dam construction and water resource supply limitations have led to water quality deterioration (Large and Petts, 1992). Algal blooms can result from the reduced water flow, and create a key water quality issue in freshwater ecosystems (Reynolds, 1992, Ha *et al.*, 1998). Many projects have researched the possible causality of phytoplankton proliferation in freshwater systems, with a substantial portion using ecological models based on machine learning (Maier *et al.*, 2001, Joo and Jeong, 2005, Mitrovic *et al.*, 2006). They generally justified this approach on the requirement for diverse input variables and the time cost for selecting the appropriate model to represent the system being studied.

Of machine learning algorithms, neural networks (NNs) are recognized as a method of choice, providing high predictability – though often at a large human time cost to determine the NN architecture (i.e., number of hidden nodes, time-delay components, variable selection, etc.), and have been widely used in phytoplankton modeling (Jeong *et al.*, 2006a). Liang (2009) proposed to overcome this with a hybrid modeling architecture, using genetic algorithms as a wrapper for neural networks, providing an alternative path to finding the optimal forecasting model (Hibon and Evgeniou, 2005). Evolving neural networks (ENN), a hybrid between these algorithms, has been successfully applied to phytoplankton modeling (Yao and Liu, 2001).

Although these authors were able to select relevant input variables and optimize model structure using ENN, this approach has limitations for developing long-term forecasting models. Pure neural networks may not be well suited to long-term forecasting (Zhang and Qi, 2005). This issue has been heavily debated, but there are still no obvious solutions, especially for seasonal forecasting models (Franses and Draisma, 1997, Nelson *et al.*, 1999, Wang *et al.*, 2006, Co and Boosarawongse, 2007, Curry, 2007, Zhang and Kline, 2007, Song and Li, 2008). One possible alternative is to pre-process the data to form moving averages (MA) from the raw dataset, shown to provide more accurate forecasts for time-series data (Wu *et al.*, 2009).

Given that algal blooms cause critically stressful conditions in aquatic ecosystems, it is important to predict their occurrence. Phytoplankton dynamics in long-term ecological data always exhibit strong seasonality combined with inter-annual variation. Various modeling methods have been applied to elucidate population and community dynamics in freshwater ecosystems. Community data are generally complex. Machine learning techniques are efficient methods to deal with complex datasets such as the long-term time-series data that arise in ecology, especially in comparison with traditional models (e.g. Chon *et al.*, 1996, Lek *et al.*, 1996, Huang and Foo, 2002, Papale and Valentini, 2003, Jeong *et al.*, 2005). Such techniques in ecological informatics as artificial neural networks (ANN) and genetic programming (GP) have proven valuable in forecasting and in revealing environment-community causal relationships (Jeong *et al.*, 2003, Jeong *et al.*, 2006b). Although these techniques have proven successful in short-term prediction, they are less successful in the prediction of long-term phytoplankton dynamics – crucial information for the developers of catchment management strategies.

Our aim, in this study, is to elucidate whether and how it might be possible to provide both short- and long-term predictions of phytoplankton dynamics, using variants of these methods. Specifically, we used preprocessing data (i.e. moving average) with recurrent neural networks, tuned by a genetic algorithm (GA-RNN) to learn and predict a long-term ecological data set to produce ecological models useful for forecasting future changes of phytoplankton.

We started from a working hypothesis, that short-term phytoplankton dynamics depends on external influences that are chaotic and unpredictable, so any attempt to learn their behavior for long-term forecasting will result in forecasts overfitted to the fluctuations. On a longer time-scale, we expect the phytoplankton dynamics to be determined by less chaotic influences. Thus we expect that the smoothed behavior of phytoplankton dynamics would be more predictable over long time ranges than the unsmoothed behavior. Thus we decided to adopt moving average (MA) methods in the data preprocessing. This gave rise to a total of four prediction problems using GA-RNN, namely short- (one week horizon) and long-term (one year horizon prediction, either of the raw values for phytoplankton concentration, or of smoothed MA values.

They were tested on the Nakdong River data. Previous research attempting to simulate the river's phytoplankton dynamics using deterministic algorithms met with limited success, because it was difficult to represent the strong biological interactions present (see Lee *et al.*, 2009). Non-linear ecological modeling could provide more accurate predictions in the short-term (e.g. 3-7 days ahead). However this is not sufficient for decision makers, who need to base some of their strategies require much longer lead times. Thus it is highly desirable to develop ecological models that can carry out (1) long-term prediction with (2) high accuracy.

MATERIALS & METHODS

The Nakdong River is the second largest river catchment in South Korea (35° ~ 37°N, 127° ~ 129° E), and is one of the major ‘regulated river’ systems of the East Asian region (Joo *et al.*, 1997). The river is approximately 525 km in length, and its basin passes through several metropolitan cities, with a total population of over 6 million, in the mid to downstream reaches. In total, over 10 million people live in the basin. The study site is approximately 27 km upstream of the river mouth, with the sampling point approximately 20 m off shore (Fig. 1).

Summer floods and winter droughts are the main sources of natural disturbance in Korean river ecosystems. The average annual rainfall is about 1,200 mm in the Nakdong River basin, with the frequency and amount of precipitation highly biased towards summer (June–mid September constitute more than 60% of total rainfall). In contrast, a severe scarcity of precipitation in winter brings about a deterioration of riverine water quality (Jeong *et al.*, 2007).

Further complicating the situation, the river system possesses reservoir-like characteristics, as multi-purpose dams and an estuarine barrage heavily regulate flow. Due to flow reductions and high demand for water resources from the large population, deterioration of water quality has accelerated with construction of artificial flow regulation structures in the river mouth during the 1970s and 80s. Cyanobacterial blooms (e.g. *Microcystis aeruginosa*) have been frequent in the lower Nakdong during summer (Ha *et al.*, 1999, Ha *et al.*, 2000). Specific diatom species (e.g. *Stephanodiscus hantzschii*) dominate every winter (Ha and Joo, 2000). The effects of the estuarine barrage are particularly important, and have been noted as a unique characteristic among regulated rivers (Murakami *et al.*, 1998, Sharp and Howe, 2000). Water samples were collected weekly at a depth of 0.5 m at the study site from 1994 to 2007. Water temperature, DO, conductivity and pH were measured by portable probes (YSI model 55 for water temperature and DO; YSI model 30 for conductivity; Orion model 250A for pH). Secchi depth was recorded using a 20 cm-diameter black and white disc. Turbidity was measured using a tubidimeter (HF Scientific model 11052) in the laboratory. Chlorophyll *a* concentration was analyzed by means of spectrophotometry using the filtered water samples (Wetzel and Likens, 1991). Total nitrogen, total

