Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

PREDICTION OF ALGAL BLOOM USING GENETIC PROGRAMMING

C. Sivapragasam[1], Nitin Muttil[2], S. Muthukumar[1] and V. M. Arun

[1]Department of Civil Engineering, Kalasalingam University, Krishnankoil, Tamil Nadu, India.
[2]School of Engineering and Science, Victoria University, Melbourne, VIC, Australia

ABSTRACT

In this study an attempt is made to mathematically model and predict algal blooms in Tolo Harbour (Hong Kong) using Genetic Programming (GP). Chlorophyll plays a vital role and is taken as a measure of algal bloom biomass and 8 other variables are taken as input for its prediction. It is observed that GP evolves multiple models with almost same values of errors – of – measure. Previous studies on GP modeling primarily focused on comparing the GP results with actual values. In contrast, in this study, the main aim is to propose a systematic procedure for identifying the most appropriate GP model from a list of feasible models (with almost same error-of measure) using physical understanding of the process aided by data interpretation. The study of the GP-evolved equations shows that they correctly identify the ecologically significant variables. Analysis of final GP evolved mathematical model indicates that of the 8 variables assumed to affect the algal bloom, the most significant effect is due to chlorophyll, total inorganic nitrogen and dissolved oxygen, as far as one week prediction is concerned. For higher lead prediction (biweekly), secchi disc depth and temperature appears as significant variables in addition to chlorophyll.

Key Words: Genetic Programming, Mathematical Modeling, Harmful Algal Bloom

INTRODUCTION

The Algal bloom phenomenon (particularly the red tide) has been widely reported and has become a serious environmental problem owing to its adverse influence on aquatic life as well as human health. The need for better understanding of the the Harmful Algal Bloom (HAB)

dynamics and the complex ecological processes involved is felt clearly over the years (Lee and Qu, 2004). In spite of extensive research already undertaken, the causality and dynamics of algal blooms are not well-understood and the prediction of algal blooms remains a very difficult problem, owing to the extremely complicated ecological dynamics. Thus, it is very desirable to obtain mathematical models that can give some insight into the physics of the process, while having the capability to predict the occurrence of algal blooms with an acceptable accuracy and lead time.

Conventionally, phytoplankton dynamics have been carried out using the process- based models by incorporating physical and biotic environmental variables in water quality model. This, however, is reported to suffer from the uncertainty of kinetic coefficients used in such models. In the recent past, many studies have reported successful application of data-driven Artificial Intelligence based techniques, particularly Artificial Neural Network (ANN) and Genetic Programming (GP). For example, as early as in 1997, Recknagel et al (1997) demonstrated that ANN is capable of modeling the non-linear and complex algal growth phenomena. Lee et al (2003) found that the algal concentration in the samples from Tolo Harbour is dependent primarily on their antecedent concentrations in the immediate past weeks, and the result was supported by interpretation of the neural networks' weights.

Coad et al (2005) observed that with antecedent chlorophyll-a, feed forward neural network with logistic function is able to predict the future chlorophyll-a reasonably well, indicating the sufficiency of historical values of chlorophyll-a in its future modeling. Muttil and Chau (2006) reported that both ANN and GP correctly identified the ecologically significant variables, and that long term algal growth can be predicted using only chlorophyll-a as input. They also observed that when 'Maximum initial tree size' and 'Maximum tree size' are restricted to 45 and 20 respectively, the evolved equation contains only 4–8 variables and thus the equation is easy to interpret. Whigham and Recknagel (1999) compared the GP evolved equations with ANN models to demonstrate the applicability of GP to non–linear processes in natural systems such as freshwater systems. They concluded that the transparent nature of GP solutions may allow

inference about underlying process to be made and also highlighted issues with scaling data for machine learning and the difficulty involved with producing understable models.

Bobbin and Recknagel (1999) discuss the application of Genetic Algorithms (GAs) for the construction of rule based models and found that GA can be used to extract and develop rules from water quality time series and that can be used for prediction and elucidation of timing and magnitudes of algal bloom events. Sam et al (2005) studied the monitoring of algal bloom in Jakarta Bay, Indonesia using Terra–Aqua MODIS (Moderate resolution Imaging spectro–radiometer) satellite data and found that a combination of high resolution of ALOS image and high repetitions of MODIS image will make the algal bloom phenomenon clearer.

Chau and Muttil (2007) studied the ecological and related water quality data from different periods from several monitoring stations in Tolo Harbour, Hong Kong by descriptive data mining techniques and the results from box plots reveals the spatial, temporal patterns, which in turn helps to find out the stations which are most susceptible to eutrophication, its nutrient source and control measures. Recknagel et al (1997) did a study on predictive potential of phytoplankton models by ANN and compared with other models such as AD HOC inductive models and found that predictive accuracy improved with increased event and time resolution of data. Recknagel et al (2002) compare the potential of ANN and GA in terms of forecasting and understanding of algal blooms in Lake Kasumigaura, Japan and found that models evolved by GA performs better than ANN models and provide more transparency for physical explanation as well. Lui et al. (2007) studied modeling of algal bloom with vector autoregressive model with exogenous variables in Hong Kong.

