Available online at www.sciencedirect.com
ScienceDirect
Engineering Applications of Artificial Intelligence 20 (2007) 735–744
Engineering Applications of
ARTIFICIAL
INTELLIGENCE
www.elsevier.com/locate/engappai

Machine-learning paradigms for selecting ecologically significant input variables

Nitin Muttil, Kwok-Wing Chau*

Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hung Hom, Hong Kong

Received 14 December 2005; received in revised form 18 October 2006; accepted 19 November 2006
Available online 16 January 2007

Abstract

Harmful algal blooms, which are considered a serious environmental problem nowadays, occur in coastal waters in many parts of the world. They cause acute ecological damage and ensuing economic losses, due to fish kills and shellfish poisoning as well as public health threats posed by toxic blooms. Recently, data-driven models including machine-learning (ML) techniques have been employed to mimic dynamics of algal blooms. One of the most important steps in the application of a ML technique is the selection of significant model input variables. In the present paper, we use two extensively used ML techniques, artificial neural networks (ANN) and genetic programming (GP) for selecting the significant input variables. The efficacy of these techniques is first demonstrated on a test problem with known dependence and then they are applied to a real-world case study of water quality data from Tolo Harbour, Hong Kong. These ML techniques overcome some of the limitations of the currently used techniques for input variable selection, a review of which is also presented. The interpretation of the weights of the trained ANN and the GP evolved equations demonstrate their ability to identify the ecologically significant variables precisely. The significant variables suggested by the ML techniques also indicate chlorophyll-*a* (Chl-*a*) itself to be the most significant input in predicting the algal blooms, suggesting an auto-regressive nature or persistence in the algal bloom dynamics, which may be related to the long flushing time in the semi-enclosed coastal waters. The study also confirms the previous understanding that the algal blooms in coastal waters of Hong Kong often occur with a life cycle of the order of 1–2 weeks.

Keywords: Harmful algal blooms; Red tides; Machine-learning techniques; Data-driven models; Artificial neural networks; Genetic programming; Water quality modelling; Tolo Harbour; Hong Kong

1. Introduction

An explosive growth and accumulation of harmful microscopic algae or phytoplankton may result in harmful algal blooms (HABs). The red tide is a well-known form of algal bloom. Owing to its negative impacts on human health and aquatic life, this widely reported phenomenon has become a serious environmental problem. It might lead to harmful effects such as beach closure, mariculture loss arising from oxygen depletion or toxic algae, anoxia or shellfish poisoning, and so on (Anderson, 1994). An increasing trend in the occurrence of HABs has been recorded throughout the world during the past decade. For example, the worst fish kill in Hong Kong's history in April 1998 as a result of a devastating red tide destroyed over 3400 tonne or 80% of cultured fish stock and the resulting economic loss was estimated to exceed HK$312 million (Chau, 2004). Hence, a capability to analyse and predict the occurrence of algal blooms precisely with sufficient lead-time would contribute significantly to fisheries and environmental management. Conventional knowledge-driven models address the physical problem by solving a highly coupled, non-linear, partial differential equation set with finite difference method, finite element method, etc. However, physical processes affecting HAB occurrence are highly complex and uncertain, and are difficult to be captured in some form of deterministic model. Moreover, the accuracy of the prediction is to a great extent dependent on the accuracy of the open boundary conditions, model parameters used, and the numerical scheme adopted.

*Corresponding author. Tel.: +852 2766 6014.
E-mail address: cekwchau@polyu.edu.hk (K.-W. Chau).

0952-1976/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2006.11.016

With the recent advancements in artificial intelligence (AI) techniques and availability of unrivalled computational power, extensive use of machine learning (ML) techniques or data-driven approaches in ecological modelling was reported (Recknagel, 2001). Amongst others, they include artificial neural networks (ANN) (Recknagel et al., 1997, 2002; Yabunaka et al., 1997; Maier et al., 1998; Scardi and Harding, 1999; Karul et al., 2000; Jeong et al., 2001; Scardi, 2001; Wei et al., 2001; Lee et al., 2003), evolutionary-based techniques (Bobbin and Recknagel, 2001; Recknagel et al., 2002; Jeong et al., 2003; Muttil and Lee, 2005), fuzzy and neuro-fuzzy techniques (Maier et al., 2001; Chen and Mynett, 2003), and so on. Whilst many of them were undertaken in freshwater environments (i.e., limnological or riverine systems), some have been applied to saltwater eutrophic areas (Scardi and Harding, 1999; Scardi, 2001; Lee et al., 2003; Muttil and Lee, 2005).

One of the most important steps in the application of a ML technique is the selection of significant model input variables. The aim is to determine a set of significant inputs from a superset of potentially useful inputs, which will lead to a superior model as measured by some optimality criterion. Since data-driven models are usually assumed to be able to determine which model inputs are critical, researchers often tend to present a large number of inputs to the model. Such inclusion of a large number of inputs leads to "the curse of dimensionality", which is associated with the following shortcomings (Bowden et al., 2005):

- As the input dimensionality increases, the computational complexity and memory requirements of the model increase, which in turn increase the time to build the models.
- As the input variables increase, the number of training samples required also increase.
- Misconvergence and poor model accuracy may result from the inclusion of irrelevant inputs due to an increase in the number of local minima present in the error surface.
- Interpreting complex models is more difficult than interpreting simple models that give comparable results.

Thus, there are obvious advantages in selecting an appropriate set of significant inputs for a data-driven model. The complexity of the problem is further increased in time-series modelling of dynamical systems, where better predictions are obtained by the use of lagged input variables also. As the maximum lag or memory length increases, so too does the number of inputs and the complexity of the model. Recently, researchers have recognized the importance of input variable selection for the application of data-driven models in ecological modelling. In their work on applying ANN for modelling of coastal algal blooms, Lee et al. (2003) noted that most of the literature they reviewed did not build in an optimal choice of input variables based on ecological considerations and many used almost all possible environmental parameters as inputs. Since the effects of some of the input variables may be duplicated (for example, during blooms, algal density and turbidity or secchi-disc depth (SD) are strongly correlated), the use of all possible input variables may present the model with noise, rather than useful information. Maier and Dandy (2000) extensively reviewed a number of journal articles from 1992 to 1998, which employed ANN for modelling and forecasting of water resources variables. They concluded that in many cases, the lack of a methodology for determining input variables raised doubt about the optimality of the inputs used and in some cases, inputs were even chosen arbitrarily.