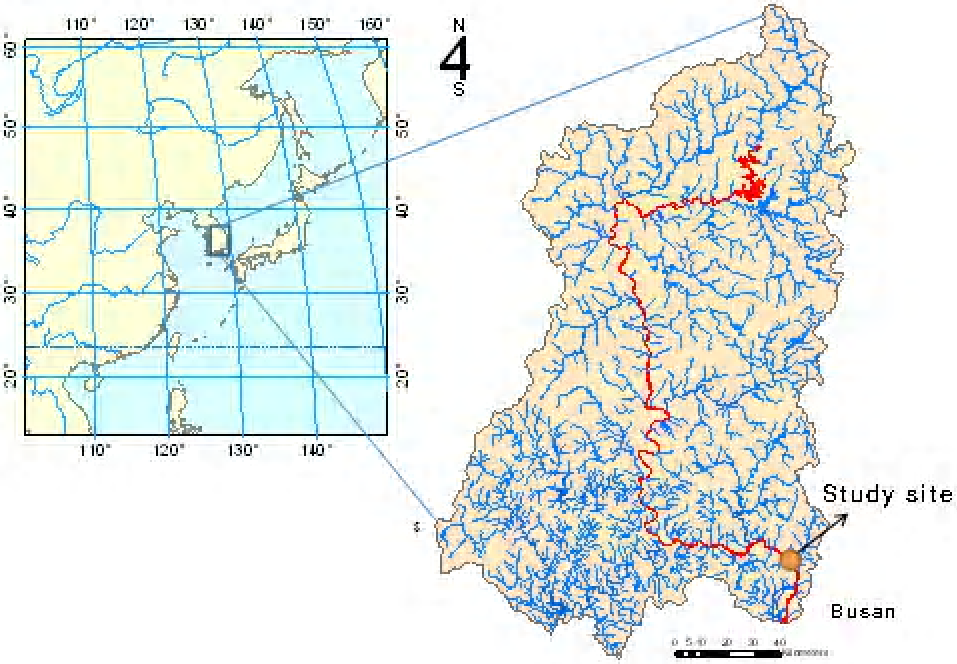


Fig. 1. Study site location (Mulgeum; 27 km from the river mouth)

phosphorus, nitrate, phosphate and silica were determined using a QuikChem Automated Ion Analyzer. Other environmental data such as precipitation, dam discharge and river flow were obtained from the Korean Meteorological Administration and the Nakdong River Flood Control Center. Summed precipitation data from five representative sites adjacent to the major dams (Andong, Daegu, Hapchon, Jinju and Miryang) were used. Dam discharge data were collected from four multi-purpose dams, Andong (AD), Imha (IH), Hapchon (HC) and Namkang (NK), while river flow data were obtained from Samrangjin, the measurement station nearest the site.

Zooplankton grazing can affect phytoplankton community dynamics in river ecosystems (Gosselain *et al.*, 1998, Kim *et al.*, 2000). For chlorophyll *a* prediction, zooplankton density was used as input data. Zooplanktons were divided into four categories: rotifer, cladoceran, nauplius and copepod (four variables). Zooplankton sampling was performed by filtering 8 liters of water with a 35-µm net, then adding 10% formalin (4% final concentration) to the specimens for preservation. Copepods and cladocerans were counted using an inverted microscope at × 25-50 magnification, while rotifers were counted at × 100-400 magnification. Zooplankton identification was determined to genus or species level using the criteria of Koste (1978), Smirnov and Timms (1983) and Einsle (1993).

In order to analyze the complex plankton data, three approaches were considered in this study. Considering that the data are time-series with marked fluctuations, an additional test was carried out using moving averages to smooth the data and reveal the trends in chlorophyll *a* dynamics. Moving averages (MA) are widely used in time-series analyses such as stock market analyses (Kimoto *et al.*, 1990, Gençay, 1996). There are few studies using MA for water quality or phytoplankton dynamics, except for auto-regressive moving average (ARMA) models, which are statistically based forecasting models (Harding and Perry, 1997, Yoo, 2002). However, these approaches are only one of many ways to predict the point values in time-series, and place limits on interpreting the interrelationship between input and output using *post hoc* analyses such as sensitivity analysis. From a management perspective, the most important thing in long-term prediction is to get an overall idea of the likely behavior in a particular season; point predictions of the phytoplankton dynamics on a particular day a year ahead are of limited value. Thus we use MA methods to identify the trend of phytoplankton community. Specifically, we used twelve-week-average ddata, aiming to reveal seasonality in chlorophyll *a* dynamics.

However MA just smoothes the data: it does not (of itself) yield predictions. In addition, we used artificial neural networks to provide the predictions of these complex plankton dynamics. Several previous studies used neural network methods such as multi-layer perceptrons to forecast plankton dynamics and ecological community patterns (Maier and Dandy, 2000, Huang and Foo, 2002, Oh *et al.*, 2007). Since we aim to use time-series data to reveal the chlorophyll dynamics, we decided to use temporal networks for prediction of the data (Chon *et al.*, 2000, Jeong *et al.*, 2001, Walter *et al.*, 2001). In this paper, a fully Recurrent Neural Network (RNN) was used for model construction (Principe *et al.*, 2000).

$$\mathrm{net}_i(n+1) = \textstyle\sum_{j<i} w_{ij} y_j(n+1) + \sum_{j\geq i} w_{ij} y_j(n) + I_i(n+1) \qquad (1)$$

$$y_i(n+1) = f(\mathrm{net}_i(n+1)) \qquad (2)$$

where *w* is weight, *I(n)* is external input to the network neurons, *y(n)* is the output. *f(x)* is the transfer function (tangent hyperbolic function) (Fig. 2). The network was trained using a Back Propagation (BP) algorithm as described by Rumelhart et al. (1986).

Most research using RNN has focused on short-term prediction (Jeong *et al.*, 2001, Walter *et al.*, 2001, Ho *et al.*, 2002), rather than long-term population dynamics. Here, we aimed to predict plankton dynamics on both short- and long-term bases. The long-term prediction will be especially useful for long-term river management planning. On the other hand, practical application gives rise to an important problem. Determining the number of nodes, and the weights in an NN requires many trial experiments, and is thus very time-consuming. In this aspect, Genetic Algorithms (GA) can effectively enhance the modeling power of ANN through their global search capability.

Genetic Algorithm based Recurrent Neural Networks (GA-RNN) were used for modeling chlorophyll *a* dynamics. We used the software package, NeuroSolutions 5.0 (Lefebvre *et al.*, 2005). The scheme for GA-RNN training is shown pictorially in Fig.s 2 and 3. Recurrent Neural Networks (RNN) of a given architecture were trained by Back-Propagation (BP), the training results being used as information in the next step, determining an appropriate network structure. Thus in addition to the BP algorithm, a GA was also used, its purpose being to optimize the network structure - to determine a good configuration

