Forecasting Algal blooms in Surface Water Systems with Artificial Neural Networks

FINAL REPORT

Prepared and submitted by:

NOAH L.L.C.

610 Lawrence Road

Lawrenceville, NJ 08684

With participation of:

NEW JERSEY AMERICAN WATER COMPANY & PASSAIC VALLEY WATER COMPANY

Sponsored by:

NEW JERSEY DEPARTMENT OF ENVIRONMENTAL PROTECTION

March 2006

DEDICATION

To Dr. Geetha Angara, consummate scientist and water professional, who helped ensure a safe and reliable water supply for hundreds of thousands of people. Her enthusiasm, dedication, and expertise will be deeply missed by all who worked with her. This report is dedicated to her life and memory, as Dr. Angara represents the highest standard of excellence for the water industry.

ACKNOWLEDGEMENTS

This work would not have been possible without funding support from the New Jersey Department of Environmental Protection, Division of Science, Research, and Technology (DSRT). The project was largely motivated by Research Scientist Mr. Manny Patel, formerly with DSRT. Dr. Tom Atherholt, Research Scientist, and subsequent Project Manager, provided excellent technical guidance and modeling suggestions that significantly improved the research. Recognition is also due to other DSRT personnel who reviewed draft documents and provided critical feedback, including Dr. Leo Korn, Research Scientist, and Thomas Belton, Research Scientist. The two participating water utilities, New Jersey American Water (NJAW), and Passaic Valley Water Commission (PVWC), also made this project possible. Ms. Carol Storm, formerly of NJAW, kindly provided the initial algae data. Mr. Kevin Kirwin and Ms. Karen Jones, also of NJAW, provided the water treatment data and technical assistance by reviewing the draft report document. Joseph Bella, Executive Director for PVWC, allocated valuable staff time; PVWC, Linda Cummings, Plant Supervisor, provided important technical feedback, and the PVWC Laboratory Staff, who performed much of the analytical work for generating the data. And last, to Ms. Linda Pasquarello, Water Quality Specialist for PVWC, who provided important background information and excellent technical guidance and suggestions.

EXECUTIVE SUMMARY

This research projected, entitled "Forecasting Algal blooms in Surface Water Systems with Artificial Neural Networks," was funded by the New Jersey Department of Environmental Protection. The original objectives of this study were: 1) to assess the feasibility of using artificial neural networks (ANNs) as a real-time tool for accurately forecasting cyanobacteria counts, more commonly knows as blue-green algae, in surface water systems; 2) through sensitivity analyses conducted with the ANNs, identify critical climate, hydrologic, and water quality factors (i.e. variables) that may influence algae levels; 3) assess the feasibility of using ANNs with formal optimization for optimizing water treatment processes for reducing algae levels; 4) assess and provide guidance for improving data collection/organization for the purpose of improving algae level forecasting capability with the ANN technology. With regard to item 1 above, as discussed in more detail later, algal forecasting was expanded to include two additional classes; chorophytes and chrysophytes.

Blue-green algal blooms are becoming an increasingly serious water quality concern around the world as they pose threats to both environmental quality and human health. In addition to causing taste and odor problems, on a more serious level, cyanobacteria blooms constitute a potentially serious human health risk. Adverse health effects due to human exposure to water with high counts of cyanobacteria are varied, and include skin rashes from dermal contact and gastrointestinal illnesses from ingestion. At least one human fatality is reported to have occurred, for a person undergoing kidney dialysis treatment. Wildlife such as waterfowl and household pets like dogs can also be adversely affected by these blooms. Algal blooms are expensive to treat, and can translate into thousands of dollars per day for a water utility. In extreme cases, as reported by Maier et al. (1998), they can cause "considerable economic and social hardship with the restriction of the use of water for domestic, agricultural, and recreational purposes and increased treatment costs."

This research was motivated by the premise that development and implementation of an accurate real-time forecasting tool could provide utilities with a means to respond to potential algal blooms proactively, rather than reactively. A proactive capability, if not actually minimizing the risk of bloom occurrence, could at least minimize the consequences, possibly reducing treatment costs while providing higher quality potable water. In this project, artificial neural network (ANN) technology, a form of artificial intelligence, was investigated as a possible forecasting tool. The ANN technology offers 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 and statistical models. However, in order to effectively learn to generalize system behavior, ANNs also require sufficient training data that covers the expected range of conditions, as well as inclusion of important predictor (i.e. causal and/or correlative) variables. The highly complex nature of the algal surface water systems, which are reported as exhibiting non-linear behavior (Recknagel and others, 1997), increases the need for historical data sets of sufficient quality and quantity. 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.

At the beginning of this project, New Jersey American Water was the sole participating utility, with the Swimming River facility serving as the test case. A preliminary research phase completed in August of 2004 used water quality and physical data collected by the utility, supplemented by weather data, to assess the feasibility of using artificial neural networks for predicting and enhancing understanding of algal blooms. Although the data sets were rather limited, particularly with regard to water quality parameters, the ANN technology demonstrated promise, and they outperformed linear models that were used for comparison. Still because of the limited data set, Passaic Valley Water Commission (PVWC) was invited to participate in the study to facilitate a more comprehensive study. Like American Water, PVWC routinely samples for a variety of parameters on their surface water supply system, including algae counts. The fact that their system consists of multiple water sources, while complicating the modeling exercise, also provided a means

for better understanding possible important predictor factors that may also be correlated with watershed conditions. As with Swimming River, data collected by the utility was supplemented by weather data from neighboring climate stations.

Consequently, in this comprehensive study, NOAH developed, tested, and assessed several hundred ANN models of a variety of types to forecast cyanobacteria counts at both Swimming River and PVWC. Many different models were developed, characterized by the input variables used, the forecasting horizons, the output variables, and for PVWC, the sampling locations. In addition, for PVWC, two additional algae groups were also modeled; chrysophyta (gold algae) and chlorophyta (green algae). These additional algae classes were added to the study as they exhibited more temporal fluctuation in counts than cyanobateria, and also present water quality and treatment concerns. Later, ANN models that classify algal counts within ranges or bins were also developed and tested for the PVWC facility. Finally, ANNs for predicting finished water quality were also developed using water quality and treatment data for the since renovated water treatment plant at Swimming River.

Based upon the modeling results for Swimming River, it was concluded by both NJDEP and NOAH that the ANN models were "over-fitting" the data. This is attributed to the relatively limited quantity of historical algae data available for this facility. Although there was a relatively large set of physical data for this facility (222 events), there were only 48 events that included water quality parameters that often serve as "limiting nutrients", including phosphorous and nitrate. This quality of this limited data set was further diminished by the fact that three station locations were modeled collectively rather than individually, due to the scarcity of the data. In addition, the modeling results tentatively indicate that for this facility, the limiting nutrients are important for accurately forecasting algal counts.

Another challenge in modeling this type of the system is the inherent impreciseness of the measured data values, and the uncertainty of how representative they are of system

conditions. Physical, chemical and biological variables in surface water systems often exhibit high spatial variability (i.e. heterogeneity) over relatively short horizontal and/or vertical distances. Further adding unquantifiable noise to the biological data are the analytical techniques used for measuring algae counts. As reported by PVWC: "Phytoplankton analytical methods are time consuming and subjective relying on observation and identification skills, which can vary by analyst. Existing sampling techniques are time consuming and subjective. Collecting a representative sample of a complex watershed is challenging at best. A one-liter sample is collected at some frequency typically once or twice per week, possibly at different depths along the water column and this is meant to represent the quality of the river." Previous literature cites that "precision of the cell count data can be ±20 percent or more" (Maier and others, 1998).

Modeling approaches for PVWC evolved to address important modeling issues that emerged during this project, including relatively limited number of historical data events, following submission of the first draft report to DEP for review. They included: 1) natural time lags in algal population dynamics, or temporal correspondence between system conditions/inputs and final measured algal counts; 2) importance of larger training data sets on model performance; 3) importance of select water quality variables on model performance; 4) the effect of so-called "correlative" water extraction input variables on model performance, and 5) feasibility of developing radial basis function nets for predicting the bin or classification range in which the final algal count falls, rather than a single numerical value estimate. The different approaches increased understanding of the surface water system and algal population dynamics, as well as enhanced awareness of data collection and design and development issues for the ANN forecasting models.

With regard to issue one above, the first or ***original*** ANN input structure or input approach consisted of model input values measured primarily at the beginning of the prediction period. The second or ***revised*** ANN input structure used model input values measured primarily at the end of the prediction period, coinciding with the final or

predicted algal count. For PVWC, the number of historical ***training*** events available for each forecasting problem ranged between 19 and 136, and averaged 65, far less than the minimum required number of 200, computed on the basis of the average number of input (i.e. 32) and output (i.e. 1) variables. To address the second data quantity versus quality issue, both the original and revised modeling approaches were assessed with two distinct data sets. The first set consisted of the smaller number of historical events, which included a higher number of input variables. The second set consisted of a larger number of historical events, and, by excluding five select water quality variables for two stations, and four for the third station, fewer model inputs. The five water quality variables, which included some of the so-called limiting nutrients, are: Biological Oxygen Demand (BOD), Total Phosphorous/Orthophosphate, Nitrite/Nitrate, Sulfate, and Total Organic Carbon (TOC), with BOD included for the third station (measured at higher frequency).

In general, both types of ANN models performed well during validation, and in many cases, accurately predicted large changes in algal populations. The level of accuracy was surprising, given the complexity of algal populations, their non-linear behavior, the expected noise in the data, and the relatively small number of historical events available for training. On the basis of validation correlation coefficients, the models that used input values measured at the beginning of the prediction period slightly outperformed those that used input values measured at the conclusion of the prediction period, with average values of 0.72 and 0.69, respectively. A more subjective visual comparison of the time-series for the validation figures appears to confirm that the original models did achieve higher performance. The models that excluded the select water quality variables for the benefit of more training events achieved a higher average validation correlation coefficient of 0.77, versus the 0.63 average value for models that included these variables. However, there was at least one case where the models that included the five select water quality variables achieved significantly higher validation performance.

That the ANN models developed with inputs measured at the beginning of the one-week and two-week ahead prediction periods accurately predicted formation and dissipation of

algal bloom events, as well as relative increase and decreases, may indicate that there are natural time lags between system conditions and algal population responses. Thus, algal populations may on average evolve predictably in response to system conditions, and the trajectory of the population over one and two-week forecast periods can be accurately forecasted on the basis of real-time measurements. This may also reflect that open water conditions do not typically change significantly in the short-term (e.g. weekly or even bi-weekly), and thus evolving algal populations are by nature predictable, and not prone to diverge significantly from trajectory paths. For example, because of the high specific capacity of water, significant temperature changes will not typically occur over one or even two-week prediction periods. Exceptions may occur with a particularly extreme weather event, such as a cold front or heavy precipitation event, which may also induce large water quality changes, but this will be atypical. Additional research is necessary to test the validity of these claims. Furthermore, the validity of any general rules concerning algal population dynamics will likely vary somewhat from system to system, or even over time as watershed characteristics change. However, one would expect certain fundamental behavior to remain fairly consistent.

That the ANN models that excluded the water quality variables on average slightly outperformed models that included them may signify less about the influence of these variables on algal population dynamics, and more about the inadequate number of training events. At the same time, it does indicate that during most time periods in the PVWC system, these variables may not be important predictors of algal populations, suggesting that they usually exist within a range of values that neither dissipate nor propagate algal blooms. This is weakly supported by the Swimming River results, where significantly better results were achieved when water quality variables were included, even though this meant a five-fold reduction in the number of data events.

It should be recognized that for some time periods, inclusion of some or all of these variables may be important. There was at least one case where inclusion of these variables significantly improved validation performance. A comparison of time-series

between nitrite/nitrate concentrations and algal counts suggest at least a correlative if not causal relationship, where extreme blooms events occurred during periods of high concentrations. Thus, with an adequate number of data events for training, inclusion of the select water quality variables could improve overall performance, and may increase the likelihood of forecasting the formation and dissipation of bloom conditions during unusual conditions.

A modeling concern that arose towards the end of the project was the effect of so-called "correlative" variables, namely volumetric extractions from water sources, on ANN forecasting performance. It was speculated that these variables may be more correlative than causative in nature, as they often reflect operational decisions instituted by PVWC in response to measured algal counts (e.g. discontinue extractions from a river during a bloom event). If indeed more correlative, it would not be appropriate to input correlative variables into models that are designed to capture the underlying mechanistic processes that govern algal population dynamics. To investigate this, volumetric extraction variables from various water sources were eliminated as input variables for select models. It was found that the ANNs still performed relatively well, particularly for the two river sampling stations where mixing of various source waters does not occur.

Finally, a different ANN forecasting paradigm was developed and tested for PVWC. Radial basis function (RBF) nets were developed to predict the pre-selected bin (i.e. (data range group or classification) within which the final measured algal count fell. Given the combination of inherent data noise and forecast uncertainty, this approach may be more appropriate when the goal is to classify water quality within ranges of conditions with relatively high probability for success. In consultation with NJDEP and PVWC, four different bins or classifications were selected for eight different modeling exercises. The ANNs, despite the limited number of historical events available for model development, performed extremely well, even for cases where significant variations in measured counts were present. For three of the eight cases, perfect classification was

achieved, with the poorest performing RBF model classifying 83 percent of the final counts into their correct bin.

The LMs, developed as an objective benchmark for comparison, did not perform as well, on average achieving significantly lower correlation coefficients and higher mean absolute errors, and in some cases, failed to predict very high count algal blooms while erroneously predicting other blooms during low count periods. It should be mentioned that the statistical distribution of input variables were not computed in this study, and consequently, data transformations performed (e.g. log normal). Had statistical transformations been made, it is possible that the LMs would have performed better. However, this also underscores one of the inherent advantages of ANNs; because of their universal non-linear modeling capability, they are not limited by the form of the data distribution(s).

To facilitate an understanding of the system and help assess model performance, three general types of sensitivity analyses were conducted. For the first analysis, the ANN models confirmed physical intuition that the three stations generally represent distinct water quality systems. This was determined on the basis of the higher model performance achieved when modeling each station individually, rather than collectively, although the later approach generated significantly larger data sets.

For the second analysis, a general ranking methodology identified the most important model predictor variables on the basis of the relative changes in root mean squared errors when the variables are excluded as inputs. Although this type of sensitivity analysis is suspect when data events are relatively spare, some consistent trends did emerge. Variables that ranked highly included reservoir and river extractions, which were presumed to be correlative rather than causal variables. Another finding was that the select "limiting nutrients" excluded from some models did not generally rank highly. In contrast, some weather/meteorological conditions like length of day did consistently rank high, with the exception for the models that excluded water extraction variables.

For the third sensitivity analysis already discussed above, exclusion of select variables was performed. First, the importance of the less frequently measured water quality variables was assessed by excluding them from various models. Statistically, there is an increase in predictive performance when these variables are excluded. Again, this may have more to do with the relative lack of historical events, than any physical relationship or lack thereof between these water quality parameters and algal populations. Still, on the basis of the other sensitivity analysis, where these variables did not generally rank highly, that during most time periods, they exist within a range of conditions that neither diminish nor propagate algal populations. Second, exclusion of the water extraction variables, as previously mentioned, was assessed for various models, and results indicate that the ANN models are not overly biased by possible correlations between utility operational decisions and algal counts.

This work provides a foundation for future modeling work as more data become available. Systematic elimination of input variables combined with additional sensitivity analyses may improve system understanding, and facilitate convergence to optimal sets of model input variables. Not only would increased ANN forecasting accuracy be achieved, but data collection strategies could be improved, which may even reduce sampling costs and efforts for the utility. Following ANN model refinement, implementation of various forecasting models in real-time would provide a more rigorous assessment of forecasting capability, as well as a comparative analysis of their performance, strengths, and limitations, which may further improve modeling methodology and forecasting capability.

For the water treatment modeling component of this project, significantly more data was available for model development, assessment, and refinement. The ANN models overall performed well in predicting finished water quality conditions, and even confirmed a non-intuitive relationship between raw water quality and temperature. However, consultation with utility personnel is necessary to assess whether the ANN models, as developed, are sufficiently robust for optimizing water treatment processes. At a

minimum, this modeling effort demonstrates how ANNs could possibly be used for water treatment optimization, and also affords the opportunity for future improvements and enhancements in accordance with utility guidance.

Because of the large number of models that were developed and assessed (several hundred), this document overviews the important general findings of this project and discusses representative results and important conclusions. To augment future efforts, and provided interested readers with additional information, an extensive collection of summary tables and figures for other ANN models are provided in the appendices. The presentation of algae modeling, results, and analysis for both Swimming River and the documentation and discussion of the original PVWC modeling efforts also appear in the appendices. Finally, the modeling data used in this study is also available in digital format.

1. INTRODUCTION

This document constitutes the final report for the algae study, entitled "Forecasting Algal blooms in Surface Water Systems with Artificial Neural Networks," funded by the New Jersey Department of Environmental Protection, with participating utilities New Jersey American Water Company and the Passaic Valley Water Commission. The objectives of this study were: 1) to assess the feasibility of using artificial neural networks (ANNs) as a real-time tool for accurately forecasting cyanobacteria counts, more commonly knows as blue-green algae, in surface water systems; 2) through sensitivity analyses conducted with the ANNs, identify critical climate, hydrologic, and water quality factors (i.e. variables) that may influence algae levels; 3) assess the feasibility of using ANNs with formal optimization for optimizing water treatment processes for reducing algae levels; 4) assess and provide guidance for improving data collection/organization for the purpose of improving algae level forecasting capability with the ANN technology. With regard to item 1 above, as discussed later in this report, the algal forecasting was expanded to include two additional classes; chorophytes and chrysophytes.

For the initial phase of this project, the Swimming River reservoir system, owned and operated by New Jersey American Water, was used as a test case for assessing the feasibility of using ANNs for forecasting algae levels of cyanobacteria based upon weather/meteorological, hydrological, biological, and water quality conditions. Despite a relatively limited data set, particularly with regard to water quality data, the ANNs demonstrated potential as a useful algae level forecasting tool. The ANNs not only outperformed linear models, but surprisingly, when limited water quality data were included, accurately reproduced variable algae levels. A concern, however, was the relatively limited data events (48) that included water quality data, which is supported by relatively low validation accuracy. Given the relatively high number of input variables, and the possible range of conditions, it was concluded that additional data would be necessary to develop robust ANN forecasting models for the reservoir. However, as discussed in the initial report (August 4, 2004), additional follow-up modeling was still to

be conducted with this data set to further assess ANN model capability, and help identify potentially important predictor variables for possible future model refinement and implementation. Following review of the algal modeling results by both NOAH and DEP personnel, it was determined that the ANNs were likely "over-fitting" the state-transition equations to the limited data. Consequently, the detailed description and findings of this algal modeling work for this facility appears in Appendix A-1, and can be obtained from the NJDEP in disk form upon request.

The Swimming River facility also provided NOAH with a rather extensive water treatment data set, consisting of daily measurement values for a variety of parameters, spanning approximately 2.5 years, from July 2001 to December 2003**.** This data set, consisting of raw water quality conditions, chemical dosages, and final water quality conditions, was used for assessing the feasibility of using ANNs for optimizing treatment processes. Thus, in accordance with study objective 3 stated above, the feasibility of using ANNs for accurately predicting finished water quality following treatment was investigated, so that their potential for reducing treatment costs via optimization while achieving water quality standards could be assessed.

Finally, to facilitate algae study objectives, the Passaic Valley Water Commission (PVWC) was invited to serve as a second test case. This utility's participation provided at least two additional benefits. First, unlike Swimming River, where relatively few sampling events precluded individual modeling of sampling locations, a larger set of historical events at the PVWC utility permitted individual modeling of sampling stations. This site specific modeling issue may not be as relevant for Swimming River, as the three sampling locations in the reservoir essentially represent the same watershed/water sources, with perhaps some variability due to wind and hydraulic dynamics. At the PVWC utility, however, two sampling stations are located on two different rivers, each part of a different watershed with distinct water quality conditions. In addition, occasional diversions from a storage reservoir can induce unique water quality changes at

the third sampling location, which is usually a combination of the two river sources. Second, the PVWC routinely samples for a higher number of water quality indicators.

For the PVWC test case, in addition to cyanobacteria, ANN modeling efforts were extended to two other algae phyla/divisions: chrysophyta and chlorophyta. Discussions with Ms. Linda Pasquarello of PVWC determined that it would be helpful if the utility could also forecast levels for these two algae phyla/divisions. Although not posing the potential health problems as cyanobacteria, their elevated presence can create undesirable taste and odor problems for consumers, as well as clog water treatment filters. The capability to forecast possible high levels of these organisms, then, would allow the utility to implement proactive measures to minimize potential negative effects. In addition, over the period of records for the PVWC, these species exhibited greater temporal variability in counts than cyanobacteria, and provided an additional test for ANN forecasting capability.

Consequently, in this final study, NOAH developed, tested, and assessed a variety of ANN models to forecast cyanobacteria counts at both Swimming River and PVWC. Many different models were developed, each differing by the input variables used, the forecasting horizons, and, for PVWC, the sampling locations. In addition, for PVWC, two additional algae groups were also modeled; chrysophyta (gold algae) and chlorophyta (green algae). These additional algae classes were added to the study as they exhibited more temporal fluctuation in counts than cyanobateria, and also present water quality and treatment concerns. ANN models that classify algal counts within ranges or bins were also developed and tested for the PVWC facility. Finally, ANNs for predicting finished water quality were also developed using water quality and treatment data for the since renovated water treatment plant at Swimming River.

Modeling approaches for PVWC evolved to address important modeling issues that emerged during this project, following submission of the first draft report to DEP for review. They included: 1) natural time lags in algal population dynamics, or temporal

correspondence between system conditions/inputs and final measured algal counts; 2) importance of larger training data sets on model performance; 3) importance of select water quality variables on model performance; 4) the effect of so-called "correlative" water extraction input variables on model performance, and 5) feasibility of developing radial basis function nets for classifying algal counts into bins or ranges, rather than a single numerical value estimate. The different approaches increased understanding of the surface water system and algal population dynamics, as well as enhanced awareness of data collection and design and development issues for the ANN forecasting models.

Two different ANN model paradigms were used in this project; prediction of a single numerical algal count value using traditional multilayered perceptron nets, and classification of algal counts into bins or ranges using radial basis function (RBF) nets. The single value predictions used two general input variable structures or approaches. The first or original approach consisted of model input values measured primarily at the beginning of the prediction period. The second or revised approach consisted of model input values measured primarily at the end of the prediction period. Modeling performance was further assessed using two distinct data sets; a larger data set (i.e. more historical events) generated by eliminating select water quality variables that were less frequently measured, and a data set that included these water quality variables, resulting in fewer historical events, and more input variables. In addition, the influence of so-called correlative water extraction variables was assessed by including and then excluding for select forecasting problems. Sensitivity analyses were also conducted to try and help identify important predictor variables and increase system understanding. The study permitted a systematic analysis of available modeling and data collection options for performing real-time forecasting, particularly with limited data sets, with a general discussion of their comparative advantages and disadvantages, as well as their feasibility under real-time conditions. In addition, the multitude of approaches permitted a more general evaluation of ANN capability, as well as helping to cross-verify results and interpretations.

Despite relatively few historical events available for ANN training and validation, ANN models in general demonstrate the potential for accurately forecasting algal population counts. Different model types generally performed well during validation, accurately reproducing higher and lower algal count periods during validation, and even accurately predicting the incidence and magnitude of a number of blooms during validation, including dissipation. The original model input approach performed slightly better than the revised input approach, and larger training sets that excluded the less frequently measured water quality variables generally improved overall performance for the ANN models that explicitly predict single algal counts. The opposite result occurred for the few classification models developed and assessed, though larger data sets (i.e. eliminating five water quality variables) produced more events that border on two bins; still, they demonstrate strong potential for accurately differentiating between low count periods and algal bloom events.

ANN models because of the ease at which they can be implemented in real-time have the potential to serve as a powerful proactive management tool for utilities seeking to minimize the problems posed by algal blooms. By accurately forecast the occurrence of problematic algal blooms, they can provide utilities with sufficient lead time to implement cost-effective mitigative measures. These forecasting models also provide value added to expensive data collection systems, and may even be used to optimize sampling strategies, potentially reducing costs. As more data becomes available, the ANN models can relatively easily be retrained in real-time to increase forecasting capability.

The entire report, including appendices, presents the following ANN modeling work, with select linear models included for comparison; 1) additional cyanobacteria forecast modeling for the Swimming River facility; 2) modeling of finished water quality conditions at the Swimming River facility due to treatment processes; 3) individual forecast modeling of cyanobacteria, chrysophyta, and chlorophyta counts at the PVWC facility. In addition to presenting and assessing modeling predictive results, sensitivity

analyses are also presented, and are analyzed within the context of the physical system being modeled.

Because of the large number of models that were developed and assessed during this research, most of the results are presented in the appendices, including some detailed analysis. Appendix A-1 includes a description of the ANN and LM modeling approaches, results, and analyses pertaining to algae prediction for Swimming River. Appendices B-1 through B-6 generally present additional ANN and LM modeling results in tabular and figure form, with some text describe modeling efforts and results not presented in this main report document. Appendices C-1 through C-3 present additional water treatment modeling results and analyses for the Swimming River facility not presented in this report (e.g. seasonal models). Please note that the appendices as well as all data used in this project are available in electronic format upon request from NJDEP.

Lastly, because of water security issues, detailed facility site maps were not be provided in this report. The schematics and text use arbitrary naming conventions for the rivers and the reservoir to protect the identity of the water source and intake point locations.

2. ALGAL BLOOM MODELING PROBLEM

Although there is consensus among scientists that the incidence of algal blooms world-wide is increasing (Smith et al, 2006), and the detrimental effects of these blooms on the environment and water supplies are well documented, controversy remains over the important factors and mechanisms responsible for their occurrence, and the most effective means for both modeling and forecasting this phenomena.

Researchers have identified fundamental "nutrient" variables, primarily nitrogen and phosphorous, as "limiting" to the growth of algae. Because nitrogen and phosphorous are often strongly correlated with the quantity of algal biomass in water systems, researchers frequently develop models that predict algal biomass as a function of one or more of these compounds. Limiting the models to such highly reduced input-output relationships, while efficient and sometimes effective, not only ignores the complexity of processes that determine nutrient levels, but also overlooks the myriad of other factors that can influence algal biomass. For example, nitrogen may originate from a number of different sources, and its chemical form and concentration is dictated by different nitrogen processes, such as nitrification, which depends upon the presence of certain bacterial organisms. In some water systems, nutrients that might otherwise be considered limiting on the resident algal populations appear to persist within a range of concentrations that does not significantly affect the organisms.

Further complicating the dynamics of algal populations are the variety of physical and biological factors that influence the formation and dissipation of algal blooms. Sunlight is essential for the development of these photosynthetic organisms, and the amount of light that penetrates the water column is controlled by a number of factors, each of which may have multiple effects upon the system. For example, precipitation not only reflects a lower sunlight factor, but also influences the amount of turbidity in the water column via sediment transfer from rainfall run-off. The degree of run-off depends not only upon the quantity of precipitation, but also the size and characteristics of the watershed.

Precipitation also increases surface water flow velocities which can stir up and suspend sediments from the bottom and scour sediments from the banks. Other factors that influence algal levels, like dissolved oxygen are similarly affected by other conditions of the system, such as water temperature, wind speed and direction, and the presence of other competing or even predatory organisms.

Given the multitude, interplay, and complexity of various weather, water quality, biological, and hydrologic factors, many of which will vary over space and/or time, some researchers argue that there are no fast and true "rules" for predicting algal biomass and blooms. As with many complex natural phenomena, developing a robust model that adequately represents the site-specific conditions and dynamics of the system, specifically for the purpose of providing accurate prediction capability, can be a daunting challenge. It is undoubtedly true that some algal biomass models, while appropriate for some systems, will be inappropriate and/or infeasible for others, dictated largely by the degree to which the fundamental assumptions and characteristics of the model conform to the essential elements of the real-world system, as well as the quality and quantity of the data. Some researchers argue that reliance upon numerical models for predicting algal biomass is insufficient, given our inadequate "'level of understanding of how these complicated ecological systems work'" (Pelley, 2005).

The challenge then, is not only to increase our understanding of the complex dynamics that govern algal populations, but to develop different sampling protocols and models that can be adapted to our scientific understanding for improving real-time forecasting capability. This goal is becoming more important as algal biomasses around the world become more prevalent, contaminating water supplies, often in areas where acceptable alternative supplies to not exist. Thus, this research not only attempts to provide insights into algal populations in the surface water system considered here, which can perhaps be extrapolated to other similar systems, but to also provide an alternative paradigm for researchers and decision-makers seeking to model and forecast these organisms.

3. RELATED ANN RESEARCH

There is previous work in the scientific literature where ANNs were developed and tested for predicting algal blooms, as reported in two journal articles, both published in the Journal of Ecological Modeling. Both papers have researchers associated with the University of Adelaide in Australia, with the first published in 1997 and the second in 1998.

The first paper, "Artificial neural network approach for modeling and prediction of algal blooms" (Rechnagel and others, 1997) applies the technology to four different freshwater systems. The paper first introduces the complexity and non-linearity of algal bloom dynamics and the severe impacts associated with them, such as water discolorations and human exposure to toxins.

The researchers used at least six and up to ten years of what appear to be weekly data consisting of limiting nutrients, water temperature, light conditions, and, in one case, density data of zooplankton groups to train the ANN models to predict phytoplankton organisms. Two years of data not used for training were then used to validate or test each model. Typical ANN input variables included silica, total nitrogen, turbidity, color, water temperature, conductivity, pH, ortho-phosphate, nitrate, photosynthetic active radiation, wind speed, secchi depth, and oxygen. For one lake system, some of the grazing organisms, namely Rotifera, Cladocera, and Copepoda, were also used as inputs. For the sole river system, flow rate was considered an important predictor variable, and was included in the model. The output variable for a single ANN consisted of all the different algal organisms measured in the water body. As the authors state, the variables used represent those that were measured consistently over time.

In general, the ANNs were able to predict the timing and magnitude of different blooms, although there were some validation years where the ANNs did not perform particularly well. The ANNs were also used to conduct sensitivity analyses to assess the relative

importance of different predictor variables. The authors conclude that the results "provide a means of defining the primary components driving they dynamics of algae species."

In the second paper, "Use of artificial neural networks for modeling cyanobacteria *Anabaena spp*. in the River Murray, South Australia" (Maier and others, 1998), seven years of weekly data, consisting of eight input variables, was used to provide forecast algae counts four weeks into the future. This river is the same used in the study presented in the first paper, and the predictor variables consisted of color, turbidity, temperature, flow, soluble phosphorous, total phosphorous, oxidized nitrogen, and total iron. The four nutrient variables were collected on a monthly basis, but linear interpolation was used to estimate weekly values. With the exception of flow, which was measured on a daily basis, the remaining variables were measured on a weekly basis.

One issue explored in this research was the importance of using lagged inputs. The researchers concluded that models that did not use lagged inputs were "relatively unsuccessful", particularly with regard to flow. Inclusion of this variable allowed the ANN models to forecast the onset and duration of algal blooms. Different combinations of input variables were used, including models with just a single input. It was found that the most important variables were flow, temperature, and color for predicting blooms. Other variables, namely nutrients, iron, and turbidity, were not important for predicting blooms. The authors concluded that adequate nutrients were available for algal growth, and hence were not limiting factors for the river system studied.

One of the most relevant findings within the context of the NJDEP sponsored study is that the ANN models were unable to predict one particular algal bloom event. The authors attribute this bloom to an unusual hydrodynamic event, where a large flood transported algae populations from shallow lagoons connected to the River. Thus, because many of the algae did not originate on the river system, this was not a true "bloom" event, but rather a mixing of two different populations. Consequently, not only

was this event outside the physical bounds of the "closed" system, there were insufficient representations of this anomaly for the ANN models to effectively learn this extreme condition, which underscores the need for data that represents the expected range of system behavior.

With respect to water treatment, numerous sources in the literature document successful ANN applications. In the work of Skipworth and others (1999), ANN technology was demonstrated to accurately predict oxidation-reduction potential under different patterns of demand and using different sources at multiple locations within a water distribution system. Yu and others (1999) accurately predicted proper coagulant dosing for a water treatment plant, significantly outperforming a regression model. Mirsepassi and others (1995) used ANNs to accurately determine alum and polymer dosages, achieving correlation coefficients for both parameters of 0.97 (between measured and predicted values) for a water treatment plant in Australia.

Baxter and others (2001) present the application of ANN technology for optimizing several treatment processes in a large water treatment plant. The first application presented is for optimizing color removal through coagulation. The ANN achieved a mean absolute error (MAE) (< 0.32 TCU) less than the instrument error used to measured color in the clarifiers. Second, an ANN model was developed for turbidity removal via coagulation, using raw water quality, operations, and dose information. For the same process, an inverse model was developed, where the turbidity becomes an input to predict the alum dosage required for treatment. The models not only achieved low errors, but also provide insights into particulate removal by enhanced coagulation, and are being used in real-time to enhance operational decisions. The third application was for water softening, where ANNs were developed to estimate the total hardness in a softening clarifier effluent and the softening lime dose requirements. High predictive accuracy was achieved with the ANN models, with correlation coefficients for the validation data ranging between 0.84 and 0.95. Lastly, an ANN model was developed for predicting filter effluent counts following filtration. As stated in the paper, "there is currently a

weak understanding of relationships among particle count removal, chemical usage, and raw water quality, since the relationships are complex and nonlinear." The model achieved a mean absolute error for the validation data of 2.3 counts/ml, with a correlation coefficient of 0.79, demonstrating "excellent predictive capacity on previously unseen data."

In summary, limited research in the literature demonstrates the potential utility of ANN technology for predicting algal blooms and optimizing water treatment processes. However, every site, particularly in regard to algae population dynamics, will offer different challenges, and perhaps the most daunting challenge with algal bloom forecasting is implementation of the model in real-time conditions. That is, the current research has used historical data to test or validate the models, where the model inputs are known a-priori. The real challenge is to develop models that can accurately forecast water quality conditions in real-time using available measured data and forecasted conditions (e.g. weather).

4. ARTIFICIAL NEURAL NETWORKS

ANN technology is a compelling alternative to the physical-based modeling approaches, and as discussed previously in the Related Research section, has been used with success for algal bloom prediction problems (Recknagel and others, 1997; Maier and others, 1998). An ANN, through proper development and training, "learns" the system behavior of interest by processing representative data patterns through its architecture. What sets an ANN apart from a physical-based model is that because it does not rely upon the governing physical laws, information regarding physical parameters is often not required for its development and operation.

Because of its empirical nature, ANN technology is sometimes erroneously referred to as an "advanced" type of regression analysis. What distinguishes ANN technology from regression is the famous Kolmogorov's Theorem (Hecht-Nielsen, 1987, Sprecher, 1965). Specifically, this theorem asserts that any continuous function, from R^m to R^n, can be represented *exactly* by a three layer feedforward 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. By contrast, regression is guaranteed to provide only an approximation by computing the best fit from a given function family. In addition, unlike regression, which treats all output variables independent of each other, the presence of common arcs in the ANN architecture allows it to identify important inter-relationships that may exist between output variables.

Figure 1 depicts a sample three-layer feedforward ANN architecture. Each ANN layer consists of individual nodes (elements), and the nodes are interconnected across layers by special non-linear (usually non-rational) transfer functions, expressed in terms of the nodal input variables and connection weights. During training, data patterns are processed through the ANN, and the connection weights are adaptively adjusted until a minimum acceptable error between the ANN predicted output and the actual output is

achieved. It is at this point that the ANN has "learned" to predict the system behavior of interest (i.e. values of output variables) in response to the values of the input variables.

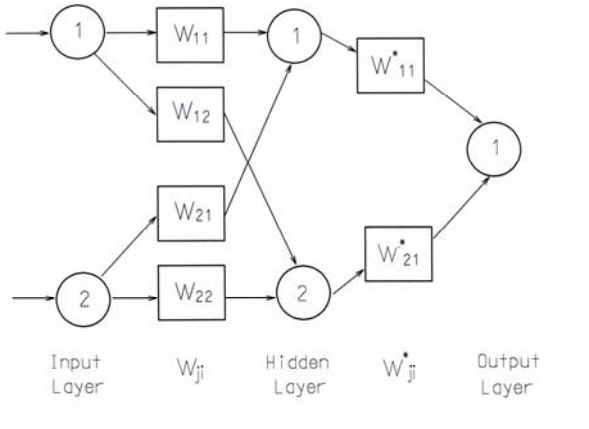


Figure 1. Architecture for a simple multi-perceptron ANN

In this work, the commonly employed non-linear hyperbolic tangent transfer function,

$$f_j(Sum_j) = \frac{e^{Sum_j} - e^{-Sum_j}}{e^{Sum_j} + e^{-Sum_j}}\ , \qquad \textbf{(1)}$$

was used, where Sum_j represents the weighted sum for a node in the hidden layer, and e denotes the basis of the natural logarithm. In Sum_j, the input value received by each node in the hidden layer is multiplied by an associated connection weight, whose value is identified during learning. This weighted sum can be formally represented as:

$$Sum_j = \sum_{i=1}^{n} w_{ji} x_i + w_{jb}, \qquad \textbf{(2)}$$

where w_{ji} represents the connection weight between the ith node in the input layer and the jth node in the hidden layer. The input x_i is known, and represents the values of the input variables for node i in the input layer. A bias unit, which helps to provide numerical stability, is merely added as the connection weight w_{jb} because it has a constant input value of 1.0.

There are various kinds of ANN learning algorithms, and the interested reader is referred to the work of Poulton (2001) for more details. In this work, a combination of back

propagation and conjugate gradient algorithms were used. The prediction accuracy of an ANN is measured by the mean squared difference between the actual and predicted output values. For a preselected ANN model and corresponding data set, this mean squared error depends only on the values of the connection weights. During learning, the ANN processes training patterns consisting of input-output patterns through the network, systematically adjusting the connection weights, so that the measure of the overall goodness of the ANN model defined as the root mean squared error (RMSE) between the ANN-estimated output values and the actual values, is minimized. The minimization learning algorithm is always iterative, and each step is considered "learning".

The RMSE is mathematically defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\gamma_k - C_k\right)^2} \qquad \textbf{(3)}$$

where γ_k is the ANN-estimated/predicted algae count value for the *k*th training event, C_k is its corresponding measured algae count, and *N* denotes the total number of such events.

In selecting the most appropriate ANN model, a variety of factors must be considered. This includes the functional form of the ANN transfer functions, the number of hidden layers and nodes, the most appropriate set of input variables, and the method used to minimize the objective function. 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 variables during initial ANN development requires a basic understanding of the governing system dynamics. However, a sensitivity analysis in conjunction with trial and error can help the modeler converge to the most appropriate 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.

During ANN development in this project, learning proceeded in a series of training and verification steps. The ANN is presented with training data during which patterns are processed through the network, and the learning algorithm adaptively adjusts the network connection weights to minimize the RMSE between actual and estimated output values. Intermittently, the training phase is interrupted, and a separate (verification) data set is processed through the ANN to verify progressive learning, as indicated by a declining RMSE value obtained with the verification data set. Verification guards against overtraining, where the ANN has memorized or over-fitted the connection weights to the training patterns. Training proceeds until the verification RMSE either stabilizes or begins to increase. At this point, ANN training is terminated, and the ANN can now be validated with a third data set not previously used for training or verification. Validation is used to determine whether the ANN has learned system behavior of interest over the range of expected conditions. To provide robust training, verification, and validation, statistically similar data sets spanning the expected range of system behavior were used for ANN development and validation. One half of the available data are used for training, one quarter for verification, and the remaining one quarter for validation or testing.

Being "data-driven" models, robust ANN development is absolutely dependent upon the quantity and quality of the data used to train the models. As discussed by Coppola and others (2003), "appropriate training set size for an ANN depends upon a number of factors, including its dimension (i.e. number of connection weights), the required ANN accuracy, the probability distribution of behavior, the level of noise in the system, and the complexity of the system." Complexity within the context of ANN modeling refers to a system where small changes in model input values produce large and even contradictory changes in model output values. A system that does not exhibit this type of complexity may then be referred to as a "well-behaved" system.

Researchers have developed a number of heuristic equations for estimating the minimum number of training data required for robust ANN model development. One such rule

uses the following mathematical relationship, where this number is a function of the number of ANN input or predictor variables, designated m, and the number of ANN output or prediction variables, designated n:

$$\text{Minimum Number of Required Training Samples} = [(1.5 \times m) + (1.5 \times n)] \times c \quad \textbf{(4)}$$

where c is some constant, typically ranging between 4 and 10. Thus, assuming an average sized model for the algae work conducted for PVWC, consisting of 32 inputs and one output, the summation in the brackets, $[(1.5 \times 32) + (1.5 \times 1)]$, computes to 50 (rounded up). Using the range of values for the constant c, the computed number of training events required for robust ANN training ranges from 200 to 500 sample events.

Because algae population dynamics are highly complex and even considered non-linear in behavior (reference), and, given the expected noise in the data, the number of required training data necessary for robust ANN model development is probably closer to the upper value approaching 500 events. It should not be overlooked that more data also permits validation of the models under a greater range of conditions.

The poor performance for the Swimming River facility is also related to the second data issue, quality, as pertaining to model inclusion of all important causal or predictor variables governing the behavior or outputs of interest. As discussed above, ideally, the modeler should have a well-founded intuitive or empirical if not theoretical understanding of the governing dynamics of the system. This understanding ensures that all necessary predictor variables are included as model inputs. An excellent illustration of this requirement, and in stark contrast to the PVWC case, is that ANN performance for the Swimming River facility declined significantly following elimination of four water quality parameters intermittently sampled for by the utility (ammonia, turbidity, nitrite/nitrate, and phosphorous). This decline in predictive performance occurred even though the sample size for ANN development and assessment increased more than five-fold, from 48 to 221 events.

As part of the quality issue, data should also span the expected range of system conditions. For example, if there is the possibility for algae blooms during cooler months, data collected should include these months, and not only warmer months when blooms occur at higher frequency. While ANNs are excellent interpolators, they generally do not extrapolate well outside of the range of conditions for which they have been trained.

As was done in this research, sensitivity analysis and selective inclusion and exclusion of input variables can help the modeler converge to an appropriate set of predictor variables. However, ideally, the modeler should have a basic understanding of the system, which eliminates the possibility of excluding an important variable, but also promotes a more expedient convergence to a robust model. This understanding can help the modeler assess the potential strengths and weaknesses of the model, and identify situations under which its predictive capability may be suspect, or where inclusion of a certain variable critical. In addition, the modeler can help design a sampling program to include important or potentially important variables that to date have been sample either infrequently or not at all. This type of analysis is an accepted modeling methodology to reduce the dimensionality of the problem, and also eliminate spurious input variables (Swingler, 1996).

Two statistical measures in particular are used to assess ANN performance in this study; mean absolute error and correlation coefficient, mathematically defined as:

$$\text{Mean Absolute Error} = 1/\text{N}\sum_{k=1}^{N}\left|Yk - Ck\right| \quad \textbf{(5)}$$

$$\text{Correlation Coefficient} = \frac{Scy}{\sqrt{SccSyy}} \quad \textbf{(6)}$$

where γ_k is the ANN-estimated/predicted algae count value for the kth event, C_k is its corresponding measured algae count, N denotes the total number of such events, S_{cc} and

S_{yy} denote the sample variances of the final predicted algal counts and final measured algal counts, respectively, and S_{cy} represents the covariance between final predicted and final measured counts. Mean absolute error is a measure of the absolute average difference between final measured and final predicted algal counts. Correlation coefficients range between -1 and +1, with values close to +1 indicating a strong positive correlation between predicted and measured values, and values closer to 0 indicating little correlation. In this analysis, no correlation values were below 0.

At the end of this project, radial basis function (RBF) neural networks were also developed for eight select cases. As discuss in more detail later, instead of forecasting a single count value, as done previously, the RBF nets were developed to predict the bins or classification range (four pre-specified bins or classes) in which the final measured algal counts would fall.

For the RBF nets used for predicting the bin or classification range of the final measured algal counts (into four pre-specified bins or classes), a slightly different learning approach are used. For this paradigm, it is assumed that it input patterns are mapped to a higher dimensional space, there is a greater chance that the problem will become linearly separable based on Cover's Theorem (Cover, 1965; Haykin, 1999). For RBF's, the input pattern is non-linearly mapped to this higher dimensional space through the use of radially symmetric functions (usually Gaussian), with similar input patterns transformed through the same RBF node. The training process starts with an unsupervised phase during which the center and width of each RBF node must be trained. The centers start with random values and for each input pattern, and the center with the minimum distance to the input pattern is updated to move closet to that input pattern. Once the center vectors are fixed, the widths of the RBFs are established based on the root-mean-squared distance to a number of nearest neighbor RBFs. Following completion of the unsupervised phase, the connection weights between the RBF layer and the output layer are trained.

As discussed above, networks with a large number of input variables can be susceptible to over-fitting and a sign of over-fitting is a network function with high curvature. High curvature of the fitted function is a result of large connection weights. One method to reduce the network complexity or curvature is to penalize large connection weights through a process of weight regularization. The method of Weigend et al. (1991) is commonly used, where a penalty term in Equation 7 is calculated and added to the error term and to the connection weight's derivative at each step. Weights with large values are penalized the most and driven to small values. Once the weights are below a user-defined threshold the input variable or the hidden processing element is eliminated if all connection weights attached to it fall below the threshold value.

$$\text{Penalty} = \lambda \Sigma w_i^2 / (w_0 + w_i^2) \qquad \textbf{(7)}$$

λ and w_0 are user specified constants.

For the RBF modeling only, this approach was used, with a threshold value of 0.5 assigned to determine when a variable or processing element should be eliminated during training.

For all modeling conducted in the project, the commonly used modeling software Statistica was used. NOAH has used this software in a variety of projects, including an EPA Water Security Project and groundwater modeling. Dr. Mary Poulton, a NOAH principle and ANN expert modeler, uses this software in her graduate level course at the University of Arizona. Both she in this course and NOAH in its work have benchmarked the software against MATLAB's neural network toolbox and NeuralWare, two other popular ANN software programs, and it has comparatively performed well. In addition, the software was also used for the linear modeling in this project, which served as a benchmark for ANN performance.

5. THE STUDY AREAS

5.1 Swimming River

The Swimming River facility is located in Tinton Falls, New Jersey, approximately 6 miles west of the Atlantic Ocean. The facility has a history of algal blooms, which normally occur in warmer months. The data set used for developing and assessing the ANN methodology included most variables that are considered potentially important for predicting algae levels. A total of 221 weekly events of physical parameters measurements were initially generated from the available data, as well as 48 water quality samples collected intermittently over the period March 2002 to November 2003. Data provided by the New Jersey American Water Company (NJAWC) for the study site included water temperature, dissolved oxygen or D.O., secchi depth, pH level, algae counts per class, Total Phosphorus (µg/L), Nitrite/ Nitrate (µg/L), Ammonia as N (µg/L), Turbidity NTU, Silica (mg/L), Iron (mg/L), and Manganese (mg/L). Much of the data was provided in paper format, which was entered by NOAH personnel, and then underwent QA/QC for transcription accuracy, by an independent assessor. For this preliminary assessment, the later three constituents were not used, as they were measured relatively infrequently. Climate data was obtained from the National Oceanic and Atmospheric Administration (NOAA) and included total daily precipitation, average daily temperature, wind speed, wind direction and sky cover, while solar radiation data were taken from stations operated and maintained by NJDEP. The weather station used for precipitation, wind speed, and direction was Belmar, which is located approximately 8 miles southeast of the study area. For the relatively few days where data was missing, a linear regression was performed using Glendola's and Belmar's data from January 1, 2002 to July 24, 2003, with which missing precipitation values were estimated. Solar radiation data was collected at the Flemington station, located approximately 44 miles from Tinton Falls. In this final phase, additional modeling efforts and refinement were restricted to the 48 events that included water quality conditions. The rationale for this is the most accurate modeling performance was achieved with this data set, even though the

"physical" data set, which did not include any of the "chemical" parameters (total phosphorous, nitrite/nitrate, ammonia, and turbidity), had significantly more data available for training.

The conclusion drawn from these preliminary modeling results, and supported by algae population dynamics, is that inclusion of water quality conditions can be important for developing robust models that can accurately forecast algal levels. As discussed previously, following review of the algal modeling results by both NOAH and DEP personnel, it was determined that the existing data set was insufficient for model development, and the ANNs were likely "over-fitting" the state-transition equations to the data. Consequently, a detailed description and analysis of this modeling effort is not included in this document, but is provided in Appendix A-1, which can be obtained upon request by interested individuals from NJDEP in disk format.

For the water treatment modeling, the data set provided by New Jersey American constituted approximately 2.5 years of daily data consisting of 60 variables, of which 57 were used as model inputs, with the remaining 3 serving as outputs. The outputs were**:** average daily turbidity, highest daily turbidity, and average daily number of readings above 0.1 NTU.

The variables used in the water treatment problem can be classified into four basic groups: physical water data, chemical water data, water treatment data, and weather data. For example, physical water data includes variables such as total daily flow and average water temperature. Chemical water data include average daily measured turbidity, pH, and average daily chlorine levels. Water treatment data was the quantity or doses of chemicals added for treatment, such as sodium hydroxide and hydryfluosilicic acid.

A variety of different modeling exercises were conducted for the water treatment problem. However, because of length concerns, only the most important and representative modeling work and results are presented in this document. Similar to the

Swimming River algal modeling, the complete description and analysis of this modeling effort is provided in Appendices C-1 through C-3, which can be obtained upon request by interested individuals from NJDEP in disk format.

5.2 Passaic Valley Water Commission (PVWC)

Passaic Valley Water Commission (PVWC) is a wholesale and retail provider of drinking water to over 750,000 customers in Northern, New Jersey. The Commission owns and operates the Little Falls Water Treatment Plant (LFWTP), located in Totowa, New Jersey. The treatment plant was recently upgraded to include high rate sand-ballasted flocculation, ozonation and biological filtration to provide multiple barriers of microbial protection and to comply with recently implemented Safe Drinking Water Act Regulations.

