Bayesian Statistical Analysis in Coastal Eutrophication Models: Challenges and Solutions

by

Farnaz Nojavan A.

Environment
Duke University

Date: _____

Approved:

_____
Song S. Qian, Co-Supervisor

_____
Gabriel Katul, Co-Supervisor

_____
Elizabeth A. Albright

_____
Marco Marani

_____
Craig Stow

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Environment
in the Graduate School of Duke University
2014

Abstract

Bayesian Statistical Analysis in Coastal Eutrophication Models: Challenges and Solutions

by

Farnaz Nojavan A.

Environment
Duke University

Date: _____

Approved:

_____
Song S. Qian, Co-Supervisor

_____
Gabriel Katul, Co-Supervisor

_____
Elizabeth A. Albright

_____
Marco Marani

_____
Craig Stow

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Environment in the Graduate School of Duke University
2014

Abstract

Estuaries interfacing with the land, atmosphere and open oceans can be influenced in a variety of ways by anthropogenic activities. Centuries of overexploitation, habitat transformation, and pollution have degraded estuarine ecological health. Key concerns of public and environmental managers of estuaries include water quality, particularly the enrichment of nutrients, increased chlorophyll *a* concentrations, increased hypoxia/anoxia, and increased Harmful Algal Blooms (HABs). One reason for the increased nitrogen loading over the past two decades is the proliferation of concentrated animal feeding operations (CAFOs) in coastal areas. This dissertation documents a study of estuarine eutrophication modeling, including modeling of major source of nitrogen in the watershed, the use of the Bayesian Networks (BNs) for modeling eutrophication dynamics in an estuary, a documentation of potential problems of using BNs, and a continuous BN model for addressing these problems.

Environmental models have emerged as great tools to transform data into useful information for managers and policy makers. Environmental models contain uncertainty due to natural ecosystems variability, current knowledge of environmental processes, modeling structure, computational restrictions, and problems with data/observations due to measurement error or missingness. Many methodologies capable of quantifying uncertainty have been developed in the scientific literature. Examples of such methods are BNs, which utilize conditional probability tables to describe the relationships among variables. This doctoral dissertation demonstrates

how BNs, as probabilistic models, can be used to model eutrophication in estuarine ecosystems and to explore the effects of plausible future climatic and nutrient pollution management scenarios on water quality indicators. The results show interaction among various predictors and their impact on ecosystem health. The synergistic effects between nutrient concentrations and climate variability caution future management actions.

BNs have several distinct strengths such as the ability to update knowledge based on Bayes' theorem, modularity, accommodation of various knowledge sources and data types, suitability to both data-rich and data-poor systems, and incorporation of uncertainty. Further, BNs' graphical representation facilitates communicating models and results with environmental managers and decision-makers. However, BNs have certain drawbacks as well. For example, they can only handle continuous variables under severe restrictions (1- Each continuous variable be assigned a (linear) conditional Normal distribution; 2- No discrete variable have continuous parents). The solution, thus far, to address this constraint has been discretizing variables. I designed an experiment to evaluate and compare the impact of common discretization methods on BNs. The results indicate that the choice of discretization method severely impacts the model results; however, I was unable to provide any criteria to select an optimal discretization method.

Finally, I propose a continuous variable Bayesian Network methodology and demonstrate its application for water quality modeling in estuarine ecosystems. The proposed method retains advantageous characteristics of BNs, while it avoids the drawbacks of discretization by specifying the relationships among the nodes using statistical and conditional probability models. The Bayesian nature of the proposed model enables prompt investigation of observed patterns, as new conditions unfold. The network structure presents the underlying ecological ecosystem processes and provides a basis for science communication. I demonstrate model development and

temporal updating using the New River Estuary, NC data set and spatial updating using the Neuse River Estuary, NC data set.

To Saed

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

AUC	Area Under Curve
BN	Bayesian Network
CAFO	Concentrated Animal Feeding Operation
cBN	Continuous Bayesian Network
Chl *a*	Chlorophyll *a*
CPT	Conditional Probability Table
DAG	Directed Acyclic Graph
DCERP	Defense Coastal/Estuarine Research Program
DIN	Dissolved Inorganic Nitrogen
DO	Dissolved Oxygen
EPA	Environmental Protection Agency
HAB	Harmful Algal Bloom
JAGS	Just Another Gibbs Sampler
MCMC	Markov chain Monte Carlo
NADP	National Atmospheric Deposition Program
NEEA	National Estuarine Eutrophication Assessment
NMP	Nutrient Management Plans
NOAA	National Atmospheric and Oceanic Administration
NPDES	National Pollutant Elemination System

NRE	New River Estuary
SSE	Sum of Squared Errors
TKN	Total Kjehldahl Nitrogen

Acknowledgements

Words cannot express my gratitude to my advisor, Professor Song Qian, for his endless support, continuous encouragement, and immense knowledge. I would like to also thank the other members of my dissertation committee, Professors Gabriel Katul, Elizabeth A. Albright, Marco Marani, and Craig Stow: for their extreme patience in the face of numerous obstacles. This dissertation would not have been possible without the unconditional support of Professor Gabriel Katul, the director of graduate studies of the Environmental Science & Policy Division at the Nicholas School of the Environment. I am also deeply grateful to other faculty at Duke University: Professors Robert Clemen, Peng Sun, Kenneth Reckhow, Jerry Reiter, Robert Winkler for their help in both research and coursework. My deepest gratitude goes to Dr. Richard Anderson, who initially got me interested in the applications of Bayesian Statistics in environmental sciences. I am also indebted to Professor Hans Paerl, Nathan Hall and Benjamin Peierls at the Institute of Marine Sciences, UNC Chapel Hill, who provided me with access to the data I used in the chapter 3 and 5 of my dissertation. My sincere gratitude goes to Professor Marc Alperin at the Department of Marine Sciences, UNC Chapel Hill, who is one of the best teachers that I have had in my life. I am sending a big thank to Professors Hugh Crumley and Douglas James for the amazing efforts in the *Preparing Future Faculty* and *Certificate in College Teaching* programs. I deeply appreciate their work at the Graduate School.

I am most grateful to Meg Stephens, whose friendly support, encouragement

and big heart helped me face all the obstacles and continue with my work. I will never forget her kindness. I thank my colleagues and friends, Ibrahim Alameddine, YoonKyung Cha, Roxolana Kashuba, Boknam Lee, Michaela Margida, and Kriss Voss, for the stimulating discussions. In particular, a big thank you to Yun Jian for her support, feedback, and friendship that helped me overcome setbacks and stay focused on my graduate study.

I express my deepest appreciation to my parents, Yadollah and Farrin, who taught me to reach for the stars. In a society full of restrictions, they taught me the sky is my limit. A special thank you to my sister, Behnaz, for being such a kind, sweet, and uplifting little sister. I would like to thank all my family whom I missed through the years of living abroad, especially my grandmother, Nana Tavous- may she rest in peace.

1

Introduction

Human population growth, particularly in the world's coastal regions, resulted in adverse changes in many coastal aquatic ecosystems, largely due to anthropogenic activities such as land use alterations, fertilizer use, industrial activity, and climatic perturbations. As population growth will likely continue in the future, a better understanding of the mechanisms of biogeochemical cycling in these vulnerable ecosystems is critical to evaluate and quantify the ecological consequences of current and predicted changes in climate and land use. Among coastal aquatic ecosystems, estuaries are unique systems to address questions of climate and land use change interactions. The close proximity to population centers makes estuaries particularly susceptible to adverse effects of anthropogenic activities that exacerbate nutrient loading and eutrophication. Furthermore, climate change due to increased greenhouse gas emissions can also influence the response of coastal aquatic ecosystems to the stress directly related to human activities. Models play a critical role in quantifying how environmental changes affect estuarine ecosystems' ecological health and water quality.

Environmental models can be categorized into deterministic and probabilistic models. Deterministic models ignore parameter variability. Therefore, a particu-

lar model input always produces the same output. Probabilistic models contain the inherent uncertainty in the environmental processes; a particular model input would produce a range of outputs- a probability distribution, due to the quantified model randomness. Mechanistic biogeochemical models have been used extensively in aquatic ecosystems research and are still a common research tool. Reliable predictions of system behavior is achieved when processes are adequately described mathematically. The fundamental assumption is that such mathematical formulation captures the dominant dynamics of the system, which is often difficult in complex ecological ecosystems. In terms of performance, mechanistic biogeochemical models for aquatic ecosystems perform well in predicting temperature and dissolved oxygen, moderately in prediction of limiting nutrients and phytoplankton, and relatively poorly for bacteria and zooplankton dynamics (Arhonditsis et al., 2004). The mechanics of eutrophication is so complex that it might be impossible to describe the system in sufficient detail mathematically; hence, probabilistic models might be more appropriate due to quantification of uncertainty when the processes of the ecosystem are unknown.

A great deal of effort has been expended to combine statistics and simple causal relationships to build water quality models with accurate uncertainty assessment since Beck (1987) highlighted the significance of uncertainty analysis in water quality modelling. An example of such recent developments is the application of Bayesian Networks (BNs) in environmental modeling. BNs are probabilistic graphical models, probabilistic models with a graphical representation of the conditional dependence between variables, suitable for uncertain and complex domains such as environmental ecosystems. BNs have several distinct strengths. The main strength of BNs lies in their knowledge updatability based on the Bayes' theorem. This is an advantage in the context of adaptive management of ecosystems. The BNs modularity enables integrating multiple system components or aspects of problem (e.g., science network

and management network in Johnson et al. (2010)). The modularity is beneficial in environmental modeling due to the complexity of natural ecosystems and the associated decision-making processes. BNs can accommodate various knowledge sources and data types (e.g., expert knowledge, previous data from the same system or other similar systems), with transparent definition of prior knowledge. Another advantage of BNs over other modeling approaches is suitability to both data-rich and data-poor systems. Environmental ecosystems often lack quality data associated with a new problem under investigation. Therefore, accommodating minimal data in conjunction with expert knowledge is a methodological advantage. The model can be developed with minimal data and, as more information becomes available, the model can be updated. Environmental modeling cannot be implemented without incorporating uncertainty, as it aims to explore complex ecosystems and provide support for management of natural resources. BNs explicitly represent uncertainty by conditional probability distributions for each node and the uncertainty is propagated through the model and presented in final results. These advantages of BNs resulted in an exponential rise in the application of BNs in ecological and environmental sciences over the last decade.

BNs also accommodate adaptive management framework well by their updatability. Adaptive management is a continuous process of decision making, where decisions are periodically revised as outcomes of management actions are observed under uncertain conditions of the ecosystem (Walters and Hilborn, 1978). The key concept in adaptive management is iterative learning. The requirements of iterative learning are (1) observing the ecosystem to gauge the impact of policies and management actions continuously; (2) communicating the ecosystem's status with policy makers and managers; (3) updating the management actions and recommendations. BNs meet such requirements (Walters, 1997). They also provide a straightforward ability to assimilate new information by using a Bayesian approach. In long-term

monitoring programs, new data become available every day/week/month. It would be greatly beneficial for managers/policy makers to update the model in time intervals depending on the frequency of sampling and the temporal resolution of the problem. Here, the posterior distribution calculated in the previous model run step would be considered an updated prior distribution. An updated posterior distribution can then be computed via Bayes' theorem using new data. Based on the updated posterior, the effectiveness of previous policies/strategies can be evaluated and new recommendations can be provided accordingly. While BNs are useful modeling tools, they have certain drawbacks such as discretization and acyclicity.

The Neuse River Estuary, North Carolina, is a shallow estuary with a history of eutrophication, HABs, and fish kill. The adoption of the Neuse Rules by the North Carolina Environmental Management Commission in 1998 expressed a regional concern over nitrogen loading in North Carolina's Neuse River Basin. One reason for the increased nitrogen loading over the past two decades is the proliferation of swine concentrated animal feeding operations (CAFOs) in the region. The synthesis in chapter 2 tracks the fate of the annual nitrogen input through food ($37.1 \pm 3.9\ Gg\ N/yr$) to swine CAFOs within the Neuse River Basin. I conducted a comprehensive literature review to assess the relative impact of nitrogen fates in swine CAFOs, and combined them into a nitrogen budget for the Neuse River Basin. I also characterized the uncertainty in the major nitrogen fates by using Monte Carlo simulation methods. I shofd that the most significant losses of nitrogen are through lagoon denitrification ($34.6\% \pm 7.7\%$), hog assimilation ($29.9\% \pm 3.2\%$), and ammonia volatilization to the atmosphere ($18.3\% \pm 5.6\%$). Our results also indicate that ,at most, $15\% \pm 9\%$ of the nitrogen export from the Neuse River Basin ($6.86\ Gg\ N/year$) could be due to swine lagoon seepage and $7\% \pm 4\%$ is due to sprayfield leaching.

I utilize a BN approach in Chapter 3 to intuitively present and quantify our current understanding of the complex physical, chemical, and biological processes

that lead to eutrophication in an estuarine ecosystem (New River Estuary, North Carolina, USA). The model is further used to explore the effects of plausible future climatic and nutrient pollution management scenarios on water quality indicators. The BN, through visualization of the network's structure, facilitates communication with managers/stakeholders who might not be experts in the underlying scientific disciplines. Moreover, the developed structure of the BN is transferable to other comparable estuaries. The BN's nodes are discretized using a new approach called moment matching method. The conditional probability tables of the variables are driven by a large dataset (four years). The results show interaction among various predictors and their impact on water quality indicators. The synergistic effects caution future management actions. This chapter provides a sufficient context for understanding the development process of the BNs and motivating the fourth chapter.

I witnessed the advantages and challenges of BNs through developing a BN for water quality modeling. In Chapter 4, I address the question of whether the choice of the discretization method impacts the final results of a BN. The fourth chapter of this doctoral dissertation presents the results of an experiment to compare different approaches of discretization during the development of BNs. BNs can only handle continuous variables under severe restrictions. The solution, thus far, to address this constraint has been discretizing variables. I designed an experiment to evaluate and compare the impact of common discretization methods on the final BN. The results indicate that the choice of discretization method severely impacts the model results. However, such an optimal method is inevitably case-specific. The conclusion of this chapter is extendable to other fields where BNs are applied. In the final chapter I focus on developing a continuous BN model.

The work presented in the final chapter is an attempt to address the problems of discretization. I propose a continuous variable BN methodology and demonstrate its

application in water quality modeling in estuarine ecosystems. The importance of uncertainty analysis, adaptive management, and science communication in ecological modeling gave rise to the application of BNs in the environmental sciences. Although BNs have gained popularity in recent years, their main restriction, discretization, has not yet been addressed. I propose a method that retains certain characteristics of BNs, while it avoids the drawbacks of discretization by specifying the relationships among the nodes using statistical and conditional probability models. The Bayesian nature of the proposed model enables prompt investigation of observed patterns, as new conditions unfold. The network structure presents the underlying ecological ecosystem processes and provides a basis for science communication. I demonstrate model development and temporal updating using the New River Estuary, NC data set and spatial updating using the Neuse River Estuary, NC data set. The proposed methodology in the final chapter is applicable to other contexts where BNs are often implemented.

2

Environmental Impacts of Swine Confined Animal Feeding Operations (CAFOs) on Atmospheric and Aquatic Resources within North Carolina's Neuse River Basin

2.1 Introduction

North Carolina has witnessed a large growth in commercial swine operations since the 1990s. In fact, North Carolina's hog inventory increased over 400% between 1987 (2.5 million hogs) and 2007 (10.1 million hogs) (USDA, 1992, 2007), while the corresponding increase in human population across roughly the same 20-year period (1990-2010) was only 44% (Census, 2010). Despite this staggering increase in hog inventory, the number of farms possessing hogs actually decreased by 59% between 1987 (6921 farms) and 2007 (2836 farms) (USDA, 1992, 2007). Consequently, the proportion of farms with more than 5000 hogs increased from just under 1% in 1987 to over 20% in 2007 (USDA, 1992, 2007). Thus, the dramatic rise in North Carolina's hog inventory has been attributed to the intensification of Concentrated Animal Feeding Operations (CAFOs), farms which stock animals at high density and

apply factory-like techniques to animal production.

The growth in swine CAFOs over the last two decades has primarily been concentrated in southeast North Carolina. Three major river basins in the state, the Neuse River Basin, Tar-Pamlico River Basin, and Cape Fear River Basin, contain roughly 80% of North Carolina's hog inventory (USDA, 2007). These rivers and the coastal estuaries to which they flow have been impacted by increased nutrient loading over the last 20 years. During the 1990s, the Neuse River estuary in particular exhibited symptoms of coastal eutrophication, including harmful algal blooms and fish kill (Burkholder et al., 1992; Paerl et al., 1998). Such observations precipitated the development of the Neuse Rules, promulgated by the North Carolina Environmental Management Commission in 1998 (15A NCAC 2B 1998). The Neuse Rules include a comprehensive nutrient reduction strategy that combines stormwater and agricultural best management practices with riparian buffer requirements. CAFOs following approved Nutrient Management Plans (NMPs), which document each CAFO's waste management protocols, are exempt from these rules. The US Environmental Protection Agency (EPA) also requires all CAFOs to prepare state-approved NMPs to be considered non-discharge operators under the National Pollutant Discharge Elimination System (NPDES) regulations (73 FR 70418 2008). These plans are, therefore, important regulatory instruments that act to protect human and ecological health.

Treatment technologies within CAFOs have been designed to reduce the environmental impact from the increased nutrient influx necessary to support large-scale animal farming. While treatment techniques vary, most facilities treat animal waste in anaerobic lagoons (Reddi, 2005) where a combination of physicochemical and biological processes release volatile nitrogen gases into the air and consolidate nitrogen into lagoon sediments. Further treatment is afforded by applying the lagoon slurry as a crop fertilizer to agricultural fields. Traditionally, most nitrogen losses in these treatment systems have been attributed to ammonia volatilization (Hatfield

et al., 1998; Reddi, 2005). These additional atmospheric inputs create significant regional air quality (Aneja et al., 2006) and water quality (Burkholder et al., 1992) concerns, due chiefly to odorous ammonia plumes and increased atmospheric deposition. Thus, significant efforts have focused on developing accurate ammonia emission factors (e.g., $kg\ NH_3 - N/kg\ hog/yr$) to quantify the impact of concentrated swine farming on regional air resources (Doorn et al., 2002b,a; Arogo et al., 2003).

Some investigators have debated whether ammonia volatilization represents the primary nitrogen loss from anaerobic lagoons, and therefore eschew the use of universal emission factors. The anoxic conditions observed in treatment lagoons have traditionally been thought to preclude nitrogen gas loss through classical nitrification/denitrification pathways (Reddi, 2005). Harper et al. (2000, 2004), however, reported significant dinitrogen gas fluxes (range: 11 - 86 $kg\ N_2/hectare/day$), which the authors interpreted as evidence of lagoon denitrification. Despite the anoxic conditions found in the lagoon, wind driven surficial oxygen transport at wind speeds characteristic of the farms studied by Harper et al. (2000), has been shown to provide sufficient oxygenation for nitrification (Ro et al., 2008). Significant denitrification enzymes, however, have not been observed in the water column (Hunt et al., 2010), suggesting that alternative biochemical pathways may be responsible for the observed dinitrogen fluxes. Nonetheless, lagoon dinitrogen fluxes, regardless of their source, may be a significant loss of nitrogen from CAFOs.

Despite the requirements that CAFOs within the Neuse River basin follow state-approved NMPs, evidence indicates that both air quality (Aneja et al., 2006) and water quality (Burkholder et al., 2007) remain threatened. In addition to atmospheric ammonia deposition, seepage from aging lagoon liners and overapplication of lagoon slurry continue to threaten regional waterbodies (Paerl, 1997; Ribaudo et al., 2003; Burkholder et al., 2007). I contend that calculations made in NMPs disregard both the variability in measured values and the uncertainty in published reference values,

resulting in unwarranted confidence that air and water resources are being adequately protected. Consequently, a literature synthesis quantifying the fates of the annual nitrogen input to swine CAFOs within North Carolina's Neuse River Basin will help ascertain the extent of these threats. Specifically, my goal is to create a nitrogen budget to track the fate of annual nitrogen input through feed for all swine CAFOs within the Neuse River Basin. I will use a mass-balance approach, conserving mass by accounting for nitrogen entering and leaving CAFOs, that quantifies all nitrogen sinks and their uncertainties. I will constrain fluxes which the literature has been unable to precisely quantify (e.g., lagoon denitrification) by assessing other sinks more accurately. Furthermore, I plan to use background atmospheric deposition data and nitrogen export from the Neuse River basin to establish the relative importance of these sinks to the watershed nitrogen budget.

2.2 Methods

I employed a mass-balance approach that quantified all nitrogen sinks and their uncertainties to create a nitrogen budget that tracks the fate of the annual nitrogen input for all swine CAFOs within the Neuse River Basin. The Neuse River Basin drains 14590 km^2 of land within 23 counties in eastern North Carolina. 533 hog farms requiring nutrient treatment are in operation in the basin (Figure 2.1).

Three main systems comprise most swine CAFOs: the confinement housing, the anaerobic lagoon, and the sprayfield. Nitrogen enters swine CAFOs primarily through feed intake in the confinement housing, and is exported by several sinks across the three compartments. Within the confinement housing nitrogen is lost to ammonia volatilization, hog assimilation, and export to the anaerobic lagoon. The major nitrogen sinks to the atmosphere from the lagoon are ammonia volatilization and denitrification. Other export terms include sludge accumulation and lagoon liner seepage, with the remainder applied to crops in the sprayfield compartment.

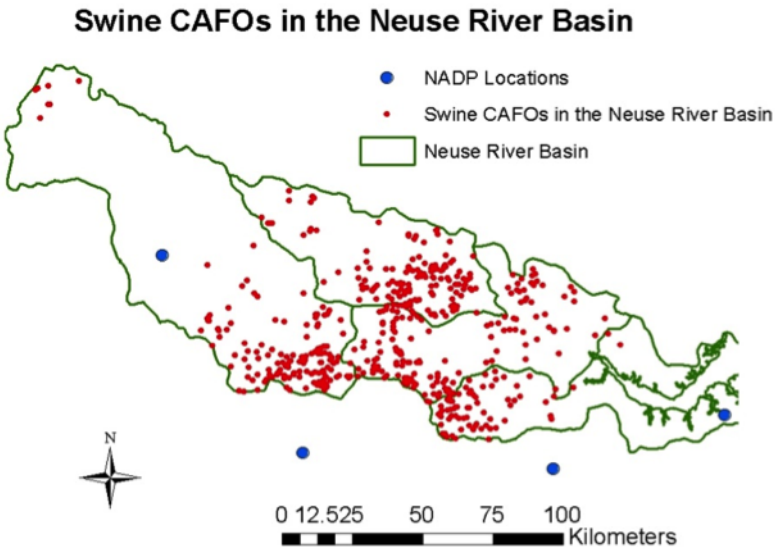


FIGURE 2.1: Swine CAFOs in the Neuse River Basin.

After nitrogen is applied in the sprayfield, it is available for plant uptake. Losses to the atmosphere include ammonia volatilization and denitrification. Other sprayfield export terms include microbial immobilization, soil nitrogen storage, and leaching.

I estimated input and export terms for nitrogen using data I gathered from a comprehensive literature review and a database of currently operating CAFOs (NC-DENR, 2011). Several sources of data were available to quantify certain sinks, and I combined all of them to estimate each sink in a way that incorporated the uncertainty inherent across studies. In most cases I estimated the uncertainty in each variable using specific distributional assumptions that matched the first two moments of the data (i.e., the mean and standard deviation) and an associated Monte Carlo simulation (n = 1000). Most of these estimates were scaled on a per unit hog mass or per unit area basis. Thus, I was able to scale up these area-based and mass-based estimates to all swine CAFOs in the Neuse River Basin by using estimates for the total hog mass, lagoon volume, and sprayfield area, which I determined from re-

gressions constructed from literature data. Specifically, I obtained a database of all current swine operations (NCDENR, 2011) and used the regressions to estimate the total hog mass, lagoon volume, and sprayfield area at each farm. The total in the basin was then calculated by summation across the estimates for each CAFO. Given the large uncertainty in the lagoon denitrification flux reported in the literature, I assessed all other sinks independently and then calculated the denitrification flux by subtraction. The associated R (R Core Team, 2014) code is in Appendix A.

2.3 Results

The entire nitrogen budget I compiled for swine CAFOs in the Neuse River Basin CAFOs is shown in Table 2.1 and Figure 2.2. The system boundaries lie at the edges of the three CAFO compartments (i.e., confinement housing, lagoon, sprayfield). Therefore, the budget I present does not directly assess transport from farms to local waterbodies because such transport relies on numerous local hydrologic factors which are difficult to synthesize (e.g., Israel et al. (2005)).

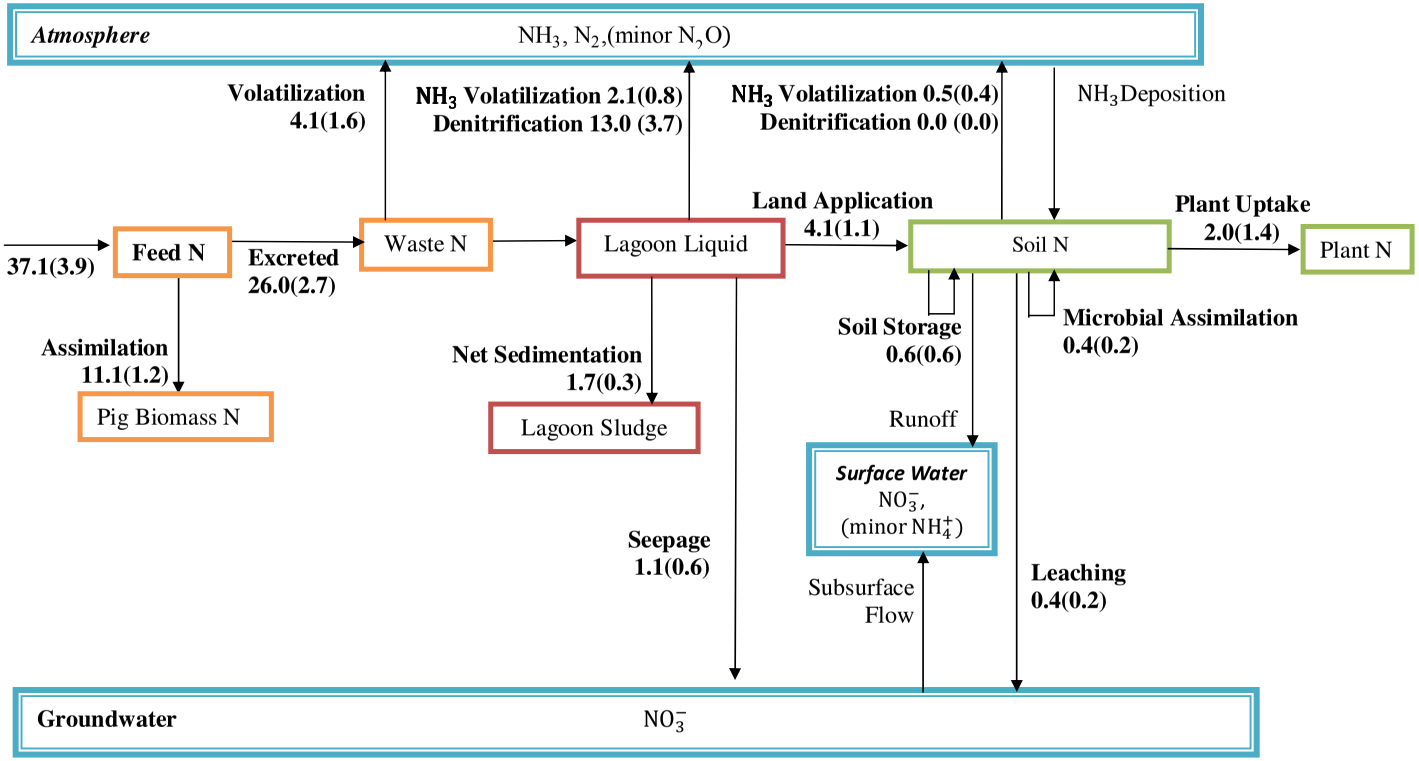


FIGURE 2.2: Nitrogen budget for swine CAFOs in the Neuse River Basin

Table 2.1: Nitrogen budget for swine CAFOs in the Neuse River Basin

	Gg N/yr		**% of Input**	
Inputs	Mean	SD	Mean	SD
Feed Input	37.1	3.9	100%	10.5%
Outputs	Mean	SD	Mean	SD
Housing				
Ammonia Volatilization	4.1	1.6	11.1%	4.3%
Hog Assimilated Nitrogen	11.1	1.2	29.9%	3.2%
Lagoon				
Sludge Sedimentation (Net)	1.7	0.3	4.5%	1.1%
Ammonia Volatilization	2.1	0.8	5.8%	2.3%
Denitrification	13.0	3.7	34.6%	7.7%
Seepage Export	1.1	0.6	2.8%	1.7%
Sprayfield				
Soil Storage	0.6	0.6	1.6%	1.7%
Microbial Immobilization	0.4	0.2	1.2%	0.6%
Plant Uptake	2.0	1.4	5.6%	3.7%
Ammonia Volatilization	0.5	0.4	1.4%	1.2%
Denitrification	0.0	0.0	0.0%	0.0%
Leaching	0.4	0.2	1.2%	0.7%

2.3.1 Inputs and Housing Sinks

533 farms possess 1.9 $\times 10^6$ hogs in North Carolina's Neuse River Basin(NCDENR, 2011). The total hog mass estimated through Monte Carlo simulation is 99.6 $\pm$ 0.5 Gg. Combining regression estimates for each farm, I calculated total food intake of 1300 $\pm$ 45 $Gg/year$. After multiplying by the percent nitrogen in feed (2.8% $\pm$ 0.3%), I estimated the annual nitrogen input to swine CAFOs in the Neuse River Basin to be 37.1 $\pm$ 3.9 $Gg\ N/yr$ (Table 2.1). In the confinement housing compartment, I estimated three major nitrogen fates: hog biomass assimilation, ammonia volatilization, and waste input to the lagoon. Several studies use a standard value of 30% of ingested nitrogen that can be assimilated into biomass (Doorn et al., 2002b,a; Aneja et al., 2008a,b). Thus, nitrogen assimilated into biomass is estimated at 11.1 $\pm$ 1.2 Gg N/year and nitrogen excreted in waste is 26.0 $\pm$ 2.7 Gg N/year (Table 2.1). Af-

ter combining estimates and uncertainties across studies, I estimated the mass-based ammonia volatilization rate to be 0.041 $\pm$ 0.017 $kg\ N/kg\ hog/year$. The annual ammonia volatilization from confinement housing is therefore estimated to be 4.1 $\pm$ 1.6 $Gg\ N/year$ (Table 2.1). The remainder (21.9 $\pm$ 3.3 $Gg\ N/year$) is exported to the lagoon compartment. In the following the detailed step by step procedure of calculating nitrogen sources/sinks within swine CAFO's housing is demonstrated.

Average Hog Mass for Each Type of Farm The detailed data on types of farms, the percentage of each hog type within the farm and the average hog mass of each hog type (Table 2.2) is available (Williams et al., 2003). The distribution of average hog mass in each operation type was calculated using Monte Carlo simulation, which enabled us to capture uncertainty.

Table 2.2: Types of CAFO operations with the Mean and Standard Deviation of Hog Mass

Type of operations	Mean	Standard Deviation
Farrow to Wean	88.9	6.9
Farrow to Feed	45.3	3.1
Farrow to Finish	54.3	3.7
Feed to Finish	61.0	6.0
Wean to Feed	13.6	1.4
Gilt	180.0	18.0
Boar	180.0	18.0

Food Intake

Food intake was calculated by developing a regression based on the data from Aneja et al. (2008a,b). The regression is as follows, and data and the developed regression is shown in Figure 2.3.

Food Intake = $13.78039 + 3.56039e-06\times$ Total Hog Mass

Percentage Nitrogen in Food Using data from Aneja et al. (2008a,b), mean and variance of percent nitrogen in food was calculated and then Monte Carlo simulation was used. To estimate a distribution for percent nitrogen in food a beta

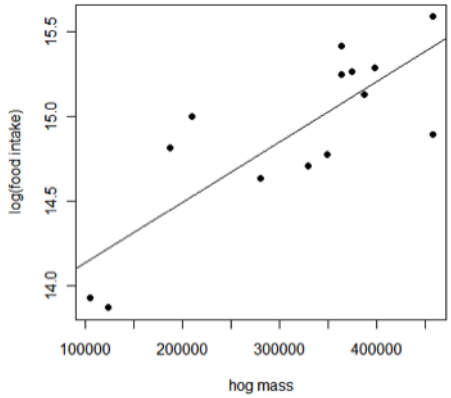


FIGURE 2.3: Relationship between Food Intake and Total Hog Mass.

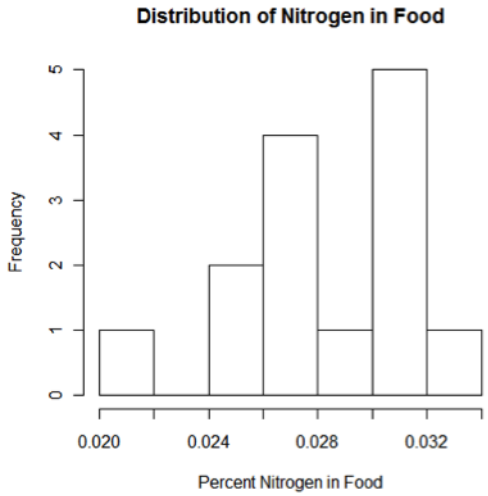


FIGURE 2.4: Distribution of Percentage Nitrogen in Food.

distribution was utilized. This enabled us to capture the uncertainty. Figure 2.4 show the result of this simulation.

Number of Hogs The data is available from NC DENR DWQ 2011:

The amount of nitrogen input to Neuse River Basin CAFOs through feed: Neuse Swine CAFO Nitrogen Feed Input = $FoodIntake\times$ % N in Food $\times$ Number of Hogs $\times$ Avg. Hog Mass Splitting up nitrogen to pig biomass and waste: From Aneja et al. (2008a,b); Doorn et al. (2002b), it is estimated that 30% of nitrogen goes to pig

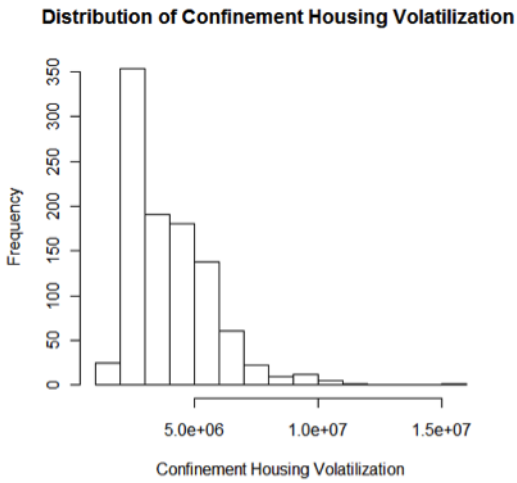


FIGURE 2.5: Distribution of Confinement Housing Ammonia Volatilization.

biomass, and the 70% remaining goes to waste.

Volatilization from Confinement Housing to Atmosphere

Using the data from (Williams et al., 2003; Aneja et al., 2008a,b; Doorn et al., 2002b) and bootstrap method, we calculated ammonia emission rates from confinement housing (Figure 2.5).

2.3.2 Lagoon Sinks

The total lagoon volume, surface area, and bottom area in the Neuse River Basin were estimated by regression to be 1.03 $\times 10^7$ m^3, 4.11 $\times 10^6$ m^2, and 4.11 $\times 10^6$ m^2 respectively. After combining estimates and uncertainties across studies, I estimate the ammonia volatilization rate to be 0.02 $\pm$ 0.008 $kg\ NH_3 - N/kg\ hog/year$. Consequently, the annual lagoon ammonia volatilization flux is 2.1 $\pm$ 0.8 $Gg\ N/year$ (Table 2.1). By combining an estimation of the mass-based sludge accumulation rate (0.0040 $\pm$ 0.0004 $m^3/kg\ hog/year$) and the total nitrogen concentration in sludge (4.3 $\pm$ 0.7 g/L), the annual nitrogen accumulated in sludge is 1.7 $\pm$ 0.3 $Gg\ N/year$ (Table1). Combining several independent sources of lagoon seepage, I estimate the total export from lagoon seepage to be 1.1 $Gg\ N/yr$. The lagoon denitrification flux,

calculated by subtracting lagoon ammonia volatilization, sludge accumulation, seepage export, and land application (See Sprayfield Section Below) from lagoon import, is 13.0 $\pm$ 3.6 $Gg\ N/year$. Multiplying the TKN (Total Kjehldahl Nitrogen) concentration in the lagoon liquid (402.10 $\pm$ 77.52 mg/L) by the total lagoon volume in the Neuse River Basin yields a lagoon nitrogen pool size of 4.1 $\pm$ 0.8 Gg N. Assuming steady state, and dividing this number by the flux through the lagoon, I estimated the mean residence time of nitrogen in the lagoon to be 58 days, a reasonable estimate. In the following the detailed step by step procedure of calculating nitrogen sources/sinks within swine CAFO's lagoon compartment is demonstrated.

Total Lagoon Volume, Surface Area, Bottom Area

Having the data for the relationship between lagoon volume and hog mass (Hunt et al., 2010; Bicudo et al., 1999), we developed a regression in order to estimate the total lagoon volume for the Neuse River basin (Figure 2.6). To calculate the surface and bottom area, we made an assumption on the lagoon geometry that all lagoons are rectangular prisms.

Lagoon Volatilization to Atmosphere

Using the data from Williams et al. (2003); Aneja et al. (2008a,b); Doorn et al. (2002b) we did a Monte Carlo simulation for different seasons, in order to capture variation of ammonia volatilization due to seasonality, and then averaged them to come up with a mean value for the lagoon volatilization to the atmosphere.

Lagoon Sludge Accumulation

In order to calculate sludge accumulation, first we estimated the sludge nitrogen concentrations, based on the data available from Bicudo et al. (1999); Williams et al. (2003) and using Monte Carlo simulation. Sludge accumulation rates were calculated from Chastain (2006). Nfluxsludge = Sludge Accumulation Rate $\times$ Average Sludge Nitrogen Concentration $\times$ Total Hog Mass

Lagoon Seepage

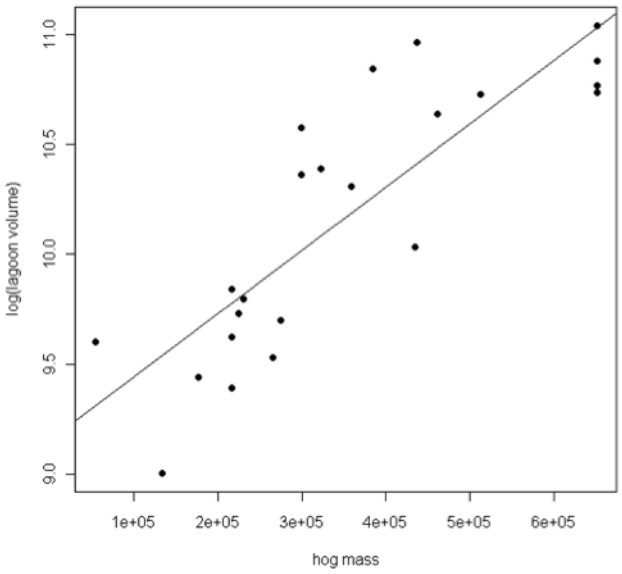


FIGURE 2.6: Relationship between log (Lagoon Volume) and Total Hog Mass

Using data from Ham (2002), we calculated lagoon seepage rate. Then we calculated average ammonia concentration in the lagoon based on the data from Bicudo et al. (1999); Hunt et al. (2010). Finally we calculated export as follows:

Lagoon Seepage Export = Average Ammonia Concentration in Lagoon × Seepage Rate

2.3.3 Sprayfield Sinks

The total sprayfield area in the Neuse River Basin was estimated to be 1.24 $\times 10^8$ m^2. The mean agronomic rate across soil types characteristic of the region weighted by my assumed crop composition (20% Corn, 40% Bermuda Grass, 40% Soybeans) is 0.016 ± 0.005 $kg\ N/m^2/application$. Assuming all farms apply biannually, the total nitrogen applied to swine CAFO sprayfields was calculated to be 4.1 ± 0.6 Gg N/year. Six major fates were assessed using rescaled percentage losses reported by Whalen and DeBerardinis (2007): plant uptake = 2 ± 1.4 $Gg\ N/year$, soil storage = 0.6 ± 0.6 $Gg\ N/year$, microbial immobilization = 0.4 ± 0.2 Gg N/year, leaching = 0.4 ±0.2

$Gg\ N/year$, ammonia volatilization = $0.5 \pm 0.4\ Gg\ N/year$, and denitrification 0.03 $\pm$ 0.02 $Gg\ N/year$ (Table 1). In the following the detailed step by step procedure of calculating nitrogen sources/sinks within swine CAFO's sprayfield compartment is demonstrated.

Sprayfield Total Area

We took two different datasets available to calculate the Sprayfield area. From 2007 Agricultural Census and NC Animal Waste Operators Certification Program, we had the total sprayfield area for different livestock, and then based on the ratios of different livestock we calculated the sprayfield area for swine CAFOs.

We also had a detailed Nutrient Management Plan for a particular farm in the North Carolina. Based on the number of hogs in the farm, type of operation and the sprayfield area, we scaled that up to all the swine CAFOs in the Neuse River Basin. Finally we averaged the two numbers together.

Sprayfield Land Application

We assumed that swine CAFOs in Neuse River basin produce corn, Bermuda grass and soybean in their sprayfields. This assumption is not far from reality since these three crop types are the most common ones in the Neuse river basin. We had the data for application rates for the three crop types (Table 2.3), and based on that we did a Monte Carlo simulation, to calculate the mean application rate, and also capture the variation in it.

Table 2.3: Distribution of Crops in the Neuse River Basin, and their Application Rates. Units for Mean of Application Rate (App. Rate) and Standard Deviation (SD) of Application Rate are both $kg/m^2/application$

	Spatial Distribution %	**App. Rate** Mean	**App. Rate** SD
Corn	40	0.013	0.003
Soybean	20	0.017	0.004
Bermuda Grass	40	0.019	0.005

Sprayfield Sinks

From Whalen and DeBerardinis (2007) we have the detailed percentage data to track the nitrogen through its sinks in the sprayfield. The major sinks in the sprayfield are plant up take, microbial activity, leaching, volatilization, and soil storage.