In the present paper, we employ two extensively used data-driven models for selecting the significant input variables, namely, ANN and genetic programming (GP) using first, a test problem from fluid mechanics and then using the biweekly water quality data from Tolo Harbour, Hong Kong. It should be noted that this work is not a hybrid combination of ANN and GP analysis in a two-stage procedure or ANN encapsulated within the GP framework. This work applies two distinct data-driven models, i.e., ANN and GP, to analyse the significant input variables in eutrophication phenomenon and the results by these two methods are compared. Since the data-driven model is itself used for input variable selection, there is no need to select any other analytical procedure. Moreover, Recknagel (2001) considered ANN and genetic algorithms to be the most innovative techniques for ecological modelling. Maier and Dandy (2000) also reported ANN to be the most widely used model in water resources variable modelling.

In the following sections, we first present a review of the input variable selection techniques used in ecological modelling applications of data-driven models. Then, details of the data and modelling approach are presented, followed by the application of ANN and GP for selection of the significant input variables.

2. Review of previous work on input selection

The main approaches that have been employed for input determination in ecological modelling literature can be broadly classified into five methods, which are presented in the following sub-sections. Other than the methods presented below, few studies have also used all available variables as inputs to their model, without considering any technique for selecting the significant variables (Recknagel et al., 1997; Karul et al., 2000; Jeong et al., 2003).

2.1. Methods based on ecological considerations

A commonly adopted approach in the choice of the initial set of input variables is to apply *a priori* knowledge of causal variables and physical/ecological insight into the problem. If important candidate inputs are not included, then some information about the system may be lost and if spurious inputs are included, it may provide the model with noise and thus may confuse the training process. Many of the papers

reviewed relied on a combination of *a priori* knowledge and analytical approaches to select the appropriate model inputs and lags of inputs (Maier et al., 1998; Chen and Mynett, 2003; Lee et al., 2003; Muttil and Lee, 2005).

2.2. Methods based on linear correlation

When the relationship to be modelled is not well understood, then an analytical technique, such as correlation analysis, is often employed to select inputs and their lags. Correlated variables introduce redundancy in the model in the sense that no additional information is gained by adding them. Lee et al. (2003) used auto-correlation of chlorophyll concentrations to select the time lags of input variables. They showed that algal dynamics in coastal waters of Hong Kong are correlated up to time lags of around 2 weeks. The major disadvantage associated with using a correlation analysis is that it is only able to detect linear dependence between two variables. Therefore, such an analysis is unable to capture any non-linear dependence that may exist between the inputs and the output, and may possibly result in the omission of important inputs that are related to the output in a non-linear fashion.

2.3. Methods based on data mining techniques

Some researchers used data mining techniques like principal component analysis (PCA), cluster analysis, etc., for selecting the significant input variables. Chen and Mynett (2003) used PCA to identify the major abiotic driving factors and to reduce the input dimensionality. In their application of fuzzy logic to model eutrophication in Taihu Lake, Chen and Mynett (2003) noted that the size of fuzzy ruleset grows exponentially with the number of variables (dimension), which results in not only redundancy, but also difficulties for interpretation and formulization. Petersen et al. (2001) also used PCA to provide a reduced description of the system using a few significant and interpretable PCA patterns which reflect the most relevant processes and which can be used in other models. Brosse et al. (2001) compared Kohonen self-organizing maps (SOM) and PCA to analyse the spatial occupancy of several European freshwater fish species in the littoral zone of a large French lake. They concluded that only SOM was able to reliably visualize the entire fish assemblage in a two-dimensional space and PCA provided irrelevant ecological information for some species. They note that this may be because the information given by PCA techniques suffers from some drawbacks in that the relationships between variables in environmental sciences are often non-linear, while the methods used in PCA are based on linear principles.

2.4. Forward selection and backward elimination methods

In this approach, an optimization of the input variables is performed. The two standard approaches are forward selection and backward elimination methods. Forward selection starts by finding the best single input and selecting it for the final model. In each subsequent step, given a set of selected inputs, the input that improves the model's performance most is added to the final model. Backward elimination (network trimming) starts with a set of all inputs, and sequentially deletes the input that reduces performance the least. Maier et al. (1998) used a forward selection approach to determine the important input variables for forecasting the concentration of the cyanobacteria Anabaena spp. in the River Murray, Australia. Lee et al. (2003) used a network trimming approach, starting with the most complicated network scenario and then removing one parameter at a time for the subsequent scenarios. Each scenario would indicate how significant the excluded parameter would affect the network performance. The main disadvantage of these approaches is that they are based on trial-and-error, and as such, there is no guarantee that they will find the globally best subsets. The forward selection approach may also fail when there is interaction amongst variables, i.e., when a variable that is useless by itself may provide a significant performance improvement when taken with others. Another disadvantage of these stepwise approaches is that they are computationally intensive.

2.5. Sensitivity analysis using trained ANN

Sensitivity analyses are the most commonly used method of extracting information from a trained ANN (Maier et al., 1998; Scardi and Harding, 1999; Lee et al., 2003). A stepwise analysis is carried out by varying each input variable, one at a time by a predetermined constant percentage while keeping the others constant. These new input data series are used to compute the predictions from the network trained with the original inputs. However, the difficulty with this approach is choosing a reasonable value to perturb the input by and selecting the appropriate cut-off point for input significance.

3. Data and modelling approach

In this section, we give an account of the test problem, the real-world case study and details of the analysis.