Algal blooms result in two basic types of treatment challenges: effective particle settling/removal and objectionable tastes and odors. PVWC has an extensive watershed water quality monitoring program in place to assist with decision making for source water selection and prediction of water quality changes. Grab and online sample data are supplemented by USGS flow and water quality monitoring stations located throughout the watershed. The existing algal monitoring program consists of analyzing key water quality parameters and correlating changes in concentrations to predictions of algal concentrations. Routine direct measurement of the odorants Methyl Isoborneol and Geosmin, both algal metabolites, at key watershed, plant and finished water locations provide additional guidance for source water selection, in particular to minimize algal related taste and odor events. However, PVWC is continually exploring innovative methods to provide early warning of changes in water quality, and in particular, presence of algal blooms with sufficient response time to adjust treatment as required for mitigating negative water quality impacts.

Figure 2 provides a schematic representation of the system used in this project, with actual site location names omitted for water security reasons. The figure depicts the two possible river sources and a surface water reservoir used to supply the utility with its source water, which enters the LFWTP intake point for treatment. River B is gravity fed to the treatment plant intake. River A flows into and becomes part of River B upstream of the LFWTP intake, but can also be directly pumped to the plant intake via Pumping Station 2, thereby influencing the blend of River A being treated at the plant. A third surface supply, Reservoir A, is available on a limited basis to dilute any negative water quality impacts that may be present in Rivers A and B. Reservoir A is recharged with water from River A via Pumping Station 1 as needed to maintain adequate raw water storage levels, and in times of abstraction, is delivered directly to the plant intake location in the canal.

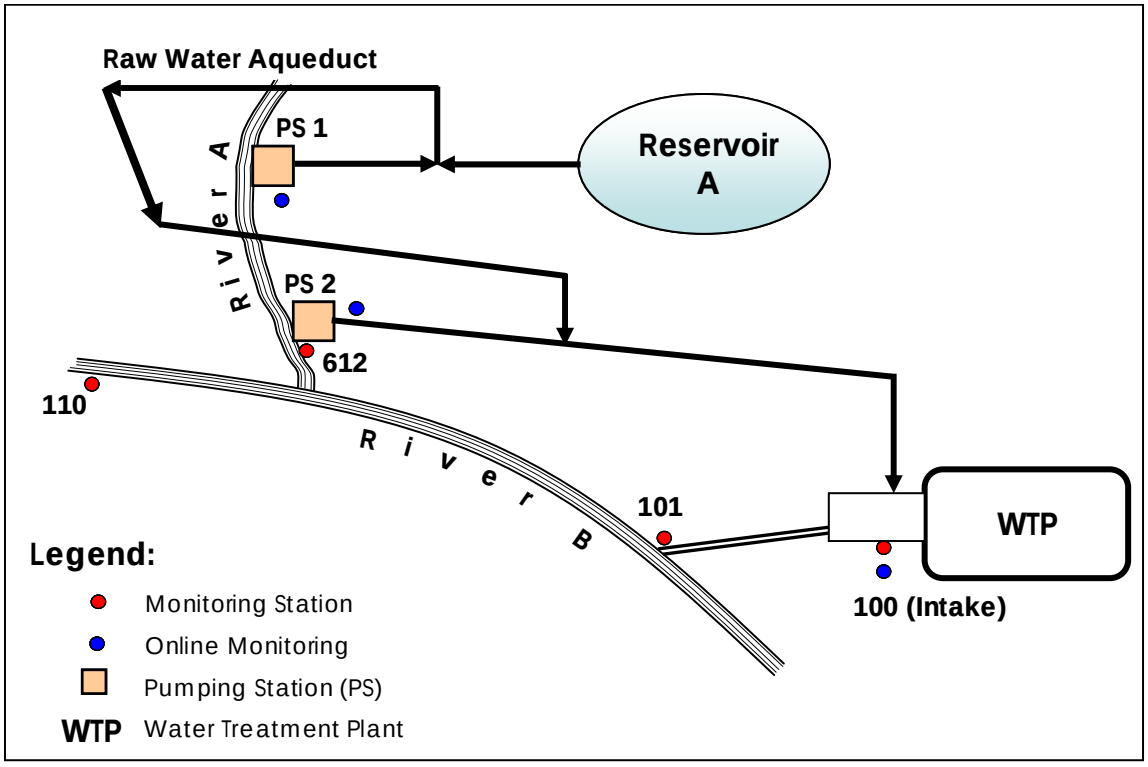


Figure 2. PVWC Raw Water Configuration

The LFWTP typically treats raw water from either one of two surface water supplies, Rivers A and River B, or some blended combination thereof. A third raw water reservoir, Reservoir A, is available on a limited basis, as needed during drought conditions and/or to minimize negative water quality impacts from either of the two primary surface water sources. These source waters are highly variable in both quality and quantity with respect to seasonal changes and precipitation related events. Source water selection is based primarily on water quality conditions followed by economic factors such as treatment and pumping costs.

Rivers A and B have historically exhibited variable and unique water quality characteristics that impart differing degrees of treatment challenges. River B is considered to be of poorer water quality because of higher organic concentrations, while River A has a higher incidence of extreme algal blooms, which have challenged the treatment plant prior to the recently upgrade treatment processes.

The three water quality monitoring or sampling stations, numbered 100, 101, and 612 shown in figure 2 were used in this study. Sampling Station 100 is located at the raw water intake point for the LFWTP, and is representative of the water sources used to supply the treatment plant. The water source(s) entering the intake point at sampling Station 100 can at any given time originate from one or more of three possible sources; River A, River B, and Reservoir A. Water quality sampling Station 612, located at the intake for Pumping Station 2, is almost always representative of water quality on River A (in past years, some low flow extreme conditions coupled with high extractions at Pumping Station 2 captured some portion of River B for short periods). The third and last water quality sampling location, Station 101, located at the mouth of the canal that leads from River B to the LFWTP, is representative of water quality conditions on River B capturing conditions after the confluence of Rivers A and B.

As part of the recent plant upgrade, online water quality monitoring stations were installed at the pumping stations 1 and 2 and at the plant intake to provide continuous

data for source water selection and for advanced warning of water quality changes. Because the largest source of water used by the utility during the study period was derived from River A, for most time periods, water quality at Station 100 (i.e. intake location) is highly correlated with water quality at Station 612 (i.e. Pumping Station 2 on River A). During most of the data collection period of this study, Station 100 intake water was 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.

The data set used for developing and assessing the ANN methodology included most variables that are considered potentially important for predicting algae levels. The final ANN models were developed for two forecast prediction periods; one-week ahead and two-weeks ahead. The two forecasting horizons were selected in part recognizing that the best available weather forecasts that provide some measure of confidence do not extend much beyond two weeks or so. In addition, these forecast horizons provide the utility with sufficient lead time to plan and implement effective proactive measures and strategies.

It was found that to improve model performance, each station should be individually modeled, rather than collectively as was done with Swimming River. This finding conforms to physical intuition, as two of the stations measure water quality conditions on different rivers, each with unique watershed characteristics. A summary of modeling results for the combined station modeling effort is presented in Appendix B-4.

A total of 553 measurement events, consisting of water quality, hydrologic, and water data collected intermittently over the period January 1999 to August 2004 were used in the study. The three individual stations were sampled at different frequencies, and thus events between stations do not always correspond in time. Station 100, located at the intake point, was sampled at the highest frequency, with as many as 270 historical events available for model training and validation. At the other extreme, Station 101, located on the least frequently abstracted river source, was sampled much less frequently, with as

few as 41 events and no more than 109 available for model training and validation. Station 612, located on the most frequently abstracted River A, also had relatively few historical events, but when the five select water quality variables were eliminated, it had the largest relative increase in sample size, from approximately 40 or 50 events to as many as 174 events.

The parameters provided by PVWC are listed in Table 1 below.

Table 1. Parameters Provided by PVWC for the Algae Modeling Study

Parameter	Unit	Parameter	Unit
Water Temperature	^{0}C	Total Phosphorus	mg/L
pH		Orthophosphate	mg/L
Turbidity[1]	NTU	Nitrite	mg/L
Alkalinity[1]	mg/L as CaCO3	Nitrate	mg/L
Hardness, Total	mg/L as CaCO3	Ammonia	mg/L
Hardness, Ca[4]	mg/L as CaCO3	Total Suspended Solids[3]	mg/L
Color[2]	Cu	Total Solids[4]	mg/L
Odor		Total Dissolved Solids[4]	mg/L
Conductivity	Umhos/cm	UV_{254}	cm^{-1}
Dissolved Oxygen	mg/L	Total Organic Carbon	mg/L
Biochemical Oxygen Demand	mg/L	Total Amorphous Materials[2]	Cells/ml
COD[4]	mg/L	Cyanobacteria counts	Cell/ml
Sulfate	mg/L	Chrysophyta counts	Cell/ml
Chloride	mg/L	Chlorophyta counts	Cells/ml
Fluoride[4]	mg/L		

1 – value estimated for Station 100, 2 – data not available to Station 100
3 – data not available to Station 101 and 612, 4 – very limited data, not used in ANN development

The data were provided to NOAH in electronic format by Ms. Linda Pasquarello, Water Quality Specialist. Climate data was obtained from National Oceanographic and Atmospheric Administration (NOAA) and included total daily precipitation, average

daily temperature, wind speed, and wind direction. There was no available data for solar radiation; hence data for sky cover, heating degree days, and length of day were also used, with values for the two first variables also obtained from NOAA, and the last obtained from sunrise and sunset tables obtained from the following website: www.jgiesen.de/GeoAstro/GeoAstro.htm.

The weather stations included Caldwell, located approximately 6 miles southeast of the site, Newark, located approximately 11.4 miles south of the site, and Tetterboro, which is located approximately 9 miles southwest of the site. Flow data were collected by the United States Geologic Survey (USGS) at the River A gauging station. PVWC also provided daily volumetric extraction data for River A, River B, Reservoir A and Pumping Station 1 and Pumping Station 2. Pumping Station 2 is at the intake point of River A.

6. WATER TREATMENT MODELING – SWIMMING RIVER RESERVOIR

6.1 Overview

As discussed above, for the water treatment modeling, the data set provided by New Jersey American spanned approximately two and a half years, consisting of 60 variables recorded for each day. The variables used in this water treatment modeling problem can be classified into four basic groups- physical water data, chemical water data, water treatment data, and weather data. For example, physical water data includes variables like total daily flow and average daily water temperature. Chemical water data include average daily measured turbidity, pH, and average daily chlorine levels. Water treatment data was the quantity or doses of chemicals added for treatment, such as sodium hydroxide and hydryfluosilicic acid. Weather variables were the most limited, consisting of only three variables; middle ambient air temperature, average daily ambient air temperature, and maximum daily ambient air temperature.

The water treatment output variables for which ANN prediction models were developed were average daily effluent turbidity reading (average turbidity), highest daily effluent turbidity reading (highest turbidity), and daily number of readings above 0.1 NTU (daily readings). All three of these variables are a measure of how effectively water treatment processes have removed suspended materials, such as sediments and organic material, like algae. A complete list of the input variables can be found in Appendix C-3, Table 1.

Somewhat less modeling was performed for the highest daily turbidity reading output variable. Because most of the variables measured represent average or total daily values, it may not be realistic to accurately model and predict a system state that is highly dependent on system changes that occur over much shorter time periods (e.g. hourly). Nevertheless, modeling of this variable was conducted for most cases. It should be noted that improved modeling performance was achieved with ANN models developed for each season, and a detailed overview of these results, and comparison with the models presented here, can be found in Appendix C-1.

6.2 Initial ANN Modeling Results

Figure 3 depicts the three prediction output variables of interest over the roughly 2.5 year period of record, from July 1, 2001 to December 31, 2003. As shown, the levels of the three variables generally increased over time, and this could be due to a reduction in the effectiveness of the water treatment processes. A plotting of the turbidity levels for the raw water does not exhibit an obvious increase with time, but instead, is correlated with seasons, which will be discussed in more detail later.

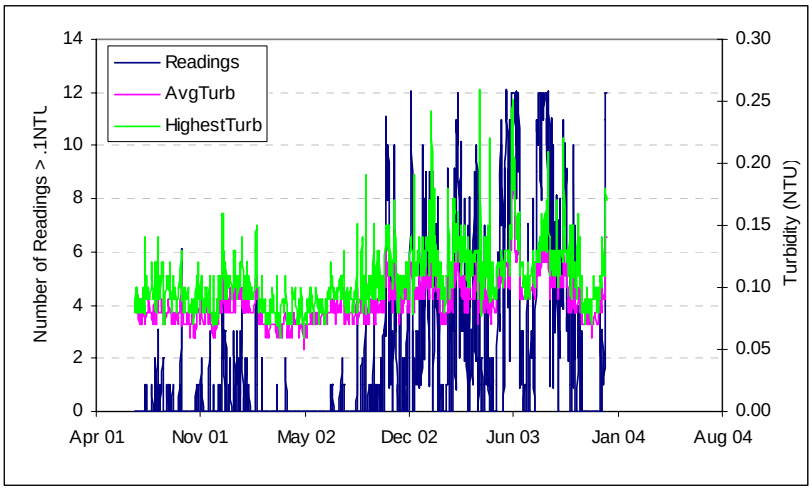


Figure 3. Daily turbidity readings above 0.1 NTU, average daily turbidity, and highest daily turbidity with respect to time over data record for Swimming River.

For the initial modeling exercise, seven different ANN models were developed and tested; two to predict average turbidity, two to predict daily readings, and three to predict highest turbidity. More than one ANN was trained for each so that, as with the Swimming River modeling efforts described above, the robustness of the models as indicated by consistency in both statistical performance and sensitivity analysis results could be assessed. Table 2 below compares the overall model performance achieved for the seven initial models.

Table 2. Overall Statistical Performance of the ANN Models for Swimming River Water Treatment

	Average Turbidity		Highest Turbidity			Daily Readings	
	First	Second	First	Second	Third	First	Second
MAE	0.01	0.01	0.025	0.03	0.01	1.48	1.49
Correlation Coefficient	0.72	0.66	0.11	0.18	0.61	0.72	0.72

As shown, the ANN models for average turbidity and daily readings performed consistently well, with relatively high correlation values. In contrast, there were inconsistent results for the three individual ANN models developed for predicting highest daily turbidity, with generally low correlations. Relatively similar performance can be observed from Table 3, which compares, for each of the best models for each variable, overall ANN model performance against performance during training and validation (i.e. testing).

Table 3. Comparison of Overall Performance versus Training and Validation for Best ANN Water Treatment Models for Swimming River

	Overall		**Training**		**Validation**	
	MAE	Correlation Coefficient	MAE	Correlation Coefficient	MAE	Correlation Coefficient
Average Turbidity	0.01	0.72	0.01	0.71	0.01	0.80
Highest Turbidity	0.01	0.61	0.01	0.66	0.02	0.58
Reading	1.53	0.72	1.27	0.81	1.56	0.66

The training and validation results are most similar for the ANN models that predict average daily turbidity and daily readings. There is more discrepancy between training and validation for the ANN model that predicts highest daily turbidity, as shown by the mean absolute errors.

Figures 4 through 6 compare measured against ANN model predicted values, for average turbidity, highest turbidity, and daily readings, respectively. As shown, for average

turbidity and daily readings, the ANN models generally follow the higher and lower readings. For average turbidity, the ANN model has a stronger tendency to under-predict extremely high readings, and over-predict very low readings. This tendency to miss extremes is likely due to relatively few extreme examples for which the ANNs can learn from. Similarly, for the highest turbidity, the models have a stronger tendency to under-predict extreme high values and over-predict extreme low values.

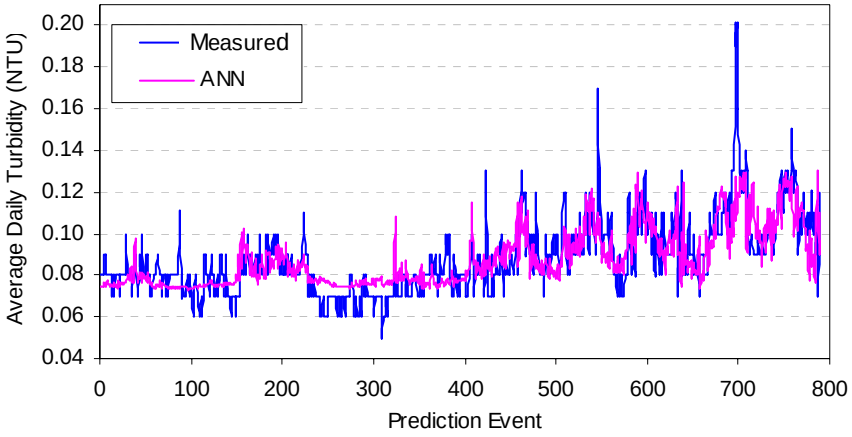


Figure 4. Measured average daily turbidity readings versus ANN predictions For Swimming River.

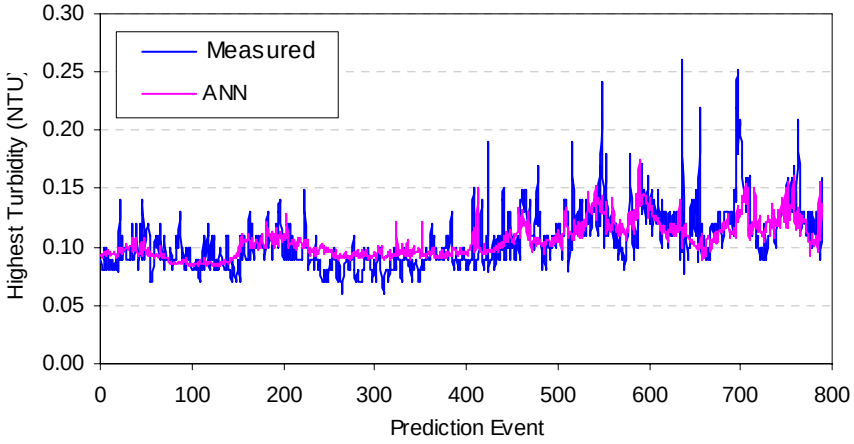


Figure 5. Measured highest daily turbidity readings versus ANN predictions for Swimming River.

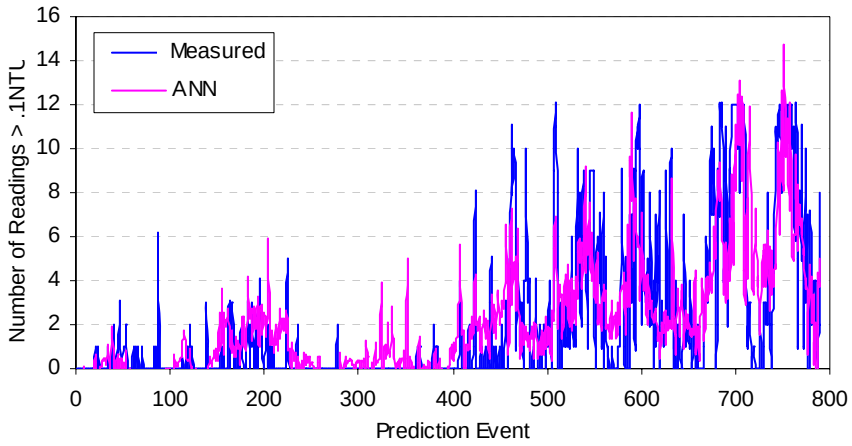


Figure 6. Measured daily turbidity readings above 0.1 NTU versus ANN predictions for Swimming River.

6.3 Sensitivity Analysis

As was done with the Swimming River algae modeling component, sensitivity analyses were conducted for the ANN water treatment models. Tables 4, 5, and 6 depict representative results for all ANN models used for the three prediction variables, average turbidity, highest turbidity, and average readings, respectively. The ten highest and lowest ranking variables for each model are displayed in the tables. Color coding has been used to facilitate recognition of identical input variables. The complete tables can be found in Appendix C-3.

Table 4. Comparison of Sensitivity Analysis Results for Average Daily Turbidity Models with top 10 and bottom 10 variables in terms of RMSE ratio values for Swimming River

Model 1				Model 2			
Variable	Input Variable No.	Ratio	Rank	Variable	Input Variable No.	Ratio	Rank
Temp Raw °C	Var45	1.087	1	Turbidity Settled NTU	Var35	1.061	1
Turbidity Settled NTU	Var35	1.036	2	Chlorine mix mg/l	Var41	1.021	2
Chlorine mix mg/l	Var41	1.034	3	Filters Washed/24H	Var3	1.015	3
Chlorine Eff mg/l	Var44	1.028	4	Chlorine Eff mg/l	Var44	1.014	4
Filters Washed/24H	Var3	1.021	5	Chlorine PPM/24H	Var24	1.012	5
Temp Eff °C	Var46	1.02	6	Temp Raw °C	Var45	1.011	6
Chlorine PPM/24H	Var24	1.016	7	Polyaluminum chloride gals/24H	Var15	1.011	7
Hydryfluosilicic acid PPM/24H	Var26	1.015	8	Sysdel SR Midd MGD	Var31	1.01	8
Fluoride avg.mg/l/24H	Var47	1.011	9	pH Settled	Var39	1.01	9
Air temp °C min/24H	Var54	1.011	10	Fluoride avg.mg/l/24H	Var47	1.01	10
Washwater MGD	Var4	0.999	48	Turbidities measured/24H	Var49	1	48
pH Mix	Var38	0.999	49	pH Mix	Var38	0.999	49
Peak flow raw MGD/24H	Var52	0.999	50	Zinc orthophosphate PPM/24H	Var27	0.999	50
Air temp °C avg/24H	Var56	0.999	51	Air temp °C min/24H	Var54	0.999	51
Chlorine lowest mg/l/24H	Var50	0.999	52	Raw MGD/24H	Var1	0.999	52
Zinc orthophosphate gals/24H	Var19	0.999	53	Air temp °C avg/24H	Var56	0.999	53
Polymer PPM/24H	Var22	0.999	54	Air temp °C max/24H	Var55	0.999	54
Settled readings >2,0 NTU/24H	Var57	0.998	55	Peak flow raw MGD/24H	Var52	0.999	55
Polymer gals/24H	Var16	0.997	56	System Delivery MGD	Var2	0.997	56
Carbon PPM/24H	Var25	0.997	57	pH Effluent	Var40	0.962	57

Table 5. Comparison of Sensitivity Analysis Results for Highest Daily Turbidity Models with top 10 and bottom 10 variables in terms of RMSE ratio values for Swimming River.

Model 1				Model 2				Model 3			
Variable	Input Variable	Ratio	Rank	Variable	Input Variable	Ratio	Rank	Variable	Input Variable	Ratio	Rank
Turbidity Settled NTU	Var35	1.003	1	Settled readings >2,0 NTU/24H	Var57	1.008	1	Turbidity Settled NTU	Var35	1.028	1
Temp Raw °C	Var45	1.001	2	Turbidity Settled NTU	Var35	1.003	2	pH Raw	Var37	1.015	2
Turbidity Raw NTU	Var34	1.001	3	Temp Raw °C	Var45	1.001	3	NaOH PPM/24H	Var23	1.015	3
Air temp °C min/24H	Var54	1.001	4	Turbidity Raw NTU	Var34	1.001	4	Chlorine mix mg/l	Var41	1.013	4
Sysdel SR Main MGD	Var30	1	5	Temp Eff °C	Var46	1.001	5	Temp Raw °C	Var45	1.012	5
Chlorine lbs/24H	Var8	1	6	Chlorine lbs/24H	Var8	1	6	Chlorine PPM/24H	Var24	1.011	6
Zinc orthophosphate gals/24H	Var19	1	7	Zinc orthophosphate PPM/24H	Var27	1	7	Raw MGD/24H	Var1	1.011	7
Hydryfluosilicic acid PPM/24H	Var26	1	8	Chlorine CW mg/l	Var43	1	8	Fluoride avg.mg/l/24H	Var47	1.01	8
Fluoride avg.mg/l/24H	Var47	1	9	Sysdel SR Main MGD	Var30	1	9	Filters Washed/24H	Var3	1.007	9
Polymer lbs/24H	Var7	1	10	Fluoride avg.mg/l/24H	Var47	1	10	Polyaluminum chloride PPM/24H	Var21	1.007	10

Table 6. Comparison of Sensitivity Analysis Results for Highest Daily Turbidity Models with top 10 and bottom 10 variables in terms of RMSE ratio values for Swimming River (continued)

Model 1				Model 2				Model 3			
Variable	Input Variable	Ratio	Rank	Variable	Input Variable	Ratio	Rank	Variable	Input Variable	Ratio	Rank
Polyaluminum chloride lbs/24H	Var6	1	48	CW Lowest level/24H	Var51	1	48	Polymer gals/24H	Var16	1	48
Carbon PPM/24H	Var25	1	49	Sodium Chlorite gals/24H	Var20	1	49	Zinc orthophosphate lbs/24H	Var12	1	49
Polymer gals/24H	Var16	1	50	Chlorine PPM/24H	Var24	1	50	Turbidity Raw NTU	Var34	1	50
Settled readings >2,0 NTU/24H	Var57	1	51	Polyaluminum chloride PPM/24H	Var21	1	51	Chlorine lbs/24H	Var8	0.999	51
Average % solids/liter	Var14	1	52	Polyaluminum chloride gals/24H	Var15	1	52	Air temp °C avg/24H	Var56	0.999	52
Chlorine PPM/24H	Var24	1	53	Washwater MGD	Var4	1	53	Settled readings >2,0 NTU/24H	Var57	0.999	53
Temp Eff °C	Var46	1	54	Carbon lbs/24H	Var5	1	54	Zinc orthophosphate PPM/24H	Var27	0.999	54
NaOH PPM/24H	Var23	1	55	Hydrofluosilicic acid lbs/24H	Var11	1	55	Ferric chloride PPM/24H	Var28	0.998	55
Chlorine mix mg/l	Var41	1	56	Chlorine Settled mg/l	Var42	1	56	Air temp °C max/24H	Var55	0.998	56
Air temp °C max/24H	Var55	1	57	pH Raw	Var37	1	57	pH Effluent	Var40	0.996	57

Table 7. Comparison of Sensitivity Analysis Results for Readings above 0.1 NTU Models with top 10 and bottom 10 variables in terms of RMSE ratio values for Swimming River.

Model 1				Model 2			
Variable	Input Variable Number	Ratio	Rank	Variable	Input Variable Number	Ratio	Rank
Sodium chlorite lbs/24H	Var9	1.148	1	Turbidity Settled NTU	Var35	1.095	1
Chlorine mix mg/l	Var41	1.107	2	Temp Raw °C	Var45	1.081	2
Air temp °C min/24H	Var54	1.076	3	Chlorine mix mg/l	Var41	1.062	3
Temp Raw °C	Var45	1.060	4	Fluoride avg.mg/l/24H	Var47	1.021	4
Turbidity Settled NTU	Var35	1.059	5	Hydrofluosilicic acid lbs/24H	Var11	1.018	5
Air temp °C max/24H	Var55	1.057	6	CW Lowest level/24H	Var51	1.011	6
Hydryfluosilicic acid PPM/24H	Var26	1.037	7	Chlorine Settled mg/l	Var42	1.010	7
Chlorine lbs/24H	Var8	1.037	8	Filters Washed/24H	Var3	1.010	8
System Delivery MGD	Var2	1.034	9	pH Raw	Var37	1.008	9
Ferric chloride lbs/24H	Var13	1.000	48	Polyaluminum chloride gals/24H	Var15	1.000	48
Sodium chlorite PPM/24H	Var29	1.000	49	Sysdel NS Midd MGD	Var33	1.000	49
Turbidities required/24H	Var48	1.000	50	Raw MGD/24H	Var1	0.999	50
NTU over limit	Var36	1.000	51	Sysdel SR Midd MGD	Var31	0.999	51
Turbidities measured/24H	Var49	1.000	52	Sysdel SR Main MGD	Var30	0.999	52
pH Effluent	Var40	1.000	53	Polyaluminum chloride PPM/24H	Var21	0.998	53
Sysdel SR Midd MGD	Var31	0.999	54	Average % solids/liter	Var14	0.998	54
Carbon lbs/24H	Var5	0.999	55	pH Settled	Var39	0.998	55
Zinc orthophosphate lbs/24H	Var12	0.999	56	Zinc orthophosphate gals/24H	Var19	0.998	56
Ferric chloride PPM/24H	Var28	0.995	57	pH Effluent	Var40	0.994	57

What is first striking is that none of the variables for any of the seven models had a particularly high ratio value. That is, no single input variable was particularly important for predicting the final output value. For predicting average turbidity, there was consistency between the two models. The two models shared seven of the ten highest ranking predictor variables, with raw water temperature, turbidity settled, and both chlorine mixed and chlorine effluent ranking right near the top. All of the ten highest ranking variables had a ratio value about 1.01. There was less similarity for the ten least important variables, all of which had a ratio value < 1.0.

For predicting highest turbidity, because three models rather than two were developed, it is less likely to have similar rankings between all models. However, for all three models, the same two input variables, turbidity settled and raw water temperature, rank within the top five. Other input variables, such as chlorine (lbs/24hr) raw water turbidity, and zinc orthophosphate, rank within the top ten most important variables for two models. Fluoride appears within the top ten for all three models. In terms of least important, no single variable occurs in all three models as ranking within the bottom 10.

The two ANN models developed for predicting daily readings share some highly ranked variables. Turbidity settled, mixed chlorine, and raw water temperature rank within the top five for both models. For the least important variables, two rank near the bottom; pH effluent and zinc orthophosphate.

One would expect that on average, out of 57 input variables, there would be some overlap between higher and lower ranking variables. However, for the three prediction output variables, two input variables ranked within the top six most important variables for all seven models; raw water temperature and settled turbidity. In terms of chemicals added, besides chlorine, no single coagulant or chemical agent appears to consistently rank high for any single finished water quality output variable. For predicting average turbidity, filters (washed/24hr) ranked highly for both models.

6.4 ANN versus Linear Models

As a comparison of ANN performance, linear models (LMs) were applied to the identical data sets. Table 8 below is a statistical performance comparison between the best ANN models against the LMs developed for each of the three output prediction variables

Table 8. Statistical Performance Comparison of LMs versus Best ANNs for Water Treatment Modeling at Swimming River

	Average Daily Turbidity	Maximum Daily Turbidity	No. of Readings > 0.1 NTU
ANN Mean Absolute Error	0.0095	0.014	1.48
LM Mean Absolute Error	0.0096	0.018	1.81
ANN Correlation Coefficient	0.72	0.60	0.72
LM Correlation Coefficient	0.65	0.14	0.41

For the best performing ANN models, and in fact, for the two different models that were used for predicting average turbidity and daily readings, the ANN models consistently outperformed the LMs. The discrepancy between the model types for average turbidity was relatively small; it was more significant for high turbidity and daily readings. However, some of the poorer performing ANN models did not statistically perform as well as the LMs for predicting highest turbidity.

Figures 7 through 11 further below compare measured versus linear model predicted values for average turbidity, highest turbidity, and average readings, respectively. What is interesting is that while on average, the LMs performed well, there are rare events where a highly erroneous output value is computed by the linear model. At least one of these erroneously high computed values was due to a data input error, which will be discussed later within the context of the sensitivity analyses results.

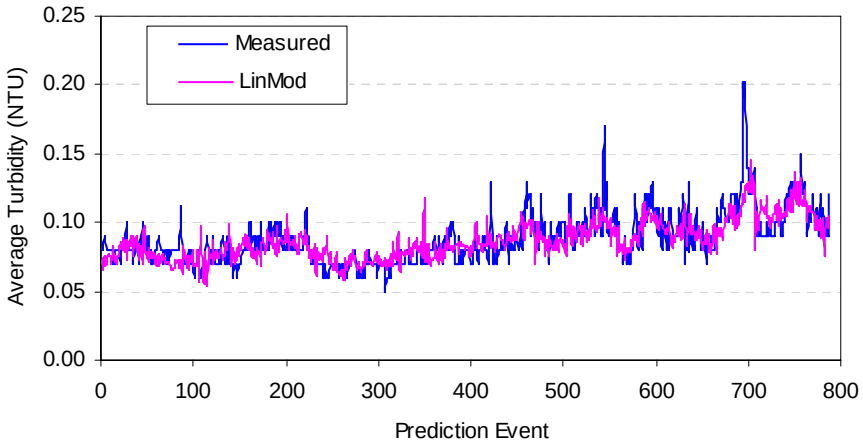


Figure 7. Comparison of measured average daily turbidity versus LM predictions for Swimming River.

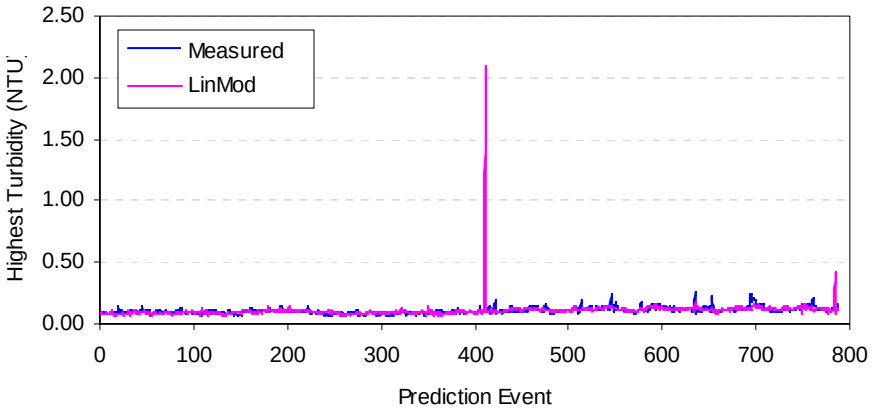


Figure 8. Comparison of measured highest daily turbidity versus LM predictions for Swimming River

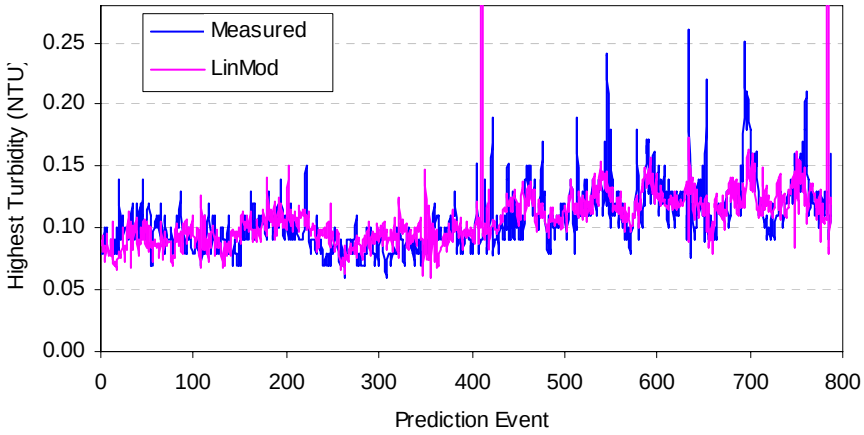


Figure 9. A magnification of Figure 19 above.

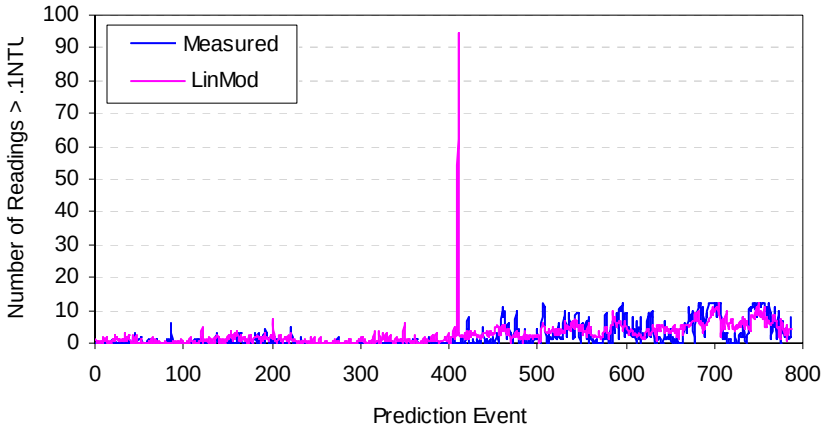


Figure 10. Comparison of daily turbidity readings above 0.1 NTU versus LM predictions

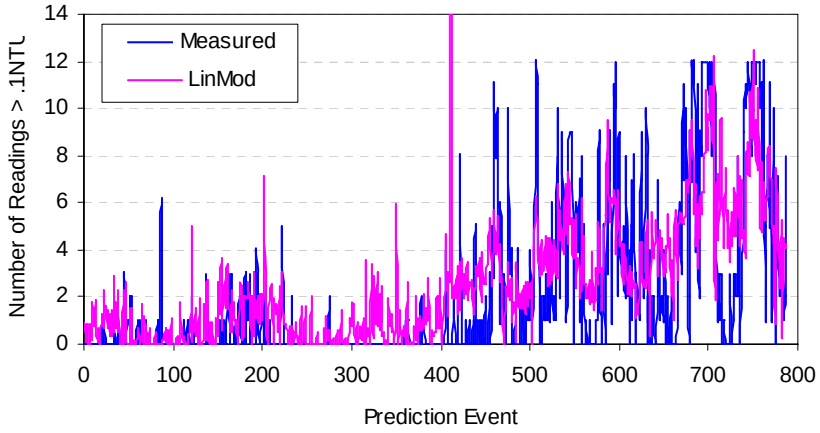


Figure 11. A magnification of Figure 21 above.

Tables in Appendix C-3 summarize the sensitivity analyses results for the LMs. Many of the variables identified as important by the ANN models were not identified as important by the LMs. For each of the three output variable prediction problems, only two variables appeared in the top 10 for both the ANN and LM; chlorine (ppm/24hr) for the average and highest turbidity outputs, and sodium chlorite (lbs/24hr) for daily readings.

As was reported in the preliminary report (August 4, 2004), the LMs have a tendency to have variables with extremely high RMSE ratio values (e.g. 10^{14}), which signifies an instability in the solution for the model coefficients. This instability has the potential to produce highly erroneous outputs. In this problem, an extremely low ratio value

achieved by the LM produced is associated with a highly erroneous output value. This occurred for prediction event 433 with both output variables highest turbidity and daily reading, where extremely erroneous high values were computed by the LMs. A review of the data showed that an infeasible value of 85 was reported and used for the input variable pH effluent, which likely should have been 8.5.

For the pH effluent input variable, the LM produced a ratio value of 0.313 for predicting highest daily turbidity and a ratio value of 0.619 for predicting daily readings above 0.1 NTU. By comparison, the ANN models for this variable had computed ratio values of 0.996 and 1.00 for the highest turbidity and daily reading outputs, respectively. In these cases, then, that the LMs may have a tendency to "over-react" or overcompensate for a single erroneous value, while the ANN models are not as affected by an erroneous, or in this case, infeasible input value for pH.

6.5 ANN Models (Reduced) with Fewer Input Variables

The next modeling exercise was to develop ANNs with only the most important predictor variables, as identified by the sensitivity analyses discussed above. The primary objective is to determine whether the ANNs can provide relatively accurate predictions with fewer variables, which could have cost savings benefits. In addition, this would help confirm the importance of or lack thereof of certain model input predictor variables.

The ten most important predictor variables as identified by the sensitivity analyses results (i.e. ten highest ranked variables) were used to develop a new ANN model for each of the three output variables, average daily turbidity, highest daily turbidity, and daily readings. Because all of these variables had computed ratio values above 1.0, they would all be expected to improve overall model performance. Table 9 compares the statistical performance of the original best ANN models with the best performance for each of the three output variables against the models that used only 10 input variables. It should be noted that the ANN models that used only 10 inputs had more events. Some days had at least one variable with a missing value, for which the model cannot be used. Thus, the

model using fewer input variables will have fewer days (i.e. events) with at least one missing value.

Table 9. Statistical Performance Comparison of Complete versus Reduced ANN Models for Water Treatment at Swimming River

	Average Turbidity	Highest Turbidity	No. of Readings > 0.1 NTU
Complete ANN Mean Absolute Error	0.00948	0.0143	1.48
Reduced ANN Mean Absolute Error	0.00933	0.0148	1.63
Complete ANN Correlation Coefficient	0.720	0.605	0.721
Reduced ANN Correlation Coefficient	0.703	0.582	0.690

Although the complete ANN models consistently outperformed the reduced ANN models, the discrepancy is relatively small. This is somewhat unexpected given the fact that in the original sensitivity analyses results, far more than ten variables in each case attained ratio values >1.0, approximately 45 total. Figures 12 through 14 compare measured versus ANN predicted values by the reduced models for average turbidity, high turbidity, and daily readings, respectively. On the other hand, probably relatively few variables are critical state and water treatment variables for the process.

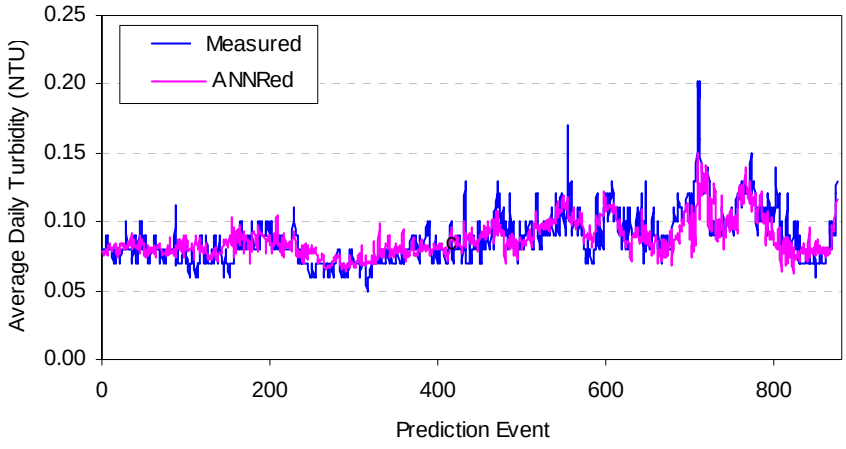


Figure 12. Average daily turbidity measurements versus reduced ANN model predictions for Swimming River

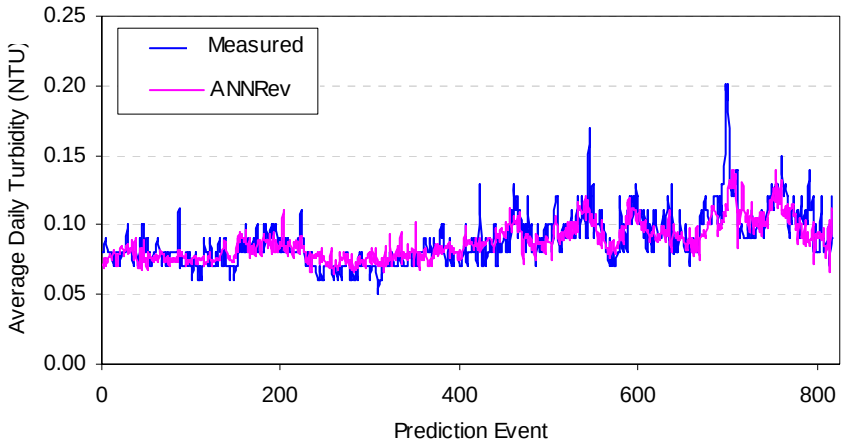


Figure 13. Highest daily turbidity measurements versus reduced ANN model predictions for Swimming River

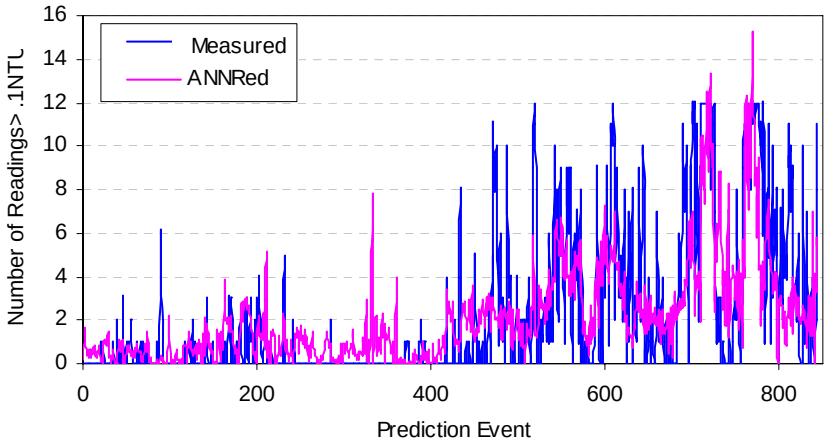


Figure 14. Daily readings measurements versus reduced ANN model predictions for Swimming River

A comparison of the sensitivity analysis results for the complete versus the reduced ANN models shows that the relative rankings of the ten variables are relatively dissimilar for the corresponding models (e.g. average turbidity, etc.). The tabulated sensitivity analyses results can be found in Appendix C-3.

6.6 More Refined ANN Models

As an additional analysis, instead of using only the ten most important predictor variables, as identified by the sensitivity analysis, as ANN model inputs, all variables that had an overall computed ratio value of at least 1.001 was used. For average turbidity, there were 32 variables; for highest turbidity, 28 variables; and for daily readings, 33 variables. Table 10 compares the statistical performance of the best performing complete ANN models against the ANN models with the refined input set.

Table 10. Statistical Performance Comparison of Complete ANN models versus Refined ANN Models for Water Treatment at Swimming River

	Turbidity	Highest Turbidity	Readings
Complete ANN Mean Absolute Error	0.00948	0.0142	1.48
Refined ANN Mean Absolute Error	0.00927	0.0144	1.52
Complete ANN Correlation Coefficient	0.720	0.605	0.721
Refined ANN Correlation Coefficient	0.722	0.6118	0.708

Statistically there is nominal difference between the refined and the complete ANN models. Not only did the refined models on average perform almost as well as the complete ANN models, but at times, their accuracy was better. This similar performance would potentially have cost implications, where marginal improvement may be offset by increased sampling costs. What is necessary is a more clear understanding of what is essential for predicting water quality variables of interest to acceptably accurate standards. What is also important to consider is the likelihood that the utility is following a consistent pattern of treatment strategies, based upon raw water quality conditions, etc., and that the ANN is consequently "keying" off surrogate variables, which implicitly represent chemical doses, etc. Figures 15 through 17 depict the modeling results.

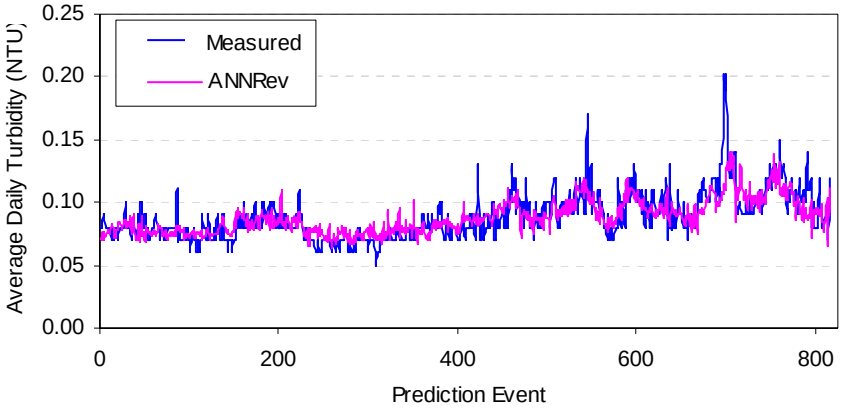


Figure 15. Comparison of average daily turbidity measurements versus ANN predictions for Swimming River with refined model.

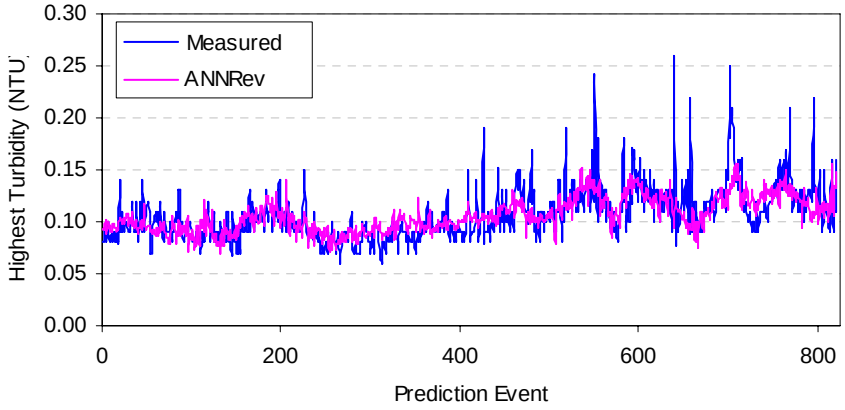


Figure 16. Comparison of highest daily turbidity measurements versus ANN predictions for Swimming River with refined model.

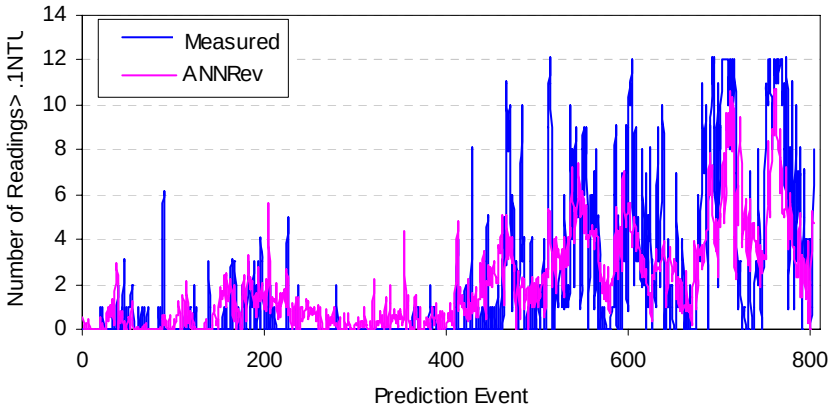


Figure 17. Comparison of daily readings above 0.1 NTU versus ANN predictions with refined model.

