Division of Science, Research and Technology

Research Project Summary

June 2006

Forecasting Algal Blooms at a Surface Water System with Artificial Neural Networks

Emery A. Coppola Jr.[1], Ph.D.
Adorable B. Jacinto[1], B.S.
Scott Lohbauer[1]
Mary Poulton[1,2], Ph.D.
Ferenc Szidarvoszky[1,3], Ph.D.
Tom Atherholt[4], Ph.D.

Abstract

Algal blooms (AB) in potable water supplies are becoming an increasingly prevalent and serious water quality problem around the world. AB events can cause taste and odor problems, damage the environment, and some algal classes like cyanobacteria (blue-green algae) may release toxins that can cause human illness or even death. There is a need to develop models that can accurately forecast algal bloom events on the basis of predictive physical, meteorological, chemical, and biological information. Such forecasting models can provide valuable lead time for water treatment systems to implement measures to minimize the consequences of the AB event, if not actually prevent it. Given the multitude, interplay, and complexity of the various controlling environmental factors, modeling and forecasting AB is a daunting challenge. This research focused on the feasibility of using artificial neural network (ANN) technology as an accurate, real-time modeling and forecasting tool. Previously-collected data from a NJ water utility served as the test case. AB forecasting periods included one-week and two-weeks prior to the event. Despite a less than ideal number of historical AB events, the high predictive accuracy achieved in this study indicates that with sufficient data, both in terms of the number of historical AB events and availability of important predictor data, ANNs can serve as reliable, accurate, real-time AB forecasting tools.

Introduction

There is consensus among scientists that the incidence of algal blooms (AB) world-wide is increasing (Smith et al, 2006). The detrimental effects of AB on the environment and water supplies are well documented but controversy remains over the important factors and mechanisms responsible for their occurrence as well as the most effective means for modeling and forecasting this phenomena. Given the multitude of factors – physical, meteorological, chemical, and biological – that can contribute to AB events, developing a robust model that represents site-specific conditions responsible for AB, especially for the purpose of providing accurate prediction capability, can be a daunting challenge.

In this project, artificial neural network (ANN) technology, a form of artificial intelligence, was investigated as a possible AB event forecasting tool. ANN technologies offer the advantage of "learning" system behavior from historical data and hence are not necessarily constrained by simplifying model assumptions inherent to mechanistic or physical-based models. To effectively learn and generalize system behavior, ANNs require a sufficient amount of historical "cause and effect" data that covers the expected range of conditions at a given site. One of the critical issues of this study, then, was a rudimentary assessment of the existing data and analysis as to what sampling strategies might improve forecasting capability.

Previous work in the literature documents the development and testing of ANN models for forecasting algal blooms in river and lake systems (Recknagel et al, 1997; Maier et al, 1998; Olden, 2000). However, the data sets were relatively large, collected over time periods approaching a decade, and with parameters that were measured relatively frequently. In this study, the number of available historical events was more limited, which is representative of most water supply systems. After examination of data from one water treatment plant (WTP) showed an insufficient amount for robust model development, a second WTP was selected for this project. Three classes of algae, cyanobacteria, chrsyophytes, and chlorophytes were included as the forecast parameters of interest.

Different ANN models were developed and tested. Most models predicted discrete algal count numbers while some predicted a "bin" or pre-selected classification range within which final algal counts would fall. The use of different modeling approaches and time discretization schemes improved system understanding. It also permitted a systematic analysis of available modeling and data collection options for performing real-time forecasting, under real-time conditions and with limited data sets.

Artificial Neural Networks

ANN architecture (Figure 1) is based upon Kolmogorov's theorem (Hecht-Nielsen, 1987; Sprecher, 1965) which asserts that any continuous function (in this case algal counts and the causative inputs) can be represented exactly by a three-layer, feed-forward neural network with n elements in the input layer, 2n+1 elements in the hidden layer, and m elements in the output layer, where n and m are arbitrary positive integers. The presence of common arcs in its architecture allows ANN to identify important inter-relationships that may exist between output variables. ANN technology is a compelling alternative to physical and statistical-based modeling approaches such as linear models. ANN "learns" system behavior by processing representative data through its architecture. ANN is different from physical-based models because it does not rely upon the governing physical laws. Information regarding physical parameters is not required for ANN development and operation nor simplifying statistical assumptions.

