The Use of Artificial Neural Networks to Describe and Predict the Presence of Harmful Algae in the Indian River Lagoon, Florida

By

Erin L. Faltin

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science
Department of Environmental Science and Policy
College of Arts and Sciences
University of South Florida St. Petersburg

Major Professor: Melanie Riedinger-Whitmore, Ph.D.
Leon Hardy, Ph.D.
Leanne Flewelling, Ph.D., Florida Fish and Wildlife Conservation Commission
David Millie, Ph.D., Michigan Technological University

Date of Approval: November 6, 2014

Keywords: chlorophyll a, *Pyrodinium bahamense*, modeling, ecosystem, gray box

DEDICATION

This work is dedicated to my parents and grandparents for their unending support, encouragement, and love. I would also like to thank Kyle for his support and encouragement, even when it occasionally bordered on harassment – I guess those were the times when I needed to be harassed! Karen A. and Becca, each in their own way, offered support and the space I needed to maintain my sanity through these years, perhaps without even realizing it.

I would like to thank my advisor Dr. Melanie Whitmore for her calming voice and steady guidance throughout the whole process. Thanks also to my committee as a whole, Drs. Whitmore, Leon Hardy, Leanne Flewelling, and Dave Millie, for pushing me to be more thorough and rigorous than I thought I could be. Through their commitment and personal excellence, I've learned more than I ever expected I would.

To the Harmful Algal Blooms group as a whole, and especially Leanne, Karen S., and Cindy, I'd like to say a most sincere thank you for giving me the boost I needed, and the environment that made it possible, to go to graduate school in the first place.

ACKNOWLEDGMENTS

Several scientists contributed their time and expertise to the creation of the dataset I used in my analyses: Jennifer Wolny and Julie Brame, both formerly of the Florida Fish and Wildlife Conservation Commission, analyzed each water sample for phytoplankton community composition, and Dr. Edward Phlips and Susan Badylak at the University of Florida conducted nutrient, pigment, and chemical analyses. The St. Johns River Water Management district provided rainfall data for each site. Financial support for the data collection and analysis was provided by the St. Johns River Water Management District and the Indian River Lagoon National Estuary Program, through grants administered by Dr. Jan Landsberg.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Harmful algal blooms are a natural phenomenon of growing global concern. Dense blooms of single celled phytoplankton can have wide reaching effects on both the aquatic ecosystem and surrounding economies. This study constructed artificial neural network models of the northern Indian River Lagoon, Florida, using an existing dataset. Models attempted to both describe and predict chlorophyll a, as an indicator of total algal biomass, or *Pyrodinium bahamense*, a dinoflagellate known to bloom and produce the paralytic shellfish toxin saxitoxin in the lagoon. Descriptive models used current data while predictive models used time-lagged data as input. Further analyses were conducted on the best fitting descriptive models of chlorophyll a and *P. bahamense* in an attempt to elucidate driving factors of phytoplankton density within the ecosystem.

Water samples were collected bimonthly for five years from six fixed sites in the northern Indian River Lagoon; a variety of environmental and hydrological parameters were collected and chemical and biological analyses done for each sample. Additional descriptive and meteorological data were collected or calculated for each site and added to other input variables. The dataset analyzed contained 645 samples, with at least 11 parameters recorded for each.

Models of total chlorophyll a were relatively successful in describing absolute values and trends, and the predictive model (NMSE = 0.135, r = 0.933) was slightly more accurate than the descriptive (NMSE = 0.167, r = 0.913). Further analysis using metadata from the best descriptive model, known as “gray box” analyses, indicated that total phosphorus had a relatively large impact on overall chlorophyll a content in the water column.

Models of *P. bahamense* attempted to describe or predict varying descriptors of density, including absolute density, density in known positive samples, relative density (high, medium, low) in known positive samples, and presence/absence. Only presence/absence classification models were relatively successful in describing or predicting *P. bahamense* density; descriptive models were accurate for 78.9% of samples while predictive models were accurate for 73% of samples. Further analysis of metadata from the best descriptive model offered very little insight beyond factors known to affect phytoplankton growth in laboratory based enrichment experiments.

INTRODUCTION

Harmful Algal Blooms

Although they are a natural occurrence, the frequency and awareness of harmful algal blooms (HABs) is widely regarded to be increasing over the past few decades (Anderson, Glibert and Burkholder 2002). These blooms have complex and wide ranging cultural, ecological, and financial implications for the areas affected. In some cases, bloom-forming organisms produce toxic compounds that can cause illness or death to marine life and humans; in others, the overgrowth of microscopic organisms causes a decrease in water quality and incident light, killing benthic plants necessary to the ecosystem or causing other trophic shifts (Van Dolah, Roelke and Greene 2001; Anderson, Glibert and Burkholder 2002).

Financially, HABs cause losses to fishing and tourism industries, as well as to local governments. Fisheries may be closed for months or years as a result of mortality to existing stock or ongoing toxicity of edible tissues, as is found in shellfish or puffer fish (Hoagland, et al. 2002). In cases where nursery areas are affected, reduced catch rates in following years might also result in income losses (Hoagland, et al. 2002). Some HABs produce toxins which aerosolize under certain conditions, causing respiratory irritation on or near beaches, and affecting tourism and beach-front businesses and properties (Van Dolah, Roelke and Greene 2001; Glibert, et al. 2005). The cost of removing fish kills, scum, or foam produced by some HABs often defaults to local governments, further impacting the local economy (Hoagland, et al. 2002). The overall costs of these combined impacts are estimated to be in the millions each year (Hoagland, et al. 2002).

Algal blooms are often most noticeable in shallow coastal waters or embayments, and in many cases these areas are among the most sensitive to changing trophic structures (Anderson, Glibert and Burkholder 2002). Blooms which do not produce toxins may cause a decrease in dissolved oxygen levels, resulting in anoxic regions and the production of anaerobic byproducts such as methane or hydrogen sulfide (Paerl 1988). These chemical shifts, along with light attenuation, can cause the death of benthic macroalgae and seagrasses (Zingone and Enevoldsen 2000). As shallow environments frequently serve as nesting or juvenile habitat, hypoxic conditions, loss of protective vegetation, loss of food sources, or increased levels of toxic byproducts can cause the death of large numbers of larvae and juvenile fish and shellfish (Anderson, Glibert and Burkholder 2002). Aside from the risk to commercial fisheries, changes in primary producers and lower trophic levels will inevitably affect which upper level consumers will be successful, and how many of those consumers the ecosystem can support.

Harmful Algal Blooms in Florida

Residents of the state of Florida are familiar with HABs and their effects; fish kills and water discoloration have been documented on the west coast of the state for over 150 years (Steidinger and Joyce 1973; Steidinger 2009). The causative organism of many of these blooms, however, is known to be *Karenia brevis* (Davis 1948). The result of this long standing familiarity is a comprehensive program for sampling both water and shellfish for the presence of *K. brevis* and its associated toxins, brevetoxins (Steidinger 2009). This experience made initiating a second monitoring program easier when Quilliam, et al. (2004) traced, for the first time, the cause of several cases of human intoxication to saxitoxin found in the tissues of puffer fish harvested from the Indian River Lagoon (IRL).

Environmental monitoring in the IRL has been ongoing since the late 1980s (Sigua, Steward and Tweedale 2000). As population and industry has grown up around the watershed, the water quality and ecological integrity of the lagoon has degraded (Sigua, Steward and Tweedale 2000). This initial monitoring included several water chemistry parameters and chlorophyll data, but did not identify the species of algae present in the water column at the time of sampling (Sigua, Steward and Tweedale 2000). Algal monitoring in the IRL was primarily event response at that point (Steidinger, et al. 1998).

With the understanding that changing water quality conditions can give rise to changing algal communities, researchers at the University of Florida began two concurrent studies in 1997 (Phlips, Badylak and Grosskopf 2002; Phlips, et al. 2004; Badylak and Phlips 2004). The first study, a five year investigation at a site near Titusville, found two species known to produce toxins, *Pseudo-nitzschia pseudodelicatissima* (Hasle) and *Pyrodinium bahamense* var. *bahamense* (Plate), the latter of which appeared to be increasing in abundance

(Phlips, et al. 2004). The second study lasted two years and compared phytoplankton assemblages and community structure at eight sites throughout the IRL (Phlips, Badylak and Grosskopf 2002; Badylak and Phlips 2004). This study also found the potentially toxic *P. bahamense*, and listed it among other species seen at bloom densities throughout the study (Badylak and Phlips 2004; Landsberg, et al. 2006).

The cases of severe food poisoning traced back to puffer fish prompted researchers at the Florida Fish and Wildlife Conservation Commission to assemble a task force comprising researchers and experts from the University of Florida and St. Johns River Water Management District (SJRWMD), with assistance from Innovative Health Applications and the Ocean Research Conservation Association (Landsberg 2010). Results of the phytoplankton surveys conducted in the late 1990s and early 2000s gave the task force valuable baseline and community structure data which helped shape a cooperative monitoring study. This study was intended to address several areas of concern, including saxitoxin puffer fish poisoning, algal blooms, dolphin and turtle diseases, and possible threats to human health (Landsberg 2010). Funding through the Indian River Lagoon National Estuary Program (IRLNEP) and St. Johns River Water Management District (SJRWMD) allowed researchers to periodically sample water, picoplankton, phytoplankton, and sediment at six fixed stations in the northern portion of the IRL over the course of five years (Phlips, et al. 2011). The resulting dataset, composed of over 600 samples with at least 10 environmental or chemical parameters recorded for each, is known as the Core Data Set.

Artificial Neural Networks

Attempts to describe and predict bloom dynamics in Florida have been undertaken since the mid-twentieth century (Chew 1956). These attempts relied on heuristic knowledge and the results of laboratory-based experimentation, which are unlikely, especially in initial stages, to account for complex ecological interactions and codependent variables (Doig, III and Martin 1974). Even more recent and relatively complex models have difficulty incorporating all the factors that are now expected to affect the growth and transport of HABs (Walsh, et al. 2001). As a result, none of these models have yet been able to accurately describe processes that drive the spatial and temporal changes in development of HABs (Walsh, et al. 2001). Machine learning techniques, such as artificial neural network (ANN) modeling, offer an alternative to traditional modeling methods, and have been used successfully in aquatic ecosystems to model water quality and other parameters (Recknagel, et al. 1997; Kuo, et al. 2007).

A neural network is a mathematical learning model consisting of nodes, or neurons, designed to mimic human brain function (Goh 1995). Each neuron represents a complex polynomial equation with a weighted term corresponding to each of the input variables or neurons in the previous layer. The basic structure of a modern neural network includes an input layer, one or more hidden layers, and an output layer (Abdi 1994). Once given the variables from the input layer, the neurons in the hidden layer determine their weighted relationships using a variety of non-linear functions, and produce the output layer (Lippmann 1987). Back propagation, in which the predicted value is compared to the actual value and the weights and equations adjusted to produce greater accuracy, is frequently used to train the model (Goh 1995).

Neural network research and development began in the 1940s with the goal of developing a machine learning algorithm that would function more like a human brain (Abdi 1994). Early versions of neural networks, first used in the late 1950s and early 1960s, were similar in structure to ones we now use, but could only learn associations between inputs and outputs if those associations were the result of linear relationships (Abdi 1994). Additionally, these “models” used only binary inputs and outputs (Abdi 1994). Advances in the late 1970s and early 1980s, both in computers and in neural networks, introduced the use of non-linear relationships and novel methods of training the hidden layer of neurons (Abdi 1994).

Advances continue to be made, and ANNs are routinely used in disciplines as varied as stock market prediction, missile guidance and detonation, speech recognition, and drug development (Widrow, Rumelhart and Lehr 1994). Though initial scientific applications of ANNs focused on medicine and molecular biology, researchers began to investigate the usefulness of ANNs for ecological research in the early 1990s (Lek and Guegan 1999). Subsequently, comparisons were drawn between results of ANN analyses and those produced by traditional linear modeling (Lek, Delacoste, et al. 1996). The ability to calculate weighted and nonlinear relationships between variables and desired outputs allowed ANN analyses to better describe ecological relationships and predict results than traditional linear methods (Lek, et al. 1996). Artificial neural networks do not require the same assumptions about the data that parametric statistics do, namely that of a normal distribution, and may be used to classify or predict data based on inputs (NeuroDimension, Inc. 2012). Additionally, they easily process very large data sets and do not require the researcher to artificially reduce the number of variables analyzed (NeuroDimension, Inc. 2012).

Despite their superior performance and flexibility, ANNs lack the mathematical transparency of traditional linear and statistical models. The hidden layers are computed and adjusted according to whichever software program is in use, often with little, if any, input from the researcher, and as such are commonly referred to as a “black box” (Millie, et al. 2012). In a large environmental data set with many variables, this can obscure those factor(s) which may be causative agents of large, persistent, or more toxic phytoplankton blooms, and therefore limit the usefulness of the model in making management decisions (Young and Weckman 2010).

Traditionally, multivariate linear regression (MLR) has been used to model algal blooms and environmental drivers, and while the method provides greater clarity throughout the process, it is not without drawbacks (Millie, et al. 2012). As with all parametric models, assumptions regarding the distribution and variance of the variables are implicit in MLR, and because environmental data rarely meets these assumptions, the resulting model may have little or no value. Furthermore, MLR requires at least basic knowledge of which input variables are appropriate predictors, and which are not, beforehand so that they may be included or excluded from the model (Millie, et al. 2012). This also assumes that the researcher(s) are aware of all the factors influencing the model and have collected the necessary data to represent them.

In an effort to find middle ground between inaccurate but mathematically transparent “white-box” models, such as linear regressions, and more accurate “black-box” models such as ANNs, researchers have begun developing “gray-box” techniques (Young and Weckman 2010). These techniques may be applied to existing trained ANNs to extract knowledge and build a less complex model that still explains the system but also allows insight into its driving factors (Young and Weckman 2010). In the case of ANNs applied to ecological systems, this reduction of complexity may help researchers determine which environmental variables have the most impact on their target analyses, and help focus their research.

There are several methods of extracting detailed knowledge from existing trained ANNs; decomposition investigates the internal “hidden” structure, whereas the pedagogical approach compares relationships between the input and output layers, and the eclectic approach combines the two (Young and Weckman 2010). Though the eclectic method of creating a gray box for a given ANN appears rarely in practice, approaches to decomposition and pedagogical methods range from fairly simple Neural Interpretation Diagrams (NIDs) to complex response surfaces and decision trees (Young and Weckman 2010). The data to construct these analyses are generated by the computer program used to generate and train the ANN, and may be accessed through metadata files (Millie and Weckman, pers. com. 2012).

The goals of this project were to use ANN models to describe and predict key ecological concerns in the north IRL, specifically chlorophyll *a* levels and *P. bahamense* density. Sensitivity analyses conducted using the modeling software would identify which input variable(s) would have the greatest effect on output, and, once best-fitting models were identified, metadata would be used to determine driving factors via gray box analyses using NIDs and connected weights analysis. The results of the sensitivity and metadata analyses could then be compared and combined to determine which input variables are most likely to drive these regional concerns.

References

Abdi, Herve. "A Neural Network primer." *Journal of Biological Systems* 2, no. 3 (1994): 247-283.

Anderson, Donald M., Patricia M. Glibert, and JoAnn M. Burkholder. "Harmful Algal Blooms and Eutrophication: Nutrient Sources, Composition, and Consequences." *Estuaries* 25, no. 4b (August 2002): 704-726.