2.4 Discussion

According to the approach I took in this study, the most significant exports of annual nitrogen import in feed are lagoon denitrification (35%), hog assimilation (30%), total ammonia volatilization (18%), and plant uptake (5%). To estimate the impact of CAFOs on the Neuse River Basin nitrogen budget, a comparison of CAFO nitrogen exports to watershed nitrogen yield is instructive. I assume that CAFOs degrade ambient water quality through sprayfield leaching and lagoon seepage. Although the nitrogen exported by these two processes may suffer other fates before reaching surface or ground water, the amounts set an upper boundary of nitrogen impact to water bodies from swine CAFOs in the Neuse Basin. The annual nitrogen yield from the Neuse River Basin, estimated by the Spatially Referenced Regression on Watershed Attributes (SPARROW) model is 4.7 $kg\ N/ha/year$, resulting in a total basin export of 6.86 $Gg\ N/year$ (McMahon et al., 2003). Thus, my results indicate that, at most 15% $\pm$ 9% of the nitrogen export is due to swine lagoon seepage and 7% $\pm$ 4% is due to sprayfield leaching. Since these nitrogen sinks can be further transformed before reaching surface or ground water, the actual impact is likely diminished. My analysis does not assess catastrophic lagoon breaches or runoff due to land application at an inappropriate time (e.g., before a rain storm) due to absence of accurate data. Detailed information about the soil type, hydrological parameters, CAFO location in regard to waterbodies, and crops in rotation would help provide a more accurate estimate.

According to my nitrogen budget, ammonia volatilization from swine CAFOs in

the Neuse basin significantly impacts the atmosphere. Combining ammonia emissions from housing, lagoon and sprayfield, the annual flux is 6.7 $\pm$ 2.1 $Gg\ N/year$. Using data from the National Atmospheric Deposition Program (NADP 2011), the mean wet ammonia deposition in the Neuse River Basin is 4.42 kg inorganic N/ha, resulting in total wet nitrogen deposition of 6.45 $Gg\ N/year$. Thus, the amount of ammonia emission from swine CAFOs is roughly equal to the wet deposition of inorganic nitrogen. Although a portion of the emission will be deposited outside the boundary of the Neuse River Basin depending on prevailing atmospheric conditions, these numbers suggest that CAFOs are significant sources of atmospheric inorganic nitrogen deposited within the basin.

I estimate the annual lagoon denitrification flux from swine CAFOs in the Neuse Basin to be 13.0 $\pm$ 3.7 $Gg\ N_2/yr$. This sink represents the most significant export (35.0%) of the yearly nitrogen input. After converting this estimate to the same units as reported by Harper et al. (2000, 2004), the distribution (87.0 $\pm$ 24.7 $kg\ N_2/ha/day$) covers the upper end of the range of reported values for North Carolina (11-86 $kg\ N_2/ha/day$), with the mean corresponding to the high endpoint. My estimate of the ratio between lagoon dinitrogen and ammonia emissions (7.6:1 +/- 5.4:1) also covers Harper's estimate from one swine farm (5.7:1). Although I am confident that lagoon denitrification represents a significant loss of nitrogen from the system, my estimate should be received with several qualifications. First of all, my estimate was calculated by subtraction, a method which assumes I have accurately assessed all other nitrogen losses. If I have underestimated or excluded any other loss, the denitrification flux will be artificially high. For example, my estimate of nitrogen applied to land assumes that CAFO operators are applying waste at agronomic rates due to absence of accurate data on violations of agronomic application rates. If significant overapplication is occurring, denitrification will then be proportionately less. Furthermore, my assessment of total sprayfield area, crop distribution, and

agronomic rates, while based on data I had at hand, oversimplifies the heterogeneity of sprayfields in the basin. Finally, given that uncertainty exists regarding the exact biological process which generates the lagoon dinitrogen flux, this loss might better be termed dinitrogen loss.

Similarly, the assessment of nitrogen fates in the sprayfield compartment is highly uncertain. My method utilizes the proportional losses of applied nitrogen as reported by Whalen and DeBerardinis (2007). These proportional losses were reported with high uncertainty, and when combined with my crude estimate of land applied nitrogen, the proportional losses exhibit high uncertainty. Although the data reported by Whalen and DeBerardinis (2007) represent the most comprehensive assessment of nitrogen sinks after application of swine lagoon slurry, I compared the values I estimated to other independent estimates of sprayfield ammonia volatilization. My estimate ($0.007 \pm 0.002\ kg\ NH_3 - N/kg\ hog/year$) was much lower than the estimates of Murray et al. (2003)($0.046 \pm 0.024\ kg\ NH_3 - N/kg\ hog/year$), but corresponded in magnitude to that reported by Doorn et al. (2002b,a) ($0.012\ kg\ NH_3 - N/kg\ hog/year$).

To create the nitrogen budget, I made several simplifying assumptions due to a lack of detailed data on farms in the Neuse River Basin. In the confinement housing compartment I assumed that 70% of the nitrogen in food goes to waste, and 30% goes to the pig biomass (Doorn et al., 2002b,a; Aneja et al., 2008a,b). Detailed information regarding the distribution of nitrogen division between pig biomass and waste was lacking, so I was unable to assess variability in this number. I also assumed that all farms in the Neuse River Basin treat waste with anaerobic lagoons even though this is true for only a majority of farms. Due to the lack of data on crop types for individual farms, I assumed that swine sprayfields grow only three types of crops (Bermuda Grass, Corn, and Soybeans). Although several other crop types are grown on swine CAFO sprayfields and the distribution among crop types varies significantly,

this assumption while reasonable may be unduly influencing my results. Calculating the total sprayfield area within the Neuse Basin was also challenging. As noted by Murray et al. (2003), estimating sprayfield area for farms across a wide geographic region can be difficult without farm-level data. The two techniques I used generally agreed, but both estimates were significantly higher than those in Williams et al. (2003).

Further work to rigorously assess CAFO nitrogen losses should be performed. I used percentages from Whalen & DeBerardinis(2007) to ascertain the relative nitrogen losses in the sprayfield; however, soil heterogeneity and hydrology influence these percentages. Future work should also examine the extent and variability of lagoon denitrification in conjunction with studies estimating the pathway responsible for such large fluxes. Assessing the extent of nitrous oxide (N_2O) leakage which occurs during denitrification would also be worthwhile.

3

A Study of Anthropogenic and Climatic Disturbance of the New River Estuary using a Bayesian Belief Network

3.1 Introduction

Over the past two decades, more estuarine ecosystems across the globe have experienced eutrophication, defined as "an increase in the rate of supply of organic matter to an ecosystem" (Nixon, 1995; Rabalais et al., 2009). Estuaries are particularly susceptible to eutrophication due to riverine nutrient inflow, efficient nutrient trapping, long flushing times, and shallow depth. Moreover, climatic and anthropogenic perturbations have exacerbated eutrophication symptoms, through higher temperatures, extreme floods/droughts, and land use alterations, which further endangers estuarine ecological health (Neff et al., 2000; Cloern, 2001; Scavia et al., 2002; Lloret et al., 2008; Armstrong, 2009; Rabalais et al., 2009; Kaushal et al., 2010).

Our goals were to quantify the impact of eutrophication on the ecological health of estuaries, facilitate decision-making processes by managers, and develop tools for clear communication with stakeholders. To this end, we investigated potential drivers

of eutrophication in an estuary using a Bayesian belief network (BBN) approach (Ryther and Dunstan, 1971; Strobl and Robillard, 2008; Conley et al., 2009; Sheldon and Alber, 2011). The term Bayesian refers to methods related to statistical inference using Bayes' theorem.

BBNs are directed acyclic graphical models, composed of nodes and links, with embedded conditional probability tables associated with each node (Jensen and Nielsen, 2007; Heckerman, 2008). The BBNs′ visual interface makes them a valuable tool to illustrate complicated connections and communicate scientific research with a wide range of stakeholders. Additionally, the BBNs′ modularity makes them transferable to other estuaries, assuming the structure of the model is generalizable (Nixon, 1995; Smith, 2003) and accounting for ecosystem specific variability (Koller and Pfeffer, 1997; Jensen and Nielsen, 2007; Johnson et al., 2010). The developed model for the former estuary would act as prior information for the later estuary. The model can then be updated using data from the later estuary due to the Bayesian nature of the BBN (posterior $\propto$ likelihood of observed data $\times$ prior). Finally, BBNs accommodate our goal of scenario investigation (Uusitalo, 2007).

In this chapter, we describe the study area, the dataset, and the BBN model construction and evaluation. Using the model, we explore potential impacts of climatic variability and management scenarios on the NRE′s water quality.

3.2 Materials and Methods

3.2.1 Study Area

Our study area (see Figure 3.1), the New River Estuary (NRE, $\sim$ 1435 km^2), located in Onslow County, North Carolina, USA, was a highly eutrophic estuary (1995-2002) with elevated levels of chlorophyll *a* (>60 $\mu g/l$), nitrogen (total dissolved nitrogen >1 mg/l), phosphorus (total dissolved phosphorus >0.1 mg/l), turbidity (secchi disk depth <1 m), occasional bottom water hypoxia (dissolved oxygen <2 mg/l), and

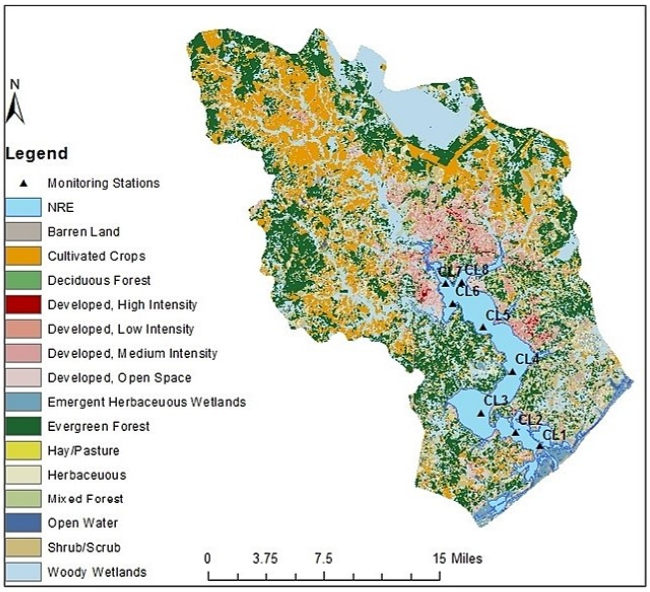


FIGURE 3.1: New River Estuary, land use and cover, and monitoring stations. For the BBN, data from CL 6, CL7, and CL8 were used.

nuisance/harmful algal blooms (Mallin et al., 2005). Even after the upgrade of the sewage treatment plant of the City of Jacksonville and the U.S. Marine Corps Base at Camp Lejeune in 1998, moderate to severe eutrophication symptoms are still observed (NOAA, 1996; Mallin et al., 2005). Poor water quality has negatively impacted the regional commercial fisheries for blue crab and shrimp (NCDMF, 1993; Tomas et al., 2007). Although high nutrient concentrations and elevated chlorophyll a production extends along the 25 Km of the NRE, eutrophication symptoms are most severe in Morgan Bay, our study area, near the head of the estuary (NOAA, 1996; Mallin et al., 2005). The NRE Bayesian belief network was developed with water quality monitoring data from the Defense Coastal/Estuarine Research Program (DCERP), Aquatic Estuarine Monitoring component (RTI, 2013), unless otherwise stated. Eight stations along the length of the NRE were sampled on a monthly basis, starting in October 2007, for a range of physical, chemical, and biological variables. In this study, we used data from October 2007 to October 2012 for the three stations

in the upper estuary (stations CL 6, CL7, and CL8 in the Morgan Bay) where eutrophication symptoms are most severe (Figure 3.1) (Hall et al., 2012). Furthermore, the NRE is functionally divided into two shallow estuaries with different ecological properties, with the upper section showing stronger relations between chlorophyll *a* and nutrient concentrations (RTI, 2013). We used the data from October 2007 to September 2011 for model development and the remainder (i.e., October 2011 to October 2012) for model validation purposes.

3.2.2 Model Construction

We followed the guidelines on developing Bayesian belief network models suggested in the literature by undergoing several cycles of model development and revision (Marcot et al., 2006; Chen and Pollino, 2012). We implemented the BBN in the Hugin Educational 7.1 software package (Madsen et al., 2003).

The first step in developing the BBN was to determine the most important factors that were believed to have an impact on the eutrophication in the NRE. The number of variables and nodes of the BBN model depend on the purpose and scope of the study. Here our objective was to quantify the impacts of anthropogenic and climatic factors on water quality indicators (i.e., chlorophyll *a* concentrations, bottom water dissolved oxygen and presence/absence of harmful algal bloom species). To this end, we developed an ecological network with surface chlorophyll *a* concentrations, bottom water dissolved oxygen and presence/absence of harmful algae (toxic algae and/or hypoxia generating and/or food web disrupting) as key water quality indicators suggested by US EPA National Coastal Condition Report′s suggestions (EPA, 2001a,b; Sheldon and Alber, 2011) (Figure 3.2). The variables within the BBN were compartmentalized into five functional components: "Physical Environment", "Chemical Environment", "Biological Environment", "Harmful Algae", and "Hypoxia/Anoxia" (Figure 3.2), to accommodate our goal of investigating the im-

pacts of anthropogenic nutrient pollution and climatic variability. One advantage of this compartmentalization is that the current components can be further expanded in the future or additional components such as land-use can later be developed and added as a sub-model to the current BBN.

We investigated a large selection of variables in the dataset (see supplementary material Table B.2) using exploratory data analysis, personal communications with local experts and scientific literature during the variable selection procedure. We included the following variables with a monthly time scale in the model to predict surface chlorophyll *a* concentrations, bottom water dissolved oxygen and presence/absence of harmful algae (toxic algae and/or hypoxia generating and/or food web disrupting): wind speed (data from State Climate Office of North Carolina, 2007, Station Name: New River MCAS, mph), photosynthetically active radiation (PAR) (data from State Climate Office of North Carolina (CRONOS, 2007), Station Name: New River MCAS, $\mu moles/sec/m^2$), temperature ($^\circ C$, representing seasonal variation), precipitation (data from State Climate Office of North Carolina 2007, Station Name: New River MCAS, cm/mo), freshwater discharge (average monthly data from USGS Station Number 02093000, latitude: $34^\circ 50' 57''$, longitude: $77^\circ 31' 10''$, m^3/s), stratification (density gradient, bottom water density minus surface water density, g/cm^3)(salinity is used in calculating stratification), light attenuation coefficient (K_d, $1/m$), dissolved inorganic nitrogen (surface DIN, $\mu g/l$), orthophosphate (surface PO_4, $\mu g/l$), chlorophyll *a* (surface algal biomass, $\mu g/l$), primary production (PPR, $mg\ of\ C/m^3/h$), growth rate (GR, $1/h$), and bottom water dissolved oxygen (O_2, mg/l).

Harmful algal genera described in the NRE include *Karlodinium veneficum* (dinoflagellate), *Chattonella*, *Fibrocapsa*, and *Heterosigma* (raphidophytes). Species composition was determined by high performance liquid chromatography (HPLC) measurements. Unique class levels pigments of 19′-hexanoyloxyfucoxanthin and vi-

olaxanthin were used to determine presence/absence of *Karlodinium veneficum* and marine raphidophytes, respectively. The presence of violaxanthin was translated into presence of raphidophytes. Chlorophytes also contain violaxanthin but they are a minor component of the biomass in the NRE (Hall et al., 2012). Other pigments such as zeaxanthin (cyanobacteria) that might be important in mid- to late summer have not been included in this study since they have not been described in the NRE. We also investigated occurrence of dinoflagellates as a group containing several harmful species by presence/absence of peridinin. In the BBN the presence of diagnostic pigments does not translate into harmful algal bloom (HAB) conditions (i.e., blooming and/or producing toxins); however, the report of absence of a pigment ascertains lack of HAB conditions.

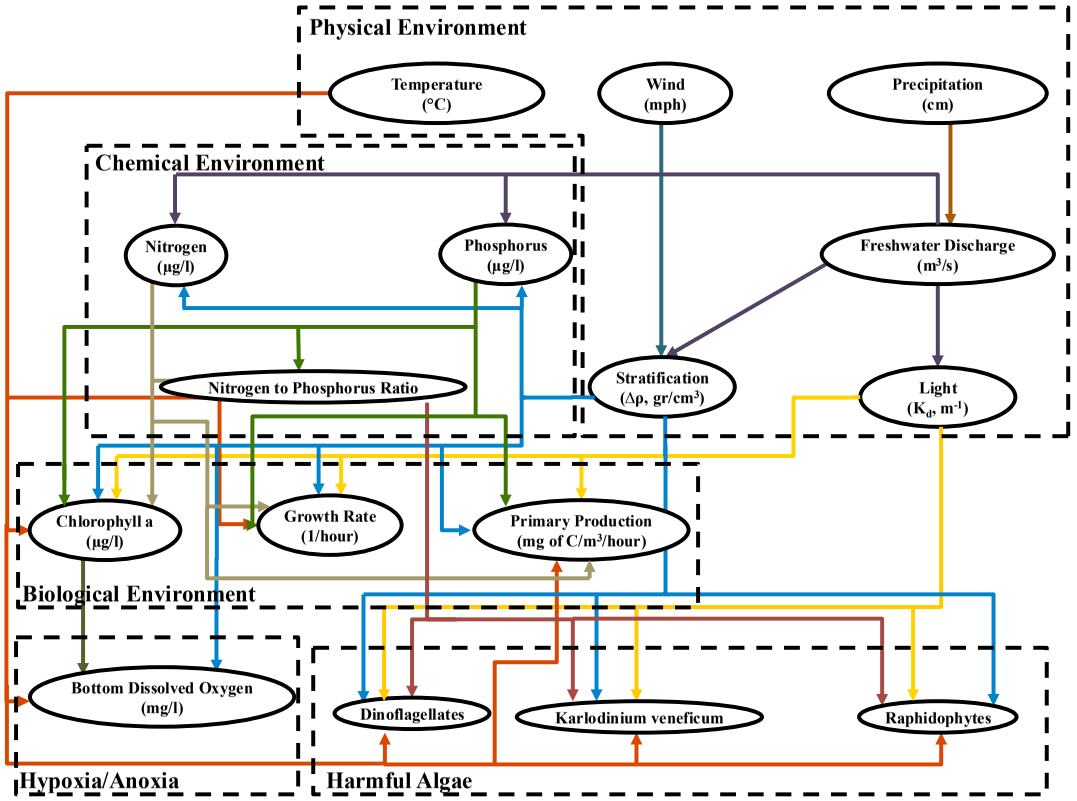


FIGURE 3.2: The Directed Acyclic Graphical (DAG) representation of the New River Estuary model and its components: "Physical Environment", "Chemical Environment", "Biological Environment", "Harmful Algae", and "Hypoxia/Anoxia". The five functional components were defined to accommodate our goal of investigating the impacts of anthropogenic nutrient pollution and climatic variability.

The main driving factors of chlorophyll *a* concentration are light, physical forces (mixing), nutrients, temperature and zooplankton grazing. All aforementioned variables have been measured for the NRE except grazing. In order to examine grazing, the growth rate was calculated based on an empirical relationship by Cloern et al. (1995), using concentration of the most limiting nutrient (DIN), half saturation constant (K_N, the value was prescribed equal to 1 μM based on Cloern et al. (1995)), spectrally averaged light attenuation coefficient (K_d), depth of mixing layer (H, calculated based on salinity profile), daily irradiance (I_Φ, calculated from PAR data) and temperature (T):

$$\begin{aligned} \text{Chlorophyll a : C} &= 0.003 + \frac{0.0154N}{K_N+N} e^{0.05T-0.059\frac{I_\phi\left(1-e^{-kH}\right)}{kH}} \\ GR &= \frac{PPR}{\text{Chlorophyll a} \times \text{Chlorophyll a:C}} \end{aligned}$$

Significant uncertainty is associated with all phytoplankton growth models; hence, the growth rate term was interpreted as lower versus higher than 0.01 to investigate top-down versus bottom-up control. The 0.01 value is an average value for growth rate calculated from studies on the Neuse River Estuary, a similar neighboring estuary (Hall et al., 2008). The growth rate node is terminal (i.e., no child nodes) and as such does not influence the model′s behavior.

The second step in developing the BBN was to determine the structure of the model. The primary structure of the BBN was developed based on a priori expected causal dependencies established on the existing literature and exploratory data analysis (see supplementary material Figure B.1 for the base model that only included chlorophyll *a* and its predecessor nodes) (RTI, 2011). The developed model was then reviewed and revised by experts through series of in person/online meetings and email communications to evolve to its final version shown in Figure 3.2. The expert comments were specifically valuable for variable choice in the harmful algae

component since they are estuarine specific. The structure was then fed into the Hugin software.

The third step of the BBN model development process is to specify the conditional probability tables (CPTs). The CPT contains the probability of a node′s values conditioned on every possible combination of its predecessor nodes′ values (for examples of CPTs see supplementary material Tables 3 to 11). The arrows in a BBN represent conditional probabilities associated with states of a node as a function of the states of its direct predecessors (parent nodes). There are several ways to specify the conditional probabilities in the BBN depending on the sources and amount of accessible data (Marcot et al., 2006; Chen and Pollino, 2012). In this study, the CPTs were estimated using the Expected-Maximization algorithms provided by the Hugin software (Dempster et al., 1977; Madsen et al., 2003) for data from field monitoring sites (see Subsection 3.2.1).

The present BBN was developed with discrete probability distributions for all variables in the model (Jensen and Nielsen, 2007; Alameddine et al., 2011), due to restrictions of continuous variables in BBNs and their specialized software. The discretization procedure remains one of the challenges in building the BBNs. Points to consider when discretizing are: the size of the available dataset, the interpretation goal of the node, the placement of the node with in the BBN (does it have any predecessor nodes?), the shape of the underlying distribution, the number of outliers, and the number of repetitive values for data points.

The two established discretization techniques for empirical datasets are equal-interval and equal-frequency (Chen and Pollino, 2012). The equal-interval method is unsuitable when the dataset is unevenly distributed or contains outliers, since it would result in sparsely populated bins. The equal-frequency method has shortcomings when dataset has repetitive values. Further, neither of these techniques preserve the original distribution of the data; hence, we discretized the BBN nodes

by exploring a new approach called moment matching method, which focuses on matching lower statistical moments of the initial distribution (i.e., mean, variance, skewness, kurtosis, etc.). Unlike the equal-interval or equal-frequency discretization methods, the moment matching method leads to a better representation of the underlying continuous distribution by matching its moments with an appropriate discrete distribution (Smith, 1993).

Another important point to address while discretizing continuous variables is the number of intervals. Large number of intervals would improve representation of the underlying distribution but increase the size of the conditional probability tables due to increase in states of predecessor nodes; hence, an optimal number of intervals for each variable should be determined. The number of intervals for each node was determined to accommodate our analysis and its application to the various scenarios that were investigated (Alameddine et al., 2011).

In this study environmental factors were discretized into four bins to accommodate detailed scenario investigations. For chlorophyll *a* and bottom water dissolved oxygen nodes, we fixed one of the intervals on roughly (the approximation is due to the limitation in discretization of BBNs, an interval endpoint should be an observed data point) 40 $\mu g/l$ and 4 mg/l respectively, to examine scenarios resulting in violation of water quality standards in North Carolina (NCDENR-DWQ, 2007). The other endpoints for chlorophyll *a* and bottom water dissolved oxygen nodes were selected to match the moments of the underlying distributions. Our purpose here was to identify conditions that are suitable for presence/absence of harmful algae; hence, for the harmful algae component, we defined two intervals, representing presence and absence of indicator pigments.

3.2.3 Model Diagnostics

The developed BBN′s performance accuracy was evaluated both qualitatively and quantitatively. As a qualitative assessment, we examined the various scenarios of chlorophyll *a* concentration, bottom water dissolved oxygen, and nutrients to assess whether the observed increased/decreased responses were consistent with the directions of such changes in the literature (see table 3.1). Quantitative evaluations were done using the Area under the Receiving Operating Characteristic Curve (AUC) to validate the BBN (Fawcett, 2006; Marcot et al., 2006; Chen and Pollino, 2012). The analysis wizard embedded in Hugin 7.1 was utilized to evaluate the BBN. The dataset from October 2011 to 2012 was used to calculate the evaluation criteria, i.e., AUC and 90 % Highest Density Interval (HDI) (Kruschke, 2010). The AUC varies between 0 and 1 and provides a diagnostic measure of models prediction accuracy, which represents the probability of a true positive outcome (the proportion of actual observations which are correctly classified) versus a false positive outcome (accuracy of data classification). A model with perfect predictions would have an AUC equal to 1. The 90% HDI contains credible values, that have higher credibility than values outside the interval, which spans 90% of the distribution. The 90% HDI is a more intuitive and meaningful summary of the posterior distribution, and hence a better evaluation criteria. Furthermore, the HDI intervals work better for skewed distributions than equal-tailed intervals, which exclude points in the compact tail with higher credibility and include points near the skewed tail.

3.3 Results: Current Conditions

The BBN for the sampling period is depicted in Figure 3.3. Following, we summarize the results under three categories of biological environment, harmful algae, and hypoxia/anoxia, the water quality indicators of interest.

Table 3.1: List of investigated scenarios. The first column represents the scenario investigated, the variable name and the range it was set for the investigated scenario. The second column lists the variables that showed a significant change under the investigated scenario. Only variables with $\Pr(\mid \mu_1 - \mu_2 \mid > 0) \geqslant 0.9$ are presented, where μ_1 and μ_2 are the means during the sampling period and under investigated scenarios, respectively. The third column represents the difference of the means during the sampling period and under investigated scenarios for the variables in the second column.

Scenario		Variables	$\mu_1 - \mu_2$
Low Precipitation	[1.37, 3.80]	Freshwater Discharge	-13.20
		Stratification	-0.23
		Chlorophyll a	-2.74
High Precipitation	[16.69, 40.60]	Freshwater Discharge	107.00
		Stratification	0.62
		Light	0.55
		Phosphorus	15.70
		Nitrogen to Phosphorus Ratio	3.17
		Chlorophyll a	5.18
Stratified	[3.61, 15.11]	Light	0.82
		Nitrogen	134.00
		Phosphorus	15.70
		Nitrogen to Phosphorus Ratio	2.01
		Primary Productivity	17.20
		Chlorophyll a	4.48
		Bottom Dissolved Oxygen	-1.26
		Dinoflagellates	-0.07
		Raphidophytes	-0.13
Mixed	[-0.04, 1.04]	Light	0.23
		Nitrogen	18.80
		Phosphorus	-2.23
		Chlorophyll a	-4.01
		Bottom Dissolved Oxygen	-0.46
Low Temperature	[5.42, 12.25]	Bottom Dissolved Oxygen	1.56
		Karlodinium veneficum	0.14
High Temperature	[25.80, 32.63]	Chlorophyll a	2.28
		Karlodinium veneficum	0.22
		Dinoflagellates	0.17
Low Nitrogen	[5.57, 56.30]	Chlorophyll a	-7.80
		Raphidophytes	0.04

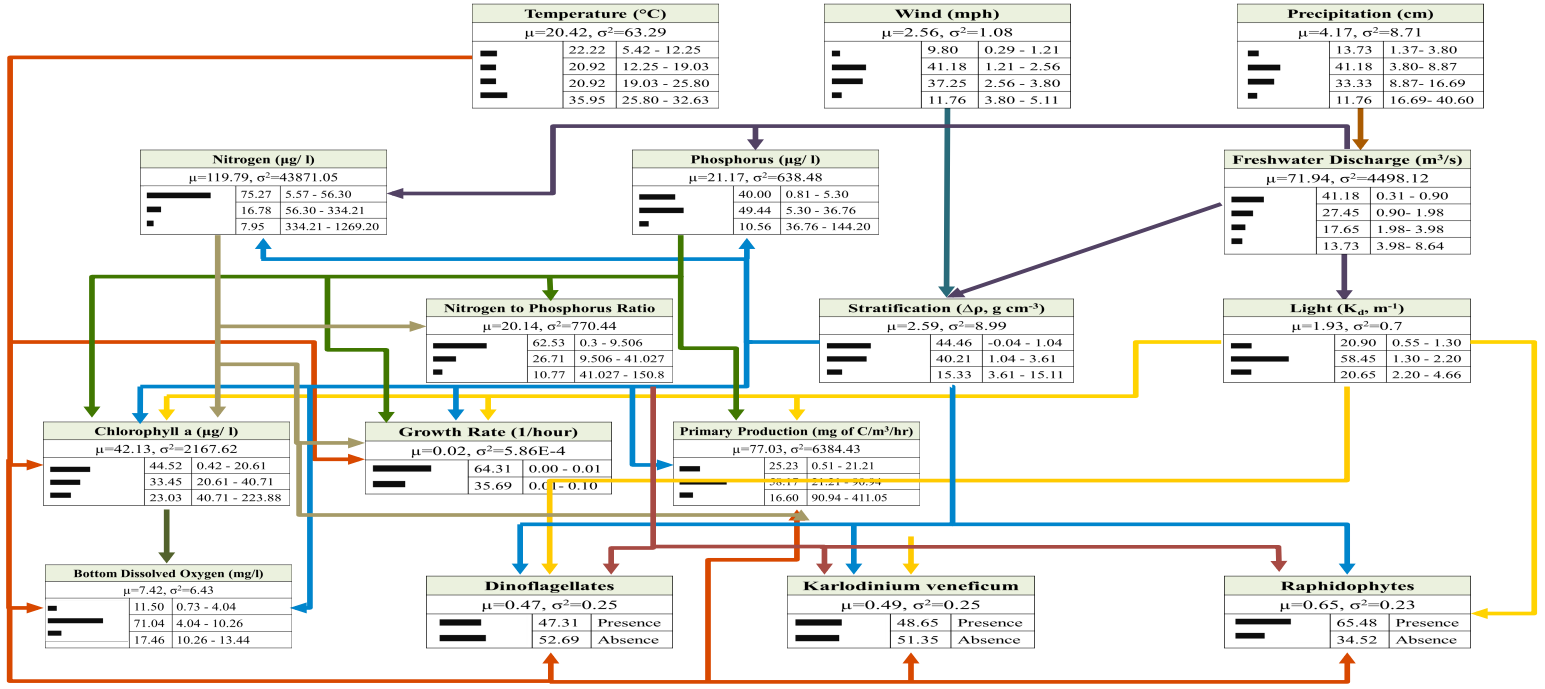


FIGURE 3.3: The New River Estuary BBN during sampling period: the nodes represent variables of interest, and a link between two nodes represent conditional dependency. The numbers next to the black bars are probabilities of the node being in a specific state and the intervals defined for the status of the nodes,respectively. For example chlorophyll a is between 40.71 $\mu g/l$ and 223.88 $\mu g/l$ 23.03% of the time; hence, the North Carolina's water quality criteria of 40 $\mu g/l$ was violated approximately 1/5 of the time during this time-series. The main output variables are water quality indicators chlorophyll a, bottom dissolved oxygen, and the three pigment presence/absence.

3.3.1 Nutrient and Biological Conditions

Nutrients are primarily delivered by freshwater discharge or recycled through sediment re-suspension; therefore, the intensity of water column stratification and freshwater discharge control nutrient delivery. Nutrient concentration is a poor predictor of nutrient availability for limiting nutrients, because those available are removed by phytoplankton rapidly and not captured by monthly sampling (Cassar et al., 2011). Although nutrient loading would be the best indicator of nutrient availability, in the absence of acceptable nutrient loading estimates, nutrient concentration was used as an indicator of nutrient availability.

The results show that chlorophyll *a* concentrations violate North Carolina′s "acceptable" water quality standard of 40 $\mu g/l$ (NCDENR-DWQ, 2007) 23.03% of the time during the sampling period (monthly frequency)(see Figure 3.3). In conjunction with chlorophyll a concentrations, we also examined primary production (PPR) and growth rate (GR) to assess whether the system is top-down, controlled by predation, (high GR, low PPR) or bottom-up, controlled by nutrients or light, (low GR, high PPR) controlled. The estimated GR is compared to the respective PPR to judge top-down versus bottom-up control, due to lack of data on grazing in the NRE. The BBN analyses indicate that during the sampling period, GR is low (<0.01 with 64% frequency) and PPR is medium to high (>21.21 with 75% frequency) (Figure 3.3), which is typical of a bottom-up eutrophic ecosystem (Laws, 2000). Growth rate and primary production distribution would change under different seasons, which are specified in the BBN by freshwater discharge, temperature, and light. Growth rate and primary production distributions rule out grazing limitation of primary production; hence, the system is either light or nutrient limited. Our analysis does not take into consideration phytoplankton vertical migration, which is known to occur in the NRE (Hall et al., 2012). This does not have an impact on our analysis, since the

NRE does not appear to be light limited.

Light limitation is common in highly turbid estuarine ecosystems, such as NRE (Cloern, 1987). In our study K_d is lower than 2.2 $1/m$ 78% of the time. Considering that the NRE has an average depth of 1.6 m, the water column is in the euphotic zone more than 78% of the time; hence, light limitation is unlikely during the sampling period. The nutrient concentrations reveal that dissolved inorganic nitrogen and orthophosphate are in the low bracket range (<56.30 and <5.30 $\mu g/l$, respectively) 80% and 40% of the time, respectively. The N to P ratio is lower than 9.506, 62.5% of the time; hence, NRE is N limited, which is consistent with previous studies showing nitrogen limitation in the NRE (Altman and Paerl, 2012). The nutrient limitation in the NRE highlights the importance of management decisions on nutrient regulation, since the system would respond promptly to nutrient enrichment.

3.3.2 Harmful Algae

Based on the model results over the dataset timeline (4 years), *Karlodinium veneficum*, other harmful dinoflagellates, and harmful raphidophytes (*Chattonella*, *Fibrocapsa*, and *Heterosigma*) were present in the NRE 48.7%, 47.3%, and 65.5% of the time during the sampling period, respectively (Figure 3.3). It should be noted that presence does not necessarily imply bloom or toxic conditions. The harmful bloom conditions can be determined by investigating chlorophyll *a* concentrations of higher than 40 $\mu g/l$ in conjunction with community composition, which is outside the scope of this study.

3.3.3 Hypoxia/Anoxia

In the BBN structure bottom water dissolved oxygen is connected to chlorophyll *a* concentrations, temperature, and stratification intensity. The relationship of chlorophyll *a* and bottom water dissolved oxygen is strongest during summer-fall (see

supplementary material, Figure B.2). Lower temperatures result in higher oxygen solubility in water; therefore, during winter/early spring there is not a strong relationship between oxygen and chlorophyll *a*. Bottom water dissolved oxygen concentrations in the NRE violated the North Carolina's water quality standard of 4 mg/l, 11.5% of the time during the sampling period(NCDENR-DWQ, 2007).

3.4 Discussion

The BBN was used to quantitatively assess the response of water quality indicators to climatic variability and nutrient management scenarios. The evidence for low and high scenarios of precipitation, stratification, temperature, and nitrogen was propagated through the model and marginal probabilities were re-calculated (Table 1). The scenarios in this chapter were set within the system′s observed variability of the physical and chemical environment. During the study period the ecosystem experienced a wide range of values for climatic and nutrient variables (RTI, 2013); hence, investigation of future climatic and nutrient management scenarios was possible. Here we only present the results with $\Pr(\mid \mu_1 - \mu_2 \mid > 0) \geqslant 0.9$, where μ_1 and μ_2 are the means - the weighted-average of the midpoints of each category, weighted by the probability of that category- of any of the variables for the sampling period and under the investigated scenarios, respectively (Kruschke, 2010). Therefore, we are comparing the water quality under an investigated scenario versus current conditions. The AUC, a measure of the model′s prediction accuracy, was 0.75 and 0.95 for chlorophyll *a* and bottom water dissolved oxygen, respectively, suggesting good performance in predicting the variables of interest. We concluded that the BBN is capable of distinguishing between different values measured for all the water quality indicators.

3.4.1 Climatic Variability

Our BBN quantifies the impact of climatic variability on water quality, which was described only qualitatively in the literature (Najjar et al., 2000; Rogers and McCarty, 2000; Rabalais et al., 2009; Kaushal et al., 2010). Following, we will present the results of low/high precipitation, stratification, and temperature scenarios (one variable manipulation per scenario). These scenarios were formed based on Najjar et al. (2000) assessment of the potential impacts of climate change on the mid-Atlantic coastal region of the United States.

Precipitation

Precipitation controls water quality indicators in our BBN through several pathways, i.e., nutrient delivery, turbidity, stratification intensity and freshwater discharge. Although time-lagged freshwater discharge and water residence time impact the phytoplankton biomass of the NRE (Peierls et al., 2012), we did not consider lagging because the freshwater discharge gauging station is very close to our monitoring sites. Furthermore, land-use/land-cover affects freshwater discharge, but we cannot directly measure the effect due to the short duration of the study. As a result freshwater discharge is only connected to precipitation. There is lagging between precipitation and freshwater discharge; however, this is negligible for the upper estuary, our study area.

Precipitation and thus freshwater discharge in our dataset includes a range of drought to flood conditions (Peierls et al., 2012). We ran the BBN under low and high precipitation scenarios to examine how the NRE may respond to extreme events such as storms and droughts. The BBN predicts that under a low precipitation scenario ($<3.78\ cm/month$), freshwater discharge decreases significantly, with flows equal or less than $0.58\ m^3/s$ occurring 71.4% of the time (see supplementary online material, Figure 1). Under low precipitation the water column is completely mixed (stratifica-

tion <1.04 g/cm^3) 50.07% of the time, which might explain the lower chlorophyll *a* concentrations. The increase in the probability of chlorophyll *a* concentration being lower than 40 $\mu g/l$ is 81.3% vs. 79.9% during the whole sampling period. The change, while subtle, is significant based on the statistical criterion ($\Pr(\mid \mu_1 - \mu_2 \mid > 0) \geqslant 0.9$). This is interpreted as lower probability of violating North Carolina′s water quality standards (Chlorophyll *a* <40 $\mu g/l$ and bottom water dissolved oxygen <4 mg/l). In our BBN model, the reduction of chlorophyll *a* under low precipitation is due to the impact of river flow on both stratification and nutrient load (Vargo, 2009)(see supplementary online material, Figure B.3).

Under the high precipitation scenario, defined as precipitation greater than 16.7 $cm/month$, the model predicts statistically significant changes in freshwater discharge, stratification, chlorophyll *a* concentration, light, phosphorus, and molar N:P ratio, under non-saturation conditions (Table 1). As we have discussed before the NRE is a nutrient limited estuary and thus not saturated. The probability of freshwater discharge being higher than 1.98 m^3/s is 88.3% - a 57.0% increase from the sampling period average of 31.3%. The probability of the NRE being stratified is 23.8%, more likely due to the density gradient induced by high freshwater discharge. The probability of chlorophyll *a* concentration exceeding North Carolina's water quality standard (22.4%) is significantly greater than the sampling period average of 20.4%. This is likely due to high nutrient concentrations as a result of high nutrient delivery associated with enhanced freshwater discharge. The change in chlorophyll *a* concentration is subtle since higher precipitation and hence very high freshwater discharge (>3.98 m^3/s) results in lower light availability due to high concentrations of colored dissolved organic matter, mixed water column, and transfer of phytoplankton to the lower estuary. The likelihood of high N:P ratio (>9.506) increases by 11.4% (from 37.5% to 48.9%) consistent with higher availability of N relative to P under high precipitation (Green and Wang, 2008). Mobilization of nitrogen and

phosphorus are further dependent on the degree of stratification and internal nutrient regeneration. The model fails to detect a significant increase in nitrogen concentration, which may be attributed to the discrete observations that could have missed the event (e.g., first flush) and fast nutrient drawdown by phytoplankton. The 15.3% increase (from 20.7% to 36.0%) in K_d of higher than 2.20 $1/m$ is associated with high levels of riverine colored dissolved organic matter inflow, which is enhanced by high freshwater discharge (see supplementary online material, Figure B.4) (Branco and Kremer, 2005). Our study quantitatively confirms previous qualitative speculations on the effects of precipitation on coastal ecosystems (Najjar et al., 2000; Scavia et al., 2002; Wetz and Paerl, 2008; Doney, 2010).

Water Column Stratification

The intensity of water column stratification in estuarine ecosystems is regulated by freshwater discharge, wind, and tidal forcings (Cloern, 2001). Tidal forcing is minimal in the upper NRE due to the microtidal regime of the NC coast and the attenuation of tidal flow by the narrow and shallow inlet and straits within the lower NRE (RTI, 2013). As a result in the NRE freshwater discharge and wind are the only determinants of stratification. We investigated the response of the system to stratification since it is one of the most important influencers of water quality in estuaries (Paerl, 1988; Diaz and Rosenberg, 2008). Stratification and freshwater discharge have a parabolic relationship; the system is mixed under low freshwater discharge (<0.90 m^3/s) but also stratification breaks down again under very high freshwater discharge (>3.98 m^3/s). In the BBN the stratification is also impacted by wind intensity; hence, it is not expected to observe the marginal impact of freshwater discharge in the BBN.

A stratified water column (stratification>3.61 g/cm^3) is associated with significant changes in primary productivity, nitrogen (N), phosphorus (P), N:P ratio, light

availability, chlorophyll *a* concentrations, bottom water dissolved oxygen and the probability of presence of dinoflagellates and harmful motile raphidophytes (*Chattonella*, *Fibrocapsa*, and *Heterosigma*) (Table 1). The important role of stratification has been noted in previous studies (Diaz and Rosenberg, 2008; Murphy et al., 2011). Stratification is in part induced by freshwater discharge, which also brings high concentrations of colored dissolved organic matter (with K_d >2.2 $1/m$, 55.9% of the time) and results in a decrease in transparency (Keller, 1989; Domingues et al., 2011; Gameiro et al., 2011). Freshwater discharge delivers high nutrient loads which stimulate phytoplankton growth and result in elevated chlorophyll *a* concentration. The likelihood of bottom water dissolved oxygen violating NC water quality standard (4 mg/l) increases by 21.75% (from 11.50% to 33.25%). The duration of stratification is an important factor in this speculation; however, it is not captured by our monthly sampling scheme. Data from two vertical profilers along the NRE with 30 minute sampling frequency could be used to further assess the duration of stratification (RTI, 2013). The likelihood of the presence of harmful raphidophytes (*Chattonella*, *Fibrocapsa*, and *Heterosigma*) and dinoflagellates decrease by 9.3% (from 65.5% to 56.2%) and 2.8% (from 47.3% to 44.5%), respectively. This is most likely due to the loss of their competitive advantage over other phytoplankton under high N:P conditions. Paerl and others have however shown that N:P ratio is not a strong forcing on harmful algae, especially under conditions where nitrogen or phosphorus are saturating (Paerl, 2009; Paerl and Scott, 2010; Lewis Jr et al., 2011). In a partially mixed water column (1.04<stratification<3.61 g/cm^3), motile harmful species often seem to thrive. The likelihood of presence of *Karlodinium veneficum*, raphidophytes and dinoflagellates increase by 3.1% (from 48.6% to 51.7%), 7.1% (from 65.5% to 72.6%), and 9.2% (from 47.3% to 56.5%), respectively.

