A Framework For Estimating Nutrient And Sediment Loads That Leverages The Temporal Variability Embedded In Water Monitoring Data

Baxter G. Miatke

University of Vermont

Follow this and additional works at: https://scholarworks.uvm.edu/graddis

Part of the Environmental Engineering Commons, and the Natural Resources Management and Policy Commons

Recommended Citation

https://scholarworks.uvm.edu/graddis/651

This Thesis is brought to you for free and open access by the Dissertations and Theses at ScholarWorks @ UVM. It has been accepted for inclusion in Graduate College Dissertations and Theses by an authorized administrator of ScholarWorks @ UVM. For more information, please contact donna.omalley@uvm.edu.
A FRAMEWORK FOR ESTIMATING NUTRIENT AND SEDIMENT LOADS THAT LEVERAGES THE TEMPORAL VARIABILITY EMBEDDED IN WATER MONITORING DATA

A Thesis Presented

by

Baxter G. Miatke

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Master of Science
Specializing in Civil and Environmental Engineering

October, 2016

Defense Date: July 27th, 2016
Thesis Examination Committee:

Donna Rizzo, Ph.D., Advisor
Andrew Schroth, Ph.D., Chairperson
Mandar Dewoolkar, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College
ABSTRACT

Rivers deliver significant macronutrients and sediments to lakes that can vary substantially throughout the year. These nutrient and sediment loadings, exacerbated by winter and spring runoff, impact aquatic ecosystem productivity and drive the formation of harmful algae blooms. The source, extent and magnitude of nutrient and sediment loading can vary drastically due to extreme weather events and hydrologic processes, such as snowmelt or high flow storm events, that dominate during a particular time period, making the temporal component (i.e., time over which the loading is estimated) critical for accurate forecasts. In this work, we developed a data-driven framework that leverages the temporal variability embedded in these complex hydrologic regimes to improve loading estimates. Identifying the “correct” time scale is an important first step for providing accurate estimates of seasonal nutrient and sediment loadings. We use water quality concentration and associated 15-minute discharge data from nine watersheds in Vermont’s Lake Champlain Basin to test our proposed framework. Optimal time periods were selected using a hierarchical cluster analysis that uses the slope and intercept coefficients from individual load-discharge regressions to derive improved linear models. These optimized linear models were used to improve estimates of annual and “spring” loadings for total phosphorus, dissolved phosphorus, total nitrogen, and total suspended loads for each of the nine study watersheds. The optimized annual regression model performed ~20% better on average than traditional annual regression models in terms of Nash-Sutcliffe efficiency, and resulted in ~50% higher cumulative load estimates with the largest difference occurring in the “spring”. In addition, the largest nutrient and sediment loadings occurred during the “spring” unit of time and were typically more than 40% of the total annual estimated load in a given year. The framework developed here is robust and may be used to analyze other units of time associated with hydrologic regimes of interest provided adequate water quality data exist. This, in turn, may be used to create more targeted and cost-effective management strategies for improved aquatic health in rivers and lakes.
CITATIONS

Material from this thesis has been submitted for publication to Journal of Great Lakes Research on August 14, 2016 in the following form:

ACKNOWLEDGEMENTS

There are a number of people who deserve more than just an acknowledgement, but deep gratitude and thanks for all of their support, guidance, and knowledge. At the top of the list is my advisor in academics and life, Donna Rizzo. My deepest thanks for everything you do; pushing me to do the M.S. program, supporting me to be the best researcher/person possible, and the numerous life lessons and stories to get me through the process. Thank you to my committee members, Andrew Schroth, Kristen Underwood and Mandar Dewoolkar, who have guided me on my journey from summer intern to graduate researcher and provided invaluable knowledge and experience along the way. I would also like to thank Braden Rosenberg for assistance with spring sampling, analysis, and teaching me about rocks. Thanks to all of my fellow graduate students, faculty, and staff in engineering at UVM, past and present, who have been the sources of criticism, motivation, food, and mostly friendship during my time at the University of Vermont (Votey Hall). I would have not made it through without all of you.