Badylak, S., and E. J. Phlips. "Spatial and temporal patterns of phytoplankton composition in a subtropical coastal lagoon, the Indian River Lagoon, Florida, USA." *Journal of Plankton Research* 26, no. 10 (2004): 1229-1247.

Chew, Frank. "A Tentative Method for the Prediction of the Florida Red Tide Outbreaks." *Bulletin of Marine Science* (University of Miami - Rosenstiel School of Marine and Atmospheric Science) 6, no. 4 (1956): 292-304.

Davis, Charles C. "*Gymnodinium brevis* sp. nov., a cause of discolored water and animal mortality in the Gulf of Mexico." *Botanical Gazette* 109, no. 3 (March 1948): 358-360.

Doig, III, Marion T., and Dean F. Martin. "The Effect of Naturally Occurring Organic Substances on the Growth of a Red Tide Organism." *Water Research* 8, no. 9 (1974): 601-606.

Glibert, Patricia M., Donald M. Anderson, Patrick Gentien, Edna Graneli, and Kevin G. Sellner. "The global, complex phenomena of harmful algal blooms." *Oceanography* 2 (June 2005): 136-147.

Goh, A. T. C. "Back-propagation neural networks for modeling complex systems." *Artificial Intelligence in Engineering* 9 (1995): 143-151.

Hoagland, P., D. M. Anderson, Y. Kaoru, and A. W. White. "The Economic Effects of Harmful Algal Blooms in the United States: Estimates, Assessment Issues, and Information Needs." *Estuaries* 25, no. 4b (August 2002): 819-837.

Kuo, Jan-Tai, Ming-Han Hsieh, Wu-Seng Lung, and Nian She. "Using artificial neural network for reservoir eutrophication prediction." *Ecological Modelling* 200 (2007): 171-177.

Landsberg, Jan H., Sherwood Hall, Jan N. Johannessen, Kevin D. White, Stephen M. Conrad, Jay P. Abbott, Leanne J. Flewelling, R. William Richardson, Robert W. Dickey, Edward L.E. Jester, Stacey M. Etheridge, Jonathan R. Deeds, Frances M. Van Dolah, Tod A. Leighfield, Yinglin Zou, Clarke G. Beaudry, Ronald A. Benner, Patricia L. Rogers, Paula S. Scott, Kenji Kawabata, Jennifer L. Wolny, and Karen A. Steidinger. "Saxitoxin Puffer Fish Poisoning in the United States, with the First Report of *Pyrodinium bahamense* as the Putative Toxin Source." *Environmental Health Perspectives* 114, no. 10 (2006): 1502-1507.

Landsberg, Jan. *Monitoring of Toxic Algae in the Indian River Lagoon, Florida, USA.* Grant final report, Fish and Wildlife Research Institute, Florida Fish and Wildlife Conservation Commission, St Petersburg, FL: Florida Fish and Wildlife Conservation Commission, 2010, 1-38.

Lek, Sovan, and J. F. Guegan. "Artificial neural networks as a tool in ecological modelling, an introduction." *Ecological Modelling* 120 (1999): 65-73.

Lek, Sovan, Marc Delacoste, Philippe Baran, Ioannis Dimopoulos, Jacques Lauga, and Stephane Aulagnier. "Application of neural networks to modelling nonlinear relationships in ecology." *Ecological Modelling* 90 (1996): 39-52.

Lippmann, Richard P. "An Introduction to Computing with Neural Nets." *IEEE Acoustics, Speech, and Signal Processing Magazine* (IEEE) 3, no. 4 (1987): 4-22.

Millie, Dave, and Gary Weckman. *Personal Communication* (10 2012).

Millie, David F., Gary R. Weckman, William A. Young II, James E. Ivey, Hunter J. Carrick, and Gary L. Fahnenstiel. "Modeling microalgal abundance with artificial neural networks: Demonstration of a heuristic 'Grey-Box' to deconvolve and quantify environmental influences." *Environmental Modelling & Software* 38 (2012): 27-39.

NeuroDimension, Inc. *NeuroSolutions Help.* 6.2. Edited by Jose Principe, Curt Lefebvre, Gary Lynn, Craig Fancourt and Dan Wooten. Gainesville, FL, 10 23, 2012.

Paerl, Hans W. "Nuisance phytoplankton blooms in coastal, estuarine, and inland waters." *Limnology and Oceanography* 33, no. 4 (1988): 823-847.

Phlips, Edward J., Susan Badylak, Mary Christman, Jennifer Wolny, Julie Brame, Jay Garland, Lauren Hall, Jane Hart, Jan Landsberg, Margaret Lasi, Jean Lockwood, Richard Paperno, Doug Scheidt, Ariane Staples, and Karen Steidinger. "Scales of temporal and spatial variability in the distribution of harmful algae species in the Indian River Lagoon, Florida, USA." *Harmful Algae* 10, no. 3 (2011): 277-290.

Phlips, Edward J., Susan Badylak, and T. Grosskopf. "Factors affecting the abundance of phytoplankton in a restricted subtropical lagoon, the Indian River Lagoon, Florida, USA." *Estuarine, Coastal and Shelf Science* 55 (2002): 385-402.

Phlips, Edward J., Susan Badylak, S. Youn, and Karen Kelley. "The occurrence of potentially toxic dinoflagellates and diatoms in a subtropical lagoon, the Indian River Lagoon, Florida, USA." *Harmful Algae* 3 (2004): 39-49.

Quilliam, Michael, Dominik Wechsler, Steven Marcus, Bruce Ruck, Marleen Wekell, and Timothy Hawryluk. "Detection and Identification of Paralytic Shellfish Poisoning Toxins in Florida Pufferfish Responsible for Incidents of Neurologic Illness." Edited by K. A. Steidinger, J. H. Landsberg, C. R. Tomas and G. A. Vargo. *Harmful Algae 2002.* St. Petersburg, FL: Florida Fish and Wildlife Conservation Commission, Florida Institute of Oceanography, and Inter-governmental Oceanographic Commission of United Nations Educational, Scientific and Cultural Organization, 2004. 116-118.

Recknagel, Friedrich, Mark French, Pia Harkonen, and Ken-Ichi Yabunaka. "Artificial neural network approach for modelling and prediction of algal blooms." *Ecological Modelling* 96 (1997): 11-28.

Sigua, Gilbert C., Joel S. Steward, and Wendy A. Tweedale. "Water-Quality Monitoring and Biological Integrity Assessment in the Indian River Lagoon, Florida: Status, Trends, and Loadings (1988-1994)." *Environmental Management* 25, no. 2 (2000): 199-209.

Steidinger, Karen A. "Historical perspective on *Karenia brevis* red tide research in the Gulf of Mexico." *Harmful Algae* 8, no. 4 (2009): 549-561.

Steidinger, Karen A., and Jr., Edwin A. Joyce. "Florida Red Tides." *State of Florida Department of Natural Resources Educational Series*, April 1973: 1-26.

Steidinger, Karen, Jan Landsberg, Earnest Truby, and Beverly Roberts. "First report of *Gymnodinium pulchellum* (dinophyceae) in North America and associated fish kills in the Indian River, Florida." *Journal of Phycology* 34 (1998): 431-437.

Van Dolah, Frances M., Daniel Roelke, and Richard M. Greene. "Health and Ecological Impacts of Harmful Algal Blooms: Risk Assessment Needs." *Human and Ecological Risk Assessment* 7, no. 5 (2001): 1329-1345.

Walsh, John J., Bradley Penta, Dwight A. Dieterle, and W. Paul Bissett. "Predictive Ecological Modeling of Harmful Algal Blooms." *Human and Ecological Risk Assessment* 7, no. 5 (2001): 1369-1383.

Widrow, Bernard, David E. Rumelhart, and Michael A. Lehr. "Neural Networks: Applications in Industry, Business and Science." *Communications of the ACM* 37, no. 3 (1994): 93-105.

Young, William A., and Gary R. Weckman. "Using a heuristic approach to derive a grey-box model through an artificial neural network knowledge extraction technique." *Neural Computing & Applications* 19 (2010): 353-366.

Zingone, Adriana, and Henrik Oksfeldt Enevoldsen. "The diversity of harmful algal blooms: a challenge for science and management." *Ocean & Coastal Management* 43 (2000): 725-748.

CHAPTER I: NEURAL NETWORK MODELING OF CHLOROPHYLL A IN THE INDIAN RIVER LAGOON

Introduction

The Indian River Lagoon (IRL) is one of a string of shallow interconnected lagoons that span over 150 miles of the eastern coast of Florida, from Ponce de Leon Inlet near New Smyrna Beach in Volusia County to Fort Pierce Inlet in St. Lucie County (Figure 1.1) (St. Johns River Water Management District 2010). The IRL is an estuarine habitat that supports over 4000 plant and animal species, 35 of which are listed as threatened or endangered (Figure 1.2) (St. Johns River Water Management District n.d.). Overall, the IRL supports the most biologically diverse estuarine ecosystem in the country (St. Johns River Water Management District 2010). This diversity is supported in part by widely varying 50% renewal times, which range from over a year to just a few days, depending on location, precipitation, and tidal rhythms (Smith 1993). In recent decades, the area has been subject to intense development, with both industrial and recreational uses putting pressure on the estuarine ecosystems (Badylak and Phlips 2004).

Widespread growth of seagrasses is generally associated with ecosystem health and diversity. Primary and secondary production in seagrass beds is very high, and many commercially and recreationally valuable species spend at least part of their life cycle in such beds (Gillanders 2006). Of the 60 seagrass species found globally, at least 10% are historically found in the IRL, one of which, Johnson's seagrass (*Halophilia johnsonii* Eiseman), is listed as threatened under the Endangered Species Act (Endangered and Threatened Species 1998; Virnstein and Carbonara 1985). Seagrass coverage has been monitored in the IRL for decades via aerial photography, and transect studies which began in the mid-1980s have provided more specific data for several locales (St. Johns River Water Management District 2013). Coverage near urban areas had dropped during the mid-1980s and early 1990s, a period which saw a great deal of growth in population and land use, likely indicating a negative anthropogenic influence (Sigua, Steward and Tweedale 2000; St. Johns River Water Management District 2013). Since that time, conservation initiatives such as storm water controls to reduce sediment influx and coastal wetland restoration projects have yielded an overall increase in coverage, until the past five years (Sigua, Steward and Tweedale 2000; St. Johns River Water Management District 2013). Since 2011, scientists at the St. Johns Water Management District have observed a 60% drop in coverage to the lowest levels recorded since monitoring began: less than 30,000 acres within the IRL as far south as the Fort Pierce inlet (St. Johns River Water Management District 2013).

Loss of seagrass may have several causal factors, including incident light limitation, mechanical damage, and nutrient enrichment leading to trophic shifts (Deegan, et al. 2002; Duarte 2002). The recent sharp decrease in seagrass coverage is thought to be the result of two dense, extensive, and long-lasting algae blooms which covered 130,000 acres of the IRL, from north of Titusville to

Eau Gallie (St. Johns River Water Management District 2013). Sunlight availability is known to be a limiting factor to seagrass growth in the IRL, and it is believed that these algal blooms shaded out benthic flora, contributing to a decline in sea grass coverage (Sigua, Steward and Tweedale 2000). The first bloom comprised picoplanktonic green algae and cyanobacteria, and was observed from March through November 2011 with a maximum density of more than 10^6 cells ml^{-1} (Phlips and Badylak 2012). The second, seen from June to August 2012, was a brown tide later identified as the pelagophyte *Aureoumbra lagunensis* (DeYoe), in densities as high as 3.3 x 10^6 cells ml^{-1} (DeYoe, et al. 1997; Phlips and Badylak 2012). This species, associated with an eight-year bloom which caused major ecological impacts in Texas, had not been known to bloom in the IRL prior to this event, though examination of historical samples confirmed its presence as far back as 2005 (Phlips and Badylak 2012).

Chlorophyll *a* measurements have long been used to represent total phytoplankton abundance, and very high (171 µg L^{-1}) levels measured during the peak of the brown tide support the use of this metric as an estimator of *A. lagunensis* density along with other species (Steele 1962; Phlips and Badylak 2012). Successful descriptive modeling of chlorophyll *a*, using other observed and measured environmental parameters as input, could yield insight into driving factors of increased plankton growth. Predictive models, if possible, could make mitigation of the harmful effects of blooms more effective. Information gained from either type of model could be very useful to researchers, in targeting their efforts, to policymakers in generating guidelines to increase the overall health of natural resources, and to managers making decisions about public use or allocation.

Use of Artificial Neural Networks to describe and predict Harmful Algal Blooms

The use of Artificial Neural Networks (ANNs) to describe and predict the behavior of Harmful Algal Blooms (HABs) has been attempted in many different natural systems since the mid-1990s (Lee, et al. 2003; Singh, et al. 2009). These studies have mainly focused on lacustrine and riverine systems, with only a handful having been conducted on coastal areas (Lee, et al. 2003). In general, freshwater studies have been successful in developing models that accurately describe seasonal variation in either target species abundance or chlorophyll levels, but their usefulness as a predictive tool has been limited by their design (Lee, et al. 2003).

In the cases where ANNs have been used to describe or predict chlorophyll *a* or species abundance in coastal regions, model accuracy has been increasing over the past decade, presumably as technology advances and the ecosystems are better understood (Barciela, Garcia and Fernandez 1999; Velo-Suarez and Gutierrez-Estrada 2007; Melesse, Krishnaswamy and Zhang 2008). Where models were built with varying numbers of input variables, models with a higher number were more accurate than those with fewer, supporting the high complexity of an ecosystem (Melesse, Krishnaswamy and Zhang 2008). Thus, data sets comprising many variables over a long time scale are good candidates for analysis via ANN, and are likely to provide more accurate and robust models.

The Core Dataset is just such a candidate, spanning five years and containing multiple analyses for over 600 water samples. To date, several analyses have been conducted on individual parameters, such as phytoplankton community analyses or nutrient change over time, but no study has attempted to use multiple parameters to describe the ecosystem more holistically (E. J. Phlips, et al. 2010; E. J. Phlips, et al. 2011; E. J. Phlips, et al. 2014). Artificial neural networks provide a platform that may be able to encompass multiple widely varying parameters for each data point and visualize how the system as a whole functions and changes.

To generate the Core Data Set, samples were collected twice monthly at six sites associated with a range of waterbody/watershed size ratios, site characteristics, and surrounding urbanization (Figure 1.1) (Phlips, et al. 2011). Average time between sampling was 15 days. Salinity and temperature were measured in the field with either a YSI or Hach/Hydrolab multi-probe, and water samples were taken using an integrated tube sample to within 0.1 m of the bottom (Phlips, et al. 2011). Samples were then split and preserved with either Lugol's solution or gluteraldehyde, with unpreserved aliquots frozen for nutrient analysis (Phlips, et al. 2011). Phytoplankton were enumerated in Lugol's preserved samples using inverted phase contrast microscopy (Phlips, et al. 2011).

As water clarity, and therefore phytoplankton abundance, is a concern in the IRL, and ANNs provide a more flexible platform to include many variables, this study attempted to combine the two. I used environmental and chemical parameters from each sample, to build an ANN which endeavored to describe the chlorophyll *a* measurements from the same or future samples. Where the network had sufficient input data and successfully described the system, the analysis of its structure provided insight into the driving factors of high chlorophyll *a* levels, and therefore phytoplankton abundance.

