

ABSTRACT

LIBERA, DOMINIC ANTHONY. Reducing Uncertainty in Predicting and Forecasting Nutrient Constituents Across the Southeastern United States. (Under the direction of Dr. Sankarasubramanian Arumugam).

Water quality observations for streams are only available continuously for water bodies that are impaired or near impairment classified using Numeric Nutrient Water Quality Criteria developed by the EPA. Frequency of samplings for non-impaired streams range from the bi-monthly to seasonal timescale because of the expensive cost in data collection and analysis. Estimation models using streamflow as predictors can be useful for estimating daily water quality constituents on non-observation days to determine long-term compliance under the nutrient criteria. Two commonly used regression models are the LOADEST model and the WRTDS model developed by the USGS. This study examines the performance of LOADEST and WRTDS in estimating total nitrogen (TN) load and concentration for 18 stations from the Water Quality Network (WQN) over the Southeast region. The role of basin area, sampling frequency, and model type in determining model performance was examined using a leave-one-out cross-validation approach. This study found that the WRTDS model performed better than the LOADEST model in predicting the observed variability of TN concentration for most of the stations (14 out of 18). Across both models, prediction of TN concentration performed much worse ($NSE < 0.4$) when compared TN load prediction ($NSE > 0.8$). To examine this phenomenon further, this study provides a non-parametric approach for assessing the performance of water quality models particularly in predicting concentration in the form of a free software toolkit. Null distributions of common performance metrics with no skill are constructed through bootstrap resampling and used to find p-values for metrics from the LOADEST and WRTDS models for determining if a sample metric belongs to the null distribution. Applying the toolkit software to the WQN stations showed that the performance in predicting the variability of

observed concentration was significantly different than performance metrics with no skill for 12 out of the 18 stations. Assessing model performance using the toolkit can identify stations that have no skill in predicting the variability and be targeted for improving sampling methods, e.g. increase sampling frequency.

An alternative to using statistical models, like WRTDS and LOADEST, that use streamflow to develop continuous nutrient estimations is to use mechanistic models which use meteorological information and land-use data to develop streamflow and nutrient estimates. Modeling complex systems like the nitrogen cycle, for estimating nitrogen load, can produce systematic bias in estimates and bias caused by imperfect calibration. This study proposes a multivariate bias correction technique based on canonical correlation analysis (CCA) that simultaneously reduces the bias in streamflow and loadings while preserving the observed moments (mean, standard deviation). CCA is shown to reduce the bias in the cross-correlation between streamflow and loadings while improving the joint probability in estimating observed streamflow and loadings. This bias-correction technique can be used for both water quantity and water quality purposes which is advantageous for use in forecasting and management strategies.

Mechanistic models like the Soil & Water Assessment Tool (SWAT) model can be adapted for producing streamflow and nutrient load forecasts simultaneously given future climate forcings from atmospheric global circulation models (AGCMs). Downscaled and disaggregated precipitation and temperature from the ECHAM4.5 AGCM can be forced with the SWAT model for long-term assessment (30 years) of 1-month ahead forecasts. Monthly forecasts using forcings from the AGCM are compared with forecasts using observed (perfect forecasts) and climatological forcings (climatological forecasts). Forecasts using future climate information are shown to produce improvement in skill over climatological forecasts for both

streamflow and nutrient load. This study provides a framework for producing long-term forecasts for water quality constituents for the ultimate use in watershed nutrient reduction strategies.

© Copyright 2018 by Dominic Anthony Libera

All Rights Reserved

Reducing Uncertainty in Predicting and Forecasting Nutrient Constituents across the
Southeastern United States

by
Dominic Anthony Libera

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Civil Engineering

Raleigh, North Carolina

2018

APPROVED BY:

Dr. Sankarasubramanian Arumugam
Committee Chair

Dr. Dan Obenour

Dr. Ranji Ranjithan

Dr. Brian Reich

DEDICATION

To my loving Mother and Father. Every day I try to emulate the caring and compassion that you give to our family. You are truly wonderful role models. I love you, Sempre Libera.

“Congrats on your master’s degree Remus! Kobi are you next?”

To my best friend who was lost too soon. I will see you again, but not yet... not yet
In loving memory of LHC III.

BIOGRAPHY

Dominic Anthony Libera was born in Durham, NC on February 3rd, 1989 to Joseph and Patricia Libera. He grew up in Elon, NC with siblings Hannah, Jacob, and Theresa. Growing up in North Carolina he enjoyed being outdoors, backpacking, cycling, and rock climbing. As a senior attending Western Alamance High School, Dominic took Environmental Science AP taught by Mrs. Shelly Pilversack. This course nurtured Dominic's passion of preserving and caring for the environment and ultimately motivated him to pursue a degree in environmental engineering. Dominic attended North Carolina State University and graduated Magna Cum Laude with a Bachelor of Science degree in Environmental Engineering in 2011. While enrolled, Dominic was a member of the Tau Beta Pi honor society, the Chi Epsilon honor society, and the University Scholars Program. In 2013, Dominic received his Master of Science degree in Environmental Engineering from North Carolina State University under the direction of Dr. Francis de los Reyes III. Upon completion of his master's program he has continued his education at the university pursuing a Doctorate in Civil Engineering studying hydrology topics under the direction of Dr. Sankar Arumugam. During his extended time with the University he has been fortunate enough to participate in competitive teaching programs like the Preparing the Professoriate Program and in competitive research programs like the Global Change Fellowship. Now that his student career is coming to an end he is excited to build an academic career in merging the climate change and human health fields. Although, Dominic is graduating with his final degree and leaving NC State after 11 years he will continue to be a diehard Wolfpack fan!

ACKNOWLEDGMENTS

First and foremost, I would like to extend a great deal of gratitude to my advisor Dr. Sankar Arumugam. You have given me invaluable advice and mentored me through rough times. I am forever grateful for your kind and compassionate leadership. It has been a pleasure working with you.

Thank you to Drs. Ranji Ranjithan, Dan Obenour, and Brian Reich for serving on my committee and providing me guidance in my research.

I would like to thank my Mother and Father for always encouraging me and for strengthening my will to succeed; I can always count on your love to calm and comfort me.

I would like to thank my Brother and Sisters for giving me love and support during difficult times.

A great thanks to my friends and colleagues from CHWR: Dr. Amirhossein Mazrooei, Dr. Seung Beom Seo, Dr. Rajarshi Das Bhowmik, Dr. Jason Patskoski, Dr. Sudarshana Mukhopadhyay, and Dr. Dol Raj Chalise.

Thank you to the National Science Foundation for supporting this project under the CAREER Grant “Climate Informed Uncertainty Analyses for Integrated Water Resources Sustainability” No. 0954405.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
Chapter 1. Introduction	1
Chapter 2. In-stream Total Nitrogen Concentration and Load Prediction across the Southeastern United States.....	9
Abstract	9
2.1. Introduction.....	11
2.2. Data Sources	13
2.2.1. WQN Measurements.....	13
2.3. Methodology	15
2.3.1. LOADEST Estimates	15
2.3.2. Leave-One-Out Cross Validation Estimates	17
2.3.3. WRTDS Estimates	18
2.3.4. Post- Load Regression Estimates for Concentration	19
2.3.5. Post-Concentration Regression Estimates for Load	20
2.3.6. Best- fit Performance Metrics	21
2.4. Results	22
2.5. Discussion and Conclusion	31
Chapter 3. A Non-Parametric Bootstrapping Framework Embedded in a Toolkit for Assessing Water Quality Model Performance	35
Abstract	35
3.1. Introduction.....	36
3.1.1. Software	36
3.1.2. Background	36
3.2. Data Sources	39
3.2.1. WQN Measurements.....	39
3.2.2. LOADEST Estimates	39
3.2.3. WRTDS Estimates	40
3.3. Methods.....	41
3.3.1. Motivation.....	41
3.3.2. Non-parametric Bootstrapping.....	42

3.3.3. Performance Metrics	44
3.3.4. P-values and Skill Score	47
3.3.5. Autocorrelation	48
3.4. Toolkit Interface.....	49
3.5. Application and Results	52
3.6. Discussion and Conclusion	57
Chapter 4. Multivariate Bias Corrections of Mechanistic Water Quality Model	
Predictions.....	58
Abstract	58
4.1. Introduction.....	59
4.2. Data Sources	64
4.2.1. Performance Measures	66
4.2.2. Total Nitrogen Loadings	67
4.2.3. SWAT Model Estimates	68
4.3. Methodology: Univariate and Multivariate Bias Correction Methods	69
4.3.1. Motivation.....	69
4.3.2. Simple Linear Regression	70
4.3.3. Multivariate Bias Correction.....	71
4.3.4. Comparison of Bias Correction Techniques over Different Time Scales.....	75
4.3.5. Comparing Cross-Correlations	77
4.3.6. Comparing Joint Likelihoods.....	78
4.4. Results: Performance of Bias Correction Techniques over different time scales	79
4.4.1. WQN Data.....	79
4.4.2. Performance of Bias Correction Techniques over Daily and Monthly time scales	85
4.5. Discussion and Concluding Remarks	89
Chapter 5. Reducing Error in Streamflow and Total Nitrogen Loadings Forecasts from the SWAT Model.....	94
Abstract	94
5.1. Introduction.....	95
5.2. Data	98
5.2.1. WQN Measurements.....	98
5.2.2. Performance Measures	99

5.2.3. Nitrogen Load Estimates.....	101
5.2.4. SWAT Model Estimates	102
5.2.5. Climate Forcings	104
5.3. Methodology	105
5.3.1. Setting up SWAT Model for Streamflow and Load Forecasting	105
5.3.2. Reducing SWAT Model Bias Using Canonical Correlation Analysis	107
5.3.3. Skill Scores and Relative Operating Characteristic	109
5.4. Results	110
5.5. Discussion and Conclusion	120
Chapter 6. Future Work and Conclusion	122
6.1. Future Work	122
6.2. Conclusions	127
REFERENCES	130

LIST OF TABLES

Table 2.1. Summary of 18 selected stations from the Water-Quality Monitoring Network (WQN).	15
Table 2.2. LOADEST regression forms available for load or concentration estimation.	17
Table 2.3. Description of Concentration Estimates	20
Table 2.4. Description of Load Estimates.....	21
Table 4.1. Summary of station data for the selected watersheds showing the number of daily observations for total nitrogen records from the Water Quality Network (WQN) database. Numbers in parenthesis in the far column represent the number of years the daily observations.....	65
Table 4.2. Performance metrics for SWAT model hydrologic and water quality calibration for the observed WQN period.....	69
Table 4.3. Performance of 30-day moving window CCA (in comparison to raw SWAT predictions) in preserving the cross-correlation between observed discharge and LOADEST loadings for the selected three watersheds.	86
Table 5.1. Summary of selected watersheds showing the number of daily observations for total nitrogen records from the Water Quality Network (WQN) database. Numbers in parenthesis in the far column represent the number of years the daily observations span.	99
Table 5.2. Parameters used in the SWAT model for hydrologic calibration of the 3 watersheds using 50 years of observed streamflow.	103
Table 5.3. Parameters used in the SWAT model for water quality calibration of the 3 watersheds using LOADEST load estimates using Model choice #7.....	104
Table 5.4. Forecast performance statistics (post CCA bias-correction) for each forecast (climatological, GCM, and perfect) compared to the observed for the 3 watersheds across the Southeast. Green cells indicate positive improvement of that statistic compared to the raw forecast (pre bias-correction); red cells indicate no improvement.....	111
Table 6.1. Scenarios for application of elemental nitrogen fertilizer.....	124

LIST OF FIGURES

Figure 2.1. Map of WQN stations (circles) considered for the prediction of TN concentration. The stations with shaded watersheds were considered for use in the SWAT model.	14
Figure 2.2. Leave-one-out cross validation approach for estimating concentration/load using the LOADEST model.	18
Figure 2.3. Performance of LOADEST Model #4 (CE1) and LOADEST Model #7 (CE2) for concentration estimation for both NSE (top) and correlation (bottom).	23
Figure 2.4. Performance of load estimation using LOADEST Model#0 (LE1) and LOADEST Model #7 (LE2) for both NSE (top) and correlation (bottom).	24
Figure 2.5. Nash-Sutcliffe Efficiencies for concentration regression (CE2, top) and load regression (LE1, bottom) using LOADEST for 18 WQN stations across the Southeastern U.S.	25
Figure 2.6. Nash-Sutcliffe Efficiencies for concentration regression (CE3, top) and load regression (LE3, bottom) using WRTDS for 18 WQN stations across the Southeastern U.S.	26
Figure 2.7. Performance for concentration estimation using LOADEST Model #7 (CE2), WRTDS (CE3), and the post-load regression technique (CE4) for both NSE (top) and correlation (bottom) metrics.	27
Figure 2.8. Performance of load estimates from LOADEST Model #0 (LE1), WRTDS (LE3) and the post-concentration regression technique (LE5) for NSE (top) and correlation (bottom).	29
Figure 2.9. Role of drainage area in explaining the spatial variability of the skill of concentration regressions from WRTDS (CE3) and LOADEST (CE2).	30
Figure 2.10. Role of drainage area in explaining the spatial variability of the skill of load regressions from WRTDS (LE3) and LOADEST (LE1).	31
Figure 2.11. Role of sampling frequency in explaining the difference of skill of NSE (left) and Pearson’s correlation (right) in concentration estimation from WRTDS (CE3) and LOADEST (CE2).	33
Figure 3.1. Re-sampling framework for developing the specified null distribution of the given performance statistic.	46
Figure 3.2. Nonparametric toolkit- input and output in detail.	50
Figure 3.3. MATLAB interface for the nonparametric toolkit.	51
Figure 3.4. Bootstrapped null distribution with kernel smoothening for two-tailed tests (green line) using $\alpha=0.05$. Black dots show the density of the distribution. Red line denotes model performance measure.	51

Figure 3.5. NSE of predicted load using the LOADEST model and the decision for the null hypothesis using p-values form the null distribution obtained from the nonparametric framework.	54
Figure 3.6. NSE of predicted concentrations using LOADEST model and the decision for the null hypothesis using p-values form the null distribution obtained from the nonparametric toolkit.	55
Figure 3.7. LOADEST performance as shown by skill scores for both Pearson’s and NSE metrics, demonstrating the usefulness of the skill score for application across different performance metrics.	56
Figure 3.8. LOADEST and WRTDS performance, as shown by the skill score, demonstrating which stations have performance significantly different than zero while also showing that WRTDS has preferred concentration estimates for the majority of stations	56
Figure 4.1. Selected three watersheds across the southeast region chosen for bias correction using the SWAT model. Shaded areas represent the drainage area and numbered points provide the stream gauge locations; (1) Tar River at Tarboro, NC (2) Ogeechee River near Eden, GA (3) Escambia River near Century, FL.	66
Figure 4.2. Canonical Correlation Analysis (CCA) Approach for bias correcting SWAT model outputs. Using a split sample approach, canonical coefficients and correlations from the calibration data set are applied to the validation data set.	74
Figure 4.3. Canonical Correlation Analysis (CCA) Approach for bias correcting daily SWAT model outputs based on 30-day moving window for training with the middle day being considered for validation.	77
Figure 4.4. Performance comparison of bias correction techniques, based on NSE, in predicting the observed discharge in the WQN database: (a) canonical correlation analysis and (b) simple linear regression.	80
Figure 4.5. Performance comparison of bias correction techniques, based on NSE, in predicting the observed loadings in the WQN database: (a) canonical correlation analysis and (b) simple linear regression.	81
Figure 4.6. Performance of bias correction techniques, CCA and LR, in preserving the observed cross-correlation between the loadings and discharge in the WQN data for the selected three watersheds.	83
Figure 4.7. Difference between the observed and estimated (left: SWAT model; middle: CCA; right: LR) joint probability of streamflow and loadings of the WQN data for (left) the Tar River at Tarboro, NC (center) Ogeechee River near Eden, GA, and (right) Escambia River near Century, FL basins.	83

Figure 4.8. Performance of 30-day moving window CCA approach in improving the performance of raw daily SWAT model predictions, evaluated by NSE, of observed discharge and LOADEST loadings.	84
Figure 4.9. Performance of 30-day CCA approach in bias correction of raw daily SWAT predictions compared to LOADEST loadings and observed streamflow for the Escambia River near Eden, GA over the period February 22 nd , 1998-March 24 th 1998, a specific window pertaining to a high flow event.	84
Figure 4.10. Performance of CCA approach, evaluated by NSE, in bias correcting monthly SWAT loadings and discharge compared to monthly LOADEST loadings and observed discharge for the Tar River at Tarboro, NC basin under split-sample validation.	87
Figure 4.11. Comparison of WQN observed loadings and discharge (hollow black marker) with raw SWAT predictions (black cross), CCA approach (filled red marker) and Simple Linear Regression (filled black marker) for the selected three river basins over the Southeast US.	88
Figure 5.1. Selected three watersheds across the southeast region chosen for developing nutrient forecasts using the SWAT model. Green shaded areas represent the drainage area and numbered points provide the stream gauge locations; (1) Tar River at Tarboro, NC (2) Ogeechee River near Eden, GA (3) Escambia River near Century, FL.	99
Figure 5.2. SWAT forecasting schematic framework. The SWAT model is run with observed precipitation and temperature for 5 years to initialize the model (light brown) and then run for 1 month using forcings from the ECHAM4.5 GCM (green).	107
Figure 5.3. A CCA model removes the bias GCM forecasts introduced by the adapted SWAT model for each month by training the CCA model using previous observations and perfect forecasts.	108
Figure 5.4. Three-by-three grid showing the number of falling into the BN, N, and AN categories for both the GCM forecasts and observations. This table is used in determining the true positive and false positive rates.	110
Figure 5.5. Performance of bias-corrected streamflow and loadings described by NSE, ρ^2 , and R-RMSE from climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Tar River at Tarboro, NC	113
Figure 5.6. Performance of bias-corrected streamflow and loadings described by NSE, ρ^2 , and R-RMSE from climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Ogeechee River near Eden, GA	114
Figure 5.7. Performance of bias-corrected streamflow and loadings described by NSE, ρ^2 , and R-RMSE from climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Escambia River near Century, FL.	114

Table 5.5. Root mean squared error skill scores (RMSE-SS) from the entire forecasting period, 1961-2010, for both streamflow and loadings for watersheds across the Southeast.	115
Figure 5.8. RMSE-SS for streamflow and loadings for each month for Tar River at Tarboro, NC.	116
Figure 5.9. RMSE-SS for streamflow and loadings for each month for Ogeechee River near Eden, GA.	116
Figure 5.10. RMSE-SS for streamflow and loadings for each month for Escambia River near Century, FL.	117
Figure 5.8. ROC curves for bias-corrected streamflow and loadings binned into the 0.33 quantile (below normal), 0.67 quantile (normal), and 1.0 quantile (above normal) for climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Tar River at Tarboro, NC.	118
Figure 5.9. ROC curves for bias-corrected streamflow and loadings binned into the 0.33 quantile (below normal), 0.67 quantile (normal), and 1.0 quantile (above normal) for climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Ogeechee River near Eden, GA.	119
Figure 5.10. ROC curves for bias-corrected streamflow and loadings binned into the 0.33 quantile (below normal), 0.67 quantile (normal), and 1.0 quantile (above normal) for climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Escambia River near Century, FL.	120
Figure 6.1. Map showing the land-use types of the Escambia River near Century, FL watershed. Agricultural row-crops (AGRR) are shaded light brown.	123
Figure 6.2. Percentage of land cover purposed for each land type in the Escambia River watershed.	123
Figure 6.3. Forecasted mean streamflow and total nitrogen load from a high flow year (1998) and a low flow year (2000) for Escambia River near Century, FL when applying 500 kg/ha of elemental nitrogen to agricultural lands.	125
Figure 6.4. Forecasted mean streamflow and total nitrogen load from a high flow year (1998) and a low flow year (2000) for Escambia River near Century, FL when applying 100 kg/ha of elemental nitrogen to agricultural lands.	126

Chapter 1. Introduction

Nitrogen and phosphorus are essential nutrients for life and they are present in the DNA of all living things. When there is an abundance of both nutrients there is opportunity for growth of organic matter and when there is a deficit, growth declines. These nutrients are the main components in fertilizers and are present in the food and waste of humans and livestock. Runoff and leaching from agricultural and livestock purposed lands are sources of nitrogen and phosphorous loads to natural water systems (Puckett, 1994). Although, these nutrients exist naturally from sources like the degradation of matter, atmospheric deposition, and mineralization from the earth (Seelig, 2000), large loads during runoff events can exceed natural levels of presence. When large amounts of nitrogen and phosphorus enter a water body, such as a lake or reservoir, from tributaries the water body becomes a prime candidate for eutrophication and is susceptible to harmful algae blooms. Algae blooms can degrade the health of water systems and cause stresses on water resource infrastructures. Algae uptake nitrogen and phosphorus to grow and reproduce, and in the process use oxygen (Puckett, 1994). Blooms reduce the amount of dissolved oxygen which can kill fish and aquatic vegetation. Eutrophication in water bodies can also cause other negative effects such as increased stress on water treatment plants by clogging influent intakes. Additionally, algae blooms can cause economic problems such as decreased recreational use of water bodies for swimming and fishing. Identifying tributaries that cause algae blooms can be very difficult for large watersheds (USEPA, 2012), prediction of in-stream nutrients in tributaries flowing into water bodies can help identify sources of impairment.

Nutrient load in streams can vary depending on the hydrological conditions of the stream. For instance, under high flow conditions we expect nutrient load to be higher than normal and under low flow conditions we expect load to be below normal. Discharge of point

sources have little effect on the downstream streamflow meaning that the main driver of nutrient load is the natural water balance model. Studies have shown that climate variability will affect future streamflow events (Oh and Sankarasubramanian, 2012), and we can expect nutrient load to also be effected by climate variability. Improving methods to predict and forecast nutrient load using climate forecasts can improve management of impaired water bodies. The utility of prediction is determined by the quality and quantity of observed nutrient records. In-stream concentration is not frequently measured as other stream characteristics such as streamflow. The United States Geological Survey (USGS) provides daily monitoring of streamflow across the United States from 1931-present. Autonomous monitoring of total nitrogen is still not financially feasible and monitoring requires personnel to physically collect and record samples (Rao et al., 2013). Expensive cost, labor, and time required for monitoring (Cohn et al., 1992) often leaves sparse and non-continuous observations for nitrogen concentration. Some watersheds across the U.S. have daily samples available for 1-3 years but they are usually limited to currently impaired or recently impaired water bodies that have had funds for monitoring programs (Smith et al., 1997). The development of the Water Quality Portal (WQP) (<http://www.waterqualitydata.us/>) has created a centralized tool for accessing historical water quality measurements across the 50 states from the United States Geological Survey (USGS), the Environmental Protection Agency (EPA), and the National Water Quality Monitoring Council (NWQMC) (Read et al., 2017). Data sources having multi-decadal observations such as the U.S. Geological Survey National Stream Water-Quality Monitoring Network (WQN) which includes 679 stations across the United States over the period 1962 to 1995 have scattered and non-continuous water quality records (Alexander et al., 1998). Budgetary constraints caused the sampling frequency of the WQN to vary over the period with most sites starting at monthly

sampling and then dropping to bimonthly and eventually to quarterly sampling starting in 1982 (Alexander et al., 1998). Management models purposed for planning under climate variability require multi-decadal records for calibration and validation. Further, prediction models examining the watershed response of nitrogen loadings to climate events such as El Nino Southern Oscillations (ENSO), which occurs about every 3-7 years, require a longer and continuous period of observations for developing water quality forecasts (Keener et al., 2010). Hence, efforts have focused on using predictions using statistical and mechanistic models over longer periods.

Methods to improve water quality predictions have been carried out using statistical models (Aulenbach, 2013; Moyer et al., 2012; Oh and Sankarasubramanian, 2012; Park and Engel, 2015) and mechanistic models to fill in the data gaps of water quality constituents (Amatya et al., 2013; Jha et al., 2010, 2007; Shrestha et al., 2008). Statistical models for estimating nutrient loadings, such as the Load Estimator (LOADEST) (Cohn et al., 1992, 1989; Cohn, 2005) and Weighted Regression on Time, Discharge, and Season (WRTDS) (Hirsch et al., 2010) model use streamflow, time, and a seasonality component to predict nutrient loadings and develop continuous records over longer time periods. The main challenge with this approach is that predictions are restricted to basins with gauged stations and to periods with observed streamflow, thereby making them not suited for predicting water quality in ungauged basins or for future periods. On the other hand, mechanistic models, such as the Soil & Water Assessment Tool (SWAT) model (Douglas-Mankin et al., 2010a), use soil and land-use information along with long-term observed meteorological records to predict streamflow, nutrient loadings and concentrations. Further, impacts from changes in land-use change, anthropogenic forcings, and management practices can be investigated using the mechanistic model (Douglas-Mankin et al.,

2010b). Complex mechanistic models like SWAT are abstractions of actual physical processes that are difficult to represent; hence, such models potentially exhibit bias in predicting the observed variables, even though these models reasonably capture the variability of observed streamflow and loadings if calibrated well. Model bias can be introduced from imperfect representations of complex natural processes, as seen in GCM modeling of atmospheric physics, or from using imprecise parameter estimates (Maraun, 2016). The SWAT model has abstract representations of actual physical processes that are difficult to represent, specifically the nitrogen cycle which involves the formation and degradation of several nitrogen species. Given the complexity of such processes, model bias can be introduced from model deficiencies in process representation and/or failure to adequately calibrate parameters. The benefit of using these models is in estimating loadings and concentrations on days without measurements, such as historical gaps in measurement or future periods. Forecasting nutrient loadings into the future at 1-2 months ahead could determine future compliance of nutrient loading strategies under high and low conditions.

This dissertation is organized into 4 chapters. **Chapter 2** is a quantitative comparison between two popular regression models (LOADEST and WRTDS) used for predicting nutrient load and concentration. Most studies using LOADEST and WRTDS focus on estimating the long-term mean of concentration or load for specific sites (Hirsch, 2014; Hirsch et al., 2010; Park and Engel, 2015) while this study focuses on the estimating the observed variability of load and concentration at a region-wide scale. Since these models are useful in filling in data gaps they are often used as surrogate observations for long-term calibration and validation of physical based models (Jha et al., 2010, 2007; Kim and Kaluarachchi, 2014). Thus, efforts need to be made for quantifying the amount of observed variability estimated using popular regression

models for the use in long-term calibrations. The framework in this chapter uses LOADEST in a leave-one-out cross-validation approach for similar comparison of estimates obtained using the WRTDS model. The comparison is applied for 18 stations, from the WQN, located in Region 3 of the Southeastern United States. Results from this chapter show that when compared to load regression models, concentration estimates consistently have difficulty in capturing the observed variability and suggest that concentration models do not follow the same underlying distribution of streamflow.

Chapter 3 of this dissertation proposes a non-parametric framework for testing whether the performance of two models (LOADEST and WRTDS) are statistically different from the mean of the observed concentration. A non-parametric significance test is provided based as an open-source toolkit available to water quality modelers estimating concentration. Common significance tests rely on normality assumptions, e.g. Student T-test, which is not the best approach when using distributions from the Nash-Sutcliffe efficiency (NSE) which has shown to change in shape based on sample size and underlying population values (Mueller and Spahr, 2006). Parametric transformations of the NSE distribution have been done to make them nearly normal, however these transformations are restricted to positive NSE values (McCuen et al., 2006). A non-parametric approach is best fit for this study given that it does not require distributional assumptions and is robust for different sample sizes. Other non-parametric approaches have been used for reducing the uncertainty in water quality trend estimates from the WRTDS model (Hirsch et al., 2015). This chapter focuses on reducing the uncertainty in estimating the observed variability of TN concentration from both the LOADEST and the WRTDS models. The toolkit constructs a null distribution of three popular metrics by constructing model estimates based on uncorrelated sets of predictand and predictors. Null

distributions centered on zero (i.e. $NSE=0$), represent models that on average capture the mean of observed concentration. The spread of the null distribution around the mean ($NSE>0$) attribute to the model's random ability in capturing the observed variability. A case study using the toolkit is applied to the same 18 stations studied in Chapter 2. The majority of stations (12 out of 18) having low performance (<0.4 NSE) show that the model performance is significantly different than the observed mean. Additionally, a new skill score is introduced to provide a more meaningful statistic for comparison across underlying performance distributions.

Chapter 4 of the dissertation uses load estimates from LOADEST as a surrogate truth for extending mechanistic model predictions beyond the observed period, which is becoming increasingly popular (Jha et al., 2010, 2007). Three of the 18 stations were selected for representation in the SWAT model. Load estimates from the LOADEST model were used to calibrate the watersheds over a 60-year period. A significant bias was present specifically for load estimates after calibration. Mechanistic models can have bias (deviations from observed values) from imperfect representations of natural processes and/or improper parametrization during calibration (Maraun, 2016). Improving calibration is the best way to reduce bias but it is not always the easiest (Ehret et al., 2012). Thus, to reduce bias from mechanistic model outputs a post-simulation bias-reduction must be used. Univariate bias-removal procedures have been used for streamflow (Stewart and Reagan-Cirincione, 1991) and nutrient load (Leisenring and Moradkhani, 2012; Windolf et al., 2011), however they do not preserve the joint moments. This chapter proposes the use of canonical correlation analysis (CCA) as a multivariate bias-correction technique for the simultaneous removal of bias in streamflow and nitrogen load. The technique proves to be advantageous in preserving the observed cross-correlation and estimating the joint-likelihood of streamflow and TN load while reducing individual bias in streamflow and

loads. Since nutrient load and streamflow are directly related, this technique is especially useful in bias correcting raw mechanistic model output.

The final chapter, **Chapter 5**, uses the findings from previous chapters in examining the skill of 1-month ahead forecasts of nutrient loads and streamflow using the SWAT model. The SWAT model is useful for forecasting since it can simultaneously produce streamflow and nutrient load forecasts while the LOADEST and WRTDS models require forecasted streamflow to make nutrient load forecasts. Physical based models, like the SWAT model, have shown to produce skillful streamflow forecasts using climate information (Mazrooei et al., 2015; Sinha and Sankarasubramanian, 2013). Less attention has been given to retrospective nutrient load forecasting with studies rather focusing on climate change studies (Serpa et al., 2017) into the future or historical estimation (Jha et al., 2007). The SWAT model is adapted for 1-month ahead forecasting where simulated state variables are updated at the beginning of each month before the forecast using a looping framework. The SWAT model is difficult to modify for updating initial conditions using the FORTRAN source code; a MATLAB version was created for updating conditions but only for streamflow, not nutrients (Sun et al., 2015). This study provides a framework for updating both streamflow and nutrient related conditions at the beginning of each month. Precipitation and temperature forecasts from the ECHAM4.5 global circulation model used as inputs to the initial-conditions updated SWAT model to develop forecasts of streamflow and load one month ahead. Using the CCA bias-correction framework, from Chapter 4, monthly forecasts of streamflow and load were bias corrected, the forecast skill is also quantified for the three selected basins from Chapter 3. Further, an analysis examining the effects of watershed wide reduction of fertilizer on downstream nutrient forecasts is included in **Chapter 6**.

Overall, this dissertation aims to reduce the uncertainty in nutrient load prediction and forecasting for predicting water quality downstream. Improving the skill of model predictions can also potentially identify sources of nutrient loads and prevent water body impairments that require BMPs for reducing loads. Given the current state of available water quality observations it is imperative to utilize statistical and physical techniques to improve predictions that rely on the sparse observation sets. This study focused on using the best estimates from current methods in addition to proposing new methods for assessing them and improving their predictions and reconstruction over the Southeastern U.S.

Chapter 2. In-stream Total Nitrogen Concentration and Load Prediction across the Southeastern United States

Abstract

Water quality observations for streams are usually not available continuously because of the expensive cost in data collection and analysis. Predictive models can be useful tools for estimating water quality constituents on non-observation days to fill in the data gaps using easily available information such as streamflow as predictors. Two commonly used regression models are the LOADEST model and the WRTDS model developed by the USGS. These models have been used to develop regressions for both nutrient load and concentration, with the skill of the former being typically higher than the latter. This is partly due to the ability of streamflow in explaining the variability in loadings and the ratio estimation effect causing concentration (loadings divided by streamflow) regression skill to be lower than the skill of load regression. This study examines the performance of LOADEST and WRTDS estimates of total nitrogen (TN) load and concentration for 18 Water Quality Network (WQN) stations over the Southeast region. Leave-one-out cross validation estimates from both the LOADEST and WRTDS models show estimates for TN load using the LOADEST model capture more observed variability, measured by Nash-Sutcliffe Efficiency (NSE), than the WRTDS model for 12 out of the 18 stations. However, when estimating TN concentration, WRTDS model estimates have higher NSE values for 14 out of the 18 stations. Further, we show that using load estimates from the LOADEST model to get concentration estimates, given observed streamflow, yield better estimates than the WRTDS model for 12 out of the 18 stations. Relationships between basin characteristics and model performance were examined showing that smaller watersheds tended

to perform better in estimating TN concentration but in predicting TN loads there were no observable relationships.

2.1. Introduction

Advances in low-cost autonomous methods for measuring water quality constituents such as temperature, pH, and dissolved oxygen have been successfully developed (Rao et al., 2013). Measurements that require filtering and/or addition of reagents, such as total nitrogen measurements, largely still rely on the collection and analysis by a trained technician. Given the cost associated with daily continuous observations, TN samplings are infrequently collected and are usually limited to impaired water bodies with nutrient reduction programs (Cohn, 2005; Quilbé et al., 2006). Analyzing limited TN observations is difficult especially for long-term studies looking at the changes in mean annual concentration due to potential climate change (Hirsch et al., 2015) or for developing seasonal load forecasts (Oh and Sankarasubramanian, 2012). Accurate predictions of in-stream concentration of total nitrogen can assist in placement of best management practices (BMP) (Bosch et al., 2013) and could provide better estimates for total maximum daily load (TMDL) programs (Kim et al., 2012). The Water Quality Network (WQN) offers TN observations for multiple decades across the United States, however the observations are not continuous with as little as 1-2 samplings every two months (Alexander et al., 1998). Unfortunately, since nutrient observations are sparse and non-continuous, it makes it extremely difficult to accurately predict concentration/load over a long period which is necessary for effective management practices (Smith et al., 1997). Predictive regression models estimate water quality constituents on non-observation days using other continuously available data such as streamflow. Two such regression models, USGS's Load Estimator (LOADEST) model (Cohn, 2005; Cohn et al., 1992, 1989) and the Weighted Regression on Time, Discharge, and Season (WRTDS) (Hirsch et al., 2010), are commonly used for load estimation (Cohn, 2005; Hirsch, 2014; Jha et al., 2007; Mueller and Spahr, 2006; Oh and Sankarasubramanian, 2012) and

for the estimation of mean annual concentration (Hirsch et al., 2015). Regional assessment of loading regression models have been conducted (Oh and Sankarasubramanian, 2012; Smith et al., 1997), but little attention has been given towards concentration regressions at the regional scale. In-stream concentration estimation is important since some states (e.g. Florida) over the Southeastern U.S. (SEUS) have adopted numeric nutrient criteria for nitrogen in rivers and streams for the Federal Clean Water Act (Public Law 92-500) purposes (US EPA, 1998). Concentration limits for total nitrogen are also issued for specific regions and are typically applied to all rivers and streams included in the region. For example, rivers and streams in the Pensacola Bay Watershed in the western panhandle of Florida, which includes Escambia River, have an annual mean total nitrogen limit of 0.67 mg/L (FL Surface Water Quality Standards, Ch. 62–302).

Prediction of concentration is usually done using a statistical regression model or a mechanistic, physical based model. Mechanistic models have certain advantages in prediction of streamflow and water quality by incorporating land-use and climate changes. Additionally, a calibrated watershed model given observed precipitation and temperature can predict discharge and water quality constituents for ungauged stations (R. Srinivasan et al., 2010). The Soil and Water Assessment Tool (SWAT) uses land cover, climate information, and have the capacity to incorporate point and non-point source contributions (Douglas-Mankin et al., 2010; Santhi et al., 2001). Streamflow records available from the USGS allow successful hydrologic calibration of SWAT models over the long periods of time (+50 years). Calibration of water quality constituents, like total nitrogen concentration, over the long period is difficult given discontinuous and sparse observations. Thus, there is a need for accurate model predictions over long periods for water quality constituents (Smith et al., 1997). Popular models such as

LOADEST and WRTDS models can provide predictions for nutrient concentration and load given streamflow and time as predictors. The LOADEST model performs regression on log concentration or loadings, while WRTDS considers only log concentration regression models. Further, the WRTDS model uses leave-one-out estimation of concentration by using unique coefficients and bias-correction factors (Hirsch, 2014). This process may improve prediction for the observed period but may result in limited skill over the long term. In this study, we apply the LOADEST model in a similar leave-one-out cross validation to get estimates using unique coefficients and standard errors for each day in the observed period. Considering estimates from the regression models as the “truth” allows for long term water quality calibration and bias-correction. We consider the following three estimates of concentration as the “truth”: (a) LOADEST concentration estimates, (b) WRTDS concentration estimates, and (c) concentration estimates from dividing LOADEST load estimates by streamflow, i.e. $(C=\hat{L}/Q$, where \hat{L} are load estimates from the LOADEST model). This study examines the performance, in NSE, of two popular regression models in estimating total nitrogen concentration and load for 18 WQN stations across the Southeast.

2.2. Data Sources

2.2.1. WQN Measurements

This study considers 18 stations from the USGS water resource Region 3 of the SEUS (Figure 2.1). These 18 stations have been considered for forecasting seasonal nutrients using climate information (Oh and Sankarasubramanian, 2012). Sampling frequencies of several sites started at monthly timescales in 1962 and reduced to bimonthly and seasonal sampling from 1982 to 1995. The number of daily observations for total nitrogen records averaged over 18 stations is about 200 days extending about 20 years (Table 2.1). These stations belong to both

the National Stream Water-Quality Monitoring Network (WQN) and the Hydro-Climatic Data Network (HCDN) (Alexander et al., 1998). Stations belonging to the HCDN have streamflow that is minimally impacted by anthropogenic influences like artificial storage or pumping preserving the climate signal in streamflow (Slack et al., 1993; Vogel and Sankarasubramanian, 2005). The WQN is a combination of two subsets of networks, the National Stream Quality Accounting Network (NASQAN) with observations from 1962 to 1995 and the Hydrologic Benchmark Network (HBN) with observations from 1973 to 1995. Total nitrogen concentration is reported as the summation of nitrate-nitrogen ($\text{NO}_3\text{-N}$), nitrite-nitrogen ($\text{NO}_2\text{-N}$), ammonia-nitrogen ($\text{NH}_3\text{-N}$) and organic nitrogen. Total nitrogen load, reported as kilogram per day, is computed as the product of concentration and streamflow on the day of observation from USGS streamgages and the appropriate conversion factor.