A sensitivity analysis comparison was also performed for the two models. Again, the input variables identified by the complete ANN as important (i.e. RMSE ratio above 1.001) were used as the input variables for the corresponding refined ANN models. Tables 11 through 13 compare the variable ratios and rankings for the variables in the refined models that ranked above 1.001. Color coding is again used to facilitate reviewer comparison. A complete listing of the sensitivity analyses results is provided in Appendix C-3.

Table 11. A Comparison of the Top 15 Variables for Complete and Refined ANN Models that Predict Average Daily Turbidity for Swimming River

Complete ANN Model				Refined ANN Model			
Temp Raw °C	Var45	1.087	1	Temp Raw °C	Var45	1.177	1
Turbidity Settled NTU	Var35	1.037	2	Temp Eff °C	Var46	1.118	2
Chlorine mix mg/l	Var41	1.035	3	Turbidity Settled NTU	Var35	1.101	3
Chlorine Eff mg/l	Var44	1.029	4	Hydryfluosilicic acid PPM/24H	Var26	1.095	4
Filters Washed/24H	Var3	1.021	5	System Delivery MGD	Var2	1.077	5
Temp Eff °C	Var46	1.020	6	Sodium hydroxide gals/24H	Var17	1.046	6
Chlorine PPM/24H	Var24	1.016	7	Chlorine mix mg/l	Var41	1.041	7
Hydryfluosilicic acid PPM/24H	Var26	1.015	8	Filters Washed/24H	Var3	1.040	8
Fluoride avg.mg/l/24H	Var47	1.012	9	Sysdel NS Main MGD	Var32	1.031	9
Air temp °C min/24H	Var54	1.011	10	Fluoride avg.mg/l/24H	Var47	1.025	10
CW Lowest level/24H	Var51	1.009	11	Polyaluminum chloride PPM/24H	Var21	1.025	11
Sodium hydroxide gals/24H	Var17	1.008	12	Turbidity Raw NTU	Var34	1.024	12
Sysdel NS Midd MGD	Var33	1.008	13	pH Settled	Var39	1.022	13
Zinc orthophosphate lbs/24H	Var12	1.007	14	Sysdel NS Midd MGD	Var33	1.022	14
Chlorine CW mg/l	Var43	1.007	15	Zinc orthophosphate PPM/24H	Var27	1.021	15

For average turbidity, of the original 32 input variables used in the refined model, 15 retained a ratio above 1.001. Both the complete and refined ANN models exhibit similarities in terms of the top ranking variables, with eight of the same variables ranking within the top fifteen for both models. Raw water temperature and turbidity settled were within the top three for both models. Also within the top eight for both models were temperature of the effluent, chlorine mixed, filters washed, and hydrofluosilic acid.

Table 12. A Comparison of the Top 16 Variables for Complete and Refined ANN Models that Predict Highest Daily Turbidity for Swimming River

Complete ANN Model				Refined ANN Model			
Turbidity Settled NTU	Var35	1.028	1	Temp Raw °C	Var45	1.123	1
pH Raw	Var37	1.015	2	Turbidity Settled NTU	Var35	1.049	2
NaOH PPM/24H	Var23	1.015	3	Hydrofluosilicic acid lbs/24H	Var11	1.045	3
Chlorine mix mg/l	Var41	1.013	4	Hydryfluosilicic acid PPM/24H	Var26	1.044	4
Temp Raw °C	Var45	1.012	5	Polyaluminum chloride PPM/24H	Var21	1.044	5
Chlorine PPM/24H	Var24	1.011	6	Raw MGD/24H	Var1	1.028	6
Raw MGD/24H	Var1	1.011	7	Temp Eff °C	Var46	1.024	7
Fluoride avg.mg/l/24H	Var47	1.010	8	Chlorine mix mg/l	Var41	1.020	8
Filters Washed/24H	Var3	1.007	9	Chlorine CW mg/l	Var43	1.017	9
Polyaluminum chloride PPM/24H	Var21	1.007	10	Sodium hydroxide lbs/24H	Var10	1.016	10
Temp Eff °C	Var46	1.006	11	Peak flow raw MGD/24H	Var52	1.015	11
Sysdel NS Midd MGD	Var33	1.006	12	Filters Washed/24H	Var3	1.014	12
Chlorine Eff mg/l	Var44	1.006	13	Air temp °C min/24H	Var54	1.010	13
pH Settled	Var39	1.005	14	pH Raw	Var37	1.009	14
Hydryfluosilicic acid PPM/24H	Var26	1.004	15	Sysdel NS Main MGD	Var32	1.009	15
Polyaluminum chloride lbs/24H	Var6	1.004	16	Sysdel NS Midd MGD	Var33	1.008	16

For predicting highest daily turbidity, 16 of the original 28 variables in the refined ANN model retained a ratio value above 1.001. Both the complete and refined ANNs had the same nine variables within the top 16. As with the ANN models that predicted average turbidity, both models had raw water temperature and settled turbidity ranked highly; in this case, within the top 5 for both models. However, compared to the average turbidity models, there were not as many shared variables that ranked near the top (i.e. within top half of the table).

Table 13. A Comparison of the Top 17 Variables for Complete and Refined ANN Models that Predict Number of Turbidity Readings > 0.1 NTU for Swimming River

Complete ANN Model				Refined ANN Model			
Sodium chlorite lbs/24H	Var9	1.148	1	Temp Raw °C	Var45	1.164	1
Chlorine mix mg/l	Var41	1.107	2	Turbidity Settled NTU	Var35	1.131	2
Air temp °C min/24H	Var54	1.076	3	Sodium hydroxide lbs/24H	Var10	1.093	3
Temp Raw °C	Var45	1.060	4	Temp Eff °C	Var46	1.087	4
Turbidity Settled NTU	Var35	1.059	5	Polyaluminum chloride PPM/24H	Var21	1.084	5
Air temp °C max/24H	Var55	1.057	6	Chlorine mix mg/l	Var41	1.079	6
Hydryfluosilicic acid PPM/24H	Var26	1.037	7	NaOH PPM/24H	Var23	1.052	7
Chlorine lbs/24H	Var8	1.037	8	Sodium hydroxide gals/24H	Var17	1.049	8
System Delivery MGD	Var2	1.034	9	Air temp °C min/24H	Var54	1.041	9
Chlorine CW mg/l	Var43	1.031	10	Raw MGD/24H	Var1	1.026	10
Temp Eff °C	Var46	1.029	11	Chlorine Settled mg/l	Var42	1.017	11
pH Raw	Var37	1.027	12	Zinc orthophosphate PPM/24H	Var27	1.017	12
CW Lowest level/24H	Var51	1.026	13	Average % solids/liter	Var14	1.014	13
Washwater MGD	Var4	1.024	14	Chlorine lbs/24H	Var8	1.010	14
Polyaluminum chloride PPM/24H	Var21	1.024	15	Hydryfluosilicic acid PPM/24H	Var26	1.010	15
pH Settled	Var39	1.022	16	pH Raw	Var37	1.009	16
Sodium Chlorite gals/24H	Var20	1.020	17	Chlorine CW mg/l	Var43	1.008	17

Of the 33 original input variables for the refined ANN model developed for daily readings, seventeen retained a ratio value above 1.001. As with the ANN models for average and highest turbidity, raw water temperature and settled turbidity ranked within the top five for both the complete and refined ANN models. Chlorine mixed also ranked within the top six variables for both models.

Overall, then, for the three predictor variables, approximately half of the original inputs for the refined models retained ratio values above 1.001. In addition, approximately half of the highest ranking variables for the refined models ranked near the top for the complete models. As was found previously for the complete ANN models, raw water temperature and settled turbidity consistently ranked near the top, and in fact both were always within the top three most important variables for the three refined ANN models.

6.10 Incremental Sensitivity Analysis

As a final modeling exercise, model simulations were run to assess the effect of select variables on computed outputs. In this exercise, as was done for the cyanobacteria count modeling, several events were selected for which the value of a single input variable was systematically changed, with all other input variables held constant, from which a new output value was computed. This exercise was repeated for several input variables to assess their effect on the output variable. For example, would an increase in raw water temperature increase or decrease average turbidity of the finished water? As previously discussed, it should be emphasized that this analysis is rather limited and simplistic, and discretion should be used when interpreting the significance of the results. Again, for complex and non-linear systems**,** such cause and effect relationships may not be straightforward. That is, the corresponding changes of the system are undoubtedly related to the state of the system with respect to other variables. In this case, similar increases in temperature for two events may product two different outcomes, depending upon the other states or inputs into the system. In addition, incremental temperature changes may induce a different range of responses, depending upon the temperature values over which these incremental changes occur. However, this exercise may also

help identify some basic tendencies, and helps to illustrate how perhaps ANN technology can be used to help optimize treatment strategies.

Two different spring ANN models were used; both used all original 57 inputs, with the first predicting average turbidity only, and the second predicting both average turbidity and average readings (model outputs). For predicting average turbidity, five different prediction events were used, each representing a different range of system conditions (i.e. season), for each of which seven different input variables were incrementally varied. Of the seven variables, five of the most important variables, as identified by the sensitivity analysis, were selected, and they are: raw water temperature, in °C; system delivery, in million gallons per day; ferric chloride added, in parts per million per 24 hours; zinc orthophosphate added, in lbs per 24 hours; and turbidity settled, in NTU. As a comparison, two of the least important predictor variables were also used; measured chlorine effluent, in mg/l, and polyaluminum chloride, in lbs/24H. For each iteration, the selected input variable was changed from the actual measured value for the event to its minimum, mean, or maximum measured values for that particular variable. Table 14 below lists these values for the seven parameters.

Table 14. Input Variables Used in the Sensitivity Analysis for the Average Turbidity Model, with their Corresponding Ratio Rankings and Minimum, Mean, and Maximum Measured Values.

Ratio Ranking	Name	Min	Average	Max
1	Temp Raw °C	3.00	11.63	20.00
2	Sysdel NS Midd MGD	1.900	5.364	10.480
3	Ferric chloride PPM/24H	5.38	11.51	15.95
4	Zinc orthophosphate lbs/24H	957.00	2752.38	3834.38
5	Turbidity Settled NTU	0.36	0.72	2.10
56	Chlorine Eff mg/l	0.5	1.5	1.6
57	Polyaluminum chloride lbs/24H	0.00	1065.85	2260.50

Bar figures 18 and 19 depict representative results for two events, which were consistent for all five events, with the other bar graphs contained in Appendix C-2. The bar graphs show that for each perturbed input variable, the corresponding computed output for the actual event value versus the corresponding computed outputs associated with the minimum, mean, and maximum input values listed above in Table 14.

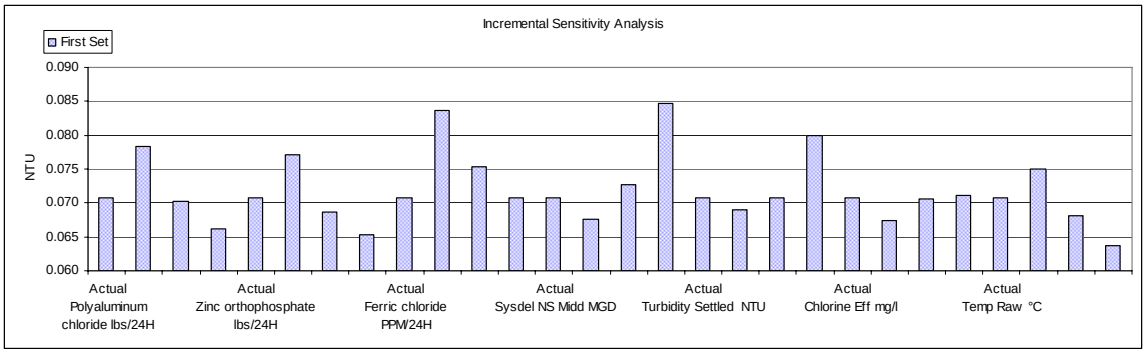


Figure 18. Predicted average daily turbidity in response to different input values for different variables for the first event for Swimming River.

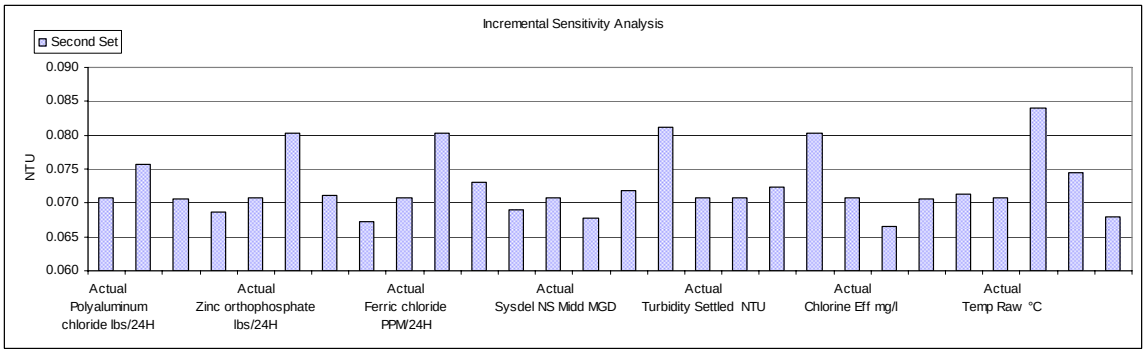


Figure 19. Predicted average daily turbidity in response to different input values for different variables for the second event for Swimming River.

Both higher water deliveries and higher settled turbidity readings both resulted in higher average turbidity levels for finished water. Both of these tendencies would appear to make sense, as both of these variables are measures of the quantity of mass in the raw water. If all other variables (e.g. chemical dosages) are held constant, turbidity removal is not increased, and thus average turbidity levels of the finished water will be elevated.

The inverse relationship between raw water temperature and average turbidity identified with the above analysis is consistent with the inverse relationship first identified when these two variables were plotted against each other. Thus, at higher temperatures, the raw water has lower turbidity. By extension, then, the treated finished water would also be expected to have lower turbidity during times when the raw water has a higher temperature, such as in summer months.

Lastly, chlorine effluent appears to be weakly positively correlated with average turbidity. This may have some physical basis, as the presence of organic material, such as algae, will increase raw water turbidity, which will increase the quantity of chlorine added. Figure 20 depicts raw water turbidity versus chlorine effluent, and does not show an obvious relationship.

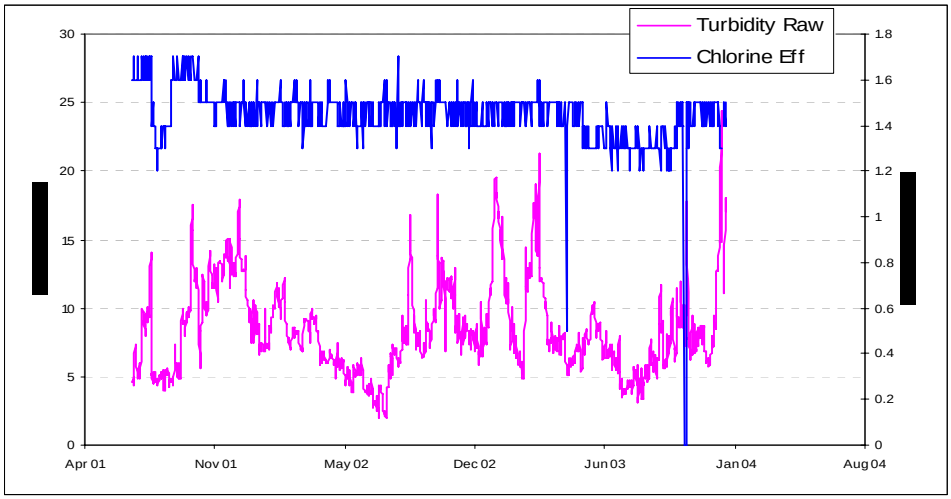


Figure 20. Raw water turbidity versus chlorine effluent concentrations for Swimming River.

The next analysis was to compare the average range over which the output values changed with respect to each of the seven input variables against the *relative* effect of each of the different input variables on the outputs. This was done using the five events. Again, because different variables can assume significantly different ranges of values, a direct comparison in computed output values using minimum and maximum possible

values provides limited information. The relative effect of each input variable on the corresponding outputs was computed using the sensitivity ratio provided previously in Equation 5.

In effect, the sensitivity ratio provides not only a measure of whether an input is positively or negatively correlated with a particular output, but also the unit change of the output per unit change of the select input variable. This understanding can be important for helping to identify treatment strategies. At the same time, as discussed earlier, it must be emphasized that the unit responses will not be uniform in a non-linear system, and may also depend upon the other states of the system. Table 15 below compares the RMSE ratio ranking against the range ranking and sensitivity ranking, with the computed values for the later two measures.

Table 15. Comparison of different variable's relative effect on average daily turbidity

Variable Name	RMSE Ratio Ranking	Range Change	Range Ranking	Sensitivity Ratio	Sensitivity Ranking
System Delivery	2	0.0162	1	0.00155	3
Raw Water Temp	1	-0.0149	2	-0.000552	4
Zinc Ortho	4	-0.0141	3	-1.47E-06	7
Ferric Chloride	3	-0.0122	4	-0.000352	5
Polyalum. Chloride	57	-0.0111	5	-2.75E-06	6
Turbidity Settled	5	0.0101	6	0.00492	1
Chlorine Effluent	56	0.00523	7	0.00436	2

For the events considered, turbidity settled and chlorine effluents have the greatest relative effect on average daily turbidity, while zinc orthophosphate has the least.

For the dual ANN model that included two output variables, average turbidity and average readings, the sensitivity analysis results were somewhat different than that for the single (output) ANN model. In this case, the five most important variables, in descending order of importance, are: raw water temperature, in °C; pH Effluent; maximum air

temperature, in °C per 24 hours; polymer added, in lbs per 24 hours, and settled readings in NTU. The two least important were Sodium hydroxide lbs/24H and pH Raw. Table 16 below reports the range of values for these seven input variables.

Table 16. Input Variables Used in the Sensitivity Analysis for the Dual Average Turbidity and Number of Readings Model, with their Corresponding Ratio Rankings and Minimum, Mean, and Maximum Measured Values

RMSE Ratio Ranking	Name	Min	Mean	Max
1	Temperature Raw °C	3.00	11.50	20.00
2	pH Effluent	7.20	7.5	7.80
3	Air temperature °C max/24H	-1.00	17.5	36.00
4	Polymer lbs/24H	45.45	204.525	363.60
5	Settled readings >2,0 NTU/24H	0.00	5	10.00
56	Sodium hydroxide lbs/24H	957.00	2395.69	3834.38
57	pH Raw	6.50	7.05	7.60

It should be noted, however, that the ratio and corresponding ranking values are computed cumulatively for the two outputs. Thus, a particular input variable that is not an important predictor for one output variable may be relatively important for the other output variable. In this case, because average turbidity and average are physically related, and hence are probably correlated**,** it is expected that there would be some consistency between these two variables. This apparent correlation is shown by the plotting of normalized average turbidity versus normalized average readings in Figure 21, where the values have been normalized between 0 and 1 to facilitate a more direct comparison.

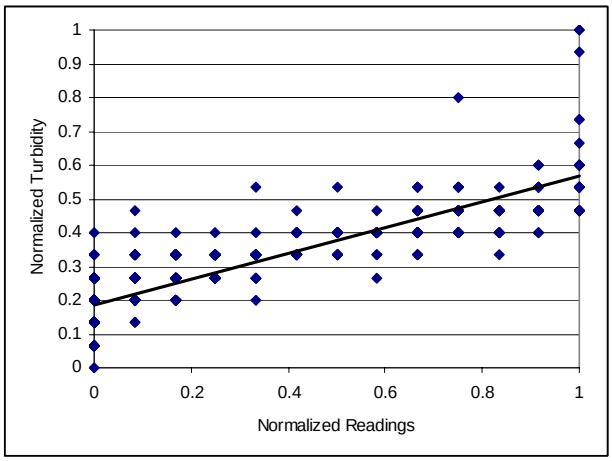


Figure 21. A regression chart comparing average daily turbidity versus daily readings about 0.1 NTU with a Line of Best Fit for Swimming River.

As the figure shows, lower average turbidity correlates with lower average reading, and similarly, higher average turbidity correlates with higher average readings.

Table 17 lists the variables in descending order relative to the computed sensitivity values, with a comparison to the ratio rankings.

Table 17. Comparison of Different Variable's Relative Effect on Average Daily Turbidity for the Dual ANN Output Model

Variable Name	RMSE Ratio Ranking	Range Change	Range Ranking	Sensitivity Ratio	Sensitivity Ranking
Temp Raw °C	1	-0.0136	1	-0.000798	3
Polymer lbs/24H	4	0.00986	2	3.1E-05	6
Air temp °C max/24H	3	0.00703	3	0.00019	5
Sodium hydroxide lbs/24H	56	0.00489	4	1.7E-06	7
pH Effluent	2	-0.00417	5	-0.00695	1
Settled readings >2,0 NTU/24H	5	0.004	6	0.0004	4
pH Raw	57	-0.00123	7	-0.00111	2

Figures 22 and 23 depict for two representative events the computed output response for average turbidity associated with individual changes in model inputs for the seven variables. Four variables consistently demonstrate a positive correlation with average daily turbidity, where increases in values for these inputs increased average daily

turbidity. The four variables include polymers added, sodium hydroxide, air temperature, and settled readings.

The single input variable that consistently exhibits an inverse relationship with average readings is raw water temperature. This inverse relationship is consistent with what was obtained with the ANN model with average turbidity as the single output. What appears to contract this relationship is how the model predicts that higher air temperatures result in a decrease of average daily turbidity. One would assume a positive correlation between air temperature and raw water temperature, and, by extension, would expect air temperature to be negatively correlated with average daily turbidity, as raw water temperature is. Both the input variables pH raw and pH effluent don't exhibit consistent behavior in terms of being positively or negatively correlated with average daily turbidity.

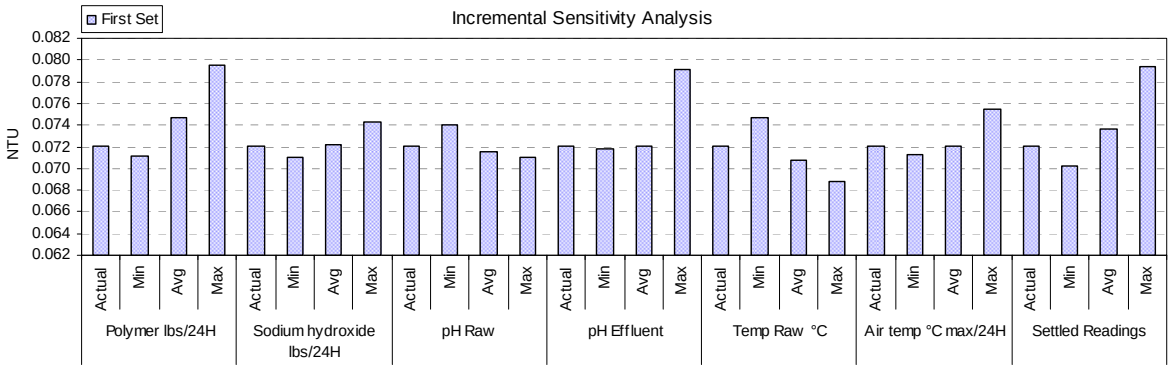


Figure 22. Predicted average daily turbidity in response to different input values for different variables for the first event for Swimming River

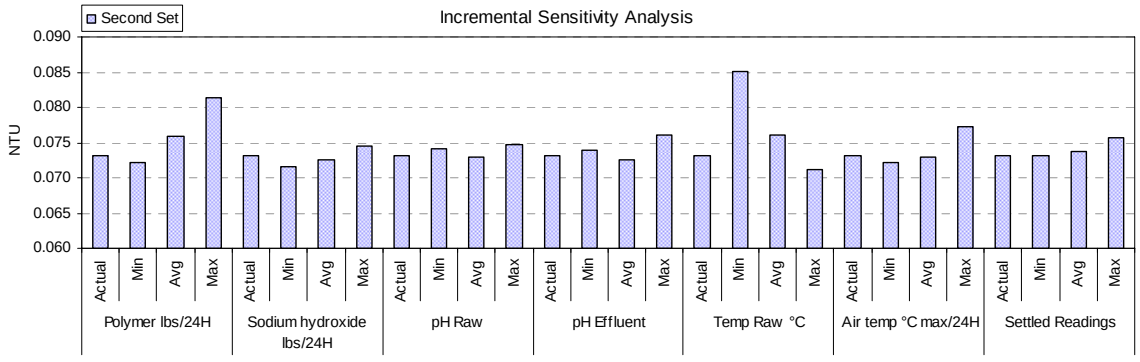


Figure 23. Predicted average daily turbidity in response to different input values for different variables for the second event for Swimming River

7. PVWC DATA

7.1 General Overview

Data utilized for the development of ANN models included the physical and chemical water data collected at three distinct sampling station locations. As discussed previously in the Study Area section, and depicted in Figure 2, Station 612 is located on the River A at the intake point for the Pumping Station 2. Station 101 is located just inside the mouth of the canal that connects River B with the PVWC treatment plant. Station 100 is located further inside the canal at the intake point for the PVWC water treatment plant. Measured conditions, then, at stations 612 and 101, are generally representative of water quality on Rivers A and B, respectively. Measured conditions at Station 100 are representative of water entering the treatment plant from the rivers and reservoir. When River B is not used for supply, hydraulic loading of water from the other source(s) minimizes its mixing at Station 100. Thus, if most of the source water for the day is extracted from either River A via the Pumping Station 2 or the Reservoir A, water quality at Station 100 is more representative of these water sources. Conversely, when River B is the predominant water source for the day, water quality conditions at Station 100 will be more representative of this river. Because of unique water quality conditions specific to each station, it was found using ANN models that the best predictive performance was achieved by modeling each station individually.

The sampling data used in this study were collected over the following time periods for the three stations; Station 100 - April 1999 to September 2004; Station 101 – May 1999 to October 2004; and Station 612 - April 1999 to September 2004. Available data sets at the three stations consisting of water quality, hydrologic, and weather were paired with corresponding counts of three different algae classes; Chrysophyta and Chlorophyta (under the kingdom Protista), and Cyanophyta (under the Kingdom Monera).

The initial data set used for developing and assessing the ANN methodology included most variables that are considered potentially important for predicting algae levels. Table 18 summarizes the number of historical data events available for modeling development and testing by station, potential prediction period, and data set (i.e. complete and reduced). From the data collected intermittently during the periods listed above, a total number of 156 sampling events were available for station 100, 41 events for station 101, and 52 events for station 612. For generating data events for model development purposes, however, incomplete data records reduced the final number of events available for ANN training and validation. That is, some sampling events had missing physical and/or chemical measurements, and in some cases, missing algal count measurements. In accordance with guidance provided by PVWC, some missing values were estimated by interpolation. For the second reduced data set used in this study, which excluded the less frequently measured water quality variables (four for Station 100, five for Stations 101 and 612), a total number of 266 sampling events were available for Station 100, 109 events for Station 101 and 173 events for Station 612. The actual number of these events available for model development and assessment depended on predicted on the length of forecasting horizons, as events separated by the prediction period (e.g. one-week ahead) had to be paired. (Note that 3-weeks ahead is included for previous modeling efforts – contained in Appendices).

Table 18. Available number of data events for each station for different modeling horizons

Station	Prediction Horizon	Original Models		Revised Models	
		Complete	Reduced Input	Complete	Reduced Input
100	One-week Ahead	156	266	182	270
	Two-week Ahead	160	249	167	252
	Three-week Ahead	150	236	142	218
101	One-week Ahead	41	109	47	96
	Two-week Ahead	40	106	35	66
	Three-week Ahead	48	110	36	71
612	One-week Ahead	52	173	51	174
	Two-week Ahead	54	139	40	140
	Three-week Ahead	56	135	45	135

As overviewed previously, data provided by the Passaic Water Valley Commission (PWVC) for the study station included water temperature, pH level, alkalinity, color, odor, conductivity, total hardness, turbidity, dissolved oxygen (DO), biochemical oxygen demand (BOD), sulfate, chloride, total phosphorus/orthophosphate, nitrite/ nitrate, ammonia, total suspended solids (TSS), UV 254, total organic carbon (TOC), algae counts for Chrysophyta, Cyanobacteria and Chlorophyta, fluoride and chemical oxygen demand (COD). Most of the data was provided in electronic format by PVWC and underwent further organization and assimilation by NOAH personnel. In this project, the later four constituents namely fluoride, COD, TDS and total solids were not used, as they were measured relatively infrequently. Climate data were obtained from NOAA and included total daily precipitation, wind speed, wind direction, sky cover, and heating degree days. Water source extraction data were provided by PVWC for Rivers A and B, Reservoir A (supplies water directly to Station 100) and Pumping Stations 1 and 2. Daily river flow data collected by the United States Geologic Survey for River A were also used.

As advised by Ms. Pasquarello of PVWC, data gaps at Station 100 for turbidity and alkalinity were estimated by taking a weighted average of values for these parameters measured at Stations 612 and 101, by proportioning their relative contribution to the PVWC treatment plant for the subject time period. Other missing data for a particular variable (e.g. temp, pH, conductivity, etc.) were estimated using linear regression with highly correlated variables; for example, turbidity was estimated from total algae counts using linear regression**.** Although less preferable to actual measurements, this is a reasonable approach for generating adequate data for this preliminary modeling assessment. In addition, this method was employed relatively infrequently; from less than 1% (one estimated value out of 166 events) for temperature and pH to 7% (12 out 166 events) for total suspended solids (TSS) for Station 100; 2% (temperature and pH) to 24% (total hardness) of the data for Station 101; and 1% (turbidity) to 17% (BOD) for Station 612.

Table 19 below presents a statistical tabulation of all the variables used in this modeling effort by Station. As was mentioned previously in section 5.2, and is discussed in more detail in the Modeling Section, a number of the less frequently measured variables, including some of the so-called "nutrients", were systematically eliminated as model inputs. In addition, so-called "correlative" or water extraction variables, were also eliminated from select models. Again, a complete presentation of the original models that used the complete set of input variables is presented in Appendices B-1 through B-4. The remainder of this section provides a detailed discussion of each of the variables used in this modeling problem. It should be noted that many of the variable types used for modeling the PVWC system were also used for modeling the Swimming River reservoir system.

Table 19. Statistical tabulation for all model variables used by Station

Parameter*	Unit	Station 100			Station 101			Station 612		
		Min	Ave	Max	Min	Ave	Max	Min	Ave	Max
Water Temperature	^oC	0.1	13.6	25.8	0.6	13.5	26.5	0.0	13.2	27.8
pH		6.4	7.4	8.8	6.3	7.3	8.5	7.0	7.6	9.8
Turbidity	NTU	0.0	4.0	22.5	2.6	9.7	22.1	1.7	7.1	55.0
Alkalinity	mg/L as CaCO3	0.0	41.6	103.5	6.4	72.7	106.0	30.0	68.1	110.0
Total Hardness	mg/L as CaCO3	38.0	101.5	194.0	30.0	116.3	192.0	60.0	108.6	220.0
Conductivity	umhos/cm	168	432	1312	203	540	1307	253	445	756
Color	umhos/cm				15.1	44.1	153.0	12.0	34.8	195.0
Dissolved Oxygen	mg/L	4.8	10.5	16.2	5.3	9.6	15.1	4.8	10.3	15.5
Biochemical Oxygen Demand	mg/L	0.7	4.2	10.2	0.4	4.2	8.1	4.8	10.3	15.5
Chloride	mg/L	25.0	85.1	283.0	52.0	109.0	192.0	44.0	87.8	174.0
Sulfate	mg/L	10.3	22.2	55.7	12.2	27.7	61.1	10.6	20.2	39.3
T.Phosphorus/Orthophos-phate	mg/L	0.0	0.3	1.7	0.0	0.7	2.5	0.0	0.2	1.5
Nitrite/Nitrate	mg/L	0.0	1.6	5.3	0.4	2.3	6.2	0.4	1.4	3.7
Ammonia	mg/L	0.0	0.1	0.4	0.0	0.1	0.4	0.0	0.1	0.2
Total Suspended Solids	mg/L	1.0	13.4	47.0						
UV-254	cm^{-1}	0.1	0.1	0.3	0.1	0.2	1.0	0.1	0.1	0.1
Total Organic Carbon	mg/L	2.7	4.8	11.3	3.1	6.2	11.4	0.0	3.8	8.5
Prediction Period's Precipitation Total	inches	0.0	0.7	3.6	0.0	0.8	4.6	0.0	0.9	8.1
Prediction Period's Lagged Precipitation Total	inches	0.0	0.9	8.2	0.0	0.8	3.0	0.0	0.6	3.0
Wind Direction	$0\text{-}360^0$	52	211	301	93	214	293	120	219	299
Wind Speed	mph	4.7	7.4	11.2	4.7	7.3	10.9	4.9	7.6	10.9
Heating Degree Days	^oC	0.0	12.5	45.6	0.0	12.6	38.2	0.0	13.0	45.5
Sky Cover		0.9	4.3	9.1	0.6	4.2	8.7	0.4	4.2	8.0
Length of Day	hour	9.2	12.5	15.1	9.3	12.6	15.1	5.3	12.2	15.1
River A Streamflow	cu.ft./sec	30	546	3684	40	375	2569	30	379	4159
River A Extraction	MGD	0.0	35.4	62.1	0.0	31.8	59.1	0.0	29.9	60.2
River B Extraction	MGD	0.0	19.0	89.7	0.0	24.6	89.7	0.0	26.7	86.4
Reservoir A Extraction	MGD	0.0	0.4	19.5	0.0	0.6	19.5	0.0	0.6	18.3
Pumping Station 1 Extraction	MGD	0.0	0.6	28.0	0.0	1.8	31.4	0.0	0.9	29.6
Total Algae Count	Cells/ml	8	167	1520	8	115	572	16	187	842
Chrysophyta Count	Cells/ml	0	84	656	0	67	200	0	88	616
Cyanophyta Count	Cells/ml	0	30	1364	0	6	180	0	47	760
Chlorophyta Count	Cells/ml	0	49	608	0	40	304	0	48	420

Min = minimum, Max = maximum, Ave = Average

* - Total Amorphous Materials were measured as Light, Medium and Heavy

7.2 Biological Data

7.2.1 Cyanobacteria

Cyanobacteria, more commonly known as blue green algae (after the color of the first known species), are prokaryotic organisms that belong to the kingdom Monera. They are relatives of bacteria because they share similar structures but they are also related to the chloroplasts present in eukaryotic organisms (algae). They are sometimes referred to as blue-green algae. Cyanobacteria occur naturally in fresh water and in marine environments and along with the eukaryotic algal phyla, comprise the important component of phytoplankton community. However under favorable conditions, they can become dominant and can turn into a phenomenon called a bloom. Cyanobacteria blooms appear mostly in summer months or during periods of hot weather, when lack of mixing of surface and deeper water layers in a river or reservoir can lead to thermal stratification. Factors that contribute to thermal stratification of water bodies include temperature, wind, solar radiation and flow. Cyanobacteria prosper in non-turbulent conditions where water column stability enhances their ability to maintain an optimum position in the water column for photosynthetic activity and growth.

Similar to other algae, cyanobacteria are photoautotrophic organism, producing their own food by using chlorophylls (and other pigments) to fix carbon as starch through the process of photosynthesis. Nutrients also play a role in their growth and reproduction. Dissolved organic carbon, as well as phosphate, nitrate, ammonia and iron, are important factors. Cyanobacteria are one of very few groups of organisms that can convert ("fix") inert atmospheric nitrogen into an organic form, such as nitrate or ammonia. Cyanobacteria reproduce asexually and photosynthesis also plays a large and important role in their reproduction and growth. The wavelength of the light available determines what form of Cyanobacteria will grow.

The taxa included in the phylum for Cyanobacteria published under the Bacteriological Code (1990 Revision) include the classes *Chroobacteria*, *Hormogoneae* and *Gloeobacteria*; the orders *Chroococcales*, *Gloeobacterales*, *Nostocales*, *Oscillatoriales*, *Pleurocapsales* and *Stigonematales*; the families *Prochloraceae* and *Prochlorotrichaceae*, and the genera *Halospirulina*, *Planktothricoides*, *Prochlorococcus*, *Prochloron*, *Prochlorothrix*.

Figure 24 compares measured cyanobacteria counts for each of the three stations with respect to time. At Station 100, over the four year period, cyanobacteria bloom episodes occurred from June to October 1999, with the highest levels measured in August and September of this year at 1072 and 1364 cells/ml, respectively. By comparison, relatively small counts of cyanobacteria occurred over the remaining years, with a minor algal bloom occurring in July 2002 with a count of approximately 200 cells/ml.

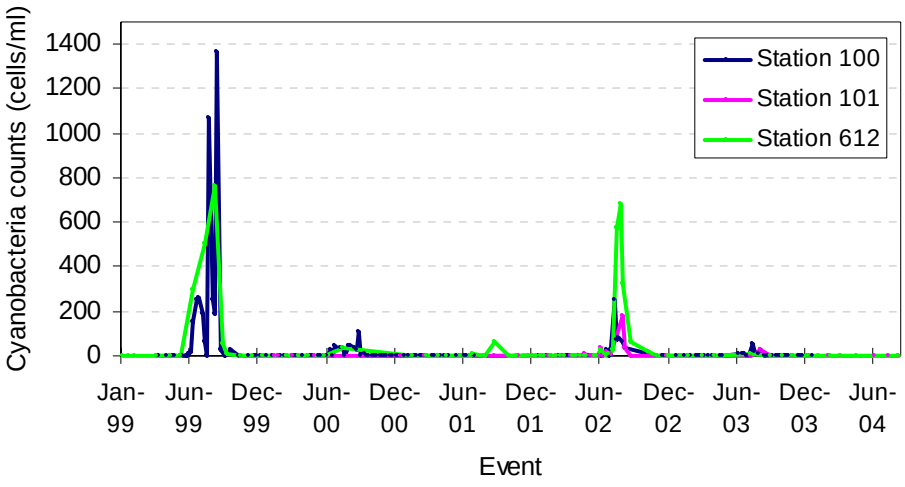


Figure 24. Comparison plots of measured cyanobacteria counts at three sampling stations

At Station 101, no data were available in 1999 to compare algae counts with Station 100. An algal bloom did occur at this station from July through August 2002, with the highest levels measured at 180 cells/ml in August.

Like Station 100, Station 612 had two cyanobacteria bloom events in the same years, the first in 1999 and the second in 2002. As shown in the figure, the bloom episodes that occurred in 1999 at this station (with the highest of 506 and 760 cells/ml in August and September, respectively) coincided with the blooms at Station 100. Similarly, the 2002 bloom incidence (highest at 684 cells/ml in August) also coincided with the bloom at Station 101. The 2002 bloom occurred over the period when water was extracted from the Point View Reservoir (July 24 to August 30). The Pumping Station 2 was temporarily turned off on July 23, 2002 and the Reservoir A was online the following day. Over this five week period, an average of 35% (13-100%) of daily water supply came from the Reservoir A. This water source was likely used to mitigate the algal bloom on the River A. By contrast, in 1999, only the River A through the Pumping Station 2 and the River B supplied water to the treatment facility.

In summary, then, the cyanobacteria bloom episode that occurred at Station 612 in July to September 1999 coincides with the bloom at Station 100. Station 101 during this period did not experience any bloom episodes. Another coincident bloom at Stations 100 and 612 occurred in July 2002; however, the algae levels at Station 100 started to decline by the end of the month, as the alternative reservoir source was used. At Station 612, the high algae counts persisted until the middle of August, around which time a bloom at Station 101 occurred. Most of the time, particularly during low cyanobacteria periods, counts at the three stations are more or less equal. However, during bloom events, cyanobacteria counts at Station 612 could be as much as eight times more than the other two stations. Thus, the River A appears to be the surface water source most subject to cyanobacteria bloom events, and extraction of this source appears to be responsible for bloom events measured at Station 100.

7.2.2 Chrysophyta

Chrysophytes are eukaryotes organisms that are abundant in freshwater, but may also be found in marine environments. These protists are particularly important in lakes where

they are the primary food source for zooplankton. Chrysophytes are also called golden algae because the yellow and brown caroteneoid (including fucoxanthin) and xantophyll accessory pigments masks the chlorophylls a and c, imbuing them with a golden color. There are more than a thousand described species of golden algae, mostly unicellular or colonial, swimming or floating in lakes and oceans as phytoplankton. In many chrysophytes, the cell walls are composed of cellulose with large quantities of silica, while some species are amoeboid forms with no cell walls. Motile species may have one or two flagella, which can be similar or dissimilar in structure. In shallow ponds that dry up in summer or freeze completely in winter, golden algae survive by forming protective cysts that can withstand the harsh conditions. When favorable conditions return, the algae emerge from the cysts.

The vast majority of golden algae are autotrophs (photosynthetic). They store food outside of the chloroplasts in the form of oil or polysaccharide laminarin, or chrysolaminarin. However, some biologists do not consider them to be truly autotrophic since they become facultatively heterotrophic (get energy from non-photosynthetic sources) in the absence of adequate light, or in the presence of plentiful dissolved food. When this occurs, the chrysoplasts atrophies and the alga may turn predator, feeding on bacteria or diatoms (www.ucmp.berkely.edu). Very few chrysophytes do not have chloroplasts, and those few live on dead organic material. However, some golden algae can use dead and organic materials, or use chloroplasts for energy, and thus switch back and forth between the two modes of nutrition, depending upon conditions (Lund and Lund).

Motile forms of golden algae reproduce by fission, while non-motile forms produce motile zoospores. There is also a special type of spore unique to this group, known as a statospore. It is spherical with a plug, which is popped out as the spore germinates. Isogamous sexual reproduction is rare. The unicellular species with a single flagellum include *Chromulina, Chrysococcus* and *Mallomonas* and the larger colonial forms comprise of *Synura, Chrysophaerella*, *Uroglena* and *Dinobryon.*

Figure 25 compares measured chrysophyta counts for each of the threes stations over time.

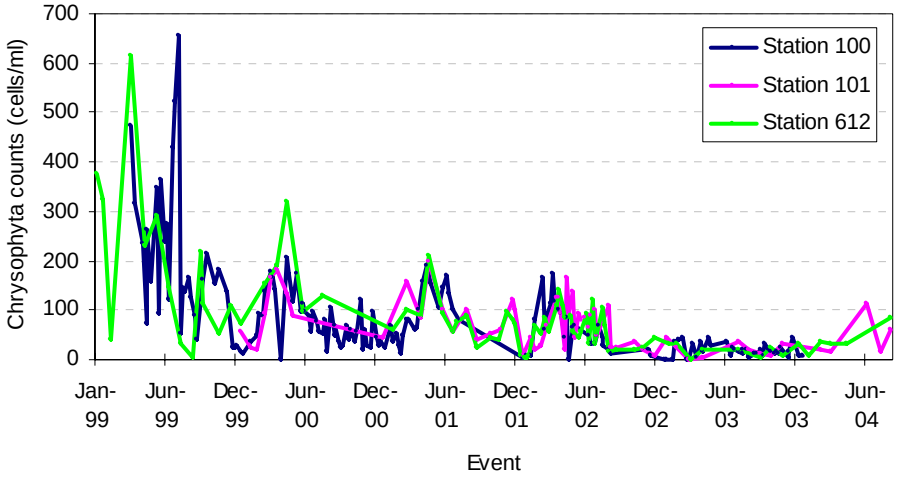


Figure 25. Comparison plots of measured chrysophyta counts at three sampling stations

At Station 100 (PVWC intake point at treatment facility), the highest counts of Chrysophytes over the period 1999 into 2004 were recorded in July and August 1999 at 520 and 656 cells/ml, respectively. With the exception of 2003, when measured concentrations were relatively low throughout the year, seasonal patterns over each of the years are evident. Higher Chrysophytes levels generally occurred in late winter/early spring (March) until the earlier to middle part of summer (June or July), while for the remainder of the year, counts were low.

At Station 101 (River B), the highest Chrysophytes counts were measured in April 2000 at 184 cells/ml and in May 2001 at 200 cells/ml. In general, a temporal pattern similar to Station 100 is evident, in which counts rise begin around the month of March until June, after which levels decline and remain low for the rest of the year. As at Station 100, low counts were also recorded at this station over 2003.

Station 612 (River A) exhibited a pattern consistent with the other two stations, with higher levels beginning in March and lasting though June or July followed by lower

levels through the end of the year. The highest Chrysophytes counts were measured in April 1999 at 616 cells/ml, and in May 2000 at 320 cells/ml.

7.2.3 Chlorophyta

Chlorophytes are eukaryotic organisms that form the largest phylum of the algae, the Chlorophyta. Commonly known as green or grass-green algae, the abundance of chlorophyll pigments in their chloroplasts gives them the bright green color. There are about 8,000 species of chlorophytes, mostly aquatic, but some are terrestrial living on damp soil or attached to land plants and even in snow and ice. These protists range in size from microscopic unicellular plankton that grow in lakes and oceans, to colonial filaments of pond scum, to leaflike seaweeds that grow along rocky and sandy intertidal areas. Some species are symbiotic, forming lichens with fungi or living with corals, while others can be found in freshwater sponges, imbuing the sponges with a bright green color, or in permanent snow banks, turning the snow reddish in color because of the secondary pigment that masks the chlorophyll. These green algae are also found on damp soil or attached to land plants (a few are parasitic).

Green algae are photosynthetic, containing organelles called chloroplasts which are characterized by clearly stacked thylakoids (any of the membranous disks of lamellae within plant chloroplasts that are composed of protein and where photochemical reactions of photosynthesis occur). Their photosynthetic pigments contain chlorophylls a and b, and various carotenoids are found in plants and in similar proportions. Food reserves are stored inside their plastids in the form of starchs, fats and oils. Most green algae have cell walls made up of two layers: an inner cellulose layer and an outer layer of pectin. Like most protists, chlorophytes have two or more flagella (tail like appendage) near the apex of the cell, at least once in their life cycle.

Green algae may reproduce vegetatively, by fragmentation and by cell division; asexually, by means of spores and zoospores that develop directly into new plants; and

sexually by the fusion of pairs of sex cells (gametes). In multicellular chlorophytes, alternation of generations is common, where the algae alternate between gametophyte and sporophyte generations.

Green algae are an extremely important source of food for other aquatic life forms. This makes green algae vital to many ecosystems on the planet. They also make a major contribution to the world's oxygen supply through photosynthesis and by fixing approximately 10^{10} tons of carbon per year. Classes of Chlorophyta include: *Chara, Chlamydomonas, Cladopjora, Coleochaete Desmid, Eremosphaera, Hydrodictyon, Oedogonium, Pandorina, Pediastrum, Spirogyra Ulothrix, Ulva and Volvox.*

Figure 26 compares measured chlorophyta counts for each of the three stations with respect to time.

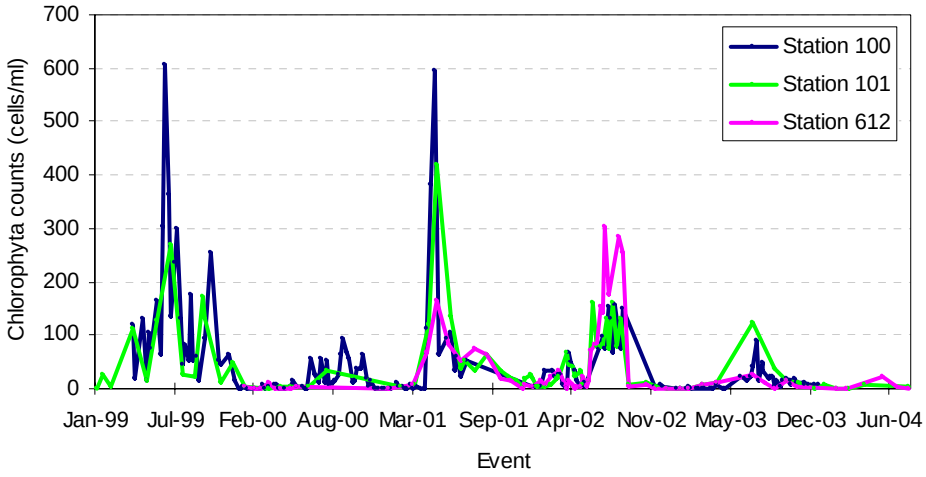


Figure 26. Comparison plots of measured chlorophyte counts at three sampling stations

At Station 100, counts were comparatively high throughout 1999, while levels over the succeeding years through 2004 counts were low, with exception to one sudden rise in May 2001. As depicted by the figure, higher levels generally occurred in the middle part of the year from approximately May to September, and the highest counts occurred in June 1999 and May 2001 with counts of 608 and 596 cells/ml, respectively. At Station

101, the highest Chlorophyte counts was measured in May 2001 at 164 cells/ml, and in 2002, high values of 304 and 284 cells/ml were measured in July and August, respectively. Station 612 exhibited more or less the same pattern as Station 100, wherein high chlorophytes levels usually occurred in May to September. The highest were measured at this station 612 were measured at 268 cells/ml in July 1999 and 420 cells/ml in May 2001. Because most of the source water at Station 100 originates from the River A, the correlation between Stations 100 and 612 is not unexpected.

7.3 Physical Data

7.3.1 Temperature

Most algae are freshwater organisms, and like most aquatic organisms, are unable to internally regulate their core body temperature. Consequently, temperature exerts a major influence on their biological activity and growth. Each species thrives within a narrow range of temperature, with some preferring cooler temperatures over warmer temperatures, while others have the opposite preference. When temperature gets too far above or below the preferred range, algae concentration decreases until there may be few or none.

Water temperature data collected at the three sampling stations were plotted against algae counts, and are presented in Figures 27 through 29. Temperature as expected exhibits a seasonal component. Different relationships between temperature and algae counts are exhibited by algae types and Station locations. At Station 100 as shown in Figure 27, higher chrysophytes level from 2000 to 2003 coincide with periods before temperatures reached their peak for each year, usually within the range of 10 to 20^{o}C, while the highest chrysophtes levels in 1999 coincide with periods when temperatures were highest (above 20^{o}C) in the year. Higher Chlorophytes counts that occurred primarily during the middle part of the year also coincide with high temperature periods above 15^{o}C. Similarly, the

two cyanobacteria bloom events in 1999 at this station coincide with periods of highest temperatures, measured from 20 to above 25°C.