Materials and Methods

In addition to the Core Data Set, daily rainfall data were obtained from the St. Johns River Water Management District for each site. These data are the result of an ongoing project by the SJRWMD using daily radar maps in comparison with in situ rainfall gauges. Rainfall data from a network of 75 rain gauges are provided to a contractor, who then also creates a radar map from several overlapping National Weather Service radar stations (St. Johns River Water Management District n.d.). The resulting radar map covers the SJRWMD area in totality, and is subdivided into pixels measuring 2 kilometers on each side. Rain gauge and radar data are combined to derive a gauge-adjusted dataset, which is then delivered to the SJRWMD where is it quality checked and added to the district's database (St. Johns River Water Management District n.d.). Global positioning system (GPS) coordinates were overlaid with maps available from the St. Johns River Water Management District website (http://webapub.sjrwmd.com/agws10/ radrain/index.htm) to identify the pixel containing each sampling site. Staff at St. Johns River Water Management district provided daily rainfall data for each pixel during the years when sampling took place, and cumulative rainfall since most recent sample date was calculated for each sample, according to site.

Additional calculations using total nitrogen (TN) and total phosphorus (TP) within the Core Data Set generated the TN/TP ratio. Total depth at each sampling site was obtained from either the St. Johns River Water Management District or National Oceanographic and Atmospheric Administration (NOAA) depth charts, and then used with secchi depth data already in the Data Set to calculate relative secchi depth for each sample. For predictive models, desired variables at each sample location were shifted by one sampling period such that data collected on one sample date would be used to predict the target variables on the following.

Building the Artificial Neural Networks

Both descriptive and predictive models were built using NeuroSolutions 6.06 (NeuroDimension, Inc. 2010). This platform allows the user to select from different model types and learning algorithms for each network built. Multi-layer perceptrons (MLP) utilizing static backpropagation, a relatively basic type of neural network, were used in each case, and learning algorithms were varied to determine which would produce a model with the best correlation to in situ measurements.

Models intended to describe and predict chlorophyll *a* using nutrient and environmental data from each sample as input were initially built using two hidden layers, with cross validation data, and trained using each of three learning algorithms commonly found in text (conjugate gradient, step, and momentum)

(Singh, et al. 2009; D. F. Millie, et al. 2013). Samples were randomized (n = 645) and 60% were used as training exemplars, 15% was used for cross validation, and the remaining 25% for testing the model. For each learning type, the number of processing elements (PEs) in each hidden layer was varied from 2 to 16 as the network trained, to produce the most efficient and accurate model. After the models were trained, each was tested and performance metrics were produced by the software. The most accurate version, as determined by the given metrics, was then re-trained as many as ten times to further refine the algorithm, and the best of these replicates was retained for sensitivity testing. Sensitivity testing is part of the NeuroSolutions software, and the program will vary the values of each of the input variable independently of the others and record the change in output for analysis. The user may specify the range of variation for each variable; in this case, I used both one and two standard deviations about the mean. Gray box analysis was performed only on the most accurate descriptive model in an effort to visualize influences contributing to chlorophyll *a* values in real time.

Descriptive models of chlorophyll *a* were built using the following variables as input: temperature, salinity, dissolved oxygen, pH, ratio of secchi depth to total depth (relative secchi depth), total nitrogen, total phosphorus, TN/TP and rainfall since last sample. Predictive models attempted to describe chlorophyll a using the variables described above collected from the previous sample in addition to the previously measured chlorophyll *a* value.

Gray Box Analysis

Gray box analysis resulting in a neural interpretation diagram and connected weights analysis was based on an Excel template provided by Dr. Gary Weckman (Russ College of Engineering and Technology, Ohio University) and used in his teaching. Using a metadata file with file extension ".bst" ("best" files), each term of the complex polynomials used to construct the model can be recreated. This file contains weights generated by the software during training for each input representing the contribution of that input to the next layer of the model. The use of multiple spreadsheets within Excel can allow the user to trace the relative importance of each input through the web of equations produced by the model and generate a NID or other graphical representation of the input variables. In this case, I generated both an NID, depicting the relative contribution of each variable to each node in each layer, and a pie chart of connected weights showing the relative contribution of each input variable to the output layer.

Results

The conjugate gradient learning algorithm produced the most accurate descriptive model initially, and re-training produced the most accurate model in the sixth iteration (MSE = 70.12, NMSE = 0.167, r = 0.913; Figure 1.3, Table 1.1). The final model contained two hidden layers with 5 nodes in the first and 12 nodes in the second. Sensitivity analysis performed using the NeuroSolutions software determined that relative secchi depth, total phosphorus, and temperature had the greatest impacts on chlorophyll *a* concentrations (Figure

1.4, Table 1.2). Gray box analysis produced an NID showing how each node in one layer relates to each in the following layer, as well as pie chart of connected weights indicating the relative importance of each input variable; connected weights analysis indicated the three most important variables were total phosphorus, relative secchi depth, and TN/TP (Figures 1.5 and 1.6).

The momentum learning algorithm produced the most accurate predictive model of chlorophyll *a*, and re-training produced the most accurate model in the fourth iteration (MSE = 48.76, NMSE = 0.135, r = 0.933; Figure 1.7 and Table 1.3). The final model contained two hidden layers, the first consisting of 7 nodes and the second of 14. Sensitivity analysis performed using the NeuroSolutions software determined that the three input variables which contributed most to future chlorophyll *a* amounts were chlorophyll *a* measurements from the previous sample, TN/TP, and temperature (Figure 1.8, Table 1.4).

Discussion

Though the use of descriptive models to gain insight into the driving forces of HABs is promising, it is not without pitfalls. The ability of a neural network to encompass a great many input variables means that it is less likely to miss trends or important influences, however, the strength and accuracy of models is greatly increased by large quantities of data, which may be difficult for researchers to collect consistently. Furthermore, understanding changing influences in ecosystems requires collection of data over relatively large time scales, which might be impractical to collect due to funding or other reasons.

The variables collected as part of the Core Data Set as a whole included several which were not included in the models, such as total phytoplankton communities and sediment analyses. These data were not included as inputs into the chlorophyll models because the overall goals of these models were to estimate future chlorophyll levels and determine driving factors of present chlorophyll levels using variables which can be collected remotely and transmitted to researchers wirelessly. Had these data been included in the models, they might have provided more insight into the effects of different grazing and recruitment from cyst populations on chlorophyll *a*, both elements suspected or known to affect phytoplankton communities (Phlips, Badylak and Grosskopf 2002).

The individual sample sites used for data collection in the Core Data Set were chosen for their variety over a number of defining characteristics. In this case, though the data set in its entirety was large, the number of samples collected at any one site was insufficient to generate a robust model. For this reason, all samples were analyzed together, and the model generated describes generalities of the northern IRL as a whole, but may not accurately reflect all the influences acting on smaller regional scales.

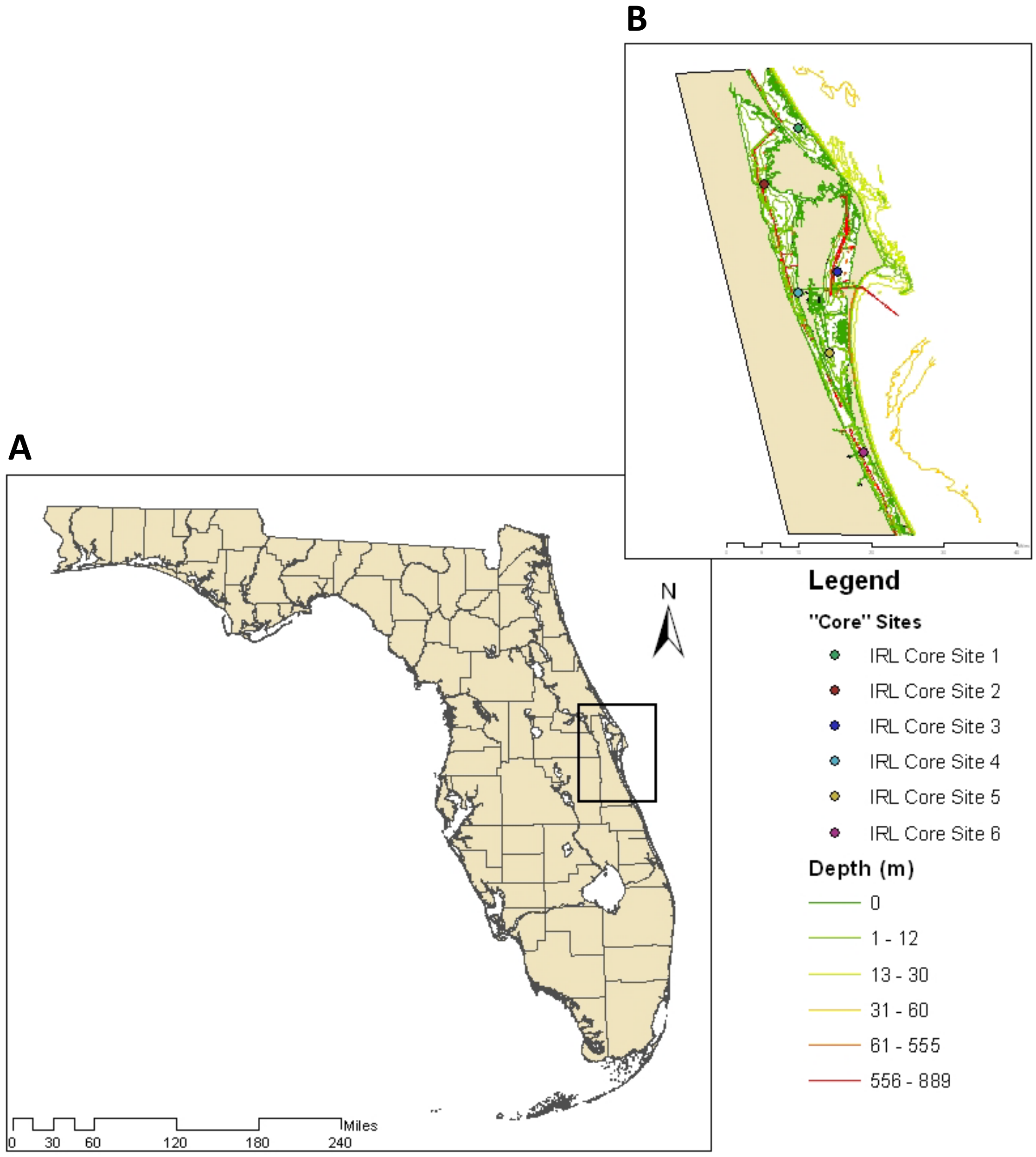
Results of the sensitivity analysis included in the NeuroSolutions software and those from the connected weights analysis emphasize the need for heuristic knowledge and the laboratory-based experiments that preceded the development of artificial intelligence. Though the two do not agree completely on which variables are most likely to be driving factors, they both emphasize the importance of relative secchi depth and total phosphorus. However, secchi depth is affected by the presence of chlorophyll *a* producing phytoplankton, and therefore this does not represent a causal relationship in the manner represented by the model. Thus, the estimation of a causal effect resulting from changes in phosphorus concentrations is the more valuable datum resulting from the model, indicating that the northern part of the estuary may be phosphorus limited. This conclusion is further supported by calculation of TN/TP, the mean of which was 20.53 for all samples, and in excess of the Redfield ratio (16:1 for N:P; Howarth 1988).

Phlips, et al. (2002) also found that phosphorus was likely to be the limiting nutrient in this region of the IRL when they analyzed both phytoplankton samples and bioassay results. However, in subsequent analyses Phlips, et al. (2010) found that temperature and nitrogen were driving factors of total phytoplankton abundance at a site not far from Core Site 2. Differences in results between Phlips' two studies, and between those and the ANNs, are likely to arise from the differing analyses used in each case. In 2002, Phlips, et al. used Pearson's Correlation Analyses alongside bioassays, while Phlips, et al. (2010) built several restricted all-possible-models regressions. These results not only underscore the value of laboratory based research, but emphasize the importance of model choice. It should also be noted that Phlips, et al. (2002; 2010) indicated that salinity, as influenced by both retention time and rainfall, affected phytoplankton density, while the ANN models did not identify either salinity or rainfall as strong drivers.

The greatest predictor of future chlorophyll *a* levels, by a large margin, was the chlorophyll *a* value. As it is uncommon for blooms to appear or dissipate in the space of two weeks, this seems intuitive, if not especially insightful into driving factors of bloom formation or of the ecosystem as a whole. Indeed, this is a common result seen in predictive models of phytoplankton production, and some researchers have used either time-lagged chlorophyll *a* or species abundance as the sole input for their models (Melesse, Krishnaswamy, & Zhang, 2008; Velo-Suarez & Gutierrez-Estrada, 2007). The next most influential variable on future chlorophyll *a* concentrations is TN/TP, which supports the conclusion drawn from the descriptive model that limiting nutrients are a driving force in phytoplankton abundance.

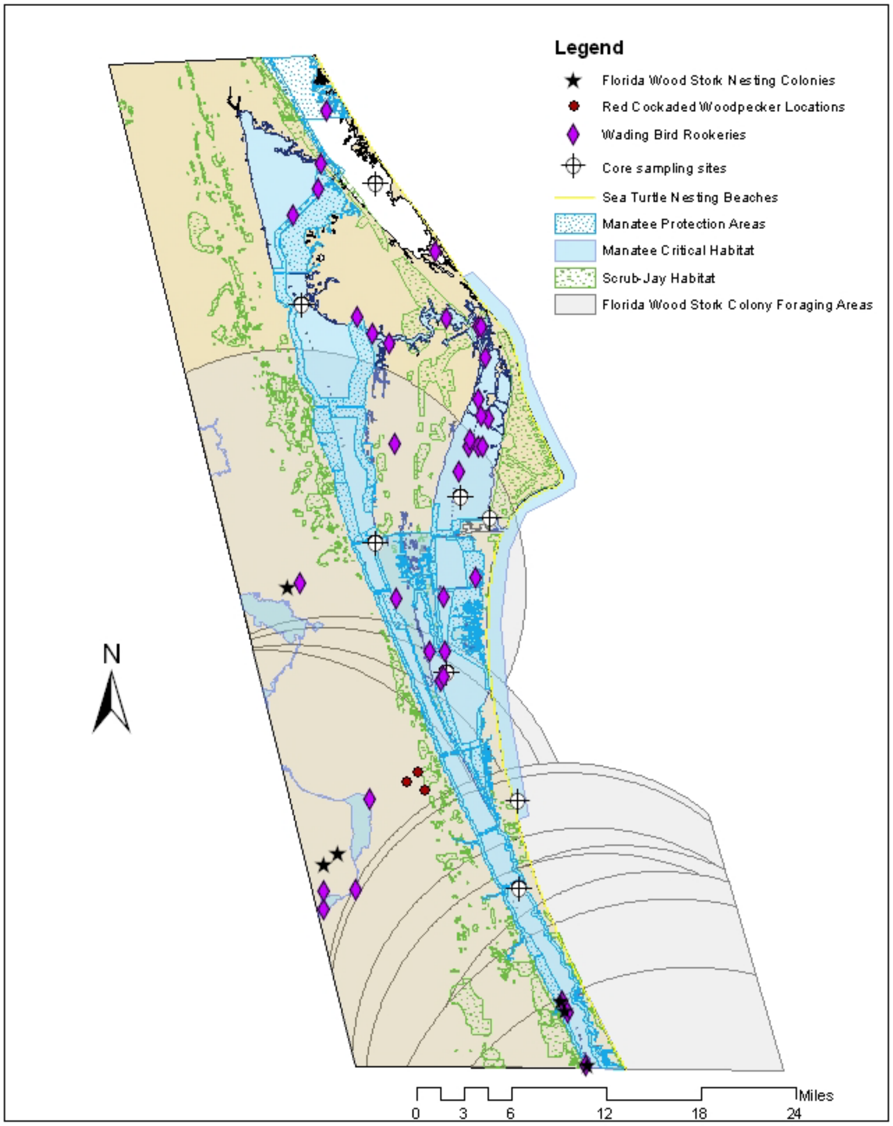
Interestingly, the descriptive model found that rainfall amounts between sample dates were a moderately important variable, while the predictive model found the same variable to have very little impact. This may indicate that nutrient influx from run-off has an immediate impact that does not last much beyond a week or two, and that other sources of nutrient enrichment in the environment, such as ground water influx, should be considered important as well. Furthermore, learning that the northern IRL is primarily phosphorus limited can help managers focus on specific types of pollution in addition to modes of pollution transport.