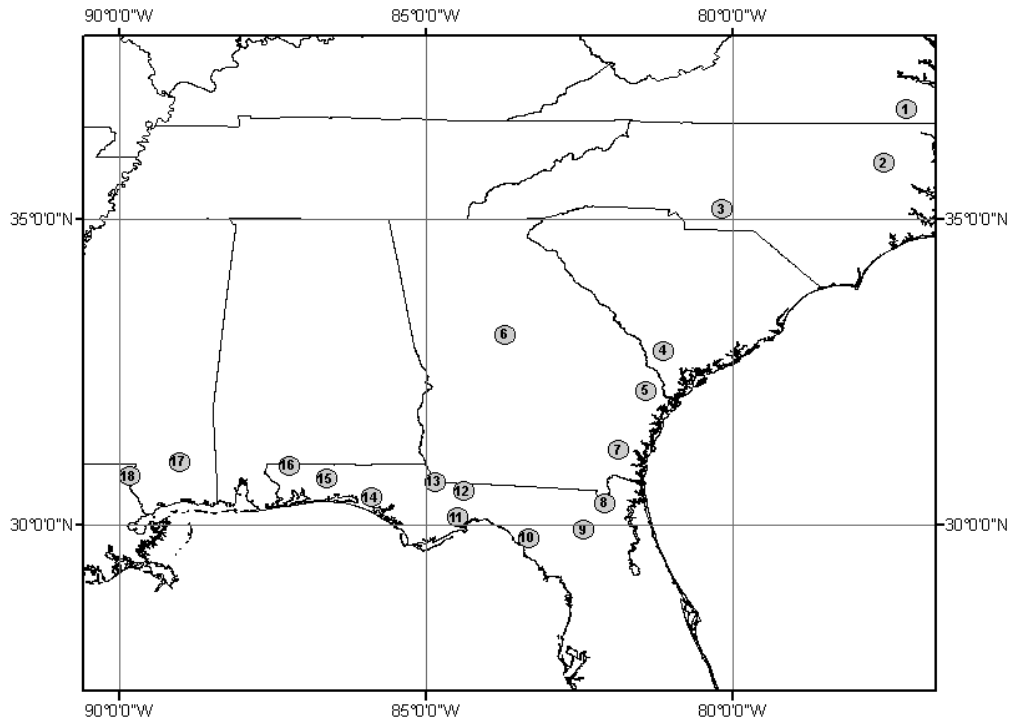


Figure 2.1. Map of WQN stations (circles) considered for the prediction of TN concentration. The stations with shaded watersheds were considered for use in the SWAT model.

Table 2.1. Summary of 18 selected stations from the Water-Quality Monitoring Network (WQN).

Station Index	Station Number	Station Name	Drainage Area (km ²)	Number of years (# of daily Obs.)	Load Model #
1	02047000	Nottoway River near Sebrell, VA	3732.17	18(102)	9
2	02083500	Tar River at Tarboro, NC.	5653.94	23(167)	9
3	02126000	Rocky River near Norwood, NC	3553.46	5(71)	4
4	02176500	Coosawhatchie River near Hampton, SC	525.77	13(101)	9
5	02202500	Ogeechee River near Eden, GA	6863.47	22(148)	9
6	02212600	Falling Creek near Juliette, GA	187	2(56)	3
7	02228000	Satilla River at Atkinson, GA	7226.07	20(123)	9
8	02231000	St. Marys River near Macclenny, FL	1812.99	14(110)	6
9	02321500	Santa Fe River at Worthington Springs, FL	1489.24	5(83)	9
10	02324000	Steinhatchee River near Cross city, FL	906.5	12(94)	8
11	02327100	Sopchoppy River near Sopchoppy, FL	264.18	5(128)	9
12	02329000	Ochlockonee River near Havana, FL	2952.59	22(136)	9
13	02358000	Apalachicola River at Chattahoochee, FL	44547.79	23(152)	9
14	02366500	Choctawhatchee River near Bruce, FL	11354.51	21(132)	2
15	02368000	Yellow River at Milligan, FL	1616.15	21(124)	9
16	02375500	Escambia River near Century, FL	9885.98	22(147)	8
17	02479155	Cypress Creek near Janice, MS	136.23	16(55)	4
18	02489500	Pearl River near Bogalusa, LA	17023.99	22(136)	9

2.3. Methodology

2.3.1. LOADEST Estimates

The USGS's Constituent Load Estimator (LOADEST) tool is a FORTRAN program available for estimating in-stream constituent loadings using streamflow, time, and season as predictors (Cohn, 2005; Cohn et al., 1992, 1989). The program offers ten preset model choices (Table 2.2) and a custom model definition option for estimating log load. Observed streamflow and concentration from the WQN series are used to fit model coefficients for the chosen model form, defined in the "header.inp" file. When model choice #0 is defined, models 1-9 are considered and the best model is chosen based on the lowest Akaike Information Criterion (AIC). Streamflow and decimal time are centered for reducing multicollinearity before calibration (Cohn, 2005). When using uncensored observations, maximum likelihood estimation (MLE) is used to determine model coefficients and standard errors. Estimates for constituent load are found using the fitted model coefficients with the time series defined in the estimation

file, “est.inp”. Log load estimates are transformed back to the original space using a bias-correction factor, defined as the function phi with LOADEST (Likeš, 1980), to yield minimum variance unbiased estimates of instantaneous load (Cohn, 2005). Performance of the LOADEST model is evaluated by comparing the concentration /load estimations with the observed WQN data.

Although, LOADEST is built for the estimation of nutrient load in the log-space using the pre-defined models, estimation of log concentration can be specified by using the custom model option (model #99). User defined streamflow and concentration is still used to compute load in model choice #99, so the model must be tricked into doing regression on concentration (Runkel et al., 2004). Setting observed streamflow equal to “0.408735” cubic feet per second, will yield a load (kg/day) equal to the value of concentration (mg/L) after streamflow is converted to the appropriate units. Actual values of observed streamflow can then be provided as additional variables in the calibration file. When estimating concentration, the best model selection feature (model #0) is disabled and must be done manually.

Two load estimates using LOADEST were considered as the “truth” for each station: load estimates using model #0 (LE1) and load estimates only using model #7 (LE2). Predictor Q_i is the centered observed daily streamflow on day i , T_i is the centered decimal time, and $\hat{\alpha}_s$ are model estimated coefficients (Table 2.2). Model LE1 represents the best-case for load estimation; each station can have unique model forms picked using AIC and are listed in Table 2.1. Model LE2 holds the model form constant across all stations. Two concentration estimates are considered: concentration estimates from model #4 (CE1) and model #7 (CE2).

Table 2.2. LOADEST regression forms available for load or concentration estimation.

Model #	Regression Model Form
0	Choose best model based on Akaike Information Criteria
1	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i)$
2	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \ln(Q_i^2)$
3	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3(T_i)$
4	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \sin(2\pi T_i) + \widehat{\alpha}_4 \cos(2\pi T_i)$
5	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \ln(Q_i^2) + \widehat{\alpha}_4(T_i)$
6	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \ln(Q_i^2) + \widehat{\alpha}_3 \sin(2\pi T_i) + \widehat{\alpha}_4 \cos(2\pi T_i)$
7	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \sin(2\pi T_i) + \widehat{\alpha}_4 \cos(2\pi T_i) + \widehat{\alpha}_5(T_i)$
8	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \ln(Q_i^2) + \widehat{\alpha}_4 \sin(2\pi T_i) + \widehat{\alpha}_5 \cos(2\pi T_i) + \widehat{\alpha}_6 \ln(T_i)$
9	$\widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \ln(Q_i^2) + \widehat{\alpha}_4 \sin(2\pi T_i) + \widehat{\alpha}_5 \cos(2\pi T_i) + \widehat{\alpha}_6 \ln(T_i) + \widehat{\alpha}_7 \ln(T_i^2)$

2.3.2. Leave-One-Out Cross Validation Estimates

Since the WRTDS model uses a leave-one-out cross validation approach to produce log concentration estimates (Hirsch, 2014), the LOADEST model was also used in a similar way for estimating both log load and log concentration. Each day in the WQN period had separate LOADEST models having unique model coefficients and standard errors to produce estimates for that day. The entire observed streamflow and concentration time series was included in the calibration file except for values on the target day (Figure 2.2). Then observed streamflow on the target day was included in the estimation file time series to retain load/concentration estimates on that target day. Model estimates for load/concentration were constructed by considering each day in the WQN period as the target day and retaining estimates specific for

each target day to build back an entire time series. Using this framework, estimates were produced without knowledge of the actual observed values on each target day.

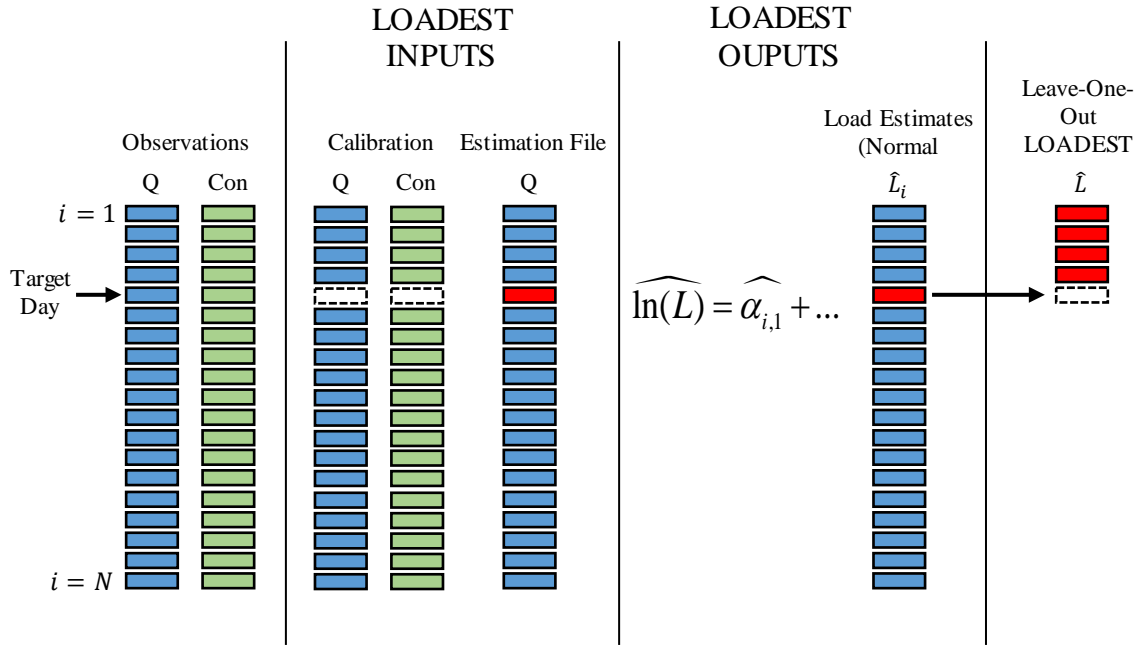


Figure 2.2. Leave-one-out cross validation approach for estimating concentration/load using the LOADEST model.

2.3.3. WRTDS Estimates

The Weighted Regression on Time, Discharge, and Season (WRTDS) model is an estimation method offered in the Exploration and Graphics for RivEr Trends (EGRET) R-package used for studying long-term changes in water quality and streamflow (Hirsch et al., 2010). Station and parameter ID numbers are used to retrieve streamflow from USGS streamgages and constituent concentration from the EPA STORET dataset. Nitrogen concentration from the WQN data set is considered for this study which differs from the EPA STORET data set, in this case the WQN has more total nitrogen observations available for the majority of stations, so WQN values had to be manually uploaded using a user-specified file.

Since the WRTDS model uses a sole regression model (2.1) for estimation of log concentration, only one model for estimation of concentration was considered (CE3).

$$\ln(c_i) = \widehat{\beta}_1 + \widehat{\beta}_2 \ln(Q_i) + \widehat{\beta}_3 \ln(T_i) + \widehat{\beta}_4 \sin(2\pi T_i) + \widehat{\beta}_5 \cos(2\pi T_i) + \widehat{\varepsilon}_i \quad (2.1)$$

where c_i is the observed daily total nitrogen concentration in mg/l on day i , Q_i is the mean daily streamflow for day i , T_i is the centered decimal time, $\widehat{\beta}$ s are model estimated coefficients, and $\widehat{\varepsilon}_i$ denotes the estimated model residual on day i . The model form in (1.1) most closely follows LOADEST's model #7 (Table 2.1), with the main difference being the technique for parameter estimation. The WRTDS model estimates log concentration for each day in the observed period by using (2.1) with fit coefficients using a weighted Tobit regression. Log concentration estimates for each day in the observed period have unique regression coefficients and standard errors using a leave-one out cross validation approach. The weight for each observation is determined by the product of individual weights for discharge, time, and season obtained from a tricube function (Hirsch et al., 2010). Estimates for each day are transformed back to the original space using respective bias-correction factors (BCF) shown in (2.2) (Hirsch et al., 2010). Load estimates from WRTDS are produced using the product of concentration estimates and observed streamflow (LE3), a significant difference from LOADEST which uses a regression model estimate load.

$$BCF = \exp\left(\frac{SE^2}{2}\right) \quad (2.2)$$

2.3.4. Post- Load Regression Estimates for Concentration

Since load is the product of concentration and streamflow, one could use a post-regression technique to estimate concentration by dividing LOADEST load estimates by observed streamflow. This technique is a simple approach for estimating concentration without

having to perform a separate regression in LOADEST. Form the LOADEST model, we considered two load estimates: LE1 and LE2. Using equation (2.3) two post-regression estimates for concentration (CE4 and CE5) can be produced by dividing LE1 or LE2, denoted by (\hat{L}), by observed streamflow (Q). This technique cannot be applied to the WRTDS model since it does not estimate log load using a regression method. A summary of the concentration estimation methods is listed in Table 2.3.

$$\hat{c} = \hat{L} / Q \quad (2.3)$$

Table 2.3. Description of Concentration Estimates

Concentration Estimates	Formula	Model Technique
CE1	Model #4: $\ln(c_i) = \hat{\alpha}_1 + \hat{\alpha}_2 \ln(Q_i) + \hat{\alpha}_3 \sin(2\pi T_i) + \hat{\alpha}_4 \cos(2\pi T_i) + \hat{\varepsilon}_i$	LOADEST
CE2	Model #7: $\ln(c_i) = \hat{\alpha}_1 + \hat{\alpha}_2 \ln(Q_i) + \hat{\alpha}_3 \sin(2\pi T_i) + \hat{\alpha}_4 \cos(2\pi T_i) + \hat{\alpha}_5(T_i) + \hat{\varepsilon}_i$	LOADEST
CE3	Equation 1: $\ln(c_i) = \hat{\beta}_1 + \hat{\beta}_2 \ln(Q_i) + \hat{\beta}_3 \ln(T_i) + \hat{\beta}_4 \sin(2\pi T_i) + \hat{\beta}_5 \cos(2\pi T_i) + \hat{\varepsilon}_i$	WRTDS
CE4	$\hat{c} = \hat{L} / Q$; where \hat{L} is LOADEST Model #0 (LE1)	Post-LOADEST estimation
CE5	$\hat{c} = \hat{L} / Q$; where \hat{L} is from LOADEST Model #7 (LE2)	Post-LOADEST estimation

2.3.5. Post-Concentration Regression Estimates for Load

Likewise, load estimates can be produced by multiplying estimates of concentration by observed streamflow using (2.4). Two model forms in LOADEST are used for estimating concentration (CE1 and CE2) (Table 2.3) thus, two estimates for load can be produced (LE4 and LE5). Load estimation from the WRTDS model is already done using this technique as shown in LE3. A summary of the load estimation models considered are listed in Table 2.4.

$$\hat{L} = \hat{c} \cdot Q \quad (2.4)$$

Table 2.4. Description of Load Estimates

Load Estimates	Formula	Model Technique
LE1	Model #0: Best Selection from #1-9	LOADEST
LE2	Model #7: $\ln(L_i) = \hat{\alpha}_1 + \hat{\alpha}_2 \ln(Q_i) + \hat{\alpha}_3 \sin(2\pi T_i) + \hat{\alpha}_4 \cos(2\pi T_i) + \hat{\alpha}_5(T_i) + \hat{\varepsilon}_i$	LOADEST
LE3	$\hat{L} = \hat{c} \cdot Q$; where \hat{c} is from WRTDS (CE3)	WRTDS
LE4	$\hat{L} = \hat{c} \cdot Q$; where \hat{c} is from LOADEST Model #4 (CE1)	Post- LOADEST estimation
LE5	$\hat{L} = \hat{c} \cdot Q$; where \hat{c} is from LOADEST Model #7 (CE2)	Post- LOADEST estimation

2.3.6. Best-fit Performance Metrics

The metrics used in this study to quantify a model's ability in capturing the observed variability of concentration/load are Pearson's Correlation Coefficient (ρ) and Nash-Sutcliffe Efficiency (NSE). NSE measures the squared deviations of the model (\hat{X}_i) to the observed values (X_i) with respect to the squared deviations of observations to the mean (\bar{X}) of observations as shown in (2.5). The NSE statistic ranges from negative infinity to one, with a value of one indicating that the model perfectly captures the observed variability. A value of zero, indicates that the observed mean is a better predictor than the proposed model (Nash and Sutcliffe, 1970).

$$NSE = 1 - \frac{\sum_{i=1}^N (X_i - \hat{X}_i)^2}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (2.5)$$

2.4. Results

Comparison between estimates CE1 and CE2 show the added benefit of including log time as a predictor in the regression model for concentration estimation in LOADEST, shown in Figure 2.3. Estimate CE2 shows noticeable improvement ($>10\%$ NSE and ρ) for 7 out of the 18 stations in the southeast. Comparison between LE1 and LE2 shows that LE2 does not improve performance ($<5\%$ NSE and ρ) for any stations, indicating that LE1 is preferred for load estimation in LOADEST. This is to be expected since model choice #7 was not chosen for any stations for load estimation using LOADEST (Table 2.1). The best concentration estimation method using LOADEST is CE1 (Figure 2.5-top) and the best load estimation method is LE1 (Figure 2.5-bottom). A consistent difference between the NSE magnitudes for concentration and load estimation is present across all stations in the SEUS when using LOADEST (Figure 2.5). Concentration (CE3) and load (LE3) estimations using WRTDS are shown in Figure 2.6. Throughout the southeastern region, concentration models perform poorly (NSE <0.4) for all 18 stations when compared to load estimation performance (NSE >0.8) regardless of the model type used (LOADEST or WRTDS). As shown in Figure 2.6 (top), WRTDS estimates (CE3) with NSE values above 0.26 are shown in Florida. Load estimates from WRTDS (Figure 2.6 -bottom) perform higher (NSE >0.6) uniformly across all stations than concentration estimates. There appears to be a better performance when the concentration estimates from WRTDS (CE3) are multiplied by observed streamflow to get load estimates (LE3). The low performance throughout the region is not seen in load estimates from WRTDS. This phenomenon suggests that variability of concentration and load estimates differ substantially, specifically for stations with limited data.

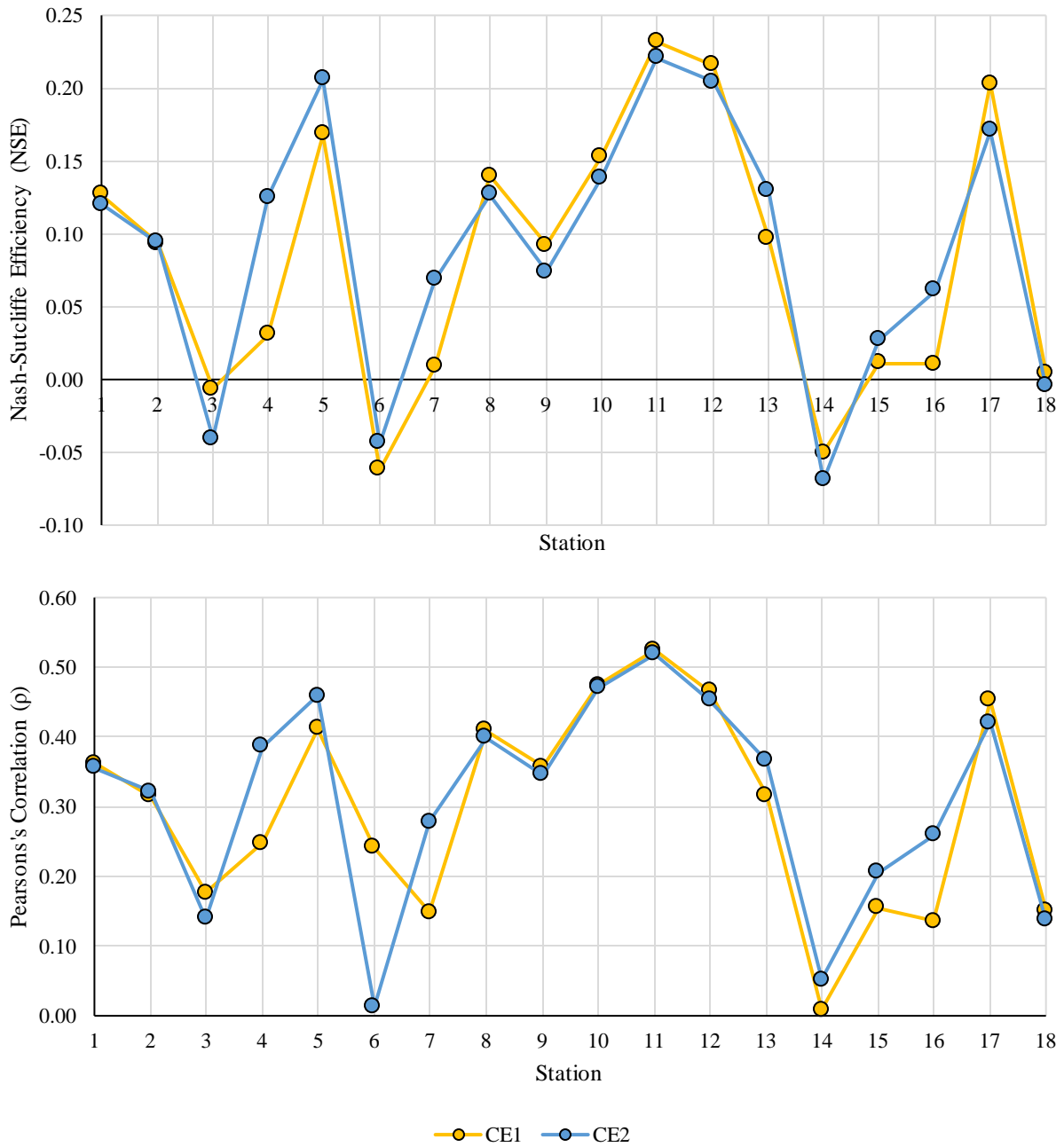


Figure 2.3. Performance of LOADEST Model #4 (CE1) and LOADEST Model #7 (CE2) for concentration estimation for both NSE (top) and correlation (bottom).

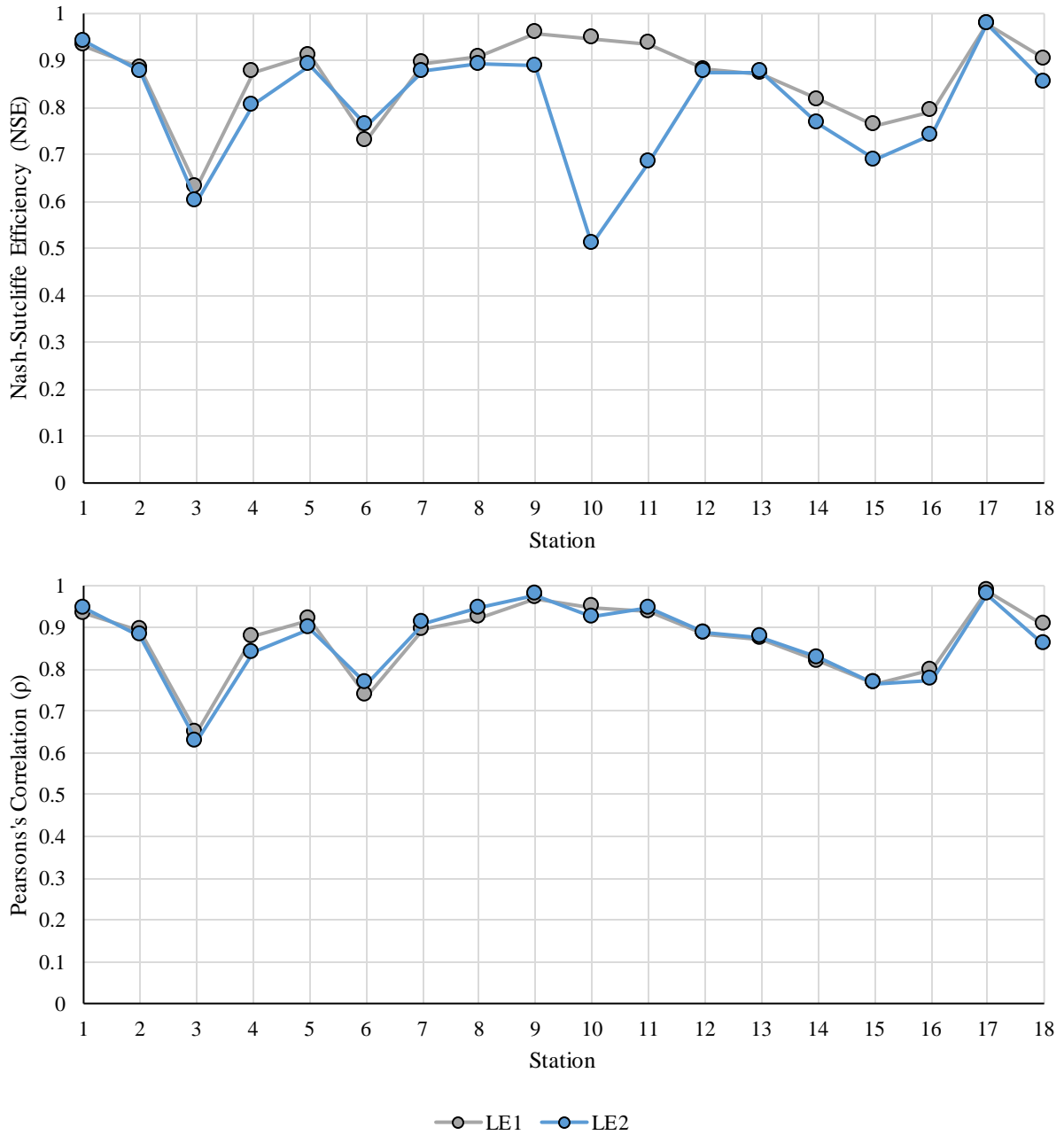


Figure 2.4. Performance of load estimation using LOADEST Model#0 (LE1) and LOADEST Model #7 (LE2) for both NSE (top) and correlation (bottom).

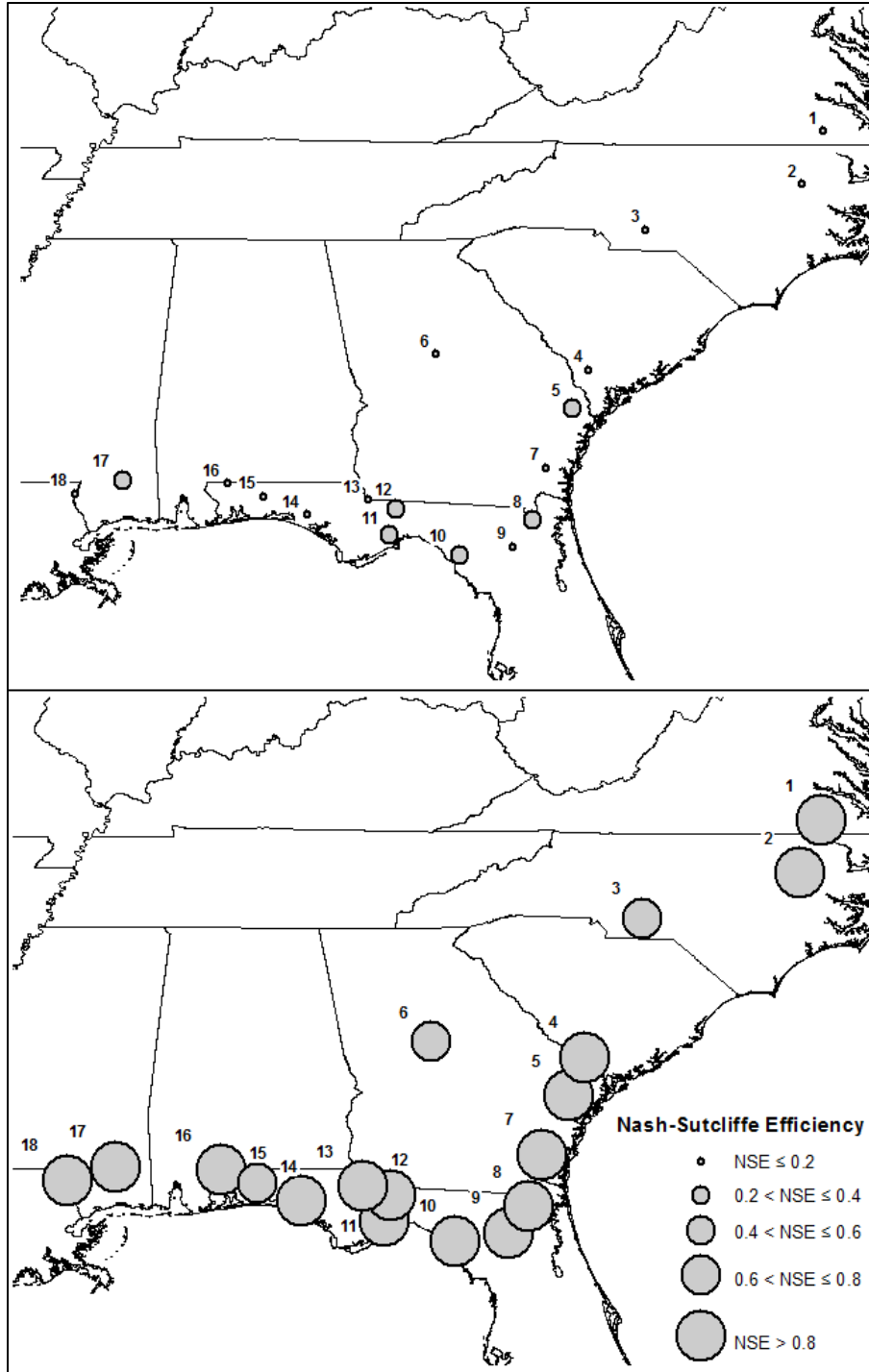


Figure 2.5. Nash-Sutcliffe Efficiencies for concentration regression (CE2, top) and load regression (LE1, bottom) using LOADEST for 18 WQN stations across the Southeastern U.S.

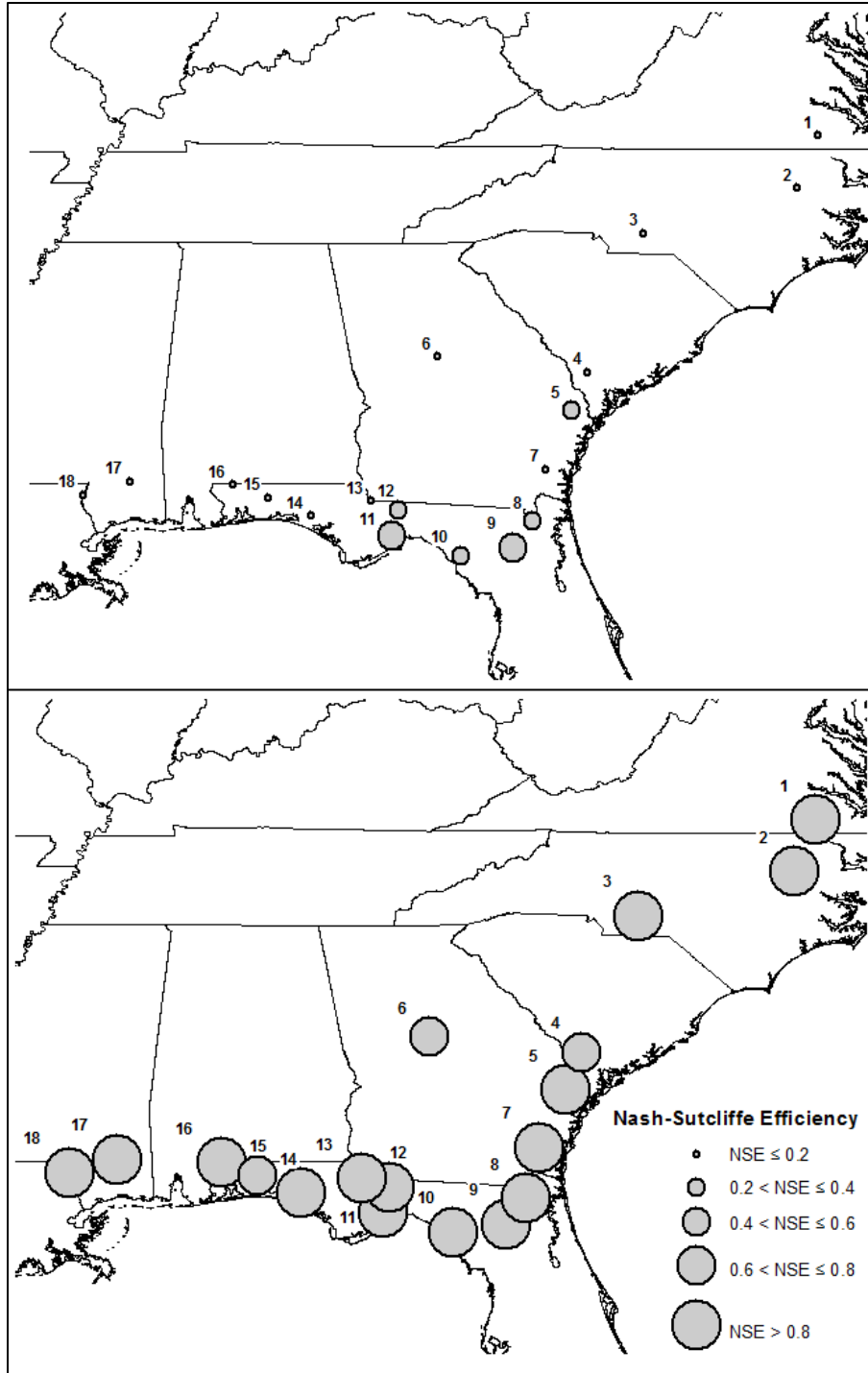


Figure 2.6. Nash-Sutcliffe Efficiencies for concentration regression (CE3, top) and load regression (LE3, bottom) using WRTDS for 18 WQN stations across the Southeastern U.S.

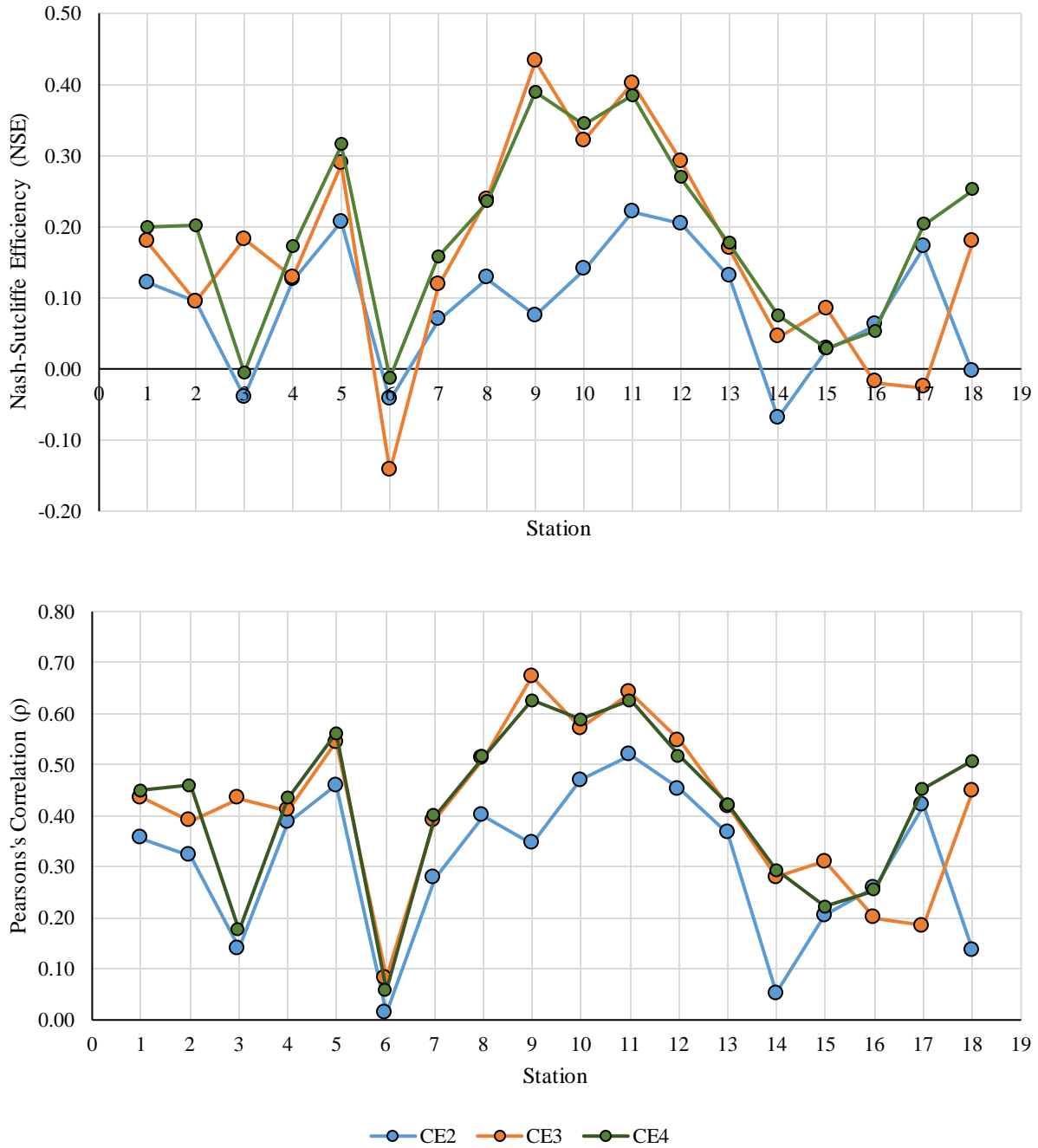


Figure 2.7. Performance for concentration estimation using LOADEST Model #7 (CE2), WRTDS (CE3), and the post-load regression technique (CE4) for both NSE (top) and correlation (bottom) metrics.

Another method to estimate in-stream concentration, illustrated in CE4, uses the best load estimates from LOADEST (LE1) and divides them by observed streamflow. This ratio estimation method provides higher performance than both WRTDS and LOADEST models for

11 out of the 18 stations shown in Figure 2.7. All but three stations showed improvement in NSE for concentration estimation when using WRTDS as compared to LOADEST (Model #7). The regression model forms for estimating concentration used in LOADEST (CE2) and WRTDS (CE3) are identical, the only difference is that WRTDS uses a semi-parametric regression where LOADEST uses a least-squares regression. Estimates from CE5 were excluded from this figure since it uses load estimates from LE2, which performs lower than LE1. We can conclude that WRTDS performs better than LOADEST in predicting concentration for the majority of stations but concentration estimates from CE4 capture more of the observed variability for most of the stations.