This research was supported by NSF Grant EPS-1101317. I would also like to thank all researchers involved with VT-EPSCoR & RACC and the UVM College of Engineering and Mathematical Sciences who have guided me along my path. I also would like to acknowledge the VTDEC for the long term water quality data set and USGS for continuous flow records provided for this work.

I finally want to extend my deepest appreciation to my family for over 23 years of support and motivation to keep at it, as well as my adventure partner in life, Victoria Rude, for every bit of her support and love. Thank you all!!
# TABLE OF CONTENTS

CITATIONS................................................................................................................................. ii  
ACKNOWLEDGEMENTS ................................................................................................................ iii  
LIST OF TABLES ............................................................................................................................ vi  
LIST OF FIGURES .......................................................................................................................... vii  

CHAPTER 1: MULTIPLE FACTORS AFFECTING ESTIMATES OF RIVER NUTRIENT AND SEDIMENT LOADS IN A WARMING CLIMATE ................................. 1  
1.1 EUTROPHICATION AND MACRONUTRIENT DELIVERY ................................................... 2  
1.2 LAND USE/COVER AND SOIL TYPES ................................................................................ 4  
1.3 SEASONALITY AND A CHANGING CLIMATE ..................................................................... 5  
1.4 LOAD ESTIMATION MODELS AND METHODS .................................................................... 7  
1.5 OVERVIEW OF THE THESIS AND SPECIFIC OBJECTIVES ............................................. 9  
  1.5.1 Spring Study Site - Missisquoi Watershed ..................................................................... 11  
1.6 THESIS OUTLINE ................................................................................................................ 11  
1.7 REFERENCES ........................................................................................................................ 12  

CHAPTER 2: MANUSCRIPT FOR JOURNAL ............................................................................. 18  
2.1 ABSTRACT .............................................................................................................................. 19  
2.2 INTRODUCTION .................................................................................................................... 20  
2.3 METHODS ............................................................................................................................ 24  
2.5 RESULTS AND DISCUSSION ............................................................................................... 31  
  2.5.1 Total Phosphorus ............................................................................................................. 33  
  2.5.2 Dissolved Phosphorus ..................................................................................................... 33  
  2.5.3 Total Nitrogen ................................................................................................................ 34  
  2.5.4 Total Suspended Solids .................................................................................................. 34  
  2.5.5 Watershed Trends and Discussion .................................................................................. 34  
  2.5.6 “Spring” Trends and Model Metrics ............................................................................... 36  
  2.5.7 Framework Estimation of Annual Loads ....................................................................... 39  
2.6 CONCLUSIONS AND RECOMMENDATIONS .................................................................... 43  
2.7 ACKNOWLEDGEMENTS ....................................................................................................... 45
LIST OF TABLES

Table 2.1 - Watershed Characteristics ................................................................. 58

Table 2.2 - Period of record and source for TP, DP, TN, and TSS for all nine study watersheds ................................................................. 58

Table 2.3 - Optimal clusters indicated by number in parentheses, shading, and vertical dividers from January to December (J to D) for the nine study watersheds partitioned by constituent TP, DP, TN, and TSS with optimal and simple annual performance metrics Adjusted $R^2$, RMSE, NSE, and Bias ................................................................. 59

Table 2.4 - Simple, traditional, and optimal model performance metrics for spring months March, April and May; Adjusted $R^2$, RMSE, NSE, and Bias, partitioned by constituent TP, DP, TN, and TSS for each of the nine study watersheds ................................. 60
LIST OF FIGURES

Figure 2.1- Location of Watersheds Delineated to VT DEC Sampling Points. Figure created using data from VCGI: Vermont Center for Geographic Information. Available from http://vcgi.vermont.gov/opendata. *Basins excluded from this analysis due to limited water quality and discharge data.................................................................51

Figure 2.2- Annual linear regressions for log(Load Estimate) vs. log(15-Minute Discharge) for Missisquoi TP and Winooski TP. Examples show adjusted R², root mean square error (RMSE), and slope (b1) and intercept (b0) coefficients reported for each regression model. ........................................................................................................52

Figure 2.3-(a) Framework for Selecting Optimal Combinations using Hierarchical Cluster Analysis with Missisquoi TP Example Data: (b) Number of Samples in Each Month 1990-2015 (c) Slope Coefficients for Each Unit of Time, and (d) Absolute Value of Intercept Coefficients for Each Unit of Time ...............................................................53

