Performance-based modeling of spatial and temporal variability of treated wastewater quality for improved nutrient management

by

Bihu Suchetana

B.E., IIEST, Shibpur, India, 2010

M.E., Indian Institute of Science, Bangalore, India, 2013

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Civil, Environmental and Architectural Engineering

2018

This thesis entitled:
Performance-based modeling of spatial and temporal variability of treated wastewater
quality for improved nutrient management
written by Bihu Suchetana
has been approved for the Department of Civil, Environmental and Architectural
Engineering

_____

Prof. Balaji Rajagopalan

_____

Prof. JoAnn Silverstein

_____

Prof. Joseph Kasprzyk

_____

Prof. Sherri Cook

_____

Prof. Sybil Sharvelle

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both
the content and the form meet acceptable presentation standards of scholarly work in the
above mentioned discipline.

Suchetana, Bihu (Ph.D., Civil Engineering)

Performance-based modeling of spatial and temporal variability of treated wastewater quality
    for improved nutrient management

Thesis directed by Prof. Balaji Rajagopalan

The United States Environmental Protection Agency (U.S.EPA) has identified nutrient pollution as the leading cause of use impairment in U.S. waters. Consequently, for improved nutrient management, the U.S.EPA recommends ensuring point sources comply with their permit limits. However, there is limited understanding of compliance with discharge limits, primarily due to great spatial and temporal variability in effluent nutrient concentrations as well as discharge permit limits. Further, the regulatory climate for nutrients is rapidly changing in several states of the country, with the adoption of more stringent discharge limits for wastewater treatment plants. This research presents a performance-based statistical modeling approach to understand the spatial and temporal variability of nutrient compliance (specifically nitrogen species) with changing regulations, in treated wastewaters of the United States.

A hierarchical model is built using Generalized Linear Models (GLMs) and Kriging, and effluent ammonia concentrations from Discharge Monthly Report (DMR) data from more than 100 wastewater treatment plants across the US. Compliance with current ammonia permit discharge limits is seen to be determined by the flow rate and its compliance history. The probability, frequency and magnitude of risk of non-compliance with ammonia discharge limits is modeled using GLMs and Extreme Value Theory (EVT). The probability, frequency and magnitude of risk of non-compliance with ammonia discharge limits is found to be determined by both the flow rate and compliance history, in addition to the fractional use of design capacity. Wastewater treatment plant compliance with decreasing ammonia discharge limits is assessed using a regression trees. Once again, the compliance history and the flow

rate are seen to affect compliance with both existing and lowered discharge limits. Some states, such as Colorado, are considering broader regulations, for all nitrogen species, by regulating levels of Total Inorganic Nitrogen (TIN) in effluent wastewaters. A Hidden Markov Model (HMM) and multinomial logistic regression based modeling framework is presented to predict TIN concentrations in treated wastewaters, using data from an operating wastewater treatment plant in Colorado, US. Effluent TIN concentrations are found to be a function of climate variables (such as minimum air temperature and precipitation), seasonality, effluent ammonia concentrations and effluent TIN concentrations in the previous month.

The performance-based models presented in this research can be beneficial to several stakeholders; they can be useful for predictive purposes or reliability analysis on both a single treatment plant or multi-plant level. While they can help individual plant operators ensure compliance with changing nutrient regulations, monitoring and enforcement efforts can now be better channelized towards frequent and egregious violators. Additionally, for various proposed (lowered) discharge limits, these models can be implemented to delineate reliable sources of demand and supply for a point source-to-point source nutrient credit trading scheme.

Dedication

To my family.

Acknowledgements

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

1.1   Background

The history of water pollution control in the United States dates back to 1899, when the Rivers and Harbors Act was passed to prevent unauthorized obstructions in the navigable waters of the country. While this legislation primarily focused on navigation, the 1948 Federal Water Pollution Control Act (FWPCA) marked the beginning of the federal government's involvement in water pollution for public health protection (P.L. 80-845, 62 Stat. 1155). The FWPCA Amendments of 1972 allowed the federal government to direct and define water pollution control programs (P.L. 92-500, 86 Stat. 816). The *"fishable, swimmable"* goal envisaged by this legislation, tried to set an interim target of achieving water quality to protect and propagate fish, shellfish, and wildlife and provide for recreation in and on the water. The FWPCA Amendments of 1972 required municipal facilities to meet secondary treatment standards by 1977.Approximately $50 billion in federal funding were spent through the Construction Grants Program to build wastewater treatment plants with secondary treatment facilities, which included removal of Biochemical Oxygen Demand (BOD) and Total Suspended Solids (TSS). The Clean Water Act's (P.L. 95-217) legislative history in 1977 envisioned nutrients as one of the constituents to be removed by advanced treatment, beyond secondary treatment requirements. This goal, however, could not be met due to federal funding shortfalls. The FWPCA legislation also resulted in the establishment of the National Permit Discharge Elimination System (NPDES), which requires all point sources

discharging pollutants into waters of the United States, to obtain a permit (U.S.EPA, 2010).

The permit writing process involves deriving necessary effluent limits to meet certain water quality standards (Section 301(b)(1)(C) of the Clean Water Act, 1977). Water quality standards are developed by individual states and revised from time to time, based on designated use. States are at liberty to develop water quality standards which are more stringent than U.S.EPA guidelines. These standards comprise of (numeric or narrative) water quality criteria to support designated use of each receiving waterbody. If a waterbody has multiple use designations, the convention is to develop criteria which support the most sensitive use (U.S.EPA, 2010). The U.S.EPA's recommended nutrient criteria are *ecoregional*, as opposed to its nationally applicable standards for other common pollutants like BOD and TSS. This allows for refined, localized nutrient criteria better suited to protect the quality of the receiving waters. Ammonia criteria, specifically, also varies seasonally according to pH, temperature, presence of salmonid species, and presence or absence of early life stages of fish. Effluent limits based on such nutrient criteria which are generally water quality based (known as Water Quality Based Effluent Limits or WQBELs), thus vary temporally and spatially.

In 2012, the U.S.EPA denied a petition by the National Resources Defense Council (NRDC) and ten other organizations to establish new technology-based nutrient limits for nutrients (Shapiro, 2012). They observed that there are several obstacles to developing a uniform technology-based nutrient standard, applicable to municipal wastewater treatment plants nationwide. This was mainly due to requirements of significant technological upgrades to meet these limits, which would be difficult due to financial and logistic (e.g. insufficient land) constraints. This decision upheld the U.S. EPA's long-held view of addressing nutrient discharges through water-quality based permitting. Yet, the nation's surface water quality has been deteriorating, with up to 78% rivers, 76% lakes/reservoirs and 100% of estuaries in certain states impaired for designated uses (https://www.epa.gov/nutrient-policy-data/waters-assessed-impaired-due-nutrient-related-causes). This has prompted states like

Colorado and Utah to consider more stringent, technology-based nutrient limits (CDPHE, 2012, UDEQ, 2014). There has also been increased interest in exploring incentive-based programs such as nutrient credit trading schemes, as a flexible and cost effective approach to control nutrient pollution in surface waters (U.S.EPA, 2013, CDPHE, 2012, VDEE, 2016, CDEEP, 2015). Thus, the nutrient regulatory climate is steadily changing, to ensure better effluent quality from wastewater treatment plants, and ultimately preserve and protect the nation's surface water.

1.2   Motivation

Traditionally, wastewater treatment process modeling has been largely mechanistic in nature (Liwarska-Bizukojc and Biernacki, 2010, Phillips et al., 2009, Vanhooren et al., 2003, Filipe et al., 2001, Fillos et al., 2000, Melcer, 1999). Mass balance or rate-based simulation models like AquaSim (Reichert, 1994) or BioWin (http://envirosim.com/ products/biowin), are dependent on detailed inputs of influent wastewater quality, as well as parameter estimates specific to a system. Biological wastewater treatment systems are dynamic in nature, and they are influenced by several external factors not accounted for by mechanistic models, such as geographical and climatological conditions, fluctuations in environmental conditions and in-plant operational variations. Performance-based modeling, on the other hand, is based on the requirements of certain measurable performance criteria, such as specific levels of ammonia or fecal coliform in the effluent. Statistical or probabilistic approaches are commonly used for performance-based modeling, enabling estimation of expected effluent quality and its variations. Such models not only provide a consistent basis for analysis of effluent quality, but also allow estimation of uncertainty and reliability.

Niku et al. (1979) made one of the first attempts towards probabilistic prediction of Biochemical Oxygen Demand (BOD) and Suspended Solids (SS) concentrations. They developed estimates of coefficient of reliability (COR), which relates mean pollutant values to treatment standards on a probability basis , using effluent data from 37 wastewater plants.

The predictions from this probabilistic model, however, did not account for the effect of non-randomness, trends and cyclical variations in effluent, making them suitable for long term predictions only. In a follow-up study, Niku and Schroeder (1981b), studied stability of activated sludge processes based on statistical measures. They found poor correlations between arithmetic mean annual flow and annual mean and standard deviation values of BOD and SS in a sample of 43 activated sludge treatment plants. Niku and Schroeder (1981a) used linear regression analysis to study the effect of operations on individual treatment plant performance. Using daily operational data from 21 wastewater treatment plants across the country, they found that a single or consistent group of factors(including influent pollutant loads, biological and operational variables) could explain individual plant variability in BOD and SS removal. Such results could be explianed by the relatively narrow range of operating conditions at individual plants. More recently, Generalized Linear Models (GLMs) have been used to assess variability in wastewater treatment performance. Weirich et al. (2011) analyzed monthly effluent concentrations of BOD, TSS, ammonia and fecal coliform from more than 200 wastewater treatment plants in the US. They used average monthly flow rate and design capacity utilization covariates; process type was explicitly ignored to quantify treatment performance as related to decentralization. They found that smaller plants are more likely to violate their discharge permits, especially for ammonia. Small treatment facilities (<4000 m^3/d), operating near or over their design flow rate had higher occurrences of permit violations, with pollutant concentrations well above the discharge limit. In a case study of the St. Vrain Creek basin, Colorado, US, using GLM-based simulations, they also found that decentralized networks were more stable and resilient when discharge limits are lax, and for stringent discharge standards, centralized networks tended to be most resilient to process upsets (Weirich et al., 2015b). They also developed a statistical model to quantify resilience and stability of wastewater treatment plants, as a function of earlier performance (as opposed to independent monthly events) and average flow rate and design capacity utilization (Weirich et al., 2015a).

1.3   Gaps in existing literature

While Niku et al. (1979)'s and Niku and Schroeder (1981a,b)'s work initiated probabilistic methods in wastewater treatment plant performance, (Weirich et al., 2011, 2015b,a) made the very first attempt to statistically quantify level of compliance, resilience and stability of wastewater treatment plants on a national scale. However, there were several other factors affecting nutrient compliance that were not considered in these studies. Non-compliance with discharge limits, which is among the leading causes of deterioration of surface water quality, also poses significant financial burden on treatment plants (in the form of enforcement actions). The probability of non-compliance or violation of discharge limits has been studied by Weirich et al. (2015a). However, a complete understanding of risk of non-compliance requires the additional knowledge of the frequency and magnitude of permit violation, which has never been researched in the past. With more stringent nutrient regulations in place, there has been very little research to assess the ability of wastewater treatment plants across the country to comply with lowered discharge limits, with their existing treatment technology and infrastructure. Even with nutrient regulations shifting focus from ammonia to TIN, there is no literature on modeling effluent TIN concentrations in wastewater treatment plants. To address these gaps in existing research, performance-based statistical models are built to better assess compliance with changing nutrient regulations. It should be noted that due to insufficient effluent phosphorus concentration data, modeling of effluent nutrient concentrations is performed using nitrogen data only (ammonia-N, nitrate-N and nitrite-N).

1.4   Dissertation Outline

For improved nutrient management, it is imperative for plant operators as well as regulatory, monitoring and enforcement agencies to develop a thorough understanding of the factors that affect compliance (or the lack there-of) with changing regulations. As discussed in the previous section, various gaps in existing literature, as well as several emergent

questions in evaluating compliance, motivated this study. A comprehensive study of the various aspects of regulatory compliance/ non-compliance has been performed by exploring the following four objectives in this dissertation:

- to model the variability in compliance with current ammonia discharge limits in the US

- to model various attributes (the probability, frequency and magnitude) of risk of non-compliance with ammonia discharge limits

- to model wastewater treatment facility compliance with decreasing ammonia discharge limits

- to model Total Inorganic Nitrogen (TIN) concentrations in treated wastewaters

In Chapters 2-5 of this dissertation, performance based modeling approaches have been developed to meet each of these objectives. Chapter 2 describes a hierarchical modeling framework, using Generalized Linear Models (GLM) and Kriging, which models the spatial and temporal variability of wastewater treatment plant compliance with existing ammonia discharge limits in the US. In Chapter 3, the probability, frequency and magnitude of ammonia permit limit violations in the US has been modeled using GLMs and the Generalized Pareto Distribution (GPD). Chapter 4 deals with a regression tree based approach to assess compliance with current and future (lowered) ammonia discharge limits in the US. Chapter 5 details a Hidden Markov Model and methodology to model TIN concentrations in effluent wastewaters of an operating facility in Colorado, US. Finally, conclusions and future work are presented in Chapter 6. Using data sets of varying spatial and temporal resolutions, this dissertation offers the first comprehensive, performance-based approach towards understanding nutrient compliance, which can provide important operational, monitoring, regulatory and enforcement strategies to concerned stakeholders.

Chapter 2

A hierarchical modeling approach to evaluate spatial and temporal variability of wastewater treatment compliance with ammonia discharge limits in the US

Abstract

A hierarchical modeling approach is used to diagnose the variability of US wastewater treatment facility compliance with existing ammonia discharge permit limits. The first level of hierarchy models monthly variability in effluent ammonia concentration, over a four-year period, using a Generalized Linear Model (GLM), which relates compliance to plant characteristics such as flow rate, capacity utilization, seasonality and compliance history. The residuals from this model are subjected to a spatial analysis using Kriging in the second level of hierarchy, to model spatial co-variability. This hierarchical approach used Discharge Monthly Report (DMR) data for the period 2004 - 2008 from a sample of 106 municipal plants. The first level of hierarchy captures most of the total variance in compliance data for

ammonia. Seasonality does not impact effluent ammonia concentrations, for which permit limits are typically seasonal. Level of compliance with ammonia discharge limits is significantly affected by the effluent concentration in the previous month. The second level of hierarchy identifies geographic regions of anomalous compliance and improves the model's predictive performance for a subset of plants with discharges closer to permit limits.

2.1   Introduction

Historically the goal of design and operation of wastewater treatment and disposal in the US has been protecting public health and surface water quality through National Pollutant Discharge Elimination System (NPDES) permits, which treat each facility as a singular entity. Since 1972, the impacts of expanded secondary wastewater treatment, resulting from national and state regulations have been significant, in terms of increased population served and water quality improvements (Fairfax and Hamilton, 2000). In addition, new knowledge of the environmental impacts of various wastewater constituents has driven understanding of biological, chemical, and physical processes and related advances in process technology. Incorporation of these factors into computational models to simulate treatment plant performance has been a valuable tool for optimizing individual plant design and operation.

However, the ability to evaluate compliance with NPDES limits on a regional or national scale is of increased interest for understanding the actual impacts of wastewater discharges on water quality and for assessing the effects of changes in discharge regulations. Of particular importance is adoption of more stringent discharge limits for nutrients, which has increased in pace since 1998 (Figure 2.1). In 1998, 39 States and the District of Columbia had no numeric water quality criteria for nutrients that covered all of a particular type of water body, e.g., lakes, rivers/streams, estuaries and coastal waters, and only three States had numeric criteria applied to all of one or two water types; only one State, Hawaii, had statewide numeric nutrient criteria for all water types. The EPA expects an increasing trend in statewide regulation of nutrients by water body type, anticipating that by the end of 2017,
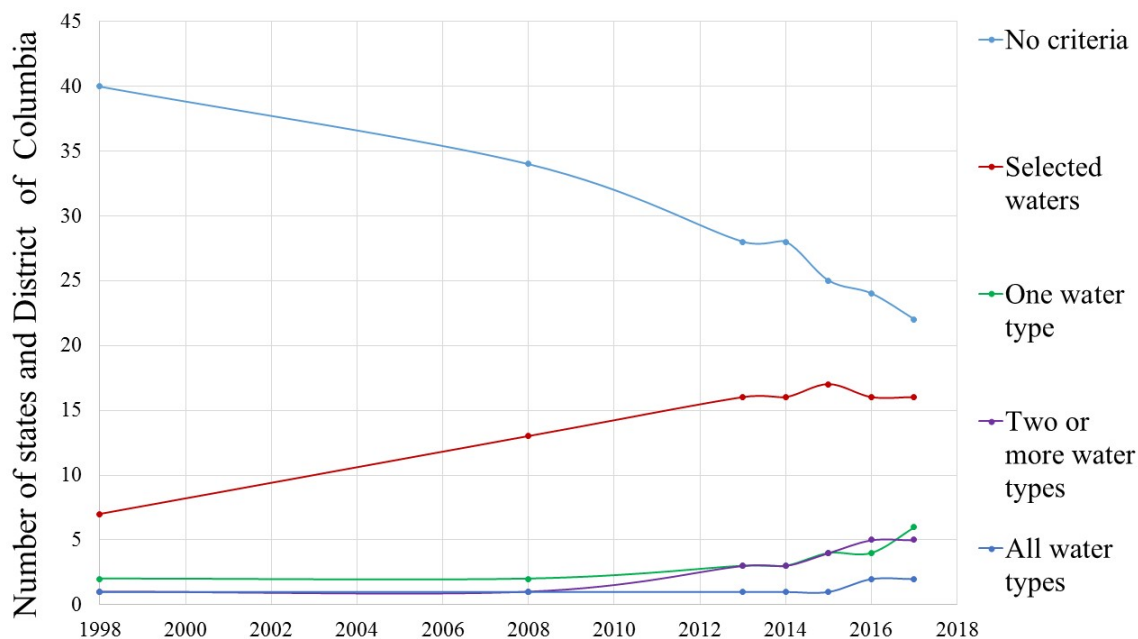
Figure 2.1: Progress in specifying water quality criteria for nitrogen species and/or phosphorus by 50 States and the District of Columbia, 1998 - 2017

11 states will have numeric criteria for nitrogen species and/or phosphorus for all of one or more water types, and at least two States will regulate both nutrients for all water types.

In 2012, the USEPA estimated costs and benefits of uniform national regulations for nitrogen and phosphorus discharges and reported that capital cost of added technology (retrofitting, new installation) could exceed $50 billion, ($5 to $63/ capita/ year) with annual operating costs over $5 billion. The burden on small utilities would be even greater given the diseconomies of scale for small systems. Furthermore, the cost of developing and obtaining approval for the new regulation would exceed $10 million and take at least five years (Shapiro, 2012). In states considering new water quality criteria and associated nutrient discharge regulations, the local stakes are also significant. Certainly, water quality degradation is an important result of non-compliance, and enforcement actions also carry significant costs to regulators, utilities and ratepayers. A few researchers have studied the potential for compliance with emerging limits for nutrients. Daigger et al. (2014) used data from 30 mechanical wastewater treatment plants and 23 lagoon facilities in Utah to pre-

dict compliance with more stringent discharge standards for nitrogen and phosphorus, using a spreadsheet-based process simulator. They found that plant capacity was an important determinant of the ability of facilities to meet higher levels of nutrient removal, with the greatest performance and cost impacts on small plants.

A recent approach to evaluate the aggregate performance of wastewater treatment facilities is based on national compliance data and selected plant characteristics in a statistical model. The motivation for this is to develop a decision support tool for planning and regulatory agencies considering systemic issues such as changes to water quality based effluent limits and enhanced monitoring of discharge permit compliance. A generalized linear model (GLM) was used to predict the probability of meeting permit standards as a function of the degree of decentralization of wastewater systems, which is associated with individual plant capacity since decentralized collection and treatment inherently consists of smaller facilities serving a given population. Four years of discharge monthly reports from 200 treatment plants in the USEPA Integrated Compliance Information System (ICIS) data base were modeled to predict the likelihood that plants would meet discharge limits for BOD, TSS, ammonia and Fecal Coliform based on design capacity and the extent of capacity utilization (Weirich et al., 2011). In a follow up study using the same data set, prior exceedance of permit limits at a plant was found to be associated with increased probability of subsequent violations, when the reported effluent concentration exceeded discharge limits for the same constituent in the previous month (Weirich et al., 2015a).

Although Weirich et al.'s (2011) model was a novel statistical initiative to model the current state of national compliance as a function of plant size and operation, there may be many other factors affecting compliance. We hypothesize that national compliance can be best modeled using two different set of covariates- large scale covariates, which remain fairly constant over space, and local variables, which affect compliance on a smaller scale. These two distinct categories of variables are modeled successively in a hierarchical framework. The first level of hierarchy consists of regression modeling using GLM, where compliance is

modeled as a function of plant capacity (flow) and hydraulic loading (similar to Weirich et al.s 2011 model). Additionally, we introduce covariates for seasonality (to test for seasonal variation in treatment performance) and Lag (to account for temporal dependence or 'persistence' in the discharge water quality data). The second level of hierarchy models spatial variations in compliance. Geographic factors such as temperature, precipitation, and water use may affect municipal wastewater characteristics and also treatment conditions on a local scale. In addition, discharge limits, notably for ammonia, are typically based on local surface water quality standards, which may have a spatial characteristic depending on factors such as in-stream dilution, aquatic life, and designated uses. Another component of compliance data may be associated with location, namely local water quality criteria and associated discharge permit limits for nutrients, particularly ammonia, which is currently the most commonly regulated nutrient. All of these spatial factors, together are incorporated using Kriging, using the residuals from the GLM. Both these tiers together model the temporal and spatial variability in discharge compliance over the United States.

2.2   Methodology

The models in the two level of hierarchy, GLM regression and Kriging, are described below.

2.2.1   Generalized Linear Models

The GLM provides a more flexible approach to regression than a standard linear regression model. In the GLM, the response variable, Y, can be determined by any distribution in the exponential family whereas standard linear models require observations to be from a Normal distribution. Effluent water quality measurements are always non-negative and typically positively skewed which violates the normality assumption and thus cannot be accurately modeled using a standard linear model. Assuming a binomial distribution for the response variable reduces the GLM to a logistic regression; a Poisson distribution makes it

a Poisson regression model, etc. (McCullagh and Nelder, 1989). The ability to model data using a variety of distributions is the major advantage of GLM.

The GLM relation is represented as follows:

$$Y \sim G(\theta) \tag{2.1}$$

where $G(.)$ is any distribution form the exponential family. A link function, η, relates the expected value of Y, and consequently θ, as a linear function of independent variables (or covariates or predictors), $\boldsymbol{X}$, as follows:

$$\eta(E(Y)) = \boldsymbol{X}\boldsymbol{\beta}^T \tag{2.2}$$

where $\eta(.)$ is the link function, $\boldsymbol{\beta}^T$ is the vector of model parameters, $\boldsymbol{X}$ is the set of covariates, $E(Y)$ is the expected value of the response (dependent) variable.

The residuals, $\epsilon = Y - E(Y)$ are assumed to be normally distributed and uncorrelated as with standard linear regression.

The response variables, Y, in this application are the discharge levels of ammonia, normalized to individual facility permit limits to allow comparability on the basis of likelihood of meeting (or violating) permit limits. Values of Y are positive and highly skewed so a Gamma distribution is appropriate and it has three possible link functions (McCullagh and Nelder, 1989)- identity, log and the canonical link function, inverse.

The GLM relations resulting from the inverse, identity and log link functions are as shown below respectively:

$$\frac{1}{E(Y)} = \boldsymbol{X}\boldsymbol{\beta}^T \tag{2.3}$$

$$E(Y) = \boldsymbol{X}\boldsymbol{\beta}^T \tag{2.4}$$

$$ln(E(Y)) = \boldsymbol{X}\boldsymbol{\beta}^T \tag{2.5}$$

The maximum likelihood method is employed for estimation the model parameters, $\boldsymbol{\beta}$, and the best model determined by the best subset of covariates and associated link function,

is selected as the combination that minimizes the Bayesian Information Criterion (BIC) (Schwarz et al., 1978)

$$BIC = k.ln(n) - 2L \tag{2.6}$$

where $k=$ number of free parameters to be estimated, $n=$ number of observations, L=log of the likelihood function of the model.

2.2.2   Kriging

The second level of hierarchy consists of a spatial model of the average value of residuals at each location, calculated from the first level GLM described above. Kriging is a spatial interpolation technique where the estimate at any location is the weighted combination of all the observations, and is a best linear unbiased estimator (Cressie, 1990, Krige, 1951). A brief description is provided below; however, details can be found in the references cited in this section.

Consider observations $Y_1, Y_2, ..., Y_N = \boldsymbol{Y}^T$, at locations $s_1, s_2, ..., s_N$ with the estimate, $\hat{Y}_0$ at any location s_0 (one of the observed or new locations) is obtained as a weighted combination of the observations as:

$$\hat{Y}_0 = \sum_{i=1}^{n} w_i Y_i = \boldsymbol{w}^T \boldsymbol{Y} \tag{2.7}$$

such that

$$\boldsymbol{w}^T \boldsymbol{1} = \sum w_i = 1 \tag{2.8}$$

where w_i are the non-negative weights obtained by minimizing the mean squared error:

$$\sigma_E{}^2 = E([Y_0 - \hat{Y}_0]^2) = Var(Y_0 - \hat{Y}_0) \tag{2.9}$$

Equation 2.9 reduces to $C_{00} + \boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w} - 2\boldsymbol{w}^T \boldsymbol{D}$ where,

$$\boldsymbol{C} = \begin{bmatrix} Var(Y_1) & \cdots & Cov(Y_1, Y_N) \\ \vdots & \ddots & \vdots \\ Cov(Y_N, Y_1) & \cdots & Var(Y_N) \end{bmatrix} = \begin{bmatrix} C_{11} & \cdots & C_{1N} \\ \vdots & \ddots & \vdots \\ C_{N1} & \cdots & C_{NN} \end{bmatrix}$$

and

$$\boldsymbol{D\{Y, Y_0\}} = \begin{bmatrix} Cov(Y_1, Y_0) \\ \vdots \\ Cov(Y_N, Y_0) \end{bmatrix} = \begin{bmatrix} C_{10} \\ \vdots \\ C_{N0} \end{bmatrix}$$

is the covariance vector of the value at the desired location and all the observations. $C_0 0$ is the observational variance.

The covariance estimates in the above matrix and vectors are obtained by a variogram - which is a function that decays with distance. This is the central aspect of Kriging. A theoretical variogram is fitted to this empirical variogram; exponential dependence is the commonly used model, which embodies the concept that points nearby have high correlation which weakens with distance, and is given by:

$$C_{i,j} = Cov(Y_i, Y_j) = \begin{cases} C_0 + C_1(1 - \exp \frac{-3|h|}{a}), |h| > 0 \\ 0, |h| = 0 \end{cases} \tag{2.10}$$

where $= s_i - s_j$, C_0 is the nugget effect caused by dissimilarity of sample values corresponding to points in the close vicinity of one another, a is the effective range or the distance at which the variogram reaches its plateau; and $C_0 + C_1$ is the sill or the maximum variogram value. The range and sill are the two parameters of the variogram that are estimated from the observations. With the appropriate variogram $\boldsymbol{C}$ and $\boldsymbol{D}$ are obtained.

