available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/ecolinf

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

Kwang-Seuk Jeong, Dong-Kyun Kim, Gea-Jae Joo *

Department of Biology, Pusan National University, Jang-Jeon Dong, Gum-Jeong Gu, Busan 609-735, South Korea

ARTICLE INFO

Article history:
Received 5 January 2005
Received in revised form
15 April 2006
Accepted 26 April 2006

Keywords:
Phytoplankton prediction model
Artificial Neural Networks
Multivariate Linear Regression
Ecological modelling
Input information

ABSTRACT

In this study, a comparison between statistical regression model and Artificial Neural Network (ANN) is given on the effectiveness of ecological model of phytoplankton dynamics in a regulated river. From the results of the study, the effectiveness of ANN over statistical method was proposed. Also feasible direction of increasing ANN models' performance was provided. A hypertrophic river data was used to develop prediction models (chlorophyll *a* (chl. *a*) 41.7±56.8 μg L^{-1}; n=406). Higher time-series predictability was found from the ANN model. Failure of statistical methods would be due to the complex nature of ecological data in the regulated river ecosystems. Reduction of ANN model size by decreasing the number of input variables according to the sensitivity analysis did not have effectiveness with respect to the predictability on testing data set (RMSE of the ANN with all 27 variables, 25.7; 47.9 from using 2 highly sensitive variables; 42.9 from using 5 sensitive variables; 33.1 from using 15 variables). Even though the ANN model presented high performance in prediction accuracy, more efficient methods of selecting feasible input information are strongly requested for the prediction of freshwater ecological dynamics.

1. Introduction

When researchers try to explain or predict the dynamics of river ecosystems, it is important to select suitable methodologies for their purposes. Modelling is a good tool for system simulation and information abstraction about the target ecosystem (see Mangel and Clark, 1988). However, the ecological data set which is used as the source of the model usually has a unique characteristic, so-called "non-linearity" or "complexity", so that the capacity of model on dealing with this complexity is a crucial point (Fielding, 1999).

Currently improved techniques enable researchers to monitor river ecosystems efficiently (e.g., cheaper and accurate detection devices). It is possible nowadays to find or collect underlying but unseen ecological information which might not be discerned previously. However, sufficient ecological information sometimes makes ecological modellers be confused in selecting adequate environmental variables to explain the interested phenomena. This would be especially raised in developing mathematical or deterministic models (i.e., it causes enormous time and cost to improve current model architecture). Even though empirical models such as Machine Learning (ML) algorithms are known as appropriate to ecological modelling because of their capacity in dealing with non-linearity (e.g., Chon et al., 1996; Lek et al., 1996; Lee et al., 2003; Huanga and Foo, 2002; Papale and Valentini, 2003; Jeong et al., 2005), they would not be free from this situation as well.

* *Corresponding author*. Tel.: +82 51 510 2258.
E-mail address: gjjoo@pusan.ac.kr (G.-J. Joo).

1574-9541/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2006.04.001

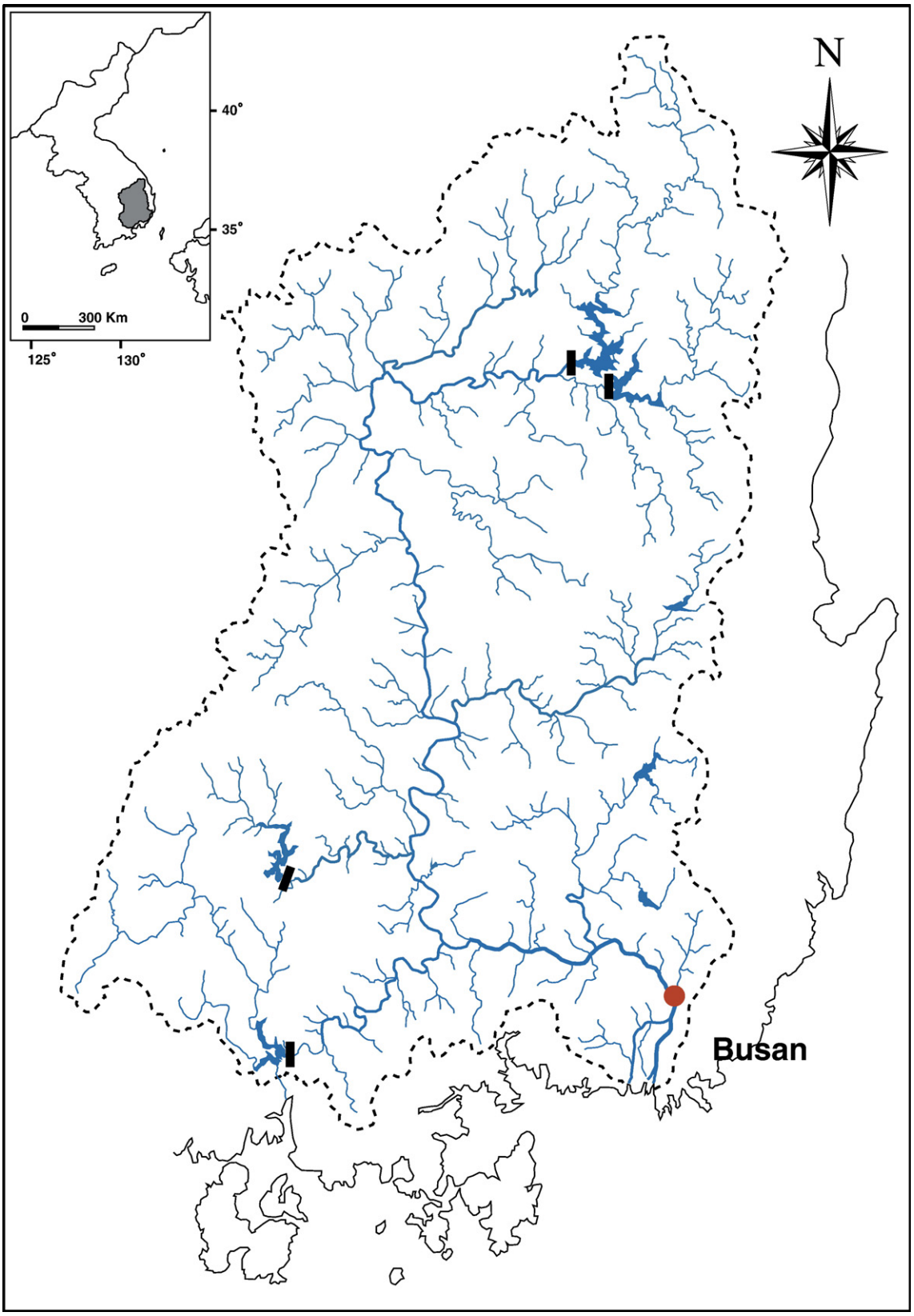

Fig. 1 – The map showing the basin of the Nakdong River and the rainfall distribution on the basin. ■, the multi-purpose dams; ●, the study site.