Figure 2.4- Dendrogram of Monthly Clusters for Missisquoi TP Example from Figure 2.2 ........................................................................................................................................54

Figure 2.5- Traditional Annual Regressions, Traditional Spring Regressions, and Optimal Spring Regressions for log(Load Estimate) vs. log(Discharge) for Missisquoi TSS and TP Example with Adjusted R², RMSE, and slope and intercept coefficients reported for each regression. .........................................................................................55

Figure 2.6- Missisquoi TP Cumulative Loading in 2014 for Optimized vs. Simple Annual Models vs. 15-minute hydrograph ...............................................................................................................................56

Figure 2.7- Missisquoi TP, DP, TN, and TSS annual flux from water years 2000-2015 using optimal seasonal models............................................................................................................................57
CHAPTER 1: MULTIPLE FACTORS AFFECTING ESTIMATES OF RIVER NUTRIENT AND SEDIMENT LOADS IN A WARMING CLIMATE

Developing better nutrient and sediment load estimation models to capture the variability inherent in seasons as well as the complex interaction of hydrology and land cover on the flux of nutrients, pollutants, and other constituents from rivers to receiving water bodies is critical for understanding aquatic ecosystem health. Rivers make significant contributions to macronutrient and sediment delivery of water bodies including oceans, lakes, and ponds throughout the year that impact overall aquatic ecosystem productivity and drive harmful algae blooms (HAB’s). Of the major nutrients, phosphorus, nitrogen, and sediment are of concern, since they promote eutrophication when supplied in excess to receiving waters causing HABs (Anderson et al., 2002). The sources of these macronutrients and sediment in rivers are directly linked to both point (e.g., wastewater treatment plants, tile drains, storm water discharge pipes) and nonpoint sources due to the erosive nature of the changing watershed land use, land cover, and soil types. The source areas and transport mechanisms vary in both space and time. Temporal aspects, such as extreme weather events and hydrological processes that dominate during particular seasons can have considerable effects on loading. Our study area comprises a northern, humid temperate climate, and as such, experiences significant spring snowmelt and runoff processes that impact nutrient and sediment loading. It is expected that climate change will likely affect the timing and magnitude of riverine nutrient and sediment loading to receiving water bodies, which makes the temporal component critical for accurate load estimates. In this research we develop a new framework for improving the seasonal nutrient
and sediment load estimates by identifying (i.e., optimally clustering) and leveraging the temporal variability embedded in high resolution discharge records.

1.1 EUTROPHICATION AND MACRONUTRIENT DELIVERY

Eutrophication is defined as excessive nutrient abundance that causes an increase in productivity and a decrease in dissolved oxygen content, thereby reducing aquatic biodiversity (Sharpley et al., 2003). It is thought to result in the formation of algae and aquatic weeds that can pose a health hazard due to the presence of cyanobacteria, which can release harmful toxins that impact water supplies. In the saltwater conditions of coastal areas, eutrophic conditions are often attributed to excessive nitrogen loading; while in freshwater conditions, conditions are often attributed to excessive phosphorus loading. Phosphorus is considered the limiting factor for primary productivity in aquatic organisms because of low concentrations of bioavailable P in rivers and lakes (Anderson et al., 2002). There are numerous examples worldwide of increases in HABs linked to increased nutrient loading (Schindler, 1977; Burkholder, 2001; Burkholder & Glibert, 2006). In January 2003, the US Environmental Protection Agency sponsored a “roundtable discussion” to develop consensus on the relationship between eutrophication and harmful algal blooms (HABs) and determined that (1) degraded water quality from increased nutrient pollution promotes the development and persistence of many HABs, which continue to expand in the U.S. and other nations and (2) management of nutrient inputs to the watershed can lead to significant reduction in HABs (Heisler et al., 2008). Studies have suggested that agricultural fertilizer application is a primary source of excess phosphorus to receiving water bodies (Ghebremichael and Watzin, 2011; Jamieson et al., 2003); and other studies
have established relationships between high flow events and sediment/nutrient delivery (Pellerin et al., 2012; Sebestyen et al., 2009).