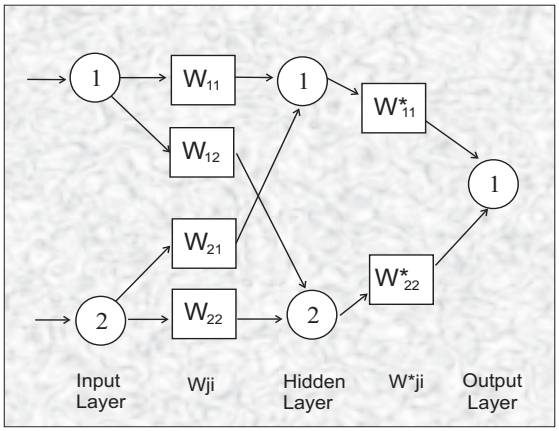

Figure 1. Architecture for a simple multi-perceptron ANN

In this study, 50% of the available data was used for "training," that is to "learn" cause and effect relationships if present. Another 25% of the data was used to "verify" the model, to guard against over-training or over-fitting the data. Following training, the remaining 25% of the data was used to validate or assess how well the model learned to generalize system behavior. During training, data patterns are processed through the ANN and "connection weights" are adaptively adjusted until a minimum acceptable error between the ANN-predicted output and the actual output is achieved. At this point, the ANN has "learned" to predict the system behavior of interest (in this case algal counts) in response to the values of the various input parameters.

There are a variety of ANN model design features and options. To design an appropriate model, a number of factors must be considered, including the functional form of the transfer functions, the number of hidden layers and nodes in the architecture, the most appropriate set of input variables, and the algorithm(s) used to minimize the objective function (i.e. training error). This process is typically conducted in an iterative manner within the context of professional judgment and modeling experience. For example, selection of an appropriate set of input parameters during initial ANN development requires a basic understanding of the governing system dynamics (e.g., factors known to influence AB). However, a "sensitivity analysis," in conjunction with trial and error, can help the modeler converge to the most appropriate and feasible set of predictor variables. The sensitivity analysis, which quantifies the relative importance of each input variable for accurately predicting each output variable, can be used in lieu of common statistical methods.

ANNs require sufficient data that spans the range of expected system conditions to achieve robust learning. Based upon the number of input and output parameters, heuristic equations were used to estimate the minimum number of training data events required for robust model development. Calculated estimates of the number of training events necessary for robust training in this study ranged from 200 to 500, depending upon the ANN model used. Because of the number and complexity of environmental factors and their interactions which control AB dynamics, and given the expected "noise" in the data, the number of required training data sets is probably closer to 500. Unfortunately, the number of data sets available in this study was well below 200. Therefore, multiple modeling iterations were performed to validate results.

Study Area and Data

Figure 2 represents the WTP modeled in this project. Two rivers and a reservoir supply water to the WTP. River A flows into River B upstream of the WTP's intake canal. River B water is gravity fed to the WTP intake by way of the canal. Station 101 water samples are River B water. River A water can be pumped directly to the plant intake location (Station 100) via Pumping Station 2 (PS 2). Station 612 samples are River A water. Reservoir A water is only available to the WTP on a limited basis during the summer to augment needed supply. When not in use, Reservoir A is recharged with River A water via Pumping Station 1 (PS 1). When needed, Reservoir A water can also be delivered directly to Station 100. Thus, Station 100 water represents the final blend of the respective supply waters (River A, River B, and/or Reservoir A) that form the WTP's intake. During most of the data collection period of this study, Station 100 intake water was