3.1. The test problem

To test the capability of ANN and GP for selecting the significant input variables, they are first tested on a test problem with known dependence. A simple example from fluid mechanics, the Bernoulli's equation is used. Ignoring any losses, the total energy head, E, can be expressed as

$$E = z + \frac{p}{\gamma} + \frac{v^2}{2g} = \text{const}, \quad (1)$$

where z is the vertical distance above a datum (m), p is the pressure (N/m^2), v *is the* velocity (m/s), γ is the specific

gravity of water (9810 N/m^3), and g is the gravitational acceleration (9.81 m/s^2). Using a standard random number generator, 1000 samples of different combinations of z, p and v are generated. The values of the energy head, E, are then computed using Eq. (1). In addition to these 3 input variables of z, p and v, 7 more variables, *ran1, ran2,…,ran7* are also randomly generated, making a total of 10 input variables. The dependent variable E, is actually dependent on only 3 input variables, namely z, p and v and the remaining 7 inputs are spurious, with no relationship with E. All the randomly generated input variables are within the range of 0–100, except z and v, which have random numbers in the range of 0–25. This restriction was imposed on these 2 variables because large values for them would highly increase their significance, especially that of v, which has a squared relationship with E.

3.2. The real-world case study

The excessive growth of aquatic plants, both attached and planktonic, to levels that are considered to be an interference with desirable water uses bring about the eutrophication phenomenon. The growth of aquatic plants results from many causes. Harmful algae are the microscopic single celled organisms that are present in the sea. These algae contain reddish pigments thus the water seems to be appearing in red colour in their presence. The red tide occurs when there is a rapid production in the single celled organisms because of the increased levels of temperature, salinity and nutrient concentration such as nitrogen, phosphorus (PO_4), etc. in the sea water, resulting in reddish water. It should be mentioned that only a few dozen of the many thousands of species of microscopic and macroscopic algae are repeatedly associated with toxic or harmful blooms. The specific type of species ''Gymnodinium Breve'' contains the neurotoxic shell fish poisoning. Similar events world wide include Florida, US (Kirkpatrick et al., 2006),Yeosu and Tongyeong, Korea (Lee, 2006),Osaka Bay, Japan (Tsujimoto et al., 2006), etc.

Tolo Harbour, a semi-enclosed bay in the northeastern coastal waters of Hong Kong, is connected to the open sea at Mirs Bay (Fig. 1). The water quality generally declines from the better flushed outer ''Channel Subzone'' towards the more enclosed and densely populated inner ''Harbour Subzone''. Over the past two decades, one of the major environmental concerns in the harbour is the nutrient enrichment arising from municipal and livestock waste discharges. Whilst point sources of organic loads are mainly derived from the two major treatment plants at Shatin and Taipo, non-point sources come from direct runoff and waste from mariculture. Several previous studies (Morton, 1988; Xu et al., 2004) demonstrated that the healthy state of the marine coastal ecosystem in the Tolo Harbour had been deteriorating continuously as a result of the nutrient enrichment in the harbour brought about by urbanization, industrialization and livestock rearing. Eutrophication phenomenon with frequent algal blooms and red tides has been reported particularly in the weakly flushed tidal inlets inshore. Consequently, occasional massive fish kills were recorded as a result of severe dissolved oxygen (DO) depletion or toxic algal blooms. Morton (1988) reported that the inner Tolo Harbour was effectively dead as a marine disaster in the late 1980s. At that time, a critical stage had been reached, which prompted the Hong Kong Government to implement an integrated Tolo Harbour Action Plan (THAP). THAP resulted in a significant reduction of pollutant loading which in turn improved the water quality. Moreover, both

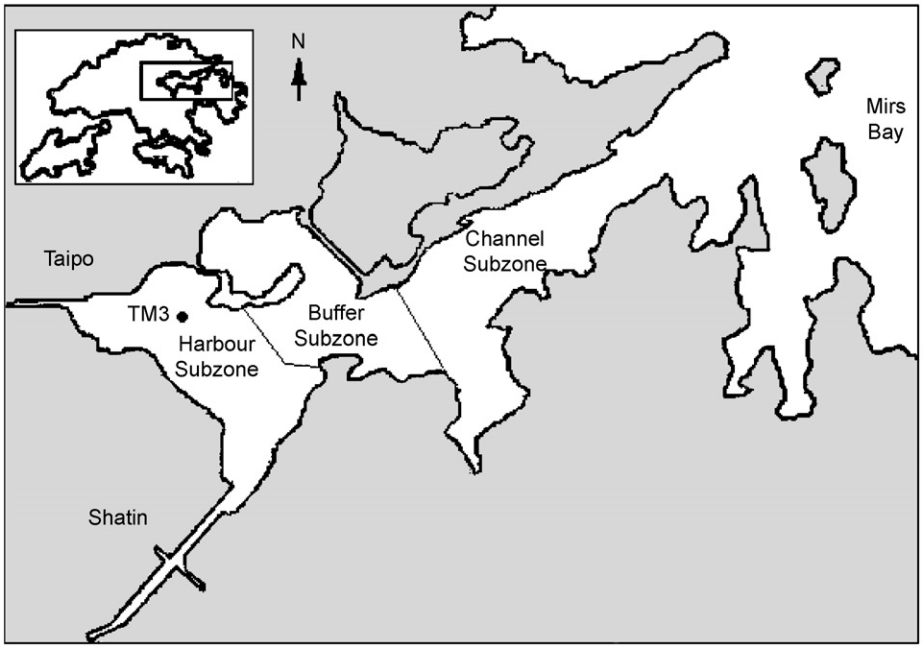

Fig. 1. Location of study site: Tolo Harbour (TM3 station).

field and process-based modelling efforts have been made to address the eutrophication phenomenon and DO dynamics in Tolo Harbour (e.g., Chan and Hodgkiss, 1987; Chau and Jin, 1998; Jin et al., 1998; Chau, 2004, 2005; Xu et al., 2004).

In this study, the depth-averaged monthly/biweekly water quality data, which were measured under a water quality monitoring program by the Environmental Protection Department of the Hong Kong government is employed. Amongst others, the data from the most weakly flushed monitoring station, TM3, are purposely chosen so as to minimize hydrodynamic effects. Preprocessing of the biweekly observed data is required to obtain the daily values by linear interpolation. Furthermore, daily meteorological data of wind speed (WS), solar radiation (SR) and rainfall supplied by the Hong Kong Observatory are employed. The data from 1988–1992 are used for training both data-driven models. For more details on the water quality data, readers are referred to Lee et al. (2003).