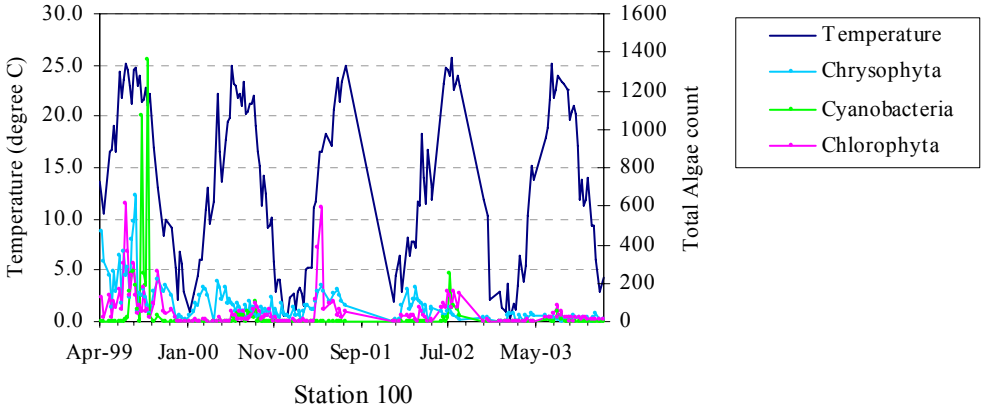


Figure 27. Total Algae counts versus Temperature measured at Station 100

At Station 101, chrysophytes counts from in 2000 through 2003 exhibited an almost identical pattern to that observed at Station 100, where higher counts coincide with periods before temperature (below 20°C) reached their annual peaks. High Chlorophytes counts that occurred in 2002 coincided with higher temperatures measured above 20°C, while in 2001, higher chlorophytes counts occurred when temperatures were lower (around 15°C). Cyanobacteria blooms at this station that occurred in 2002 coincide with periods of high temperatures, measured above 25°C. The relationship between temperature and high cyanobacteria and chrysophytes counts at this station was observed to be consistent with that of Station 100.

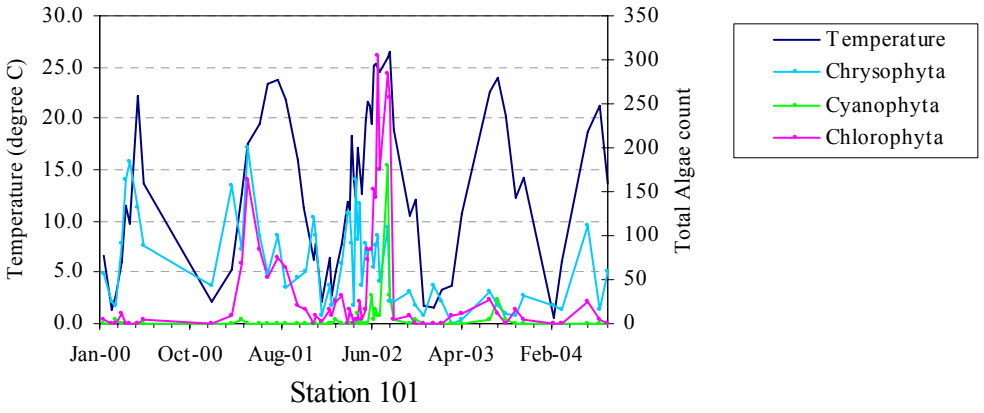


Figure 28. Total Algae counts versus Temperature measured at Station 101

At Station 612, the three algae types exhibited somewhat similar behavior to those from the other two stations. Most high Chrysophytes levels occurred during relatively warmer periods, around 10 to 20°C, as temperature warmed up to their peaks (>20°C) towards the summer period. As for the chlorophytes, higher levels during the years 1999 and 2002 coincide with periods of higher temperatures measured above 20°C. In 2001, however, the highest algae levels occurred when temperatures were lower, around 15°C. Cyanobacteria exhibited a consisted pattern at all three stations, in which blooms occurred during times when temperatures were highest (above 20°C in 1999 and above 25°C in 2002).

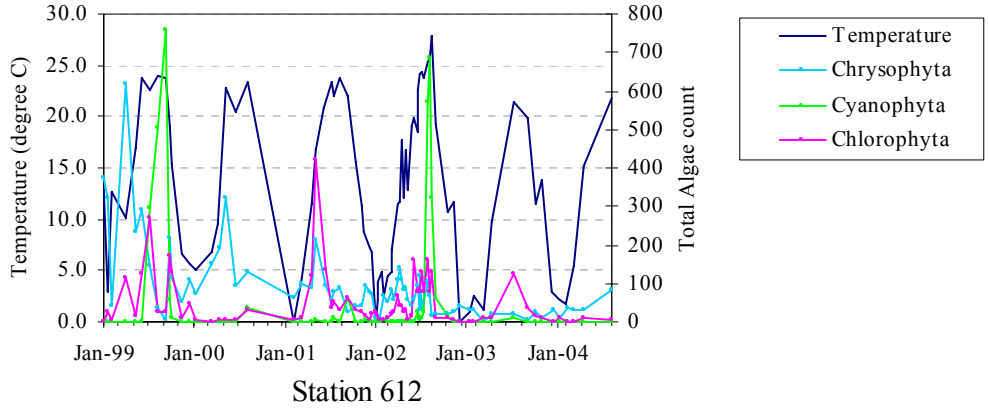


Figure 29. Total Algae counts versus Temperature measured at Station 612

7.3.2 pH

The pH is by definition a measure of concentration of hydrogen ions in a water sample (negative log of the concentration). During photosynthesis, algae consume hydrogen molecules, diminishing the hydrogen ion concentration, and therefore increasing the pH. Consequently, pH is generally higher during daylight hours or during the growing seasons, while the process of respiration and decomposition decreases pH.

As depicted in Figures 30, 31 and 32, pH levels fluctuate from 6.4 to 8.8 at station 100, 6.3 to 8.5 at station 101 and from 7 to 9.8 at station 612, respectively. High algae levels coincide with high pH levels at the corresponding sampling stations. Chrysophyta,

chlorophyta, and cyanobacteria are organisms that produce their own food through photosynthesis, thereby increasing the pH level during the process.

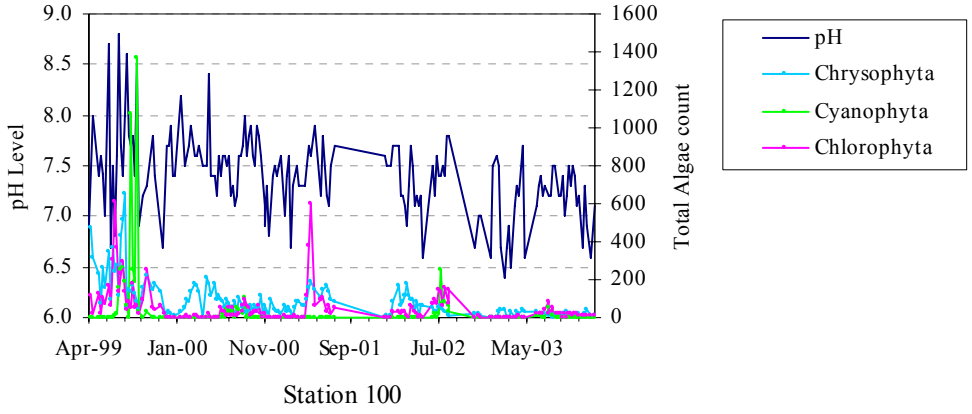


Figure 30. Total Algae counts versus pH measured at Station 100

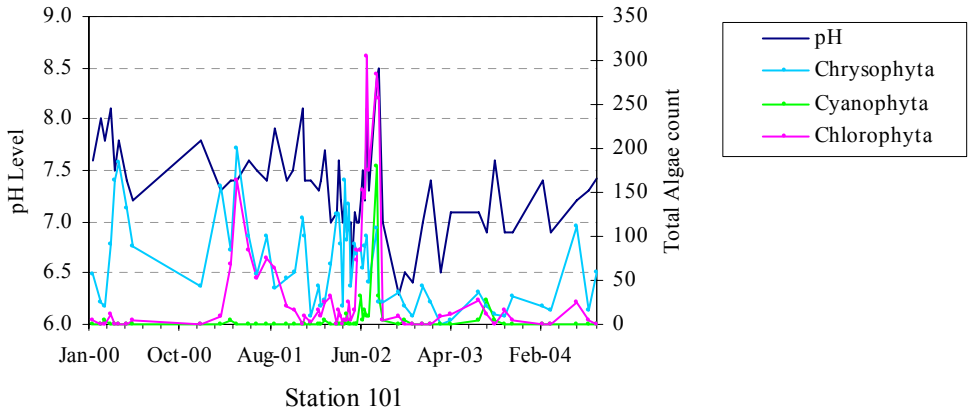


Figure 31. Total Algae counts versus pH measured at Station 101

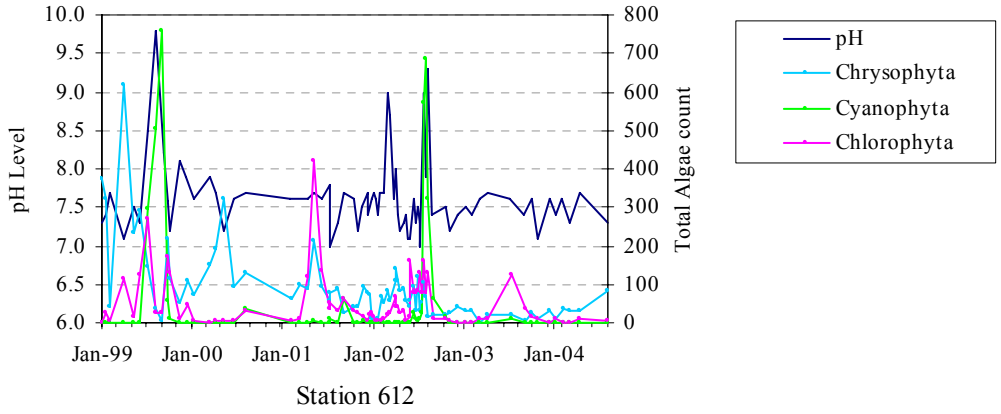


Figure 32. Total Algae counts versus pH Level measured at Station 612

7.3.3 Turbidity

Turbidity refers to "solids and organic matter that do not settle out of water" (Driscoll, 1986), and can include constituents ranging from clay, silt, and plankton to industrial wastes and sewerage. It is an important factor in algae prediction because it affects the light penetration into the water column, which regulates photosynthesis of organisms as well as the temperature of water that affects their biological activities. However, the measured turbidity may also be a measure of the amount of algae in the water column.

Figures 33 through 35 depict the comparison plots of turbidity levels against total algae counts at the three sampling stations. At Station 100 (water treatment intake point), no clear relationship between algae levels and of turbidity levels emerged. At Station 101 (River B) however, higher levels of turbidity coincide with higher levels of chrysophyta chlorophyta and cyanobacteria, most evident of which was in August 2002. During this month, the highest recorded turbidity level was measured at 22 NTU, which coincided with a cyanobacteria bloom, as well as the highest overall chlorophytes level measured at this station. In addition, the chrysophytes count was also relatively high during this period.

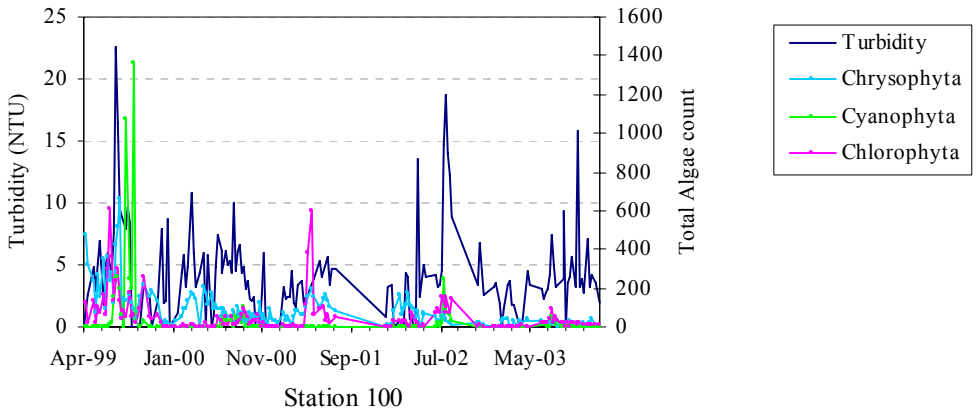


Figure 33. Total Algae counts versus Turbidity measured at Station 100

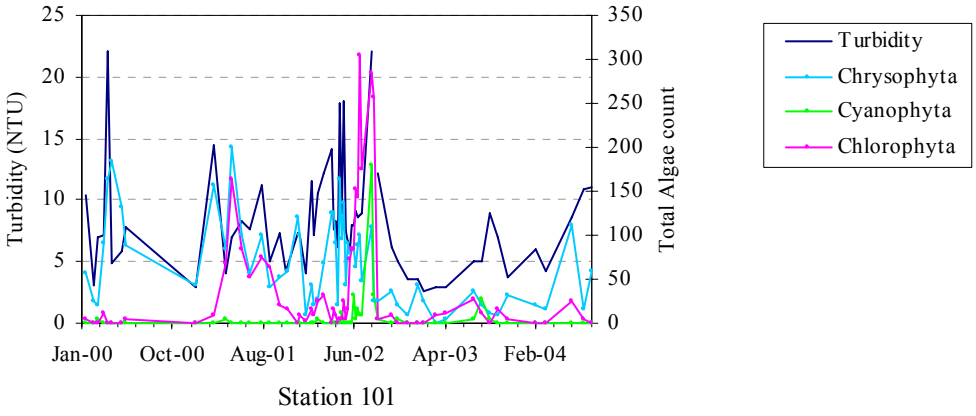


Figure 34. Total Algae counts versus Turbidity measured at Station 101

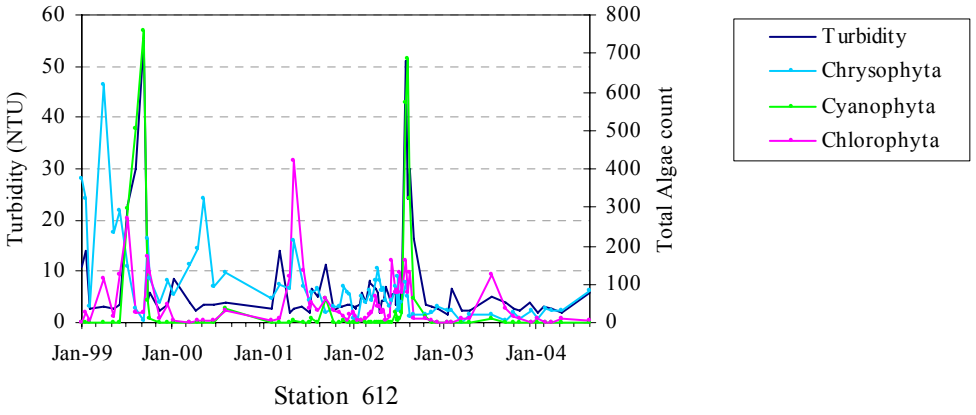


Figure 35. Total Algae counts versus Turbidity measured at Station 612

At Station 612, the highest turbidity levels over the 4-year records (51 and 55 NTU in September 1999 and July 2002, respectively) coincide almost perfectly with cyanobacteria bloom incidences. The other two algae types, chrysophyta and chlorophyta, exhibited different behavior, in which higher levels occurred during times when turbidity levels were lower.

In conclusion, there may be some threshold at which a higher turbidity levels during the right conditions is indicative of or corresponds with higher algae levels and/or limiting nutrients and minerals. However, above this threshold, elevated turbidity may reflect a high presence of material (e.g. sediments) which, by limiting sunlight penetration, inhibits algae growth.

7.3.4 Alkalinity

Alkalinity is a measure of the capacity of water to neutralize acids, as it measures how much acid can be added to water without causing a significant change in pH (i.e. buffering capacity). If any chemical changes are made to the water that could raise or lower the pH value, alkalinity acts as a buffer, protecting the water and its life forms from sudden shifts in pH.

Alkalinity is important in rivers and lakes because it buffers pH changes that occur naturally during photosynthetic cycles, or un-naturally by addition of acids to water via acid rain. The normal pH value of natural surface water bodies is usually between 6.5 and 8.5 but raising the alkalinity almost always raises the pH, which in turn affect aquatic life. If the alkalinity is too high, the water can become too cloudy and inhibit photosynthesis.

At the PVWC utility, as shown by Figures 36, 37 and 38, contradictory behavior was observed at the different stations. At Station 100, algae thrived during times when alkalinity was low. The highest alkalinity at this station was measured in July 2000 at 103 mg/l, and algae levels were correspondingly low**.** However, at Stations 101 and 612, higher algae levels generally occurred during periods of high alkalinity. The high alkalinity level of 98 mg/l in August 2002 at station 101 (highest at 106 mg/l in July) and the highest alkalinity level of 110 mg/l at stations 612 measured in July 2002, coincided with cyanobacteria bloom episodes at their respective stations. Moreover, at station 101, the highest chlorophytes counts occurred in July and August 2002, at exactly the same periods when alkalinity levels were highest. Another obvious coincidence between high chlorophytes and chrysophytes counts and high alkalinity occurred at this station in May 2001. At station 612, the highest chlorophytes counts occurred in May 2001 during relatively high alkalinity, while another cyanobacteria bloom episode in September 1999 at this station also coincide with high alkalinity above 70 mg/l.

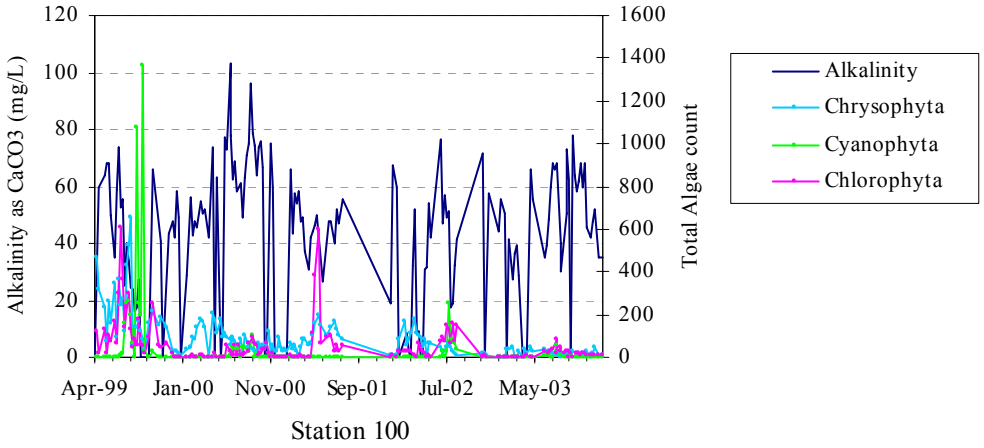


Figure 36. Total Algae counts against Alkalinity measured at Station 100

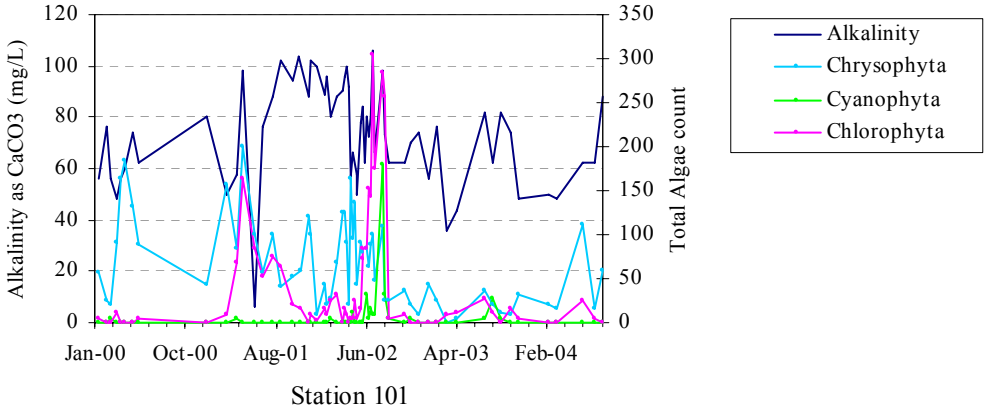


Figure 37. Total Algae counts versus Alkalinity measured at Station 101

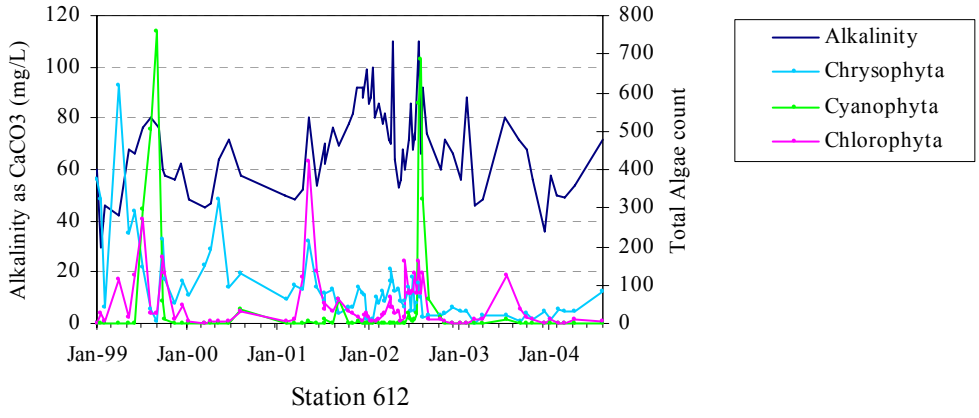


Figure 38. Total Algae counts versus Alkalinity measured at Station 612

7.3.5 Total Hardness

Hardness is a measure of the concentration of divalent cations dissolved in water, and the principal components in most natural water systems are typically calcium and magnesium ions. Hardness is generally defined by the following equation;

$$\text{Hardness} = 2.5(Ca^{2+}) + 4.1(Mg^{2+}) \quad \textbf{(6)}$$

where the Ca^{2+} and Mg^{2+} ion concentrations are measured in mg/l. The hardness represents the equivalent concentration of $CaCO_3$ that would produce an equivalent effect in terms of forming a soft precipitates or reacting in boilers to form a solid scale precipitate. The factors 2.5 and 4.1 represent the ratio of $CaCO_3$ formula mass to Ca and Mg atomic masses, respectively. Water with hardness below 75 mg/l is considered "soft" and water with hardness above 150 mg/l is considered hard.

At Station 100 as shown in Figure 39, the cyanobacteria bloom episodes in August and September 1999 occurred when corresponding total hardness was measured within the range 107 to 144 mg/l of $CaCO_3$, which by standard can be considered as moderately hard to hard water. The highest chrysophyta and chlorophyta counts occurred in the same year when water was hard (total hardness is above 150 mg/l of $CaCO_3$). Figure 40 depicts the comparison plot of different algae counts against total hardness at Station 101. As shown, the highest levels in all three algae types occurred during times when water was hard (above 120 mg/l of $CaCO_3$). As depicted by the figure, other higher chrysophyta concentrations coincided with moderately hard water as well. At Station 612, the two cyanobacteria bloom events and the highest chrysophyta and chlorophyta event coincided with moderately hard water (below 120 mg/l of $CaCO_3$). Figure 41 depicts these trends.

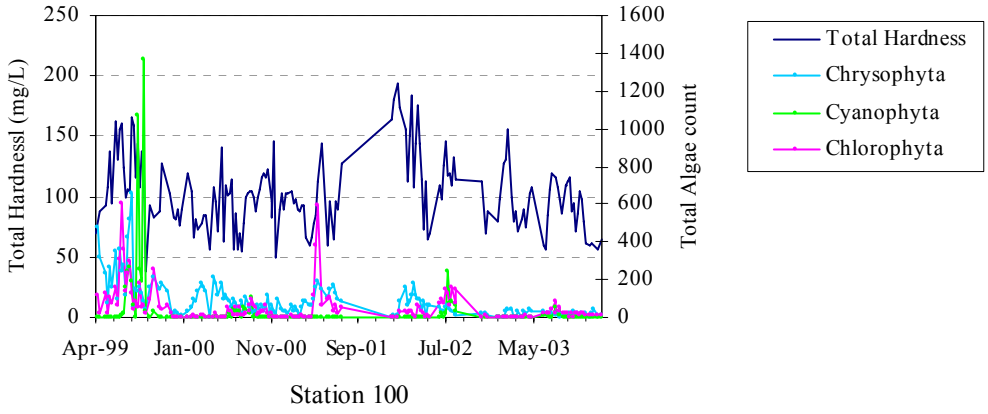


Figure 39. Total Algae counts versus Total Hardness measured at Station 100

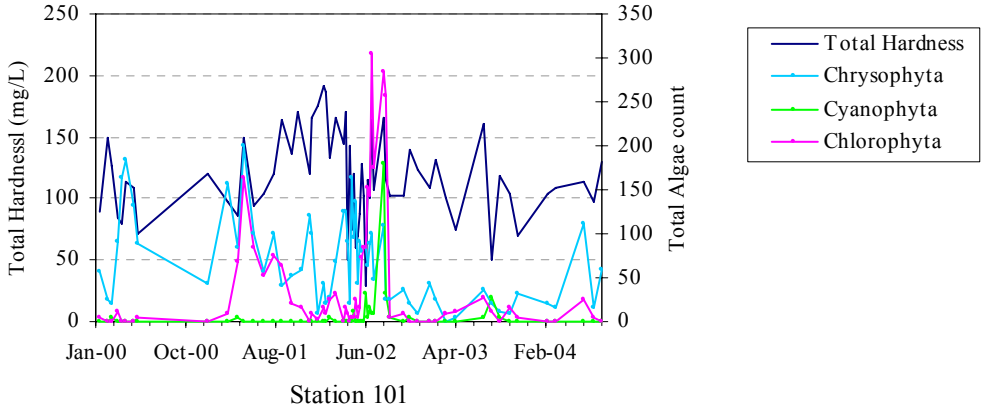


Figure 40. Total Algae counts versus Total Hardness measured at Station 101

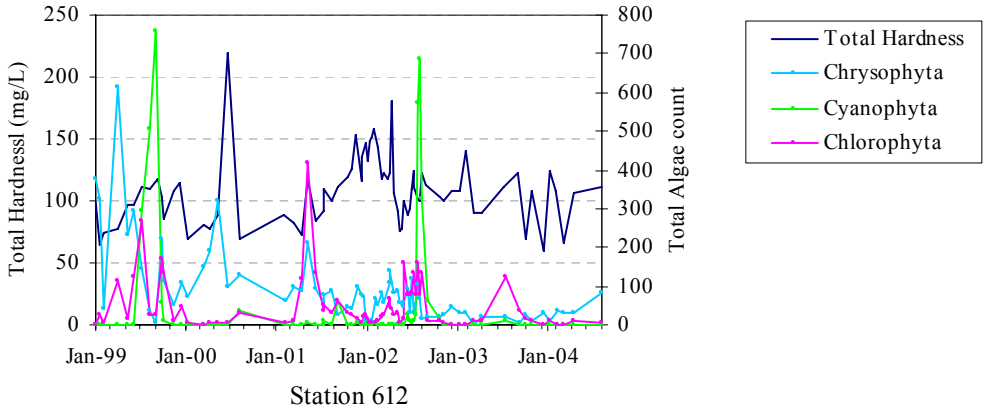


Figure 41. Total Algae counts versus Total Hardness measured at Station 612

7.3.6 Conductivity

Conductivity is a meaure of electrical current flow through a solution, is expressed in units of microSiemens (*u*S). Because conductivity increases nearly linearly with increasing ion concentration, it is generally found to be good measure of the concentration of total dissolved solids in water, which may contain nutrients available for algal growth, or other ions, such as chloride, which may inhibit it.

Figures 42 through 44 depict the algae counts against the conductivity measurements at the three sampling stations. All three stations exhibited the same distinct patterns: high algae levels corresponding with high conductivity measurement, and low algae levels with low conductivity measurements. The highest conductivity values measured at three stations occurred during the same measurement event, January 2002, at 1312 umhos/cm for Station 100, 1307 umhos/cm for Station 101, and 756 umhos/cm for Station 612.

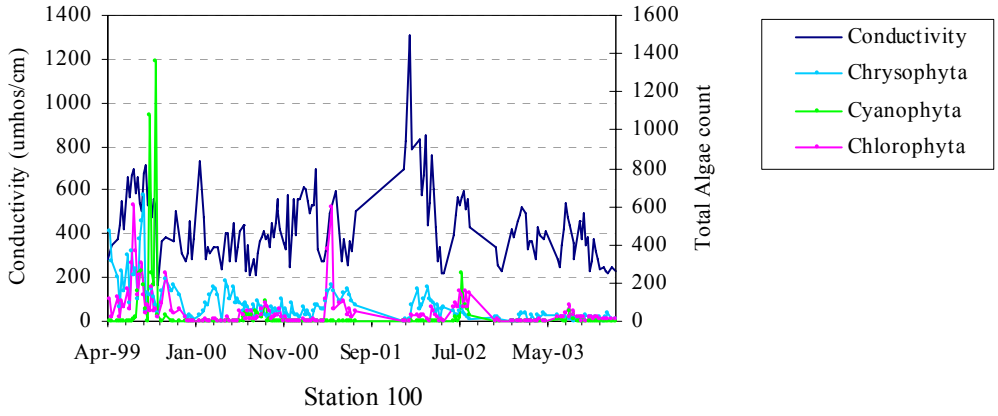


Figure 42. Total Algae counts versus Conductivity measured at Station 100

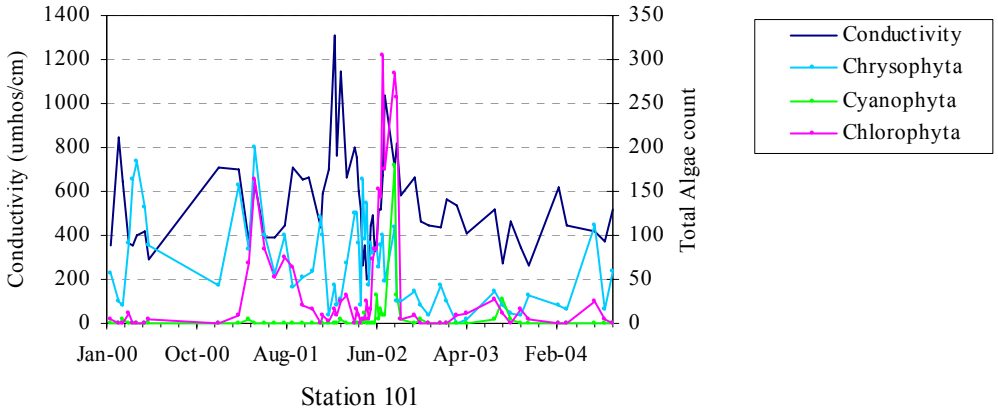


Figure 43. Total Algae counts versus Conductivity measured at Station 101

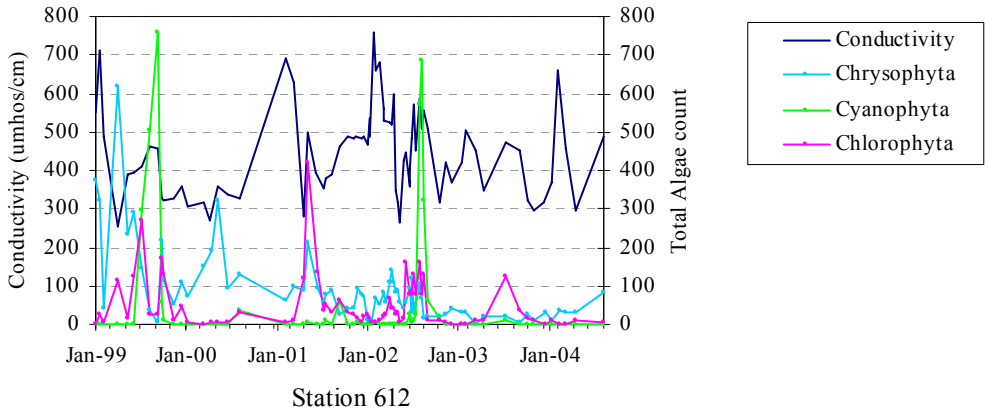


Figure 44. Total Algae counts versus Conductivity measured at Station 612

Algae levels during this event were correspondingly low at all three stations. At station 100, the cyanobacteria bloom episodes in August and September 1999, as well as the high chrysophytes and chlorophytes counts occurred during periods when conductivity were moderately high, from 593 to 717 756 umhos/cm. At Station 101, the single cyanobacteria bloom episodes in August 2002 coincide with conductivity level of 701 umhos/cm, while the highest chrysophytes counts that occurred in July to August 2002 occurred during periods when conductivity was considerably high, within the range of 637 to 1037 umhos/cm. High chlorophytes counts, on the other hand, coincide with low to moderate conductivity. At Station 612, the two cyanobacteria bloom episodes in September 1999 and August 2002 coincide with moderate conductivity level, from over 400 and over 500 umhos/cm, respectively.

7.3.7 Total Suspended Solids

Total Suspended Solids (TSS) consist of sand, silt and fine organic matters such as leaves, pieces of woods, etc. suspended in streams or lakes. High flows can increase TSS as water erodes banks and prevent the suspended solids from settling to the river bottom. High suspended solids can also exacerbate chemical degradation of water quality. Pesticides and bacteria that attach to them are more readily transportable, and their advection downstream can kill plants and animals, while making the water less safe or even undrinkable to humans and wildlife.

TSS data were available only for Station 100. As depicted by Figure 45, the four-year record at this station shows that the highest TSS at 47 mg/l was measured in March and April 2001. It was observed that the cyanobacteria bloom episodes that occurred in August and September 1999 coincide with higher TSS values measured at 16 and 34 mg/l, respectively. During lower periods of TSS, all three algae levels were correspondingly low.

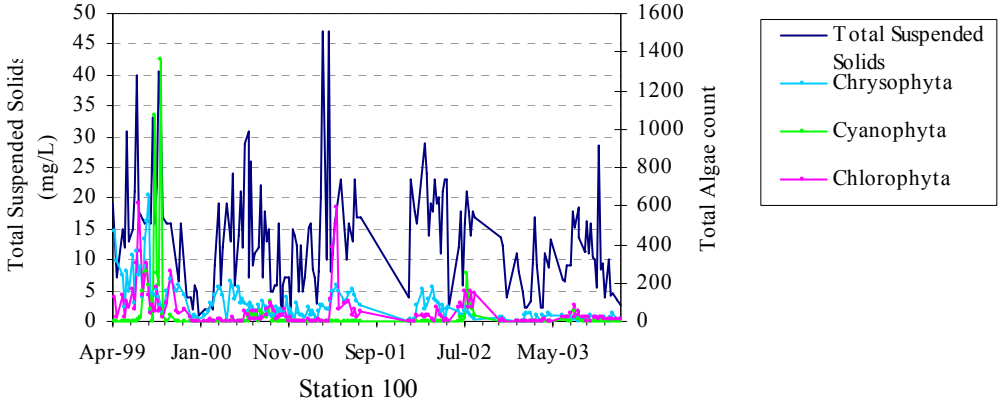


Figure 45. Total Algae counts versus Total Suspended Solids measured at Station 100

7.3.8 Total Amorphous Materials

Aggregate particles not passing through the 0.45 µm pore-size filter during rapid processing of water samples with materials along streams and rivers are more often

described as being amorphous. They consist of various unidentifiable fragments of organic and inorganic matter bound in an organic matrix. Sometimes the components are identifiable and may consist of diatom frustules or animal remain.

Particles passing the 0.45 μm pore-size filter are considered discrete materials and can be further classified into dissolved organic matter (DOM) and particulate organic matter (POM).

7.3.9 Color

Suspended sediments and organic matter (e.g. algae) affect the color of water in rivers or reservoirs. The green water that promotes fish production comes from billions of suspended microscopic algae. Eventually, the algae are consumed and/or simply die-off and the color of water will change. Sediments washed into rivers after heavy rains will also change the color of water, and as settling proceeds, the water color will return to normal, which may take several days. Obviously water that is too dark will inhibit photosynthesis necessary for algae to flourish.

There is no available data for color measurements at Station 100. At Station 101, the two highest color units were both recorded in 2002, at 96 Cu in June and at 153 Cu in April, while the lowest reading of 15 Cu, also in 2002, was recorded in August. As shown in Figure 46, high algae counts at this station occurred during periods of moderate color units, from 65 to 75 Cu. At Station 612 however, a different relationship was observed, where cyanobacteria bloom episodes in 1999 and 2002 corresponded with the two highest color measurements of 92 and 195 Cu, respectively. The two other algae types did not exhibit any consistent correlation with color levels. Plot of the three algae counts against the color measured at Station 612 is presented in Figure 47.

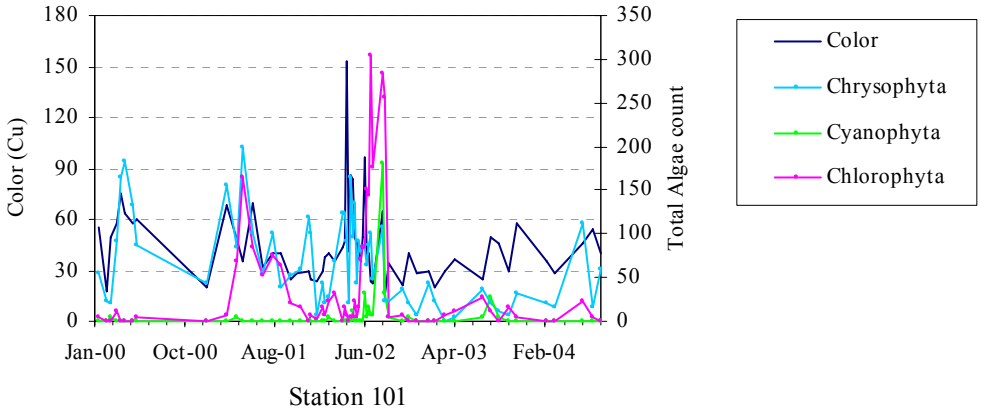


Figure 46. Total Algae counts against Color measured at Station 101

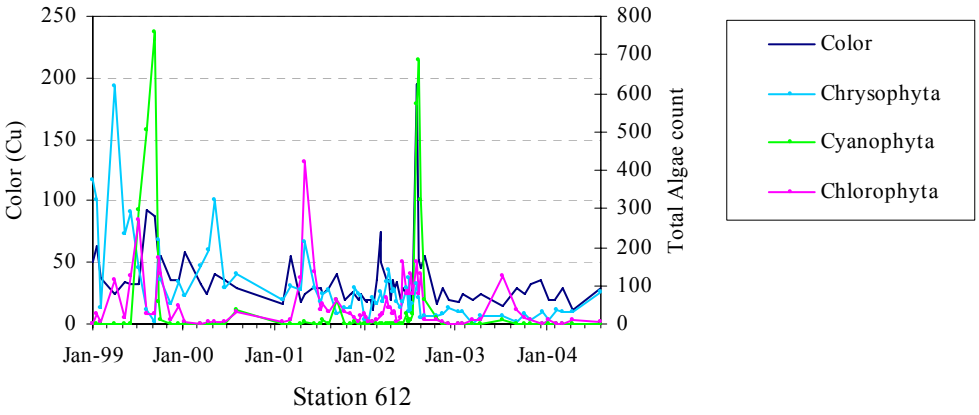


Figure 47. Total Algae counts against Color measured at Station 612

7.3.10 Odor

Odor in water is usually caused by organic compounds, inorganic salts, or dissolved gases as well as microscopic organisms such as algae. Seasonal occurrences of musty/moldy or earthy odors in rivers are due to naturally occurring algal and fungal by-products. Compounds known as Methyl-Isoborneol (MIB) and Geosmin are algal metabolites and can be released to the water during algae die off. These compounds are usually stable and difficult to remove.

7.4 Chemical Data

7.4.1 Dissolved Oxygen

Chrysophytes, chlorophytes and cyanobacteria like plants use the energy of sunlight to make their own food through the process called photosynthesis, and produce oxygen as byproduct of the process. Dissolved oxygen (DO) is consumed by algae through respiration and during decomposition as the algae are broken down by bacteria.

Figures 48, 49 and 50 depict DO measurements at the three sampling stations plotted against their corresponding algae levels. Station 100 exhibited a more distinct pattern in which generally higher DO levels, ranging from 13.8 to 16.2 mg/l, were measured from the end of the year to beginning of the following year. Most of the cyanobacteria blooms at this station, including the two particularly high count events in August and September 1999, coincided with period of lower DO levels. However, the bloom events in July and early August 1999 occurred during periods of relatively high DO levels, at 10.4 to 12.5 mg/l, respectively. Most of the higher counts of chrysophyta and chlorophyta also coincide with periods of relatively low DO levels. At Station 101 no distinct pattern between algae and DO levels emerges from the data, as shown by the figure. Higher algae levels occurred during both higher and lower DO levels. The highest DO level at this station was measured at 15.1 mg/l in February 2000, while the lowest at 5.34 and 5.52 mg/l were measured in August and June of the same year, respectively**.** At station 612, DO levels fluctuate from 4.8 to 13 mg/l during the years 1999 to 2000. However, at the beginning of 2001, DO levels were high and progressively increased towards the end of the year. A high spike (15.4 mg/l) in August 2002 coincided with a cyanobacteria bloom at this station. The other cyanobacteria bloom at Station 612, in 1999 coincide with high DO level, however high counts of chryosphyta and chlorophyta coincide with periods of relatively low DO levels.

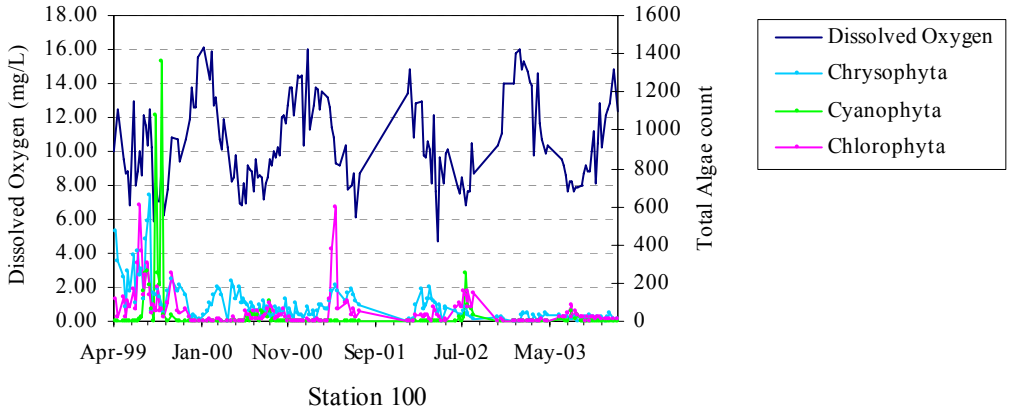


Figure 48. Total Algae counts versus Dissolved Oxygen measured at Station 100

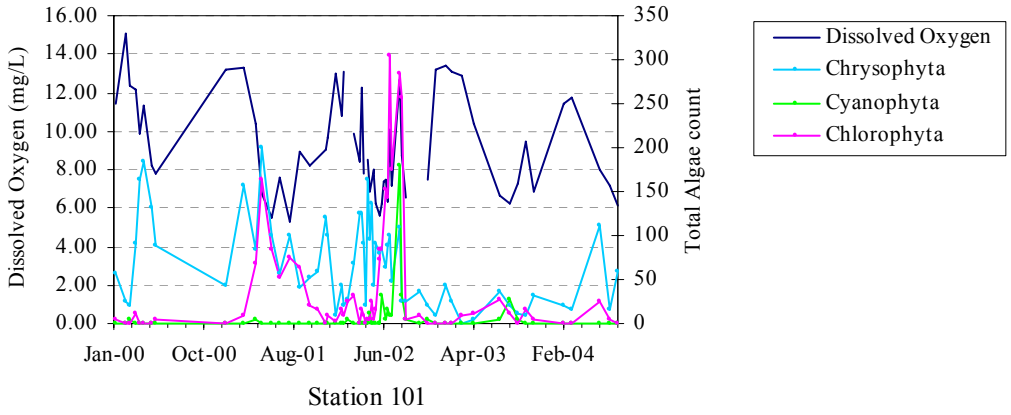


Figure 49. Total Algae counts versus Dissolved Oxygen measured at Station 101

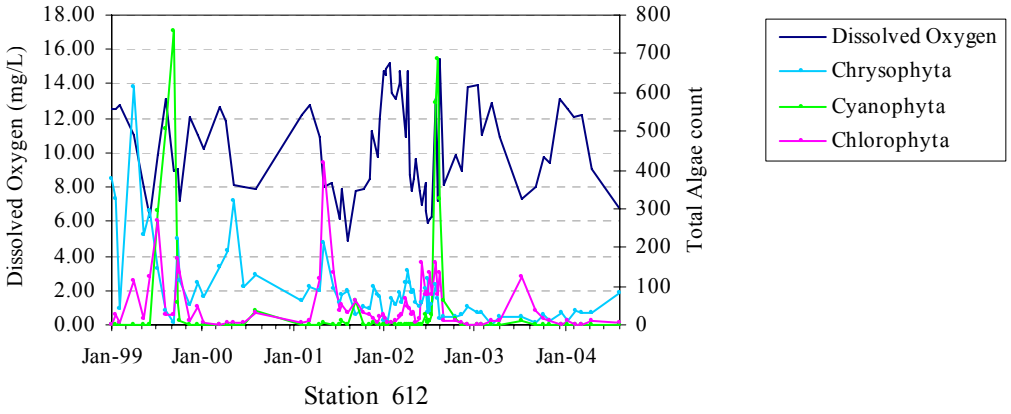


Figure 50. Total Algae counts versus Dissolved Oxygen measured at Station 612

7.4.2 Biochemical Oxygen Demand

Biochemical Oxygen Demand (BOD) is a measure of the amount of dissolved oxygen in the water required by the aerobic organisms to breakdown organic matter. It is sometimes referred to biological oxygen demand. Higher amounts of organic material in the water body results in higher oxygen consumption by the organisms for aerobic oxidation. In extreme cases, an overabundance of organic material can result in depletion of DO, which can stress or kill other aquatic life.

As depicted by Figure 51, BOD levels at Station 100 ranged from 0.70 to 10.2 mg/l, and fluctuated on a weekly basis over the 4-year record. Algae levels do not appear to correlate either positively or negatively with BOD levels. Similarly, at Station 101, no obvious relationship between BOD and algae levels can be discerned as depicted by Figure 52. The BOD levels at this station ranged from 0.40 mg/l August 2004 to 8.10 mg/l in June 2001. Incidences of high algae concentrations occurred for all different levels of BOD. At Station 612 and shown in Figure 53, BOD levels over the 4-year record showed no distinct pattern, although relationships between BOD levels and algae counts do emerge.

The maximum level was measured at 10 mg/l in August 1999, while the lowest at 1.58 mg/l was measured in June 2002. At this station, high algae levels occurred mostly during the periods of high BOD. The two cyanobacteria bloom episodes that occurred in 1999 and 2002 as well as the highest chlorophyta levels in 2001 coincide with high BOD levels. The same positive correlation can be observed for chrysophyta, where higher counts correspond with higher BOD, with exception to April 1999, when the highest measured chrysophyta event occurred during a low BOD period.