A mixed water column (stratification<1.04 g/cm^3) results in a significant change in light availability, nitrogen, phosphorus, chlorophyll *a*, and bottom water dissolved

oxygen compared to the sampling period (Table 1). A 4.0% decrease (from 20.65% to 16.65%) in K_d less than 2.2 $1/m$ is associated with lower chlorophyll *a*. In a mixed water column the likelihood of the bottom water dissolved oxygen violating state criteria decreases by 10.9% (from 11.50% to 0.6%) (see Figure 3.3 and supplementary online material, Figure B.7). An 8.9% increase (from 78.0% to 86.9%) in frequency of chlorophyll *a* concentrations lower than 40.71 $\mu g/l$ are associated with a mixed water column. A 10% increase (from 75.3% to 85.3%) in DIN<56.30 $\mu g/l$, and 15.1% increase (from 40.0% to 55.1%) in PO_4<5.30 $\mu g/l$ are associated with higher freshwater discharge, water column mixing, and sediment resuspension.

The likelihood of presence of potentially harmful algal species under mixed water column conditions does not show a significant change; in a partially mixed water column, all harmful algal species increase significantly; whereas in a stratified water column mixed responses are observed. One potential explanation is that the highest stratification intensities are likely related to high flow events, which may reduce residence time and limit harmful algae development. Well-mixed conditions are more likely to occur when freshwater inputs are low which also corresponds to low input of riverine nutrients. Additionally, if there is any stratification, it is likely that the flagellates are capable of vertically migrating and this may give them an advantage over other groups (Hall and Paerl, 2011) (see supplementary online material, Figure B.5- B.7).

Temperature Effects

Low temperatures (<12.25 $^{\circ}C$) result in a 9.2% increase (from 88.5% to 97.7%) in bottom water dissolved oxygen greater than 4 mg/l due to increased solubility of oxygen, lower productivity during cooler months, and lower respiration rates at lower temperatures. An increase in the presence of *Karlodinium veneficum* (8.7% (from 48.7% to 57.4%)) is observed under low temperatures (<12.25 $^{\circ}C$). Chloro-

phyll *a* does not show a significant change under a low temperature scenario (<12.25 $^{\circ}C$) compared to the whole sampling period. This might be due to shifts in phytoplankton community composition during lower temperature periods or an increase in chlorophyll *a* to phytoplankton carbon in cell ratio. Furthermore, grazing activity is also impacted by temperature. A reduction in grazing pressure at low temperature could result in weak sensitivity of chlorophyll *a* to temperature (see supplementary online material, Figure B.8).

High temperatures (>25.8 $^{\circ}C$) are associated with an increase in the presence of *Karlodinium veneficum*, dinoflagellates, and harmful raphidophytes by 16.4% (from 48.7% to 65.1%), 4.3% (from 47.3% to 51.6%), 16.6% (from 65.5% to 82.1%), respectively. Previous studies on *Karlodinium veneficum* high abundance at temperatures of 5-15 $^{\circ}C$ and 25-30 $^{\circ}$C) confirmed our observed and quantified pattern (Zhang et al., 2008) (see supplementary online material, Figure B.9).

Nutrient Availability

The NRE is nitrogen limited (Altman and Paerl, 2012); hence, for the nutrient management scenarios the BBN was run only under varying nitrogen concentrations, rather than phosphorus, to examine nutrient management scenarios and their impact on water quality indicators.

A significant decrease in chlorophyll *a* concentration is observed under a low DIN scenario (<56.29 $\mu g/l$). The likelihood of the NRE chlorophyll *a* concentration violating the state criteria decreases to 18.7%. If nutrient concentrations are low, the N:P ratio might be a good predictor for harmful species. The probability of harmful raphidophytes slightly increases by 3.0%, consistent with raphidophytes thriving under low N:P and low nitrogen conditions (Hodgkiss and Ho, 1997).

High DIN concentration scenario (>334.20 $\mu g/l$) is concurrent with very high freshwater discharge, which results in a mixed water column. Dinoflagellates lose

their competitive advantage (motility) over other phytoplankton; hence, their likelihood of presence decreases by 8.2%, under a high DIN. K_d is predicted to increase significantly with likelihood of K_d >2.2 $1/m$ increasing from 20.7% to 65.9% of the time. This is likely the result of the correlation of high DIN concentration with freshwater discharge and resulting elevated colored dissolved organic matter content.

BBN can be used for backward propagation analysis to explore a multitude of scenarios, including violations of NC water quality standards. As an example, chlorophyll *a* would consistently violate the NC water quality standard when freshwater discharge in the bracket above 0.90 m^3/s increases by 6.9% (from 58.8% to 65.7%), stratification greater than 1.04 g/cm^3 increases by 16.1% (from 55.5% to 71.6%) and DIN and PO_4 concentrations of higher than 56.30 and 5.30 $\mu g/l$ increase by 6.5% (from 24.7% to 31.2%) and 11.0% (from 60.0% to 71.0%) respectively (see supplementary online material, Figures B.10 & B.11).

3.5 Applications

The Bayesian belief network model presented in this chapter provides structure to understand and communicate the key dynamics and factors driving water quality and potential state standard violations in the New River estuary. As an aid to estuarine informed management decisions, the BBN can explicitly make predictions of water quality standard violations under varying scenarios. The North Carolina Division of Water Resources (NCDWR), a division of NCDENR, is charged with developing a basin-wide management plan for each basin in the state. Each plan must examine the effects of pollution on the water bodies in terms of their designated uses (e.g., primary recreation, supporting aquatic life) and involve stakeholders in its development. The management of New River estuary, situated in the White Oak basin, includes a diverse set of stakeholders and water managers such as the City of Jacksonville and the Camp Lejeune Marine Corps Base, along with multiple state agencies.

Several scholars have developed guidelines for good practices of incorporating scientific information into decision-making processes, such as basin-wide planning. These guidelines suggest: (1) use of a transparent and simple modeling process (Korfmacher, 1998; Maguire, 2003) ; (2) inclusion of stakeholders in the modeling process (Grayson et al., 1994; Korfmacher, 1998; Maguire, 2003); (3) explicitly addressing and accounting for scientific uncertainty (Reckhow, 1994; Ragas et al., 1999; Huang and Xia, 2001; Borsuk et al., 2001; Maguire, 2003; McDaniels and Gregory, 2004) and (4) use of an adaptive management approach to decision making (Walters, 1997; Failing et al., 2004; Shindler and Cheek, 1999; Borsuk et al., 2001; Maguire, 2003; Smith and Bosch, 2004). The BBN points to the driving forces behind chlorophyll *a* concentrations in the estuary, providing a transparent and fairly simple modeling approach, digestible to the array of stakeholders present in the basin. By manipulating nitrogen concentrations and climatic conditions, along with other factors, the likelihood of water quality standard violations can be calculated. The BBN can be updated as new data is collected, allowing for an adaptive management approach in basin-wide planning. While providing insight into the dynamics of the New river estuary on which management decisions can be based, the BBN also offers a communication tool to interested stakeholders.

3.6 Conclusion

The BBN-quantified effects of nutrient input and likely future climate change on the NRE eutrophication are in agreement with qualitative descriptions from previous studies (Wetz and Paerl, 2008; Conley et al., 2009; Paerl and Scott, 2010). The BBN further highlights the potential impacts of extreme climatic events as well as nutrient management scenarios on the ecological condition of the NRE. Our results also confirm the importance of nutrient input reduction to minimize the presence of harmful algae and avoid violating water quality standards.

The accuracy of the BBN's predictions for the NRE was evaluated with AUC. The moment matching method improved on data discretization by a better representation of the underlying continuous distribution, although discretization still constitutes a limitation of the BBN approach. The BBN presented in this chapter is an initial step that can be followed by a network-based model with continuous variables. The continuous model would build upon the structure of the BBN but further improve the description of relationships between nodes. Furthermore, BBNs' acyclicity cannot handle feedback relationships. This can be addressed using Dynamic Object Oriented Bayesian Networks (OOBN) with each OOBN representing a time step; Implementation of such a model would require a longer dataset to populate conditional probability tables for each time slice of the DOOBN. To further improve the model, separate sub-networks should be developed for the "Chemical Environment component" (including DIN, PO_4 and N:P ratio) with inorganic and organic nutrient composition. Additional improvements, based on existing models such as the Bayesian SPARROW (Qian et al., 2005), can be achieved by developing a sub-BBN on land use/land cover and its impact on nutrient composition and freshwater discharge levels in the NRE. Finally, although temperature indirectly reflects seasonality in our current model, future modeling efforts should account for seasonal variations in the direction and magnitude of relationships.

4

A Comparison of Discretization Methods for Bayesian Networks

4.1 Introduction

Bayesian Networks (BNs) are directed acyclic graphical (DAG) models, which are causal networks that consist of nodes and directed links. The relationships between the variables are described using conditional probability tables (CPTs). BNs are promising tools to aid reasoning and decision making under uncertainty. The term Bayesian Network was first introduced by Pearl (1982) and Spiegelhalter and Knill-Jones (1984) in the field of expert systems. Some of the early appearances of BNs in environmental modeling were by Varis and Kuikka (1997), Varis (1997), and Reckhow (1999).

BNs have several distinct strengths. The main strength of BNs lies in their knowledge updatability based on Bayes's theorem, which is important in the context of adaptive management. The BNs modularity enables integrating multiple system components or aspects of problems (e.g., science network and management network in Johnson et al. (2010)). This is beneficial in environmental modeling due to the

complexity of natural ecosystems and the associated decision-making processes. BNs can accommodate various knowledge sources and data types (e.g., expert knowledge, previous data from same system or other similar systems), with transparent definition of prior knowledge. Among environmental modeling approaches, suitability to both data-rich and data-poor systems is another advantage over other modeling approaches. As new environmental problems arise, monitoring plans start or adapt to accommodate the data requirements. Hence, accommodating minimal data in conjunction with expert knowledge is a methodological advantage. The model can be developed with minimal data and as more information becomes available the model can be updated. Environmental modeling cannot be implemented without incorporating uncertainty from natural ecosystems variability, current knowledge of environmental processes, modeling structure, computational restrictions, and problems with data/observations (due to measurement error or missingness), as it aims to explore complex ecosystems and provide support for the management of natural resources. BNs explicitly represent uncertainty by conditional probability distributions for each node and the uncertainty is propagated through the model and presented in final results.

These advantages of BNs resulted in a large number of applications in ecological and environmental sciences over the last decade, including natural resources management (McCann et al., 2006; Castelletti and Soncini-Sessa, 2007; Dorner et al., 2007; Farmani et al., 2009), ecological risk assessment (Borsuk et al., 2004; Pollino et al., 2007; Barton et al., 2008; Malekmohammadi et al., 2009), and integrated models (Bromley et al., 2005; Croke et al., 2007; Johnson et al., 2010; Kragt et al., 2011).

Aguilera et al. (2011) examined 118 papers published between 1990 to 2010 related to the applications of BNs in environmental sciences. Among these papers, 62 (52.6%) used discrete data and 32 (30.7%) used some form of discretization method to convert continuous data; however, 48.6% of the papers did not include any de-

scription about the process of discretization, 25.7% used experts to discretize the continuous data into intervals, 2.9% used equal interval, 2.9% used equal frequency, and 2.9% used default method of the software (Aguilera et al., 2011). Of all the applied BN papers published in the field of environmental sciences, from January 1990 to December 2010, 34.2% used Netica (`http://www.norsys.com/`) and 20.2% used Hugin (`http://www.hugin.com/`, Madsen et al. (2005)) (Aguilera et al., 2011). Netica provides tools to facilitate discretization of continuous variables. The input data file does not have to be discretized; however, the intervals for each variable should be defined. On the other hand, Hugin software is not capable of allocating continuous data into intervals; hence, input data must be either previously discretized or manually done so during the model development phase in the Graphical User Interface.

Discretization is a process that can result in loss of information but since BNs can handle continuous variables only under severe constraints (1- Each continuous variable be assigned a (linear) conditional Normal distribution; 2- No discrete variable have continuous parents (Nielsen and Jensen, 2009)), data is usually discretized to develop BNs. I am interested in how discretization may affect the resulting model, since different discretization methods will lead to different characterization of variable distributions. In this study, I use a simple example with a large data set to examine the effects of discretization methods on the final model.

4.2 Material and methods

4.2.1 Study design

I designed a study to assess the effects of discretization methods and number of intervals on the developed BN models. The BN presented in this study is a simple one. It consists of three nodes. Figure 4.2 shows the dependency relationships among variables using a DAG model. Two decisions must be made when discretizing a

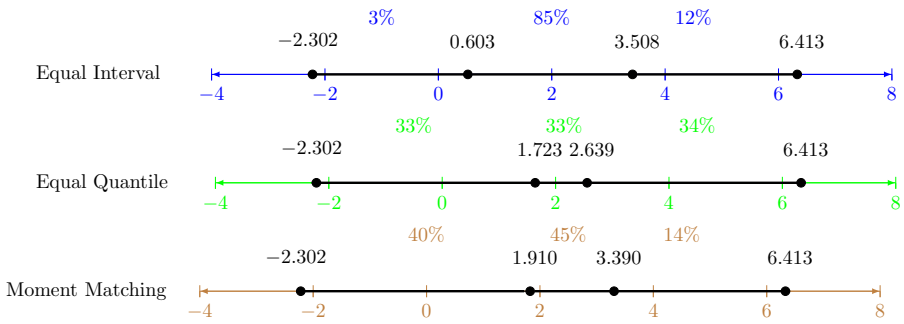


FIGURE 4.1: The figure depicts original data for chlorophyll *a* concentrations from -2.302 to 6.411, discretized using the equal interval, equal quantile, and moment matching methods. The numbers in black show the break points of each method and the percentages show the frequency of observation in each interval.

continuous data set: (1) the discretization method and (2) the number of intervals. Our study assesses the impact of the three most common discretization techniques (described below) and three most common number of intervals on the developed BNs. I categorized the original continuous data set into three (four or five) sets and I named the categories as Low, Medium, and High ("Low, Medium Low, Medium, High" or "Low, Medium Low, Medium, Medium High, and High") (Figure 4.1). Nine BNs were developed, each corresponding to one of the nine combinations of discretizing method and number of intervals. The BN was then fit to the training discretized data using the `bnlearn` package in R (Scutari, 2010; Nagarajan et al., 2013; R Core Team, 2014).

Discretization methods

Equal length interval is a discretization method using which the distribution is divided into equal intervals between the minimum and maximum observed values. It is used frequently because of its simplicity. This discretization method can be problematic in cases where there are outliers in the data set. For example, in the data

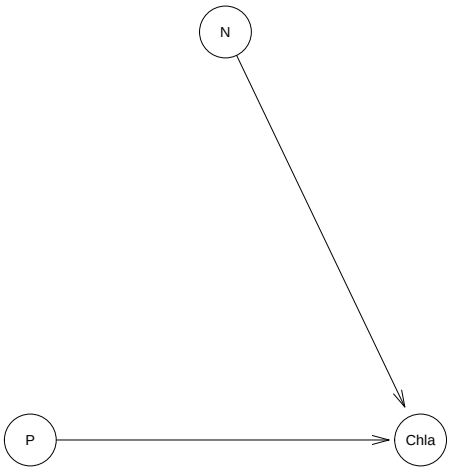


FIGURE 4.2: Directed Acyclic Graph

set I used, several unusually low and high nitrogen concentration values appear after log-transformation. The discretization when the potential outlier values are included results in the following break points: 3.434, 5.665, 7.896, 10.127. The low nitrogen includes ten observations (0.05% of all observations). In the case of chlorophyll *a* including outliers, the low would include only one observation, whereas a minimum of 6 observations in each interval is recommended (Liu et al., 2002). The discretization when the outlier value is removed results in the following break points: 4.500, 5.818, 7.137, 8.455. The discretization is transformation dependent. The definition of low nitrogen changes from < 5.665 to < 5.818 (high from $7.896 <$ to $7.137 <$) in the log scale. Furthermore, when the data is log-normally distributed as is the case with most chlorophyll *a* and nutrient data, equal interval would result in some intervals with high and some with low probabilities. Considering the chlorophyll *a* concentrations (not log-transformed), the discretization will have the following break points using the equal interval method: 5.0×10^{-11}, 1.3×10^{2}, 2.6×10^{2}, 3.9×10^{2}. The low, medium, and high will contain 99.3%, 0.006%, and 0.004% of data respectively.

Equal quantile (a.k.a. equal frequency) discretization method is based on the

frequency of observed values. It divides data into categories of (approximately) equal sample size. The equal quantile method can result in assignment of the same value to different intervals if there are multiple occurrences of the same value (in variables such as secchi disk depth). In equal quantile, the order of the observation is important; however, in equal interval the relative spacing among observations is also critical. Hence, equal interval discretization is transformation-dependent, whereas equal interval discretization changes as the data is log-transformed.

The third method I use in our experiment is moment matching. This method matches the moments of the discretized distribution with the moments of the continuous distribution. As the number of the moments being matched increases, the discrete distribution becomes a more accurate approximation of the continuous distribution. However, as the number of moments to be matched grows, the problem becomes more complex and computationally interactable for more than five intervals.

Number of intervals

Although a model may be more precise as the number of intervals increases, the model is not necessarily more accurate (Marcot et al., 2006). The conditional probability tables, especially in models with more than three layers, will be complicated, as for every state of a given variable, the probability of its occurrence must be assigned given every combination of its parent nodes. Even in a simple causal network, as the one described in this study, the difference between three states and five states for chlorophyll *a* would be assigning nine versus 25 conditional probabilities.

4.2.2 Comparison

Our criteria to compare the developed BNs discretized using different methods were based on Marcot (2012). I used sum of squared errors (SSE), model accuracy, and Area Under Curve (AUC) to assess and compare how well the developed models

predicted the test data. SSE is calculated as the squared discrepancy between the observed data and the mid point of the predicted interval. Model accuracy is calculated as the percentage of total number of cases for which actual interval and predicted interval are equal using the confusion matrix (Marcot et al., 2006; Chen and Pollino, 2012), which is a table with a row and column for each defined interval, whereby each element of the matrix is the number of cases for which the actual interval is the row and the predicted interval is the column. The area under the receiving operating characteristic curve (AUC) varies between 0 and 1. It provides a diagnostic measure of model's prediction accuracy, which represents the probability of a true positive outcome (the proportion of actual observations which are correctly classified) versus a false positive outcome (accuracy of data classification). A model with perfect predictions would have an AUC equal to 1.

4.2.3 Study area

The goal in this chapter was not to study a specific ecosystem or to model certain process. However, for the purpose of demosntrating the impact of discretization on resulting BNs, I used lake monitoring data from Finland reported by Malve and Qian (2006). The large number of lakes in Finland, coupled with long-term monitoring of Finnish lakes, resulted in a rich data-set. I used 19248 July and August observations from a Finnish Lakes data set for total nitrogen (N), total phosphorus (P), chlorophyll *a* (Chl*a*) from 1988 to 2004 (Malve and Qian, 2006). The data set covers 2289 Finnish Lakes which are categorized into nine different types based on the guidelines of the Finnish Environment Institute (SYKE) (Lepisto et al., 2002).

I examine the effect of discretizing method on a model's predictive accuracy using a cross-validation procedure. The Finnish Lakes data set was randomly divided into two subsets for training and testing purposes. The training data set holds 90% of the original data and the testing subset holds 10% of the original data. This process

was repeated 100 times.

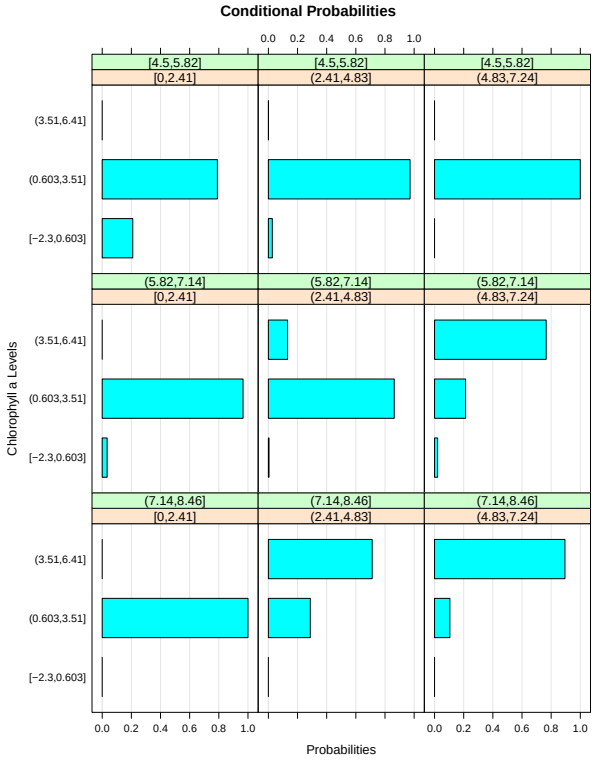
4.3 Results

4.3.1 Conditional probability tables

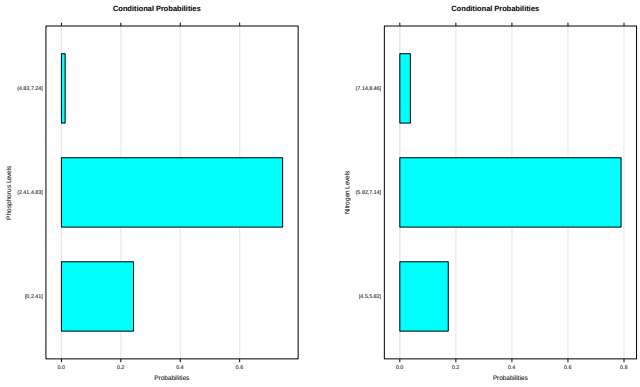
Conditional probability tables (CPTs) represent the probability of a node taking any of its discrete states, given the states of its parent nodes. CPTs describe the relationship among the nodes. Table 4.1 shows the CPT for chlorophyll *a* in the equal quantile discretized BN. I developed the BNs with the same data set; hence, the defined relationships (CPTs) should be similar as well. This is not the case when I discretize the data set with different methods. The probabilities of chlorophyll *a* states under medium nitrogen and low phosphorus are shown in Figures 4.3(a) to 4.5(a) and summarized in Table 4.2. Chlorophyll *a* has a 3% chance of being low under such conditions in the BN discretized with the equal interval methods, whereas it has a 60% and 31% chance of being low in the BN discretized with equal quantile and moment matching methods, respectively. Further, the definition of low, medium, and high is different in each discretization methods, which results in a communication problem. High chlorophyll *a* is defined as concentrations larger than 3.51, 2.64, and 3.39 ($\mu g/l$ in log scale) in BNs discretized using equal interval, equal quantile, and moment matching, respectively. I will further discuss the conceptual and application difficulty as a result of difference among CPTs in subsection 4.3.2.

4.3.2 Prediction

BNs discretized using different methods result in different future predictions, as well. I used the developed BNs using the training data set (90% of the original data set) to predict the testing data set (10% of the original data set). Consider the confusion matrices in table 4.3. The equal interval method predicts Chlorophyll *a* to be in low, medium, and high states with probabilities of 0%, 96%, and 4%,

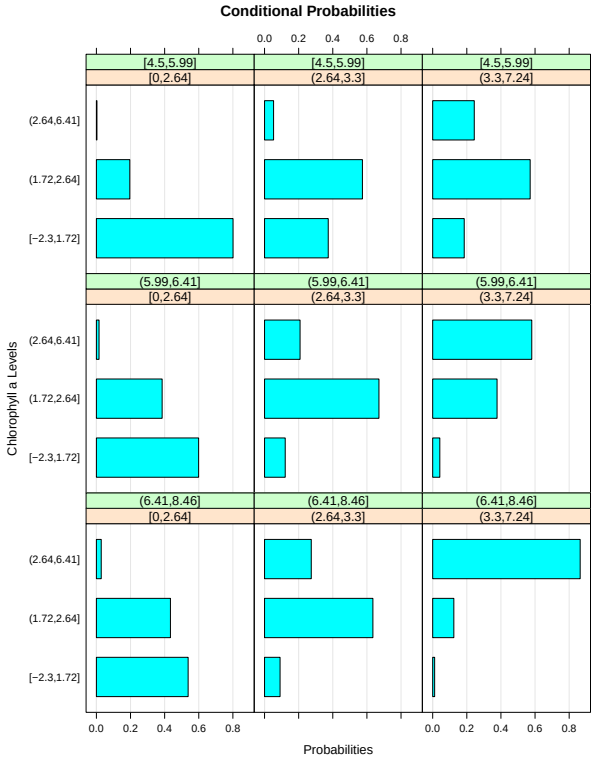


(a) Chlorophyll a

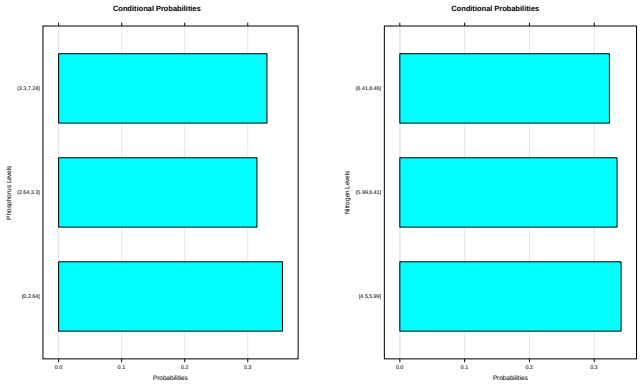


(b) Phosphorus (c) Nitrogen

FIGURE 4.3: Conditional Probability Table for Phosphorus, Nitrogen, and Chlorophyll a - Equal Interval

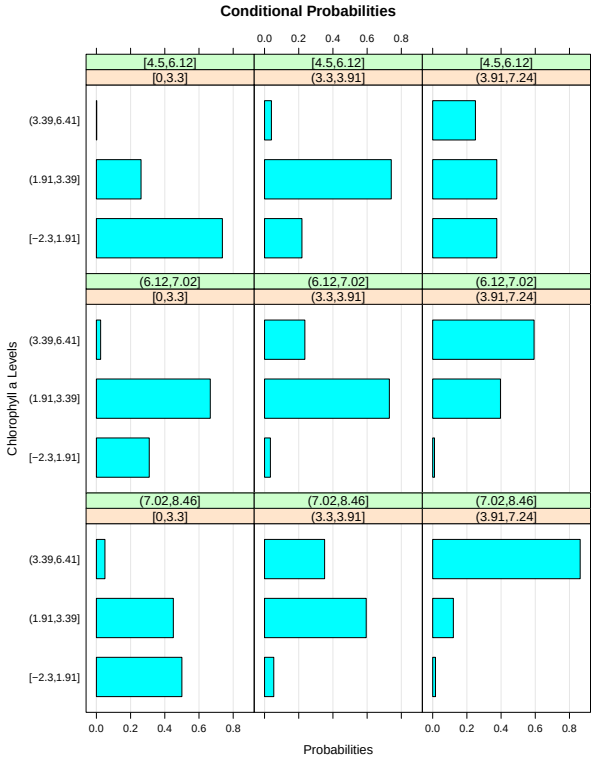


(a) Chlorophyll a

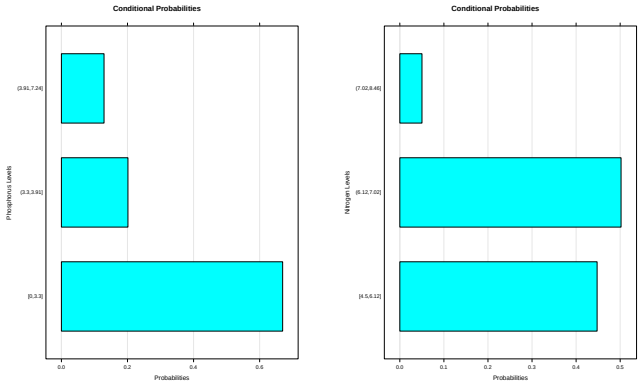


(b) Phosphorus (c) Nitrogen

FIGURE 4.4: Conditional Probability Table for Phosphorus, Nitrogen, and Chlorophyll a - Equal Quantile



(a) Chlorophyll a



(b) Phosphorus (c) Nitrogen

FIGURE 4.5: Conditional Probability Table for Phosphorus, Nitrogen, and Chlorophyll a - Moment Matching

respectively. The equal quantile method predicts Chlorophyll *a* to be in low, medium, and high states with probabilities of 37%, 32%, and 31%, respectively. The moment matching method predicts Chlorophyll *a* to be in low, medium, and high states with probabilities of 41%, 46%, and 13%, respectively. Our predictions for different states of chlorophyll *a* changes significantly from one method to the other. Furthermore, I should note that the definition of states changes from one method to another.

I used multiple measures of performance to compare predictions for the testing data, without one method outperforming the others. Table 4.4 summarizes model comparison results using SSE, Model Accuracy, and AUC as criteria. The data set is large and the considered variables and their underlying relationships is simple and known (4.6). Furthermore, I am optimally fitting the data to each model and often the goodness-of-fit is not a good source of information to differentiate models (Qian and Cuffney, 2012). Hence, the differences shown in Table 4.4 are small. There is no conclusion based on chosen criteria that one method/number of intervals outperforms the others; however, the models differ significantly in the relationships and CPTs as well as scenario investigation (described in subsection 4.3.3).

4.3.3 Management application

BNs are tools for managers and policy makers to assess the impact of their decisions/policies on the ecosystem prior to implementation. Consider a case where the policy makers are assessing the impact of lowering phosphorus on chlorophyll *a*. Table 4.5 summarizes the results of low phosphorus on chlorophyll *a* distribution. Different BNs discretized using different methods result in different conclusions in such analyses. The BN discretized using equal interval method might not conclude that lowering phosphorus is effective, while the BN discretized using equal quantile finds lowering phosphorus effective since it results in low chlorophyll *a* concentrations 66% of the time.

As another case, consider managers targeting policies that would result in low or medium chlorophyll *a* concentrations (avoid high chlorophyll *a*). While the BNs discretized with equal quantile and moment matching methods recommend low and medium phosphorus and nitrogen concentrations, the BN discretized using equal interval recommends medium nitrogen and phosphorus concentrations (see Table 4.6).

4.4 Discussion

The application of BNs in the environmental sciences is justified by their many advantages. It has been argued in the literature though that discretization of data in BNs results in loss of information. Furthermore, there are no guidelines provided on the process of discretization. My goal in this paper was to investigate the effect of different methods of discretization on BNs. If so, the decisions made based on various BNs would consequently be different. I compared nine combinations of discretizing method and number of intervals in this study. The resulting CPTs changed from one method of discretization to the other. The CPTs provide the basis and define the relationships in a BN; hence, any calculation based on them would also be different. The predictions and scenario investigations were different among methods. As discussed in subsection 4.3.3 management recommendations were also different among the developed BNs. In the following paragraphs, I summarize the main drawbacks of discretization.

Firstly, as our results in subsection 4.3.2 showed, our quantified comparison criteria, SSE, model accuracy, and AUC were not able to provide a sound reasoning in favor of one discretization method. Hence, I was unable to provide guidelines without one method outperforming the others in the the defined criteria.

Secondly, I discussed in section 4.2.1 that some of the discussed discretization methods are log-transformation dependent. This is of importance in environmental

sciences, as the data is often log-normally distributed. As the methods are transformation dependent, another question rises about whether the data should be log-transformed or not.

Thirdly, each discretization method resulted in a different definition of categories (low, medium, and high). In our 3-interval case study, the definition of low, medium, and high changes between equal interval, equal quantile, and moment matching methods. The distinct categorical definitions can be a source of miscommunication, as different parties involved in the decision analysis might have different interpretation of low, medium, and high. The categorical/discrete variables as opposed to continuous variables would specially be problematic when expert elicitation rather than data is used to develop the CPTs, as interpretation of defined categories varies among experts. For example, definitions of low chlorophyll *a* varies among scientists and in different contexts.

Finally, the discrepancy in the management recommendations is the main drawback of the discretized BNs. As discussed in the results, the BN discretized using equal interval did not find the lowering phosphorus as effective as the other two methods. If the BN discretized with equal intervals was used to provide recommendations, then the management might decide lowering phosphorous is not cost-effective, although that finding is only the result of discretization. I would caution managers about making decisions based on models for which the outputs vary by a choice (discretization method) that does not have a solid basis.

4.5 Conclusions

The BNs are effective in quantifying uncertainty and valuable tools in environmental modeling. I highlighted the main drawback of the BNs, discretization. I argued that unless solid reason addresses a certain method's superior performance, a continuous data set should not be discretized. However, with the current softwares available for

BNs and the restrictions that come with them, it is not possible to avoid discretization. Future work should focus on developing BNs using continuous data sets.

Table 4.1: Conditional probability table developed using Expected-Maximization algorithm for chlorophyll *a* node in the BN discretized using equal quantile method. Each number represents the probability of chlorophyll *a* taking any of its discrete states, given the states of nitrogen and phosphorus. For example, the last number in the lower left of the table, 0.86, is the probability of chlorophyll *a* concentrations between 2.64 and 6.41 $\mu g/l$ given that nitrogen concentrations are between 6.41 and 8.46 and phosphorus concentrations are between 3.3 and 7.24

	Nitrogen	[3.43,5.99]		
	Phosphorus	[0,2.64]	(2.64,3.3]	(3.3,7.24]
Chlorophyll *a*	[-2.3,1.72]	0.80	0.37	0.18
	(1.72,2.64]	0.20	0.57	0.57
	(2.64,6.41]	0.00	0.05	0.24
	Nitrogen		(5.99,6.41]	
	Phosphorus	[0,2.64]	(2.64,3.3]	(3.3,7.24]
Chlorophyll *a*	[-2.3,1.72]	0.60	0.12	0.04
	(1.72,2.64]	0.39	0.67	0.38
	(2.64,6.41]	0.02	0.21	0.58
	Nitrogen			(6.41,8.46]
	Phosphorus	[0,2.64]	(2.64,3.3]	(3.3,7.24]
Chlorophyll *a*	[-2.3,1.72]	0.54	0.09	0.01
	(1.72,2.64]	0.43	0.64	0.12
	(2.64,6.41]	0.03	0.27	0.86

Table 4.2: Chlorophyll *a* probabilities under low phosphorus and medium nitrogen concentrations scenario under three different discretization methods in a 3-interval BN. For example, the equal interval BN, chlorophyll *a* concentrations will be low, medium, and high with probabilities of 0.03, 0.09, and 0.00, respectively.

	Chlorophyll *a*		
	Low	Medium	High
Equal Interval	0.03	0.97	0.00
Equal Quantile	0.60	0.38	0.02
Moment Matching	0.31	0.67	0.02

Table 4.3: Confusion matrix for chlorophyll *a* in BN discretized using equal interval method and 3-interval. Each element of the matrix is the number of cases for which the actual interval is the row and the predicted interval is the column.

		Predicted		
	Equal Interval	[-2.3,0.603]	(0.603,3.51]	(3.51,6.41]
Observed	[-2.3,0.603]	0	65	0
	(0.603,3.51]	0	1637	17
	(3.51,6.41]	0	165	59

		Predicted		
	Equal Quantile	[-2.3,1.72]	(1.72,2.64]	(2.64,6.41]
Observed	[-2.3,1.72]	533	105	6
	(1.72,2.64]	180	389	107
	(2.64,6.41]	11	121	491

		Predicted		
	Moment Matching	[-2.3,2.56]	(2.56,3.37]	(3.37,6.41]
Observed	[-2.3,1.91]	642	155	1
	(1.91,3.39]	154	657	71
	(3.39,6.41]	2	90	171

Table 4.4: Comparison of predictive accuracy among different discretization methods using SSE, Accuracy, and AUC as criteria with 3 intervals and 5 intervals

	3 Intervals		
	SSE	**Accuracy**	**AUC**
Equal Interval	1818.545	0.871	0.844
Equal Quantile	3681.574	0.728	0.824
Moment Matching	3228.382	0.751	0.807
	4 Intervals		
	SSE	**Accuracy**	**AUC**
Equal Interval	1962.192	0.700	0.803
Equal Quantile	2852.654	0.624	0.772
Moment Matching	2480.924	0.621	0.765
	5 Intervals		
	SSE	**Accuracy**	**AUC**
Equal Interval	1325.392	0.698	0.902
Equal Quantile	2141.934	0.549	0.725
Moment Matching	818.166	0.702	0.843

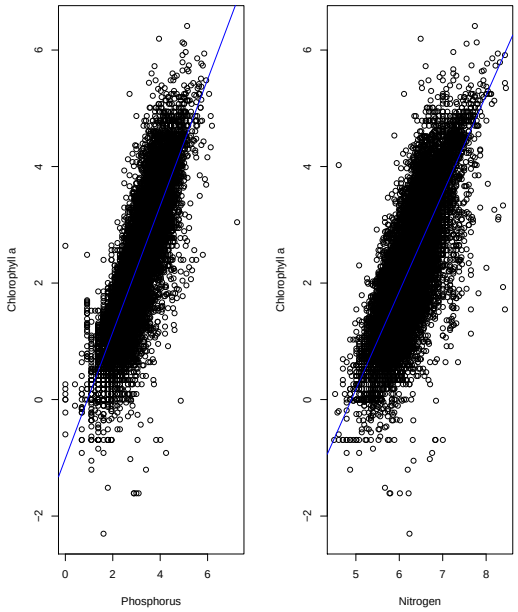


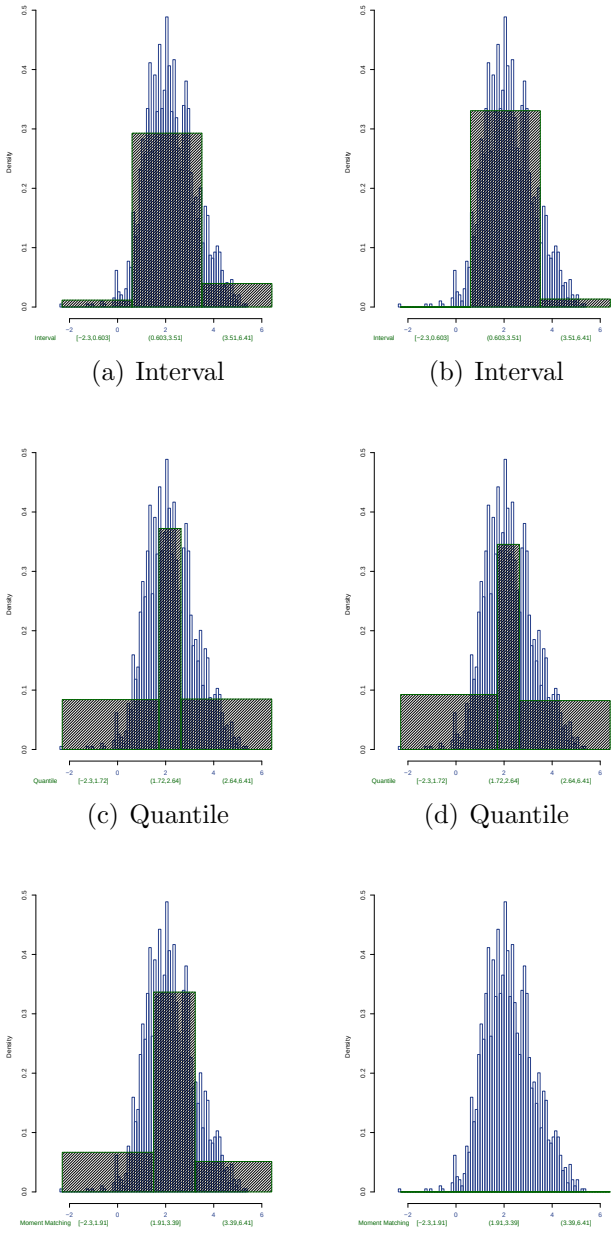
FIGURE 4.6: Log-transformed chorophyl *a* distribution versus log-transformed nitrogen and phosphorus. The blue line depicts the fitted linear regression model between chlorophyll *a* and nutrients

Table 4.5: Probability Table for Chlorophyll *a* under low phosphorus scenario for models discretized using three different methods.

	Chlorophyll *a*		
Method	Low	Medium	High
Equal Interval	0.07	0.93	0.00
Equal Quantile	0.66	0.32	0.02
Moment Matching	0.73	0.26	0.01

Table 4.6: Probability Table for phosphorus and nitrogen under a scenario where chlorophyll *a* concentrations do not exceed medium.

	Phosphorus			**Nitrogen**		
Method	Low	Medium	High	Low	Medium	High
Equal Interval	0.27	0.72	0.01	0.19	0.79	0.02
Equal Quantile	0.46	0.36	0.18	0.41	0.33	0.26
Moment Matching	0.44	0.49	0.07	0.48	0.48	0.04



(a) Interval (b) Interval

(c) Quantile (d) Quantile

(e) Moment Matching (f) Moment Matching

FIGURE 4.7: (a)Left figures: Evaluation data in continuous and discrete form using the equal interval, equal quantile, and moment matching method, respectively from top to bottom, respectively. (b) right figures: Evaluation data in continuous and predicted values in discrete form using the equal interval, equal quantile, and the moment matching method, respectively from top to bottom, respectively.

5

A Continuous Variable Bayesian Network Model for Water Quality Prediction under Uncertainty

5.1 Introduction

Human population growth, particularly in the world's coastal regions, resulted in adverse changes in aquatic ecosystems, largely due to anthropogenic activities such as land use alterations, fertilizer use, industrial activity, and climatic perturbations. One of the central issues in coastal ecosystems, which demands particular attention of managers and policymakers, is eutrophication. Estuaries are particularly susceptible to eutrophication due to riverine nutrient inflow, efficient nutrient trapping, long flushing times, and shallow depth. Key concerns of public and environmental managers for estuaries include water quality, particularly the enrichment of nutrients causing elevated chlorophyll *a* concentrations and depressed oxygen levels (i.e. hypoxia/anoxia) (Kiddon et al., 2003). As human-induced pressures coupled with climate-driven variability will likely continue in the future, models, as decision-making tools, gain more utility for effective environmental management and development of adaptation strategies and restoration plans. An updatable, adapt-

able, communicable, and transferable model would best serve water quality modeling with the purpose of policy/decision analysis, climate/human impact evaluation, and broad stakeholder participation. We propose the continuous variable Bayesian network modeling (cBN) approach to integrate the most valuable attributes of Bayesian Networks (BNs), empirical, and causal/mechanistic modeling approaches.