The Core Data Set provides a much broader view of the northern IRL than previous data sets have been able to, merely because of the amount of data collected from each sample. The ANNs built using this 5 year data set have highlighted the processes that influence phytoplankton abundance on that relatively short time scale. While this information can be beneficial in day to day or year to year management, decade scale changes will only be observed through decade scale data. Networks built using a longer term data set will be required to elucidate processes active over large time scales. The short term processes revealed by these networks may be compared with others, as determined using analyses of either subsequent short term data sets or combined long term data sets, and managers will have a much more complete tool box for improving the long term health of this critical estuary.



E. Faltin, 5.4.2011
Data courtesy Florida Fish & Wildlife Conservation Commission

Figure 1.1: Sampling locations and bathymetry for Core Data Set sites in the Indian River Lagoon. The Indian River Lagoon in relation to the east coast of Florida (A); colored dots represent sampling sites while graduated colored lines mark depth (B).


Legend
Florida Wood Stork Nesting Colonies
Red Cockaded Woodpecker Locations
Wading Bird Rookeries
Core sampling sites
Sea Turtle Nesting Beaches
Manatee Protection Areas
Manatee Critical Habitat
Scrub-Jay Habitat
Florida Wood Stork Colony Foraging Areas
N
0
3
6
12
18
24
Miles

E. Faltin, 5.4.2011
Data courtesy Florida Fish & Wildlife Conservation Commission

Figure 1.2: Known nesting sites and foraging areas of several protected species within the study area, demonstrating the density of critical habitats within the IRL.

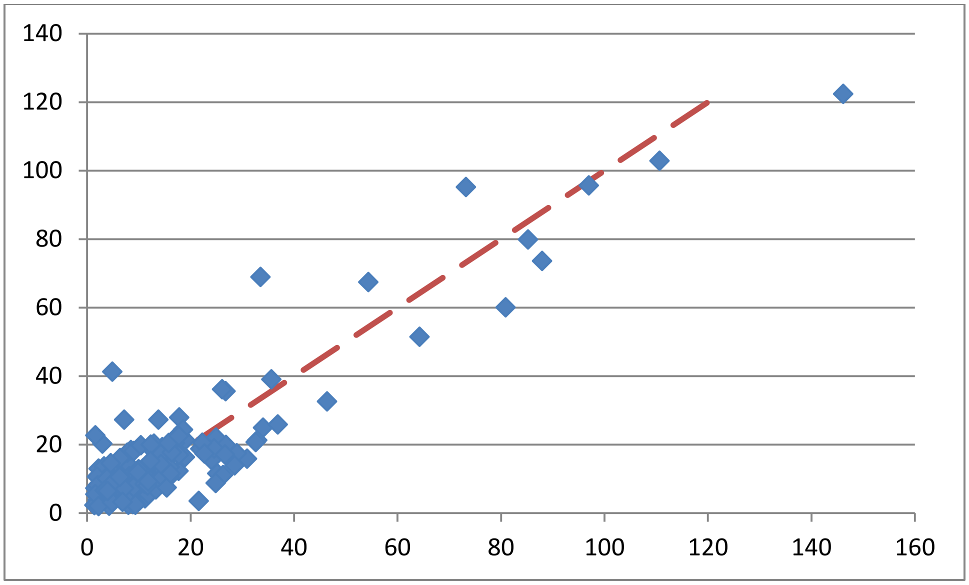


Figure 1.3: Chlorophyll *a* descriptive model performance: predicted values are plotted against observed values, red line has the equation $y = x$ and is shown for reference.

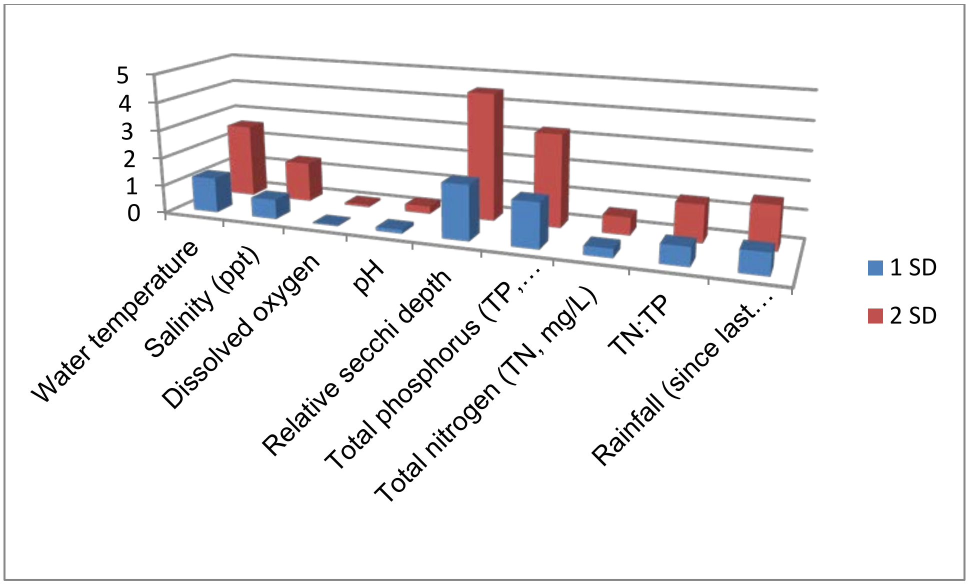


Figure 1.4: Sensitivity analysis results for Chlorophyll *a* descriptive model; each variable was altered by 1 and then 2 standard deviations around the mean value while other variables were held constant; changes to the model output are represented by the bar graph below.

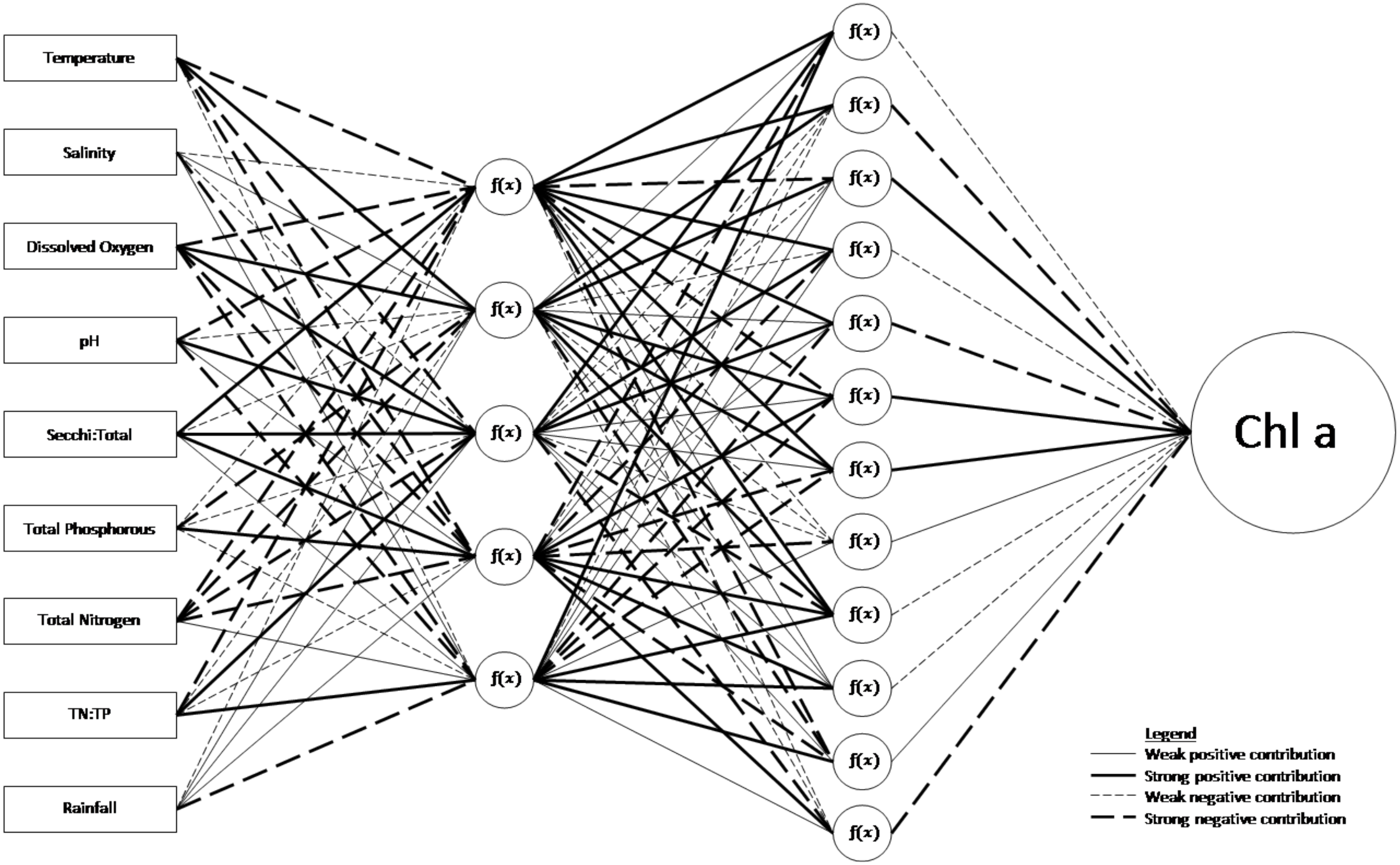

Temperature
Salinity
Dissolved Oxygen
pH
Secchi:Total
Total Phosphorous
Total Nitrogen
TN:TP
Rainfall
f(x)
Chl a
Legend
Weak positive contribution
Strong positive contribution
Weak negative contribution
Strong negative contribution

Figure 1.5: Neural Interpretation Diagram for Chlorophyll *a* descriptive model, generated using metadata produced by NeuroDimensions software in the course of training the ANN.

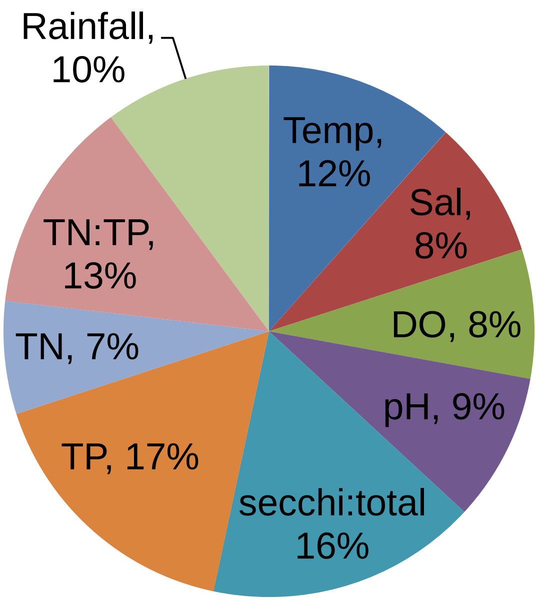


Figure 1.6: Connected weights pie chart for Chlorophyll *a* descriptive model showing contribution of each input variable to output result as determined by tracing the weights of each layer through the ANN.

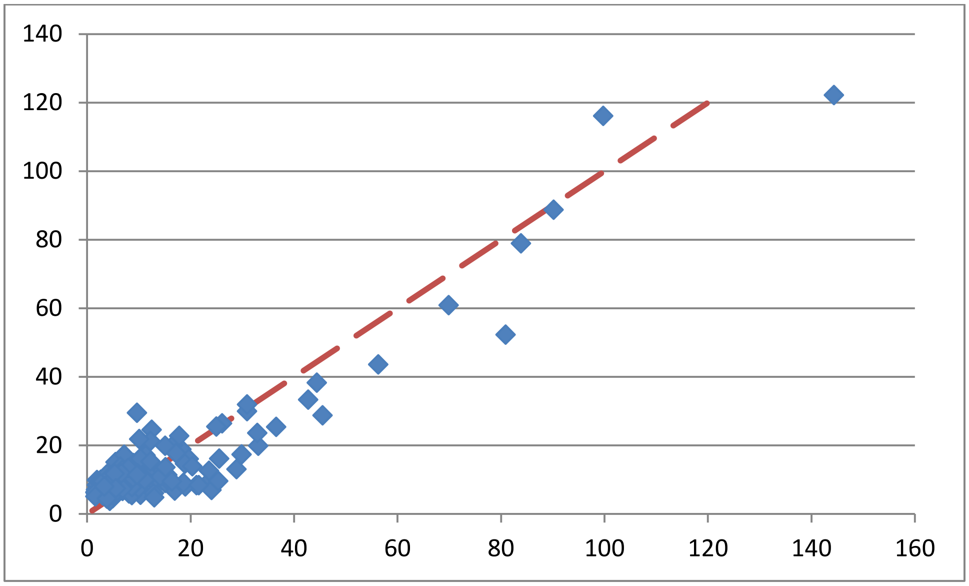


Figure 1.7: Chlorophyll *a* predictive model performance: predicted values are plotted against observed values, red line has the equation $y = x$ and is shown for reference.

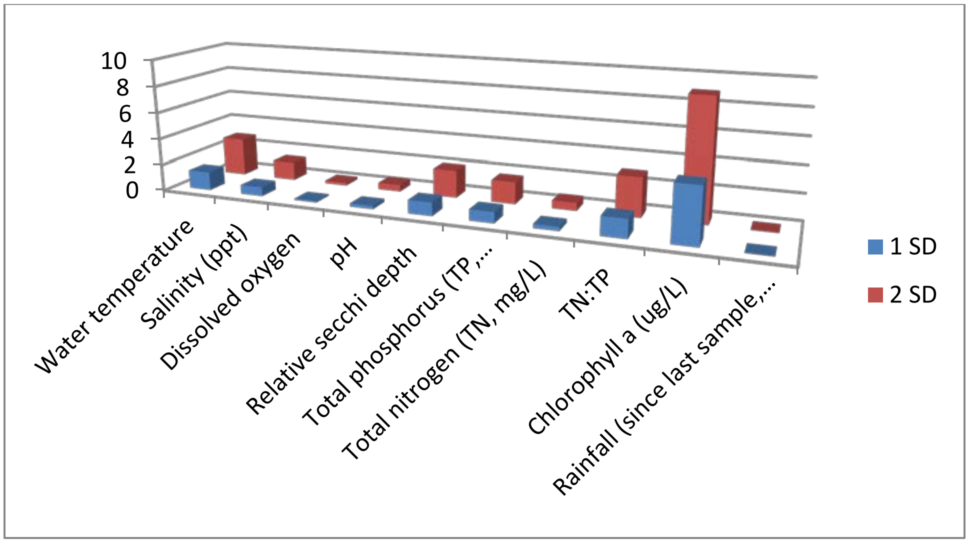


Figure 1.8: Sensitivity analysis results for Chlorophyll *a* predictive model; each variable was altered by 1 and then 2 standard deviations around the mean value while other variables were held constant; changes to the model output are represented by the bar graph below.