Most of the reported works on HAB studies by ANN and GP focuses primarily on prediction of algal growth and compare the potential of each other. The present study focuses on developing GP based mathematical models with an emphasis on the procedure to select the best model which can ensure the best prediction performance for the extreme values. In addition, whenever possible, an attempt is also made to interpret or at least get some insight into the algal bloom process with the GP evolved models. Mathematical models are developed for weekly and

biweekly forecast for the water quality data from Tolo Harbour, Hong Kong. The next section briefly describes the GP approach. This is followed by the study area description. Then model development and analysis of the results are presented. Finally conclusions are arrived at.

GENETIC PROGRAMMING

GP is very similar to using a GA, being an evolutionary algorithm based on Darwinian theories of natural selection and survival of the fittest. However, GP operates on parse trees, rather than on bit strings as in a GA, to approximate the equation (in symbolic form) that best describes how the output relates to the input variables. The algorithm considers an initial population of randomly generated programs (equations), derived from the random combination of input variables, random numbers and functions. The functions can include arithmetic operators (plus, minus, multiply, divide), mathematical functions (sin, cos, exp, log), logical/comparison functions (OR/AND) etc., which have to be appropriately chosen based on some understanding of the process. This population of potential solutions is then subjected to an evolutionary process and the 'fitness' (a measure of how well they solve the problem) of the evolved programs are evaluated; individual programs that best fit the data are then selected from the initial population.

The programs that best fit are selected to exchange part of the information between them to produce better programs through 'crossover' and 'mutation', as used in GAs (to mimic the natural reproduction process). Here, exchanging the parts of best programs with each other is called crossover, copied exactly into the next generation is called reproduction and randomly changing programs to create new programs is called mutation (Koza 1992). The user must decide a number of GP parameters before applying the algorithm to model the data, such as population size, number of generations, crossover and mutation probability, etc. The programs that fitted the data less well are discarded. This evolution process is repeated over successive generations and is driven towards finding symbolic expressions describing the data, which can be scientifically interpreted to derive knowledge about the process being modeled.

Tree Based Genetic Programming

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

The primitives of GP, the function and terminal nodes, must be assembled into a structure before they may be executed. Three main types of structure exist: tree, linear and graph. In this study, the method utilized for modeling algal bloom is a tree-based genetic programming (TGP) approach. TGP was introduced by Koza as an extension of the GA, in which programs are represented as tree structures and expressed in the functional programming language, LISP (Koza, 1992). A comprehensive description of GP is beyond the scope of this paper. Details on GP can be obtained from Koza (1992) and from Babovic and Keijzer (2000) for explanations from a water resources perspective.

In this study, GPKernel, developed by DHI Water and Environment (Babovic and Keijzer, 2000) is used for implementing GP. GPKernel is a command line based tool for finding functions on data. For a detailed explanation of various features of GPKernel, the reader is referred to Babovic and Keijzer (2000).

CASE STUDY DESCRIPTION

Tolo Harbour is a semi-enclosed bay in the Northeastern coastal waters of Hong Kong (Figure 1). It is connected to the open sea at Mirs Bay. 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 1), non-point sources from runoff and direct rainfall. 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 have shown that the ecosystem health state of the Tolo Harbour had been progressively deteriorating since the early 1970s up to late 1980s. Tolo Harbour had reached a critical stage in the late eighties, which resulted in the development of an integrated action plan, the Tolo Harbour Action Plan (THAP), by the Hong Kong Government. The implementation of THAP in 1988 achieved significant effectiveness on the reduction of pollutant loading and on the improvement of the water quality.

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

A number of field- and process-based modeling studies on eutrophication and dissolved oxygen dynamics of this harbour have been reported 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 (EPD), are used in this study as a basis for the modeling. 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 1), are used. The ecological variables are all depth-averaged. The biweekly observed data are 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 are used. It should be noted that whenever a coarse sampling frequency (biweekly in this study) is used, the value of any input variable at an interval shorter than the monitoring interval must be interpolated from data. This inevitably introduces some of the actual observations (which we seek to predict) into the modelling process. To avoid this problem, one approach can be that the algal dynamics is predicted with a lead-time of the minimum monitoring interval of the original observations, which is biweekly in the data used in this study. We have done the modeling for biweekly prediction (t + 14 days) in this study. In addition, we also preferred to include 7 day lead prediction (given the limitations in the data we had) in order to study the model performance.

