Short Term Forecasting of Algal Blooms in Drinking Water Reservoirs using Artificial Neural Networks

A thesis submitted for the award of Doctor of Philosophy

Hugh Edward Campbell Wilson
Discipline of Environmental Biology
School of Earth and Environmental Sciences
The University of Adelaide

April 2004

Abstract

Artificial neural networks (ANNs), trained to make short term forecasts of algal blooms in lakes and rivers, are potentially useful decision making tools for the operational management of eutrophication. This thesis addresses the question of whether a standardised, generic ANN model representation can be developed to achieve this goal. It is argued that four requirements need to be addressed; i) compatibility of models with existing water quality monitoring regimes, ii) stability and repeatability of training outcomes, iii) realistic and meaningful estimates of model performance and iv) explanation of predictions.

ANN model inputs were represented as *summary statistics* of *sliding time windows*. This approach was shown to increase the compatibility of typical time-series ANN model structures with datasets compromised by missing values and uneven sampling intervals. To improve stability, models were represented as an ensemble of ANNs trained on bootstrap samples of data (ie *bagging* (Breiman, 1994)). It was shown that the average prediction of the bagging ensemble was relatively unaffected by variance of the individual member models. Validation set representation was maximised by use of leave-k-out methods. Comparative error measures were devised to illustrate model performance characteristics relative to "naive" controls. A *sensitivity analysis through time* approach was utilised to explain the relative importance of input variables and to account complex interactions between variables.

Training data was available from six sites including Lake Biwa (Japan), Burrinjuck Dam (NSW, Australia), Darling River (NSW, Australia), Lake Kasumigaura (Japan), Myponga Reservoir (SA, Australia) and Lake Soyang (South Korea). These datasets were found to differ significantly from each other in terms of environmental characteristics and data availability. Models were developed to make one and two week forecasts. Predicted variables included chlorophyll *a* concentration and cell counts of the three most abundant algal species for each dataset. Experimental results showed that site/output specific input layers lead to better performance than site/output generic models. Furthermore, it is evident that ANNs capable of non-linear processing generalise better over local (short term) time scales, whereas perceptron models constrained to linear decision boundaries perform better over global (long term) scales.

Contents