There has been a significant amount of research on the cause and effect of eutrophication in Lake Erie, which annually struggles with severely degraded water quality and large algal blooms. Studies showed that high flow-events are drivers of total phosphorus loading to Lake Erie, primarily driven by the spring storm period (Richards et al., 2008). As such, the magnitude of spring runoff in watersheds draining into Lake Erie has recently been determined to be the most consistent predictor of the severity of summer algal blooms (Michalak et al., 2013). Throughout the Lake Erie watershed, abundant agricultural activity, a major economic sector in the region, is the likely source for much of the total phosphorus delivered to the lake, because point source discharges from urban centers have decreased significantly due to mandated reductions in wastewater treatment plant discharges and the use of phosphorus-based detergents. Lake Champlain, situated between New York, and Vermont in the US and Quebec in Canada also struggles annually with algal blooms, which have increased in size and frequency throughout the century and continue to be a serious problem (Isles et al., 2015; Galbraith, 2015; Silberman, 2016). However, few studies in Vermont have investigated the relationship between the temporal variability associated with weather and stream flows throughout a calendar year and the complex interaction with land use/cover on the flux of river nutrient/sediment loadings.
1.2 LAND USE/COVER AND SOIL TYPES

The source and magnitude of nutrient loading along rivers has been linked to both point and nonpoint pollution from different land use types, primarily agriculture, urban activity, and industry (Carpenter et al., 1998). In less urbanized regions, the major contributor is nonpoint source pollution; and it is more of a concern in this work since a large portion of the Vermont land use involves agriculture, dairy or crop production. It is well known that the flow of water through soils with manure and fertilizer increase the export of nutrients including phosphorus from agricultural fields (Carpenter et al., 1998; Dupas et al., 2015), and that some practices, such as tilling and the installation of tile drains, can significantly alter hydrologic pathways, resulting in increased erosion and increased export of farm runoff directly to rivers (Edwards & Hooda, 2008; Hively et al., 2005). By contrast, pristine forested environments have significantly lower concentrations of both dissolved and sediment-bound phosphorus due to vegetation that anchors soil to the land surface and decreases erosion relative to some agricultural environments (Wang et al., 2008).

Land use/cover can modify, promote and/or control erosion and transport of sediment and nutrients; but soil characteristics differ widely across watersheds and range from thick silts and clays that do not erode easily to fine sands that are easily erodible. Phosphorus can adsorb to soil particles that erode and get transported to receiving waters, such as rivers and lakes (Sharpley et al., 2013; Newcomb, 2007). Hydric soils, for instance, which are permanently or seasonally saturated by water due to rainfall and flooding, can increase runoff nutrient and sediment transport (Skorupa, 2013). Sediment may be
mobilized from different environments (e.g., land surface, stream banks and the stream bed itself) during high-flow events and/or react differently during different seasons (Bayard et al., 2005). Furthermore, frozen/saturated soils, where nutrients and sediments accumulate, either in the snow, or below the snowpack in the near surface soil are known to effect surface and shallow subsurface flows as ice and snow melt (Bayard et al., 2005). The spring runoff period in temperate climates is characterized by variably frozen and saturated soils, which are known to isolate water at the ground surface and within the upper soil horizons that are actively melting (Bayard et al., 2005; Groffman et al., 2001; Hardy et al., 2001). The land surface and soil type characteristics play a direct role in the functions of water retention, sedimentation, and biogeochemical cycling of nutrient loading in river systems. As land surface conditions change with human use and the seasons, so will riverine loading, and the timing of these changes will have major impacts on receiving water bodies.

1.3 SEASONALITY AND A CHANGING CLIMATE

The extent and magnitude of nutrient/sediment loadings can vary considerably due to extreme weather events and different hydrologic processes that dominate during a particular season, making the temporal scale critical for estimating accurate loadings (Royer et al., 2006; Danz et al., 2010; Adhikari et al., 2010). Snowmelt and heavy spring rains have also been suggested as critical drivers of HAB severity (Daloglu et al., 2012; Stottlemeyer, 2001). Runoff over the frozen ground surface prior to a spring thaw can promote high concentrations of macronutrients and suspended sediment in stream networks that are then delivered to receiving waters, where the nutrients have potential to generate HABs as waters warm during the summer months. In the Northeastern United States, the
spring-time period is specifically important for quantifying nutrient loadings, especially in watershed dominated by high agriculture land use, because high flows generated by melting snowpack and the high number of storm events occurring during this season (Miatke, 2015; Rosenberg, 2016). This interesting dynamic between soils heavily impacted by the application of manure and fertilizers and the projected increases in magnitude and frequency of precipitation events during spring (Guilbert et al., 2015) motivates the need for data-driven models that better estimate nutrient and sediment loadings by leveraging the temporal variability embedded in high resolution climate and discharge records.