Modeling algal biomass basically involves estimating the chlorophyll at any particular future time, by giving the chlorophyll, chy (μg/L) along with other input variables such as total inorganic nitrogen, TIN (mg/L); phosphorus, PO_4 (mg/L); dissolved oxygen, DO (mg/L); secchi-disc depth, SD (m); water temperature, Temperature (°C); daily rainfall, Rain (mm); daily solar radiation, SR (MJ/m^2) and daily average wind speed, MWS (m/s) at time (t) influence the eutrophication.
The data used in this study is available from 1988 to 1996.

PERFORMANCE MEASURES

The forecast performance is evaluated using two goodness-of-fit measures, the root-mean-square - error (RMSE) and the correlation coefficient (CC) as defined below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(X_m)_i - (X_s)_i\right]^2} \qquad (1$$

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

$$CC = \sum_{i=1}^{n} \frac{\left[\left(X_m\right)_i - \left(\overline{X_m}\right)\right]\left[\left(X_s\right)_i - \left(\overline{X_s}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left[\left(X_m\right)_i - \left(\overline{X_m}\right)_i\right]^2}\sqrt{\sum_{i=1}^{n}\left[\left(X_m\right)_i - \left(\overline{X_s}\right)_i\right]^2}} \qquad (2)$$

where X is any variable that is being forecasted; the subscripts m and s represent the measured and simulated values; the average value of the associated variable is represented with a 'bar' above it; and n is the total number of training records.

MODEL DEVELOPMENT

The chlorophyll prediction at 7 day lead period and 14 day lead period can be functionally represented as

$$chy_{t+7} = f\left(TIN_t, chy_t, PO_t, DO_t, SD_t, Temp_t, SR_t, MWS_t\right) \qquad (3)$$

$$chy_{t+14} = f\left(TIN_t, chy_t, PO_t, DO_t, SD_t, Temp_t, SR_t, MWS_t, chy_{t+7}\right) \qquad (4)$$

where the variable holds the meaning as described earlier. Chy_{t+14} is the 14 day ahead prediction, Chy_{t+7} is the 7 day ahead prediction and Chy_t is the current value of chlorophyll. GP is run for 17 experiments with its parameters, namely, crossover rate, mutation rate, number of generations, population size, etc, which are optimized by trial and error, are presented in Table 1. The functional set consists of simple arithmetic functions, trigonometric functions, logarithmic function and exponential functions. The procedure adopted in this study for selecting the best model from the list of various models evolved by GP is described as below:

(a) Identification of the maximum and minimum value of Chy_t in the time series.

(b) Separate the time series into two categories viz., (i) for training the GP in a specified range of Chy_t and (ii) Validating GP outside this range i.e. those containing Chy_t close to the low and high extremes.

(c) The GP is trained with those input vectors which contain intermediate values of Chy_t.

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

(d) The GP evolved models for each experiment is validated separately for both intermediate values of Chy_t and values which are close to the lower and higher extremes. This is to test how various models perform for Chy_t values extrapolated outside the training range of Chy_t.

(e) From the models obtained in step (d) above, the best models with almost equal error measures are selected. These are then analyzed to see their meaningfulness in explaining the physics of the process.

(f) The best model obtained from (e) above is subjected to sensitivity analysis to identify the significance of the input variables.

The above procedure is applied to 7 lead day prediction and the GP evolved models are listed in Table 2. For biweekly prediction, two different input vectors are considered:

(i) Direct Prediction: Input vectors as adopted in Eq (4) i.e. with the known values of input variables at time 't', direct prediction is aimed for t+14.

(ii) Sequential Prediction: Input vectors include the predicted Chy_{t+7} (from the best model) as a new variable in addition to the input variables used in Eq (4).

The procedure for selecting the best model, as described above is also adopted to select the best model for predicting Chy_{t+14}.

RESULTS AND DISCUSSIONS

GP model was run with 65% of the data as training set and 20% as test set and the remaining as production set. Table 2 lists 10 best models from various runs of GP for predicting chlorophyll at 7-lead day. The maximum and minimum value of Chyt is 0.2 µg/l and 40 µg/l. In this study, GP is trained for chlorophyll value in the range 4 to 20 µg/l. The models evolved are validated for chlorophyll value within this range and outside this range i.e. less than $0.2 < Chyt < 4$ and $20 < Chyt < 40$. The RMSE values are shown in Table 2 for the different models. Surely it will be interesting to know what will happen if the training is done on the extremes and testing on medium range. However, in this study, the number of data for extreme range is too few compared

to the medium range. This will obviously result in a poor training and therefore poor output. Therefore, this is not used in this study. The suggested approach is superior to traditional temporally based selection approach because traditional methods do not guarantee good performance on extremes (as observed in many other hydrological applications such as flood predictions etc).