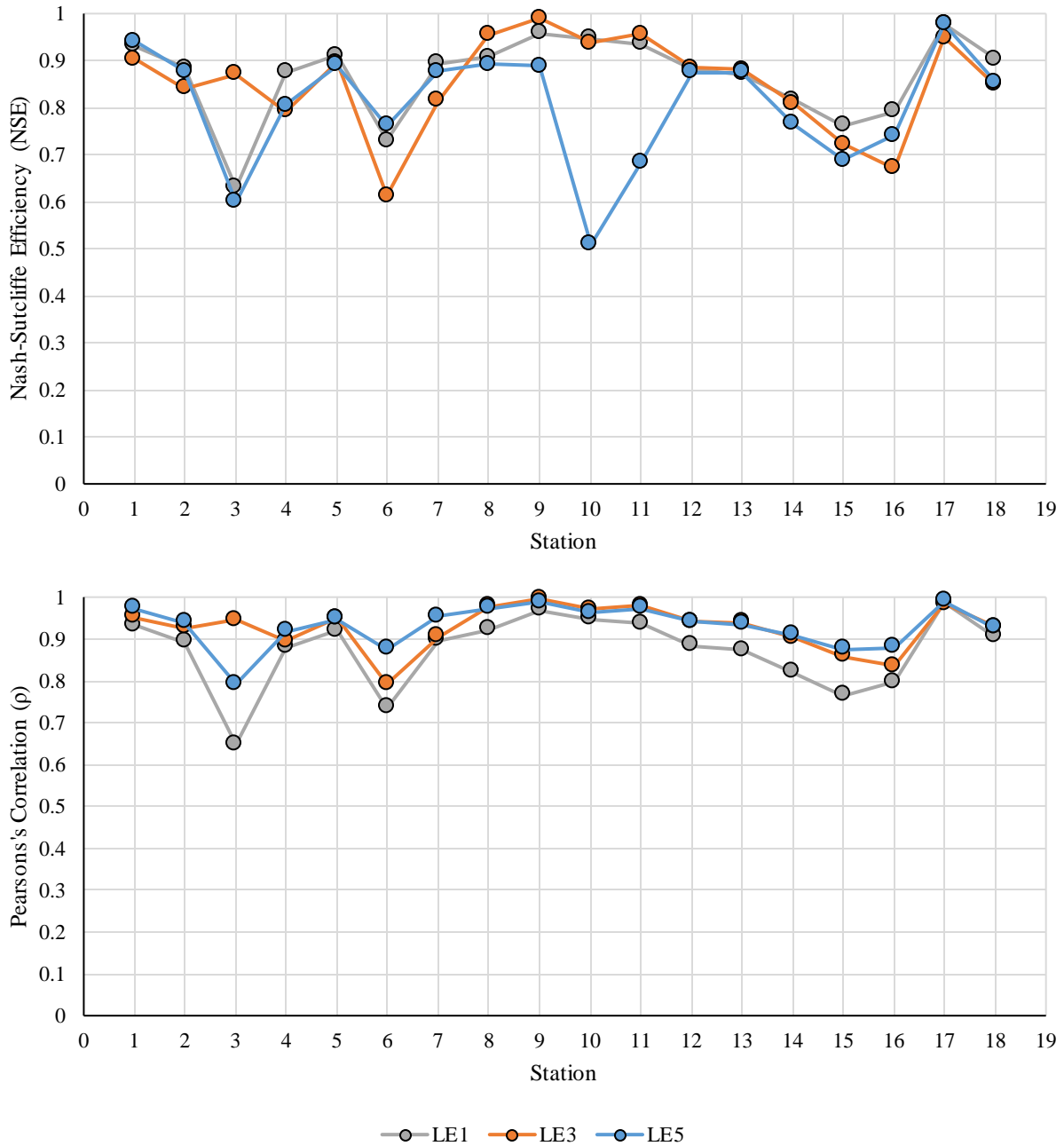


Figure 2.8. Performance of load estimates from LOADEST Model #0 (LE1), WRTDS (LE3) and the post-concentration regression technique (LE5) for NSE (top) and correlation (bottom).

Load estimates from LOADEST Model #0 (LE1) and WRTDS (LE3) are shown in Figure 2.8. The LOADEST model has higher NSE values for 12 out of the 18 stations, indicating it is the preferred model for load estimation. Estimates from LE5 and LE3 do not

perform better than estimates from LE1, showing that estimates from regressions on log load perform better than using concentration estimates and post multiplying by streamflow to get load estimates. We do not show LE4, since it uses concentration estimates from CE1 which perform worse than the concentration estimates from CE2 used in LE5.

Watershed drainage area can play a significant role in the magnitude and shape of streamflow time series and can potentially influence loadings too. As shown in Figure 2.9, we see that smaller watersheds tend to perform better for concentration models than larger watersheds irrespective of the model used. Smaller watersheds are less likely to have large influences from TN transport as the storage effects are lesser thereby explaining the better variability in observed TN concentration. Loading regressions does not seem to be a function of watershed drainage area (Figure 2.10) as most watersheds perform well ($NSE > 0.6$).

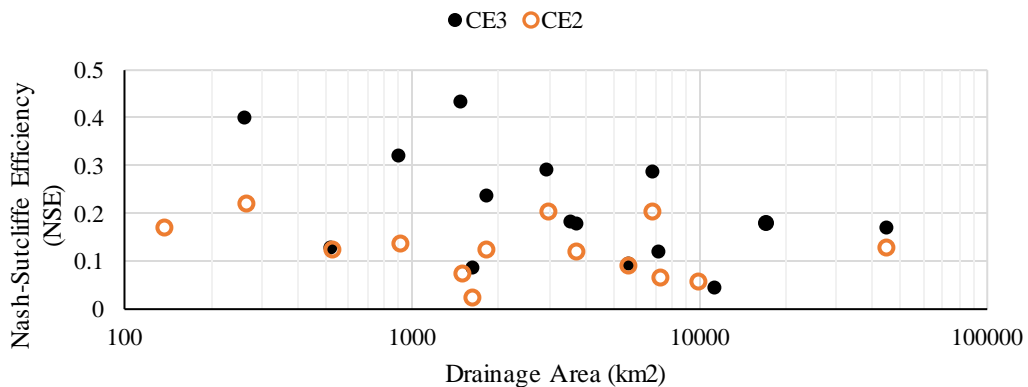


Figure 2.9. Role of drainage area in explaining the spatial variability of the skill of concentration regressions from WRTDS (CE3) and LOADEST (CE2).

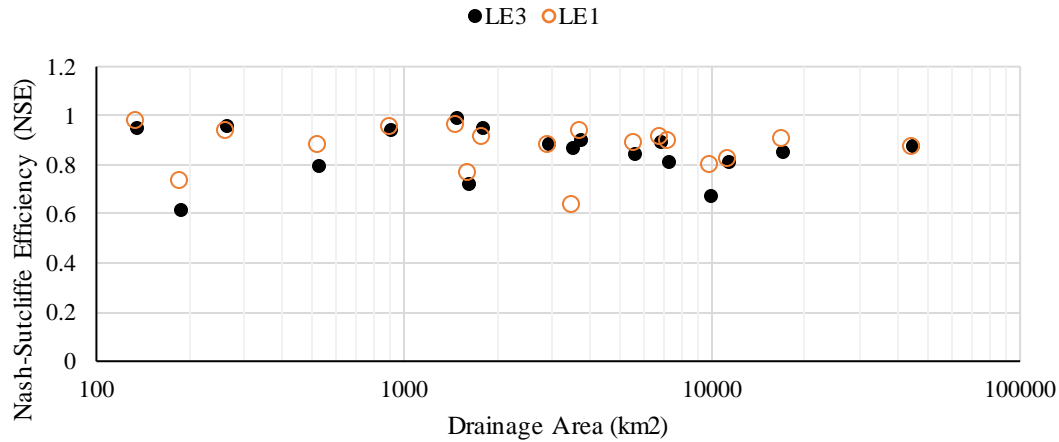


Figure 2.10. Role of drainage area in explaining the spatial variability of the skill of load regressions from WRTDS (LE3) and LOADEST (LE1).

2.5. Discussion and Conclusion

This study investigates the challenges in estimating TN concentration and loadings given the infrequent and sparse observation data sets over the SEUS. To improve water quality models in the future there is a need for more frequent and regular measurements of in-stream concentrations. More specifically, the poor performance of concentration estimation can be masked by the better performance in load estimation performance, which is shown by multiplying concentration estimates by observed streamflow. Depending on the constituent being modeled, concentration versus load, the performance in capturing the observed variability can vary in orders of magnitude given the same observed data set. This suggests an underlying difference in the performance distributions between concentration and load. For instance, if one assumes streamflow to be log-normal, then load also follows log-normal, but concentration being the ratio of two log-normal variates will follow a Cauchy distribution, whose moment-generating function is undefined. This poses difficulty in developing regression models for concentration.

The WRTDS model performs better in estimating concentration compared to LOADEST. However, the LOADEST models performs better in estimating load when compared to WRTDS.

Using this latter observation, one can get a better performing estimate for concentration than WRTDS by using load estimates from LOADEST and observed streamflow. Since this method uses observed streamflow there is no error contributed by the denominator in the ratio. As discussed earlier, when transforming concentration estimates to load estimates there is a dilution effect where the poor performance in concentration is hidden by better performance in load estimates. Since the WRTDS model uses a localized weighted regression that assigns higher weights to nearby observations, stations having frequent observations can be expected to have better concentration estimates from WRTDS. Figure 2.11 shows the difference in NSE and Pearson's correlation performance between WRTDS and LOADEST for each station as a function of the average number of days in-between each sampling and the total number of observations. Stations that have a higher total number of observations tend to have a lower average number of days in-between samplings, giving the trend a negative linear trend in the figure (2.11). As stations follow the negative trend we expect an increase in performance from WRTDS, indicating a warmer color on the color map. Although the figure suggests that concentration estimates from WRTDS does better in capturing the observed variability when the frequency of sampling increases by some points becoming warmer, there are a few outliers in the figure that restrict these claims. There were no observable differences in NSE and Pearson's correlation performance of load estimates as a function of sampling frequency, thus are not shown.

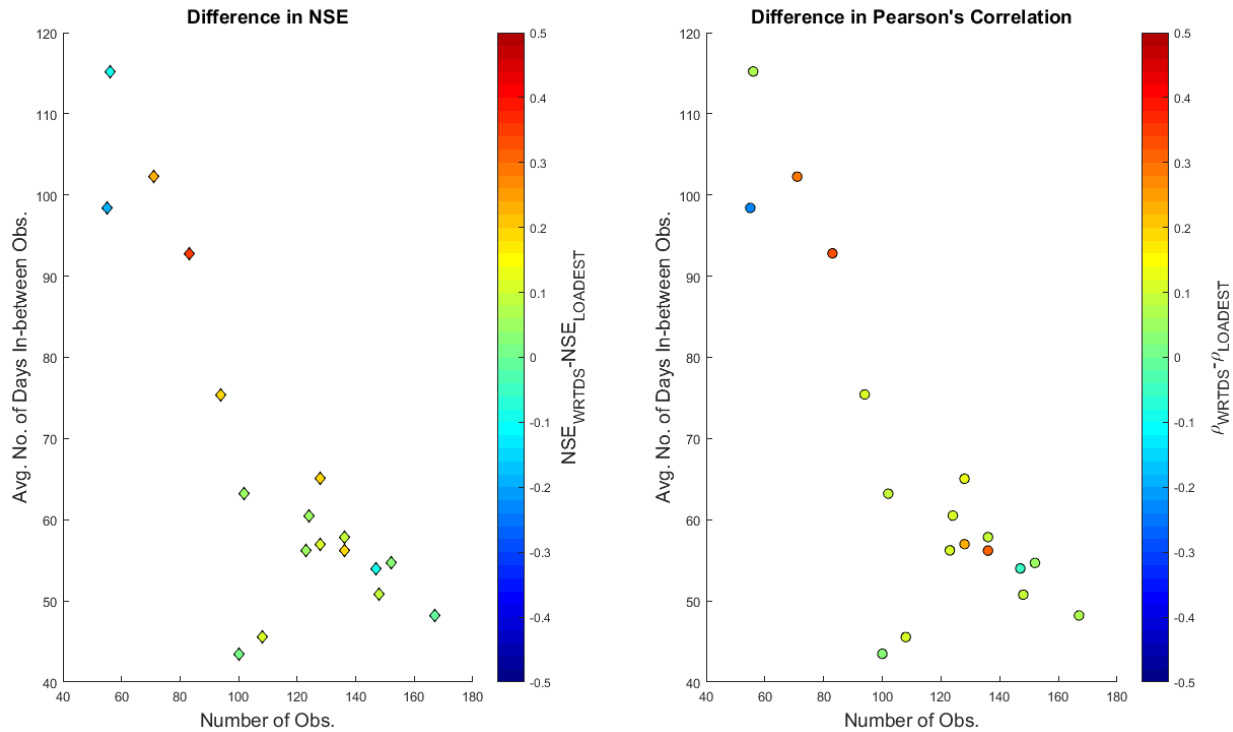


Figure 2.11. Role of sampling frequency in explaining the difference of skill of NSE (left) and Pearson's correlation (right) in concentration estimation from WRTDS (CE3) and LOADEST (CE2).

In conclusion, this study has shown that readily and common estimation tools like LOADEST and WRTDS which are popular estimation tools do not perform as well in estimating concentration as load. Looking at the standard deviation in the observed data, we see that over the observed period there is only variation between about 0.5 mg/l to 1.5 mg/l in most watersheds. This is a very low standard deviation, compared to say the standard deviation of streamflow which occurs on a magnitude of 1×10^3 cfs. High sampling variance is to be expected when having a lower number of observations, thus making it difficult for concentration estimates to have high NSE values. Thus, this requires proper statistical testing on whether the estimated skill of NSE is statistically significant. When looking at observed loadings the trend closely follows the trend of observed streamflow further confirming that concentration performance can be masked. These sites with limited data, streamflow contributes the majority of variance seen

in loadings making it easier for models to capture the observed variability. Accurate estimates of regional concentration and load can be very useful in determining nutrient criteria for states across the southeast. States like Georgia have plans to adopt nitrogen and phosphorus criteria for river and streams by the year 2020. Using the best models available to inform numeric criteria for loadings and concentration can ultimately improve nutrient reduction management strategies and improve stream water quality.

Chapter 3. A Non-Parametric Bootstrapping Framework Embedded in a Toolkit for Assessing Water Quality Model Performance

Abstract

Assessing the ability to predict nutrient concentration in streams is important for determining compliance with the Numeric Nutrient Water Quality Criteria for Nitrogen in the U.S.A. Evaluation of the LOADEST and the WRTDS models in predicting total nitrogen (TN) loads for 18 stations in the southeast show good performance (Nash-Sutcliffe Efficiency (NSE) >0.8) in capturing the observed variability even with limited data. However, both models captured only 40% of the observed variance in TN concentration (NSE < 0.4). Thus, the models performed differently in predicting two attributes –load and concentration – questioning the skill of the models. This study proposes a non-parametric approach for assessing the performance of water quality models particularly in predicting concentration. Null distributions of common performance metrics with no skill are constructed through bootstrap resampling and used to find p-values for metrics from the models for determining if the sample metrics belong to the null distributions.

3.1. Introduction

3.1.1. Software

The Toolkit for Assessing Performance in Concentration Regression Models (TAP-CRM) is available for free online at the following url: https://github.com/chwr-ncsu/water_quality_toolkit. The program is designed to run as a windows executable (.exe) and communicates with the LOADEST and WRTDS program. The MATLAB and R codes are available for individual download. Details for the software requirements and developer contact information can be found in the “README.md” file available on the site.

3.1.2. Background

Water quality predictions, particularly non-point pollutants such as total nitrogen (TN), typically focus on estimating load and concentration across various spatio-temporal scales. The primary challenge in assessing water quality models’ performance over a large spatial extent is due to the limited availability of observed values (Oh and Sankarasubramanian, 2012). Data networks with observed water quality measurements available at regional scale, such as USGS’s Water Quality Network (WQN) (Alexander et al., 1998), have observations over many years but are infrequent with some months having only one observed value. As a result, measurements such as TN concentration are available for less than 200 days spanning a 20-year period. While mechanistic models, like the SWAT model (Arnold et al., 2012), have potential to provide continuous values, they are difficult to calibrate and validate using the sparse and non-consecutive observations of water quality constituents; a problem that doesn’t exist for streamflow predictions (Smith et al., 1997). Apart from limited and infrequent water quality samples, nutrient loading predictions from regression and mechanistic models are able to capture the observed variability well (Nash-Sutcliffe Efficiency (NSE) >0.6-0.7) (Smith et al.,

1997)(Douglas-Mankin et al., 2010; Kim and Kaluarachchi, 2014). Thus, model predictions using streamflow as a predictor tend to easily capture the observed variability in loadings. Loadings is typically the target constituent for regulation management as seen in Total Maximum Daily Load (TMDL) programs. Equally important are concentration estimates which are required for assessing compliance with instream concentration limits. Water quality models like the Weighted Regression on Time, Discharge, and Season (WRTDS) (Hirsch et al., 2010) and the Load Estimator (LOADEST) (Cohn et al., 1992, 1989; Cohn, 2005) which can be adapted for concentration estimation use observed discharge and time as predictors for predicting concentration. However, this study shows that these empirical models have difficulty attaining higher accuracies ($NSE > 0.4$) in predicting observed concentration which could be potentially interpreted as insignificant particularly for stations with limited data. Even though observed loads are simply the product of observed concentration and streamflow, lower values of NSE in predicting concentration suggests that different accuracy thresholds apply for determining the statistical significance of concentration estimates. Thus, there is a need for an alternate way of evaluating the skill in predicting concentration by water quality models. One such way is proposed here.

The intent of this study is to develop a robust nonparametric toolkit that does not depend on distributional assumptions for determining if a sample performance metric is likely drawn from a population of metrics having no skill. The framework used in the toolkit can be applicable to a wide range of performance metrics and model types, but in this study the framework is applied to two types of water quality prediction models. The toolkit provides null distributions, p-values, and skill scores for assessing model performances. A tool for studying the uncertainty in water quality trends from the WRTDS model, named the WRTDS Bootstrap Test

(WBT), was developed by Hirsch et. al 2015. WBT was built to address the uncertainty in water quality trends, for example it could provide the 95% confidence interval in the rise of annual mean flow normalized concentration. Similar to this, the TAP-CRM toolkit focuses on determining if concentration predictions from the LOADEST and WRTDS models are statistically different from models with no skill. Although the differences in model performance between the LOADEST and WRTDS models are presented, the purpose of this study is not on comparing the models, which has already been discussed comprehensively (Hirsch, 2014). Popular statistical measures, Pearson's correlation (ρ) and the NSE, are considered for assessing water quality model performance as part of the toolkit. Instead of just reporting the model performance on NSE or ρ , the toolkit also provides the p-value in predicting concentration. However, to estimate p-values for a given sample statistic, distributional assumption on the water quality data is required (Gotway et al., 1994). For instance, if one assumes the data follows normal distribution, then NSE is assumed to follow a F-distribution as a ratio of chi-square distribution (Larsen and Marx, 1986). Given the limited, discontinuous water quality data, such distributional assumption may not be desirable. Hence, we propose a non-parametric re-sampling method that eliminates distributional assumption and is robust for evaluating model performance using smaller sample sizes. Additionally, a new skill score is introduced that provides a numerical measurement for comparing the accuracy across models and their performances as opposed to using just the p-value which represents a binary response (reject null, fail to reject null). This article describes the algorithm for determining the nonparametric distribution of performance statistics, NSE and ρ , and how they are useful in finding p-values and respective skill scores, using the toolkit software. A case study comparing concentration

regression models for 18 locations across the southeast using metrics from the toolkit is also presented.

3.2. Data Sources

3.2.1. WQN Measurements

This study considers 18 stations from Region 3 of the Southern Eastern United States, sites that have also been studied in forecasting seasonal nutrients using climate information (Oh and Sankarasubramanian, 2012). Frequency of sampling across sites started at the monthly timescale in 1962 and reduced to bimonthly and seasonal sampling from 1982 to 1995. The number of daily observations for total nitrogen records for the 18 stations averages about 200 days extended over about 20 years. These stations belong to both the National Stream Water-Quality Monitoring Network (WQN) and the Hydro-Climatic Data Network (HCDN) (Alexander et al., 1998). Stations belonging to the HCDN have streamflow that is minimally impacted by anthropogenic influences like artificial storage or pumping preserving the climate signal in streamflow (Slack et al., 1993; Vogel and Sankarasubramanian, 2005). The WQN is a combination of two subsets of networks, the National Stream Quality Accounting Network (NASQAN) with observations from 1962 to 1995 and the Hydrologic Benchmark Network (HBN) with observations from 1973 to 1995. The nonparametric framework proposed in this study is only applied to models predicting total nitrogen concentration but can be applicable to other water quality constituents having limited data such as total phosphorus or dissolved oxygen.

3.2.2. LOADEST Estimates

Observed streamflow and total nitrogen concentrations from the WQN records were used to calibrate the USGS's constituent load estimator, LOADEST (Cohn, 2005; Cohn et al., 1992,

1989). Model performance is evaluated using concentration or load estimations from each day in the observed period. Model regression form number 4 was used for load estimation for all 18 stations using maximum likelihood estimation (MLE) to provide context to the inherent difference between concentration and load estimation performance. The LOADEST model was modified for concentration estimation by manually defining the model form in the header.inp file, using the technique outlined in the manual. The internal model selection feature, using the Akaike Information Criterion (AIC), is disabled when altering the model form. Predetermined model forms (1-10) for estimating concentration using MLE were tested for goodness-of-fit using NSE as the criteria. The final model form used in this study is shown in (3.1), which is very similar to the preset model form number 4 for load regression,

$$\ln(c_i) = \widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \sin(2\pi dtime) + \widehat{\alpha}_4 \cos(2\pi dtime) + \widehat{\varepsilon}_i \quad (3.1)$$

where c_i is the observed daily total nitrogen concentration on day i ; Q_i is the observed daily streamflow on day i ; $dtime$ is the centered decimal time; $\widehat{\alpha}_1$ - $\widehat{\alpha}_4$ are model estimated coefficients; and $\widehat{\varepsilon}_i$ denotes the estimated model residual on day i .

3.2.3. WRTDS Estimates

The WRTDS model is an estimation method for water quality constituents included in the Exploration and Graphics for RivEr Trends (EGRET) package used for studying long-term changes in water quality and streamflow. Concentration measurements from the WQN data set are manually uploaded into the WRTDS model. Streamflow data from the USGS streamgage location is loaded from the data retrieval process internal to the model package. The WRTDS package is already set up for concentration estimation using the model form in (3.2).

$$\ln(c_i) = \widehat{\beta}_1 + \widehat{\beta}_2 \ln(q_i) + \widehat{\beta}_3 \ln(T) + \widehat{\beta}_4 \sin(2\pi T) + \widehat{\beta}_5 \cos(2\pi T) + \widehat{\varepsilon}_i \quad (3.2)$$

where c_i is the observed daily total nitrogen concentration in mg/l on day i ; q_i is the mean daily streamflow for day i ; T is the decimal time; $\widehat{\beta}_1$ - $\widehat{\beta}_4$ are model estimated coefficients; and $\widehat{\varepsilon}_i$ denotes the estimated model residual on day i . The regression form shown in (3.2) is fit for each observation using a weighted Tobit regression in a leave-one out cross validation approach. The weight for each observation is determined by the product of individual weights for discharge, time, and season found using a tricube function (Hirsch et al., 2010). Model estimates for each day in the observed period have corresponding model estimated coefficients and standard errors. Estimates for concentration on each day are transformed back to the original space using a bias-correction factor (BCF) shown in (3.3) (Hirsch et al., 2010).

$$BCF = \exp\left(\frac{SE^2}{2}\right) \quad (3.3)$$

3.3. Methods

3.3.1. Motivation

Performance metrics like R-squared and NSE are commonly used for assessing water quantity prediction models (Ahl et al., 2008; Alansi et al., 2009; Kim and Kaluarachchi, 2014; Wang et al., 2014). Although, specific criteria does not exist for determining “poor” and “excellent” models using these performance metrics when values of NSE become negative or close to zero, they are considered unusable predictions, instead one can simply use the mean value of the observed values (Santhi et al., 2001). Hydrologic models have shown that even calibrated models can result in different values of the NSE under different time scales, hence a context should be provided when describing the model’s performance. Application of WRTDS and LOADEST models have shown that popular concentration estimation models consistently have model predictions with NSE values less than 0.4 (Figure 3.1). To provide reference to the

WRTDS and LOADEST models' performance, we propose a non-parametric performance assessment toolkit for assessing if the NSE and ρ estimated by the water quality models are significantly different than metrics having no skill (i.e. NSE=0, $\rho=0$) (Jain and Sudheer, 2008; Schaeffli and Gupta, 2007). Non-parametric hypothesis tests based on re-sampling allows for the estimation of the underlying null distribution that may not follow common forms, such as bi-modal distributions. The null distributions of certain performance measures such as the Pearson's correlation (ρ) and NSE, depend on sample size and the underlying distribution of streamflow. Standard normal approximations have been used (McCuen et al., 2006) to calculate the p-values of the NSE statistic using a transformation very similar to the Fisher Z transformation (Fisher, 1992) that is used for approximating the distribution of Pearson's correlation. These transformations are limited to positive NSE values, given that NSE can be negative, this approach is not useful in calculating p-values for all underlying population NSE values. Under these conditions, it is advantageous to consider non-parametric hypothesis tests for region wide studies where distributions may change between watersheds and even time periods.

3.3.2. Non-parametric Bootstrapping

This toolkit estimates the null distribution for three performance measures to evaluate whether a metric (e.g., NSE) from either LOADEST or WRTDS is significantly different than models having no skill in predicting the observed variability. Null distributions for the metrics used in the toolkit will be centered close to zero although, in most cases there will be some positive bias in each metric coming from the model's random ability to capture the observed variance. The p-value provides the criteria for testing whether a performance statistic (e.g., NSE) belongs to a null distribution, at a certain alpha level, having "no skill" in predicting the

observed variability (Gronewold et al., 2009). Specifically, the null distribution is comprised of performance metrics from estimates that have no skill in estimating the variability of predictands (i.e., concentration) using the given predictors (streamflow). For instance, if we are testing whether the NSE from the LOADEST regression model ($NSE_{LOADEST}$) is statistically significant from a null model whose predictors and predictands do not co-vary, thereby the NSE corresponding to no skill (equation (3.4)). The corresponding alternative hypothesis is shown in equation (3.5). This study refrains, from defining the null distribution as being equal to zero (e.g. $H_0: NSE_{LOADEST}=0$) since the nonparametric resampling creates null distributions that are not exactly centered on zero.

$$H_0 : NSE_{LOADEST/WRTDS} = NSE_{no\ skill} \quad (3.4)$$

$$H_A : NSE_{LOADEST/WRTDS} \neq NSE_{no\ skill} \quad (3.5)$$

The null distribution of performance measures is created based on “N” realizations of LOADEST or WRTDS model runs which are fitted uncorrelated predictand and predictor sets. Each realization having a sample length of ‘n’ is created by randomly sampling with replacement from the observed concentration dataset with each observation being equally likely. Realizations are created by randomly sampling from the concentration data since the regression models both take concentration as the calibration data set to fit their respective models. This resampling with replacement scheme randomizes the predictand set while keeping the day and observed discharge fixed within the dataset. This process removes the time dependency between the predictand (concentration) and predictors (day and streamflow), thereby creating realizations whose mean correlation between the predictand and predictor sets approaches zero. The detailed outline for the hypothesis testing framework is shown in Figure 3.1. Using the uncorrelated predictand and predictors of length ‘n’, both WRTDS and LOADEST models are fitted first and then the

performance of the fitted model were evaluated using NSE, ρ , and Index of Agreement (IOA) by comparing with the observed predictand. This process is repeated for N realizations ($N_{LOADEST}=100$ and $N_{WRTDS}=30$) to develop the null distribution of NSE, ρ , and IOA. Since the predictors and predictands are uncorrelated, the performance measures from the null model represent the water quality models' random ability to capture the observed variability, thereby providing a basis for checking the statistical significance of the water quality model's performance ($NSE_{LOADEST/WRTDS}$).

3.3.3. Performance Metrics

The three performance measures considered in this toolkit are the Pearson's correlation (ρ), Nash-Sutcliffe efficiency (NSE), and Index of Agreement (IOA). Pearson's correlation measures the linear correlation between the model estimates (P_i) and the observations (O_i), shown in (3.6). Correlation values range from -1 to 1, with a value of 1 signifying perfect correlation and a value of 0 indicating no correlation between model and observed values.

$$\rho = \frac{\sum_{i=1}^L (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^L (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^L (P_i - \bar{P})^2}} \quad (3.6)$$

Nash-Sutcliffe Efficiency measures the squared deviations of the model values (P_i) to the observed values (O_i) with respect to the squared deviations of observations to the mean of observations (\bar{O}) as shown in (3.7). NSE is basically defined as one minus the sum of squared residuals normalized by the observed variance (Nash and Sutcliffe, 1970). NSE values range from negative infinity to 1, with negative values indicating that the mean value of the observations is a better predictor than the model (Nash and Sutcliffe, 1970). Models that just

predict the mean of the observations and do not capture any of the observed variability will result in an NSE value of zero, in this case we will call this a model as one with no skill.

$$NSE = 1 - \frac{\sum_{i=1}^L (O_i - P_i)^2}{\sum_{i=1}^L (O_i - \bar{O})^2} \quad (3.7)$$

The reason for presenting both Pearson's correlation and NSE is to examine the amount of bias (deviation between the model and observation); since the difference between the two is just the square of the bias. As the bias reduces, square of the Pearson's correlation approaches NSE (Krause et al., 2005). In hydrologic predictions the NSE can be very sensitive to large deviations in observed and predicted values for high flow which can cause negative values (Legates and McCabe, 1999). The index of agreement (IOA) was proposed to overcome the sensitivities of the NSE statistic by using the potential error in the denominator (Willmott, 1984), seen in (3.8). The range of IOA is from 0 to 1, with a value of 0 indicating no correlation.

$$IOA = 1 - \frac{\sum_{i=1}^L (O_i - P_i)^2}{\sum_{i=1}^L (|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \quad (3.8)$$

When using the NSE statistic for comparison certain conditions will cause the null distribution to center around negative one rather than zero. If the expectation of the model estimates, \hat{Y}_t , from LOADEST or WRTDS approach the observed mean with a variance close to zero, then the NSE value will approach zero as shown in (3.9). However, if the variance of the model estimates approach that of the observations the null distribution will be centered on negative 1, while still showing a Pearson's correlation distribution centered on zero shown in (3.10).

$$\left. \begin{array}{l} E(\hat{Y}_i) = \mu_{obs} \\ Var(\hat{Y}_i) = 0 \end{array} \right\} \longrightarrow NSE = 0 \quad (3.9)$$

$$\left. \begin{array}{l} E(\hat{Y}_i) = \mu_{obs} \\ Var(\hat{Y}_i) = \sigma_{obs}^2 \end{array} \right\} \longrightarrow NSE = -1 \quad (3.10)$$

Essentially, the numerator in (3.7) will become twice the denominator and the value will become negative one. The variance of the WQN data sets for each station are much less than 1 (~0.09) and the variance of model estimates are an order of magnitude smaller essentially being close to zero (~0.009). Using the bootstrapping framework to produce concentration estimates results in a scenario shown in (3.9) where the NSE is centered near zero. Load estimates from LOADEST or WRTDS not only reproduce the observed mean but also the observed variance while still having no correlation. Under this scenario, the NSE will be centered on negative 1 shown in (3.10). The index of agreement has been included as another performance metric since it does not have this issue and is provided for a more robust analysis.

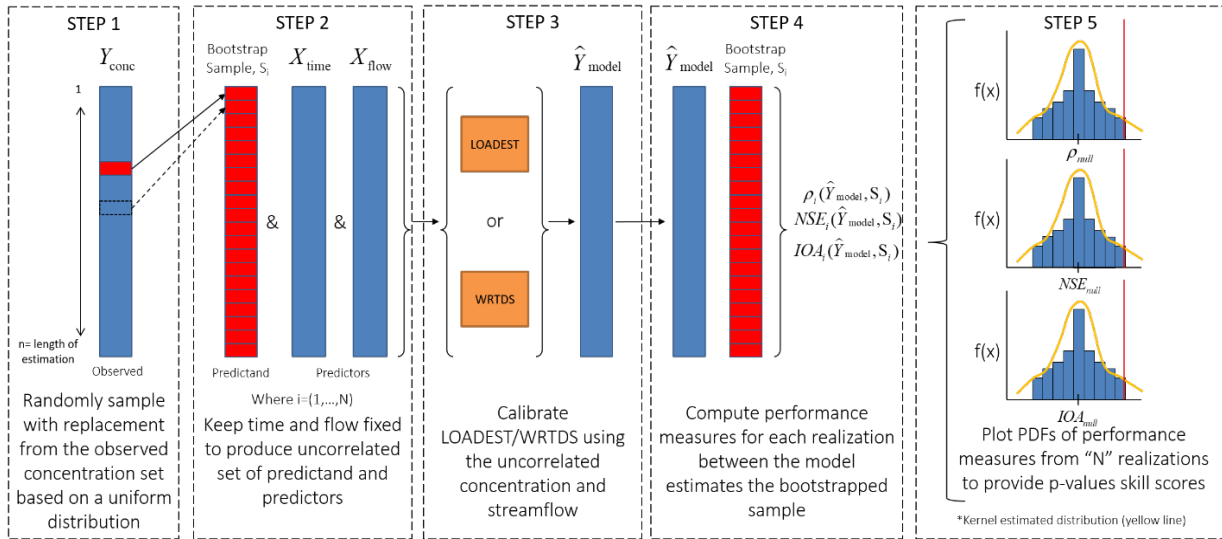


Figure 3.1. Re-sampling framework for developing the specified null distribution of the given performance statistic.

3.3.4. P-values and Skill Score

Using a non-parametric bootstrap, outlined in Figure 3.1, allows for the realizations to have null distributions centered towards zero for Pearson's correlation, NSE, and IOA. Since this is a non-parametric re-sampling technique, the p-value for the performance statistics are computed empirically instead of assuming an underlying distribution for model estimates and observations. Although, ρ and NSE can be positive or negative, the p-value is calculated using a one-tailed test. Recall that negative values of ρ and NSE indicate models estimates that are worse than just the observed mean so the user should only be interested when the performance statistic is positive. Using a significance level (α) of 5%, any test statistic having a p-value less than 0.05 can be considered not belonging to the null distribution. A critical value for the performance statistic can be calculated based on α , which specifies the threshold value on the right tail of the null distribution. Statistic values that are to the left of the threshold value are considered to part of the null distribution for that given α thereby insignificant. Furthermore, this study introduces a skill score, shown in (3.11), which can be useful for comparison between two different models that have different null distributions for the same performance statistics. PS denotes the performance statistic and $PS_{critical}$ denotes the right tailed critical value of the performance statistic for the chosen α . The skill score ranges from 0% to 100%; with a score of 0% indicating that performance statistic is centered on the critical value and a score of 100% indicating a perfect performance metric.

$$SS_{PS} = \frac{PS_{model} - PS_{critical}}{1 - PS_{critical}} \times 100\% \quad (3.11)$$

Since a value of 1 indicates a perfect performance metric for all three metrics discussed in this paper, the form of equation (3.11) does not change. Negative values for SS_{PS} indicate that

the performance metric belongs to the null distribution and will have a corresponding p-value greater than α . Since p-values are primarily used for hypothesis testing and not for indicating the strength of performance metrics, the SS_{PS} was developed to provide a standard scale for comparing performance metrics between different models.

3.3.5. Autocorrelation

Using data sets that have significant autocorrelation should preserve the autocorrelation in the resampling technique. Concentration measurements from the WQN are from the monthly to seasonal scale over multiple decades and any significant lag correlation can be interpreted as purely coincidentally. In some cases, where stations have records from the same month they are never on consecutive days or do not occur over the entire period enough to create significant lag correlation. If using a data set with consecutive daily measurements or with reasonable lag correlation, we recommend moving-block bootstrap sampling. In the described framework, the block size for resampling is set to 1 meaning that the only one observation is resampled at a time from the observations. The framework code is provided and allows for the manipulation of the block size for data sets with significant lag correlation. A test for autocorrelation is recommended for each data set before using the toolkit. The moving-block bootstrap option used in this study is very similar to the framework used to preserve lag correlation in studies for addressing uncertainty in trend analyses for concentration and load (Hirsch et al., 2015; Vogel and Shallcross, 1996). For example, say daily concentration values for Tar River at Tarboro, NC had significant lag-3 correlation then the resampling technique would resample 3 neighbors of observed concentration in a single sample for creating a realization. So, to construct observed concentration of length 'n' by resampling, we then resample only n/3 times. Further details of the moving-block bootstrap can be found with the source code but is not discussed further as

most observed water quality data sets are discontinuous with daily observations being far apart (e.g., once in two months).

3.4. Toolkit Interface

A performance toolkit for the purpose of testing the significance and comparing models was developed for open source distribution to academics and commercial water quality modelers. Although the toolkit is constructed for assessing concentration regression models outlined in Section 3.2 the bootstrapping framework was applied for testing the significance of LOADEST load estimations to show that performance was significant. Using the non-parametric re-sampling technique outlined and then applying a kernel smoothing function to the histogram reveals a shape and features that may not be so easily discoverable otherwise. Figure 3.2 shows the general framework for uploading data and the options users can choose to evaluate model performance. Users first upload observed streamflow and concentration data using an excel or .csv file. The number of simulations and the alpha level can be specified, with recommended values of $N=100$ when using LOADEST and $N=30$ when using WRTDS and with an $\alpha=0.05$. The toolkit does not run the LOADEST and WRTDS models internally, rather it communicates with the FORTRAN executable and an R compiler respectively for each model program. The WRTDS model takes longer to run and requesting a large number of simulations (>30) can greatly increase computation time. The user can choose to define null distributions for three statistics: Pearson's Correlation (ρ), Nash-Sutcliffe Efficiency (NSE), and Index of Agreement (IOA). After the toolkit runs the re-sampling method outlined in Figure 3.1, the nonparametric distribution is plotted with visible kernel smoothing curves (shown as blue lines) and black density dots displayed along the bottom based on the histogram as shown in Figure 3.4. A summary of all performance statistics, p-values, and skill scores are provided in a result

after the distributions are plotted. The toolkit first runs LOADEST realizations, since these runs are quicker, then a model comparison option is offered where the user can choose to run the WRTDS model and compare using the p-values, skill scores, Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) all presented in a model comparison table as shown in Figure 3.2 (Akaike, 1981). The toolkit interface was built using the GUIDE feature in MATLAB, shown in Figure 3.3, and is offered as a windows executable that can be used stand still.

