International Journal of Environment and Pollution, Vol. 28, No. 3-4, 2006, pp. 223-238

NEURAL NETWORK AND GENETIC PROGRAMMING FOR MODELLING COASTAL ALGAL BLOOMS

Nitin Muttil[1] and Kwok-wing Chau[2, *]

[1]Research Associate, Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hung Hom, Hong Kong

[2]Associate Professor, Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hung Hom, Hong Kong

[*]Corresponding author. Tel.: +852- 2766 6014; Email: cekwchau@inet.polyu.edu.hk

Abstract

In the recent past, machine learning (ML) techniques such as artificial neural networks (ANN) have been increasingly used to model algal bloom dynamics. In the present paper, along with ANN, we select genetic programming (GP) for modelling and prediction of algal blooms in Tolo Harbour, Hong Kong. The study of the weights of the trained ANN and also the GP-evolved equations shows that they correctly identify the ecologically significant variables. Analysis of various ANN and GP scenarios indicates that good predictions of long-term trends in algal biomass can be obtained using only chlorophyll-a as input. The results indicate that the use of biweekly data can simulate long-term trends of algal biomass reasonably well, but it is not ideally suited to give short-term algal bloom predictions.

Keywords: Harmful algal blooms; Machine learning techniques; Artificial neural networks; Genetic programming; Water quality modelling; Hong Kong.

Biographical notes

Dr. Nitin Muttil is currently a Research Associate in the Department of Civil and Structural Engineering of The Hong Kong Polytechnic University. He obtained his Bachelor of Engineering degree from Engineering College, Jamia Millia Islamia, New Delhi in 1995. He then continued his studies in Indian Institute of Technology and National University of Singapore and obtained his Master of Technology and Doctor of Philosophy in 1997 and 2003, respectively. He has been employed by the University of Hong Kong as a Research Associate working on projects related to algal bloom prediction, monitoring, etc. and also on hydrological modeling.

Dr. Kwok-wing Chau is currently an Associate Professor in the Department of Civil and Structural Engineering of The Hong Kong Polytechnic University. He received both his BSc(Eng) and MPhil in Civil Engineering from the Department of Civil Engineering, The University of Hong Kong, Hong Kong. He received his PhD in Civil engineering from the Department of Civil Engineering, The University of Queensland, Australia. He is very active in undertaking research works and the scope of his research interest is very broad, covering numerical flow modeling, water quality modeling, hydrological modeling, knowledge-based system development and knowledge management.

1. Introduction

Harmful algal blooms (HABs) refer to the explosive growth and accumulation of harmful microscopic algae (phytoplankton). The well-known form of algal bloom – the red tide – has been widely reported and has become a serious environmental problem due to its negative impacts on human health and aquatic life. Harmful effects of red tides include beach closure, mariculture loss due to oxygen depletion or toxic algae, anoxia or shellfish poisoning (Anderson, 1994). In the past two decades there is an increasing trend in the occurrence of harmful algal blooms throughout the world. In particular, in April 1998, a devastating red tide resulted in the worst fish kill in Hong Kong's history - it destroyed over 80 percent (3400 tonnes) of cultured fish stock, with estimated loss of more than HK$312 million (Lee and Qu, 2004). Thus, a capability to analyze and predict the occurrence of algal blooms with an acceptable accuracy and lead-time would apparently be very beneficial to fisheries and environmental management.

Traditionally, models of phytoplankton dynamics are based on theories of the dependence of growth and decay factors on physical and biotic environmental variables (e.g. solar radiation, nutrients, flushing) – expressed mathematically and incorporated in advective diffusion equations in a water quality model. Such deterministic models are normally referred to as process-based models. A limitation of the process-based models is the significant uncertainty of many kinetic coefficients adopted in the water quality model. In the recent past, with the development of artificial intelligence (AI) techniques and easy availability of computer-aided analysis, machine learning (ML) techniques have been extensively used in ecological modelling (Recknagel, 2001).

ML techniques are ideally suited to model the algal dynamics since such models can be set up rapidly and is known to be effective in handling dynamic, non-linear and noisy data, especially when underlying physical relationships are not fully understood, or when the required input data needed to drive the process-based models are not available. In the present study, we employ artificial neural networks (ANN) and genetic programming (GP) for analysis of water quality data from Tolo Harbour. In the following sections, we first present an introduction to the key principles of ANN and GP, followed by its application to modeling of algal dynamics.

2. Machine Learning (ML) modelling techniques

During the past two decades, researchers have at their disposal, many fourth generation ecological models, ranging from numerical, mathematical and statistical methods to techniques based on AI. ML is an area of computer science, a sub-area of AI concentrating on the theoretical foundations (Solomatine, 2002). A ML technique is an algorithm that estimates hitherto unknown mapping (or dependency) between a system's inputs and its outputs from the available data (Mitchell, 1997). Once a dependency is discovered, it can be used to predict (or effectively deduce) the system's future outputs from the known input values. The growing development of computer-aided analysis, which is easily accessible to all researchers, has facilitated the application of various ML techniques in ecological modelling (Recknagel, 2001). These techniques include ANN (Recknagel et al., 1997, 2002; Yabunaka et al., 1997; Maier et al., 1998; Scardi and Harding, 1999; Jeong et al., 2001; Scardi, 2001; Wei et al., 2001; Chau and Cheng, 2002; Lee et al., 2003; Chau, 2004a&b; Cheng et al., 2005; Chau, 2005), fuzzy and neuro-fuzzy techniques (Maier et al., 2001; Chen and Mynett, 2003), evolutionary based techniques (Bobbin and Recknagel, 2001; Recknagel et al., 2002; Jeong et al., 2003; Muttil et al., 2004), etc. Although most of these studies are applied to freshwater environments (i.e., limnological or riverine systems), a few have been

to saltwater eutrophic areas (Scardi and Harding, 1999; Scardi, 2001; Lee et al., 2003; Muttil et al., 2004).