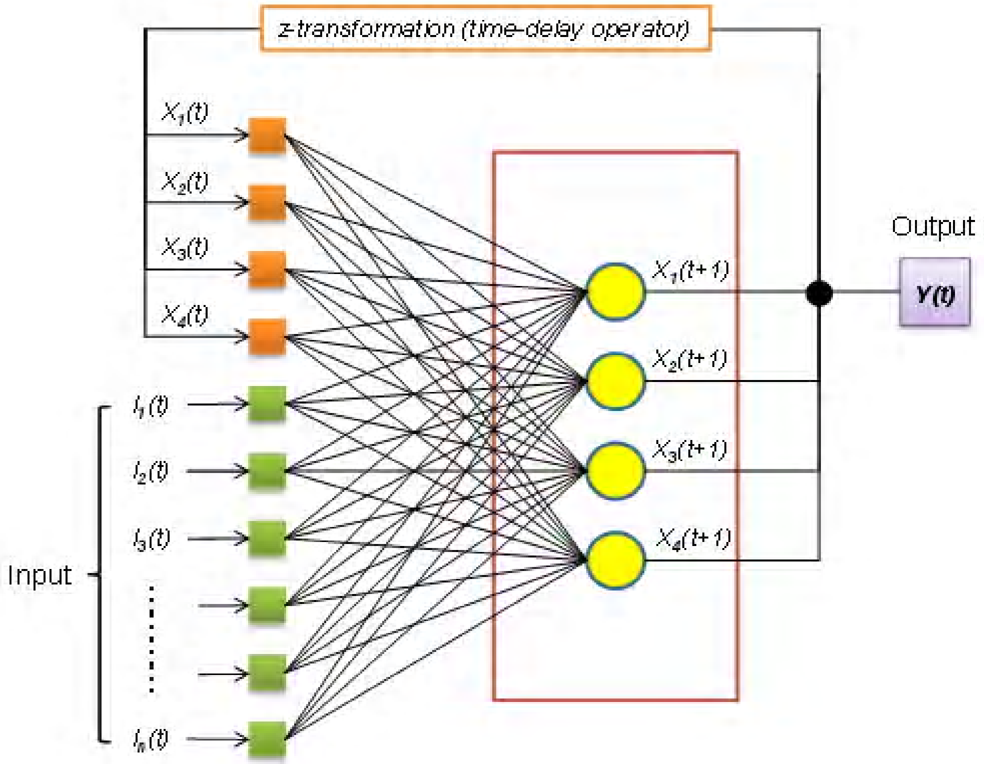


Fig. 2. The basic structure of the fully recurrent neural network

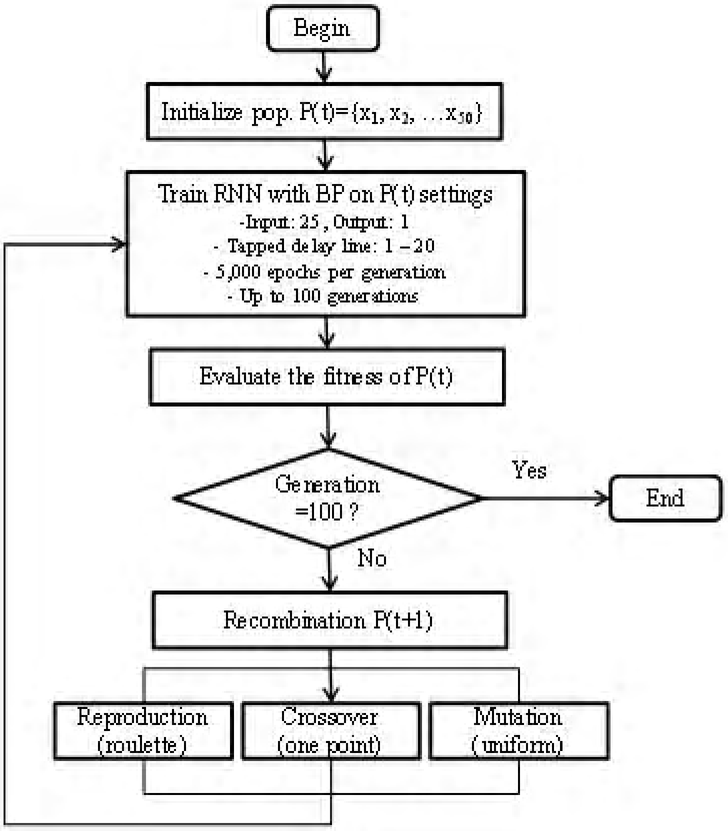

Begin
Initialize pop. P(t)={x_1, x_2, …x_{50}}
Train RNN with BP on P(t) settings
-Input: 25 , Output: 1
- Tapped delay line: 1 – 20
- 5,000 epochs per generation
- Up to 100 generations
Evaluate the fitness of P(t)
Generation =100 ?
Yes
End
No
Recombination P(t+1)
Reproduction (roulette)
Crossover (one point)
Mutation (uniform)

Fig. 3. The computational process of the evolutionary recurrent neural networks

for the components (Blanco *et al.*, 2000). A GA can be used to evolve RNNs by searching for the RNN structure (number of nodes in the hidden layer, momentum values between each layer, learning rate, number of time-steps) that lead to the best learning, in the sense of best generalization to new data. In this experiment, the number of hidden layers was fixed at one, and the taps (i.e. number of successive inputs) were varied from one to twenty weeks through the GA. The GA also evolved the size of the hidden layer, within the range one to five. The crossover and mutation rates in the GA were fixed at 90% and 1% respectively as a default. Other network parameters were also set at suitable defaults (i.e. input layer; time-delay neural network, learning momentum; 0.7, maximum epochs (for NN) and generations (for GA); 5000 and 100 and termination criterion: after 500 epochs without improvement).

Both short-term and long-term data were used for forecasting in this study. The weekly data from June 1994 to December 2006 (n=657) were used to develop non-Moving Average (MA) predictive models for short- and long-term. For model training, data from June 1994 to 2003 was used, with cross-validation from 2004-2005 data. 2006 data was used for testing. Data from June 1994 to June 2006 (n=628) was separately used for long-term predictive model development. Training used data from June 1994 to 2001, cross-validation exploited data from 2002 to 2004 and testing used data from 2005 to June 2006. With respect to long term forecasting, prediction results using data from June 1995 through June 2007 were presented. For those reasons, data numbers differed between short- and long-term sets.

We conducted a sensitivity analysis (SA) on all variables used in the models. The NeuroSolutions 5.0 package we used directly supports SA on each variable (Lefebvre *et al.*, 2005). In SA, one initially trains the network as normal, then fixes the weights. The SA step consists of perturbing, one at a time, each channel of the input vector, starting from Mean – Standard Deviation (S.D.) and ranging through to Mean + S.D., then measuring the change in corresponding output. The sensitivity, S_k for input k may be expressed as

$$S_k = \frac{\sum_{p=1}^{P} (y_p - \bar{y}_p)^2}{\sigma_k^2} \qquad (3)$$

where y_p is the original measured data, ¢$_p$ is the output obtained with the fixed weights for the *p*th pattern, P is the number of patterns and σ^2_k is the variance of the input variables (Principe *et al.*, 2000, Lefebvre *et al.*, 2005).

RESULTS & DISCUSSION

Four GA-RNN models were developed to forecast chlorophyll *a* concentration in the lower Nakdong River (short- and long-term, MA and non-MA). The greatest predictive performance in each case was selected from the GA-RNN of five (non-MA) and six taps (MA) for short-term forecasting, and of nine (non-MA) and two taps (MA) for long-term forecasting. These models captured peak concentrations of chlorophyll *a* reasonably faithfully, although the magnitude of chlorophyll *a* concentration was somewhat underestimated in some years.