3.3. Interpolation effect

As mentioned above, the biweekly water quality data is linearly interpolated to get the daily values. We would like to point out that when interpolation is applied to produce time series from longer sampling frequency to a shorter time step, future observations are used to drive the predictions, as pointed out by Lee et al. (2003). An approach to avoid this use of "future data" in the predictions is to predict the algal dynamics with a lead-time equal to or greater than the sampling frequency of the original observations.

3.4. Input variables and time lags

The following nine variables are chosen as the initial set of input variables: chlorophyll-*a* (Chl-*a*) (μg/L); total inorganic nitrogen (TIN) (mg/L); PO_4 (mg/L); DO (mg/L); SD (m); water temperature, Temp (°C); daily rainfall, rain (mm); daily SR (MJ/m^2) and daily average WS (m/s). These variables are found to exhibit significant effects on the algal dynamics of Tolo Harbour according to several previous field and modelling studies in the weakly flushed embayment (Chau et al., 1996; Chau, 2002, 2005; Lee et al., 2003). Although it was found in previous studies that the variation of salinity in the seawater might result in reddish water, its value within this almost isolated water body is nearly a constant. Its variation mainly comes from the precipitation and hence daily rainfall value is included instead. The model output is Chl-*a*, which is an indicator of the algal biomass.

In Tolo Harbour and coastal waters of Hong Kong, algal blooms often occur with a life cycle of the order of 1–2 weeks. Lee et al. (2003) used auto-correlation of measured daily-averaged chlorophyll concentration from a monitoring station (Kat O) in north-eastern coastal waters of Hong Kong, to indicate that the algal dynamics is significantly correlated up to time lags of around 2 weeks. Based on this consideration of the ecological process, a 1-week lead-time is used for the predictions to identify the significant input variables. But, as this lead-time is less than the data sampling frequency of 2 weeks, the predictions would be affected by the interpolation effect, as discussed in the previous sub-section. Since a biweekly or more lead-time prediction would be free of this "interpolation effect", a biweekly prediction is also carried out. In the 1-week predictions, a time lag of 7–13 days is introduced for each of the input variables and in the biweekly predictions, a time lag of 14–20 days is introduced, whereas for both the predictions, Chl-*a* is predicted at time *t*.

For both the 1-week and biweekly predictions, each of the nine input variables comprise 7 time-lagged variables, amounting to a total of 9×7 ($= 63$) input variables. The significant input variables are to be selected from amongst them. Relationships between the Chl-*a* concentration at time *t* and the time-lagged input variables are developed by training ANN networks and evolving GP equations, details of which are presented in the following section.

4. Proposed techniques

In this section, we present the application of ANN and GP for input variable selection. Since these two data-driven models are themselves used for significant input variable selection, there is no need for going for any other analytical procedure for the same. Moreover, these models overcome some of the limitations associated with the commonly used selection techniques presented in Section 2. They can learn problems involving very non-linear and complex data and can identify correlated patterns between input data sets and corresponding target values. Both ANN and GP can take into account the interaction amongst variables and thus identify variables that may not be significant by itself, but are significant in combination with other variables. Thus, these data-driven models are ideally suited for identifying significant variables in ecological processes, which are known to be very complex and often non-linear.

4.1. Artificial neural networks (ANN)

An ANN is tailored to mimic natural neural networks in terms of computing paradigm (Haykin, 1999). Amongst many types of ANNs, the most widely used is the feed-forward neural network, multi-layer perceptron (MLP) or back-propagation network. The MLP is organized as layers of computing elements, known as neurons, which are connected between layers via weights. Apart from an input layer receiving inputs from the environment and an output layer generating the network's response, one or more intermediate hidden layers occur in between.

When triggered, each neuron will compute a response from the weighted sum of its inputs from neurons connected to it, on the basis of a predetermined activation function. In turn, its output will become the inputs of other

neurons located in the next layer. Whilst there are many commonly used activation functions, the sigmoid and the hyperbolic-tangent (tanh) functions are amongst the most popular. A back-propagation method is used in the training process to adjust the connection weights in the network in order to best match the network's response with the desired response. The optimization is executed by using an approximation to a gradient descent method (Haykin, 1999; Tarassenko, 1998).

4.2. Significant input variables based on ANN weights

A MLP neural network, which is trained using a back propagation algorithm with a momentum term, is employed. It comprises three layers, namely, an input, a hidden and an output layer. An interpretation of the connection weights from the input layer to the hidden layer of the trained network is undertaken. The inputs with the largest weight values denote the most significant input variables. An input significance measure S_n of the input variable n is defined as follows:

$$S_n = \sum_{j=1}^{H} |w_{jn}|, \tag{2}$$

where H is the number of hidden nodes, w_{jn} is the weight from input variable n to the hidden layer j. The summation of absolute values of weights is employed because some weight values may be positive and others negative. Belue and Bauer (1995) proposed using the sum of the squared weights as a saliency metric, but we found that such a measure magnifies the significance of inputs with larger weight values and diminishes the significance of those with smaller weights.

A point that needs attention is the over-training problem which might produce incorrect results. The over-training of a network depends on the optimal number of nodes in the hidden layer. In this study, the number of nodes in the hidden layer is determined by a trial and error procedure. The number is increased gradually from 3 to a maximum value dictated by the rule of thumb, i.e., the number of samples in the training set should at least be greater than the number of synaptic weights in the network (Tarassenko, 1998). The network training is terminated after a fixed number of epochs, which is also found by trial and error.

The neural network is first trained using the data from the test problem of Bernoulli's equation, which has 10 input variables and E is the dependent variable. The optimal number of nodes in the hidden layer was found to be 8. Trial and error method is also employed to determine the learning rate parameter and the momentum term. Their finally adopted values are 0.05 and 0.5, respectively. For both hidden and output layers, the hyperbolic-tangent function is adopted as the activation function. The network training is terminated after 1000 epochs. Fig. 2 shows the input significance of all the 10 inputs using the weights of the trained network. It is observed that the input significance corresponding to the variables z, p and v are significantly larger in magnitude than other spurious input variables correctly.