Out of the several possible ML techniques, we considered ANN and GP. We selected ANN because it is the most widely used method in water resources variable modelling (Maier and Dandy, 2000), and GP because it has an advantage in that it generates equations or formulae relating input and output variables, which might shed physical insight into the ecological processes involved. Further, Recknagel (2001) reported that ANN and genetic algorithms (of which GP is an extension) currently appear to be most innovative for ecological modelling. We present brief introductions to these two techniques in the following sub-sections.

2.1. Artificial neural network (ANN)

An ANN is a computing paradigm designed to mimic natural neural networks (Haykin, 1999). Although there are many types of ANNs, by far the most widely used is the feed-forward neural networks or the multi-layer perceptron, which is organized as layers of computing elements (called neurons) connected via weights between layers. Typically, there is an input layer that receives inputs from the environment, an output layer that produces the network's response, and one or more intermediate hidden layers.

The action of each neuron is to compute a response from the weighted sum of its inputs from neurons connected to it, using a predetermined activation function. The output is routed to become the inputs of other neurons in the following layer. Many activation functions are in use, with the most popular being the sigmoid and the hyperbolic-tangent (*tanh*) functions.

The feed-forward network is also known as the (error) back-propagation network because of the method used in its training. Training is a process of adjusting the connection weights in the network so that the network's response best matches the desired response. Although this can be addressed as an optimization method, the back-propagation method avoids this costly exercise by using an approximation to a gradient descent method. More details on ANN can be found in, for example, Haykin (1999).

2.2. Genetic programming (GP)

The basic search strategy behind genetic programming (Koza, 1992) is a genetic algorithm (Goldberg, 1989), which imitates biological evolution. It differs from this traditional genetic algorithm in that it typically operates on parse trees instead of bit strings. A parse tree is build from a terminal set (the variables in the problem) and a function set. Suppose the terminal set consists of a single variable x and some constants, and the function set consist of the operators for multiplication, division, addition and subtraction, the space of available parse trees constitute all polynomials of any form over x and the constants. An example can be found in Figure 1, where the parse tree for the model: $y = -0.2x + 0.3$ is shown.

As a genetic algorithm, genetic programming proceeds by initially generating a population of random parse trees, calculate their fitness - a measure of how well they solve the given problem - and subsequently selects the better parse trees for reproduction and recombination to form a new population. This process of selection and reproduction iterates until some stopping criterion is satisfied. The recombination takes place by crossover: randomly swapping sub-trees of the parse trees between selected individuals. A more comprehensive presentation of GP can be found in Babovic and Abbott (1997) and Babovic and Keijzer (2000).

The main advantage of GP for the modeling process is its ability to produce models that build an understandable structure, i.e., a formula or equation. Thus, for "data rich, theory

poor" instances, GP may offer advantages over other techniques since GP can self modify, through the genetic loop, a population of function trees in order to finally generate an "optimal" and physically interpretable model.

3. Modelling approach and application

The application of ANN and GP for real-time algal bloom prediction at Tolo Harbour in the northeastern coastal waters of Hong Kong is presented. In this section, we give an account of the nature of the data used and details of the modelling.

3.1. The study site and data

Tolo Harbour is a semi-enclosed bay in the northeastern coastal waters of Hong Kong (Figure 2). It is connected to the open sea at Mirs Bay; in general the water quality improves from the more enclosed and densely populated inner “Harbour Subzone” towards the better flushed outer “Channel Subzone”.

The nutrient enrichment in the harbour due to municipal and livestock waste discharges has been a major environmental concern over the past two decades. The organic loads are derived from the two major treatment plants at Shatin and Taipo (Figure 2), non-point sources from runoff and direct rainfall, and waste from mariculture. The eutrophication has resulted in frequent algal blooms and red tides, particularly in the weakly flushed tidal inlets inshore, with occasional massive fish kills due to severe dissolved oxygen depletion or toxic red tides. Various studies (Morton, 1988; Xu et al., 2004) have shown that the ecosystem health state of the Tolo Harbour had been progressively deteriorating since the early 1970s up to late 1980s. During this period, the nutrient enrichment in the harbour due to urbanization, industrialization and livestock rearing had caused serious stresses to the marine coastal ecosystem. The situation became worst in the late 1980s with frequent occurrences of red tides and associated fish kills. Morton (1988) referred to the Tolo Harbour as “Hong Kong’s First Marine Disaster” and pointed out that the inner harbour was effectively dead. In the late 1980s, Tolo harbour had reached a critical stage, which resulted in the development of an integrated action plan, Tolo Harbour Action Plan (THAP) by the Hong Kong Government. The implementation of THAP in 1988 had achieved significant effectiveness on the reduction of pollutant loading and on the improvement of the water quality. A number of field and process-based modeling studies on eutrophication and dissolved oxygen dynamics of this harbour have been reported (e.g., Chan and Hodgkiss, 1987; Chau and Jin, 1998; Lee and Arega, 1999; Chau, 2004c; Xu et al., 2004).

The monthly/biweekly water quality data, collected as part of the routine water quality monitoring program of the Hong Kong government’s Environmental Protection Department, is used as a basis for the modelling. In order to isolate the ecological process from the hydrodynamic effects as much as possible, the data from the most weakly flushed monitoring station, TM3 (Figure 2), are used. The ecological variables are all depth-averaged. The biweekly observed data is linearly interpolated to get the daily values. In addition, daily meteorological data (thus no interpolation required) of wind speed, solar radiation and rainfall recorded by the Hong Kong Observatory is used. The data from 1988 – 1992 are used for training and data from 1993 – 1996 are used for testing the models. More details on the water quality data can be found elsewhere (Lee and Arega, 1999; Lee et al., 2003).

3.2. Objective function and model performance criterion

The objective function used for the evolution of the GP models is the minimization of Root-Mean-Square-Error (RMSE) of the prediction over the training period. The performance of the predictions for both ANN and GP is evaluated by two goodness-of-fit

measures: the RMSE and correlation coefficient (CC). Time series plots are used for visual comparison.