It can be observed that all the models give almost same RMSE for chlorophyll value inside the range 4 to 20 μg/l. It can be inferred that for a given process (for sample data), more than one model is feasible. This can be explained through Figure 2, where one can clearly see how various linear and non-linear models can be fit for a given range of data value. However, the predictions from different models deviate when extrapolated either below or above the training range. So, in order to choose the best model from the ones listed in Table 2, the RMSE outside training range has to be compared. It is surely possible in some cases that training based on a range of 4 till 20, and validating based on the rest of the data will provide the best models, that have better generalization ability based on the data under consideration, and not necessarily because they are physically relevant. However, given the complexity of algal bloom modeling, the chances of the physical relevance are much more enhanced when the data is validated outside the training range (as explained from Figure 2). Most of the earlier approaches on GP training have focused on the entire range, but this does not guarantee the performance of the model, particularly when we are interested in future prediction (where the possibility of values falling outside range are surely high). Since our main intention is to predict the algal growth at a larger lead time, this approach seems more meaningful than the traditionally adopted approach.

It can be seen in Table 2 that model 9 and model 10 have RMSE of about 2.58 μg/l and all other models have higher value. Therefore, these two models can be expected to describe the process better given the limitations on data availability. Due to linear interpolation of biweekly data to obtain daily data, the performance of naïve model (model 1 in Table 2 evolved by GP itself) appears as good as some of the other models evolved by GP for 7 lead day prediction. Therefore, this model is not considered in the analysis.

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

Since model 1, 3 and 6 have almost equal RMSE and not too much different from models 9 and 10, it can be argued that as long as all inputs considered are physically relevant, excluding some models on physics basis cannot be justified. However, our contention here is that at closer observation, a particular model seems to explain the process better than the other (though the inputs to both the models are physically relevant) under some given conditions (in this case, for example, presence of TIN seems to be more meaningful, than a model without this input). Therefore, we chose to focus on more physically relevant models.

A in depth analysis of model 9 and 10 are discussed here. Although they have approximately same RMSE value, the variables presumed to describe the process are not the same. For example, model 9 is governed by Chy_t, DO_t and TIN_t, whereas model 10 is governed only by Chy_t and SD_t. The presence of a nutrient variable (TIN) is more meaningful because the growth and reproduction of phytoplankton are dependent on the availability of various nutrients. Similarly, DO level is also very important for algal growth as it enhances certain chemical reactions as well as required for the respiration of the organisms. Under these considerations, it is more appropriate to choose model 9 as the best evolved model for 7 lead day chlorophyll prediction:

$$chy_{t+7} = \sqrt{\{\sqrt{[DO_t + \sqrt{(chy_t + TIN_t)} + \log(DO_t)] * chy_t}\} * chy_t} \qquad (5)$$

However, it is important to note that Eq (5) looks very difficult to physically interpret because of presence of triple-square-root function. One reason for this could be that both power and square root functions are used in GP modeling (besides lack of sufficient data coupled with multi-variable inputs), which played a role to make the square root function have a dominating effect. However, we still prefer to choose this model to the other because of revealing vital information such as the presence of TIN and DO as influencing variables.

In order to investigate the importance of different variables in Eq (5) above, a sensitivity test is carried out and the results are summarized in Table 3. As seen from the table, chlorophyll is most sensitive as even 5% error in its estimation affects the prediction considerably. Similar conclusions were arrived by Coad et al (2005) on the importance of chlorophyll. So, chlorophyll has to be very accurately estimated. Though the other two variables are more stable even with

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

15% error in their estimation, we prefer to keep these variables in the model because of their physical relevance in the algal growth.

Figure 3 shows the plot of prediction obtained with model 9 in Table 2 (presented as Eq (5) above) and the actual Chy_{t+7}. As seen from the figure, the model captures most of the peaks without phase error with a few exceptions especially in the under prediction of the maximum chlorophyll level (40.1 μg/l). On closer examination, for lower range of chlorophyll values, some phase lag is observed.