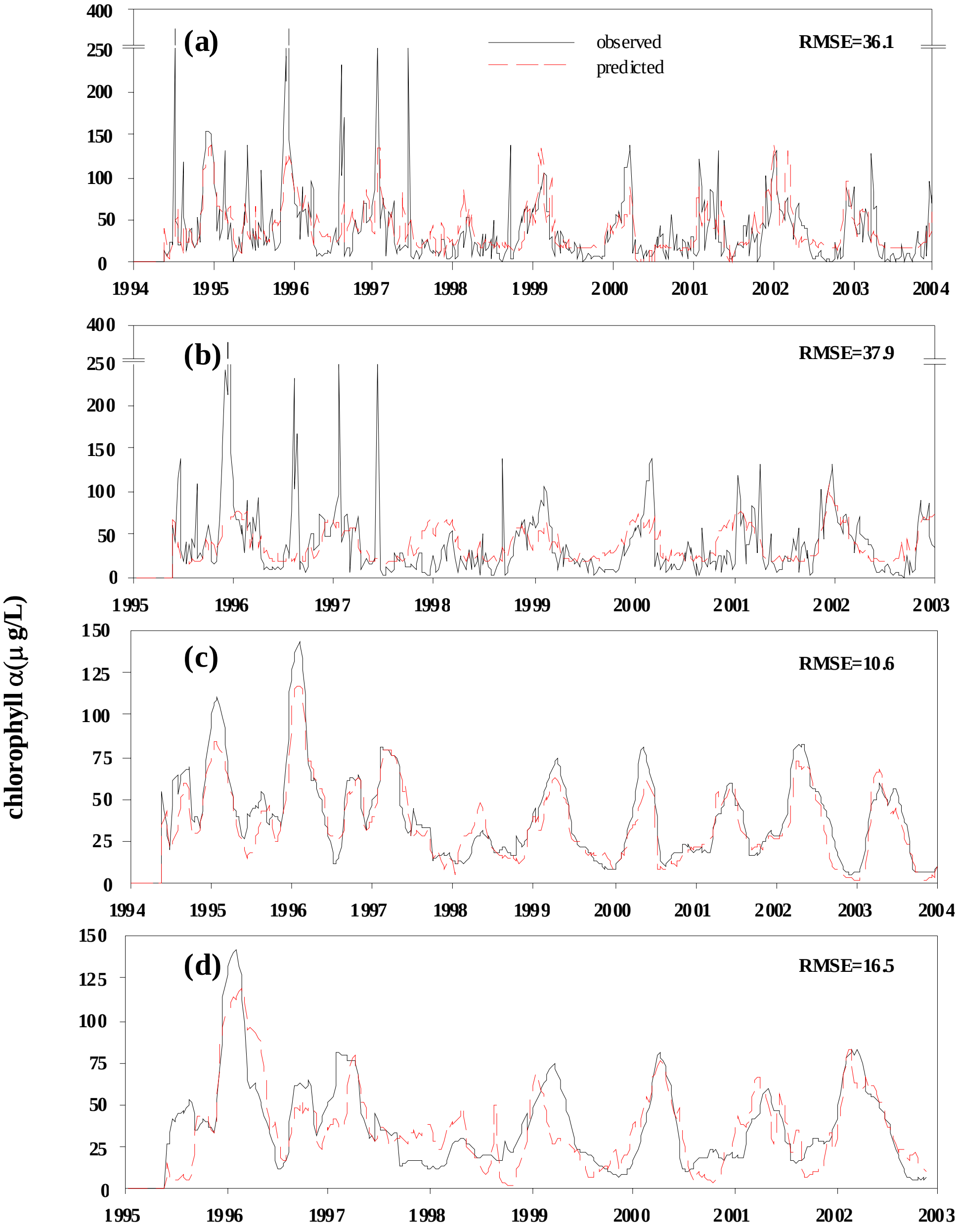
The results of these experiments are presented in detail in Table 1. For models using as-sampled (i.e. non-MA) data, the GA-RNN predicted the timing effectively, although some peaks of chlorophyll *a* concentration were missed. Root Mean Squared Error (RMSE) for training were 36.1 (r^2=0.34), 37.9 (0.20), 10.6 (0.86) and 16.5 (0.62) (Fig. 4), for cross-validation were 30.5 (0.44), 27.2 (0.52), 16.9 (0.76) and 8.5 (0.89) (Fig. 5), and for testing were 20.5 (0.59), 30.5 (0.60), 9.4 (0.88) and 13.7 (0.80) (Fig. 6), for one-week non-MA, one-year non-MA, one-week MA and one-year MA, respectively. Overall, the MA preprocessed models displayed greater short- and long-term predictability in terms both of timing and of magnitude.

The results demonstrate the feasibility of using machine learning to generate ecological models for the long-term forecasting of water quality dynamics, especially for management purposes. Data preprocessing made an important contribution to the accuracy of prediction. As is well known, field data in ecological science are often coarse and complex (Fielding, 1999), and data for phytoplankton dynamics (i.e., chlorophyll *a* dynamics) is highly variable, as illustrated in the weekly changes in chlorophyll *a* (Fig.s 4 to 6 (a), (b)). Liang (2009) provided a brief literature review summarizing the advantages of statistical methods of adaptation in data preprocessing, prior to neural network modeling. MA approaches have occasionally been used in statistical time-series modeling, mainly to reduce the effects of short-term perturbations and noise, especially when the original data was sampled at irregular and/or coarse scales (Lillo *et al.*, 2000, Schumann and Lauener, 2005, Wu *et al.*, 2009). The high accuracy in long-term prediction obtained by the GA-RNN model is consistent with this previous research.

Although MA is capable of dramatically improving model predictability, its use must be adapted appropriately to the objectives of the model prediction, and to the level of fluctuation in the output variables. Otherwise, using a MA may risk introducing time lags in the prediction of parameter changes. From a

Table 1. Performance results of the GA-RNN models

RMSE (r^2)	Short-term (one week prediction)			Long-term (one year prediction)		
	Training	Cross-Validation	Test	Training	Cross-Validation	Test
Non-MA	36.1 (0.34)	30.5 (0.44)	20.5 (0.59)	37.9 (0.20)	27.2 (0.52)	30.5 (0.60)
MA	10.6 (0.86)	16.9 (0.76)	9.4 (0.88)	16.5 (0.62)	8.5 (0.89)	13.7 (0.80)



Training

Fig. 4. Prediction results of model's training in (a) one-week non-MA, (b) one-year non-MA, (c) one-week MA, (d) one-year MA

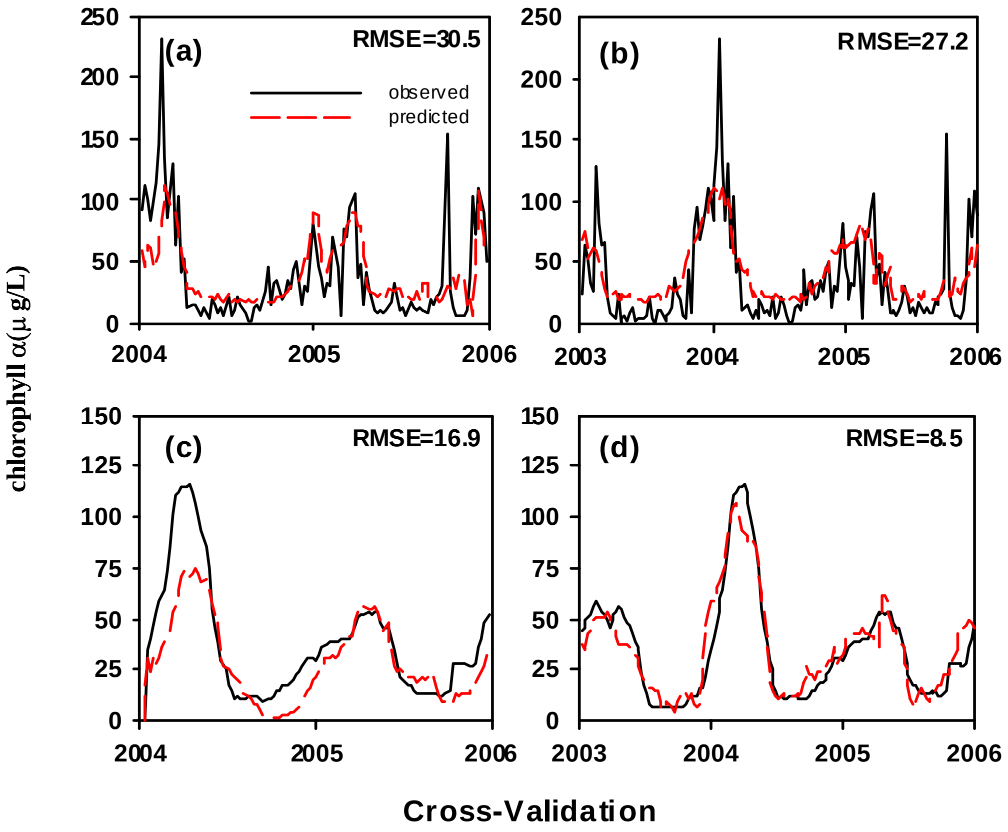


Fig. 5. Prediction results of model's cross-validation in (a) one-week non-MA, (b) one-year non-MA, (c) one-week MA, (d) one-year MA