3.3. Selection of model inputs and output

Based on previous field and modelling studies in weakly flushed embayment in Tolo Harbour (Chau et al., 1996; Lee and Arega, 1999; Lee at al., 2003), the following nine input variables are selected to be the most influential on the algal dynamics of Tolo Harbour: chlorophyll-a, Chl-a (μg/L); total inorganic nitrogen, TIN (mg/L); phosphorus, PO_4 (mg/L); dissolved oxygen, DO (mg/L); secchi-disc depth, SD (m); water temperature, Temp ($^{\circ}$C); daily rainfall, Rain (mm); daily solar radiation, SR (MJ/m^2) and daily average wind speed, WS (m/s).

Chlorophyll-a, an indicator of the algal biomass, is taken as the model output. Based on considerations of the ecological process at Tolo Harbour and practical constraints of data collection, a one-week prediction of algal blooms is set as the modelling target. For each of the input variables, a time lag of 7-13 days is introduced. The time lag starts from 7 because of the requirement of one-week lead-time of the prediction.

4. Selection of significant input variables

In any ML technique, the selection of appropriate model input variables is extremely important. The choice of input variables is generally based on *a priori* knowledge of causal variables and physical/ecological insight into the problem. Moreover, the use of lagged input variables also leads to better predictions in a dynamical system. Maier and Dandy (2000) have reviewed 43 international journal papers, which used ANN for modelling and forecasting of water resources variables, published between 1992 and 1998. They concluded that in many papers, the modelling process is carried out incorrectly and one of the main areas of concern included arbitrary selection of model inputs, which we address in this section.

ANN networks are trained and GP equations are evolved to develop a relationship between the chlorophyll-a concentration at time *t* and the nine input variables with the time lag of 7-13 days. Thus, for each of the nine input variables, we have seven time-lagged variables, making a total of 9 × 7 (= 63) input variables, out of which the significant input variables are to be selected.

4.1. Significant variables based on ANN weights

In this study, a fully connected feed forward MLP neural network trained with a back propagation algorithm with momentum term was used for forecasting the algal blooms. A single hidden layer was considered in this study, and thus the resulting MLP neural network structure consisted of three layers: an input layer, a hidden layer and an output layer.

The input layer had 63 nodes, which were the input variables determined previously. Determination of the optimal number of nodes in the hidden layer is an important factor, which affects the performance of the trained network. In general, networks with fewer hidden nodes are preferable as they usually have better generalization capability and less over fitting problem. The use of computational time is also less with fewer numbers of nodes. But if the number of nodes is not enough to capture the underlying behaviour of the data then the performance ability of the network decreases. In this study, a trial and error procedure was carried out by gradually varying the number of nodes in the hidden layer from 2 to 10 and the optimal number was found to be 6. The output layer of the networks contained only one neuron, which is the chlorophyll-a concentration that is to be predicted.

The learning rate parameter and the momentum term were also determined by trial and error. The optimal values for the learning rate parameter and the momentum term for the

neural network runs were found to be 0.05 and 0.1 respectively. The hyperbolic-tangent function (*tanh*) was used as the activation function for both the hidden and the output layer.

The above described network structure (see Figure 3) is used to predict the algal biomass with a one-week lead-time. All simulations are carried out in accordance with the model formulation and training details described above. The back propagation training was stopped after 500 epochs in all the simulations. This number was selected by trial and error, as it was observed that 500 epochs were enough to train the network and there was no over-training.

In order to select the significant input variables, an analysis of the network weights was done. In the trained network, the connection weights along the paths from the input nodes to the hidden nodes demonstrate the relative predictive importance of the independent variables. Now, in our network, there are 63 input nodes, and in order to measure the importance of any one variable in predicting the network's output, relative to the other input variables in the same network, we define a term called the contribution factor. The contribution factor of the n^{th} variable, CF_n is defined as below:

$$CF_n = \frac{\sum_{j=1}^{nH} ABS(w_{jn})}{\sum_{i=1}^{nV}\sum_{j=1}^{nH} ABS(w_{ji})} \times 100 \qquad (1)$$

where nH denotes the number of hidden nodes, nV is the number of input variables, w_{ji} are the weights from input layer i to the hidden layer j (see Figure 3) and ABS is the absolute function.

Using Eqn. (1), the contribution factor of each of the 63 input variables is calculated, and this is presented in Table 1. The sum of the contribution factors of all the 63 input variables should add up to 100%, which can be seen in Table 1. The contribution factor is so defined that the higher its value for a variable, the more that variable is contributing to the prediction. From this analysis, which is based on the weights of the trained neural network, Chl-a at (t-7), with a contribution factor of 9.17%, is the most significant in predicting the one-week ahead algal biomass. In Table 1, the variables with a contribution factor greater than 2.00% are shaded and are considered relatively more significant. The other variables suggested to be significant are PO_4, TIN, DO and SD.

4.2. Significant variable using GP equations

An attempt is made to use the evolutionary search capabilities of GP for selecting the significant input variables. The GPKernel parameters used for all the GP runs are presented in Table 2. The parameters "Maximum initial tree size" and "Maximum tree size" indicate the maximum size of the tree of the initial population and of the population from subsequent generations, respectively. The values of "Maximum initial tree size" and "Maximum tree size" are constrained to 45 and 20, respectively. This restriction is necessary since GP has the tendency to evolve uncontrollably large trees, if the tree size is not limited. A maximum tree size of 20 evolves simple expressions that are easy to interpret. It is observed that when "Maximum tree size" is restricted to 20, the evolved equation contains only 4 to 8 variables. Thus, we are allowing the evolutionary process to select only about 4 to 8 variables from the total of 63 variables that are used as input.

In order to avoid a functional relationship comprising of dimensionally non-homogenous terms within the evolved GP model, all the variables are normalized or non-dimensionalized initially. They are non-dimensionalized by dividing all the variables by their respective maximum values.

The GP tool software used in this study is GPKernel, which is developed by DHI Water & Environment and is available at http://www.d2k.dk. GPKernel is a command line based tool for finding functions on data. All computations were performed on a Pentium PC with 1021 MB RAM; for each adopted function and variable set, GPKernel is run for 30 CPU minutes to obtain the optimal solution.