In this study, the predictability of Artificial Neural Network (ANN) was evaluated on the changes of river phytoplankton proliferations with a variety of environmental variables as a case study. Different number of environmental variables was used step-wise, and the ANN models' performance was shown in predicting time-series changes of phytoplankton proliferation. Comparatively, a linear statistical model of Multivariate Linear Regression (MLR) was adapted to compare the prediction accuracy of algal proliferation, with the same environmental variables. The results may provide the following information: (1) comparison of performance between MLR and ANN in river phytoplankton prediction model, (2) changes of predictability of ANN model along with different number of environmental variables, and (3) significance of the number of input variables with respect to the predictability of ANN model on phytoplankton proliferation (i.e., selection of input variables). Following discussion will suggest the feasible direction of efficient ML approach in ecological model.

2. Materials and methods

2.1. *Description of the study site*

The Nakdong River basin (35–37°N, 127–129°E) lies in the monsoon region of eastern Asia (S. Korea) (Fig. 1). There are four distinct seasons with concentrated rainfall in the early summer (mid June to mid July) caused by the monsoon climate. After the monsoon rainfall, there are several typhoon events that bring large amounts of precipitation over the remainder of summer. The annual mean precipitation across the river basin is about 1200 mm, but about 60% of the annual rainfall is concentrated during the summer season (June–August). The annual mean water temperature in the lower section of the river is 13.7 °C. The mean water temperature during the coldest month (January) is 2.2 °C and in the warmest month (August) is 25.9 °C.

The main channel of the river is 526 km long, and the catchment area occupies about 25% of the S. Korea comprising 23,817 km^2 that covers parts of 7 cities. The study site (Mulgum: water intake station; maximum depth of ca. 11 m; mean depth of ca. 4 m; river width of 250–300 m) is located 27.4 km above the river mouth barrage.

Modification of the river channel has progressed for the last 3 decades due to high water demand. The Nakdong River has 4 multipurpose dams in the upper part of the river and an estuarine barrage for preventing salt water intrusion. About 10 million people depend on the river for drinking, agricultural and industrial water supply. Physical alterations have accelerated eutrophication of the lower part of the river (Joo et al., 1997; Kim et al., 1998).

2.2. *Sampling strategy and data collection*

The precipitation data during the study period was obtained from the 5 representative meteorological stations within the Nakdong River basin (Andong, Daegu, Hapchun, Jinju, and Miryang). Discharge data of the study period was gained from the Flood Control Center. Data of irradiance, wind velocity, cloudiness and evaporation were collected from the Busan Local Meteorological Station, which is the nearest station to the study site. The hydrological characters of four major dams (Andong, Imha, Hapchon and Namgang dams; abbreviated AD, IH, HP and NG respectively) were represented by the dam water storage and discharge out of the dams. These data were supported by the Korean Water Resources Corporation. All the supported meteorological and hydrological data were daily averaged. To prepare the dataset with limnological variables, those environmental variables were 3-day-ahead averaged (discharge, wind velocity, cloudiness and dam properties) or 3-day-ahead cumulated (irradiance and precipitation) from the limnological monitoring dates.

Weekly water samples were obtained at 0.5 m depth at the study site from 1994 to 2001. To determine the limnological characteristics of the study site, physico-chemical variables (water temperature, concentration of dissolved oxygen (DO),

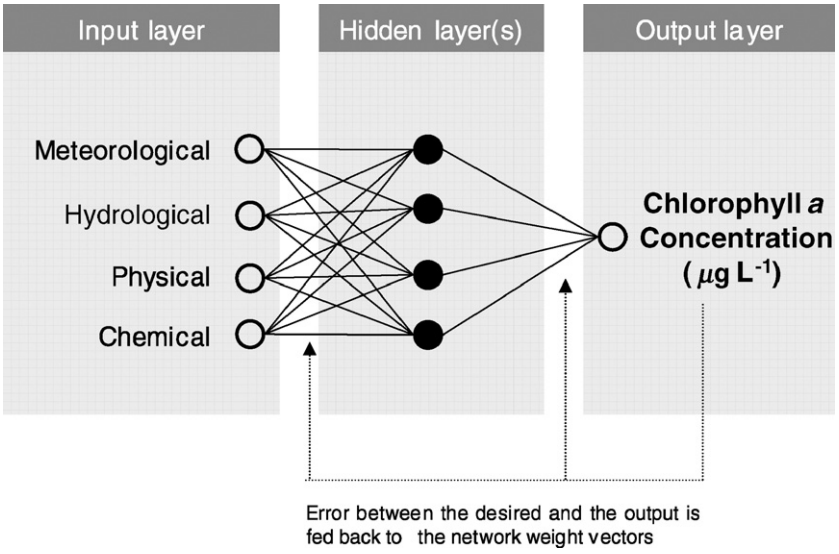

Fig. 2 – The applied modelling architecture using MLP in this study.

Secchi transparency, pH, turbidity, nitrate (NO_3^-–N), ammonia (NH_4^+–N), phosphate (PO_4^{3-}–P), and silica (SiO_2) concentration) and biological parameter (chl. *a*) were investigated. Water temperature, and dissolved oxygen (DO) were determined with an YSI DO meter (Model 58). Transparencies were determined using a 20 cm-diameter Secchi disk. An Orion pH meter (Model 250A) was used to measure pH, and turbidity (NTU) was detected by a Turbidimeter (Model 11052). Water samples were filtered using 0.45 μm Whatman GF/C glass filters to determine the nutrient concentrations and these filtrates were frozen and analyzed by a QuikChem Automated Ion Analyzer (NO_3^-–N, No. 10-107-04-1-O; NH_4^+–N, No. 10-107-06-1-B; PO_4^{3-}–P, No. 10-115-01-1-B; SiO_2, No. 10-114-27-1-A). Chlorophyll *a* (chl. *a*) concentration was spectrophotometrically determined using extraction methods described by Wetzel and Likens (1991).

2.3. *Mode development for the time-series prediction*

Two types of algorithms for the time-series data process were adopted in this study. Firstly, for the linear prediction model, an MLR equation was developed using the collected and supported limnological database. This statistical process was popularly adapted previously for the prediction of system behavior (Ball et al., 2000; Giraudel and Lek, 2001). The environmental variables were used as input data, and the chl. *a* was the target. The result of prediction was used to compare the predictability for the chl. *a* concentration with the following ANN model.