Table 1.1: Chlorophyll *a* descriptive model performance metrics; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***Chl a***
MSE	70.11705506
NMSE	0.167310819
MAE	5.893172361
Min Abs Error	0.018140696
Max Abs Error	36.39797849
r	0.91300188

Table 1.2: Sensitivity analysis results for Chlorophyll *a* descriptive model; numerical values calculated to describe sensitivity of the model output to changes within 1 and 2 standard deviations of the mean for each input variable.

Sensitivity 1 SD around mean	***Chl a***
TEMP	1.236817804
SALINITY	0.685720596
DISSOLVED_O2	0.030549802
pH	0.127367292
sechhi:total depth	1.927409241
TP (mg/L)	1.579842633
TN (mg/L)	0.308209417
TN:TP	0.648790715
Rainfall (since last sample, inches)	0.753631954

Sensitivity 2 SD around mean	***Chl a***
TEMP	2.601655056
SALINITY	1.414581405
DISSOLVED_O2	0.076471032
pH	0.262769831
sechhi:total depth	4.495305491
TP (mg/L)	3.271331235
TN (mg/L)	0.623704303
TN:TP	1.313219392
Rainfall (since last sample, inches)	1.547770059

Table 1.3: Chlorophyll *a* predictive model performance metrics; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***Chl a next sample***
MSE	48.76381201
NMSE	0.134985059
MAE	5.256872977
Min Abs Error	0.046971993
Max Abs Error	28.6687964
r	0.932963157

Table 1.4: Sensitivity analysis results for Chlorophyll *a* predictive model; numerical values calculated to describe sensitivity of the model output to changes within 1 and 2 standard deviations of the mean for each input variable.

Sensitivity 1 SD around mean	***Chl a next sample***
TEMP	1.376300375
SALINITY	0.670326623
DISSOLVED_O2	0.099810743
pH	0.238913915
sechhi:total depth	1.003571761
TP (mg/L)	0.820453129
TN (mg/L)	0.293183092
TN:TP	1.406726239
CHLa est (ug/L)	4.162075881
Rainfall (since last sample, inches)	0.019577222

Sensitivity 2 SD around mean	***Chl a next sample***
TEMP	2.83803242
SALINITY	1.373120928
DISSOLVED_O2	0.198922208
pH	0.491863986
sechhi:total depth	2.05675856
TP (mg/L)	1.653035687
TN (mg/L)	0.585252392
TN:TP	2.946528966
CHLa est (ug/L)	8.95384258
Rainfall (since last sample, inches)	0.042639534

References

Badylak, S., and E. J. Phlips. "Spatial and temporal patterns of phytoplankton composition in a subtropical coastal lagoon, the Indian River Lagoon, Florida, USA." *Journal of Plankton Research* 26, no. 10 (2004): 1229-1247.

Barciela, Rosa M., Emilio Garcia, and Emilio Fernandez. "Modelling primary production in a coastal embayment affected by upwelling using dynamic ecosystem models and artificial neural networks." *Ecological Modelling* 120, no. 2-3 (1999): 199-211.

Deegan, Linda A., et al. "Nitrogen loading alters seagrass ecosystem structure and support of higher trophic levels." *Aquatic Conservation* 12, no. 2 (2002): 193-212.

DeYoe, Hudson R., et al. "Description and Characterization of the Algal Species *Aureoumbra lagunensis* gen. et sp. nov. and Referral of *Aureoumbra* and *Aureococcus* to the Pelagophyceae." *Journal of Phycology* 33, no. 6 (1997): 1042-1048.

Duarte, Carlos M. "The future of seagrass meadows." *Environmental Conservation* 29, no. 2 (2002): 192-206.

Endangered and Threatened Species. "Threatened Status for Johnson's Seagrass." *Federal Register*, September 14, 1998: 63 Fed. Reg. 49035-49041.

Gillanders, Bronwyn M. "Seagrasses, Fish, and Fisheries." Chap. 21 in *Seagrasses: Biology, Ecology, and Conservation*, edited by Anthony W. D. Larkum, Robert J. Orth and Carlos M. Duarte, 503-536. Dordrecht: Springer, 2006.

Howarth, Robert W. "Nutrient Limitation of Net Primary Production in Marine Ecosystems." *Annual Review of Ecology and Systematics* 19 (1988): 89-110.

Lee, Joseph H.W., Yan Huang, Mike Dickman, and A.W. Jayawardena. "Neural network modelling of coastal algal blooms." *Ecological Modelling* 159 (2003): 179-201.

Melesse, Assefa M., Jayachandran Krishnaswamy, and Keqi Zhang. "Modeling Coastal Eutrophication at Florida Bay using Neural Networks." *Journal of Coastal Research* 24, no. 2B (2008): 190-196.

Millie, David F., et al. "Coastal 'Big Data' and nature-inspired computation: Prediction potentials, uncertainties, and knowledge derivation of neural networks for an algal metric." *Estuarine, Coastal and Shelf Science* 125 (2013): 57-67.

Phlips, Edward J., and Susan Badylak. *Phytoplankton Abundance and Composition in the Indian River Lagoon.* Special Publication, Fisheries and Aquatic Sciences, University of Florida, Palatka: St. Johns River Water Management District, 2012, 1-31.

Phlips, Edward J., Susan Badylak, Mary Christman, Jennifer Wolny, Julie Brame, Jay Garland, Lauren Hall, Jane Hart, Jan Landsberg, Margaret Lasi, Jean Lockwood, Richard Paperno, Doug Scheidt, Ariane Staples, and Karen Steidinger. "Scales of temporal and spatial variability in the distribution of harmful algae species in the Indian River Lagoon, Florida, USA." *Harmful Algae* 10, no. 3 (2011): 277-290.

Phlips, Edward J., Susan Badylak, and T. Grosskopf. "Factors affecting the abundance of phytoplankton in a restricted subtropical lagoon, the Indian River Lagoon, Florida, USA." *Estuarine, Coastal and Shelf Science* 55 (2002): 385-402.

Phlips, Edward J., Susan Badylak, Mary C. Christman, and Margaret A. Lasi. "Climatic Trends and temporal Patterns of Phytoplankton Composition, Abundance, and Succession in the Indian River Lagoon, Florida, USA." *Estuaries and Coasts* 33 (2010): 498-512.

Phlips, Edward J., Susan Badylak, Margaret A. Lasi, Robert Chamberlain, Whitney C. Green, Lauren M. Hall, Jane A. Hart, Jean C. Lockwood, Janice D. Miller, Lori J. Morris, Joel S. Steward.. "From Red Tides to green and Brown Tides: Bloom Dynamics in a Restricted Subtropical Lagoon Under Shifting Climatic Conditions." *Estuaries and Coasts*, 2014: (in press).

Sigua, Gilbert C., Joel S. Steward, and Wendy A. Tweedale. "Water-Quality Monitoring and Biological Integrity Assessment in the Indian River Lagoon, Florida: Status, Trends, and Loadings (1988-1994)." *Environmental Management* 25, no. 2 (2000): 199-209.

Singh, Kunwar P., Ankita Basant, Amrita Malik, and Gunja Jain. "Artificial neural network modeling of the river water quality - A case study." *Ecological Modelling* 220 (2009): 888-895.

Smith, Ned P. "Tidal and Nontidal Flushing of Florida's Indian River Lagoon." *Estuaries* 16, no. 4 (1993): 739-746.

St. Johns River Water Management District. *Aquatic grasses are a vital part of the water world.* April 5, 2013. http://www.floridaswater.com/aquaticgrasses/ (accessed March 9, 2014).

—. *Indian River Lagoon National Estuary Program.* May 14, 2010. http://www.floridaswater.com/itsyourlagoon/index.html (accessed April 21, 2011).

—. *The Health and Future of this Estuary of National Significance.* n.d. http://www.sjrwmd.com/irlinsert/index.html (accessed April 5, 2011).

—. *The Indian River Lagoon: an estuary of national significance.* October 4, 2013. http://floridaswater.com/itsyourlagoon/ (accessed March 9, 2014).

—. *The 'perfect storm' and 2011 superbloom.* May 16, 2013. http://floridaswater.com/itsyourlagoon/2011superbloom.html (accessed March 14, 2014).

Steele, J. H. "Environmental Control of Photosynthesis in the Sea." *Limnology and Oceanography* 7, no. 2 (1962): 137-150.

Velo-Suarez, L., and J. C. Gutierrez-Estrada. "Artificial neural network approaches to one-step weekly prediction of *Dinophysis acuminata* blooms in Huelva (Western Andalucía, Spain)." *Harmful Algae* 6, no. 3 (2007): 361-371.

Virnstein, Robert W., and Patricia A. Carbonara. "Seasonal Abundance and Distribution of Drift Algae and Seagrasses in the Mid-Indian River Lagoon, Florida." *Aquatic Botany* 23, no. 1 (1985): 67-82.

CHAPTER 2: USING ARTIFICIAL NEURAL NETWORKS TO DESCRIBE AND PREDICT *PYRODINIUM BAHAMENSE* BLOOMS IN THE INDIAN RIVER LAGOON

Introduction

Pyrodinium bahamense

Pyrodinium bahamense, a bioluminescent photosynthetic dinoflagellate found in tropical and subtropical marine environments, was first described from a sample collected in the Bahamas in 1906 (Seliger, et al. 1970; Plate 1906). Since then, *P. bahamense* has been divided into two varieties based on morphology: var. *compressum* and var. *bahamense* (Steidinger, Tester and Taylor 1980). The two varieties of *P. bahamense* do not appear to have overlapping geographical ranges; *P. bahamense* var. *compressum* has only been collected from the Pacific Ocean, while *P. bahamense* var. *bahamense* is primarily found in the western Atlantic (Balech 1985).

Since the early 1970s, *Pyrodinium bahamense* var. *compressum* has been known to produce saxitoxins, which can accumulate in shellfish and cause Paralytic Shellfish Poisoning (PSP) (Worth, Maclean and Price 1975). Cases of human intoxication via contaminated shellfish were first recorded in Papua New Guinea, but have since spread to many other countries in the Indo-Pacific region (Azanza and Taylor 2001; Worth, Maclean and Price 1975). Saxitoxin presents a serious health risk: mild cases of poisoning cause symptoms within 30 minutes, and extreme cases may cause death by respiratory paralysis anywhere from two to 24 hours after exposure (Azanza and Taylor 2001). Between 1976 and 1999, over 3100 cases of PSP were reported in south-east Asia, with at least 178 fatalities (Azanza and Taylor 2001). In culture, an isolate of *P. bahamense* var. *compressum* collected from Malaysia produced five saxitoxin variants, the quantities of which varied with salinity, temperature, and light intensity (Usup, Kulis and Anderson 1994).

Persistent dense blooms of *P. bahamense* var. *bahamense* have been found in the West Indies, Bahamas, and Puerto Rico for over sixty years, and the ongoing bioluminescence has even become a tourist attraction (Seliger, et al. 1970). However, despite the close relationship between the two varieties, until 2002 toxicity had only been attributed to *P. bahamense* var. *compressum*, and PSP toxicity from *P. bahamense* was not a concern in the Atlantic (Azanza and Taylor 2001).

Pyrodinium bahamense in the Indian River Lagoon

Food poisoning incidents and the subsequent discovery of potentially toxic HAB species spurred the cooperative project which produced the Core Data Set. Between the start of 2002 and mid-year 2004, 28 cases of food poisoning were linked to puffer fish harvested in the Indian River Lagoon, on the eastern coast of Florida (Landsberg, et al. 2006). The symptoms were identical to those of traditional puffer fish poisoning (PFP), common to Japan, caused by tetrodotoxin (TTX) naturally found in some species of puffer fish (Landsberg, et al. 2006).

Analysis of the remainder of one of the fillets associated with a poisoning event by liquid chromatography-mass spectrometry (LC-MS) revealed the presence of the paralytic shellfish poisoning (PSP) toxins saxitoxin (STX), decarbamoylsaxitoxin (dcSTX), and gonyautoxin-5 (GTX5), but failed to yield any TTX (Quilliam, et al. 2004). However, these are some of the same toxins found in the culture of *P. bahamense* var. *compressum* studied by Usup, et al. (1994). Puffer fish harvesting in the IRL was subsequently banned by the Florida Fish and Wildlife Conservation Commission (FWC) (Landsberg, et al. 2006). Further monitoring of wild-caught puffer fish that spanned three years revealed lasting toxicity in multiple tissues. Captive-held fish also maintained toxicity of skin secretions for over a year (Landsberg, et al. 2006).

Because saxitoxins had not previously been found in Florida marine environments, a rigorous survey of potential toxin producing organisms was initiated in the IRL (Landsberg, et al. 2006). From phytoplankton collected during this survey, eleven *P. bahamense* cultures were established and, when tested, provided evidence of the source of the toxins identified previously as responsible for the cases of human intoxication (Landsberg, et al. 2006). Concurrent with Landsberg’s study was a five year effort by Phlips et al. (2004) to characterize phytoplankton species and dynamics in the IRL. In addition to what appeared to be an increasing abundance of *P. bahamense*, the diatom *Pseudo-nitzschia pseudodelicatissima* was also found (E. J. Phlips, et al. 2004).   *P. pseudodelicatissima* is known to produce the neurotoxin domoic acid (DA), the cause of amnesiac shellfish poisoning (ASP), elsewhere in the United States, and was found in several samples at concentrations high enough to exceed risk guidelines set forth in several other countries (E. J. Phlips, et al. 2004). Though, much like saxitoxin prior to 2002, there is no evidence to date to support the production of domoic acid in the IRL.

The cases of PSP linked to puffer fish prompted the formation of a task force and the subsequent decision to begin a cooperative study to investigate algal blooms in the area and their relation to saxitoxin puffer fish poisoning, dolphin and turtle diseases, and possible threats to human health (J. Landsberg 2010). The surveys conducted by Landsberg, et al. (2006) and Phlips, et al. (2004) provided valuable information and guidance for the cooperative study. The result of this study is the Core Data Set, a collection of over 600 water samples spanning five years and comprising at least 10 environmental or chemical parameters for each sample.

To generate the Core Data Set, samples were collected twice monthly at six sites associated with a range of waterbody/watershed size ratios, site characteristics, and surrounding urbanization (Figure 2.1) (E. J. Phlips, et al. 2011). Average time between sampling was 15 days. Salinity and temperature were measured in the field with either a YSI or Hach/Hydrolab multi-probe, and water samples were taken using an integrated tube sample to within 0.1 meter of the bottom (E. J. Phlips, et al. 2011). Samples were then split and preserved with either Lugol’s solution or gluteraldehyde, with unpreserved aliquots frozen for nutrient analysis (E. J. Phlips, et al. 2011). Phytoplankton were enumerated in Lugol's preserved samples using inverted phase contrast microscopy (E. J. Phlips, et al. 2011).