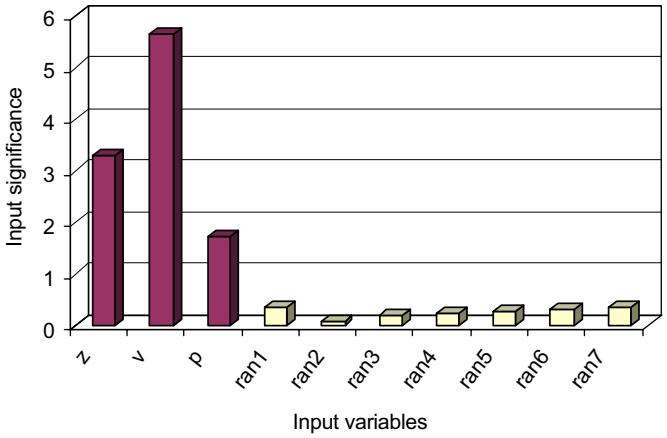

Fig. 2. Input significance, calculated using the weights of the trained ANN (using Eq. (2)) for the test problem.

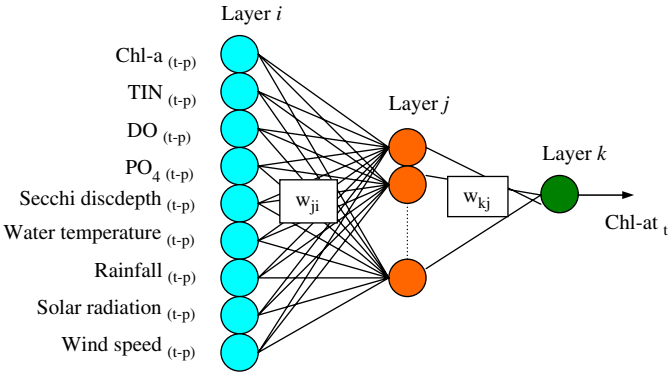

Fig. 3. General neural network for the prediction of algal blooms, $p = 7, \ldots, 13$ for 1-week predictions and $p = 14, \ldots, 20$ for biweekly predictions.

It is then applied to the real-world case study of predicting the algal biomass. Neural networks are trained for both 1-week and biweekly predictions. For both the predictions, there are 63 nodes in the input layer and the predicted Chl-*a* concentration is the only neuron in the output layer. Fig. 3 shows the network structure for both predictions. The optimal number of hidden nodes is found to be 6 and 4 for 1-week and biweekly predictions, respectively. The learning rate and the momentum term values are found to be 0.05 and 0.1, respectively and as before, the hyperbolic-tangent function was used as the activation function. The stopping criterion for the back propagation training is 500 epochs, which is obtained by trial and error.

Figs. 4 and 5 show the input significance of each input variable for 1-week and biweekly predictions, respectively. If all input variables have equal significance, then each input will have a significance of 1/63 (= 1.58%) of the total value of 137.23. We assume that those variables

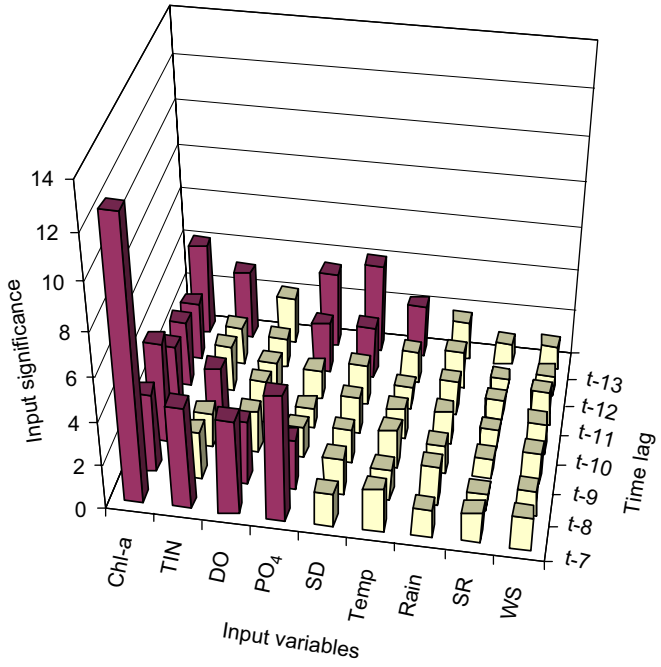

Fig. 4. Input significance, calculated using the weights of the trained ANN (using Eq. (2)) for 1-week predictions.

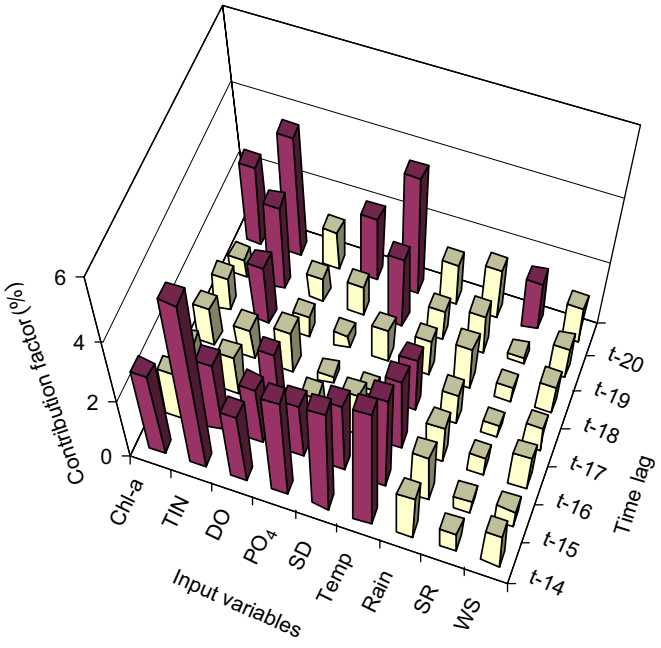

Fig. 5. Input significance, calculated using the weights of the trained ANN (using Eq. (2)) for biweekly predictions.

having an input significance greater than 1.58% are relatively more significant, and columns representing these variables are shaded dark in Figs. 4 and 5. It can be observed that Chl-*a* is the most significant input variable. In the 1-week ahead algal biomass prediction, the effect of Chl-*a* reduces with an increase in time lag and in the biweekly predictions, the significance of Chl-*a* is up to the time lag of (*t*-14). Apart from Chl-*a*, TIN, PO_4, DO and SD are found to be significant. For the biweekly predictions, Temp is significant and SR at (*t*-20) is slightly significant.