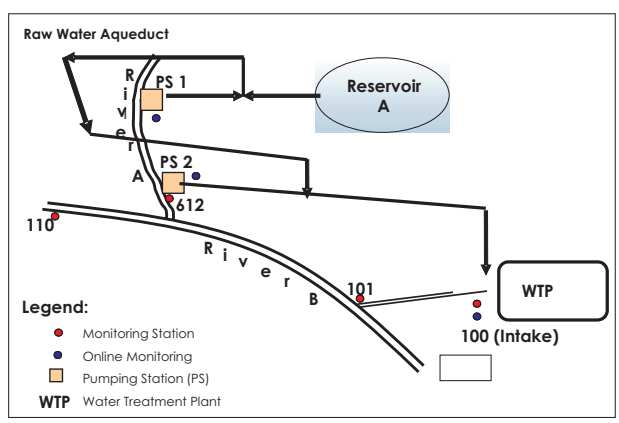

Figure 2. Water Treatment Plant Raw Water Configuration

a blend of River A (100 – 60%) and River B (0 – 40%) water. On rare occasions, it was a blend of River B and Reservoir A water.

Rivers A and B have historically exhibited variable and unique water quality characteristics that impart different treatment challenges. River B is considered to be of lower water quality because of more numerous upstream contaminant sources. However, River A has a higher incidence of AB events.

Previously collected (1999-2004) water quality data from Stations 100, 101, and 612 were used. Temporally corresponding meteorological data from nearby weather stations was also collected. Data that was used included: total algae, chrysophytes, chlorophytes, cyanophytes, water temperature, pH, turbidity, alkalinity, hardness, conductivity, color, odor, dissolved oxygen level, biochemical oxygen demand level, chloride, sulfate, total phosphorous/ortho-phosphate, ammonia, nitrite/nitrate, total suspended solids, total amorphous material, UV-254 absorbance, total organic carbon, precipitation, length of day, wind speed and direction, heating degree days, cloud cover, streamflow, and extraction volume from each water source. The non-algae parameters were considered important or potentially important for predicting algal population counts. Sampling dates and frequencies differed by both station and parameter. Thus, prediction events (data sets) between stations often do not correspond in time. For regulatory reasons Station 100 is sampled at the highest frequency and as many as 270 historical events were available from this station. Station 101, located on the less-frequently used river source, was sampled less often with between 32 and 108 events available (depending on ANN model used) while Station 612 contained data from 40 up to 172 events. Stations 100, 101 and 612 were each modeled individually (rather than pooled into a single "location") as each source has different water quality characteristics.

Modeling Methodology

Because reliable weather forecasts do not extend beyond one- to two-week time horizons, ANN models were developed for one-week and two-week ahead forecasting periods. These periods are long enough to allow a WTP sufficient time to plan and implement countermeasures for predicted AB events. Two model input structures were used. The first, referred to as "original", consisted of input values measured at the beginning of the prediction periods. The second, referred to as "revised", used input values measured at the end of the prediction period, coinciding with the final or predicted algal count. Under the original approach, under real-time forecasting conditions the input values would be known a-priori. The revised approach would have to forecast or assume input values corresponding to the future prediction day.

Both the original and revised approaches were assessed using two distinct data sets. The first events consisted of a smaller number of temporally coincident data events, but which included a higher number of input parameters. The second set, by excluding several less-frequently-sampled water quality parameters (total phosphorous/ortho-phosphate, nitrite/nitrate, sulfate, and total organic carbon for all stations and biological oxygen demand for Stations 101 and 612), consisted of a larger number of data set events but with fewer parameters per set. The dual approach provided insights into algal population dynamics and also addressed the issue of dependence of ANN model performance on the quantity of data

The classification "bin" models, which consisted of Radial Basis Function (RBF) nets, used the original set of input parameters but excluded the source water extraction volume parameter. This was done to assess whether ANN learning and predictions were biased by correlations between changing intake water quality and resulting WTP operational decisions (i.e., changing the source of supply water). Four pre-selected bins or classification ranges for algal counts were used as model outputs: 0 to 10, 11 to 50, 51 to 200, and > 200 organisms per milliliter.