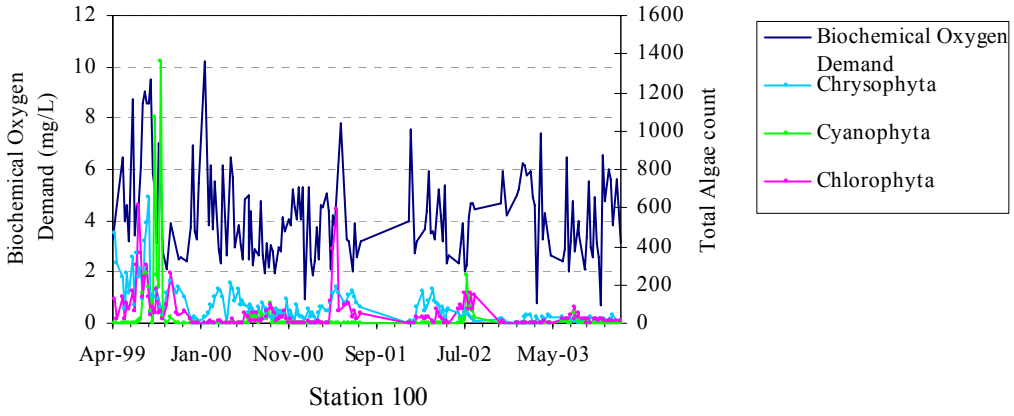


Figure 51. Total Algae counts versus Biochemical Oxygen Demand measured at Station 100

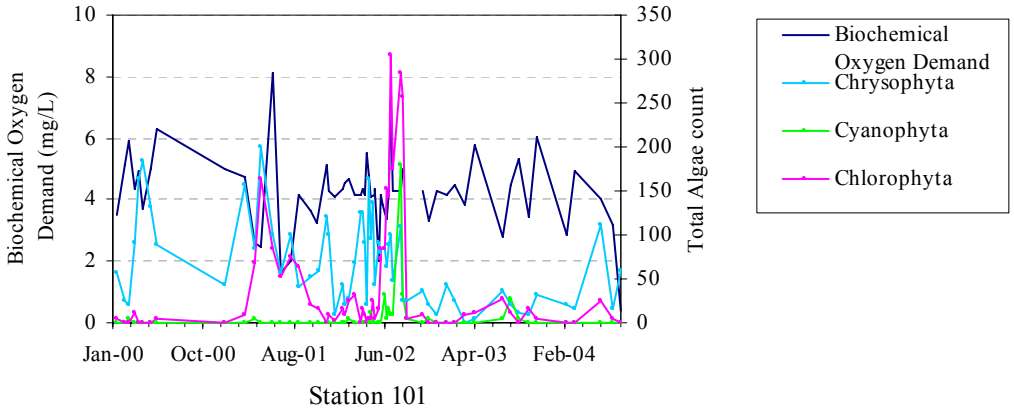


Figure 52. Total Algae counts versus Biochemical Oxygen Demand measured at Station 101

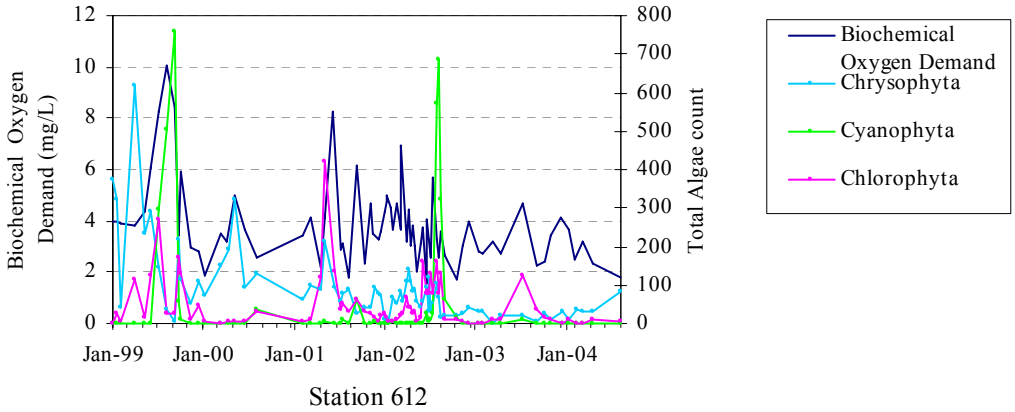


Figure 53. Total Algae counts versus Biochemical Oxygen Demand measured at Station 612

7.4.3 Nitrogen Compounds

Nitrogen, as an integral component of proteins and nucleic acids (like carbon), is required by all organisms to grow and reproduce. Dissolved nitrogen in water bodies may originate from a number of different sources. Organic forms may enter streams from surface runoff or groundwater sources, while proteins that contain organic nitrogen which are released by decomposing organic matters are broken down by bacteria to form ammonium (NH_4^+). Under very alkaline conditions (pH > 9), ammonium is converted to ammonia (NH_3). Inorganic forms of nitrogen, namely nitrate (NO_3), nitrite (NO_2), ammonia (NH_3) and nitrogen gas (N_2), are products of different nitrogen processes. Ammonium is the preferred nitrogen source for most algae.

Nitrite and nitrate are formed through the process of nitrification, in which ammonium is oxidized (combined with oxygen) by specialized bacteria and transformed into nitrate (NO_3) with nitrite (NO_2) as an intermediate product. The two steps of nitrification are performed by different bacterial species, the ammonium oxidizers and nitrate oxidizers, which convert ammonium to nitrite (NH_4^+ to NO_2) and nitrite to nitrate (NO_2 to NO_3), respectively. Nitrate is usually the most prevalent form of nitrogen in lakes that can be used by most aquatic plants and algae, as nitrite is relatively short-lived in water due to its rapid conversion to nitrate by bacteria. Nitrate and/or nitrite can be harmful to humans and wildlife at high concentration.

As depicted by Figure 54, no distinct pattern of ammonia concentrations can be observed at Station 100. High concentrations were measured four times over four different years; 0.43 mg/l in November 1999; 0.42 mg/l in December 2000; 0.41 mg/l in April 2002; and 0.33 mg/l in February 2003. The cyanobacteria bloom episodes in August and September 1999 as well as high cyanobacteria counts in July 2002, all coincide with low ammonia concentration below 0.05 mg/l. The same behavior was observed for chrysophyta and chlorophyta, that is, high algal counts occurred during low ammonia concentration periods usually below 0.05 mg/l.

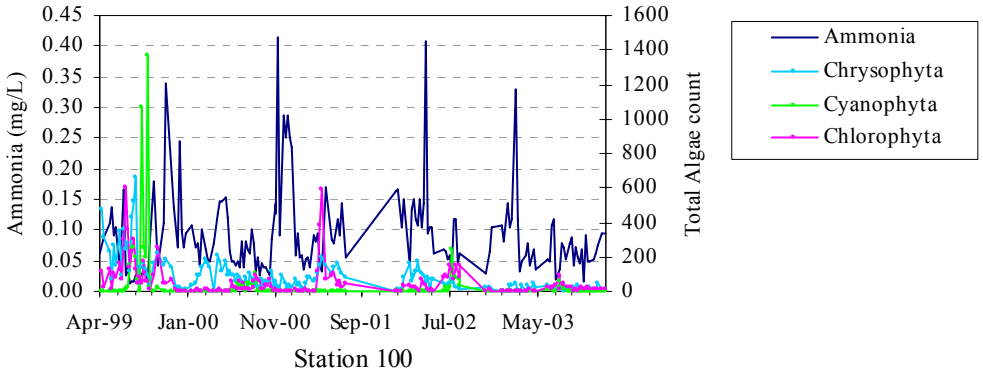


Figure 54. Total Algae counts versus Ammonia measured at Station 100

As depicted by Figure 55 below at Station 101, ammonia concentrations fluctuated over time, from a low 0.02 mg/l in July and August 2002 and November 2003 to a high of 0.43 mg/l in January 2002. The three algae classes behaved differently in response to ammonia concentration. The lone cyanobacteria bloom episode at this station occurred at the lowest ammonia concentration of 0.02 mg/l. Similarly, the highest chlorophytes counts also coincide with the lowest concentration. On the contrary several higher chlorophytes counts coincided with relatively high ammonia concentrations that ranged from 0.02 to 0.10 mg/l. Chrysophyta counts on the other hand, coincide with moderate ammonia concentrations ranging from 0.07 to more than 0.16 mg/l.

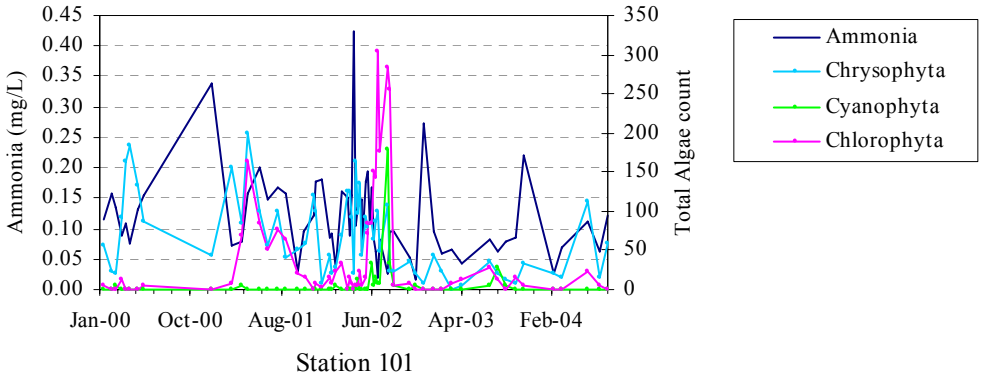


Figure 55. Total Algae counts versus Ammonia measured at Station 101

At Station 612, as with the other two stations, ammonia concentrations varied over time, ranging from a low of 0.01 mg/l in July and August 1999 to a high of 0.19 mg/l in January 2002. During 2000 and 2001, there were more dramatic weekly fluctuations as

compared to the years 1999, 2003, and 2004. Similar to the other two stations, the two cyanobacteria bloom episodes that occurred in July to September 1999 and July to August 2002 coincide with low ammonia levels less than 0.05 mg/l. As with cyanobacteria, the highest chrysophytes counts coincide with low ammonia levels below 0.05 mg/l, while the other higher counts occurred with ammonia levels above 0.05 and 0.17 mg/l. The highest chlorophytes counts occurred with ammonia concentrations at 0.05 mg/l. Comparison plot of nitrite/nitrate measured at Station 612 against the three algae counts is shown below in Figure 56.

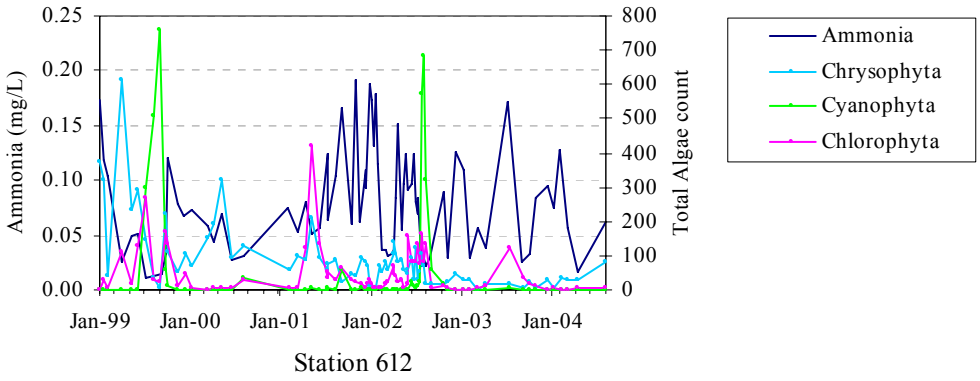


Figure 56. Total Algae counts versus Ammonia measured at Station 612

As also depicted by Figures 57 through 59, nitrite/nitrate concentrations, like ammonia, did not exhibit any consistent patterns at the three stations with respect to time or algae counts. At station 100, measured concentrations ranged from a low of 0.03 mg/l measured in November 2000 to a high of 5.34 mg/l measured in January 2002. Over the five-year record, particularly high nitrite/nitrate concentrations were observed in 1999, 2000 and 2002, during which the three algae types exhibited conflicting patterns. The highest nutrient concentrations in 1999 measured from June to September 1999 coincide with the highest counts of all three algae types. Similarly, the highest algal counts for 2000 also coincided with the highest concentrations, measured in December 2000. In contrast, all three algae classes had relatively low measured counts in January to April 2002 when nitrite/nitrate was again the highest during the year. In fact, measured cyanobacteria counts for this period were zero, chrysophyta counts were moderate and chlorophyta counts were relatively low.

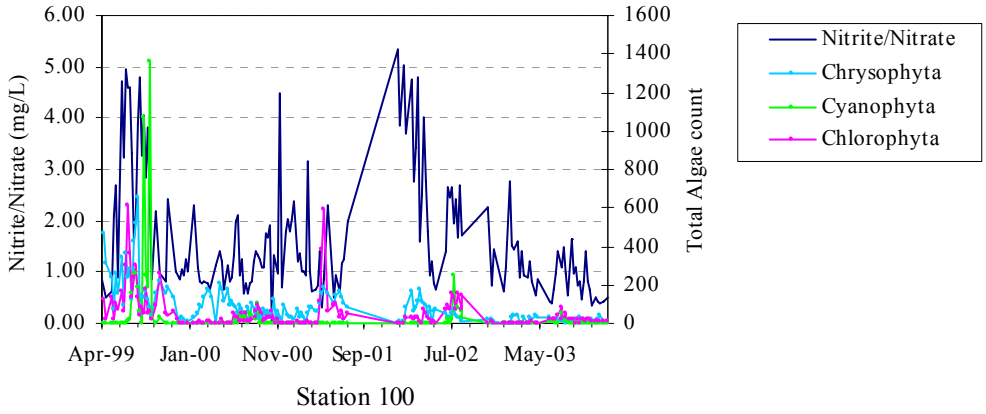


Figure 57. Total Algae counts versus Nitrite/Nitrate measured at Station 100

At Station 101, the sampling events in October 2001 and January 2002 yielded the two highest concentrations for this station with values measured at 5.94 and 6.15 mg/l, respectively. As depicted by the Figure 58, high algae levels coincide predominantly with low nitrite/nitrate concentrations. A notable exception occurred during the period July to September 2002, when the highest chlorophyta and cyanobacteria levels as well as moderate levels of chrysophyta coincide with fairly high nitrite/nitrate concentrations. Figure 59 shows that at Station 612, the highest nitrite/nitrate concentrations were measured in November 2001 at 3.70 mg/l, with moderately high concentrations also occurring in January, February and April of 2002. As can be seen in the figure, the three algae types behaved differently in response to variable nitrite/nitrate concentrations. High levels of chrysophyta occurred during lower concentration events, while high chlorophyta levels occurred during both lower and higher nitrite/nitrate concentration events. In contrast, Cyanobacteria bloom events in September 1999 and August 2002 coincide with moderate level of nitrite/nitrate concentrations.

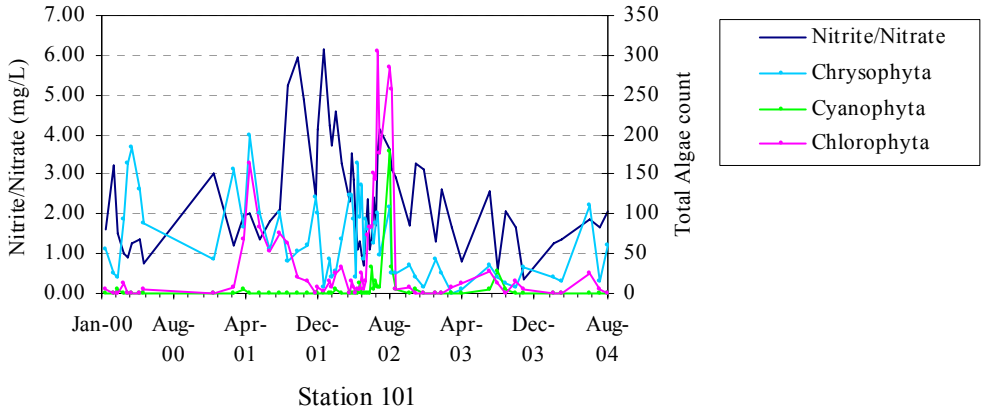


Figure 58. Total Algae counts versus Nitrite/Nitrate measured at Station 101

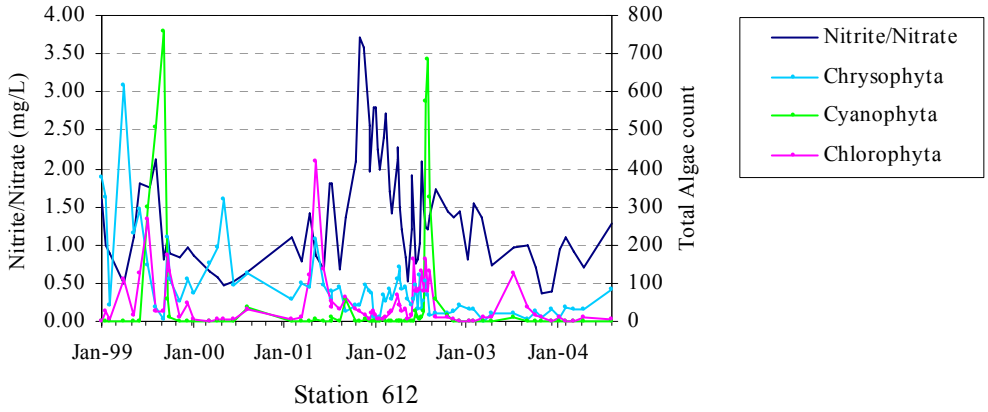


Figure 59. Total Algae counts versus Nitrite/Nitrate measured at Station 612

Because nitrogen compounds are important nutrients for algae growth, higher algae levels might be expected immediately following or during periods of nitrogen concentrations. However, the different observed relationships between the three types of algae types and the nitrite/nitrate levels at the three stations underscores the complex algae population dynamics that is certainly affected by multiple factors and conditions.

7.4.4 Total Phosphorus/Orthophosphate

Phosphorus as a structural component of nucleic acids and a key component of a cell's regulatory machinery (of signal transduction enzymes, nucleotides, factors and co-factors, etc.) is an essential nutrient that stimulates the growth of algae and aquatic plants.

It is considered as the limiting factor for algae and plants growth, as both will not grow if there is not enough phosphorus, even when plenty of other nutrients are present in water. The simplest form of phosphorus found in water is orthophosphate, which algae most readily consume.

Figure 60 shows that at station 100, no distinct pattern can be observed in the total phosphorous/orthophosphate concentrations over the study period. The measured total phosphorous/orthophosphate concentrations ranged from 0.003 to 1.71 mg/l. The three algae classes, all with the highest counts measured in 1999, behaved differently in terms of the nutrient concentrations. The two particularly high cyanobacteria counts in August and September 1999 coincide with moderately high nutrient concentrations from 0.7 to 0.82 mg/l. Similarly, the highest Chrysophyta count also coincides with moderate nutrient concentration of 0.84 mg/l. However, several higher chrysophyta counts also occurring during this year coincide with a range of concentrations for this variable, from a low 0.07 to as high as 1.27 mg/l. Chlorophyta showed somewhat more consistency; the highest count in 1999 and a fairly high count in 2001 both occurred when concentration were low, at 0.003 and 0.02 mg/l, respectively.

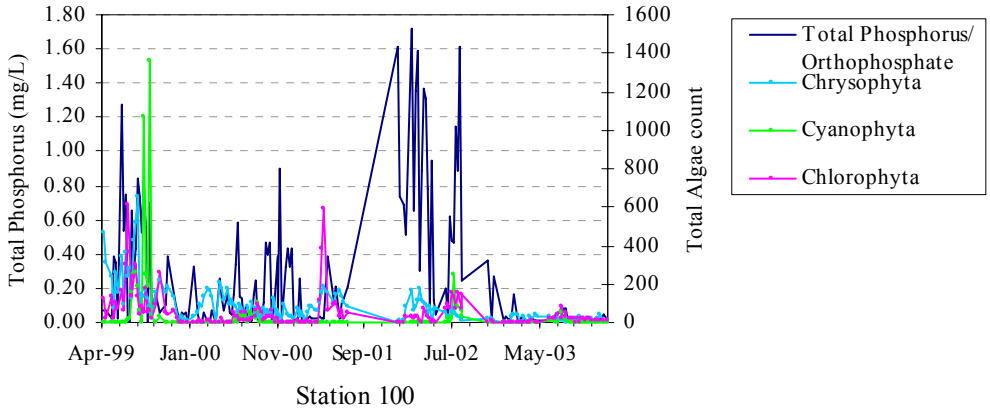


Figure 60. Total Algae counts versus Total Phosphorous/Orthophosphate measured at Station 100

As depicted by Figure 61, at station 101, a number of events of high total phosphorous/orthophosphate concentrations were measured from November 2001 (when the highest concentration was measured at 2.49 mg/l) to November 2002. Incidentally,

the highest chlorophyta levels, the cyanobacteria bloom episode and higher level of chrysophyta occurred simultaneously from July to September 2002, which coincide with the period of high total phosphorus/orthophosphate concentrations at this station. Station 612 also experienced several high total phosphorus/orthophosphate concentrations during the 4-year study period as shown by Figure 62. Among the peaks, the highest was in September 2001 at 1.58 mg/l. The two cyanobacteria bloom episodes in September 1999 and August 2002 at this station coincide with high total phosphorus/orthophosphate concentrations, while high levels of chrysophyta and chlorophyta occurred during periods of both low and high levels for this chemical.

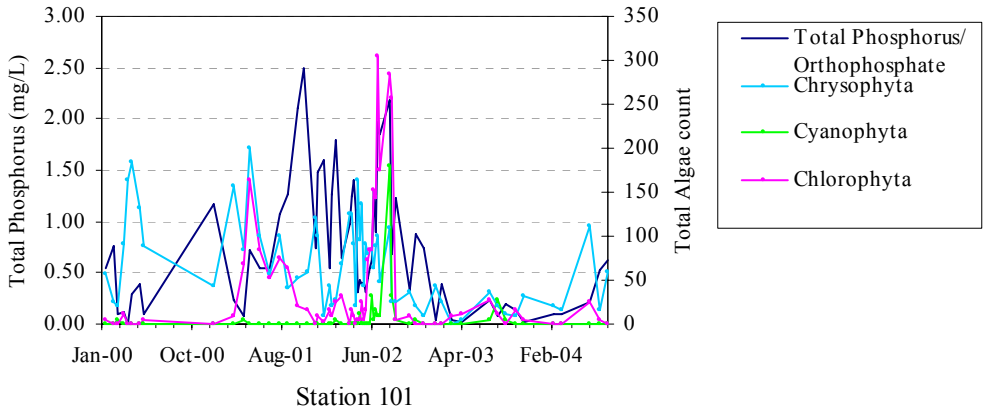


Figure 61. Total Algae counts versus Total Phosphorous/Orthophosphate measured at Station 101

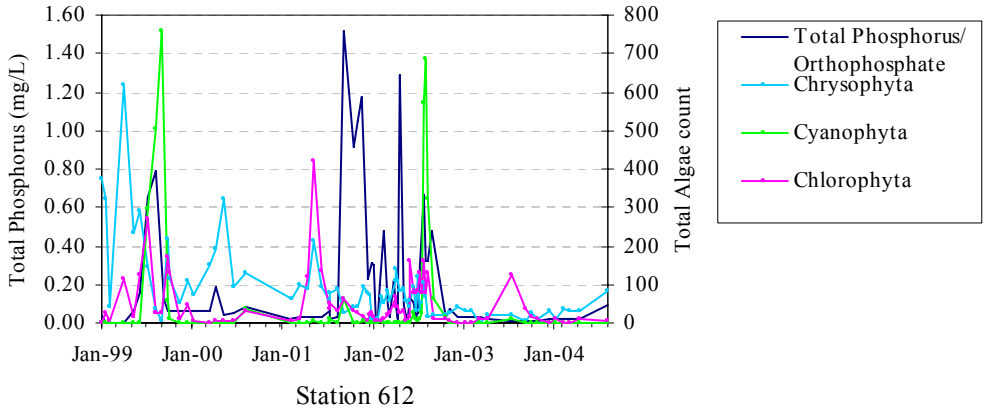


Figure 62. Total Algae counts versus Total Phosphorous/Orthophosphate measured at Station 612

7.4.5 Total Organic Carbon

Total Organic Carbon (TOC) is the total amount of organic matter found in natural water, including suspended particulate and dissolved organic carbon, which are essential components of the carbon cycle. Organic matter plays a major role in aquatic systems because it affects biogeochemical processes, nutrient cycling, biological availability, chemical transport and interactions. It also has direct implications in the planning of wastewater treatment and drinking water treatment. The TOC of a water body is affected by several factors, including vegetation, climate and treated sewage.

At station 100 as depicted by Figure 63 below, the four-year record showed no distinct pattern. There were several high TOC events, the highest of which was measured in May 2002 at 11.34 mg/l. High algae counts at this station coincide with TOC values between approximately 4 and 6 mg/l. Figure 64 below shows that at Station 101, the highest TOC level measured in August 2002 at 17.2 mg/l coincide with a cyanobacteria bloom episode and the second highest chlorophytes count. The highest chlorophytes counts as well as high chrysophytes counts occurred when TOC levels were below 6 mg/l. At station 612, there were little variations in TOC levels from January 1999 to October 2002. As depicted by Figure 65, the highest TOC was measured in August 2004 at 8.48 mg/l. Measured high counts of all three algae types occurred during periods when TOC was between 3 and 5 mg/l.

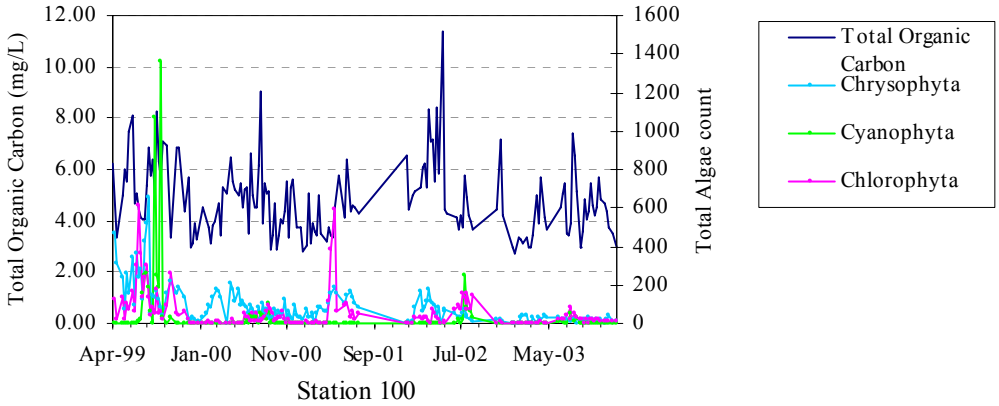


Figure 63. Total Algae counts versus Total Organic Carbon measured at Station 100

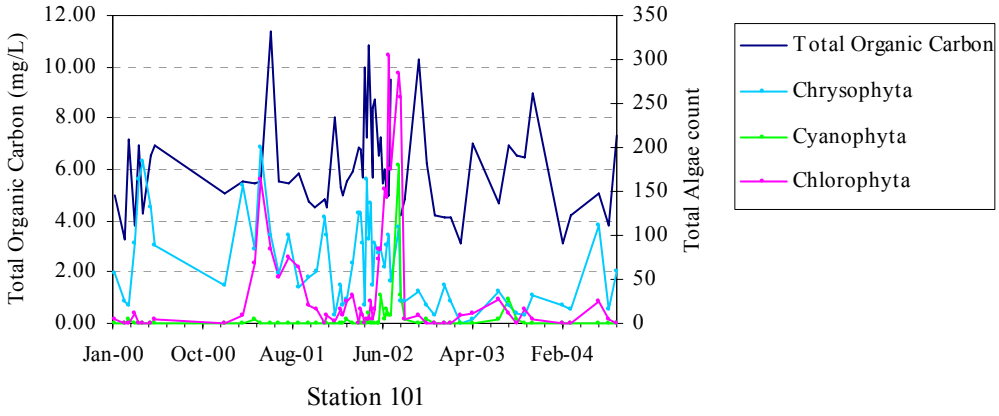


Figure 64. Total Algae counts versus Total Organic Carbon measured at Station 101

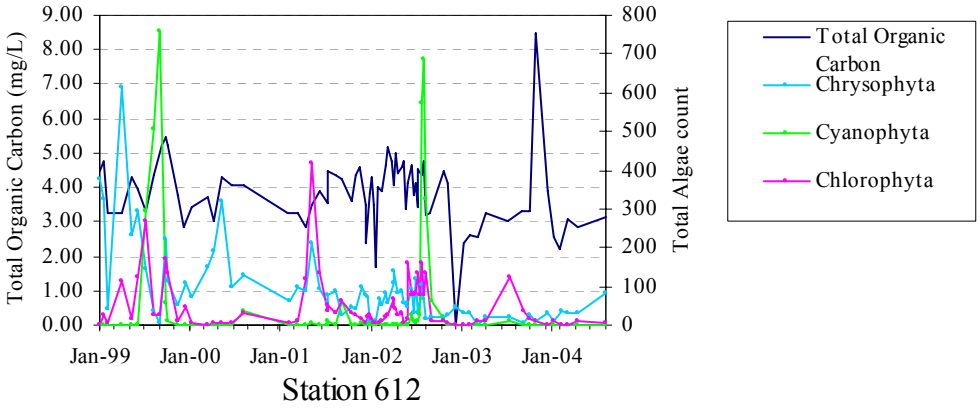


Figure 65. Total Algae counts versus Total Organic Carbon measured at Station 612

7.4.6 Chloride

Chlorides are binary compounds of chlorine chemically combined with a metal. The presence of chloride in water bodies where it does not occur naturally indicates possible pollution, along with excessive nutrients or bacteria. Sources of chloride include septic systems, wastewater treatment plant effluent, animal waste, potash fertilizer, and drainage from rock salt applied to roads, parking lots and sidewalks to lower the melting point of ice The normal range of chloride in fresh surface water systems is 45-155 mg/L, while USEPA has set a public drinking water standard of 250 mg/L.

Elevated chlorides can kill plants and wildlife, and may affect other organisms present in water including algae**.** Small amounts of chlorides are necessary for normal cell functions in plant and animal life. However, fish and aquatic animals can not live in high levels of chlorides.

Chloride concentrations over time at stations 100, 101 and 612 are shown in Figures 66, 67 and 68, respectively. At Station 100, a temporal pattern of chloride concentrations for the year 2000 to 2003 is evident, with peak concentrations occurring in late winter/early spring (January to March), after which concentrations decline and remain low through summer months. The highest chloride level at 283 mg/l was measured in January 2002. A notable exception to this otherwise consistent temporal pattern occurred in 1999, when the highest measured concentration of 164 mg/l occurred in July. High algae levels occurred mostly during moderate chloride concentrations. At Station 101, four high chloride events were measured from 2000 to 2002 with values ranging from 172 (February 2002) to 192 mg/l (February 2000). In 2000, another high chloride concentration occurred in August to September which coincides with the highest chlorophyta level and algal bloom incidence. Chrysophyta level on the other hand was moderate during this period. At Station 612, peak concentrations were observed in January to March with values ranging from 127 mg/l in March 2003 to 174 mg/l in January 2002. At this station however, most of high algae levels coincide with periods of moderate chloride concentrations.

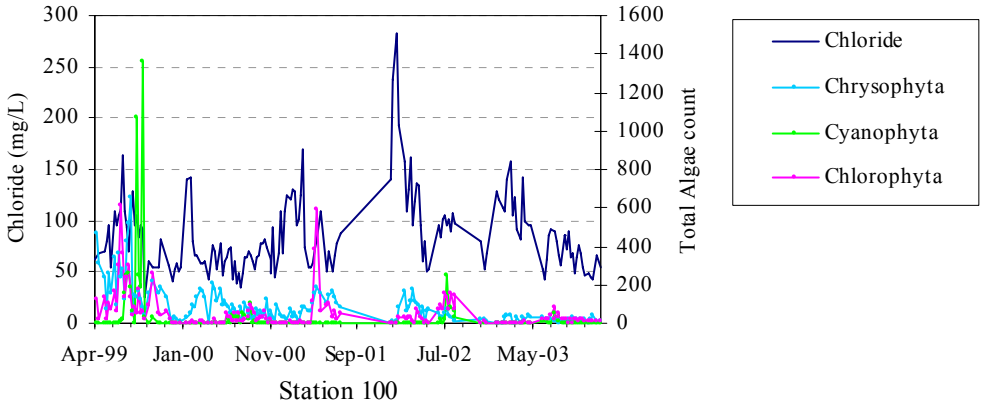


Figure 66. Total Algae counts versus Chloride measured at Station 100

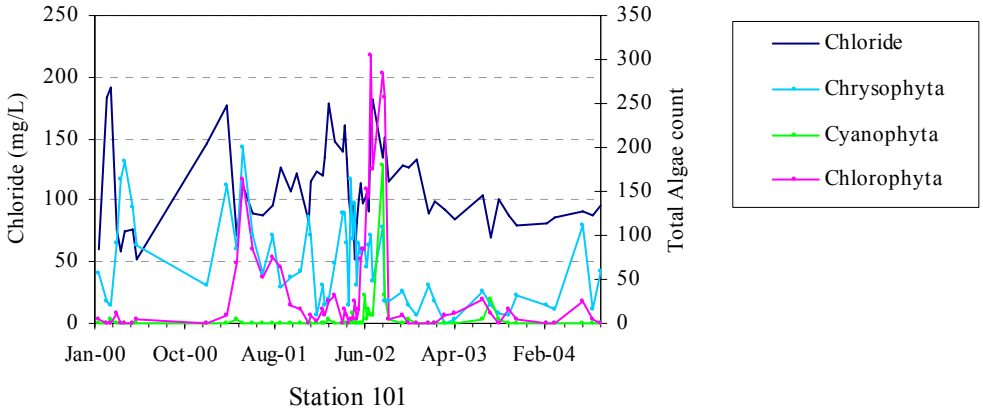


Figure 67. Total Algae counts versus Chloride measured at Station 101

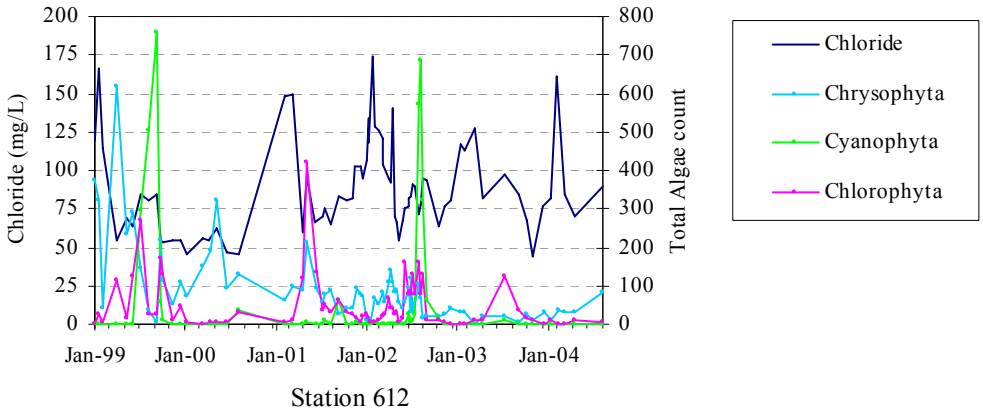


Figure 68. Total Algae counts versus Chloride measured at Station 612

7.5 Climate Data

7.5.1 Precipitation

In most algal bloom studies, precipitation is generally not considered one of the major controlling factors. However, as blooms occur in response to a variety of natural conditions, it is important that the potential affect of precipitation on these events not be excluded. During relatively high precipitation events, when run-off is generated, large amounts of nutrients are delivered to surface water bodies. This nutrient loading promotes the growth of aquatic plants and algae. Another impact due to precipitation, when it is minimal or absent, is a decline in surface water levels with an associated

decrease in flow rate. Low flow conditions increase the retention time of the water along any stretch of river, and thus increase the time available for algae to grow. Low flow conditions also promote water column stability, which enhances the ability of the algae to optimally locate themselves within the water column for photosynthesis.

Three weather stations located in proximity to the study area and operated by NOAA were used for the ANN development. These stations include Caldwell, located approximately 6 miles to the southwest, Newark, located 11.4 miles to the south, and Tetterboro, located 9 miles to the southeast. Data collected during the study period from 1999 to 2004, showed that distributions of average monthly precipitation totals from all three weather stations were similar. That is, the lowest precipitation of about 2 inches occurred in January and February, the highest precipitation of more than 5 inches occurred in June and September while the rest of the months, all stations received an average monthly precipitation totals of around 3 to 4 inches. For this study, the averages of daily precipitation totals from the three stations were used.

7.5.2 Wind Speed/Direction

Wind direction and speed have influence on the movement and/or mixing of water in the river or lake. As the wind become stronger, it creates waves and underwater currents below the waves. The underwater currents tend to move water particles horizontally and in irregular swirling motion known as turbulence. Nutrients are distributed vertically by turbulence, hence facilitating the recycling of nutrients from sediments and deeper water, which in turn will be available for algae suspended in water. At the other extreme, resuspension of sediments due to turbulence also increase the turbidity of the water and reduce light penetration which can affect the biochemical activity of algae such as photosynthesis.

Available wind speed and direction data used for developing the ANNs were taken from the three NOAA weather stations located in Caldwell, Newark and Tetterboro, all in proximity to the study area.

Computed average wind speeds over the week at the three stations were 7.4 miles per hour (mph) at station 100, 7.3 mph at station 101 and 7.6 mph at station 612. Based upon the data used in the modeling efforts, it was observed that most of high algae levels (all three) occurred during periods when wind speeds were around the average values between 6 to 8 mph. With exception to station 612 in April 1999, when the corresponding wind speed to the highest chrysophytes count was computed to 9.6 mph.

7.5.3 Heating degree days/Sky Cover/Length of Day

For any individual day, heating degree days indicate how far that day's average temperature departed from 65 degrees Fahrenheit, measuring heating energy demand. Sky cover indicates the total cloud cover throughout the entire day, 0.0 indicates a nearly clear sky, where high thin clouds may not be included in the total sky cover for some locations.

As in precipitation and wind direction/speed, data for heating degree days and sky cover used in this study were taken from the three weather stations maintained and operated by NOAA. Again, the stations used were Caldwell, Newark and Tetterboro.

Length of day is the time difference between sunrise and sunset. Sunrise and sunset are defined as instants in the morning/evening under ideal meteorological conditions, with a standard refraction of the sun's rays, when the upper edge of the sun's disk is coincident with an ideal horizon. Length of day has been shown to be an important variable for many higher plants that have “critical photoperiods”, during which growth regulators respond to length of day. The data used in this study was downloaded from a public domain provided as a public service by Horizon Network Security.

8. MODELING APPROACH AND RESULTS

Because the three water quality sampling stations each have distinct water quality and hydrologic conditions (e.g. unique watershed, etc.), each was modeled independently of the others. The validity of this approach was verified by preliminary ANN modeling results presented to DSR and PVWC during an October, 2004 meeting, showing superior modeling performance when each station was modeled individually rather than collectively. This finding was further verified with a comprehensive modeling effort, with the results for modeling all the stations collectively for each prediction period and algae type summarized in Appendix B-4. As previously discussed, many different ANN models were developed and tested for the PVWC, which evolved to address important modeling issues that emerged during this project, following submission of the first draft report to DEP for review.

After organizing the corresponding measurements for each variable by date, the values for the input and output variables were assigned or computed for each prediction event. For many variables, the input value is a single measurement, such as water temperature measured at a particular station on a particular date. For other variables, such as the prediction period's cumulative precipitation, daily values were summed over the period of interest. For other variables, such as wind direction and water temperature, the average daily values were used.

ANNs that explicitly predicted final measured count values constituted the great majority of forecasting models in this study, but eight RBF classification nets were also developed and assessed. Model forecast performances were assessed based upon a number of criteria, including standard statistical measures, such as absolute mean errors and correlation coefficients, statistical discrepancies between training and validation results, the ability of the models to forecast bloom and non-bloom events, the ability of the models to accurately predict relative changes in population counts, and visual comparison of time series. Again, the relatively few number of historical validation events limits the

strength with which conclusions can be drawn, particularly since there are relatively few events where populations change from bloom to non-bloom events, and vice versa. As a final benchmark of performance, LMs were also developed for comparison against ANN performance.

Sensitivity analyses was performed, using both the ratio and ranking method described in the water treatment modeling section, as well as the exclusion of select water quality and water extraction variables. The hope was that certain fundamental trends or tendencies would emerge that might help identify important predictor variables for improving future data collection and modeling strategies.

Most of the modeling results are presented in Appendices B-1 through B-6, with representative results for three test cases presented here. A detailed discussion of the original and revised modeling approaches is provided, followed by the systematic comparison of different modeling approaches that evolved out of the original modeling effort to address important data and modeling issues.

8.1 Original and Revised ANN Modeling Approaches

For the original modeling work, 27 different cases were modeled; the three stations, for the three different algal groups, for three prediction periods consisting of real-time, one-week ahead, and two-weeks ahead. As discussed previously, and presented in the appendices, the original set of ANN models developed for PVWC, which utilized all of the variables listed in Table 19, did not consistently perform well during validation, particularly for cases with smaller data sets. This finding was not unexpected, given the low number of data events, the relatively high number of input variables, and the complexity of algal population dynamics. Another potentially complicating factor considered was the original model inputs reflect conditions existing at or preceding the beginning of the prediction period, which may not be strongly causal and/or correlative factors of algal counts one-week or more into the future. That is, system changes (e.g.

precipitation event(s), hot spell, etc.) over the prediction period of interest, which is not captured or represented by the model inputs, could largely determine the algal populations as measured at the end of the prediction period. Lastly, the original models did not include length of day as an input variable, which ranked fairly high in importance for many of the models.

To further assess the possible disconnection between real-time system conditions and algal counts at some future date, the revised models were developed, with the measurement date of the input values generally corresponding to the prediction date. In a sense, then, these models are developed in a more correlative manner, though, as discussed in more detail later, for real-time operation, the modeler would have to forecast or predict final system conditions. Tables 20 and 21 below specify how the value for each model input type was computed/assigned for the original and revised models, respectively, using sample calendar dates. Again, the most recently developed original and revised models used forecasting periods of one-week and two-weeks ahead, but some three week models were developed in previous work, contained in the Appendices.

Table 20. Input Value Assignments/Computations for Three Prediction Periods for Original ANN Models

Model Variable Type	July 14 (one-week ahead)	July 21 (two-weeks ahead)	July 28 (three-weeks ahead)
Physical Water Parameters	Single Value Measured July 7	Single Value Measured July 7	Single Value Measured July 7
Water Quality Parameters	Single Value Measured July 7	Single Value Measured July 7	Single Value Measured July 7
Streamflows/ River Extractions	Mean Value July 1 – July 7	Mean Value July 1 – July 7	Mean Value July 1 – July 7
Prediction Period's Precipitation Two Week Lag	Total Cumulative June 24– June 30	Total Cumulative June 24 – June 30	Total Cumulative June 24 – June 30
Prediction Period's Precipitation One Week Lag	Total Cumulative July 1 – July 7	Total Cumulative July 1 – July 7	Total Cumulative July 1 – July 7
Other Weather Conditions	Mean Value July 1 – July 7	Mean Value July 1 – July 7	Mean Value July 1 – July 7
Initial Algae Count	July 7	July 7	July 7

Table 21. Input Value Assignments/Computations for Three Prediction Periods for Revised ANN Models

Model Variable Type	July 14 (one-week ahead)	July 21 (two-weeks ahead)	July 28 (three-weeks ahead)
Physical Water Parameters	Single Value Measured July 14	Single Value Measured July 21	Single Value Measured July 28
Water Quality Parameters	Single Value Measured July 14	Single Value Measured July 21	Single Value Measured July 28
Streamflows/ River Extractions	Mean Value July 7 – July 14	Mean Value July 7 – July 21	Mean Value July 7 – July 28
Prediction Period's Lagged Precipitation	Total Cumulative July 1 – July 7	Total Cumulative June 24 – July 7	Total Cumulative June 16 – July 7
Prediction Period's Precipitation	Total Cumulative July 7 – July 14	Total Cumulative July 7 – July 21	Total Cumulative July 7 – July 28
Other Weather Conditions	Mean Value July 7 – July 14	Mean Value July 7 – July 21	Mean Value July 7 – July 28
Initial Algae Count	July 7	July 7	July 7

As depicted by Tables 20 and 21, the major functional difference between the original and revised modeling approaches is the temporal correspondence of the input values with the final predicted algal counts. For the original models, the input values for future prediction periods are generally measured at the beginning of the prediction period, and hence, under real-time conditions, would be explicitly known a-priori. Thus, there is no uncertainty with regard to the values of the input variables. In contrast, for the revised models, input values generally correspond to the end of the prediction period, and hence, under real-time conditions, would have to be assumed or forecasted.

Another important distinction is that, unlike for the revised modeling approach, the original ANN models, for predicting algal counts n days into the future, use exactly the same input values, regardless of the forecast period (e.g. one-week ahead, two-weeks ahead, etc.). This may seem counterintuitive to use identical input values for different prediction periods, but a different model was developed for each different prediction period. It may be, as is discussed in more detail later, that there is a natural and therefore repeatable temporal transition of algal populations from initial state conditions, and thus

the one-week ahead ANN model would "learn" a different rate of algal population changes than the two-week ahead model.

The obvious limitation of the revised modeling approach is the uncertainty regarding model input values, each of which constitutes a prediction problem in itself. However, a number of these variables, such as water temperature, D.O., and conductivity, are related to seasons, and consequently, input values can be estimated or assumed with reasonable accuracy based upon measured historical conditions. Weather forecasts can be obtained from NOAA, and hydrologic and water use conditions, such as river flow and water extractions, can also be estimated or forecasted. Thus, using historical records and weather forecasts, uncertainty with regard to input values can be reduced.

Another factor that further reduces uncertainty is that some variables will generally not exhibit large changes over the relatively short prediction periods. Water, because of its high specific heat capacity, is characterized by gradual temperature changes. Exceptions may occur during freezing or thawing periods, but the modeler can combined historical records, weather forecast information, with physical intuition to estimate expected changes over prediction periods of interest. For example, if mean daily temperatures are expected to remain fairly constant over the next prediction period, relatively small water temperature changes may be expected. However, if a hot or cold spell is forecasted, the modeler may assume more dramatic change in water temperature.

To investigate at a very basic level the feasibility of this approach, the range and mean changes of representative input variables over one-week, two-weeks, and even three-weeks ahead forecast periods are presented below in Tables 22, 23, and 24 for Stations 100, 101, and 612, respectively. Both the actual and absolute value changes with respect to minimum, mean, maximum, and standard deviations are presented.

As the tables show, on average, relatively small absolute value changes in the values of the predictor variables occur over the three different periods. In some cases, even the

maximum observed changes for some variables is relatively small. What would be important in future modeling work is to more systematically quantify how algae counts predicted by the ANN models change in response to expected changes in input values over the prediction period of interest.