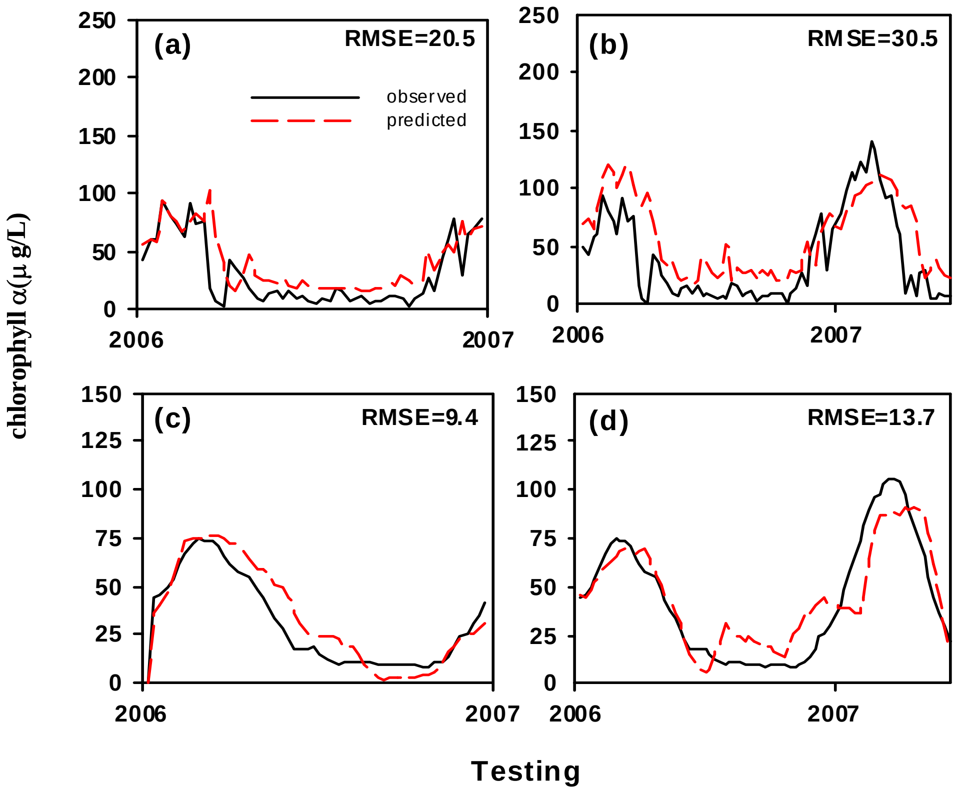


Fig. 6. Prediction results of model's testing in (a) one-week non-MA, (b) one-year non-MA, (c) one-week MA, (d) one-year MA

management perspective, short-term MA predictions may be of questionable value: knowing a 13-week moving average accurately one week ahead may have little impact on decision making. However in long-term predictions, the value may be much greater. Accurate predictions of the overall seasonal behavior one year in advance will allow important long-term decisions to be made with a much greater level of assurance. Future work in this area will study different time lags in the MA process, to determine the lags maximizing predictability and revealing periodicity in the system; we also aim to investigate in more detail the management utility of different scales of prediction.

Twenty-five variables were used as inputs for the GA-RNN. Sensitivity analysis was performed after developing the most effective model. A genetic algorithm was used to select suitable parameters and determine a feasible architecture, as explained earlier. Sensitivity values are shown in Fig. 7. Although the general trends were similar, differences were observed in the importance of variables between the one-week and one-year predictions, and between non-MA and MA. The number of influential variables was slightly larger in the non-MA models when used for long-term predictions. Overall, pH was the most influential variable. In the non-MA models, other variables became more influential, especially DO and nitrate.

Across the four GA-RNN models, chlorophyll *a* dynamics was heavily influenced by variations in water temperature, and in DO, pH and nitrogen concentration. The degree of influence varied across the different models. While pH was most influential in MA models, other factors, particularly DO and nitrate, were more influential in non-MA. In the one-week forecasting, chlorophyll *a* was most sensitive to DO (2.90) followed by pH (1.85) and nitrate (1.80) for non-MA models, and to pH (2.16) followed by total nitrogen (1.10) and water temperature (0.74) in MA. For one-year forecasting, the most sensitive parameters were nitrate (3.82) in non-MA and pH (2.84) in MA results.

Overall, nitrogen and pH were most influential in determining the variation of chlorophyll *a*. In models using the original sampled data without MA, however, the sensitivity to DO was also high. It was also notable that in the long-term MA model, substantially more variables were found to be influential (e.g. dam discharges, other nutrients, etc).