Then the constrained minimization problem is solved using Lagrange multipliers, a standard method for solving such optimization problems (Bivand et al., 2013), and the weights $\boldsymbol{w}$ are obtained as

$$\boldsymbol{w} = \boldsymbol{C}^{-1}[D - \lambda \boldsymbol{1}] \tag{2.11}$$

with the Lagrange multiplier $\lambda = \frac{\boldsymbol{D}^T \boldsymbol{C}^{-1}\mathbf{1}-1}{\mathbf{1}^T \boldsymbol{C}^{-1}-1}$

The variance of the Kriging estimate (i.e. standard error) is obtained from equation by substituting the expression for $\boldsymbol{w}$ as:

$$\sigma_O K^2 = C_{00} - \boldsymbol{w}^t \boldsymbol{D} - \lambda \tag{2.12}$$

Thus Kriging provides the estimate and the corresponding variance. Fitting an appropriate variogram is generally the difficult part of Kriging as the theoretical options are limited and the data often do not confirm to these options (Bivand et al., 2013). Variations of the above Kriging method, also known as Ordinary Kriging, have been proposed to incorporate external variables in addition to distance (Bivand et al., 2013).

Kriging is an attractive and easy to implement method to model spatial fields especially if the data are irregularly spread out in space. The advantages of Kriging include the ability to assign weights based on its proximity to the point of estimation. Kriging also helps compensate the effect of data clustering, by considering points in a cluster as a single point. In our study, the plant locations are clustered at some regions and this makes Kriging the more suitable spatial interpolation technique. Also, during the estimation process, Kriging provides uncertainty estimates which can be helpful for mapping estimation error or for simulation of Kriging estimates.

2.3   Data

Data from the Environmental Protection Agencys Integrated Compliance Information System (ICIS) database for period 2004 - 2008 is used in this analysis. The ICIS database contains enforcement and compliance information based on discharge monthly report (DMR) data from more than 10,000 facilities with NPDES permits in 28 states and US territories. To reduce computational expense, we randomly select every 20th plant entry (629 facilities) providing a good spatial coverage across the country. This dataset is further filtered to

remove facilities with missing data. The independent variable in the hierarchical models is the relative ammonia concentration, defined as the ratio of the reported ammonia concentration to the permit limits for ammonia during that month. Relative discharge values were computed for each facility in each reporting month, with values less than one indicating compliance with a permit limit and greater than one indicating a violation. Use of normalized discharge concentrations allows for comparison of plant performance, and, given the relation between permit limits and receiving water quality, compliance is also an indirect measure of water quality impacts of facilities.

Data points with relative constituent concentration outside the range of 0.01-2 are eliminated so as to preclude extraordinarily low relative constituent concentrations and extremely high violations in permit values. Similarly, data points where wastewater inflows are reported to be less than 5% of plant hydraulic capacity and greater than 200% were also eliminated. This was done to ensure that the GLM implementation in R (www.r-project.org), an open-source statistical computing and graphing package (R Core Team, 2017) was not skewed by extreme values of the predictor and response variables. For ammonia, where the number of extreme violations is higher, it resulted in removal of about 11% of the data. Extreme value analysis (Towler et al., 2010), which can model large excursions, is being considered for future research.
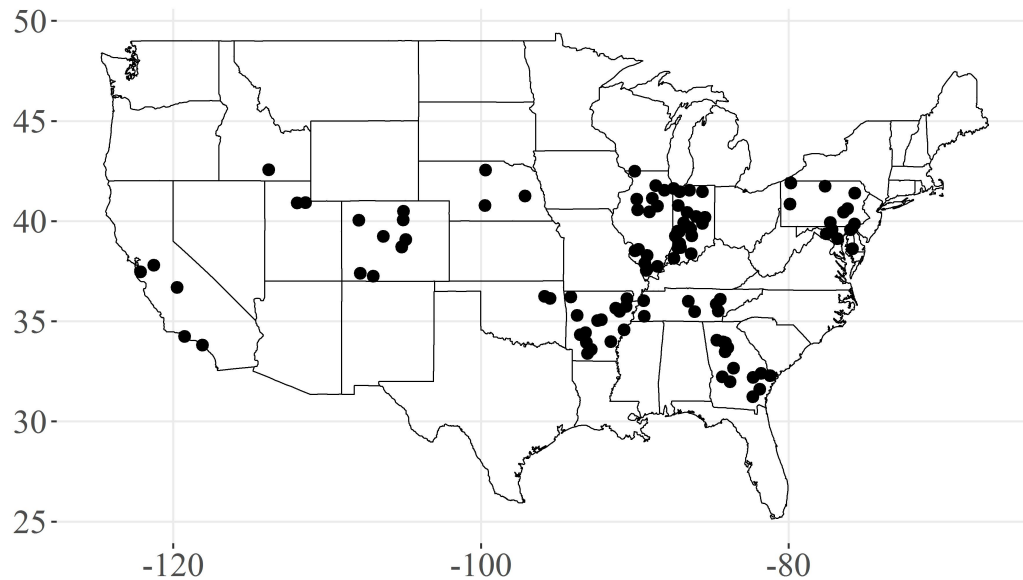
Figure 2.2: Location of 106 facilities with DMR data for ammonia between 2004 and 2008

The final data set for analysis contains an average of 40 monthly values of each variable for each treatment facility in the period between 2004 and 2008 (Figure 2.2). We note that the spatial distribution of permitted Publicly Owned Treatment Works (POTWs) in the US is not uniform. The number of permitted facilities per capita on a statewide basis is highest in the Midwest region, near the Ohio and Mississippi River drainage regions resulting in a higher concentration of facilities in this region in the randomly selected sample population (U.S.EPA, 2008).

GLM model inputs include the natural log of the average monthly flow rate into the plant (A), and the capacity utilization (C), which is the reported monthly average flow rate divided by the design flow rate. In addition to A and C, the product AC was added as a covariate (Weirich et al., 2011), to capture non-linear interactions between A and C. In the research by Weirich et al. (2011), these variables were used in a GLM to test the

hypothesis that decentralization of wastewater infrastructure, necessarily associated with smaller treatment facilities, might have an impact on performance measured by compliance with permit discharge limits.

However, the residuals from this GLM were high in magnitude, auto-correlated and predictive ability of this model was limited. To improve on the model skill, we introduced two additional independent variables. The first variable, referred to as 'Lag', which is the relative ammonia concentration in the previous month, was used by Weirich et al. to incorporate 'persistence' or 'inertia' in system performance and eliminate autocorrelation (Weirich et al., 2015a). In addition to flow and historic compliance characteristics, we hypothesized that seasonality is a second variable which could account for a portion of compliance variability. Seasonal treatment operations can reflect temperature effects on biological transformation rates (Cornel and Weber, 2004). In order to capture seasonality effects, two cyclical co-variates: sine component ($\sin(\frac{2\pi*month}{12})$) (sc) and cosine component ($\cos(\frac{2\pi*month}{12})$) (cc) of monthly relative values are computed for each month at each location for which the DMR data are present. The sine and cosine cyclical covariates have been widely used to effectively capture seasonality in stochastic daily weather generators using GLMs (Kleiber et al., 2012, 2013). While the sine component incorporates the cyclical changes in response variable, the cosine component takes into account the phase shift together they can model any arbitrary seasonal cycle pattern. Thus, the covariates A, C, AC, sine terms, cosine terms and the Lag variable constitute the complete predictor set considered in the GLM analysis.

2.4   Results and Discussions

The results from the first and second level of hierarchy are presented here. The best link function for the Gamma family and associated coefficients for the GLM are selected using the minimum BIC criterion and are summarized in Table 2.1. The best GLM is evaluated using the 'stepAIC' function in the MASS package (Venables and Ripley, 2013) of R. In this, BIC values are computed for all combination of predictors and the model with the least BIC

Covariate	Coefficient
Intercept	0.048
A	-0.007
Lag	0.857
Link function	Identity

Table 2.1: Covariates and coefficients for the best GLM in the first level of the hierarchical model for ammonia, as determined by BIC. The variables C, AC, sc and cc do not show up in the best GLM
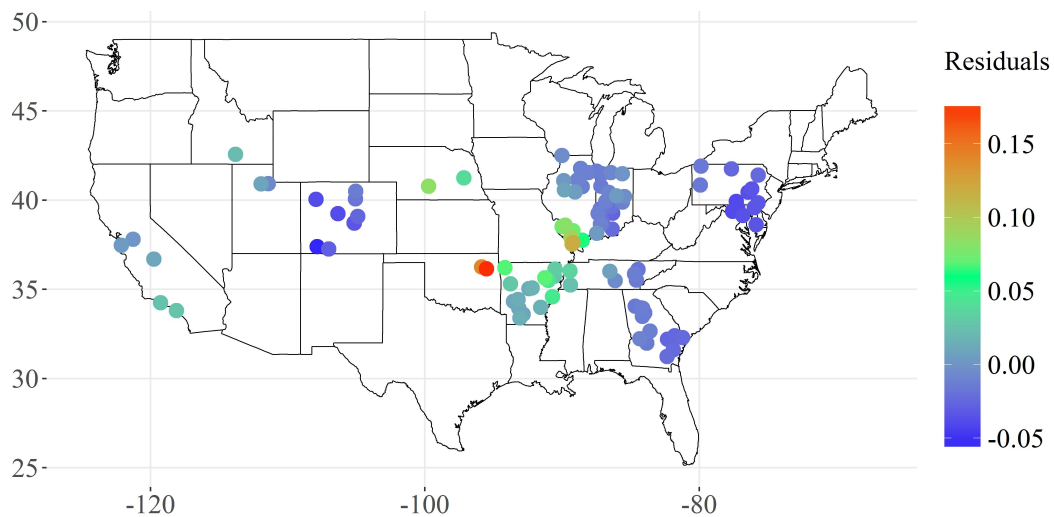
Figure 2.3: Spatial map of average of residuals (at each location) obtained from GLM for relative ammonia, plotted using Kriging

value is selected as the best model.

The best GLM for ammonia consists of the variables, A and Lag, with the p-values of the coefficients significant at 99% confidence level or higher. Once again, the positive

coefficient for the lag variable indicates persistence of discharge levels from one month to another, including higher likelihood of successive violations. It is interesting to note that seasonality was not found to affect the relative ammonia concentration. This could result from the practice that ammonia is typically regulated by seasonal discharge limits, which tend to be higher in winter. Low temperatures particularly affect nitrification, and with uniformly low limits it might be expected to see a strong influence of the seasonal variable in the GLM. However, the leeway in ammonia discharge levels provided by seasonal permit limits may explain why this variable did not have a significant effect on compliance.

GLM for ammonia is spatially plotted using Kriging. Figure 2.3 shows the spatial patterns of the residuals. Once again, the magnitude of residuals is close to zero across the country, except for a small region in northeast Oklahoma. This is also the region with high relative ammonia concentration (0.5- 1.8), indicating that the GLM is unable to capture high relative concentration values. These are the locations, where the second level of hierarchy explains the remainder of the variance. The predictions from Kriging are added to the average of the GLM predictions at each location, to get the hierarchical model predictions.

2.5   Model Validation

The predictive performance of the hierarchical model is evaluated in a drop 10% cross validation mode. This is done by randomly dropping 10% of the plants in the data set and fitting a GLM and Kriging model on the remaining 90% plants. Hierarchical model estimates are obtained as the sum of the GLM and Kriging model predictions. The predictions are compared to the corresponding observations and the R^2 value is computed. This is repeated 500 time to obtain 500 R^2 values which are shown as boxplots. Such a cross validation approach stresses the model and provides estimate of the predictive capability. Unlike traditional approach which uses a training and validation data set, the performance is beholden to these two data partitions. The drop-10% cross validation approach have been used in water quality modeling (Towler et al., 2009, Zachman et al., 2007) and streamflow

forecasting (Grantz et al., 2005, Regonda et al., 2006), among other applications.
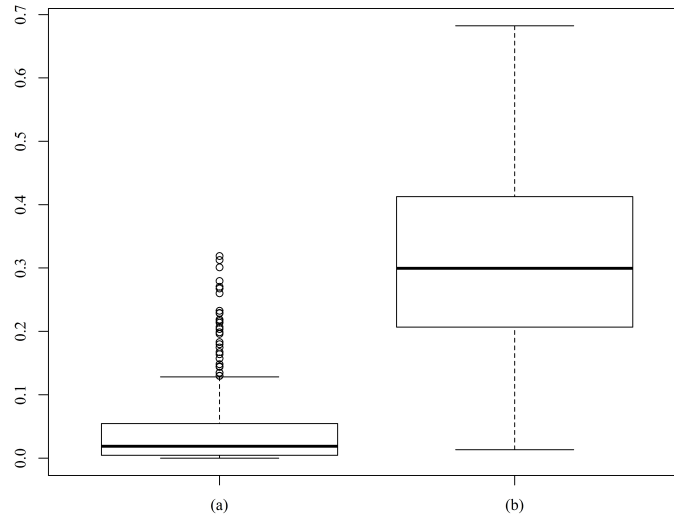
Figure 2.4: Boxplot of R^2 values obtained by comparing observed and predicted relative ammonia concentrations in the drop 10% cross validation mode, when (a) the parent covariate set consisted of A, C and AC (b) the parent covariate set consisted of A, C, AC, seasonality and Lag
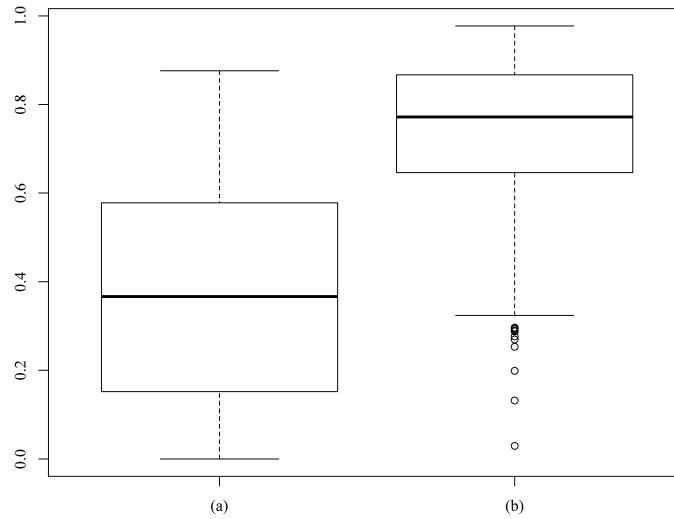
Figure 2.5: Boxplot of R^2 values obtained by comparing observed and predicted relative ammonia concentrations in the drop 10% cross validation mode, at locations where the relative ammonia concentration exceeds 0.5, when (a) only the first level of hierarchy, GLM, is used for prediction (b) both levels of the hierarchical model are used for prediction

In this process, we first test the predictive performance of the first level of hierarchy i.e., the GLM due to the addition of seasonality and Lag to A, C and AC. From the covariate set consisting of A, C and AC, the best GLM is selected using the minimum BIC criterion. Figure 2.4 (a) shows the boxplot of R^2 values from drop 10% cross validation, when the best GLM is chosen from a predictor set consisting of A, C and AC and, Figure 2.4 (b) shows the same using a predictor set of A, C, AC, seasonality and Lag.

For Figure 2.4 (a), the best model consists of only A, while for Figure 2.4 (b), the best model consists of A and Lag (first level of hierarchy, refer to Table 2.1). We observe that the addition of seasonality and Lag to the GLM parent covariate set results in the increase of the median R^2 value from about 0.05 to 0.3, thus offering improved predictive performance.

Next we evaluate the improvement in predictive performance of the GLM models above

with a second level of hierarchy using the spatial. From the spatial maps of the residuals it was evident that this GLM captured most of the variability in the data except at locations where the relative concentrations were high (exceeding 0.5 for both BOD and ammonia). So, we evaluated the predictive performance of the GLM and the hierarchical model GLM with Kriging model at these specific locations. Boxplots of R^2 values are shown in Figures 2.5 (a) and (b). The two-level hierarchical model offers considerable improvement in the predictive performance over the first level of hierarchy, median R^2 increases from 0.3 to 0.75.

2.6   Conclusions

We report the results of a novel hierarchical modeling approach to capture the temporal and spatial variability of wastewater treatment compliance to ammonia. Data from 106 plants for ammonia was used to build the hierarchical model, which enables diagnosis of treatment plants' compliance with existing discharge limits based on flow, hydraulic loading, seasonality and prior performance. The advantage of this two-tiered hierarchical approach is that it individually models two different sets of factors which affect compliance- variables which act as major drivers of compliance variability along with local covariates i.e. latitude and longitude. In the first level of hierarchy, addition of Lag to the GLM consisting of A, C and AC greatly enhances the predictive performance. For most regions across the country, where relative concentrations of pollutants are low (<0.5), the first level of hierarchy captures up to 70% of the variability in the data. Unlike earlier statistical models of wastewater treatment reliability (Niku and Schroeder, 1981a, Oliveira and Von Sperling, 2008), process technology was not considered in the GLM. However, the result of our study: that up to 70% of the variability in ammonia discharge levels relative to permit limits is accounted for by global operating characteristics (flow, capacity loading, and prior performance), suggests that the influence of these variables, exerted over the facility lifetime, should be considered in wastewater treatment compliance evaluation.

For regions with higher relative pollutant concentrations, the GLM alone may not be

enough. Addition of a spatial component helps account for some local variability in the data, providing further improvement in the predictive performance. The spatial maps of residuals also can act as a quick tool for regulatory agencies to identify regions of non-compliance, as the regions of high residuals correspond to the regions of high relative pollutant concentrations. Thus the hierarchical set up is more efficient in diagnosing and predicting non-compliance with ammonia discharge limits, which continue to be regulated on a local basis (Shapiro, 2012). Given that the geographical distribution of POTWs is not uniform with respect to population, this hierarchical framework also assuages concerns about the sensitivity of the model predictions to the spatial density of observations.

The hierarchical approach presented here can also be applied under other scenarios. The predictions from this hierarchical model can provide insight into how treatment plant compliance with nutrient limits might change if more stringent technology-based limits were imposed on a statewide, regional or national scale. Further, by incorporating other probability distributions in the first level of hierarchy, some other aspects of compliance may be modeled. For example, a binomial distribution in the first level of hierarchy can be used to model the probability of regulatory threshold exceedance; a Poisson distribution would enable hierarchical modeling of the number of permit violations expected at various locations across the country. These scenarios are explored in subsequent chapters.

Chapter 3

Modeling risk attributes of wastewater treatment plant violations of Ammonia Discharge Limits in the United States

A version of this chapter has been submitted to the journal of Stochastic Environmental Risk and Research Assessment.

Abstract

A performance-based modeling framework is presented to estimate three attributes of the risk of ammonia permit limit violations: probability, magnitude and frequency of violations. Discharge Monthly Report (DMR) data from a sample of 106 US municipal treatment plants for the period of 2004-2008 is used for the analysis. The probability and frequency of violations are modeled in a Generalized Linear Model (GLM) framework, using logistic and Poisson regressions respectively. The magnitude of violations is modeled as a non-stationary Generalized Pareto Distribution (GPD), based on Extreme Value Theory. The models are fit using plant flow rate, fractional use of design capacity, seasonality and prior performance as covariates. The GLM and GPD models identify the covariates which can predict the probability, frequency and magnitude of ammonia permit violations. Model performances are validated; the logistic regression model offers a median Brier Skill Score of 0.25 in prediction while the Poisson regression model provides a median of 65% correct prediction. The GPD model provides an accurate fit to the observations, with equivalent model and empirical quantiles. The model predictions for the probability and frequency of

ammonia permit violations are combined to obtain two composite risk indices- the Estimated Violation Index (EVI), given as the product of the average probability of violation and the frequency of violation at each location, and Estimates Severity Index (ESI), given as the product of the EVI and the two-year return level of discharge ammonia concentrations above the regulatory limit. Spatial maps of EVI and ESI are constructed as easy visual guides for regulatory agencies to identify risks from sources of high levels of ammonia discharge, and prioritize compliance and enforcement efforts.

3.1   Introduction

Chapter 2 discusses a hierarchical framework for modeling compliance with current ammonia discharge limits. Non-compliance with these limits can result in an unanticipated increase in nutrient loading from wastewater treatment plants. Thus, it is imperative to develop a better understanding of the factors affecting regulatory non-compliance, and a first methodological attempt in this direction has been made in this chapter. Discharge of ammonia in treated wastewaters is regulated by in the US through Water Quality Based Effluent Limits or WQBELs (U.S.EPA, 2010) which typically vary geographically and seasonally. WQBELs are set by state regulators, usually to incorporate specific characteristics of the receiving water body, including level of impairment, designated uses, dilution factors and local aquatic life conditions. Seasonal variability in WQBELs allows for changes in streamflow, water temperature and sensitivity of the developmental stages of aquatic organisms (Rossman, 1989). As a result, on a national scale these discharge limits vary greatly- between 0.5 to 215.5 mg/l, according to the U.S.EPA's ICIS/ ECHO database (www.echo.epa.gov). A histogram of these discharge limits is shown in Figure 3.1 (a). For the sake of clarity, discharge limits up to 50 mg/l are shown in this figure only. The largest fraction (40%) of the discharge limits lie between 0.5 to 2 mg/l ammonia and almost 70% of the limits are below 5 mg/l. Figure 3.1 (b) is a plot of the observed effluent ammonia concentration values compared with the discharge limits for each DMR. 13% of the plants (14 plants) had

an average of one permit violation per year, while 11% of the plants (12 plants) had more than one (ranging between 2 to 8) permit violations per year. The magnitude of ammonia violations range from 0.5-800% above the permit limit, with a median violation 73% above the discharge limit . Given this range, for any treatment plant, the report of a permit limit violation alone is not a reliable indicator of risk to receiving water quality. More comprehensive risk characterization also depends on the frequency, or return period, of violations and their magnitude.
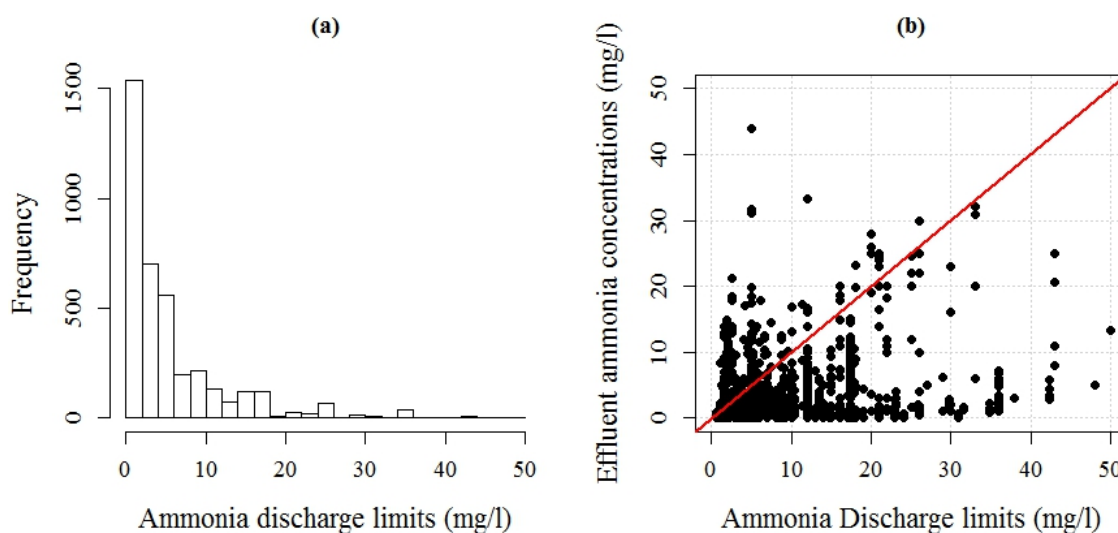
Figure 3.1: (a) Histogram of ammonia permit discharge limits for each DMR in the data set (b) monthly effluent ammonia concentrations vs. ammonia discharge limits for 106 treatment plants over the period of 2004-2008

Performance-based analysis of risk, reliability and resilience provide important measures of infrastructure effectiveness, security and sustainability applicable to wastewater treatment systems (Bogardi and Kundzewicz, 2002, Hashimoto et al., 1982, Kjeldsen and Rosbjerg, 2004). The reliability of individual wastewater treatment plants in meeting effluent water quality criteria has been evaluated using performance data, technology type and a variety of statistical methods and artificial neural networks (Hao et al., 2013, Niku et al.,

1979, Oliveira and Von Sperling, 2008, Santin et al., 2015). Weirich et al. (2015a) used Generalized Linear Model (GLM) regression and synthetic time series analysis of DMR data from over 200 US wastewater facilities to simulate the likelihood, frequency and duration of BOD, total suspended solids and ammonia permit violations in order to examine the effect of decentralization on wastewater treatment performance.

However, probabilistic attributes are only one component of overall risk. To assess the impact of treatment performance on receiving water quality the purpose of NPDES effluent limits, knowing the distribution of the magnitude of non-compliance events is critical, and to date there are no studies of wastewater system performance in a complete risk framework. To address this need, we model three attributes of risk associated with wastewater system performance: the probability, frequency and magnitude of permit violations using the ICIS/ECHO data subset described above. Each risk attribute is modeled separately, with modifications for fitting an Extreme Value distribution to model the magnitude of effluent ammonia concentrations above the regulatory limit. These modeled risk attributes are combined to generate composite risk metrics, to provide a performance-based risk assessment of the water quality impacts treated wastewater ammonia, accounting for variability in effluent water quality.

3.2   Data

To model the probability, frequency and magnitude of ammonia permit limit violations, we analyze the U.S.EPA's Integrated Compliance Information System (ICIS) database for the period of 2004-2008, as described in Section 2.3. This data set consists of 3888 DMR records from 106 plants, where each record contains the average influent flow rate, design flow rate, monthly average effluent ammonia concentration and permit average effluent ammonia concentration for a given month. The geographic distribution of these plants is shown in Figure 2.2.

3.3   Model Development

3.3.1   Variable Selection

The probability of violating the ammonia permit discharge limit is modeled using logistic regression, where the dependent or response variable is the state of compliance at each location during the period of 2004-2008 — 0 denoting a state of compliance (effluent concentrations of ammonia below or equal to the permit limit) and 1 denoting a state of non-compliance or violation (effluent concentrations above the permit limit). As described in Chapter 2, the covariates or predictor variables consist of A, C and AC, seasonality- sine component ($\sin(\frac{2\pi*month}{12})$) (sc) and cosine component ($\cos(\frac{2\pi*month}{12})$) (cc). Additionally, we use Lag as a covariate, which in this case is the state of compliance/ non-compliance in the previous month. The logistic regression model is fitted using the 'glm' function from the 'stats' package in R (R Core Team, 2017). The best model is selected as the combination of covariates which minimizes the Bayesian Information Criteria (BIC). This is implemented in R using the 'stepAIC' function, as in Chapter 2.