Multi-Layer Perceptron (MLP) was utilized to predict the time-series changes of the chl. *a* concentration. The ANN is originally inspired from the animal neural systems. Even though there are some differences between the real nerve systems and the ANN, this algorithm is known as a suitable methodology for the prediction as well as elucidation of the ecological theories and phenomena.

The MLP consists of several layers which deal with the data input, internal calculations and output. In this study, a total of three layers of MLP were constructed, as input, hidden and output layers (Fig. 2). The Backpropagation (BP) algorithm was used to reduce the error of the calculation of the MLP. For the input layer, the 27 environmental variables were placed which were used in the statistical model, and training iteration was given as 1500. In the training, learning rule adopted momentum process, and any layer momentum value was fixed at 0.7. Learning rate was determined at level of 1 on the hidden layer and 0.1 at output layer. The number of nodes in the hidden layers varied from 2 to 31, and the best-predicting network was selected after five networks as replicate were developed at each number of nodes, on the basis of Root Mean Squared Error (RMSE) with the best weight vector after training.

Training of both MLR and MLP was implemented with the 7 years dataset (1995–2001), and the testing was undertaken on the year 1994's phytoplankton dynamics (chlorophyll *a*). List of the input variables is shown in Table 1. The number of training data was 361 for each of input variable, and that of testing data was 37. Two percents of the training data which were randomly

Table 1 – Descriptive statistics of three data sets used to train MLR and MLP models

Variables	Units	Overall (n=406)			Training (n=361)			Cross validation (n=8)			Testing (n=37)		
		Mean±S.D.	Max.	Min.	Mean±S.D.	Max.	Min.	Mean±S.D.	Max.	Min.	Mean±S.D.	Max.	Min.
Air temp.	°C	15.4±8.0	29.0	−1.0	15.2±8.0	28.9	−1.0	17.1±6.7	26.6	8.0	17.4±8.4	29.0	2.2
Wind vel.	m s^{-1}	3.8±1.0	7.8	0.9	3.8±1.0	7.8	0.9	3.2±1.1	4.7	1.8	4.1±1.0	7.1	2.3
Cloudiness		4.7±2.6	10	0.0	4.7±2.6	10.0	0.0	6.5±2.7	9.8	3.1	4.4±2.4	9.4	0.7
AD Storage	×10^6 ton	612±199	1201	253	626±197	1201	254	631±307	995	253	473±139	882	327
AD discharge	m^3 s^{-1}	30±25	157	0	30±26	157	0	22±11	40	9	34±18	66	5
IH storage	×10^6 ton	221±84	534	126	226±86	534	126	212±100	346	128	174±44	299	128
IH discharge	m^3 s^{-1}	18±25	278	0	18±26	278	0	20±29	87	1	11±7	29	2
HC storage	×10^6 ton	356±133	750	172	67±132	750	173	380±178	581	172	250±64	497	190
HC discharge	m^3 s^{-1}	19±16	98	0	20±17	98	1	16±10	31	6	13±15	76	0
NG storage	×10^6 ton	44±21	153	13	45±22	153	18	37±14	64	19	34±13	60	13
NG discharge	m^3 s^{-1}	68±159	1453	4	70±160	1453	6	117±227	666	12	36±78	471	4
Rainfall	mm	9.7±22.7	237.5	0.0	9.8±23.3	237.5	0.0	19.8±31.1	93.5	0.0	6.2±12.7	44.2	0.0
Evaporation	mm	9.7±3.7	20.4	1.8	9.6±3.6	19.9	1.8	7.6±3.7	15.8	4.3	10.8±4.5	20.4	2.2
Irradiance	MJ m^{-2}	39±15	80.3	5.7	39±15	76	6	28±12	50	12	43±17	80	15
River discharge	m^3 s^{-1}	562±644	6168	9	572±669	6166	9	840±871	2782	248	405±57	575	312
Water temp.	°C	16.8±8.9	34.4	1.1	16.6±8.8	32.7	1.1	16.6±8.7	28.6	4.8	19.1±9.6	34.4	3.0
DO	mg L^{-1}	10.9±3.7	20	3.7	11.1±3.7	20.0	3.7	9.7±4	17.3	5.3	9.9±3.9	19.0	3.8
DO saturation	%	10.9±31	237.8	8.4	109±28	183	8	103±37	183	60	109±49	238	37
pH		8.3±0.8	10.2	6.4	8.2±0.8	9.8	6.4	8.4±1.1	9.4	6.7	8.6±0.9	10.2	7.0
Secchi trans.	cm	79±32	262	5	80±33	262	5	70±29	120	23	74±21	120	43
Turbidity	NTU	14.6±42.0	648.0	1.8	15.2±44.5	648.0	1.8	11.8±9.4	32.5	4.8	9.3±6.6	34.6	3.6
Conductivity	μs cm^{-1}	318±130	670	78	319±133	670	78	354±141	515	140	309±93	640	188
Alkalinity	mg L^{-1}	49±21	102	0	49±21	102	0	54±21	84	30	54±13	84	36
Nitrate-N	mg L^{-1}	2.8±1.0	5.6	0.0	2.9±0.9	5.6	0.0	3.0±0.8	4.4	2.2	1.9±0.9	4.1	0.2
Ammonia-N	mg L^{-1}	0.5±0.6	4.0	0.0	0.5±0.6	4.0	0.0	0.3±0.3	0.7	0.0	0.3±0.3	1.6	0.0
Phospate P	μg L^{-1}	40.7±30.7	238.6	0.0	41.4±31.5	238.6	0.0	46.7±25.2	91.4	10.0	32.7±22	101.0	5.0
Silica	mg L^{-1}	5.2±4.1	21.6	0.0	5.3±4.1	16.3	0.0	6.6±7.0	21.6	0.5	3.5±2.4	9.2	0.0
Chlorophyll	μg L^{-1}	41.7±56.8	573.8	0.0	40.1±53.6	573.8	0.0	35.0±35.5	107.6	6.3	59.0±83.5	479.0	7.0

Air temp. indicates air temperature, wind vel. for wind velocity, and Secchi trans. for Secchi transparency.

chosen (i.e., 8 of data) were used as cross-validation. To avoid over-fitting of the network, training was stopped if there was no improvement in cross-validation process for 100 iterations.

Recurrent proliferation of phytoplankton is a serious problem in the Nakdong River system, and causes problems on water purification system (Jeong et al., 2003). Therefore it is strongly requested to develop ecological models that can predict severe algal blooms, such as proliferations in year 1994. On the other hand, Jeong et al. (2001) have successfully constructed a neural network model in predicting phytoplankton biomass in the Nakdong River system. However, they did not involve dam hydrology which was regarded as an important factor of controlling algal proliferations. Testing on this circumstance may reveal the possible relationship between environmental variables and severe phytoplankton blooms in the river system. In 1994, there was significant proliferations of phytoplankton (*Microcystis* spp. and *Stephanodiscus* spp.) occurred in summer and winter (Ha et al., 1998, 1999).