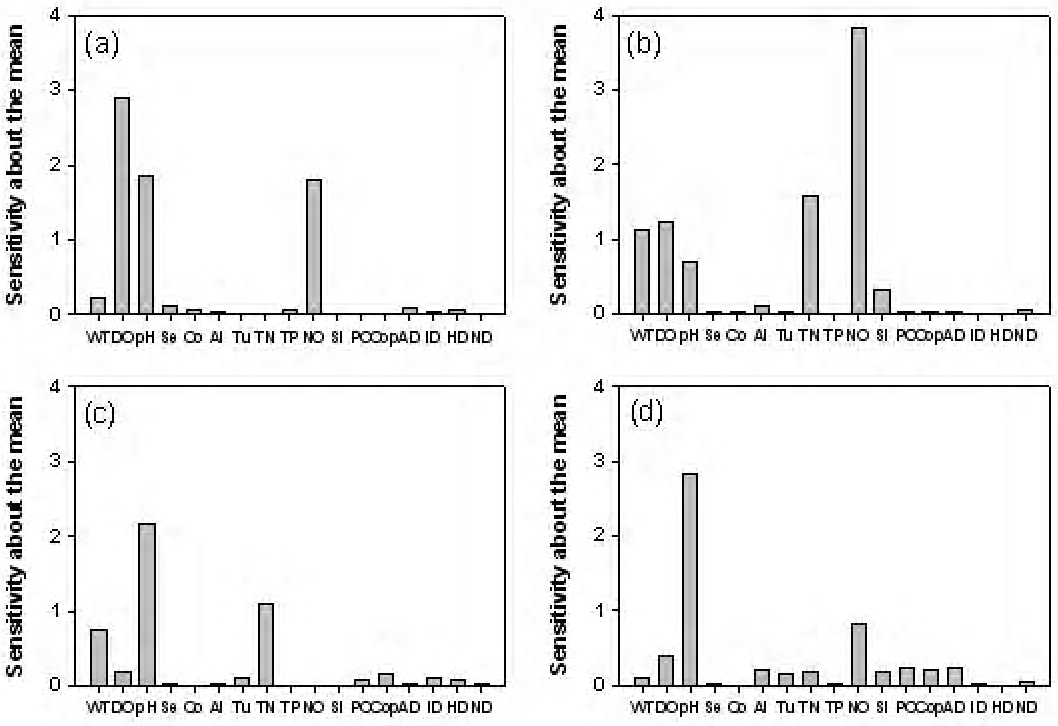


Fig. 7. Sensitivity comparison of variables in determining chlorophyll *a* dynamics in (a) one-week non-MA, (b) one-year non-MA, c) one-week MA, d) one-year MA. (Water Temperature; WT, Dissolved Oxygen; DO, Secchi depth; Se, Conductivity; Co, Alkalinity; Al, Turbidity; Tu, Total Nitrogen; TN, Total Phosphorus; TP, Nitrate; NO, Silica; Si, Phosphate; PO, Copepods; Cop, Each dam's discharge; AD, ID, HD, ND)

Time-delayed recurrent neural networks differ from multi-layer perceptrons (MLP) in using a range of consecutive input values. The delay in input data through the 'memory' effect, between the input layer and the internal inputs provided by the recurrent network structure, enables the model to explore the nature of phytoplankton biomass time-series. Recurrent neural networks (RNN) may be more suitable than MLP for the prediction of time-series data like chlorophyll *a* dynamics (Jeong *et al.*, 2008).

Finding the optimal network architecture is always difficult, and may require a great deal of redundant iterative trials and evaluations (Yao, 1999). The complexity is even greater for recurrent NNs than for MLPs, because of the increased degrees of freedom. While artificial neural network models have been used over a wide range of ecological research (Joy and Death, 2004, Recknagel *et al.*, 2006b, Goethals *et al.*, 2007), only a few studies have used an evolutionary NN for time-series prediction in phytoplankton biomass (Yao and Liu, 2001). To avoid time-consuming effort to find optimal architecture, we used a genetic algorithm (GA) to optimize the recurrent neural networks. GAs are highly capable at global search, being able to find a near optimal NN architecture without human intervention (Yao and Liu, 2001). The GA framework also supports checking the importance of an input variable in determining the output data. Thus RNNs, combined with an evolutionary algorithm, the so-called evolutionary recurrent neural network (ERNN), can relieve model designers of the onerous task of iterative trial-and-error training (e.g. adjusting and varying the number of hidden nodes and taps).

This study demonstrated that techniques from ecological informatics could be effective for both short- and long-term prediction. The evolutionary recurrent neural network (ERNN) models exhibited higher predictivity on MA-preprocessed data, but this predictivity must be balanced against utility. Although chlorophyll *a* concentration was better predicted over a short-term (a week), the long-term (a year) prediction also yielded an acceptable range (Fig.s 4 to 6); conversely, the utility of long-term MA predictions may be much greater than short-term. Neural networks have previously been used successfully in various short-term prediction problems in water quality modeling (Walter *et al.*, 2001, Whigham and Recknagel, 2001, Recknagel *et al.*, 2006a, Velo-Suárez and Gutiérrez-Estrada, 2007) using non-MA data. However our study showed the value of MA preprocessing in improving the predictability of chlorophyll *a* for both short- and long-term predictions over the non-MA case (Figs. 4 to 6).

The predictability of GA-RNN was greater for models based on MA-preprocessed data than for the original data. MA preprocessing is able to smoothe fluctuations and reduce noise arising in field sampling and measurement. Although finding specific values is important in the short term, trend detection can also be important for algal dynamics prediction, especially in long-term forecasting. In this case, the non-MA long-term model was too inaccurate for management application, while the MA long-term model yielded usable management predictions. This is important because long-term prediction of general trends, as in one-year forecasting, is generally more valuable for management decision-making than short-term prediction, because of the time lags involved in major management decisions.

In the sensitivity analysis, we saw some important differences between non-MA and MA, and between short- and long-term models (Fig. 7). Nitrogen components were highly sensitive parameters for chlorophyll *a* concentration in non-MA models, with pH playing the same role in MA models. Nitrogen plays a key role in algal population growth, and particularly from the sensitivity analysis, it appears that chlorophyll *a* was more sensitive to nitrogen than to phosphorus concentration in the lower Nakdong River. However nitrogen is subject to short-term fluctuations; when these fluctuations are smoothed, its long term relative influence is reduced. We think that the importance of pH variation could be related to the dissociation kinetics of dissolved carbonates, which are obviously of major importance to photosynthetic organisms depending on the availability of inorganic carbon for photosynthesis. In the lower Nakdong River, due to extreme diatom proliferation during the winter (Ha *et al.*, 2003), pH sometimes rises above 10.0 – extreme in comparison with most major rivers.

It is natural to expect dissolved oxygen to figure prominently among parameters for models of chlorophyll *a*, since it may be formed as a product of algal photosynthesis. It was found to be a high sensitivity variable, particularly in the non-MA short-term model. DO is rapidly altered by the fluctuation of phytoplankton biomass (Lampert and Sommer, 2007). Thus in the non-MA short term model, we may be seeing some degree of confusion of cause and effect. By contrast, in the MA long-term model, the sensitivities to other parameters such as dam discharge, zooplankton abundance, alkalinity, turbidity and other nutrients increased. Some degrees of sensitivity of chlorophyll *a* to those parameters were observed, although it was not as high as to the preceding variables. It leads to the conjecture that algal dynamics over the long-term may be linked to a greater diversity of environmental components. For instance, the sensitivity to AD dam discharge increased substantially in the MA long-term model, in comparison with the other models. In the regulated Nakdong River basin, dam operation can be a significant factor in determining hydrological impacts upon algal dynamics lower down the river. Reinforcing this, Jeong et al. (2007) described the results of cross-correlation analysis, finding that dam discharge has a significant relationship to algal species abundance over a 24-month period. The study reported by Kim et al. (2007) also showed the AD dam discharge as one of the most influential hydrological factors for winter diatom blooms. This is particularly important from a management perspective, because dam operation is the model component most directly available for management control.

In summary, we found that environmentally related parameters could be influential in determining algal population dynamics, based on four different models derived from Genetic Algorithm based Recurrent Neural Networks (GA-RNN). In particular, we were able to develop a fairly reliable long-term predictive model using MA data preprocessing. Such a model can be much more practically useful for water resources management than a short-term model - though in combination, they may be more useful still in determining appropriate management of algal blooms. The benefit of generating four different models (short- and long-term, MA and non-MA data) was also demonstrated through the enhanced interpretability of the combined results.

CONCLUSION

In this paper, short- and long-term forecasting time-series models of algal dynamics were developed, using both the raw data and moving average preprocessing. The models were extracted from the data using evolutionary recurrent neural networks. Moving average data generated substantially more reliable and useful predictive models for long-term prediction. For short-term prediction, moving average data also generated more accurate predictions, but the trade-off here may be substantially reduced utility. From a management perspective, the most useful combination may be short-term non-MA models for short-term decision making, combined with long-term MA models for long-term decision making. Overall, the research confirmed the value of recurrent neural networks for generating predictive time-series models, and the value of MA preprocessing for improving the accuracy of fit.