For the GP runs, 4 different function sets are used (Table 3), and for each function set, 20 GP equations are evolved with different initial seeds. Thus, 80 GP equations were evolved for the one-week predictions. As seen in Table 3, small and simple function sets are used. This is so because GP is very creative at taking simple functions and creating what it need by combining them (Banzhaf et al., 1998). In fact, GP often ignores the more sophisticated functions in favor of the simple ones during evolution. A simple function set also leads to evolution of simple GP models, which are easy to interpret. Now, since GP has the ability to select input variables that contribute beneficially to the model, it is expected that the GP evolved equations would contain the most significant of the 63 input variables. In other words, a measure of the significance of a variable is the number of times the variable is selected. For each of the 63 input variables, the number of times it is selected in each of the 80 evolved equations are summed up and are presented in Table 4. The significant variables as indicated by GP are shaded in this table. These shaded variables are those whose number of terms are more than 2% of the total number of terms in the 80 GP equations. In this particular analysis, the total number of terms in the 80 equations is 790 and variables that contribute more than 2% of 790 (i.e., 16 or more terms) are shaded. It can be seen from this table that the Chl-a value itself plays a significant role in predicting its one-week ahead value. The other variables indicated as significant are PO_4, DO, TIN and SD.

4.3. Discussion on suggested significant variables

From the analysis presented above using both ML techniques, ANN and GP, it is clearly observed that Chl-a values during the past 1 – 2 weeks are significant, with Chl-a at (t-7) being the most significant in predicting itself. This suggests an auto-regressive nature for the algal dynamics. Geophysical time series frequently exhibit such auto-regressive nature or "persistence" because of inertia or carryover process in the physical system. Modelling can contribute to understanding the physical system by revealing something about the process that builds persistence into the series. In the present study, the auto-regressive nature (or persistence) of chlorophyll dynamics suggested by the ML techniques may be related to the long flushing time (residence time) in the semi-enclosed coastal water. This was also suggested in a recent ANN study of algal dynamics in Tolo Harbour (Lee et al., 2003). The tidal currents are very small with the average current velocity being 0.04 m/s in the inner Harbour Subzone and 0.08 m/s in the outer Channel Subzone (EPD, 1999). Thus, the landlocked nature of the estuary leads to weak tidal flushing; the flushing times in inner Harbour Subzone have been estimated to be in the order of 1 month (Lee and Arega, 1999).

Both the ML techniques also suggest that the nutrients (PO_4 and TIN) along with DO and SD (to a lesser extent) to be significant, although these variables are not as highly significant as Chl-a. The importance of nutrients is understandable, since the growth and reproduction of phytoplankton is dependent on the availability of various nutrients. In subtropical coastal waters with mariculture activities, the DO level is also intimately related to algal growth dynamics. DO is important for the respiration of these organisms and for some chemical reactions. Moreover, Xu et al. (2004) pointed out that the general trends for PO_4, TIN and DO increased from the early 1970s to the later 1980s or the early 1990s; and then decreased from the early 1990s to the later 1990s (the decrease can be attributed to the implementation of THAP in 1988, as discussed in Section 3.1). In the present study, the

significance of PO_4, TIN and DO seem to be understandable, since our training data is from 1988 – 1992, a period when there was an increasing trend in these variables.

As mentioned earlier, the biweekly water quality data is linearly interpolated to get the daily values, which is used in this study. It should be noted that when interpolation is applied to produce time series from longer sampling frequency to short time step, future observations are used to drive the predictions (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 the time interval of the original observation. In the present case, a biweekly or more lead-time prediction would be free of this "interpolation effect". Thus, a biweekly prediction is carried out using the same data and the significant input variables are identified using ANN and GP, as was done using the one-week predictions. In the biweekly predictions, a time lag of 14 - 20 days is introduced for each of the input variables and Chl-a is predicted at time *t*. The significant variables are presented in Table 5 and Table 6 for the ANN and GP predictions, respectively. As before, the shaded variables are those whose contribution is more than 2% in terms of ANN weights and number of terms in GP equations, respectively. From these tables, it is observed that for biweekly predictions, the significant input variables are the same as that for the one-week predictions, except for Temp, which is also indicated as significant in the biweekly analysis. Since the biweekly predictions are free of the interpolation effect, we can conclude that the significant input variables from one-week predictions are reasonable. The one-week predictions, although to some extent being driven by interpolation of data, still have physical knowledge in them and it seems that they are based on cause-effect relationship between the time-lagged input variables and future algal biomass.

In the next section, the one-week ahead predictions using both ANN and GP are presented. The variables suggested as significant, namely, Chl-a, PO_4, TIN, DO and SD are used as input for the predictions. For each of these significant input variables, 7 – 13 days of time-lagged inputs are used.

5. The one-week ahead predictions

5.1. Predictions using ANN

In this section, different neural network runs are carried out with different combinations of the significant input variables. As before, the concentration of Chl-a is predicted with a one-week lead-time. These neural network simulations are carried out in accordance with the model formulation and training details described for the input variable selection in Section 4.1.

Figure 4 shows the comparison of the predicted Chl-a with observed values for both the training and testing periods, for the best ANN prediction, which is with only time-lagged Chl-a as inputs. A blow-up of the same predictions for a period (May 1993 to September 1994) from the testing data is presented in Figure 5.

5.2. Predictions using GP

Using as input, the significant variables indicated in the previous section, GP models are evolved for Chl-a prediction for one-week lead-time. Five different GP runs are conducted with each of the 4 function sets presented in Table 3. The parameters for the GP runs are the same as those used for the input variable selection (presented in Table 2), with the exception of "Maximum tree size", which is increased to 45, because a bigger tree size would give lower RMSE.

The best GP model (with minimum RMSE) was evolved with the function set consisting of the basic math operators (+, -, *, /), which confirms our using simple function sets (see Table 3) based on the understanding that GP requires simple function sets to create

models with best predictive capability. Figure 4 also shows the comparison of the predicted chlorophyll-a with observed values for both the training and testing periods, again for the best GP predictions (using only time-lagged Chl-a as inputs). Figure 5 also presents a blow-up of the predictions, similar to that presented for the ANN prediction.