Sensitivity analysis was conducted with the best-predicting MLP model. Fifty values for each of input variable were linearly produced, ranged within mean±standard deviation. They were fed to an input variable while the others were fixed at their average levels. Standard deviation of the output from the model was then calculated, and the values were used as sensitivity of the input variable onto the changes of chlorophyll *a* (unit, $\mu g\ L^{-1}$).

Reduction of the number of input variables was given in constructing additional MLP models to find the changes in predictability. There was no specified criterion in determining how the variables were more or less sensitive, therefore the number of input variables for the subsequent models were decided according to the sensitivity result of the main MLP model. Firstly, two variables which had the highest sensitivity in the prior MLP model (i.e., pH and ammonia-N; sensitivity over 4.0 $\mu g\ L^{-1}$) were used to make Additional Model 1. The Additional Model 2 used 5 variables, including those two highly sensitive variables along with evaporation, nitrate-N, and water temperature (sensitivity over 1.0 $\mu g\ L^{-1}$). The third additional model (i.e., Additional Model 3) utilized a total of 15 variables (involves those five sensitive variables and cloudiness, wind velocity, DO, HC dam discharge, NG dam storage, Secchi transparency, AD dam discharge, alkalinity, DO% saturation, and silica concentration; sensitivity over 0.1 $\mu g\ L^{-1}$). The prediction accuracy (RMSE and visual comparison) was compared with the prior MLP model. Those subsequent models were also evaluated by sensitivity analysis. In this case, it does not make sense to compare the values themselves between sensitivity results, so that the relative importance of the input variables was compared.

All the MLR equation was developed by using the Time-Series Toolbox from Whigham and Keukelaar (2001). The ANN models were constructed by the NeuroSolution for Excel 4.0.

3. Results

3.1. River limnology and descriptive statistics of data sets

Table 1 summarized descriptive statistics of environmental variables used to develop models. The lower Nakdong River had moderate water temperature variations, and experienced serious eutrophication (mean phosphate-P>40 $\mu g\ L^{-1}$, mean chlorophyll a>40 $\mu g\ L^{-1}$). The multi-purpose dams and river hydrology showed high variations in their water impoundment and discharge, which are directly related to the climate variability such as monsoon and typhoon events. Due to the eutrophication and agitation of water body, Secchi transparency and turbidity had large variations as well.

The averages of two data sets (i.e., training and cross validation) had small differences in each variable, compared with testing set. Because the testing data set consisted of the limnological data from the year 1994 when severe drought occurred, dam impoundment, discharge and river flow had relatively low values. In contrast, temperature-related variables such as air and water temperature, irradiance and evaporation showed slightly higher values compared with training and cross validation data sets. Other variables had slight differences but nutrient concentrations were lower than training and validation sets.

3.2. Prediction by linear and non-linear models

When comparing the prediction capacity between linear (MLR) and non-linear (MLP) models, the latter case produced good prediction results (Fig. 3A). The timing of peaks and the magnitude of each peak could be well detected by the MLP model (Fig. 3A). The MLR model prediction had small fluctuations and varied only between 150–250 $\mu g\ L^{-1}$ of chl. *a* even though the timing was recognized. The MLR equation is shown in Eq. (1):

$$\begin{aligned}\text{Chl. } a = {} & -0.533 \times \text{AT} - 0.0417 \times \text{Wind} - 0.025 \times \text{Cloud} \\ & + 0.0257 \times A_{st} - 0.002 \times A_{dis} + 0.006 \times I_{st} + 0.033 \\ & \times I_{dis} - 0.001 \times H_{st} + 0.473 \times H_{dis} - 0.002 \times N_{st} - 0.001 \\ & \times N_{dis} + 0.013 \times \text{Rain} + 1.964 \times \text{Eva} - 0.006 \\ & \times \text{Irr} - 0.001 \times \text{Dis} + 0.597 \times \text{WT} + 4.547 \\ & \times \text{DO} - 0.003 \times \text{DO\%} + 6.029 \times \text{pH} - 0.001 \times \text{Sec} \\ & + 0.084 \times \text{Turb} + 0.054 \times \text{Cond} + 0.213 \\ & \times \text{Alk} - 0.049 \times \text{Nit} - 0.070 \times \text{Amm} - 0.001 \times \text{Pho} \\ & + 0.360 \times \text{Sil} \end{aligned} \quad (1)$$

where AT is air temperature, Wind for wind velocity, A_{st} and A_{dis} for dam storage and discharge of AD, I_{st} and I_{dis} for dam storage and discharge of IH, H_{st} and H_{dis} for dam storage and discharge for HC, N_{st} and N_{dis} for dam storage and discharge for NG, Rain for rainfall, Eva for evaporation, Irr for irradiance, Dis for river discharge, WT for water temperature, DO for dissolved oxygen, DO for percent saturation of DO, Sec for Secchi depth, Turb for turbidity, Cond for conductivity, Alk for alkalinity, Nit for nitrate, Amm for ammonia, Pho for phosphate and Sil for Silica concentrations.

In the case of MLP model (Fig. 3B), the best-predicting network model consisted of 22 hidden nodes. Root Mean Squared Error (RMSE) from the best weight vector after training was 25.7 (r^2=0.77). When the model was tested onto 1994's phytoplankton proliferation, all of the spring–summer–winter peaks were well recognized, and the model produced good magnitude of peaks against the observed dataset of the year 1994. Slight over-estimation was detected during autumn increase of the chl. *a*.

3.3. *Sensitivity of MLP model*

The developed MLP model used all the environmental variables listed in the Table 1. Sensitivity of each variable in this model is shown in Fig. 4A. Among the variables, mostly physico-chemical environments were important to the changes of phytoplankton dynamics: i.e., pH and ammonia-N concentration were the most sensitive to the changes of chlorophyll *a*, followed by evaporation, nitrate-N concentration and water temperature (sensitivity over 1 μg L^{-1}). Meteorological (i.e., cloudiness, wind velocity, air temperature, and irradiance) and hydrological variables (i.e., HC dam discharge, NG dam storage, AD dam discharge, and rainfall) had relatively small sensitivity on chlorophyll *a* (between 0.1 and 1.0 μg L^{-1}). In this range, most of limnological variables such as Secchi transparency, turbidity and alkalinity were involved. The remaining dam and river hydrology had almost no sensitivity (less than 0.1 μg L^{-1}).