4.3. Genetic programming (GP)

GP (Koza, 1992) is an offshoot of genetic algorithm (Goldberg, 1989), which imitates biological evolution. The major distinction is that it operates on parse trees whilst genetic algorithm operates on bit strings. A function set and a terminal set together build up the parse tree. The function set comprises various operators (for example, multiplication, division, addition, subtraction and so on), whilst the terminal set comprises input variables and constants. The solution space of the available parse trees is constituted by all polynomials of any form over the input variables and constants. Fig. 6 shows an example of a parse tree for the model: $y = 2e^{-0.5x} - 0.6$.

GP starts with an initial generation of a population of random parse trees, calculation of their fitness in solving the problem domain and selection of the better parse trees for reproduction and evolution to a new generation. These processes iterate until a certain stopping criterion is met. The crossover operation takes place by randomly swapping sub-trees between the selected individuals (Babovic and Abbott, 1997; Babovic and Keijzer, 2000). An apparent advantage of using GP for the modelling process is its ability to produce models that are in the form of an interpretable equation (or formula). This is particularly useful for "data rich, theory poor" cases, since it can self evolve, through the genetic loop, a population of function trees so as to generate an "optimal" model that facilitates physical interpretation. Since these GP evolved equations or formulae relating input and output variables might shed physical insight into the ecological processes involved, they are used to identify the significant variables.

In this study, the GP software, GPKernel, developed by DHI Water and Environment is employed. It is a command line based tool to determine functions based on the available data. The optimal solution is obtained in

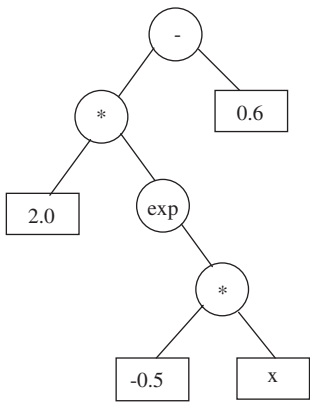

Fig. 6. Example of GP parse tree representing $2e^{-0.5x} - 0.6$.

Table 1
Values of control parameters used in GP runs

Parameter	Value
Maximum initial tree size	45
Maximum tree size	20
Crossover rate	1
Mutation rate	0.05
Population size	500
Elitism used	Yes

all GPKernel runs after it is run for 30 CPU minutes on a Pentium 4 1.4 GHz PC with 1021 MB RAM.

4.4. Significant input variables using GP equations

Table 1 shows the GPKernel parameters adopted for all GP runs for the selection of significant input variables. The value of "maximum initial tree size", representing the maximum size of the tree of the initial population, is constrained to 45. The value of "maximum tree size", representing the population of subsequent generations, is constrained to 20. It is found that, with these constraints, the evolved equations are easy to interpret and contain only 4–8 significant variables. All variables are preprocessed and normalized by dividing with their respective maximum values to avoid the occurrence of any dimensionally non-homogenous terms in the functional relationship of the evolved GP model.

Similar to the case of ANN, GP is first applied for input variable selection on the test problem with known dependence between the output and input variables. Using the simple math operators (+, −, *, /) as the function set, 10 different GP models are evolved using different initial seed for each GP run. Fig. 7 presents the number of times each of the input variables is selected in the 10 GP equations, which is the measure of its significance. It is observed that the evolved equations are mainly made up of 3 significant input variables of z, p and v, clearly demonstrating that GP can detect their significance in predicting the dependent variable, E.

Next, GP is applied on the problem of algal bloom prediction. GP models are evolved using 4 different function sets, which are presented in Table 2. For each of the 4 function sets, 20 GP equations are evolved with different initial seeds, resulting in 80 equations each, for 1-week and biweekly predictions. Small and simple function sets are employed since GP is very creative and effective at taking simple functions and evolving easily interpretable equations by combining them (Banzhaf et al., 1998). Figs. 8 and 9 show the total number of times of selection of the 63 input variables in the 80 evolved equations for 1-week and biweekly predictions, respectively. As in the ANN analysis, columns representing the significant variables are shaded dark, which are those with

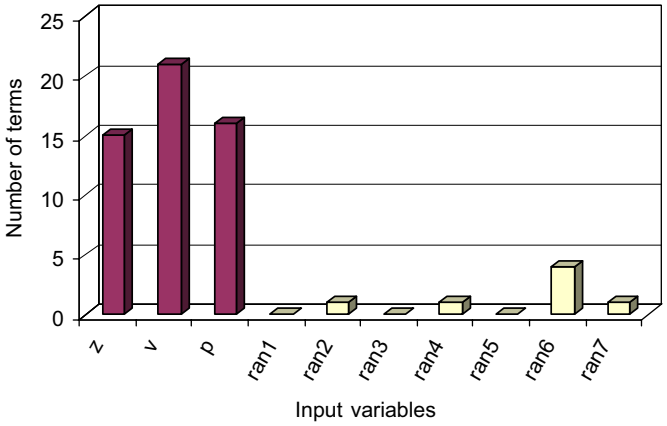

Fig. 7. Number of input variable selections in 10 GP runs for the test problem.