Table 22. Statistical Summary of Input Value Changes for Representative Variables for Three Prediction Periods at Station 100

Parameter	Statistics	One-week Interval Changes		Two-week Interval Changes		Three-week Interval Changes	
		Actual	Absolute	Actual	Absolute	Actual	Absolute
Temperature	Std. Dev.	2.80	1.85	3.29	2.19	3.86	2.52
	Min	-7.30	0.00	-9.10	0.00	-10.30	0.00
	Ave	0.08	2.10	0.28	2.46	0.43	2.95
	Max	9.40	9.40	11.00	11.00	11.80	11.80
pH	Std. Dev.	0.60	0.44	0.55	0.40	0.55	0.36
	Min	-2.60	0.00	-2.90	0.00	-2.30	0.00
	Ave	0.02	0.41	0.01	0.38	0.00	0.41
	Max	3.20	3.20	2.20	2.90	1.70	2.30
Turbidity	Std. Dev.	6.40	5.41	6.08	5.01	6.19	5.08
	Min	-34.90	0.00	-33.50	0.00	-33.50	0.00
	Ave	-0.15	3.42	-0.23	3.44	-0.25	3.55
	Max	30.33	34.90	37.86	37.86	38.44	38.44
Alkalinity	Std. Dev.	23.60	18.01	25.29	18.13	27.07	18.90
	Min	-86.08	0.00	-86.68	0.00	-89.00	0.00
	Ave	0.83	15.24	0.65	17.61	0.30	19.34
	Max	89.67	89.67	87.23	87.23	91.10	91.10
Hardness	Std. Dev.	25.14	16.94	29.00	17.97	32.15	19.71
	Min	-88.00	0.00	-109.00	0.00	-110.00	0.00
	Ave	-1.05	18.58	0.63	22.73	1.93	25.42
	Max	110.00	110.00	62.00	109.00	106.00	110.00
Conductivity	Std. Dev.	113.31	76.64	139.35	93.68	146.62	95.06
	Min	-558.00	0.50	-468.00	0.00	-472.50	3.50
	Ave	-7.87	83.67	1.63	102.97	2.17	111.40
	Max	324.50	558.00	675.50	675.50	436.50	472.50
Dissolved Oxygen	Std. Dev.	1.63	1.12	1.80	1.24	1.89	1.21
	Min	-6.56	0.00	-7.59	0.00	-7.32	0.02
	Ave	-0.01	1.18	-0.06	1.30	-0.12	1.45
	Max	7.57	7.57	6.62	7.59	4.99	7.32
Biochemical Oxygen Demand	Std. Dev.	2.19	1.32	2.20	1.41	2.45	1.52
	Min	-5.27	0.02	-5.83	0.00	-8.07	0.00
	Ave	-0.02	1.74	-0.10	1.68	-0.16	1.93
	Max	7.24	7.24	6.27	6.27	7.54	8.07
Chloride	Std. Dev.	23.24	16.77	26.74	18.17	31.77	21.92
	Min	-91.00	0.00	-111.00	0.00	-120.00	0.00
	Ave	0.21	16.06	0.65	19.59	1.39	22.98
	Max	128.00	128.00	86.00	111.00	116.00	120.00
Ammonia	Std. Dev.	0.06	0.04	0.06	0.05	0.07	0.05
	Min	-0.26	0.00	-0.28	0.00	-0.26	0.00
	Ave	0.00	0.04	0.00	0.04	0.00	0.05
	Max	0.36	0.36	0.31	0.31	0.32	0.32
Total Suspended Solids	Std. Dev.	24.85	23.36	18.66	17.12	20.53	18.90
	Min	-264.00	0.00	-41.00	0.00	-52.63	0.00
	Ave	-0.33	8.44	0.46	7.42	0.68	8.03
	Max	263.00	264.00	253.00	253.00	263.00	263.00
UV254	Std. Dev.	0.04	0.03	0.04	0.03	0.05	0.03
	Min	-0.11	0.00	-0.19	0.00	-0.16	0.00
	Ave	0.00	0.03	0.00	0.03	0.00	0.03
	Max	0.13	0.13	0.17	0.19	0.16	0.16
Total Organic Carbon	Std. Dev.	1.32	0.84	1.34	0.85	1.44	0.94
	Min	-4.15	0.00	-4.07	0.01	-4.35	0.02
	Ave	-0.05	1.02	-0.06	1.03	-0.05	1.10
	Max	3.90	4.15	5.35	5.35	3.99	4.35

Table 23. Statistical Summary of Input Value Changes for Representative Variables for Three Prediction Periods at Station 101

Parameter	Statistics	One-week Interval Changes		Two-week Interval Changes		Three-week Interval Changes	
		Actual	Absolute	Actual	Absolute	Actual	Absolute
Temperature	Std. Dev.	3.26	2.13	3.92	2.55	4.53	2.98
	Min	-8.24	0.00	-7.90	0.10	-7.54	0.20
	Ave	0.55	2.52	1.10	3.16	1.15	3.58
	Max	9.94	9.94	10.20	10.20	13.09	13.09
pH	Std. Dev.	0.56	0.40	0.58	0.40	0.49	0.32
	Min	-2.20	0.00	-1.70	0.00	-1.10	0.00
	Ave	0.02	0.39	0.00	0.41	0.05	0.37
	Max	1.70	2.20	1.60	1.70	1.30	1.30
Turbidity	Std. Dev.	5.50	4.31	6.01	4.59	5.87	4.55
	Min	-18.90	0.00	-18.20	0.00	-17.10	0.00
	Ave	-0.20	3.41	-0.33	3.86	-0.49	3.72
	Max	17.90	18.90	19.50	19.50	17.00	17.10
Alkalinity	Std. Dev.	14.21	8.30	17.41	11.74	22.91	15.83
	Min	-30.00	0.00	-52.00	0.00	-84.60	0.00
	Ave	2.28	11.70	6.78	14.45	5.48	17.33
	Max	39.60	39.60	53.60	53.60	62.60	84.60
Hardness	Std. Dev.	39.01	25.41	41.20	25.59	41.92	25.51
	Min	-120.00	0.00	-98.00	0.00	-96.00	0.00
	Ave	3.47	29.65	5.80	32.56	7.94	33.97
	Max	130.00	130.00	112.00	112.00	100.00	100.00
Color	Std. Dev.	28.36	19.80	31.41	20.81	30.19	20.97
	Min	-99.00	0.00	-84.00	0.00	-84.00	0.00
	Ave	0.12	20.20	-0.85	23.37	-4.32	21.99
	Max	106.00	106.00	97.11	97.11	79.00	84.00
Conductivity	Std. Dev.	121.14	79.67	136.98	71.57	169.29	105.53
	Min	-542.00	3.00	-271.00	3.00	-486.00	6.00
	Ave	9.14	91.24	35.31	121.17	46.79	139.52
	Max	219.00	542.00	265.00	271.00	506.00	506.00
Dissolved Oxygen	Std. Dev.	2.41	1.70	2.63	1.91	2.22	1.47
	Min	-7.38	0.01	-10.66	0.00	-5.08	0.03
	Ave	0.16	1.70	0.04	1.79	-0.12	1.66
	Max	6.73	7.38	8.60	10.66	8.44	8.44
Chloride	Std. Dev.	23.66	15.21	25.46	13.46	34.90	23.33
	Min	-92.00	0.00	-60.00	0.00	-88.00	0.00
	Ave	1.82	18.12	5.98	22.26	9.93	27.62
	Max	54.00	92.00	54.00	60.00	131.00	131.00
Ammonia	Std. Dev.	0.08	0.05	0.07	0.04	0.08	0.06
	Min	-0.15	0.00	-0.15	0.00	-0.27	0.00
	Ave	0.00	0.06	0.00	0.05	-0.01	0.06
	Max	0.32	0.32	0.13	0.15	0.25	0.27
UV254	Std. Dev.	0.06	0.04	0.08	0.06	0.08	0.06
	Min	-0.19	0.00	-0.25	0.00	-0.21	0.00
	Ave	0.00	0.04	0.00	0.05	0.00	0.06
	Max	0.15	0.19	0.28	0.28	0.29	0.29

Table 24. Statistical Summary of Input Value Changes for Representative Variables for Three Prediction Periods at Station 612

Parameter	Statistics	One-week Interval Changes		Two-week Interval Changes		Three-week Interval Changes	
		Actual	Absolute	Actual	Absolute	Actual	Absolute
Temperature	Std. Dev.	2.94	1.88	3.68	2.37	3.68	2.37
	Min	-8.20	0.00	-10.00	0.00	-10.00	0.00
	Ave	0.26	2.26	0.72	2.89	0.72	2.89
	Max	9.30	9.30	12.50	12.50	12.50	12.50
pH	Std. Dev.	0.36	0.26	0.39	0.29	0.39	0.29
	Min	-1.50	0.00	-1.50	0.00	-1.50	0.00
	Ave	-0.02	0.25	-0.03	0.27	-0.03	0.27
	Max	1.60	1.60	1.40	1.50	1.40	1.50
Turbidity	Std. Dev.	5.65	4.93	3.48	2.65	3.48	2.65
	Min	-14.00	0.00	-15.00	0.00	-15.00	0.00
	Ave	0.18	2.76	-0.35	2.26	-0.35	2.26
	Max	56.00	56.00	7.40	15.00	7.40	15.00
Alkalinity	Std. Dev.	11.18	7.68	12.12	7.20	12.12	7.20
	Min	-45.00	0.00	-36.00	0.00	-36.00	0.00
	Ave	0.37	8.11	1.26	9.79	1.26	9.79
	Max	40.00	45.00	28.00	36.00	28.00	36.00
Hardness	Std. Dev.	19.93	14.47	22.32	14.14	22.32	14.14
	Min	-80.00	0.00	-68.00	0.00	-68.00	0.00
	Ave	-0.82	13.68	0.84	17.23	0.84	17.23
	Max	84.00	84.00	80.00	80.00	80.00	80.00
Color	Std. Dev.	11.88	8.50	11.58	8.36	11.58	8.36
	Min	-49.00	0.00	-60.00	0.00	-60.00	0.00
	Ave	0.11	8.28	-0.87	8.03	-0.87	8.03
	Max	40.00	49.00	30.00	60.00	30.00	60.00
Conductivity	Std. Dev.	82.56	55.14	105.23	67.10	105.23	67.10
	Min	-328.00	0.00	-330.00	1.00	-330.00	1.00
	Ave	-3.35	61.36	-3.20	80.83	-3.20	80.83
	Max	249.00	328.00	310.00	330.00	310.00	330.00
Dissolved Oxygen	Std. Dev.	1.61	1.08	1.81	1.22	1.81	1.22
	Min	-4.78	0.00	-5.62	0.00	-5.62	0.00
	Ave	-0.09	1.20	-0.35	1.38	-0.35	1.38
	Max	7.16	7.16	4.43	5.62	4.43	5.62
Chloride	Std. Dev.	18.43	13.54	21.34	15.23	21.34	15.23
	Min	-58.00	0.00	-69.00	0.00	-69.00	0.00
	Ave	-0.45	12.47	-0.98	14.92	-0.98	14.92
	Max	79.00	79.00	88.00	88.00	88.00	88.00
Ammonia	Std. Dev.	0.09	0.08	0.10	0.09	0.10	0.09
	Min	-0.49	0.00	-0.47	0.00	-0.47	0.00
	Ave	0.00	0.05	0.01	0.05	0.01	0.05
	Max	0.81	0.81	0.70	0.70	0.70	0.70
UV254	Std. Dev.	0.60	0.59	0.62	0.61	0.62	0.61
	Min	-4.74	0.00	-4.76	0.00	-4.76	0.00
	Ave	0.00	0.10	-0.02	0.11	-0.02	0.11
	Max	4.79	4.79	4.76	4.76	4.76	4.76

8.2 Representative Modeling Results

Because of the large number of models previously developed, as well as the high possible number of prediction cases, for illustration purposes, six select cases are presented in this document for the single value algal count forecast models. The cases include each of the three different algal classes and stations, with each presenting different characteristics in terms of system conditions, algal populations, prediction periods, and the number of historical events. The general results are consistent with other modeling work performed, some of which is presented in the appendices. These case studies are considered sufficient for achieving the important objective of demonstrating the feasibility of using ANNs to accurately forecast algal blooms, and important modeling and data acquisition and processing issues are fully addressed. This is achieved with the comparisons between the original and revised modeling approaches, inclusion versus exclusion of select water quality inputs, inclusion versus exclusion of water volume extraction variables, and smaller versus larger number of data events. For the classification models, eight select test cases were used, which is discussed in detail in the corresponding section.

For the cyanobacteria forecasting example, Station 612, located on River A at the intake point for the Pumping Station 2, was selected. Not only is this the largest source of water for the utility for most time periods, but it also has the highest frequency of algal bloom events for this algae class. A two-week ahead prediction period was used for this location. For the chlorophytes class, Station 100 was selected, which is the location of the water intake point for the treatment facility. Based upon the available data, station 100 has the highest frequency of chlorophytes bloom incidences among the three sampling stations, and a one-week ahead prediction period was selected for this location. Lastly, for chrysophytes, Station 101, located just outside the mouth of the canal that connects the River B with the water treatment facility, was selected. A two-week ahead prediction period was selected for this modeling case**.** The other three modeling cases include the one-week ahead cyanobacteria predictions at Station 100, two-week ahead chlorophytes predictions at station 101 and lastly, the one-week ahead chrysophytes

predictions at Station 612. The figures, descriptive statistics and sensitivity analyses results of the six representative models are presented in the remainder of this section.

8.2.1 Original Modeling Paradigm – Larger Data Sets with Fewer Inputs versus Smaller Data Sets with More Inputs

Recognizing the inherent advantage of the original modeling approach, where input values are known a-priori, an effort was made to improve generalization capability of these models by simultaneously generating larger number of data events by reducing the number of input variables by excluding the following less frequently measured water quality variables: Biological Oxygen Demand (BOD), Total Phosphorous/Orthophosphate, Nitrite/Nitrate, Sulfate, and Total Organic Carbon (note BOD not excluded for Station 100). For Station 100, the increase in the number of data events increased on average approximately 1.6 times; for Station 101, data set sizes increased approximately 2 to 2.5 times; and for Station 612, the increase was approximately 2.4 to 3.5 times. At the same time, for all models, data set sizes for training fell short of the minimum computed value of 200 training events. This was most pronounced for Stations 101 and 612, when all input variables were included, with as little as 19 and 20 events available for training, respectively. Station 100, with by far the most historical data, was the only station that ever had more than 100 events available for training, with a maximum number of 136. However, even this station had less than 100 events available for training when its four less frequently water quality variables were included.

Thus, in general, training set sizes were on average one quarter to one half less than the minimum 200 required, computed in accordance with the number of input and output variables. As discussed in the ANN Background section, ANN generalization capability generally increases with larger training sets consisting of more historical events. Also, an ANN with a larger number of input variables will generally require more training patterns to effectively search the higher dimensional error surface during learning. In a complex

and non-linear system, it may be expected that even more data events are required for robust model development and validation.

A comparison between representative original ANN models, which included these water quality variables, and the models that excluded them, is shown below for the six representative cases. Tables 25 and 26 summarize the statistical performances of the models developed with complete and reduced input variable sets, respectively, including the number of data events available for model training and validation, the number of bloom events, and the number of false positive and false negatives with respect to algal blooms. In terms of mean absolute errors and correlation coefficients, there does not appear to be a significant discrepancy in overall model performance between the complete and reduced input models. A similar lack of discrepancy can be observed when comparing the occurrences of false positives and false negatives. However, it should also be noted that because the models that used the reduced input variable set had larger data sets, there was more variability in system conditions, or algal counts, as measured by the standard deviation. Thus, the reduced models were also subject to a greater range of conditions during validation.

As an additional measure of model performance, Table 27 compares the percentage accuracy of the models in terms of predicting relative increases or decreases from the initial to final measured counts. Please note that for cases where the initial and final measured counts were 0, ANN predictions less than 100 were deemed correct (i.e. non-bloom event). This was done to correct for what would otherwise require an exact prediction of 0, and as shown later in the figures, for most of these cases, the ANNs predicted counts significantly less than 100. Using this measure, all ANN models achieve 56% accuracy or higher in predicting increasing or decreasing counts. The original ANN models that included the less frequently measured water quality variables statistically outperformed the models that excluded these variables for four of the six cases, though advantages were relatively small, with average correct percentages of 78 and 71.5%, respectively, for the two methods. On the other hand, the models that

excluded the select water quality inputs on average achieved a higher correlation coefficient during validation; 0.80 versus 0.64 for the models that included these variables. Thus, the impact of excluding the five select water quality variables for the purpose of generating larger data sets is not clear in this case. At the same time, unlike Swimming River, it appears that excluding select water quality variables does not significantly compromise forecasting performance.

Table 25. Comparison of statistical performance of Original ANN Models for Predicting the three different algae classes at different modeling horizons using all inputs

	Station 612 – Two week Ahead Cyanobacteria Predictions			Station 100 - One week Ahead Chlorophytes Predictions			Station 101 – Two week Ahead Chrysophytes Predictions		
	Overall	Training	Validate	Overall	Training	Validate	Overall	Training	Validate
Data Mean	52.000	45.143	104.000	45.455	54.910	38.667	75.400	73.000	77.600
Data S.D.	174.835	157.523	260.059	82.410	96.075	68.815	61.635	76.230	39.159
Error Mean	0.072	0.511	1.517	-4.433	-10.906	-1.519	-1.311	-4.838	-1.737
Error S.D.	72.343	11.389	145.360	65.237	81.245	54.315	47.906	59.438	30.982
Abs E. Mean	25.607	8.643	72.119	33.485	41.627	30.026	32.075	36.307	27.853
S.D. Ratio	0.414	0.072	0.559	0.792	0.846	0.789	0.777	0.780	0.791
Correlation	0.939	0.997	0.915	0.612	0.538	0.618	0.638	0.670	0.657
No. of Events	54	28	13	156	78	39	40	20	10
No. of Blooms	5	2	2	17	12	5	9	5	2
False Positives	0	0	0	3	2	1	4	1	2
False Negatives	1	0	1	5	7	3	2	2	0
	Station 100 – Two week Ahead Cyanobacteria Predictions			Station 101 - One week Ahead Chlorophytes Predictions			Station 612 – One week Ahead Chrysophytes Predictions		
	Overall	Training	Validate	Overall	Training	Validate	Overall	Training	Validate
Data Mean	34.152	52.200	17.128	52.293	41.714	26.800	63.769	67.077	57.846
Data S.D.	144.995	196.960	45.363	76.158	70.260	25.111	55.032	65.636	38.769
Error Mean	5.133	2.349	21.386	5.034	2.735	10.720	-0.032	0.509	-0.926
Error S.D.	49.080	44.384	61.905	48.849	40.327	47.876	27.086	21.487	40.909
Abs E. Mean	20.457	20.251	28.864	35.939	26.729	42.159	19.529	15.147	31.480
S.D. Ratio	0.338	0.225	1.365	0.641	0.574	1.907	0.492	0.327	1.055
Correlation	0.944	0.974	0.903	0.780	0.828	0.640	0.871	0.945	0.122
No. of Events	158	80	39	41	21	10	52	26	13
No. of Blooms	11	7	2	7	3	0	9	5	2
False Positives	5	3	2	5	1	2	0	0	0
False Negatives	3	2	0	0	0	0	4	2	2

Table 26. Comparison of statistical performances of Original ANN Models for Predicting the three different algae classes at different modeling horizons excluding five water quality inputs.

	Station 612 – Two week Ahead Cyanobacteria Predictions			Station 100* - One week Ahead Chlorophytes Predictions			Station 101 – Two week Ahead Chrysophytes Predictions		
	Overall	Training	Validate	Overall	Training	Validate	Overall	Training	Validate
Data Mean	34.849	20.000	28.941	49.455	48.515	44.939	69.736	59.037	74.615
Data S.D.	157.155	93.832	157.271	77.028	72.579	70.967	67.014	53.632	70.217
Error Mean	-1.414	0.855	2.074	3.697	-2.165	18.461	-0.300	6.022	-12.226
Error S.D.	32.336	22.685	16.580	52.465	40.655	68.847	54.353	40.884	52.119
Abs E. Mean	12.816	10.933	7.446	33.569	27.683	42.556	37.639	30.523	38.308
S.D. Ratio	0.206	0.242	0.105	0.681	0.560	0.970	0.811	0.762	0.742
Correlation	0.979	0.975	0.995	0.772	0.828	0.807	0.590	0.649	0.707
No. of Events	139	71	34	266	134	66	106	54	26
No. of Blooms	8	4	1	39	18	8	26	10	6
False Positives	0	0	0	11	4	4	7	4	0
False Negatives	2	2	0	19	6	4	7	4	1
	Station 100*– Two week Ahead Cyanobacteria Predictions			Station 101 - One week Ahead Chlorophytes Predictions			Station 612 – One week Ahead Chrysophytes Predictions		
	Overall	Training	Validate	Overall	Training	Validate	Overall	Training	Validate
Data Mean	30.297	45.600	9.871	44.367	50.982	31.259	80.231	78.667	86.698
Data S.D.	121.847	162.598	34.542	75.159	83.132	49.556	82.210	82.743	83.477
Error Mean	6.100	-6.193	26.298	2.782	0.053	8.697	-2.623	-1.869	-5.852
Error S.D.	59.382	60.479	71.099	50.770	58.035	34.239	58.721	58.456	64.108
Abs E. Mean	31.144	33.885	31.982	30.809	32.441	28.385	39.477	37.550	44.684
S.D. Ratio	0.487	0.372	2.058	0.675	0.698	0.691	0.714	0.706	0.768
Correlation	0.873	0.937	0.847	0.738	0.716	0.789	0.700	0.708	0.656
No. of Events	249	125	62	109	55	27	173	87	43
No. of Blooms	19	14	1	15	8	3	46	20	15
False Positives	5	2	3	8	6	1	9	6	3
False Negatives	10	8	0	1	1	0	22	7	10

* - excluded only four less frequently measured water quality variables

Table 27. Percentage accuracy of the Original ANN Models for predicting the three different algae classes at different modeling horizons in terms of predicting relative increases or decreases from the validation data sets' Initial to Final measured counts

	Prediction Events	No. of Events	Accuracy			
			Correct	%	Incorrect	%
Original Models with All inputs	612-2wk Ahead Cyanobacteria	13	12	92	1	8
	100-1wk Ahead Chlorophyta	39	27	69	12	31
	101-2wk Ahead Chrysophyta	10	9	90	1	10
	100-2wk Ahead Cyanobacteria	39	37	95	2	5
	101-1wk Ahead Chlorophyta	10	6	60	4	40
	612-2wk Ahead Chlorophytes	13	8	62	5	38
Original Models with Fewer Inputs (exclude select water quality variables)	612-2wk Ahead Cyanobacteria	34	31	91	3	9
	100*-1wk Ahead Chlorophyta	66	37	56	29	44
	101-2wk Ahead Chrysophyta	26	16	62	10	38
	100*-2wk Ahead Cyanobacteria	62	58	94	4	6
	101-1wk Ahead Chlorophyta	27	17	63	10	37
	612-2wk Ahead Chlorophytes	43	27	63	16	37

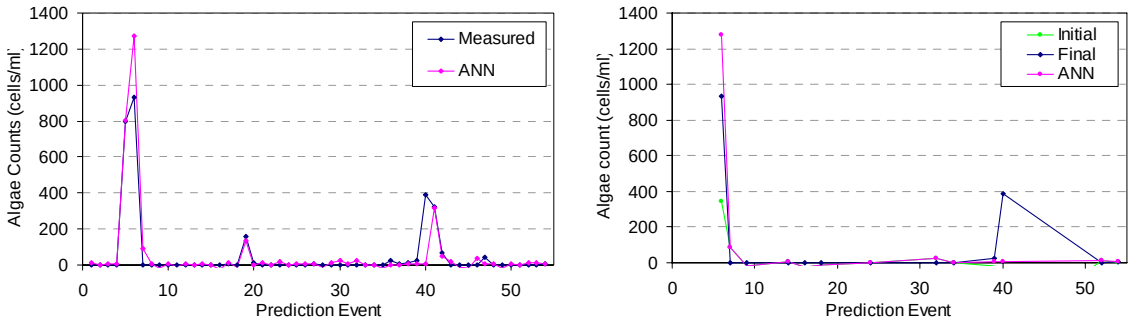
* - excluded only four less frequently measured water quality variables

This finding suggests that for this particular system, unlike Swimming River, the range of values under which the excluded variables exist over the modeling history, while having some influence, may not significantly affect the algal populations. Similar findings have been reported in the literature (Maier and others, 1997), and it may be that the constituents exist within a range of values that do not significantly diminish or propagate the organisms. Again, however, given the size deficiency of data sets, this possible explanation must be taken with guarded skepticism.

What does lend support, however, is the opposite result was found with the Swimming River Reservoir, where significantly smaller data sets that included nutrient variables significantly increased performance. There may even be a physical basis for these differences in model sensitivities to nutrient conditions in the two water utility systems. A reservoir, by nature a closed system, is hydraulically less dynamic, and short-term

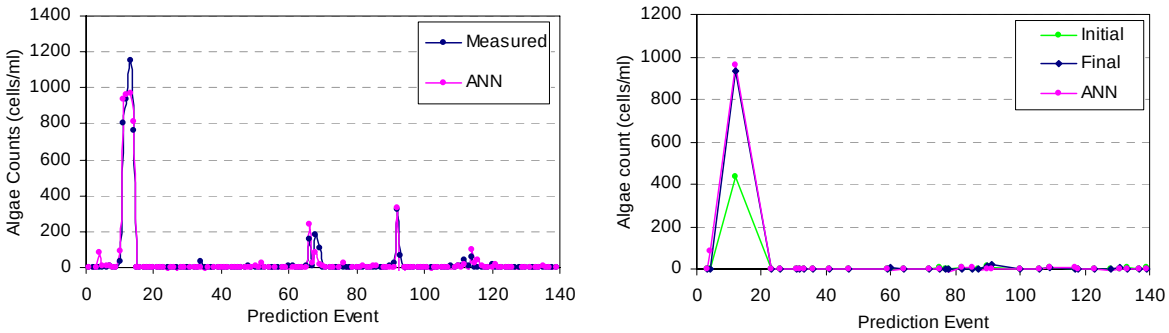
water quality changes may induce stronger influences over algal populations during relatively short time periods. A river system by its flowing nature is more dynamic hydraulically, and may comprise an algal ecosystem that is less influenced by nutrient changes over relatively short time periods.

In order to provide a visual comparison, Figures 69 through 80 below compare the overall and validation modeling results for the different models.



(a) (b)

Figure 69. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data sets at Station 612 (Original Model with all inputs)



(a) (b)

Figure 70. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data set at Station 12 (Original Model excluding five water quality inputs)

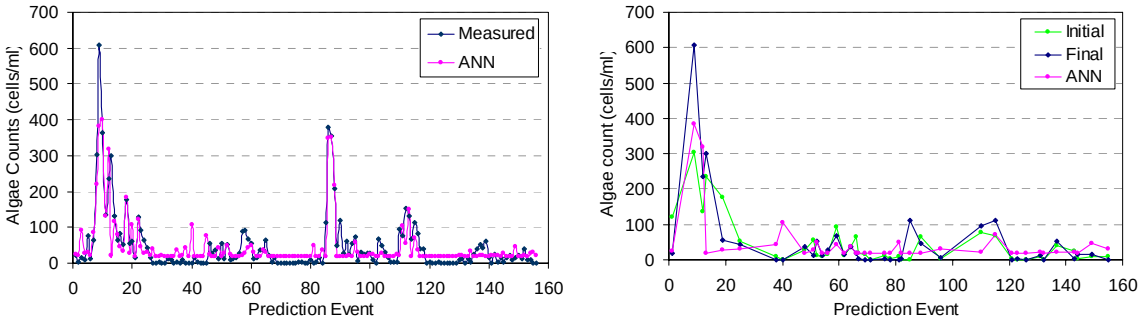


Figure 71. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 100 (Original Model with all inputs)

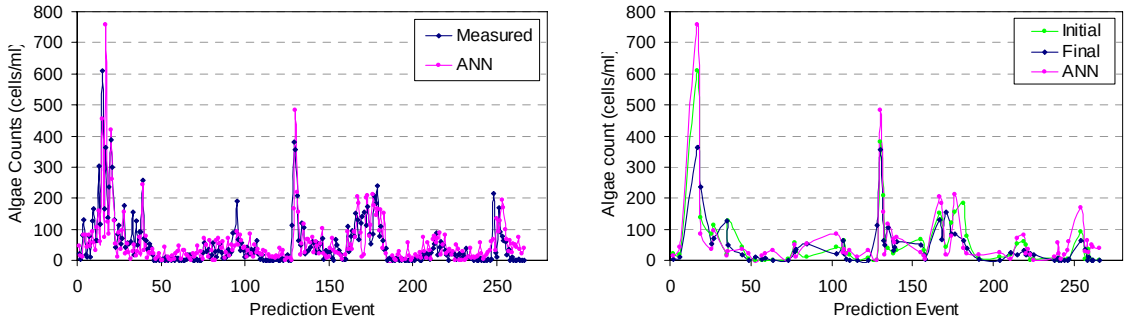


Figure 72. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 100 (Original Model excluding five water quality inputs)

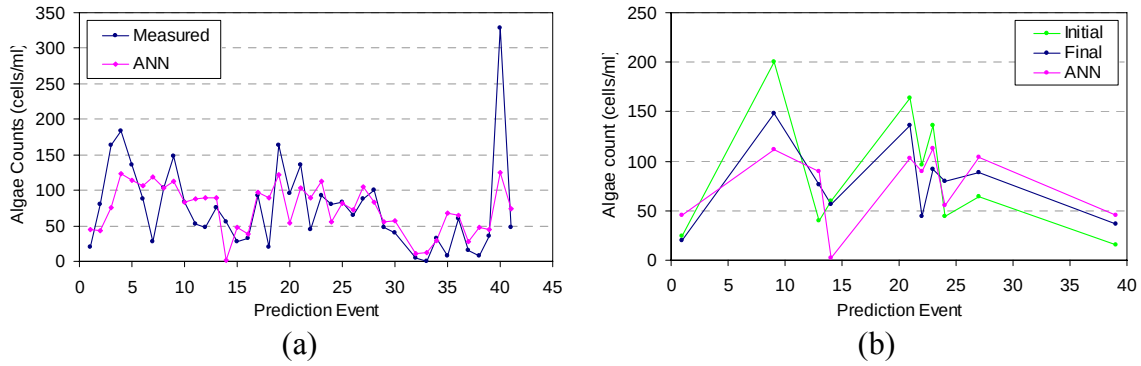
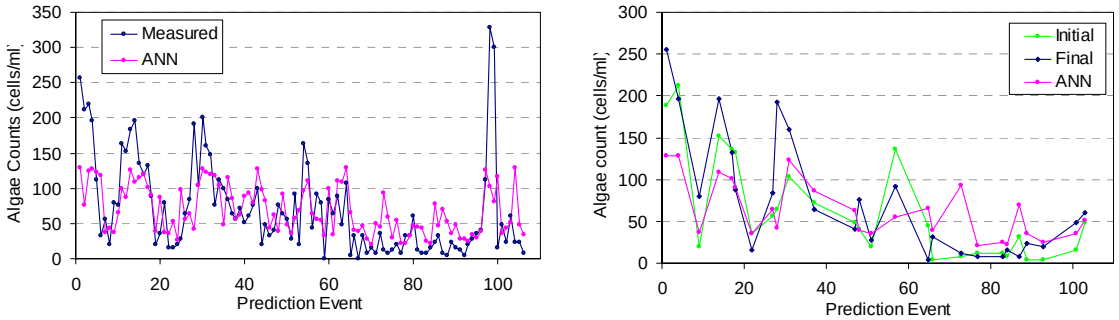
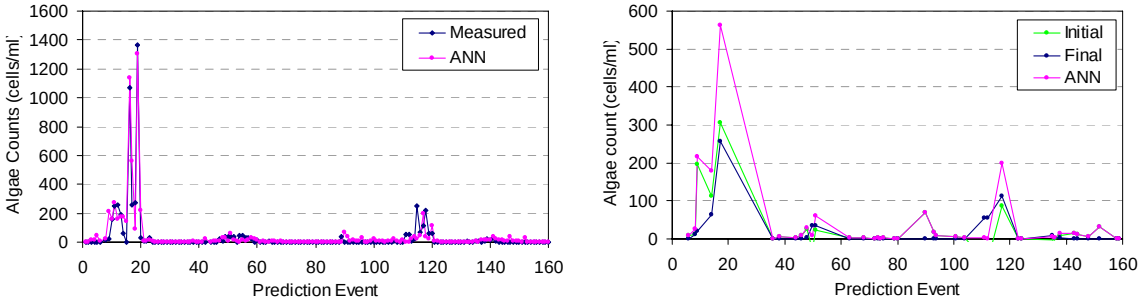


Figure 73. Time-series plots of measured Chrysophytes counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data set at Station 101 (Original Model using all inputs)



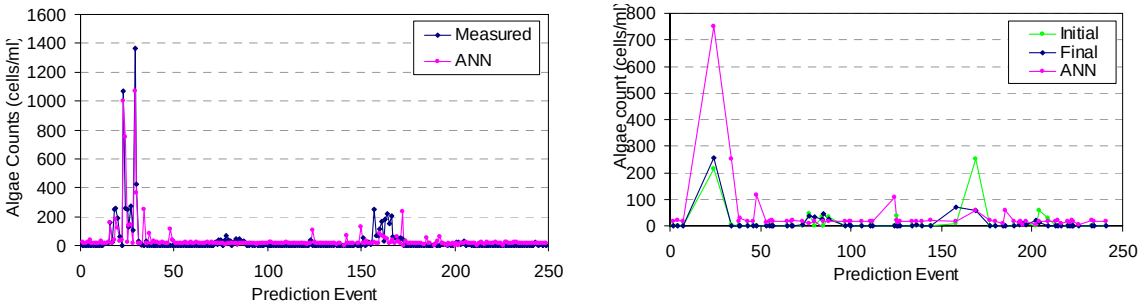
(a) (b)

Figure 74. Time-series plots of measured Chrysophytes counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data set at Station 101 (Original Model excluding five water quality inputs)



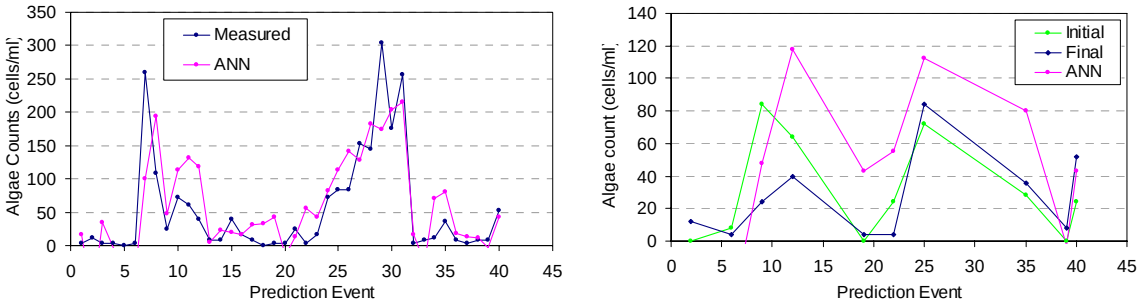
(a) (b)

Figure 75. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data set at Station 100 (Original Model using all inputs)



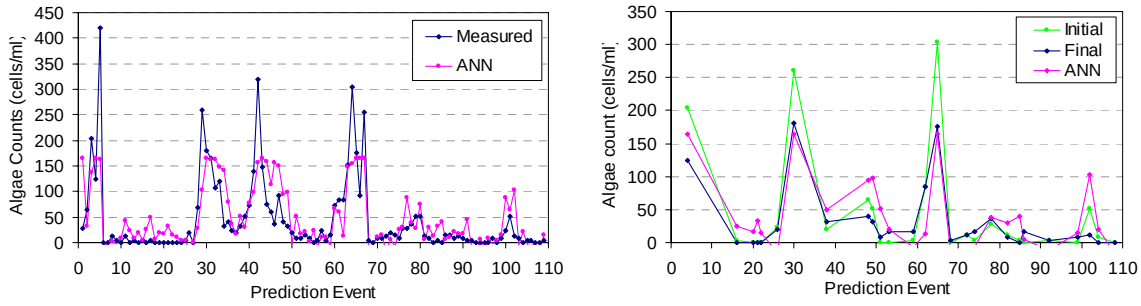
(a) (b)

Figure 76. Time-series plots of measured Cyanobactera counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data set at Station 100 (Original Model excluding five water quality inputs)



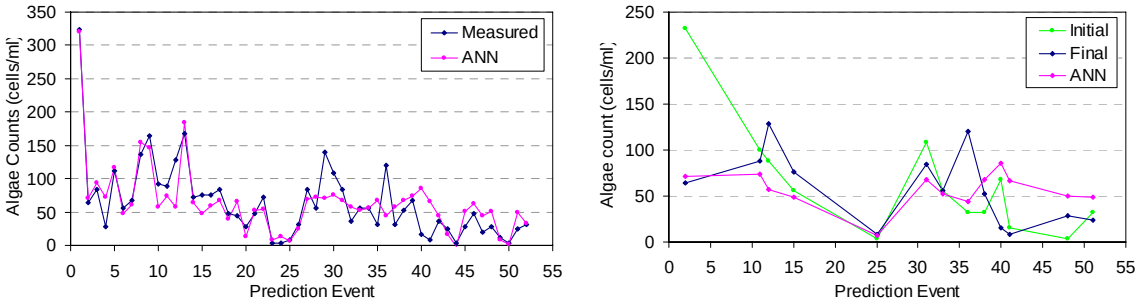
(a) (b)

Figure 77. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 101 (Original Model using all inputs)



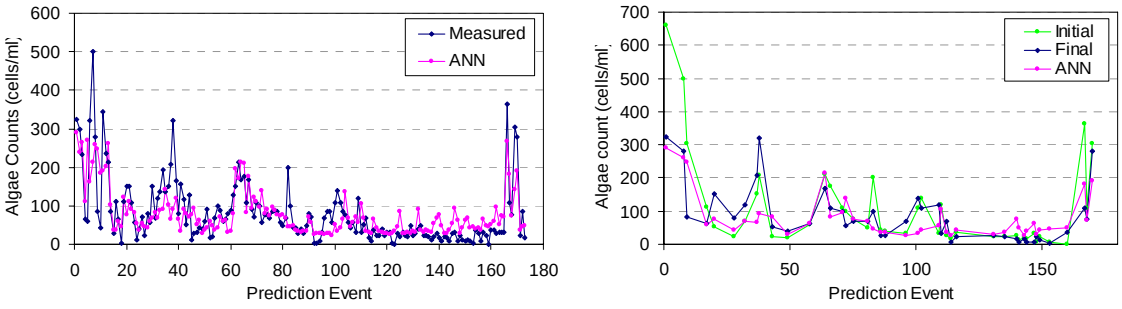
(a) (b)

Figure 78. Time-series plots of measured Chlrophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 101 (Original Model excluding five water quality inputs)



(a) (b)

Figure 79. Time-series plots of measured Chrysophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 612 (Original Model using all inputs)



(a) (b)

Figure 80. Time-series plots of measured Chrysophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 612 (Original Model excluding five water quality inputs)

A visual comparison of the model predictions, particularly the validation figures, is helpful for assessing ANN performance. For depicting validation results, the initial algal counts measured at the beginning of the prediction period are plotted along with the measured and ANN predicted final counts so as to provide a more transparent assessment of forecasting performance. For example, on Figure 80b, the first validation event had an initial chrysophytes algal count around 680 organisms (i.e. cells). At the end of the corresponding one-week prediction event, the final measured count was just over 300 organisms, while the ANN predicted a count just below 300.

Overall, given the sparse data, many models perform surprisingly well in predicting dramatic changes in algal populations during validation. For example, for the two-week ahead prediction period for cyanobacteria at Station 612 using all inputs (Figure 69b), the ANN model accurately predicted a 600 count increase for this organism. The ANN model developed for predicting chlorophytes at Station 101 one-week ahead (Figure 77b) using all inputs validated extremely well, accurately predicting relative increases and decreases in counts. The same was achieved with the mode developed for predicting chrysophytes counts one week ahead for Station 612 with the reduced inputs. The model accurately predicted very large decreases in algal counts (i.e. 200 or more) for three events, and generally reproduced higher and lower count events. Even a model trained with very few data events (Chrysophytes two-week ahead predictions at Station 101 with

complete inputs) accurately predicted relative increases and decreases in algal counts during validation (Figure 73b). Interestingly, the corresponding model that used the reduced input set for the same prediction problem did not predict relative changes as well, as shown by Figure 74b, and seems to be "keying" off the initial counts. In this case, it suggests that at least some of the water quality variables may have been important for this prediction problem.

As inferred from statistical measures, the figures do not reveal an obvious superiority between the complete and reduced input models. Thus, increasing the number of historical events did not significantly improve ANN forecasting ability, although the models developed with more data events generally achieved higher correlation coefficients (0.80 versus 0.64). By extension, excluding the five select water quality variables may not generally compromise development of robust ANN forecasting models, although there appears to be at least one exception.

Perhaps most importantly, the modeling results suggest that on average, the initial system conditions measured at the beginning of the prediction period capture the dynamics that govern algal population changes over the prediction periods of interest. This could be due to relatively small changes in system conditions over the short prediction periods, and/or natural and consistent transitions from initial system conditions, measured at the beginning of the period, to end of the prediction period, when final algal counts are predicted. In other words, there may be a natural and consistent time lag in algal population responses to system conditions that, except under extreme conditions (e.g. sudden and dramatic temperature change), evolve in a fairly consistent. The extremely important implication is that development of robust ANN models that use input values measured at the beginning of the prediction period appears to be quite feasible.

8.2.2 Sensitivity Analysis Results – Original Model Paradigms

This section compares the sensitivity analyses results obtained for the two original modeling approaches. The analysis is done using both the ratio and ranking method, described previously in the water treatment modeling section.

Table 28 below presents the sensitivity analysis results for the original model that included all input variables for the two-week ahead prediction of cyanobacteria at Station 612. As depicted by the table, odor and total algal counts ranked first and second in terms of importance, respectively, both with ratio value over 1.2. Other variables that ranked highly in terms of importance with ratio value over 1.1. These variables include length of day, total amorphous materials and Reservoir A extraction. At the other extreme, a total of 14 of the 34 variable inputs achieved low ratio values of less than 1.0. The three lowest ranking were pH, sky cover, and BOD.

Table 28. Sensitivity Analysis for Original ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 612 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Odor	1.530	1	Total Hardness	1.010	18
Total Algal Count	1.214	2	Nitrite/Nitrate	1.008	19
Temperature	1.129	3	Pumping Station 1 Extraction	1.004	20
Length of Day	1.109	4	Chloride	0.998	21
Total Amorphous Material	1.108	5	Initial Cyanobacteria counts	0.998	22
Reservoir A Extraction	1.097	6	River A Extraction	0.993	23
Wind Direction	1.088	7	Dissolved Oxygen	0.991	24
Sulfate	1.063	8	UV254	0.989	25
Initial Chrysophytes counts	1.046	9	Turbidity	0.983	26
Heating Degree Days	1.035	10	Prediction Period's Precipitation Total	0.975	27
Ammonia	1.032	11	Color	0.975	28
Alkalinity	1.031	12	TotalPhosphorus/Orthophosphate	0.967	29
Conductivity	1.029	13	River B Extraction	0.964	30
River A Streamflow	1.028	14	Prediction Period's Lagged Precipitation Total	0.946	31
Initial Chlorophytes counts	1.025	15	Biochemical Oxygen Demand	0.938	32
Wind Speed	1.024	16	Sky Cover	0.936	33
Total Organic Carbon	1.012	17	pH	0.864	34

For the original model with reduced input variables for the two-week ahead prediction of cyanobacteria at Station 612, as presented in Table 29, extraction from Reservoir A was the most important variable with a ratio value of 2.89. Odor and extraction from River A also ranked high in importance, both with ratio values over 1.8. At the other extreme, a number of low ranking variables had ratio values below 1.0, five of which had ratio values of 0.89 and below, including total hardness, dissolved oxygen, alkalinity, pH and conductivity.

Table 29. Sensitivity Analysis for Original ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 612 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Reservoir A Extraction	2.895	1	Total Algal counts	1.025	16
Odor	1.835	2	River A Streamflow	1.024	17
River A Extraction	1.802	3	Initial Chlorophytes counts	1.023	18
Length of Day	1.602	4	Initial Chrysophytes counts	1.010	19
River B Extraction	1.451	5	UV254	0.998	20
Turbidity	1.278	6	Pump Station 1 Extraction	0.992	21
Sky Cover	1.229	7	Prediction Period's Precipitation Total	0.984	22
Prediction Period's Lagged Precipitation	1.189	8	Initial Cyanobacteria counts	0.974	23
Heating Degree Days	1.133	9	Chloride	0.933	24
Wind Direction	1.091	10	Conductivity	0.895	25
Total Amorphous Material	1.068	11	pH	0.893	26
Ammonia	1.064	12	Alkalinity	0.873	27
Wind Speed	1.044	13	Dissolved Oxygen	0.818	28
Temperature	1.033	14	Total Hardness	0.818	29
Color	1.032	15			

The sensitivity analyses results for the original ANN model with all the variable inputs for one-week ahead prediction of chlorophytes counts at Station 100 is shown below in Table 30. Initial chlorophytes counts ranked first in importance, followed by heating degree days and length of day. Ratio values, however, were relatively low. That is, most of the variables on top of the ranking including the highest ranking initial chlorophytes counts had ratio values of just a little over 1.0. At the other extreme, several variables

achieved ratio values of less than 1.0, with the three lowest ranking being odor, ammonia and initial cyanobacteria counts.

Table 30. Sensitivity Analysis for Original ANN Model for One-week Ahead Predictions of Chlorophytes at Station 100 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chlorophytes counts	1.043	1	UV-254	1.002	18
Heating Degree Days	1.024	2	Total Suspended Solids	1.001	19
Length of Day	1.020	3	pH	1.001	20
Wind Direction	1.017	4	River B Extraction	1.001	21
River A Extraction	1.017	5	Dissolved Oxygen	1.000	22
Total Algal counts	1.013	6	River A Streamflow	1.000	23
Temperature	1.011	7	Pumping Station 1 Extraction	1.000	24
Nitrite/Nitrate	1.010	8	Prediction Period's Lagged Precipitation Total	1.000	25
Wind Speed	1.010	9	Conductivity	1.000	26
Total Organic Carbon	1.008	10	Total Hardness	0.999	27
Initial Chrysophytes counts	1.006	11	Chloride	0.999	28
Sulfate	1.006	12	Reservoir A Extraction	0.999	29
T.Phosphorus/Orthophosphate	1.005	13	Alkalinity	0.998	30
Prediction Period's Precipitation Total	1.005	14	Initial Cyanobacteria counts	0.996	31
Sky Cover	1.004	15	Ammonia	0.991	32
Turbidity	1.003	16	Odor	0.964	33
Biochemical Oxygen Demand	1.002	17			

Table 31 below presents the sensitivity analysis results for the similar model with reduced inputs. Initial chlorophytes counts ranked first in terms of importance with a ratio value of 1.3, followed by odor, and then extractions from Rivers A and B. At the bottom of the ranking, a total of seven variables had ratios less than 1.0. Among these lower ranking variables were initial chrysophytes counts, previous and current week's precipitation totals, and dissolved oxygen.

Table 31. Sensitivity Analysis for Original ANN Model for One-week Ahead Predictions of Chlorophytes at Station 100 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chlorophytes	1.304	1	Conductivity	1.006	16
Odor	1.241	2	River A Streamflow	1.004	17
River A Extraction	1.181	3	Total Organic Carbon	1.004	18
River B Extraction	1.181	4	Initial Cyanobacteria counts	1.002	19
Biochemical Oxygen Demand	1.059	5	Alkalinity	1.001	20
Sky Cover	1.045	6	Reservoir A Extraction	1.001	21
Turbidity	1.036	7	pH	1.001	22
Chloride	1.029	8	Heating Degree Days	0.999	23
Ammonia	1.026	9	UV254	0.999	24
Total Algal Counts	1.022	10	Pump Station 1 Extraction	0.999	25
Wind Direction	1.018	11	Dissolved Oxygen	0.998	26
Temperature	1.017	12	Prediction Period's Precipitation Total	0.998	27
Wind Speed	1.011	13	Prediction Period's Lagged Precipitation Total	0.995	28
Length of Day	1.009	14	Initial Chrysophytes counts	0.994	29
Total Hardness	1.008	15			

Table 32 below presents the sensitivity analysis results for the original model that included all input variables for the two-week ahead prediction of chrysophytes at Station 101. As with the other original models developed and assessed with the complete input set, variables achieved relatively low ratio values, including most of the top ranked variables with ratio values of just over 1.0. A number of variables at the bottom of the ranking had ratio values of less than 1.0, and the three lowest ranking variables were extraction from Pumping Station 1, conductivity, and chloride.

Table 32. Sensitivity Analysis for Original ANN Model for Two-week Ahead Predictions of Chrysophytes at Station 101 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chrysophytes count	1.091	1	Prediction Period's Precipitation Total	1.004	18
Color	1.067	2	River A Streamflow	1.003	19
Length of Day	1.043	3	Nitrite/Nitrate	1.003	20
Wind Direction	1.035	4	Wind Speed	1.002	21
Ammonia	1.016	5	Turbidity	1.001	22
Total Algal Counts	1.012	6	T.Phosphorus/Orthophosphate	1.000	23
Alkalinity	1.010	7	Reservoir A Extraction	1.000	24
Initial Cyanobacteria counts	1.010	8	Biochemical Oxygen Demand	1.000	25
pH	1.009	9	UV254	1.000	26
River B Extraction	1.009	10	Dissolved Oxygen	0.999	27
River A Extraction	1.009	11	Sulfate	0.999	28
Initial Chlorophytes counts	1.008	12	Total Hardness	0.998	29
Heating Degree Days	1.006	13	Temperature	0.997	30
Odor	1.006	14	Prediction Period's Lagged Precipitation Total	0.997	31
Total Organic Carbon	1.005	15	Chloride	0.994	32
Total Amorphous Material	1.005	16	Conductivity	0.994	33
Sky Cover	1.005	17	Pumping Station 1 Extraction	0.992	34

Similar to the ANN model for two-week ahead chrysophytes predictions at Station 101 using the complete input set, the sensitivity analysis result for the model with reduced inputs exhibited low overall ratio values, as shown in Table 33 below. In particular, the four highest ranking variables, total algal counts, initial chrysophytes counts, length of day, and pH, had ratio values of just over 1.0. . Most of the remaining variables had ratio values around 1.0, and the seven lowest ranking variables all had ratio values less than 1.0, with the three lowest ranking being total amorphous materials, turbidity and previous week's precipitation.

Table 33. Sensitivity Analysis for Original ANN Model for Two-week Ahead Predictions of Chrysophytes at Station 101 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Total Algal counts	1.026	1	Color	1.002	16
Initial Chrysophytes counts	1.022	2	UV254	1.001	17
Length of Day	1.014	3	Chloride	1.001	18
pH	1.010	4	Sky Cover	1.001	19
Pump Station 1 Extraction	1.009	5	Heating Degree Days	1.000	20
Reservoir A Extraction	1.008	6	River A Streamflow	1.000	21
Prediction Period's Precipitation Total	1.007	7	Ammonia	1.000	22
Initial Cyanobacteria counts	1.005	8	River B Extraction	0.999	23
Conductivity	1.004	9	Initial Chlorophytes counts	0.999	24
Alkalinity	1.004	10	Total Hardness	0.999	25
Odor	1.003	11	River A Extraction	0.998	26
Dissolved Oxygen	1.003	12	Prediction Period's Lagged Precipitation Total	0.995	27
Temperature	1.002	13	Turbidity	0.994	28
Wind Direction	1.002	14	Total Amorphous Material	0.990	29
Wind Speed	1.002	15			

The sensitivity analysis result for original modeling for two-week ahead cyanobacteria counts at station 100 with the complete input set is shown in Table 34. In terms of importance, the two highest ranking variables were pH and nitrite/nitrate, both with ratio value over 1.6. Oother variables that ranked highly with ratio values above 1.3 include odor, sky cover, total organic carbon, River B extractions and intial chrysophytes counts. At the other extreme, a total of nine variables achieved ratio values less than 1.0, with the lowest ranking being total suspended solids, wind speed, and length of day.

Table 34. Sensitivity Analysis for Original ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 100 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
pH	1.667	1	Total Algal Counts	1.043	18
Nitrite/Nitrate	1.638	2	Prediction Period's Lagged Precipitation Total	1.034	19
Odor	1.430	3	UV-254	1.030	20
Sky Cover	1.396	4	Chloride	1.020	21
Total Organic Carbon	1.362	5	Biochemical Oxygen Demand	1.009	22
River B Extractions	1.321	6	Reservoir A Extractions	1.006	23
Initial Chrysophyta counts	1.301	7	Initial Cyanophytes counts	1.004	24
T.Phosphorus/Orthophosphate	1.240	8	Prediction Period's Precipitation Total	0.999	25
Wind Direction	1.227	9	Alkalinity	0.984	26
Sulfate	1.217	10	Pumping Station 1 Extractions	0.981	27
Heating Degree Days	1.206	11	Conductivity	0.978	28
River A Extractions	1.203	12	River A Streamflow	0.953	29
Turbidity	1.194	13	Temperature	0.943	30
Dissolved Oxygen	1.170	14	Length of Day	0.941	31
Total Hardness	1.116	15	Wind Speed	0.927	32
Initial Chlorophytes counts	1.101	16	Total Suspended Solids	0.896	33
Ammonia	1.092	17			

The sensitivity analysis results for the similar prediction model but with the reduced input set are presented in Table 35. Extractions from River B ranked first in terms of importance with ratio value of 1.52, followed by sky cover, total organic carbon, and biochemical oxygen demand, all with ratio values of over 1.2. The lower ranking variables, all with ratio values less than 1.0, include the prediction period's precipitation total, water temperature, and extractions from Pumping Station 1.

Table 35. Sensitivity Analysis for Original ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 100 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
River B Extraction	1.519	1	Prediction Period's Lagged Precipitation Total	1.136	16
Sky Cover	1.327	2	Ammonia	1.119	17
Total Organic Carbon	1.299	3	River A Streamflow	1.062	18
Biochemical Oxygen Demand	1.297	4	Length of Day	1.051	19
Initial Chrysophytes counts	1.292	5	Initial Cyanobacteria counts	1.051	20
Wind Direction	1.281	6	Total Hardness	1.043	21
Turbidity	1.269	7	Total Algal Counts	1.029	22
Heating Degree Days	1.233	8	UV254	1.020	23
River A Extraction	1.220	9	Chloride	1.019	24
Dissolved Oxygen	1.213	10	Total Suspended Solids	1.010	25
Initial Chlorophytes counts	1.196	11	Conductivity	1.009	26
pH	1.182	12	Reservoir A Extraction	1.005	27
Odor	1.171	13	Pump Station 1 Extraction	0.995	28
Alkalinity	1.168	14	Temperature	0.994	29
Wind Speed	1.162	15	Prediction Period's Precipitation Total	0.952	30

The overall sensitivity analysis results for one-week ahead chlorophytes predictions with complete input set at Station 101 are summarized in Table 36. The three highest ranking variables were the initial chlorophytes counts, wind speed, and direction, and all input variables exhibited low ratio values, with only the initial chlorophytes counts having a ratio value above 1.07 (1.2). Most of the variables had ratio values just over 1.0, with and 13 having ratio values less than 1.0. The three lowest ranking variables were odor, heating degree days, and water temperature.