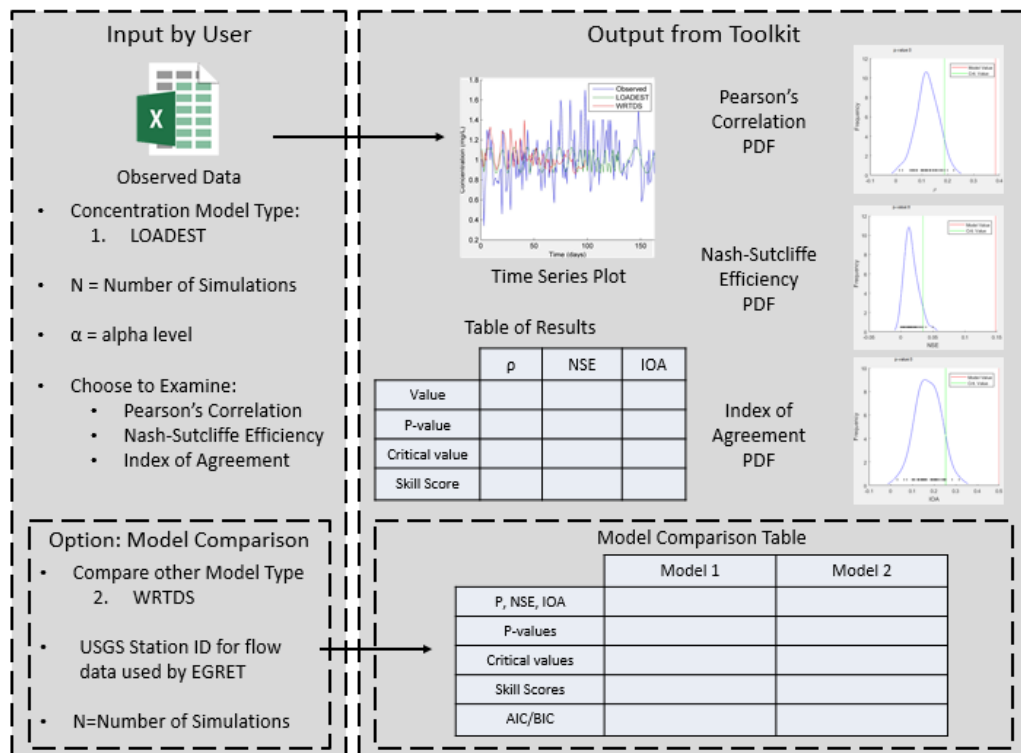


Figure 3.2. Nonparametric toolkit- input and output in detail.

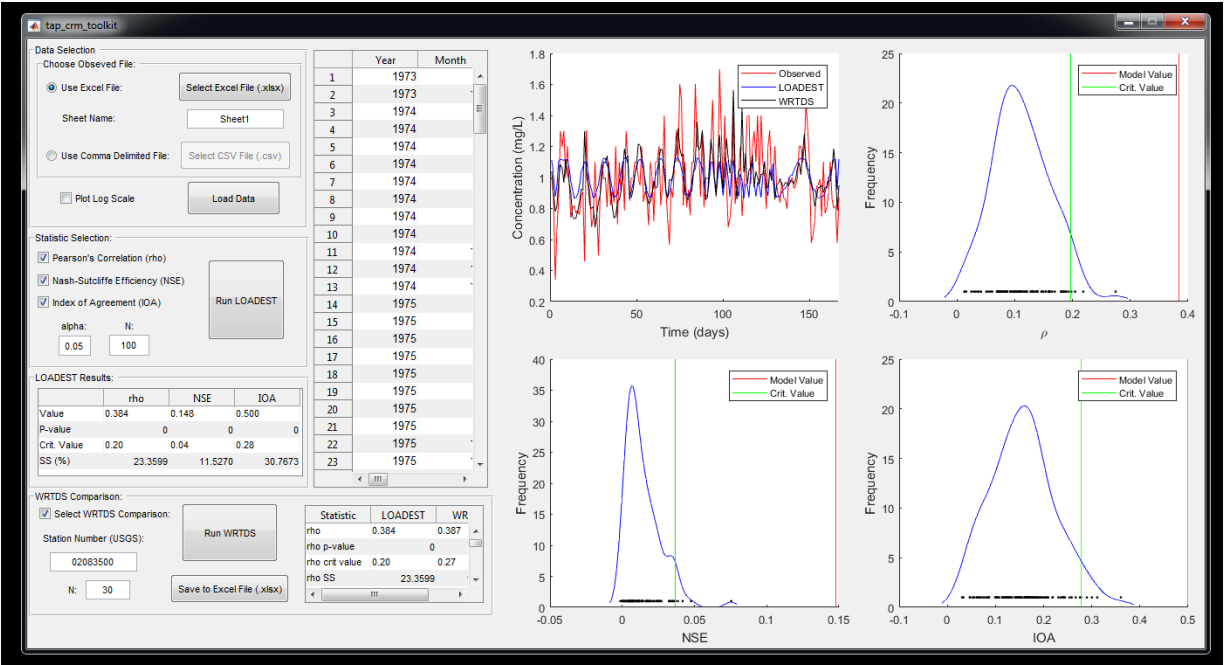


Figure 3.3. MATLAB interface for the nonparametric toolkit.

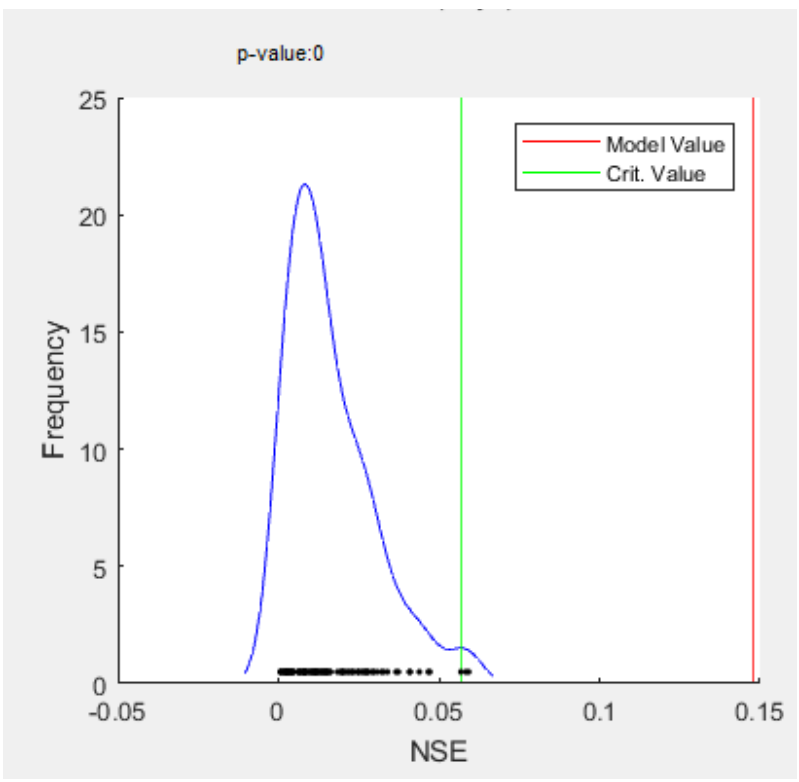


Figure 3.4. Bootstrapped null distribution with kernel smoothing for two-tailed tests (green line) using $\alpha=0.05$. Black dots show the density of the distribution. Red line denotes model performance measure.

3.5. Application and Results

LOADEST and WRTDS regression models were used for predicting TN concentrations over 18 stations belonging to the WQN over the Southeastern United States. For additional details on the 18 stations, see Figure 2.1 and Table 2.1. Application of the LOADEST model in estimating load (concentration) for 18 stations in Figure 3.5 (Figure 3.6) show the NSE varies from 0.79-0.98 (0.01-0.17) indicating performance of load estimation being better than that of concentration estimates in explaining the observed variability of the respective attributes. However, p-values obtained from the application of the nonparametric toolkit show that the NSE values are statistically different from zero for 12 of the 18 sites (i.e. reject the null hypothesis) (Figure 3.6). Thus, a rigorous hypothesis testing of the reported skill is necessary for reporting the water quality model performance. The developed nonparametric toolkit is useful in discerning whether the skill is statistically significant without requiring any distribution assumption to assess the model performance. Given that the re-sampling approach relies on no lag dependency in the model data, caution should be placed while using the modeled data with continuous daily values. Since observations are often discontinuous in the water quality data and the performance statistic estimation requires comparison with observed data, the lag-correlations are not going to be significant in this case. In case if the lag-correlations are significant, we suggest considering moving-block approach for re-sampling (Hirsch et al., 2015; Vogel and Shallcross, 1996). Application also demonstrated the usefulness of skill score (3.11) in comparing the model performance using different performance metrics (Figure 3.7) and in comparing the performance of two different models (Figure 3.8), LOADEST and WRTDS, estimated using the same metric (NSE). Lower values for the skill score indicate that they are approaching the threshold for failing to reject the null distribution and higher values mean the

performance statistics are approaching a value of 1. Skill scores (Figure 3.7) comparison between correlation and NSE obtained in predicting concentration shows tremendous variability across the region, but the performance of two statistics are consistent. Stations having negative skill scores all correspond to having p-values above 0.05 and thus fail to reject the null hypothesis. Specifically, station 7 (Satilla River at Atkinson, GA) has a negative skill score for ρ but a slightly positive skill score for NSE. Since the positive skill score for station 7 is so close to the threshold (<1%), this station is considered to have insignificant performance. The skill scores for stations 6 (Falling Creek near Juliette, GA) for both performance metrics have magnitudes larger than -10% indicating the individual metrics are more toward the center of the null distribution. Skill scores comparing the two concentration models (Figure 3.8), LOADEST and WRTDS, show that WRTDS model perform slightly better in estimating the observed concentration across all stations except station 17. Station 17 has the shortest record of observations (55) among the 18 stations and the WRTDS has difficulty making meaningful predictions using short records. The primary advantage of using the skill score is that the water quality models' performance can be compared and evaluated across performance statistics as well as across models.

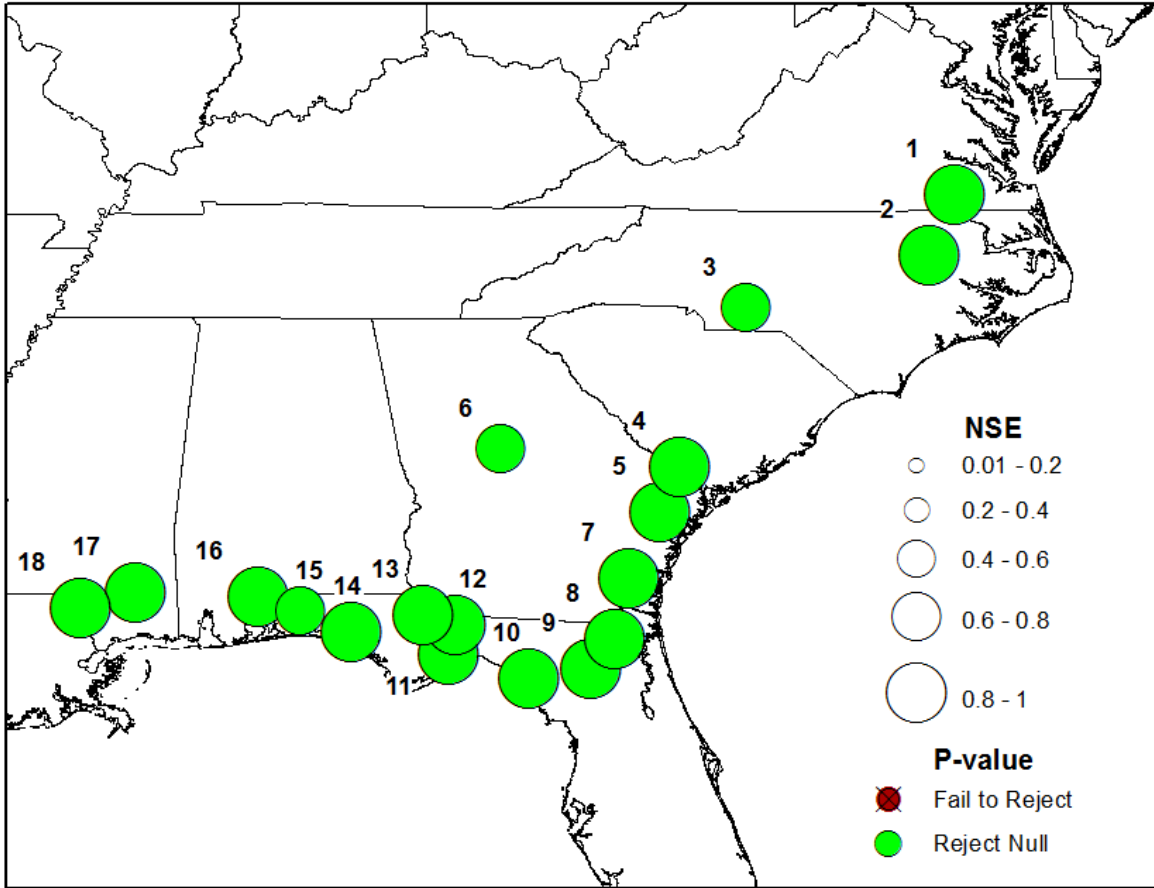


Figure 3.5. NSE of predicted load using the LOADEST model and the decision for the null hypothesis using p-values from the null distribution obtained from the nonparametric framework.

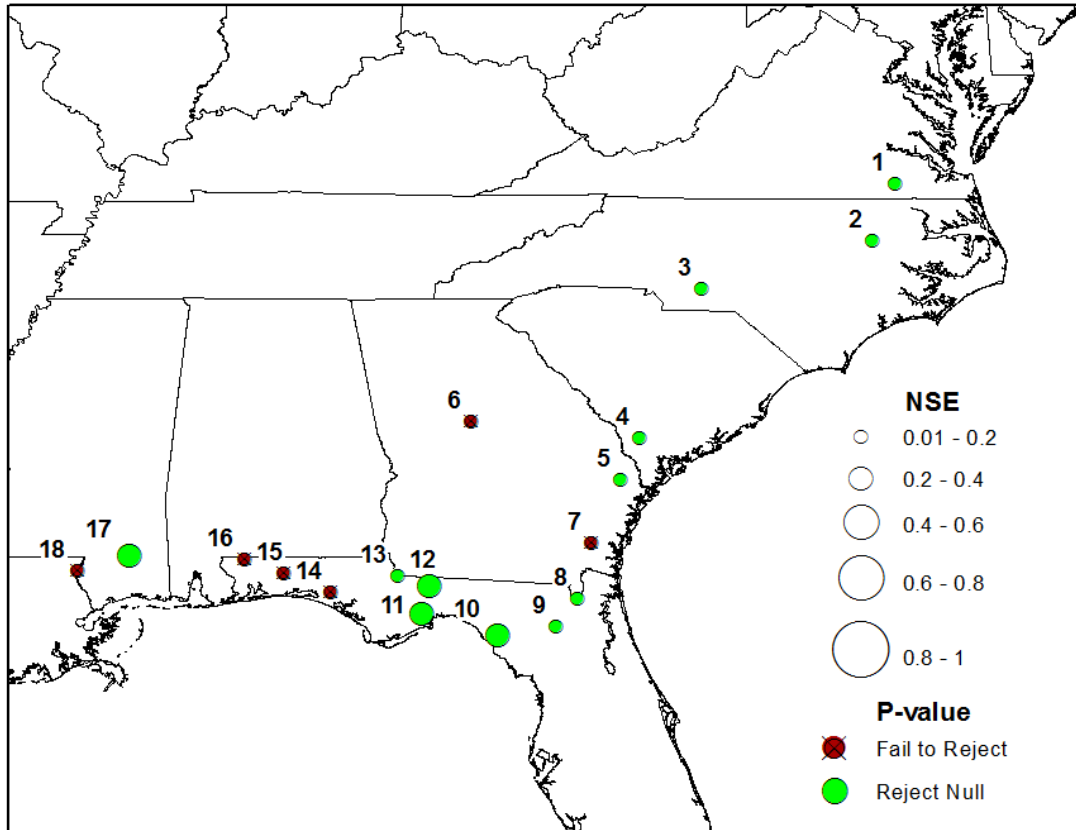


Figure 3.6. NSE of predicted concentrations using LOADEST model and the decision for the null hypothesis using p-values from the null distribution obtained from the nonparametric toolkit.

Skill Scores for ρ and NSE on concentration regressions from the LOADEST model

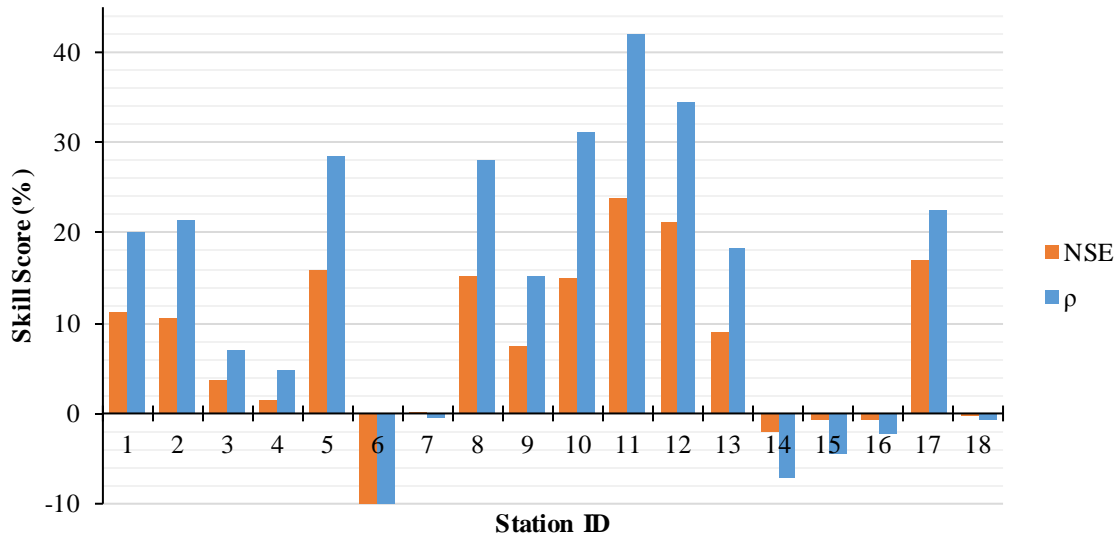


Figure 3.7. LOADEST performance as shown by skill scores for both Pearson’s and NSE metrics, demonstrating the usefulness of the skill score for application across different performance metrics.

Skill Scores for NSE from concentration regressions using the LOADEST and WRTDS model

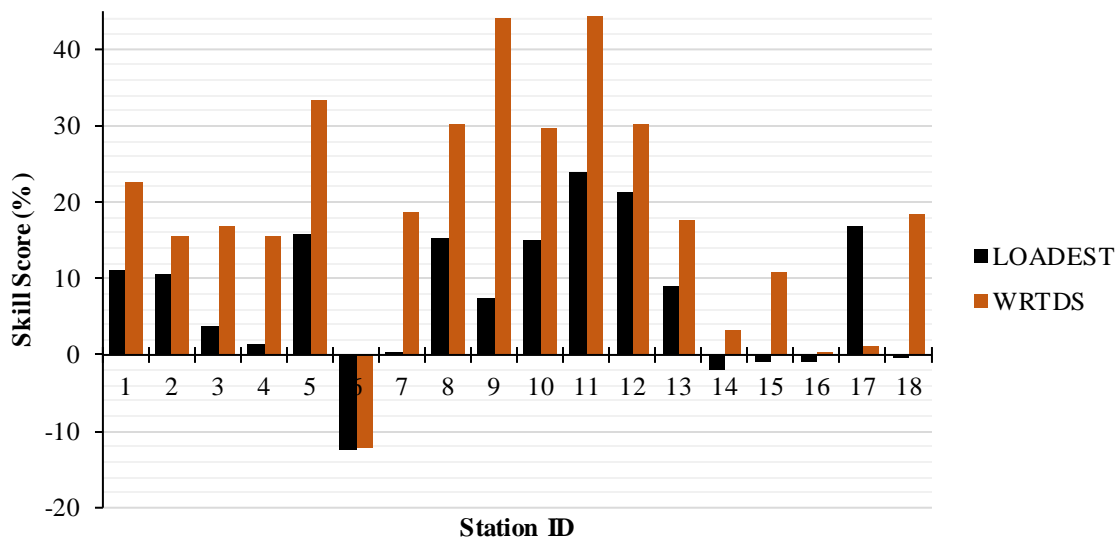


Figure 3.8. LOADEST and WRTDS performance, as shown by the skill score, demonstrating which stations have performance significantly different than zero while also showing that WRTDS has preferred concentration estimates for the majority of stations

3.6. Discussion and Conclusion

This study recognizes the inherent problem of data availability when dealing with water quality observations and stresses the need for longer continuous data sets. This study promotes the use of a non-parametric re-sampling approaches as opposed to using parametric tests for evaluating model performance as it is difficult to accurately quantify the underlying model with infrequent and limited observations. This version of the toolkit was developed for water quality concentration regression models and does not easily apply to load estimation models. Specifically, this pertains mostly to the calibration and estimation files for LOADEST load estimation which are formatted differently. A toolkit for the purposes of assessing loading regression models may not be as useful considering that load estimation models perform much higher than concentration estimation models, as proven by the case study. If the TAP-CRM toolkit is used for data sets that have significant autocorrelation options to use a moving-block bootstrapping approach are available in the MATLAB source code and the block size should be modified. Using the non-parametric re-sampling approach for hypothesis testing, the toolkit and the proposed skill scores provides a means of comparing a given model's performance with different performance statistics (as seen in Figure 3.7) as well as in comparing multiple models' performance under a given statistic (as seen in Figure 3.8). It is reassuring to see that given the low values of NSE, the model's skill is still statistically significant. The evidence provided by the toolkit helps establish expected results for modelers focusing on concentration, which is a critical variable for downstream habitats. As more waters become impaired due to anthropogenic influences, we can expect more observations to be available for modeling and expect the values of performance to increase, but until then water quality modelers must discover ways on how to deal with the currently available data to make the best-informed decisions.

Chapter 4. Multivariate Bias Corrections of Mechanistic Water Quality Model Predictions

Abstract

Water quality networks usually do not include observations on a continuous timescale over a long period. Statistical models that use streamflow and mechanistic models that use meteorological information and land-use are commonly employed to develop continuous streamflow and nutrient records. Given the availability of long meteorological records, mechanistic models have the potential to develop continuous water quality records, but such predictions suffer from systematic biases on both streamflow and water quality constituents. This study proposes a multivariate bias correction technique based on canonical correlation analysis (CCA) - a dimension reduction technique based on multivariate multiple regression - that reduces the bias in both streamflow and loadings simultaneously by preserving the cross-correlation. We compare the performance of CCA with linear regression (LR) in removing the systematic bias from the SWAT model forced with precipitation and temperature for three selected watersheds from the Southeastern US. First, we compare the performance of CCA with LR in removing the bias in SWAT model outputs in predicting the observed streamflow and total nitrogen (TN) loadings from the Water Quality Network (WQN) dataset. We also evaluate the potential of CCA in removing the bias in SWAT model predictions at daily and monthly time scales by considering the LOADEST model predicted loadings as the predictand for CCA and LR. Evaluation of CCA with the observed dataset and at daily and streamflow time scales shows that the proposed multivariate technique not only reduces the bias in the cross-correlation between streamflow and loadings, but also improves the joint probability of estimating observed streamflow and loadings. Potential implications of the proposed bias-correction technique, CCA, in water quality forecasting and management are also discussed.

4.1. Introduction

Water quality measurements available over a continuous period are usually limited to watersheds that have implemented monitoring programs for tracking water quality impairments under the US Clean Water Act of 1972 (PUBLIC LAW 92-500). Cost and labor requirements limit these daily continuous observations to shorter time periods, i.e. 1-2 years, from the start time of impairment (Quilbé et al., 2006; Rao et al., 2013). However, studies focusing on the regional and long-term variability of nutrient loadings to climate have been limited to smaller sample lengths and sparse sampling for interpreting and calibrating water quality models (Smith et al., 1997). Data sources having multi-decadal observations such as the U.S. Geological Survey National Stream Water-Quality Monitoring Network (WQN) which includes 679 stations across the United States over the period 1962 to 1995 have scattered and non-continuous water quality records (Alexander et al., 1998). Frequency of WQN sampling was determined by the type of water quality constituent. For instance, total nitrogen sampling typically ranged from 4 to 12 times per year. Budgetary constraints caused the sampling frequency to vary over the period with the majority of sites starting at monthly sampling and then dropping to bimonthly and eventually quarterly sampling starting in 1982 (Alexander et al., 1998). Load prediction models used for water quality management require long-term measurements on the monthly and seasonal scale for model development. Hence, efforts have focused on filling the data gaps in observations using both statistical and mechanistic models.

Water quality modeling and predictions have been carried out using statistical based models (Aulenbach, 2013; Moyer et al., 2012; Oh and Sankarasubramanian, 2012; Park and Engel, 2015) and studies have also used physically based mechanistic models for predicting water quality constituents (Amatya et al., 2013; Jha et al., 2010, 2007; Shrestha et al., 2008).

Statistical based models for estimating nutrient loadings, like the LOADEST (Cohn, 2005; Cohn et al., 1992, 1989) and WRTDS (Hirsch et al., 2010) model use streamflow, time, and a seasonality component to predict nutrient loadings and develop continuous records over longer time periods. The main challenge with this approach is that predictions are restricted to basins with gauged stations and to periods with observed streamflow, thus making them not suited for predicting water quality in ungauged basins. On the other hand, mechanistic models, such as the Soil & Water Assessment Tool (SWAT) model (Douglas-Mankin et al., 2010), use soil and land-use information along with long-term observed meteorological records to predict streamflow, nutrient loadings and concentrations. Further, impacts from changes in land-use change, anthropogenic forcings, and management practices can be investigated using the mechanistic model (Douglas-Mankin et al., 2010).

One of the mechanistic models, the SWAT model, has been used widely in studies for streamflow prediction and forecasting (Ahl et al., 2008; Alansi et al., 2009) and to understand the impacts of land use changes on streamflow and pollutant loadings (Marhaento et al., 2017; Serpa et al., 2017; Tong et al., 2009; Wang et al., 2014). Complex mechanistic models like SWAT are abstractions of actual physical processes that are difficult to represent; hence, such models can exhibit bias in predicting the observed variables, even though these models reasonably capture the variability of observed streamflow and loadings if calibrated well. Model bias can be introduced from imperfect representations of complex natural processes, as seen in GCM modeling of atmospheric physics, or from using improper parameter conditions (Maraun, 2016). The SWAT model has sub-representations of actual physical processes that are difficult to represent, specifically the nitrogen cycle which involves the formation and degradation of several nitrogen species. Given the complexity of such processes, model bias can be introduced

from model deficiencies in formula representation and/or failure to adequately calibrate parameters. In this study we will define bias as the systematic deviation between model and observed moments: mean, variance, and cross-correlation. Improving model calibration can greatly reduce model bias and is the most preferred form of bias correction but not the easiest (Ehret et al., 2012). Using time independent variables, such as sparse or non-continuous WQN records, for calibrating models makes it difficult to identify if sources of bias result from model deficiencies or incorrect calibration. Thus, the type of bias correction considered for this study is a post-processing technique to reduce the deviation between the model and observed moments: mean, variances, and cross-correlation.

Model predictions with systematic deviations from the observed streamflow and loadings should be corrected prior to application (Santhi et al., 2001). Under these situations, for model applications, univariate bias correction techniques using regressions are commonly employed for removing bias in streamflow (Stewart and Reagan-Cirincione, 1991) and loadings (Leisenring and Moradkhani, 2012; Windolf et al., 2011). Studies examining the long-term or future impacts of climate on streamflow and water quality also use forcings from general circulation models, whose outputs need to be bias-corrected (precipitation and temperature) before forcing them in the SWAT model. In this context, univariate and multi-variate bias correction procedures have been applied for bias correction in climate forcings (Das Bhowmik et al., 2017; Fang et al., 2015; Mazrooei et al., 2015; Wood et al., 2004). On the other hand, bias correction techniques for improving water quality predictions from mechanistic models like the SWAT model is very limited (Windolf et al., 2011).

The hydrological community using precipitation and temperatures from GCMs have largely used bias correction as a pre-process step to remove deviations prior to forcing the

variables in prediction models. Prediction models using observed precipitation and temperature may use bias correction as a post-process application to remove bias caused by model deficiencies (Maraun, 2016). A common post-processing technique for bias correction is to apply bias removing coefficients to model outputs in the form of a regression equation with a multiplicative coefficient representing the slope and an additive coefficient being an intercept based on regression (Stewart and Reagan-Cirincione, 1991). A thorough search for the cause of model bias should be done prior to using a post-processing bias correction technique.

Mechanistic models use internal variables that describe actual physical processes and identifying the parameters that introduce bias to predictions can be very helpful in understanding the modeling process. Incorrect use of bias correction types may result in the covering up of model deficiencies rather than actually reducing bias (Reichert and Mieleitner, 2009). In this study, applying post-process bias correction techniques to mechanistic model predictions is appropriate since identifying the cause of bias when using sparse non-continuous variables is difficult.

Most bias correction procedures are univariate with the regression relationship being developed separately between observed and model output for one variable at a time. Linear post-processing has shown to be useful in reducing bias in the mean and variance of model predictions, however reducing bias present in other moments may require further processes (Vannitsem, 2011). Das Bhowmik (2016) showed that univariate bias correction techniques such as quantile mapping and linear regression do not preserve the cross-correlation between the bias-corrected variables. To address this, Das Bhowmik et. al., (2017) suggested multivariate bias correction based on asynchronous canonical correlation analysis (ACCA) for bias correcting both monthly precipitation and monthly temperature from GCMs. The term asynchronous indicates the GCMs projections under climate change do not have time correspondence with the

observed climatic variables. Another multivariate downscaling technique that is quite popular is the multivariate adaptive constructed analogues (MACA) approach which provides spatially disaggregated time series of precipitation and temperature inside GCM simulations (Abatzoglou and Brown, 2012; Hidalgo et al., 2008).

In the context of water quality predictions, mechanistic models that predict both streamflow and nutrient loadings are strongly correlated and inherently exhibit bias with the observed attributes. One could consider univariate bias correction procedures such as simple regression or quantile mapping for removing the bias in the estimation of loadings. Since streamflow and loadings are strongly dependent on each other, it is important to perform bias correction that better preserves the cross-correlation between the two variables, thereby improving the joint likelihood of estimating observed streamflow and loadings. In other words, any efforts to bias correct them separately will result in underestimation of the cross-correlation between streamflow and loadings. This study provides a multivariate bias correction technique using canonical correlation analysis (CCA) that simultaneously reduces the bias in streamflow and nutrient loadings while reducing the bias in estimating the cross-correlation between the two variables. The key differences of the proposed approach from the ACCA (Das Bhowmik et al., 2017) is that here CCA is applied synchronously and as a post-processing step. Keeping time correspondence between the SWAT model predictions and the observed variables is paramount since our interest is in improving monthly streamflow and water quality predictions.

This paper compares two bias correction techniques – Univariate Regression and Multivariate CCA – on their ability to reduce the bias in estimating the observed mean, standard deviation of streamflow and loadings and also reduce the bias in the estimation of cross-correlation between streamflow and loadings. We also discuss how the reduced bias in the

estimation of moments translate to improving the joint probability of estimating observed streamflow and loadings. Model values of predicted streamflow and total nitrogen loadings come from the calibrated SWAT model for three watersheds chosen across the southeastern U.S. The univariate and multivariate bias correction techniques are first evaluated at the daily time scale by comparing with the observed moments of streamflow and loadings from the WQN dataset. Further, we also compare the two bias correction techniques at monthly time scale using the aggregated total nitrogen (TN) loadings estimated from the LOADEST model. The ability of both bias correction techniques in preserving the joint probability of observed streamflow and loadings is also presented along with potential application of the multivariate techniques in streamflow and water quality forecasting. The paper is organized as follows: first a description of the data sources and SWAT model construction is presented, which is followed by a discussion of two bias correction techniques. Next, we evaluate the performance of two bias correction techniques by applying it for the WQN data and at daily and monthly time scales. Finally, we summarize the findings along with potential applications of multivariate bias correction techniques for water quality forecasting.

4.2. Data Sources

This study considers three watersheds chosen from Region 3 of the Southern Eastern United States (SEUS) (Figure 4.1). Oh and Sankarasubramanian (2012) considered 18 watersheds from the SEUS for seasonal nutrient forecasting using climate information. These three watersheds were selected from the 18 watersheds based on the length of nutrient and streamflow observations. This study was limited to just three basins across the southeast, in varying basin size, given the calibration and computation time of the SWAT model. Table 4.1 shows a summary of the total number of daily observations for nitrogen loadings for all three

watersheds with each having less than 200 days extended over about 20 years. Further, these watersheds are part of the USGS Hydro-Climatic Data Network (HCDN) and the National Stream Water-Quality Monitoring Networks (WQN) (Alexander et al., 1998), whose streamflow and water quality observations are minimally impacted by anthropogenic influences. The HCDN is a subset of USGS streamgages that have been identified as watersheds where anthropogenic activities such as pumping and having artificial storage are minimal/absent so that climate signal can be studied in streamflow (Slack et al., 1993; Vogel and Sankarasubramanian, 2005). The WQN is a combination of two networks, the National Stream Quality Accounting Network (NASQAN) with observations scattered from 1962 to 1995 and the Hydrologic Benchmark Network (HBN) with observations scattered from 1973 to 1995. The framework proposed in this study can be applicable to all constituents but we limit our analyses of bias correction to only streamflow and total nitrogen loadings. Observed nitrogen loadings is calculated by multiplying the observed total nitrogen concentration from the WQN by the average streamflow on that day from USGS stream gages.

Table 4.1. Summary of station data for the selected watersheds showing the number of daily observations for total nitrogen records from the Water Quality Network (WQN) database. Numbers in parenthesis in the far column represent the number of years the daily observations span.

Station Index	USGS Station Number	Station Name	Drainage Area (sq. miles)	Number of Daily Obs. (# of years)
1	02083500	Tar River at Tarboro, NC	2183	148 (22)
2	02202500	Ogeechee River near Eden, GA	2650	167 (23)
3	02375500	Escambia River near Century, FL	3817	147 (22)

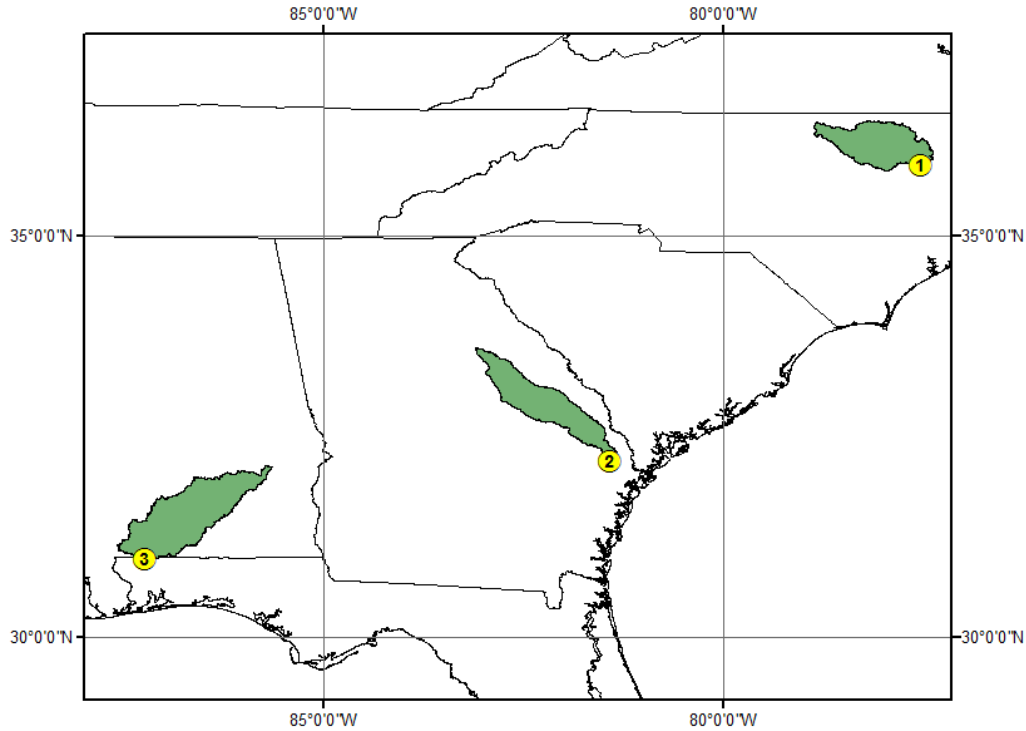


Figure 4.1. Selected three watersheds across the southeast region chosen for bias correction using the SWAT model. Shaded areas represent the drainage area and numbered points provide the stream gauge locations; (1) Tar River at Tarboro, NC (2) Ogeechee River near Eden, GA (3) Escambia River near Century, FL.

4.2.1. Performance Measures

The performance statistic used to quantify model performance is Nash-Sutcliffe Efficiency (NSE). Nash-Sutcliffe Efficiency measures the squared deviations of the model value, \hat{X}_i , to the observed value, X_i , with respect to the squared deviations of the observed values to the mean of the observations, \bar{X} , as shown in equation (4.1) (Krause et al., 2005); here the values could be for either streamflow or loadings. NSE describes how much of the observed variance is captured by the model values and ranges from negative infinity to one. A NSE value of 1 indicates that the model is perfect and capturing all of the observed variance while a NSE

value that tends towards less than and close to zero is considered a poor model (Santhi et al., 2001).

$$NSE = 1 - \frac{\sum_{i=1}^N (X_i - \widehat{X}_i)^2}{\sum_{i=1}^N (X_i - \overline{X})^2} \quad (4.1)$$

Percent bias (PBIAS) is a term used in this paper to quantify the bias between the observed ($Q_{o,i}$) and model values ($Q_{m,i}$); equation (4.2) shows an example for calculating PBIAS of modeled streamflow from the SWAT model.

$$PBIAS = \left(\frac{\sum_{i=1}^n (Q_{o,i} - Q_{m,i})}{\sum_{i=1}^n Q_{o,i}} \right) \times 100\% \quad (4.2)$$

These two metrics were chosen based on their wide use in the SWAT modeling community and availability of satisfactory performance ratings for both streamflow and nutrient load (D. N. Moriasi et al., 2007).

4.2.2. Total Nitrogen Loadings

Observed streamflow and total nitrogen concentration from the WQN records were used to calibrate the USGS's constituent load estimator, LOADEST (Cohn, 2005; Cohn et al., 1992, 1989). To apply the proposed CCA bias correction procedure at the daily time scale, daily nitrogen loadings were estimated for the period 1951 to 2010 using the LOADEST model shown in equation (4.3). This model version was chosen from the 10 other predetermined models within the LOADEST program by using Akaike Information Criterion (AIC) (Akaike, 1974). The selected LOADEST model is

$$\ln(L_i) = \widehat{\alpha}_1 + \widehat{\alpha}_2 \ln(Q_i) + \widehat{\alpha}_3 \sin(2\pi dtime) + \widehat{\alpha}_4 \cos(2\pi dtime) + \widehat{\varepsilon}_i \quad (4.3)$$

where L_i is the observed daily total nitrogen loadings on day i ; Q_i is the observed daily streamflow on day i ; $dtime$ is the centered decimal time; $\widehat{\alpha}_1$ - $\widehat{\alpha}_4$ are model estimated coefficients; and $\widehat{\varepsilon}_i$ denotes the estimated model residual on day i .