Results and Discussion

Several hundred ANN models were developed and assessed during this study. Despite the low number of historical data events available for training, many of the ANN models performed well during validation, often achieving relatively high correlation coefficients and accurately predicting sudden and significant changes in algal populations. The models developed with both one-week and two-week ahead prediction periods often accurately predicted formation and dissipation of AB events, as well as the relative increase and decrease in cell counts. This indicates that there are natural time lags between system conditions and algal population responses. That is, the trajectory of algal counts over one and two-week forecast periods can be accurately forecasted on the basis of real-time measurements. Statistical analyses of the data also reflect the fact that water conditions, as influenced by external factors like weather, do not change significantly in the short-term (e.g., weekly or bi-weekly) in most cases. Thus, evolving algal populations are generally not prone to abrupt deviations from trajectory paths.

On the basis of validation correlation coefficients, the ANN models that used inputs measured at the beginning of the prediction period slightly outperformed the models that used inputs measured at the conclusion of the prediction periods, but the difference was not significant ($r = 0.72$ vs. 0.69).

The models that forecasted algal count values (instead of bin classifications) achieved the highest performance in most cases when the less-frequently measured water quality parameters were excluded as input variables. That is, the models that excluded the phosphate, nitrate, sulfate, TOC and BOD parameters produced a higher average correlation coefficient than did the models that included these variables (0.77 versus 0.63). This may have been due to the larger number of historical events that were available for training, after excluding these variables, rather than the relative lack of influence of these parameters on algal populations. However, it may be that the excluded parameters, at least some of which are considered by most scientists to

be critical for algal growth, usually existed within a range of concentrations that neither limited nor promoted AB events in this particular system.The latter possibility was weakly supported by sensitivity analysis results and time-series plotting of input parameters vs. algal counts.

Figures 3, 4, and 5 provide a visual assessment of model performance for three cases, where the validation data show the initial algal count corresponding to the prediction event. The term “initial” in the figures designates the initial count measured at the beginning of the prediction period, “final” designates the final algal count measured at the conclusion of the prediction period (i.e., that which is being predicted), and “ANN” designates the final count predicted by the ANN model.

The Station 101 and 612 (River’s B and A water) models that predicted algal concentration ranges (i.e., classification nets or “bins”) rather than actual counts also achieved high forecasting performance. Three of the eight models that included all of the input parameters achieved 100 percent classification accuracy. The worst-performing net correctly classified 83 percent of the events. For this approach, the models that included the less-frequently-sampled inputs (phosphate, nitrate, etc.) slightly outperformed those that did not, with correct classification percentages of 96 and 92 percent, respectively. However, the models that excluded these parameters had approximately three times the number of available data and hence had more events which bordered two adjacent classification bins. All incorrect classifications for all models occurred within an adjacent bin (e.g., a measured count of 8 which placed the count in the 0-10 bin, while the predicted bin was 11-50). Given the inherent imprecision of algal counts (Maier et al, 1998) this performance is impressive.

The Station 101 and 612 models, which excluded the water extraction volumes as input parameters, performed well. The excluded parameters were likely correlative at these stations. At Station 100 (the WTP intake), equivalent models showed diminishment after excluding these parameters. Source water mixing occurs at Station 100 and hence the excluded parameters would be considered more causal at this location.