Jacobs et al. (2005) argues that the limited communication of scientific research to broader audience is mainly due to the complex nature of scientific findings; other studies highlight the critical role of visualization in environmental communication (Cox, 2012). Hence, modeling tools are preferred that are not only powerful in their predictive ability but are also readily communicable to variety of stakeholders. Among such modeling approaches are BNs, with their graphical structure that makes them communicable to a wide variety of stakeholders.

BNs are directed acyclic graphs (DAGs) composed of nodes and links, with embedded conditional probability tables (CPTs) associated with each node. BNs are models with structures based on scientific understanding of the underlying processes and/or empirical investigation of field data (Nielsen and Jensen, 2009). The visual nature of BNs facilitates model communication with decision/policy makers and stakeholders who might not be experts in the underlying scientific disciplines. However, the main drawback of BNs is the requirement of data discretization (Nielsen and Jensen, 2009; Alameddine et al., 2011; Nojavan A. et al., 2014). The discretization captures only certain characteristics of the original underlying distribution. Depending on the method used, discretization often leads to loss of information. The scientific literature currently does not make any recommendations on selection of discretization methods; furthermore, the majority of applied BN papers in the field of environmental sciences do not justify the selection process of their method of discretization (Aguilera et al., 2011).

Mechanistic models represent our best understanding of the causal relationship;

however, a typical mechanistic model often provides too much detail to be adequately calibrated. Hence, results from a mechanistic model are likely uncertain. Further, the calibration process would fit a mechanistic model to a specific data set. In a mechanistic model, casual relationships are based on scientific knowledge; however, the underlying statistical relationships are not justified well.

The commonly argued shortcoming of statistical models, representing correlation but not necessarily causation, is addressed by combining the network-based attribute of BNs with empirical relationships. The BN graphical structure is used for developing and presenting the underlying causal relationship. The links among nodes of a BN imply dependencies and the direction of a link corresponds to the the direction of causality. The network-based structure also reduces the complexity of the model fitting process by fitting each sub-model separately.

Smith (2003) and others (Nixon, 1995; Jørgensen and Richardson, 1996) suggest a common global pattern in aquatic and coastal ecosystems' eutrophication, whereas other studies suggest unique ecosystem specific patterns which may substantially differ in magnitude and trajectory, reflecting complex non-linear and estuary-specific ecological interactions (Cloern, 2001). The opposing views of common versus unique patterns in estuaries have resulted in models developed either for a single estuary (Borsuk et al., 2003) or multiple estuaries (Smith, 2006). In a model developed for a single estuary, the information gained from other similar ecosystems is disregarded; in a multiple ecosystem model, each estuary loses its individual specifications. Our method is a compromise between the two opposing approaches, as it can use another system's model output but it is also ecosystem specific. We provide a general framework to integrate data from other estuaries in two ways described in section 5.3 under spatial model updating and section 5.4.

While some estuaries have been monitored extensively for long periods of time, others have been monitored minimally, if at all. For example, the National Estuar-

ine Eutrophication Assessment (NEEA) lists only eight of the Southern California Bight′s 76 estuaries as study sites and only two of those has adequate data to make an assessment of eutrophic status (McKee et al., 2011). Furthermore, many monitoring programs are initiated when problems occur in a water body; as a result, we may not have adequate data right away for developing models. However, environmental managers should take action immediately and cannot wait for long-term data set availability. The data integration and the Bayesian attribute of the proposed modeling approach enables usage of data from other well-studied estuaries to make informed decisions promptly and not be restricted by data availability.

Data integration and model updating can be done spatially or temporally in the proposed methodology. The spatial model updating is done among a set of comparable estuaries. This is useful when developing models for estuaries with no or limited data and it can be achieved by introducing models developed from similar estuaries as a prior model. The model from the former ecosystem(s) would act as a prior for estuaries with no or limited data. The model can then be updated using data from the latter estuary through Bayes' theorem as it becomes available (posterior $\propto$ likelihood of observed data $\times$ prior). As for temporal model updating, it is done on one estuary as new data becomes available. This is beneficial for studying gradual environmental changes such as climate/land use change, as well as implementing adaptive management. For more details on spatial and temporal model updating, refer to sections 5.3.2 and 5.3.3.

In this chapter, we explore a continuous variable Bayesian network modeling (cBN) approach, which provides a general framework that can be used in single or multiple estuarine ecosystems to investigate eutrophication and the effects of climate and land use variation on water quality. Our goal in this chapter is to propose a methodology and demonstrate its performance with emphasis on management. We are not aiming at understanding a specific system. We demonstrate an application

of the proposed modeling approach for the New River Estuary, North Carolina, USA as a case study.

5.2 Materials and Procedures

5.2.1 Materials - Data set

The New River Estuary (NRE, $\sim$ 1435 km^2) located in North Carolina, USA is a shallow broad lagoon. The NRE begins upstream of the city of Jacksonville, NC and after approximately 30 km, it discharges to the Atlantic Ocean. The National Atmospheric and Oceanic Administration (NOAA) listed the NRE as one of the four estuaries with high expression of eutrophic conditions within the south Atlantic region (Bricker et al., 1999). The proposed model in the present study was developed with water quality monitoring data from the Defense Coastal/Estuarine Research Program (DCERP), Aquatic Estuarine Monitoring component (RTI, 2013). Eight stations along the axis of the NRE (Figure 5.1) were sampled on a monthly basis, starting in October 2007, for a range of physical, chemical, and biological variables. In this study, we used data from October 2007 to September 2011 to build the model and data from October 2011 to October 2012 to assess the model.

The variables in the model included: temperature ($^{\circ}C$, representing seasonal variation), stratification (density gradient, bottom water density minus surface water density, g/cm^3), salinity (used in calculating stratification, psu), light attenuation coefficient ($K_d, 1/m$), dissolved inorganic nitrogen (surface DIN, $\mu g/l$), total dissolved nitrogen (surface TDN, $\mu g/l$), orthophosphate (surface PO_4, $\mu g/l$), chlorophyll a (surface algal biomass in $\mu g/l$), and bottom water dissolved oxygen (O_2, mg/l).

Ott (1995), using the central limit theorem, demonstrates that environmental concentration variables are log-normal, which justifies log-transformation of all nutrient and chlorophyll a concentration data prior to statistical analyses in the present chapter; we note that the interpretation of regression model coefficients are different

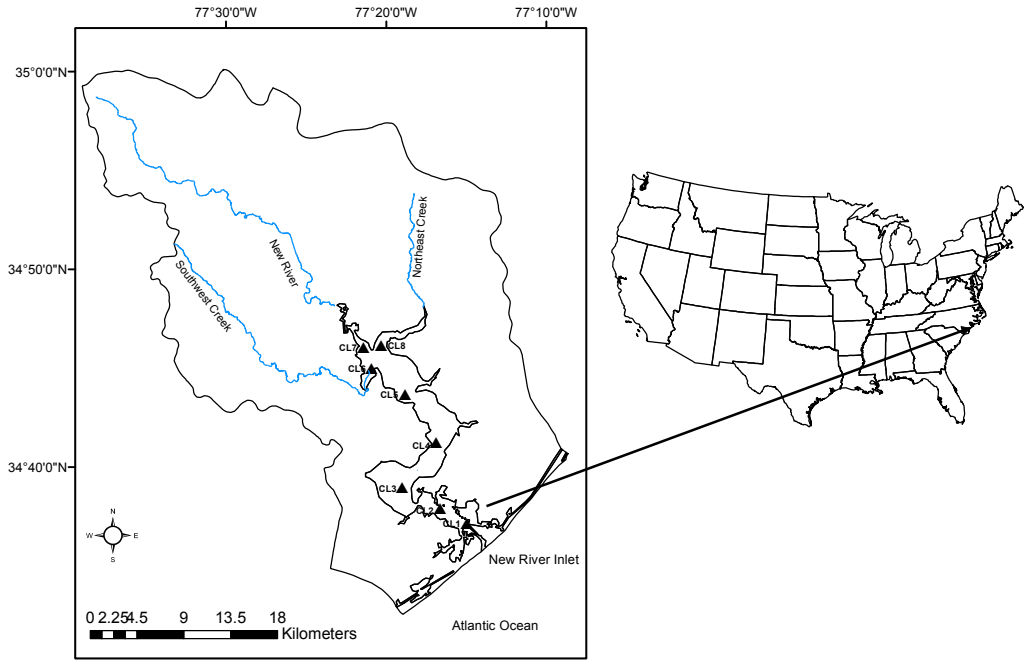


FIGURE 5.1: The New River Estuary (NRE), study area, is located in Onslow County, North Carolina, USA. High chlorophyll *a* concentration, low bottom water dissolved oxygen, and harmful algal blooms is an ongoing problem in the NRE.

when log-transformed (Qian, 2010). Further, all predictors in our case study data set were scaled based on the discussion of Gelman and Hill (2007) and Gelman (2008) on scaling predictors to simplify the interpretation of the intercept when predictors cannot be set equal to zero. Scaling also improves the interpretation of coefficients in models with interacting terms (e.g., in the present model the interaction between salinity and nitrogen, see subsection 5.2.2), and coefficients can be interpreted on approximately a common scale. Weisberg (2005) also demonstrates that centered predictors would result in uncorrelated regression model coefficients.

5.2.2 Procedures – Rationale

The BN model in Chapter 3 was the first statistical modeling effort using the DCERP data set for the NRE since the sampling started in October 2007 (RTI, 2013). The

BN was an initial step toward quantifying our understanding of the eutrophication process and factors leading to high chlorophyll *a* and low oxygen concentrations in the NRE. The BN provided a platform to communicate the results with the stakeholders and investigate management scenarios. The results from Chapter 3 are in agreement with qualitative descriptions in the literature (Wetz and Paerl, 2008; Conley et al., 2009; Paerl and Scott, 2010). The BN further highlights the potential impacts of extreme climatic events as well as nutrient management scenarios on the ecological condition of the NRE. Our results also confirm the importance of nutrient input reduction to minimize the presence of harmful algae and avoid violating water quality standards. However, the discretization process essential for developing BNs resulted in loss of information and limited the model application. In this section, we describe some of the drawbacks of discretization and provide reasoning for proposing a continuous variable Bayesian Network model (cBN).

Many software packages (e.g., Hugin) limit the discretizing break point values to be among the observed values. For example, when discretizing bottom dissolved oxygen values into three intervals using moment matching methods, we have low defined as bottom dissolved oxygen from 0.73 to 4.04 (mg/l), medium from 4.04 to 10.26 (mg/l), and high from 10.26 to 13.44 (mg/l). The North Carolina state criterion for the bottom dissolved oxygen is 4 (mg/l) (EPA, 2001b); however, we were unable to fix the break point at 4 since it was not among the observed values. Also, future samples might not fall within the current observed range from 0.73 to 13.44 (mg/l), which would result in an error during the model update procedure, depending on the software being used to develop the model. Another drawback of discretization is that once the model is developed and the break points are fixed, we no longer have the flexibility of working with continuous data. For example, in the developed BN, we are not able to examine the probabilities associated with anoxia, bottom dissolved oxygen concentrations lower than 2 (mg/l), because 2 (mg/l) is not

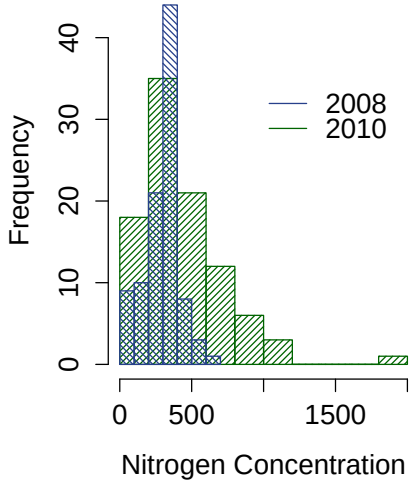


FIGURE 5.2: Histogram of nitrogen concentration in the New River Estuary for 2008 and 2010. The figure depicts the different ranges of observed nitrogen concentrations during the two years due to different precipitation patterns. Hence, the definition of low, medium, and high would be different in 2010 compared to 2008.

a break point in the developed model.

Once a continuous variable is discretized, the resulting categories (e.g., low, medium, high) may be interpreted differently under different circumstances. For example, using our data, the nitrogen concentration in the BN is discretized into three categories: [5.57-56.30), [56.30-334.21), and [334.21-1269.20] ($\mu g/l$), representing low, medium, and high, respectively. Because nitrogen concentration distributions in NRE in 2008 and 2010 are very different (Figure 5.2), the meanings of low, medium, and high are also different between 2008 and 2010.

Our proposed model addresses these discretization problems by developing empirical models among connected variables to replace CPTs. We illustrate the process of building a cBN model by developing a model for the NRE, using chlorophyll *a* concentration and bottom water dissolved oxygen as indicator variables, to demonstrate the proposed model's development process.

Network Structure

Our proposed methodology begins with developing a graphical representation of the key environmental variables of interest (the response variables, chlorophyll *a* concentrations and bottom water dissolved oxygen) and factors affecting them (predictors, nitrogen, phosphorus, light, salinity, stratification, temperature, season, and section) in the studied ecosystem. The predictor variables for chlorophyll *a* and bottom water dissolved oxygen were established based on a combination of previous findings and exploratory data analysis (e.g., scatterplot matrices and multivariate conditional scatterplots). It has been shown in the literature that chlorophyll *a* concentration in estuaries are affected mainly by light, nutrients, water column mixing, temperature, and grazing (Ryther and Dunstan, 1971; Cloern, 1987; Koseff et al., 1993; Conley et al., 2009). Furthermore, oxygen levels in the water column are also influenced by water column mixing, temperature, and chlorophyll *a*. Based on the exploratory analysis described in the appendix, we present our model graphically in the form of a DAG (Figure 5.3).

Model Formulation

The relationship between chlorophyll *a*, bottom dissolved oxygen and their predictors is examined using simple linear regression model as an initial step. We compared different empirical relations based on goodness-of-the-fit statistics such as R^2. The simple linear regressions are expressed in terms of conditional probability distributions. The conditional probability distributions are then combined based on the DAG model. For example, in the oxygen component shown in Figure 5.5, the bottom water dissolved oxygen is the variable we are interested in predicting. The predictors based on the analysis described in Section 5.2.2 are chlorophyll *a*, stratification, and temperature. The log-transformed bottom dissolved oxygen has a normal distribution. The mean is calculated as a regression model between the important predictors

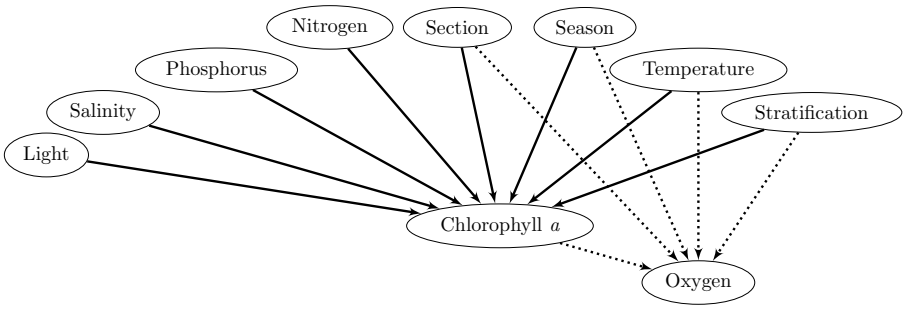
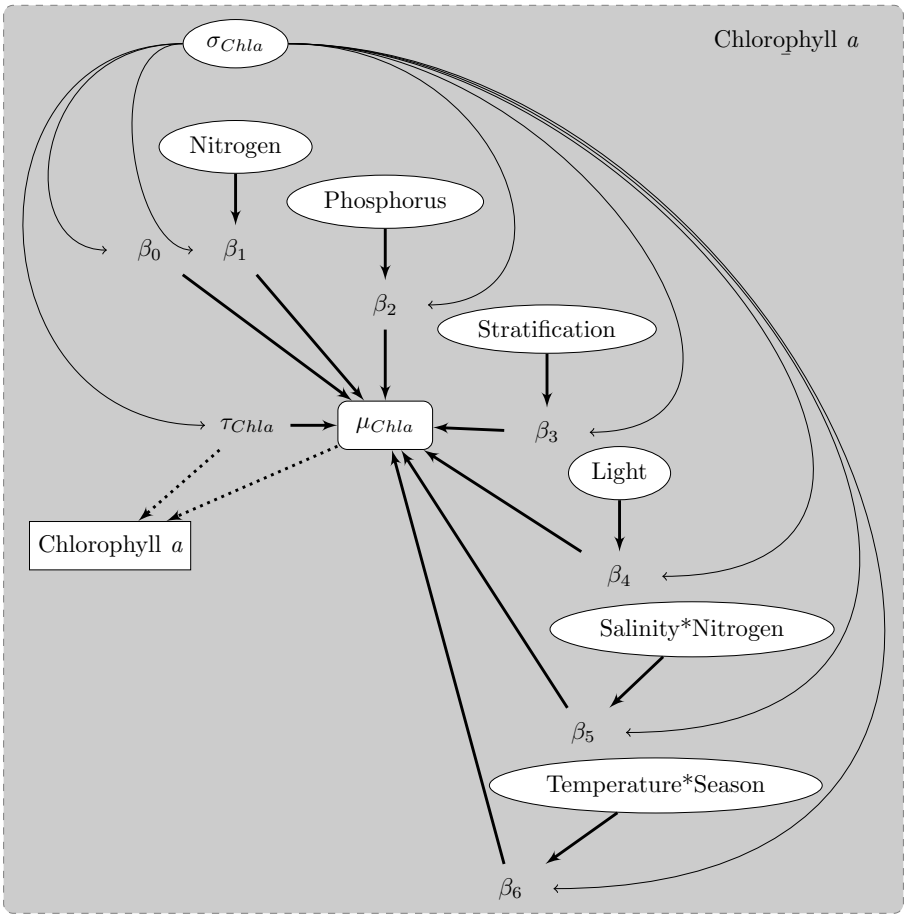


FIGURE 5.3: Directed Acyclic Graphical (DAG) model depicts the variables of interest, chlorophyll *a* and bottom dissolved oxygen, and their predictors.

(see Figure 5.3). The standard deviation of the normal distribution, as well as each parameter in the regression model, is then assigned non-informative prior distributions. When fitting the oxygen component model separately, observed chlorophyll *a* is used as the predictor. Figures 5.5 and 5.4 describe the chlorophyll *a* and bottom water dissolved oxygen network-based models. The equations are further described in the appendix.

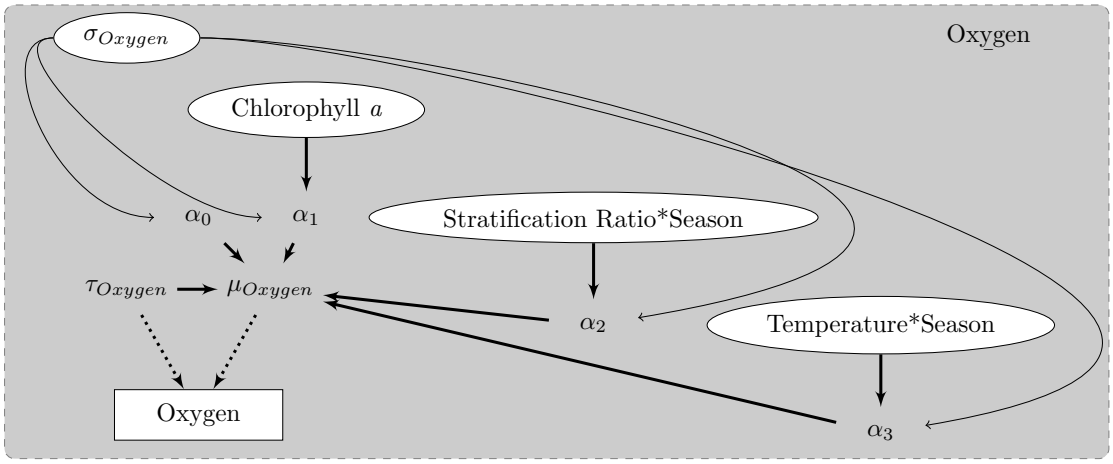
Connecting Components

The nodes (and the variables represented by the nodes) in the combined model are classified into three groups: (1) forcing nodes (nodes without parents, e.g., temperature), (2) intermediate nodes (with both parents and child, e.g., chlorophyll *a*), and (3) terminal nodes (without child, e.g., oxygen). The pre-specified causal network, DAG (see Figure 5.3) specifies the marginal distributions of forcing nodes. The con-



$$Chlorophyll\ a_i \sim \mathcal{N}(\mu_{Chlorophyll\ a}, \tau^2_{Chlorophyll\ a})$$
$$\mu_{Chlorophyll\ a} = \beta_0[Section_i] + \beta_1 Nitrogen_i + \beta_2 Phosphorus_i + \beta_3 Stratification_i + \beta_4 Light_i + \beta_5 Salinity_i \times Nitrogen_i + \beta_6[Season_i] \times Temperature_i$$

FIGURE 5.4: The chlorophyll a model shows chlorophyll a as the response variable with its predictors. The equation below the figure describes the distribution of log-transformed chlorophyll a as normal. The corresponding mean is calculated from a regression model developed for chlorophyll a and its predictors. Uninformative priors are placed on the coefficients. For detailed information on the equations, please refer to the appendix.



$$
\begin{aligned}
&Oxygen_i \sim \mathcal{N}(\mu_{Oxygen}, \tau^2_{Oxygen}) \\
&\mu_{Oxygen} = \alpha_0[Section_i] + \alpha_1 Chlorophyll\ a_i + \alpha_2[Season_i] \times Rstratification_i \\
&\qquad\qquad +\alpha_3[Season_i] \times Temperature_i
\end{aligned}
$$

FIGURE 5.5: The oxygen model shows bottom dissolved oxygen as the response variable with its predictors. The equation below the figure describes the distribution of log-transformed bottom dissolved oxygen as normal. The corresponding mean is calculated from a regression model developed for bottom dissolved oxygen and its predictors. Uninformative priors are placed on the coefficients. For detailed information on the equations, please refer to the appendix.

ditional probability distributions of all intermediate and final nodes are based on the regression analysis in subsection 5.2.2. The final model is depicted in Figure 5.6.

The goal in this section is to define the joint distribution of all variables. The $A = \{\text{Nitrogen}, \text{Phosphorus}, \cdots, \text{Light}, \text{Chlorophyll}a, \text{Oxygen}\}$ is the collection of all variables. The causal diagram defines conditional dependency of the component variables. The forcing nodes have marginal distributions (e.g., $Pr(\text{Nitrogen})$), whereas the intermediate or terminal nodes have conditional probability distributions (e.g., $Pr(\text{Oxygen})|Chlorophyll a, Stratification \times Season, Temperature \times Season$). Using the marginal distributions and conditional probability rule, we can assemble the joint distribution for A. The conditional distributions are relatively easy to find; hence, our approach simplifies computation.

The joint distribution can be characterized using a Markov chain Monte Carlo (MCMC) simulation method (Qian et al., 2003), once the set of full conditionals is specified. We created a Markov process and ran it long enough to approximate the joint probability distribution. The joint probability distribution was then used to predict 2011-2012 conditions. It can help answer questions such as which conditions would lead to high chlorophyll *a* and hypoxia/anoxia and help develop strategies to avoid undesirable conditions. Statistical analyses were performed using R 3.0.2 (R Core Team, 2014) and JAGS (Plummer et al., 2003).

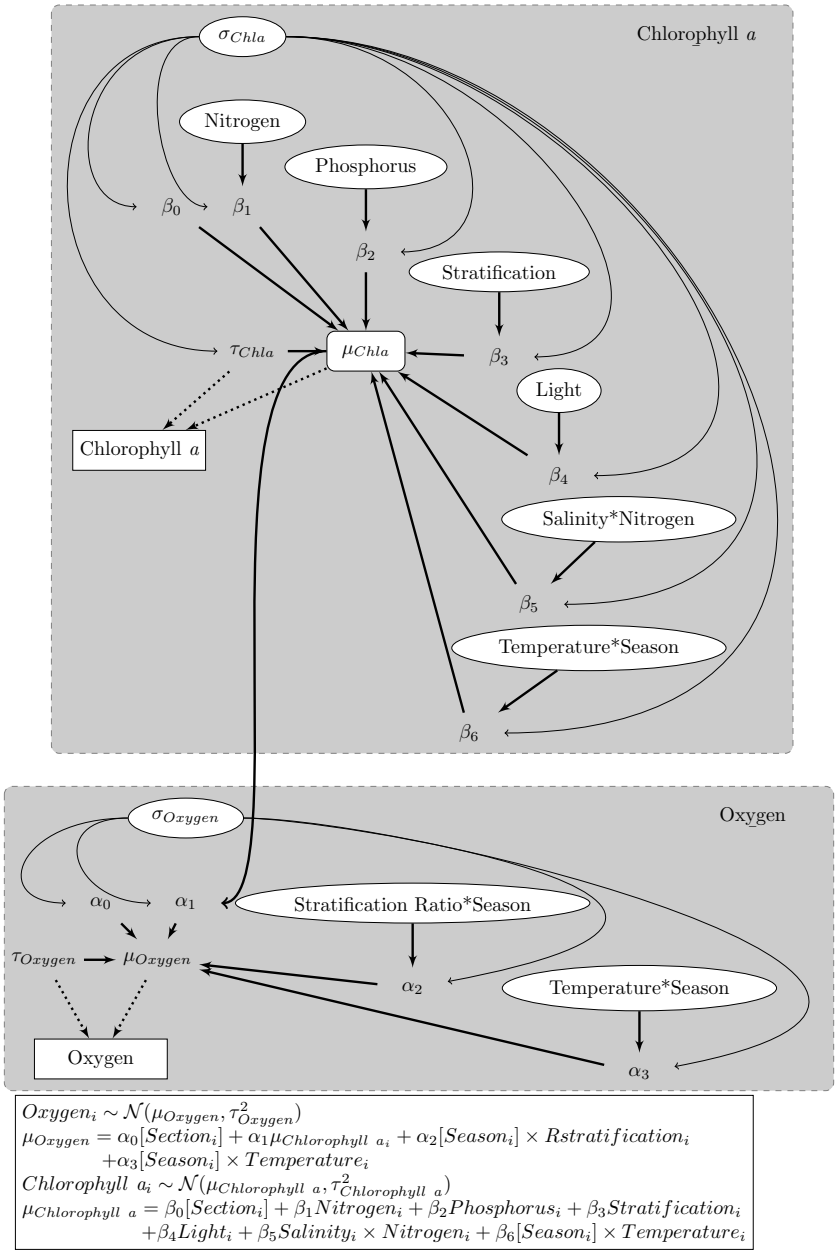
5.3 Assessment

In the assessment section, we describe the following:

- Fitted versus observed (for assessing the goodness-of-the-fit)
- Predicted versus observed (not used in model fitting)
- Temporal updating
- Spatial updating

5.3.1 Model Performance

We developed the model using the NRE data from October 2007 to September 2011. The developed model was then used to predict the observations from October 2011 to October 2012. First, the posterior distribution for coefficients $\alpha s'$ and $\beta s'$ (Table 5.2) were calculated using the developed model. The predictive distribution for chlorophyll *a* and dissolved oxygen levels for the period of 2011 to 2012 were then calculated and compared to the observed values. The comparison was done by examining the model's ability to predict the mean, median, violation of state criteria for chlorophyll *a* (chlorophyll *a*$> 40\ \mu g/l$) and dissolved oxygen concentrations (DO$< 4\ mg/l$) (Table 5.1).



$Oxygen_i \sim \mathcal{N}(\mu_{Oxygen}, \tau^2_{Oxygen})$
$\mu_{Oxygen} = \alpha_0[Section_i] + \alpha_1 \mu_{Chlorophyll\ a_i} + \alpha_2[Season_i] \times Rstratification_i$
$+\alpha_3[Season_i] \times Temperature_i$
$Chlorophyll\ a_i \sim \mathcal{N}(\mu_{Chlorophyll\ a}, \tau^2_{Chlorophyll\ a})$
$\mu_{Chlorophyll\ a} = \beta_0[Section_i] + \beta_1 Nitrogen_i + \beta_2 Phosphorus_i + \beta_3 Stratification_i$
$+\beta_4 Light_i + \beta_5 Salinity_i \times Nitrogen_i + \beta_6[Season_i] \times Temperature_i$

FIGURE 5.6: The combined continuous variable Bayesian network model combines individual models in Figures 5.5 and 5.4 to develop the final model. The nodes (and the variables represented by the nodes) in the combined model are classified into forcing nodes (nodes without parents- e.g., temperature), intermediate nodes (with both parents and child- e.g., chlorophyll *a*), and terminal nodes (without child- e.g., oxygen). Operationally, the combination process is complete when observations for intermediate nodes are replaced by their respective means.

Table 5.1: Results of predictive capability of the original model for the New River Estuary data (2011-2012). As evaluation criteria, we compare the predictive value versus the observed value for chlorophyll a violation (chlorophyll $a > 40\ \mu g/l$), bottom dissolved oxygen violation (bottom dissolved oxygen $< 4\ mg/l$) and their means and medians.

Variable	Predicted (95% CI)	Observed	Error (%)
$\Pr(Chla > 40\ \mu g/l)$	(0.07,0.11)	0.11	-19
$Chla$ Mean	(1.32, 2.64)	2.45	-18
$Chla$ Median	(1.22, 2.73)	2.47	-20
$\Pr(Oxygen < 4\ mg/l)$	(0.09, 0.21)	0.17	-15
$Oxygen$ Mean	(1.59, 2.13)	1.73	9
$Oxygen$ Median	(1.62, 2.21)	1.93	1

Table 5.2: Means, standard deviations, 2.5% quantile, and 97.5% quantile of coefficients of the developed, temporally updated, and spatially updated models

	Original Model				**Temporal Update**				**Spatial Update**			
	mean	**sd**	**2.5%**	**97.5%**	**mean**	**sd**	**2.5%**	**97.5%**	**mean**	**sd**	**2.5%**	**97.5%**
$\alpha_0[1]$	2.12	0.08	1.97	2.28	2.13	0.07	1.99	2.27	1.41	0.18	1.07	1.76
$\alpha_0[2]$	2.06	0.07	1.92	2.21	2.07	0.07	1.94	2.19	1.62	0.17	1.29	1.95
$\alpha_0[3]$	2.04	0.05	1.94	2.14	2.05	0.05	1.96	2.14	1.52	0.15	1.23	1.82
α_1	-0.07	0.03	-0.13	-0.01	-0.08	0.03	-0.13	-0.03	-0.06	0.07	-0.21	0.07
$\alpha_2[1]$	-0.07	0.03	-0.13	-0.00	-0.05	0.03	-0.11	0.01	-0.41	0.05	-0.50	-0.32
$\alpha_2[2]$	-0.26	0.03	-0.31	-0.21	-0.26	0.03	-0.31	-0.21	-0.68	0.05	-0.77	-0.58
$\alpha_2[3]$	-0.24	0.22	-0.66	0.19	-0.39	0.07	-0.51	-0.26	-0.42	0.05	-0.52	-0.32
$\alpha_2[4]$	0.00	0.03	-0.05	0.05	0.01	0.03	-0.04	0.06	-0.07	0.07	-0.20	0.06
$\alpha_3[1]$	-0.23	0.04	-0.31	-0.16	-0.23	0.04	-0.30	-0.15	-0.57	0.07	-0.70	-0.44
$\alpha_3[2]$	-0.29	0.03	-0.35	-0.23	-0.30	0.03	-0.36	-0.24	-0.76	0.05	-0.86	-0.66
$\alpha_3[3]$	-0.40	0.20	-0.79	-0.01	-0.53	0.09	-0.70	-0.36	-0.72	0.09	-0.89	-0.55
$\alpha_3[4]$	-0.28	0.03	-0.33	-0.22	-0.27	0.03	-0.32	-0.22	-0.50	0.04	-0.59	-0.42
$\beta_0[1]$	2.24	0.11	2.02	2.46	2.35	0.10	2.16	2.54	2.42	0.05	2.32	2.51
$\beta_0[2]$	2.14	0.12	1.90	2.37	2.21	0.11	2.00	2.43	2.37	0.03	2.30	2.44
$\beta_0[3]$	1.74	0.11	1.52	1.96	1.83	0.10	1.63	2.03	2.41	0.05	2.31	2.52
β_1	0.15	0.09	-0.03	0.33	0.18	0.08	0.02	0.34	-0.07	0.04	-0.14	-0.00
β_2	0.08	0.07	-0.06	0.23	0.07	0.07	-0.06	0.20	0.02	0.03	-0.04	0.08
β_3	0.10	0.06	-0.03	0.23	0.06	0.06	-0.05	0.17	-0.00	0.02	-0.04	0.04
β_4	0.27	0.06	0.14	0.39	0.28	0.06	0.16	0.39	0.36	0.03	0.31	0.41
β_5	0.06	0.11	-0.16	0.28	0.08	0.10	-0.12	0.28	-0.19	0.05	-0.29	-0.08
$\beta_6[1]$	0.04	0.11	-0.18	0.25	-0.00	0.10	-0.21	0.20	-0.23	0.06	-0.34	-0.12
$\beta_6[2]$	0.08	0.62	-1.14	1.29	0.74	0.20	0.35	1.13	0.05	0.07	-0.09	0.20
$\beta_6[3]$	-0.01	0.09	-0.19	0.18	0.03	0.08	-0.13	0.20	-0.06	0.03	-0.12	0.01
$\beta_6[4]$	0.06	0.05	-0.04	0.15	0.06	0.04	-0.03	0.15	0.21	0.02	0.16	0.26

5.3.2 Temporal Model Updating

As discussed in section 5.1, ecosystem managers and policy makers need tools that can help learning from experience and enable them to manage the ecosystem as new knowledge becomes available. Several studies have called for adaptive management of eutrophication (Rabalais et al., 2002; Stow et al., 2003). In this section, a Bayesian model updating has been implemented after additional data for the period of 2011 to 2012 was acquired. The Bayesian model updating is based on the repeated use of the Bayes' theorem, whereby the posterior of the model developed in the methods section with non-informative priors and the data from September 2007 to October 2011 is used as the prior for the Bayesian model updating step. The regression model coefficients are modeled by multivariate-normal distributions with means equal to a vector that consists of the means of posterior distributions of α's and covariance matrices equal to the covariance matrices multiplied by 100 from the combined model runs, respectively. $\sigma_{Chlorophyll\ a}$ and σ_{Oxygen} have scaled inverse χ^2 distributions (informative prior) with parameters based on posterior distribution of σ_{Oxygen} ($\sigma_{Chlorophylla}$) in the combined model run. The additional data from 2011 to 2012 is used to form the likelihood function at the Bayesian model updating step. The fitted model can be updated with the new data one year (or other time periods depending on the frequency of monitoring and the need for updated information) at a time.

Apart from the benefit of updating the model to evaluate the efficacy of the new management strategies, the output from the updated model can be used to evaluate the validity of the network model and parameterized relationship in the previous step. This comparison was done using QQ plots to detect any significant changes between the distribution of coefficients before and after the temporal model updating step. A QQ plot is a visual way to compare the distribution of coefficients prior and

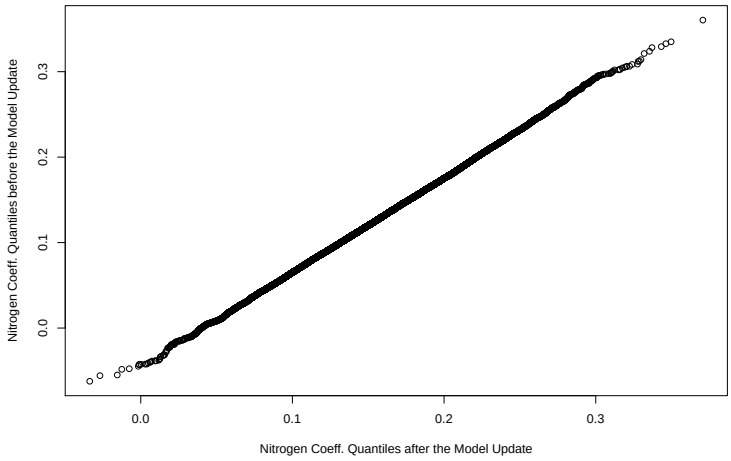


FIGURE 5.7: QQ plot for nitrogen coefficient before and after temporal model updating.

post the update. The QQ plots do not show any changes for model coefficient during the temporal updating process (Figure 5.7). However, if the QQ plots detect any significant changes, further investigation is required to decide the reason behind the change. One speculation of such a change would be that the model is no longer adequate to describe the system. This might have been the result of a change in the system, such as a nutrient management scenario that has altered the dynamics of the estuary. The discrepancy between the prior and posterior distributions offers further insights into the prediction capability of the model.

5.3.3 Spatial Model Updating

We built the original model with the goal of transferability to other similar estuaries. The developed model could have been further simplified if we were only targeting the NRE. For example, phosphorus was included while the NRE specific data set did not confirm a significant role for the phosphorus in predicting chlorophyll *a*; however, it has been shown in the scientific literature that phosphorus is a critical nutrient for phytoplankton growth (Conley et al., 2009). In the following section, we will

describe spatial model updating steps for another neighboring estuary.

The spatial model updating step is demonstrated using the Neuse River Estuary data set. The Neuse River Estuary has similar dynamics and concerns of eutrophication as of the NRE. The Neuse River Estuary is a shallow drowned river valley estuary located in central coastal North Carolina, USA, with a length of approximately 70 km. The Neuse River Estuary, similar to the NRE, is a mesotrophic to eutrophic ecosystem. It has a history of hypoxia/anoxia, harmful algal blooms, and fish kills. The stratification is stronger in comparison to the NRE; however, the wind pattern and tidal regimes are similar.

The spatial model updating steps and procedure is similar to temporal model updating. We used the data sampled between January 2007 and December 2012. Again, we divide the data into two subsets: a training dataset (January 2007 to December 2011) and a verification data set (Jan., 2012 to Dec, 2012). The training dataset was used to update the model from the NRE. The verification data was used to predict chlorophyll a and bottom dissolved oxygen using the updated model. We then compared the predictions to the observed values. Table 5.3 shows the results of the comparison.

Table 5.3: Results of predictive capability of the spatially updated model for the Neuse River Estuary data (2011-2012). As evaluation criteria, we compare the predictive value versus the observed value for chlorophyll a violation (chlorophyll $a > 40$ $\mu g/l$), bottom dissolved oxygen violation (bottom dissolved oxygen < 4 mg/l) and their means and medians.

Variable	Predicted (95% CI)	Observed	Error (%)
Pr($Chla > 40$ $\mu g/l$)	0.02	0.01	34
$Chla$ Mean	2.33	2.37	-2
$Chla$ Median	2.32	2.29	1
Pr($Oxygen < 4$ mg/l)	0.53	0.51	3
$Oxygen$ Mean	1.22	0.88	39
$Oxygen$ Median	1.31	1.37	-5

5.4 Discussion

The proposed cBN retains the advantages of BNs. It has the graphical network-based structure, which eases the communication of science and also depicts the causal relationships and the dependencies among variables of interest. The cBN is also suitable for any sample size data sets. Furthermore, different knowledge sources can be used to form the prior distributions of the model. The BNs discretize data sets into intervals/categories, which results in loss of information. There are different methods of discretization described in the literature; however, no guideline on usage of discretization methods has been provided. The proposed cBN avoids the pitfalls of discretization and uses empirical modeling to establish the connections among nodes. The relation among variables in the proposed cBN is described by empirical models, unlike BNs where the relation among variables are described by CPTs. However, the cBN can accomodate process-based models by using the equations from a process-based model and then describing priors for the coefficients.

The proposed cBN facilitates temporal updating. As new information becomes available for the ecosystem, the Bayesian attribute of the proposed model facilitates temporal updating. We demonstrated an example of temporal model updating in subsection 5.3.2. The model can further be used in other ecosystems with similar dynamics. We called this procedure spatial updating. As an example, we used the model from the NRE for the Neuse River Estuary, which is a similar non-tidal estuary.

Complex ecosystem dynamics and uncertain conditions require more than merely ecosystem monitoring and single step models. Managers and policy makers must modify their strategies and action plans as new information becomes available. The utility of the proposed model is in its capacity to study the patterns promptly, as new conditions unfold. The updatable nature of the Bayesian model captures the essence

of adaptive management by enabling learning from experience (i.e., data). The key concept in adaptive management is iterative learning. The requirements of iterative learning are (1) observing the ecosystem to gauge the impact of policies and management actions, continuously, (2) communicating the ecosystem's status with policy makers and managers, (3) updating the management actions and recommendations. Our proposed model facilitates the latter two requirements. Our proposed model simplifies communication with policy makers and managers with its network-based structure. It also provides a straightforward ability to assimilate new information by using a Bayesian approach. In long-term monitoring programs, new data become available every day/week/month. It would be greatly beneficial for managers/policy makers to update the model in some time intervals depending on the frequency of sampling and the temporal resolution of the problem. Here, the posterior distribution calculated in the previous model run step would be considered an updated prior distribution. An updated posterior distribution can then be computed via Bayes' theorem using new data. Based on the updated posterior, effectiveness of previous policies/strategies would be evaluated and new recommendations would be provided.

Adaptive management is also intertwined with uncertainty. The hypothesis in adaptive management is that our decisions have uncertain outcomes and managers should update their understanding of the ecosystem as they learn from the consequences of their actions (Ellison, 1996). Despite decades of study, uncertainty still exists in our understanding of the eutrophication, and hence, in the consequences of policies and management recommendations. Our proposed method, through probability calculus, provides an explicit expression of the amount of uncertainty in our knowledge.

As a concluding remark, we used a continuous variable Bayesian network model to predict water quality indicators in a non-tidal estuarine ecosystem. The statistical approach proposed here can be further applied in water quality problems other than

eutrophication and in other ecosystems (e.g., lakes). However, this methodology should also work with some modifications for a broad range of aquatic pollution. We encourage researchers to take and implement the proposed continuous variable Bayesian network model in other contexts.