Table 36. Sensitivity Analysis for Original ANN Model for One-week Ahead Predictions of Chlorophytes at Station 101 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chlorophytes counts	1.211	1	Reservoir A Extractions	1.001	18
Wind Speed	1.070	2	Conductivity	1.000	19
Wind Direction	1.069	3	Dissolved Oxygen	1.000	20
River A Extraction	1.047	4	Chloride	1.000	21
Total Organic Carbon	1.035	5	Biochemical Oxygen Demand	0.999	22
Total Hardness	1.027	6	Total Amorphous Material	0.996	23
River A Streamflow	1.025	7	Length of Day	0.994	24
Initial Cyanobacteria counts	1.022	8	Ammonia	0.993	25
Sulfate	1.018	9	River B Extractions	0.992	26
Total Algal Count	1.014	10	Pumping Station 1 Extractions	0.992	27
UV254	1.012	11	Alkalinity	0.990	28
Color, Cu	1.007	12	T.Phosphorus/Orthophosphate	0.990	29
Turbidity	1.006	13	Sky Cover	0.989	30
Prediction Period's Lagged Precipitation Total	1.006	14	Initial Chrysophytes counts	0.987	31
pH	1.005	15	Temperature	0.970	32
Nitrite/Nitrate	1.001	16	Heating Degree Days	0.966	33
Prediction Period's Precipitation Total	1.001	17	Odor	0.961	34

As with the previous model, the one-week ahead chlorophytes predictions at Station 101 with reduced input set exhibited low sensitivity ratio values, as presented in Table 37. Even the highest ranking variable achieved a value of little just over 1.0 and six low ranking variables have ratio value less than 1.0. In terms of importance, initial chlorophytes counts, total algal counts, and the prediction period's precipitation ranked first, second and third, respectively, while the three lowest ranking variables were sky cover, wind speed and wind direction.

Table 37. Sensitivity Analysis for Original ANN Model for One-week Ahead Predictions of Chlorophytes at Station 101 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chlorophytes counts	1.110	1	River B Extraction	1.002	16
Total Algal counts	1.093	2	Turbidity	1.001	17
Prediction Period's Precipitation Total	1.050	3	Conductivity	1.001	18
River A Extraction	1.021	4	Ammonia	1.001	19
Total Hardness	1.009	5	UV254	1.001	20
Initial Chrysophytes counts	1.007	6	Chloride	1.001	21
Temperature	1.006	7	Pump Station 1 Extraction	1.000	22
Dissolved Oxygen	1.006	8	pH	1.000	23
Reservoir A Extraction	1.006	9	Length of Day	0.999	24
Color	1.005	10	Prediction Period's Lagged Precipitation Total	0.997	25
Heating Degree Days	1.005	11	Odor	0.997	26
River A Streamflow	1.004	12	Sky Cover	0.996	27
Alkalinity	1.004	13	Wind Speed	0.995	28
Initial Cyanobacteria counts	1.003	14	Wind Direction	0.990	29
Total Amorphous Material	1.002	15			

The sensitivity analysis results for the original ANN model for one-week ahead chrysophytes predictions at Station 612 with complete inputs is shown below in Table 38. Initial chrysophytes counts was the highest ranking variable, with a ratio value of 1.65. Other variables that also ranked highly were length of day, wind speed, and total amorphous materials. Of the seven low ranking variables that achieved ratio value of less than 1.0, the three lowest ranking were initial cyanobacteria counts, total algal counts, and turbidity.

Table 38. Sensitivity Analysis for Original ANN Model for One-week Ahead Predictions of Chrysophytes at Station 612 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chrysophytes counts	1.658	1	Reservoir A Extractions	1.013	18
Length of Day	1.295	2	Sky Cover	1.013	19
Wind Speed	1.104	3	Temperature	1.013	20
Total Amorphous Material	1.100	4	Biochemical Oxygen Demand	1.009	21
UV254	1.094	5	pH	1.008	22
TotalPhosphorus/Orthophosphate	1.088	6	Chloride	1.002	23
River B Extractions	1.078	7	Total Organic Carbon	1.000	24
Initial Chlorophytes counts	1.075	8	Prediction Period's Precipitation Total	1.000	25
Conductivity m	1.062	9	River A Streamflow	1.000	26
Total Hardness	1.059	10	Prediction Period's Lagged Precipitation Total	1.000	27
Alkalinity	1.051	11	Nitrite/Nitrate	0.998	28
River A Extractions	1.046	12	Color	0.997	29
Odor	1.031	13	Wind Direction	0.994	30
Ammonia	1.022	14	Pumping Station 1 Extractions	0.994	31
Dissolved Oxygen	1.020	15	Turbidity	0.993	32
Heating Degree Days	1.017	16	Total Alga Count	0.991	33
Sulfate	1.013	17	Initial Cyanobacteria counts	0.965	34

Table 39 presents the sensitivity analysis results for one-week ahead predictions of Chrysophytes at Station 612 with the reduced input set. As with the similar model with complete input set, initial chrysophytes counts was the highest ranking variable; however, this variable achieved a relatively low ratio value of only 1.1, as compared with the 1.6 with the other model. Other variables that ranked near the top, with low ratio value just over 1.0, include odor, temperature and chloride. The three lowest ranking variables, all with ratio values below 1.0, were sky cover, pH, and initial chlorophytes counts.

Table 39. Sensitivity Analysis for Original ANN Model for One-week Ahead Predictions of Chrysophytes at Station 612 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chrysophytes counts	1.146	1	Turbidity	1.005	16
Odor	1.046	2	UV254	1.003	17
Temperature	1.029	3	Pump Station 1 Extraction	1.002	18
Chloride	1.021	4	River A Extraction	1.002	19
Heating Degree Days	1.018	5	Conductivity	1.000	20
Length of Day	1.017	6	Alkalinity	1.000	21
Total Hardness	1.014	7	Reservoir A Extraction	1.000	22
Total Algal counts	1.014	8	Prediction Period's Precipitation Total	1.000	23
Color	1.009	9	River B Extraction	0.999	24
Ammonia	1.008	10	Prediction Period's Lagged Precipitation	0.998	25
Dissolved Oxygen	1.006	11	River A Streamflow	0.997	26
Wind Direction	1.006	12	Initial Chlorophytes counts	0.997	27
Wind Speed	1.006	13	pH	0.995	28
Total Amorphous Material	1.005	14	Sky Cover	0.986	29
Initial Cyanobacteria counts	1.005	15			

8.2.3 Revised Modeling Paradigm – Larger Data Sets with Fewer Inputs versus Smaller Data Sets with More Inputs

As was done with the original modeling approach, a comparison between the larger number of historical data events that excluded the select water quality variables versus inclusion of these variables with fewer events was performed with the revised models. As was found with the original modeling approach, increasing the number of data events by excluding the five select water quality variables did not necessarily produce superior results. Tables 40 and 41 summarize the statistical results for these two data sets for the six representative forecasting cases. Half of the models that excluded the five select water quality variables achieved higher correlations during validation, and half did not. What is interesting is that the revised models, regardless of the set of input variables, generally avoided false positives and false negatives.

Table 40. Comparison of statistical performances of Revised ANN Models for Predicting the three different algae classes at different modeling horizons using all inputs

	Station 612 – Two week Ahead Cyanobacteria Predictions			Station 100 - One week Ahead Chlorophytes Predictions			Station 101 – Two week Ahead Chrysophytes Predictions		
	Overall	Training	Validate	Overall	Training	Validate	Overall	Training	Validate
Data Mean	11.900	21.400	4.400	55.582	70.630	31.467	75.771	78.316	70.500
Data S.D.	51.244	70.782	10.651	95.147	113.067	49.750	52.424	44.016	61.284
Error Mean	-1.194	0.160	-5.419	3.204	1.584	7.182	-4.116	-0.314	-28.581
Error S.D.	8.324	7.853	11.346	51.870	57.877	28.108	31.599	7.393	53.261
Abs E. Mean	4.019	4.313	6.539	33.123	39.016	21.007	19.655	5.488	52.268
S.D. Ratio	0.162	0.111	1.065	0.545	0.512	0.565	0.603	0.168	0.869
Correlation	0.987	0.994	0.342	0.839	0.860	0.825	0.824	0.986	0.581
No. of Events	40	20	10	182	92	45	35	19	8
No. of Blooms	1	1	0	30	21	4	10	5	2
False Positives	0	0	0	9	5	1	2	0	1
False Negatives	0	0	0	11	9	0	3	1	1
	Station 100 – Two week Ahead Cyanobacteria Predictions			Station 101 - One week Ahead Chlorophytes Predictions			Station 612 – One week Ahead Chrysophytes Predictions		
	Overall	Training	Validate	Overall	Training	Validate	Overall	Training	Validate
Data Mean	31.760	44.988	17.171	53.277	39.360	90.545	81.020	98.815	38.333
Data S.D.	140.813	189.279	47.124	77.348	64.693	88.087	67.909	76.824	37.003
Error Mean	-0.773	-3.300	0.211	9.530	-1.356	39.375	15.904	9.073	46.659
Error S.D.	54.234	63.786	43.339	36.275	11.082	61.744	39.417	43.281	21.648
Abs E. Mean	24.456	28.531	19.361	19.995	9.176	51.255	34.964	35.164	48.131
S.D. Ratio	0.385	0.337	0.920	0.469	0.171	0.701	0.580	0.563	0.585
Correlation	0.923	0.942	0.393	0.908	0.985	0.803	0.814	0.826	0.842
No. of Events	167	85	41	47	25	11	51	27	12
No. of Blooms	12	6	3	7	2	4	15	9	2
False Positives	8	2	0	1	0	1	2	2	0
False Negatives	2	4	2	0	0	0	5	3	0

Table 41. Comparison of statistical performances of Revised ANN Models for Predicting the three different algae classes at different modeling horizons excluding five water quality inputs

	Station 612 – Two week Ahead Cyanobacteria Predictions			Station 100* - One week Ahead Chlorophytes Predictions			Station 101 – Two week Ahead Chrysophytes Predictions		
	Overall	Training	Validate	Overall	Training	Validate	Overall	Training	Validate
Data Mean	28.886	33.829	10.857	51.167	57.360	32.537	80.606	77.529	64.250
Data S.D.	142.553	143.274	53.858	83.488	85.256	52.854	72.390	64.389	72.477
Error Mean	0.540	1.876	-2.077	1.810	0.561	9.598	-2.259	0.854	-3.943
Error S.D.	26.645	23.973	36.732	49.504	51.883	44.289	36.326	15.916	51.068
Abs E. Mean	10.973	12.160	12.131	34.560	36.144	29.930	26.909	12.983	41.603
S.D. Ratio	0.187	0.167	0.682	0.593	0.609	0.838	0.502	0.247	0.705
Correlation	0.983	0.987	0.897	0.805	0.794	0.603	0.865	0.969	0.722
No. of Events	140	70	35	270	136	67	66	34	16
No. of Blooms	7	4	1	40	23	6	21	8	5
False Positives	0	0	0	6	2	1	2	1	0
False Negatives	2	1	0	23	13	3	3	0	1
	Station 100* – Two week Ahead Cyanobacteria Predictions			Station 101 - One week Ahead Chlorophytes Predictions			Station 612 – One week Ahead Chrysophytes Predictions		
	Overall	Training	Validate	Overall	Training	Validate	Overall	Training	Validate
Data Mean	29.921	44.635	17.143	47.542	49.417	49.333	79.885	81.636	75.814
Data S.D.	121.225	165.547	40.938	83.662	86.946	84.614	82.099	90.686	61.425
Error Mean	2.015	-0.605	5.752	-1.505	0.717	-2.126	-1.291	1.208	-5.225
Error S.D.	36.711	42.937	34.892	38.167	31.260	57.128	47.770	45.794	46.045
Abs E. Mean	19.328	22.188	19.558	21.032	18.116	28.920	35.973	34.487	34.685
S.D. Ratio	0.303	0.259	0.852	0.456	0.360	0.675	0.582	0.505	0.750
Correlation	0.954	0.966	0.793	0.891	0.933	0.761	0.814	0.867	0.684
No. of Events	252	126	63	96	48	24	174	88	43
No. of Blooms	20	13	3	14	8	4	46	23	12
False Positives	4	3	1	1	0	1	12	6	1
False Negatives	8	5	1	1	0	1	17	7	5

* - excluded only four less frequently measured water quality variables

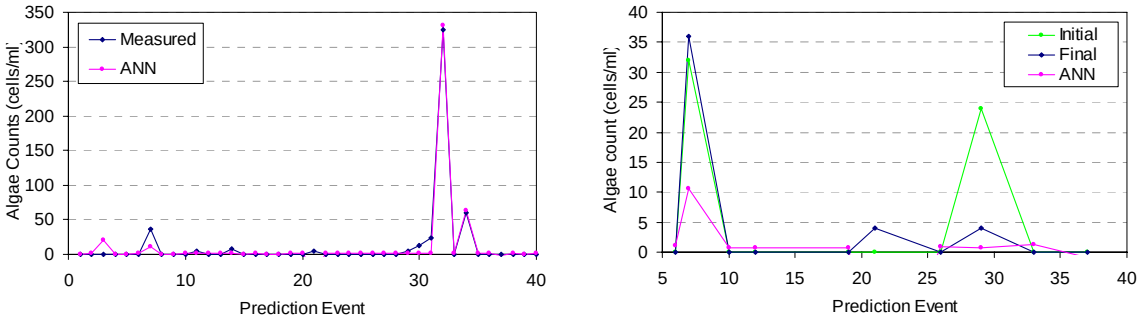
As done for the original models, the percentage accuracy for predicting relative increases or decreases from the initial to final measured counts were compared between the revised models with complete and reduced input sets, shown below in Table 42 below. Unlike the original models, the revised ANN models that excluded the five select input variables slightly outperformed the models that included them, with average correct forecast percentages of 77.5% and 75%, respectively. In a head to head comparison, it was a draw, with three of each outperforming the corresponding model. In terms of average correlation coefficient, the discrepancy between the revised models that excluded and included the select variables was larger, with computed values of 0.74 and 0.63, respectively.

Table 42. Percentage accuracy of the Revised ANN Models for predicting the three different algae classes at different modeling horizons in terms of predicting relative increases or decreases from the validation data sets' Initial to Final measured counts

	Models	No. of Events	Accuracy			
			Correct	%	Incorrect	%
Revised Models with All inputs	612-2wk Ahead Cyanobacteria Predictions	10	8	80	2	20
	100-1wk Ahead Chlorophyta Predictions	45	34	76	11	24
	101-2wk Ahead Chrysophyta Predictions	8	5	62	3	38
	100- 2wk Ahead Cyanobacteria Predictions	41	38	93	3	7
	101- 2wk Ahead Chlorophyta Predictions	8	5	62	3	0.38
	612-1wk Ahead Cyanobacteria Predictions	12	9	75	3	25
Revised Models with Fewer Inputs (exclude select water quality variables)	612-2wk Ahead Cyanobacteria Predictions	35	33	94	2	6
	100*-1wk Ahead Chlorophyta Predictions	67	52	78	15	22
	101-2wk Ahead Chrysophyta Predictions	16	9	56	7	44
	100*- 2wk Ahead Cyanobacteria Predictions	63	57	90	6	10
	101- 2wk Ahead Chlorophyta Predictions	16	12	75	4	25
	612-1wk Ahead Cyanobacteria Predictions	43	31	72	12	28

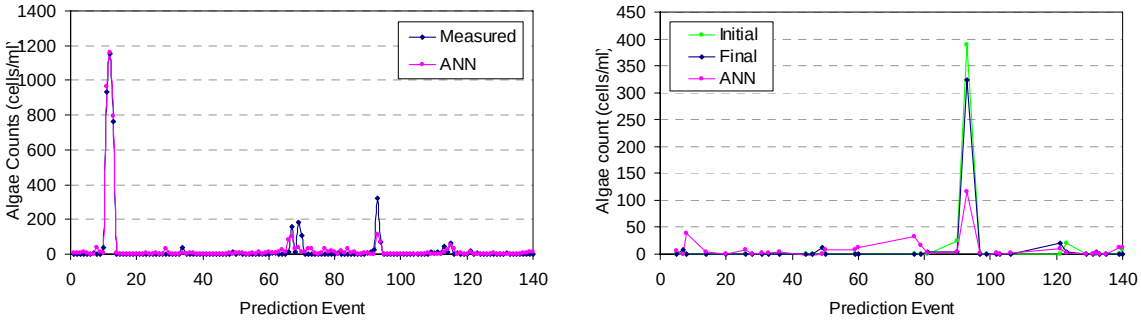
* - excluded only four less frequently measured water quality variables

Figures 81 through 92 depicting the various models are presented below for all representative test cases, showing both the overall and validation results.



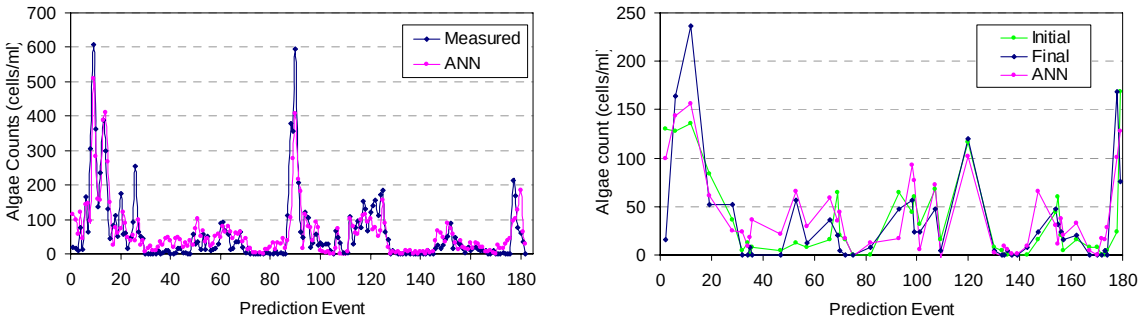
(a) (b)

Figure 81. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data sets at Station 612 (Revised Model using all inputs)



(a) (b)

Figure 82. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data set at Station 612 (Revised Model excluding five water quality inputs)



(a) (b)

Figure 83. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data sets at Station 100 (Revised Model using all inputs)

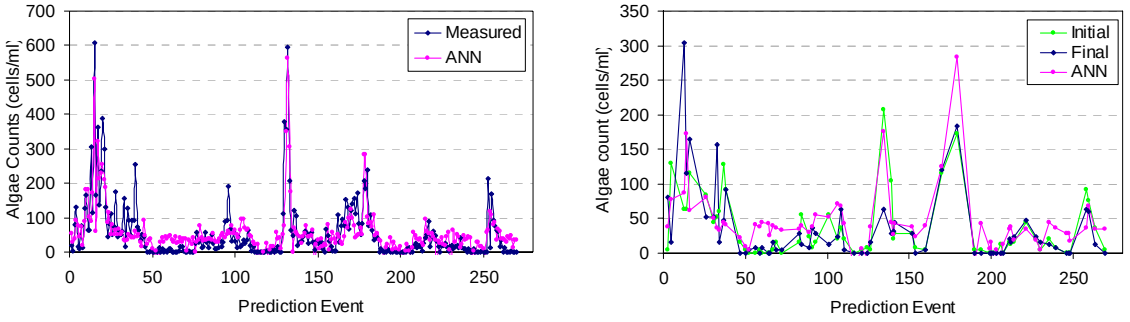


Figure 84. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 100 (Revised Model excluding four water quality inputs)

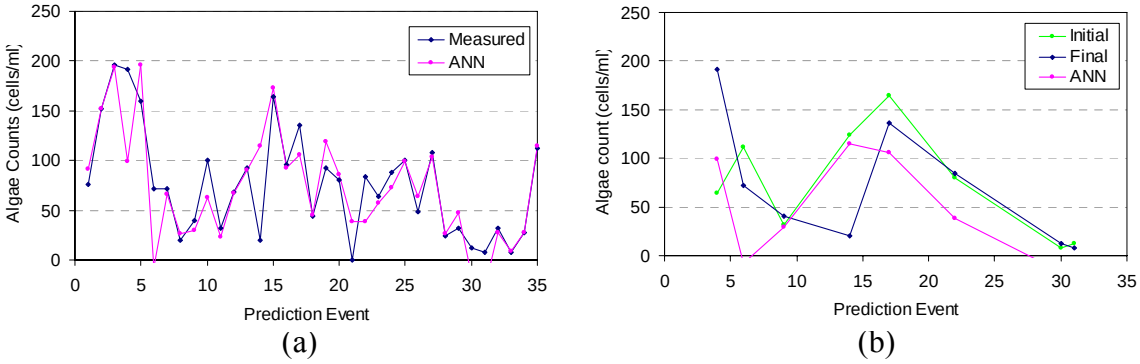


Figure 85. Time-series plots of measured Chrysophytes counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data sets at Station 101 (Revised Model using all inputs)

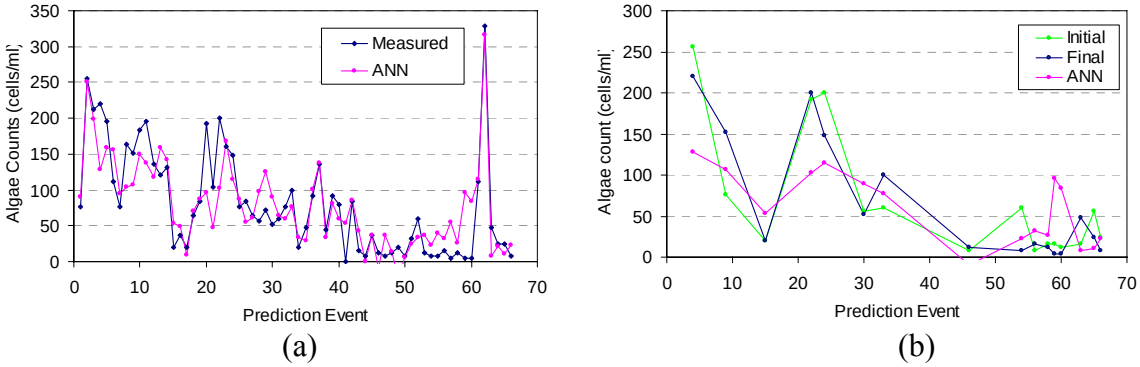
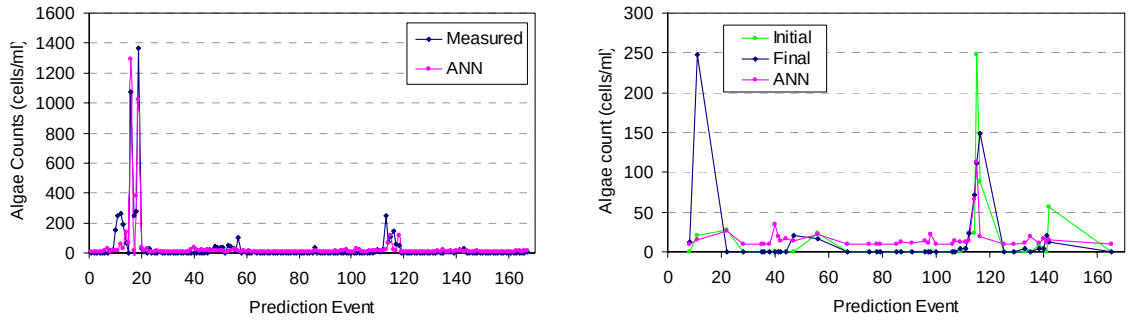
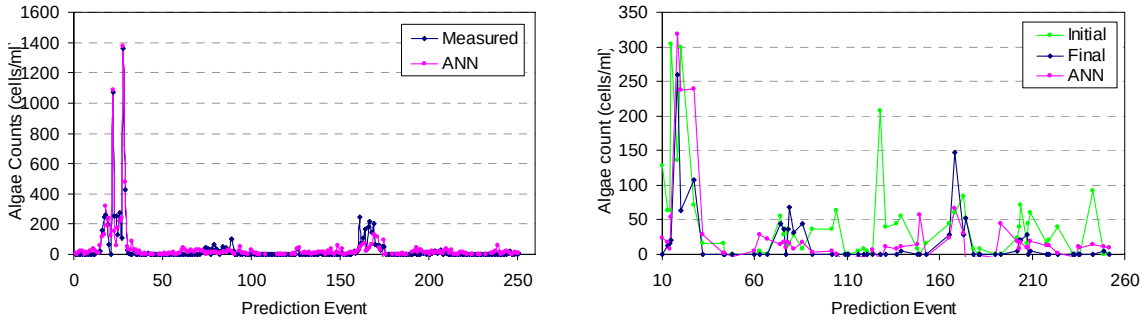


Figure 86. Time-series plots of measured Chrysophytes counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data sets at Station 101 (Revised Model excluding five water quality inputs)



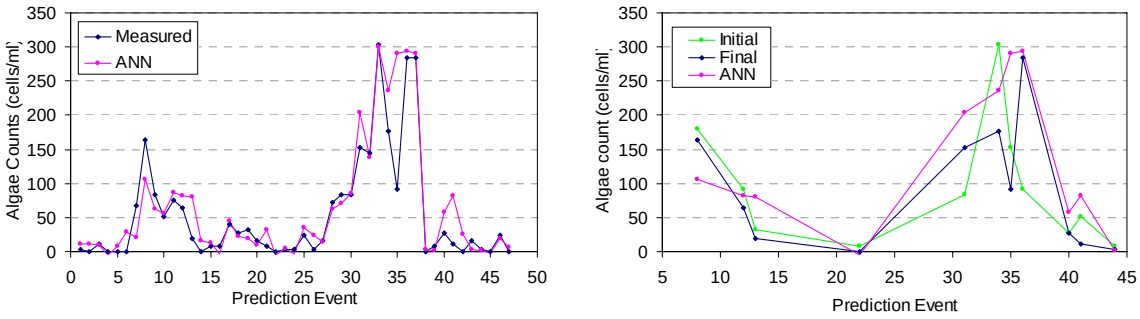
(a) (b)

Figure 87. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data sets at Station 100 (Revised Model using all inputs)



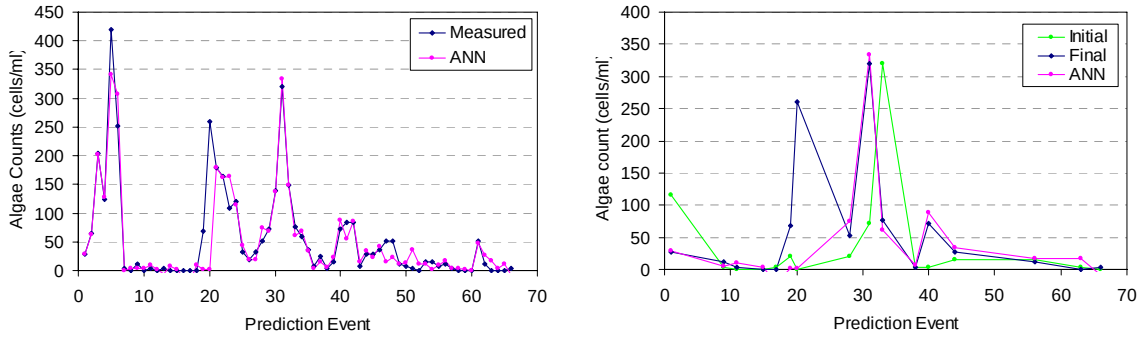
(a) (b)

Figure 88. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data sets at Station 100 (Revised Model excluding four water quality inputs)



(a) (b)

Figure 89. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data sets at Station 101 (Revised Model using all inputs)



(a) (b)

Figure 90. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data sets at Station 101 (Revised Model excluding five water quality inputs)

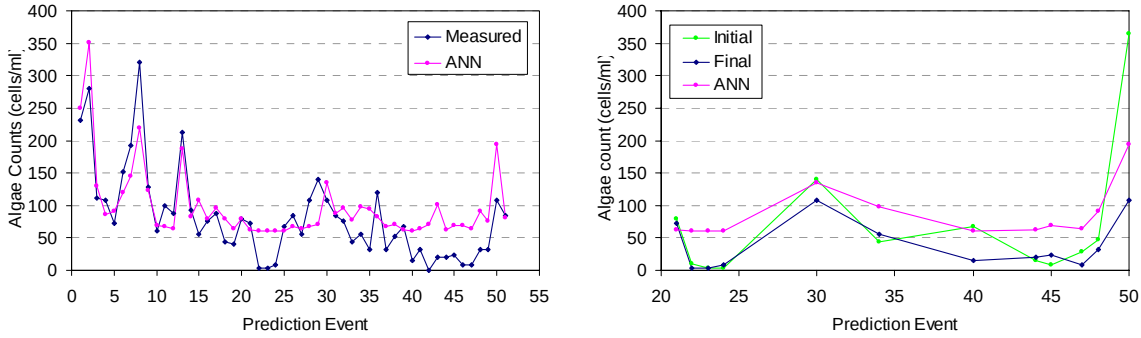
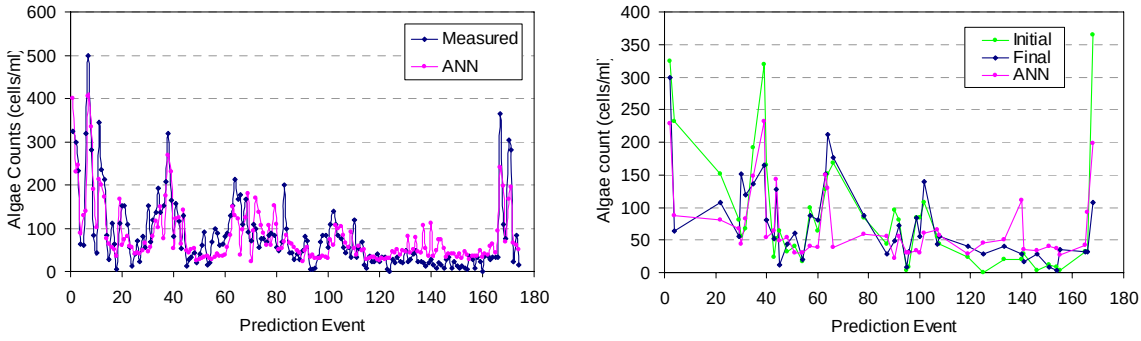


Figure 91. Time-series plots of measured Chrysophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data sets at Station 612 (Revised Model using all inputs)



(a) (b)

Figure 92. Time-series plots of measured Chrysophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data sets at Station 612 (Revised Model excluding five water quality inputs)

Compared to the original models, the validation performance of the revised models are similar, but perhaps slightly less accurate (coefficient value of 0.69 is slightly below the 0.72 value achieved with the former model type). There are some examples where the models during validation accurately predicted very large algal count changes, from bloom to non-bloom events, and vice-versa. For example, the revised model developed with the reduced inputs for predicting chlorophytes one-week ahead at Station 101 (Figure 90b) accurately predicted the evolution of a bloom to non-bloom condition of with a count below 50 for the first validation event. Around event 30, the model accurately forecasted the evolution from an initial non-bloom condition of 60 counts to a bloom condition of approximately 300 counts. Impressively, the model then predicts dissipation of this bloom condition to a non-bloom condition of approximately 60 counts. The model did fail to predict one significant bloom event during validation, but overall, it performed extremely well.

The model developed for predicting chrysophytes one-week ahead at Station 612 with reduced inputs also performed well during validation, accurately forecasting large increases and decreases in algal counts (Figure 92b). Other models performed reasonably well; for example, predicting chlorophytes one-week ahead at Station 100 with all inputs, where evolution from a non-bloom to bloom condition and other relatively large changes were accurately predicted (Figure 83b). For this same prediction problem, the revised model with reduced inputs did not perform quite as well, but did generally manage to predict relative changes (Figure 84b). At the same time, there were models that while correctly predicting a certain condition, may have underestimated or overestimated the magnitude by a fairly large number. The ANN model developed for predicting cyanobacteria two-weeks ahead at Station 612 using the reduced input set is an excellent example (Figure 82b). From an initial count close to 400, it accurately predicts a decrease, but significantly undershoots the measured final value of approximately 270 counts to around 110 counts (Figure 82b).

For the revised models, exclusion of the five select water quality variables at the benefit of larger data sets for ANN development appears to have slightly improved forecasting performance. It is likely that for some conditions, however, inclusion of at least some of these variables would be important, but this needs to be assessed with larger data sets. Again, ideally, the utility will continue to monitor these variables at some regular and perhaps increased frequency, and perhaps, in the future, their potential contribution under a variety of conditions can be more fully investigated.

8.2.4 Sensitivity Analyses Results – Revised Model Paradigm

Table 43 presents the sensitivity analysis result for two-week ahead cyanobacteria predictions with complete set at Station 612. Extraction from PVWC ranked first in terms of importance with relatively high ratio value of 3.7, followed by sky cover and initial cyanobacteria counts, both with ratio values over 2.9. Other variables that ranked near the top also had high ratio values exceeding 2.0. These variables include turbidity, heating degree days, prediction period's precipitation, and total phosphorus/orthophosphate. At the bottom of the ranking, a number of variables achieved low ratio values of less than 1.0, with the three lowest ranking variables being extraction from River B, length of day, and temperature.

Table 43. Sensitivity Analysis for Revised ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 612 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Reservoir A Extraction	3.662	1	Prediction Period's Precipitation Total	1.076	18
Sky Cover	2.915	2	Sulfate	1.075	19
Initial Cyanobacteria counts	2.901	3	Initial Chlorophytes counts	1.065	20
Turbidity	2.637	4	Total Hardness	1.051	21
Heating Degree Days	2.557	5	Chloride	1.050	22
Prediction Period's Lagged Precipitation Total	2.275	6	Wind Direction	1.036	23
T.Phosphorus/Orthophosphate	2.201	7	Alkalinity	1.014	24
Biochemical Oxygen Demand	1.732	8	Conductivity	0.999	25
Wind Speed	1.537	9	Total Organic Carbon	0.991	26
pH	1.381	10	Initial Chrysophytes	0.989	27
Initial Total Algal Counts	1.356	11	River A Streamflow	0.987	28
River A Extraction	1.320	12	Total Amorphous Materials	0.979	29
UV254	1.308	13	Pumping Station 1 Extraction	0.974	30
Color	1.287	14	Odor	0.949	31
Dissolved Oxygen	1.123	15	Temperature	0.947	32
Nitrite/Nitrate	1.112	16	Length of Day	0.935	33
Ammonia	1.078	17	River B Extraction	0.922	34

The overall sensitivity analysis result for the similar revised models but with reduced input variables is shown below in Table 44. As shown, color, turbidity and initial cyanobacteria counts were the three highest ranking, all with ratio values over 2.5. Extraction from Reservoir A and odor also rank high with ratio values around 2.0. At the other extreme, three variables achieved a ratio value less than 1.0, and were total amorphous materials, initial cholorophytes counts, and sky cover.

Table 44. Sensitivity Analysis for Revised ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 612 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Color	2.614	1	Initial Chrysophytes counts	1.091	15
Turbidity	2.601	2	Prediction Period's Lagged Precipitation Total	1.061	16
Initial Cyanobacteria counts	2.581	3	Alkalinity	1.058	17
Heating Degree Days	2.352	4	Prediction Period's Precipitation Total	1.058	18
Reservoir A Extraction	2.241	5	Length of Day	1.052	19
Odor	1.960	6	Total Hardness	1.045	20
River A Streamflow	1.452	7	Conductivity	1.038	21
Wind Direction	1.353	8	Chloride	1.038	22
River A Extraction	1.269	9	Initial Total Algal Counts	1.029	23
pH	1.199	10	Pumping Station 1 Extraction	1.002	24
Wind Speed	1.187	11	Ammonia	1.000	25
Dissolved Oxygen	1.180	12	UV254	1.000	26
River A Extraction	1.093	13	Sky Cover	0.991	27
Temperature	1.092	14			

Table 45 below presents the overall results of sensitivity analysis for one-week ahead prediction of chlorophytes at Station 100. As depicted by the table, initial chlorophytes counts and odor ranked first and second in importance, respectively, with ratio values above1.2, with total organic carbon ranking third with a ratio value of 1.1. All other variables exhibited relatively low ratio values, ranging from just less than 1.0 to just greater than 1.0. Turbidity, wind direction, and extraction from Pumping Station 1 were the three lowest ranking variables.

Table 45. Sensitivity Analysis for Revised ANN Model for One-week Ahead Predictions of Chlorophytes at Station 100 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chlorophytes counts	1.268	1	Wind Speed	1.010	18
Odor	1.212	2	Biochemical Oxygen Demand	1.009	19
Total Organic Carbon	1.104	3	Sulfate	1.008	20
Sky Cover	1.050	4	Dissolved Oxygen	1.007	21
Heating Degree Days	1.045	5	Prediction Period's Precipitation Total	1.006	22
T.Phosphorus/Orthophosphate	1.044	6	Chloride	1.006	23
UV-254	1.038	7	River A Extraction	1.005	24
Temperature	1.025	8	Prediction Period's Lagged Precipitation Total	1.003	25
Initial Total Algal Counts	1.024	9	Initial Chrysophytes counts	1.003	26
Ammonia	1.022	10	Reservoir A Extraction	1.003	27
Initial Cyanobacteria counts	1.020	11	pH	1.002	28
Length of Day	1.020	12	River A Streamflow	1.002	29
Total Hardness	1.018	13	Total Suspended Solids	1.001	30
Conductivity	1.016	14	Pumping Station 1 Extraction	1.000	31
River B Extraction	1.014	15	Wind Direction	0.999	32
Nitrite/Nitrate	1.013	16	Turbidity	0.984	33
Alkalinity	1.011	17			

Table 46 below presents the overall sensitivity analysis result for the ANN revised model with reduced inputs for one-week ahead chlorophytes predictions at Station 100. As shown, heating degree days ranked first in terms of importance, with a ratio value of 1.2, followed by initial chlorophytes counts, odo,r and initial total algal counts, all with ratio values around 1.1 For the lower ranking variables, several variables achieved ratio values less than 1.0, with the three lowest ranking being River A streamflow, wind direction, and UV254.

Table 46. Sensitivity Analysis for Revised ANN Model for One-week Ahead Predictions of Chlorophytes at Station 100 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Heating Degree Days	1.224	1	Total Organic Carbon	1.015	16
Initial Chlorophytes Counts	1.163	2	Prediction Period's Lagged Precipitation Total	1.014	17
Odor	1.118	3	Ammonia	1.012	18
Initial Total Algal Counts	1.100	4	Turbidity	1.011	19
River A Extractions	1.099	5	Alkalinity	1.006	20
Total Hardness	1.095	6	Reservoir A Extraction	1.005	21
Length of Day	1.080	7	Initial Cyanobacteria Counts	1.003	22
Biochemical Oxygen Demand	1.048	8	Total Suspended Solids	1.000	23
Initial Chrysophytes Counts	1.045	9	Sky Cover	1.000	24
Dissolved Oxygen	1.041	10	pH	1.000	25
Temperature	1.033	11	Wind Speed	0.998	26
Conductivity	1.024	12	Pump Station 1 Extraction	0.997	27
River B Extraction	1.022	13	UV254	0.995	28
Prediction Period's Precipitation Total	1.017	14	Wind Direction	0.994	29
Chloride	1.015	15	River A Streamflow	0.993	30

As for the two-week chryosphytes predictions at Station 101, Table 47 below presents the the sensitivity analysis results. As shown by the table, extraction from River B ranked as the most important predictor variable, with a ratio value of 1.7. Other variables that ranked near the top include odor, color, sulfate, and dissolved oxygen. At the other extreme, only three variables had ratio values less than 1.0, and they were extraction from Pumping Station 1, wind speed and biochemical oxygen demand.

Table 47. Sensitivity Analysis for Revised ANN Model for Two-week Ahead Predictions of Chrysophytes at Station 101 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
River B Extraction	1.725	1	T.Phosphorus/Orthophosphate	1.081	18
Odor	1.424	2	pH	1.078	19
Color	1.270	3	River A Streamflow	1.077	20
Sulfate	1.239	4	Sky Cover	1.075	21
Dissolved Oxygen	1.229	5	Prediction Period's Lagged Precipitation	1.063	22
Ammonia	1.218	6	Total Amorphous Materials	1.057	23
UV254	1.206	7	Length of Day	1.038	24
Heating Degree Days	1.200	8	Wind Direction	1.028	25
Conductivity	1.187	9	Nitrite/Nitrate	1.028	26
Prediction Period's Precipitation	1.166	10	Initial Cyanobacteria counts	1.025	27
Initial Total Algal Counts	1.162	11	Initial Chlorophytes counts	1.024	28
Turbidity	1.134	12	Reservoir A Extraction	1.012	29
Initial Chrysophytes	1.129	13	Total Organic Carbon	1.007	30
River A Extraction	1.123	14	Total Hardness	1.003	31
Alkalinity	1.121	15	Biochemical Oxygen Demand	1.000	32
Temperature	1.102	16	Wind Speed	0.978	33
Chloride	1.081	17	Pumping Station 1 Extraction	0.924	34

The result of sensitivity analysis for the similar revised model but with reduced input is shown in Table 48. Length of day ranked first with a ratio value of 1.45, followed by total algal counts and odor, both with ratio values over 1.2. At the bottom of the ranking, three variables achieved ratio values less than 1.0, indicating that they were the least important predictor variables in this particular prediction case. These variables include DO, wind direction, and initial chrysophytes counts.

Table 48. Sensitivity Analysis for Revised ANN Model for Two-week Ahead Predictions of Chrysophytes at Station 101 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Length of Day	1.450	1	pH	1.050	15
Initial Total Algal Count	1.279	2	Turbidity	1.047	16
Odor	1.220	3	Color	1.042	17
Wind Speed	1.172	4	Temperature	1.036	18
Prediction Period's Lagged Precipitation Total	1.152	5	Conductivity	1.016	19
Alkalinity	1.145	6	UV254	1.014	20
River B Extraction	1.134	7	River A Extraction	1.010	21
Initial Cyano Counts	1.092	8	Pump Station 1 Extraction	1.003	22
Chloride	1.089	9	Total Amorphous Materials	1.003	23
Ammonia	1.078	10	Initial Chlorophytes Counts	1.002	24
River A Streamflow	1.065	11	Total Hardness	1.002	25
Prediction Period's Precipitation Total	1.061	12	Initial Chrysophytes Counts	0.997	26
Heating Degree Days	1.052	13	Wind Direction	0.978	27
Sky Cover	1.051	14	Dissolved Oxygen	0.971	28

Table 49 presents the sensitivity analysis results for the revised ANN model two-week ahead cyanobacteria predictions with complete inputs at Station 100. River A extractions and turbidity were the two highest ranking variables in terms of importance, with ratio value of 1.94 and 1.9, respectively. Other high ranking variables achieved high ratio values of from 1.3 to over 1.6, including River B extractions, wind direction, and heating degree days. Several low variables attained ratio values of less than 1.0, including dissolved oxygen, UV254, and length of day.

Table 49. Sensitivity Analysis for Revised ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 100 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
River A Extraction	1.940	1	Initial Cyanobacteria counts	1.033	18
Turbidity	1.905	2	Total Hardness	1.019	19
River B Extraction	1.637	3	Prediction Period's Precipitation Total	1.016	20
Wind Direction	1.375	4	Alkalinity	1.012	21
Heating Degree Days	1.368	5	Sulfate	1.009	22
Odor	1.280	6	Initial Chrysophytes counts	1.008	23
Wind Speed	1.223	7	Biochemical Oxygen Demand	1.006	24
Initial Total Algal Counts	1.197	8	Chloride	1.003	25
Initial Chlorophytes counts	1.190	9	Pumping Station 1 Extraction	1.002	26
Total Suspended Solids	1.166	10	Reservoir A Extraction	0.998	27
Temperature	1.164	11	Conductivity	0.995	28
Ammonia	1.140	12	pH	0.995	29
River A Streamflow	1.076	13	Nitrite/Nitrate	0.995	30
Total Organic Carbon	1.066	14	Length of Day	0.991	31
Prediction Period Lagged Precipitation Total	1.053	15	UV-254	0.988	32
Sky Cover	1.047	16	Dissolved Oxygen	0.973	33
T.Phosphorus/Orthophosphate	1.040	17			

As for the similar model with the complete input set, input variables for the ANN model with the reduced inputs achieved relatively high ratio values, as presented in Table 50. It is interesting that both models have the same three highest ranking variables, though in different orders; turbidity and extractions from Rivers A and B. For the reduced input set model, the first two variables had ratio value just over 2.0 and the third had a ratio value of 1.92. Other high ranking variables include wind speed, wind direction, and dissolved oxygen. At the other extreme, five variables at the bottom of the ranking attained ratio values less than 1.0, with the three lowest ranking being initial chrysophytes counts, biochemical oxygen and total organic carbon.

Table 50. Sensitivity Analysis for Revised ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 100 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Turbidity	2.494	1	River A Streamflow	1.026	16
River A Extractions	2.101	2	Alkalinity	1.023	17
River B Extraction	1.920	3	Temperature	1.014	18
Wind Speed	1.429	4	Prediction Period's Precipitation Total	1.014	19
Wind Direction	1.407	5	pH	1.007	20
Dissolved Oxygen	1.386	6	Chloride	1.006	21
Odor	1.386	7	Total Hardness	1.002	22
Initial Total Algal Counts	1.248	8	Pump Station 1 Extraction	1.002	23
Heating Degree Days	1.205	9	Reservoir A Extraction	1.001	24
Initial Chlorophytes Counts	1.190	10	Total Suspended Solids	1.000	25
Ammonia	1.187	11	UV254	0.999	26
Initial Cyanobacteria Counts	1.155	12	Conductivity	0.997	27
Length of Day	1.070	13	Total Organic Carbon	0.997	28
Sky Cover	1.040	14	Biochemical Oxygen Demand	0.992	29
Prediction Period's Lagged Precipitation Total	1.035	15	Initial Chrysophytes Counts	0.987	30

For the ANN developed for two-week ahead chlorophytes predictions at Station 101, Table 51 below shows that initial cyanobacteria counts and Reservoir A extractions were the two highest ranking variables, with ratio values of 1.8 and 1.65, respectively. The next two highest ranking variables were initial chlorophytes counts and conductivity, both with ratio values just over 1.1. At the other extreme, the three lowest ranking variables were heating degree days, temperature, and ammonia. These three variables were among the ten low ranking variables that achieved ratio values less than 1.0.

Table 51. Sensitivity Analysis for Revised ANN Model for One-week Ahead Predictions of Chlorophytes at Station 101 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Cyanobacteria counts	1.810	1	Chloride	1.006	18
Reservoir A Extraction	1.650	2	Initial Chrysophytes counts	1.004	19
Initial Chlorophytes counts	1.189	3	pH	1.004	20
Conductivity	1.109	4	Turbidity	1.003	21
Wind Speed	1.094	5	Prediction Period's Lagged Precipitation Total	1.002	22
Prediction Period's Precipitation Total	1.060	6	Total Hardness	1.001	23
Initial Total Algal Counts	1.040	7	Sky Cover	1.001	24
UV254	1.037	8	Sulfate	0.996	25
Colro	1.026	9	Dissolved Oxygen	0.995	26
Biochemical Oxygen Demand	1.025	10	Total Amorphous Materials	0.994	27
River A Extraction	1.021	11	Pump Station 1 Extraction	0.992	28
Wind Direction	1.020	12	Odor	0.990	29
Total Organic Carbon	1.019	13	T.Phosphorus/Orthophosphate	0.984	30
River A Streamflow	1.018	14	Nitrite/Nitrate	0.977	31
River B Extraction	1.016	15	Ammonia	0.975	32
Length of Day	1.015	16	Temperature	0.968	33
Alkalinity	1.015	17	Heating Degree Days	0.954	34

Table 52 below presents the overall sensitivity analysis results for one-week ahead predictions of Chlorophytes a at Station 101 with reduced inputs. In terms of importance, alkalinity, wind direction, and length of day were the three highest ranking variables, all with ratio values just over 1.1. Other variables near the top of the ranking include pH, initial chlorophytes counts, and sky cover. At the bottom of the ranking, four variables achieved ratio values of less than 1.0, including extraction from Pumping Station 1, color, conductivity, and turbidity.

Table 52. Sensitivity Analysis for Revised ANN Model for One-week Ahead Predictions of Chlorophytes a at Station 101 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Alkalinity	1.188	1	Prediction Period's Precipitation Total	1.014	15
Wind Direction	1.184	2	Heating Degree Days	1.013	16
Length of Day	1.162	3	Initial Total Algal Count	1.012	17
pH	1.127	4	Initial Cyano Counts	1.012	18
Initial Chlorophytes Counts	1.120	5	Ammonia	1.009	19
Sky Cover	1.089	6	Chloride	1.007	20
Odor	1.077	7	Total Amorphous Materials	1.007	21
River A Streamflow	1.069	8	Prediction Period's Lagged Precipitation Total	1.005	22
Temperature	1.064	9	Initial Chrysophytes Counts	1.001	23
Wind Speed	1.048	10	River A Extraction	1.000	24
Total Hardness	1.036	11	Turbidity	0.999	25
UV254	1.022	12	Conductivity	0.998	26
Dissolved Oxygen	1.019	13	Color	0.997	27
River B Extraction	1.018	14	Pumping Station 1 Extraction	0.997	28

As for the one-week chryosphytes predictions at Station 612, Table 53 below presents the overall results for the sensitivity analysis. As shown, the prediction period's lagged precipitation total, extraction from River A, and temperature were the first three top ranking variables. The first had a ratio value of 1.2 while the other two had ratio values just over 1.1. A total of eleven variables achieved ratio values of less than 1.0, and the three lowest ranking variables were total amorphous material, pH, and heating degree days.