Linear models (LMs) were also developed to predict algal

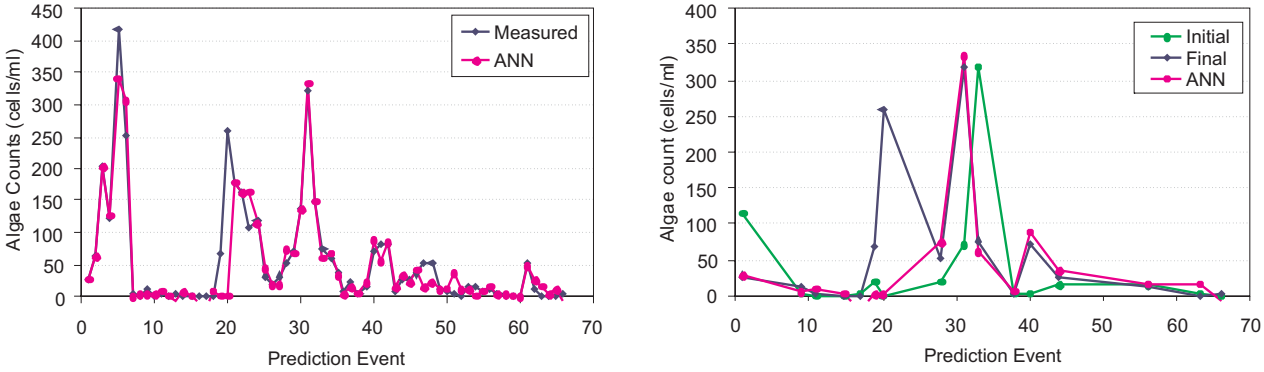

Figure 3. Time-series plots of measured Chlorophyte counts and ANN One-week Ahead predicted values for (a) complete and (b) validation data sets at Station 101 (Revised Model excluding five water quality inputs)

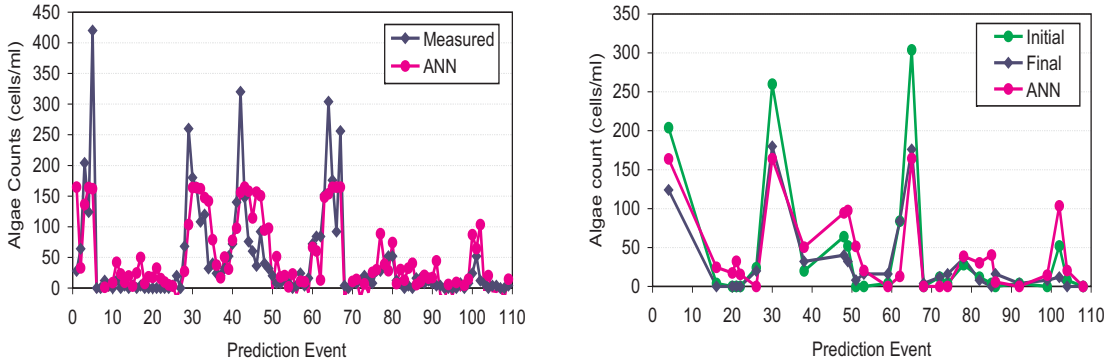

Figure 4

Figure 4. Time-series plots of measured Chlrophyte counts and ANN One-week Ahead predicted values for (a) complete and (b) validation data sets at Station 101 (Original Model excluding five water quality inputs

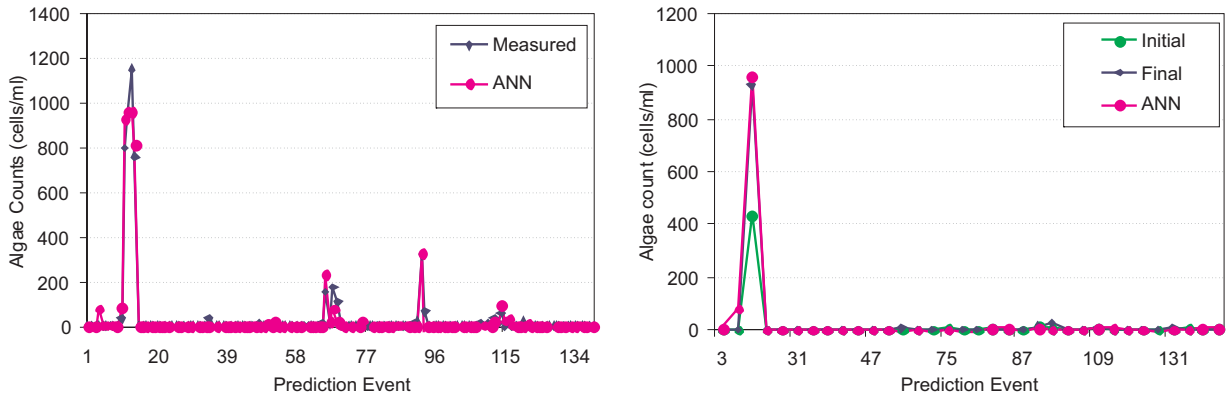

Figure 5. Time-series plots of measured Cyanobacteria counts and ANN Two-week Ahead predicted values for (a) complete and (b) validation data sets at Station 612 (Original Model excluding five water quality inputs)