4.2.3. SWAT Model Estimates

The watershed models were built in ArcGIS version 10.1 using ArcSWAT version 2012 (Arnold and Fohrer, 2005; P. W. Gassman et al., 2007). Delineation of the watersheds was done using the watershed delineator internal to ArcSWAT using digital elevation data from the National Elevation Dataset (NED) using 1/3-arc second resolution (USGS, 2009). Land cover information was obtained from the 2011 National Land Cover Database and soil type information from the SWAT US Soils database (Homer et al., 2015). Gridded 1/8th degree observed daily precipitation and temperature was used for the period 1949 to 2010 (Livneh et al., 2013). Total nitrogen loadings were calculated by summing outputs for organic nitrogen, ammonia, nitrite, and nitrate and multiplying by the streamflow at the stream gage location. The SWAT model was calibrated to maximize the Nash-Sutcliffe efficiency (NSE) in predicting observed monthly streamflow over the 60-year period (2 years were removed as a startup period) by changing the following parameters: initial SCS runoff curve number for moisture condition II (CN2), available water capacity of first soil layer (SOL_AWC), Soil evaporation compensation factor (EPCO), and plant uptake compensation factor (ESCO). NSE was chosen as the objective metric for calibration based on the suggested procedure for SWAT model calibration presented in Santhi et al. 2001. A summary of the hydrologic calibration performance in predicting monthly streamflow for the three watersheds is shown in Table 4.2. All of the NSE values are above 0.5 and all Pearson's Correlation Squared (ρ^2) are above 0.6. Nitrogen loadings from the SWAT model were calibrated using daily WQN observed loadings over the short time scale

(<200 days). Considering the simulation period (>20,000 days) is much longer than the observed time scale, low performance in calibration was expected. The following nitrogen related parameters were manipulated to improve the NSE of loadings: benthic P source rate coefficient (RS3), organic N settling rate coefficient (RS4), nitrogen uptake distribution (N_UPDIS). Other parameters related to the nitrogen cycle were initially considered for manipulation, but calibration runs showed that the NSE of loadings was not sensitive to changes in the excluded variables. Furthermore, if the parameters listed did not result in a noticeable difference (>0.1) in NSE performance after 30 iterations manual calibration was halted; a summary of performance metrics for loadings calibration is shown in Table 4.2.

Table 4.2. Performance metrics for SWAT model hydrologic and water quality calibration for the observed WQN period.

Watershed	Streamflow Calibration			Loadings Calibration		
	ρ^2	NSE	PBIAS	ρ^2	NSE	PBIAS
Tar River	0.76	0.66	23%	0.4	-0.11	-38%
Ogeechee River	0.78	0.56	43%	0.68	0.66	21%
Escambia River	0.69	0.65	13%	0.48	0.28	47%

4.3. Methodology: Univariate and Multivariate Bias Correction Methods

4.3.1. Motivation

Daily streamflow is available over the entire modeling period, hence performance evaluation with observed streamflow is quite possible. However, over the entire prediction period loadings performance values for NSE were much lower compared to hydrologic calibration NSE values, as shown in Table 4.2. Pearson’s Correlation Squared values turned out to be only slightly lower (~0.2) in value for loadings calibration when compared to hydrologic calibration, also shown in Table 4.2. NSE and ρ^2 both approach a value of 1 for perfect model prediction but as the deviation between model and observed values increases the two values,

NSE and ρ^2 , begin to vary in value (Krause et al., 2005). Considering that loadings calibration for the Tar and Escambia have moderate values for ρ^2 (~0.4) but have poorer values for NSE (< 0.2) suggests that there is considerable bias in model estimates; this is also indicated by high values of percent bias (PBIAS) shown in Table 4.2. Based on this, bias correction is needed to reduce the amount of bias present between the observed and SWAT model values. Since observed loadings are limited, which restricts the period for bias correction, daily loadings from LOADEST are considered when bias-correcting the SWAT model over the entire simulation period. The LOADEST model is able to capture more than 80% of the variability in observed loadings from the WQN data using equation (4.3). The Nash-Sutcliffe efficiency (NSE) for the LOADEST model in predicting the observed loadings is 0.91 for Tar River, 0.93 for Ogeechee River, and 0.81 for Escambia River. This study proposes a multivariate regression technique known as canonical correlation analysis (CCA) that reduces the bias in streamflow and loadings prediction. The performance of CCA is first compared with a simple linear regression in reducing the bias as well as in preserving the cross-correlation between observed streamflow and loadings from the WQN data. Then, CCA is used in reducing the bias in predicting daily streamflow and daily loadings over the entire period, considering observed streamflow and LOADEST loadings estimates. Lastly, bias is removed using CCA from monthly streamflow and monthly loadings predictions considering observed monthly streamflow and monthly LOADEST estimates for the considered 60-year period (1951-2010).

4.3.2. Simple Linear Regression

CCA-based bias-corrected values of streamflow and loadings are compared with a simple linear regression (LR) approach for the WQN data using the same training and validation data used for developing CCA approach. Equation (4.4) shows the form of the linear adjustment with

$\widehat{\beta}_0$ denoting a constant, $\widehat{\beta}_1$ denoting the slope of the observed to model fit and $\widehat{\varepsilon}_t$ the random error term. These linear adjustment parameters are fit by minimizing the sum of squared residuals, $\widehat{\varepsilon}_t$, between the model estimates and Y_{obs} . Linear regressions are done separately for streamflow and loadings.

$$Y_{obs,t} = \widehat{\beta}_0 + \widehat{\beta}_1 Y_{SWAT,t} + \widehat{\varepsilon}_t \quad (4.4)$$

The residuals of the fitted values must be approximately normal to satisfy the assumptions that a linear model is appropriate. Normality was checked by using the probability plot correlation coefficient (PPCC) (Vogel, 1986); which calculates the correlation of the normality plot for the residuals. The PPCC values for regressions on streamflow and loadings, respectively for each watershed are: Tar River (0.90, 0.93), Ogeechee River (0.99, 0.98), and Escambia River (0.86, 0.95). The high PPCC values indicate that the residuals are approximately normal.

4.3.3. Multivariate Bias Correction

Canonical correlation analysis (CCA) is a regression based technique where multiple predictors (SWAT model streamflow and loadings) and multiple predictands (observed streamflow and loadings) are rotated in such a way to maximize the correlation between two linear combinations of the variables. CCA has been employed for bias correction of precipitation and temperature from GCM outputs (Das Bhowmik et al., 2017). It's been used in attributing the key sources in forecasting using GCMs (Barnett and Preisendorfer, 1987) and also for identifying high correlated model fields for predicting precipitating anomalies (Tippett et al., 2003). Anytime you have multiple model values that are related like, precipitation and temperature or streamflow and nutrient load, and they have significant cross-correlation it makes sense to bias correct them simultaneously. However, significant cross-correlation is not a

requirement for using CCA it will simultaneously reduce individual bias amount the variates. In this study, we will be using CCA as a multivariate regression to reduce the bias from SWAT model outputs.

CCA rotates centered multivariate arrays of observations, denoted \mathbf{Y} , of streamflow and loadings from WQN data and the corresponding modeled values from SWAT, denoted \mathbf{X} , and maximizes the correlation between the rotated \mathbf{X} and \mathbf{Y} matrices. Both observed (\mathbf{Y}) and model predicted values (\mathbf{X}) are first transformed into the log space to avoid negative values of discharge and loadings after bias correction. The matrices will be of size $(n \times p)$, with n being the number of observations or estimations and p being the number of variables, in this case $p=2$. Canonical coefficients \mathbf{A}, \mathbf{B} of size $(p \times p)$ are used to rotate the centered matrices \mathbf{X} and \mathbf{Y} into orthogonal space shown in (4.5) and (4.6). Matrix \mathbf{U} corresponds to the variable \mathbf{X} in rotated space and matrix \mathbf{V} corresponds to the variable \mathbf{Y} in rotated space. The canonical correlations \mathbf{r} , of size $(p \times 1)$, relate the columns between rotated variables \mathbf{U} and \mathbf{V} ; for example, r_2 is the correlation between U_2 and V_2 . The canonical correlations can then be used to relate the rotated model matrix \mathbf{U} to the rotated observed matrix \mathbf{V} as shown in (4.7), a process similarly done in (Das Bhowmik et al., 2017), with \mathbf{V}' representing the rotated model matrix related to the observed matrix. Afterwards, \mathbf{V}' is rotated back to log-space from orthogonal space using the \mathbf{B} matrix shown in (4.8), the observed mean is added back in log-space, and then transformed back to normal space.

$$\mathbf{U} = (\mathbf{X} - \mu_{\mathbf{X}}) \mathbf{A} \quad (4.5)$$

$$\mathbf{V} = (\mathbf{Y} - \mu_{\mathbf{Y}}) \mathbf{B} \quad (4.6)$$

$$\mathbf{V}' = \mathbf{U} \mathbf{r} \quad (4.7)$$

$$X' = V'inv(B) \quad (4.8)$$

CCA is evaluated in a split-sample validation approach by using a \mathbf{Y} matrix of observed monthly streamflow and monthly LOADEST estimates with an \mathbf{X} matrix of SWAT monthly streamflow and loadings for a given month (January, February, etc.) covering the entire period. The first half of the data (month 1- month 30) is used as training data set and the second half (month 31-month 60) is used as a validation data set. For example, the training data set for the January CCA model would include Jan 1951-Jan1980, and the validation data set would include Jan 1981- Jan 2010. The training data set (denoted by subscript t) is transformed into log-space and centered using μ_t , it is then rotated into orthogonal space into \mathbf{U}_t and \mathbf{V}_t matrices using canonical coefficients $\mathbf{A}_t, \mathbf{B}_t$. Canonical correlations (r_t) relate \mathbf{U}_t and \mathbf{V}_t of the training data set. The validation data set, \mathbf{X}_v , (denoted by subscript v) is log-transformed, centered, and then rotated using its own means (μ_v) and canonical coefficients ($\mathbf{A}_v, \mathbf{B}_v$) shown in (4.9). Then \mathbf{V}_v' is produced for the validation set using \mathbf{U}_v and \mathbf{r}_t as shown in (4.10). Using \mathbf{B}_t from the training data set, \mathbf{V}_v' can be rotated back to log-space to \mathbf{X}'_v using (4.11). Finally, the mean from the training data set, μ_t , is added back in log space to the validation set in (4.12) and transformed into normal space. This process is repeated for each month of the year and is outlined in (Figure 4.2).

$$U_v = (X_v - \mu_{x_v})A_v \quad (4.9)$$

$$V_v' = U_v r_t \quad (4.10)$$

$$X_v' = V_v' inv(B_t) \quad (4.11)$$

$$X_v = X_v' + \mu_{x_t} \quad (4.12)$$

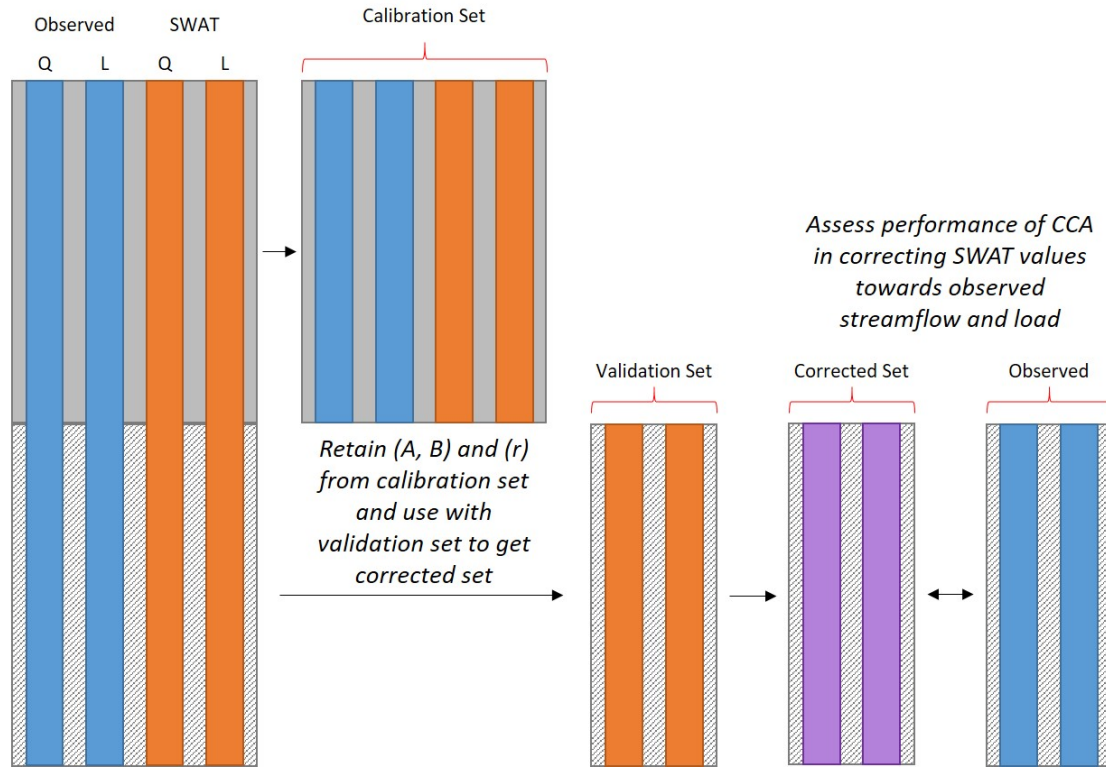


Figure 4.2. Canonical Correlation Analysis (CCA) Approach for bias correcting SWAT model outputs. Using a split sample approach, canonical coefficients and correlations from the calibration data set are applied to the validation data set.

The second canonical correlation, r_2 , is much lower (< 0.3) than the first canonical correlation (~ 0.8) across all three watersheds, thus, we decided to retain only the first rotated components of X and Y for further analysis. In this case, the model and observed matrices are composed of streamflow and loadings; with streamflow being the main driver of loadings. The primary reason for the second canonical correlation being so low is because the first component explains most of the variance leaving little unexplained variance outside the first component; the first component comes mainly from variance explained in streamflow. For a more in depth explanation of canonical correlation analysis, please see (Wilks, 1995).

4.3.4. Comparison of Bias Correction Techniques over Different Time Scales

This section first presents a general description of CCA for application to the observed WQN period. Following that, we discuss setting up of monthly and daily bias prediction models and provide how both models are evaluated under split-sample validation at monthly and daily time scales. The performance of the two techniques – CCA and simple regression (described in section 4.3.2) – is compared by their ability in preserving the cross-correlation as well as in estimating the joint probability of the observed streamflow and observed/LOADEST loadings.

WQN data: To begin with, application of CCA is first compared with simple regression for the observed WQN data (i.e. < 200 days) to illustrate its utility in preserving the observed cross-correlation from the WQN data. Since the WQN data consists of non-consecutive days and has such a short sample length, model validation was not performed for WQN data. The performance of the simple regression and CCA approach was evaluated in reducing the bias in the observed cross-correlation between loadings and streamflow of the WQN data. Since the data length is short, we did not perform any cross-validation under WQN data. We next compare the CCA and simple regression for bias correcting daily and monthly predictions of streamflow and loadings from the SWAT model.

Bias Correction of Monthly SWAT Predictions: For bias correcting at monthly time scale, a separate CCA model for each calendar month was developed using the observed monthly streamflow and LOADEST estimated monthly loadings as Y for bias correction for the 60-year period. Monthly loadings from the LOADEST model was obtained by aggregating the daily loadings estimated from daily streamflow for the 60-year period (1951-2010). Separate bias correction models for each month is advantageous because calibration is done only using months that are experiencing the similar inter-annual variability, which should improve bias correction.

For example, the CCA model for January would have 60 monthly values, one for each year over the entire period. Given we have 60-year length for each month, we can also evaluate the monthly bias correction model under split-sample validation with 30 years for model fitting and 30 years for validation. This type of validation structure allows for application to a forecasting mode. Historical observations and simulations can train the bias correction model without knowledge of the future for use with forecasting simulations.

Bias Correction of Daily SWAT Predictions: Given the data length is quite long with 60 years of daily predictions from the SWAT model, a 30-day moving window is considered for bias correcting daily streamflow and loadings for training (Figure 4.3). For example, to bias-correct streamflow and loadings for Jan 1, 1990 (denoted as the target day), the training data set would include streamflow and loadings from December 17th-31st 1989 and January 2nd-16th 1990. CCA techniques are trained over the 30-day moving window with the left-out day being added in the validation dataset. In this case the observed matrix, \mathbf{Y} , consists of daily streamflow and daily LOADEST estimates and the model matrix, \mathbf{X} , consisting of daily SWAT streamflow and loadings. The moving window approach dynamically changes the training data set to the previous and subsequent 15 days (resulting in a 30-day training period) for every day in the 60-year period; the first and last 15 days of the period are thrown-out for comparison. Canonical coefficients ($\mathbf{A}_t, \mathbf{B}_t$) and correlations (\mathbf{r}_t) are retained from the 30-day training data set after matrices \mathbf{X}_t and \mathbf{Y}_t are centering and rotated to make \mathbf{U}_t and \mathbf{V}_t . Canonical correlations are used with rotated matrix from the validation data set (\mathbf{U}_v) over the last 30 years, which includes model values from the 30-day window plus the target day (resulting in a 31-day validation period), to relate to the \mathbf{V}_v matrix and canonical coefficient, \mathbf{B}_t , is used to rotate to \mathbf{X}_v' in a fashion shown in (4.11). Lastly, the mean from the training data set is added back to the validation data set. The

bias-corrected value for streamflow and loadings only on the target day is kept and the process is repeated for each day over the entire period; this process is outlined in (Figure 4.3). In this case, we are using past and present information to bias correct, respective to the day being corrected. This structure would be appropriate for correcting long periods of historical simulations.

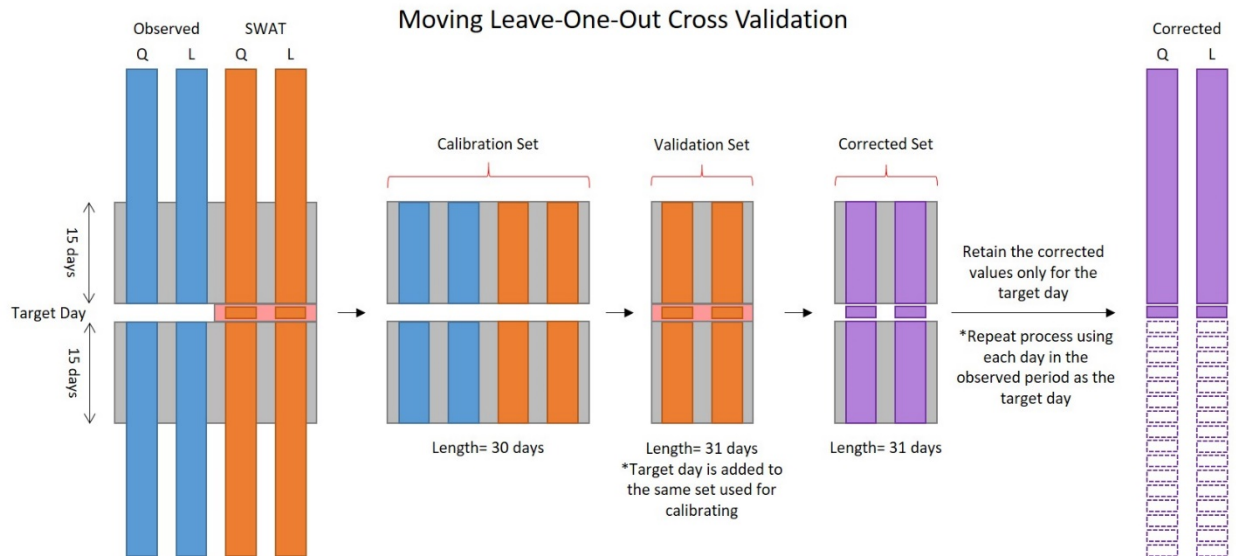


Figure 4.3. Canonical Correlation Analysis (CCA) Approach for bias correcting daily SWAT model outputs based on 30-day moving window for training with the middle day being considered for validation.

4.3.5. Comparing Cross-Correlations

Fisher's Z-transformation test is used to determine whether the model/bias-corrected cross-correlation between the loadings and nutrients is statistically equal to the observed cross-correlation. Given that correlations do not follow normality for small sample sizes or population with high correlation, Fisher (1992) suggested a transformation of correlation shown in equation (4.14) that results in a normal distribution for hypothesis testing on correlation (Fisher, 1992). The z-transformation has a standard error shown in equation (4.15) with a sample length of n (Fisher, 1992). Since the z-transformation follows a normal distribution, we can use the standard normal z-table to establish significance. Given two cross-correlations, we first find the

difference in their respective z-transformations and divide it by the standard error of the difference to create a standard normal z-score, Z' , using equation (4.16). The variance of the difference in z-transformations is the summation of reciprocals of the sample sizes ($n-3$) in this study, the sample lengths of the observed and models are equal, so the variance just becomes $2/(n-3)$ (Fisher, 1992). Taking the square root of the variance gives the standard error of the difference in z-transformations and (4.16) becomes (4.17). A standard normal z-score can be looked up using an alpha level of say 5% and compared to the Z' ; if $Z' < Z_{\alpha/2}$ then the difference between the two cross-correlations are not significant.

$$Z = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) \quad (4.14)$$

$$\sigma_z = \frac{1}{\sqrt{n-3}} \quad (4.15)$$

$$Z' = \frac{Z_{obs} - Z_{model}}{\sigma_{Z_{obs} - Z_{model}}} \quad (4.16)$$

$$Z' = \frac{Z_{obs} - Z_{model} (\sqrt{n-3})}{\sqrt{2}} \quad (4.17)$$

4.3.6. Comparing Joint Likelihoods

In addition to preserving cross-correlation, the performance of bias correction techniques was also evaluated for their ability in predicting the joint likelihood of streamflow and loadings using observed moments. Streamflow and water quality constituents have both been known to follow log-normal distributions (Bowers et al., 2012; Van Buren et al., 1997). Hence, we assume the joint likelihood of streamflow and loadings follow a bivariate lognormal distribution, denoted as “multilog”. Observed means of streamflow and loadings, $\boldsymbol{\mu}_{obs}$, and the observed covariance matrix, \boldsymbol{COV}_{obs} , were used to estimate joint probabilities of observed streamflow and loadings in

equation (4.18). Likewise, the joint probabilities of CCA bias-corrected values and LR bias-corrected values were estimated using the observed moments based on equations 19 and 20 respectively. The joint probability using equation (4.18) is expected to be higher than the bias-corrected streamflow and loadings given in equations (4.19) and (4.20).

$$(Q_{Obs}, L_{Obs}) \sim \text{MultiLOG}(\mu_{obs}, COV_{obs}) \quad (4.18)$$

$$(Q_{CCA}, L_{CCA}) \sim \text{MultiLOG}(\mu_{obs}, COV_{obs}) \quad (4.19)$$

$$(Q_{LR}, L_{LR}) \sim \text{MultiLOG}(\mu_{obs}, COV_{obs}) \quad (4.20)$$

The estimated joint probabilities from the above three equations were compared based on box-plots.

4.4. Results: Performance of Bias Correction Techniques over different time scales

4.4.1. WQN Data

Linear bias correction using the regression model was compared with CCA bias correction by correcting daily swat model values only for days corresponding with WQN observations. The Nash-Sutcliffe Efficiency (NSE) of the bias-corrected streamflow for both CCA and linear regression is shown in (Figure 4.4). NSE values for the raw SWAT model predictions are shown in the parentheses. Linear bias correction (Figure 4.4b) shows better performance in improving the NSE of SWAT discharge in predicting the observed discharge compared to the CCA approach in two basins (Figure 4.4a). CCA bias correction shows a reduction in NSE from the NSE of the SWAT model in Tar and Escambia River basins, while Ogeechee River is showing a slight improvement, going from 0.56 NSE to 0.64. Overall, the absolute differences in NSE between the raw SWAT values and CCA bias-corrected values, for discharge, are less than 0.1.

Bias correction of SWAT loadings using CCA (Figure 4.5a) shows more improvement in NSE values from raw SWAT loadings for both Tar and Escambia River in comparison to the performance of LR (Figure 4.5b). In the case of Ogeechee River, CCA bias-corrected loadings performance (NSE=0.53) is slightly lower than the LR corrected values (NSE=0.68) (Figure 4.5). From the three watersheds, Ogeechee River was the best calibrated watershed in terms of loadings (NSE=0.66) and the worst calibrated watershed in terms of streamflow (NSE=0.56), yet after CCA bias correction for Ogeechee River basin performs the best for streamflow (NSE=0.64) and the worst in loadings (NSE=0.53). Effectively, CCA being a multivariate multiple regression, bias corrects both streamflow and loadings by trading off the explained variance in bias-corrected streamflow and loadings.

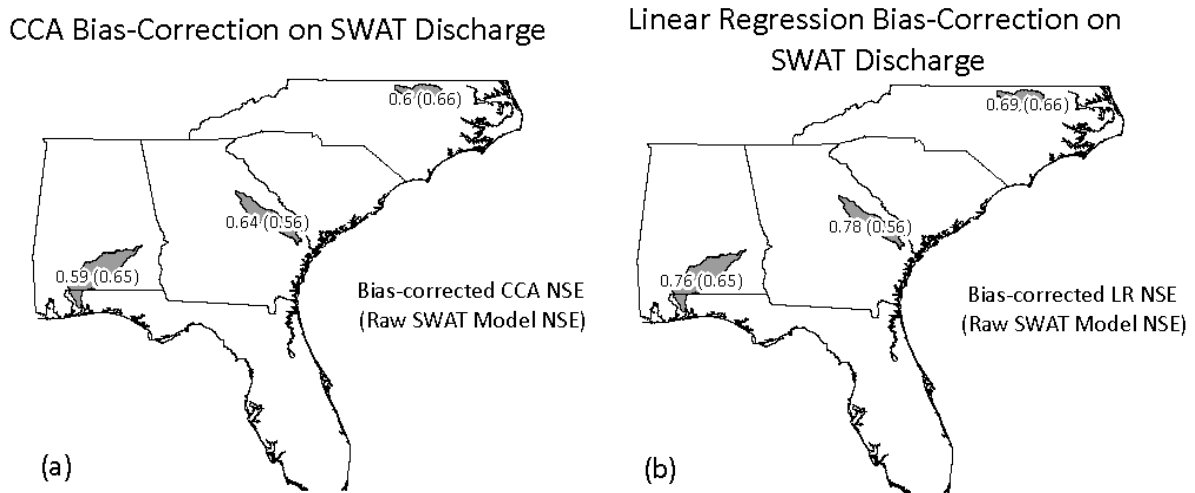
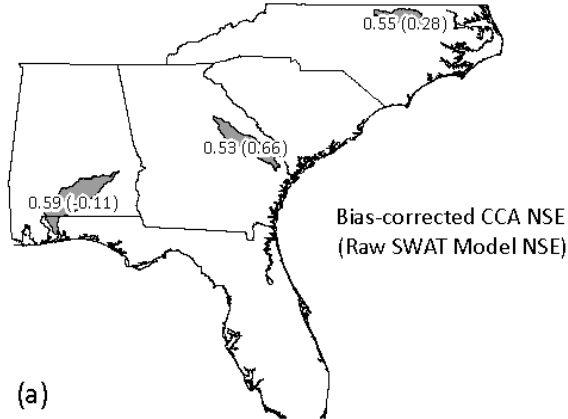


Figure 4.4. Performance comparison of bias correction techniques, based on NSE, in predicting the observed discharge in the WQN database: (a) canonical correlation analysis and (b) simple linear regression.

CCA Bias-Correction on SWAT Loadings



Linear Regression Bias-Correction on SWAT Loadings

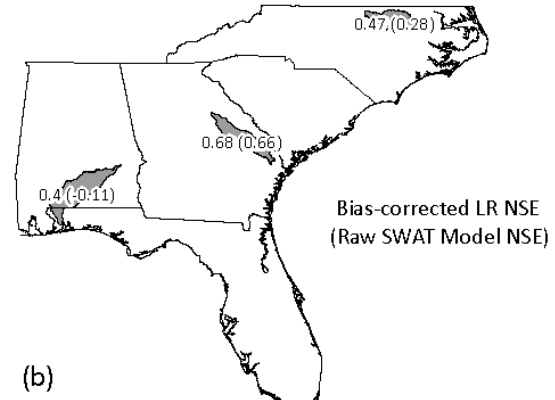


Figure 4.5. Performance comparison of bias correction techniques, based on NSE, in predicting the observed loadings in the WQN database: (a) canonical correlation analysis and (b) simple linear regression.

Cross-correlation is better preserved under CCA approach in comparison to using the LR approach. Correlation between the bias-corrected streamflow and loadings using CCA are all within +/- 0.01 of the observed cross-correlation for each watershed, as shown in (Figure 4.6). Using a Fisher Z-transformation test at an alpha level of 5%, reveals that the cross-correlations of the CCA corrected values are not significantly different than the observed. Specifically, all the Z' values are below 1.96 with Tar, Ogeechee, and Escambia River having scores of 0.97, 1.03, and 0.36 respectively. Cross-correlation of the raw SWAT model values from the Ogeechee River watershed was initially $\rho_{cross} = 0.93$, which is not significantly different than the observed correlation being $\rho_{cross} = 0.95$, using a Fisher Z-transformation. However, (Figure 4.6) shows that the CCA correction method is still able to improve the cross-correlation resulting in $\rho_{cross} = 0.94$. Cross-correlations of bias-corrected values using LR did not improve from the raw model values as shown in (Figure 4.6). Tar and Ogeechee River both have high cross-correlations (>0.90) while Escambia, the largest watershed, has a lower value of $\rho_{cross} = 0.88$.

Preserving the cross-correlation during bias correction is important specifically for management periods (5-10 years) where streamflow is understood to change over time.

In addition to better preserving the observed cross-correlation, CCA also does a better job in predicting the joint likelihood of bias-corrected streamflow and loadings given the observed moments. Each observation over the WQN period has an associated probability using the bivariate log normal distribution. Similarly, the probability of CCA and LR bias-corrected values also have an associated probability for each day over the period. The differences in observed probabilities to the probabilities from the raw SWAT model, CCA, and LR for each day are shown as boxplots in (Figure 4.7) for all basins. The mean of the CCA boxplot shows to be centered very close to zero, while LR and the raw SWAT model are shifted below zero; suggesting that CCA better preserves in predicting the joint likelihood of streamflow and loadings for the Tar River and Escambia River basins. The difference in the observed standard deviation to the standard deviation of CCA is 3.9×10^{-8} , and 9.2×10^{-8} for LR indicating that CCA captures the observed variability of the WQN data for the Tar River basin. CCA is the preferred method over LR for predicting the mean and standard deviation of the observed joint likelihood for Escambia River as well as shown in (Figure 4.7). Estimates from the SWAT model were able to capture the mean and standard deviation relatively well before bias correction, thus both CCA and LR are not able to improve prediction of the joint likelihood beyond the model estimates.

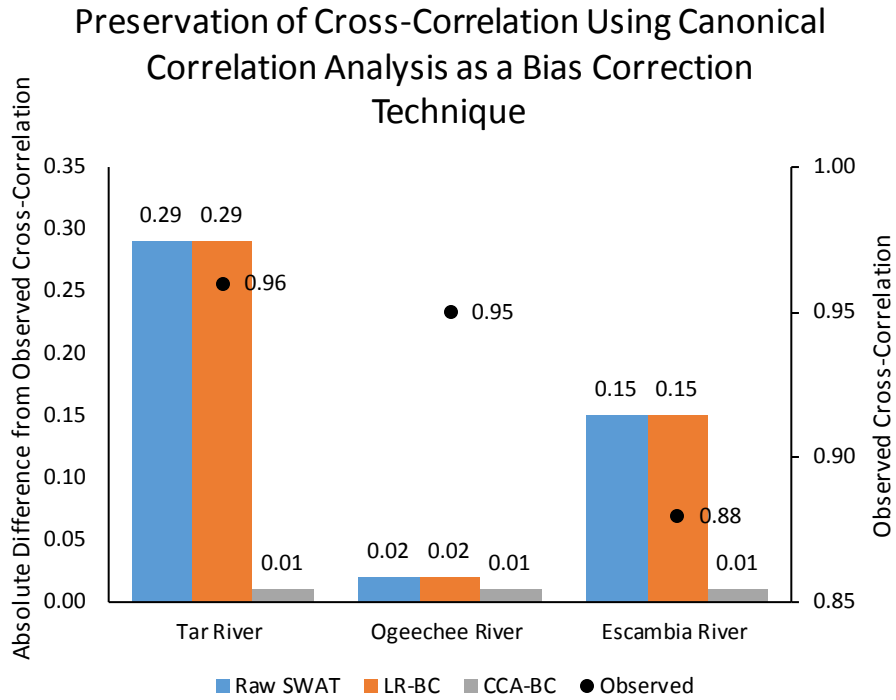


Figure 4.6. Performance of bias correction techniques, CCA and LR, in preserving the observed cross-correlation between the loadings and discharge in the WQN data for the selected three watersheds.

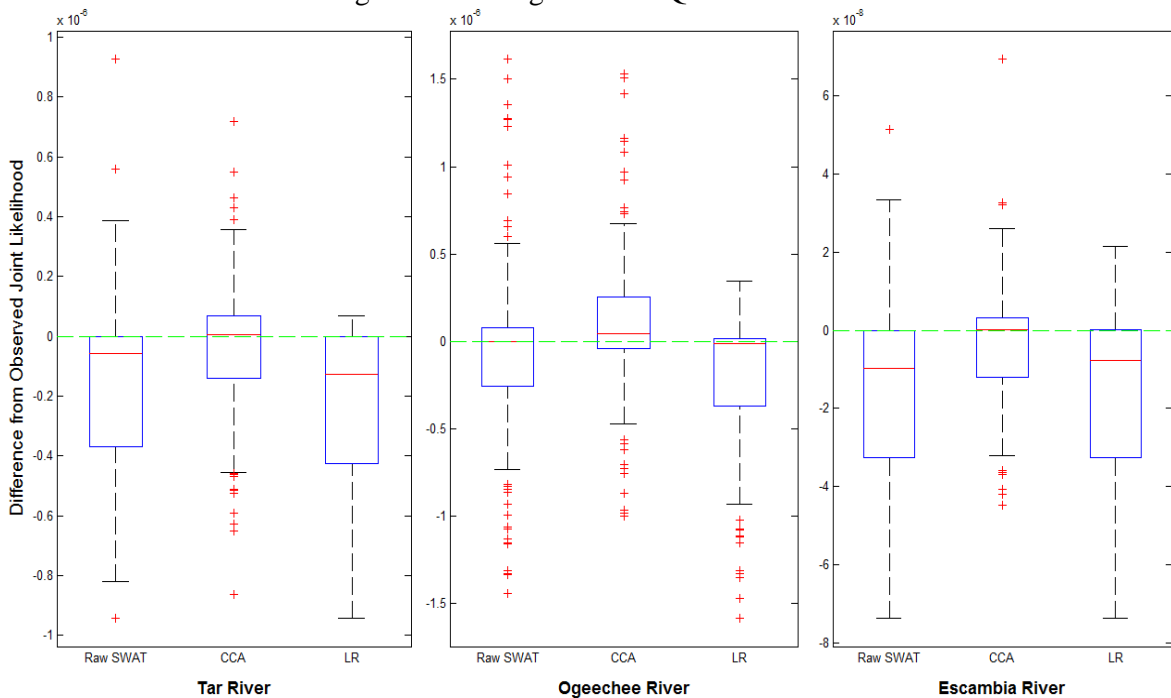
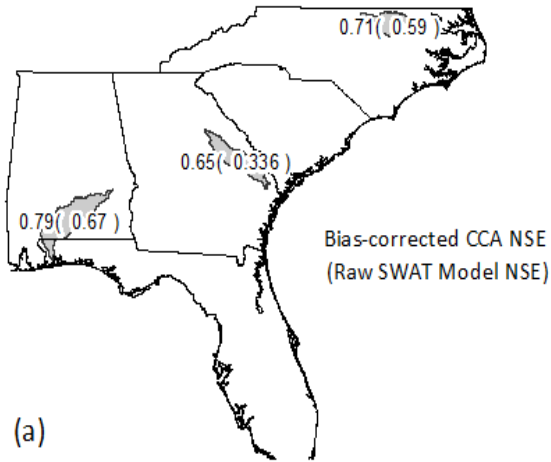


Figure 4.7. Difference between the observed and estimated (left: SWAT model; middle: CCA; right: LR) joint probability of streamflow and loadings of the WQN data for (left) the Tar River at Tarboro, NC (center) Ogeechee River near Eden, GA, and (right) Escambia River near Century, FL basins.

30 Day Moving Window Bias-Correction
on SWAT Discharge



30 Day Moving Window Bias-Correction
on SWAT Loadings

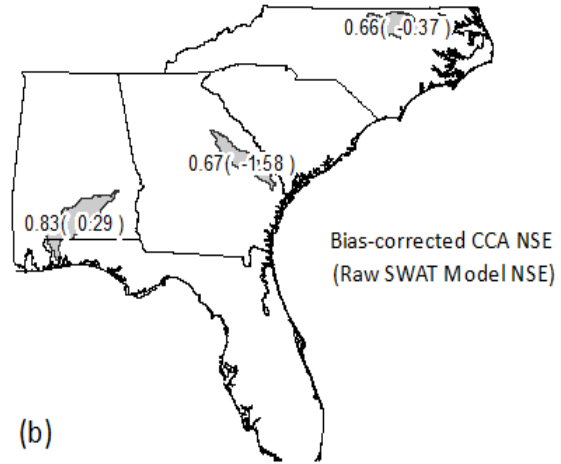


Figure 4.8. Performance of 30-day moving window CCA approach in improving the performance of raw daily SWAT model predictions, evaluated by NSE, of observed discharge and LOADEST loadings.