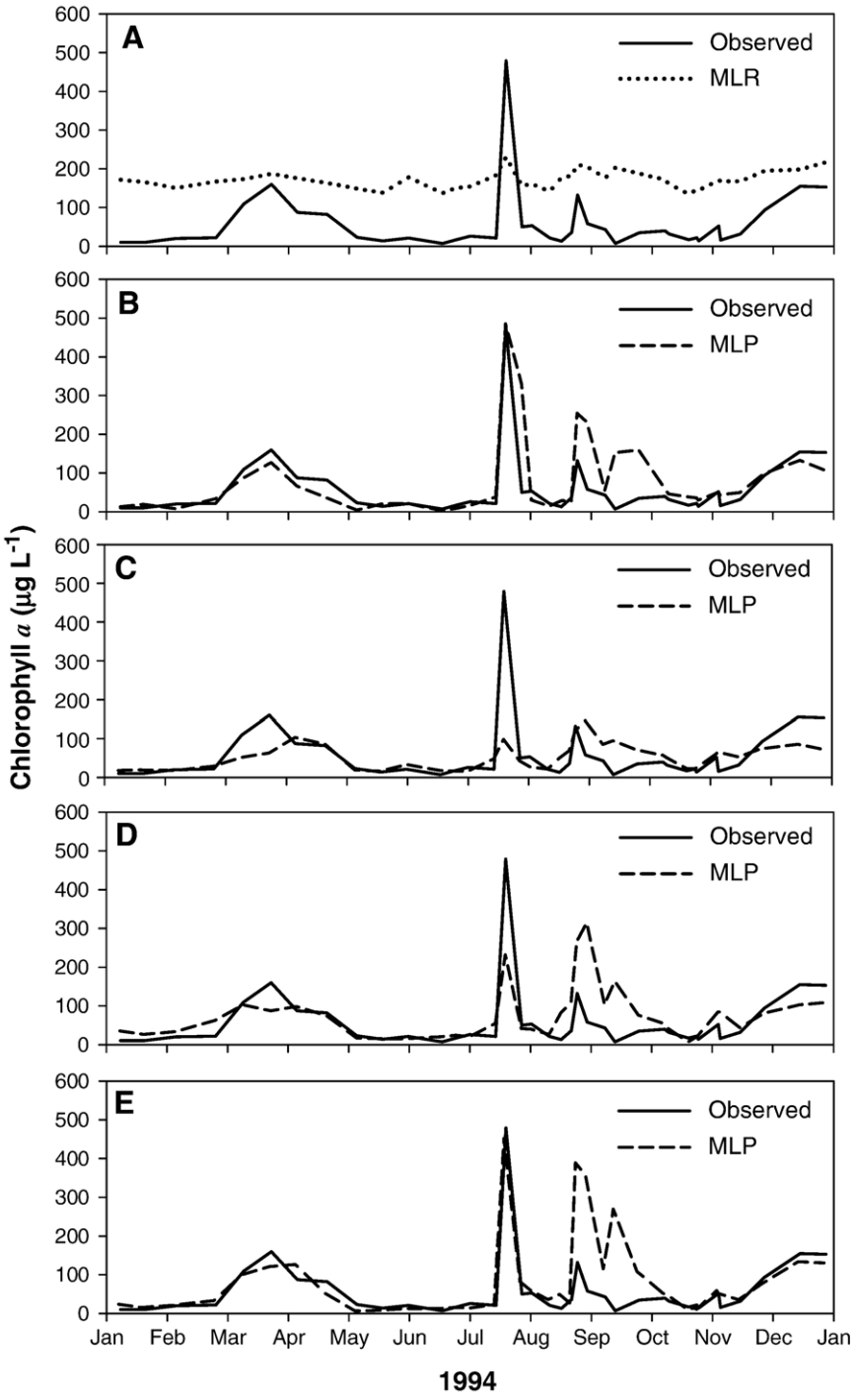

Fig. 3 – The results of prediction through the MLR and MLP against the observed values of 1994 chl. *a*. A, prediction by MLR; B, main MLP; C, Additional Model 1; D, Additional Model 2; E, Additional Model 3.