ACKNOWLEDGEMENT

We are grateful to anonymous reviewers in providing valuable comments on the manuscript. This research was accomplished under the D. K. Kim's Postdoc Programme (355-2008-1-C00047) supported by Korea Research Foundation, and also conducted under the Long Term Ecological Research Program of the Ministry of Environment, Republic of Korea.

REFERENCES

Blanco, A., Delgado, M. and Pegalajar, M. C. (2000). A genetic algorithm to obtain the optimal recurrent neural network. International Journal of Approximate Reasoning, **23,** 67-83.

Chon, T. S., Park, Y. S. and Cha, E. Y. (2000). Patterning of community changes in bentic macroinvertebrates collected from urbanized streams for the short term prediction by temporal artificial neuronal networks. (In S. Lek and J.F. Guegan (Eds.), Artificial Neuronal Networks: Application to ecology and evolution. (pp. 99-114). Berlin: Springer.)

Chon, T. S., Park, Y. S., Moon, K. H. and Cha, E. Y. (1996). Patterning communities by using an artificial neural network. Ecological Modelling, **90,** 69-78.

Co, H. C. and Boosarawongse, R. (2007). Forecasting Thailand's rice export: Statistical techniques vs. artificial neural networks. Computers and Industrial Engineering, **53 (4),** 610-627.

Curry, B. (2007). Neural networks and seasonality: Some technical considerations. European Journal of Operational Research, **179 (1),** 267-274.

Einsle, U. (1993). Crustacea, Copepoda, Calanoida und Cyclopoida. Süßwasserfauna von Mitteleuropa. (Stuttgart: Gustav Fisher Verlag).

Fielding, A. (1999). An introduction to machine learning methods. (In A. Fielding (Ed.), Machine Learning Methods for Ecological Applications. (pp. 1-35). Massachusetts: Kluwer Academic Publishers.)

Franses, P. H. and Draisma, G. (1997). Recognizing changing seasonal patterns using artificial neural networks. Journal of Econometrics, **81 (1),** 273-280.

Gençay, R. (1996). Non-linear prediction of security returns with moving average rules. Journal of Forecasting, **15 (3),** 165-174.

Goethals, P. L. M., Dedecker, A. P., Gabriels, W., Lek, S. and Pauw, N. D. (2007). Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology, **41 (3),** 491-508.

Gosselain, V., Descy, J.-P., Viroux, L., Joaquim-Justo, C., Hammer, A., Métens, A. and Schweitzer, S. (1998). Grazing by large river zooplankton: a key to summer potamoplankton decline? The case of the Meuse and Moselle rivers in 1994 and 1995. Hydrobiologia, **369/370,** 199-216.

Ha, K., Cho, E.-A., Kim, H. -W. and Joo, G. -J. (1999). *Microcystis* bloom formation in the lower Nakdong River, South Korea: importance of hydrodynamics and nutrient loading. Marine and Freshwater Research*,* **50,** 89-94.

Ha, K., Jang, M. -H. and Joo, G. -J. (2003). Winter *Stephanodiscus* bloom development in the Nakdong River regulated by an estuary dam and tributaries. Hydrobiologia*,* **506/509,** 221-227.

Ha, K. and Joo, G. -J. (2000). Role of silica in phytoplankton succession: an enclosure experiment in the downstream Nakdong River (Mulgum), Korean Journal of Ecology*,* **23 (4),** 299-307.

Ha, K., Kim, H.-W., Jeong, K.-S. and Joo, G.-J. (2000). Vertical distribution of *Microcystis* population in the regulated Nakdong River, Korea, Limnology*,* 1**,** 225-230.

Ha, K., Kim, H. W. and Joo, G. J. (1998). The phytoplankton succession in the lower part of hypertrophic Nakdong River (Mulgum), South Korea. Hydrobiologia*,* **370,** 217-227.

Harding, L. W. and Perry, E. S. (1997). Long-term increase of phytoplankton biomass in Chesapeake Bay, 1950-1994. Marine Ecology Progress Series*,* **157,** 39-52.

Hibon, M. and Evgeniou, T. A. (2005). Simple procedure for reliability of repairable systems. International Journal of Forecasting*,* **21,** 15-24.

Ho, S. L., Xie, M. and Goh, T. N. (2002). A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. Computers and Industrial Engineering, **42 (2-4),** 371-375.

Huang, W. and Foo, S. (2002). Neural network modeling of salinity variation in Apalachicola River. Water Research, **36,** 356-362.

Jørgensen, S. E. (1992). Integration of Ecosystem Theories: A Pattern. (Dordrecht: Kluwer)

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

Jeong, K. -S., Kim, D. -K., Chon, T. -S. and Joo, G. -J. (2005). Machine learning application to the Korean freshwater ecosystems. Korean Journal of Ecology*,* **28 (6),** 405-415.

Jeong, K. -S., Kim, D. -K. and Joo, G. -J. (2006a). River phytoplankton prediction model by Artificial Neural Network: Model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system. Ecological Informatics*,* **1 (3),** 235-245.

Jeong, K. -S., Kim, D. -K. and Joo, G. -J. (2007). Delayed influence of dam storage and discharge on the determination of seasonal proliferations of *Microcystis aeruginosa* and *Stephanodiscus hantzschii* in a regulated river system of the lower Nakdong River (South Korea). Water Research*,* **41 (6),** 1269-1279.

Jeong, K. -S., Kim, D. -K., Jung, J. -M., Kim, M. -C. and Joo, G. -J. (2008). Non-linear autoregressive modelling by Temporal Recurrent Neural Networks for the prediction of freshwater phytoplankton dynamics. Ecological Modelling, **211(3-4),** 292-300.

Jeong, K. -S., Kim, D. -K., Whigham, P. and Joo, G. -J. (2003). Modelling *Microcystis aeruginosa* bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach, Ecological Modelling, **161,** 67-78.

Jeong, K. -S., Recknagel, F. and Joo, G. -J. (2006b). Prediction and Elucidation of Population Dynamics of a Blue-green Algae (*Microcystis aeruginosa*) and Diatom (*Stephanodiscus hantzschii*) in the Nakdong River-Reservoir System (South Korea) by Artificial Neural Networks. (In F. Recknagel (Ed.), Ecological Informatics: Scope, Techniques and Applications. (pp. 255-273). Berlin: Springer.)

Joo, G. -J. and Jeong, K. -S. (2005). Modelling community changes of cyanobacteria in a flow regulated river (the lower Nakdong River, S. Korea) by means of a Self-Organizing Map (SOM). (In S. Lek and M. Scardi and P.F.M. Verdonschot and J.-P. Descy and Y.-S. Park (Eds.), Modelling Community Structure in Freshwater Ecosystems. (pp. 273-287). Berlin: Springer.)

Joo, G. -J., Kim, H. -W., Ha, K. and Kim, J. -K. (1997). Long-term trend of the eutrophication of the lower Nakdong River, Korean Journal of Limnology*,* **30 (supplement),** 472-480.

Joy, M. K. and Death, R. G. (2004). Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neual networks, Freshwater Biology, **49,** 1036-1052.

Kim, D. -K., Jeong, K. -S., Whigham, P. A. and Joo, G. -J. (2007). Winter diatom blooms in a regulated river in South Korea: explanations based on evolutionary computation. Freshwater Biology*,* **52 (10),** 2021-2041.