The frequency of ammonia discharge limit violations is modeled using a Poisson regression, where the response variable is the average number of permit violations per year at each location over the four-year study period. At any location, this is calculated by dividing the total number of permit violations by the number of years of observations available. As we have one value of response variable at each location, the covariates consist of the values of A, C and AC, averaged over the period of records available for each location. The 'glm' function from the 'stats' package is used, with family 'poisson' and the canonical link function, log.

The magnitude of permit violations is modeled by fitting a non-stationary Generalized Pareto Distribution (GPD) to the relative ammonia concentrations, defined (in Chapter 2) as the ratio of the reported effluent ammonia concentration to the ammonia permit limit for that month. A GPD is fit such that one of the distribution parameters (the scale factor) is non-stationary, varying as a function of A, C, AC and seasonality as defined previously,

Additionally, we included the Lag covariate, the relative pollutant concentration in the previous month. The threshold for fitting the GPD model is one, which denotes the regulatory threshold, i.e., a relative ammonia concentration which is equal to the discharge permit limit. The significance of the threshold in fitting a GPD is explained later in section 3.4.2. The GPD is fit using the 'fevd' function from the 'extRemes' package in R (Gilleland and Katz, 2011). The Maximum Likelihood Estimator (MLE) is used to fit the GPD, and the best covariate set for the non-stationary scale factor is determined using the likelihood ratio ('lr.test' function in R from the same package) as well as the BIC. The distribution thus fit is used to evaluate the two-year return level, i.e., a vector of relative ammonia concentrations with probability of exceedance equal to 0.5, which serves as GPD model estimates for the magnitude of permit violations.

3.3.2   Validation

The predictive performance of both the logistic and Poisson regression models is evaluated in a drop 10% cross-validation method. The logistic regression model is validated by calculating the Brier Skill Score (Wilks, 2011), while the Poisson regression model is validated by comparing the percentage of predictions which are the same as the observations, both in the drop 10% cross validation mode. As the magnitude of exceedance is computed by fitting a distribution, instead of fitting a regression model, the above mentioned cross-validation approach is not applicable. Instead, we assess the goodness of model fit graphically from Quantile-Quantile plots.

3.4   Statistical Methods

3.4.1   Generalized Linear Models

Generalized Linear Models (GLMs) have already been introduced in Chapter 2. For the probability of permit violation at any location, the binary response variable (value of

0 if in compliance, and 1 if in violation) is assumed to be from a Binomial distribution, associated with a logistic regression model using the canonical or default link function, logit, in the GLM to model the probability of violations.

$$Y \sim Binom(p) \tag{3.1}$$

$$ln(\frac{E(Y)}{1 - E(Y)}) = \boldsymbol{X}\boldsymbol{\beta}^T + \epsilon \tag{3.2}$$

where $E(Y) = p$ is the probability of violation, ϵ is the residual, given as the difference between the observed value and predicted probability of exceedance, $\boldsymbol{\beta}^T$ is the vector of model parameters and $\boldsymbol{X}$ is the set of covariates.

The average number of permit violations per year at any location is a discrete number which can be modeled using the Poisson distribution with the canonical link function, log, expressed as below:

$$Y \sim Poisson(\lambda) \tag{3.3}$$

$$ln(E(Y)) = \boldsymbol{X}\boldsymbol{\beta}^T + \epsilon \tag{3.4}$$

where $E(Y) = \lambda$ is the expected value of the response variable, Y, and the remaining terms are as explained above. Model parameter estimation in both of these models is done using the Maximum Likelihood method. The best set of covariates or independent variables are selected as the subset of covariates which minimizes the Bayesian Information Criterion (Schwarz et al., 1978), also discussed in Chapter 2.

Researchers have used logistic regression to study land-use change (Wang et al., 2013), cropland degradation (Dubovyk et al., 2013), habitat studies (Pearce and Ferrier, 2000) etc. Poisson regression has found many applications in vehicular crash studies (Li et al., 2013,

Dong et al., 2014),epidemiology studies (Hogan et al., 2012), ecological modeling (Towler et al., 2013), and air pollution studies (Best et al., 2000).

3.4.2 Generalized Pareto Distribution

Classical statistical methods tend to focus on the mean or average behavior of a stochastic process with distributions that have thin tails, with the result that the extreme or rare events, which lie in the tails of the distribution are poorly modeled or often underestimated. Extreme value distributions borne from Extreme Value Theory (Coles et al., 2001, Katz et al., 2002) are used to model the maximum values of a large set of data (e.g., seasonal or annual maxima) or the observations exceeding a threshold level, known as Peak Over Threshold (POT). While the former is modeled using Generalized Extreme Value (GEV distribution)(Jenkinson, 1955), the latter is modeled using a Generalized Pareto Distribution (GPD) (Pickands III, 1975). The random variable Y has a Generalized Pareto Distribution GPD (k, σ) if the cumulative distribution function of Y is given as:

$$F(y; k, \sigma) = \begin{cases} (1 - (1 - \frac{ky}{\sigma})^{\frac{1}{k}}), & k \neq 0, \sigma > 0 \\ 1 - e^{\frac{-y}{\sigma}} & k = 0, \sigma > 0 \end{cases} \tag{3.5}$$

where σ and k are scale and shape parameters, $x > 0$ for $k \leq 0$ and $x \in [0, \frac{\sigma}{k}]$ for $k > 0$ (Castillo and Hadi, 1997). When $k = 0$, the GPD becomes the exponential distribution, and when $k = 1$, it becomes the uniform distribution (Choulakian and Stephens, 2001). POT analysis has been used to model temperature spells (Furrer et al., 2010), water demand (Haagenson et al., 2013) and streamflow reconstruction (Lukas et al., 2007).

In both the GEV and the POT approach, the distribution parameters can be modeled as function of set of covariates (Katz et al., 2002), resulting in a non-stationary model. The shape parameter, k is usually fixed as it is noisy, and the scale parameter, σ, is modeled as a function of covariates, $\boldsymbol{X}$, as with a GLM (Katz et al., 2002).

$$Y \sim GPD(k, \sigma) \tag{3.6}$$

$$ln(\sigma) = \boldsymbol{X}\boldsymbol{\beta}^T + \epsilon \tag{3.7}$$

where the terms are as described in previous sections. Covariates have been incorporated in extreme value analysis of climate and hydrology, including precipitation (El Adlouni et al., 2007, Katz et al., 2002, Maraun et al., 2009), temperature (Cooley, 2009), and water quality variables (Towler et al., 2010). The parameter estimation is done using the Maximum Likelihood Estimation method while the best subset of covariates is selected using the likelihood ratio test (Coles et al., 2001, Reiss and Thomas, 2007). This study is the first application of non-stationary GPDs to wastewater compliance risk modeling.

3.5   Results

3.5.1   Probability of permit violation

The covariates and coefficients of the best logistic regression model are shown in Table 3.1. The p values for all coefficients are significant at the 99% confidence level or higher. Only A and Lag are selected in the best fitted model. A has a negative coefficient, indicating that smaller facilities have a higher probability of a permit violation. The coefficient for Lag, the state of compliance in the previous month, has the highest positive value among all variables, indicating a strong dependence between the state of compliance in the previous and current months. These results are consistent with the findings from Chapter 2, where A and Lag are seen to be the only covariates, in the first level of hierarchy, necessary for prediction of relative ammonia concentrations in effluent wastewaters.

Covariate	Coefficient
Intercept	-3.93
A	-0.33
Lag	3.43

Table 3.1: Covariates and coefficients for the Logistic regression model, as determined by BIC
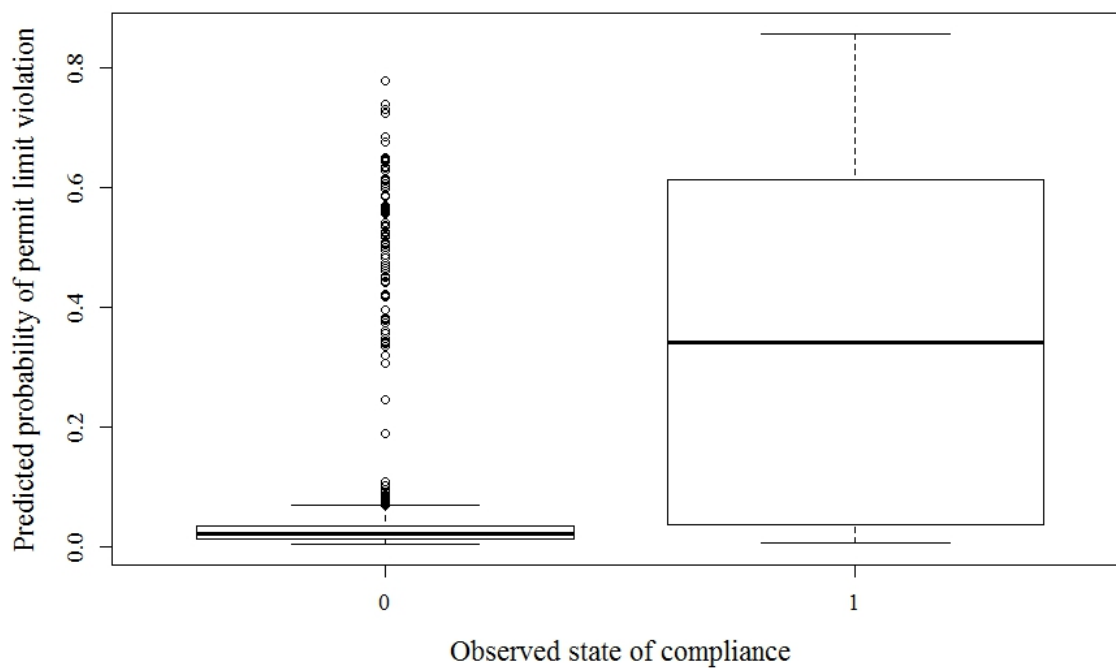
Figure 3.2: Boxplots for average predicted probability of ammonia permit limit violations for the period of 2004-2008, as predicted by the logistic regression

For each state of compliance (0 for compliance, 1 otherwise), a boxplot for the predicted probability of ammonia permit limit violation shown in Figure 3.2. For an observed probability of exceedance value of 0 (compliance), the median predicted probability is 0.02; the median predicted probability is 0.34 for an observed probability of 1 (violation).

Validation

The predictive performance of the logistic regression model is assessed in the drop 10% cross validation mode and the resulting boxplot of BSS values is shown in Figure 3.3. The median BSS value obtained from this model is about 0.25. A positive BSS value indicates that the model predictions are better than the climatological forecast. A high positive coefficient for Lag (Table 3.1) indicates that for any month, the model prediction is the same as the state of compliance/ non-compliance in the previous month. This is because for majority of the observations (3693 of 3888), the state of compliance in any given month is the same as the state in the previous month. For example, a state of compliance (0) in the current month is preceded by a state of compliance (0) in the previous month as well, or a state of non-compliance (1) in the current month is preceded by a state of non-compliance (1) in the previous month. However, for only 195 of 3888 observations, a change in the compliance state between consecutive months is seen i.e., non-compliance (1) is followed by compliance (0), or vice-versa. For this data set, Lag is inversely related to model predictions. The model shown in Table 3.1 is unable to capture this inverse relation, which is why the model predictions have relatively lower BSS values.
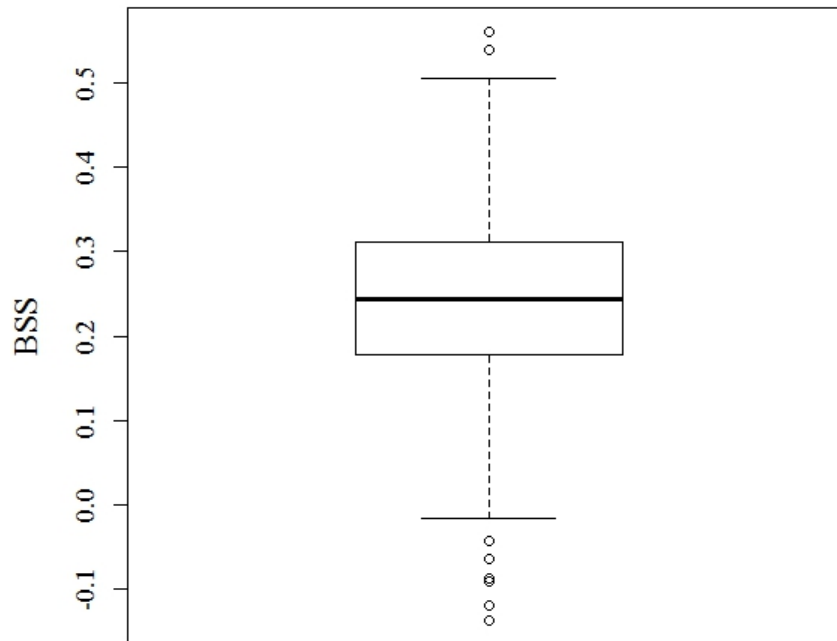
Figure 3.3: Boxplot of BSS values obtained from the logistic regression model in the drop 10% cross validation mode

3.5.2   Frequency of permit violation

The covariates and coefficients for the best Poisson regression model are shown in Table 3.2. A is the only significant variable in predicting the average number of permit limit violations per year, at each of the 106 wastewater treatment facilities. As with the logistic regression model, the negative coefficient on A suggests an inverse relation between the response variable (permit violation frequency) and plant flow rate; that is, facilities with lower average monthly flow rates are likely to experience a higher average number of permit limit exceedances per year. The p values for all coefficients are significant at the 99% confidence level or higher.

Covariate	Coefficient
Intercept	-1.26
A	-0.36

Table 3.2: Covariates and coefficients for the best Poisson regression model, as determined by BIC

Bar plots of the observed and predicted (by the Poisson regression model) number of average annual number of permit violations is shown in Figure 3.4 . The Poisson regression model can successfully predict violation frequency at locations where there are 0-3 permit violations per year. However, it is unable to predict higher number of violations per year, such as 4 and 8. Clearly, the Poisson regression model works better on a narrower range of response variable values.

Observed   **Predicted**
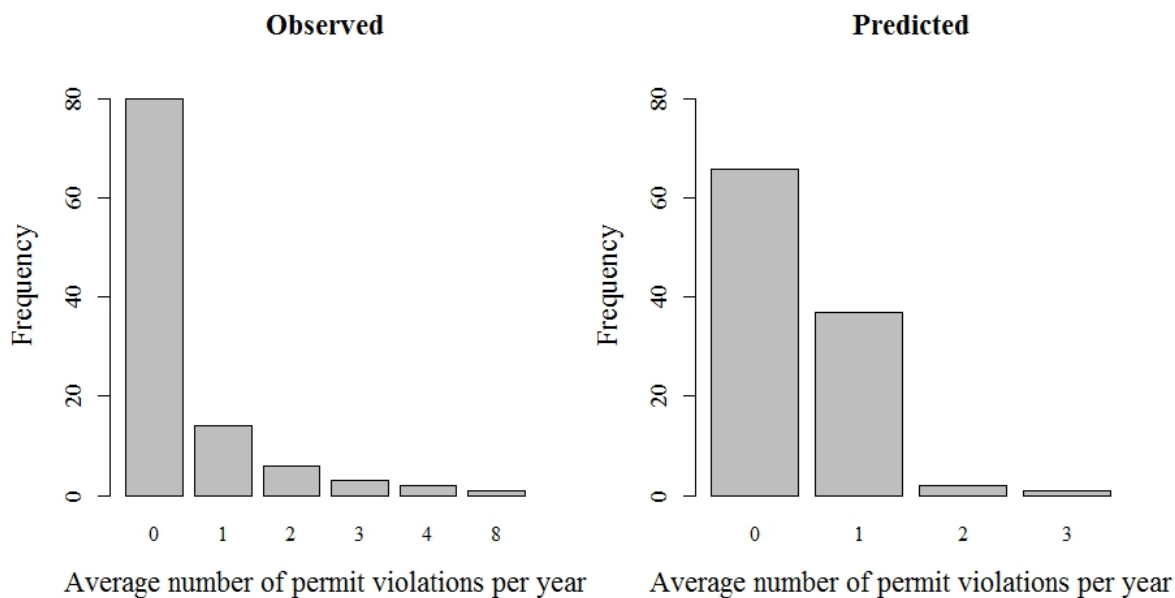
Figure 3.4: Observed and predicted (by the Poisson regression model) values of average number of permit violations occurring per year for the period of 2004-2008

Validation

The predictive performance of the Poisson regression model is evaluated in the drop 10% cross validation mode. Figure 3.5 shows the boxplot of the percentage of correct predictions obtained in each iteration. The median prediction rate of the correct number of average annual ammonia permit limit violations is about 65% with a maximum of 100%. Clearly, the Poisson regression model successfully captures most of the variability in the frequency of permit violations across 106 wastewater treatment facilities in the US.
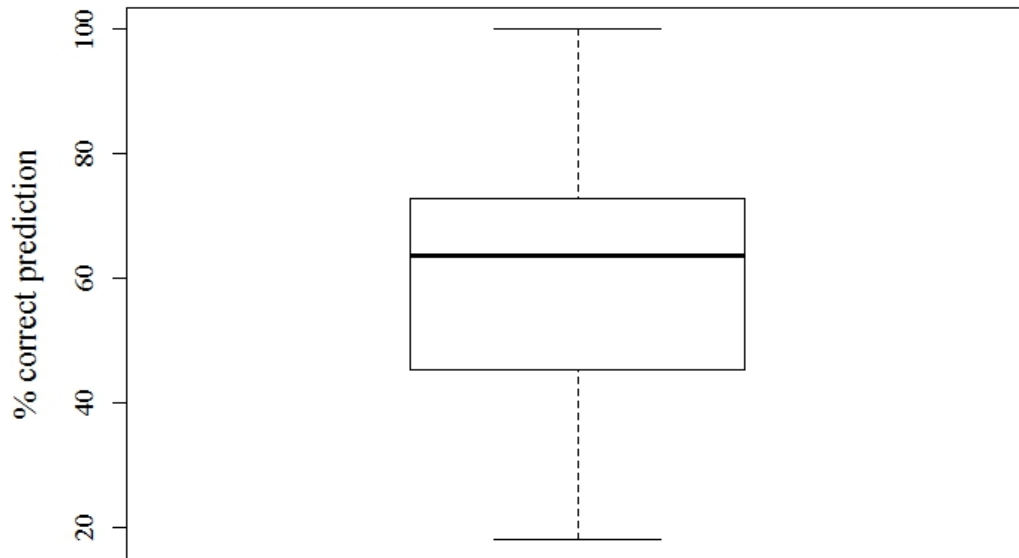
Figure 3.5: Boxplot of the percentage of observations same as the predictions obtained from the Poisson regression model in the drop 10% cross validation mode

3.5.3 Magnitude of permit violation

The non-stationary GPD is fit using a threshold relative ammonia concentration value of 1 (i.e, when effluent ammonia concentration is exactly equal to the regulatory threshold), to model the severity of permit limit violations with respect to the regulatory threshold. More than 10 combinations of covariates (A, C, AC, sc, cc and Lag) as candidate models are used for fitting the non-stationary GPD model. Comparing likelihood ratios as well as the BIC values, the coefficients and covariates for the best scale factor GLM are summarized in Table 3.3. The shape factor value is also provided in the same table.

Parameter	Covariate	Coefficient
Scale factor, σ	Intercept	1.32
	A	-0.06
	AC	0.09
	Lag	0.76
Shape factor, k	–	0.13

Table 3.3: Covariates and coefficients for the best scale factor GLM and shape factor used for fitting the non-stationary Generalized Pareto Distribution model

Figure 3.6 shows a plot of the observed relative ammonia concentrations vs. the predicted 2 year return levels (relative ammonia concentrations with a 0.5 probability of exceedance). the return levels can also be conceptualized as the median predictions from the GPD. The 2 year return levels are seen to be generally higher in magnitude than the observed values.
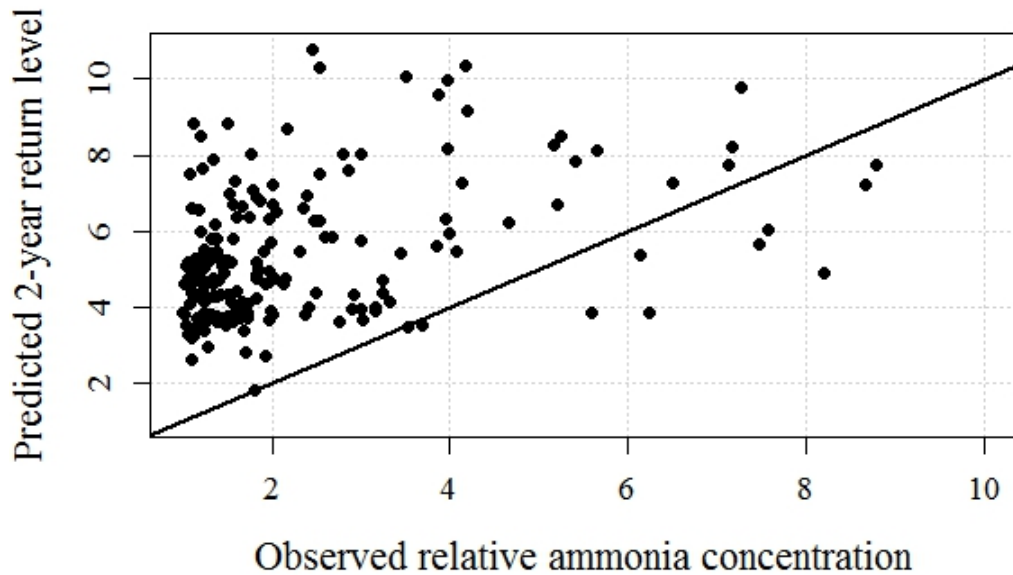
Figure 3.6: Predicted two-year return levels of relative ammonia concentrations greater than 1, as obtained from the GPD model, for the period of 2004-2008

Model Fit

To evaluate how well the model fits to the relative ammonia concentration values, above a threshold of 1, we compare the empirical and model probabilities in Figure 3.7 (a). We can see that the model probabilities are very consistent with the empirical probabilities, as these values all lie very close to the 1:1 line. Similarly, from the quantile-quantile plot (Figure 3.7 (b)), we find that the empirical and model quantiles also align along the 1:1 line. The Q-Q plots indicate good model fit to the observed relative effluent ammonia data.
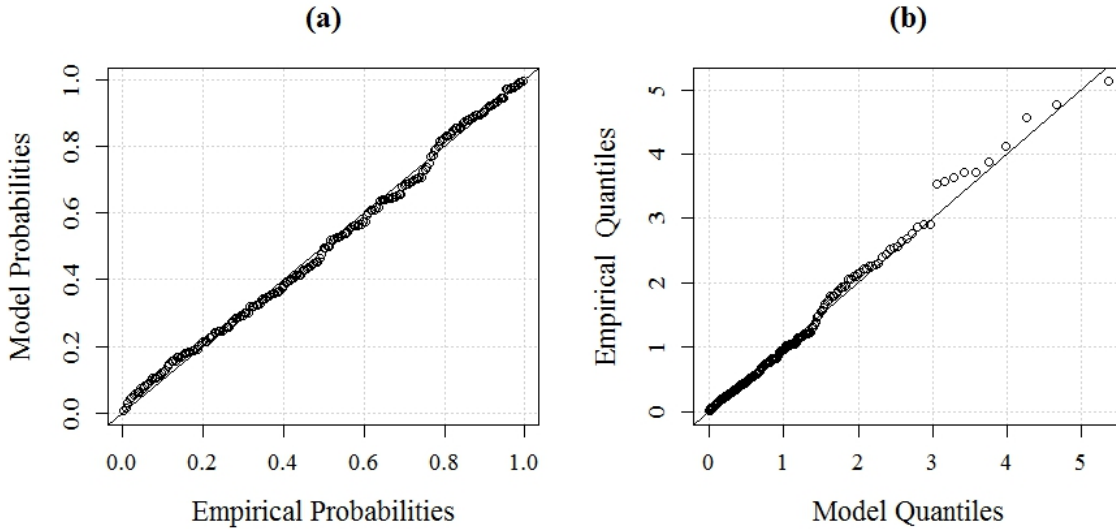
Figure 3.7: (a) Plot of model probabilities vs. empirical probabilities (b) Q-Q plot, as obtained from the non-stationary GPD model

3.5.4   Composite Risk Estimates

Individual risk attributes, as described above, can be combined to formulate composite risk attributes, to provide a quick yet holistic understanding of the nature of violation at each plant. We define two composite risk indices in this section, the Estimated Violation Index (EVI) and the Estimated Severity Index (ESI). EVI gives the permit violation profile at a treatment plant, and is defined as the product of the average probability of violation at any plant and the predicted average number of annual violations at that plant. This index helps delineate plants which rarely and habitually violate their permit limits. The probability of permit violation lies in the range of 0 to 1, while the predicted average annual number of violations ranges between 0 to 12. As a result, the value of EVI, a product of these two quantities, ranges between 0 to 12. An EVI value of 0 indicates that the wastewater treatment plant is in compliance throughout the year; higher the EVI value, the worse is its violation profile, with a higher probability and frequency of violation. The Estimated Severity Index (ESI) is given as the product of the EVI and the average 2 year return level

relative ammonia concentration at each plant. ESI enhances the information given by EVI, by providing an additional estimate of the severity of the violation. While ESI can have a minimum value of 0, the maximum value depends on the relative ammonia concentration return level, which is a positive, non-zero quantity. Clearly, monitoring and enforcement agencies would be interested in investigating treatment plants with a non-zero EVI and ESI value.

A spatial map of the computed non-zero EVI values at each location is shown in Figure 3.8 (a). For monitoring and enforcement agencies, this map provides a quick estimate of the more likely and habitual violators. Figure 3.8 (b) provides additional information about the severity of violations at each of these treatment plant. Once the habitual violators have been identified from Figure 3.8 (a), monitoring and enforcement actions can be prioritized based on the severity of violation given from the ESI map in Figure 3.8 (b). Thus, when used in conjunction, EVI and ESI can serve as quick, simple and reliable indices for monitoring and enforcement agencies .