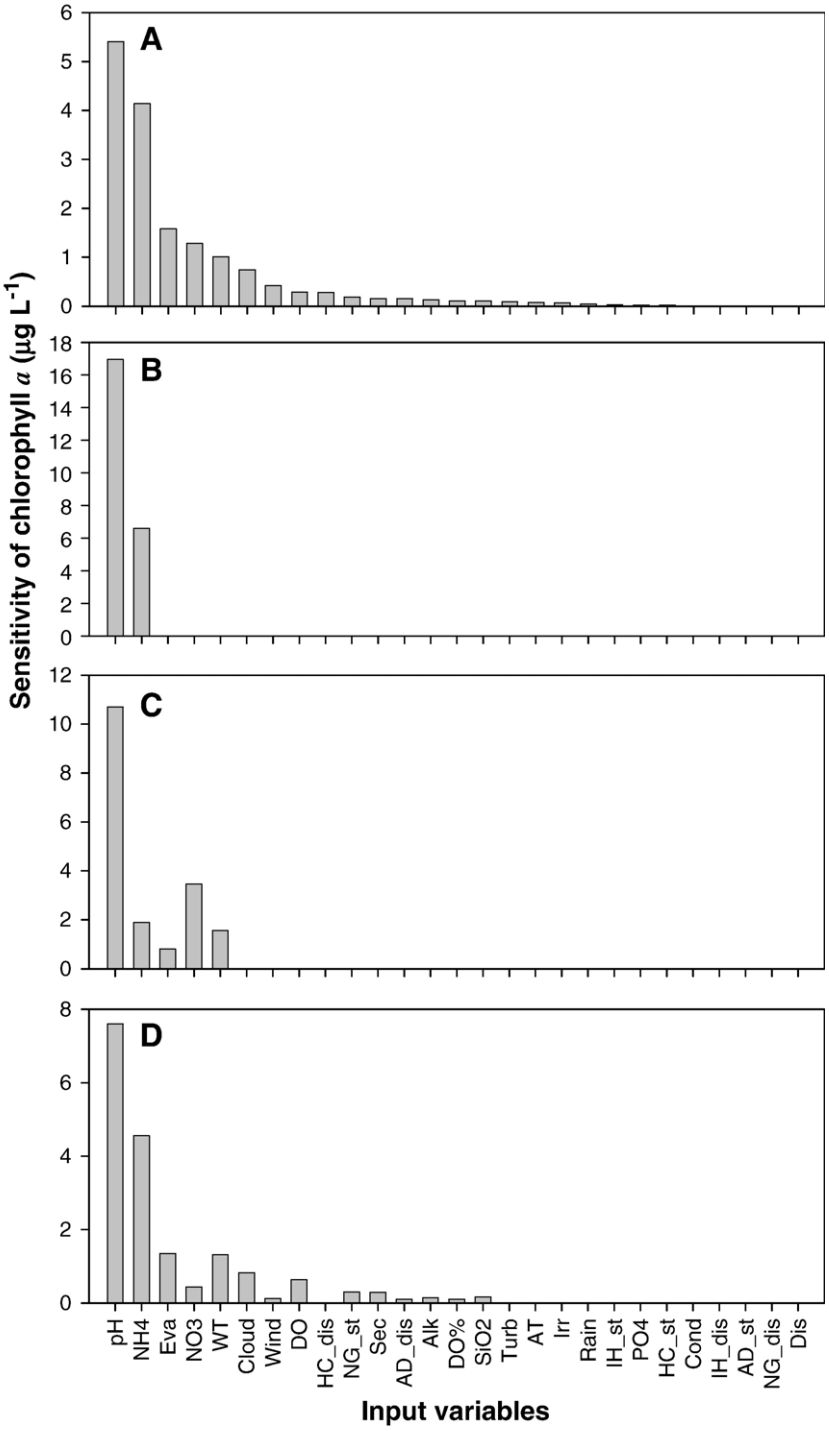

Fig. 4 – Sensitivity analysis on the developed MLP models. A, Main MLP model; B, Additional Model 1; C, Additional Model 2; D, Additional Model 3.

3.4. *Capacity of prediction with different number of input variables*

The values of RMSE onto training data set were inversely produced with the increase in the number of input variables from subsequent MLP models (RMSE 72.8 from Additional Model 1, 74.6 from Additional Model 2, and 83.1 from Additional Model 3). However, onto testing data, effective prediction on the algal proliferation could be observed with increase of the number of input variables (RMSE 47.9 from Additional Model 1, 42.9 from Additional Model 2, and 33.1 from Additional Model 3). The prediction accuracy for both training and testing data set was highest in the main MLP model.

Fig. 3 depicted the testing results of three subsequent MLP models. The best-predicting result was from the main MLP

model, and the Additional Model 3 had relatively similar predictability to the main model. Among the subsequent models, the more input variables were used in the MLP models, the higher predictability could be observed. The Additional Model 1 failed to predict exact magnitude of phytoplankton biomass, even though the timing of proliferation was recognized weakly (Fig. 3C). Fig. 3D was from the Additional Model 2, and this model recognized the timing of proliferation as well. But the largest summer peak was not well predicted, while over-estimation in autumn peak was observed. Overall changing pattern of the predicted chlorophyll *a* in the Additional Model 3 was similar to the main MLP model, and summer peak was well recognized (Fig. 3E). However this model produced two large peaks in autumn.

In decreasing the number of used input variables, sensitivity values tended to increase, however, overall aspect of sensitivity was similar to each model (Fig. 4). There was some difference in the importance of variables among the main and subsequent models. Compared with the main MLP model, the Additional Model 1 made a similar aspect of sensitivity changes to the main MLP mode (Fig. 4B). Between two variables of the Additional Model 1, relative importance of pH increased largely compared with the main MLP model. In the Additional Model 2, nitrate-N played a more important role on determining chlorophyll *a* than evaporation and ammonia-N concentration (Fig. 4C). In the case of the Additional Model 3 (Fig. 4D), nitrate-N, wind velocity and HC dam discharge were less sensitive to chlorophyll *a* compared with the main MLP model (Fig. 4A).

4. Discussion

4.1. *Performance of the neural network model*

From the presented results, it would be obvious that the neural network model has superiority in prediction of chlorophyll *a* concentration in a regulated river system. High complexity of the limnological database would be the primary cause of failure in prediction by the statistical model. Different from natural river systems, the regulated rivers experience two phases in turn: i.e., river phase and reservoir phase (Kim et al., 1998). Two system characteristics exist in an ecosystem seasonally, which may increase complexity of the system. Even though the statistical models (either they are parametric or non-parametric) were successful to analyze the dynamics in ecosystems, the models have sufficient explanatory ability if the data tend to be linear. It is known that biological as well as environmental dynamics in ecosystems are not expressed in the linear manner in fields. Gevrey et al. (2003) emphasized two defects of regression models in analyzing field data: i.e., incapacity to take into account non-linear relationships between dependent and independent variables, and inability to explain the characteristics of ecosystem dynamics from the model. Statistical methods might be adequate to find overall patterns of ecological systems, but the non-linear behavior of ecosystems could not be efficiently reviewed by conventional linear methods (Jeong et al., 2005). Otherwise, most algorithms in machine learning techniques contain the models to deal with non-linear data based on adaptive or heuristic methods.

An alternative mechanism of predicting phytoplankton proliferation is deterministic architecture of ecological model. Well-known frameworks of phytoplankton dynamics (e.g., relationship between algal cell growth rates and temperature) are adapted in this architecture. Compared with traditional deterministic models, there have been notable advances in deterministic models (e.g., Hamilton and Schladow, 1997; Yang et al., 2000; Bonnet and Wessen, 2001; Lewis et al., 2002; Håkanson and Boulion, 2003). However, they are less flexible compared with machine learning techniques due to the restriction of ecosystem explanation which would be available within the information contained in the frameworks (Jeong et al., 2005).

The advantage of ANN models for the phytoplankton in regulated river systems can be embossed from this point of view. Various environmental factors and their couplings affect algal proliferations in rivers, which cause large variations in ecological information. Moreover, the control of river hydrology by humans (i.e., water regulation) may impose the complexity of phytoplankton proliferations as well (Ha et al., 1998; Jeong et al., 2003). Variables related to dam hydrology which had small sensitivity (i.e., less than 1.0 $\mu g\, L^{-1}$; see Fig. 4) would increase the prediction accuracy in the MLP model. It can be thought that the neural network model found the information regarding regulation of river flow and phytoplankton proliferations. It is believed that ANN models provide flexibility in data processing and capacity of dealing with non-linearity, which is thought to be necessary for ecological models (e.g., Guan et al., 1997; Karul et al., 1998; Moatar et al., 1999; Growns and Growns, 2001; Collier, 2002). Also prediction or explanation of phytoplankton dynamics in freshwater systems was popularly conducted in recent decade (Scardi, 2001; Maier et al., 2001a,b; Joo and Jeong, 2005).

In some cases, the combination between linear and non-linear approaches could increase the performance of modelling. The data pre-manipulation could reinforce the prediction of MLR (e.g., Monte Carlo method) (Whitehead and Hornberger, 1984). However, the data premanipulation can reduce the dimension of dataset, and it causes sometimes loss of important and crucial information of the ecosystems. When they put the focus of model on the prediction itself, it is worthwhile to utilize those types of approaches. The ecological models should perform ecological information extraction as well as prediction, so that machine learning approaches can be preferable compared with those methods.