The biweekly prediction of chlorophyll (Chy_{t+14}) is also attempted at. Initially, the input vector to GP is kept same as that used for weekly prediction i.e. Eq (3). The best ten GP evolved models are listed in Table 4 with RMSE value for both inside and outside the training range of chlorophyll. The results show that the best predicted model has a RMSE as high as 5.32 μg/l.

$$chy_{t+14} = \log(\log(DO_t)) + \left[\left(SD_t * \log(DO_t) \right) + \left(\frac{chy_t}{SD_t} \right) \right] \qquad (6)$$

The 14 day prediction model as obtained above shows that Chy_t, DO_t and SD_t governs the process. In an attempt to improve the prediction, GP input vector is modified by including the 7 day prediction of chlorophyll as obtained above. This modified chlorophyll prediction model for the 14 day lead period prediction is shown in Eq (4):
The best results from various GP runs are listed in Table 5. The best model obtained is model 10, which is presented below as Eq (7):

$$chy_{t+14} = chy_{t+7} \left| 1 - \tanh\left(chy_t - \frac{3SD_t^{\,2} + chy_t}{SD_t * Temp_t} \right) \right| \qquad (7)$$

It is also interesting to note from Table 4 and Table 5 that the RMSE of the evolved models is nearly same when the models are validated for Chy within the training range. However, the

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

results drastically differ when validated for Chy outside the training range. Further, inclusion of Chy_{t+7} has reduced the RMSE of the best model to 3.98 µg/l. Further, the temperature of the water body also seems to affect the process. Because of the presence of Chy_{t+7}, the effect of DO and TIN are automatically included. Thus, in addition to Chy, TEMP and SD affect the process. Similar conclusions on the variables affecting algal bloom modeling has been arrived at by other researchers, for example, Recknagel et al. (2002), Bobbin and Recknagel (1999), Lee et al. (2003) and Muttil and Lee (2005).

Figure 4 shows the plot of prediction obtained with model 10 in Table 5 (presented as Eq (7) above) and the actual Chy_{t+14}. As seen from this figure, a phase error up to 1 week is observed for both lower and higher values of chlorophyll and also with few under predictions.

CONCLUSION

The following conclusions can be arrived at based on the present study:

(a) The procedure outlined in this study illustrates a simple way to select the best model evolved from various GP runs (in terms of selection of training and validating set). It is very clear that for a better confidence in the use of GP model, it is better to select that model which gives best performance when validated outside the training range.

(b) The prediction of Chy_{t+7} most significantly depends on Chy_t. Though TIN and DO also affect the process, they are more stable with respect to measurement error.

(c) The prediction of Chy_{t+14} most significantly depends on Chy_{t+7} and Chy_t. In addition, TEMP and SD are also found to influence the process.

(d) Due to linear interpolation of biweekly data to obtain daily data, the performance of naïve model appears as good as one of the other models evolved by GP for 7 lead day prediction. However, such models do not appear for biweekly predictions.

(e) It is strongly believed that more meaningful insight can be obtained to predict the complex algal bloom process using GP only once the suggested methods are validated on other data sets.

Acknowledgement

The authors would like to acknowledge and thank the Environmental Protection Department of the HKSAR for providing the data used in this study. The authors also wish to thank DHI Water and Environment for providing the GP software, GPKernel and also wish to acknowledge the financial support received from Council of Scientific and Industrial Research (CSIR), Government of India for this work. The authors wish to express their special thanks to the anonymous reviewers who have given very vital inputs to improve the quality of the manuscript.

REFERENCES

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

Bobbin, J. and Recknagel, F. 1999 Mining water quality time series for predictive rules for algal blooms by genetic algorithms, MODSIM 1999 International Congress on Modelling and Simulation, Modeling and Simulation Society of Australia and New Zealand, 6-9 December, 1999, University of Waikato, New Zealand.

Chau, K.W. and Muttil, N. 2007 Data mining and multivariate statistical analysis for ecological system in coastal, Journal of Hydroinformatics, 9 (4), 305–317.

Coad, P.B., Cathers, D. and Senden, V. 2005 Predicting estuarine algal blooms utilizing neural network modeling- A Preliminary Investigation, In Zerger, A. and Argent, R.M. (eds) MODSIM 2005 International Congress on Modeling and Simulation, Modeling and Simulation Society of Australia and New Zealand, 2373 - 2379. ISBN: 0-9758400-2-9.

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

Lee, J.H.W. and Qu, B. 2004 Hydrodynamic tracking of the massive spring 1998 red tide in Hong Kong, Journal of Environmental Engineering, ASCE, 130 (5), 535–550.

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

Lui, G.C.S., Li, W.K., Leung, K.M.Y., Lee, J.H.W., Jayawardena, A.W. 2007 Modelling algal blooms using vector autoregressive model with exogenous variables and long memory filter, Ecological Modelling, 200 (1-2), 130-138.

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

Muttil, N. and Chau, K.W. 2006 Neural network and genetic programming for modeling coastal algal blooms, International Journal of Environment and Pollution, 28 (3-4), 223–238.

Recknagel, F., French, M., Harkonen, P. and Yabunaka, K.I. 1997 Artificial neural network approach for modeling and prediction of algal blooms, Ecological Modeling, 96 (3), 11-28.

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, Journal of Hydroinformatics, 4, 125–134.