3.6   Conclusions

A comprehensive and novel framework for modeling three different attributes of risk- the probability, frequency and magnitude of permit violation, has been provided in this research. While the risk attributes provide valuable information about the state of compliance of 106 wastewater treatment plants in the country over the period of 2004-2008, the individual model covariates can help plant operators and managers gain a better understanding of permit violations. The probability of violation attribute, obtained from predictions of the logistic regression model, can help monitoring and enforcement agencies identify which plants are more likely to violate their permits. It would be beneficial to channel (limited) resources into monitoring plants with a higher probability of violation. The state of compliance/ non-compliance is the only relevant covariate in predicting the probability of violation. This means that the compliance history of the plant is predictive of future violations. Using
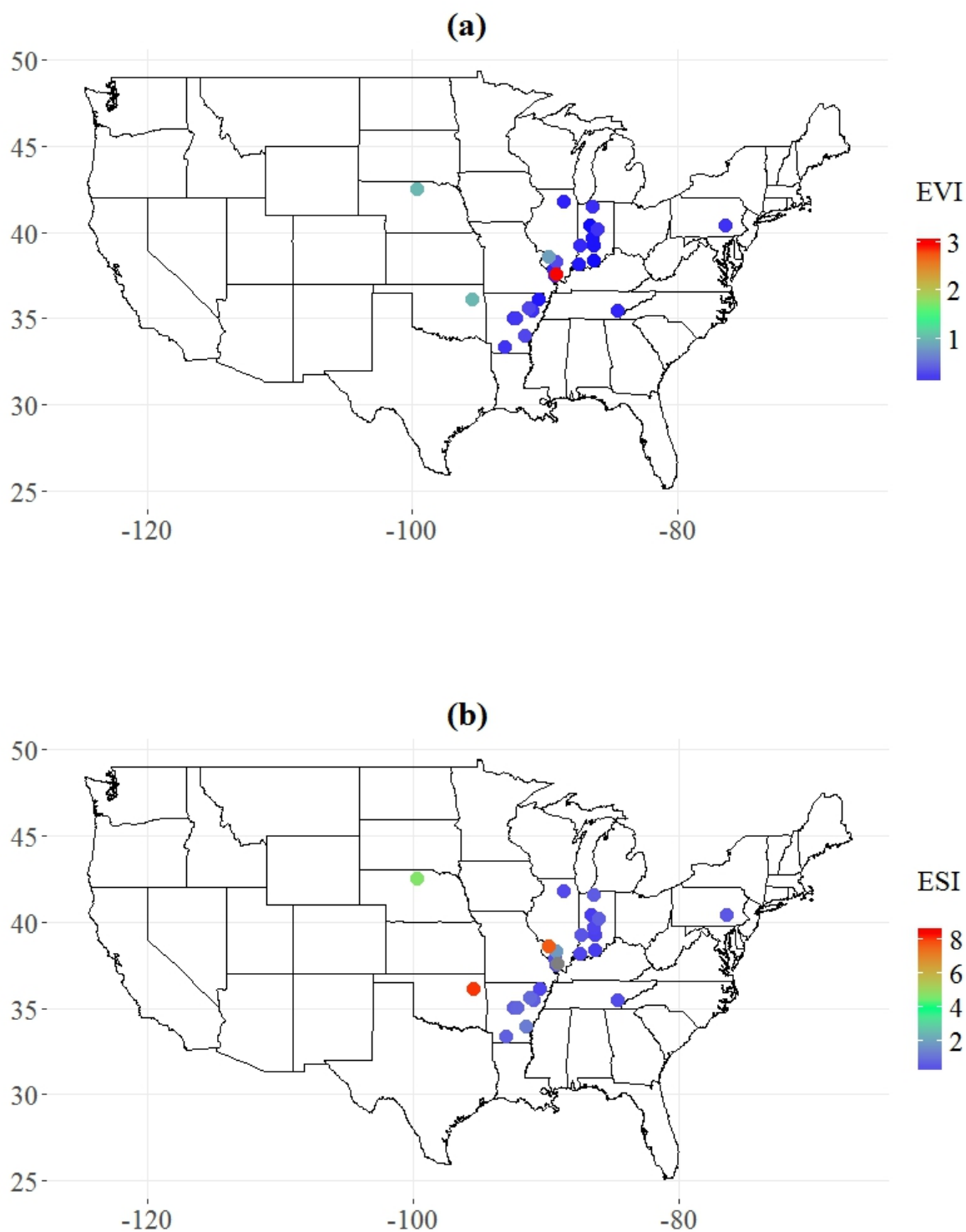
Figure 3.8: (a) Estimated Violation Index (EVI) map (b) Estimated Severity Index (ESI) map for ammonia permit limit violations across various locations for the period of 2004-2008. Only non-zero EVI and ESI values have been shown here

this information, plant operators can frequently monitor effluent ammonia concentrations and tackle process upsets immediately to prevent violations. In addition to the probability, enforcement agencies might be interested in how often these violations occur annually. The answer to this question can be provided using predictions from the Poisson regression model. Once again, plant inflow is the only covariate affecting the frequency of permit violations. Plant operators should monitor low flows for possible permit violations, as low flows are found to be associated with higher number of monthly violations per year. Finally, the severity or magnitude of violations can be obtained from relevant quantile estimates from the GPD model. The relevant covariates are plant inflow, capacity utilization and relative ammonia concentration in the previous month. Once again it provides operators an idea of which variables to monitor to prevent high exceedances above the regulatory threshold.

The risk attributes are combined to evaluate the composite risk indices- EVI and ESI, which are mapped spatially. These maps provide quick, visual aids for monitoring and enforcement agencies to identify habitual and egregious violators.

Chapter 4

Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits using a regression tree model

Abstract

A regression tree-based diagnostic approach is developed to evaluate factors affecting US wastewater treatment plant compliance with ammonia discharge permit limits, using Discharge Monthly Report (DMR) data from a sample of 106 municipal treatment plants for the period of 2004-2008. Predictor variables used to fit the regression tree are selected using random forests, and consist of the previous month's effluent ammonia concentration, influent flow rates and plant capacity utilization. The tree models are first used to evaluate compliance with existing ammonia discharge standards and then applied assuming more stringent discharge limits, under consideration in many states. The model predicts that the ability to meet both current and future limits depends primarily on the previous month's treatment performance. With more stringent discharge limits, predicted ammonia concentration rel-

ative to the discharge limit increases. In-sample validation shows that the regression trees can provide a median classification accuracy of more than 70%. The regression tree model is validated using ammonia discharge data from an operating wastewater treatment plant and is able to accurately predict the observed ammonia discharge category approximately 80% of the time, indicating that the regression tree model can be applied to predict compliance for individual treatment plants providing practical guidance for utilities and regulators with an interest in controlling ammonia discharges. The proposed methodology is also used to demonstrate how to delineate reliable sources of demand and supply in a point source-to-point source nutrient credit trading scheme, as well as how planners and decision makers can set reasonable discharge limits in future.

4.1   Introduction

In the United States, the discharge of pollutants from wastewater treatment facilities to surface waters is regulated by the National Permit Discharge Elimination System (NPDES) created in 1972, under the Clean Water Act. NPDES permits are typically issued and enforced by individual States to control pollutant levels in wastewater discharges by setting effluent concentration limits for regulated constituents. These limitations may be technology based or water quality based (U.S.EPA, 2010). Technology based effluent limits (TBELs) are based on conventional secondary treatment technologies, and set national maximum limits for Biochemical Oxygen Demand (BOD), Total Suspended Solids (TSS) and pH in discharging waters. Water Quality Based Effluent Limits (WQBELs) are set at the state level to achieve desired water quality standards in a particular receiving water, and typically include limits for effluent nitrogen and phosphorus, varying geographically and often seasonally. Spatial variability in WQBELs typically depends on local aquatic life conditions, designated uses of receiving waters and dilution factors while seasonal variations are due to seasonal changes in streamflow, water temperature and sensitivity of the developmental stages of aquatic organisms (Rossman, 1989).

After the USEPA's decision to decline issuing TBELs for nutrients (Shapiro, 2012), many states have considered new uniform statewide permit limits for nutrients, e.g. Colorado (CDPHE, 2012) and Utah (Daigger et al., 2014). Concern about the ability of plants to comply with lowered permit limits using current infrastructure and the cost of implementation of nutrient removal technology has led some states to consider exemptions and allow a transition period before new limits take effect. For example, the State of Colorado exempts plants with design capacity less than one million gallons per day (MGD) (~ 4000 m^3/d) from new nitrogen and phosphorus limits (CDPHE, 2012).

The effectiveness of lower nutrient discharge limits in achieving improved water quality is dependent on compliance with these limits as well as adequate enforcement measures. Unfortunately, a number of studies indicate appreciable levels of non-compliance with existing limits and spotty enforcement across the country (PIRG, 2004, U.S.EPA, 2011). Moreover, the USEPA expects available funds for NPDES permit enforcement to decrease in the coming years (U.S.EPA, 2015). Beyond fiscal limits, there are often political and legal obstacles to both public agency and citizen-initiated enforcement actions (Andreen, 2007, Steinzor, 2003). Given the uncertainty around compliance with permit limits and challenges to enforcement measures across the country, achieving compliance with new, more stringent nutrient permits and associated water quality improvement may be a challenging task.

Uncertainty in compliance can also affect nutrient trading, a market-based strategy for watershed-scale nutrient management consistent with the Total Maximum Daily Load (TMDL) framework used by the EPA and State regulators (Branosky et al., 2011, Hoag and Hughes-Popp, 1997, Jones et al., 2010, Lal, 2010, Ribaudo et al., 2005, VDEE, 2016). Studies of the feasibility of nutrient credit trading and attempts to implement a market approach have focused on trades between point (wastewater discharges) and non-point (primarily agriculture) sectors and results have been generally disappointing. Factors that limit inter-sector trading are differing economic incentives, regulatory structures, and even differences in the ability to verify credits (Hoag and Hughes-Popp, 1997, King and Kuch, 2003).

The most acknowledged success of emissions trading is the cap-and-trade system to reduce acid rain by limiting sulfur dioxide (SO_2) emissions from coal-fired power plants enabled by the Clean Air Act Amendments of 1990. Factors associated with the success of the acid rain program include similar incentives to control emissions across the market and willingness of stakeholders, especially policymakers and regulators to allow flexibility in reaching environmental goals (Chan et al., 2012). A similar market-based approach for nutrient credit trading among wastewater treatment plants will rely on a sufficient pool of potential sellers plants that reliably discharge well-below permit limits, and potential buyers plants that are at risk for violating permits. However, uncertainty in nutrient discharges from both buyers and sellers is a major challenge to establishing reliable markets for nutrients (Dennison et al., 2012, U.S.EPA, 2009). We propose a diagnostic framework, which can evaluate compliance to nutrient limits as a function of readily measurable variables, can provide timely guidelines for estimating the reliability of individual treatment plants with benefits for utilities and policy makers.

Recent research using statistical models based on long-term effluent data has identified the sources of variability in compliance with effluent limits (Suchetana et al., 2016, Weirich et al., 2011, 2015a). Chapters 2 and 3 describe performance-based models to evaluate variability in compliance/non-compliance with existing discharge regulations. While these models were well suited for predictive purposes, they could not offer any operational or diagnostic insights to improve compliance. Further, with the current trend of lowering nutrient discharge limits, these models could not compare and contrast the factors affecting compliance with old and new regulations. This necessitated building a new data-driven diagnostic framework that could distinguish which variables affect compliance to both existing and more stringent ammonia discharge standards, as well as promote strategies for individual wastewater treatment facility operators to improve compliance and achieve water quality standards for nutrients at a reasonable cost. To this end, we develop a regression tree-based diagnostic framework to evaluate compliance of wastewater treatment plants to

current and future ammonia discharge standards. The new approach in this study is the first methodological attempt to evaluate compliance with ammonia limits using statistical modeling based on past performance data with the advantage of enabling consideration of more variables than previous approaches and output that can guide both treatment facility operators and environmental policy makers.

4.2   Methodology

The primary objective of the regression tree analysis is to propose a simple diagnostic framework to serve two related goals to aid water quality management decisions. First is to assess how more stringent discharge limits affect the fraction of dischargers able to comply with lower limits, in this case, for ammonia. Second is to test the relationship between the level of discharge limits and the creation of pools of "winners" (over-compliant dischargers) and "losers" (likely violators) who might form a market for nutrient trading to achieve water quality goals more cost effectively. Conventional regression methods are not able to achieve this objective in a direct manner. Moreover, application of random forest analysis enables determination of the relative importance of independent variables in achieving compliance with both current and future ammonia limits. The pruned regression tree provides a graphical identification of the relative importance of covariates and associated threshold values. A brief description of regression trees and random forests is provided below.

4.2.1   Regression Tree

Classification and Regression trees (CART) (Breiman et al., 1984) are local regression models, which work in a recursive manner to partition the predictor space by delineating regions where the predictions of the dependent or response variable to a set of predictors is homogeneous. These trees first group the predictors and then a particular model is assigned to each of these groupings- the simple average is most commonly used (Elith et al., 2008). Local models have the advantage of representing the true relation between the covariates

(i.e. predictors) and the response variable in each of the groups in the tree, as compared to global models where a single equation is used for the entire data set. If the response variable is continuous, then it is called a regression tree; if categorical, then it is called a classification tree. We briefly describe the CART algorithm below.

The objective of the regression tree is to minimize the sum of squared errors, S, given by the following equation:

$$S = \sum_{c \in leaves(T)} \sum_{i \in c} (y_i - \hat{y}_c)^2 \tag{4.1}$$

where the prediction of the dependent variable for the cth category (also referred to as "leaf"), given by the simple average of the observations in that category, is $\hat{y}_c = \frac{1}{n_c} \sum_{i \in c} y_i$. Equation 4.1 can be rewritten as below:

$$S = \sum_{c \in leaves(T)} n_c V_c \tag{4.2}$$

where V_c denotes the variance within the leaf c, and n_c is the number of observations in c. The splits in the tree are chosen so as minimize the value of sum of squared errors, S. The following steps provide a summary of the algorithm used to form the regression trees.

1. Using the entire data set, a threshold value of a split is determined for all of the predictors, on a basis of a selected criterion e.g., minimizing the sum of squared errors in regression trees (Equation 4.2).

2. The best predictor is chosen using the applicable criterion and constitutes the first level of split or the first node. Each node gives rise to two leaves and the value represented by each leaf is the simple average of the response variable values falling under that category.

3. Each node thus developed is further expanded following steps 1-2. The maximum size of the tree is determined by the condition used for splitting the tree. For regression

trees, the maximum size of the tree is reached when splitting a node does not reduce the sum of squared errors any further.

Trees grown to their full size tend to over-fit the data i.e. they tend to fit the training data so well that they cannot extrapolate efficiently (Elith et al., 2008). To overcome this problem, the size of the tree often needs to be limited by removing the least important splits based on some objective criterion; this process is called "pruning". In this analysis the deviance criterion is used, so the pruned 'best' tree is the one which has the least mean squared error. Unlike ordinary regression analyses, CART has the unique ability to capture the complex higher order interactions among the input variables as well as provide a visual representation of these interactions. CART has been widely used in various fields like smart energy efficient building systems (May-Ostendorp et al., 2013), wastewater treatment performance modeling (Atanasova and Kompare, 2002), landslide susceptibility studies (Youssef et al., 2016), data mining (Chrysos et al., 2013) etc. The predictors used in the regression tree are selected by determining variable importance using random forests, which is explained below.

4.2.2   Random Forests

According to Breiman et al. (2001), random forests are a combination of tree predictors where the values of an independently sampled random vector determine the configuration of each tree, which has the same distribution as all trees in the forest. So, instead of growing a single tree (like in CART) where the best split at each node is determined using all predictors, a large number of trees (or a forest) are grown by splitting each node in a tree using a randomly selected subset of predictors. This random selection of predictors ensures less correlation among the trees in the forest as well as a decrease in variance, leading to higher accuracy in predictions. Although the working principle behind random forests is as stated above, the statistical mechanism which ensures high prediction accuracy

is still being explored (Biau, 2012, Biau et al., 2008). Nevertheless, random forests have been widely recognized as a fast classifier, yielding generalization error rates comparable to the best statistical and machine learning methods (Biau et al., 2008). They have even been categorized among the most accurate general-purpose classifiers available (Biau et al., 2008). Random forests have been used in various fields like Alzheimers disease studies (Gray et al., 2013), land-cover classification (Rodriguez-Galiano et al., 2012), stock trading (Booth et al., 2014), and mortgage payment prediction (Liang and Lin, 2014) among others.

In addition to accurate prediction, random forests have the unique ability to rank the predictors by their relative importance (Breiman, 2001). Bootstrapping with replacement in a data set leaves out about 37% of the observations (Breiman, 1996). These observations are called Out-Of-Bag (OOB) observations, which are used to estimate an index for the relative importance of variables (Breiman, 2001). For N input variables, each tree is constructed after which the value of the nth variable in the OOB observations is randomly permuted. Classifications from the tree with the permuted variable are compared to the classifications from the tree using the OOB observations with all variables intact. The variable which results in the highest increase in misclassification rate due to random permutation, is defined as the most important variable. Various other measures have been used to assess variable importance like Gini importance (Friedman et al., 2001) and Permutation importance (Strobl et al., 2008). The convenience offered by random forests in variable selection has prompted several researchers to use random forests exclusively for variable selection (Diaz-Uriarte and de Andrés, 2005, Sandri and Zuccolotto, 2006). Consequently, in this research random forests alone have been used for variable selection. A regression tree is fit using the selected (continuous) variables, which is pruned to obtain the best trees. This framework efficiently combines the variable selection capability of random forests with the ease of interpretation of regression trees.
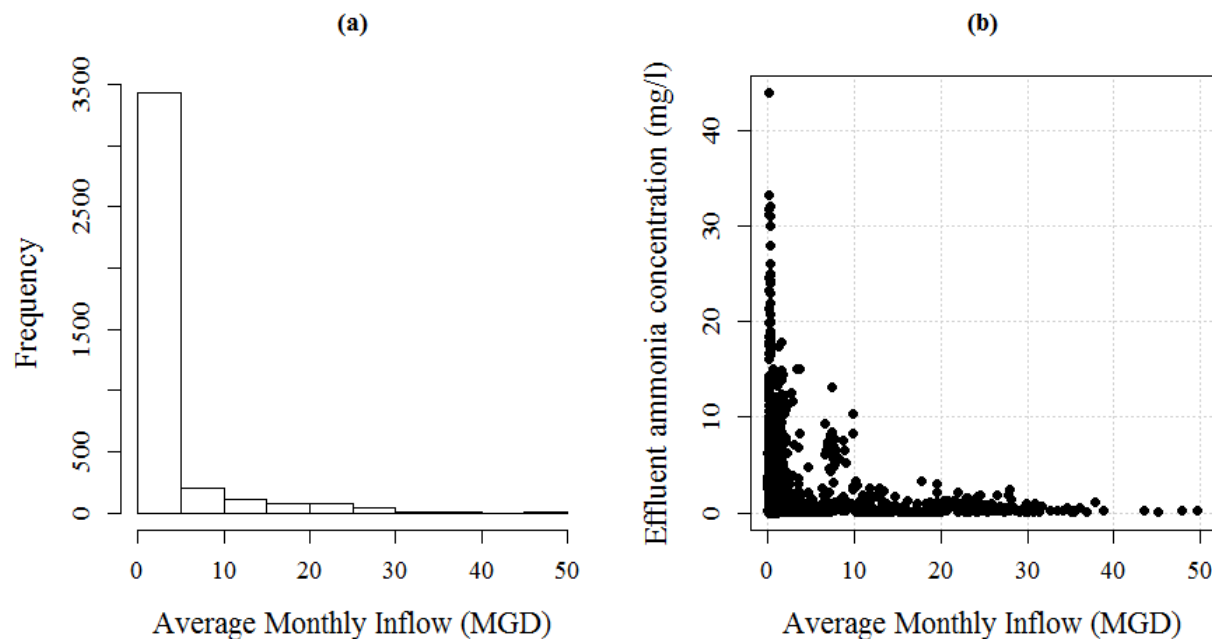
Figure 4.1: (a): Distribution of the average monthly inflow rate for 106 sample facilities (b) the DMR effluent ammonia concentration vs. average monthly plant inflow rate data

4.3 Data

The data set used in Chapter 2 is again used for this analysis. Figure 4.1 (a) shows a histogram of the average monthly plant inflow rate while 4.1 (b) shows a plot of the average monthly effluent ammonia concentration as a function of the average monthly plant inflow rate. Clearly, plants with influent flow rates < 5 MGD are far more numerous and account for a large fraction of the higher values of the DMR effluent ammonia concentrations.

The plot of monthly average effluent ammonia concentration vs. their permit limit for the 106 sample wastewater treatment facilities is shown in Figure 3.1 (b). For the sake of clarity, the observations with permit limits up to 50 mg/l have been plotted only, leaving out 0.5% of the observations with higher permit limits, as high as 215 mg/l. Most of the plants report monthly effluent ammonia well below the permit limit (shown by points lying below the 1:1 slope line), while approximately 5% of the DMR values are above the permit

limit, with increasing frequency when the limit is less than 10 mg/l ammonia-nitrogen. In this study we investigate factors distinguishing between discharges within or significantly below ammonia discharge limits and ones with effluent near or over limits and further test whether those factors are predictive of compliance with more stringent ammonia discharge limits being considered in many states.

Regression tree model validation is performed using data from a Colorado municipal wastewater treatment facility serving a population of approximately 100,000. This facility treated about 47,320 m^3/d (12.5 MGD) over the sample period and has a design capacity of 94,635 m^3/d (25 MGD). The validation data set consists of daily observations of treated wastewater quality over the period of January 1, 2010- March 31, 2014. Monthly averages for flow and effluent ammonia concentration are used for validation.

4.4   Data Analyses

The dependent variable used for fitting the regression tree is the relative concentration of ammonia in the treated effluent, defined as the ratio of the reported effluent concentration to the permit discharge limit. Relative concentration allows comparison of the absolute effluent concentration as a fraction of the discharge limit, providing a ready measure of compliance. For example, relative concentrations ≤ 1 indicate that the effluent quality meets the permit discharge limits, while a relative concentration > 1 indicates non-compliance. The data set of reported monthly average ammonia discharges is divided into two ranges of relative ammonia concentration - 0 to 0.75 (compliant plant performance) and > 0.75 (plant performance close to or above permit limit); each range is modeled individually to separately identify the factors associated with compliance or the lack thereof. It should be noted that a relative ammonia concentration of 0 indicates effluent ammonia concentrations below detection level. Of the 3888 reported ammonia discharges from 106 treatment plants across the country over the period of 2004 - 2008, 3570 observations corresponded to relative effluent ammonia concentrations in the range of 0 - 0.75, while 318 observations corresponded

to relative ammonia concentrations in the range greater than 0.75.

The predictors or independent variables include the natural logarithm of the average monthly flow rate into the plant (A), the capacity utilization, defined as the ratio of the average monthly inflow to the design capacity (C) and the product of the two (AC), which captures the non-linear interactions between A and C (Weirich et al., 2011). The relative ammonia concentration in the previous month (Lag) is used as a covariate to incorporate "persistence" in the system (Suchetana et al., 2016). These variables have been reported to be the significant determinants of ammonia concentrations relative to existing permit limits (Suchetana et al., 2016).

The analysis was performed in R (R Core Team, 2017) using the relevant packages. Random forests are used for variable selection using the 'VSURF' package (Genuer et al., 2013). The variable importance plots for each predictor variable and for both ranges of relative ammonia values are obtained using this function. The variable with the maximum importance value is selected along with those whose importance value is greater than half the maximum importance value. The selected variables are used to fit a regression tree using the 'tree' package (Ripley, 2005) which is pruned using the 'prune.tree' function, also in the 'tree' package, to get the best tree for each range of relative concentrations. The regression tree is validated using the 2010 - 2014 data from the Colorado facility.

In the above analysis, the relative ammonia concentrations are first fitted based on the existing permit limits to diagnose factors associated with current levels of compliance. However, the ability of plants under current operating conditions to comply with uniformly lower discharge limits for ammonia being considered in many states, e.g., Colorado (CDPHE, 2012), is of significant interest to utilities anticipating future costs and regulators charged with enforcement of the Clean Water Act. Hence, we also evaluate a scenario where the discharge limit for ammonia-nitrogen is lowered to 1.3 mg/l year round. The future limit is chosen as the 10th percentile of the existing discharge limits for the plants in the sample. The relative monthly average ammonia concentrations are then recalculated for each plant

	Existing standards (Until 11/30/2017) (mg/l)	Future standards (From 12/01/2017) (mg/l)
January	13.9	4.8
February	10.9	2.8
March	5.3	3.3
April	9.7	2.9
May	11.1	8.8
June	8.9	3.8
July	9.2	2.1
August	11.1	1.9
September	11.1	2.2
October	11.1	4.4
November	13.9	9.0
December	13.9	7.8

Table 4.1: Existing and future monthly permit discharge standards for the validation wastewater treatment plant

using the lower limit, and as before the data set of reported monthly discharges is divided into two ranges based on the new relative ammonia concentration values in each DMR - 0 - 0.75 and greater than 0.75. The categories now signify observations when the plant was well within the new discharge limit or when the plant was operating close to or above it, respectively. The fitted regression trees identify the variables affecting the level of compliance with more stringent ammonia discharge limits. 2714 effluent ammonia concentration values corresponded to relative ammonia concentrations in the range of 0 - 0.75, while 1126 effluent ammonia concentration values corresponded to relative ammonia concentrations in the range greater than 0.75. Lowering discharge limits leads to an increase of almost 255% of the observations falling in the near or actual violation category. As before, the predictors used in this case are A, C, AC and Lag. Suchetana et al. (2016) found that seasonality did not affect ammonia compliance when effluent ammonia limits reflected seasonal factors. However, when uniform lower limits are imposed across the entire year, the effect of seasonal factors such as temperature on nitrification are expected to result in higher relative effluent ammonia levels for many plants. To test this hypothesis, covariates for seasonality are introduced (as in Chapter 2)- the sine ($\sin(\frac{2\pi*month}{12})$) (sc) and cosine ($\cos(\frac{2\pi*month}{12})$) (cc) components. In summary, the predictor variable set for analyzing compliance with lowered ammonia discharge limits consists of A, C, AC, Lag, sine component (sc) and cosine component (cc). As before, variable selection is performed by random forest followed by fitting the regression tree and pruning.

The regression tree model based on lower anticipated ammonia limits is validated using DMR data from a 1.1 m^3/s (25 MGD) treatment facility in Colorado incorporating actual new discharge limits to take effect December 1, 2017 (CDPHE, 2012). These limits are significantly lower than the existing discharge limits, as shown in Table 4.1. Validation of the regression tree using actual discharge limits instead of the uniform value of 1.3 mg/l provides a realistic challenge for validation of the predictive performance of the regression tree model. Because we are interested in the ability of plants to meet lower discharge limits

under current operating conditions, we assume that wastewater flow rates and treated effluent concentrations remain the same as DMR values over 2010 - 2014. It should be noted that in many plants, higher levels of ammonia conversion can be achieved by changing process operations; however, even if possible, such changes would typically entail additional costs.