5.3. Discussion on prediction results

The goodness-of-fit measures for both the ANN and GP predictions for the different scenarios are presented in Table 7. From the results presented in this table, it is clearly observed for both ANN and GP predictions that the accuracy of the predictions worsen with increase in the number of input variables. In fact, the best predictions are when only time-lagged Chl-a is used as input. This significance of Chl-a in predicting itself is in contrast to various previous studies, which have tended to include many more input variables. For example, Jeong et al. (2003) used 19 input variables, but concluded that only 4 variables are required to predict cyanobacteria biovolume with a high accuracy; Jeong et al. (2001) used 16 variables to predict time series changes of algal biomass with a time-delayed recurrent neural network; Wei et al. (2001) included eight environmental factors to estimate the evolution of four dominant phytoplankton genera using ANN; Recknagel et al. (1997) used 10 input variables to a feed-forward ANN for the prediction of algal bloom in lakes in Japan, Finland and Australia; Yabunaka et al. (1997) used 10 environmental parameters as ANN inputs to predict the concentration of five freshwater phytoplankton species. This observation of algal biomass alone being sufficient for predicting itself may also throw doubt on the advantage of deploying expensive equipment (like automatic nutrient analyzers for ammonia and nitrate nitrogen) in algal bloom warning systems in coastal waters.

From the time series plots of the predictions (Figure 4), it is seen that the prediction can track the algal dynamics with reasonable accuracy. But on closer examination (the blow-up of the predictions is presented in Figure 5), a phase error of about one-week can be noted in the predictions. Thus, it can be concluded that the use of this biweekly data may not be ideally suited for short-term predictions of algal blooms. The use of higher frequency data should be a solution for improving the accuracy of the predictions. We would also point out that biweekly predictions using the significant input variables were also carried out, but those results are not presented here as the phase error in them is much more than that for the one-week predictions.

6. Conclusions

This study presents the analysis of algal dynamics data from a coastal monitoring station using two ML techniques, ANN and GP. The interpretation of ANN weights and GP equations appear to be able to identify key input variables that are in accordance with ecological reasoning. The study reveals that chlorophyll-a itself is enough as input for predicting itself, suggesting an auto-regressive nature of the algal dynamics in the semi-enclosed coastal waters. The results for the prediction of chlorophyll-a suggest that the use of biweekly data can simulate long-term trends of algal biomass reasonably well, but it is not ideally suited to give short-term algal bloom predictions. The use of higher frequency data is suggested for improving the short-term predictions.

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

1. Anderson, D. M., 1994. Red tides, Scientific American, 271: 62-68.
2. Babovic, V. and Keijzer, M., 2000. Genetic Programming as a model induction engine, J. Hydroinform., 2 (1): 35-60.
3. Babovic, V. and Abbott, M. B., 1997. The evolution of equations from hydraulic data, part I: Theory, J. Hydraul. Res., 35 (3): 397-410.
4. Banzhaf, W, Nordin, P., Keller, R.E. and Francone, F.D., 1998. Genetic programming, an Introduction: On the Automatic Evolution of Computer Programs and its Applications, Morgan Kaufmann, San Francisco, CA, USA.
5. Bobbin, J. and Recknagel, F, 2001. Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms. Ecol. Model. 146: 253-262.
6. Chan, B.S.S. and Hodgkiss, I.J., 1987. Phytoplankton productivity in Tolo Harbour. Asian Mar. Biol. 4: 79-90.
7. Chau, K.W. and Cheng, C.T., 2002. Real-time prediction of water stage with artificial neural network approach. Lecture Notes in Artificial Intelligence, 2557: 715-715.
8. Chau, K.W., 2004a. River stage forecasting with particle swarm optimization. Lecture Notes in Computer Science. 3029: 1166-1173.
9. Chau, K.W., 2004b. Rainfall-runoff correlation with particle swarm optimization algorithm. Lecture Notes in Computer Science. 3174: 970-975.
10. Chau, K.W., 2004c. A three-dimensional eutrophication modeling in Tolo Harbour. Applied Mathematical Modelling. 28 (9): 849-861.
11. Chau, K.W., 2005. A split-step PSO algorithm in prediction of water quality pollution. Lecture Notes in Computer Science, 3498: 1034-1039.
12. Chau, K.W. and Jin, H.S., 1998. Eutrophication model for a coastal bay in Hong Kong. Journal of Environmental Engineering, ASCE. 124 (7): 628-638.
13. Chau, K.W., Jin, H.S. and 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.
14. Chen, Q. and Mynett, A. E., 2003. Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. Ecol. Model. 162: 55-67.
15. Cheng, C.T., Chau, K.W., Sun, Y.G. and Lin, J.Yi., 2005. Long-term prediction of discharges in Manwan Reservoir using artificial neural network models. Lecture Notes in Computer Science. 3498: 1040-1045.
16. EPD, 1999. Marine Water Quality in Hong Kong: Results for 1998 from the Marine Monitoring Program of the Environmental Protection Department, Hong Kong Government, Hong Kong.
17. Goldberg, D.E., 1989. Genetic Algorithms for Search, Optimization and Machine Learning. Addison-Wesley Publishing Co., Reading, Mass.
18. Haykin, S., 1999. Neural Networks: a Comprehensive Foundation. Second Edition, Upper Saddle River, New Jersey.
19. Jeong, K. S., Kim, D. K., Whigham, P. and Joo, G. J., 2003. Modelling Microcystis aeruginosa bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. Ecol. Model. 161: 67-78.
20. Jeong, K.S., Joo, G.J., Kim, H.W., Ha, K. and Recknagel, F., 2001. Prediction and elucidation of algal dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. Ecol. Model. 146: 115-129.
21. Koza, J., 1992. Genetic Programming: On the Programming of Computers by Natural Selection. MIT Press, Cambridge, MA.
22. Lee, J.H.W. and Qu, B., 2004. Hydrodynamic tracking of the massive spring 1998 red tide in Hong Kong. J. Environ. Eng., ASCE. 130 (5): 535-550.