Toxin-producing blooms of *P. bahamense* can threaten human health and have negative effects on local industry. As such, the ability to describe the dynamics of this species within a complex ecosystem is highly desirable, and would allow preemptive measures to minimize such impacts. While the Core Data Set is large and diverse enough to support analysis using any number of model types, Artificial Neural Network (ANN) modeling requires very little data manipulation prior to application and does not require variables to fit assumptions, such as normality. Additionally, because the model determines the weight of each input variable, little to no knowledge of the ecological system as a whole is required of the user prior to building the model (Millie, et al. 2012). Though remote monitoring was not used in generating the Core Data Set, much of the ecological and chemical data in the dataset could have been collected remotely via autonomous sampling platforms, as in Millie, et al. (2013), and transmitted using a satellite uplink to researchers off-site. Thus, I have used ANNs to estimate current *P. bahamense* density, using input variables which could be available remotely, as well as future density, using time-lagged data as input.

Materials and Methods

Samples for phytoplankton enumeration were collected as described previously; briefly, water was collected at each site using a vertical integrating tube sampler to within 0.1 m of the bottom, split and immediately preserved with either Lugol's solution or gluteraldehyde for later analysis (E. J. Phlips, et al. 2011). Lugol's preserved samples were settled in cylindrical chambers (diameter = 19mm) and analyzed using a Leica phase contrast inverted microscope (Phlips, et al., 2011; Figure 2.2). Phytoplankton were identified and enumerated at 400x; a minimum of 30 grids were counted and if 100 cells of a single taxon were not identified, and counting continued until either 100 cells were observed or 100 grids were counted, whichever occurred first (E. J. Phlips, et al. 2011). If identification was difficult at 400x, additional techniques, such as the squash method and scanning electron microscopy, were used (Phlips, et al., 2011; Figure 2.2). *P. bahamense* (Figure 2.2) enumeration data for each sample were used, along with nutrient and environmental data, to build the ANNs.

Ongoing phytoplankton surveys, in contrast to event response data, are often zero-heavy and therefore fail tests for normal distribution; in the Core Data Set, *P. bahamense* cells were absent from 56% of samples (E. J. Phlips, et al. 2004). For further confirmation, basic descriptive statistics were generated for *P. bahamense* count data using SPSS Statistics 22 (IBM Corporation 2013).

Building the Artificial Neural Networks

Both descriptive and predictive models of *P. bahamense* were constructed. Models intended to describe densities of *P. bahamense* concurrent with nutrient and environmental data from each sample were initially built using either one or two hidden layers, both with and without cross validation data, and trained using each of the four learning algorithms most commonly found in text (Levenberg-Marquart, conjugate gradient, step, and momentum) (Singh, et al.

2009; Millie, et al. 2013). Going forward, the more complex model structure with two hidden layers was used, along with cross validation data sets, and three of the four learning algorithms (conjugate gradient, step, and momentum) were tested for each model in an attempt to find the greatest agreement. Samples were randomized (n = 650) and 60% used for training the model, 15% for cross-validation, and 25% for testing the model's performance. For each model, the number of processing elements (PEs) in each hidden layer was varied from 2 to 16 as the network trained, to produce the most efficient and accurate model. After the models were trained, each was tested and performance metrics were produced by the software. The most accurate version, as determined by the given metrics, was then re-trained as many as ten times to further refine the algorithm, and the best of these replicates was retained for sensitivity testing. Sensitivity testing is included in the software, and the program will vary the values of each of the input variables independently of the others and record the change in output for analysis; when used on a model that describes the dataset well, this will show which input variable(s) have the greatest impact on the output. The user may specify the range of variation for each variable; as before, I used both one and two standard deviations about the mean. Gray box analysis was performed only on the most accurate descriptive model in an effort to visualize influences contributing to *P. bahamense* density in real time.

Descriptive models of *P. bahamense* counts were built using the following variables as input: temperature, salinity, dissolved oxygen, pH, ratio of secchi depth to total depth, total nitrogen (TN), total phosphorus (TP), ratio of total nitrogen to total phosphorous (TN/TP), chlorophyll *a*, and rainfall since last sample. As models often have difficulty describing zero-heavy datasets, a classification network was also built in an attempt to describe the simple presence/absence of *P. bahamense* cells. Another model deleted the zero count values from the data set, reducing it by roughly half (n = 286), in an effort to use the same input variables to describe cell density. Finally, the data set, less the zero values, was binned into low, medium, and high cell densities (low = 333-1,000 c/L, medium = 1,001-12,600 c/L, high = 12,601-1,451,300 c/L), each designation containing approximately one third of the samples, and a classification network built to describe the resulting dataset.

Predictive models used the more complex structure with two hidden layers and cross-validation data, and followed the same development as descriptive models. As before, the desired output was shifted one sample period, such that models attempted to describe counts using the chemical and environmental data from the previous sample date, less *P. bahamense* counts. Average time between samples at each site was 15 to 16 days. Predictive models were developed for all alternate datasets constructed for descriptive models (presence/absence, counts only, binned counts).

Gray Box Analysis

Gray box analysis was done only for the descriptive presence/absence classification model. Gray box analysis was adapted from an Excel template provided by Dr. Gary Weckman (Russ College of Engineering and Technology, Ohio University) to suit the new model structure: two output variables instead of one. The “best” file, with weights for each input as it relates to the next model layer, was used to reconstruct the complex web of interactions within the model. I generated both a Neuro Interpretation Diagram (NID), depicting the relative contribution of each variable to each node in each layer, and a pie chart showing the relative contribution of each input variable to the output layer.

Results

Values for skewness and kurtosis were positive (11.3 and 178.4, respectively; Table 2.1) and varied substantially from those expected when the data is normally distributed, supporting the use of a distribution-free modeling method such as an ANN.

The most accurate description of count values, including all zero values, was produced using the step learning algorithm with two hidden layers and cross-validation data; both hidden layers in this model contained 16 nodes (MSE = 11,032,919,172, NMSE = 0.601, r = 0.640; Figure 2.3 and Table 2.2). The trends observed in this first model were used as a basis for building all following models: two hidden layers described the data better than one, and the use of a cross-validation data set improved model performance. Despite these aids, attempts to describe both specific count and zero values were largely unsuccessful. This was seen also in the predictive models, in which case the best correlation was obtained using the momentum learning algorithm, having 16 nodes in each of the two hidden layers (MSE = 1,574,486,616, NMSE = 1.036, r = 0.385; Figure 2.4 and Table 2.3).

Greater success was obtained when the count data was reduced to presence/absence values and modeling was attempted using a classification structure. The momentum learning algorithm produced the most accurate descriptive model in the fourth iteration, having 10 nodes in the first hidden layer and 8 in the second (Present: MSE = 0.160, r = 0.597, % correct = 74.67; Absent: MSE = 0.160, r = 0.598, % correct = 84.09; Table 2.4 and Table 2.5). Sensitivity analysis done using the NeuroSolutions software indicated that the three input variables that most affected both the presence and absence of *P. bahamense* were temperature, pH, and dissolved oxygen (Figure 2.5 and Table 2.6). Gray box analysis produced an NID showing how each input and node relates to the following layer, as well as a pie-chart of connected weights indicating the relative importance of each input variable to the output layer (Figure 2.6). In contrast to the sensitivity analysis performed within NeuroSolutions, the analysis of connected weights indicated that the most important influences on the final output were relative secchi depth, pH, dissolved oxygen, and the TN/TP ratio (Figure 2.7).

Predictive models using the presence/absence classification dataset were also relatively successful, achieving 65.7% and 78.5% correct values for presence and absence, respectively. These results were obtained using the momentum learning algorithm in the seventh iteration; the resulting model had eleven nodes in the first hidden layer and seven in the second (Present: MSE = 0.185, r = 0.532, % correct = 65.7; Absent: MSE = 0.185, r = 0.531, % correct = 78.5; Table 2.7 and Table 2.8).

When zero values were eliminated and the model attempted to describe only samples in which *P. bahamense* are present, efforts were largely unsuccessful. For both descriptive and predictive models, MSE was high and r values indicated that the model did not follow the data trends well. The conjugate-gradient learning algorithm produced the most accurate descriptive model, having 18 nodes in the first layer and three in the second (MSE = 4,689,232,952, NMSE = 0.784, r = 0.686; Figure 2.8 and Table 2.9). The most accurate predictive model was produced by the same algorithm, with 15 nodes in the first hidden layer and 13 in the second (MSE = 3,347,175,751, NMSE = 1.122, r = 0.662; Figure 2.9 and Table 2.10).

Removing zero values and binning *P. bahamense* counts into low, medium, and high ranges (low = 333-1000 cells/L, n = 98; medium = 1001-12,600 cells/L, n = 93; high = 12,601-1,451,300 cells/L, n = 95), then using a classification model structure produced better results, but still did not describe the data very well. The step learning algorithm based model which produced the best fit had 14 nodes in each of the two hidden layers (Low: MSE = 0.180, r = 0.515, % correct = 59.3; Medium: MSE = 0.210, r = 0.202, % correct = 50.0; High: MSE = 0.184, r = 0.411, % correct = 60.9; Table 2.11 and Table 2.12). The most accurate predictive model was produced by the momentum learning algorithm and had six nodes in the first hidden layer and 16 in the second (Low: MSE = 0.190, r = 0.380, % correct = 42.9; Medium: MSE = 0.233, r = 0.125, % correct = 53.8; High: MSE = 0.159, r = 0.537, % correct = 70.8; Table 2.13 and Table 2.14).

Discussion

These results indicate that attempting to describe the dynamics of an individual species within a complex ecosystem, even with a relatively large data set, is not simple. The parameters measured and calculated as part of the Core Data Set were determined by heuristic knowledge, which may be incomplete in this case. The input variables did not include any measurements or observations of lateral or higher trophic levels, and therefore cannot account for predation or competition pressure. Philps, et al. (2010) also found poor model agreement ($R^2 = 0.43$) when using similar input variables in an all-possible-models regression of total dinoflagellate biovolume. Input parameters for the ANN models were chosen based on those which may be collected remotely via an autonomous monitoring platform and transmitted wirelessly back to researchers, or those available using other forms of remote sensing. Additionally, different sampling sites may have different factors influencing the density and growth rate of *P. bahamense*, as demonstrated by Phlips, et al. (2010), that these larger scale models do not accurately represent.

Despite these limitations, one of the models, that which described whether *P. bahamense* was present or absent from samples, was remarkably accurate, producing the correct answer for 75% and 84% of exemplars, present and absent, respectively (Table 2.5). While the model cannot determine how dense the population will be, the ability to remotely estimate presence and absence would be of immense value to managers in targeting sampling efforts.

Sensitivity analysis for this classification model indicated that primary factors influencing the presence or absence of *P. bahamense* are temperature, pH, and dissolved oxygen. Gray box analysis produced connected weights that suggest the primary influencers are relative secchi depth, pH, and dissolved oxygen. Similarly, restricted all-possible-models regressions identified temperature and secchi depth for total dinoflagellate biovolume at a site near Core Site 2 (E. J. Phlips, et al. 2010). These results provide little novel insight into driving factors of *P. bahamense* presence or density. Laboratory-based enrichment experiments have demonstrated that dinoflagellate growth is limited by temperature and pH; that this is found as a primary influence merely indicates that, in the sample areas, these environmental parameters occasionally deviate from those optimal for logarithmic growth (Usup, Kulis and Anderson 1994; Hansen, Lundholm and Rost 2007). Relative secchi depth and dissolved oxygen may both be affected by high densities of phytoplankton, so rather than illustrating a causal relationship between those inputs and *P. bahamense* densities, these analyses more likely indicate a correlation. Interestingly, factors found by other studies to impact phytoplankton abundance were not identified as drivers by these models: Phlips, et al. found that weather conditions and nutrient availability in the IRL were most likely to impact phytoplankton community structure and abundance (2002), and total dinoflagellate biovolume (2010). It is most likely these differences are a result of the different analyses and modeling techniques that were used for each study, as well as differing input variables.

In my attempt to build models based on parameters which can be collected remotely, I omitted several variables from the input layer that were collected as part of the Core Data Set. Among these were total plankton community counts and cyst data from sediment samples, which could have provided an estimate of zooplankton grazing or recruitment. Phlips et al. found that grazing and competition affect total phytoplankton biomass (2002) and total dinoflagellate biovolume (2010) in the IRL, and *P. bahamense* resting spores, or cysts, have long been known to contribute to pelagic communities (Wall and Dale 1969).

Though none of the factors indicated as influential by sensitivity or gray box analyses are likely to be so, the model is still relatively accurate. In all likelihood, this is because the driving factor(s) that remain hidden are either driven by or covary with those inputs which were identified as significant. This may also indicate why one model type, the presence/absence classification, was more accurate than any other which attempted to describe or predict *P. bahamense* densities. These results, when viewed together, indicate that further research is necessary to begin to understand *P. bahamense* bloom dynamics, and that in doing so, researchers must look beyond parameters traditionally considered important to phytoplankton growth to find underlying driving factors.

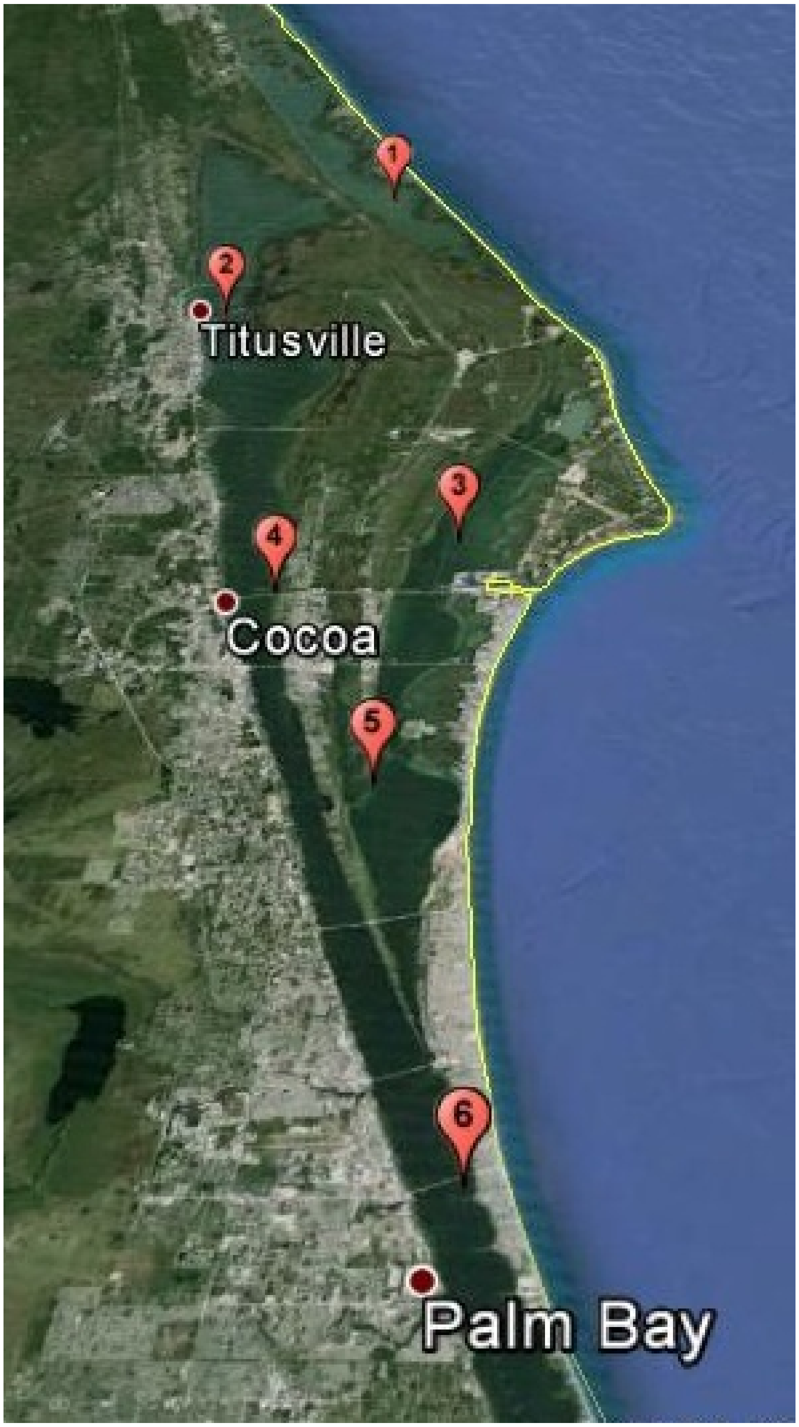


Figure 2.1: Core Data Set fixed sampling locations, showing the range of water body sizes, relation to areas of high population density, and relation to saltwater inputs.