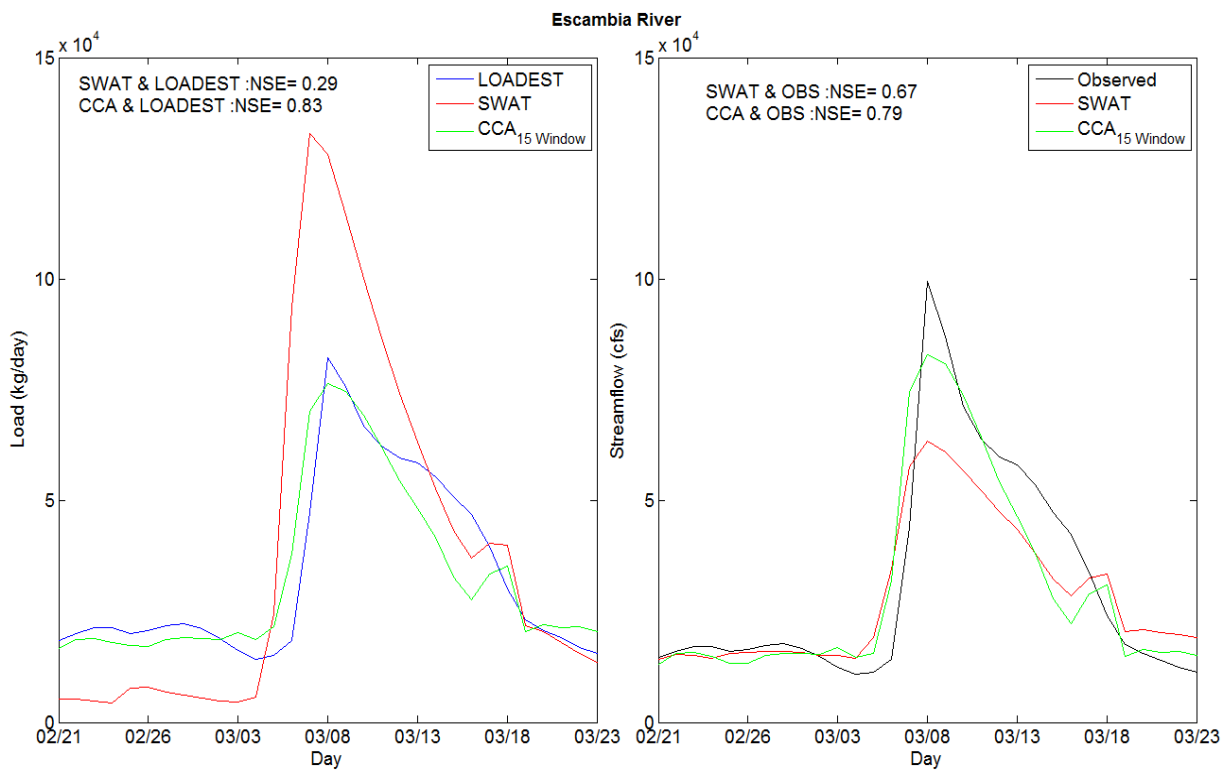


Figure 4.9. Performance of 30-day CCA approach in bias correction of raw daily SWAT predictions compared to LOADEST loadings and observed streamflow for the Escambia River near Eden, GA over the period February 22nd, 1998-March 24th 1998, a specific window pertaining to a high flow event.

4.4.2. Performance of Bias Correction Techniques over Daily and Monthly time scales

Given the advantages in preserving the observed cross-correlation and joint likelihood when using the CCA approach for the observed WQN data, LR was not compared for bias-correcting daily and monthly values. Bias correction for the daily time scale for discharge and loadings using the moving window CCA approach for the selected three watersheds was compared to the original raw SWAT model performance in (Figure 4.8). Improvement from the raw SWAT performance with CCA bias-corrected NSE values being larger than 0.6 (Figure 4.8a). CCA bias-corrected performance shows an even larger increase in NSE values for loadings, shown in (Figure 4.8b), with all the raw SWAT performance NSE values being lower than 0.3 and bias-corrected NSE being all above 0.6. Bias-corrected streamflow performance for Escambia River is $NSE=0.79$ which is an improvement from the raw SWAT performance of $NSE=0.67$. Tar River basin ($NSE=+0.12$) and Ogeechee River basin ($NSE=+0.25$) also improved the NSE on the bias-corrected streamflow shown in (Figure 4.8a). Considerable performance improvement is seen for bias-corrected loadings with a NSE value of 0.83 coming from a raw SWAT NSE value of 0.29 in Escambia River. Raw SWAT model performance in predicting LOADEST loadings at daily time scale was $NSE=-0.23$, but after CCA NSE improved to 0.84. Ogeechee River did not exhibit any increase in performance for daily bias-corrected loadings for CCA, in fact NSE dropped by a value of 0.06. Recall, that for the daily timescale we are using daily LOADEST estimates for loadings in the \mathbf{Y} matrix of CCA. Given that observed streamflow is a predictor in determining the LOADEST loadings estimates, shown in (3), the observed cross-correlation can be expected to be very high (Table 4.3). Observed cross-correlation between daily streamflow and daily LOADEST loadings are 0.99 which is close to 1, the upper limit of correlation, for Tar and Escambia River. CCA is shown to perfectly preserve

the observed cross-correlation (Table 4.3) under extreme values of correlation which demonstrates the robust utility of CCA.

Table 4.3. Performance of 30-day moving window CCA (in comparison to raw SWAT predictions) in preserving the cross-correlation between observed discharge and LOADEST loadings for the selected three watersheds.

Cross-correlation	30 Day Moving Window		
	Tar River	Ogeechee River	Escambia River
Observed	0.99	0.70	0.99
Raw SWAT	0.72	0.67	0.67
Post-CCA	0.99	0.69	0.99

Since the entire daily bias-corrected data set is too long to show in one window, a hand-picked 30-day window containing a high flow event in early March of 1998 is shown for examining advantages of CCA (Figure 4.9). The NSE performance over the full period is shown in the top left-hand corners of the figure for reference (Figure 4.9). Bias-corrected loadings in (Figure 4.9 left) from CCA (green line) follow more closely with the LOADEST loadings (blue line) better than the raw SWAT model estimates (red line). Likewise, the bias-corrected streamflow (Figure 4.9 right) from the CCA approach (green) follows more closely with the observed streamflow (black line). CCA bias-corrected values for both streamflow and loadings in (Figure 4.9) shows that are shifted away from the raw SWAT values (red lines) towards observed/LOADEST values. This 30-day window shows that the SWAT model over estimates the observed high flow event and underestimates the LOADEST loadings estimate. For this case, CCA is able to reduce the overestimation in streamflow and underestimation in loadings. Considering the entire 60-year period, CCA is able to explain more observed variance than the raw SWAT model estimates shown by the NSE terms in the upper left-hand corner. These two arguments illustrate the ability of CCA to shift the magnitude and variability of model values

towards observations. Since this CCA bias correction technique for daily values uses local data points it is most suited for bias-correcting predictions for continuous historical periods.

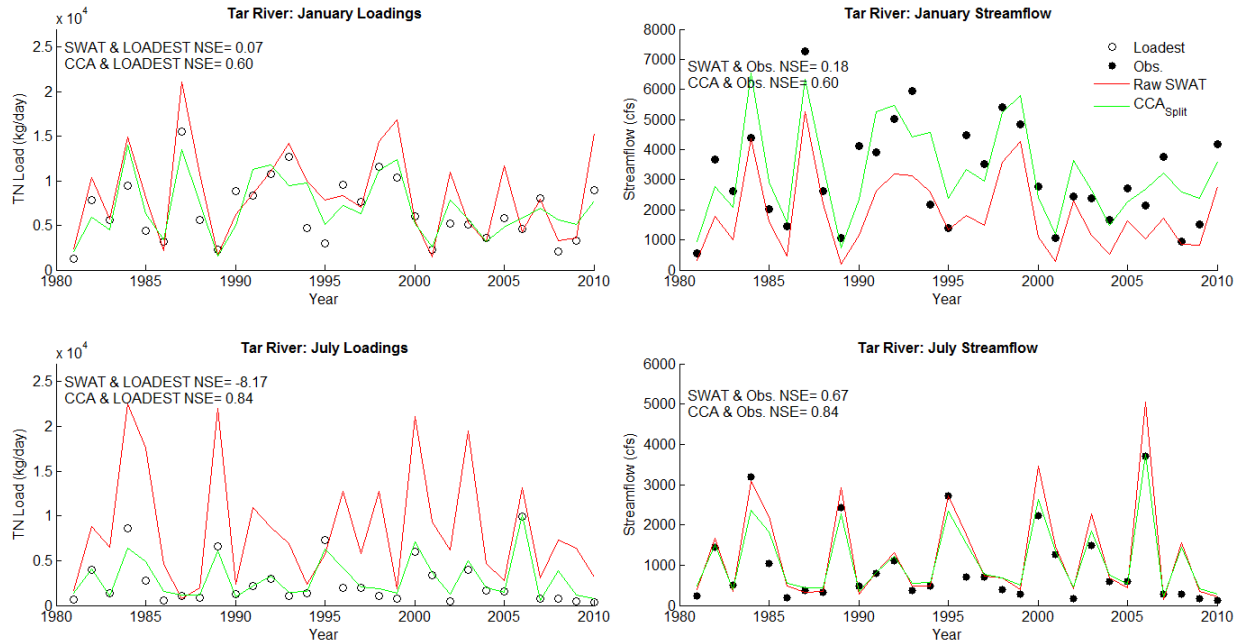


Figure 4.10. Performance of CCA approach, evaluated by NSE, in bias correcting monthly SWAT loadings and discharge compared to monthly LOADEST loadings and observed discharge for the Tar River at Tarboro, NC basin under split-sample validation.

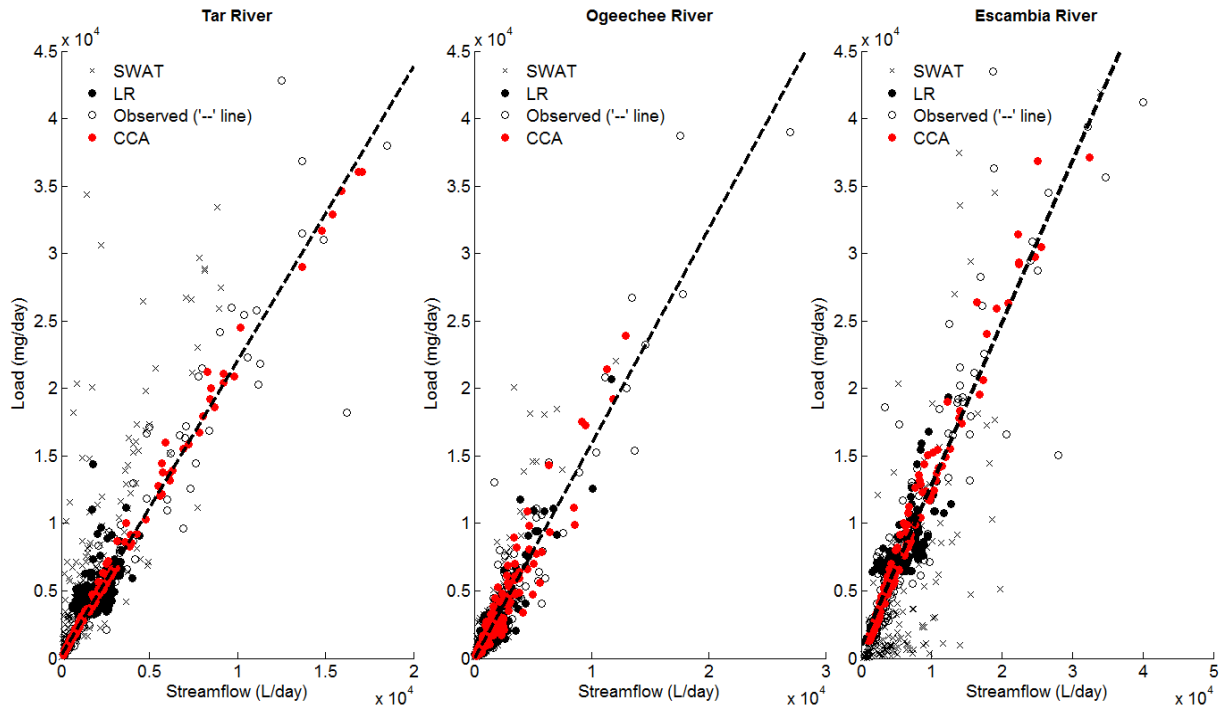


Figure 4.11. Comparison of WQN observed loadings and discharge (hollow black marker) with raw SWAT predictions (black cross), CCA approach (filled red marker) and Simple Linear Regression (filled black marker) for the selected three river basins over the Southeast US.

Monthly bias correction using CCA was conducted for every month of the year but discussion is summarized for only two months – January and July – with one from the winter season and another from the summer season for the Tar River basin. Left column of (Figure 4.10) shows bias correction for loadings and the right column shows bias correction for streamflow in January (top) and in July (bottom). The red lines show the original raw SWAT simulations while the green lines show the CCA bias-corrected values. In the loading plots (Figure 4.10 left) monthly LOADEST values are shown as empty circles and in the streamflow plots (Figure 4.10 right) the monthly observed means are shown as filled in black dots. Bias-corrected loadings for both January and July show improvement from the raw SWAT performance, which is reflected in the overall NSE values with increases from 0.02 to 0.60 in the January and from -8.2 to 0.84 in July. Loadings from CCA bias correction shifts from the raw

SWAT values and towards the LOADEST values, seen most prominently in January for the years 1994 and 1995 and for almost every year in July shown in (Figure 4.10 left). Streamflow also shows the same shift for CCA bias-corrected values, moving away from SWAT values to observed monthly streamflow (black dots) for almost every January from the years 1981-2010. Thus, CCA provides a better and robust methodology for bias correcting both streamflow and loadings from the SWAT model at daily and monthly time scales.

4.5. Discussion and Concluding Remarks

Modeling water quality constituents in mechanistic models is generally more difficult than streamflow modeling given the complex interactions between different constituents. For instance, total nitrogen concentration is the sum of organic, ammonium, nitrite, and nitrate species all which have equations estimating their respective concentrations on a daily time scale. Thus, the matrix of parameters determining total nitrogen estimates is very large and adjusting every possible parameter is time consuming and unwieldy. Further, availability of limited and discontinuous records makes it hard to validate the predictions. On the other hand, calibrating and validating the SWAT model for streamflow is relatively easy and often results in NSE values greater than 0.6 mainly due to the long observation period from USGS streamgages dating back before 1950. In addition, capturing the observed variability in streamflow estimates comes mainly from the explained variability in precipitation inputs. Irrespective of the long record of streamflow, it is always required to remove the systematic bias in the estimation of streamflow by the SWAT model using commonly available bias correction techniques (Koch and Smillie, 1986). Given the vast literature on bias correcting the streamflow, we suggest here a multivariate bias correction technique that removes bias in the estimation of observed moments of streamflow and loadings including the observed cross-correlation between streamflow and loadings.

Common bias correction techniques used in hydrological studies are univariate given that they aim to remove bias from streamflow alone. These univariate techniques are commonly using simple linear regressions or quantile regressions that relate the probability of exceedence between the observed and model predicted values. In studies where more than one variables are of interest in prediction, multivariate techniques are more appropriate. Multivariate bias correction of precipitation and temperature by way of ACCA or MACCA has shown to be effective for using climate change projections. The purpose of this paper was to propose a multivariate technique, specifically CCA, for bias-correcting streamflow and loadings simultaneously for improving water quality predictions.

This study has shown that the multivariate bias correction, CCA, has potential in reducing individual variables' biases from the SWAT model and also improves the bias in estimating the cross-correlation between streamflow and loadings even with limited samples available from the WQN data. Observed variability is also better explained in terms of NSE for bias-corrected values in comparison to raw model estimates. This partly stems from the reduced PBIAS for the bias-corrected predictions. To summarize, CCA bias correction is preferable to simple linear regression because the multivariate technique can preserve the observed cross-correlation, while univariate bias correction has limited ability in preserving the cross-correlation. The joint likelihood of estimating observed streamflow and loadings is also higher under CCA compared to univariate technique. Using the moving window CCA bias correction technique is also useful for extending daily records over long periods. CCA has also shown potential in reducing bias in moments including cross-correlation at monthly time scale.

Water quality models (e.g., LOADEST) that provide estimations of loadings also can offer estimations of concentrations to fill in sparse observations on the concentration side. For

this study, we were mainly interested in bias-correcting loadings and did not evaluate the performance of bias correction techniques in estimating TN concentration. However, plotting bias-corrected loadings versus bias-corrected streamflow for the observed WQN period can show how well the concentration was estimated. Figure 4.11 compares the performance of CCA and LR in bias-correcting SWAT model predicted loadings and streamflow for the selected three river basins. From Figure 4.11, we clearly see that the streamflow and loadings bias corrected by CCA follow closely with the observed loadings and streamflow indicating better prediction of observed concentration. A noticeable shift from the raw SWAT predictions could also be observed for both the CCA and LR approach. Thus, Figure 4.11 shows that CCA-based bias correction improves the prediction of both streamflow and loadings resulting in overall improved estimates of predicted TN concentration.

This study recognizes that there is an inherent issue with calibrating and validating a water quality model given the infrequent and sparse observation data sets. Apart from using statistical models like LOADEST and WRTDS for extending the data record, this study uses a mechanistic model to develop continuous water quality records and also suggests a multivariate bias correction procedure that preserves the cross-correlation structure between streamflow and loadings. By considering the LOADEST estimations as the “truth”, we evaluated the performance of the bias correction procedures at daily and monthly time scales for the three selected watersheds. Using LOADEST estimates as a surrogate in the observed matrix \mathbf{Y} have potential downfalls as the LOADEST estimates have errors associated with. Hence, any application of bias correction techniques with LOADEST model as the truth should consider the performance of the LOADEST model. Oh and Sankarasubramanian (2012) consider the R^2 of LOADEST model and the TN forecasting model for potential application towards seasonal

forecasting. The primary advantage of using mechanistic model like SWAT with the proposed bias correction technique is in exploiting the long meteorological records to develop streamflow and loadings estimates for watersheds subjected to limited anthropogenic influence. Further, bias correcting mechanistic model outputs could also provide streamflow and loadings under potential climate and land-use changes.

Improved bias correction techniques such as CCA are especially useful for water quality modelers that provide improve joint probability of estimating observed streamflow and loadings. Our analyses over daily and monthly time scales has shown CCA flexibility in preserving the cross-correlation structure between streamflow and loadings. Bias correction on daily SWAT estimates is useful for monitoring downstream loadings from wastewater treatment plants that discharge loadings of effluents with different concentrations (i.e., primary, secondary and tertiary) on a daily timescale. The proposed bias correction framework can also be adapted for use in nutrient (nitrogen or phosphorus) loadings forecasting that could be used in management practices (e.g., nutrient reduction plans) to plan on the monthly timescale using climate forecasts.

The mechanistic or physical based models can be very useful for nutrient reduction plans by forecasting monthly-to-seasonal water quality. Incorporating monthly bias-corrected streamflow and loadings is important for reservoir managers, so that potential eutrophic conditions could be forecasted using climate forecasts. Forcing forecasted precipitation and temperature from climate models with the SWAT model, one could obtain yield monthly forecasted loadings and streamflow for locations of interest. Bias correction can be done using hindcasted periods of streamflow and loadings as the training data set and used the retained coefficients and correlations to develop real-time 1-3 month ahead streamflow and loadings forecasts (i.e. 1 to 3 months ahead). Thus, the demonstration of the multivariate bias correction

approach based on CCA can be useful in improving streamflow and water quality predictions at daily, monthly and decadal time scales, which have relevant applications from water quality management perspective.

Chapter 5. Reducing Error in Streamflow and Total Nitrogen Loadings Forecasts from the SWAT Model

Abstract

It is widely known that regression models can forecast streamflow with moderate skill in predicting the observed variability using previous streamflow and GCM forcings as predictors. Considering that the variability in nutrient load is carried largely by the streamflow variability one could expect nutrient load forecasts to also have similar skill as streamflow forecasts. Popular regression models, like the LOADEST and WRTDS models, have been used to predict nutrient load with great skill ($\rho^2 > 0.8$) using streamflow, time, and season as predictors. These regression models require observed streamflow and are not easily applied to future periods without forecasted streamflow. Mechanistic models like the SWAT model can be adapted for producing forecasts of streamflow and nutrient load simultaneously given future forcings from atmospheric global circulation models (AGCMs). Nutrient measurements spanning multiple decades currently exist as sparse and non-continuous data sets, like that from the Water Quality Network (WQN), limited by cost and labor. Using nutrient load estimates from the LOADEST model as the best estimates for observed nutrient load allow for long-term calibration and validation of the SWAT model. Downscaled and disaggregated precipitation and temperature from the ECHAM4.5 AGCM can be forced with the SWAT model for decadal skill assessment of 1-month ahead forecasts. Monthly forecasts are compared with forecasts using observed (perfect forecasts) and climatological forcings (climatological forecasts) after a post-simulation bias-correction technique is applied to reduce the systematic SWAT model bias.

5.1. Introduction

Surface waters containing pollutant concentrations beyond the criteria levels set in US Clean Water Act of 1972 (PUBLIC LAW 92-500) are classified as impaired waters and are required to set restoration goals. The National Strategy for Development of Nutrient Criteria of 1998 helped define criteria levels specific to water body type and climate by urging states to adopt Numeric Nutrient Water Quality Criteria to reduce nutrient pollution (US EPA, 1998). Water bodies violating the imposed criteria then must adopt watershed rules for reducing the nutrient levels. Larger water bodies that serve as reservoirs for drinking water and recreation are at greatest risk of becoming impaired given the large amount of point sources, streams and tributaries, and low retention time. Although, some watershed nutrient strategies have been aimed towards reducing nutrient levels in lakes by mixing and pumping (Herman et al., 2017), mostly reduction strategies are designed to reduce the amount of nutrient flux entering upstream (Arheimer et al., 2005; Stow et al., 2003). Up-stream nutrient load is estimated as the product of observed streamflow and nutrient concentration. Since studies have shown that climate variability will affect future streamflow events (Oh and Sankarasubramanian, 2012) we can expect nutrient load to also be effected by climate variability. Forecasting the upstream nutrient load into waterbodies can inform watershed managers how current nutrient reduction strategies will operate under high and low flow conditions to reservoirs.

Forecasting nutrient load requires accurate streamflow and concentration estimates for calibrating and validating statistical and physical based forecasting models. Long-term streamflow estimates are available from USGS streamflow gages from 1931 to present. Water quality measurements, like total nitrogen concentration (mg-N/L), with continuous samplings are limited to streams that flow into impaired reservoirs or water bodies at risk of impairment.

Monitoring programs of streams from pristine watersheds, like records from the Water Quality Network (WQN), have sparse and infrequently samplings over multiple decades given high cost and labor requirements (Alexander et al., 1998; Cohn et al., 1992). Load estimation models, like the USGS's Load Estimator (LOADEST) and Weighted Regression on Time, Discharge, and Season (WRTDS) models, can use streamflow and seasonality as predictors to estimate nutrient load on days without observations. Predictions from these models are limited to the historical period where observed streamflow is required. Using the LOADEST or WRTDS model to forecast nutrient loadings would require forecasted streamflow. Statistical based methods have shown to produce skillful streamflow forecasts (Vrugt et al., 2008), one method is to use antecedent streamflow and precipitation as predictors in a regression based forecast (Garen, 1992). Using streamflow forecasts as predictors in the LOADEST or WRTDS model to forecast nutrient load could produce compounded error from using multiple estimates ultimately reducing forecasting skill. Statistical based methods have been used to forecast nutrient load such as principal component regression (PCR) using retrospective precipitation forecasts from the ECHAM 4.5 global circulation model as predictors (Oh and Sankarasubramanian, 2012). Other nonlinear methods for estimating nutrient load, such as using a Bayesian approach have been formulated for the use as a forecasting tool (Qian et al., 2005).

An alternative to using statistical based models for nutrient forecasting is to use a mechanistic model like the Soil and Water Assessment Tool (SWAT) (Douglas-Mankin et al., 2010). An advantage of the SWAT model is that can simultaneously forecast streamflow and nutrient load given adequate hydrological and water quality calibration. Using observed streamflow and nutrient load estimates from the LOADEST model as the "truth" a SWAT model can be calibrated for up to 30+ years. The SWAT model has shown to produce reliable daily and

seasonal streamflow and nutrient load predictions using observed climate information (Jha et al., 2007; Libera and Sankarasubramanian, 2018). Retrospective climate forecasts of precipitation and temperature from global circulation models (GCMs) can be forced with a mechanistic model to produce retrospective forecasts of both streamflow and nutrient load. Other mechanistic models like the VIC model have shown to produce just skillful streamflow forecasts using retrospective climate forecasts (Sinha et al., 2014). Although, some studies have used SWAT for forecasting nutrient load under future climate change scenarios (years 2071-2100) (Serpa et al., 2017); studies focusing on forecasting nutrient load using retrospective climate forecasts are generally lacking. This study uses the SWAT model to forecast streamflow and nutrient load simultaneously using retrospective climate forecasts of precipitation and temperature from the ECHAM 4.5 global circulation model (GCM). A 1-month ahead forecasting scheme is built external to the original SWAT model FORTRAN code so that initial conditions can be updated at the beginning of each forecast. External SWAT models have been built before but only using components from the daily water balance, not from nutrient cycling (Sun et al., 2015). Additionally, a systematic bias from the SWAT model is removed to using a multivariate approach (Libera and Sankarasubramanian, 2018) prior to quantifying the skill of climate forecasts. The skill of streamflow and total nitrogen load forecasts from three watersheds from Region 3 of the Southeastern U.S. are compared using three types of climate information to make retrospective forecasts: climatological averages of P & T, observed P & T, and forecasted P & T from the GCM. This study examines the role of using future climate information from a GCM in forecasting both streamflow and nutrient load. The framework of the forecasting scheme is designed to be useful for informing watershed managers monthly compliance with watershed rules of impaired waterbodies.

5.2. Data

5.2.1. WQN Measurements

This study considers three watersheds chosen from Region 3 of the Southern Eastern United States (Figure 5.1). These three watersheds were considered in a previous chapter for predicting streamflow and nutrients using observed forcings. Selection of these watersheds from the region were based on the length of nutrient observation and basin size. Total nitrogen observations from National Stream Water-Quality Monitoring Networks (WQN) (Alexander et al., 1998) range from 1962 to 1995 with sampling frequencies varying from monthly to seasonally. Total nitrogen concentration is reported as the summation of nitrate-nitrogen ($\text{NO}_3\text{-N}$), nitrite-nitrogen ($\text{NO}_2\text{-N}$), ammonia-nitrogen ($\text{NH}_3\text{-N}$) and organic nitrogen. Observed nitrogen load is calculated by multiplying the observed total nitrogen concentration from the WQN by the average streamflow on that day from USGS stream gages with appropriate conversion factors. Table 5.1 shows a summary of the total number of daily observations for nitrogen loadings for all three watersheds with each having less than 200 days spanning over 20 years. Further, these watersheds are part of the USGS Hydro-Climatic Data Network (HCDN), a subset of USGS streamgages, whose streamflow are minimally impacted by anthropogenic influences (Slack et al., 1993). The HCDN is useful for forecasting studies since no storage or pumping damper the climate signal (Oh and Sankarasubramanian, 2012; Vogel and Sankarasubramanian, 2005). Streamflow and nitrogen load forecasts are the only constituents estimated from the SWAT model, but the forecasting framework can be applied to other water quality variables, such as total phosphorus load.

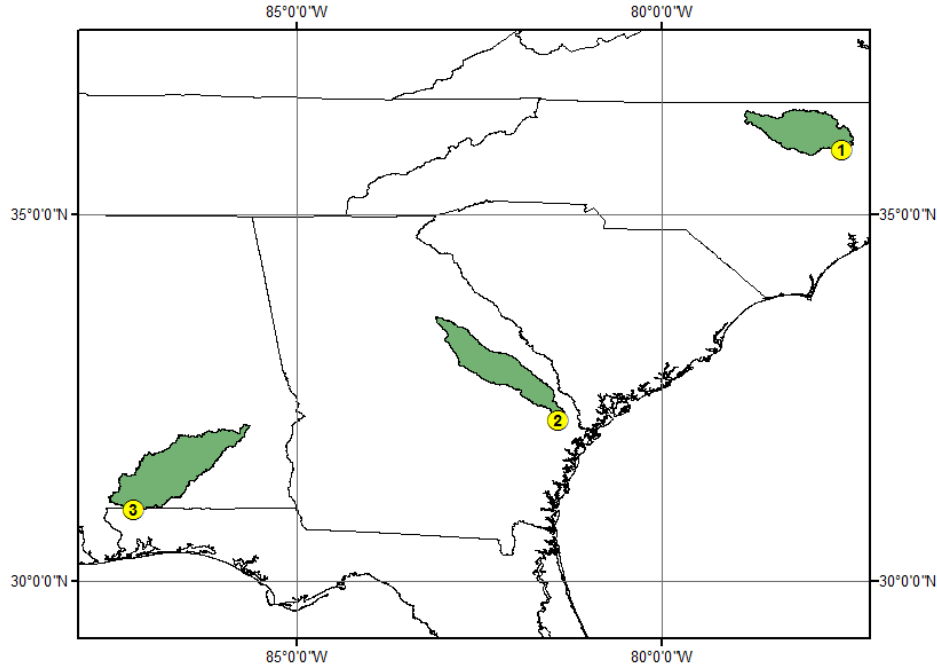


Figure 5.1. Selected three watersheds across the southeast region chosen for developing nutrient forecasts using the SWAT model. Green shaded areas represent the drainage area and numbered points provide the stream gauge locations; (1) Tar River at Tarboro, NC (2) Ogeechee River near Eden, GA (3) Escambia River near Century, FL.

Table 5.1. Summary of selected watersheds showing the number of daily observations for total nitrogen records from the Water Quality Network (WQN) database. Numbers in parenthesis in the far column represent the number of years the daily observations span.

Station Index	Station Number	Station Name	Drainage Area (sq. miles)	Number of Daily Obs. (# of yrs.)
1	02083500	Tar River at Tarboro, NC.	2183	167 (23)
2	02202500	Ogeechee River near Eden, GA	2650	148 (22)
3	02375500	Escambia River near Century, FL	3817	147 (22)

5.2.2. Performance Measures

The performance statistic used to quantify model performance is the Nash-Sutcliffe Efficiency (NSE). Nash-Sutcliffe Efficiency measures the squared deviations of the model value, \hat{X}_i , to the observed value, X_i , with respect to the squared deviations of the observed values to the mean of the observations, \bar{X} , as shown in equation (5.1) (Krause et al., 2005); here

the values could be for either streamflow or loadings from the SWAT or the LOADEST model. NSE describes how much of the observed variance is captured by the model values and ranges from negative infinity to one. A NSE value of 1 indicates that the model is perfect and capturing all of the observed variance while a NSE value that tends towards less than and close to zero is considered a poor model (Santhi et al., 2001).

$$NSE = 1 - \frac{\sum_{i=1}^N (X_i - \widehat{X}_i)^2}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (5.1)$$

Percent bias (PBIAS) is used to quantify the deviation between the observed and model values normalized by the observations (5.2). NSE and PBIAS were both chosen to quantify the SWAT model performance based on their wide use in the SWAT modeling community and availability of satisfactory performance ratings for both streamflow and nutrient load (D. N. Moriasi et al., 2007).

$$PBIAS = \left(\frac{\sum_{i=1}^n (X_i - \widehat{X}_i)}{\sum_{i=1}^n X_i} \right) \times 100\% \quad (5.2)$$

Further, Pearson's correlation coefficient squared, denoted as ρ^2 , is used to quantify the linear correspondence between model forecasts and observations, and shown in (5.3).

Correlation values range from 0 to 1, with a value of 1 signifying perfect correlation and a value of 0 indicating no correlation between model and observed values.

$$\rho^2 = \frac{\left(\sum_{i=1}^n (X_i - \bar{X})(\widehat{X}_i - \widehat{\bar{X}}) \right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (\widehat{X}_i - \widehat{\bar{X}})^2} \quad (5.3)$$

Relative root mean square error (R-RMSE) is used to quantify the accuracy present between model forecasts and observations, normalized by the mean of the

observations (4). Values of R-RMSE range from 0 to infinity with a value of 0 indicating no deviation between forecasts and observations.

$$R - RMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \hat{X}_i)^2}}{\bar{X}} \quad (5.4)$$

5.2.3. Nitrogen Load Estimates

The USGS's Constituent Load Estimator (LOADEST) tool is a FORTRAN program available for estimating in-stream constituent loadings using streamflow, time, and season as predictors (Cohn, 2005; Cohn et al., 1992, 1989). Streamflow and decimal time are centered for reducing multicollinearity during calibration (Cohn, 2005). When using uncensored observations, maximum likelihood estimation (MLE) is used to determine model coefficients and standard errors. Estimates for log constituent load are found by using the fitted model coefficients with the time series defined in the estimation file. Log load estimates are transformed back to the original space using a bias-correction factor, defined as the function phi within LOADEST (Likeš, 1980), to yield minimum variance unbiased estimates of instantaneous load (Cohn, 2005). Load estimates from model #7 of LOADEST were used to calibrate nitrogen loads from the SWAT model. Observed streamflow and concentration from the WQN dataset were used to fit model coefficients in (5.3); where L_i is the observed total nitrogen load on day i , Q_i is the observed streamflow on day i , $dtime$ is the centered decimal time, $\hat{\alpha}_1 - \hat{\alpha}_4$ are model estimated coefficients, and $\hat{\epsilon}_i$ is the estimated model residual on day i . Previously discussed in Chapter 2, Model forms #0 and #7 performed better in capturing the observed variability than any other preset models. A Fisher Z-transformation showed that Model choice #0 did not provide a significant increase in the Pearson's correlation squared (ρ^2) from Model choice #7. Model choice #7 was used given that it had a lower dimension of predictors.

Performance of the LOADEST model in predicting the observed variability of nutrient load, measured in Nash-Sutcliffe Efficiency (NSE), for each watershed is : Tar River at Tarboro, NC (0.89), Ogeechee near Eden, GA (0.90), Escambia near Century, FL (0.76).

$$\ln(L_i) = \hat{\alpha}_1 + \hat{\alpha}_2 \ln(Q_i) + \hat{\alpha}_3 \sin(2\pi dtime) + \hat{\alpha}_4 \cos(2\pi dtime) + \hat{\varepsilon}_i \quad (5.3)$$

5.2.4. SWAT Model Estimates

The watershed models were built in ArcGIS version 10.1 using ArcSWAT version 2012 (Arnold and Fohrer, 2005; P. W. Gassman et al., 2007). Delineation of the watersheds were done using the watershed delineator internal to ArcSWAT using digital elevation data from the National Elevation Dataset (NED) using 1/3-arc second resolution (USGS, 2009). Land cover information was obtained from the 2011 National Land Cover Database and soil type information from the SWAT US Soils database (Homer et al., 2015). Gridded 1/8th degree observed daily precipitation and temperature from the period 1956 to 2010 was used for calibration (Livneh et al., 2013). Total nitrogen loadings were calculated by summing output concentrations of organic nitrogen, ammonia, nitrite, and nitrate and post multiplying by the streamflow at each WQN monitoring location. The SWAT model was calibrated to maximize the Nash-Sutcliffe efficiency (NSE) in predicting observed monthly streamflow over the 50-year period (2 years were removed as a startup period) by changing the hydrological parameters in Table 5.2. The objective metric, NSE, was used based on the suggestion by Santhi et al. 2001 for SWAT model calibration. Monthly NSE and percent bias (PBIAS) for hydrological calibration of each watershed are respectively: Tar River at Tarboro, NC (0.68, 21%), Ogeechee River near Eden, GA (0.53, 31%), Escambia River near Century, FL (0.82, 11%). All watersheds have NSE and PBIAS statistics that are considered to have a satisfactory

performance rating, except Ogeechee River that has a slightly high PBIAS value. A post-simulation bias removal, explained in Chapter 3, using CCA for reducing systematic bias.

Table 5.2. Parameters used in the SWAT model for hydrologic calibration of the 3 watersheds using 50 years of observed streamflow.

Parameters	Database	Units	Range	Default Value	Calibrated Value		
					Tar River	Ogeechee River	Escambia River
SOL_AWC	.sol	(mm/mm)	0-1	0.11	0.6	0.5	0.6
EPCO	.hru	(-)	0-1	1	0.8	0.8	0.5
ESCO	.hru	(-)	0-1	0.95	0.7	0.4	0.4
CANMX	.hru	(mm H2O)	0-100	0	5	0	1
OV_N	.hru	(-)	0.01-30	0.3	0.05	1	1
ICANAL	.rte	(-)	0-1	0	0.4	0.4	0.4
ALPHA_BNK	.rte	(days)	0-1	0	0.4	0.4	0.4
SURLAG	.bsn	(days)	1-24	4	2	4	1
GW_REVAP	.gw	(-)	0.02-0.2	0.02	0.2	0.2	0.02
CH_K2	.rte	(mm/hr)	-0.01-500	0	2.5	2.5	2.5
CN2	.mgt	(-)	35-98	60	85	80	80

Nitrogen loadings from the SWAT model were calibrated using aggregated monthly LOADEST estimates from the period 1961-2010. Water quality related parameters, shown in Table 5.3, were manipulated to improve the NSE of nitrogen loads. In some cases, NSE was not sensitive to changes of certain parameters so the default value was kept. Furthermore, if the parameters listed did not result in a noticeable difference (> 0.1 NSE) in performance after 30 iterations then manual calibration was halted; a summary of calibrated parameters used for loadings calibration are shown in Table 5.3. Efficiencies and biases, respectively, from water quality calibration for each watershed are as follows: Tar River at Tarboro, NC (0.01, -50%), Ogeechee River (0.50, 3%), and Escambia River near Eden, GA (0.35, -97%). Tar River and Escambia River both receive unsatisfactory performance ratings for calibration. Further inspection shows that the ρ^2 for both watersheds are 0.5 or greater. NSE and ρ^2 both approach a value of 1 for perfect model prediction but as the bias between modeled and observed values increase, the two statistics begin to depart in value (Krause et al., 2005). The deviation between

NSE and ρ^2 for Tar and Escambia River suggests that there is a considerable bias in model estimates; this is also indicated by high values of PBIAS.

Table 5.3. Parameters used in the SWAT model for water quality calibration of the 3 watersheds using LOADEST load estimates using Model choice #7.

Parameters	Database	Units	Range	Default Value	Calibrated Value		
					Tar River	Ogeechee River	Escambia River
ERORGN	.hru	ratio	1.1-5	calculated	1.1	5	2
RS3	.swq	mg N/m ² day	0.62-0.84	0.5	0.5	0.05	0.5
RS4	.swq	day-1	0.001-0.05	0.05	0.05	0.05	0.005
BC1	.swq	day-1	0.48-0.94	0.55	0.55	0.55	0.48
BC2	.swq	day-1	0.47-1.22	1.1	1.1	1.1	1.1
BC3	.swq	day-1	0.20-0.7	0.21	0.21	0.21	0.21
CH_ONCO	.rte	ppm	0.004-0.015	0	0	0	0
RCN	.bsn	mg N/liter	0.44	1	0	0.2	0.2
CMN	.bsn	ratio	0.0016	0.0003	0.003	0.003	0.003
CDN	.bsn	ratio	2.8	1.4	1.4	1.4	1.4
SDNCO	.bsn	ratio	0.13	0.05	1	1.1	0.05
N_UPDIS	.bsn	scaling constant	65	20	20	20	20
NPERCO	.bsn	coefficient	0.32	0.2	0.2	0.2	0.2
All	.wwq	mg N mg-1 algae	0.08	0.08	0.08	0.08	0.08

5.2.5. Climate Forcings

Monthly updated precipitation and temperature forecasts available through the International Research Institute (IRI) of Climate and Society data library (Li and Goddard, 2005) were retrieved from the ECHAM4.5 AGCM. This particular GCM was chosen based on the long available record of retrospective precipitation and temperature forecasts from 1967. Additionally, climate forecasts from this GCM have proven to produce useful skill in streamflow forecasting (Mazrooei et al., 2015; Sinha et al., 2014). Forecasts were averaged over the 24 available ensembles for the period 1961-2010. Statistical downscaling was used to reduce the spatial resolution of forecasts from $\sim 2.8^\circ$ to $1/8^\circ$. Monthly forecasts were issued at the beginning of each month. Monthly precipitation and temperature from four nearby $\sim 2.8^\circ$ grids were used as predictors in predicting observed precipitation and temperature at $1/8^\circ$ using

principal component regression (PCR). The SWAT model requires maximum and minimum temperature records for running simulations. Separate PCR models were used for downscaling maximum and minimum temperature at 1/8° resolution using the same predictors with different predictands, observed maximum and minimum temperature. Precipitation and temperature (maximum & minimum) were temporally disaggregated to the daily time scale using a kernel-nearest neighbor approach (K-NN)(Prairie et al., 2007). The two closest monthly observed means (called neighbors, K=2) to each downscaled monthly mean were used to create daily weighted ensembles. Daily weighted ensembles are averaged and scaled so that the monthly means equaled that of the observed. For a more in depth explanation of the downscaling and disaggregation application, see (Sinha and Sankarasubramanian, 2013) and (Sinha et al., 2014).