4.2. *Effectiveness of selecting input variables*

Due to the improvement of technology, ecologists can use nowadays a variety of environmental information in constructing ecological model. Development of computers enables researchers to use more sophisticated approaches to explore ecosystems as well (Chon et al., 1996). Thus it is possible to use as many of environmental variables as input sources of ANN models. A dilemma exists here between increase of training efficiency and model performance of ANN. From the results of this study, some input variables that had high sensitivity were not sufficient to provide reasonable accuracy of prediction for phytoplankton proliferations. In other words, miscellaneous input variables (e.g., dam

hydrology in this study) had their own importance in phytoplankton dynamics.

From the subsequent models of neural network, it was not possible to generate a model that produced similar predictability to the main MLP model by using a small number of input variables. As shown in Fig. 3, the Additional Model 3 had the most similar prediction accuracy, however, the model produced two large over-estimations in autumn proliferations. These autumn over-estimations were well suppressed in the main MLP model (i.e., using all 27 variables). The variables which were not involved in the Additional Model 3 were mostly related to hydrological circumstances (see Fig. 4) such as river discharge, storage and discharge of upper dams. River discharge is generally believed to have negative relationship to phytoplankton dynamics. Therefore, it could be assumed that the neural network model recognized the importance of hydrological terms in predicting phytoplankton proliferations in a regulated river system. Even though the predictability was best at using all the input variables, it is obvious that 27 input variables will have redundancy in model development Therefore, an effective variable-selecting process is strongly requested in ANN modelling for the prediction of the ecological dynamics in regulated river systems.

It is known that application of many input variables make the model become inefficient (Jørgensen, 1997). Substantial information costs either too long time in ANN models' convergence or a large size of the models. However, ecological models should be able to represent the target ecosystem reasonably.

For the selection of input variable in the development of neural networks, constructive or destructive methods are widely adapted traditionally (Blanco et al., 2000). They are similar to step-forward or -backward algorithms in regression approach, respectively. Less sensitive variables would be excluded from the network models, with respect to the prediction accuracy. However, two fatal flaws exist: i.e., possibility of staying in local minima and ignorance of important information which is unseen in the development process. Relatively different importance of each input variable in the subsequent models might be due to the former reason. Randomly prepared weight vectors in the subsequent MLP models seemed to cause the changes of sensitivity aspect compared with main MLP model. The latter reason would be explained with the results from this study. It was obvious that destructive process might neglect less sensitive variables, and consequently the ANN model would loose its predictive power.

Two suggestions could be derived from the study. This could be applied not only for phytoplankton modelling but also other ecological sciences. The first is to utilize efficient input variables in cost–benefit point of view. Input environmental information would be categorized into two types either the easy-to-obtain (e.g., temperature or DO) or difficult-to-obtain (e.g., satellite imagery). Fig. 5 illustrates modelling capacity according to the utilized ecological information (modified from Jørgensen, 1997). As shown in Fig. 5A, easy sampling variables such as water temperature, DO, etc. requires lower cost for sampling, but suggests important information to the model (might be involved in shade I). The biological abundances such as cell numbers usually take longer time to make the dataset (higher cost) and can be

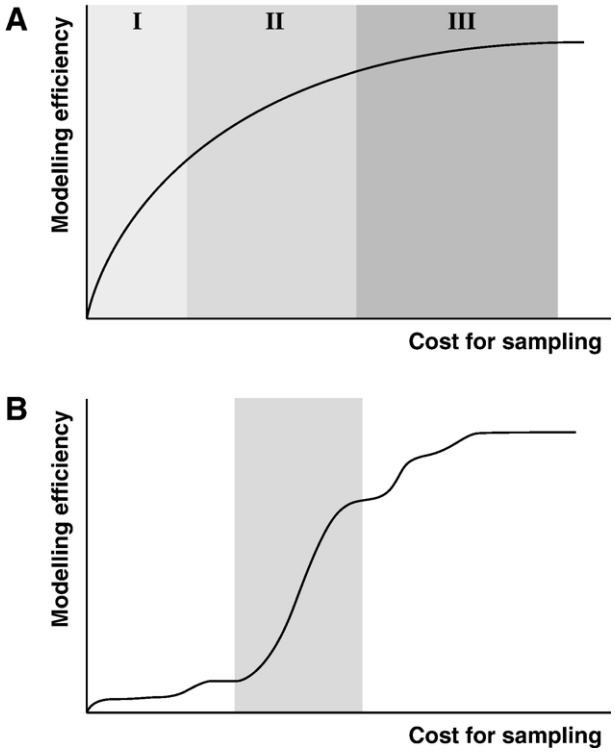

Fig. 5 – The relationship between sampling cost and modelling efficiency. A, Generally expected relationship between cost for collecting information and model performance (modified from Jørgensen, 1997); B, importance of 'firing' variables for the ecological model. I, variables which require low cost to observe, but generate high model performance; II, variables which require moderate cost and return appropriate model capacity; III, variables which request high cost but may not be necessary for modelling.

included in shade II. The shade III means that the variables require too much cost for sampling, but the increase of efficiency is not much satisfactory. Therefore, even though machine learning can deal with every type of input variables, it is necessary to select the suitable variables to achieve the best performance.

Among the difficult-to-obtain variables (e.g., shade II in Fig. 5B), there would be 'firing' variables that make the ANN models become more accurate. It is sure that the development of technology will allow ecologists to use these types of information easily in future. Among that information, ecological model developers should identify the most cost–beneficial information.

The other suggestion is to adopt an algorithm of containing selection mechanism among input information during the development of ANN models. Evolving Neural Network (ENN) (see Yao, 1999; Yao and Liu, 2000) can be thought as a good example. This algorithm was primarily proposed to provide efficient 'generalization' of ANN models. Also local minima problem would be avoided through global search process of Genetic Algorithm. Blanco et al. (2000) emphasized that Genetic Algorithm was possible to select feasible input information out of the provided variables to ANN models. Computational evolution may select the 'firing' information adaptively.

5. Conclusions

Multivariate Linear Regression and Artificial Neural Network were compared to find effective modelling architecture on the prediction of phytoplankton proliferation in a regulated river system. Neural network model produced high accuracy in predicting both magnitude and timing of algal proliferations. For the ANN model, reduction of the number of input variables caused decrease of model performance. Highly sensitive information (i.e., pH, ammonia-N, evaporation etc. in the developed model) was not sufficient to make model become efficient to explain the algal dynamics. It is strongly required to discover effective mechanism to guarantee both performance of modelling and training efficiency during the development of ecological model.