23. Lee, J. H. W., Huang, Y., Dickman, M. and Jayawardena, A. W., 2003. Neural Network Modelling of Coastal Algal Blooms. Ecol. Model. 159: 179-201.
24. Lee, J.H.W. and Arega, F., 1999. Eutrophication dynamics of Tolo Harbour. Hong Kong. Mar. Pollut. Bull. 39 (1-12): 187-192.
25. Maier, H.R., Sayed, T. and Lence, B.J., 2001. Forecasting cyanobacterium Anabaena spp. in the River Murray, South Australia, using B-spline neurofuzzy models. Ecol. Model. 146: 85-96.
26. Maier, H. R. and Dandy, G. C., 2000. Neural networks for the predication and forecasting of water resources variables: a review of modelling issues and applications. Env. Modelling & Software. 15:101-124.
27. 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. Ecol. Model. 105: 257-272.
28. Mitchell, T. M., 1997. Machine Learning. McGraw-Hill, New York.
29. Morton, B., 1988. Editorial: Hong Kong's first marine disaster. Mar. Pollut. Bull. 19: 299-300.
30. Muttil, N., Lee, J.H.W. and Jayawardena, A.W., 2004. Real-time Prediction of Coastal Algal Blooms using Genetic Programming. Proc. 6th Int. Conference on Hydroinformatics. (Eds. Liong, Phoon & Babovic), June 21-24, 2004, Singapore, 890-897.
31. Recknagel, F., French, M., Harkonen, P. and Yabunaka, K., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96: 11-28.
32. Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecol. Model. 146: 303-310.
33. Recknagel, F., Bobbin, J., Whigham, P. and Wilson, H., 2002. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. J. Hydroinformatics. 4 (2): 125-134.
34. Scardi, M., 2001. Advances in neural network modelling of phytoplankton primary production. Ecol. Model. 146: 33-45.
35. Scardi, M. and Harding, L.W., 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. Ecol. Model. 120: 213-223.
36. Solomatine, D., 2002. Data-driven modelling: paradigm, methods, experiences. Proc. 5th Int. Conference on Hydroinformatics., July 1-5, 2002, Cardiff, UK, 757-763.
37. Wei, B., Sugiura, N. and Maekawa, T., 2001. Use of artificial neural network in the prediction of algal blooms. Wat. Res. 35 (8): 2022-2028.
38. Xu, F. L., Lam, K. C., Zhao, Z. Y., Zhan, W., Chen, Y. D. and Tao, S., 2004. Marine coastal ecosystem health assessment: a case study of the Tolo Harbour, Hong Kong, China. Ecol. Model. 173: 355-370.
39. Yabunaka, K., Hosomi, M. and Murakami, A., 1997. Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. Wat. Sci. Tech. 36 (5): 89-97.

List of Tables:

Table 1 : Contribution factor, calculated using the weights of the trained ANN (see Eqn. (1)) for one-week predictions

Table 2 : Values of control parameters used in GP runs

Table 3 : Function sets used for the GP runs

Table 4 : Number of input variable selections in 80 GP runs for one-week predictions

Table 5 : Contribution factor, calculated using the weights of the trained ANN (see Eqn. (1)) for biweekly predictions

Table 6 : Number of input variable selections in 80 GP runs for biweekly predictions

Table 7 : Goodness-of-fit measures for the ANN and GP one-week predictions

Table 1. Contribution factor, calculated using the weights of the trained ANN (see Eqn. (1)) for one-week predictions

Input variables	Contribution factors of the input variables (%) *							Sum
	(t-7)	(t-8)	(t-9)	(t-10)	(t-11)	(t-12)	(t-13)	
Chl-a	9.17	2.62	3.35	2.27	2.18	1.89	3.05	24.53
TIN	3.37	1.54	1.18	1.68	1.50	1.23	2.27	12.77
DO	3.12	2.06	1.37	1.41	1.13	1.05	1.55	11.68
PO4	4.11	1.62	1.04	0.72	1.04	1.69	2.55	12.77
SD	1.08	1.20	1.11	1.22	1.37	1.70	2.98	10.66
Temp	1.42	1.01	1.28	1.00	0.78	1.02	1.70	8.20
Rain	0.95	1.31	0.97	1.08	1.15	1.18	1.31	7.95
SR	1.00	0.56	0.69	0.65	0.68	0.42	0.69	4.70
WS	1.09	0.89	1.06	1.05	1.13	0.73	0.79	6.74
Sum of contribution factors of all variables =								100

* Shaded variables have a contribution factor greater than 2%

Table 2. 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

Table 3. Function sets used for the GP runs

Function set
+, -, *, /
+, -, *, /, e^x
+, -, *, /, x^2
+, -, *, /, x^y

Table 4. Number of input variable selections in 80 GP runs for one-week predictions

Input variables	Number of terms of time-lagged input variables *							Total terms
	(t-7)	(t-8)	(t-9)	(t-10)	(t-11)	(t-12)	(t-13)	
Chl-a	229	53	65	58	51	46	23	525
TIN	14	7	10	1	3	2	5	42
DO	18	14	5	5	10	6	14	72
PO4	38	10	9	4	2	4	1	68
SD	13	17	4	11	2	1	2	50
Temp	4	3	2	1	5	7	5	27
Rain	0	0	0	0	0	1	1	2
SR	0	0	1	0	0	2	1	4
WS	0	0	0	0	0	0	0	0
Total number of terms in 80 GP models =								790

* Shaded variables contribute to more than 2% of the 790 terms in the 80 GP models

Table 5. Contribution factor, calculated using the weights of the trained ANN (see Eqn. (1)) for biweekly predictions