Appendix A

R Code for Chapter 2

```
#########################################################################
##Functions
#########################################################################
mulognorm=function(par){
    ##par[1] is Ex, par[2] is SDx
    log(par[1])-1/2*log(1 + par[2]^2/par[1]^2)
}

sdlognorm=function(par){
    sqrt(log(1 + par[2]^2/par[1]^2))
}

#########################################################################
## data import
#########################################################################
```

```
library(MCMCpack)
farm=read.csv('Swine_CAFO_Neuse_table.csv',header=T)
View(farm)
colnames(farm)
str(farm)
keep=which(farm$Number_of>0)
farm=farm[keep,]
nrow(farm)

table(farm$Regulate_1)

sum(farm$Number_of)
#########################################################################
## Average hog mass for each type of farms
## Function to calculate average hog mass for each hog type
#########################################################################
avgmass=function
(pgest,plact,pboar,pftow,pwtof,pftof,n.sims =1000,CV1 =0.1){

wtgest = rnorm(n.sims,181,(181*CV1))
wtlact = rnorm(n.sims,181,(181*CV1))
wtboar = rnorm(n.sims,181,(181*CV1))
wtftow = rnorm(n.sims,4.5,(4.5*CV1))
wtwtof = rnorm(n.sims,13.6,(13.6*CV1))
wtftof = rnorm(n.sims,61.2,(61.2*CV1))
wtavg = pgest*wtgest + plact*wtlact
        + pboar*wtboar + pwtof*wtwtof
```

```
        + pftow*wtftow + pftof*wtftof
mc.avgmean=mean(wtavg)
mc.avgsd=sd(wtavg)
return(c(mc.avgmean,mc.avgsd))
}



## farrow to finish
pgest = .068
plact = .016
pboar = .004
pftow = .097
pwtof = .254
pftof = .561
avgmass_fartofin=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[1]
sdmass_fartofin=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[2]

## farrow to wean
pgest = 81/220
plact = 19/220
pboar = 5/220
pftow = 115/220
pwtof = 0
pftof = 0
avgmass_ftow=
```

```
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[1]
sdmass_ftow=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[2]

## wean to feed
pgest = 0
plact = 0
pboar = 0
pftow = 0
pwtof = 1
pftof = 0
avgmass_wtof=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[1]
sdmass_wtof=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[2]

## feed to finish
pgest = 0
plact = 0
pboar = 0
pftow = 0
pwtof = 0
pftof = 1
avgmass_feedtofin=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[1]
sdmass_feedtofin=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[2]
```

```
## farrow to feed
pgest = 81/521
plact = 19/521
pboar = 5/521
pftow = 115/521
pwtof = 301/521
pftof = 0
avgmass_fartofeed=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[1]
sdmass_fartofeed=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[2]

## gilts
pgest = 1
plact = 0
pboar = 0
pftow = 0
pwtof = 0
pftof = 0
avgmass_g=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[1]
sdmass_g=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[2]

## boar
pgest = 0
```

```
plact = 0
pboar = 1
pftow = 0
pwtof = 0
pftof = 0
avgmass_b=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[1]
sdmass_b=
    avgmass(pgest,plact,pboar,pftow,pwtof,pftof)[2]



#########################################################################
##Feed intake
#########################################################################
no=c(4392,3727,7611,5784,4221,4373
    ,6332,6095,3386,2680,1249,1485
    ,9507,10248)
mass=c(104.3,88.5,52.3,67,82.7,48.0
    ,59.2,59.7,55.4,104.7,98.5,70.3
    ,38.3,44.7)
intake=c(12.84,12.59,10.99,12.37,11.93
    ,14.41,12.89,13.21,15.44,16.27
    ,16.27,14.47,10.03,11.02)
totmass=mass*no
totintake=no*intake*52
plot(totmass,log(totintake),main='kg feed intake per kg hog'
     ,xlab='Hog mass (kg)', ylab='Feed intake(kg)',pch=19)
```

```
logfit.intake=lm(log(totintake)~totmass)
summary(logfit.intake)
plot(totmass,log(totintake),pch=19
     ,xlab='hog mass',ylab='log(food intake)')
abline(coef(logfit.intake)[1],coef(logfit.intake)[2])

recode.list<-farm$Regulate_1
farm$type <- as.numeric(recode.list)
farm <- farm[farm$type!=9,]

meanmass <- c(avgmass_b,avgmass_fartofeed,avgmass_fartofin
              ,avgmass_ftow,avgmass_feedtofin,avgmass_g
              ,0,avgmass_wtof)
sdmass <- c(sdmass_b,sdmass_fartofeed,sdmass_fartofin
            ,sdmass_ftow,sdmass_feedtofin,sdmass_g,1
            ,sdmass_wtof)

#########################################################################
## N goes to pork produced
#########################################################################
## kg pork produced in NC, NC agricultural census
NCpork=3815438*1000/2.2
## number of pigs in 2007 in NC
NCpig=10134004
## number of pigs in 2007 in Neuse
Neusepig=1894057
## pork produced in the Neuse
```

```
NeusePork=Neusepig/NCpig*NCpork


## nitrogen % in swine
##= protein % in swine * N% in protein (which is 6.25)
##  N% in protein found in book, nutrient requirement of swine, CH. 2
## protein% found in swine nutrition
## 0.95:  change the amount of pork produced in the Neuse to empty e
## pig mass becausprotein % is based on empty body weights.
Protein=rnorm(1000,0.16,sd=0.02)
hist(Protein)
biomassN=NeusePork*0.95*Protein/6.25
hist(biomassN)
#########################################################################
## Total Lagoon Volume, Surface Area, Bottom Area Calculations
#########################################################################
mass=c(225090,230518,216808,216808,298910,216808,436846,460977
    ,651175,651175,651175,651175,384453,512604,434367)
vol=c(16790,17973,11994,18802,31611,15076,57763,41797,47570
    ,53103,62435,45928,51252,45616,22738)


new.mass=c(1000*54,9200*54,4500*61,4360*61,4900*61,2900*61
    ,5280*61,2200*61,5880*61)
new.area=c(1.89,0.54,0.92,1.25,2.68,0.58,1.58,0.58,1.32)
new.depth=c(0.78,1.66,1.77,1.10,1.46,2.17,2.06,1.40,2.27)
new.volume=new.area*new.depth*10000
plot(mass,vol,pch=19,xlim=c(0,700000),ylim=c(0,70000))
points(new.mass,new.volume,col='green',pch=19)
```

```
mass=c(mass,new.mass)
vol=c(vol,new.volume)
plot(mass, vol,xlim=c(0,600000),ylim=c(0,60000))
lagoon.fit=lm(vol[-17]~mass[-17])
plot.lm(lagoon.fit)
loglagoon.fit=lm(log(vol[-17])~mass[-17])
plot.lm(loglagoon.fit,which=4)
summary(loglagoon.fit)
plot(mass[-17], log(vol[-17]),pch=19,xlab='hog mass'
     ,ylab='log(lagoon volume)')
abline(coef(loglagoon.fit)[1],coef(loglagoon.fit)[2])

n.sims <- 100
tot.mass <- numeric()
intake <- numeric()
total.intake <- numeric()
Neuse.totalmass <- numeric()
lagvol <- matrix(NA,nrow=nrow(farm),ncol=n.sims)
for (j in 1:n.sims){
    for (k in 1:nrow(farm)){
        tot.mass[k] <- farm$Number_of[k]*
            rnorm(1,meanmass[farm$type[k]],sdmass[farm$type[k]])
        new=data.frame(totmass=tot.mass[k])
        pred=predict(logfit.intake,new,se.fit=T)
        intake[k] <- exp(
            rnorm(1,mean=pred$fit
```

```
                    ,sd=sqrt(pred$se.fit^2+pred$residual.scale^2)))
        new2=data.frame(mass=tot.mass[k])
        pred2=predict(loglagoon.fit,new2,se.fit=T)
        lagvol[k,j] <- exp(
            rnorm(1,mean=pred2$fit
                  ,sd=sqrt(pred2$se.fit^2+pred2$residual.scale^2)))
    }
    total.intake[j] <- sum(intake)
    Neuse.totalmass[j] <- sum(tot.mass)
    print(j)
}

hist(total.intake)
mean(total.intake)
sd(total.intake)

hist(Neuse.totalmass)
mean(Neuse.totalmass)
sd(Neuse.totalmass)

farm$lagvolmean <- apply(lagvol,1,mean)
farm$lagvolsd <- apply(lagvol,1,sd)
lagvoltotal <- sum(farm$lagvolmean)

depth.hunt <- c(0.78,1.66,1.77,1.10,1.46,2.17,2.06,1.4,2.27)
depth.ave <- (15*3.05 + sum(depth.hunt))/24
```

```
lagSAtotal <-lagvoltotal/depth.ave
#########################################################################
##perNfood
#########################################################################
Nfeed = c(3.14,2.52,3.02,2.95,2.78,3.24,2.56,2.67
    ,3.01,2.15,2.76,3.08,2.79,3.17)
Nfeed = Nfeed/100
hist(Nfeed, main="Distribution of Nitrogen in Food"
     , xlab="Percent Nitrogen in Food")
meanperN = mean(Nfeed)
sd(Nfeed)
n = length(Nfeed)
varperN = (n-1)/n*var(Nfeed)
alpha = meanperN*((meanperN*(1-meanperN)/varperN) - 1)
beta = (1-meanperN)*((meanperN*(1-meanperN)/varperN) - 1)
alpha
beta

N_per_feed=rbeta(1000,alpha,beta)
#########################################################################
#Yearly Feed N Intake
#########################################################################
total.intake <- rnorm(1000,mean(total.intake),sd(total.intake))
Nintake.yr <- total.intake*N_per_feed

mean(Nintake.yr)
sd(Nintake.yr)
```

```
#########################################################################
## Yearly N to Biomass
#########################################################################
N.Biomass <- Nintake.yr*0.3
mean(N.Biomass)
sd(N.Biomass)
#########################################################################
## Yearly N to Waste
#########################################################################
N.Waste <- Nintake.yr*0.7
mean(N.Waste)
sd(N.Waste)
#########################################################################
## N excretion
#########################################################################
Nexcrete=c(0.091,0.219,0.113,0.138,0.153)
mean(Nexcrete)
sd(Nexcrete)

NH3_RTI=1.61/150*2.2/17*14
NH3_doorn=c(0.0351,0.0137,0.0141)/17*14
mean(NH3_doorn)
sd(NH3_doorn)
#########################################################################
## Confinement housing Yearly Ammonia Emission Rate from
## (Aneja et al. 2008b)
#########################################################################
```

```
NH3_spr=c(0.31,0.23,0.21,0.39)
NH3_sum=c(0.58,0.23,1.07)
NH3_aut=c(0.07,0.33,0.35)
NH3_win=c(0.12,0.81)
NH3=numeric()
for (i in 1:1000){
NH3[i]=mean(c(sample (NH3_spr,1),sample (NH3_sum,1),
  sample (NH3_aut,1),sample (NH3_win,1)))/1000*52}

mean(NH3)
sd(NH3)
#########################################################################
## Confinement Housing Ammonia Emissions
#########################################################################
NH3.1=c(0.34,0.49,0.57,1.29,0.98,1.15,0.16,0.008,0.12,0.52,
    0.07,0.75)/1000*52
mean(NH3.1)
sd(NH3.1)

NH3.2 <- c(rep(1.81,70),rep(1.21,87),rep(1.81,15),rep(2.98,13)
           ,rep(2.98,331))/150*2.2/17*14
mean(NH3.2)
sd(NH3.2)

NH3_doorn=0.059/17*14
0.01/17*14
```

```
mu <- c(0.03,0.055,0.028,0.049)
sd <- c(0.0087,0.020,0.002,0.008)

trans.mu <- numeric()
trans.sd <- numeric()
for (i in 1:4){
trans.mu[i] <- mulognorm(c(mu[i],sd[i]))
trans.sd[i] <- sdlognorm(c(mu[i],sd[i]))
}

NH3.studies <- rlnorm(1000,trans.mu,trans.sd)
mean(NH3.studies)
sd(NH3.studies)
hist(NH3.studies)

totalhogmass <- rnorm(1000,mean(Neuse.totalmass)
                      ,sd(Neuse.totalmass))
confine.vol <- NH3.studies*totalhogmass
hist(confine.vol, xlab="Confinement Housing Volatilization"
     , main="Distribution of Confinement Housing Volatilization ")
mean(confine.vol)
sd(confine.vol)
#########################################################################
## Lagoon Yearly Ammonia Emission Rate  (Aneja et al. 2008a)
## avg.rt: ug N-NH3/m2 /min
## NH3vol: kg N-NH3/m2/year
#########################################################################
```

```
error_rate <- 0.227
sum.rt <- rlnorm(1000,0.117*30+4.474,error_rate)
fall.rt <- rlnorm(1000,0.117*11.6+4.474,error_rate)
win.rt <- rlnorm(1000,0.117*12.1+4.474,error_rate)
spr.rt <- rlnorm(1000,0.117*24.7+4.474,error_rate)
avg.rt <- (sum.rt + fall.rt + win.rt + spr.rt)/4

NH3vol <- avg.rt*60*24*365/1e9
mean(NH3vol)
sd(NH3vol)

## change to hog based emission rate
hogPerm2=
    ((7611*52.3+5784*67)/2/17150+(4392*104.3+3727*88.5)/2/15170)/2
mean(NH3vol/hogPerm2)
sd(NH3vol/hogPerm2)
#########################################################################
## Average Emission Rate from Lagoon
#########################################################################
lagNH3.2 <- c(0.017,0.010)
lagNH3.3 <- c(0.029,0.004)
lagNH3.4 <- c(0.021,0.006)

lagNH3<-rbind(lagNH3.2,lagNH3.3,lagNH3.4)

CV <- lagNH3[,2]/lagNH3[,1]
CV
```

```
mean(CV)

lagNH3.1 <- c(0.019,0.019*mean(CV))

lagNH3 <- rbind(lagNH3,lagNH3.1)

mu <- lagNH3[,1]
sd <- lagNH3[,2]

trans.mu <- numeric()
trans.sd <- numeric()
for (i in 1:4){
trans.mu[i] <- mulognorm(c(mu[i],sd[i]))
trans.sd[i] <- sdlognorm(c(mu[i],sd[i]))
}

NH3lag.studies <- rlnorm(1000,trans.mu,trans.sd)
hist(NH3lag.studies)
mean(NH3lag.studies)
sd(NH3lag.studies)

NH3lagvol <- NH3lag.studies*totalhogmass
mean(NH3lagvol)
sd(NH3lagvol)
#########################################################################
## Lagoon sludge N concentration mg/L
#########################################################################
```

```
par=matrix(NA, nrow=15,ncol=2)
par[,1]=musludge = c(2.8,3.3,3.2,3.4,2.4,1.2,6.2,4.2,4.5,4.2,3.2
            ,3.3,5.8,5.7,4.2)*1000
par[,2]=sdsludge =c(0.3,0.3,0.5,0.3,0.8,0.4,2.5,0.6,0.6,0.8,0.6
            ,0.9,3.5,0.6,1.1)*
  1000/1.96/sqrt(10) #var of mean = var/n
logpar=matrix(NA, nrow=15,ncol=2)
for (i in 1: 15){
logpar[i,1]=mulognorm(par[i,])
logpar[i,2]=sdlognorm(par[i,])
}

Nsludge=numeric()
for ( i in 1:1000){
Nsludge[i]=mean(rlnorm(15, logpar[,1],logpar[,2]))}
hist(Nsludge)
mean(Nsludge)
sd(Nsludge)
Nsludgeavg <- rlnorm(375,mulognorm(c(5000,700))
                     ,sdlognorm(c(5000,700)))
Nsludgeavg <-c(Nsludgeavg,
               rlnorm(625,mulognorm(c(3837,54))
,sdlognorm(c(3837,54))))
hist(Nsludgeavg)
mean(Nsludgeavg)
sd(Nsludgeavg)
#########################################################################
```

```
##Lagoon Sludge Accumulation
##(Chastain 2006 + Bicudo et al. 1999) m3/kg hog /year
#########################################################################
par=matrix(NA, nrow=6,ncol=2)
par[,1]=u_sludge = c(0.00783, 0.003,0.00353,0.00343,0.002,0.0045)
par[,2]=sd_sludge = c(0.00247,0.00054,0.00078,0.00092,0.0015,0.0015)

logpar=matrix(NA, nrow=6,ncol=2)
for (i in 1: 6){
logpar[i,1]=mulognorm(par[i,])
logpar[i,2]=sdlognorm(par[i,])
}

sludgeacc=numeric()
for ( i in 1:1000){
sludgeacc[i]=sum(c(rlnorm(6, logpar[,1],logpar[,2])
             ,0.00492,0.0022))/8}
hist(sludgeacc)
mean(sludgeacc)
sd(sludgeacc)

Nfluxsludge <- sludgeacc*Nsludgeavg*totalhogmass/1000
hist(Nfluxsludge)
mean(Nfluxsludge)
sd(Nfluxsludge)
#########################################################################
## Lagoon Seepage Rate
```

```
## Data from Ham (2002) need to check standard deviation
## mm/d
#########################################################################
seepagert =c(0.6,0.8,0.8,0.8,0.8,0.9,0.9,1,1.3,1.3,1.4,1.5,1.7,2)
par=c(mean(seepagert),sd(seepagert))
hist(seepagert,freq=F)

mulognorm(par)
sdlognorm(par)
seeprt=rlnorm(1000,mulognorm(par),sdlognorm(par))
mean(seeprt)
sd(seeprt)
#########################################################################
## Lagoon Yearly Seepage Loss *assuming* new liner
## Lagoon seepage export  Ham and DeSutter 2000, kg / m2 /year
#########################################################################

seepageexp=c(0.522,0.500,0.451,0.289,0.229,0.218,0.131)
hist(seepageexp)
mean(seepageexp)
sd(seepageexp)

seepageexp <- runif(1000,0.1,0.6)



#########################################################################
## Lagoon liquid NH3 concentration
```

```
#########################################################################

par=matrix(NA, nrow=15,ncol=2)
par[,1]=NH3_mu=c(350,350,350,310,340,350,290,330,440,470,410,490
            ,280,500,570)
par[,2]=NH3_sd=c(30,30,30,30,20,30,20,30,30,40,30,40,40,50,60)
/1.96/sqrt(10)
logpar=matrix(NA, nrow=15,ncol=2)
for (i in 1: 15){
logpar[i,1]=mulognorm(par[i,])
logpar[i,2]=sdlognorm(par[i,])
}

NH3_con=numeric()
for ( i in 1:1000){
NH3_con[i]=mean(c(rlnorm(15, logpar[,1],logpar[,2])))}
mean(NH3_con)
sd(NH3_con)

NH3lag=numeric()
for ( i in 1:1000){
NH3lag[i]=mean(rlnorm(15, logpar[,1],logpar[,2]))}
hist(NH3lag)
mean(NH3lag)
sd(NH3lag)
NH3lagavg <- rlnorm(375,mulognorm(c(349,136)),sdlognorm(c(349,136)))
NH3lagavg <-c(NH3lagavg,rlnorm(625,mulognorm(c(389,1.4))
```

```
                                ,sdlognorm(c(389,1.4))))
hist(NH3lagavg)
mean(NH3lagavg)
sd(NH3lagavg)

export2 <- NH3lagavg*seeprt*365/1e6
hist(export2)
mean(export2)
sd(export2)

seep.export.est <- rlnorm(500,mulognorm(c(mean(export2),sd(export2)))
                           ,sdlognorm(c(mean(export2),sd(export2))))
seep.export.est <-c(seep.export.est,runif(500,0.1,0.6))
hist(seep.export.est)
mean(seep.export.est)
sd(seep.export.est)

totalexport <- seep.export.est*lagSAtotal
hist(totalexport)
mean(totalexport)
sd(totalexport)
#########################################################################
## Lagoon TKN
#########################################################################
par=matrix(NA, nrow=15,ncol=2)
par[,1]=TKN_mu=c(410,420,420,370,390,400,350,390,490,520,450,560,
            340,570,650)
```

```
par[,2]=TKN_sd=c(30,40,40,30,30,30,20,40,30,40,30,60,30,50,70)
/1.96/sqrt(10)
logpar=matrix(NA, nrow=15,ncol=2)
for (i in 1: 15){
logpar[i,1]=mulognorm(par[i,])
logpar[i,2]=sdlognorm(par[i,])
}

TKN_con=numeric()
for ( i in 1:1000){
TKN_con[i]=mean(c(rlnorm(15, logpar[,1],logpar[,2])))}
mean(TKN_con)
sd(TKN_con)

TKNlag=numeric()
for ( i in 1:1000){
TKNlag[i]=mean(rlnorm(15, logpar[,1],logpar[,2]))}
hist(TKNlag)
mean(TKNlag)
sd(TKNlag)
TKNlagavg <- rlnorm(375,mulognorm(c(416,157)),sdlognorm(c(416,157)))
TKNlagavg <-c(NH3lagavg,rlnorm(625,mulognorm(c(449,1.7))
                               ,sdlognorm(c(449,1.7))))
hist(TKNlagavg)
mean(TKNlagavg)
sd(TKNlagavg)
```

```
lagoonNpool <- TKNlagavg/1000*lagvoltotal
hist(lagoonNpool)
mean(lagoonNpool)
sd(lagoonNpool)
#########################################################################
## Total Sprayfield Area
## 2007 Ag. census & NC animal waste operation certification program
## total lagoon liquid
#########################################################################
chicken_lagoon=(0.885*7.3+0.04*22.3+0.075*25.2)*6484314
cow_lagoon=(0.43*1946+0.51*6570+0.06*9490)*49456
hog_lagoon=(0.16*191+0.7*927+0.13*3203+0.03*3861+0.02*10481)*1894057
perhog=hog_lagoon/(hog_lagoon+chicken_lagoon+cow_lagoon)
totalsprayarea1=37767*perhog*4047
## management plan for Boknam's farm
totalsprayarea2=121.4/(88.9*4800)*9.96e7*4047
totalsprayarea=(totalsprayarea2+totalsprayarea1)/2

## old values
## break.val <- c(0,seq(750,21750,by=1500))
## spray.hist <- hist(farm$Number_of,breaks=break.val)
## spray.hist$breaks



##totalsprayarea<-
##(57*5+117*10+149*20+104*30+50*40
```

```
## +28*50+7*60+7*70+8*80+5*90+100)*4047
##totalsprayarea

## new values
##break.val <- c(0,3750,6250,21000)
##spray.hist <- hist(farm$Number_of,breaks=break.val)
##spray.hist$breaks
##spray.hist$counts
##totalsprayarea<- (323*50/3+143*100/3+67*260/3)*4047
##totalsprayarea
#########################################################################
##Land Application Rate
#########################################################################
corn.app <- c(104,141,70,67,71,130,138,128,125,137,148,138,98,130
              ,82,125,78,124,135,135,80,68,133,131,128,128,137,134
              ,136,150,128,135,130,81,80,139,90,119)/2.2/4046
mean(corn.app)
sd(corn.app)

soy.app <- c(119,191,97,86,79,155,192,175,172,156,176,173,124,174
             ,134,164,101,174,178,178,102,109,173,164,161,161,176
             ,172,161,195,156,170,170,97,108,163,109,135)/2.2/4046
mean(soy.app)
sd(soy.app)

bg.app <-c(196,216,165,123,180,108,210,198,194,207,230,126,155,161
           ,95,230,99,151,172,172,115,166,168,230,225,225,230,225
```

```
           ,95,165,154,151,108,108,161,165,192,151)/2.2/4046
mean(bg.app)
sd(bg.app)

hist(corn.app)
hist(soy.app)
hist(bg.app)

app.rt <- c(sample(corn.app,400,replace=TRUE)
            ,sample(soy.app,200,replace=TRUE)
            ,sample(bg.app,400,replace=TRUE))

hist(app.rt)
mean(app.rt)
sd(app.rt)
app.rt <- rnorm(1000,mean(app.rt),sd(app.rt))

landapp <- app.rt*totalsprayarea*2
hist(landapp)
mean(landapp)
sd(landapp)
#########################################################################
## Sprayfield Ammonia emission factors RTI(2003)
#########################################################################
sprRTI <- c(1.96,7.27,2.83,3.26,3.84)/150*2.2*14/17
mean(sprRTI)
sd(sprRTI)
```

```
mean(sprRTI[-2])
sd(sprRTI[-2])
#########################################################################
## Sprayfield % of several nitrogen sinks in the sprayfield.
##Summary of Data in Whalen & DeBerardinis
#########################################################################
sf.denit <- c(2,1,1,1,1,1)
sf.plantup <- c(25,46,73,68,117,52)
sf.micimm <- c(22,14,13,9,14,8)
sf.leach <- c(14,15,19,35,6,9)
sf.vol <- c(5,6,17,17,13,14)
sf.soilsto <- c(50,30,10,10,14,9)
sf.tot <- sf.denit+sf.plantup+sf.micimm+sf.leach+sf.vol+sf.soilsto
sf.tot
mean(sf.denit)
sd(sf.denit)
mean(sf.plantup)
sd(sf.plantup)
mean(sf.micimm)
sd(sf.micimm)
mean(sf.leach)
sd(sf.leach)
mean(sf.soilsto)
sd(sf.soilsto)
mean(sf.tot)
sd(sf.tot)
19/128*102
```

```
#########################################################################
##Application rates in the sprayfield Summary in Read et al.
#########################################################################

##lowapp <- c(28.,33.3,37.0,34.3,25.8,35,27.5)
##highapp <- c(29.5,35.5,3,30,46,32,31.8)
##mean(lowapp)
##sd(lowapp)
##mean(highapp)
##sd(highapp)

corn.app <- c(104,141,70,67,71,130,138,128,125,137,148,138,98,130,82
              ,125,78,124,135,135,80,68,133,131,128,128,137,134,136
              ,150,128,135,130,81,80,139,90,119)/2.2/4046
mean(corn.app)
sd(corn.app)

soy.app <- c(119,191,97,86,79,155,192,175,172,156,176,173,124,174,134
             ,164,101,174,178,178,102,109,173,164,161,161,176,172,161
             ,195,156,170,170,97,108,163,109,135)/2.2/4046
mean(soy.app)
sd(soy.app)

bg.app <-c(196,216,165,123,180,108,210,198,194,207,230,126,155,161,95
           ,230,99,151,172,172,115,166,168,230,225,225,230,225,95,165
           ,154,151,108,108,161,165,192,151)/2.2/4046
```

```
mean(bg.app)
sd(bg.app)

#########################################################################
##Background Denitrification Sprayfield
#########################################################################

##back.denit <- 210*24*365/1e9
##back.denit

### check the percentage with other data NH3 volatilization
mean(0.163*landapp/totalhogmass)
sd(0.163*landapp/totalhogmass)
windows()
hist(0.163*landapp/totalhogmass)
points(density(rlnorm(1000,mulognorm(c(0.035,0.010))
                      ,sdlognorm(c(0.035,0.010)))),type='l')
hist(rlnorm(1000,
            mulognorm(c(0.035,0.010)),sdlognorm(c(0.035,0.010))))
windows()
hist( rlnorm(1000,mulognorm(c(0.012,0.01/0.035*0.012))
             ,sdlognorm(c(0.012,0.010/0.035*0.012))))


margmean <- c(0.163,0.635,0.135,0.205,0.010,0.133)
margmean <- margmean/sum(margmean)
##margsd<- c(0.10,0.31,0.05,0.17,.005,0.05)
```

```
##margsd <- margsd/sum(margmean)

betapars <- function(m,s,n){
v <- (n-1)/n*s^2
alpha = m*((m*(1-m)/v) - 1)
beta = (1-m)*((m*(1-m)/v) - 1)
return(c(alpha,beta))
}

pars <- numeric()
for (i in 1:6){
alphbet <- betapars(margmean[i],margsd[i],6)
pars <- rbind(pars,alphbet)
}


vol <- landapp*rbeta(1000,pars[1,1],pars[1,2])
uptake <- landapp*rbeta(1000,pars[2,1],pars[2,2])
leach <- landapp*rbeta(1000,pars[3,1],pars[3,2])
soil <- landapp*rbeta(1000,pars[4,1],pars[4,2])
denitr <- landapp*rbeta(1000,pars[5,1],pars[5,2])
microb <-landapp*rbeta(1000,pars[6,1],pars[6,2])

mean(vol)
sd(vol)

mean(uptake)
```

```
sd(uptake)

mean(leach)
sd(leach)

mean(soil)
sd(soil)

mean(denitr)
sd(denitr)

mean(microb)
sd(microb)
#########################################################################
##Lagoon Denitrifcation by Difference
#########################################################################

lagdenit <- N.Waste-confine.vol-Nfluxsludge
-totalexport-NH3lagvol-landapp

hist(lagdenit)
mean(lagdenit)
sd(lagdenit)
#########################################################################
## N export from the Neuse River Basin
## Total area of Neuse River basin 1458936 hectare
#########################################################################
```

```
N_neuse=1458936*4.7
N_neuse

#########################################################################
## potential N lost to water lfrom CAFO
#########################################################################
mean(leach)/N_neuse
sd(leach)/N_neuse
mean(totalexport/N_neuse)
sd(totalexport/N_neuse)
#########################################################################
## atm deposition
#########################################################################
N_neuse=(3.26+4.89+3.62+5.91)/4*1458936 ## kg/year
N_neuse
mean(NH3lagvol+confine.vol+vol)/N_neuse
mean(NH3lagvol+confine.vol+vol)
sd(NH3lagvol+confine.vol+vol)

#########################################################################
##percentages
#########################################################################
mean(confine.vol/Nintake.yr)
sd(confine.vol/Nintake.yr)
mean(N.Biomass/Nintake.yr)
sd(N.Biomass/Nintake.yr)
```

```
mean(totalexport/Nintake.yr)
sd(totalexport/Nintake.yr)
mean(NH3lagvol/Nintake.yr)
sd(NH3lagvol/Nintake.yr)
mean(lagdenit/Nintake.yr)
sd(lagdenit/Nintake.yr)
mean(Nfluxsludge/Nintake.yr)
sd(Nfluxsludge/Nintake.yr)


mean(uptake/Nintake.yr)
sd(uptake/Nintake.yr)
mean(soil/Nintake.yr)
sd(soil/Nintake.yr)
mean(leach/Nintake.yr)
sd(leach/Nintake.yr)

mean(vol/Nintake.yr)
sd(vol/Nintake.yr)
mean(microb/Nintake.yr)
sd(microb/Nintake.yr)
```

Appendix B

Supplementary Material for Chapter 3

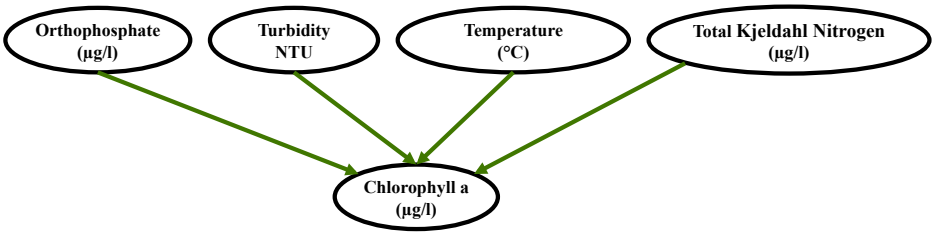


FIGURE B.1: Structure of the New River Estuary base model.

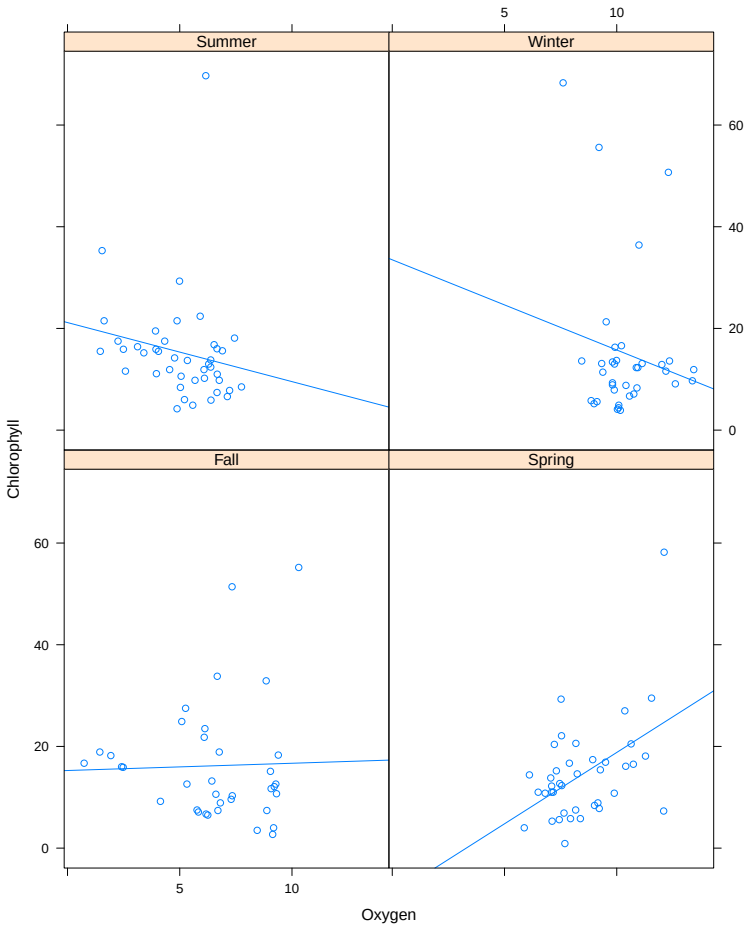


FIGURE B.2: Bivariate scatter plot chlorophyll a versus bottom dissolved oxygen with respect to seasons.

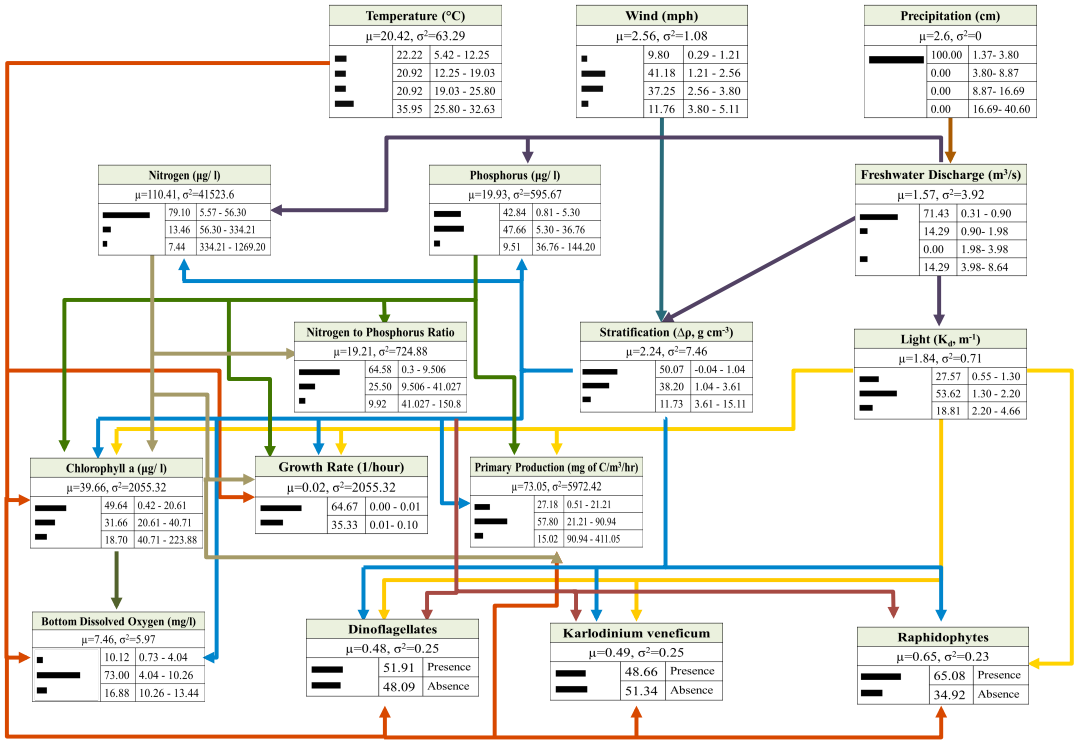


FIGURE B.3: A low precipitation scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

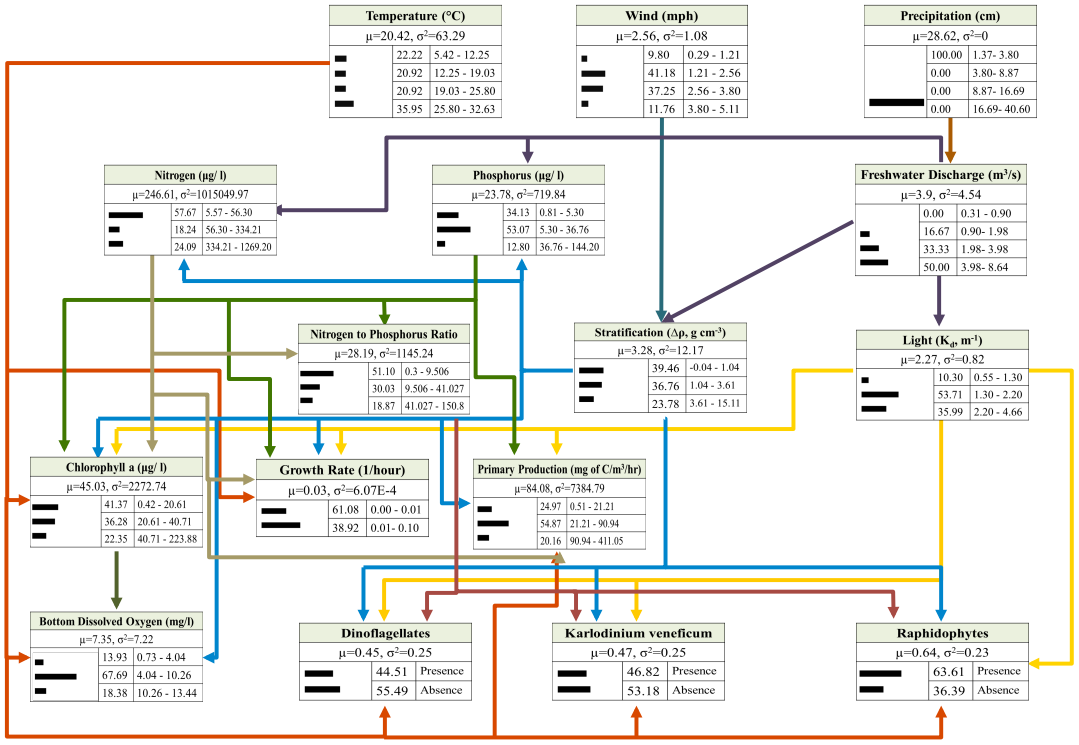


FIGURE B.4: A high precipitation scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

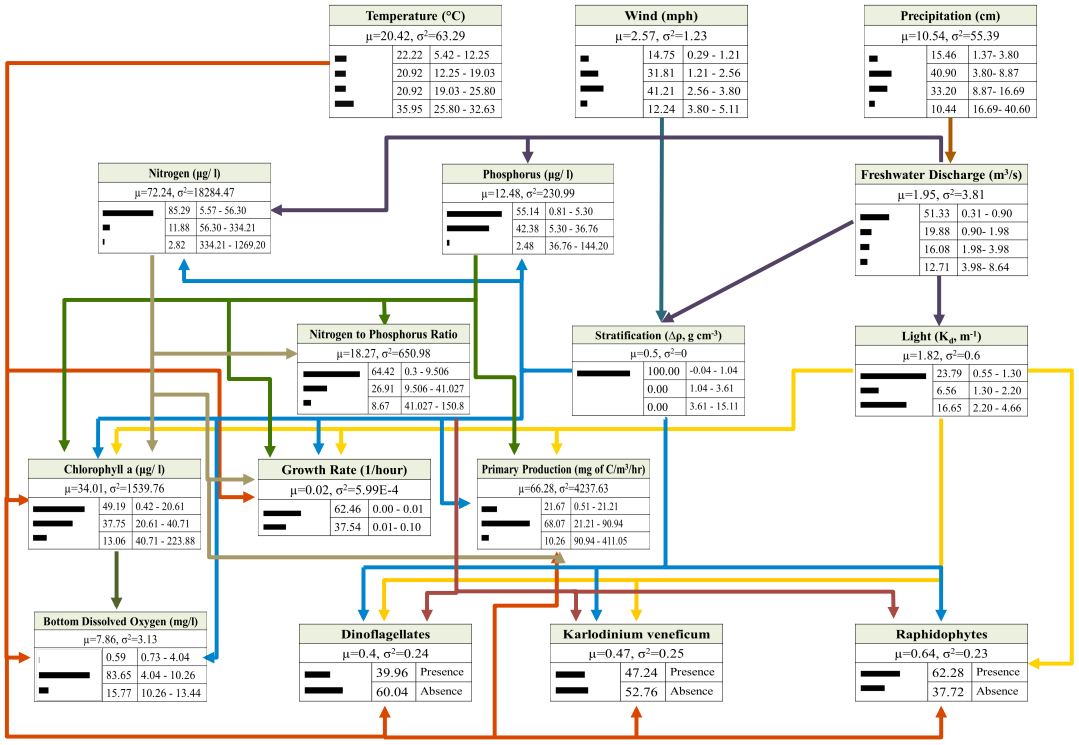


FIGURE B.5: A mixed water column scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

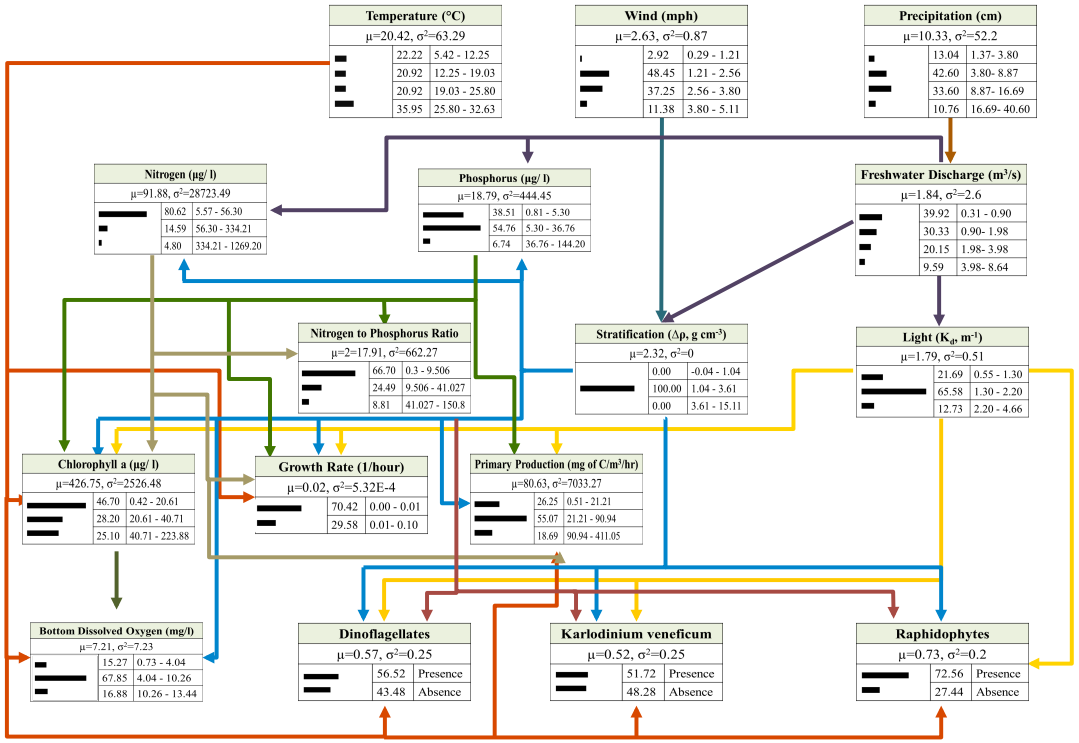


FIGURE B.6: A partially-mixed water column scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