5.3. Methodology

5.3.1. Setting up SWAT Model for Streamflow and Load Forecasting

The SWAT model is a continuous simulation model based in the FORTRAN language that uses a closed loop, cold start-up simulation at the daily time step. Initial state variables (e.g. soil moisture, storage) cannot be defined for each hydraulic response unit (HRU) at the beginning of the simulation. A few simulations are required for the state variables to initialize hence, the first couple of years are considered as start-up years and discarded. The FORTRAN source code (.f) is available for download (<https://swat.tamu.edu/software/swat-executables/>), although the complexity and size of the source code makes it difficult to adapt the SWAT framework for 1-month ahead forecasting where observed state variables must be updated at the beginning of each month for each HRU. Alternatively, a calibrated SWAT model can be forced with observed precipitation and temperature for a spin-up period (5 years) on which state variables will approach that of the observed. Using those updated initial conditions, the SWAT

model can be forced with forecasted precipitation and temperature from ECHAM4.5 starting at the first of each month, to develop 1-month ahead streamflow and load forecasts (Figure 5.2). Forecasts are forced at one month at a time, meaning that this process must be repeated for each month in the evaluation period. For instance, to get forecasted streamflow and loadings from SWAT for January 1961, the model is first fed observed climate forcings from January 1956 to December 1960 and then forced with GCM precipitation and temperature (downscaled and disaggregated) on January 1961. The mean of streamflow and loadings from each forecasted month in the SWAT loop is retained as the 1-month ahead forecast.

Forcing the calibrated SWAT model in the forecasting loop with observed precipitation and temperature instead of climate forecasts yield a “perfect” forecast from the hydrological model. Perfect forecasts can be thought of as a simulation that uses perfect climate information. Observed forcings were used in the adapted SWAT model to get perfect forecasts from 1956 through 2010. The perfect forecast period extends 5 year prior to the period of GCM forecasts to provide simulations for training the bias correction technique discussed in the next section. Comparison of perfect forecasts to forecasts obtained using ECHAM4.5 climate forecasts provide the associated error introduced into streamflow and loadings forecasts due to imprecise GCM forecasts. Furthermore, forecasts obtained using climatological values of precipitation and temperature are known as climatological forecasts which can be compared to forecasts using GCM forcings to show the added value of using GCM forecasts. Climatological forecasts are created by forcing the adapted SWAT model with observed daily averages of precipitation and temperature from the historical period from 1961 to 2010 with updates of the initial conditions every month. Thus, climatological forecast also provide information on the role of initial conditions.

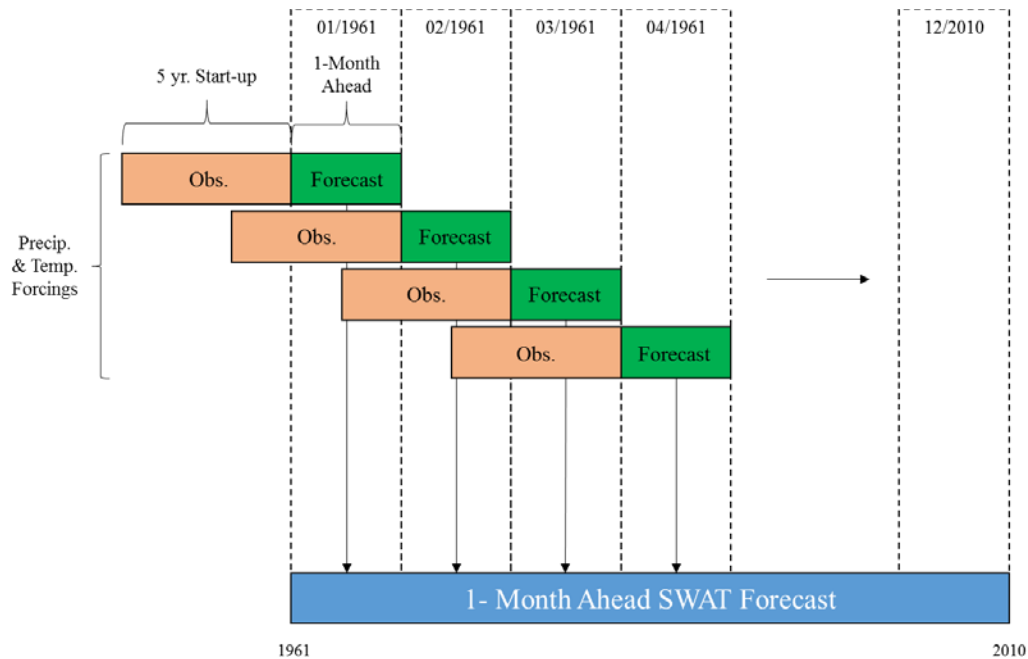


Figure 5.2. SWAT forecasting schematic framework. The SWAT model is run with observed precipitation and temperature for 5 years to initialize the model (light brown) and then run for 1 month using forcings from the ECHAM4.5 GCM (green).

5.3.2. Reducing SWAT Model Bias Using Canonical Correlation Analysis

Adapting the SWAT model for forecasting can introduce a systematic bias in addition to the bias already present from complications in calibration previously discussed in Section 5.2.4. Calibration bias reported is mainly present in loading estimates coming from SWAT simulations (no loop adaption). This bias in loadings estimates can be amplified when adapting SWAT using the loop framework. Canonical correlation analysis, as shown in depth in Chapter 4, is a multivariate bias correction technique that has been proven to simultaneously remove individual bias from streamflow and loadings from SWAT model predictions. Here CCA will be used in a similar way to remove bias from SWAT model forecasts. The goal of using CCA is to remove the bias introduced from the SWAT model. Realistically each monthly forecast should be bias corrected using only prior information at the time of the forecast, a process that mimics what a water resources manager might do in real time. Thus, a moving window approach used in

Chapter 4 is not appropriate since future information is used for training the model. This chapter suggests a leave-one-out “growing” window approach that uses all perfect forecasts and observations previous to the time of the GCM forecast to train the CCA model. Previous information from the same calendar month were used to train the model. For instance if February 1961 is being corrected then 5 perfect forecasts (February 1956-1960) and 5 observations are used to train the CCA model. The length of the training period “grows” or increases as each month in the forecasting evaluation period increases. The target month for correction is left out of the \mathbf{X} and \mathbf{Y} matrices used for training, this way only prior information is used (Figure 3). After training the CCA model and retaining the canonical coefficients, \mathbf{A} & \mathbf{B} , and canonical correlation, \mathbf{r}_1 , the target month is rotated and added to the training matrix \mathbf{U} , to create the validation set. Using the canonical correlation the validation set is related to the observed matrix, \mathbf{V} , and then transformed to normal space. For a more in-depth description of the CCA process see, Chapter 4: Section Methodology.

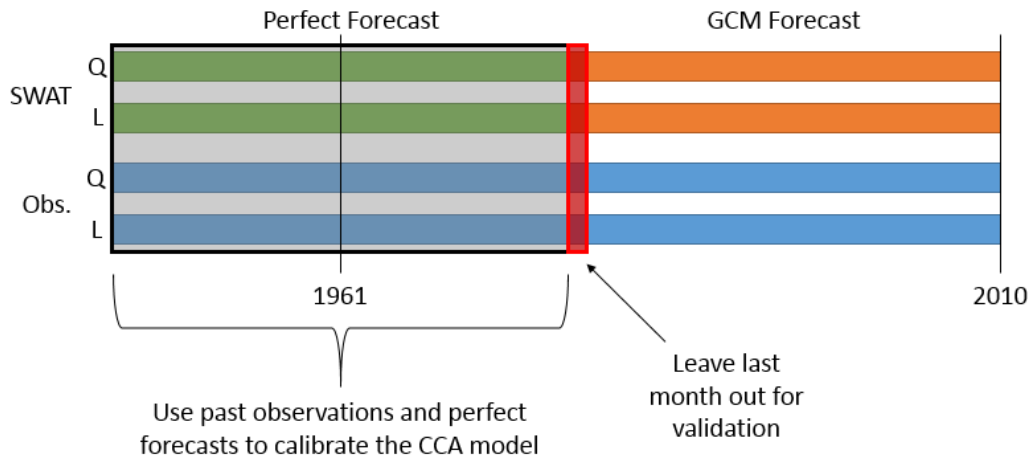


Figure 5.3. A CCA model removes the bias GCM forecasts introduced by the adapted SWAT model for each month by training the CCA model using previous observations and perfect forecasts.

5.3.3. Skill Scores and Relative Operating Characteristic

A common forecast verification metric used to quantify the skill of a forecast is the skill score (SS). Skill scores measure the difference in an accuracy metric from a forecast, A_{forecast} , to a reference forecast, A_{ref} , normalized by the difference in accuracy of that reference forecast to a perfect forecast, A_{perfect} , shown in (5.4). The skill score of a forecast is the amount of improvement over a reference forecast relative to the total possible improvement. A SS value of 1 for a forecast indicates a perfect improvement over the reference, while a value of 0 indicates no improvement. Pearson's correlation (ρ) and root mean squared error (RMSE) are popular accuracy measures that have been used for showing the skill score of forecasts. In this study, we will use the root mean squared error skill score (RMSE-SS) to quantify the improvement of using the GCM forecast over the climatological forecast relative to the perfect forecast.

$$SS = \frac{A_{\text{forecast}} - A_{\text{ref}}}{A_{\text{perfect}} - A_{\text{ref}}} \quad (5.4)$$

The RMSE-SS can be used to quantify the improvement of the forecast over the entire forecast period or individually for each month. An alternative to quantifying skill based on continuous forecasts is to quantify the skill based on the forecast's ability to predict certain categorical bins, such as above normal (AN) events. Observations can be converted to three categories: above normal (AN) events, normal (N) events, and below normal (BN) events using the 0.33, 0.67 and 1.0 quantiles. Similarly, forecasts can be converted to the same categories using the quantiles established from the observations. The true positive rate (aka. hit rate) of a forecast is the probability that a forecast will be predict the correct event given that the event occurred. The false positive rate (aka. false-alarm rate) is the probability that the forecast predicts an event given that the event did not occur. Figure 5.4 shows a three-by-three matrix with three categories (AN, N, BN) of streamflow forecasts for Escambia River using GCM

forcings on the y-axis and observed categories on the x-axis. Using this figure the true positive rate for the streamflow forecasts would be $195 / (195+5) = 0.975$ and the false positive rate would be $(140+50) / (2+58+140+33+117+50) = 0.475$. Ideally, reliable forecasts will have high true positive rates and low false positive rates. The relative operating characteristic (ROC) curve plots the true positive rate versus the false positive rate and connects the points pertaining to each bin. The ROC curve can be useful for comparing both rates simultaneously and can provide a visual comparison between the types of forecast used (i.e. perfect, GCM, climatology). Forecasts that have ROC curves that fall below the 1-1 line indicate that the false positive rates are higher than the true positive rates and can be considered unreliable.

Escambia River Streamflow

GCM Forecasted Events	AN	0	2	33
	N	5	58	117
	BN	195	140	50
		BN	N	AN
		Observed Events		

Figure 5.4. Three-by-three grid showing the number of falling into the BN, N, and AN categories for both the GCM forecasts and observations. This table is used in determining the true positive and false positive rates.

5.4. Results

The CCA bias-correction approach allows for simultaneous reduction of bias in streamflow and loadings specifically introduced by the SWAT model. All three stations from

the Southeast showed positive improvement in NSE for both streamflow and loadings (Table 5.4) over the forecast period after bias correction. Since bias correction is done simultaneously using a multivariate approach reduction of bias in one variable (e.g. streamflow) might be sacrificed with slight increase in bias for another variable (e.g. loadings). Performance metrics that did not improve after bias correction are shown as red cells in Table 5.4.

Table 5.4. Forecast performance statistics (post CCA bias-correction) for each forecast (climatological, GCM, and perfect) compared to the observed for the 3 watersheds across the Southeast. Green cells indicate positive improvement of that statistic compared to the raw forecast (pre bias-correction); red cells indicate no improvement.

	Climatological Forecast					
	Discharge (cfs)			Loadings (kg/day)		
	NSE	ρ^2	PBIAS	NSE	ρ^2	PBIAS
Tar	0.10	0.37	-51	-0.01	0.28	-55
Ogeechee	0.40	0.52	-32	0.20	0.41	-48
Escambia	-0.06	0.44	-59	-0.19	0.29	-68

	GCM Forecast					
	Discharge (cfs)			Loadings (kg/day)		
	NSE	ρ^2	PBIAS	NSE	ρ^2	PBIAS
Tar	0.37	0.42	-21	0.25	0.33	-29
Ogeechee	0.59	0.65	-18	0.35	0.58	-41
Escambia	0.47	0.58	-25	0.23	0.51	-44

	Perfect Forecast					
	Discharge (cfs)			Loadings (kg/day)		
	NSE	ρ^2	PBIAS	NSE	ρ^2	PBIAS
Tar	0.73	0.73	-3	0.65	0.69	-11
Ogeechee	0.71	0.71	1	0.58	0.67	-27
Escambia	0.83	0.82	-4	0.64	0.74	-28

As expected, perfect forecasts have the highest performance among all three forecasting schemes (Table 5.4). Values of NSE, ρ^2 , and PBIAS from the perfect forecast are all within acceptable performance ranges established by a literature review shown in (D. N. Moriasi et al.,

2007). Although PBIAS does not improve during bias correction for the perfect forecast, it still receives a “Good” performance rating, using according to ratings established by Moriasi et al 2007. These performance ratings were established using observed forcings and cannot be translated easily to the GCM forecasts. Forecasts using future climate information can add considerable error in streamflow and are expected to have lower skill than perfect forecasts. Looking at the post bias correction performance (Table 5.4), the GCM forecasts have higher performance for all metrics across all watersheds when compared to metrics from the climatological forecasts. This suggests that forecasts using future climate information are better in capturing the observed variability than using the observed climatology. Tar and Escambia River both have negative NSE values for loadings performance from climatological forecasts indicating that the historical mean of loadings is a better predictor than using the forecast (Table 5.4). These negative values in NSE are caused by large biases between the modeled and observed values as indicated by the large negative PBIASs. Further, when bias exists, large differences can be seen between the values of ρ^2 and NSE. Conversely, when bias is reduced ρ^2 and NSE values converge to the same value as seen in streamflow forecasts using perfect forcings.

Monthly performance statistics are shown for NSE, ρ^2 , and R-RMSE for both streamflow (top row) and loadings (bottom row) for Tar River (Figure 5.5), Ogeechee River (Figure 5.6), and Escambia River (Figure 5.7). In general, the perfect forecast (green line) is higher than the GCM forecast (red line) and climatological forecast (blue line) for monthly NSE and ρ^2 values (Figures 5.5-5.7). Perfect forecasts have much lower values of R-RMSE than that of the GCM forecast and climatological forecast. Since the bias-correction technique uses a “growing” window, meaning the length of the training period increases as the forecasting evaluation period

increases, the first couple of months corrected use only a few data points (~5) for training. This can cause the first year or two to have poor bias-correction values which will degrade monthly performance statistics and cause the perfect, GCM, and climatological forecasts lines in Figures 5.5-5.7 to overlap. Realistically, a GCM forecast cannot perform better than a perfect forecast which uses perfect climate information. A more robust way to compare forecasts is to use the long-term performance metrics which are less sensitive to deviations in individual data points. However, looking at the individual monthly statistics can show which forecasts are more useful for each months or season. The R-RMSE and ρ^2 lines in Figure 5 show that for September the GCM forecasts line and climatological line are almost matching indicating that the GCM forecast does not provide much additional skill over the climatological forecast for Tar River.

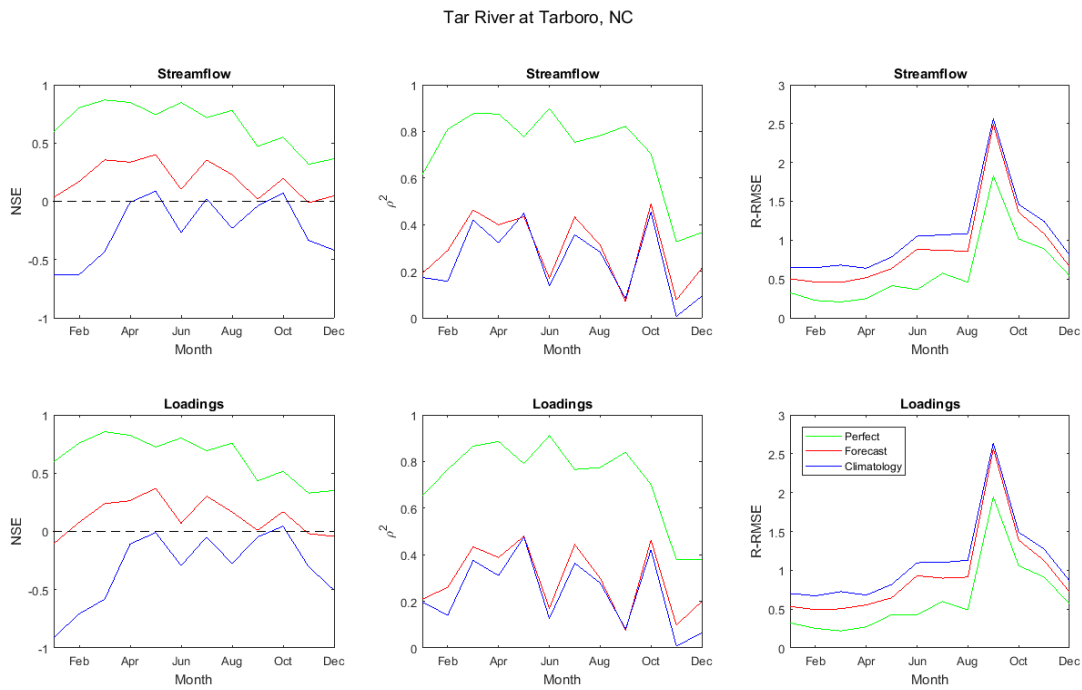


Figure 5.5. Performance of bias-corrected streamflow and loadings described by NSE, ρ^2 , and R-RMSE from climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Tar River at Tarboro, NC



Figure 5.6. Performance of bias-corrected streamflow and loadings described by NSE, ρ^2 , and R-RMSE from climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Ogeechee River near Eden, GA

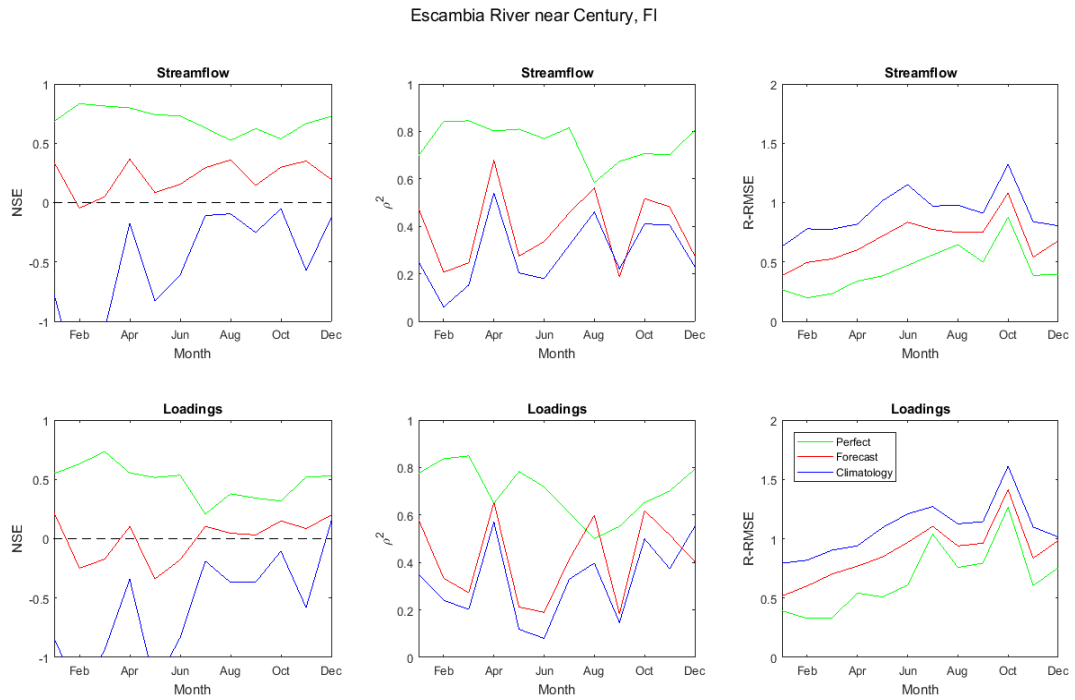


Figure 5.7. Performance of bias-corrected streamflow and loadings described by NSE, ρ^2 , and R-RMSE from climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Escambia River near Century, FL.

The skill score is a useful metric for quantifying the fraction of improvement of the GCM forecast over the climatological forecast. Table 5.5 lists a summary of the RMSE-SS for each watershed. All watersheds show improved skill in forecasting both streamflow and loadings using climate forecasts from the GCM. The RMSE-SS was also calculated for each individual month to show how the performance of the GCM forecast changes intra-annually (Figures 5.8-5.10). As discussed earlier, streamflow and loadings forecasts using forecasted precipitation and temperature from the GCM do not provide much additional information over the climatological forecast for Tar River in September. This can be inferred from Figure 5.8 where the lowest value of RMSE-SS is shown for September. Fall months (Nov-Dec) show the best improvement for the GCM forecast in Tar River for both streamflow and loadings (Figure 5.8). Ogeechee River shows peaks of RMSE-SS in the summer months for both streamflow and loadings. The January RMSE-SS in Figure 5.9 has a value of 1, because the R-RMSE lines overlap each other in Figure 5.6, which implies that the R-RMSE from the GCM forecast is actually lower than that of the perfect forecast. There is uniformly high RMSE-SS across the months for Escambia River except for the lower performance in the December RMSE-SS (Figure 5.10).

Table 5.5. Root mean squared error skill scores (RMSE-SS) from the entire forecasting period, 1961-2010, for both streamflow and loadings for watersheds across the Southeast.

	RMSE-SS (full period)		
	Tar River	Ogeechee River	Escambia River
Streamflow	0.37	0.57	0.49
Loadings	0.34	0.36	0.44

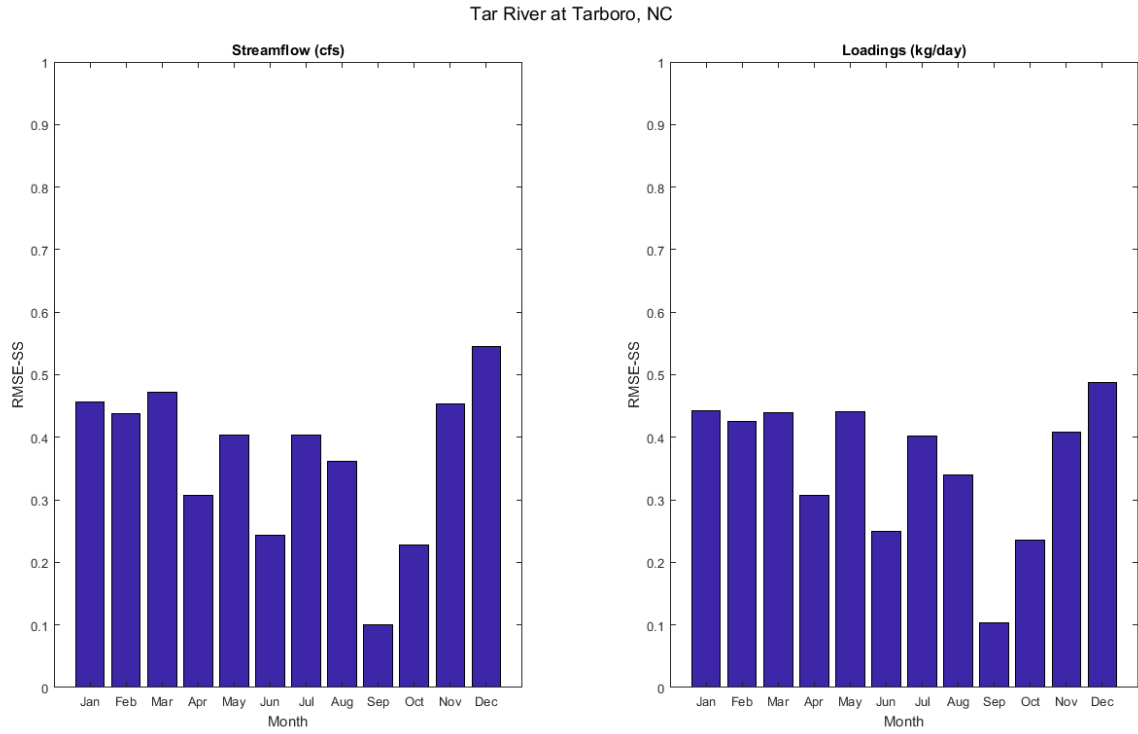


Figure 5.8. RMSE-SS for streamflow and loadings for each month for Tar River at Tarboro, NC.

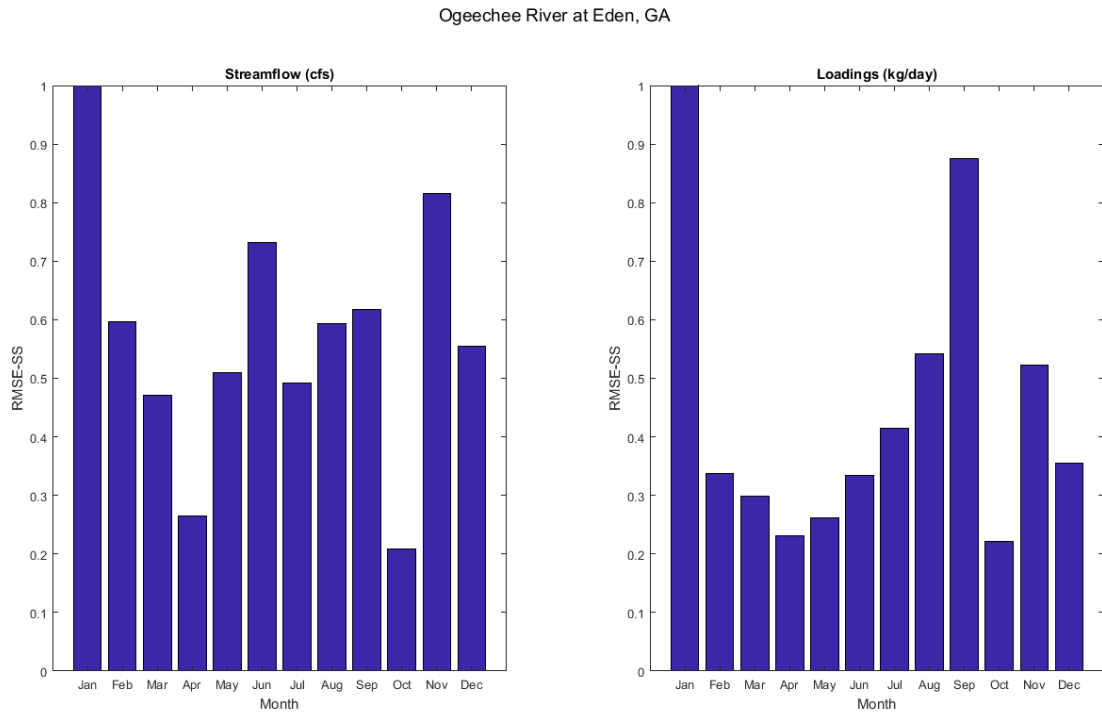


Figure 5.9. RMSE-SS for streamflow and loadings for each month for Ogeechee River near Eden, GA.

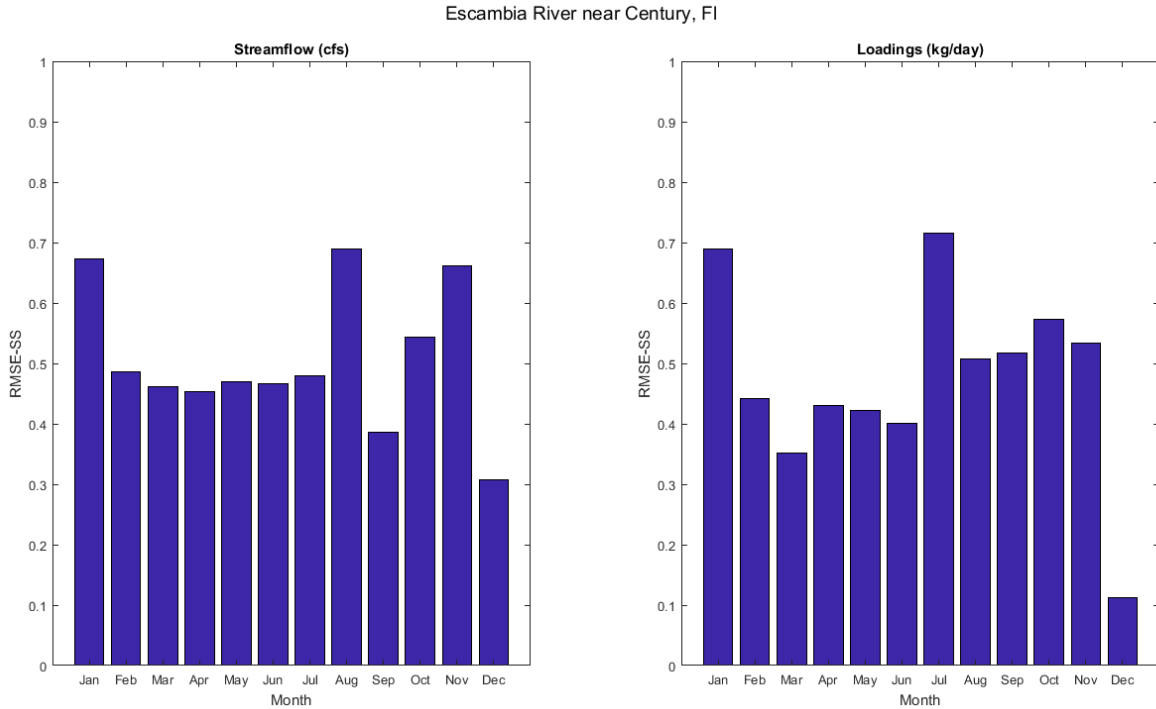


Figure 5.10. RMSE-SS for streamflow and loadings for each month for Escambia River near Century, FL.

The ROC curve of all three forecasts (perfect, GCM, and climatological) for both streamflow and loadings for Tar River (Figure 5.8), Ogeechee River (Figure 5.9), and Escambia River (Figure 5.10) show the ability of forecasts in predicting categorical events. In general, the perfect forecasts (green line) have higher true positive rates than the climatological forecasts (blue line) and GCM forecasts (red line). The below normal (BN) category for Tar River (Figure 5.8) has a higher true positive rate and a higher false alarm rate bringing the blue line above that of the GCM and perfect forecast. Although the BN (labeled 0.33 in Figure 5.8) category appears to be better than the GCM forecast, the majority of the climatological forecast curve is closer to the 1-1 line indicating their poor performance. Ideally, the area under the curve helps to quantify the overall score off the ROC curve.

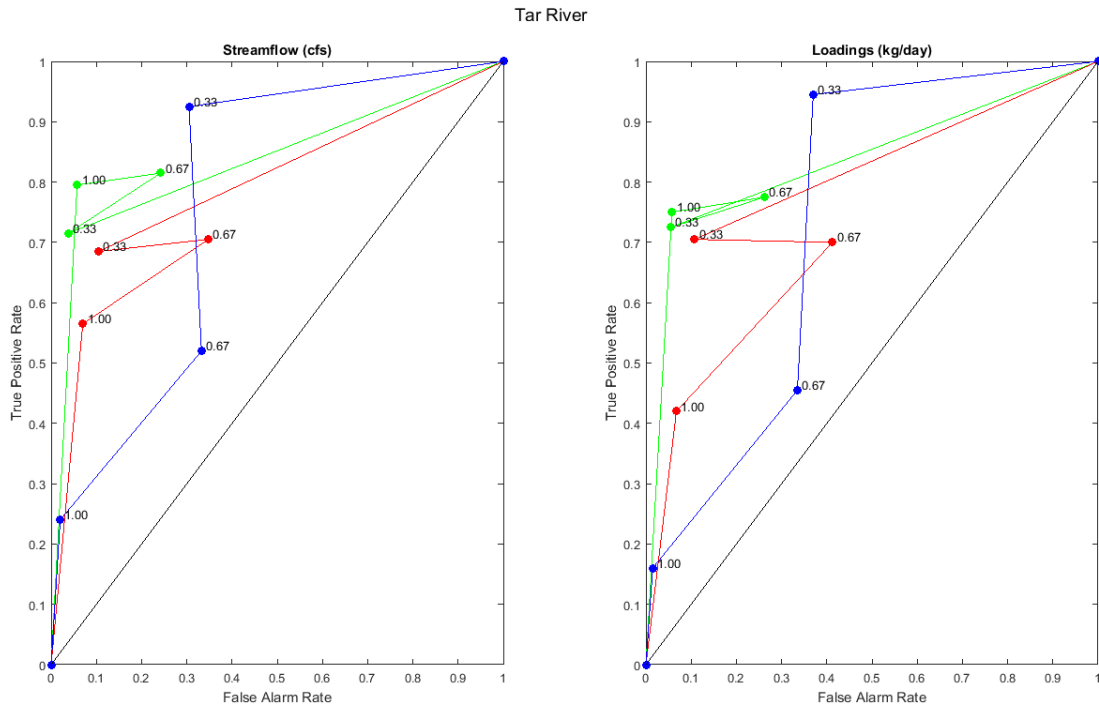


Figure 5.8. ROC curves for bias-corrected streamflow and loadings binned into the 0.33 quantile (below normal), 0.67 quantile (normal), and 1.0 quantile (above normal) for climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Tar River at Tarboro, NC

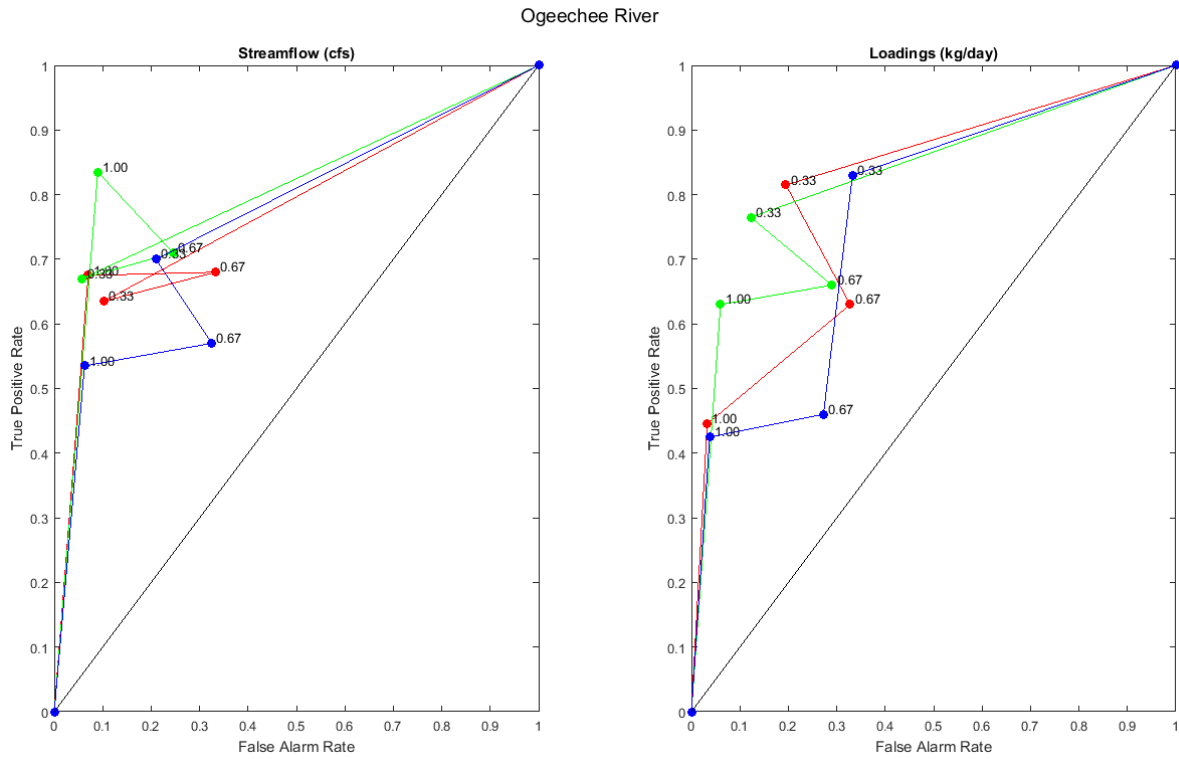


Figure 5.9. ROC curves for bias-corrected streamflow and loadings binned into the 0.33 quantile (below normal), 0.67 quantile (normal), and 1.0 quantile (above normal) for climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Ogeechee River near Eden, GA.

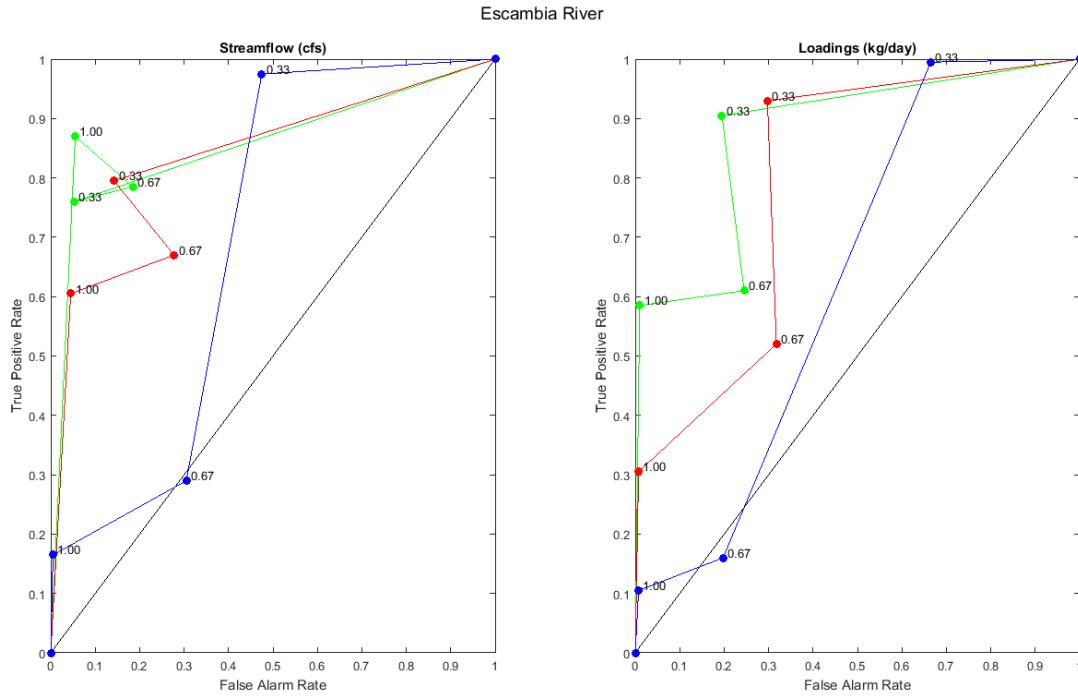


Figure 5.10. ROC curves for bias-corrected streamflow and loadings binned into the 0.33 quantile (below normal), 0.67 quantile (normal), and 1.0 quantile (above normal) for climatological forecasts (blue line), GCM forecasts (red line), and perfect forecasts (green line) for Escambia River near Century, FL.