Table 2
Function sets used for the GP runs

Function set
+, −, ∗, /
+, −, ∗, /, e^x
+, −, ∗, /, x^2
+, −, ∗, /, x^y

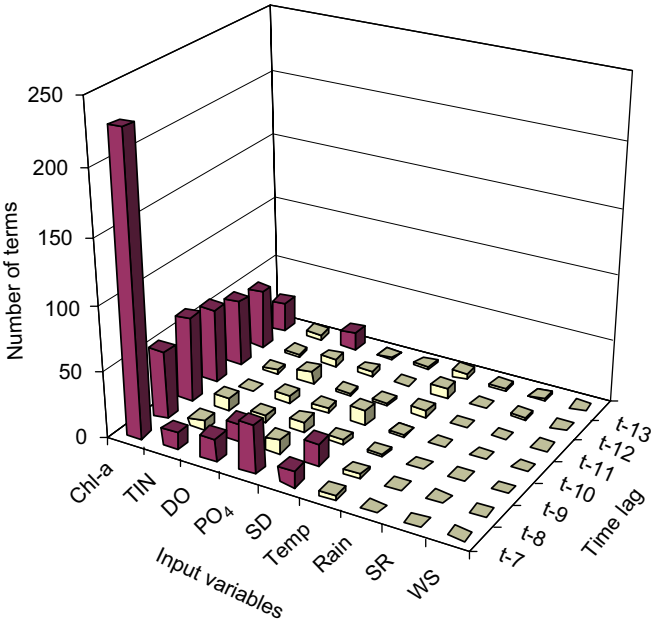

Fig. 8. Number of input variable selections in 80 GP runs for 1-week predictions.

number of terms more than 1.58% of the total number of terms in the 80 GP equations. For example, there are 790 terms in the 80 GP equations for 1-week predictions. In Fig. 8, those input variables having more than 13 (= 1.58% of 790) terms are assumed to be significant.

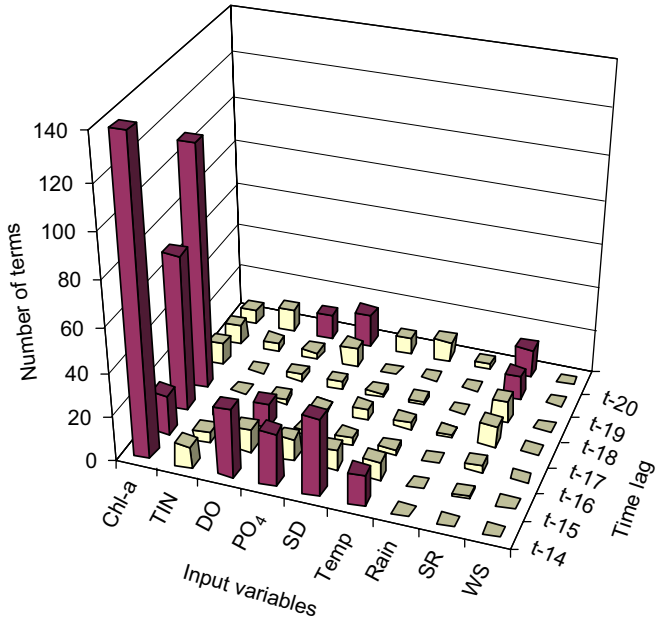

Fig. 9. Number of input variable selections in 80 GP runs for biweekly predictions.

From Fig. 8, it can be observed that the Chl-*a* value plays a key role in predicting its value 1-week ahead. Moreover, the effect of Chl-*a* reduces with the increase in the time lag, which aligns with the ANN analysis. PO_4, DO, SD and TIN are other significant variables. For biweekly predictions, the significance of Temp increases slightly and 2 terms of SR have slight significance.

5. Discussion on results

It is evident that Chl-*a* values are significant in predicting itself since both ML techniques gave the same conclusion. Chl-*a* at (*t*-7), with an input significance of 12.58 in the ANN weight analysis and with 229 terms out of a total of 790 in the GP equation analysis, is the most significant in algal biomass prediction 1-week ahead. The effect of Chl-*a* reduces with the increase in time lag and in biweekly predictions, Chl-*a* is significant up to a time lag of (*t*-14) in the ANN analysis and up to (*t*-17) in the GP analysis. This indicates *n* auto-regressive nature or "persistence" of the algal dynamics. This phenomenon is frequently exhibited in geophysical time series owing to inertia or carryover process in the physical system. In fact, modelling is a way in contributing to understanding the physical system as well as the process that builds persistence into the series. In this case, the auto-regressive nature of chlorophyll dynamics may be related to the long residence time in the semi-enclosed coastal waters. The average tidal current velocity is merely 0.04 m/s in the inner Harbour Subzone and 0.08 m/s in the outer Channel Subzone (EPD, 2005). It is justifiable that the landlocked nature of the estuary results in weak tidal flushing, which adds persistence into the algal dynamics. Furthermore, the two factors, i.e., the persistence of Chl-*a* reduces with the increase in time-lag and significant persistence is observed only up to a time lag of about (*t*-14), confirm our understanding that algal blooms in coastal waters of Hong Kong often occur with a life cycle of the order of 1–2 weeks.

Apart from Chl-*a*, both ML techniques suggest that the nutrients (PO_4 and TIN), DO and SD (to a lesser extent) are significant. It is reasonable that nutrients are significant, since the growth and reproduction of phytoplankton mainly rely on their availability. The significance of DO is also justifiable in sub-tropical coastal waters with mariculture activities, since it contributes to the production and respiration of algal organisms as well as to some chemical reactions.

In general, it is observed that the significant input variables from 1-week predictions are basically similar to those from biweekly predictions. The only exception to this is Temp and SR, which have slight significance in biweekly predictions. Thus, it can be concluded that the significant input variables from 1-week predictions are not completely driven by the interpolation effect. It appears that they exhibit at large cause–effect relationship between the time-lagged input variables and future algal biomass.

It should be mentioned that the results of this study regarding the significance of Chl-*a* in predicting itself are quite in contrast to several previous studies, which conventionally use a number of input variables (Jeong et al., 2003; Jeong et al., 2001; Wei et al., 2001; Recknagel et al., 1997; Yabunaka et al., 1997). The result that the use of previous data of algal biomass alone is good enough for future prediction might reduce the dependency on expensive equipment, such as automatic nutrient analyzers for ammonia and nitrate nitrogen, in algal bloom warning systems in coastal waters. This can attain significant cost savings.

6. Conclusion

This paper presents the prototype application of two distinct ML techniques (ANN and GP) for the selection of significant input variables, first using a test problem with known input–output dependence and then using data from a monitoring station in coastal waters of Hong Kong. It is evident that the identification of the key input variables are feasible with the interpretation of the trained ANN weights or of the evolved GP equations, which is basically in line with ecological reasoning. It is found that chlorophyll is the most significant variable in predicting algal blooms. The auto-regressive nature of the algal bloom dynamics in this semi-enclosed coastal water body is justifiable owing to the long flushing time. The result that the use of previous data of algal biomass alone is good enough for future prediction might reduce the dependency on expensive equipment in algal bloom warning systems in coastal waters.