4.5   Results

4.5.1   Compliance with existing discharge limits

The variable importance plots, as shown in Figure 4.2 (a) and (b), are used for variable selection. For relative effluent ammonia concentration in the range ≤ 0.75, A, AC and Lag are selected as the predictor variables, while in the relative concentration range > 0.75, A and Lag are chosen as the predictors. The remaining predictors have a variable importance value less than half the maximum importance value, and are thus discarded.
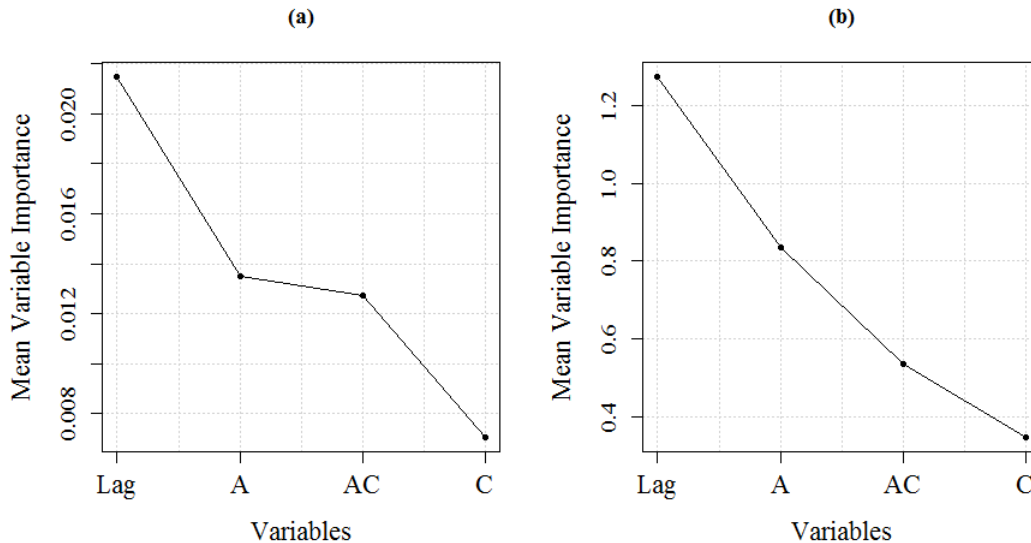
Figure 4.2: Variable importance plots from random forest analysis when existing discharge limits have been used to calculate relative ammonia concentration in the ranges of (a) 0 - 0.75 and (b) greater than 0.75

The tree can be interpreted as a series of "if-then" questions, wherein the left branch denotes the decision when the answer to the question, in this case, inequality, is true and the right branch leads to the decision when the answer to the question is false. From Figure 4.3 (a), compliance with existing discharge standards can be described as a function of 'Lag' only i.e. a function of the previous month's relative effluent ammonia concentration. The regression tree shows that when the Lag, or the previous month's relative ammonia concentration is less than 0.19, the predicted relative concentration is 0.09, or 9% of the permit limit. When Lag is greater than 0.19 but less than 0.34, the predicted relative concentration is 0.24, while it is 0.37 when Lag exceeds 0.34. Thus, the predicted relative ammonia concentration in the current month follows the trend in the previous month's discharge level. When the effluent relative ammonia is > 0.75 (Figure 4.3 (b)), the predicted relative effluent concentration is 1.44 when Lag, or the previous month's effluent relative ammonia concentration is less than 1.48. When Lag exceeds 1.48 but is less than 2.78, the predicted relative ammonia concentration is 2.75, while it is 4.32 when Lag is greater than 2.78. As before, an increase Lag values is directly related to an increase in the predicted relative ammonia concentration. It is interesting that in this higher range of relative effluent ammonia the model predicts that even plants that are in compliance the previous month (0.75 < Lag < 1) are likely to exceed their permit limit by almost 50% in the next month. It should be noted that although additional predictors are used for fitting the regression trees (as obtained from the random forests), after pruning only Lag takes precedence over the other variables. This indicates that the performance of the treatment plant in the previous month is the sole criterion in determining whether or not a plant will be in compliance with its discharge limits in the current month.
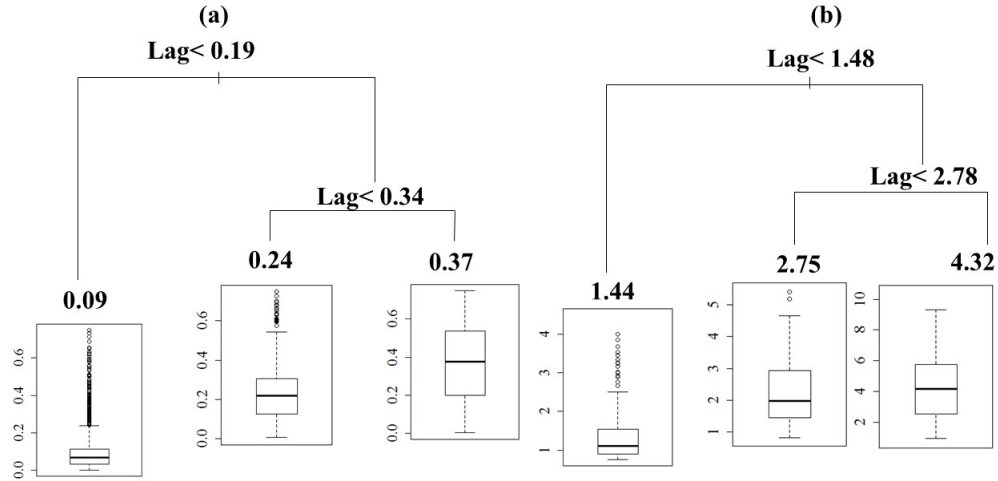
Figure 4.3: Best regression trees to model compliance with current ammonia discharge limits based on categories with previous months compliance (a) discharging ammonia with average relative concentration up to 0.75 and (b) discharging ammonia with average relative concentration greater than 0.75

4.5.2 Compliance with lowered discharge limits

As before, the variable importance plots determined by random forests are shown in Figures 4.4 (a) and (b). For relative concentrations in the range ≤ 0.75 (absolute effluent ammonia concentration $\leq 0.75*1.3$ mg/l$= 0.98$ mg/l), (Figures 4.4 (a)), Lag, A, AC and C are selected as the predictor variables for the regression trees with the lower discharge limit of 1.3 mg/l ammonia-nitrogen. Similarly, in the relative concentration range > 0.75 (absolute effluent ammonia concentration greater than 0.98 mg/l) (Figures 4.4 (b)), AC, A and Lag are selected as the predictor variables. Contrary to our hypothesis, seasonality terms (sine and cosine terms) have very low variable importance value and are not selected for fitting the regression trees.

The selected variables are used to fit a regression tree for each range of relative ammonia concentration, and these trees are pruned. The resulting trees are shown in Figures 4.5 (a)

and (b).
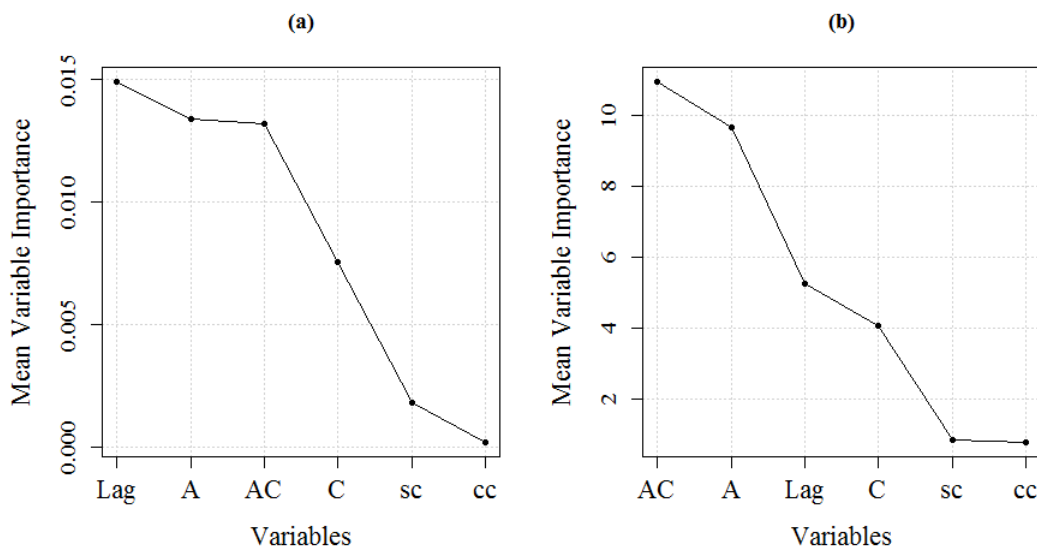
Figure 4.4: Random forest variable importance plots when a uniform discharge limit of 1.3 mg/l has been used to calculate relative ammonia concentration in the range of (a) 0-0.75 and (b) greater than 0.75

For plants discharging ammonia at up to 75% of the more stringent hypothetical uniform permit limit of 1.3 mg/l ammonia-nitrogen (up to 0.98 mg/l), Lag or the previous month's relative ammonia and A are predictive of continued compliance, although at overall higher relative ammonia values than predicted for current limits. When Lag is less than 0.12, the predicted relative ammonia concentration is 0.16. When Lag exceeds 0.12 and A is less than 2.94 (flow less than 86380 m^3/d or 19 MGD), the predicted relative ammonia concentration is 0.29, while it is 0.49 when A exceeds 2.94. Comparing Figures 4.5 (a) and 4.3 (a), the new predicted relative concentration values are higher: 0.16, 0.29, and 0.49, compared with 0.09, 0.24 and 0.37, respectively, when the discharge limits are lowered. For plants discharging at greater than 75% of the 1.3 mg/l limit (greater than 0.98 mg/l), Lag is predictive of the performance in the current month (Figure 4.5 (b)). When Lag is less

than 0.73, the predicted relative ammonia concentration is 2.55, else it is 6.02. Comparing Figures 4.5 (b) and 4.3 (b) we can see that the predicted relative ammonia concentration is significantly higher in case of the former, when the discharge limits have been lowered. There is an increase of almost 40% in the maximum predicted value of relative ammonia concentration. Once again, even though there are additional predictors for the regression tree, only Lag and A appear in the best tree in Figure 4.5 (a) while only Lag appears in the best tree in Figure 4.5 (b). As before, even when discharge limits have been significantly lowered, Lag plays an important role in determining whether or not a treatment plant will be in compliance with the reduced national regulations of 1.3 mg/l in the current month.
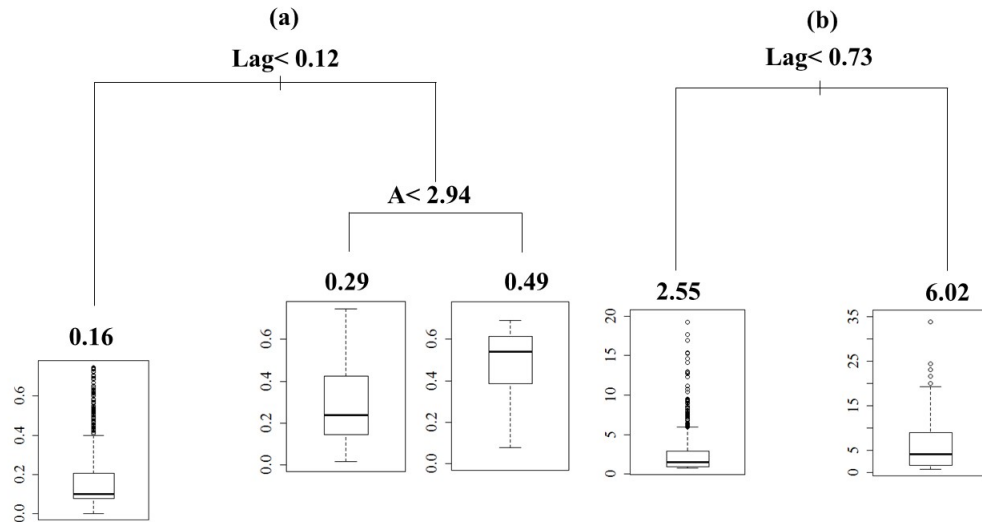
Figure 4.5: Best regression trees when a uniform discharge limit of 1.3 mg/l is used to calculate relative ammonia concentration for DMR ammonia data in the ranges of (a) 0 to 0.75 and (b) greater than 0.75

4.6 Model validation

4.6.1 In-sample validation

As the name suggests, in-sample validation is performed on the data used to fit the regression trees in Figures 4.3 and 4.5. To test the predictive performance each model shown in Figures 4.3 and 4.5 is validated in a drop one cross validation mode, Methodologically, this is similar to the drop 10% cross-validation method described in Chapter 2, except instead of dropping a randomly selected set of 10% of the observations, we drop all observations corresponding to one treatment plant. The best tree is fit and pruned using the remaining observations and this tree is used to predict the relative ammonia concentration for the dropped observations. This is repeated 500 times, randomly selecting a different treatment plant for prediction each time, and the predictions are compared with the observations. Conventional regression models provide a unique prediction for every observation, which facilitates comparison of observed and predicted quantities using measures like R^2. But regression trees provide only one relative ammonia prediction for each predicted category. This necessitates a slightly different approach for comparing the observed and predicted values. To that end, the observed and predicted values of relative ammonia concentration are divided into categories. For example, for the regression tree in Figure 4.3 (a), the observed relative ammonia concentrations in the validation data set are categorized into 3 groups: Category 1 for relative effluent ammonia concentration less than or equal to 0.1 mg/l, Category 2 for relative ammonia concentration greater than 0.1 but less than or equal to 0.3 and Category 3 for relative ammonia concentration greater than 0.3 but less than or equal to 0.75. The best regression tree is used to predict relative ammonia concentrations on the dropped points and based on the value of the predictions, they are classified into one of the 3 categories mentioned above. This is repeated 500 times and the percentage of correct categorization is plotted in a boxplot for each of the trees. For the regression tree in Figure 4.3 (b), Category 1 indicates a relative ammonia concentration less than or equal to

1.5 while Category 2 indicates relative ammonia concentration greater than 1.5 but less than or equal to 3. Category 3 indicates relative ammonia concentrations greater than 3. Using this validation method, the boxplot for percentage correct categorization corresponding to regression trees in Figures 4.3 (a) and (b) are shown in Figure 4.6 (a) and (b) respectively. The regression tree model shown in Figure 4.3 (a) provides a median of about 70% correct categorization in Figure 4.6 (a) and the model in Figure 4.3 (b) provides about 75% correct categorization in Figure 4.6 (b), and a maximum of 100% correct categorization in both cases.
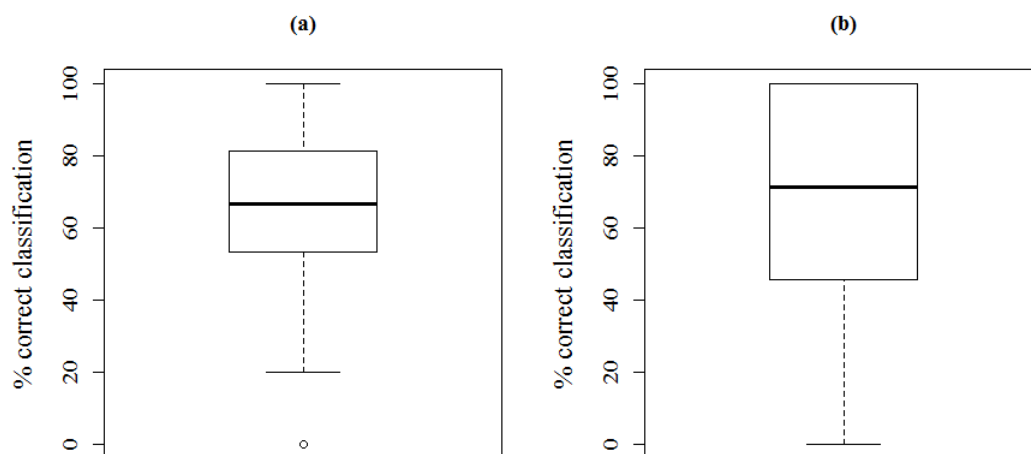
Figure 4.6: Boxplots for percentage of correct categorization obtained in the drop one plant cross validation mode for plants with current ammonia discharge limit in the relative concentration range of (a) 0-0.75 and (b) greater than 0.75

The percentage of correct categorization boxplots for the trees in Figure 4.5 (a) and (b) are shown in Figure 4.7. For the regression tree shown in Figure 4.5 (a), Category 1 indicates a relative ammonia concentration less than or equal to 0.2, Category 2 indicates a relative ammonia concentration greater than 0.2 but less than or equal to 0.35 while

Category 3 indicates a relative ammonia concentration greater than 0.35 but less than or equal to 0.75. For the regression tree shown in Figure 4.5 (b), Category 1 indicates a relative ammonia concentration less than or equal to 3 while Category 2 indicates relative ammonia concentration greater than 3. The median percentage of correct categorization is about 65% and 99% for the trees shown in Figure 4.5 (a) and (b) respectively, with a maximum value of 100%. This indicates that the regression tree models shown in Figures 4.3 (a) and (b) and Figures 4.5 (a) and (b) are robust tree regression models.
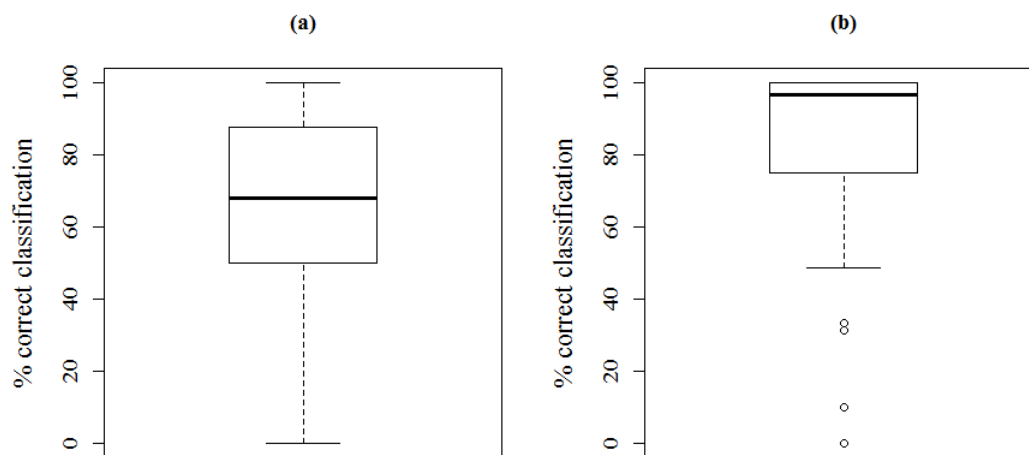
Figure 4.7: Boxplots for percentage of correct categorization obtained in the drop one plant cross validation mode when a uniform discharge limit of 1.3 mg/l is used to calculate relative ammonia concentration for DMR ammonia data in the ranges of (a) 0 to 0.75 and (b) greater than 0.75

4.6.2   Out-of-sample validation

As the name suggests, out-of-sample validation is performed on a dataset outside the one used for fitting the regression trees. This is done to test how well the regression trees perform in prediction mode using out-of-sample data from an individual treatment plant.

Out-of-sample validation of the regression tree models is performed using four years of effluent ammonia and flow data from an operating wastewater treatment facility in Colorado. Out-of-sample evaluation of forecasting accuracy is fairly popular (Campbell and Thompson, 2008, Fisher et al., 2015, Meese and Rogoff, 1983), mainly because forecasting errors are understated by in-sample errors and the best in-sample fit may not be able to accurately predict post-sample data (Bartolomei and Sweet, 1989, Pant and Starbuck, 1990, Tashman, 2000). Fildes and Makridakis (1995) were of the opinion that model performance on out-of-sample data was the benchmark for its utility in all applications. Thus, in addition to in-sample validation, out of sample validation of our regression tree would help obtain a more robust and comprehensive estimate of the predictive ability of the model.

The relative effluent ammonia concentration for the 2010 - 2014 validation data set was in the range of 0-0.75 when existing discharge limits are used as well as when lower proposed discharge limits are used. Thus out of sample model validation is performed using only regression trees shown in Figures 4.3 (a) and 4.5 (a). Seasonally varying discharge limits were used for calculating the relative ammonia concentration while validating the prediction accuracy for the regression tree in Figure 4.5 (a). This was done to evaluate the accuracy of the predictions from the regression tree for plants where stricter discharge limits different from 1.3 mg/l ammonia-nitrogen are to be implemented in the near future. The model performance thus evaluated will provide us with a more conservative estimate of the predictive performance and will test its predictive ability when actual enforceable discharge limits are used. The observed and predicted categories are same as the ones used for generating Figures 4.6 (a) and 4.7 (a).

The regression tree in Figure 4.3 (a) was validated using relative ammonia concentrations from the Colorado wastewater treatment facility, calculated using current permit limits. The percentage accuracy of correct classification was found to be 90%. Similarly, the regression tree in Figure 4.5 (a) was validated using relative ammonia concentrations from the Colorado facility calculated using future (lowered) permit limits. In this case the

percentage accuracy of correct classification was 77%. Clearly, the regression trees in Figures 4.3 (a) and 4.5 (a) can be described as fairly robust diagnostic models, which provide more than 77% median predictive accuracy in in-sample as well as out-of-sample prediction for individual treatment plant data where lowered discharge standards are to be implemented.

4.7   Discussion

The importance of Lag as the chief predictor in determining compliance/ non-compliance with current and future nutrient regulations, although intuitive, is a significant finding of this study. It demonstrates that the compliance history of a treatment plant is the most important driver in determining its ability to meet with nutrient regulations. For utilities of all sizes, this finding could mean that continuous monitoring of plant performance should help ensure compliance in the subsequent months. For enforcement agencies looking to predict which facilities could be in violation of nutrient regulations, monitoring the compliance history of the plant could provide a reasonable indication of future violations. Even if nutrient regulations are significantly lowered nationally, closely monitoring the compliance history of a plant and quickly responding to process upsets can help ensure plant compliance.

No doubt factors such as hydraulic and solids retention time (HRT and SRT) play a role in treatment performance (Oliveira and Von Sperling, 2008). However, the intent of the study was to examine whether other factors which operate across a broader range of process design and operation conditions also are important in determining compliance. Niku and Schroeder (1981a) advanced the concept of treatment "stability" which they analyzed using the standard deviation values for effluent BOD and suspended solids data for 43 one-year data sets from US activated sludge plants. They concluded that plant capacity was not a significant determinant of stability; although extended aeration plants, which tend to serve smaller communities, had low mean and standard deviation values for both constituents. Conversely, using a Generalized Linear Model (Weirich et al., 2011) found that small plants were more likely to violate permit limits for BOD and ammonia than larger plants. Our

regression tree analysis confirms that other factors such as plant size (represented by flow rate), prior performance and seasonality explain a significant portion of the variability in discharge compliance data for ammonia. Moreover, validation using both within-sample and outside sample data sets indicate that the regression tree models provide a robust diagnosis of high and low-performing dischargers based on the pruned set of independent variables.

(a)

(b)
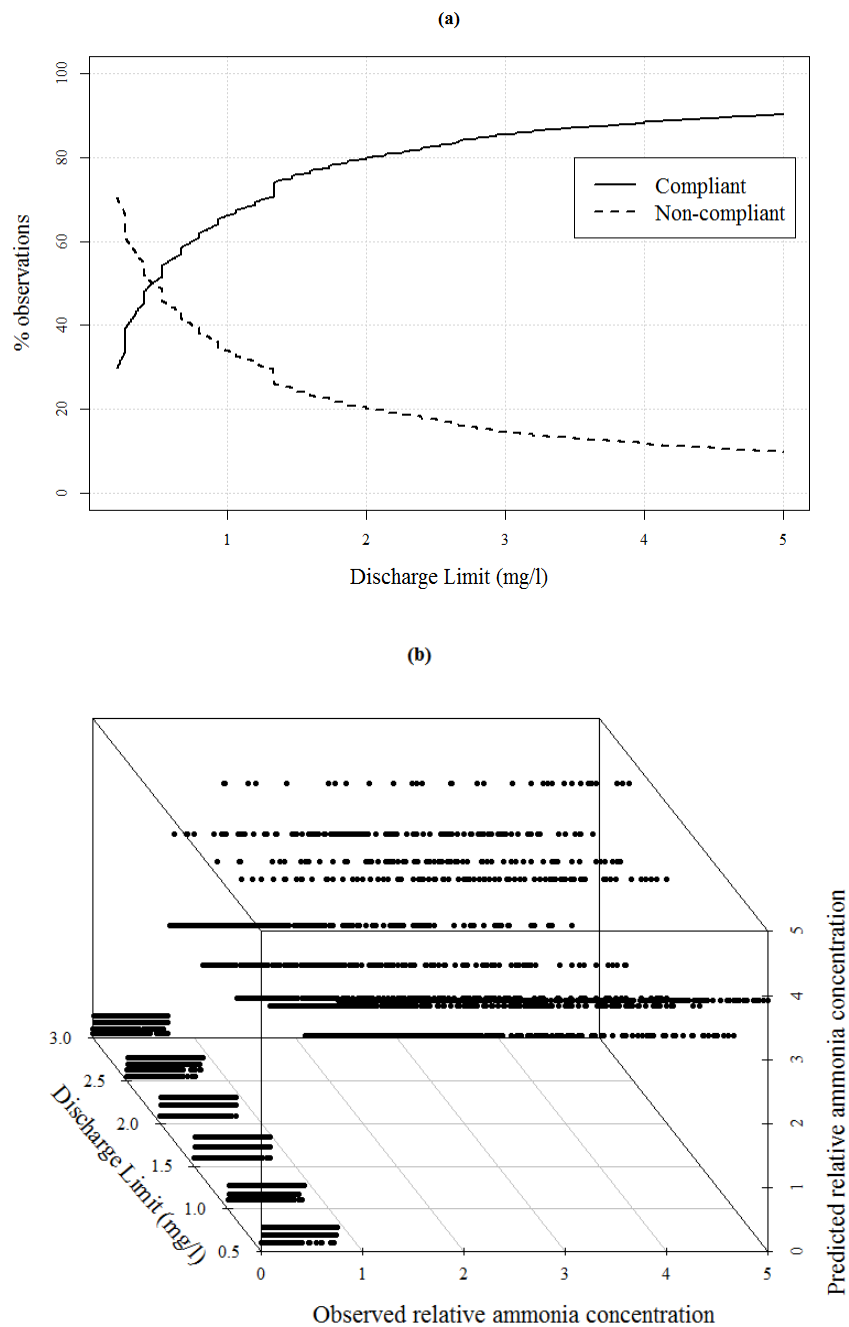
Figure 4.8: (a): Percentage of monthly observations in the compliant (relative ammonia concentration less than or equal to 0.75) and non-compliant categories (relative ammonia concentration greater than 0.75) versus universal discharge limit (b) 3D scatterplot of predicted and observed relative ammonia concentration versus universal discharge limit

The regression tree model can be potentially used as a tool to assess the feasibility of a point source-to-point source nutrient credit trading system on a watershed or statewide scale. Ammonia discharges in the relative concentration range of 0 to 0.75 indicate "over-compliance" with temporal consistency, a potential basis for the supply side of a nutrient trading market. Dischargers with relative concentrations exceeding 0.75 are potentially non-compliant, which could act as the demand side of the trading scheme. Dennison et al. (2012) describe the importance of documentation and reliability to insure reliable markets. Uncertainty in meeting changing nutrient limits can jeopardize the reliability of such markets, so it is important to know how plant performance would be affected by lowered nutrient limits and which variables would significantly impact compliance under proposed new regulations. Figure 4.8 (a) shows the percentage of observations in the compliant (relative ammonia concentration less than or equal to 0.75) and non-compliant (relative ammonia concentration greater than 0.75) categories for various hypothetical uniform national discharge limits. For a linear change in discharge limits, the percentage of observations falling in each category changes almost exponentially. This graph gives us a good estimate of how the demand and supply population would change for different discharge limits. Another potential use of the proposed regression tree methodology in facilitating point source-to-point source is demonstrated in Figure 4.8 (b). For each hypothetical uniform discharge limit in the range of 0.5-3 mg/l, the relative ammonia concentrations are calculated and the regression tree is fit against A, C, AC and Lag (as seasonality has been shown to have negligible variable importance value). For each case, the fitted tree is pruned and the best regression tree is used to predict relative ammonia concentrations for the relative concentration ranges of 0 to 0.75 and 0.75 and higher. The predicted relative ammonia concentration is plotted against the observed relative ammonia concentration and the hypothetical uniform national discharge limit in a three-dimensional scatterplot. The plot clearly delineates the observations in the 'demand' side (predicted relative ammonia concentration greater than 0.75) and a 'supply' side (predicted relative ammonia concentration less than or equal to 0.75) for a

nutrient trading scheme with wastewater utility participants. Additionally, the predicted relative ammonia concentration values for each category can help assess the reliability of a 'demand' or 'supply' sources. For example, observations where the predicted relative ammonia concentration consistently lies below the threshold level of 0.75 for various discharge limits between 0.5-3 mg/l, can be classified as a reliable supply for nutrient credits, while those consistently exceeding 0.75 can be categorized as a reliable demand for credits. It should be noted that as Lag is one of the most important variables in the regression tree, it indicates that there is a homogeneity or persistence in plant performance. Consequently, the wastewater treatment plants whose observations lie in the reliable supply category, can be characterized as a consistent nutrient credit supply source. Similarly, treatment plants with observations in the non-compliant category can be characterized as the nutrient trade demand source. In addition to indicating the feasibility of a nutrient trading scheme, the regression tree framework can support a trade-off approach to establishing future ammonia discharge limits, which would allow an acceptable degree of compliance that protects water quality within the capacity of existing infrastructure.

