Comment on 'Comparative application of artificial nueral networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes'

Joseph H. W. Lee, Y. Huang and A. W. Jayawardena

INTRODUCTION

The prediction of the dynamics of algal blooms—the nonlinear oscillations in the concentration of algal species, with attendant implications on water quality, is a notoriously difficult problem in ecological sciences and environmental engineering. This problem is important because of the practical need to develop early warning systems for harmful algal blooms. There is an enormous literature on eutrophication modeling, in which the prediction of algal dynamics often plays a central role. In the past two decades, most of the major advances were made using deterministic water quality models that mimic the phytoplankton growth and nutrient cycles (e.g. Thomann and Mueller 1987; Di Toro 2001). When properly calibrated against extensive field data, these models can predict trends reasonably well and are often adequate for water quality management purposes (e.g. Lee and Lee 1995). On the other hand, the use of data-driven methods to model algal dynamics appeared only very recently.

Recknagel *et al.* (2002) compare the performance of an artificial neural network model (ANNA) with the SALMO model with parameters estimated by a genetic algorithm and a linguistic model built by a genetic algorithm. The data used in the paper are from Lake Kasumigaura. The main conclusions from this paper are: (i) that multivariate modelling by machine learning techniques is superior to deterministic modelling techniques; (ii) artificial neural networks can be powerful short-term predictors of the timing of algal bloom events; and (iii) causal knowledge can be discovered from the evolutionary algorithms presented.

Although the paper outlines an interesting approach to algal bloom modelling, we believe that certain essential details of the data analysis are missing. In the absence of the needed details, the conclusions of the paper are perhaps pre-mature and over-stated, and indeed can be misleading if the results are taken at face value. The following points are discussed in relation to a well-established body of knowledge of algal dynamics—derived from many field and modeling studies by many investigators.

PREDICTABILITY OF ALGAL BLOOM EVENTS

Based on the comparison of seven-days-ahead prediction of Chlorophyll-*a* and cell counts with data in Figure 3, the authors suggest that the data-driven methods presented are good short-term predictors of algal blooms. We find it difficult to share this observation in several respects.

1. First, the reasonableness of the model structure of both ANNA and GAMA can be questioned (as outlined in Figures 1 and 2). It appears the algal biomass (output node) at the current time is linked with the nutrient (nitrate and orthophosphate) concentration and secchi depth (a measure of the transparency of the water) as input nodes *at the same time*. While this approach would offer insight into the controlling variables, it cannot be used as a forecasting tool. These measurements are seldom available in real time (e.g. nitrate or secchi depth). In a realistic situation, both of these are exactly the critical unknowns we are looking for. As an example, it is well known in phytoplankton dynamics (e.g. Lee *et al*. 1991) that high Chlorophyll-*a* levels are often associated with low

levels of the limiting nutrient concentration (due to nutrient uptake by algae). As another example, it is well known that light extinction in water depends on the algal concentration—the 'self-shading' effect. To within the typical scatter in the field data, the extinction coefficient is inversely proportional to the secchi depth, with $Ke = C/Zs$, where Zs = secchi depth, and C = constant, 1.7–1.9 (Thomann and Mueller 1987). On the other hand, the extinction coefficient can be related to the Chlorophyll-*a* concentration (P)—e.g. the Riley equation gives:

$$Ke = Ke' + 0.0088P + 0.054P^{2/3}$$

where Ke = extinction coefficient; Ke' = background value (the value when algal concentration is very low due to e.g. suspended solids). The above well-verified relation implies that if the secchi disk depth is known, the chlorophyll concentration can already be estimated to reasonable accuracy! The above equation applies to normal situations when near surface light transparency is highly correlated with algal biomass, and of course excludes special scenarios or episodic events such as turbidity plumes or fronts of high suspended solids concentrations.

2. In the models, the lead-time of the output (or, the time lag of the inputs) is not clearly defined. Without this, the quality of the model predictions can hardly be judged. It is also curious to note that the antecedent information of either Chlorophyll-*a* or the cell count of *Microcystis* did not enter as one of the inputs. One would expect that the algal dynamics depends on temporal feedback mechanisms (Huang 2001).
3. The paper lacks a detailed description of the data and its analysis, without which it is difficult to judge the effectiveness of the predictive models. For example, what is the temporal frequency of the data? It is also not clear whether the data has been interpolated. Without this information, it is impossible to know what the authors are actually accomplishing with their seven-days-ahead predictors. If the frequency were weekly, we would be dealing with a one time-step-ahead predictor. If the frequency were less than weekly, we would then be relying on the interpolations made by the model. Likewise, the models were judged by RMSE. Yet, it is unclear if the RMSE was calculated from the sample that it was estimated on, or on the 1986 and 1993 validation only. In Figure 3(b), in discussing the performance of ANNA in predicting the abundance of *Microcystis* in 1986, the authors state that '. . . the timing of fast algal growth towards the large summer peak in 1986 was predicted very well, even though the peak was reached a few days in advance.' From Figure 3(b), it appears however that the predicted peak is out of phase with the observations by *1–2 months*. Unless the prediction by the techniques presented by the authors are compared with observations and predictions by deterministic methods on the same time scale, and objective error measures given, the statement that the proposed techniques succeed where traditional modelling techniques fail remains highly problematic.
4. Moreover, the GAMASalmo model essentially applies a genetic algorithm to the estimation of the parameters of the SALMO model. There are many other techniques for estimating these parameters, most notably nonlinear least squares. The genetic algorithm is actually quite an inefficient search method; it is normally used to reduce the dimension of the input space, and its strong point lies in optimization over a complicated parameter space. However, the SALMO model does not appear to be overly complicated, and the application of the genetic algorithm to this problem needs justification.