Table 53. Sensitivity Analysis for Revised ANN Model for One-week Ahead Predictions of Chrysophytes at Station 612 with complete input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Prediction Period's Lagged Precipitation Total	1.201	1	Length of Day	1.005	18
River A Extractions	1.124	2	Prediction Period's Precipitation Total	1.002	19
Temperature	1.109	3	Biochemical Oxygen Demand	1.001	20
Odor	1.084	4	T.Phosphorus/Orthophosphate	1.001	21
Initial Chlorophytes counts	1.082	5	Conductivity	1.000	22
UV254	1.037	6	Sulfate	1.000	23
Wind Speed	1.027	7	Nitrate/Nitrite	0.999	24
Pumping Station Extraction	1.027	8	Turbidity	0.999	25
River A Streamflow	1.026	9	Initial Cyanobacterial counts	0.995	26
Chloride	1.022	10	Ammonia	0.994	27
Sky Cover	1.018	11	River B Extractions	0.992	28
Total Organic Carbon	1.016	12	Reservoir A Extractions	0.992	29
Color	1.016	13	Total Algal Counts	0.987	30
Initial Chrysophytes counts	1.014	14	Dissolved Oxygen	0.986	31
Total Hardness	1.013	15	Heating Degree Days	0.985	32
Alkalinity	1.012	16	pH	0.984	33
Wind Direction	1.008	17	Total Amorphous Materials	0.970	34

The sensitivity analysis result for revised ANN model for one-week ahead predictions of chrysophytes at Station 612 with reduced input set is presented in Table 54. Initial chrysophytes counts ranked first with a ratio value of 1.28, followed by length of day and odor, both with ratio value just over 1.1. At the other extreme, only sky cover achieved a ratio value of less than 1.0. Other low ranking variables were wind direction, total algal counts, and ammonia.

Table 54. Sensitivity Analysis for Revised ANN Model for One-week Ahead Predictions of Chrysophytes at Station 612 with reduced input set

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chrysophytes counts	1.282	1	Chloride	1.014	16
Length of Day	1.162	2	Turbidity	1.008	17
Odor	1.130	3	Pumping Station 1	1.007	18
River A Extractions	1.075	4	Dissolved Oxygen	1.006	19
UV254	1.058	5	Wind Speed	1.006	20
Color	1.054	6	Total Hardness	1.006	21
Temperature	1.037	7	Conductivity	1.005	22
Total Amorphous Materials	1.034	8	pH	1.005	23
Prediction Period's Precipitation Total	1.030	9	Reservoir A Extractions	1.003	24
River B Extractions	1.027	10	Prediction Period's Lagged Precipitation	1.002	25
River A Streamflow	1.023	11	Ammonia	1.002	26
Alkalinity	1.016	12	Total Algal Counts	1.001	27
Initial Cyanobacteria counts	1.015	13	Wind Direction	1.000	28
Initial Chlorophytes counts	1.015	14	Sky Cover	0.999	29
Heating Degree Days	1.015	15			

8.2.5 ANN Models without Water Extraction Input variables

A concern that emerged during the study was whether the ANN models were at least partly biased by operational input variables of a more correlative relationship with algal counts, rather than causal. This correlative relationship or bias would be problematic for a real-time forecasting method that is assumed to capture the mechanistic processes that govern algal population dynamics. The variables in question, water extractions from the rivers and reservoir, often reflect operational decisions made in response to formation and/or dissipation of bloom events, and except for Station 100, a mixing point, may not have significant causative effects on algal counts at the water quality river stations. As the example given previously, during an algal bloom event on River A, the utility may discontinue use of this source and make up the deficit with reservoir extraction. Because the reservoir extraction may have relatively little effect on the algae counts on River A,

unlike a weather or chemical variable, but can be correlated with this variable, there is the possibility that the ANN model is "keying" off this relationship. A simple relationship or correlation that the ANN might learn is: "If River A extraction drops to 0, algal counts are relatively high at Station 612, and if River A extraction is high, algal counts are relatively low at Station 612."

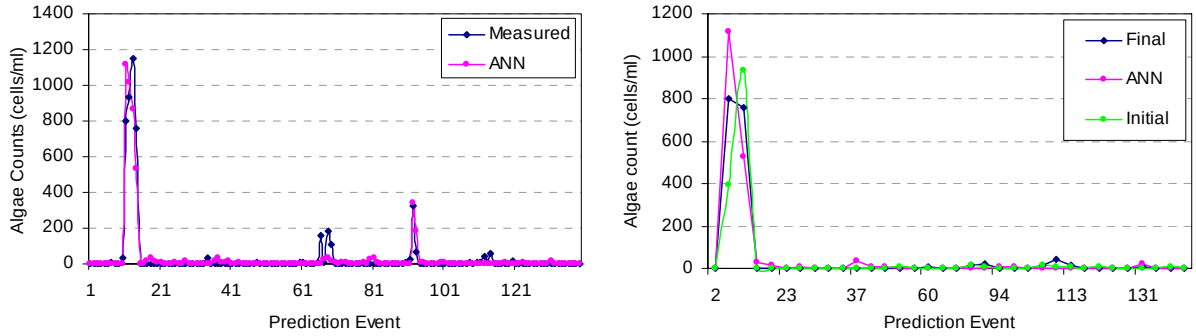
Ideally, the real-time forecasting method should be based upon variables that are independent of any operational actions that the utility may implement in response to measured algal counts, such as water extraction from the reservoir. It should be qualified, however, that river extractions may induce some causal effect on river algal populations. For example, a lower extraction from River A during a bloom event is expected to produce different hydrodynamic and chemical conditions on the river that may effect populations, such as higher turbulence and sediment suspension (i.e. less sunlight penetration). Some overlap between causal and correlative is naturally present for other variables; for example, some water quality parameters were observed to correlate to some degree with measured algal counts. The difference is that these variables are not direct human responses to algal counts, although the human responses manifested in water extraction rates may also influence these variables (e.g. lower river extraction may produce lower dissolved oxygen levels).

To assess the possible influence of these so-called correlative extraction variables on model performance, they were eliminated as inputs for select modeling cases. In consultant with NJDEP and PVWC, the four following forecasting cases were selected, using the original data sets (inputs generally measured at beginning of prediction period); one-week ahead chlorophytes counts predictions for both Stations 100 and 101; two-week ahead cyanobacteria predictions for Stations 100 and 612.

The statistical summaries for these four models are shown in Table 55, with accompanying figures 93 though 96 that depict measured versus predicted values for both the overall and validation data sets.

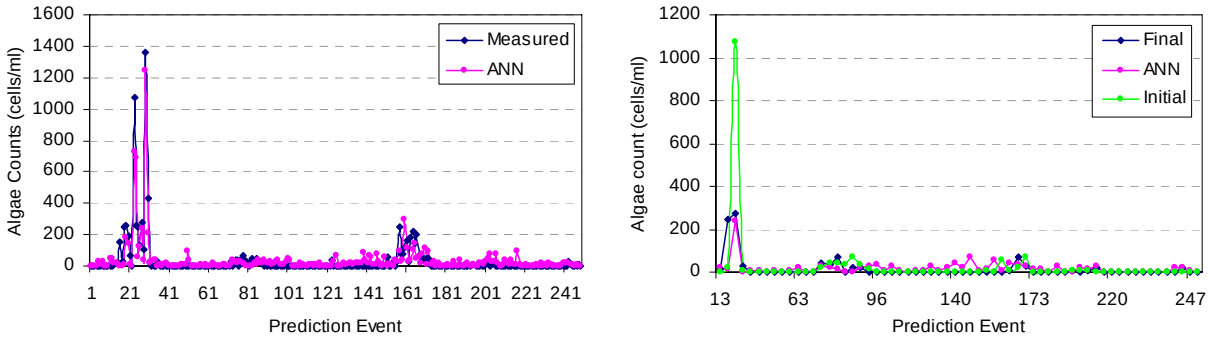
Table 55. Comparison of statistical performances of ANN Models for predicting the two different algae classes at different modeling horizons excluding water extraction inputs

	Station 612 – Two week Ahead Cyanobacteria Predictions			Station 100 - Two week Ahead Cyanobacteria Predictions		
	Overall	Training	Validation	Overall	Training	Validation
Data Mean	34.849	27.380	49.294	30.297	43.360	14.710
Data S.D.	157.155	118.630	182.950	121.847	162.515	47.664
Error Mean	-2.985	-2.476	2.218	1.913	1.255	2.262
Error S.D.	48.151	33.808	68.044	55.346	65.822	35.942
Abs E. Mean	16.479	14.335	24.108	25.473	28.838	17.157
S.D. Ratio	0.306	0.285	0.372	0.454	0.405	0.754
Correlation	0.952	0.964	0.945	0.891	0.915	0.657
No. of Events	139	71	34	249	125	62
No. of Blooms	8	5	2	19	13	2
False Positives	1	1	0	1	0	0
False Negatives	3	3	0	7	3	1
	Station 100 – One week Ahead Chlorophytes Predictions			Station 101 - One week Ahead Chlorophytes Predictions		
	Overall	Training	Validation	Overall	Training	Validation
Data Mean	49.455	51.552	51.848	44.367	55.200	37.185
Data S.D.	77.028	75.136	73.359	75.159	75.743	64.165
Error Mean	5.518	-0.780	13.493	3.651	-0.942	7.800
Error S.D.	50.863	44.149	62.370	47.343	42.079	40.628
Abs E. Mean	32.425	29.475	39.456	27.709	25.155	27.723
S.D. Ratio	0.660	0.588	0.850	0.630	0.556	0.633
Correlation	0.783	0.811	0.788	0.778	0.833	0.775
No. of Events	266	134	66	109	55	27
No. of Blooms	39	23	10	15	11	2
False Positives	21	7	7	4	2	1
False Negatives	17	8	5	1	1	0



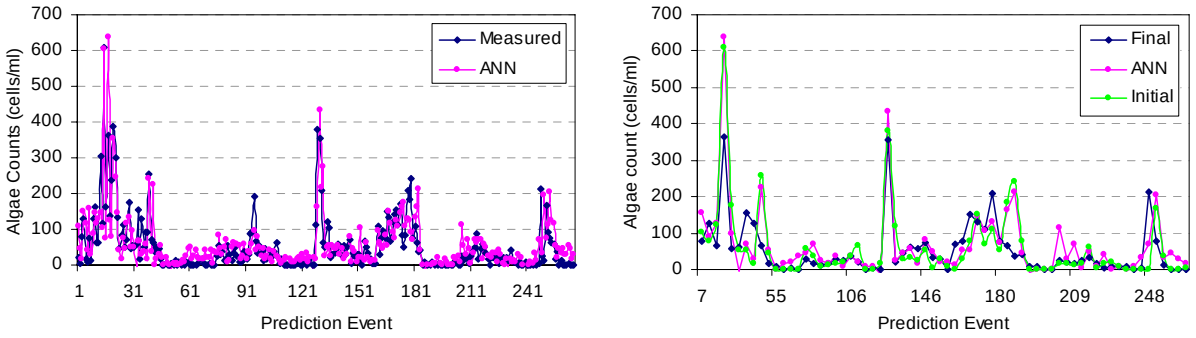
(a) **(b)**

Figure 93. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data set at Station 612



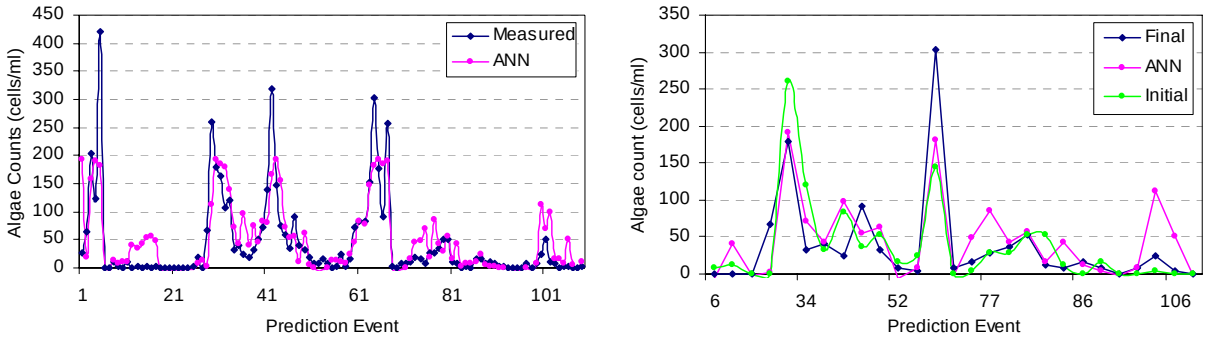
(a) **(b)**

Figure 94. Time-series plots of measured Cyanobacteria counts against ANN Two-week Ahead predicted values for (a) complete and (b) validations data set at Station 100



(a) **(b)**

Figure 95. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 100



(a) **(b)**

Figure 96. Time-series plots of measured Chlorophytes counts against ANN One-week Ahead predicted values for (a) complete and (b) validations data set at Station 101

As expected, the models performed best for Stations 101 and 612, given that mixing of various source waters does not occur, and extractions are not assumed to be strongly causative mechanisms on algal populations. In comparison, there was relatively larger diminishment in forecasting performance for Station 100, where mixing of various source waters does occur, and hence extractions are considered more causal. For the two-week ahead forecasting problem for cyanobacteria at Station 612, the ANN model correctly predicted blooms during both validation events, with no false positives. Similarly for predicting chlorophytes one-week ahead at Station 101, the model correctly forecasted both blooms during validation, with just one false positive. For both stations, the statistical measure for the models that excluded extractions compare favorably with the models that included extractions, with slightly lower correlation coefficients for the former (0.945 versus 0.985 for Station 612, and 0.775 versus 0.788 for Station 101).

For predicting chlorophytes one-week ahead at Station 100, the ANN model clearly appears to have keyed off the initial chlorophytes counts, which as presented further below, had the highest sensitivity ratio for the model. The two-week ahead cyanobacteria forecasting model for Station 100 performed better, and correctly forecasted one of the two validation algal bloom events. However, this model had a relatively large discrepancy in its correlation coefficient between training (0.915) and validation (0.657).

In contrast, the model for this forecasting problem that included extraction as an input achieved a correlation coefficient of 0.847 during validation.

The sensitivity analyses for the four models are presented below in Tables 56 through 59.

Table 56. Sensitivity Analysis for ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 612 without water extraction inputs

Variable	Ratio	Rank	Variable	Ratio	Rank
Turbidity	2.840	1	Conductivity	1.022	14
Initial Cyanobacteria counts	1.802	2	Sky Cover	1.011	15
Total Algal counts	1.521	3	Dissolved Oxygen	1.010	16
Color	1.331	4	Alkalinity	1.000	17
Odor	1.228	5	Chloride	0.999	18
Wind Direction	1.186	6	Prediction Period's Lagged Precipitation Total	0.994	19
River A Streamflow	1.129	7	Prediction Period's Precipitation Total	0.991	20
Total Amorphous Material	1.120	8	Initial Chlorophytes counts	0.990	21
Heating Degree Days	1.104	9	pH	0.982	22
Wind Speed	1.089	10	Total Hardness	0.982	23
Ammonia	1.063	11	Length of Day	0.980	24
UV254	1.034	12	Temperature	0.972	25
Initial Chrysophytes counts	1.029	13			

Table 57. Sensitivity Analysis for ANN Model for Two-week Ahead Predictions of Cyanobacteria at Station 100 without water extraction inputs

Variable	Ratio	Rank	Variable	Ratio	Rank
Odor	1.752	1	Temperature	1.026	14
Total Algal Counts	1.268	2	UV254	1.021	15
Initial Cyanobacteria counts	1.238	3	Alkalinity	1.021	16
Sky Cover	1.200	4	Chloride	1.020	17
Wind Direction	1.188	5	Initial Chrysophytes counts	1.009	18
Total Organic Carbon	1.182	6	Initial Chlorophytes counts	1.007	19
Biochemical Oxygen Demand	1.150	7	Total Hardness	1.004	20
Dissolved Oxygen	1.141	8	Wind Speed	1.003	21
Conductivity	1.135	9	Prediction Period's Precipitation Total	1.001	22
River A Streamflow	1.124	10	Length of Day	0.999	23
Ammonia	1.078	11	Total Suspended Solids	0.997	24
Prediction Period's Lagged Precipitation Total	1.029	12	Turbidity	0.997	25
pH	1.028	13	Heating Degree Days	0.997	26

Table 58. Sensitivity Analysis for ANN Model for Two-week Ahead Predictions of Chlorophytes couts at Station 100 without water extraction inputs

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chlorophytes counts	1.261	1	UV254	1.002	14
Odor	1.216	2	Total Suspended Solids	1.001	15
Chloride	1.200	3	Turbidity	1.000	16
Conductivity	1.152	4	Length of Day	0.995	17
Biochemical Oxygen Demand	1.065	5	Prediction Period's Lagged Precipitation Total	0.995	18
River A Streamflow	1.044	6	pH	0.994	19
Sky Cover	1.039	7	Total Hardness	0.991	20
Dissolved Oxygen	1.016	8	Temperature	0.991	21
Wind Speed	1.015	9	Alkalinity	0.990	22
Ammonia	1.012	10	Prediction Period's Precipitation Total	0.989	23
Total Organic Carbon	1.008	11	Heating Degree Days	0.987	24
Wind Direction	1.006	12	Initial Chrysophytes counts	0.985	25
Initial Cyanobacteria counts	1.006	13	Total Algal Counts	0.972	26

Table 59. Sensitivity Analysis for ANN Model for Two-week Ahead Predictions of Chlorophytes couts at Station 101 without water extraction inputs

Variable	Ratio	Rank	Variable	Ratio	Rank
Initial Chlorophytes counts	1.183	1	Initial Cyanobacteria counts	0.999	14
Prediction Period's Precipitation Total	1.092	2	Total Hardness	0.998	15
Total Algal counts	1.084	3	Length of Day	0.997	16
Ammonia	1.037	4	River A Streamflow	0.997	17
pH	1.026	5	Wind Direction	0.997	18
Alkalinity	1.021	6	Initial Chrysophytes counts	0.996	19
Prediction Period's Lagged Precipitation Total	1.019	7	Heating Degree Days	0.995	20
Color	1.010	8	Total Amorphous Material	0.994	21
Turbidity	1.004	9	Sky Cover	0.991	22
Temperature	1.003	10	Dissolved Oxygen	0.989	23
Wind Speed	1.002	11	Chloride	0.988	24
UV254	1.000	12	Conductivity	0.985	25
Odor	1.000	13			

While there is some similarity, there is some marked inconsistency between the earlier comparable ANN models and these. For example, length of day ranked consistently high for the previous models, but for this case, ranked near the bottom, in all cases having a ratio value < 1.0. Again, given the relatively few events available for ANN development, the sensitivity analyses must be regarded with skepticism.

Excluding volumetric water extractions considered to be correlative variables did not significantly decrease forecasting performance for the two models developed for Stations 101 and 612. In contrast, overall there was more diminishment in performance for the two models developed for Station 100. This supports the conclusion that the previously developed ANN models appear to be learning the underlying system behavior, and for Stations 101 and 612, are not overly biased in correlating extractions with counts. The results also agree with physical intuition, and indeed was anticipated that extraction variables would be more relevant for Station 100, where the mixing of various source waters occurs. Thus, despite the limited data sets, the ANN models appear to have

learned the general underlying behavior in this complex system for the range of conditions measured. This is further supported by the classification modeling results presented in the following section.

8.2.6 ANN Classification Models

As an alternative to developing ANN models that explicitly predict final measured algal counts, RBF nets were developed to predict the pre-specified bins or classification ranges in which the final measured algal counts fall. For this exercise, the following four bins or classification ranges were selected: 0 to 10 counts, 11 to 50 counts, 51 to 200 counts, and 201 and above counts. The original model input sets were used and included both the complete and reduced parameter sets, and as only Stations 101 and 612 were used, water volume extraction variables were excluded as inputs. Station 101 was used for predicting chlorophytes bins one-week ahead and chrysophytes bins two-weeks ahead. Station 612 was selected for predicting chrysophytes bins one-week ahead and chlorophytes bins two-weeks ahead. The four classification bins and stations/algae classes and eight modeling exercise were selected in consultation with NJDEP and PVWC.

Below in Tables 60 and 61 are the tabulated results for two representative cases, comparing the measured versus class predicted values for the entire data set for the models that included and excluded the select water quality variables. Overall, the ANNs did surprisingly well in accurately classifying algal counts into their classification bins. The other results can be found in Appendix B-5.

Table 60. Comparison of measured Chrysophytes counts against ANN One-week Ahead Class-predicted values at Station 101 without Chemical Inputs

Event	Measured	Bin	Accuracy	Event	Measured	Bin	Accuracy
1*	76	51-200	Correct	34	20	11-50	Correct
2	256	>200	Correct	35	48	11-50	Correct
3	212	>200	Correct	36	92	51-200	Correct
4	220	>200	Correct	37	136	51-200	Correct
5*	196	51-200	Correct	38	44	11-50	Correct
6	112	51-200	Correct	39	92	51-200	Correct
7*	76	51-200	Correct	40	80	51-200	Correct
8	164	51-200	Correct	41	0	0-10	Correct
9	152	51-200	Correct	42	84	51-200	Correct
10	184	51-200	Correct	43*	16	0-10	Incorrect
11	196	51-200	Correct	44	8	0-10	Correct
12	136	51-200	Correct	45	36	11-50	Correct
13	120	51-200	Correct	46	12	11-50	Correct
14	132	51-200	Correct	47	8	0-10	Correct
15	20	11-50	Correct	48	12	11-50	Correct
16	36	11-50	Correct	49	20	11-50	Correct
17	20	11-50	Correct	50	8	0-10	Correct
18	64	51-200	Correct	51	32	11-50	Correct
19	84	51-200	Correct	52	60	51-200	Correct
20	192	51-200	Correct	53	12	11-50	Correct
21*	104	51-200	Correct	54*	8	11-50	Incorrect
22	200	>200	Correct	55	8	0-10	Correct
23	160	51-200	Correct	56	16	11-50	Correct
24	148	51-200	Correct	57	4	0-10	Correct
25	76	51-200	Correct	58	12	11-50	Correct
26	84	51-200	Correct	59	4	0-10	Correct
27	64	51-200	Correct	60	4	0-10	Correct
28	56	51-200	Correct	61	112	51-200	Correct
29	72	51-200	Correct	62	328	>200	Correct
30	52	51-200	Correct	63	48	11-50	Correct
31	60	51-200	Correct	64	24	11-50	Correct
32	76	51-200	Correct	65	24	11-50	Correct
33	100	51-200	Correct	66	8	0-10	Correct

* - validation event

Table 61. Comparison of measured Chrysophytes counts against ANN One-week Ahead Class-predicted values at Station 101 with Chemical Inputs

Event	Measured	Bin	Accuracy
1	76	51-200	Correct
2	152	51-200	Correct
3	196	51-200	Correct
4	192	51-200	Correct
5	160	51-200	Correct
6	72	51-200	Correct
7	72	51-200	Correct
8	20	11-50	Correct
9	40	11-50	Correct
10	100	51-200	Correct
11	32	11-50	Correct
12	68	51-200	Correct
13	92	51-200	Correct
14	20	11-50	Correct
15	164	51-200	Correct
16	96	51-200	Correct
17	136	51-200	Correct
18	44	11-50	Correct
19	92	51-200	Correct
20	80	51-200	Correct
21	0	0-10	Correct
22	84	51-200	Correct
23	64	51-200	Correct
24	88	51-200	Correct
25	100	51-200	Correct
26	48	11-50	Correct
27	108	51-200	Correct
28	24	11-50	Correct
29	32	11-50	Correct
30	12	11-50	Correct
31	8	11-50	Correct
32	32	11-50	Correct
33	8	0-10	Correct
34	28	11-50	Correct
35	112	51-200	Correct

As the two representative cases demonstrate, the models correctly predicted the classification range or bin for the final measured algal counts with high accuracy. In the few cases that the classification was incorrect, it always occurred with an adjacent class. For example, for event 54 (a validation) shown in Table 60 above, the ANN model predicted the final measured algae count of 8 would fall into the bin or classification range of 11 to 50 counts. Given the impreciseness of the measured algal counts, where previous researchers estimate a measurement error of approximately ±20% (excludes spatial variability), the classification performance is high.

Table 62 below summarizes the overall performance of the different classification models by percentage classification accuracy, with the number of events provided, and the inclusion or exclusion of the five select water quality variables indicated.

Table 62. Overall Percentage Accurary for the Eight RBF Nets

Station	Organism	Select Chemical Inputs	Total Number Correct	Total Number Events	Percentage Classification Accuracy (%)
101	Chrysophyta	Included	35	35	100.0
101	Chrysophyta	Excluded	64	66	97.0
101	Chlorophyta	Included	47	47	100.0
101	Chlorophyta	Excluded	95	96	99.0
612	Chrysophyta	Included	51	51	100.0
612	Chrysophyta	Excluded	144	174	82.8
612	Chlorophyta	Included	38	44	86.4
612	Chlorophyta	Excluded	125	140	89.3

Note: Original Input Structure Used for All Models, and extraction variables excluded.

On average, the models that included the five select chemical inputs slightly outperformed those that did not, with correct classification percentages of 96 and 92 percent, respectively. However, the models that excluded these five water quality variables had approximately three times the number of events, and hence had more events that bordered on two classification bins. Three of the eight models, which included all

chemical inputs, achieved 100 percent classification accuracy. The lowest performing net correctly classified 83 percent of the events, and all incorrect classifications for all models occurred within an adjacent bin (e.g. measured count of 8, predicted range of 11 to 50).

Finally, as described in the ANN section, RBF models excluded variables that reduced classification accuracy during learning. Table 63 below presents the forecasting problems and corresponding variables that were excluded by the various RBF nets.

Table 63. List of Input variables excluded by various RBF nets during Training

Station	Organism	Prediction Horizon	Select Chemical Inputs	Variables Excluded by the Nets
101	Chrysophyta	Two-week Ahead	Included	Conductivity
101	Chrysophyta	Two-week Ahead	Excluded	Ammonia
101	Chlorophyta	One-week Ahead	Included	Total Phosphorus/ Orthophosphate, Initial Chlorophytes counts, Odor, Total Amorphous Materials
101	Chlorophyta	One-Week Ahead	Excluded	Chloride, Odor, Total Amorphous Materials
612	Chrysophyta	One-Week Ahead	Included	pH, Heating Degree Days, Length of Day, Total Amorphous Material, Initial Cyanobacteria counts, Initial Chlorophytes counts
612	Chrysophyta	One-Week Ahead	Excluded	All variables ***except*** Alkalinity, Conductivity, Wind Speed, Length of Day, Total Algal Counts, Initial Chrysophytes counts, Initial Chlorophytes counts
612	Chlorophyta	Two-week Ahead	Included	All variables ***except*** Dissolved Oxygen, UV254, Heating Degree Days, Sky Cover, Length of Day
612	Chlorophyta	Two-week Ahead	Excluded	No variable excluded

Again, because of the sparse data sets, it is difficult to draw conclusions with high confidence. It is, however, interesting to note that three of the four ANNs that initially included the select water quality parameters generally eliminated these as inputs. In addition, length to day was included for seven of eight models, including the two models that eliminated most variables, and two nets for Station 612 eliminated most inputs.

8.2.7 Linear Models versus ANNs

In order to further assess the performance of the ANN technology, as was done in the preliminary research for Swimming River, linear models were also applied to the PVWC site. Although transformation of the data, for example, logarithmically, may have improved LM performance, the main objective of this analysis was to provide an objective assessment of ANN performance with an alternative model paradigm using the identical data sets. The linear models used in this research are analogous to traditional linear regression equations, but utilize a superior algorithm to find the "best fitting" coefficients. As with the ANN modeling, the software Statistica was used to perform this work.

Linear models were developed to predict the three algae types at the three stations used as representative models. Of the twelve prediction problems, the ANN models provided a lower MAE error eleven times, often significantly smaller. The only time the LM model provided lower MAE's was for predicting chrysophytes two week-ahead at Station 101 with reduced input set measured primarily at the beginning of the prediction period.

Tables 64 and 65 provide a statistical comparison between the two model types for the two modeling approaches and data sets as presented by the three representative models.

With respect to a model's ability to accurately predict the incidence and magnitude of algal blooms while avoiding false positives, the discrepancy between the ANNs and the LMs is often quite significant. The ANNs were generally able to reproduce the rising and falling algae count levels, predicting the incidence and magnitude of algal blooms, while minimizing false positives. In contrast, the LMs sometimes fail in both measures, particularly for cyanophyta.

Table 64. Statistical Measure Comparison between ANN and LM for Original Models with two types of input variables

		Complete Input Set		Reduced Input Set	
		ANN	LM	ANN	LM
Station 612: Two-week Ahead Cyanobacteria Prediction	Mean Absolute Error	25.607	110.536	12.816	44.179
	Correlation Coefficient	0.939	0.592	0.979	0.823
Station 100*: One-week Ahead Chlorophytes Prediction	Mean Absolute Error	33.485	67.271	33.569	34.755
	Correlation Coefficient	0.612	0.139	0.772	0.719
Station 101: Two-week Ahead Chrysophytes Prediction	Mean Absolute Error	32.075	38.995	37.639	32.792
	Correlation Coefficient	0.638	0.404	0.590	0.662

Table 65. Statistical Measure Comparison between ANN and LM for Revised Models with two types of input variables

		Complete Input Set		Reduced Input Set	
		ANN	LM	ANN	LM
Station 612: Two-week Ahead Cyanobacteria Prediction	Mean Absolute Error	4.019	9.622	10.973	35.487
	Correlation Coefficient	0.987	0.932	0.983	0.863
Station 100*: One-week Ahead Chlorophytes Prediction	Mean Absolute Error	33.123	53.680	34.560	40.959
	Correlation Coefficient	0.839	0.596	0.805	0.702
Station 101: Two-week Ahead Chrysophytes Prediction	Mean Absolute Error	19.655	30.568	26.909	49.810
	Correlation Coefficient	0.824	0.491	0.865	0.361

* - excluded only four less frequently measured water quality variables for the reduced input set

Figures 97 and 98 compare the ANN and LM performance for predicting cyanophyta counts at station 612 two-week ahead and chlorophyta counts at Station 101 one-week ahead prediction periods, respectively. The figure show the original and revised modeling approach using the two type of data set (i.e. larger data set with lesser input variables and smaller data set with more input variables). Similar figures for the remaining prediction events (i.e. combinations of algae types, stations, and prediction periods) can be found in Appendix B-6.

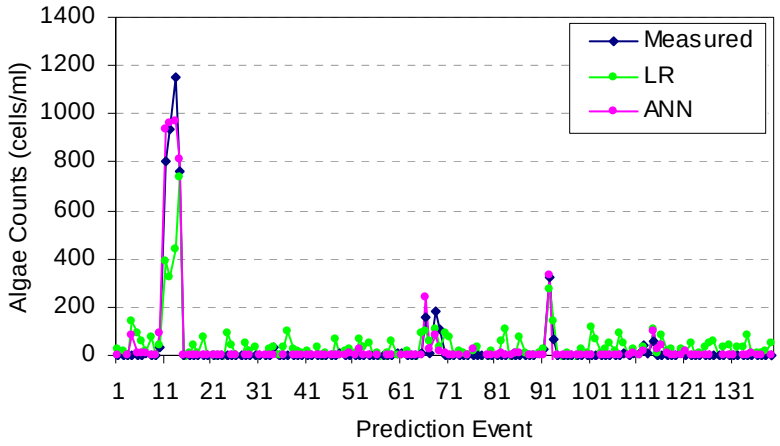


Figure 97. Comparison of Original ANN and LM performance for two week-ahead predictions of cyanobacteria at Station 612 without the five chemical variables

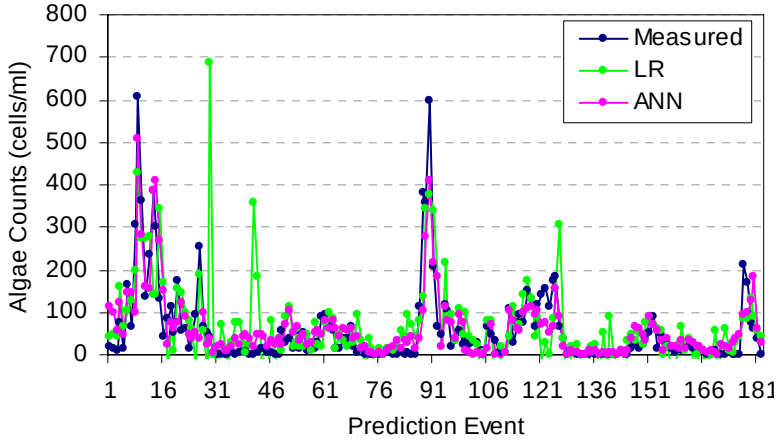


Figure 98. Comparison of Revised ANN and LM performance for One-week predictions of chlorophyta at Station 100 with complete input set

As shown by Figure 97, for the two-week ahead prediction of cyanobacteria counts at Station 612, the LM seriously under-predicted the three highest count events. The LM predicted just 388, 320 and 434 for algal blooms of 800, 932 and 1152 counts, respectively. 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 positives. Similarly for the other prediction cases using the same modeling approach, LMs have the tendency to under-predict most of the high count events. In particular, for one-week ahead prediction of chlorophyta counts at Station 101, the LM predicted just 80 and 253

for 304 and 608 count events, respectively. Similarly, in predicting two-week ahead chrysophyta counts at Station 101, LM again under-predicted the two particularly high count events. Figure 98 shows that for one-week ahead prediction of chlorophyta at Station 100, and as shown LM provided a total of 15 false positives, with three serious ones that predicted 688, 355 and 184 to a 44, zero and four count events, respectively. By contrast, the ANN accurately tracked down the high and low count events and provided just minor false positives. As the other figures in Appendix B-6 show, the LM in general was less reliable in avoiding false positives and negatives.

Overall then, the generally inferior LM performance supports the assumed non-linearity of this modeling problem. In addition, it also provides some degree of confidence that the ANN technology was, despite limited data, still able to learn to some degree seemingly subtle and complex relationships between the predictor variables and the algae count.

Tables for the sensitivity analysis for the LMs can be found in Appendix B-6. Because of the inferior performance of the LMs and the limited data sets, a formal evaluation of the sensitivity analysis results is not presented.

9. DISCUSSION AND CONCLUSIONS

This primary objective of this research project was to investigate the feasibility of using ANN technology for forecasting algal bloom events in surface water systems used for potable supply. ANN technology differs from traditional mechanistic modeling approaches that attempt to explicitly represent the governing laws with physical-based equations for forecasting system states of interest. Instead, ANNs represent a learning paradigm approach, where by processing representative historical events through their architecture, state-transition equations that predict system responses as a function of input predictor variables are obtained. Although previous work has been conducted in this area, every open water system presnts its own unique set of challenges and problems. In addition, unlike previous work in the literature, the number of available historical events was rather limited in this study, which likely will be the case for most facilities. Recognizing this, one of the objectives, then, was to attempt to identify important predictor variables for effective model performance so that sampling strategies could be improved. In addition, classification nets were also developed and tested as an alternative forecasting approach. Lastly, the feasibility and benefits of using ANN technology for modeling water treatment processes was also investigated.

Two water utilities served as test cases for forecasting algal population counts, New Jersey American Water, at its Swimming River facility, and PVWC, with the former utility also serving as the test case for the water treatment component. Each represents a unique system; Swimming River consists of a single reservoir within a watershed, while PVWC obtains its water supply from two different rivers that are part of two different watersheds with distinct characteristics. Hence, for Swimming River, water quality samples collected at three locations were treated as one location, while for PVWC, the best results were achieved when each of the three sampling stations were each modeled individually. While cyanobacteria, a recognized threat to human health and the environment, was the initial focus of this algae forecasting modeling effort, the study was expanded to include two other algae types measured at PVWC, chlorophyta and chrysophyta.

Several hundred ANN models of different types were developed and tested for the algal bloom forecasting problem, and their predictive performance was compared against measured values. In addition, linear models (LMs) were also developed with the same data sets as an additional benchmark for ANN performance. For Swimming River, a single one week-ahead prediction period was used, while for PVWC, one-week and two-week ahead prediction periods were primarily used. Input predictor variables included chemical, physical, hydrological, weather/meteorological, and water extraction data.

Unlike the PVWC facility, it was determined on the basis of the validation results that the ANN models were over-fitting the data for the Swimming River facility. Although the ANNs did significantly outperform LMs for the Swimming River data, the discrepancy between ANN training and validation results were relatively large. In addition, based upon sensitivity analyses, it was determined that water quality variables measured relatively infrequently by the utility are important predictors of cyanobacteria populations in the reservoir system. Consequently, the great majority of the ANN algae modeling effort were directed to the PVWC data.

For the PVWC system, two different input structures were initially used to assess possible time lags in algal system dynamics, and to address the feasibility of real-time implementation. The first or original input structure used input values measured primarily at the beginning of the prediction period, while the second or revised modeling structure used model input values measured primarily at the end of the prediction period, coinciding with the final or predicted algal count. Thus, unlike the original approach, where, under real-time forecasting conditions, the input values would be known a-priori, the revised modeling approach would have to forecast or assume input values corresponding to the future prediction day.

The number of historical training events available for each forecasting problem ranged between 19 and 136, and averaged 65, far below the minimum required number of 200, computed on the basis of the number of input and output variables. To address the data

quantity issue, both the original and revised modeling approaches were assessed with two distinct data sets. The first set consisted of the smaller number of historical events, which included a higher number of input variables, while the second set consisted of a larger number of historical events by excluding select water quality variables that were measured less frequently than other input variables. The five water quality variables included some of the so-called limiting nutrients, and are: Biological Oxygen Demand (BOD), Total Phosphorous/Orthophosphate, Nitrite/Nitrate, Sulfate, and Total Organic Carbon (TOC) (BOD was not excluded for Station 100).

In general, ANN models performed well during validation, and in many cases, accurately predicted large changes in algal populations. The level of accuracy was surprising, given the complexity and non-linear behavior of algal populations, the inherent data noise, and the relatively small number of historical events available for training. On the basis of validation correlation coefficients, the models (not classification) that used input values measured at the beginning of the prediction period slightly outperformed those that used input values measured at the conclusion of the prediction period, with average values of 0.72 and 0.69, respectively. A more subjective visual comparison of the time-series for the validation figures appears to confirm that the original models did achieve higher performance. The models that excluded the less frequently measured water quality variables with the benefit of more training events produced a higher average correlation coefficient of 0.77, to the 0.63 value for models that included these variables. However, there was at least one case where the models that included the select water quality variables achieved significantly higher validation performance.

The ANN models developed with inputs measured at the beginning of the one-week and two-week ahead prediction periods accurately predicted formation and dissipation of algal bloom events, as well as relative increase and decreases, indicating that there are natural time lags between system conditions and algal population responses. That is, algal populations may on average evolve predictably in response to system conditions, and the trajectory of algal counts over one and two-week forecast periods can be

accurately forecasted on the basis of real-time measurements. This may also reflect that open water conditions as influenced by external factors like weather do not typically change significantly in the short-term (e.g. weekly or even bi-weekly), and thus evolving algal populations are not prone to abrupt deviations from trajectory paths. The relatively small changes in conditions over prediction periods is supported by the statistical analyses of the data, and the fact that the revised models, which primarily used final measured input values, also performed relatively well.

There is also some physical foundation for this hypothesis. For example, because of the high specific capacity of water, significant water temperature changes will not typically occur over one and two-week prediction periods. Exceptions may occur with a particularly extreme weather event, which may also induce large water quality changes, but this will be atypical. Recognizing the complexity of these systems, additional research is necessary to test the validity of these claims for this system. Furthermore, because algal population dynamics vary from system to system, different time lag scales may be observed for different systems.

That the ANN models that excluded select water quality variables on average slightly outperformed models that included them (not classification) may signify less about the influence of these variables on algal population dynamics, and more about the inadequate number of training events. At the same time, it does indicate that during most time periods in the PVWC system, these variables may not be important predictors of algal populations, suggesting that they usually exist within a range of values that neither inhibit nor stimulate algal blooms. This is weakly supported by additional sensitivity analysis, as well as the Swimming River results, where significantly better results were achieved when water quality variables were included, even though this meant a five-fold reduction in the number of data events.

It should be recognized that for some time periods, inclusion of some or all of these variables may be important. There was at least one case where inclusion of these

variables significantly improved validation performance. A comparison of time-series between nitrite/nitrate concentrations and algal counts suggest at least a correlative if not causal relationship, where some large blooms occurred during periods of high concentrations. Thus, it may be that with an adequate number of data events for training, inclusion of the select water quality variables will improve overall performance, and may increase the likelihood of forecasting the formation and dissipation of bloom conditions during unusual conditions.

In comparison, the LMs did not perform as well, achieving significantly lower correlation coefficients and higher mean absolute errors, and in some cases, failed to predict very high count algal blooms while erroneously predicting other blooms during low count periods. It should be mentioned that the statistical distribution of the variables were not formally determined for this study, and consequently, data transformations were not performed (e.g. log normal). Had statistical transformations been made, the LMs would most likely have performed better. However, this also underscores one of the inherent advantages of ANNs; because of their universal non-linear modeling capability, they are not limited by the form of the data distribution(s).

The small number of historical data events limits the accuracy of the sensitivity analyses performed by measuring the relative increase in RMSE by excluding each input variable. However, some basic trends did emerge, with the most important possibly being the relative non-importance of the select water quality variables excluded from some models. In particular, the two "limiting nutrients", total phosphorous/orthophosphate and nitrite/nitrate, generally did not rank high as important predictor variables. This relative non-importance is weakly supported by the better performance of the models that excluded these variables. The time-series comparison of these parameters versus algal population also does not reveal an obvious relationship between concentrations and counts. Other variables like river and reservoir extractions ranked highly, and this is not surprising, given that water extraction volumes are often modified by PVWC in response to measured algal counts.

For example, a high extraction from Reservoir A, a backup source which normally is not used, occurs during periods of algal blooms on River A, where extraction from this river source is minimized. This potentially overly correlative relationship or model bias would be problematic for a forecasting method that is predicated on capturing the underlying mechanistic processes that govern algal population dynamics. When volumetric water extractions were excluded for four select test cases, forecasting performance remained high for models developed for Stations 101 and 612. In contrast, there was observable diminishment in performance for models developed for Station 100, which is a mixing point for various water sources, and thus extractions would be considered more causal in nature (i.e. hydraulic mixing of various source waters with differing levels of algal counts). The possibility for some causative relationship between water extractions and algal populations at river stations should not be dismissed. Variable extraction rates undoubtedly induce certain physical (e.g. different river stages) and water quality changes on the river that influences algal population dynamics.

As a final ANN modeling paradigm, RBF classification nets were developed with the single output variable representing the predicted "bin" or classification range of values for final measured algal counts, with four such possible bins; 0 to 10; 11 to 50; 51 to 200; and > 200. For the eight test cases, which used the original models with both the complete and reduced input sets, the ANN models classified the counts into the correct bins with very high accuracy, despite limited data. As with the previous models developed for predicting single count values, the highest performing classification nets excluded the select water quality variables. Even the poorest performing RBF net correctly classified 83 percent of the events, with all incorrect classifications falling within an adjacent bin (e.g. measured count of 8, predicted bin of 11 to 50).

There are a number of possible ways in which the ANN algal population forecasting models can be improved with additional work, and include:

1. The systematic elimination of input variables to further distinguish between critical and non-critical ANN inputs. This would help offset the relatively small number of historical events by reducing the dimensionality problem of the error space.
2. Increased monitoring of certain "limiting" nutrients, such as nitrite/nitrate and total phosphorous/ortho-phosphate, which would further define their importance on algal population dynamics.
3. Inclusion of other potentially important causal variables as model inputs, such as biological organisms that graze on algae.
4. Use of time lags for select predictor variables, such as streamflows and algal counts, which have been shown in a previous study (Maier, 1998) to significantly increase model performance.
5. A possible hybrid of the two modeling approaches, where some combination of existing/historical and future conditions is used as inputs. The obvious example of future conditions would include weather, where weather forecasts could be used as inputs to account for possible significant short-term effects on algal populations.
6. Collection of additional data to generate a larger number of events for model development and testing.
7. Following development of robust models, a perturbation sensitivity analyses that quantifies how different changes in input values affect algal population changes.

To increase data set sizes used in future modeling efforts, it is also possible to exploit the recently automatic data collection system recently implemented at the facility. Specifically, water quality conditions at Station 100, the intake point for the water treatment plant, is essentially the product of the relative contribution of the different water sources, which for most time periods is the rivers. At the very least, real-time water quality data measured at Station 100 could be used to estimate real-time water quality conditions at the water sources, and vice versa. This type of spatial-relational modeling capability could in the long-run be used by the utility to make more informed management decisions while offsetting sampling costs. This type of analysis also has a source water protection component to it, in that water quality conditions can be

continuously assessed in real-time. In terms of existing sampling methodology, it is recommended that the five water quality variables be sampled at a higher frequency.

The modeling results taken in their totality strongly indicate that the ANN models developed for the PVWC system learned some underlying relationship between select hydrologic, weather/meteorological, water quality, and extraction inputs and counts for the three algal classes. This is supported by a number of study outcomes, including; 1) relatively high model accuracy and overall consistency between training and validation results; 2) consistency in performance for different types of models, including original, revised, and classification models; 3) consistency between modeling results and physical intuition/system understanding; and 4) relatively poor performance of LMs.

Future research should focus on evaluating the use of ANN as a forecasting tool to predict single algal species. For example, odor problems are often produced by a single algae specie. However, it is not cost effective for most utilities to monitor populations at the species-specific level. Currently, taste and odor events attributed to algal productivity are best predicted by direct measurement of the odorants, with budgets for analytical services to measure odorant concentrations sometimes running into the thousands of dollars per year. ANN forecasting of specific odorant producing algal species could prove to be a more cost effect approach for predicting algal related taste and odor events. Research should be conducted to evaluate the cost savings that could be realized by replacing a 'traditional' grab and online monitoring program with an ANN generated forecast. Utilities would benefit from being able to provide the same level of monitoring while minimizing the costs of consumables and staff hours needed to maintain a monitoring program. Related to this, guidance is needed for utilities to develop sampling matrices that will generate sufficient data to adequately train and validate the ANN and reduce sampling costs. Identification of important predictor variables for effective model performance is also needed to improve sampling strategies and minimize analytical costs. Thus, in final summary for the algal bloom modeling:

- Despite a very limited number of available data sets, 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 inherent importance of these parameters to 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 water quality and meteorological 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.

For the water treatment study component, the ANN technology accurately learned to predict finished water quality conditions, namely average daily turbidity, highest daily

turbidity, and number of turbidity counts exceeding 0.1 NTU, based upon raw water conditions and treatment processes. The ANN models verified a seasonal component to treatment conditions and outcomes, and also confirmed a non-intuitive negative correlation between raw water temperature and raw water turbidity. It is believed that future consultation with water treatment experts could improve model performance by identifying the most important predictor variables based in accordance with professional experience and knowledge. As with the algae forecasting models, another possible way for improving model performance may be with inclusion of time lags for certain input variables.

10. REFERENCES

Baxter, C. W., S. J. Stanley, Q. Zhang, and D. W. Smith (2000). Developing Artificial Neural Network Models: A Guide for Drinking Water Utilities. 6th Environmental Engineering Specialty Conference of the CSCE & 2nd Spring Conference of the Geoenvironmental Division of the Canadian Geotechnical Society. London, Ontario.

Driscoll, F. G. (1986). Groundwater and Wells. Johnson Division St., St. Paul, Minnesota.

Hecht-Nielsen, R. (1987). Counterpropagation networks: *Proc. of the Int. Conf. on Neural Networks*, II, 19-31. New York, New York: IEEE Press.

Maier, H. R., G. C. Dandy, and M. D. Burch (1997). Artificial Neural Networks for Modeling and Prediction of Algal Blooms. J. of Ecological Modeling, Elsevier Press, No. 96, pp. 11-28.

Mirsepassi, A., B. Cathers, and H. Dharmappa (1995). Application of Artificial Neural Networks to the Real Time Operation of Water Treatment Plants. Proceedings of International Conference on Neural Networks. Vol. 1, Nov – Dec Perth, Australia.

Pelley, J. (2005). Environmental Science & Technology. January, pp. 37A-38A.

Poulton, M. (2001). *Computational Neural Networks for Geophysical Data Processing*. Amsterdam: Pergamon Press Ltd.

Recknagel, F, M. French, P. Harkonen, and K-I. Yabunaka (1997). Artificial Neural Networks for Modeling and Prediction of Algal Blooms. J. of Ecological Modeling, Elsevier Press, No. 96, pp. 11-28.

R.P. Lippmann, J.E. Moody and D.S. Touretzky (Eds.) Advances in Neural Information Processing Systems 3, 875-882. San Mateo, CA: Morgan Kaufmann.

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.

Skipworth, P., A. Saul, J. Machel (1999). Predicting Water Quality in Distribution Systems Using Artificial Neural Networks. Proceedings of the Institution of Civil Engineers. Water, Maritime, and Energy. Vol. 136, No. 1, pp. 1-8.

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

Statistica Software (2001). StatSoft, Inc. 2300 East 14th Street. Tulsa, OK, 74104.

Weigend, A.S., Rumelhart, D.E. and Huberman, B.A. (1991). Generalization by weight-elimination with application to forecasting.

Yakowitz, S. and F. Szidarovkszy (1989). *An Introduction to Numerical Computations.* New York/London: Macmillan College Division.

Yu, R., S. Kang, S. Liaw, M. Chen (1999). Application of Artificial Neural Network to Control the Coagulant Dosing in Water Treatment Plant. Water Science and Technology, Elsevier Science Ltd., Vol. 42., No. 3-4, pp. 403-408.