Researchers have studied economic explanations for over-compliance with wastewater discharge limits. Shimshack and Ward (2008) found that U.S.EPA enforcement actions were associated with reductions in discharges by paper and pulp processing plants already in compliance with limits for BOD and total suspended solids. They attributed the incentive to over-comply to two factors: volatility in stochastic effluent BOD along with greater perceived risk of fines due to more enforcement activity on non-compliant plants; and jointness, which is a result of common treatment technology for BOD and TSS producing simultaneous over-compliance for both constituents at relatively low cost. Thus, in addition to providing economic incentive as a nutrient credit seller, over-compliance can also obviate enforcement actions. The proposed regression tree framework can provide easy operational guidelines to treatment facilities looking to avoid the economic risks associated with anticipated consequences of enforcement actions, by ensuring consistent over-compliance.

4.8   Conclusions

The regression tree model results suggest that prior compliance (previous months efflu-
ent ammonia relative to permit limits) is a significant determinant of subsequent ability to
meet existing and future lower discharge limits. Plants currently discharging less than 75%
of their ammonia permit limit over a range of < 1 to 215 mg/l ammonia-nitrogen are likely
to remain in compliance even when a significantly lower limit of 1.3 mg/l ammonia-nitrogen
is imposed. Plants currently discharging at greater than 75% of the ammonia permit limit
are likely to exceed permit limits in the following reporting month by more than 50%. When
the discharge limit is lowered to 1.3 mg/l ammonia-nitrogen, the magnitude of permit ex-
ceedences in this category increases by approximately 50%. In- and out-of-sample validation
of regression tree models had prediction accuracy of over 70%. Possible applications of
the regression tree framework include identification of potential buyers and sellers in nutri-
ent trading scheme with wastewater utility participants and assessment of the adequacy of
wastewater treatment infrastructure to meet a range of limits.

Chapter 5

Modeling Total Inorganic Nitrogen in Treated Wastewater Using Non-Homogeneous Hidden Markov and Multinomial Logistic Regression Models

A version of this chapter has been submitted to Science of the Total Environment as a research article.

Abstract

Total Inorganic Nitrogen (TIN) in treated wastewaters, obtained as the sum of efflu-ent ammonia-nitrogen, nitrate-nitrogen and nitrite-nitrogen concentrations, is the common regulatory indicator of the nitrogen removal performance of a treatment process. Recent regulations in some states have set lower discharge limits for TIN, recognizing the environ-mental and health impacts of these species. Although permit limits may consist of annual average or median values, temporal variability in effluent TIN concentration is common. We created a modeling framework using Hidden Markov Models and multinomial logistic regres-sion using data from an operating wastewater treatment facility located in Colorado, USA, over the period of January 1, 2010- March 31, 2014. In the two-step modeling approach, the Hidden Markov Models capture regime shifts in effluent TIN concentrations and multinomial logistic regression identifies prominent factors responsible for the regime shifts. Simulations indicated that climate factors, seasonality, prior performance and effluent ammonia concen-tration are predictive of variability in effluent TIN concentration. The predicted temporal

variability suggests that TIN limits based on annual average or median values may not be effective in protecting receiving water quality.

5.1   Introduction

Discharge of nitrogen constituents in treated wastewater is typically regulated on the basis of local water quality conditions for protection of aquatic life and domestic water supplies. Historically, limits for total ammonia nitrogen (TAN) reflect risks of toxicity to aquatic life and contributions to eutrophication and dissolved oxygen consumption. Removal of ammonia is readily achieved in conventional aerated secondary treatment processes. Chapters 2-4 of this dissertation discuss performance-based modeling strategies to assess compliance with current and future (lowered) ammonia discharge limits. More recently, broader regulation of all nitrogen species is being considered recognizing the diverse impacts of nitrate and other oxidized nitrogen species on drinking water treatment as well as aquatic environment. Total inorganic nitrogen (TIN) is typically measured as the sum of TAN, nitrite-N and nitrate-N, and nitrate usually is the principal component in secondary effluent for plants achieving nitrification (Henze, 1991, Sattayatewa et al., 2009). Pocernich and Litke (1997) examined the relationship between effluent TAN and nitrate+nitrite-nitrogen for treatment plants discharging into the South Platte River in Colorado using linear regression for data from 40 plants with highly variable levels of effluent ammonia (1 to 18 mg/l) and nitrate+nitrite (0.02 - 19.8 mg/l) and found a negative relation between the two variables ($R^2 = 0.78$). However the relatively small data set (n = 12) produced large prediction errors (95% confidence interval estimated to be $\pm$ 6 mg/l nitrate and nitrite-N).

For more advanced treatment to reduce total dissolved nitrogen species, removal of nitrate requires maintenance of an anoxic region in biological treatment, and often addition of an organic carbon source. Thus achieving low levels of TIN in treated wastewater requires additional equipment, more complex operation, and additional costs (Symbiont, 2011, Oleszkiewicz and Barnard, 2006, CDPHE, 2011).

Because of the increased regulatory focus on control of TIN discharged by wastewater treatment facilities in the US and other countries, in this chapter, we were motivated to investigate the feasibility of a performance-based model to predict effluent TIN concentration using data from an operating wastewater treatment facility in Colorado, United States over the period of January 1, 2010- March 31, 2014. This facility has treated an average flow of 47,320 m^3/d (12.5 MGD) over the sample period and has a design capacity of 94,635 m^3/d (25 MGD), serving a population of approximately 114,000. This data set consists of daily average influent flow rates and measures of effluent concentrations of total ammonia-N (TAN), nitrate-N, and nitrite-N from daily grab samples. Previous research has shown a strong temporal persistence in wastewater plant treatment behavior on a monthly average basis (Suchetana et al., 2016, 2017, Weirich et al., 2015a). Because of a number of missing daily data, we were able to analyze persistence in effluent dissolved nitrogen species on a weekly average basis. TIN concentration is calculated as the sum of 7-day averages of TAN, nitrite and nitrate concentrations. Figure 5.1 shows the time series plot of average weekly effluent TIN and ammonia concentrations. The effluent TIN concentration time series (denoted by the solid black line) exhibits regime-like behavior (TIN concentration trend indicated by solid red line), with consecutive weeks when the effluent TIN concentrations show a (positive or negative) departure from the mean or average concentration (denoted by the horizontal black line). Effluent ammonia concentrations are indicated by the solid blue line. Regulation #85 TIN limit is the running annual median < 15 mg/l and 95th percentile of samples < 20 mg/l . However, local use-protected water quality standards also apply to wastewater discharge limits, with the lower standard determining the actual permit limit. For the facility in our study, beginning December 1, 2017 the a daily maximum nitrate-N limit is 14.7 mg/l, and the effluent 30-day average TAN limit varies from 1.9 to 5.2 mg/l and associated daily maximum ranges from 1.9 to 9.0 mg/l. Associated daily maximum values for TIN range from a low of 16.6 (August) to a monthly maximum of 23.7 mg/l (November). On a seasonal basis, during low flow months which constitute 50% of the year (February April,

and July - September) the daily maximum TIN limit for the plant ranges from 16.6 to 18.0 mg/l, significantly more stringent than the annual median and 95th percentile requirements in the statewide regulation. It should be noted that whenever TIN concentrations exceed the mean concentration for 6 consecutive weeks or more , we find a corresponding spike in effluent ammonia concentrations as well. This could indicate a possible relation between effluent TIN and effluent ammonia concentrations.
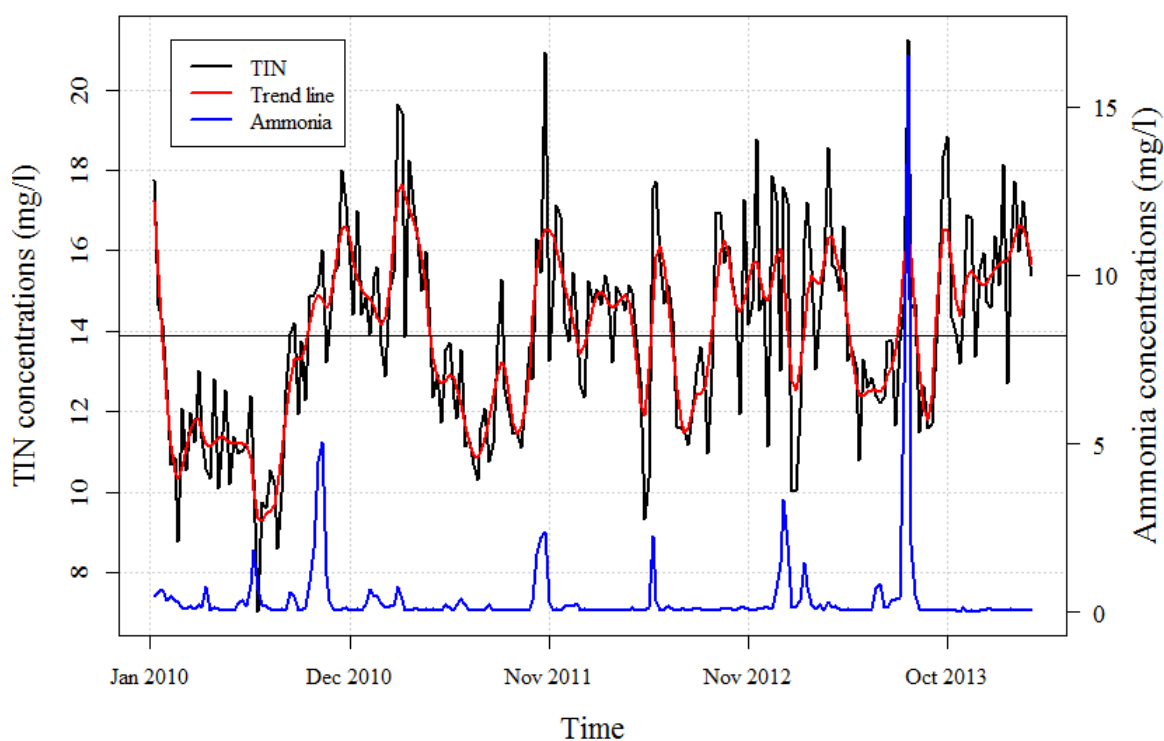
Figure 5.1: Average weekly effluent TIN (indicated in black) and ammonia (indicated in blue) concentrations, as obtained from an operating wastewater treatment facility in Colorado, US over the period of January 1, 2010- March 31, 2014. The solid horizontal line represents the average effluent TIN concentration for the entire period. The solid red line represents the (smoothed) TIN concentration trend line

Traditional time series (Tao and Delleur, 1976, Salas, 1980, Bras and Rodriguez-

Iturbe, 1985, Carlson et al., 1970, O'Connell, 1974) models- Auto-Regressive Moving Average (ARMA) methods cannot incorporate these extended high and low concentration regimes, due to their short term memory and stationarity (Bracken et al., 2014). Furthermore, ARMA models assume normal distribution of the errors, thus, only producing smoothed and stationary spectral characteristics (Salas, 1980, Stedinger and Taylor, 1982, Lawrance and Kottegoda, 1977). Nonparametric time series methods applied for streamflow modeling such as block bootstrap (Ouarda et al., 1997), kernel density-based methods (Tarboton et al., 1998, Sharma et al., 1997), K-nearest neighbor bootstrap (Lall and Sharma, 1996, Nowak et al., 2010), modified K-NN (Prairie et al., 2006) etc. address the assumption of normal distribution well, but cannot capture nonstationarity. Recently, Erkyihun et al. (2016) developed wavelet-based methods coupled with large scale climate features to model and simulate nonstationarity in streamflow.

Hidden Markov Models (HMMs) have been shown to be quite effective in modeling regime-like behavior, nonstationarity and also non-normality. They have been applied to a variety of time series applications such as - rainfall (Greene et al., 2011, Jackson, 1975, Mehrotra and Sharma, 2005, Zucchini and Guttorp, 1991), streamflow simulation (Akintug and Rasmussen, 2005, Gelati et al., 2010), speech recognition (Abushariah et al., 2010, Huang et al., 1990, Rabiner, 1989), finance (Bulla, 2011, Hamilton, 2010), bioinformatics (Wheeler et al., 2014, Shihab et al., 2013, Narasimhan et al., 2016) etc. Bracken et al. (2014) developed a logistic regression to model the time-varying characteristics of two regime transitions making it a non-homogeneous Markov Chain, and incorporated Gamma probability density function (PDF) for the regimes. They developed this framework for stochastic simulation of streamflow in the Upper Colorado River Basin. They found that this methodology was not only able to model the regime shifting behavior of the streamflow data, but could also successfully capture all the distributional statistics and spectral properties of the data.

We propose a methodology based on the combination of Hidden Markov Models and multinomial logistic regression in the first methodological attempt to build a simulation

strategy to capture variations in effluent TIN concentration as well as the factors responsible for the variability. TIN concentrations have not been modeled to date, primarily because nitrogen discharge limits have traditionally been based on ammonia. However, the State of Colorado introduced a new uniform statewide discharge limit for existing and new wastewater treatment plants with flows greater than 3,785 m³/d (1.0 MGD) (CDPHE, 2012). Previous studies have found that pollutants like Biochemical Oxygen Demand (BOD), Total Suspended Solids (TSS) and ammonia can be modeled using performance-based variables (like inflow, fractional use of design capacity) (Suchetana et al., 2016, 2017, Weirich et al., 2011, 2015a), effluent constituent concentration in the previous month (Suchetana et al., 2016, 2017, Weirich et al., 2015a), seasonality (Suchetana et al., 2016) and climate (Suchetana et al., 2016). Additionally, it is of interest to understand the relationship between variations in effluent ammonia concentrations and regime shifts in effluent TIN concentrations in order to allocate the portion of TIN removal accounted for by nitrification versus denitrification. Using this information, we first fit a Hidden Markov Model to effluent TIN concentrations to estimate the regime shifting patterns. Subsequently, we fit a multinomial logistic regression model to investigate the dependence between the regime shifts and the various possible causes. The proposed model not only helps simulate TIN concentrations, but also brings to light several issues regarding reliability of performance and compliance with discharge limits.

5.2   Methodology

5.2.1   Hidden Markov Model

Hidden Markov models (HMMs) (also known as Markov switching models, Markov mixture models or dependent mixture models) are models where an observation is generated by a distribution which depends on the state of an underlying and unobserved Markov process (Zucchini and MacDonald, 2009). A brief description of the Hidden Markov Model methodology is provided here, but readers should refer to Zucchini and MacDonald (2009)

for details. A Hidden Markov model is said to be of order 'm' when it switches or transitions between 'm' hidden states or regimes according to a discrete Markov chain. The conditional probabilities of switching or transitioning between the 'm' hidden states is given by the Transition Probability Matrix (TPM), $\boldsymbol{\Gamma}$. Each state can be described by a probability distribution, whose parameters are dependent on the state of the Markov process. This probability distribution is referred to as a component distribution. For an observed sequence X_t, where $t = 2, 3, ..., T$, the HM model is represented in the simplest form, as below:

$$Pr(S_t|\boldsymbol{S}^{t-1}) = Pr(S_t|S_{t-1}), t = 2, 3, ..., T \tag{5.1}$$

$$Pr(X_t|\boldsymbol{X}^{t-1}, \boldsymbol{S}^{(t)}) = Pr(X_t|S_t), t \in \mathbb{N} \tag{5.2}$$

where S_t is the unobserved or "hidden" sequence which follows a simple first-order Markov process, $\boldsymbol{S}_t$ is the sequence of "hidden" states $S_1, S_2, S_3, ..., S_T$. $\boldsymbol{\Gamma}$ is the Transition Probability Matrix given by:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{bmatrix} \tag{5.3}$$

where any element of $\boldsymbol{\Gamma}$ is given by $\gamma_{jk} = Pr(S_{i+1} = k|S_i = j)$. When $\boldsymbol{\Gamma}$ changes from one time-step to another, the model is referred to as a non-homogeneous Hidden Markov model. Hidden Markov models have commonly used both normal (Akintug and Rasmussen, 2005, Gelati et al., 2010) and gamma component distribution (Al-Saleh and Agarwal, 2007, Bracken et al., 2014, Evin et al., 2011). However, our model deals with positive non-zero

TIN concentrations; hence a gamma component distribution is better suited. The probability distribution of the gamma function is given by:

$$p_i(x) = g(x; k_i, \theta_i) = \frac{\theta_i^{k_i}}{\Gamma(k_i)} x^{k_i-1} e^{-\theta_i x} \qquad for \quad x \geq 0 \qquad (5.4)$$

where k_i is the state-dependent shape parameter, θ_i is the state-dependent rate parameter and Γ is the gamma function ($\Gamma(k) = (k-1)! \; \forall k \geq 0$ and $\theta > 0$).

5.2.2   Parameter Estimation

Various methods are used for parameter estimation of Hidden Markov Models, e.g. direct maximization of the likelihood function (Akintug and Rasmussen, 2005, Zucchini and MacDonald, 2009) and Bayesian estimation methods (Thyer and Kuczera, 2003, 2000). We have used another popular method, the Expectation Maximization (EM) algorithm (Dempster et al., 1977) for maximum likelihood estimation when some data (the hidden states, in this case) are are missing. In the context of HMMs, the EM algorithm is known as the Baum-Welch algorithm (Baum et al., 1970), which is used for parameter estimation of an HMM where the Markov chain is homogeneous but not necessarily stationary. Thus, the parameters estimated by the Baum-Welch algorithm are the state-dependent distributions, the transition probability matrix $\boldsymbol{\Gamma}$ as well as the initial distribution $\boldsymbol{\delta}$. The assumption for stationarity, is not used here; assuming non-stationarity allows the expected state to change in time. The criteria for initial guesses of the above mentioned parameters, to be used in the EM algorithm, were as estimated by Bracken et al. (2014). The analysis was performed in R (R Core Team, 2017) using relevant packages. The Hidden Markov model is fit using relevant functions from the 'HiddenMarkov' library (Harte, 2016).

5.2.3   Model Order selection

Selection of optimal model order, i.e., the number of hidden states or component distributions, can be done using the Akaike Information Criteria (AIC) (Akaike, 1974) or the

Bayesian Information Criteria (Schwarz et al., 1978). BIC penalizes models having more parameters, encouraging parsimony; hence, the best model is selected as the one with the minimum BIC value.

5.2.4   Global decoding

A very brief account of the decoding method is provided here; details can be found in Zucchini and MacDonald (2009). The hidden state sequence is obtained by using the Viterbi algorithm (Forney, 1973), which recursively maximizes the conditional probability of the sequence, given the observations.

$$Pr(\boldsymbol{S}^{(T)} = \boldsymbol{s}^{(T)} | \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}) \qquad (5.5)$$

We thus obtain the most likely state sequence $s_1, s_2, s_3, ..., s_T$, known as the global decoding. This is performed in R, using relevant functions from the 'HiddenMarkov' library (Harte, 2016).

5.2.5   Multinomial Logistic Regression

Using the globally decoded state sequence, we fit a multinomial logistic regression model (Friedman et al., 2001). The predictions from the model fit provide us with the probability that the system is in a particular state. When the state sequence consists of only 2 states, as in our case, the multinomial logistic regression reduces to a binary logistic regression (Friedman et al., 2001). The appropriate Generalized Linear Model for the binary logistic regression is:

$$\eta_t = ln\frac{\pi_t}{1 - \pi_t} = \beta_0 + \beta.\boldsymbol{x}_t \qquad (5.6)$$

where η_t, the logit link function, is the probability that the system was in state 2 in week t, β_0 is the intercept term, β is a vector of the logistic regression parameters and $\boldsymbol{x}_t$ is a vector

of covariates in week t. The probabilities are estimated by rearranging the terms in equation 5.6:

$$\pi_t = \frac{\exp(\beta_0 + \beta.\boldsymbol{x}_t)}{1 + \exp(\beta_0 + \beta.\boldsymbol{x}_t)} \tag{5.7}$$

For this analysis, binary logistic regression is performed using relevant functions from the 'nnet' library (Ripley, 2002), which fits multinomial log-linear models using neural networks. The best set of covariates is selected using the 'stepAIC' function from the MASS package (Ripley, 2002), which performs a step-wise model selection using AIC or BIC (we have used BIC, due to its parsimonious selection of parameters).

5.2.6   Simulation from the HMM- Logistic Regression Model

Traditional simulation from HMMs involves using the transition probabilities to generate state sequences and randomly generating values from the corresponding probability distributions. However, to fully capture the variability present in the covariates, we generate simulations from the state model (multinomial logistic regression) predictions instead. Robertson et al. (2004) state that incorporation of covariates for predicting the expected state sequence, effectively allows the transition probabilities to change in time. Thus, this kind of model set up is comparable to a non-homogeneous HMM (Robertson et al., 2004). To obtain an ensemble of simulations from this model, first we generate a uniform random number, $r = U(0, 1)$. From the state probabilities predicted by the multinomial logistic regression model, the sequence of partial sums is calculated for each time t, as $C_j = \sum_{i=0}^{j} \pi_{i,t}$. State estimates $(\hat{s}_t)$ can be calculated by determining the sample rank of the first element in the vector $[r, C_1, C_2, ..., C_m]$. Based on the value of $(\hat{s}_t)$, a random number is drawn from the corresponding component distribution. These steps are repeated multiple times (1000 times for this study) to get an ensemble of simulations.

5.3   Data

As discussed in Chapter 4, data from a wastewater treatment facility in Colorado has been used for this analysis. The Hidden Markov Model is fitted using TIN concentrations and a gamma component distribution. The multinomial logistic regression model is fit using a set of covariates which include the average monthly flow rate into the plant, in MGD (A), the capacity utilization (C), defined as the ratio of the average monthly inflow to the design capacity, and the product of the two (AC), which captures the non-linear interactions between A and C (Weirich et al., 2011, Suchetana et al., 2016, 2017). The effect of climate on TIN removal is incorporated by the maximum and minimum average weekly temperature (T_{max} and T_{min} respectively) and the average weekly precipitation terms (Precip). Seasonal variations in nitrification are incorporated using the sine ($\sin(\frac{2\pi * week}{52})$) and cosine ($\cos(\frac{2\pi * week}{52})$) components ('sc' and 'cc' respectively), which capture cyclical changes and phase shifts in the response variable (Kleiber et al., 2012, 2013, Suchetana et al., 2017, 2016). The "persistence" or "inertia" in the system is captured using the previous week's simulated state obtained from the Hidden Markov model (Lag) (Weirich et al., 2015a, Suchetana et al., 2016, 2017). As demonstrated in section 5.1, to explore the effect of effluent ammonia on effluent TIN concentrations, we include effluent TAN concentration as the ninth covariate (Amm).

5.4   Results

A second order model i.e., two gamma distributions, is selected based on fitting the Hidden Markov model to effluent TIN concentration data and application of the BIC criteria. Model parameters are shown in Table 5.1 along with the Transition Probability Matrix and initial distribution. These two component distributions are shown in Figure 5.2 along with the histogram and distribution of the data.