Acknowledgement

Authors appreciate the members of the Limnology Lab., Pusan National University, for their efforts on field survey. Also we are indebted to Profs. Tae-Soo Chon (Pusan National University), Sovan Lek (Universite Paul Sabatier), Friedrich Reckangel (University of Adelaide) and Peter Whighams (University of Otago) for their valuable comments during the logic development and paper writing. This work was supported by the programme of "Grants for Pre-Doctoral Students" (Project No. C00027) by the Korean Research Foundation, and No. 46 manuscript of the "Long-Term Ecological Research of the Nakdong River" from the Limnology Lab., Pusan National University. This study was partially supported by the "Long-Term Ecological Research Programme at the Nakdong River" from the Ministry of Environment, South Korea.

REFERENCES

Ball, G.R., Palmer-Brown, D., Mills, G.E., 2000. A comparison of Artificial Neuronal Network and conventional statistical techniques for analyzing environmental data. In: Lek, S. (Ed.), Artificial Neuronal Networks. Springer-Verlag, Berlin, pp. 165–183.

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

Bonnet, M.P., Wessen, K., 2001. ELMO, a 3-D water quality model for nutrients and chlorophyll: first application on a lacustrine ecosystem. Ecological Modelling 141, 19–33.

Chon, T.S., Park, Y.S., Moon, K.H., Cha, E.Y., 1996. Patternizing communities by using an Artificial Neural Network. Ecological Modelling 90, 69–78.

Collier, K.J., 2002. Effects of flow regulation and sediment flushing on instream habitat and benthic invertebrates in a New Zealand river influenced by a volcanic eruption. River Research and Applications 18, 213–226.

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

Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in Artificial Neural Network models. Ecological Modelling 160, 249–264.

Giraudel, J.L., Lek, S., 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecological Modelling 146, 329–339.

Growns, I.O., Growns, J.E., 2001. Ecological effects of flow regulation on macroinvertebrate and periphytic diatom assemblages in the Hawkesbury-Nepean River, Australia. Regulated Rivers: Research and Management 17, 275–293.

Guan, B.T., Gertner, G.Z., Parysow, P., 1997. A framework for uncertainty assessment of mechanistic forest growth models: a neural network example. Ecological Modelling 98, 47–58.

Ha, K., Kim, H.W., Joo, G.J., 1998. The phytoplankton succession in the lower part of hypertrophic Nakdong River (Mulgum), S. Korea. Hydrobiologia 369/370, 217–227.

Ha, K., Cho, E.A., Kim, H.W., Joo, G.J., 1999. *Microcystis* bloom in the lower Nakdong River in South Korea: Development of the 1994 summer bloom. Marine and Freshwater Research 50, 89–94.

Håkanson, L., Boulion, V.V., 2003. A general dynamic model to predict biomass and production of phytoplankton in lakes. Ecological Modelling 165, 285–301.

Hamilton, D.P., Schladow, S.G., 1997. Prediction of water quality in lakes and reservoirs: Part I. Model description. Ecological Modelling 96, 91–110.

Huanga, W., Foo, S., 2002. Neural network modeling of salinity variation in Apalachicola River. Water Research 36, 356–362.

Jeong, K.S., Joo, G.J., Kim, H.W., Ha, K., Recknagel, F., 2001. Prediction and elucidation of phytoplankton dynamics in the lower Nakdong River (Korea) by means of an Artificial Neural Network. Ecological Modelling 146, 115–129.

Jeong, K.S., Recknagel, F., Joo, G.J., 2003. Prediction and elucidation of population dynamics of the blue–green algae *Microcystis aeruginosa* and *Stephanodiscus hantzschii* in the Nakdong river-reservoir system (South Korea) by Artificial Neural Networks. In: Recknagel, F. (Ed.), Ecological Informatics: Understanding Ecology by Biologically Inspired Computation. Springer, Germany, pp. 195–213.

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

Joo, G.J., Jeong, K.S., 2005. Summer cyanobacterial blooms in the lower Nakdong River: pattern cognition by the Self-Organizing Map. In: Lek, S., Scardi, M., Verdonshot, P.F.M., Descy, J.-P., Park, Y.S. (Eds.), Modelling Community Structure in Freshwater Ecosystems. Springer, Berlin, pp. 273–287.

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

Jørgensen, S.E. (Ed.), 1997. Integration of Ecosystem Theories: A Pattern, 2nd ed. Kluwer Academic Publishers, Dordrecht. 388 pp.

Karul, C., Soyupak, S., Germen, E., 1998. A new approach to mathematical water quality modeling in reservoirs: neural networks. International Review of Hydrobiology 83, 689–696.

Kim, H.W., Ha, K., Joo, G.J., 1998. Eutrophication of the lower Nakdong River after the construction of an estuarine dam in 1987. International Review of Hydrobiology 83, 65–72.

Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecological Modelling 159, 179–201.

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

Lewis, D.M., Elliott, J.A., Lambert, M.F., Reynolds, C.S., 2002. The simulation of an Australian reservoir using a phytoplankton community model: PROTECH. Ecological Modelling 150, 107–116.

Maier, H.R., Burch, M.D., Bormans, M., 2001a. Flow management strategies to control blooms of the cyanobacterium, *Anabaena circinalis*, in the River Murray at Morgan, South Australia. Regulated Rivers: Research and Management 17, 637–650.

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

Mangel, M., Clark, C.W. (Eds.), 1988. Dynamic Modeling in Behavioral Ecology. InPrinceton University Press, NJ. 308 pp.

Moatar, F., Fessant, F., Poirel, A., 1999. pH modelling by neural networks: application of control and validation data series in the Middle Loire River. Ecological Modelling 120, 141–156.

Papale, D., Valentini, R., 2003. A new assessment of European forests carbon exchanges by eddy fluxes and Artificial Neural Network spatialization. Global Change Biology 9, 525–535.

Scardi, M., 2001. Advances in neural network modeling of phytoplankton primary production. Ecological Modelling 146, 33–45.

Wetzel, R.G., Likens, G.E. (Eds.), 1991. Limnological Analyses. Springer-Verlag, New York, p. 391.

Whigham, P.A., Keukelaar, J., 2001. Evolving structure-optimizing content. Proceedings of the Congress on Evolutionary Computation 2001, pp. 1228–1235.

Whitehead, P., Hornberger, G., 1984. Modelling algal behaviour in the River Thames. Water Research 18, 945–953.

Yang, M.D., Sykes, R.M., Merry, C.J., 2000. Estimation of algal biological parameters using water quality modeling and SPOT satellite data. Ecological Modelling 125, 1–13.

Yao, X., 1999. Evolving Artificial Neural Networks. Proceedings of the IEEE, vol. 87, pp. 1423–1447.

Yao, X., Liu, Y., 2001. Evolving neural networks for chlorophyll-*a* prediction. Proceedings of Fourth International Conference on Computational Intelligence and Multimedia Applications, pp. 185–189.