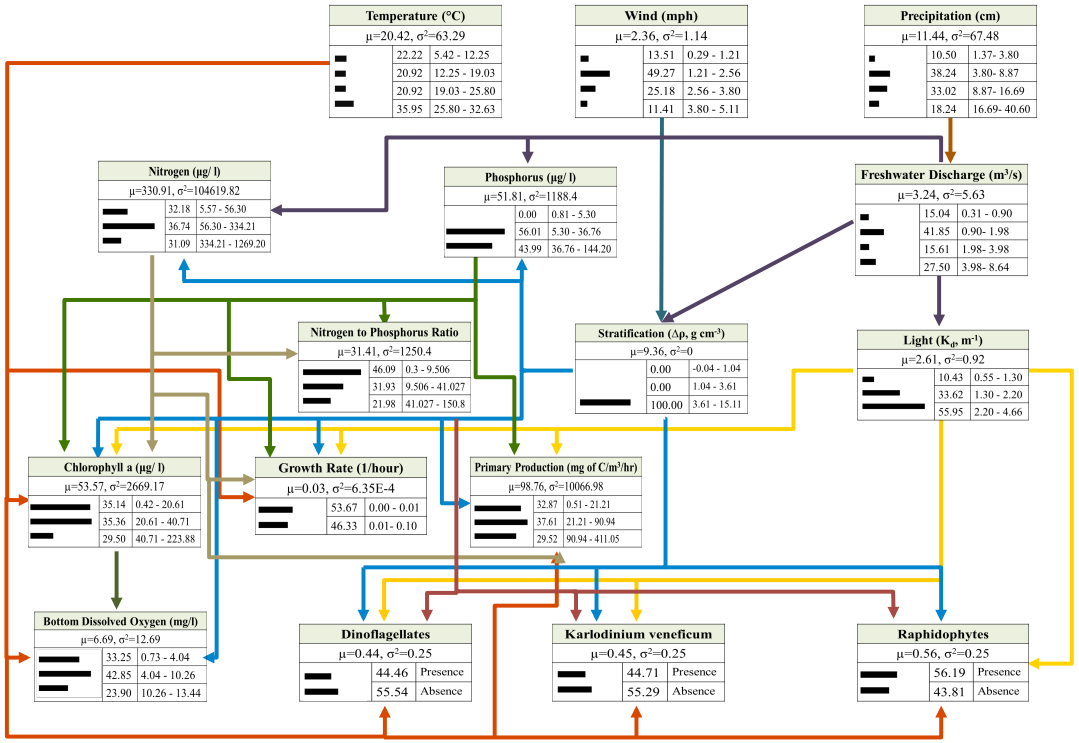


FIGURE B.7: A stratified water column scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

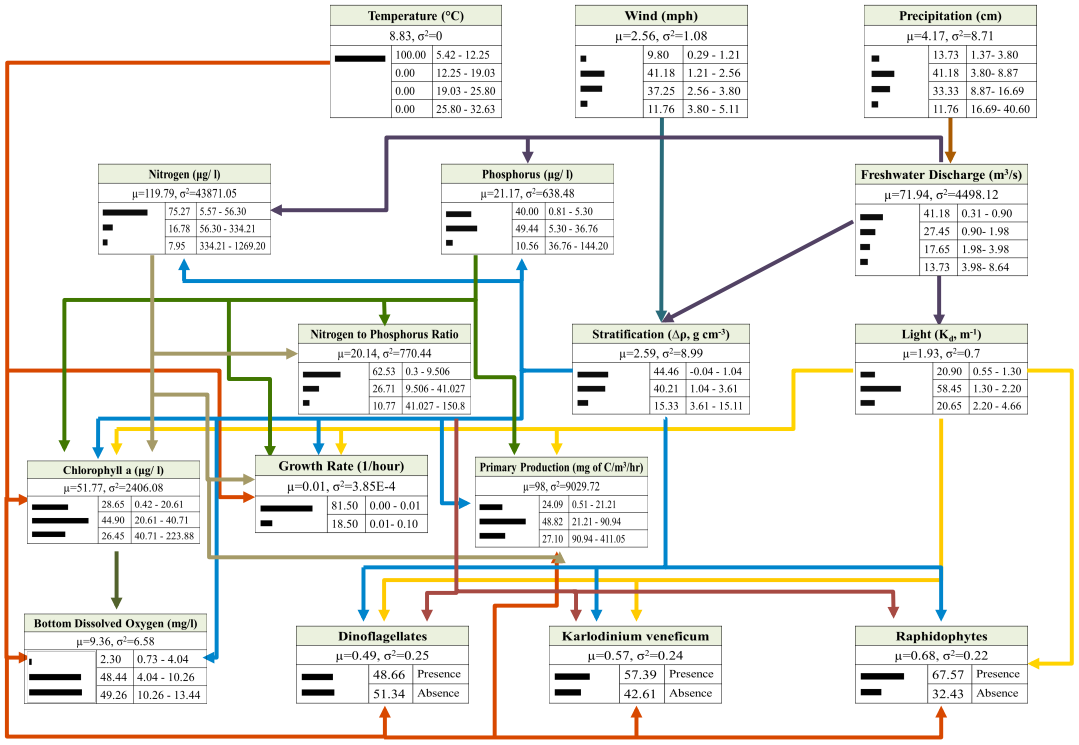


FIGURE B.8: A low temperature scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

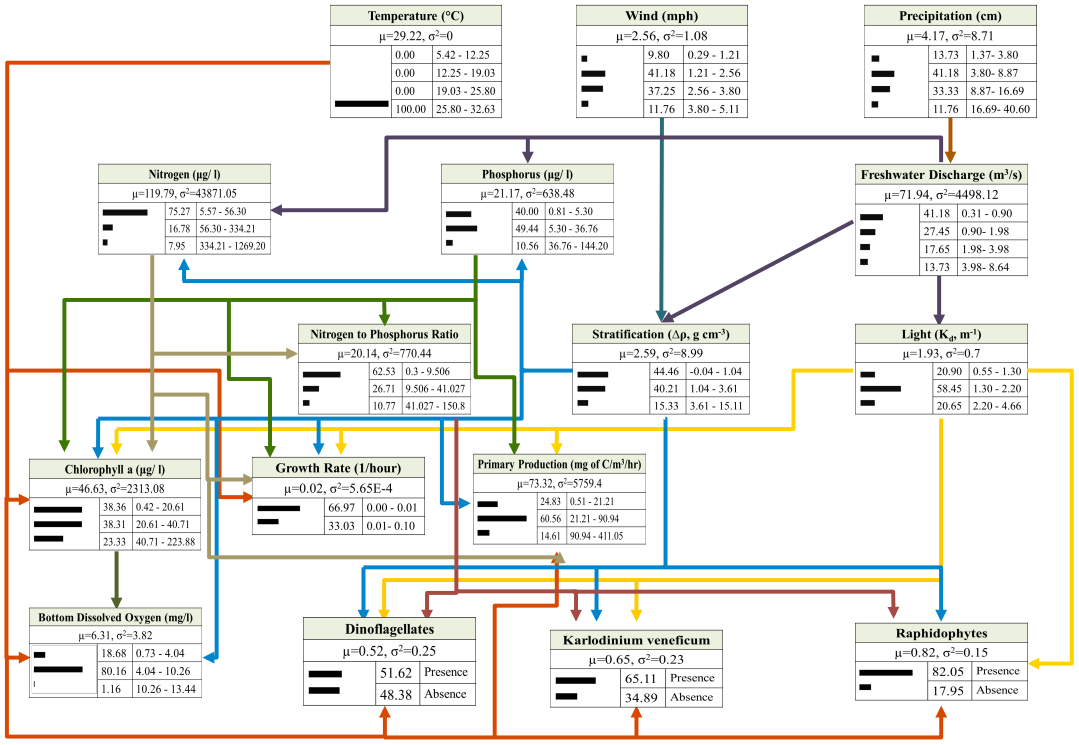


FIGURE B.9: A high temperature scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

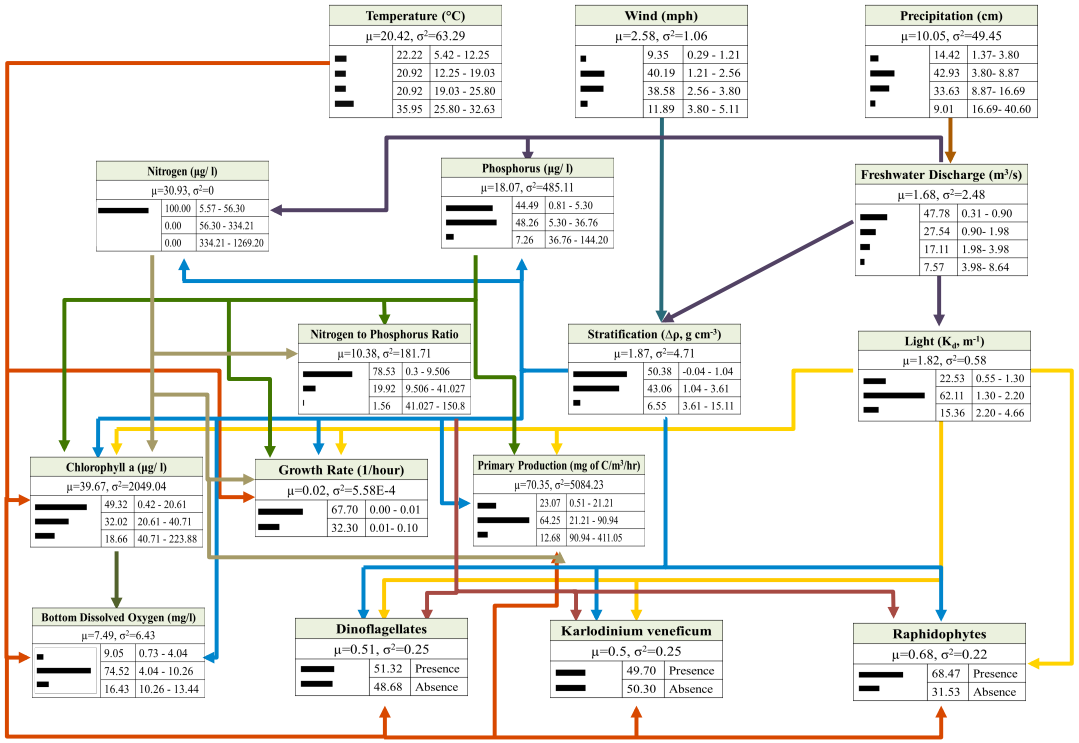


FIGURE B.10: A low nitrogen scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

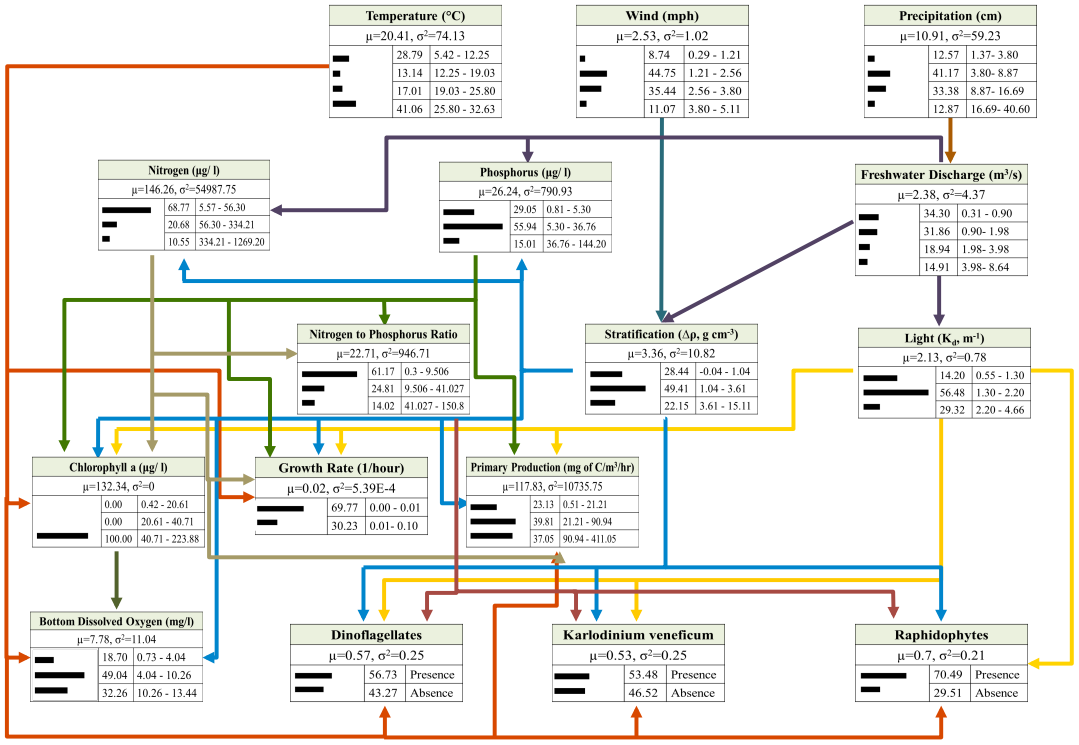


FIGURE B.11: A chlorophyll a water criteria violation scenario for the NRE. Blue bars represent probability of each defined state; red bars are evidence for a dominant state for scenarios of interest; the numbers next to bars are probability of the variable being in that state, and the numbers represented as intervals to the right of the probabilities are the interval defined for the status of the variables of interest.

Table B.1: The station information for data downloaded on precipitation.

State Climate Office of North Carolina **North Carolina State University**	
Data Base	CRONOS
Station ID	KNCA
Station Type	AWOS
Station Name	New River MCAS
City, State	Jacksonville, NC
County	Onslow County
Latitude	34.7073361
Longitude	-77.4451639

Table B.2: List of investigated physical, chemical, and biological variables

Parameter
Wind
Freshwater Discharge
Precipitation
Stratification (Density gradient)
Stratification (Density ratio)
Station name
Season
Temperature
Salinity
Dissolved oxygen
pH
Turbidity
Chlorophyll a
Secchi depth
Diffuse light attenuation coefficient
Total suspended solids
Particulate organic carbon
Particulate nitrogen
Carbon to nitrogen molar ratio
Nitrate/Nitrite
Ammonium
Dissolved inorganic nitrogen
Dissolved organic nitrogen
Orthophosphate
Total dissolved phosphorus
Nitrogen to phosphorus molar ratio
Silica
Primary productivity
Growth rate Peridinin
Fucoxanthin
19'-Hexanoyloxyfucoxanthin
Violaxanthin
Gyroxanthin

Table B.3: Conditional probability table for the temperature node with states 5.42-63.29 (Low), 12.25-19.03 (Medium), 19.03-25.80 (Medium High) and 25.80-32.63 (High).

	Interval	**Probability**
Temperature	5.42-63.29	0.2222
	12.25-19.03	0.2092
	19.03-25.80	0.2092
	25.80-32.63	0.3595
	Experience	153

Table B.4: Conditional probability table for the wind node with states 0.29-1.21 (Low), 1.21-2.56 (Medium), 2.56-3.80 (Medium High) and 3.80-5.11 (High).

	Interval	**Probability**
Wind	0.29-1.21	0.0980
	1.21-2.56	0.4118
	2.56-3.80	0.3725
	3.80-5.11	0.1176
	Experience	153

Table B.5: Conditional probability table for the precipitation node with states 1.37-3.80 (Low), 3.80-8.87 (Medium), 8.87-16.69 (Medium High) and 16.69-40.60 (High).

	Interval	**Probability**
Precipitation	1.37-3.80	0.1373
	3.80-8.87	0.4118
	8.87-16.69	0.3333
	16.69-40.60	0.1176
	Experience	153

Table B.6: Conditional probability table for the freshwater discharge node with states 0.31-0.90 (Low), 0.90-1.98 (Medium), 1.98-3.98 (Medium High), and 3.98-8.64 (High).

	Precipitation	1.37-3.80	3.80-8.87	8.87-16.69	16.69-40.60
Freshwater Discharge	0.31-0.90	0.7143	0.4286	0.4118	0.0000
	0.90-1.98	0.1429	0.3333	0.2941	0.1667
	1.98-3.98	0.0000	0.1905	0.1765	0.3333
	3.98-8.64	0.1429	0.0476	0.1176	0.5000
	Experience	21	63	51	18

Table B.7: Conditional probability table for the stratification node with states -0.04-1.04 (Stratified), 1.04-3.61 (Partially-Mixed), 3.61-15.11 (Mixed).

	Freshwater Discharge			0.31-0.90	
	Wind	0.29-1.21	1.21-2.56	2.56-3.80	3.80-5.11
Stratification	-0.04-1.04	1.0000	0.4848	0.5833	0.3333
	1.04-3.61	0.0000	0.4545	0.3333	0.6667
	3.61-15.11	0.0000	0.0606	0.8333	0.0000
	Experience	3	33	24	3
	Freshwater Discharge			0.90-1.98	
	Wind	0.29-1.21	1.21-2.56	2.56-3.80	3.80-5.11
Stratification	-0.04-1.04	0.5555	0.1667	0.4286	0.3333
	1.04-3.61	0.2222	0.5	0.4762	0.3333
	3.61-15.11	0.2222	0.3333	0.0952	0.3333
	Experience	9	12	21	0
	Freshwater Discharge			1.98-3.98	
	Wind	0.29-1.21	1.21-2.56	2.56-3.80	3.80-5.11
Stratification	-0.04-1.04	0.3333	0.1667	0.5000	1.0000
	1.04-3.61	0.3333	0.5833	0.5000	0.0000
	3.61-15.11	0.3333	0.2500	0.0000	0.0000
	Experience	0	12	12	3
	Freshwater Discharge			3.98-8.64	
	Wind	0.29-1.21	1.21-2.56	2.56-3.80	3.80-5.11
Stratification	-0.04-1.04	0.3333	0.5000	0.3333	0.4167
	1.04-3.61	0.0000	0.3333	0.3333	0.1667
	3.61-15.11	0.6667	0.1667	0.3333	0.4167
	Experience	3	6	0	12

Table B.8: Conditional probability table for the light node.

	Freshwater Discharge		0.31-0.90	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
	0.55-1.30	0.3823	0.3600	0.0000
	1.30-2.20	0.5000	0.6000	0.5000
Light	2.20-4.66	0.1176	0.0400	0.5000
	Experience	34	25	4
	Freshwater Discharge		0.90-1.98	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
	0.55-1.30	0.0625	0.1111	0.1250
	1.30-2.20	0.6875	0.8333	0.3750
Light	2.20-4.66	0.2500	0.5556	0.5000
	Experience	16	18	8
	Freshwater Discharge		1.98-3.98	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
	0.55-1.30	0.1818	0.0769	0.3333
	1.30-2.20	0.7273	0.6923	0.6667
Light	2.20-4.66	0.0909	0.2308	0.0000
	Experience	11	13	3
	Freshwater Discharge		3.98-8.64	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
	0.55-1.30	0.0000	0.2500	0.0000
	1.30-2.20	0.7500	0.2500	1.0000
Light	2.20-4.66	0.2500	0.5000	1.0000
	Experience	9	4	8

Table B.9: Conditional probability table for the nitrogen node.

	Freshwater Discharge		0.31-0.90	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
Nitrogen	5.57-56.30	0.8529	0.9200	0.7500
	56.30-334.21	0.1471	0.0800	0.2500
	334.21-1269.20	0.0000	0.0000	0.0000
	Experience	34	25	4
	Freshwater Discharge		0.90-1.98	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
Nitrogen	5.57-56.30	1.0000	0.7778	0.3750
	56.30-334.21	0.0000	0.2222	0.3750
	334.21-1269.20	0.0000	0.0000	0.2500
	Experience	16	18	8
	Freshwater Discharge		1.98-3.98	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
Nitrogen	5.57-56.30	0.8182	0.7692	0.333
	56.30-334.21	0.1818	0.2308	0.6667
	334.21-1269.20	0.0000	0.0000	0.0000
	Experience	11	13	3
	Freshwater Discharge		3.98-8.64	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
Nitrogen	5.57-56.30	0.6667	0.5000	0.0000
	56.30-334.21	0.1111	0.0000	0.2500
	334.21-1269.20	0.2222	0.5000	0.7500
	Experience	9	4	8

Table B.10: Conditional probability table for the phosphorus node.

	Freshwater Discharge		0.31-0.90	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
Phosphorus	0.81-5.30	0.5942	0.4000	0.0000
	5.30-36.76	0.4058	0.5340	0.5000
	36.76-144.20			
	Experience	34	25	4
	Freshwater Discharge		0.90-1.98	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
Phosphorus	0.81-5.30	0.5311	0.2782	0.0000
	5.30-36.76	0.3439	0.6655	0.3750
	36.76-144.20	0.1250	0.0563	0.6250
	Experience	16	18	8
	Freshwater Discharge		1.98-3.98	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
Phosphorus	0.81-5.30	0.6121	0.4615	0.0000
	5.30-36.76	0.3879	0.5385	1.000
	36.76-144.20	0.0000	0.0000	0.0000
	Experience	11	13	3
	Freshwater Discharge		3.98-8.64	
	Stratification	-0.04-1.04	1.04-3.61	3.61-15.11
Phosphorus	0.81-5.30	0.3333	0.5000	0.0000
	5.30-36.76	0.6667	0.2500	0.6250
	36.76-144.20	0.0000	0.2500	0.3750
	Experience	9	4	8

Table B.11: Conditional probability table for the nitrogen to phosphorus ratio node.

	Nitrogen		5.57-56.30	
	Phosphorus	0.81-5.30	5.30-36.76	36.76-144.20
N:P	0.30-9.51	0.6059	0.9397	0.8575
	9.51-41.03	0.3941	0.0494	0.0000
	41.03-150.80	0.0000	0.0109	0.1424
	Experience	54.4415	55.8958	5.6626
	Nitrogen		56.30-334.21	
	Phosphorus	0.81-5.30	5.30-36.76	36.76-144.20
N:P	0.30-9.51	0.0000	0.0000	1.0000
	9.51-41.03	0.2857	0.7692	0.0000
	41.03-150.80	0.7143	0.2308	0.0000
	Experience	7	13	5
	Nitrogen		334.21-1269.20	
	Phosphorus	0.81-5.30	5.30-36.76	36.76-144.20
N:P	0.30-9.51	0.3333	0.0000	0.2000
	9.51-41.03	0.3333	0.1429	0.8000
	41.03-150.80	0.3333	0.8571	0.0000
	Experience	0	7	5

Appendix C

R Code for Chapter 4

```
#####################################################################
# Load Packages
#####################################################################
require(bnlearn)
require(moments)
 require(ROCR)
require(caTools)
source("http://bioconductor.org/biocLite.R")
biocLite("RBGL")
require(gRbase)
require(gRain)
require(pROC)
require(xtable)
#####################################################################
# Set the Working Directory
```

```
#####################################################################
#Mac
base <- "~/Thesis/Discretization"
setwd(base)
getwd()
#####################################################################
# Remove Everything
#####################################################################
rm(list = ls(all = TRUE))
#####################################################################
#Finnish Lake Data-set - Data
#####################################################################
Lake.Data <- read.csv("summerAll.csv", header=TRUE, sep=",")
colnames(Lake.Data)
View(Lake.Data)
dim(Lake.Data)
colnames(Lake.Data) <- c('P', 'Chla', 'Type', 'Lake', 'Year', 'N'
, 'Month', 'Depth', 'Surface Area', 'Color')
#####################################################################
# Log-transform the data
#####################################################################
Lake.Data['P'] = log(Lake.Data['P'])
Lake.Data['Chla'] = log(Lake.Data['Chla'])
Lake.Data['N'] = log(Lake.Data['N'])

# Remove Outliers
# Remove one data point with log(Chla) = -23
```

```
Lake.Data <- Lake.Data[!Lake.Data[,'Chla']==min(Lake.Data[,'Chla']),]

# Remove one data point with log(N) > 9
Lake.Data <- Lake.Data[!Lake.Data[,'N']>9,]
Lake.Data <- Lake.Data[!Lake.Data[,'N']<4,]

hist(Lake.Data[,'P'])
hist(Lake.Data[,'N'])
hist(Lake.Data[,'Chla'])

#####################################################################
# Sample 10\%/90\% of eash lake type
# We will use 90\% of data for model development
# & 10\% of data for model evaluation
#####################################################################
for (i in 1:9){
assign(paste("Lake.",i,sep=""), Lake.Data[which(Lake.Data$Type==i),])
}
Type.Size <- c(nrow(Lake.1), nrow(Lake.2), nrow(Lake.3), nrow(Lake.4)
              , nrow(Lake.5), nrow(Lake.6), nrow(Lake.7), nrow(Lake.8)
              , nrow(Lake.9))

for (j in 1:10){
  set.seed(j+10)
  assign(paste("Model.",j,sep=""), 0)
  assign(paste("Evaluation.",j,sep=""), 0)
```

```
  for(i in 1:9){
      assign(paste("L.",i,".M.",j,sep="")
      , get(noquote(paste('Lake.', i, sep=""))))
[sample(1:Type.Size[i]
      , size= round(0.9*Type.Size[i]), replace=FALSE),]
    )
      assign(paste("L.", i,".M.E.",j, sep="")
      , get(noquote(paste('Lake.', i, sep=""))))
[-sample(1:Type.Size[i]
      , size= round(0.9*Type.Size[i]), replace=FALSE),]
    )
    # Keeping variables: N, P, and Chla
      Vars <- c('P', 'Chla', 'N')
      assign(paste("L.",i,".M.",j,sep="")
      , get(noquote(paste('L.',i,".M.",j, sep=""))))[,Vars])
    # Keeping variables: N, P, and Chla
      Vars <- c('P', 'Chla', 'N')
      assign(paste("L.",i,".M.",j,sep="")
      , get(noquote(paste('L.',i,".M.",j, sep=""))))[,Vars])
      assign(paste("L.",i,".M.E.",j,sep="")
      , get(noquote(paste('L.',i,".M.E.",j, sep=""))))[,Vars])

      assign(paste("Model.",j,sep="")
      , rbind(get(noquote(paste("Model.",j,sep="")))
      , get(noquote(paste("L.",i,".M.",j,sep="")))))
      assign(paste("Evaluation.",j,sep="")
      , rbind(get(noquote(paste("Evaluation.",j,sep=""))))
```

```
      , get(noquote(paste("L.",i,".M.E.",j,sep=""))))
  }
}

write.table(Model.1, file="Model1.csv", sep=",")
#####################################################################
#BN discretized with equal interval
#####################################################################
# Break Points
Range.Chla <- range(Lake.Data[,'Chla'])[2]
-range(Lake.Data[,'Chla'])[1]
Breaks.I.Chla = c(min(Lake.Data[,'Chla'])
, min(Lake.Data[,'Chla'])+Range.Chla/3
, max(Lake.Data[,'Chla'])-Range.Chla/3
, max(Lake.Data[,'Chla']))

Range.P <- range(Lake.Data[,'P'])[2]-range(Lake.Data[,'P'])[1]
Breaks.I.P = c(min(Lake.Data[,'P'])
, min(Lake.Data[,'P'])+Range.P/3
, max(Lake.Data[,'P'])-Range.P/3
, max(Lake.Data[,'P']))

Range.N <- range(Lake.Data[,'N'])[2]-range(Lake.Data[,'N'])[1]
Breaks.I.N = c(min(Lake.Data[,'N'])
, min(Lake.Data[,'N'])+Range.N/3
, max(Lake.Data[,'N'])-Range.N/3
, max(Lake.Data[,'N']))
```

```
for (j in 1:10){
assign(paste("Interval.",j,sep="")
, data.frame(P = cut(get(noquote(paste("Model.",j,sep="")))$P
, breaks = Breaks.I.P
, ordered = TRUE, include.lowest=TRUE, include.highest=TRUE)
          , Chla = cut(get(noquote(paste("Model.",j,sep="")))$Chla
          , breaks = Breaks.I.Chla
          , ordered = TRUE, include.lowest=TRUE, include.highest=TRUE)
          , N = cut(get(noquote(paste("Model.",j,sep="")))$N
          , breaks = Breaks.I.N
          , ordered = TRUE
          , include.lowest=TRUE, include.highest=TRUE)))
assign(paste("Interval.",j,sep="")
, get(noquote(paste("Interval.",j,sep=""))))
[complete.cases(get(noquote(paste("Interval.",j,sep="")))),])
assign(paste("Interval.E.",j,sep="")
, data.frame(P = cut(get(noquote(paste("Evaluation.",j,sep="")))$P
, breaks = Breaks.I.P, ordered = TRUE, include.lowest=TRUE
, include.highest=TRUE)
, Chla = cut(get(noquote(paste("Evaluation.",j,sep="")))$Chla
, breaks = Breaks.I.Chla, ordered = TRUE
, include.lowest=TRUE, include.highest=TRUE)
          , N = cut(get(noquote(paste("Evaluation.",j,sep="")))$N
          , breaks = Breaks.I.N, ordered = TRUE
          , include.lowest=TRUE, include.highest=TRUE)))
assign(paste("Interval.E.",j,sep="")
```

```
, get(noquote(paste("Interval.E.",j,sep="")))
[complete.cases(get(noquote(paste("Interval.E.",j,sep="")))),])

Model.Interval <-
empty.graph(names(get(noquote(paste("Interval.",j,sep="")))))
modelstring(Model.Interval) <- "[P],[N],[Chla|N:P]"
assign(paste("fit.Interval.",j,sep="")
, bn.fit(Model.Interval, get(noquote(paste("Interval.",j,sep="")))))
#   Model Evaluation for Equal Interval
assign(paste("logLik.Interval.",j,sep="")
, logLik(get(noquote(paste("fit.Interval.",j,sep="")))
, get(noquote(paste("Interval.",j,sep="")))))
#   Predict Class on Training Sample
assign(paste("Interval.Pred.ChlaT.",j,sep="")
, predict(get(noquote(paste("fit.Interval.",j,sep="")))
, "Chla", get(noquote(paste("Interval.",j,sep="")))))
#   AUC for Evaluation Sample
assign(paste("Interval.Pred.ChlaE.",j,sep="")
, predict(get(noquote(paste("fit.Interval.",j,sep="")))
, "Chla", get(noquote(paste("Interval.E.",j,sep="")))))
#   Develop Confusion Matrix for Training Sample
assign(paste("Interval.CM.ChlaE.",j,sep="")
   , table(get(noquote(paste("Interval.E.",j,sep="")))
[,'Chla'],get(noquote(paste("Interval.Pred.ChlaE.",j,sep="")))))
#   Compute Accuracy for Training Sample
assign(paste("Interval.Accuracy.",j,sep="")
, sum(diag(table(get(noquote(paste("Interval.E.",j,sep="")))[,'Chla']
```

```
,get(noquote(paste("Interval.Pred.ChlaE.",j,sep=""))))
/length(get(noquote(paste("Interval.Pred.ChlaE.",j,sep=""))))
assign(paste("Interval.Pred.ChlaE.",j,sep="")
, data.frame(get(noquote(paste("Interval.Pred.ChlaE.",j,sep="")))))
assign(paste("AUC.Interval.",j,sep="")
, auc(as.matrix(get(noquote(paste("Interval.Pred.ChlaE.",j,sep="")))))
, get(noquote(paste("Interval.E.",j,sep="")))[ ,'Chla']))
}

# Plot a Sample ROC
plot.roc(as.matrix(Interval.Pred.ChlaE.1), Interval.E.1[ ,'Chla'])

# Print a Sample Confusion MAtrix
CM.Table <- xtable(Interval.CM.ChlaE.1, digits=3)
print(CM.Table)

# Sum of Square Errors
SSE= NULL
for (i in 1:1943){
  if(Interval.Pred.ChlaE.1[i,1]=="(3.51,6.41]") {SSE[i]=4.96}
  if(Interval.Pred.ChlaE.1[i,1]=="(0.603,3.51]") {SSE[i]=2.0565}
  if(Interval.Pred.ChlaE.1[i,1]=="(-2.3,0.603]") {SSE[i]=-0.8485}
}

Square= NULL
 for (i in 1:1943){Square[i]=(SSE[i]-Evaluation.1[i+1,"Chla"])^2
                   }
```

```
sum(Square, na.rm=TRUE)
#####################################################################
# LATEX tables and figures
#####################################################################
# LATEX OUTPUT by xtable 1.7-3 package
Interval <- matrix(NA, 0, 5)

for(j in 1:10){
  assign(paste("Interval",sep="")
         , rbind(get(noquote(paste("Interval",sep=""))))
         , cbind(get(noquote(paste("Interval.Accuracy.",j,sep=""))))
         , get(noquote(paste("AUC.Interval.",j,sep="")))))))
}

colnames(Interval) = c( "Accuracy", "AUC")

Interval <- rbind(Interval
, cbind(mean(Interval[,"Accuracy"]), mean(Interval[,"AUC"])))



Interval.Table <- xtable(Interval, digits=3)
print(Interval.Table)

# CPTs to LATEX N
print(xtable(fit.Interval.1$N$prob), floating=FALSE)
# CPTs to LATEX P
```

```
print(xtable(fit.Interval.1$P$prob), floating=FALSE)
# CPTs to LATEX Chla
print(xtable(fit.Interval.1$Chla$prob[1:3, 1:3, 1]), floating=FALSE)
print(xtable(fit.Interval.1$Chla$prob[1:3, 1:3, 2]), floating=FALSE)
print(xtable(fit.Interval.1$Chla$prob[1:3, 1:3, 3]), floating=FALSE)

pdf("Plot-Interval.pdf", width=8, height=10)
hist(Model.1[ ,"Chla"], breaks=100, xlim=c(-3,7)
, xlab="Chlorophyll a", main="")
axis(3, at= Breaks.I.Chla, col="black"
, col.ticks="black", col.axis="black", line=-2)
mtext("Interval", 3, line=-2, at=-3.5, col="black")
axis(3, at= Breaks.Q.Chla, col="royalblue4"
, col.ticks="royalblue4", col.axis="royalblue4", line=0)
mtext("Quantile", 3, line=0, at=-3.5, col="royalblue4")
axis(3, at= Breaks.MM.Chla, col="darkgreen"
, col.ticks="darkgreen", col.axis="darkgreen", line=2)
mtext("Moment Matching", 3, line=2, at=-3.5, col="darkgreen")

plot(Model.Interval)

bn.fit.barchart(fit.Interval.1$P, ylab="Phosphorus Levels")
bn.fit.barchart(fit.Interval.1$N, ylab="Nitrogen Levels")
bn.fit.barchart(fit.Interval.1$Chla, ylab="Chlorophyll a Levels")

# bn.fit.dotplot(fit.Interval.1$P, ylab="Phosphorus Levels")
# bn.fit.dotplot(fit.Interval.1$N, ylab="Nitrogen Levels")
```

```
# bn.fit.dotplot(fit.Interval.1$Chla, ylab="Chlorophyll a Levels")

# Histogram of data for model building: continuous vs discretized
hist(Model.1[ ,"Chla"], breaks=100, xlim=c(-3,7), freq = FALSE
, main="", xlab="", border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
hist(Model.1[ ,"Chla"]
, breaks=c(-2.31, 0.6027629, 3.5081109, 6.42)
, main="",  add=T, border="darkgreen"
,  density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(
levels(Interval.1[, "Chla"])
, 1, line=2,at=c(-1,2,5),col="darkgreen")

mtext("Interval", 1,line=2,at=c(-3),col="darkgreen")

# Histogram of data for model evaluation: continuous vs discretized
hist(Evaluation.1[ ,"Chla"], breaks=100, xlim=c(-3,7), freq = FALSE
, main="", xlab="", border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
hist(Evaluation.1[ ,"Chla"]
, breaks=c(-2.31, 0.6027629, 3.5081109, 6.42)
, main="",  add=T, border="darkgreen"
, density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(levels(Interval.E.1[, "Chla"])
, 1, line=2,at=c(-1,2,5),col="darkgreen")
mtext("Interval", 1,line=2,at=c(-3),col="darkgreen")
```

```
# Histogram of continuous data for model evaluation
# versus discretized model prediction
hist(Evaluation.1[ ,"Chla"], breaks=100, xlim=c(-3,7)
, freq = FALSE, main="", xlab="", border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
pred.1 <-
  runif(
  length(Interval.Pred.ChlaE.1[Interval.Pred.ChlaE.1=="[-2.3,0.603]"])
  , min = -2.3, max = 0.6)
pred.2 <-
  runif(
  length(Interval.Pred.ChlaE.1[Interval.Pred.ChlaE.1=="(0.603,3.51]"])
, min = 0.604, max = 3.51)
pred.3 <-
  runif(
  length(Interval.Pred.ChlaE.1[Interval.Pred.ChlaE.1=="(3.51,6.41]"])
, min = 3.52, max = 6.41)
pred <- c(pred.1, pred.2, pred.3)
hist(
pred, breaks=c(-2.31, 0.6027629, 3.5081109, 6.42)
, main="", add=T, border="darkgreen"
, density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(levels(Interval.E.1[, "Chla"])
, 1, line=2,at=c(-1,2,5),col="darkgreen")
mtext("Interval", 1,line=2,at=c(-3),col="darkgreen")
invisible(dev.off())
```

```
#####################################################################
# BN discretized with equal quantile
#####################################################################
# Break Points
Breaks.Q.Chla=c(quantile(Lake.Data[,"Chla"]
, probs = seq(0, 1, by = 1/3)))
Breaks.Q.P=c(quantile(Lake.Data[,"P"], probs = seq(0, 1, by = 1/3)))
Breaks.Q.N=c(quantile(Lake.Data[,"N"], probs = seq(0, 1, by = 1/3)))

for (j in 1:10){
assign(paste("Quantile.",j,sep=""), data.frame(P =
  cut(get(noquote(paste("Model.",j,sep="")))$P
  , breaks = Breaks.Q.P, ordered = TRUE
  , include.lowest=TRUE, include.highest=TRUE)
  , Chla = cut(get(noquote(paste("Model.",j,sep="")))$Chla
  , breaks = Breaks.Q.Chla, ordered = TRUE
  , include.lowest=TRUE, include.highest=TRUE)
  , N = cut(get(noquote(paste("Model.",j,sep="")))$N
  , breaks = Breaks.Q.N, ordered = TRUE
  , include.lowest=TRUE, include.highest=TRUE)))
assign(paste("Quantile.",j,sep="")
, get(noquote(paste("Quantile.",j,sep=""))))
[complete.cases(get(noquote(paste("Quantile.",j,sep="")))),])

assign(paste("Quantile.E.",j,sep=""), data.frame(P =
  cut(get(noquote(paste("Evaluation.",j,sep="")))$P
```

```
  , breaks = Breaks.Q.P, ordered = TRUE
  , include.lowest=TRUE, include.highest=TRUE)
  , Chla = cut(get(noquote(paste("Evaluation.",j,sep="")))$Chla
  , breaks = Breaks.Q.Chla, ordered = TRUE
  , include.lowest=TRUE, include.highest=TRUE)
  , N = cut(get(noquote(paste("Evaluation.",j,sep="")))$N
  , breaks = Breaks.Q.N, ordered = TRUE
  , include.lowest=TRUE, include.highest=TRUE)))
  assign(paste("Quantile.E.",j,sep="")
  , get(noquote(paste("Quantile.E.",j,sep="")))
  [complete.cases(get(noquote(paste("Quantile.E.",j,sep="")))),])

Model.Quantile <-
empty.graph(names(get(noquote(paste("Quantile.",j,sep="")))))
modelstring(Model.Quantile) <- "[P],[N],[Chla|N:P]"
 assign(paste("fit.Quantile.",j,sep="")
 , bn.fit(Model.Quantile
 , get(noquote(paste("Quantile.",j,sep="")))))
#   Model evaluation for equal quantile
#   AUC for evaluation sample
assign(paste("Quantile.Pred.ChlaE.",j,sep="")
, predict(get(noquote(paste("fit.Quantile.",j,sep="")))
, "Chla", get(noquote(paste("Quantile.E.",j,sep="")))))
#   make confusion matrix for training sample
  assign(paste("Quantile.CM.ChlaP.",j,sep="")
   , table(get(noquote(paste("Quantile.E.",j,sep="")))
[,'Chla'],get(noquote(paste("Quantile.Pred.ChlaE.",j,sep="")))))
```

```
#   compute accuracy for training sample
assign(paste("Quantile.Accuracy.",j,sep="")
  , sum(diag(table(get(noquote(paste("Quantile.E.",j,sep=""))))
  [,'Chla'],get(noquote(paste("Quantile.Pred.ChlaE.",j,sep=""))))))
  /length(get(noquote(paste("Quantile.Pred.ChlaE.",j,sep="")))))))
assign(paste("Quantile.Pred.ChlaE.",j,sep="")
   , data.frame(get(noquote(paste("Quantile.Pred.ChlaE.",j,sep="")))))
assign(paste("AUC.Quantile.",j,sep="")
   , auc(as.matrix(get(noquote(
paste("Quantile.Pred.ChlaE.",j,sep="")))))
, get(noquote(paste("Quantile.E.",j,sep="")))[ ,'Chla']))
}

Quantile.Pred.ChlaE.Trial= predict(fit.Quantile.1, "Chla"
, Quantile.E.1[Quantile.E.1[,3]=="[3.43,5.99]",])
Interval.Pred.ChlaE.Trial= predict(fit.Interval.1, "Chla"
, Interval.E.1[Interval.E.1[,3]=="[3.43,5.11]",])
MM.Pred.ChlaE.Trial= predict(fit.MM.1, "Chla"
, MM.E.1[MM.E.1[,3]=="[3.43,5.7]",])

CM.Table <- xtable(Quantile.CM.ChlaP.1, digits=3)
print(CM.Table)

# Sum of Square Errors
SSE.Q= NULL
for (i in 1:1943){
  if(Quantile.Pred.ChlaE.1[i,1]=="[-2.3,1.72]") {SSE.Q[i]=-0.29}
```

```
  if(Quantile.Pred.ChlaE.1[i,1]=="(1.72,2.64]") {SSE.Q[i]=2.18}
  if(Quantile.Pred.ChlaE.1[i,1]=="(2.64,6.41]") {SSE.Q[i]=4.525}
}

Square.Q= NULL
for (i in 1:1943){Square.Q[i]=(SSE.Q[i]-Evaluation.1[i+1,"Chla"])^2
}

sum(Square.Q, na.rm=TRUE)
#####################################################################
# LATEX tables and figures
#####################################################################
# LATEX OUTPUT by xtable 1.7-3 package
Quantile <- matrix(NA, 0, 5)

for(j in 1:10){
  assign(paste("Quantile",sep="")
         , rbind(get(noquote(paste("Quantile",sep=""))))
         , cbind(get(noquote(paste("Quantile.Accuracy.",j,sep=""))))
         , get(noquote(paste("AUC.Quantile.",j,sep="")))))))
}

colnames(Quantile) = c("Accuracy", "AUC")

Quantile <- rbind(Quantile
, cbind(mean(Quantile[,"Accuracy"]), mean(Quantile[,"AUC"])))
```

```
Quantile.Table <- xtable(Quantile, digits=3)
print(Quantile.Table)

# CPTs to LATEX N
print(xtable(fit.Quantile.1$N$prob), floating=FALSE)
# CPTs to LATEX P
print(xtable(fit.Quantile.1$P$prob), floating=FALSE)
# CPTs to LATEX Chla
print(xtable(fit.Quantile.1$Chla$prob[1:3, 1:3, 1]), floating=FALSE)
print(xtable(fit.Quantile.1$Chla$prob[1:3, 1:3, 2]), floating=FALSE)
print(xtable(fit.Quantile.1$Chla$prob[1:3, 1:3, 3]), floating=FALSE)

pdf("Plot-Quantile.pdf", width=8, height=10)
bn.fit.barchart(fit.Quantile.1$P, ylab="Phosphorus Levels")
bn.fit.barchart(fit.Quantile.1$N, ylab="Nitrogen Levels")
bn.fit.barchart(fit.Quantile.1$Chla, ylab="Chlorophyll a Levels")

# Histogram of data for model building: continuous vs discretized
hist(Model.1[ ,"Chla"], breaks=100, xlim=c(-3,7), freq = FALSE
, main="", xlab="", border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
hist(Model.1[ ,"Chla"]
, breaks=c(-2.3025851, 1.722767, 2.639057, 6.4134590)
, main="",  add=T, border="darkgreen"
,  density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(levels(Quantile.1[, "Chla"])
, 1,line=2,at=c(-1,2,5),col="darkgreen")
```

```
mtext("Quantile", 1,line=2,at=c(-3),col="darkgreen")

# Histogram of data for model evaluation: continuous vs discretized
hist(Evaluation.1[ ,"Chla"], breaks=100, xlim=c(-3,7), freq = FALSE
, main="", xlab="", border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
hist(Evaluation.1[ ,"Chla"]
, breaks=c(-2.3025851, 1.722767, 2.639057, 6.4134590)
, main="",  add=T, border="darkgreen"
, density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(levels(Quantile.E.1[, "Chla"])
, 1,line=2,at=c(-1,2,5),col="darkgreen")
mtext("Quantile", 1,line=2,at=c(-3),col="darkgreen")

# Histogram of continuous data for model evaluation
# versus discretized model prediction
hist(Evaluation.1[ ,"Chla"], breaks=100, xlim=c(-3,7), freq = FALSE
, main="", xlab="", border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
pred.1 <-
  runif(length(
  Quantile.Pred.ChlaE.1[Quantile.Pred.ChlaE.1=="[-2.3,1.72]"])
, min = -2.3025851, max = 1.722767)
pred.2 <-
  runif(length(
  Quantile.Pred.ChlaE.1[Quantile.Pred.ChlaE.1=="(1.72,2.64]"])
, min = 1.722767, max = 2.639057)
```

```
pred.3 <-
          runif(length(Quantile.Pred.ChlaE.1[
          Quantile.Pred.ChlaE.1=="(2.64,6.41]"])
, min = 2.639057, max = 6.4134590)
pred <- c(pred.1, pred.2, pred.3)
hist(pred, breaks=c(-2.3025851, 1.722767, 2.639057, 6.4134590)
, main="",  add=T, border="darkgreen"
, density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(levels(Quantile.E.1[, "Chla"])
, 1,line=2,at=c(-1,2,5),col="darkgreen")
mtext("Quantile", 1,line=2,at=c(-3),col="darkgreen")
invisible(dev.off())
#####################################################################
# Break Points: BN discretized with moment matching method for P
#####################################################################
mu.P = mean(Lake.Data[,'P'])
sd.P = sd(Lake.Data[,'P'])
s.P = skewness(Lake.Data[,'P'], na.rm = FALSE)
k.P = kurtosis(Lake.Data[,'P'], na.rm = FALSE)
f.P = moment(Lake.Data[,'P'], order = 5, central=TRUE
, na.rm = FALSE)/sd(Lake.Data[,'P'])^5
c0.P = (f.P-2*s.P*k.P+s.P^3)/(k.P-s.P^2-1)
c1.P = (s.P*f.P-k.P^2+k.P-s.P^2)/(k.P-s.P^2-1)
c2.P = (-f.P+s.P*k.P+s.P)/(k.P-s.P^2-1)

# polyroot: Find zeros of a real or complex polynomial.
root.P = polyroot(c(c0.P, c1.P, c2.P, 1))
```

```
Im(root.P)
root.P = Re(root.P)

pa.P =
(1+root.P[2]*root.P[3])/((root.P[2]-root.P[1])*(root.P[3]-root.P[1]))
pb.P =
(1+root.P[1]*root.P[3])/((root.P[1]-root.P[2])*(root.P[3]-root.P[2]))
pc.P =
(1+root.P[1]*root.P[2])/((root.P[1]-root.P[3])*(root.P[2]-root.P[3]))
a.P = mu.P+sd.P*min(root.P)
b.P = mu.P+sd.P*median(root.P)
c.P = mu.P+sd.P*max(root.P)

c(a.P,b.P,c.P)
c(pa.P,pb.P,pc.P)
hist(Lake.Data[,'P'])

Break.MM.P.1 = min(Lake.Data[,'P'])
Break.MM.P.2 = a.P+(pa.P/(pa.P+pb.P))*(b.P-a.P)
Break.MM.P.3 = b.P+ (pb.P/(pb.P+pc.P))*(c.P-b.P)
Break.MM.P.4 = max(Lake.Data[,'P'])

Breaks.MM.P <- c(Break.MM.P.1, Break.MM.P.2
 , Break.MM.P.3, Break.MM.P.4)
#####################################################################
# Break Points: BN discretized with moment matching method for N
#####################################################################
```

```
mu.N = mean(Lake.Data[,'N'])
sd.N = sd(Lake.Data[,'N'])
s.N = skewness(Lake.Data[,'N'], na.rm = FALSE)
k.N = kurtosis(Lake.Data[,'N'], na.rm = FALSE)
f.N = moment(Lake.Data[,'N'], order = 5, central=TRUE
, na.rm = FALSE)/sd(Lake.Data[,'N'])^5
c0.N = (f.N-2*s.N*k.N+s.N^3)/(k.N-s.N^2-1)
c1.N = (s.N*f.N-k.N^2+k.N-s.N^2)/(k.N-s.N^2-1)
c2.N = (-f.N+s.N*k.N+s.N)/(k.N-s.N^2-1)

# polyroot: Find zeros of a real or complex polynomial.
root.N = polyroot(c(c0.N, c1.N, c2.N, 1))
Im(root.N)
root.N = Re(root.N)

pa.N =
(1+root.N[2]*root.N[3])/((root.N[2]-root.N[1])*(root.N[3]-root.N[1]))
pb.N =
(1+root.N[1]*root.N[3])/((root.N[1]-root.N[2])*(root.N[3]-root.N[2]))
pc.N =
(1+root.N[1]*root.N[2])/((root.N[1]-root.N[3])*(root.N[2]-root.N[3]))
a.N = mu.N+sd.N*min(root.N)
b.N = mu.N+sd.N*median(root.N)
c.N = mu.N+sd.N*max(root.N)

c(a.N,b.N,c.N)
hist(Lake.Data[,'N'])
```

```
Break.MM.N.1 = min(Lake.Data[,'N'])
Break.MM.N.2 = a.N+(pa.N/(pa.N+pb.N))*(b.N-a.N)
Break.MM.N.3 = b.N+ (pb.N/(pb.N+pc.N))*(c.N-b.N)
Break.MM.N.4 = max(Lake.Data[,'N'])

Breaks.MM.N <- c(Break.MM.N.1, Break.MM.N.2
, Break.MM.N.3, Break.MM.N.4)
#####################################################################
# Break Points: BN discretized with moment matching method for Chla
#####################################################################
mu.Chla = mean(Lake.Data[,'Chla'])
sd.Chla = sd(Lake.Data[,'Chla'])
s.Chla = skewness(Lake.Data[,'Chla'], na.rm = FALSE)
k.Chla = kurtosis(Lake.Data[,'Chla'], na.rm = FALSE)
f.Chla = moment(Lake.Data[,'Chla'], order = 5
, central=TRUE, na.rm = FALSE)/sd(Lake.Data[,'Chla'])^5
c0.Chla = (f.Chla-2*s.Chla*k.Chla+s.Chla^3)/(k.Chla-s.Chla^2-1)
c1.Chla = (s.Chla*f.Chla-k.Chla^2+k.Chla-s.Chla^2)/(k.Chla-s.Chla^2-1)
c2.Chla = (-f.Chla+s.Chla*k.Chla+s.Chla)/(k.Chla-s.Chla^2-1)

# polyroot: Find zeros of a real or complex polynomial.
root.Chla = polyroot(c(c0.Chla, c1.Chla, c2.Chla, 1))
Im(root.Chla)
root.Chla = Re(root.Chla)

pa.Chla = (1+root.Chla[2]*root.Chla[3])
```

```
/((root.Chla[2]-root.Chla[1])*(root.Chla[3]-root.Chla[1]))
pb.Chla = (1+root.Chla[1]*root.Chla[3])
/((root.Chla[1]-root.Chla[2])*(root.Chla[3]-root.Chla[2]))
pc.Chla = (1+root.Chla[1]*root.Chla[2])
/((root.Chla[1]-root.Chla[3])*(root.Chla[2]-root.Chla[3]))
a.Chla = mu.Chla+sd.Chla*min(root.Chla)
b.Chla = mu.Chla+sd.Chla*median(root.Chla)
c.Chla = mu.Chla+sd.Chla*max(root.Chla)

Break.MM.Chla.1 = min(Lake.Data[,'Chla'])
Break.MM.Chla.2 = a.Chla+(pa.Chla/(pa.Chla+pb.Chla))*(b.Chla-a.Chla)
Break.MM.Chla.3 = b.Chla+ (pb.Chla/(pb.Chla+pc.Chla))*(c.Chla-b.Chla)
Break.MM.Chla.4 = max(Lake.Data[,'Chla'])
Breaks.MM.Chla <- c(Break.MM.Chla.1, Break.MM.Chla.2
, Break.MM.Chla.3, Break.MM.Chla.4)
#####################################################################
# BN discretized with moment matching method
#####################################################################
for (j in 1:10){
assign(paste("MM.",j,sep=""), data.frame(P
= cut(get(noquote(paste("Model.",j,sep="")))$P
, breaks = Breaks.MM.P, ordered = TRUE
, include.lowest=TRUE, include.highest=TRUE)
, Chla = cut(get(noquote(paste("Model.",j,sep="")))$Chla
, breaks = Breaks.MM.Chla, ordered = TRUE
, include.lowest=TRUE, include.highest=TRUE)
, N = cut(get(noquote(paste("Model.",j,sep="")))$N
```

```
, breaks = Breaks.MM.N, ordered = TRUE
, include.lowest=TRUE, include.highest=TRUE)))
assign(paste("MM.",j,sep=""), get(noquote(paste("MM.",j,sep="")))
  [complete.cases(get(noquote(paste("MM.",j,sep="")))),])

assign(paste("MM.E.",j,sep="")
, data.frame(P = cut(get(noquote(paste("Evaluation.",j,sep="")))$P
, breaks = Breaks.MM.P, ordered = TRUE
, include.lowest=TRUE, include.highest=TRUE)
, Chla = cut(get(noquote(paste("Evaluation.",j,sep="")))$Chla
, breaks = Breaks.MM.Chla, ordered = TRUE
, include.lowest=TRUE, include.highest=TRUE)
, N = cut(get(noquote(paste("Evaluation.",j,sep="")))$N
, breaks = Breaks.MM.N, ordered = TRUE
, include.lowest=TRUE, include.highest=TRUE)))

assign(paste("MM.E.",j,sep="")
, get(noquote(paste("MM.E.",j,sep="")))
[complete.cases(get(noquote(paste("MM.E.",j,sep="")))),])

Model.MM <- empty.graph(names(get(noquote(paste("MM.",j,sep="")))))
modelstring(Model.MM) <- "[P],[N],[Chla|N:P]"
assign(paste("fit.MM.",j,sep="")
, bn.fit(Model.MM, get(noquote(paste("MM.",j,sep="")))))

  #   AUC for evaluation sample
  assign(paste("MM.Pred.ChlaE.",j,sep="")
```

```
   , predict(get(noquote(paste("fit.MM.",j,sep=""))))
, "Chla", get(noquote(paste("MM.E.",j,sep=""))))))
  #   make confusion matrix for testing sample
assign(paste("MM.CM.ChlaT.",j,sep="")
   , table(get(noquote(paste("MM.E.",j,sep=""))))
[,'Chla'],get(noquote(paste("MM.Pred.ChlaE.",j,sep=""))))))
  #   compute accuracy for training sample
  assign(paste("MM.Accuracy.",j,sep="")
   , sum(diag(table(get(noquote(paste("MM.E.",j,sep=""))))
[,'Chla'],get(noquote(paste("MM.Pred.ChlaE.",j,sep=""))))))
/length(get(noquote(paste("MM.Pred.ChlaE.",j,sep="")))))))
  assign(paste("MM.Pred.ChlaE.",j,sep="")
   , data.frame(get(noquote(paste("MM.Pred.ChlaE.",j,sep="")))))
  assign(paste("AUC.MM.",j,sep="")
   , auc(as.matrix(get(noquote(paste("MM.Pred.ChlaE.",j,sep="")))))
, get(noquote(paste("MM.E.",j,sep="")))[ ,'Chla']))
}



CM.Table <- xtable(MM.CM.ChlaT.1, digits=3)
print(CM.Table)

# Sum of Square Errors
SSE.MM= NULL
for (i in 1:1943){
  if(MM.Pred.ChlaE.1[i,1]=="[-2.3,1.91]")
   {SSE.MM[i]=-0.195} # -2.3+((1.91+2.3)/2)
```

```
  if(MM.Pred.ChlaE.1[i,1]=="(1.91,3.39]")
   {SSE.MM[i]=2.65} # 1.91+((3.39-1.91)/2)
  if(MM.Pred.ChlaE.1[i,1]=="(3.39,6.41]")
   {SSE.MM[i]=4.9} # 3.39+((6.41-3.39)/2)
}

Square.MM= NULL
for (i in 1:1943){Square.MM[i]=(SSE.MM[i]-Evaluation.1[i+1,"Chla"])^2
}

sum(Square.MM, na.rm=TRUE)
#####################################################################
# LATEX tables and figures
#####################################################################
# LATEX OUTPUT by xtable 1.7-3 package
MM <- matrix(NA, 0, 5)

for(j in 1:10){
  assign(paste("MM",sep="")
         , rbind(get(noquote(paste("MM",sep="")))
                 , cbind(get(noquote(paste("MM.Accuracy.",j,sep="")))
                         , get(noquote(paste("AUC.MM.",j,sep=""))))))
}

colnames(MM) = c("Accuracy", "AUC")

MM <- rbind(MM,
```

```
            cbind(mean(MM[,"Accuracy"]), mean(MM[,"AUC"])))


MM.Table <- xtable(MM, digits=3)
print(MM.Table)

# CPTs to LATEX N
print(xtable(fit.MM.1$N$prob), floating=FALSE)
# CPTs to LATEX P
print(xtable(fit.MM.1$P$prob), floating=FALSE)
# CPTs to LATEX Chla
print(xtable(fit.MM.1$Chla$prob[1:3, 1:3, 1]), floating=FALSE)
print(xtable(fit.MM.1$Chla$prob[1:3, 1:3, 2]), floating=FALSE)
print(xtable(fit.MM.1$Chla$prob[1:3, 1:3, 3]), floating=FALSE)

pdf("Plot-MM.pdf", width=8, height=10)
bn.fit.barchart(fit.MM.1$P, ylab="Phosphorus Levels")
bn.fit.barchart(fit.MM.1$N, ylab="Nitrogen Levels")
bn.fit.barchart(fit.MM.1$Chla, ylab="Chlorophyll a Levels")

# Histogram of data for model building: continuous vs discretized
hist(Model.1[ ,"Chla"], breaks=100, xlim=c(-3,7)
, freq = FALSE, main="", xlab=""
, border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
hist(Model.1[ ,"Chla"], breaks=c(-2.3025851, 1.497835, 3.231148, 6.4134590)
, main="",  add=T, border="darkgreen"
```

```
,  density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(levels(MM.1[, "Chla"]), 1,line=2,at=c(-1,2,5),col="darkgreen")
mtext("Moment Matching", 1,line=2,at=c(-3),col="darkgreen")

# Histogram of data for model evaluation: continuous vs discretized
hist(Evaluation.1[ ,"Chla"], breaks=100, xlim=c(-3,7)
, freq = FALSE, main="", xlab=""
, border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
hist(Evaluation.1[ ,"Chla"]
, breaks=c(-2.3025851, 1.497835, 3.231148, 6.4134590)
, main="",  add=T, border="darkgreen"
,  density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(levels(MM.E.1[, "Chla"]), 1,line=2,at=c(-1,2,5),col="darkgreen")
mtext("Moment Matching", 1,line=2,at=c(-3),col="darkgreen")

# Histogram of continuous data for model evaluation
# versus discretized model prediction
hist(Evaluation.1[ ,"Chla"], breaks=100, xlim=c(-3,7)
, freq = FALSE, main="", xlab="", border="royalblue4", xaxt = 'n')
axis(1, col.axis="royalblue4", , xlab="Chlorophyll a")
pred.1 <-
  runif(length(MM.Pred.ChlaE.1[MM.Pred.ChlaE.1=="[-2.3,2.56]"])
, min = -2.3025851, max =  2.564949)
pred.2 <-
  runif(length(MM.Pred.ChlaE.1[MM.Pred.ChlaE.1=="(2.56,3.37]"])
, min =  2.564949, max = 3.367296 )
```

```
pred.3 <-
    runif(length(MM.Pred.ChlaE.1[MM.Pred.ChlaE.1=="(3.37,6.41]"])
, min =  3.367296, max = 6.4134590)
pred <- c(pred.1, pred.2, pred.3)
hist(pred, breaks=c(-2.3025851, 1.497835, 3.231148, 6.4134590)
, main="",  add=T, border="darkgreen"
,  density=t(t(c(20,20,20))), angle=t(t(c(45,45,45))))
mtext(levels(MM.E.1[, "Chla"]), 1,line=2,at=c(-1,2,5),col="darkgreen")
mtext("Moment Matching", 1,line=2,at=c(-3),col="darkgreen")
invisible(dev.off())
```