counts at each station using the same data so that their performance could be compared to ANN model performance. The LMs did not perform as well as the ANN models, achieving significantly lower correlation coefficients and higher mean absolute errors. Of twelve prediction scenarios, the ANN models provided a lower mean absolute error (MAE) eleven times, were often significantly lower. As shown by Figure 6a, for the two-week ahead prediction of cyanobacteria counts at Station 612, the LM seriously under-predicted the three highest count events (comes close to fourth highest count). The LM predicted just 388, 320 and 434 for algal bloom counts of 800, 932 and 1152 counts, respectively. For relatively lower count events, the LM generally over-predicted, as shown by Figure 6b. By contrast, the ANN model accurately predicted six of the eight bloom events, and for the entire data record produced just two relatively minor false-positive events, while reproducing lower algal count events. Similarly, for the other prediction cases using the same modeling approach, LMs had the tendency to under-predict most of the high count events.

It should be noted however that the LMs were not optimized. That is, data distributions were not analyzed and data transformations such as log transformations or rankings were not attempted in LM development. However, this underscores yet another advantage of ANNs; because of their universal non-linear modeling capability, they are not limited by the form of the data distributions.

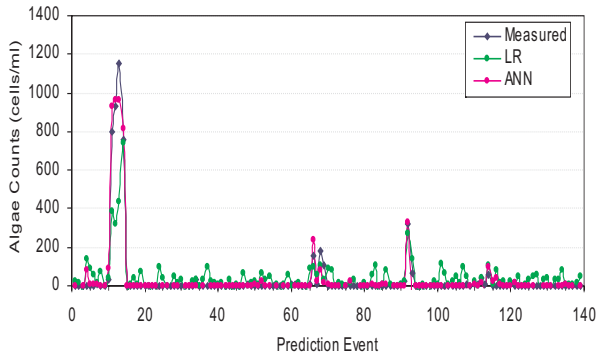

Figure 6a. Comparison of Original ANN and LM performance for two week-ahead predictions of cyanobacteria at Station 612 without the five chemical variables

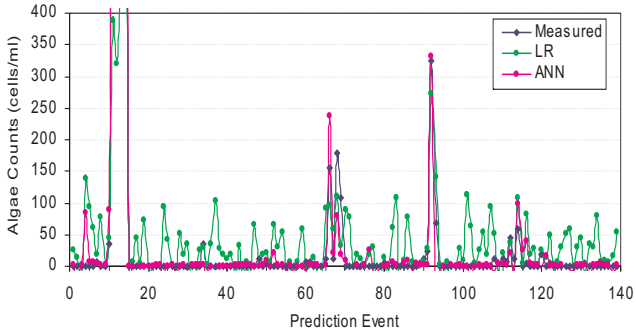

Figure 6b. Enlargement of the lower count events shown in Figure 6a.

Conclusions

- Despite a very limited number of available data events, the ANN models performed well in most cases during validation, accurately predicting large changes in algal cell populations. The degree of accuracy was surprising, given the complexity and non-linear behavior of algal populations, inherent data "noise", and the relatively small number of historical events available for model training.

- The ANN models that forecasted algal count values (instead of classification ranges) achieved the highest performance when the less-frequently measured water quality variables (phosphate, nitrate, sulfate, TOC and BOD) were excluded as input variables. This may be due to a data quantity issue rather than a lack of causative effect of these parameters on algal cell growth, but it could also be that, at the concentrations at this WTP, these parameters were not "limiting" algal growth.

- Like the cell count models, the Radial Basis Function classification net models classified the counts into the correct concentration ranges with very high accuracy, averaging 94 percent.

- Linear models did not perform as well as the ANN models, however the LM models were not optimized.