Sam, W., Tan, C.K., Ishizaka, J., Son, T.P.H., Ransi, V., Tarigan, S. and Sediadi, A. 2005 Monitoring of algal blooms and massive fish kill in the Jakarta Bay, Indonesia using satellite imageries, Estuarine, Coastal and Shelf Science, 74, 1841–1847.

Whigham, P. and Recknagel, F. 1999 Predictive modelling of plankton dynamics in freshwater lakes using genetic programming, MODSIM 1999 International Congress on Modelling and Simulation, Modeling and Simulation Society of Australia and New Zealand, 6-9 December, 1999, New Zealand, 691–696.

NOTATIONS

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

GP	- Genetic Programming
Chy	- Chlorophyll
Chyt+7	- Chlorophyll at day 7
TIN	- Total Inorganic Nitrogen
DO	- Dissolved Oxygen
HAB	- Harmful Algal Bloom
ANN	- Artificial Neural Network
GA	- Genetic Algorithm
MODIS	- Moderate Resolution Imaging Spectro–Radiometer
ALOS	- Advanced Land Observation Satellite
TGP	- Tree-based Genetic Programming
EPD	- Environmental Protection Department
PO_4	- Phosphorus
SD	- Secchi-disc Depth
SR	- Daily Solar Radiation
MWS	- Daily Average Wind Speed
RMSE	- Root-Mean-Square-Error
CC	- Correlation Coefficient

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

Table 1: Values of parameters used in GP runs

Parameter	**Value**
Population Size	
Maximum equation size	
Crossover rate	
Mutation rate	
Stopping criterion	
Elitism used	Yes

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

Model No.	GP evolved equations	RMSE (Within the Training Range)	RMSE (Outside the Training Range)
1	$chy_{t+7} = chy_t$	3.857	2.915
2	$chy_{t+7} = \sqrt{\{\sqrt{\sqrt{\sqrt{chy_t}} + \log(\sqrt{DO_t})} + \{\log(chy_t) * (chy_t + (chy_t * \sqrt{chy_t}\}}$	3.82	3.604
3	$chy_{t+7} = \{[SD_t * (SD_t - 4.294)] + 4.29\} + chy_t$	3.806	3.05
4	$chy_{t+7} = \log(chy_t) * \log[(chy_t)^2 + \sqrt{chy_t}]$	3.785	7.137
5	$chy_{t+7} = \sqrt{chy_t * \sqrt{(chy_t * DO_t) * [\sqrt{chy_t + \sqrt{chy_t}} + chy_t]}}$	3.788	12.613
6	$chy_{t+7} = chy_t + \log\left(\frac{DO_t}{SD_t}\right) + \log(TIN_t + 2.66) - \log(chy_t)$	3.803	2.9
7	$chy_{t+7} = SD_t + \tanh(\tanh(chy_t)) + \frac{chy_t}{\tanh(SD_t)}$	3.785	3.42
8	$chy_{t+7} = \frac{chy_t * [\tanh(\tanh(SD_t))]}{\tanh(\tanh(\tanh(SD_t))) + \tanh(0.2)} + \frac{SD_t}{\tanh(\tanh(SD_t))} + \tanh(\tanh(\tanh(chy_t)))$	3.779	21.44
9	$chy_{t+7} = \sqrt{\{\sqrt{[DO_t + \sqrt{(chy_t + TIN_t)} + \log(DO_t)] * chy_t}\} * chy_t}$	3.790	2.58
10	$chy_{t+7} = \left(\frac{chy_t}{sd_t}\right)^{1/4} + (chy_t)^{15/16}$	3.791	2.59

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

Table 3: Sensitivity analysis for GP evolved best model for Chyt+7

Variables	**Error measure**	**Actual value**	**5%**	**10%**	**15%**
Chlorophyll	CC	0.931	.931	.931	.931
	RMSE (µg/l)	2.583	2.63	2.72	2.84
TIN	CC	0.931	.931	.931	.931
	RMSE (µg/l)	2.583	2.583	2.583	2.583
DO	CC	0.931	.931	.931	.931
	RMSE (µg/l)	2.583	2.59	2.6	2.609

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

Model No.	GP evolved equations	RMSE (Within the Training Range)	RMSE (Outside the Training Range)
1	$chy_{t+14} = SD_t + \{(0.676 + SD_t) + \frac{CHY_t + (0.361 * SD_t)}{chy_t}\}$	5.576	6.25
2	$chy_{t+14} = \log(DO_t) + \log(SD_t) + \left(SD_t + \frac{chy_t}{SD_t}\right) PO_t$	5.568	5.65
3	$chy_{t+14} = \log(\log(DO_t)) + \left[(SD_t * \log(DO_t)) + \left(\frac{CHY_t}{SD_t}\right)\right]$	5.54	5.32
4	$chy_{t+14} = \log(SD_t) + 2SD_t + (\frac{CHY_t}{SD_t})$	5.58	6.323
5	$chy_{t+14} = \log(\log(CHY_t)) + 2SD_t + (\frac{CHY_t}{SD_t})$	5.572	6.37