Input variables	Contribution factors of the input variables (%) *							Sum
	(t-14)	(t-15)	(t-16)	(t-17)	(t-18)	(t-19)	(t-20)	
Chl-a	2.79	1.54	1.23	1.24	1.07	0.54	2.46	10.87
TIN	5.53	2.34	1.27	1.02	1.91	2.68	3.71	18.46
DO	2.42	1.96	1.86	1.39	0.63	0.80	1.19	10.26
PO4	3.35	1.92	0.85	0.28	0.42	0.93	2.04	9.78
SD	3.50	2.38	1.38	0.20	1.06	2.27	3.66	14.45
Temp	4.04	3.09	2.39	1.78	1.18	0.94	1.36	14.79
Rain	1.89	1.55	1.23	1.08	1.38	1.26	1.57	9.96
SR	0.70	0.45	0.59	0.40	0.47	0.27	1.59	4.46
WS	1.19	0.63	1.14	0.83	1.00	1.02	1.15	6.95
Sum of contribution factors of all variables =								100

* Shaded variables have a contribution factor greater than 2%

Table 6. Number of input variable selections in 80 GP runs for biweekly predictions

Input variables	Number of terms of time-lagged input variables *							Total terms
	(t-14)	(t-15)	(t-16)	(t-17)	(t-18)	(t-19)	(t-20)	
Chl-a	140	18	71	111	13	9	6	368
TIN	10	5	4	1	0	4	10	34
DO	31	10	11	2	3	3	11	71
PO4	24	10	2	1	3	9	15	64
SD	35	9	3	5	2	0	7	61
Temp	14	8	2	3	2	1	9	39
Rain	0	0	0	1	0	0	2	3
SR	0	1	3	11	10	11	13	49
WS	0	0	0	1	0	1	0	2
Total number of terms in 80 GP models =								691

* Shaded variables contribute to more than 2% of the 691 terms in the 80 GP models

Table 7. Goodness-of-fit measures for the ANN and GP one-week predictions

Input variables *	Training		Testing	
	RMSE	CC	RMSE	CC
For ANN modelling				
Chl-a, PO_4, TIN, DO, SD	1.87	0.96	4.02	0.82
Chl-a, PO_4, TIN, DO	2.16	0.94	4.49	0.86
Chl-a, PO_4, TIN	2.14	0.95	3.00	0.91
Chl-a, PO_4	2.42	0.94	2.76	0.92
Chl-a	2.55	0.93	2.24	0.95
For GP modelling				
Chl-a, PO_4, DO, SD	2.67	0.92	2.54	0.93
Chl-a, PO_4, DO	2.55	0.92	2.50	0.93
Chl-a, PO_4	2.37	0.93	2.32	0.94
Chl-a	2.55	0.93	1.99	0.95

* All input variables are of 7-13 days time lag

List of Figures:

Figure 1 : Example of GP parse tree representing -0.2x + 0.3

Figure 2 : Location of study site: Tolo Harbour (TM3 station)

Figure 3 General neural network for the prediction of algal blooms, p = 7,…,13 for one-week predictions and p = 14,…,20 for biweekly predictions

Figure 4 : One-week predictions of chlorophyll-a at TM3 using ANN and GP (for training and testing periods)

Figure 5 : Blow-up of the one-week ANN and GP predictions indicating the phase errors (from part of testing period)

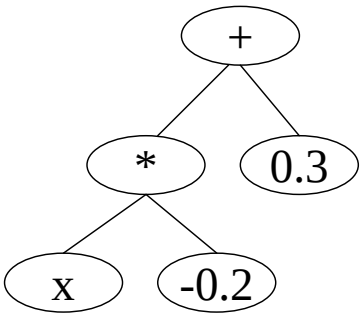

Figure 1. Example of GP parse tree representing -0.2x + 0.3

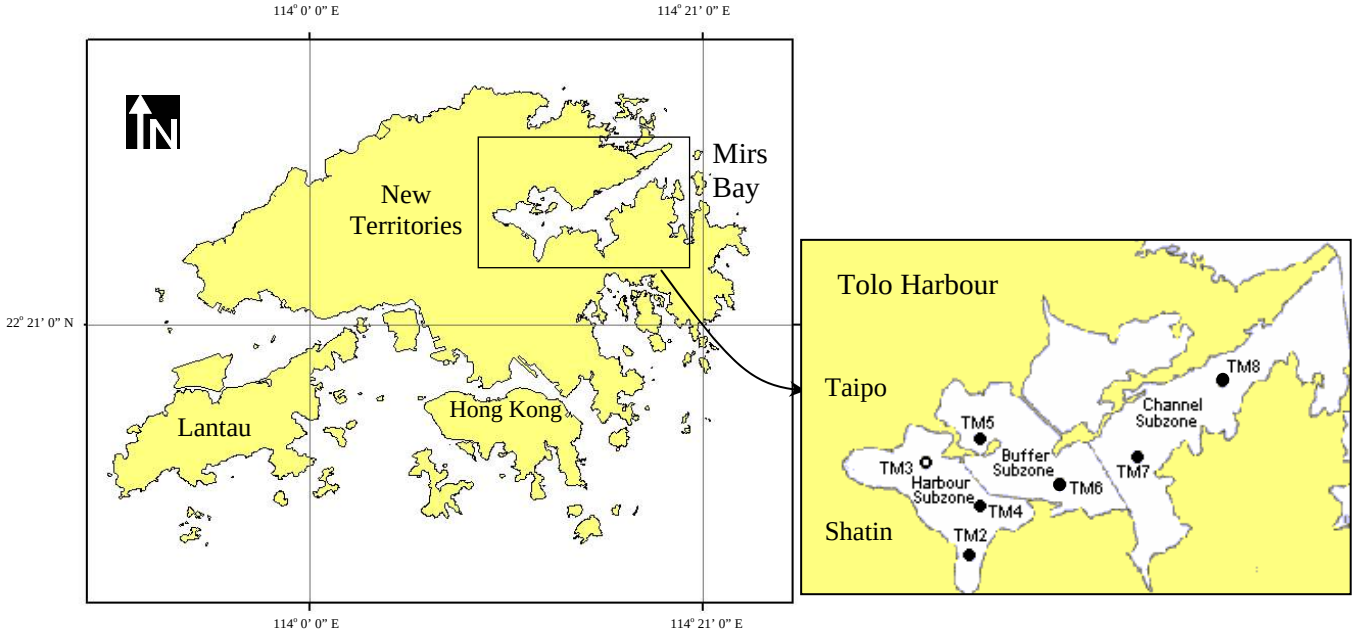

Figure 2. Location of study site: Tolo Harbour (TM3 station)