DISCOVERY OF PREDICTIVE RULES

It is suggested that the rule discovery model is proven to be a powerful tool for discovery and testing of causal knowledge. We are skeptical of the authors' claim that the GAMARules model yields results that facilitate the accumulation of causal knowledge.

1. It is claimed that GAMA is able to extract from the time series a compact multiple nonlinear equation

'which allows the approximate prediction of the seasonal dynamics of algal biomass for two unseen years'. On closer examination, the equation relates the algal concentration to the water temperature, secchi depth, and nitrate nitrogen concentration:

$$\text{Chl-a} = \frac{\cosh(T)}{6S + 2N + 2.251}$$

As discussed above, the inverse dependence of Chl-*a* on secchi depth and nutrient concentration is well known; the monotonic increase of algal growth rate on water temperature is also well-established. However, the equation (effectively a correlation) really does not allow one to do forecasting as often the nutrient concentration is *unknown.* In addition, any time lags in the dynamics are not included. It is also clear that, contrary to the authors' claims, this equation is incapable of predicting short-term algal dynamics. For example, in warm waters, at the tail or beginning of a bloom, the algal concentration is by definition very low. This directly contradicts the above equation, which reflects more the seasonal behaviour of higher Chl-*a* concentration at higher water temperature. Thus it appears that the evolutionary algorithm probably has its greatest utility in updating the parameters of a known relation, rather than discovering unknown relations.

2. The estimated GAMARules model failed to extract the knowledge as claimed. By looking at the model, we are confused by the string of IF–THEN–ELSE–EXCEPT IF–OR statements. Consider some rules obtained from the model:

   IF $P >= 81.7$ μg/l AND $P <= 126$ μg/l THEN Microcystis = 500,000 cells

   IF $P >= 15.6$ μg/l AND P$<=$ 116 μg/l THEN Microcystis = 100,000 cells

   First the rule defies any sensible ecological interpretation—why should the algae stop to grow if the nutrient concentration stays within a prescribed range? Obviously what is happening here is that as the genetic algorithm evolves, it is refining the model by applying in essence new filters. But this throws the ability of the technique to provide any causal knowledge into serious doubt. By specifying a different method of rule selection, one could have obtained a different predictor with the same level of accuracy. Put another way, we believe the rule set generated by the GAMARules model is not unique, and hence the model is not a robust discoverer of actual causal knowledge. Moreover, other input variables such as solar radiation, rotifiers and copepod concentrations are apparently missing from the GAMARules model. If we are to believe the claim that we can extract causal knowledge from the GAMARules model, then perhaps we should believe that these variables are unimportant to *Microcystis* blooms.

   Second, the GAMARules model was the only model to correctly predict the lack of a *Microcystis* bloom in 1993—i.e. it is the only model that appears to differentiate the biomass of total phytoplankton (as measured by Chlorophyll-*a*) and cell counts of specific species. It is not at all clear why the stated rule set can enable such predictive power. Although the GAMARules model is shown to be a useful predictor, the onus is on the authors to demonstrate the validity of the causal knowledge embedded in it. In fact, we would argue that there is just as much causal knowledge contained within the GAMARules model as the artificial neural network.

CONCLUSIONS

In summary, Recknagel *et al.* (2002) present some tantalizing results for the prediction of algal blooms. However, we believe that the conclusions need to be tested more thoroughly against more years of independent data and properly interpreted against the large body of process-based modeling and field observation experience. We also believe that their research was compromised to some extent by the confusing presentation. Nevertheless, it is certainly important research that deserves a higher level of attention.

Joseph H. W. Lee (corresponding author)
A. W. Jayawardena
Department of Civil Engineering,
The University of Hong Kong,
Hong Kong,
China

Y. Huang
Water Engineering and Management Engineering Faculty,
The University of Twente,
The Netherlands

REFERENCES

Di Toro, D. M. 2001 *Sediment Flux Modelling*. John Wiley and Sons, New York.

Huang, Y. 2001 Neural network modelling of coastal algal blooms. *M.Phil. Thesis*, The University of Hong Kong.

Lee, J. H. W., Wu, R. S. S. & Cheung, Y. K. 1991 Forecasting of dissolved oxygen in marine fish culture zone. *J. Environ. Engng, ASCE,* **117**(6), 816–833.

Lee, H. S. & Lee, J. H. W. 1995 Continuous monitoring of short term dissolved oxygen and algal dynamics. *Wat. Res.*, **29**(12), 2789–2796.

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. *J. Hydroinformatics* **4**(2), 125–133.

Thomann, R. V. & Mueller, J. A. 1987 *Principles of Surface Water Quality Modeling and Control*. Harper and Row, New York.