Table 4: Best ten GP evolved models for predicting Chyt+14 without including Chyt+7

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

Table 5: Best ten GP evolved models for predicting Chyt+14 with Chyt+7 as additional input

Model No.	GP evolved equations	RMSE (Within the Training Range)	RMSE (Outside the Training Range)
1	$chy_{t+14} = \frac{1.32 \times chy_{t+7}}{TIN_t \times SD_t} + (3TIN_t + 2SD_t - 2)$	5.54	11.66
2	$chy_{t+14} = chy_{t+7} + \frac{MWS_t}{2 * chy_{t+7}}$	5.66	4.997
3	$chy_{t+14} = SD_t + \frac{chy_{t+7}}{\sqrt{SD_t}} + 1$	5.565	4.86
4	$chy_{t+14} = 2 * \frac{\log(SD_t)}{SD_t} + SD_t + \frac{chy_{t+7}}{SD_t}$	5.583	5.593
5	$chy_{t+14} = SD_t + \frac{chy_{t+7} + (SD_t + chy_{t+7})}{SD_t + 0.92}$	5.564	4.736
6	$chy_{t+14} = 1.13 * \left(SD_t + \frac{chy_{t+7}}{SD_t} \right) + (SD_t + PO_t)$ $chy_{t+14} = 0.761 + SD_t + PO_t \left(\frac{chy_t}{SD_t} + SD_t \right)$	5.55 5.585	6 6.596
7	$chy_{t+14} = \log \left(\frac{chy_{t+7}}{chy_t} \right) + \left(2SD_t + \frac{chy_{t+7}}{SD_t} \right) \frac{chy_t}{\sqrt{SD_t}}$ $chy_{t+14} = \sqrt{SD_t + \sqrt{2SD_t \times DO_t}} + chy_t \times \left[\frac{chy_{t+7}}{SD_t} + \frac{chy_t}{\sqrt{SD_t}} \right]$	5.56 5.568	5.396 5.64
8	$chy_{t+14} = chy_{t+7} \times \frac{(\tanh(\tanh[\tanh(chy_{t+7}) + SD_t])) + SD_t}{SD_t}$ $chy_{t+14} = SD_t + \left(\frac{chy_t}{SD_t} \right) \tanh(\log(DO_t * SD_t))$	5.58 5.568	4.567 5.658
9	$chy_{t+14} = \log(chy_t) * SD_t + \frac{chy_{t+7}}{SD_t \left(\frac{chy_t}{\sqrt{chy_{t+7}}} + \frac{SD_t}{chy_{t+7}} \right)}$ $chy_{t+14} = \log(\log(chy_t)) + SD_t + \frac{chy_{t+7}}{SD_t \left(\frac{chy_t}{\sqrt{chy_{t+7}}} + \frac{SD_t}{chy_{t+7}} \right)}$	5.36 5.572	4.41 6.379
10	$chy_{t+14} = chy_{t+7} \left[\frac{DO_t \times 2.7}{1 - 2.352 chy_t - \left(\frac{chy_t}{SD_t * Temp_t} \right)} + 2SD_t + \frac{chy_t}{SD_t} + chy_t \right]$	5.57	3.98 6.766