Appendix D

Supplementary Material for Chapter 5

D.1 Model Formulation

The following equations describe the chlorophyll *a* model, also depicted in Figure 5. Chlorophyll *a* has a normal distribution. The corresponding mean is calculated from a regression model developed for chlorophyll *a* and its predictors. The prior for all coefficients and precision is specified afterwards.

$$\begin{aligned}\text{Chlorophyll } a_i \sim\ & \mathcal{N}((\beta_0[Section_i] + \beta_1 Nitrogen_i + \beta_2 Phosphorus_i \\ & + \beta_3 Stratification_i + \beta_4 Light_i + \beta_5 Salinity_i \times Nitrogen_i \\ & + \beta_6[Season_i] \times Temperature_i), \sigma^2_{Chlorophyll\ a})\end{aligned}$$

Specifying Priors

$$
\begin{aligned}
&\beta_0[j] \sim \mathcal{N}(\mu_{\beta_0}, \sigma^2_{\beta_0}), \text{where } j = 1, 2, 3 \\
&\mu_{\beta_0} \sim \mathcal{N}(0, 0.0001) \\
&\sigma_{\beta_0} \sim \mathcal{U}(0, 100) \\
&\beta_k \sim \mathcal{N}(0, 0.0001), \text{where } k = 1, 2, 3, 4, 5 \\
&\beta_6[j] \sim \mathcal{N}(\mu_{\beta_6}, \sigma^2_{\beta_6}), \text{where } j = 1, 2, 3, 4 \\
&\mu_{\beta_6} \sim \mathcal{N}(0, 0.0001) \\
&\sigma_{\beta_6} \sim \mathcal{U}(0, 100) \\
&\sigma_{Chla} \sim \mathcal{U}(0, 100)
\end{aligned}
\tag{D.1}
$$

The following equations describe the oxygen model, also depicted in Figure 4. Chlorophyll a has a normal distribution. The corresponding mean is calculated from a regression model developed for oxygen and its predictors. The prior for all coefficients and precision is specified afterwards.

$$
\begin{aligned}
\text{Oxygen}_\text{i} \;\sim\; \mathcal{N}((&\alpha_0[Section_i] + \alpha_1 Chla_i + \alpha_2[Season_i] \times Rstratification_i \\
&+\alpha_3[Season_i] \times Temperature_i), \sigma^2_{oxygen})
\end{aligned}
$$

Specifying Priors

$$
\begin{aligned}
&\alpha_0[j] \sim \mathcal{N}(\mu_{\alpha_0}, \sigma^2_{\alpha_0}), \text{where } j = 1, 2, 3 \\
&\mu_{\alpha_0} \sim \mathcal{N}(0, 0.0001) \\
&\sigma_{\alpha_0} \sim \mathcal{U}(0, 100) \\
&\alpha_1 \sim \mathcal{N}(0, 0.0001) \\
&\alpha_2[j] \sim \mathcal{N}(\mu_{\alpha_2}, \sigma^2_{\alpha_2}), \text{where } j = 1, 2, 3, 4 \\
&\mu_{\alpha_2} \sim \mathcal{N}(0, 0.0001) \\
&\sigma_{\alpha_2} \sim \mathcal{U}(0, 100) \\
&\alpha_3[j] \sim \mathcal{N}(\mu_{\alpha_3}, \sigma^2_{\alpha_3}), \text{where } j = 1, 2, 3, 4 \\
&\mu_{\alpha_3} \sim \mathcal{N}(0, 0.0001) \\
&\sigma_{\alpha_3} \sim \mathcal{U}(0, 100) \\
&\sigma_{Oxygen} \sim \mathcal{U}(0, 100)
\end{aligned}
\tag{D.2}
$$

The following equations describe the combined model, also depicted in Figure 6. The nodes (and the variables represented by the nodes) in the combined model are classified into forcing nodes (nodes without parents, e.g. temperature), intermediate nodes (with both parents and child, e.g. chlorophyll *a*), and terminal nodes (without child, e.g. oxygen). Operationally, the combination process is complete when observations for intermediate nodes (i.e. chlorophyll *a*) are replaced by their respective means. The prior for all coefficients and precision is specified afterwards.

$$
\begin{aligned}
&\text{Oxygen}_{\text{i}} \sim \mathcal{N}(\mu_{Oxygen}, \sigma^2_{Oxygen}) \\
&\mu_{Oxygen_i} = \mathbf{AX_{Oxygen}} \\
&\text{Chlorophyll } a_{\text{i}} \sim \mathcal{N}(\mu_{Chlorophyll\ a}, \sigma^2_{Chlorophyll\ a}) \\
&\mu_{Chlorophyll\ a_i} = \mathbf{BX_{Chlorophyll\ a}} \\
&\mathbf{A} = \begin{pmatrix} \alpha_0[k] & \alpha_1 & \alpha_2[j] & \alpha_3[j] \end{pmatrix}_{1\times 7} \\
&\mathbf{B} = \begin{pmatrix} \beta_0[k] & \beta_1 & \cdots & \beta_5 & \beta_6[j] \end{pmatrix}_{1\times 7}
\end{aligned}
$$

$$\alpha_0[k] \;\&\; \beta_0[k] = \begin{cases} 1 & \text{if } Section_i = k, \text{ where k=1, 2, 3;} \\ 0 & \text{if } Section_i \neq k. \end{cases}$$

$$\alpha_2[j] \;\&\; \alpha_3[j] \;\&\; \beta_6[j] = \begin{cases} 1 & \text{if } Season_i = j, \text{ where j=1, 2, 3,4;} \\ 0 & \text{if } Season_i \neq j. \end{cases}$$

$\mathbf{X_{Oxygen}}$ and $\mathbf{X_{Chlorophyll\ a}}$ are design matrices of predictors for the oxygen and chlorophyll a model. We have 408 observations from 2007 to 2011. For the oxygen (chlorophylla a) model, we have three (6) predictors, hence $\mathbf{X_{Oxygen}}$ ($\mathbf{X_{Chlorophyll\ a}}$) is a matrix of 408 rows and 4 (7) columns (the first column is a vector of 1s). $\mathbf{A}$ and $\mathbf{B}$ are regression model coefficient vectors both modeled by multivariate-normal distributions with means $\mu_{\mathbf{A}} = \{\alpha_0[k], \alpha_1, \alpha_2[j], \alpha_3[j]\}$ and $\mu_{\mathbf{B}} = \{\beta_0[k], \beta_1, \cdots, \beta_5, \beta_6[j]\}$ and covariance matrices $100 \times \Sigma_{\mathbf{A}}$ and $100 \times \Sigma_{\mathbf{B}}$ from the individual model runs, respectively. $\sigma_{Chlorophyll\ a}$ and σ_{Oxygen} have scaled inverse χ^2 distributions with parameters based on posterior distribution of σ_{Oxygen} ($\sigma_{Chlorophylla}$) in the individual model runs.

The following equations describe the temporal model updating.

$$\begin{aligned}
&\text{Oxygen}_j \;\sim\; \mathcal{N}(\mu_{Oxygen_j}, \sigma^2_{Oxygen}) \\
&\mu_{Oxygen_j} \;=\; \mathbf{A}\mathbf{X_{Oxygen}} \\
&\text{Chlorophyll } a_j \;\sim\; \mathcal{N}(\mu_{Chlorophyll\ a_j}, \sigma^2_{Chlorophyll\ a}) \\
&\mu_{Chlorophyll\ a} \;=\; \mathbf{B}\mathbf{X_{Chlorophyll\ a}} \\
&\mathbf{A} \;=\; \begin{pmatrix} \alpha_0[k] & \alpha_1 & \alpha_2[j] & \alpha_3[j] \end{pmatrix}_{1\times 7} \\
&\mathbf{B} \;=\; \begin{pmatrix} \beta_0[k] & \beta_1 & \cdots & \beta_5 & \beta_6[j] \end{pmatrix}_{1\times 7}
\end{aligned}$$

$$\alpha_0[k] \;\&\; \beta_0[k] = \begin{cases} 1 & \text{if } Section_i = k, \text{ where k=1, 2, 3;} \\ 0 & \text{if } Section_i \neq k. \end{cases}$$

$$\alpha_2[j] \;\&\; \alpha_3[j] \;\&\; \beta_6[j] = \begin{cases} 1 & \text{if } Season_i = j, \text{ where j=1, 2, 3,4;} \\ 0 & \text{if } Season_i \neq j. \end{cases}$$

$\mathbf{X_{Oxygen}}$ and $\mathbf{X_{Chlorophyll\ a}}$ are design matrices of predictors for the oxygen and chlorophyll a model. We have 104 observations. For the oxygen (chlorophylla a)

model, we have three (6) predictors, hence $\mathbf{X_{Oxygen}}$ ($\mathbf{X_{Chlorophyll\ a}}$) is a matrix of 104 rows and 4 (7) columns (the first column is a vector of 1s). σ_{Oxygen} ($\sigma_{Chlorophyll\ a}$) has a scaled inverse χ^2 distribution with parameters based on posterior distribution of σ_{Oxygen} ($\sigma_{Chlorophyll a}$) in the combined model. $\mathbf{A}$ and $\mathbf{B}$ are regression model coefficient vectors both modeled by multivariate-normal distributions with means $\mu_{\mathbf{A}} = \{\alpha_0[k], \alpha_1, \alpha_2[j], \alpha_3[j]\}$ and $\mu_{\mathbf{B}} = \{\beta_0[k], \beta_1, \cdots, \beta_5, \beta_6[j]\}$ and covariance matrices $\Sigma_{\mathbf{A}}$ and $\Sigma_{\mathbf{B}}$, respectively.

D.2 Figures

Recursive partitioning, a data mining tool, was used to explore the structure of the data. Figures D.1 and D.2 depict the pruned tree model; the tree models were pruned to avoid overfitting the data. It has been shown in the literature that chlorophyll *a* concentration in estuaries are affected mainly by light, nutrients, water column mixing, temperature, and grazing. Apart from grazing that is not measured in the NRE, figures D.3 to D.7 depict scatterplot matrices and conditional scatterplots for chlorophyll *a* and its predictor variables. Figure D.8 shows the interacting influence of salinity and nitrogen on chlorophyll *a* concentration. Figures D.9 and D.10 show the interacting effect of stratification and season and temperature and season on bottom dissolved oxygen.

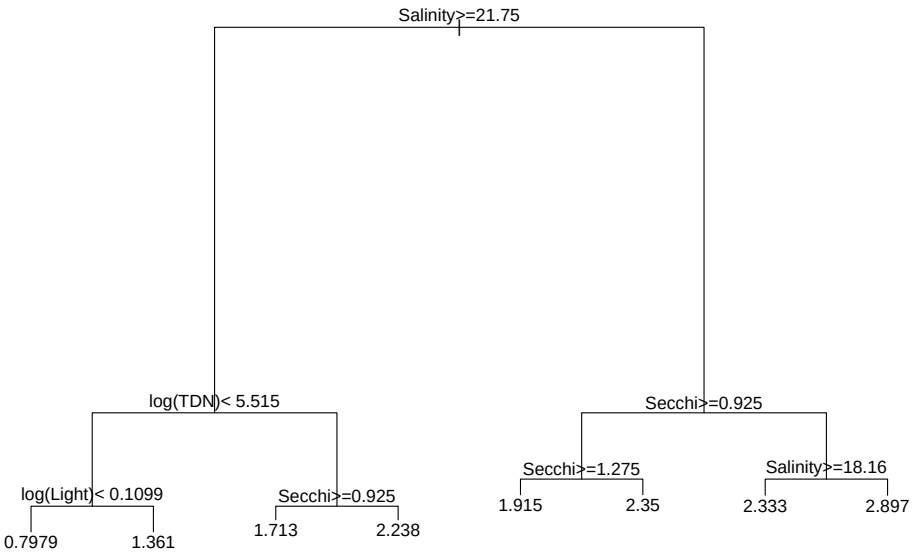


FIGURE D.1: A recursive partitioning (RP) method was applied to the NRE data set from 2007 to 2011. The figure depicts the selected classification tree for predicting chlorophyll *a*.

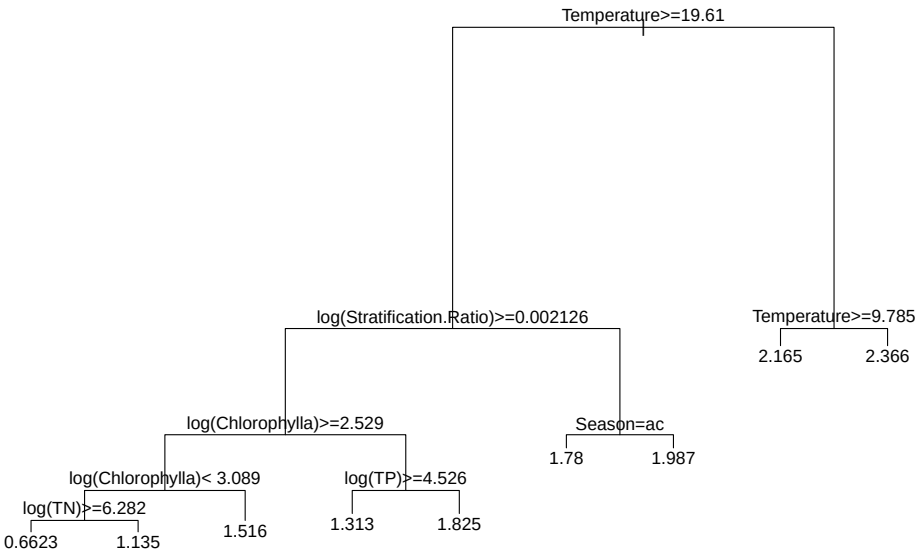


FIGURE D.2: A recursive partitioning (RP) method was applied to the NRE data set from 2007 to 2011. The figure depicts the selected classification tree for predicting bottom dissolved oxygen.

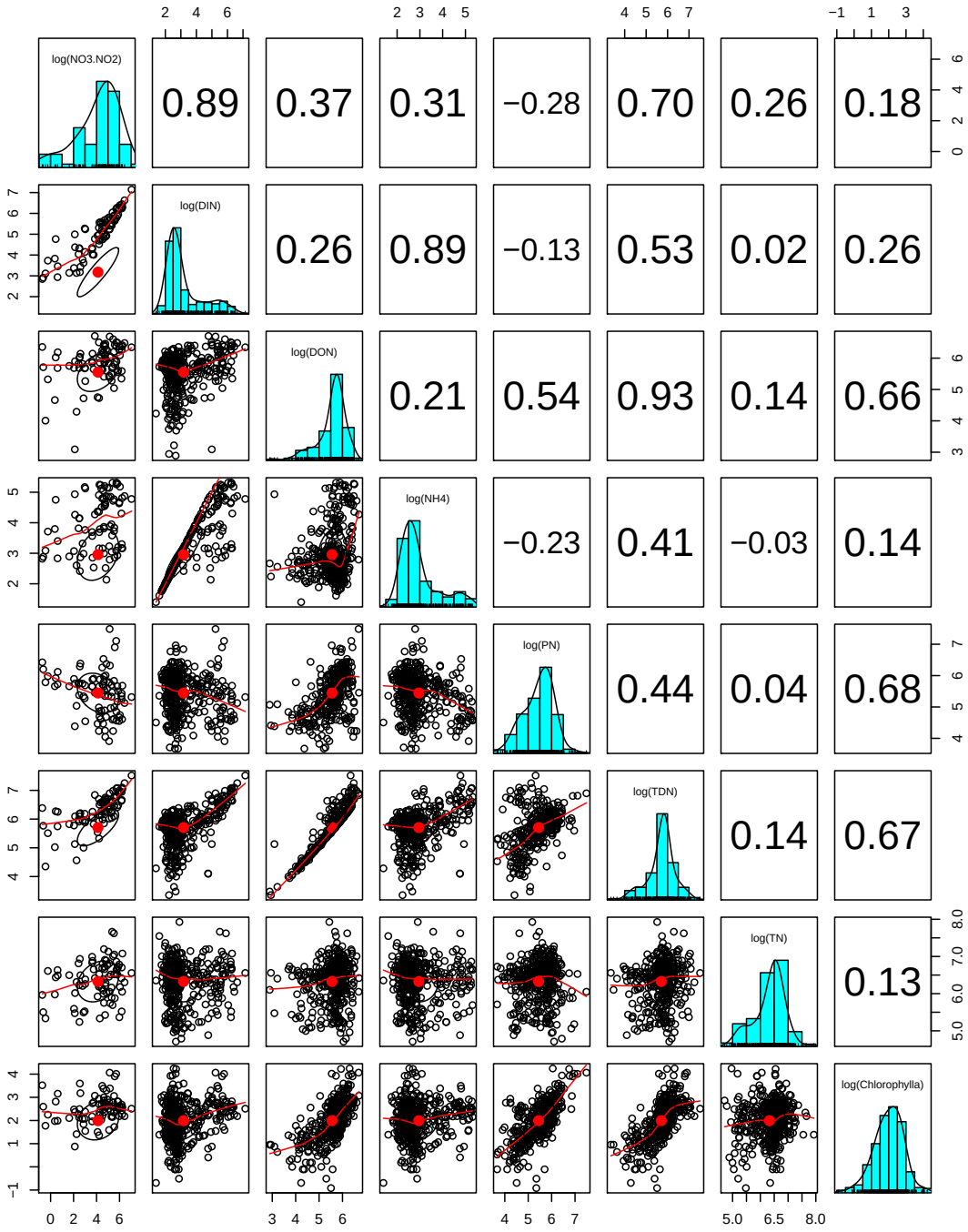


FIGURE D.3: Log-transformed chlorophyll *a* versus log-transformed dissolved/particulate and organic/inorganic nitrogen concentration is depicted.

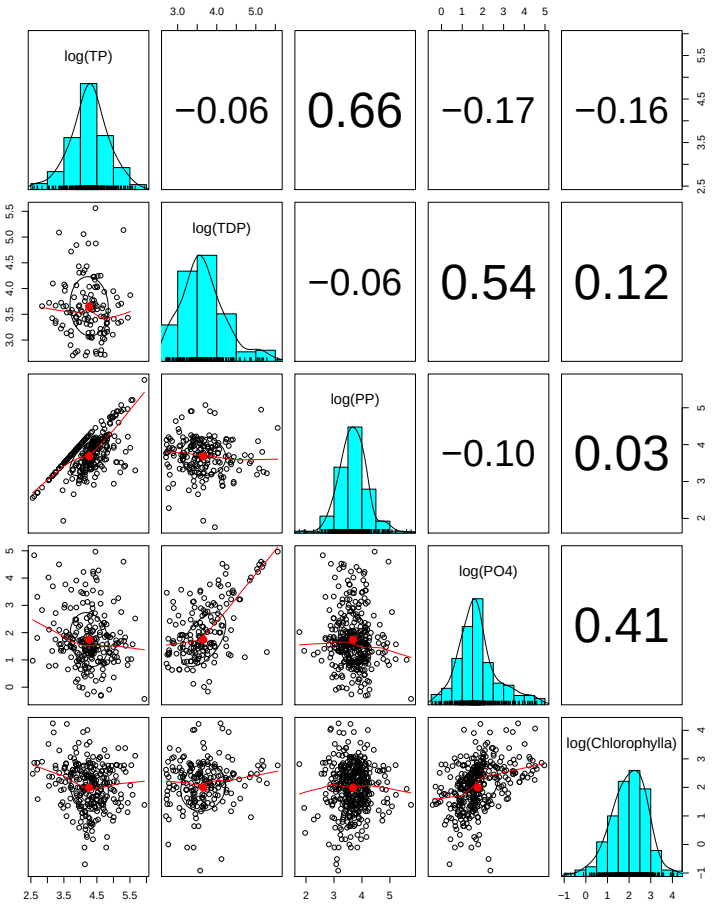


FIGURE D.4: Log-transformed chlorophyll *a* versus log-transformed dissolved/particulate phosphorus concentration.

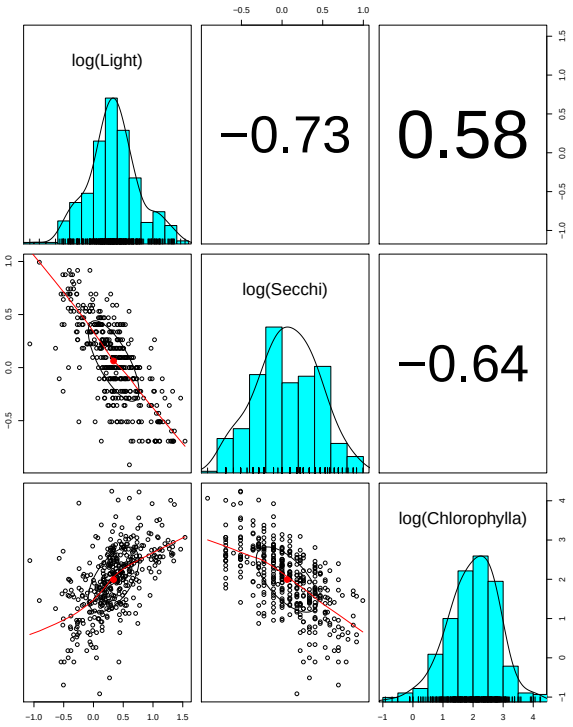


FIGURE D.5: Log-transformed chlorophyll *a* versus log-transformed light attenuation coefficient and Secchi disk depth.

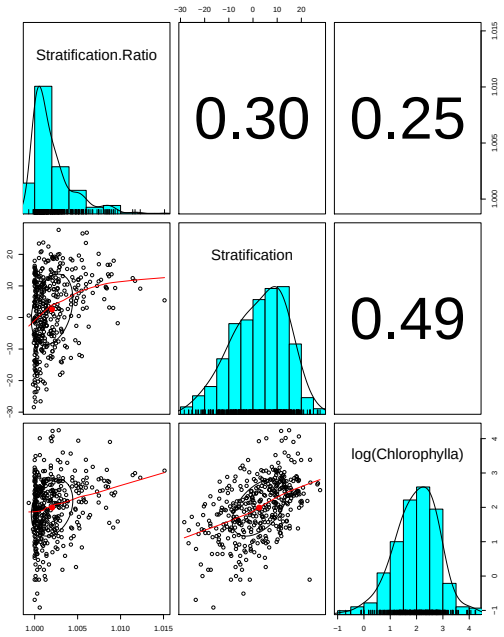


FIGURE D.6: Log-transformed chlorophyll *a* versus stratification, defined as surface and bottom water density gradient, and stratification ratio, defined as surface and bottom water density ratio.

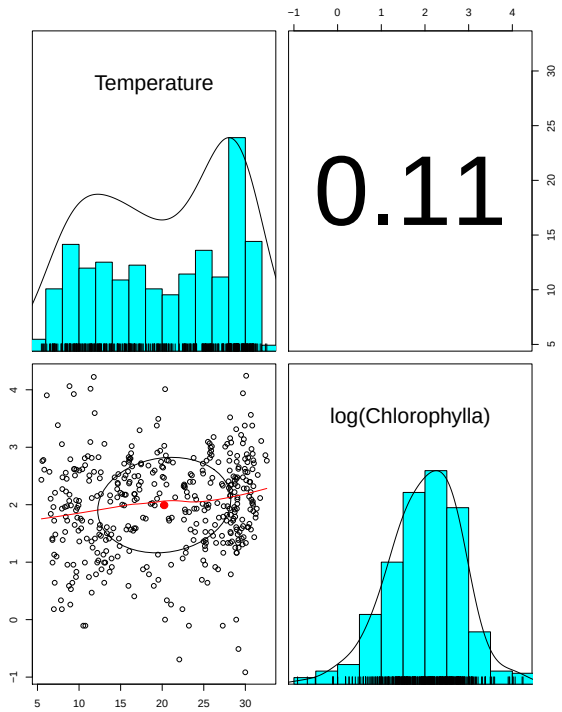


FIGURE D.7: Log-transformed chlorophyll *a* versus temperature.