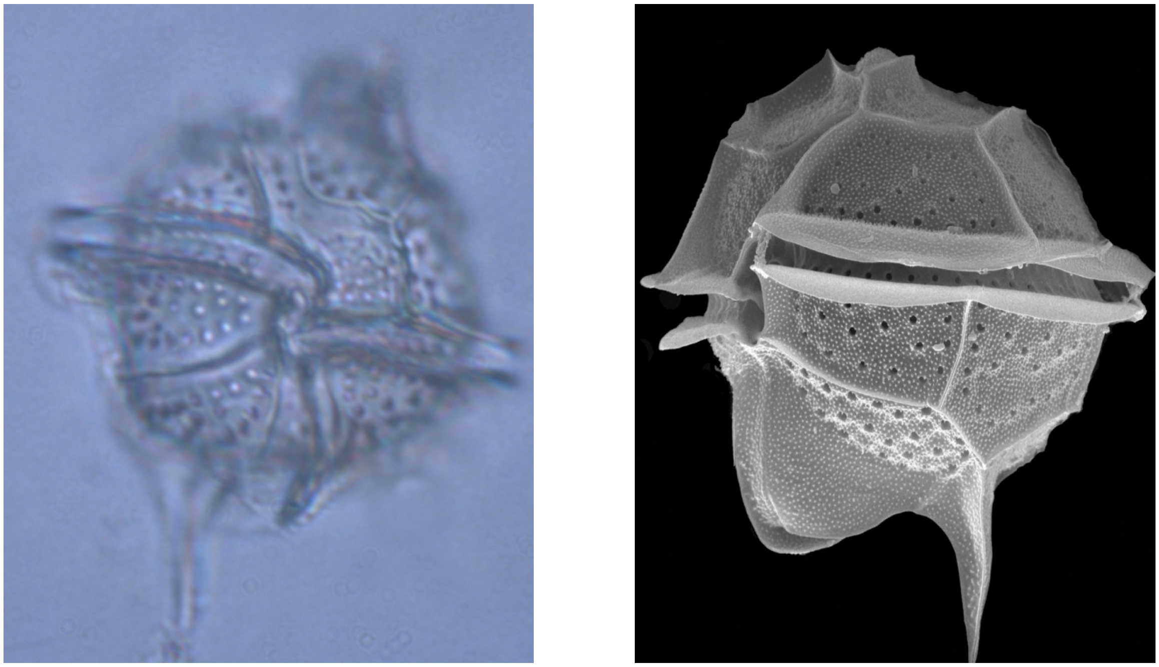


Figure 2.2: Light and Scanning Electron Microscope images of *Pyrodinium bahamense* var. *bahamense* (images courtesy Florida Fish & Wildlife Conservation Commission)

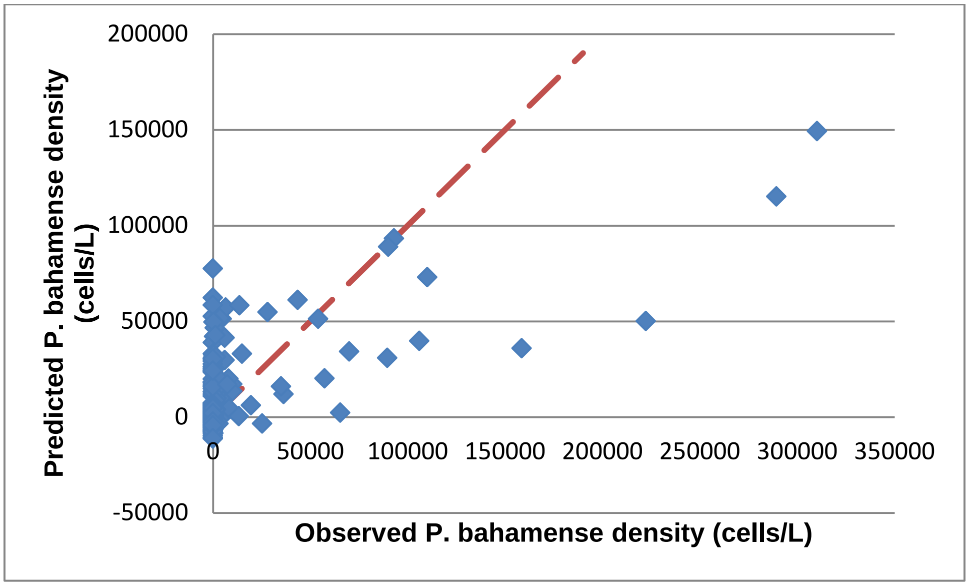


Figure 2.3: *P. bahamense* count descriptive regression model performance: predicted values are plotted against observed values; red line has the equation $y = x$ and is shown for reference.

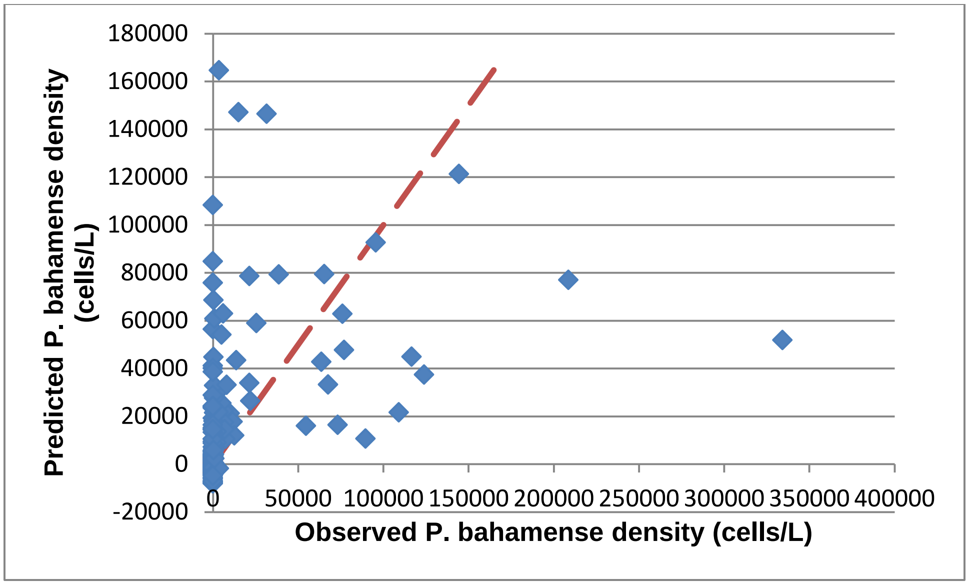


Figure 2.4: *P. bahamense* count predictive regression model performance: predicted values are plotted against observed values; red line has the equation $y = x$ and is shown for reference.

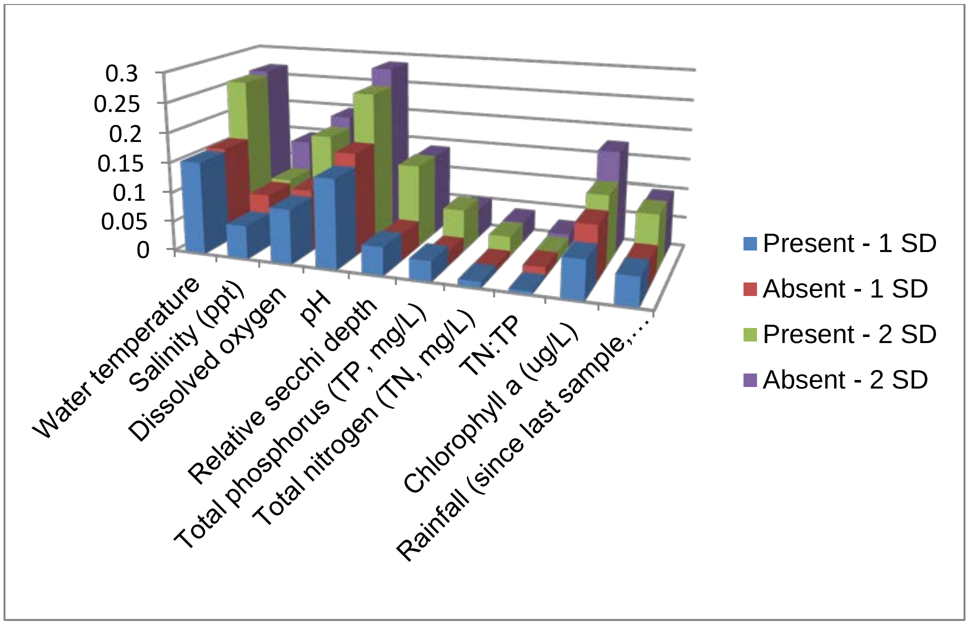


Figure 2.5: Sensitivity analysis results for descriptive classification model. Each variable was altered by 1 and then 2 standard deviations around the mean value while other variables were held constant; changes to the model output are represented by the bar graph below.

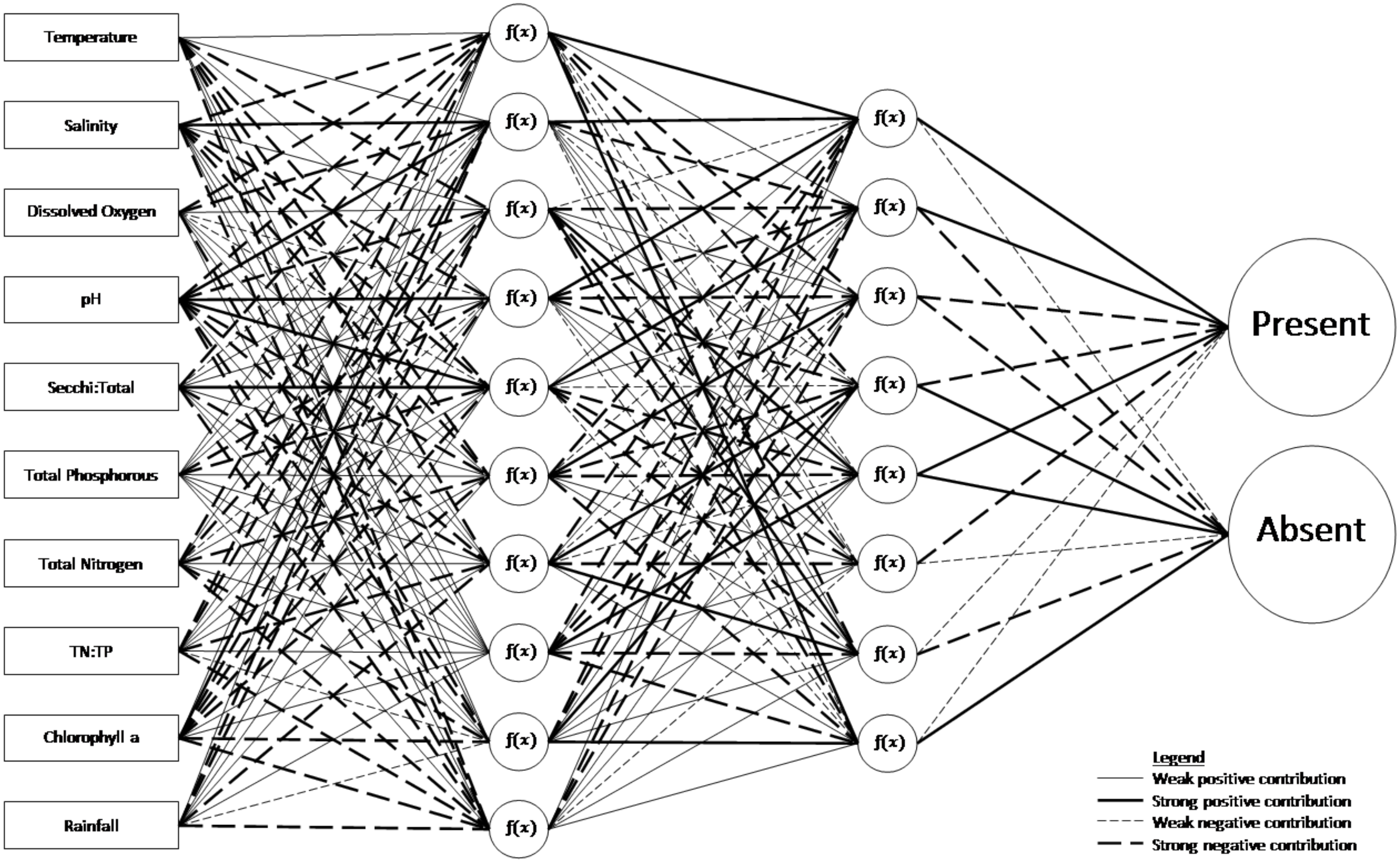


Figure 2.6: Neural Interpretation Diagram for descriptive presence/absence model generated using metadata produced by NeuroDimensions software in the course of training the ANN.

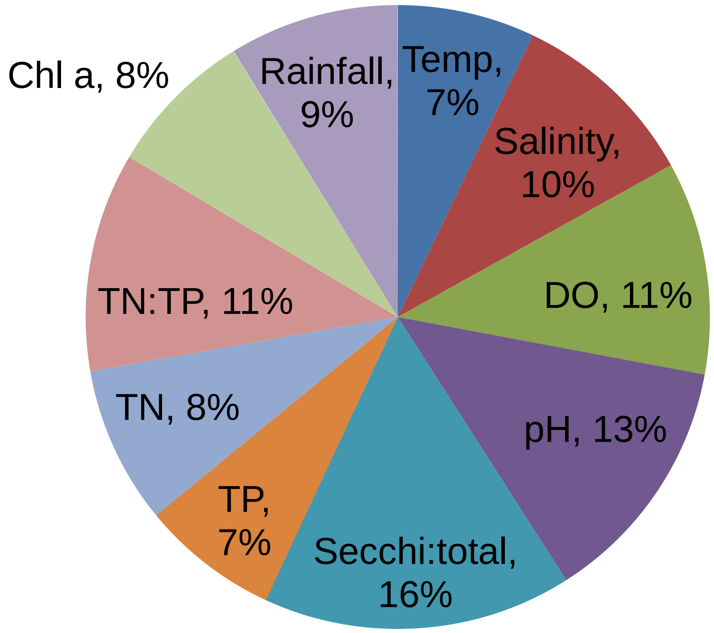


Figure 2.7: Connected weights pie chart for descriptive presence/absence model showing contribution of each input variable to output result as determined by tracing the weights of each layer through the ANN.