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

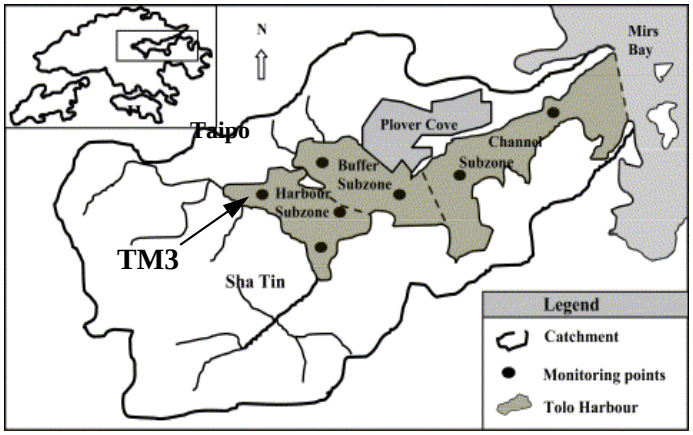

Figure 1**:** Location of study site: Tolo Harbour

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

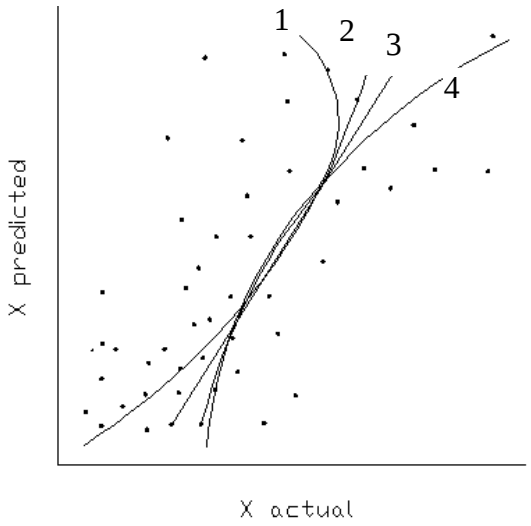

Figure 2: Demonstration sketch for illustrating fitting of different models for a given sample data

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

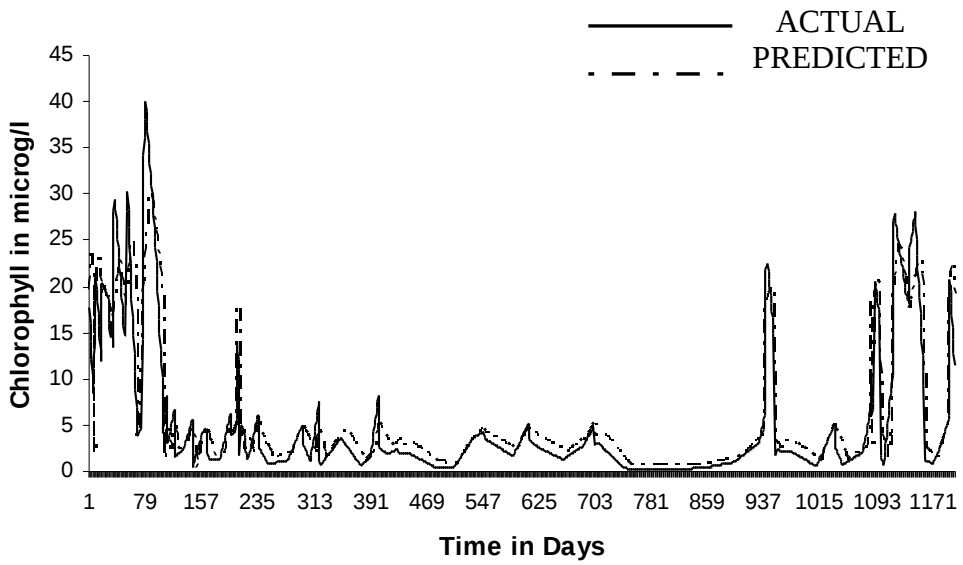

Figure 3: Comparison of actual and predicted value of chlorophyll at 7-lead day (using the model presented in Eq (5))

Marine Pollution Bulletin, Vol. 60 (10), 2010, pp. 1849 – 1855.
Published version can be downloaded from: http://dx.doi.org/10.1016/j.marpolbul.2010.05.020

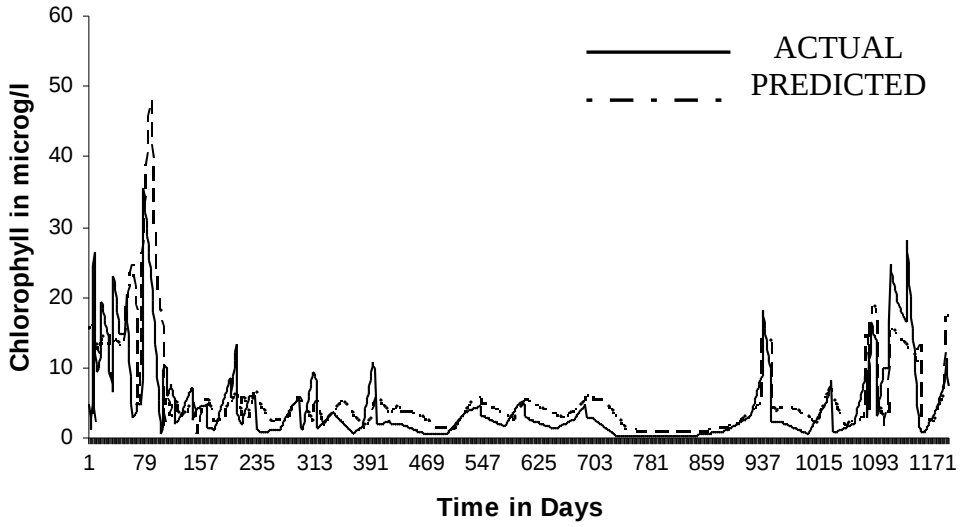

Figure 4: Comparison of actual and predicted value of chlorophyll at 14-lead day (using the model presented in Eq (7))