- While not definitive, the results strongly indicate that the ANN models learned some underlying relationship between select physical, meteorological, chemical, and biological parameters, and algal cell concentrations at this WTP. This is supported by: 1) relatively high model accuracy and overall consistency between training and validation results; 2) consistency in performance for different types of models (single value outputs and classification) and input structures (original and revised); 3) consistency between modeling results and physical intuition/system understanding; and 4) comparatively poor performance of linear models.

Recommendations

There are several ways in which the ANN AB forecasting models can be improved for this system in the future, and include:

1. Systematic elimination of input parameters to further distinguish between critical and non-critical ANN inputs.
2. Modeling and prediction of specific algal species (rather than algal classes) and/or algae-produced chemicals (e.g., odorant compounds) that are of particular concern to water utilities.
3. Increased monitoring of certain "limiting" nutrients such as nitrite/nitrate and total phosphorous/orthophosphate to further define their importance in AB events.
4. Inclusion of other potentially important input parameters, for example monitoring concentrations of certain biological organisms such as protozoa or viruses that feed on or lyse algae.
5. Use of time lags for select input parameters, such as streamflows and algal counts, which have been shown in a previous study (Maier et al, 1998) to significantly increase model performance.
6. A possible hybrid of the two modeling approaches, where some combination of existing/historical and future conditions is used as inputs.
7. Following development of robust models, a perturbation sensitivity analyses that quantifies how different changes in input values affect algal population changes.

Bibliography

Hecht-Nielsen, R. 1987. Counterpropagation networks. Proc. Int. Conf. on Neural Networks, II, 19-31. New York, IEEE Press.

Maier, H., R., G. C. Dandy, and M. D. Burch. 1998. Use of artificial neural networks for modeling cyanobacteria Anabaena spp. In the River Murray, South Australia. Ecol. Modelling, 105 (2-3): 257-272.

NOAH, 2006. Forecasting Algal Blooms in Surface Water Systems with Artificial Neural Networks. New Jersey Department of Environmental Protection. www.dep.state.nj.us/dsr

Olden, J.D. 2000. An artificial neural network approach for studying phytoplankton succession. Hydrobiologia 436: 131-143.

Recknagel, F., M. French, P. Harkonen, and K.-I. Yabunaka. 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Modelling, 96 (1-3): 11-28.

Smith, V.H., S.B. Joye, and R.W. Howarth. 2006. Eutrophication of freshwater and marine ecosystems. Limnol. Oceanogr. 51 (1, part 2): 351-355.

Sprecher, D. 1965. On the structure of continuous functions of several variables. Trans. Am. Math. Soc. 115: 340-355.

Funding

This study was funded by the Division of Science, Research and Technology. The final report (NOAH, 2006) is available at www.dep.state.nj.us/dsr. General information regarding other research efforts may be found on the website www.state.nj.us/dep/dsr.

Prepared By

Emery A. Coppola Jr.,[1] Ph.D., corresponding author, Mary Poulton,[1,2] Ph.D., Ferenc Szidarovsky,[1,3] Ph.D., Adorable B. Jacinto,[1] B.S., and Scott Lohbauer,[1] Principal Investigators, and Thomas Atherholt,[4] Ph.D., Project Manager.

1. NOAH LLC, 610 Lawrence Road, Lawrenceville, New Jersey, 08648; emerynoah@comcast.net.
2. Department of Mining and Geological Engineering, University of Arizona, Tucson, AZ, 85721-0012.
3. Department of Systems and Industrial Engineering, Univ. Arizona, 85721-0020.
4. Division of Science, Research, and Technology, New Jersey Department of Environmental Protection, Trenton, NJ, 08625-0409; Tom.Atherholt@dep.state.nj.us.

STATE OF NEW JERSEY
Jon S. Corzine, Governor

Department of Environmental Protection
Lisa P. Jackson, Commissioner

Division of Science, Research & Technology
Dr. Eileen Murphy, Director

Please send comments or requests to:
Division of Science, Research and Technology
P.O.Box 409, Trenton, NJ 08625
Phone: 609 984-6070
Visit the DSRT web site @ www.state.nj.us/dep/dsr

RESEARCH PROJECT SUMMARY