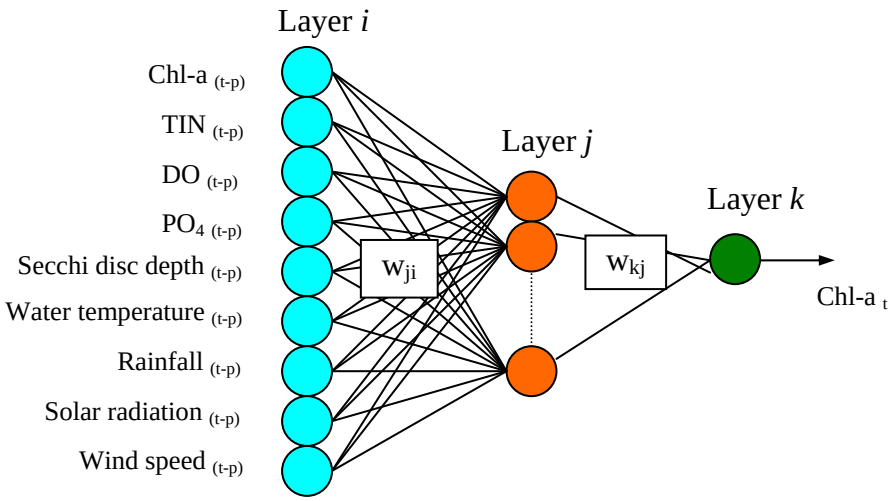

Figure 3. General neural network for the prediction of algal blooms, p = 7,…,13 for one-week predictions and p = 14,…,20 for biweekly predictions

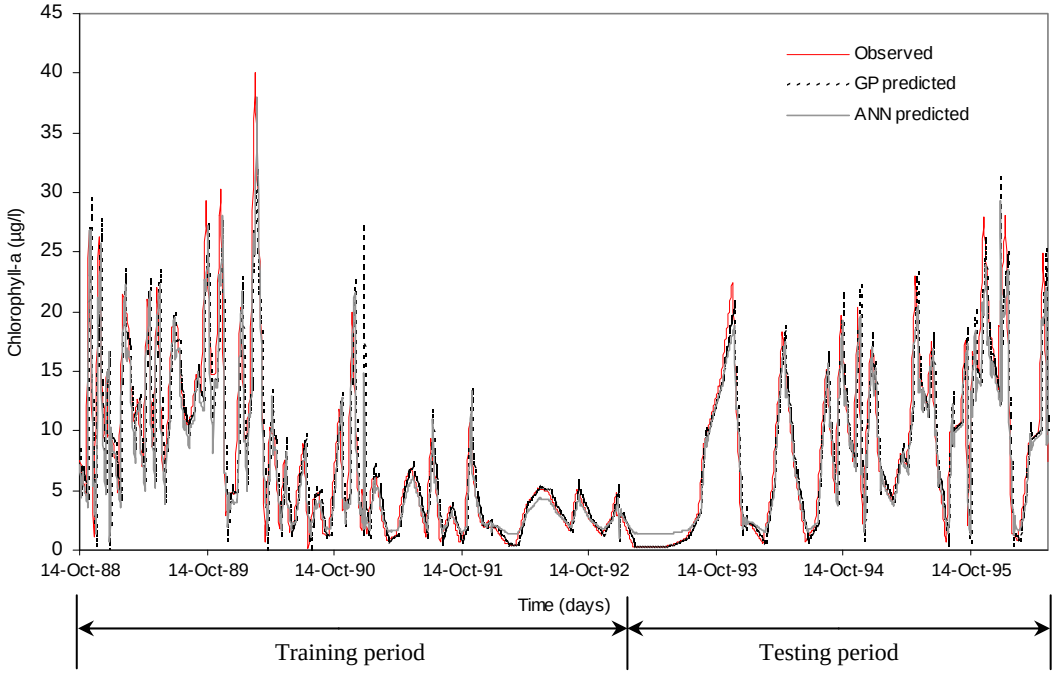

Figure 4. One-week predictions of chlorophyll-a at TM3 using ANN and GP (for training and testing periods)

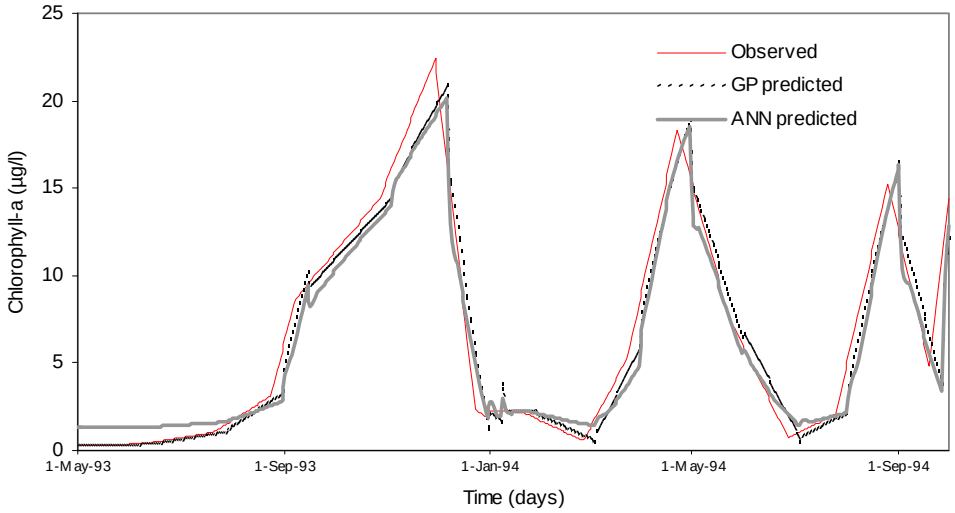

Figure 5. Blow-up of the one-week ANN and GP predictions indicating the phase errors (from part of testing period)