5.5. Discussion and Conclusion

Using climate information from GCMs along with updating initial conditions has produced skillful streamflow forecasts. This study has shown that using retrospective climate forecasts in the SWAT model produces reasonable skill in forecasting total nitrogen loads across the Southeastern U.S. The advantages of using a mechanistic model for forecasting purposes is that it simultaneously estimates streamflow and nutrient load based on climate forcings; a process not possible using load estimating regression models. Adapting the SWAT model for forecasting using climate information allows forecasts to be updated with initial conditions on a monthly basis. Using the CCA bias-correction allows for simultaneous reduction of any bias introduced using the looping SWAT model. Bias reduction is done only using historical

information to provide a technique that could be useful for water managers who are interesting in using hydrological models for forecasting. Only retrospective forecasts are used from ECHAM4.5, although the same GCM has available real-time forecasts which could be used to develop 1-month to 2-month ahead forecasts of total nitrogen loads to forecast loads into impaired water bodies. There have been studies that look at the degradation of skill in streamflow forecasting with increasing forecast lag. Forecasting in this study has only been done on a 1-month ahead basis but it would be useful to see how much the forecasting skill drops when extending forecasting to the 2-month ahead and seasonal timescale. This would be potentially useful for managers in making watershed rules in advance to plan for upcoming wet or dry seasons, where nutrient loads might change drastically.

Chapter 6. Future Work and Conclusion

6.1. Future Work

Extending the SWAT forecasting framework to include nutrient management operations can help identify watershed rules that effectively reduce the downstream nutrient loading of streams. Watersheds considered in Chapter 5 are relatively pristine basins and do not include large influences from anthropogenic sources, so management operations implemented would focus on reducing the amount of nutrients in overland run-off. The sensitivity of upstream fertilizer reduction rules on downstream nutrient loads would be especially useful for managers in choosing watershed rules to reach upcoming target reductions (e.g. 20% reduction by May 2018).

The Escambia River near Century, FL watershed was chosen out of the three watersheds from Chapter 5, based on the uniform high performance in forecasting streamflow and total nitrogen load over the evaluation period. A map showing the spatial variability of land-use type over the Escambia watershed is shown in Figure 6.1. Given that this watershed is relatively pristine we see from Figure 6.1 that the watershed is mostly forested areas (FRSE and FRSD), with small amounts of agricultural land (AGRR). Further, we see from Figure 6.2 that only 14% of the watershed area is purposed for agricultural. The SWAT model has management operations already built in and can be applied specifically to land-use types.

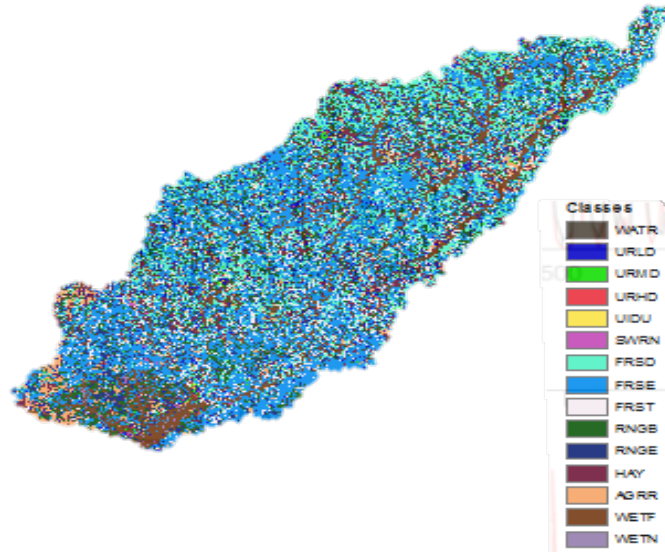


Figure 6.1. Map showing the land-use types of the Escambia River near Century, FL watershed. Agricultural row-crops (AGRR) are shaded light brown.

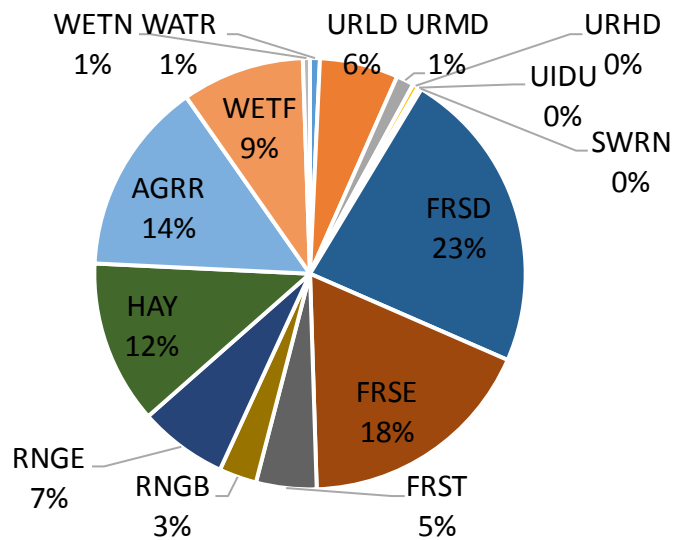


Figure 6.2. Percentage of land cover purposed for each land type in the Escambia River watershed.

The SWAT model was a built-in management operation tool that allows for the regulation of fertilizer application over certain land-use types. Scenarios (Table 6.1) applying different amounts of elemental nitrogen over the agricultural purposed lands were applied using the 1-month ahead forecasting SWAT framework from Chapter 5. Observed precipitation and temperature were forced in the calibrated SWAT model and a post-simulation CCA bias-

correction was applied. Fertilizer application occurred only on growing days which was triggered when a default heat index threshold was reached. Incremental reduction of fertilizer in steps of 100 kg/ha starting at 500 kg/ha were run to see the sensitivity in downstream nitrogen reduction. Results from Scenario “Manage9” and “Manage13” are shown respectively in Figures 6.3 and 6.4. Mean monthly forecasted streamflow and total nitrogen load is shown for a high flow year (1998) and low flow year (2010) in Figures 6.3 and 6.4. Differences in mean nutrient are not observable in these plots, suggesting that the changes in management operations had little effect on reducing the downstream nutrient loadings. As a proof of concept, the management operation was extended to all land-use types and significant changes in magnitude for downstream nutrient load were observed.

It is possible that since the operations were applied to pristine basins that downstream effects were not observable given the small amount of agricultural land present. Extending these operation experiments to the other two basins (Tar River and Ogeechee River) could provide interesting results to whether using the SWAT management could be useful for water managers wanted to forecast nutrient loadings. Additionally, other management operations could be tested such as the scheduling of BMPs implementation to upstream areas.

Table 6.1. Scenarios for application of elemental nitrogen fertilizer

Scenario Name:	Amount of fertilizer applied to HRU (kg/ha)
Manage9	500
Manage10	400
Manage11	300
Manage12	200
Manage13	100

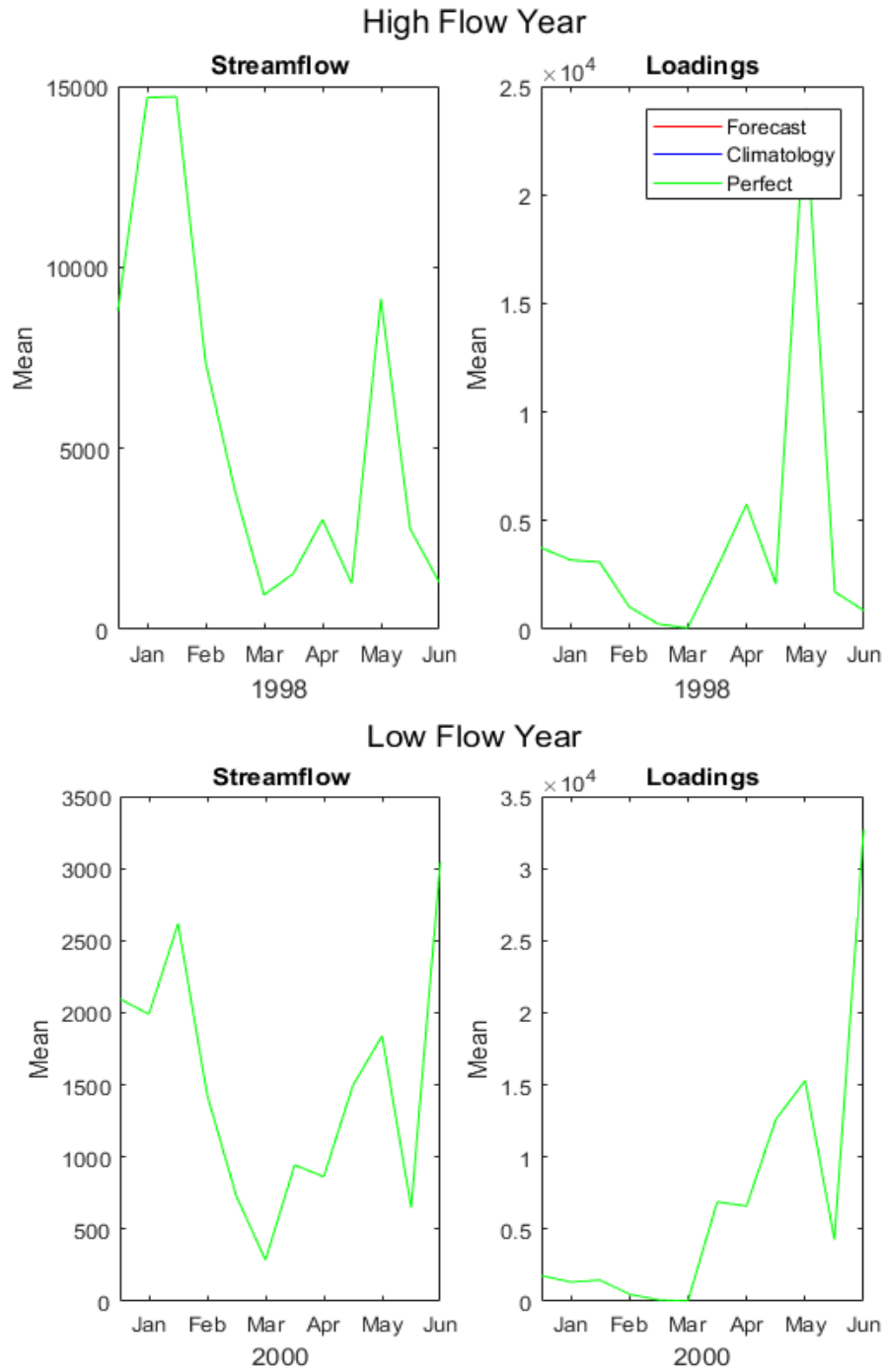


Figure 6.3. Forecasted mean streamflow and total nitrogen load from a high flow year (1998) and a low flow year (2000) for Escambia River near Century, FL when applying 500 kg/ha of elemental nitrogen to agricultural lands.

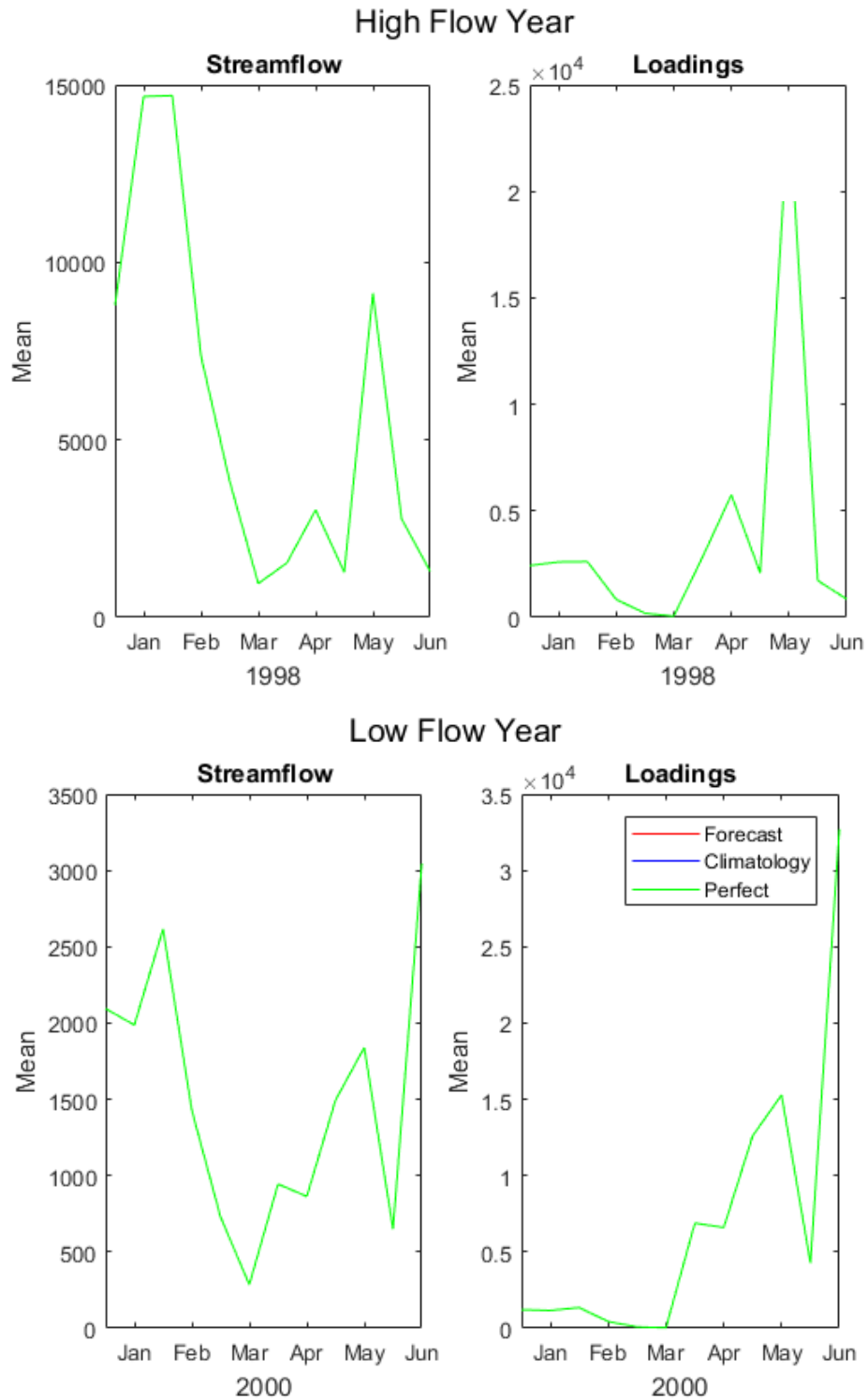


Figure 6.4. Forecasted mean streamflow and total nitrogen load from a high flow year (1998) and a low flow year (2000) for Escambia River near Century, FL when applying 100 kg/ha of elemental nitrogen to agricultural lands.

6.2. Conclusions

The main goal of this dissertation was to reduce the uncertainty in water quality estimates by improving methods for predicting and forecasting nutrient constituents. Currently long-term monitoring programs for nutrient constituents only exist for already impaired or near impaired streams. Using more accurate methods for estimating in-stream nutrient constituents using available data can help improve the health and quality of the stream and downstream systems. In Chapter 2 of this dissertation, a quantitative comparison of two water quality prediction models, the LOADEST and WRTDS models, was done across the Southeastern US. This study quantified the amount of observed variability captured in long-term estimates from both models. Although these prediction models do well in predicting the observed monthly to seasonal means of nutrient concentrations they lack skill in predicting the long-term variability. This chapter shows that for 18 stations from the SEUS that the WRTDS model in general does better in predicting the observed variability of concentration while the LOADEST model does better in predicting load. Further, we show that using LOADEST load estimates and dividing by observed streamflow can provide the better estimates for observed concentration. Quantifying this skill is of particular interest when using concentration or load estimates for long-term calibration or validation of water quality models, which is becoming more widespread in the modeling community.

In Chapter 3 of this dissertation we provide a free non-parametric software for testing if concentration estimates from the two models (LOADEST and WRTDS) are statistically different than models that just predict the observed mean. The framework uses a non-parametric resampling technique to create null distributions instead of assuming the distribution of the performance statistic. Data availability can change the shape and center of the underlying

performance metric distribution. Since the stations from the SEUS have such sparse data sets, we feel that this toolkit provides a more accurate way of testing significance opposed to a simple F-test. Twelve out of 18 having low performance (<0.4 NSE) show that the model performance is significantly different than the mean. This toolkit can inform state programs which monitoring locations need more frequent sampling to improve estimates from the LOADEST and WRTDS models.

Mechanistic models, like the SWAT model, can introduce systematic bias caused by imprecise calibration of parameters or inaccurate representations of watershed characteristics (e.g. storage). Removal of systematic bias from mechanistic models is important for identifying the skill attributable to the climate forcings. Calibration of three watersheds from the SEUS was done using load estimates from LOADEST as a surrogate truth. Canonical correlation analysis (CCA) was used in a new way for simultaneously removing of bias in streamflow and nitrogen loads. Observed cross-correlation and joint-likelihood between streamflow and loadings were preserved in estimates which is paramount in water quality studies. Ultimately, this preserves the relationship between loadings and concentration which can be lost when using other bias-correction techniques ordinary least-squares regression.

Nutrient load forecasts from the SWAT model can provide watershed managers with insight to compliance of future conditions to the Clean Water Act. Climate forecasts of precipitation and temperature from the ECHAM4.5 global circulation models can provide meaningful skill in streamflow and nutrient load forecasts. Extending this framework to the 2-month, 3-month, and seasonal forecasting scheme can provide managers the threshold for meaningful skill. This information can provide watershed managers the timeline for rule implementation. We hope that these results and contributions from this dissertation improve the

prediction and use of water quality estimates for remediating and preserving the health of streams and downstream water bodies.

REFERENCES

- Abatzoglou, J.T., Brown, T.J., 2012. A comparison of statistical downscaling methods suited for wildfire applications. *Int. J. Climatol.* 32, 772–780. doi:10.1002/joc.2312
- Ahl, R.S., Woods, S.W., Zuuring, H.R., 2008. Hydrologic Calibration and Validation of SWAT in a Snow-Dominated Rocky Mountain Watershed, Montana, U.S.A. 1. *JAWRA J. Am. Water Resour. Assoc.* 44, 1411–1430. doi:10.1111/j.1752-1688.2008.00233.x
- Akaike, H., 1981. Likelihood of a model and information criteria. *J. Econom.* 16, 3–14. doi:10.1016/0304-4076(81)90071-3
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi:10.1109/TAC.1974.1100705
- Alansi, A.W., Amin, M.S.M., Abdul Halim, G., Shafri, H.Z.M., Aimrun, W., 2009. Validation of SWAT model for stream flow simulation and forecasting in Upper Bernam humid tropical river basin, Malaysia. *Hydrol. Earth Syst. Sci. Discuss.* 6, 7581–7609. doi:10.5194/hessd-6-7581-2009
- Alexander, R.B., Slack, J.R., Ludtke, A.S., Fitzgerald, K.K., Schertz, T.L., 1998. Data from selected U.S. Geological Survey National Stream Water Quality Monitoring Networks. *Water Resour. Res.* 34, 2401–2405. doi:10.1029/98WR01530
- Amatya, D.M., Jha, M.K., Williams, T.M., Edwards, A.E., Hitchcock, D.R., 2013. SWAT Model Prediction of Phosphorus Loading in a South Carolina Karst Watershed with a Downstream Embayment. *J. Environ. Prot. (Irvine., Calif.)* 04, 75–90. doi:10.4236/jep.2013.47A010
- Arheimer, B., Löwgren, M., Pers, B.C., Rosberg, J., 2005. Integrated Catchment Modeling for Nutrient Reduction: Scenarios Showing Impacts, Potential, and Cost of Measures. *AMBIO A J. Hum. Environ.* 34, 513–520. doi:10.1579/0044-7447-34.7.513

- Arnold, J.G., Fohrer, N., 2005. SWAT2000: current capabilities and research opportunities in applied watershed modelling. *Hydrol. Process.* 19, 563–572. doi:10.1002/hyp.5611
- Aulenbach, B.T., 2013. Improving regression-model-based streamwater constituent load estimates derived from serially correlated data. *J. Hydrol.* doi:10.1016/j.jhydrol.2013.09.001
- Barnett, T.P., Preisendorfer, R., 1987. Origins and Levels of Monthly and Seasonal Forecast Skill for United States Surface Air Temperatures Determined by Canonical Correlation Analysis. *Mon. Weather Rev.* 115, 1825–1850. doi:10.1175/1520-0493(1987)115<1825:OALOMA>2.0.CO;2
- Bosch, N.S., Allan, J.D., Selegean, J.P., Scavia, D., 2013. Scenario-testing of agricultural best management practices in Lake Erie watersheds. *J. Great Lakes Res.* 39, 429–436. doi:10.1016/j.jglr.2013.06.004
- Bowers, M.C., Tung, W.W., Gao, J.B., 2012. On the distributions of seasonal river flows: Lognormal or power law? *Water Resour. Res.* 48, 1–12. doi:10.1029/2011WR011308
- Cohn, T.A., 2005. Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data. *Water Resour. Res.* 41, 1–13. doi:10.1029/2004WR003833
- Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., Summers, R.M., 1992. The validity of a simple statistical model for estimating fluvial constituent loads: An Empirical study involving nutrient loads entering Chesapeake Bay. *Water Resour. Res.* 28, 2353–2363. doi:10.1029/92WR01008
- Cohn, T.A., Delong, L.L., Gilroy, E.J., Hirsch, R.M., Wells, D.K., 1989. Estimating constituent loads. *Water Resour. Res.* 25, 937–942. doi:10.1029/WR025i005p00937
- D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, T. L. Veith, 2007.

- Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Trans. ASABE* 50, 885–900. doi:10.13031/2013.23153
- Das Bhowmik, R., Sankarasubramanian, A., Sinha, T., Patskoski, J., Mahinthakumar, G., Kunkel, K.E., 2017. Multivariate Downscaling Approach Preserving Cross-Correlations across Climate Variables for Projecting Hydrologic Fluxes. *J. Hydrometeorol.* JHM-D-16-0160.1. doi:10.1175/JHM-D-16-0160.1
- Douglas-Mankin, K.R., Srinivasan, R., Arnold, J.G., 2010. Soil and Water Assessment Tool (SWAT) model: Current developments and applications. *Trans. ASABE* 53, 1423–1431. doi:10.13031/2013.34915
- Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., Liebert, J., 2012. *HESS Opinions* “Should we apply bias correction to global and regional climate model data?” *Hydrol. Earth Syst. Sci.* 16, 3391–3404. doi:10.5194/hess-16-3391-2012
- Fang, G.H., Yang, J., Chen, Y.N., Zammit, C., 2015. Comparing bias correction methods in downscaling meteorological variables for a hydrologic impact study in an arid area in China. *Hydrol. Earth Syst. Sci.* 19, 2547–2559. doi:10.5194/hess-19-2547-2015
- Fisher, R.A., 1992. Statistical methods for research workers, in: *Breakthroughs in Statistics*. Springer, pp. 66–70.
- Garen, D.C., 1992. Improved Techniques in Regression-Based Streamflow Volume Forecasting. *J. Water Resour. Plan. Manag.* 118, 654–670. doi:10.1061/(ASCE)0733-9496(1992)118:6(654)
- Gronewold, A.D., Qian, S.S., Wolpert, R.L., Reckhow, K.H., 2009. Calibrating and validating bacterial water quality models: A Bayesian approach. *Water Res.* 43, 2688–2698. doi:10.1016/j.watres.2009.02.034

- Herman, B.D., Eberly, J.O., Jung, C.M., Medina, V.F., 2017. Review and Evaluation of Reservoir Management Strategies for Harmful Algal Blooms Environmental Laboratory 32.
- Hidalgo, H.G., D., D.M., Cayan, D.R., 2008. Downscaling with constructed analogues: Daily precipitation and temperature fields over the United States, California Climate Change. doi:CEC-500-2007-123
- Hirsch, R.M., 2014. Large Biases in Regression-Based Constituent Flux Estimates: Causes and Diagnostic Tools. JAWRA J. Am. Water Resour. Assoc. 50, 1401–1424. doi:10.1111/jawr.12195
- Hirsch, R.M., Archfield, S.A., De Cicco, L.A., 2015. A bootstrap method for estimating uncertainty of water quality trends. Environ. Model. Softw. 73, 148–166. doi:10.1016/j.envsoft.2015.07.017
- Hirsch, R.M., Moyer, D.L., Archfield, S.A., 2010. Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs1. JAWRA J. Am. Water Resour. Assoc. 46, 857–880. doi:10.1111/j.1752-1688.2010.00482.x
- Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., Megown, K., 2015. Completion of the 2011 National Land Cover Database for the Conterminous United States – Representing a Decade of Land Cover Change Information. Photogramm. Eng. Remote Sensing 81, 345–354. doi:10.14358/PERS.81.5.345
- Jain, S.K., Sudheer, K.P., 2008. Fitting of Hydrologic Models: A Close Look at the Nash–Sutcliffe Index. J. Hydrol. Eng. 13, 981–986. doi:10.1061/(ASCE)1084-0699(2008)13:10(981)
- Jha, M.K., Gassman, P.W., Arnold, J.G., 2007. Water quality modeling for the Raccoon River

- watershed using SWAT. *Trans. ASABE* 50, 479–493.
- Jha, M.K., Schilling, K.E., Gassman, P.W., Wolter, C.F., 2010. Targeting land-use change for nitrate- nitrogen load reductions in an agricultural watershed 65, 342–352.
doi:10.2489/jswc.65.6.342
- Kim, D., Kaluarachchi, J., 2014. Predicting streamflows in snowmelt-driven watersheds using the flow duration curve method. *Hydrol. Earth Syst. Sci.* 18, 1679–1693. doi:10.5194/hess-18-1679-2014
- Kim, J., Engel, B.A., Park, Y.S., Theller, L., Chaubey, I., Kong, D.S., Lim, K.J., 2012. Development of Web-based Load Duration Curve system for analysis of total maximum daily load and water quality characteristics in a waterbody. *J. Environ. Manage.* 97, 46–55.
doi:10.1016/j.jenvman.2011.11.012
- Koch, R.W., Smillie, G.M., 1986. Bias in hydrological prediction using log-transformed regression models. *Water Resour. Bull.* 22, 717–723. doi:http://doi.org/10.1175/JCLI-D-12-00821.1
- Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.* 5, 89–97. doi:10.5194/adgeo-5-89-2005
- Legates, D.R., McCabe, G.J., 1999. Evaluating the use of “goodness-of-fit” Measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* 35, 233–241.
doi:10.1029/1998WR900018
- Leisenring, M., Moradkhani, H., 2012. Analyzing the uncertainty of suspended sediment load prediction using sequential data assimilation. *J. Hydrol.* 468–469, 268–282.
doi:10.1016/j.jhydrol.2012.08.049
- Li, S., Goddard, L., 2005. Retrospective Forecasts with the ECHAM4.5 AGCMI. IRI Tech. Rep.

05-02.

- Libera, D.A., Sankarasubramanian, A., 2018. Multivariate bias corrections of mechanistic water quality model predictions. *J. Hydrol.* 564, 529–541. doi:10.1016/j.jhydrol.2018.07.043
- Likeš, J., 1980. Variance of the MVUE for Lognormal Variance. *Technometrics* 22, 253. doi:10.2307/1268465
- Livneh, B., Rosenberg, E.A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K.M., Maurer, E.P., Lettenmaier, D.P., 2013. A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States: Update and Extensions*. *J. Clim.* 26, 9384–9392. doi:10.1175
- Maraun, D., 2016. Bias Correcting Climate Change Simulations - a Critical Review. *Curr. Clim. Chang. Reports* 2, 211–220. doi:10.1007/s40641-016-0050-x
- Marhaento, H., Booij, M.J., Rientjes, T.H.M., Hoekstra, A.Y., 2017. Attribution of changes in the water balance of a tropical catchment to land use change using the SWAT model. *Hydrol. Process.* 31, 2029–2040. doi:10.1002/hyp.11167
- Mazrooei, A., Sinha, T., Sankarasubramanian, A., Kumar, S., Peters-Lidard, C.D., 2015. Decomposition of sources of errors in seasonal streamflow forecasting over the U.S. Sunbelt. *J. Geophys. Res. Atmos.* 120, 11,809-11,825. doi:10.1002/2015JD023687
- McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe Efficiency Index. *J. Hydrol. Eng.* 11, 597–602. doi:10.1061/(ASCE)1084-0699(2006)11:6(597)
- Moyer, D., Hirsch, R., Hyer, K., 2012. Comparison of two regression-based approaches for determining nutrient and sediment fluxes and trends in the Chesapeake Bay watershed: U.S. Geological Survey Scientific Investigations Report 2012-5244.
- Mueller, D.K., Spahr, N.E., 2006. Nutrients in Streams and Rivers Across the Nation — 1992-

2001. Usgs 44.
- Nash, J.E., Sutcliffe, J. V, 1970. River Flow Forecasting Through Conceptual Models Part I-a Discussion of Principles*. *J. Hydrol.* 10, 282–290. doi:10.1016/0022-1694(70)90255-6
- Oh, J., Sankarasubramanian, a., 2012. Interannual hydroclimatic variability and its influence on winter nutrient loadings over the Southeast United States. *Hydrol. Earth Syst. Sci.* 16, 2285–2298. doi:10.5194/hess-16-2285-2012
- P. W. Gassman, M. R. Reyes, C. H. Green, J. G. Arnold, 2007. The Soil and Water Assessment Tool: Historical Development, Applications, and Future Research Directions. *Trans. ASABE* 50, 1211–1250. doi:10.13031/2013.23637
- Park, Y.S., Engel, B.A., 2015. Analysis for Regression Model Behavior by Sampling Strategy for Annual Pollutant Load Estimation. *J. Environ. Qual.* 44, 1843. doi:10.2134/jeq2015.03.0137
- Prairie, J., Rajagopalan, B., Lall, U., Fulp, T., 2007. A stochastic nonparametric technique for space-time disaggregation of streamflows. *Water Resour. Res.* 43, n/a-n/a. doi:10.1029/2005WR004721
- Puckett, L.J., 1994. Nonpoint and Point Sources of Nitrogen in Major Watersheds of the United States. Director 1–8.
- Qian, S.S., Reckhow, K.H., Zhai, J., McMahon, G., 2005. Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach. *Water Resour. Res.* 41, 1–10. doi:10.1029/2005WR003986
- Quilbé, R., Rousseau, A.N., Duchemin, M., Poulin, A., Gangbazo, G., Villeneuve, J.-P., 2006. Selecting a calculation method to estimate sediment and nutrient loads in streams: Application to the Beaurivage River (Québec, Canada). *J. Hydrol.* 326, 295–310.

doi:10.1016/j.jhydrol.2005.11.008

R. Srinivasan, X. Zhang, J. Arnold, 2010. SWAT Ungauged: Hydrological Budget and Crop Yield Predictions in the Upper Mississippi River Basin. *Trans. ASABE* 53, 1533–1546.

doi:10.13031/2013.34903

Rao, A.S., Marshall, S., Gubbi, J., Palaniswami, M., Sinnott, R., Pettigrovet, V., 2013. Design of low-cost autonomous water quality monitoring system, in: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, pp. 14–19.

doi:10.1109/ICACCI.2013.6637139

Reichert, P., Mieleitner, J., 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resour. Res.* 45, 1–19.

doi:10.1029/2009WR007814

Runkel, R.L., Crawford, C.G., Cohn, T. a, 2004. Load Estimator (LOADEST): A FORTRAN program for estimating constituent loads in streams and rivers. *Tech. Methods. U.S. Geol. Surv. U.S. Dep. Inter.* 4, 69.

Santhi, C., Arnold, J.G., Williams, J.R., Dugas, W. a., Srinivasan, R., Hauck, L.M., 2001.

VALIDATION OF THE SWAT MODEL ON A LARGE RWER BASIN WITH POINT AND NONPOINT SOURCES. *J. Am. Water Resour. Assoc.* 37, 1169–1188.

doi:10.1111/j.1752-1688.2001.tb03630.x

Schaefli, B., Gupta, H. V., 2007. Do Nash values have value? *Hydrol. Process.* 21, 2075–2080.

doi:10.1002/hyp.6825

Seelig, B., 2000. Diffuse Sources of Nitrogen Related to Water Quality Protection in the Northern Great Plains.

Serpa, D., Nunes, J.P., Keizer, J.J., Abrantes, N., 2017. Impacts of climate and land use changes

- on the water quality of a small Mediterranean catchment with intensive viticulture. *Environ. Pollut.* 224, 454–465. doi:10.1016/j.envpol.2017.02.026
- Shrestha, S., Kazama, F., Newham, L.T.H., 2008. A framework for estimating pollutant export coefficients from long-term in-stream water quality monitoring data. *Environ. Model. Softw.* 23, 182–194. doi:10.1016/j.envsoft.2007.05.006
- Sinha, T., Sankarasubramanian, A., 2013. Role of climate forecasts and initial conditions in developing streamflow and soil moisture forecasts in a rainfall–runoff regime. *Hydrol. Earth Syst. Sci.* 17, 721–733. doi:10.5194/hess-17-721-2013
- Sinha, T., Sankarasubramanian, A., Mazrooei, A., 2014. Decomposition of Sources of Errors in Monthly to Seasonal Streamflow Forecasts in a Rainfall–Runoff Regime. *J. Hydrometeorol.* 15, 2470–2483. doi:10.1175/JHM-D-13-0155.1
- Slack, J.R., Landwehr, J.M., Lumb, A.M., 1993. Hydroclimatic data network (HCDN): A U.S. Geological Survey streamflow data set for the United States for the study of climate variation, 1874–1988.
- Smith, R. a., Schwarz, G.E., Alexander, R.B., 1997. Regional interpretation of water-quality monitoring data. *Water Resour. Res.* 33, 2781. doi:10.1029/97WR02171
- Stewart, T.R., Reagan-Cirincione, P., 1991. Coefficients for debiasing forecasts. *Am. Meteorol. Soc.* 119, 2047–2051.
- Stow, C. a., Roessler, C., Borsuk, M.E., Bowen, J.D., Reckhow, K.H., 2003. Comparison of Estuarine Water Quality Models for Total Maximum Daily Load Development in Neuse River Estuary. *J. Water Resour. Plan. Manag.* 129, 307–314.
- Sun, L., Nistor, I., Seidou, O., 2015. Streamflow data assimilation in SWAT model using Extended Kalman Filter. *J. Hydrol.* 531, 671–684. doi:10.1016/j.jhydro1.2015.10.060

- Tippett, M.K., Barlow, M., Lyon, B., 2003. Statistical correction of central Southwest Asia winter precipitation simulations. *Int. J. Climatol.* 23, 1421–1433. doi:10.1002/joc.947
- Tong, S.T.Y., Liu, A.J., Goodrich, J. a., 2009. Assessing the water quality impacts of future land-use changes in an urbanising watershed. *Civ. Eng. Environ. Syst.* 26, 3–18. doi:10.1080/10286600802003393
- US EPA, 1998. National Strategy for the Development of Regional Nutrient Criteria. United States Environ. Prot. Agency, Off. Water 822-R-98–0, 53.
- USEPA, 2012. Appendix A : Nitrogen Deposition from the Atmosphere to the Earth ' s Surface. React. Nitrogen United States An Anal. Inputs, Flows, Consequences, Manag. Options.
- USGS, U.S.G.S., 2009. National Elevation Dataset (NED).
- Van Buren, M.A., Watt, W.E., Marsalek, J., 1997. Application of the log-normal and normal distributions to stormwater quality parameters. *Water Res.* 31, 95–104. doi:10.1016/S0043-1354(96)00246-1
- Vannitsem, S., 2011. Bias correction and post-processing under climate change. *Nonlinear Process. Geophys.* 18, 911–924. doi:10.5194/npg-18-911-2011
- Vogel, R.M., 1986. The Probability Plot Correlation Coefficient Test for the Normal, Lognormal, and Gumbel Distributional Hypotheses. *Water Resour. Res.* 22, 587–590. doi:10.1029/WR022i004p00587
- Vogel, R.M., Sankarasubramanian, A., 2005. Monthly Climate Data for Selected USGS HCDN Sites, 1951-1990. doi:10.3334/ormlaac/810
- Vogel, R.M., Shallcross, A.L., 1996. The moving blocks bootstrap versus parametric time series models. *Water Resour. Res.* 32, 1875–1882. doi:10.1029/96WR00928
- Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of

- input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resour. Res.* 44, 1–15. doi:10.1029/2007WR006720
- Wang, G., Yang, H., Wang, L., Xu, Z., Xue, B., 2014. Using the SWAT model to assess impacts of land use changes on runoff generation in headwaters. *Hydrol. Process.* 28, 1032–1042. doi:10.1002/hyp.9645
- Wilks, D.S., 2000. Diagnostic verification of the Climate Prediction Center long-lead outlooks, 1995–98. *J. Clim.* 13, 2389–2403. doi:10.1175/1520-0442(2000)013<2389:DVOTCP>2.0.CO;2
- Wilks, D.S., 1995. *Statistical Methods in the Atmospheric Sciences*: Academic Press, New York.
- Willmott, C.J., 1984. On the Evaluation of Model Performance in Physical Geography, in: Gaile, G.L., Willmott, C.J. (Eds.), *Spatial Statistics and Models*. Springer Netherlands, Dordrecht, pp. 443–460. doi:10.1007/978-94-017-3048-8_23
- Windolf, J., Thodsen, H., Troldborg, L., Larsen, S.E., Bøgestrand, J., Ovesen, N.B., Kronvang, B., 2011. A distributed modelling system for simulation of monthly runoff and nitrogen sources, loads and sinks for ungauged catchments in Denmark. *J. Environ. Monit.* 13, 2645. doi:10.1039/c1em10139k
- Wood, A.W., Leung, L.R., Sridhar, V., Lettenmaier, D.P., 2004. Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs. *Clim. Change* 62, 189–216. doi:10.1023/B:CLIM.0000013685.99609.9e