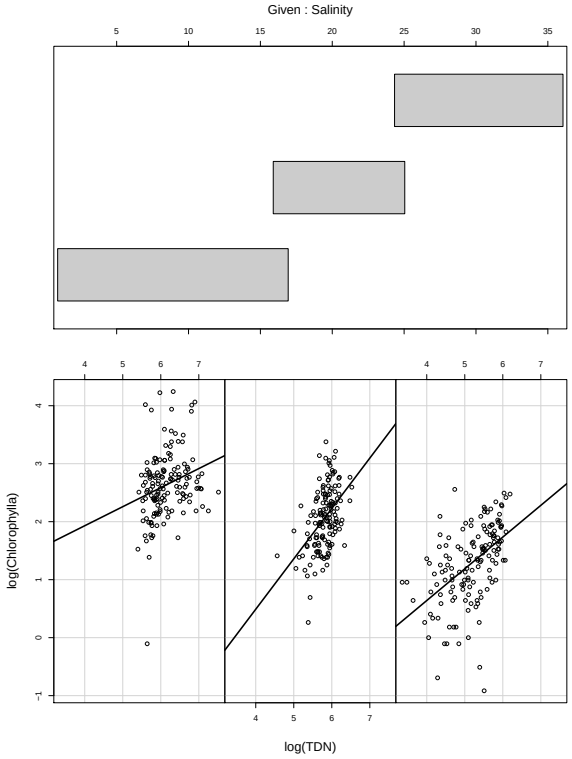


FIGURE D.8: Log-transformed chlorophyll *a* versus total dissolved nitrogen under different salinity.

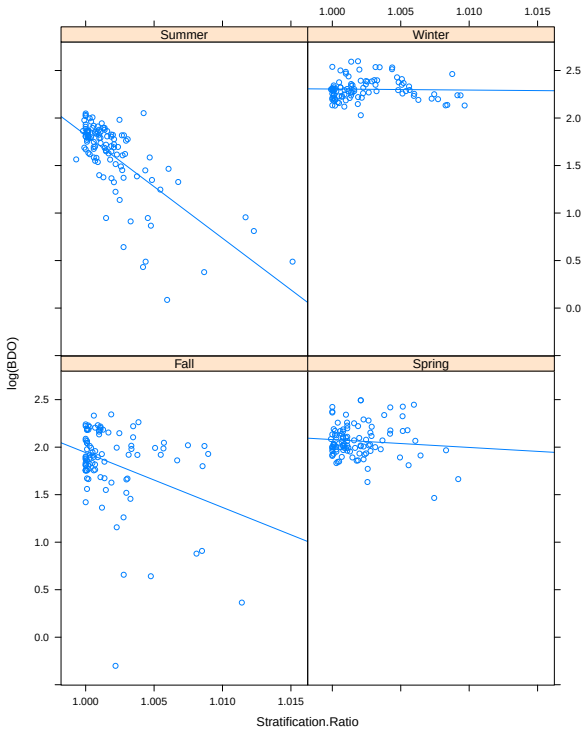


FIGURE D.9: Log-transformed bottom dissolved oxygen versus stratification under different Seasons.

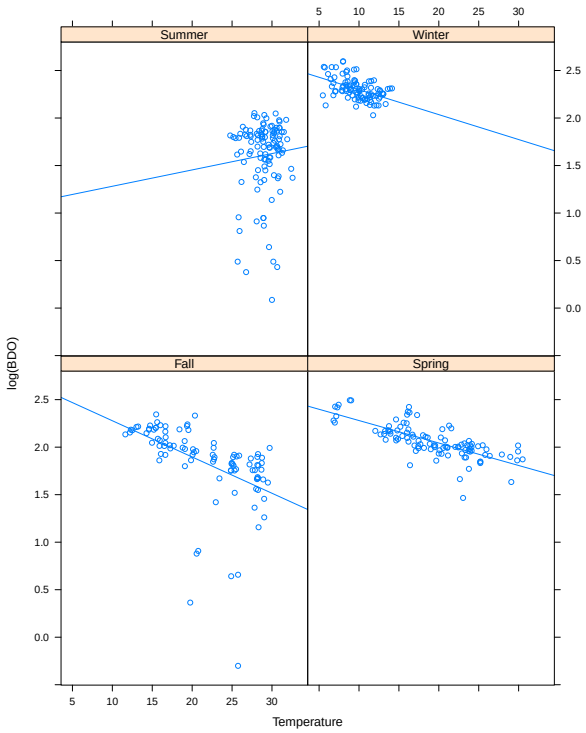


FIGURE D.10: Log-transformed bottom dissolved oxygen versus temperature under different Seasons.

Appendix E

R Code for Chapter 5

E.1 Chlorophyll *a* Model JAGS Code

```
    for (i in 1:408){
        Chla[i] ~ dnorm((beta0[Section[i]]
                        +beta1*Nitrogen[i]
                        +beta2*Phosphorus[i]
                        +beta3*Stratification[i]
                        +beta4*Light[i]
                        +beta5*Salinity[i]*Nitrogen[i]
                        +beta6[Season[i]]*Temperature[i]), tau.Chla)
    }
##Specifying Priors
tau.Chla <- pow(sigma.Chla.hat.2, -1)
sigma.Chla.hat.2 ~ dgamma(10, 1)
##beta0
```

```
for (j in 1:3){
    beta0[j] ~ dnorm(mu.beta0, tau.beta0)
}
mu.beta0~ dnorm(0, 0.0001)
sigma.beta0.hat.2 ~ dgamma(10,1)
tau.beta0 <- pow(sigma.beta0.hat.2,-1)
##beta1
beta1 ~ dnorm(0, 0.0001)
##beta2
beta2 ~ dnorm(0, 0.0001)
##beta3
beta3 ~ dnorm(0, 0.0001)
##beta4
beta4 ~ dnorm(0, 0.0001)
##beta5
beta5 ~ dnorm(0, 0.0001)
##beta6
for (j in 1:4){
    beta6[j] ~ dnorm(mu.beta6, tau.beta6)
}
mu.beta6 ~ dnorm(0, 0.0001)
tau.beta6 <- pow(sigma.beta6.hat.2, -1)
sigma.beta6.hat.2 ~ dgamma(10, 1)
}
```

E.2 Oxygen Model JAGS Code

```
model {
```

```
    for (i in 1:408){
        Oxygen[i] ~ dnorm((beta0[Section[i]]
                           +beta1*Chla[i]
                           +beta2[Season[i]]*RStratification[i]
                           +beta3[Season[i]]*Temperature[i]),tau.Oxygen)
  }
##Specifying Priors
tau.Oxygen <- pow(sigma.Oxygen.hat.2, -1)
sigma.Oxygen.hat.2 ~ dgamma(10, 1)
##beta0
for (j in 1:3){
    beta0[j] ~ dnorm(mu.beta0, tau.beta0)
}
mu.beta0~ dnorm(0,0.0001)
tau.beta0 <- pow(sigma.beta0.hat.2, -1)
sigma.beta0.hat.2 ~ dgamma(10, 1)
##beta1
 beta1~ dnorm(0, 0.0001)
##beta2 & beta3
for (j in 1:4){
    beta2[j] ~ dnorm(mu.beta2, tau.beta2)
    beta3[j] ~ dnorm(mu.beta3, tau.beta3)
  }
mu.beta2 ~ dnorm(0, 0.0001)
tau.beta2 <- pow(sigma.beta2.hat.2, -1)
sigma.beta2.hat.2 ~ dgamma(10, 1)
mu.beta3 ~ dnorm(0, 0.0001)
```

```
tau.beta3<- pow(sigma.beta3.hat.2, -1)
sigma.beta3.hat.2 ~ dgamma(10, 1)
}
```

E.3 Chlorophyll *a*-Oxygen Model JAGS Code

```
model {
    for (i in 1:408){
    Oxygen[i] ~ dnorm(mu.Oxygen[i],tau.Oxygen)
    mu.Oxygen[i]<- Alpha[]%*% X[,i]
    Chla[i] ~ dnorm(mu.Chla[i], tau.Chla)
    mu.Chla[i] <- Beta.Chla[] %*% X.Chla.P[,i]

    X[1:3,i] <- X.Oxygen.P[1:3,i]
    X[4,i] <- mu.Chla[i]
    X[5:12,i] <-  X.Oxygen.P[4:11,i]
}
##Specifying Priors
##tau.Oxygen
tau.Oxygen <- pow(sigma.Oxygen.hat, -2)
sigma.Oxygen.hat ~ dunif(0,3)
Alpha ~ dmnorm(mu.alpha.Oxygen[], omega.alpha.Oxygen[,])
omega.alpha.Oxygen <- inverse(var.alpha.Oxygen[,])
##tau.Chla
##tau.Chla <- pow(sigma.Chla.hat, -2)
sigma.Chla.hat ~ dunif(1,5)
Beta.Chla ~ dmnorm(mu.beta.Chla[], omega.beta.Chla[,])
omega.beta.Chla <- inverse(var.beta.Chla[,])
```

```
##Predicted Chl a & Oxygen values for observed data from 2011-2012
for(j in 1:104){
    Chla.Pred[j] ~ dnorm(mu.Chla.Pred[j], tau.Chla)
    mu.Chla.Pred[j] <- Beta.Chla[] %*% X.Chla[,j]
    Oxygen.Pred[j] ~ dnorm(mu.Oxygen.Pred[j], tau.Oxygen)
    mu.Oxygen.Pred[j] <- Alpha[] %*% X.Pred[,j]
    X.Pred[1:3,j] <- X.Oxygen[1:3,j]
    X.Pred[4,j] <- mu.Chla.Pred[j]
    X.Pred[5:12,j] <-  X.Oxygen[4:11,j]
  }
}
```

E.4 Temporal Model Update JAGS Code

```
#Model Updating with New Data from 2011-2012
model {
  for (i in 1:104){
    Oxygen[i] ~ dnorm(mu.Oxygen[i],tau.Oxygen)
    mu.Oxygen[i]<- Alpha[] %*% X[,i]
    Chla[i] ~ dnorm(mu.Chla[i], tau.Chla)
    mu.Chla[i]<- Beta[] %*% X.Chla[,i]
    X[1:3,i] <- X.Oxygen[1:3,i]
    X[4,i] <- mu.Chla[i]
    X[5:12,i] <-  X.Oxygen[4:11,i]
  }
  #Specifying Priors
  # The Precision
  tau.Oxygen ~ dchisq(6.7)
```

```
  tau.Chla ~ dchisq(4.4)
  #The Coefficients
  Alpha ~ dmnorm(mu.alpha[], omega.alpha[,])
  omega.alpha <- inverse(var.alpha[,])
  Beta ~ dmnorm(mu.beta[], omega.beta[,])
  omega.beta <- inverse(var.beta[,])
}
```

E.5 Spatial Model Update JAGS Code

```
#Model Updating with Neuse River Data from 2007-2012
model {
  for (i in 1:873){
    Oxygen[i] ~ dnorm(mu.Oxygen[i],tau.Oxygen)
    mu.Oxygen[i]<- Alpha[] %*% X[,i]
    Chla[i] ~ dnorm(mu.Chla[i], tau.Chla)
    mu.Chla[i]<- Beta[] %*% X.Chla.P[,i]
    X[1:3,i] <- X.Oxygen.P[1:3,i]
    X[4,i] <- mu.Chla[i]
    X[5:12,i] <-  X.Oxygen.P[4:11,i]
  }
  #Specifying Priors
  # The Precision
  tau.Oxygen <- pow(sigma.Oxygen, -2)
  sigma.Oxygen ~ dnorm(mu.sigma.Oxygen, omega.sigma.Oxygen)
  omega.sigma.Oxygen <- pow(var.sigma.Oxygen,-1)
  tau.Chla <- pow(sigma.Chla, -2)
  sigma.Chla ~ dnorm(mu.sigma.Chla, omega.sigma.Chla)
```

```
  omega.sigma.Chla <- pow(var.sigma.Chla,-1)
  #The Coefficients
  Alpha ~ dmnorm(mu.alpha[], omega.alpha[,])
  omega.alpha <- inverse(var.alpha[,])
  Beta ~ dmnorm(mu.beta[], omega.beta[,])
  omega.beta <- inverse(var.beta[,])
  #Predicted Chla a & Oxygen values for observed data from 2011-2012
  for(j in 1:151){
    Chla.Pred[j] ~ dnorm(mu.Chla.Pred[j], tau.Chla)
    mu.Chla.Pred[j] <- Beta[] %*% X.Chla[,j]
    Oxygen.Pred[j] ~ dnorm(mu.Oxygen.Pred[j], tau.Oxygen)
    mu.Oxygen.Pred[j] <- Alpha[] %*% X.Pred[,j]
    X.Pred[1:3,j] <- X.Oxygen[1:3,j]
    X.Pred[4,j] <- mu.Chla.Pred[j]
    X.Pred[5:12,j] <-  X.Oxygen[4:11,j]
  }
}
```

Bibliography

Aguilera, P., Fernández, A., Fernández, R., Rumí, R., and Salmerón, A. (2011), "Bayesian Networks in Environmental Modelling," *Environmental Modelling & Software*, 26, 1376–1388.

Alameddine, I., Cha, Y., and Reckhow, K. H. (2011), "An evaluation of Automated Structure Learning with Bayesian Networks: An Application to Estuarine Chlorophyll Dynamics," *Environmental Modelling & Software*, 26, 163–172.

Altman, J. C. and Paerl, H. W. (2012), "Composition of inorganic and organic nutrient sources influences phytoplankton community structure in the New River Estuary, North Carolina," *Aquatic Ecology*, 46, 269–282.

Aneja, V. P., Schlesinger, W. H., Nyogi, D., Jennings, G., Gilliam, W., Knighton, R. E., Duke, C. S., Blunden, J., and Krishnan, S. (2006), "Emerging national research needs for agricultural air quality," *Eos, Transactions American Geophysical Union*, 87, 25–29.

Aneja, V. P., Arya, S. P., Kim, D.-S., Rumsey, I. C., Arkinson, H., Semunegus, H., Bajwa, K., Dickey, D., Stefanski, L., Todd, L., et al. (2008a), "Characterizing ammonia emissions from swine farms in eastern North Carolina: part 1conventional lagoon and spray technology for waste treatment," *Journal of the Air & Waste Management Association*, 58, 1130–1144.

Aneja, V. P., Arya, S. P., Rumsey, I. C., Kim, D.-S., Arkinson, H., Semunegus, H., Bajwa, K., Dickey, D., Stefanski, L., Todd, L., et al. (2008b), "Characterizing Ammonia Emissions from Swine Farms in Eastern North Carolina: Part 2Potential Environmentally Superior Technologies for Waste Treatment," *Journal of the Air & Waste Management Association*, 58, 1145–1157.

Arhonditsis, G. B., Brett, M. T., et al. (2004), "Evaluation of the current state of mechanistic aquatic biogeochemical modeling," *Marine Ecology Progress Series*, 271, 13–26.

Armstrong, A. (2009), "Water pollution: Urban waste," *Nature Geoscience*, 2, 748–748.

Arogo, J., Westerman, P., and Heber, A. (2003), "A review of ammonia emissions from confined swine feeding operations," *Transactions of the ASAE*, 46, 805–817.

Barton, D., Saloranta, T., Moe, S., Eggestad, H., and Kuikka, S. (2008), "Bayesian Belief Networks as a Meta-modelling Tool in Integrated River Basin Management-Pros and Cons in Evaluating Nutrient Abatement Decisions under Uncertainty in a Norwegian River Basin," *Ecological Economics*, 66, 91–104.

Beck, M. B. (1987), "Water quality modeling: a review of the analysis of uncertainty," *Water Resources Research*, 23, 1393–1442.

Bicudo, J., Safley, L., and Westerman, P. (1999), "Nutrient content and sludge volumes in single-cell recycle anaerobic swine lagoons in North Carolina," *Transactions of the ASAE*, 42, 1087–1093.

Borsuk, M., Clemen, R., Maguire, L., and Reckhow, K. (2001), "Stakeholder values and scientific modeling in the Neuse River watershed," *Group Decision and Negotiation*, 10, 355–373.

Borsuk, M. E., Stow, C. A., and Reckhow, K. H. (2003), "Integrated Approach to Total Maximum Daily Load Development for Neuse River Estuary using Bayesian Probability Network Model (Neu-BERN)," *Journal of Water Resources Planning and Management*, 129, 271–282.

Borsuk, M. E., Stow, C. A., and Reckhow, K. H. (2004), "A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis," *Ecological Modelling*, 173, 219–239.

Branco, A. B. and Kremer, J. N. (2005), "The relative importance of chlorophyll and colored dissolved organic matter (CDOM) to the prediction of the diffuse attenuation coefficient in shallow estuaries," *Estuaries*, 28, 643–652.

Bricker, S. B., Clement, C. G., Pirhalla, D. E., Orlando, S. P., and Farrow, D. R. (1999), "National Estuarine Eutrophication Assessment: Effects of Nutrient Enrichment in the Nation's Estuaries," .

Bromley, J., Jackson, N. A., Clymer, O., Giacomello, A. M., and Jensen, F. V. (2005), "The use of Hugin to Develop Bayesian Networks as an Aid to Integrated Water Resource Planning," *Environmental Modelling & Software*, 20, 231–242.

Burkholder, J., Libra, B., Weyer, P., Heathcote, S., Kolpin, D., Thorne, P. S., and Wichman, M. (2007), "Impacts of waste from concentrated animal feeding operations on water quality," *Environmental health perspectives*, 115, 308.

Burkholder, J. M., Noga, E. J., Hobbs, C. H., and Glasgow, H. B. (1992), "New'phantom'dinoflagellate is the causative agent of major estuarine fish kills," .

Cassar, N., DiFiore, P., Barnett, B., Bender, M., Bowie, A., Tilbrook, B., Petrou, K., Westwood, K., Wright, S., and Lefevre, D. (2011), "The influence of iron and light on net community production in the Subantarctic and Polar Frontal Zones," *Biogeosciences*, 8, 227–237.

Castelletti, A. and Soncini-Sessa, R. (2007), "Bayesian Networks and Participatory Modelling in Water Resource Management," *Environmental Modelling & Software*, 22, 1075–1088.

Census (2010), "US Census Bureau (2010)," .

Chen, S. H. and Pollino, C. A. (2012), "Good practice in Bayesian network modelling," *Environmental Modelling & Software*, 37, 134–145.

Cloern, J. E. (1987), "Turbidity as a control on phytoplankton biomass and productivity in estuaries," *Continental Shelf Research*, 7, 1367–1381.

Cloern, J. E. (2001), "Our Evolving Conceptual Model of the Coastal Eutrophication Problem," *Marine ecology progress series*, 210, 223–253.

Cloern, J. E., Grenz, C., and Vidergar-Lucas, L. (1995), "An empirical model of the phytoplankton chlorophyll: carbon ratio-the conservation factor between productivity and growth rate," *Limnology and Oceanography*, 40, 1313–1321.

Conley, D. J., Paerl, H. W., Howarth, R. W., Boesch, D. F., Seitzinger, S. P., Havens, K. E., Lancelot, C., Likens, G. E., et al. (2009), "Controlling eutrophication: nitrogen and phosphorus," *Science*, 323, 1014–1015.

Cox, R. (2012), *Environmental Communication and the Public Sphere*, Sage.

Croke, B., Ticehurst, J., Letcher, R., Norton, J., Newham, L., and Jakeman, A. (2007), "Integrated Assessment of Water Resources: Australian Experiences," *Water Resources Management*, 21, 351–373.

CRONOS (2007), "State Climate Office of North Carolina, NCSU," .

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38.

Diaz, R. J. and Rosenberg, R. (2008), "Spreading dead zones and consequences for marine ecosystems," *Science*, 321, 926–929.

Domingues, R. B., Anselmo, T. P., Barbosa, A. B., Sommer, U., and Galvão, H. M. (2011), "Light as a driver of phytoplankton growth and production in the freshwater tidal zone of a turbid estuary," *Estuarine, Coastal and Shelf Science*, 91, 526–535.

Doney, S. C. (2010), "The growing human footprint on coastal and open-ocean biogeochemistry," *Science*, 328, 1512–1516.

Doorn, M., Natschke, D., Thorneloe, S., and Southerland, J. (2002a), "Development of an emission factor for ammonia emissions from US swine farms based on field tests and application of a mass balance method," *Atmospheric Environment*, 36, 5619–5625.

Doorn, M. R., Natschke, D. F., and Meeuwissen, P. C. (2002b), *Review of emission factors and methodologies to estimate ammonia emissions from animal waste handling*, Environmental Protection Agency.

Dorner, S., Shi, J., and Swayne, D. (2007), "Multi-objective Modelling and Decision Support using a Bayesian Network Approximation to a Non-point Source Pollution Model," *Environmental Modelling & Software*, 22, 211–222.

Ellison, A. M. (1996), "An introduction to Bayesian inference for ecological research and environmental decision-making," *Ecological Applications*, pp. 1036–1046.

EPA (2001a), "National Coastal Condition Report," Tech. Rep. EPA-620/R-01/005, Office of Research and Development and Office of Water, Washington, DC.

EPA (2001b), "Nutrient Criteria Technical Guidance Manual - Estuarine and Coastal Marine Waters," Tech. Rep. EPA- 822/B-01/003, Office of Research and Development and Office of Water, Washington, DC.

Failing, L., Horn, G., and Higgins, P. (2004), "Using expert judgment and stakeholder values to evaluate adaptive management options," *Ecology and Society*, 9, 13.

Farmani, R., Henriksen, H. J., and Savic, D. (2009), "An Evolutionary Bayesian Belief Network Methodology for Optimum Management of Groundwater Contamination," *Environmental Modelling & Software*, 24, 303–310.

Fawcett, T. (2006), "An introduction to ROC analysis," *Pattern Recognition Letters*, 27, 861–874.

Gameiro, C., Zwolinski, J., and Brotas, V. (2011), "Light control on phytoplankton production in a shallow and turbid estuarine system," *Hydrobiologia*, 669, 249–263.

Gelman, A. (2008), "Scaling Regression Inputs by Dividing by Two Standard Deviations," *Statistics in Medicine*, 27, 2865–2873.

Gelman, A. and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New York.

Grayson, R., Doolan, J., and Blake, T. (1994), "Application of AEAM (adaptive environmental assessment and management) to water quality in the Latrobe River catchment," *Journal of Environmental Management*, 41, 245–258.

Green, M. B. and Wang, D. (2008), "Watershed flow paths and stream water nitrogen-to-phosphorus ratios under simulated precipitation regimes," *Water Resources Research*, 44.

Hall, N. S. and Paerl, H. W. (2011), "Vertical migration patterns of phytoflagellates in relation to light and nutrient availability in a shallow microtidal estuary," *Marine Ecology Progress Series*, 425, 1–19.

Hall, N. S., Litaker, R. W., Fensin, E., Adolf, J. E., Bowers, H. A., Place, A. R., and Paerl, H. W. (2008), "Environmental factors contributing to the development and demise of a toxic dinoflagellate (Karlodinium veneficum) bloom in a shallow, eutrophic, lagoonal estuary," *Estuaries and Coasts*, 31, 402–418.

Hall, N. S., Paerl, H. W., Peierls, B. L., Whipple, A. C., and Rossignol, K. L. (2012), "Effects of climatic variability on phytoplankton community structure and bloom development in the eutrophic, microtidal, New River Estuary, North Carolina, USA," *Estuarine, Coastal and Shelf Science*.

Harper, L. A., Sharpe, R. R., and Parkin, T. B. (2000), "Gaseous nitrogen emissions from anaerobic swine lagoons: Ammonia, nitrous oxide, and dinitrogen gas," *Journal of Environmental Quality*, 29, 1356–1365.

Harper, L. A., Sharpe, R. R., Parkin, T. B., De Visscher, A., Van Cleemput, O., and Byers, F. M. (2004), "Nitrogen cycling through swine production systems," *Journal of environmental quality*, 33, 1189–1201.

Hatfield, J., Brumm, M., and Melvin, S. (1998), "Swine manure management," *Agricultural uses of municipal, animal, and industrial byproducts*, 44, 78–90.

Heckerman, D. (2008), *A tutorial on learning with Bayesian networks*, Springer.

Hodgkiss, I. and Ho, K. (1997), "Are changes in N:P ratios in coastal waters the key to increased red tide blooms?" in *Asia-Pacific Conference on Science and Management of Coastal Environment*, pp. 141–147, Springer.

Huang, G. H. and Xia, J. (2001), "Barriers to sustainable water-quality management," *Journal of Environmental Management*, 61, 1–23.

Hunt, P., Matheny, T., Ro, K., Vanotti, M., and Ducey, T. (2010), "Denitrification in anaerobic lagoons used to treat swine wastewater," *Journal of environmental quality*, 39, 1821–1828.

Israel, D. W., Showers, W. J., Fountain, M., and Fountain, J. (2005), "Nitrate movement in shallow ground water from swine-lagoon-effluent spray fields managed under current application regulations," *Journal of environmental quality*, 34, 1828–1842.

Jacobs, K., Garfin, G., and Lenart, M. (2005), "More than Just Talk: Connecting Science and Decisionmaking," *Environment: Science and Policy for Sustainable Development*, 47, 6–21.

Jensen, F. V. and Nielsen, T. D. (2007), *Bayesian networks and decision graphs*, Springer.

Johnson, S., Fielding, F., Hamilton, G., and Mengersen, K. (2010), "An integrated Bayesian network approach to Lyngbya majuscula bloom initiation," *Marine Environmental Research*, 69, 27–37.

Jørgensen, B. B. and Richardson, K. (1996), *Eutrophication in Coastal Marine Ecosystems*, vol. 52, American Geophysical Union.

Kaushal, S. S., Pace, M. L., Groffman, P. M., Band, L. E., Belt, K. T., Meyer, P. M., and Welty, C. (2010), "Land use and climate variability amplify contaminant pulses," *EOS, Transactions American Geophysical Union*, 91, 221–222.

Keller, A. A. (1989), "Modeling the effects of temperature, light, and nutrients on primary productivity: An empirical and a mechanistic approach compared." *Limnology and Oceanography*, 34, 82–95.

Kiddon, J. A., Paul, J. F., Buffum, H. W., Strobel, C. S., Hale, S. S., Cobb, D., and Brown, B. S. (2003), "Ecological Condition of US Mid-Atlantic Estuaries, 1997–1998," *Marine Pollution Bulletin*, 46, 1224–1244.

Koller, D. and Pfeffer, A. (1997), "Object-oriented Bayesian networks," in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pp. 302–313, Morgan Kaufmann Publishers Inc.

Korfmacher, K. S. (1998), "Water quality modeling for environmental management: lessons from the policy sciences," *Policy Sciences*, 31, 35–54.

Koseff, J. R., Holen, J. K., Monismith, S. G., and Cloern, J. E. (1993), "Coupled effects of vertical mixing and benthic grazing on phytoplankton populations in shallow, turbid estuaries," *Journal of Marine Research*, 51, 843–868.

Kragt, M., Newham, L. T., Bennett, J., and Jakeman, A. J. (2011), "An Integrated Approach to Linking Economic Valuation and Catchment Modelling," *Environmental Modelling & Software*, 26, 92–102.

Kruschke, J. (2010), *Doing Bayesian data analysis: A tutorial introduction with R*, Academic Press.

Laws, E. A. (2000), *Aquatic pollution: An introductory text*, John Wiley & Sons.

Lepisto, L., Pietilainen, O.-P., Rissanen, J., and Vuoristo, H. (2002), "Finnish draft for typology of lakes and rivers," *Typology and Ecological Classification of Lakes and Rivers*, p. 42.

Lewis Jr, W. M., Wurtsbaugh, W. A., and Paerl, H. W. (2011), "Rationale for control of anthropogenic nitrogen and phosphorus to reduce eutrophication of inland waters," *Environmental science & technology*, 45, 10300–10305.

Liu, H., Hussain, F., Tan, C. L., and Dash, M. (2002), "Discretization: An enabling technique," *Data mining and knowledge discovery*, 6, 393–423.

Lloret, J., Marín, A., and Marín-Guirao, L. (2008), "Is coastal lagoon eutrophication likely to be aggravated by global climate change?" *Estuarine, Coastal and Shelf Science*, 78, 403–412.

Madsen, A. L., Lang, M., Kjærulff, U. B., and Jensen, F. (2003), "The Hugin tool for learning Bayesian networks," in *Symbolic and quantitative approaches to reasoning with uncertainty*, pp. 594–605, Springer.

Madsen, A. L., Jensen, F., Kjaerulff, U. B., and Lang, M. (2005), "The Hugin Tool for Probabilistic Graphical Models," *International Journal on Artificial Intelligence Tools*, 14, 507–543.

Maguire, L. A. (2003), "Interplay of science and stakeholder values in Neuse River total maximum daily load process," *Journal of Water Resources Planning and Management*, 129, 261–270.

Malekmohammadi, B., Kerachian, R., and Zahraie, B. (2009), "Developing Monthly Operating Rules for a Cascade System of Reservoirs: Application of Bayesian Networks," *Environmental Modelling & Software*, 24, 1420–1432.

Mallin, M. A., McIver, M. R., Wells, H. A., Parsons, D. C., and Johnson, V. L. (2005), "Reversal of eutrophication following sewage treatment upgrades in the New River Estuary, North Carolina," *Estuaries*, 28, 750–760.

Malve, O. and Qian, S. S. (2006), "Estimating Nutrients and Chlorophyll a Relationships in Finnish Lakes," *Environmental Science & Technology*, 40, 7848–7853.

Marcot, B. G. (2012), "Metrics for Evaluating Performance and Uncertainty of Bayesian Network Models," *Ecological Modelling*, 230, 50–62.

Marcot, B. G., Steventon, J. D., Sutherland, G. D., and McCann, R. K. (2006), "Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation," *Canadian Journal of Forest Research*, 36, 3063–3074.

McCann, R. K., Marcot, B. G., and Ellis, R. (2006), "Bayesian Belief Networks: Applications in Ecology and Natural Resource Management," *Canadian Journal of Forest Research*, 36, 3053–3062.

McDaniels, T. L. and Gregory, R. (2004), "Learning as an objective within a structured risk management decision process," *Environmental Science & Technology*, 38, 1921–1926.

McKee, L., Gilbreath, A., Beagle, J., Gluchowski, D., Hunt, J., and Sutula, M. (2011), "Numeric Nutrient Endpoint Development for San Francisco Bay Estuary: Literature Review and Data Gaps Analysis," *Southern California Coastal Water Research Project, Research Report*, 644.

McMahon, G., Alexander, R. B., and Qian, S. (2003), "Support of total maximum daily load programs using spatially referenced regression models," *Journal of Water Resources Planning and Management*, 129, 315–329.

Murphy, R. R., Kemp, W. M., and Ball, W. P. (2011), "Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading," *Estuaries and Coasts*, 34, 1293–1309.

Nagarajan, R., Scutari, M., and Lebre, S. (2013), *Bayesian Networks in R with Applications in Systems Biology*, Springer, New York, ISBN 978-1461464457.

Najjar, R. G., Walker, H. A., Anderson, P. J., Barron, E. J., Bord, R. J., Gibson, J. R., Kennedy, V. S., Knight, C. G., Megonigal, J. P., O'Connor, R. E., et al. (2000), "The potential impacts of climate change on the mid-Atlantic coastal region," *Climate Research*, 14, 219–233.

NCDENR (2011), "North Carolina Department of Environment and Natural Resources Division of Water Quality," Accessed: 2011.

NCDENR-DWQ (2007), "Surface Waters and Wetlands Standards," .

NCDMF (1993), "Description of North Carolinas Coastal Fishery Resources, 1972-1991," Tech. rep.

Neff, R., Chang, H., Knight, C. G., Najjar, R. G., Yarnal, B., and Walker, H. A. (2000), "Impact of climate variation and change on Mid-Atlantic Region hydrology and water resources," *Climate Research*, 14, 207–218.

Nielsen, T. D. and Jensen, F. V. (2009), *Bayesian Networks and Decision Graphs*, Springer.

Nixon, S. W. (1995), "Coastal Marine Eutrophication: A Definition, Social Causes, and Future Concerns," *Ophelia*, 41, 199–219.

NOAA (1996), "NOAA's Estuarine Eutrophication Survey. Volume 1: South Atlantic Region," Tech. rep.

Nojavan A., F., Qian, S. S., Paerl, H. W., Reckhow, K. H., and Albright, E. A. (2014), "A study of Anthropogenic and Climatic Disturbance of the New River Estuary Using a Bayesian Belief Network," *Marine Pollution Bulletin*.

Ott, W. (1995), *Environmental Statistics and Data Analysis*, Lewis Publishers, Boca Raton.

Paerl, H. W. (1988), "Nuisance phytoplankton blooms in coastal, estuarine, and inland waters," *Limnology and Oceanography*, pp. 823–847.

Paerl, H. W. (1997), "Coastal eutrophication and harmful algal blooms: Importance of atmospheric deposition and groundwater as" new" nitrogen and other nutrient sources," *Limnology and oceanography*, 42, 1154–1165.

Paerl, H. W. (2009), "Controlling eutrophication along the freshwater-marine continuum: dual nutrient (N and P) reductions are essential," *Estuaries and Coasts*, 32, 593–601.

Paerl, H. W. and Scott, J. T. (2010), "Throwing fuel on the fire: Synergistic effects of excessive nitrogen inputs and global warming on harmful algal blooms," *Environmental Science & Technology*, 44, 7756–7758.

Paerl, H. W., Pinckney, J. L., Fear, J. M., and Peierls, B. L. (1998), "Ecosystem responses to internal and watershed organic matter loading: consequences for hypoxia in the eutrophying Neuse River Estuary, North Carolina, USA," *Marine Ecology Progress Series*, 166, 17.

Pearl, J. (1982), "Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach," in *AAAI*, pp. 133–136.

Peierls, B. L., Hall, N. S., and Paerl, H. W. (2012), "Non-monotonic responses of phytoplankton biomass accumulation to hydrologic variability: a comparison of two coastal plain North Carolina estuaries," *Estuaries and Coasts*, 35, 1376–1392.

Plummer, M. et al. (2003), "JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pp. 20–22.

Pollino, C. A., Woodberry, O., Nicholson, A., Korb, K., and Hart, B. T. (2007), "Parameterisation and Evaluation of a Bayesian Network for Use in an Ecological Risk Assessment," *Environmental Modelling & Software*, 22, 1140–1152.

Qian, S. (2010), *Environmental and Ecological Statistics with R*, Chapman and Hall/CRC Press.

Qian, S. S. and Cuffney, T. F. (2012), "To threshold or not to threshold? That's the question," *Ecological Indicators*, 15, 1–9.

Qian, S. S., Stow, C. A., and Borsuk, M. E. (2003), "On monte carlo methods for Bayesian inference," *Ecological Modelling*, 159, 269–277.

Qian, S. S., Reckhow, K. H., Zhai, J., and McMahon, G. (2005), "Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach," *Water Resources Research*, 41.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rabalais, N. N., Turner, R. E., and Scavia, D. (2002), "Beyond Science into Policy: Gulf of Mexico Hypoxia and the Mississippi River," *BioScience*, 52, 129–142.

Rabalais, N. N., Turner, R. E., Díaz, R. J., and Justić, D. (2009), "Global change and eutrophication of coastal waters," *ICES Journal of Marine Science: Journal du Conseil*, 66, 1528–1537.

Ragas, A. M., Etienne, R. S., Willemsen, F. H., and Van De Meent, D. (1999), "Assessing model uncertainty for environmental decision making: A case study of the coherence of independently derived environmental quality objectives for air and water," *Environmental Toxicology and Chemistry*, 18, 1856–1867.

Reckhow, K. H. (1994), "Water quality simulation modeling and uncertainty analysis for risk assessment and decision making," *Ecological Modelling*, 72, 1–20.

Reckhow, K. H. (1999), "Water Quality Prediction and Probability Network Models," *Canadian Journal of Fisheries and Aquatic Sciences*, 56, 1150–1158.

Reddi, L. N. (2005), *Animal waste containment in lagoons*, no. 105, ASCE Publications.

Ribaudo, M., Gollehon, N., and Agapoff, J. (2003), "Land application of manure by animal feeding operations: Is more land needed?" *Journal of Soil and Water Conservation*, 58, 30–38.

Ro, K. S., Szogi, A. A., Vanotti, M. B., and Stone, K. C. (2008), "Process model for ammonia volatilization from anaerobic swine lagoons incorporating varying wind speeds and gas bubbling," *Transactions of the ASAE (American Society of Agricultural Engineers)*, 51, 259.

Rogers, C. E. and McCarty, J. P. (2000), "Climate change and ecosystems of the Mid-Atlantic Region," *Climate Research*, 14, 235–244.

RTI, I. (2011), "Defense Coastal/Estuarine Research Program (DCERP) Final Research Report Prepared for Strategic Environmental Research and Development Program, Department of Defense," Tech. rep.

RTI, I. (2013), "Defense Coastal/Estuarine Research Program (DCERP1) Final Research Report Prepared for Strategic Environmental Research and Development Program, Department of Defense," Tech. rep.

Ryther, J. H. and Dunstan, W. M. (1971), "Nitrogen, phosphorus, and eutrophication in the coastal marine environment," *Science*, 171, 1008–1013.

Scavia, D., Field, J. C., Boesch, D. F., Buddemeier, R. W., Burkett, V., Cayan, D. R., Fogarty, M., Harwell, M. A., Howarth, R. W., Mason, C., et al. (2002), "Climate change impacts on US coastal and marine ecosystems," *Estuaries*, 25, 149–164.

Scutari, M. (2010), "Learning Bayesian Networks with the bnlearn R Package," *Journal of Statistical Software*, 35, 1–22.

Sheldon, J. E. and Alber, M. (2011), "Recommended indicators of estuarine water quality for Georgia," in *Proc. 2011 Georgia Water Resources Conference*.

Shindler, B. and Cheek, K. A. (1999), "Integrating citizens in adaptive management: A propositional analysis," *Conservation Ecology*, 3, 9.

Smith, C. and Bosch, O. (2004), "Integrating disparate knowledge to improve natural resource management," in *Proceedings of the 13th International Soil Conservation Organisation Conference*.

Smith, J. E. (1993), "Moment methods for decision analysis," *Management Science*, 39, 340–358.

Smith, V. H. (2003), "Eutrophication of freshwater and coastal marine ecosystems a global problem," *Environmental Science and Pollution Research*, 10, 126–139.

Smith, V. H. (2006), "Responses of Estuarine and Coastal Marine Phytoplankton to Nitrogen and Phosphorus Enrichment," *Limnology and Oceanography*, 51, 377–384.

Spiegelhalter, D. J. and Knill-Jones, R. P. (1984), "Statistical and Knowledge-based Approaches to Clinical Decision-support Systems, with an Application in Gastroenterology," *Journal of the Royal Statistical Society. Series A (General)*, pp. 35–77.

Stow, C. A., Roessler, C., Borsuk, M. E., Bowen, J. D., and Reckhow, K. H. (2003), "Comparison of Estuarine Water Quality Models for Total Maximum Daily Load Development in Neuse River Estuary," *Journal of Water Resources Planning and Management*, 129, 307–314.

Strobl, R. O. and Robillard, P. D. (2008), "Network design for water quality monitoring of surface freshwaters: A review," *Journal of Environmental Management*, 87, 639–648.

Tomas, C. R., Peterson, J., and Tatters, A. O. (2007), "Harmful algal species from Wilson Bay, New River, North Carolina: Composition, nutrient bioassay and HPLC pigment analysis," *Water Resources Research Institute of the University of North Carolina*.

USDA (1992), "Census of Agriculture," .

USDA (2007), "2007 Census of Agriculture," .

Uusitalo, L. (2007), "Advantages and challenges of Bayesian networks in environmental modelling," *Ecological Modelling*, 203, 312–318.

Vargo, G. A. (2009), "A brief summary of the physiology and ecology of Karenia brevis Davis (G. Hansen and Moestrup comb. nov.) red tides on the West Florida Shelf and of hypotheses posed for their initiation, growth, maintenance, and termination," *Harmful Algae*, 8, 573–584.

Varis, O. (1997), "Bayesian Decision Analysis for Environmental and Resource Management," *Environmental Modelling & Software*, 12, 177–185.

Varis, O. and Kuikka, S. (1997), "Joint Use of Multiple Environmental Assessment Models by a Bayesian Meta-model: the Baltic Salmon Case," *Ecological Modelling*, 102, 341–351.

Walters, C. (1997), "Challenges in adaptive management of riparian and coastal ecosystems," *Conservation Ecology*, 1, 1.

Walters, C. J. and Hilborn, R. (1978), "Ecological optimization and adaptive management," *Annual review of Ecology and Systematics*, pp. 157–188.

Weisberg, S. (2005), *Applied Linear Regression*, Wiley.

Wetz, M. S. and Paerl, H. W. (2008), "Estuarine phytoplankton responses to hurricanes and tropical storms with different characteristics (trajectory, rainfall, winds)," *Estuaries and Coasts*, 31, 419–429.

Whalen, S. and DeBerardinis, J. (2007), "Nitrogen mass balance in fields irrigated with liquid swine waste," *Nutrient Cycling in Agroecosystems*, 78, 37–50.

Williams, C. M., Murray, B. C., Van Houtven, G. L., Deerhake, M. E., Dodd, R. C., Lowry, M. M. I., Yao, C., Miles, A. M., Bowman, E. J., Bruhn, M. C., et al. (2003), "Benefits of Adopting Environmentally Superior Swine Waste Management Technologies in North Carolina: An Environmental and Economic Assessment," .

Zhang, H., Litaker, W., Vandersea, M. W., Tester, P., and Lin, S. (2008), "Geographic distribution of Karlodinium veneficum in the US east coast as detected by ITS-ferredoxin real-time PCR assay," *Journal of Plankton Research*, 30, 905–922.

Biography

Farnaz Nojavan Asghari is a Ph.D. candidate in Environmental Science & Policy at the Nicholas School of the Environment, Duke University. She was born in 1983 in Oroumieh, Iran. Prior to joining Duke, she received the following degrees:

University of Florida, Gainesville, FL *August 2007-December 2008*

M.Sc. in Industrial & System Engineering (Operations Research)

University of Tehran, Tehran, Iran *August 2005-August 2007*

M.Sc. in Socio-economic System Engineering

Sharif University of Technology, Tehran, Iran *August 2001-August 2005*

B.Sc. in Industrial Engineering - System Analysis