Kim, H. -W., Hwang, S. -J. and Joo, G. -J. (2000). Zooplankton grazing on bacteria and phytoplankton in a regulated large river (Nakdong River, Korea). Journal of Plankton Research*,* **22,** 1559-1577.

Kimoto, T., Asakawa, K., Yoda, M. and Takeoka, M. (1990). Stock market prediction system with modular neural networks. (Paper presneted at International Joint Conference on Neural Networks, IEEE, San Diego)

Koste, W. (1978). Rotatoria. Die Raderiere Mitteleuropas. Ein Bestimmungswerk begrundet von Max Voift. (Stuttgart: Borntrager)

Lampert, W. and Sommer, U. (2007). Limnoecology. (New York: Oxford University Press Inc.)

Large, A. R. and Petts, G. E. (1992). Rehabilitation of river margins. (Oxford: Blackwell Scientific Publication).

Lee, S., Choi, J. -M. and Jeong, K. -S. (2009). Water quality simulation at Mulgeum station (Nakdong River) using zooplankton community data. Journal of Korean Society of Water Quality*,* **25 (6),** 832-839.

Lefebvre, C., Fancourt, C., Principe, J., Gerstenberger, J., Samson, D., Euliano, N., Wooten, D., Lynn, G., Geniesse, G., Allen, M., Lucas, M. and Marossero, D. (2005). NeuroSolution 5.0. NeuroDimension, Inc.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. and Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling, **90 (1),** 39-52.

Liang, Y. H. (2009). Combining seasonal time series ARIMA method and neural networks with genetic algorithms for predicting the production value of the mechanical industry in Taiwan. Neural Computing and Applications, **18 (7),** 833-841.

Lillo, M. P. Y., Perez-Correa, R., Latrille, E., Fernandez, M., Acuna, G. and Agosin, E. (2000). Data processing for solid substrate cultivation bioreactors. Bioprocess Engineering, **22 (4),** 291-297.

Maier, H. R. and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software, **15,** 101-124.

Maier, H. R., Sayed, T. and Lence, B. J. (2001). Forecasting cyanobacterium *Anabaena* spp. in the River Murray, South Australia, using B-spline neurofuzzy models. Ecological Modelling, **146,** 85-96.

Millie, D. F., Weckman, G. R., Paerl, H. W., Pinckney, J. L., Bendis, B. J., Pigg, R. J. and Fahnenstiel, G. L. (2006). Neural net modeling of estuarine indicators: Hindcasting phytoplankton biomass and net ecosystem production in the Neuse (North Carolina) and Trout (Florida) Rivers, USA. Ecological Indicators, **6,** 589-608.

Mitrovic, S. M., Chessman, B. C., Bowling, L. C. and Richard H. Cooke (2006). Modelling suppression of cyanobacterial blooms by flow management in a lowland river. River Research and Applications, **22 (1),** 109-114.

Moss, B. (1998). Ecology of Fresh Waters: Man and Medium, Past to Future. (Osney Mead: Blackwell Science Ltd.)

Murakami, T., Kuroda, N. and Tanaka, T. (1998). Effects of a rivermouth barrage on planktonic algal development in the lower Nagara river, central Japan, Japanese Journal of Limnology, **59 (3),** 251-262.

Nelson, M., Hill, T., Remus, W. and O'connor, M. (1999). Time series forecasting using neural networks: Should the data be deseasonalized first?, Journal of Forecasting, **18 (5),** 359-367.

Oh, H. -M., Ahn, C. -Y., Lee, J. -W., Chon, T. -S., Choi, K. H. and Park, Y. -S. (2007). Community patterning and identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using artificial neural networks. Ecological Modelling, **203 (1-2),** 109-118.

Papale, D. and Valentini, R. (2003). A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. Global Change Biology, **9 (4),** 525-535.

Principe, J. C., Euliano, N. R. and Lefebvre, W. C. (2000). Neural and adaptive systems: Fundamentals through simulations. (New York: John Wiley and Sons, Inc.)

Recknagel, F., Cao, H., Kim, B., Takamura, N. and Welk, A. (2006a). Unravelling and forecasting algal population dynamics in two lakes different in morphometry and eutrophication by neural and evolutionary computation. Ecological Informatics, **1 (2),** 133-151.

Recknagel, F., Talib, A. and Van Der Molen, D. (2006b). Phytoplankton community dynamics of two adjacent Dutch lakes in response to seasons and eutrophication control unravelled by non-supervised artificial neural networks. Ecological Informatics, **1 (3),** 277-285.

Reynolds, C. S. (1992). Algae. (In P. Calow and G.E. Petts (Eds.), The River Handbook: Hydrological and Ecological Principles. (pp. 195-215). Oxford: Blackwell Scientific Publication.)

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by backpropagation errors. Nature, **323,** 533-536.

Schumann, G. and Lauener, G. (2005). Application of a degree-day snow depth model to a Swiss glacierised catchment to improve neural network discharge forecasts. Nordic Hydrology, **36 (2),** 99-111.

Sharp, J. J. and Howe, L. J. (2000). The Sarawak River barrage - hydrotechnical and geotechnical aspects, Proceeding of the Institution of Civil Engineers. Water, Maritme and Energy, **142 (2),** 87-96.

Smirnov, N. N. and Timms, B. V. (1983). A revision of the Australian Cladocera (Crustacea). Records of the Australian Museum Supplement, **1,** 1-132.

Song, H. and Li, G. (2008). Tourism demand modelling and forecasting - A review of Recent research. Tourism Management, **29 (2),** 203-220.

Velo-Suárez, L. and Gutiérrez-Estrada, J. C. (2007). Artificial neural network approaches to one-step weekly prediction of *Dinophysis acuminata* blooms in Huelva (Western Andalucía, Spain), Harmful Algae.

Walter, M., Recknagel, F., Carpenter, C. and 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),** 97-113.

Wang, H., Wang, J. P. and Tian, W. F. (2006). A seasonal GRBF network for nonstationary time series prediction. Measurement Science and Technology, **17 (10),** 2806-2810.

Wetzel, R. G. (1983). Limnology. (Philadelphia: Saunders College Publishing)

Wetzel, R. G. and Likens, G. E. (1991). Limnological Analysis. (New York: Springer-Verlag)

Whigham, P. A. and Recknagel, F. (2001). Predicting chlorophyll-*a* in freshwater lakes by hybridising process-based models and genetic algorithms. Ecological Modelling,**146,** 243-251.

Wu, C. L., Chau, K. W. and Li, Y. S. (2009). Methods to improve neural network performance in daily flows prediction. Journal of Hydrology,**372 (1-4),** 80-93.

Yao, X. (1999). Evolving Artificial Neural Networks. Proceedings of the IEEE, **87,** 1423-1447.

Yao, X. and Liu, Y. (2001). Evolving neural networks for chlorophyll-a prediction. (Paper presneted at Fourth International Conference on Computational Intelligence and Multimedia Applications Yokusika City, Japan).

Yoo, H. -S. (2002). Statistical analysis of factors affecting the Han River water quality. Journal of Korean Society of Environmental Engineers, **24 (12),** 2139-2150.

Zhang, G. P. and Kline, D. M. (2007). Quarterly time-series forecasting with neural networks. IEEE Transactions on Neural Networks, **18 (6),** 1800-1814.

Zhang, G. P. and Qi, M. (2005). Neural network forecasting for seasonal and trend time series. European Journal of Operational Research, **160 (2),** 501-514.