Parameter	Value
k (shape parameter)	$\begin{bmatrix} 67.49 \\ 76.98 \end{bmatrix}$
θ (rate parameter)	$\begin{bmatrix} 5.74 \\ 4.99 \end{bmatrix}$
Γ (Transition Probability Matrix)	$\begin{bmatrix} 0.92 & 0.08 \\ 0.06 & 0.94 \end{bmatrix}$
δ (initial distribution)	$[0,1]$

Table 5.1: Component distribution parameters for States 1 and 2, transition probability matrix and initial distribution used for fitting the Hidden Markov Model
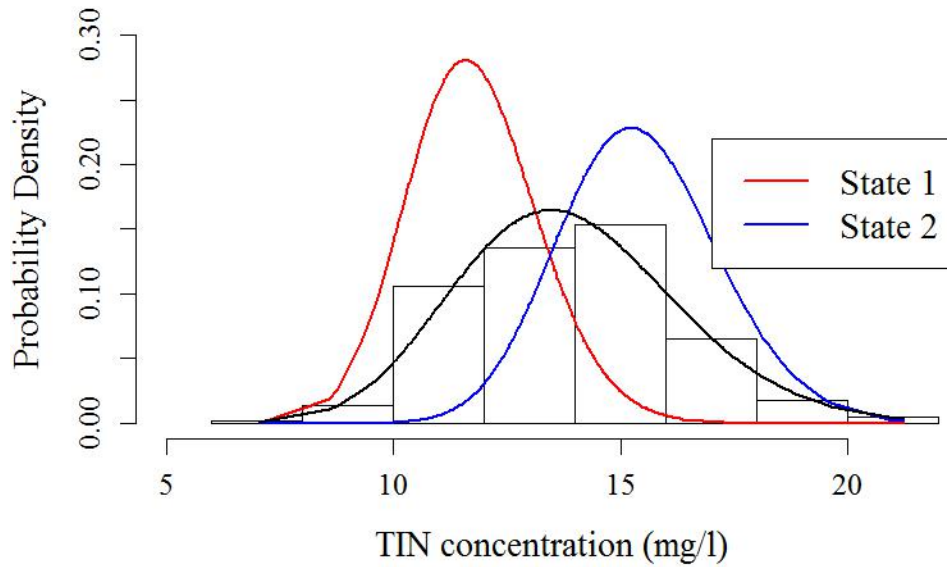
Figure 5.2: Component gamma distributions of the Hidden Markov model. The solid black line represents the gamma distribution fit on all effluent TIN data (k=32.16, θ =34.77)

The expected state sequence obtained from global decoding is shown in Figure 5.3. A weekly average time series plot of TIN concentration is overlain, to illustrate the TIN concentrations corresponding to each state. From Figure 5.3, we find evidence of regime shifting behavior, with effluent TIN concentrations alternating between 'high' and 'low' states. While state 1 or the 'low' state corresponds to TIN concentrations less than or equal to 15 mg/l, state 2 or the 'high' state corresponds to TIN concentrations greater than 15 and exceeding the limit set in Colorado Regulation #85 (CDPHE, 2012). Suchetana et al. (2016, 2017), Weirich et al. (2015a) have reported the importance of an 'inertia' (referred to as 'Lag') in predicting ammonia concentrations on a monthly time scale. We find similar patterns of 'inertia' or persistence in effluent TIN concentrations on the weekly time scale, as well, with a given state generally persisting for 3 weeks, but to up to 36 weeks, on occasion.
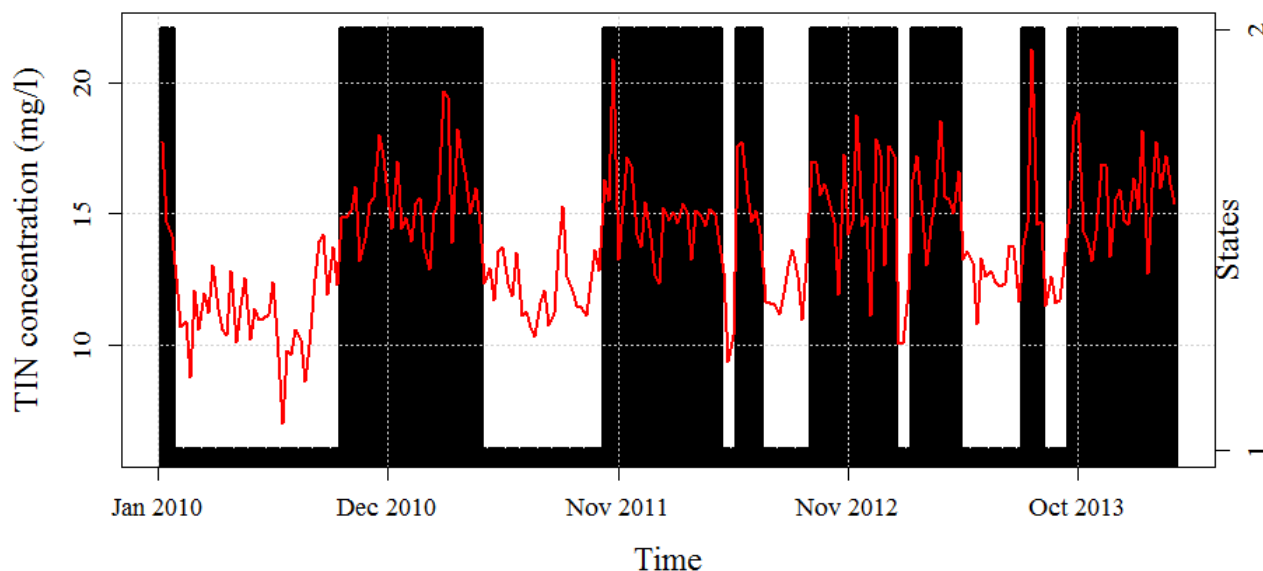
Figure 5.3: Expected state sequence obtained from global decoding. The solid red line indicates the time series of weekly average TIN concentrations

A binary logistic regression model is fit using the expected state sequence for TIN as the response variable and the covariate set described above in the previous section. The best set of covariates is selected using the BIC criteria. Table 5.2 lists the covariates and the corresponding coefficients for the best binary logistic regression model. T_{min}, Precip, sine component, Lag and Amm. appear in the best logistic regression model. It is interesting to note that none of the flow variables (A, C or AC) appear in the selected logistic regression equation, possibly because of low variability in the values of these variables as compared to other covariates. Appearance of T_{min} (and not T_{max}) in the best equation indicates that the minimum air temperature has a more significant effect on effluent TIN levels, than the maximum air temperature. This is not surprising, since the minimum air temperatures were as low as -21°C (or -6°F) which is outside the tolerance range of nitrifying bacteria. A negative sign on the logistic regression coefficient for T_{min} indicates that lower minimum temperatures result in higher effluent TIN concentrations. Precipitation and effluent TIN

concentrations are also inversely related (logistic regression coefficient=-5.81), as heavy precipitations often cause dilution of TIN concentrations in wastewater. Seasonal variations in flow are demonstrated by the sine component. The positive relation between effluent TAN and TIN (logistic regression coefficient = 2.78) supports the finding of Pocernich and Litke (1997), which reported effluent ammonia was a predictor of other nitrogen species .

Covariates	Coefficients
Intercept	-7.24
T_{min} (Weekly average minimum temperature degF)	-0.08
Precip. (Weekly average precipitation, inches)	-5.81
sine component (seasonality)	-1.56
Lag (expected state in the previous week)	6.98
Amm. (average weekly ammonia concentration in mg/l)	2.78

Table 5.2: Covariates and coefficients of the best binary logistic regression model, as determined by BIC

The binary logistic regression model predictions are the probabilities of each observation belonging to either state. The Brier Skill Score (BSS) is calculated for this model (Epstein, 1969, Weigel et al., 2007) with negative BSS values indicating that the model predictions are undesirable, as they are no better than the climatological forecast. Positive BSS

values are desirable; the closer the BSS value is to 1, better the model predictions. The BSS for our model is 0.84, indicating the model predictions are better than the climatological forecast, and are very close to the observed categories.
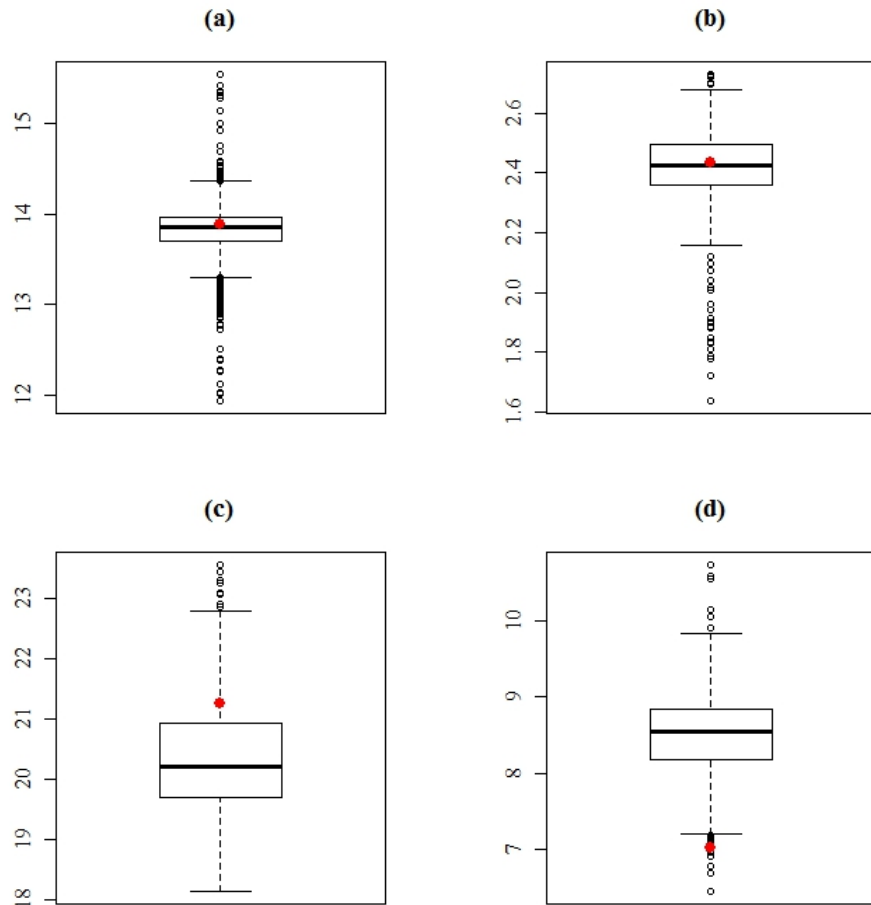
Figure 5.4: Boxplots indicating the spread of (a) mean (b) standard deviation (c) maximum (d) minimum of the simulations. Red dot indicates the corresponding observed statistics

The model predictions are used to simulate 1000 values for each observation, using the procedure described previously. Various simulation statistics including the mean, standard deviation, maximum and minimum, are calculated for this ensemble. The resulting boxplots are shown in Figure 5.4. The observed mean and standard deviation are captured accurately

by the simulations. The maximum and minimum values however, are underestimated and overestimated respectively, indicating that the model is not capturing the extreme values. Figure 5.5 shows the boxplots of the Probability Density Function (PDF) of the simulations and the observed PDF, which match closely in shape and scale. Thus, the combined Hidden Markov Model and binary logistic regression model successfully captures the regime shifting behavior of effluent TIN at this wastewater treatment plant.
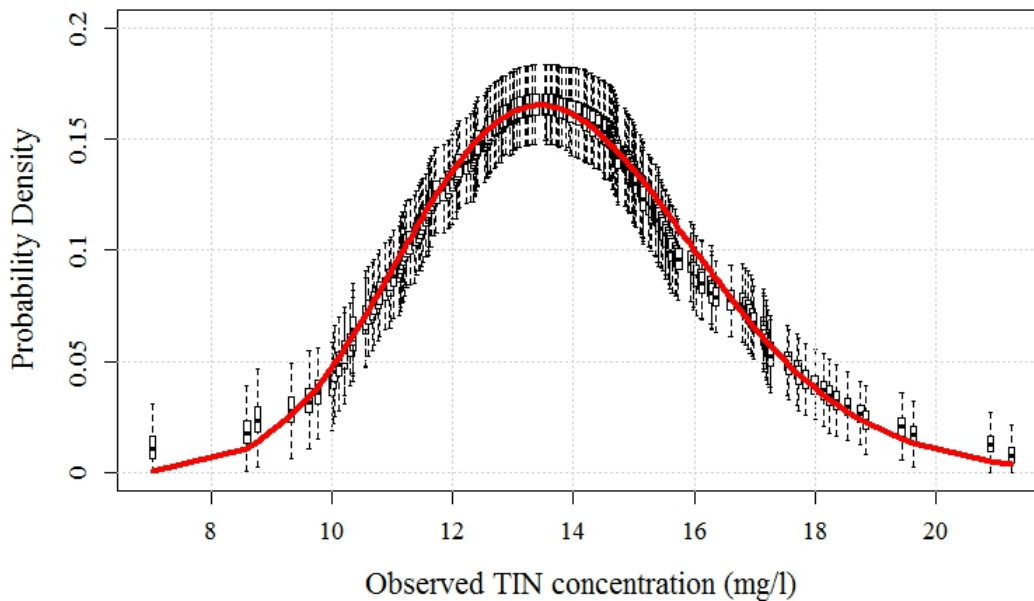
Figure 5.5: Observed (solid red line) and boxplots of simulated PDF using the binary logistic regression model

5.5   Discussion

5.5.1   Relation between TIN and ammonia discharge limit violations

The wastewater treatment facility under consideration is scheduled to adopt more stringent ammonia discharge limits from December 1, 2017. Table 4.1 provides the current and future monthly ammonia discharge limits. It may be beneficial to know whether events

of ammonia permit limit violations can be related to violations in TIN discharge limits. Currently, there are monthly discharge limits for ammonia but annual discharge limits for TIN. In the absence of ammonia and TIN discharge limit information of the same temporal resolution, in this section, we only provide an outline of the methodology to use the HMM and multinomial logistic regression framework, to relate TIN and ammonia discharge limit violations.

First, we calculate the relative concentration of TIN, which is defined as the ratio of the absolute monthly average concentration of TIN to the discharge limits for that month (current regulations do not specify monthly TIN discharge limits). Relative concentrations ≤ 1 indicate compliance, and > 1, permit limit violations. We fit a Hidden Markov Model to the monthly relative TIN concentrations and obtain the expected state sequence. As before, a multinomial logistic model is fit using the same predictor variables as described in section 5.3, except that effluent TAN is transformed to relative TAN the monthly discharge limit (Table 4.1, columns 2 & 3 to be used to calculate ammonia concentrations relative to existing discharge limits and future discharge limits, respectively). We expect that if relative ammonia concentration appears in the best regression equation, it should be predictive of compliance monthly (assumed) TIN discharge limits.

5.5.2   Compliance with Colorado Regulation #85

As of September 30, 2012, Colorado Regulation #85 specifies a maximum permissible discharge limit for TIN, with a running annual median (median of all samples taken in the last 12 months) and a 95th percentile value of 15 mg/l and 20 mg/l respectively (CDPHE, 2012). However, for the purpose of evaluation, we calculate the annual median and 95th percentile of TIN concentrations individually for each of the years from 2010-2014, shown in Figure 5.6 (a). The boxplots show the spread of simulated TIN concentrations for individual years, while the red dots indicate the 95th percentile calculated from the simulated TIN concentrations. The solid and dashed horizontal lines indicate the annual median and 95th

percentile limits specified by Colorado Regulation #85, respectively. It should be noted that for the year 2014, we only have data from January to March (winter months), and as a result we find higher simulated values of effluent TIN concentrations. The simulated annual median effluent TIN as well as the 95th percentile concentrations, consistently lie below the Colorado limits, with the exception of 2014, when the annual median is equal to permissible limit. Figure 5.6 (b) shows the weekly time series of effluent TIN concentrations, and as before, the solid and dashed horizontal lines indicate the annual median and 95th percentile limits specified by Regulation #85 respectively. Comparing Figures 5.6 (a) and (b) we note that even though on an annual scale the wastewater treatment plant is in compliance, there are frequent excursions on a weekly scale. For 73 weeks during the period of January 2010 - March 2014, which constitutes 33% of the observations, TIN concentrations in the effluent exceeded the annual median discharge limit of 15 mg/l, with effluent TIN concentrations exceeding the 95th percentile discharge limit of 20 mg/l twice during the available period of record. Clearly , depending on stream flow conditions, relying on annual median or average discharge limits for TIN may not be sufficient to preserve water quality for designated uses.

5.5.3   Reliability Estimates

Reliability of a system, α, can be defined as the frequency or probability of being in a satisfactory state, S (Hashimoto et al., 1982)

$$\alpha = Prob[X_t \in S] \tag{5.8}$$

where X_t denotes the systems output state or status at time $t = 1, 2, 3, ...$ In our analysis, reliability denotes how often the TIN concentrations remain below a threshold. For example, we can try to estimate the range of reliability of the wastewater treatment plant meeting the limit of 15 mg/l on a weekly scale. For each set of simulations for TIN concentrations, we calculate the frequency of TIN concentrations below 15 mg/l and pro-
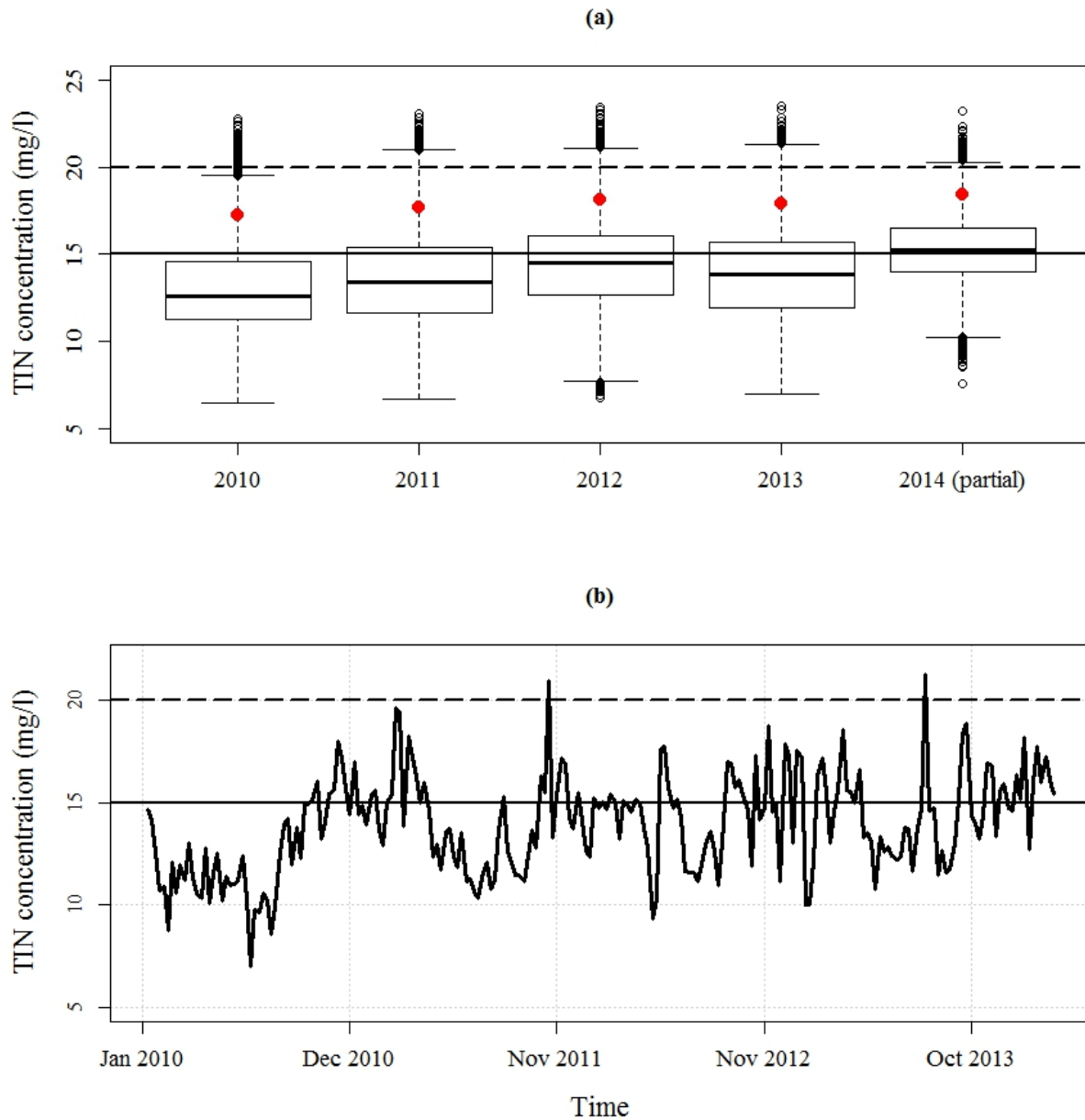
Figure 5.6: (a): Boxplots of simulated TIN concentrations (in mg/l) for 2010-2014 (data from January-March available for 2014). The red dots indicate the 95th percentile TIN concentrations calculated from the simulations (b) Time series plot of weekly averaged TIN concentrations. The solid and dashed horizontal lines indicate the annual median and 95th percentile TIN discharge limits respectively, as specified by Regulation #85

duce a boxplot of the values obtained for all 1000 simulations, as shown in Figure 5.7. The red dot indicates the observed frequency of meeting the annual median discharge limit on a weekly basis. The median reliability of meeting the target is 0.67, the same as the observed value, which is equivalent to a 1 in 3 (0.33) probability of exceeding the target. The highest value of reliability estimates obtained from the simulations is 0.95, while the lowest is 0.39, corresponding to an exceedance probability of 0.05 and 0.61 respectively. Reliability estimates further reinforce the idea that annual compliance with the discharge limits of 15 mg/l TIN does not account for excursions on a daily, weekly or even monthly time scale. Since the overall goal of TIN regulation is to protect the quality of receiving waters, it may be pertinent to rethink the relevance of annual discharge limits for TIN.
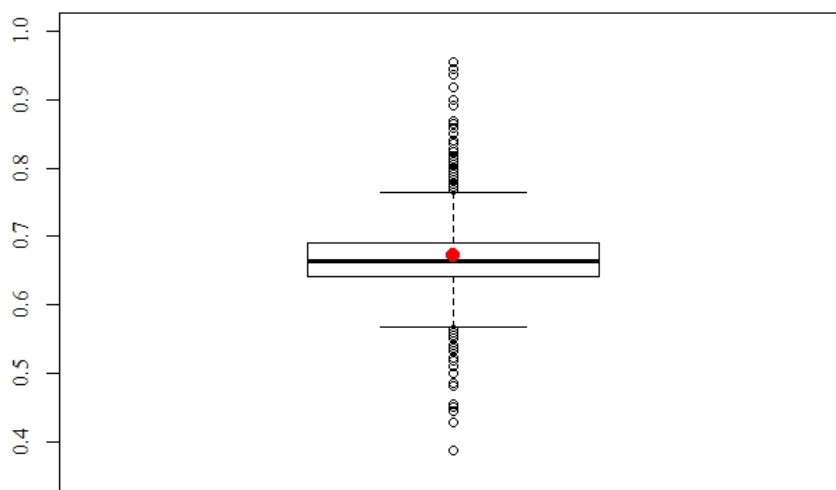
Figure 5.7: Boxplot indicating the spread of reliability estimates. Red dot indicates the reliability value obtained from observations

5.6   Conclusions

We provide a framework for stochastic simulation of average weekly TIN concentrations using a Hidden Markov Model with gamma component distributions and a subsequent binary logistic regression model. The binary logistic regression model indicates the influence of climate, seasonality, past performance and effluent ammonia concentration in predicting the effluent TIN state sequence. The simulations successfully capture the observed mean, standard deviation, maximum and minimum effluent TIN values observed for the years 2010 2014 at an operating activated sludge treatment plant with an average annual flow of 47,320 m^3/d (12.5 MGD).

An important application of this simulation framework is to investigate the relation between ammonia and TIN discharge limit violations. The discussed methodology is used to predict future violations in TIN discharge limits, based on compliance with effluent ammonia discharge limits. The approach is further used to obtain estimates of the degree of compliance with discharge limits and anticipating the ability to meet changing limits , including their application over shorter time periods, which requires greater nitrogen removal process reliability.

Chapter 6

Conclusions and Future Work

6.1   Summary

Motivated by the need to develop a better understanding of municipal wastewater treatment plant compliance with changing nutrient regulations, several performance-based models have been developed to model compliance with existing and new nutrient regulations. In *"A hierarchical modeling approach to evaluate spatial and temporal variability of wastewater treatment compliance with ammonia discharge limits in the US"* (Chapter 2) a hierarchical framework has been developed which can model variability in compliance with ammonia discharge limit at various location in the country, over the period of 2004-2008. For the first time, national compliance has been modeled using two sets of covariates- large scale variables, which remain fairly homogeneous over space and time, and local variables which affect compliance on a smaller scale. These two sets of covariates are modeled using Generalized Linear Models (GLMs) and Kriging respectively, in a two-tiered hierarchical approach. In the first level of hierarchy, GLM, plant flow rate and the compliance history of the plant are seen to appear in the best model. The residuals are modeled in the second level of hierarchy using Kriging, as a function of the latitude and longitude of the plant. Hierarchical model predictions are obtained as the sum of model predictions from each level of hierarchy. This methodology provides significant improvement in the predictive ability, with the hierarchical model offering median R^2 values of 0.8 in prediction.

Traditionally, in wastewater treatment, the risk of violating discharge limits has been

quantified using the probability of violation only (Weirich et al., 2015a). *"Modeling risk attributes of wastewater treatment plant violations of ammonia discharge limits in the United States"*(Chapter 3) presents the first methodological attempt to model three different attributes of the risk of non-compliance with ammonia discharge limits- the probability, frequency and magnitude of violations. The probability of non-compliance is modeled using logistic regression, the frequency using Poisson regression and the magnitude using non-stationary Generalized Pareto Distributions. Individually, these risk attributes provide important information about how likely a plant is to violate its permit, and how frequent and severe these violations might be. Additionally, these risk attributes may be combined to obtain composite risk estimates- the Estimated Violation Index (EVI) and the Estimated Severity Index (ESI). The relative values of these indices at various treatment plants across the country provide an estimate of the risk profile of that plant during the period of 2004-2008.

In Chapter 4 titled *"Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits"*, a diagnostic framework has been provided to evaluate factors affecting compliance with current and future ammonia discharge limits. While the variables affecting compliance with existing discharge limits are known, from Chapter 2, the relative importance of these variables has been explored in this chapter. Also, variables and their relative importance in predicting compliance with future discharge limits are investigated in this chapter. Random forests are used to perform variable selection, and the selected variables are used to fit regression trees. Previous month's treatment performance is found to be the most important variable affecting compliance with current and future discharge limits. This approach is validated using daily effluent data from an individual wastewater treatment plant in Colorado, US. The results indicate that the regression tree models can accurately predict compliance (or the lack thereof) with existing and future discharge limits for individual plants as well as on a national scale. This methodology also marks the first statistical attempt to delineate reliable sources of demand and supply for a

point source-to-point source nutrient credit trading scheme.

Finally, as broader regulations for nitrogen species in effluent wastewaters are being considered, in Chapter 5 titled *Modeling Total Inorganic Nitrogen in treated wastewater using non-homogeneous Hidden Markov and Multinomial Logistic Regression Models*, the first statistical attempt has been made to model variations in effluent Total Inorganic Nitrogen (TIN) concentrations using Hidden Markov and Multinomial Logistic Regression Models. Weekly effluent data from a wastewater treatment plant in Colorado, US has been used for this analysis. First, the Hidden Markov Model captures the regime shifts in effluent TIN concentrations, followed by the multinomial logistic regression model which identifies factors associated with these regime shifts. The best regression model indicates that prior performance, seasonality, effluent ammonia concentrations and climate variables (like minimum temperature and precipitation) affect effluent TIN concentrations. Model simulations can capture the sample statistics such as mean and standard deviation very well. Reliability estimates have also been developed from the simulations and the adequacy of using annual nutrient discharge limits to protect surface water quality has been discussed.

6.2   Discussion

This research makes significant strides in advancing existing knowledge about performance-based statistical modeling of variability in compliance with nutrient discharge limits. By analyzing data of varying spatial (national level as well as single plant level) and temporal resolution (weekly and monthly effluent data), as well as analyzing existing and future nutrient regulations, an in-depth study of regulatory compliance has been conducted. These modeling strategies can be used for robust planning and decision-making efforts, on an individual plant level as well as on a regional or national scale.

6.3   Future Work

Several extensions of these performance-based statistical models can be envisaged. Some of them are presented here:

(i) The statistical models developed in this dissertation can be extended to other nutrients, like phosphorus, and other pollutants like Biochemical Oxygen Demand (BOD), Total Suspended Solids (TSS). It would be interesting to note which set of covariates affect the effluent concentrations of these pollutants.

(ii) In this research, due to the availability of spotty spatial data, point estimates for effluent ammonia concentrations and risk of permit exceedance have been provided. If effluent data of finer spatial resolution is available, the existence of spatial patterns in regulatory compliance can be investigated and depicted in regional or national surface maps. Similar spatial maps for the risk of permit violation can also be developed. These maps can be used by monitoring and enforcement agencies in the identification of 'high-risk' zones, i.e., regions with a higher risk of violating their permit limit. Also, agencies looking to set-up a new wastewater treatment plant can take into account any regional risks of non-compliance while designing the plant.

(iii) Several biological nitrogen removal processes such as Modified Ludzack-Ettinger or four-stage Bardenpho utilize the biodegradable organic substrate present in the wastewater as an electron donor for denitrification (Grady Jr et al., 2011). Thus the probability distributions of effluent BOD and Total Nitrogen concentrations are expected to be related. By estimating the joint distribution and the joint probability of exceedance of BOD and Total Nitrogen, not only can we predict the joint behavior of these pollutants in wastewater, but also predict how a process upset can alter the effluent concentrations of these pollutants.