Acknowledgements

The authors wish to thank DHI Water & Environment for providing the GP software, GPKernel. This research

was supported by the Research Grants Council of Hong Kong (PolyU5132/04E).

References

Anderson, D.M., 1994. Red tides. Scientific American 271, 62–68.

Babovic, V., Abbott, M.B., 1997. The evolution of equations from hydraulic data, Part I: Theory. Journal of Hydraulic Research 35 (3), 397–410.

Babovic, V., Keijzer, M., 2000. Genetic Programming as a model induction engine. Journal of Hydroinformatics 2 (1), 35–60.

Banzhaf, W, Nordin, P., Keller, R.E., Francone, F.D., 1998. Genetic Programming, an Introduction: on the Automatic Evolution of Computer Programs and its Applications. Morgan Kaufmann, San Francisco, CA, USA.

Belue, L.M., Bauer, K.W., 1995. Determining input features for multilayer perceptrons. Neurocomputing 7, 111–121.

Bobbin, J., Recknagel, F., 2001. Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms. Ecological Modelling 146, 253–262.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. Journal of Hydrology 301, 75–92.

Brosse, S., Giraudel, J.L., Lek, S., 2001. Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. Ecological Modelling 146, 159–166.

Chan, B.S.S., Hodgkiss, I.J., 1987. Phytoplankton productivity in Tolo Harbour. Asian Marine Biology 4, 79–90.

Chau, K.W., 2002. Field measurements of SOD and sediment nutrient fluxes in a land-locked embayment in Hong Kong. Advances in Environmental Research 6 (2), 135–142.

Chau, K.W., 2004. A three-dimensional eutrophication modeling in Tolo Harbour. Applied Mathematical Modelling 28 (9), 849–861.

Chau, K.W., 2005. An unsteady three-dimensional eutrophication model in Tolo Harbour, Hong Kong. Marine Pollution Bulletin 51 (8–12), 1078–1084.

Chau, K.W., Jin, H.S., 1998. Eutrophication model for a coastal bay in Hong Kong. Journal of Environmental Engineering, ASCE 124 (7), 628–638.

Chau, K.W., Jin, H.S., Sin, Y.S., 1996. A finite difference model of 2-d tidal flow in Tolo Harbour, Hong Kong. Applied Mathematical Modelling 20 (4), 321–328.

Chen, Q., Mynett, A.E., 2003. Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. Ecological Modelling 162, 55–67.

EPD, 2005. Marine Water Quality in Hong Kong: Results for 2004 from the Marine Monitoring Program of the Environmental Protection Department. Hong Kong Government, Hong Kong.

Goldberg, D.E., 1989. Genetic Algorithms for Search, Optimization and Machine Learning. Addison-Wesley Publishing Co., Reading, MA.

Haykin, S., 1999. Neural Networks: a Comprehensive Foundation, second ed. Upper Saddle River, New Jersey.

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

Jeong, K.S., Kim, D.K., Whigham, P., 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.

Jin, H.S., Egashira, S., Chau, K.W., 1998. Carbon to chlorophyll-a ratio in modeling long-term eutrophication phenomena. Water Science and Technology 38 (11), 227–235.

Karul, C., Soyupak, S., Cilesiz, A.F., Akbay, N., Germen, E., 2000. Case studies on the use of neural networks in eutrophication modelling. Ecological Modelling 134, 145–152.

Kirkpatrick, B., Fleming, L.E., Backer, L.C., Bean, J.A., Tamer, R., Kirkpatrick, G., Kane, T., Wanner, A., Dalpra, D., Reich, A., Baden, D.G., 2006. Environmental exposures to Florida red tides: effects on emergency room respiratory diagnoses admissions. Harmful Algae 5 (5), 526–533.

Koza, J., 1992. Genetic Programming: On the Programming of Computers by Natural Selection. MIT Press, Cambridge, MA.

Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural Network Modelling of Coastal Algal Blooms. Ecological Modelling 159, 179–201.

Lee, Y.S., 2006. Factors affecting outbreaks of high-density *Cochlodinium polykrikoides* red tides in the coastal seawaters around Yeosu and Tongyeong, Korea. Marine Pollution Bulletin 52 (10), 1249–1259.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the predication and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling & Software 15, 101–124.

Maier, H.R., Dandy, G.C., Burch, M.D., 1998. Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. Ecological Modelling 105, 257–272.

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

Morton, B. (Ed.), 1988. Editorial: Hong Kong's first marine disaster, Marine Pollution Bulletin 19, 299–300.

Muttil, N., Lee, J.H.W., 2005. Genetic programming for analysis and real-time prediction of coastal algal blooms. Ecological Modelling 189 (3-4), 363–376.

Petersen, W., Bertino, L., Callies, U., Zorita, E., 2001. Process identification by principal component analysis of river water-quality data. Ecological Modelling 138, 193–213.

Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecological Modelling 146, 303–310.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecological Modelling 96, 11–28.

Recknagel, F., Bobbin, J., Whigham, P., Wilson, H., 2002. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. Journal of Hydroinformatics 4 (2), 125–134.

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

Scardi, M., Harding, L.W., 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. Ecological Modelling 120, 213–223.

Tarassenko, L., 1998. A Guide to Neural Computing Applications. Arnold Publishers, London.

Tsujimoto, A., Nomura, R., Yasuhara, M., Yamazaki, H., Yoshikawa, S., 2006. Impact of eutrophication on shallow marine benthic foraminifers over the last 150 years in Osaka Bay, Japan. Marine Micropaleontology 60 (4), 258–268.

Wei, B., Sugiura, N., Maekawa, T., 2001. Use of artificial neural network in the prediction of algal blooms. Water Research 35 (8), 2022–2028.

Xu, F.L., Lam, K.C., Zhao, Z.Y., Zhan, W., Chen, Y.D., Tao, S., 2004. Marine coastal ecosystem health assessment: a case study of the Tolo Harbour, Hong Kong, China. Ecological Modelling 173, 355–370.

Yabunaka, K., Hosomi, M., Murakami, A., 1997. Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. Water Science and Technology 36 (5), 89–97.