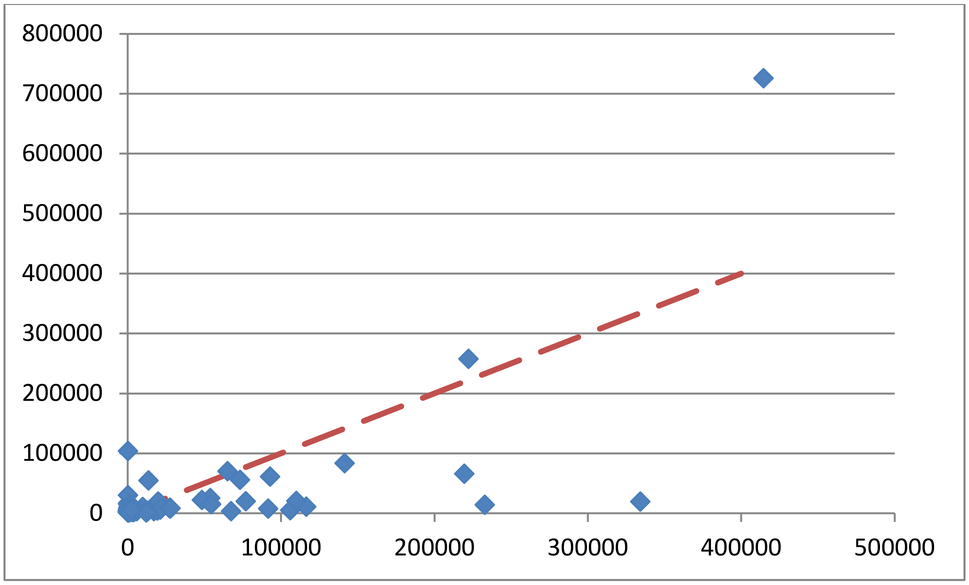


Figure 2.8: *P. bahamense* zero-removed counts descriptive regression model performance: predicted values are plotted against observed values; red line has the equation $y = x$ and is shown for reference.

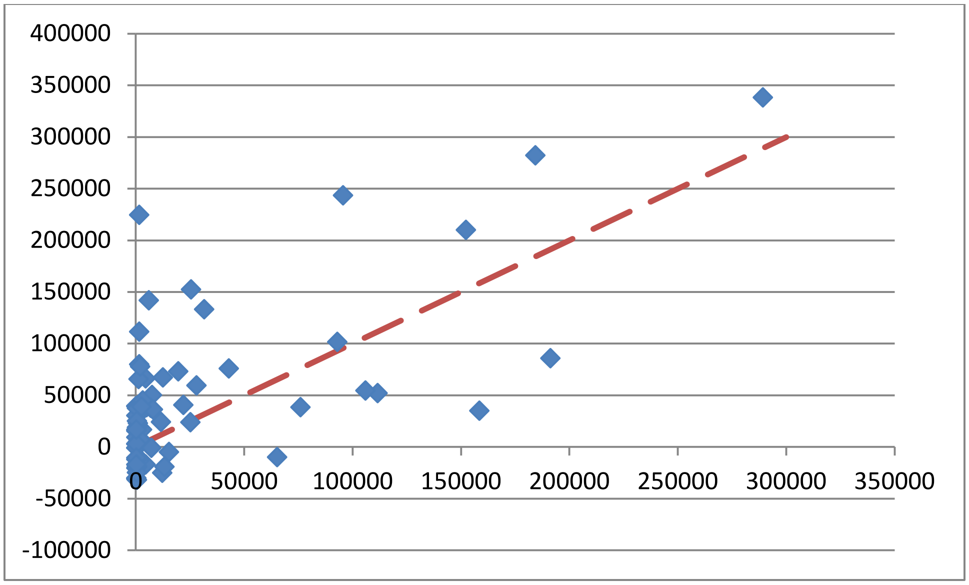


Figure 2.9: *P. bahamense* zero-removed counts predictive regression model performance: predicted values are plotted against observed values; red line has the equation $y = x$ and is shown for reference.

Table 2.1: Descriptive statistics generated for all *P. bahamense* count values. High positive values for skewness and kurtosis indicate deviation from a normal distribution.

			Statistic	Std. Error
Count	Mean		17927.17	3096.735
	95% Confidence Interval for Mean	Lower Bound	11846.36	
		Upper Bound	24007.99	
	5% Trimmed Mean		5846.03	
	Median		0	
	Variance		6.243E+09	
	Std. Deviation		79012.265	
	Minimum		0	
	Maximum		1451300	
	Range		1451300	
	Interquartile Range		2000	
	Skewness		11.344	0.096
	Kurtosis		178.407	0.191

Table 2.2: *P. bahamense* count descriptive regression model performance metrics; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***Count***
MSE	1103291972
NMSE	0.600863602
MAE	18514.98776
Min Abs Error	16.64680134
Max Abs Error	174281.5936
r	0.639872115

Table 2.3: *P. bahamense* count predictive regression model performance metrics; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***Count at next sample date***
MSE	1574486616
NMSE	1.035728648
MAE	19436.21704
Min Abs Error	42.12807662
Max Abs Error	282419.8915
r	0.385469674

Table 2.4: Output vs. Desired results of descriptive classification model; desired values are in columns, model predicted values are in rows. Model prediction is correct for 79.8% of samples.

	Desired: Present	*Desired: Absent*
Output: Present	56	14
Output: Absent	19	74

Table 2.5: Performance metrics of descriptive classification model; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***Present***	***Absent***
MSE	0.160312423	0.160126533
NMSE	0.645354662	0.644606342
MAE	0.312350685	0.308842543
Min Abs Error	0.00097396	0.00696486
Max Abs Error	1.030765554	1.029867956
r	0.596867716	0.597941272
Percent Correct	74.66666667	84.09090909

Table 2.6: Sensitivity analysis results for descriptive classification model; numerical values calculated to describe sensitivity of the model output to changes within 1 and 2 standard deviations of the mean for each input variable.

Sensitivity 1 SD around mean	***Present***	***Absent***
TEMP	0.155040267	0.160030225
SALINITY	0.056822164	0.087636136
DISSOLVED_O2	0.092953591	0.103365106
pH	0.149867694	0.171482146
sechhi:total depth	0.048361016	0.04462895
TP (mg/L)	0.035110689	0.024204439
TN (mg/L)	0.011152582	0.010602528
TN:TP	0.005797622	0.017332029
CHLa est (ug/L)	0.065911198	0.095786105
Rainfall (since last sample, inches)	0.04936855	0.047455105

Sensitivity 2 SD around mean	***Present***	***Absent***
TEMP	0.260227977	0.267216609
SALINITY	0.093270542	0.142770228
DISSOLVED_O2	0.17627705	0.19485387
pH	0.255003834	0.284816976
sechhi:total depth	0.138606999	0.136602685
TP (mg/L)	0.071410903	0.050334828
TN (mg/L)	0.034503226	0.035631222
TN:TP	0.01759006	0.02302226
CHLa est (ug/L)	0.120913401	0.17123185
Rainfall (since last sample, inches)	0.09753528	0.094160839

Table 2.7: Output vs. Desired results of predictive classification model; desired values are in columns, model predicted values are in rows. Model prediction is correct for 73.0% of samples.

	Desired: present	*Desired: absent*
Output: present	46	20
Output: absent	24	73

Table 2.8: Performance metrics of predictive classification model; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***present***	***absent***
MSE	0.184800606	0.18500588
NMSE	0.754219249	0.755057023
MAE	0.302513393	0.301382952
Min Abs Error	0.002894105	0.000516863
Max Abs Error	1.002910177	0.996266062
r	0.53233458	0.530819039
Percent Correct	65.71428571	78.49462366

Table 2.9: *P. bahamense* zero-removed counts descriptive regression model performance metrics; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***Count***
MSE	4689232952
NMSE	0.783686942
MAE	30925.85216
Min Abs Error	58.38421847
Max Abs Error	315353.3774
r	0.685841923

Table 2.10: *P. bahamense* zero-removed counts predictive regression model performance metrics; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***count at next sample date***
MSE	3347175751
NMSE	1.122395325
MAE	41219.13204
Min Abs Error	1113.274814
Max Abs Error	222732.858
r	0.662484184

Table 2.11: Output vs. Desired results of descriptive binned counts classification model; desired values are in columns, model predicted values are in rows. Model prediction is correct for 56.9% of samples.

	Desired: Low	*Desired: Mid*	*Desired: High*
Output: Low	16	4	2
Output: Mid	4	11	7
Output: High	7	7	14

Table 2.12: Performance metrics of descriptive binned counts classification model; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***Low***	***Mid***	***High***
MSE	0.17960198	0.210296879	0.183737797
NMSE	0.766301783	0.991071835	0.845161261
MAE	0.326472983	0.414795133	0.367586985
Min Abs Error	0.006736412	0.081347739	0.006113535
Max Abs Error	0.984232019	0.878320082	0.948885766
r	0.514621675	0.201607425	0.411171822
Percent Correct	59.25925926	50	60.86956522

Table 2.13: Output vs. Desired results of predictive binned counts classification model; desired values are in columns, model predicted values are in rows. Model prediction is correct for 56.3% of samples.

	Desired: low	*Desired: mid*	*Desired: high*
Output: low	9	6	1
Output: mid	10	14	6
Output: high	2	6	17

Table 2.14: Performance metrics of predictive binned counts classification model; MSE = mean squared error, NMSE = normalized mean squared error, MAE = mean absolute error.

Performance	***low***	***mid***	***high***
MSE	0.189676055	0.232988136	0.159349371
NMSE	0.910625708	1.003840338	0.712127817
MAE	0.307199183	0.441045704	0.319086521
Min Abs Error	0.002407694	0.043941965	0.000942164
Max Abs Error	0.958516836	0.841595932	1.013773925
r	0.379798083	0.12516893	0.537496042
Percent Correct	42.85714286	53.84615385	70.83333333

References

Azanza, Rhodora V., and F.J.R. Max Taylor. "Are Pyrodinium Blooms in the Southeast Asian Region Recurring and Spreading? A View at the End of the Millennium." *Ambio* 30, no. 6 (2001): 356-364.

Balech, Enrique. "A Revision of *Pyrodinium bahamense* Plate (Dinoflagellata)." *Review of Palaeobotany and Palynology* (Elsevier Science Publishers B.V.) 45 (1985): 17-34.

Hansen, P. J., N. Lundholm, and B. Rost. "Growth limitation in marine red-tide dinoflagellates: effects of pH versus inorganic carbon availability." *Marine Ecology Progress Series* 334 (2007): 63-71.

Landsberg, Jan H., et al. "Saxitoxin Puffer Fish Poisoning in the United States, with the First Report of *Pyrodinium bahamense* as the Putative Toxin Source." *Environmental Health Perspectives* 114, no. 10 (2006): 1502-1507.

Landsberg, Jan. *Monitoring of Toxic Algae in the Indian River Lagoon, Florida, USA.* Grant final report, Fish and Wildlife Research Institute, Florida Fish and Wildlife Conservation Commission, St Petersburg, FL: Florida Fish and Wildlife Conservation Commission, 2010, 1-38.

Millie, David F., et al. "Coastal 'Big Data' and nature-inspired computation: Prediction potentials, uncertainties, and knowledge derivation of neural networks for an algal metric." *Estuarine, Coastal and Shelf Science* 125 (2013): 57-67.

Millie, David F., Gary R. Weckman, William A. Young II, James E. Ivey, Hunter J. Carrick, and Gary L. Fahnenstiel. "Modeling microalgal abundance with artificial neural networks: Demonstration of a heuristic 'Grey-Box' to deconvolve and quantify environmental influences." *Environmental Modelling & Software* 38 (2012): 27-39.

Phlips, Edward J., Susan Badylak, S. Youn, and Karen Kelley. "The occurrence of potentially toxic dinoflagellates and diatoms in a subtropical lagoon, the Indian River Lagoon, Florida, USA." *Harmful Algae* 3 (2004): 39-49.

Phlips, Edward J., Susan Badylak, Mary Christman, Jennifer Wolny, Julie Brame, Jay Garland, Lauren Hall, Jane Hart, Jan Landsberg, Margaret Lasi, Jean Lockwood, Richard Paperno, Doug Scheidt, Ariane Staples, and Karen Steidinger. "Scales of temporal and spatial variability in the distribution of harmful algae species in the Indian River Lagoon, Florida, USA." *Harmful Algae* 10, no. 3 (2011): 277-290.

Phlips, Edward J., Susan Badylak, and T. Grosskopf. "Factors affecting the abundance of phytoplankton in a restricted subtropical lagoon, the Indian River Lagoon, Florida, USA." *Estuarine, Coastal and Shelf Science* 55 (2002): 385-402.

Phlips, Edward J., Susan Badylak, Mary C. Christman, and Margaret A. Lasi. "Climatic Trends and temporal Patterns of Phytoplankton Composition, Abundance, and Succession in the Indian River Lagoon, Florida, USA." *Estuaries and Coasts* 33 (2010): 498-512.

Plate, L. "*Pyrodinium bahamense* n. g., n. sp. die Leucht-Peridinee des "Feuersees" von Nassau, Bahamas." *Archiv fur Protistenkunde* 7 (1906): 411-428.

Quilliam, Michael, Dominik Wechsler, Steven Marcus, Bruce Ruck, Marleen Wekell, and Timothy Hawryluk. "Detection and Identification of Paralytic Shellfish Poisoning Toxins in Florida Pufferfish Responsible for Incidents of Neurologic Illness." Edited by K. A. Steidinger, J. H. Landsberg, C. R. Tomas and G. A. Vargo. *Harmful Algae 2002.* St. Petersburg, FL: Florida Fish and Wildlife Conservation Commission, Florida Institute of Oceanography, and Inter-governmental Oceanographic Commission of United Nations Educational, Scientific and Cultural Organization, 2004. 116-118.

Seliger, H. H., J. H. Carpenter, M. Loftus, and W. D. McElroy. "Mechanisms for the Accumulation of High Concentrations of Dinoflagellates in a." *Limnology and Oceanography* 15, no. 2 (1970): 234-245.

Singh, Kunwar P., Ankita Basant, Amrita Malik, and Gunja Jain. "Artificial neural network modeling of the river water quality - A case study." *Ecological Modelling* 220 (2009): 888-895.

Steidinger, K. A., L. S. Tester, and F. J. R. Taylor. "A redescription of *Pyrodinium bahamense* var. *compressa* (Bohm) stat. nov. from Pacific red tides." *Phycologia* 19 (1980): 329-334.

Usup, Gires, David M. Kulis, and Donald M. Anderson. "Growth and Toxin Production of the Toxic Dinoflagellate *Pyrodinium bahamense* Var. *compressum* in Laboratory Cultures." *Natural Toxins* 2 (1994): 254-262.

Wall, David, and Barrie Dale. "The "hystrichosphaerid" resting spore of the dinoflagellate *Pyrodinium bahamense*, Plate, 1906." *Journal of Phycology* 5 (1969): 140-149.

Worth, G. K., J. L. Maclean, and M. J. Price. "Paralytic Shellfish Poisoning in Papua New Guinea, 1972." *Pacific Science* 29, no. 1 (1975): 1-5.