(iv) As demonstrated in Chapter 5, effluent concentrations of TIN vary with changes

in climate variables such as temperature and precipitation. Thus, in the face of changing climate, it can be expected that the efficiency of nitrogen removal systems, and biological wastewater treatment systems in general, will change. Substituting relevant values of temperature and precipitation from climate model projections in the regression equation shown in Table 5.2, effluent TIN (or other pollutant) concentrations can be estimated. This information can be crucial to plant managers and operators, who are looking to ensure compliance with discharge limits under altered climatic conditions.

(v) Chapter 4 discusses a methodology to delineate reliable sources of demand and supply for a nutrient credit trading scheme. Using these sources of demand and supply, the feasibility of an actual point source-to-point source nutrient trading scheme can be demonstrated for a given region, for example, the dischargers along the South Platte River in Colorado, US. From treatment cost data at the supply sources, unit credit pricing can be derived. Using effluent data from 'under-performers' (i.e. plants discharging above their permit limit), the pollutant mass loading, and subsequent nutrient credits needed by these dischargers can be estimated. Effluent data from 'over-performers' (i.e. wastewater treatment plants discharging below their permit limit) can be used to estimate number of credits available for sale. The net pollutant loading can be calculated, to demonstrate whether the point source-to-point source nutrient trading scheme can indeed help improve the quality of the receiving water.

Bibliography

Abushariah, M. A., Ainon, R. N., Zainuddin, R., Elshafei, M., and Khalifa, O. O. Natural speaker-independent arabic speech recognition system based on hidden markov models using sphinx tools. In Computer and Communication Engineering (ICCCE), 2010 International Conference on, pages 1–6. IEEE, 2010.

Akaike, H. A new look at the statistical model identification. IEEE transactions on automatic control, 19(6):716–723, 1974.

Akintug, B. and Rasmussen, P. A markov switching model for annual hydrologic time series. Water resources research, 41(9), 2005.

Al-Saleh, J. A. and Agarwal, S. K. Finite mixture of gamma distributions: A conjugate prior. Computational statistics & data analysis, 51(9):4369–4378, 2007.

Andreen, W. L. Motivating enforcement: Institutional culture and the clean water act. Pace Envtl. L. Rev., 24:67, 2007.

Atanasova, N. and Kompare, B. Modelling of wastewater treatment plant with decision and regression trees. 3 rd Workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI2002), 23(9):3, 2002.

Bartolomei, S. M. and Sweet, A. L. A note on a comparison of exponential smoothing methods for forecasting seasonal series. International Journal of Forecasting, 5(1):111–116, 1989.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The annals of mathematical statistics, 41(1):164–171, 1970.

Best, N. G., Ickstadt, K., and Wolpert, R. L. Spatial poisson regression for health and exposure data measured at disparate resolutions. Journal of the American statistical association, 95(452):1076–1088, 2000.

Biau, G. Analysis of a random forests model. Journal of Machine Learning Research, 13 (Apr):1063–1095, 2012.

Biau, G., Devroye, L., and Lugosi, G. Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9(Sep):2015–2033, 2008.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. Applied Spatial Data Analysis with R. Springer, 2013.

Bogardi, J. J. and Kundzewicz, Z. W. Risk, reliability, uncertainty, and robustness of water resource systems. Cambridge University Press, 2002.

Booth, A., Gerding, E., and Mcgroarty, F. Automated trading with performance weighted random forests and seasonality. Expert Systems with Applications, 41(8):3651–3661, 2014.

Bracken, C., Rajagopalan, B., and Zagona, E. A hidden markov model combined with climate indices for multidecadal streamflow simulation. Water Resources Research, 50 (10):7836–7846, 2014.

Branosky, E., Jones, C., and Selman, M. Comparison tables of state nutrient trading programs in the chesapeake bay watershed. World Resources Institute, Washington, DC, 2011.

Bras, R. L. and Rodriguez-Iturbe, I. Random functions and hydrology. Courier Corporation, 1985.

Breiman, L. Out-of-bag estimation, 1996.

Breiman, L. Random forests. Machine learning, 45(1):5–32, 2001.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. Classification and regression trees. CRC press, 1984.

Breiman, L. et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3):199–231, 2001.

Bulla, J. Hidden markov models with t components. increased persistence and other aspects. Quantitative Finance, 11(3):459–475, 2011.

Campbell, J. Y. and Thompson, S. B. Predicting excess stock returns out of sample: Can anything beat the historical average? Review of Financial Studies, 21(4):1509–1531, 2008.

Carlson, R. F., MacCormick, A., and Watts, D. G. Application of linear random models to four annual streamflow series. Water Resources Research, 6(4):1070–1078, 1970.

Castillo, E. and Hadi, A. S. Fitting the generalized pareto distribution to data. Journal of the American Statistical Association, 92(440):1609–1620, 1997.

CDEEP. An act concerning the sustainibiliy of the nitrogen credit exchange program, public act no. 15-38. Technical report, Connecticut Department of Energy & Environmental Protection, 2015.

CDPHE. Nutrient cost-benefit study results and conclusions: Water quality control division prehearing statement - exhibit 5. Technical report, Colorado Department of Public Health and Environment, 2011.

CDPHE. Nutrient management control regulation: Regulation #85. Technical report, Colorado Department of Public Health and Environment, 2012.

Chan, G., Stavins, R., Stowe, R., and Sweeney, R. The so2 allowance trading system and the clean air act amendments of 1990: reflections on twenty years of policy innovation. Technical report, National Bureau of Economic Research, 2012.

Choulakian, V. and Stephens, M. Goodness-of-fit tests for the generalized pareto distribution. Technometrics, 43(4):478–484, 2001.

Chrysos, G., Dagritzikos, P., Papaefstathiou, I., and Dollas, A. Hc-cart: A parallel system implementation of data mining classification and regression tree (cart) algorithm on a multi-fpga system. ACM Transactions on Architecture and Code Optimization (TACO), 9(4):47, 2013.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. An introduction to statistical modeling of extreme values, volume 208. Springer, 2001.

Cooley, D. Extreme value analysis and the study of climate change. Climatic change, 97(1): 77–83, 2009.

Cornel, P. and Weber, B. Water reuse for irrigation from waste water treatment plants with seasonal varied operation modes. Water Science and Technology, 50(2):47–53, 2004.

Cressie, N. The origins of kriging. Mathematical geology, 22(3):239–252, 1990.

Daigger, G. T., Datta, T., Stensel, H. D., Whitlock, D. D., and Mackey, J. K. Evaluating the role of point source discharges informs statewide nutrient control policy in utah. Water Environment Research, 86(6):559–572, 2014.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977.

Dennison, W., Helfrich, M., Michelsen, E., Pritzlaff, R., and Tutman, F. Senior scientists and policymakers for the bay nutrient trading sub-committee: Nutrient trading, preliminary investigation. Technical report, Chesapeake Bay Action Plan, 2012.

Diaz-Uriarte, R. and de Andrés, S. A. Variable selection from random forests: application to gene expression data. arXiv preprint q-bio/0503025, 2005.

Dong, C., Richards, S. H., Clarke, D. B., Zhou, X., and Ma, Z. Examining signalized intersection crash frequency using multivariate zero-inflated poisson regression. Safety science, 70:63–69, 2014.

Dubovyk, O., Menz, G., Conrad, C., Kan, E., Machwitz, M., and Khamzina, A. Spatio-temporal analyses of cropland degradation in the irrigated lowlands of uzbekistan using remote-sensing and logistic regression modeling. Environmental monitoring and assessment, 185(6):4775–4790, 2013.

El Adlouni, S., Ouarda, T., Zhang, X., Roy, R., and Bobée, B. Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. Water Resources Research, 43(3), 2007.

Elith, J., Leathwick, J. R., and Hastie, T. A working guide to boosted regression trees. Journal of Animal Ecology, 77(4):802–813, 2008.

Epstein, E. S. A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology, 8(6):985–987, 1969.

Erkyihun, S. T., Rajagopalan, B., Zagona, E., Lall, U., and Nowak, K. Wavelet-based time series bootstrap model for multidecadal streamflow simulation using climate indicators. Water Resources Research, 52(5):4061–4077, 2016.

Evin, G., Merleau, J., and Perreault, L. Two-component mixtures of normal, gamma, and gumbel distributions for hydrological applications. Water Resources Research, 47(8), 2011.

Fairfax, V. and Hamilton, V. Progress in water quality: an evaluation of the national investment in municipal wastewater treatment. Washington, DC: US EPA, 2000.

Fildes, R. and Makridakis, S. The impact of empirical accuracy studies on time series analysis and forecasting. International Statistical Review/Revue Internationale de Statistique, pages 289–308, 1995.

Filipe, C. D., Crawford, G., Johnson, B., and Daigger, G. T. Integrating asm 2d models into whole plant mass balance simulators. Proceedings of the Water Environment Federation, 2001(13):696–710, 2001.

Fillos, J., Katehis, D., Ramalingam, K., Carrio, L., and Gopalakrishnan, K. Determination of nitrifier growth rates in nyc water pollution control plants. Proceedings of the Water Environment Federation, 2000(13):585–601, 2000.

Fisher, A. J., Coley, R. Y., and Zeger, S. L. Fast out-of-sample predictions for bayesian hierarchical models of latent health states. arXiv preprint arXiv:1510.08802, 2015.

Forney, G. D. The viterbi algorithm. Proceedings of the IEEE, 61(3):268–278, 1973.

Friedman, J., Hastie, T., and Tibshirani, R. The elements of statistical learning, volume 1. Springer series in statistics Springer, Berlin, 2001.

Furrer, E. M., Katz, R. W., Walter, M. D., and Furrer, R. Statistical modeling of hot spells and heat waves. Climate Research, 43(3):191–205, 2010.

Gelati, E., Madsen, H., and Rosbjerg, D. Markov-switching model for nonstationary runoff conditioned on el niño information. Water resources research, 46(2), 2010.

Genuer, R., Poggi, J., and Tuleau-Malot, C. Vsurf: Variable selection using random forests, 2013.

Gilleland, E. and Katz, R. W. New software to analyze how extremes change over time. Eos, Transactions American Geophysical Union, 92(2):13–14, 2011.

Grady Jr, C. L., Daigger, G. T., Love, N. G., and Filipe, C. D. Biological wastewater treatment. CRC press, 2011.

Grantz, K., Rajagopalan, B., Clark, M., and Zagona, E. A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. Water Resources Research, 41(10), 2005.

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., Rueckert, D., Initiative, A. D. N., et al. Random forest-based similarity measures for multi-modal classification of alzheimer's disease. NeuroImage, 65:167–175, 2013.

Greene, A. M., Robertson, A. W., Smyth, P., and Triglia, S. Downscaling projections of indian monsoon rainfall using a non-homogeneous hidden markov model. Quarterly Journal of the Royal Meteorological Society, 137(655):347–359, 2011.

Haagenson, E., Rajagopalan, B., SUMMERS, S., and ROBERSON, J. Projecting demand extremes under climate change using extreme value analysis. Journal: American Water Works Association, 105(2), 2013.

Hamilton, J. D. Regime switching models. In Macroeconometrics and Time Series Analysis, pages 202–209. Springer, 2010.

Hao, R., Li, S., Li, J., Zhang, Q., and Liu, F. Water quality assessment for wastewater reclamation using principal component analysis. Journal of Environmental Informatics, 21(1), 2013.

Harte, D. HiddenMarkov: Hidden Markov Models. Statistics Research Associates, Wellington, 2016. URL ftp://ftp.gns.cri.nz/pub/davidh/sslib/r-repo/. R package version 1.8-7.

Hashimoto, T., Stedinger, J. R., and Loucks, D. P. Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. Water resources research, 18 (1):14–20, 1982.

Henze, M. Capabilities of biological nitrogen removal processes from wastewater. Water Science and Technology, 23(4-6):669–679, 1991.

Hoag, D. L. and Hughes-Popp, J. S. Theory and practice of pollution credit trading in water quality management. Review of Agricultural Economics, pages 252–262, 1997.

Hogan, J. N., Daniels, M. E., Watson, F. G., Conrad, P. A., Oates, S. C., Miller, M. A., Hardin, D., Byrne, B. A., Dominik, C., Melli, A., et al. Longitudinal poisson regression to evaluate the epidemiology of cryptosporidium, giardia, and fecal indicator bacteria in coastal california wetlands. Applied and environmental microbiology, 78(10):3606–3613, 2012.

Huang, X. D., Ariki, Y., and Jack, M. A. Hidden Markov models for speech recognition, volume 2004. Edinburgh university press Edinburgh, 1990.

Jackson, B. B. Markov mixture models for drought lengths. Water Resources Research, 11 (1):64–74, 1975.

Jenkinson, A. F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quarterly Journal of the Royal Meteorological Society, 81(348): 158–171, 1955.

Jones, C., Branosky, E., Selman, M., and Perez, M. How nutrient trading could help restore the chesapeake bay. World Resources Institute Working Paper, 2010.

Katz, R. W., Parlange, M. B., and Naveau, P. Statistics of extremes in hydrology. Advances in water resources, 25(8):1287–1304, 2002.

King, D. M. and Kuch, P. J. Will nutrient credit trading ever work? an assessment of supply and demand problems and institutional obstacles. Environmental Law Reporter News and Analysis, 33(5):10352–10368, 2003.

Kjeldsen, T. R. and Rosbjerg, D. Choice of reliability, resilience and vulnerability estimators for risk assessments of water resources systems/choix destimateurs de fiabilité, de résilience et de vulnérabilité pour les analyses de risque de systèmes de ressources en eau. Hydrological sciences journal, 49(5), 2004.

Kleiber, W., Katz, R. W., and Rajagopalan, B. Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes. Water Resources Research, 48(1), 2012.

Kleiber, W., Katz, R. W., Rajagopalan, B., et al. Daily minimum and maximum temperature simulation over complex terrain. The Annals of Applied Statistics, 7(1):588–612, 2013.

Krige, D. A statistical approach to some mine valuations and allied problems at the witwatersrand. University of Witwatersrand, 1951.

Lal, H. Nutrient credit tradinga market-based approach for improving water quality. Advances in Nitrogen Management for Water Quality (Delgado JA, Follett RF, eds.). Ankeny, IA: Soil and Water Conservation Society, 2010.

Lall, U. and Sharma, A. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resources Research, 32(3):679–693, 1996.

Lawrance, A. and Kottegoda, N. Stochastic modelling of riverflow time series. Journal of the Royal Statistical Society. Series A (General), pages 1–47, 1977.

Li, Z., Wang, W., Liu, P., Bigham, J. M., and Ragland, D. R. Using geographically weighted poisson regression for county-level crash modeling in california. Safety science, 58:89–97, 2013.

Liang, T.-H. and Lin, J.-B. A two-stage segment and prediction model for mortgage prepayment prediction and management. International Journal of Forecasting, 30(2):328–343, 2014.

Liwarska-Bizukojc, E. and Biernacki, R. Identification of the most sensitive parameters in the activated sludge model implemented in biowin software. Bioresource technology, 101 (19):7278–7285, 2010.

Lukas, J. J., Wade, L., and Rajagopalan, B. Paleohydrology of the lower colorado river basin and implications for water supply availability. Completion report (Colorado Water Institute); no. 223, 2007.

Maraun, D., Rust, H. W., and Osborn, T. J. The annual cycle of heavy precipitation across the united kingdom: a model based on extreme value statistics. International Journal of Climatology, 29(12):1731–1744, 2009.

May-Ostendorp, P. T., Henze, G. P., Rajagopalan, B., and Corbin, C. D. Extraction of supervisory building control rules from model predictive control of windows in a mixed mode building. Journal of Building Performance Simulation, 6(3):199–219, 2013.

McCullagh, P. and Nelder, J. Generalized Linear Models. Chapman & Hall, 1989.

Meese, R. A. and Rogoff, K. Empirical exchange rate models of the seventies: Do they fit out of sample? Journal of international economics, 14(1-2):3–24, 1983.

Mehrotra, R. and Sharma, A. A nonparametric nonhomogeneous hidden markov model for downscaling of multisite daily rainfall occurrences. Journal of Geophysical Research: Atmospheres, 110(D16), 2005.

Melcer, H. Full scale experience with biological process modelscalibration issues. Water science and technology, 39(1):245–252, 1999.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. Bcftools/roh: a hidden markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics, page btw044, 2016.

Niku, S. and Schroeder, E. D. Factors affecting effluent variability from activated sludge processes. Journal (Water Pollution Control Federation), pages 546–559, 1981a.

Niku, S. and Schroeder, E. D. Stability of activated sludge processes based on statistical measures. Journal (Water Pollution Control Federation), pages 457–470, 1981b.

Niku, S., Schroeder, E. D., and Samaniego, F. J. Performance of activated sludge processes and reliability-based design. Journal (Water Pollution Control Federation), pages 2841–2857, 1979.

Nowak, K., Prairie, J., Rajagopalan, B., and Lall, U. A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. Water Resources Research, 46 (8), 2010.

O'Connell, P. A simple stochastic modelling of hurst's law. Mathematical models in hydrology, 1, 1974.

Oleszkiewicz, J. A. and Barnard, J. L. Nutrient removal technology in north america and the european union: a review. Water Quality Research Journal of Canada, 41(4):449–462, 2006.

Oliveira, S. C. and Von Sperling, M. Reliability analysis of wastewater treatment plants. Water research, 42(4):1182–1194, 2008.

Ouarda, T. B., Labadie, J. W., and Fontane, D. G. Indexed sequential hydrologic modeling for hydropower capacity estimation, 1997.

Pant, P. N. and Starbuck, W. H. Innocents in the forest: Forecasting and research methods. Journal of Management, 16(2):433–460, 1990.

Pearce, J. and Ferrier, S. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological modelling, 133(3):225–245, 2000.

Phillips, H., Sahlstedt, K., Frank, K., Bratby, J., Brennan, W., Rogowski, S., Pier, D., Anderson, W., Mulas, M., Copp, J., et al. Wastewater treatment modelling in practice: a collaborative discussion of the state of the art. Water Science and Technology, 59(4): 695–704, 2009.

Pickands III, J. Statistical inference using extreme order statistics. the Annals of Statistics, pages 119–131, 1975.

PIRG. Troubled waters: An analysis of clean water act compliance, january 2000june 2003. US Public Interest Research Group, 2004.

Pocernich, M. and Litke, D. W. Nutrient concentrations in wastewater treatment plant effluents, south platte rwer basin. JAWRA Journal of the American Water Resources Association, 33(1):205–214, 1997.

Prairie, J. R., Rajagopalan, B., Fulp, T. J., and Zagona, E. A. Modified k-nn model for stochastic streamflow simulation. Journal of Hydrologic Engineering, 11(4):371–378, 2006.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286, 1989.

Regonda, S. K., Rajagopalan, B., Clark, M., and Zagona, E. A multimodel ensemble forecast framework: Application to spring seasonal flows in the gunnison river basin. Water Resources Research, 42(9), 2006.

Reichert, P. Aquasim–a tool for simulation and data analysis of aquatic systems. Water Science and Technology, 30(2):21–30, 1994.

Reiss, R.-D. and Thomas, M. Statistical analysis of extreme values with applications to insurance. Finance, Hydrology and Other Fields, Birkha {ü} ser, Basel, Switzerland, 3, 2007.

Ribaudo, M. O., Heimlich, R., and Peters, M. Nitrogen sources and gulf hypoxia: potential for environmental credit trading. Ecological Economics, 52(2):159–168, 2005.

Ripley, B. tree: Classification and regression trees. R package version, pages 1–0, 2005.

Ripley, B. D. Modern applied statistics with S. Springer, 2002.

Robertson, A. W., Kirshner, S., and Smyth, P. Downscaling of daily rainfall occurrence over northeast brazil using a hidden markov model. Journal of climate, 17(22):4407–4424, 2004.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67:93–104, 2012.

Rossman, L. A. Risk equivalent seasonal waste load allocation. Water Resources Research, 25(10):2083–2090, 1989.

Salas, J. D. Applied modeling of hydrologic time series. Water Resources Publication, 1980.

Sandri, M. and Zuccolotto, P. Variable selection using random forests. In Data analysis, classification and the forward search, pages 263–270. Springer, 2006.

Santin, I., Pedret, C., Meneses, M., and Vilanova, R. Artificial neural network for nitrogen and ammonia effluent limit violations risk detection in wastewater treatment plants. In System Theory, Control and Computing (ICSTCC), 2015 19th International Conference on, pages 589–594. IEEE, 2015.

Sattayatewa, C., Pagilla, K., Pitt, P., Selock, K., and Bruton, T. Organic nitrogen transformations in a 4-stage bardenpho nitrogen removal plant and bioavailability/biodegradability of effluent don. Water research, 43(18):4507–4516, 2009.

Schwarz, G. et al. Estimating the dimension of a model. The annals of statistics, 6(2): 461–464, 1978.

Shapiro, M. EPA reply to petition from ann alexander and nrdc, et al. for new technology-based nutrient limits. Technical report, United States Environmental Protection Agency, 2012.

Sharma, A., Tarboton, D. G., and Lall, U. Streamflow simulation: A nonparametric approach. Water resources research, 33(2):291–308, 1997.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N., and Gaunt, T. R. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. Human mutation, 34(1):57–65, 2013.

Shimshack, J. P. and Ward, M. B. Enforcement and over-compliance. Journal of Environmental Economics and Management, 55(1):90–105, 2008.

Stedinger, J. R. and Taylor, M. R. Synthetic streamflow generation: 1. model verification and validation. Water Resources Research, 18(4):909–918, 1982.

Steinzor, R. Subcommittee on fisheries, wildlife, and water hearing: Oversight of the clean water act. Technical report, United States Environmental Protection Agency, 2003.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. Conditional variable importance for random forests. BMC bioinformatics, 9(1):307, 2008.

Suchetana, B., Rajagopalan, B., and Silverstein, J. Hierarchical modeling approach to evaluate spatial and temporal variability of wastewater treatment compliance with biochemical oxygen demand, total suspended solids, and ammonia limits in the united states. Environmental Engineering Science, 33(7):514–524, 2016.

Suchetana, B., Rajagopalan, B., and Silverstein, J. Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits using a regression tree model. Science of The Total Environment, 598:249–257, 2017.

Symbiont. Evaluation of practical technology-based effluent standards for phosphorus in illinois, project no. w111034. Technical report, Illinois Association of Wastewater Agencies, 2011.

Tao, P.-C. and Delleur, J. W. Seasonal and nonseasonal arma models in hydrology. Journal of the Hydraulics Division, 102(10):1541–1559, 1976.

Tarboton, D. G., Sharma, A., and Lall, U. Disaggregation procedures for stochastic hydrology based on nonparametric density estimation. Water resources research, 34(1):107–119, 1998.

Tashman, L. J. Out-of-sample tests of forecasting accuracy: an analysis and review. International journal of forecasting, 16(4):437–450, 2000.

Thyer, M. and Kuczera, G. Modeling long-term persistence in hydroclimatic time series using a hidden state markov model. Water Resources Research, 36(11):3301–3310, 2000.

Thyer, M. and Kuczera, G. A hidden markov model for modelling long-term persistence in multi-site rainfall time series 1. model calibration using a bayesian approach. Journal of Hydrology, 275(1):12–26, 2003.

Towler, E., Rajagopalan, B., Seidel, C., and Summers, R. S. Simulating ensembles of source water quality using a k-nearest neighbor resampling approach. Environmental science & technology, 43(5):1407–1411, 2009.

Towler, E., Rajagopalan, B., Gilleland, E., Summers, R. S., Yates, D., and Katz, R. W. Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. Water Resources Research, 46(11), 2010.

Towler, E., Roberts, M., Rajagopalan, B., and Sojda, R. S. Incorporating probabilistic seasonal climate forecasts into river management using a risk-based framework. Water Resources Research, 49(8):4997–5008, 2013.

UDEQ. Utah nutrient strategy: technology limits. Technical report, Utah Department of Environmental Equality, 2014.

U.S.EPA. Discharge monitoring report (DMR) pollutant loading tool. Technical report, United States Environmental Protection Agency, 2008.

U.S.EPA. Water quality trading toolkit for permit writers. Technical report, United States Environmental Protection Agency, 2009.

U.S.EPA. NPDES permit writers' manual. Technical report, United States Environmental Protection Agency, 2010.

U.S.EPA. EPA must improve oversight of state enforcement. Technical report, Office of the Inspector General, United States Environmental Protection Agency, 2011.

U.S.EPA. Testimony of michael h. shapiro before the subcommittee on water and wildlife. Technical report, United States Environmental Protection Agency, 2013.

U.S.EPA. Justification of appropriation estimates for the committee on appropriations. Technical report, United States Environmental Protection Agency, 2015.

Vanhooren, H., Meirlaen, J., Amerlinck, Y., Claeys, F., Vangheluwe, H., and Vanrolleghem, P. A. West: modelling biological wastewater treatment. Journal of Hydroinformatics, 5 (1):27–50, 2003.

VDEE. Virginia nonpoint source nutrient credit registry. Technical report, Virginia Department of Environmental Equality, 2016.

Venables, W. N. and Ripley, B. D. Modern applied statistics with S-PLUS. Springer Science & Business Media, 2013.

Wang, N., Brown, D. G., An, L., Yang, S., and Ligmann-Zielinska, A. Comparative performance of logistic regression and survival analysis for detecting spatial predictors of land-use change. International Journal of Geographical Information Science, 27(10):1960–1982, 2013.

Weigel, A. P., Liniger, M. A., and Appenzeller, C. The discrete brier and ranked probability skill scores. Monthly Weather Review, 135(1):118–124, 2007.

Weirich, S. R., Silverstein, J., and Rajagopalan, B. Effect of average flow and capacity utilization on effluent water quality from us municipal wastewater treatment facilities. Water research, 45(14):4279–4286, 2011.

Weirich, S. R., Silverstein, J., and Rajagopalan, B. Resilience of secondary wastewater treatment plants: prior performance is predictive of future process failure and recovery time. Environmental Engineering Science, 32(3):222–231, 2015a.

Weirich, S. R., Silverstein, J., and Rajagopalan, B. Simulation of effluent biological oxygen demand and ammonia for increasingly decentralized networks of wastewater treatment facilities. Environmental Engineering Science, 32(3):232–239, 2015b.

Wheeler, T. J., Clements, J., and Finn, R. D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden markov models. BMC bioinformatics, 15(1):7, 2014.

Wilks, D. S. Statistical methods in the atmospheric sciences, volume 100. Academic press, 2011.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., and Al-Katheeri, M. M. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at wadi tayyah basin, asir region, saudi arabia. Landslides, 13(5):839–856, 2016.

Zachman, B. A., Rajagopalan, B., and Summers, R. S. Modeling nom breakthrough in gac adsorbers using nonparametric regression techniques. Environmental Engineering Science, 24(9):1280–1296, 2007.

Zucchini, W. and Guttorp, P. A hidden markov model for space-time precipitation. Water Resources Research, 27(8):1917–1923, 1991.

Zucchini, W. and MacDonald, I. L. Hidden Markov models for time series: an introduction using R, volume 22. CRC press Boca Raton, 2009.