A phosphorus load study in Lake Erie found that the frequency of extreme storm events (defined as above the 85th percentile) since 1970 has been significantly greater in the spring compared to the fall (Daloglu et al., 2012). This same study showed that the frequency of such events has increased dramatically over the past decade, coinciding with both spring and fall fertilizer seasons, and demonstrates the importance of human-induced impacts in increasing nutrient load estimates as well. Guilbert et al., (2015) analyzed more than fifty years of climate data across the Northeastern US and showed that high rainfall events are increasing with the largest increase occurring in April. Thus, the onset and duration of the spring runoff season in this region is likely to vary significantly in the coming years causing changes in snow magnitude due to warmer winters and temporal shifts in the spring runoff conditions (Betts et al. 2014; Crossman, 2013; Betts, 2011). Our Lake Champlain study area is also likely to experience significant seasonal change during the twenty-first century as the annual snowfall and the number of days below 0°C are expected to decrease by 50% and 45 days, respectively (Guilbert et al., 2014).
changing weather patterns will have dramatic impacts on the temporal and spatial drivers of springtime nutrient loading in the Lake Champlain Basin. In Vermont’s Missisquoi River Basin specifically and similar watersheds, spring ice break and snowmelt create a bio-geochemically distinct event for reactive phosphorus and important hydrological period for nutrient loading to receiving waters (Rosenberg, 2016; Miatke, 2015). As seasonal and hydrologic conditions change, so will riverine loading, and the timing of this change will have major impacts on receiving water bodies.

1.4 LOAD ESTIMATION MODELS AND METHODS

Models are generally created to take estimates from historical data records and make future predictions. Process-based models are rooted in physics and seek primarily to describe observed data patterns (i.e., internal structure, rules, and behavior) embedded in key mechanisms using physical-based equations that conserve system processes such as mass, energy and/or momentum. In contrast, empirical models seek principally to describe the statistical relationships (patterns) among observed data with little regard to underlying theory or physical-processes governing the system. Given the complexity of physical systems governing the transport of sediment and nutrients through watersheds to receiving water bodies, empirical models are often used for estimating nutrient loads; and these load estimates are based on measuring concentrations and streamflow discharges (Cohn, 1992). The most common approach for predicting sediment and nutrient concentrations is to develop an empirical model that relates observations of concentration and discharge using a linear regression model (Vogel et al., 2005). In other studies, concentration is replaced with load for a tighter relationship (i.e., load-discharge) linear regression model (Labeau
et al., 2015). Since both load and discharge are functionally related, there tends to be spurious correlation in some models (Vogel et al. 2005). Despite the increased spurious correlation in load-discharge models, its use can be advantageous with proper bias correction factors (Kenney, 1982) because they are simple to use and apply, and represent a relatively coarse-level estimate and standardized approach, wherein differences between seasons (or between watersheds) can be examined. In this work, the linear regression slope and intercept coefficients are analyzed from multiple load-discharge models to develop our clustering framework for improved/optimized load estimates.

Despite the importance of accounting for temporal variability in estimating nutrient loads throughout a given year, relatively few studies use seasonal regressions with near continuous discharge measurements, but instead rely on annual regressions with average daily discharge values (Johnes, 2007; Smeltzer et al., 2012; LaBeau et al., 2015). Smeltzer et al. (2012) and Medalie (2014) estimated nutrient loads for 18 rivers and streams in the Lake Champlain Basin with a focus on phosphorus and nitrogen. Smeltzer et al. (2012) generated total phosphorus load estimates using annual regression relationships to predict concentrations and calculate daily load estimates from 1991-2008 using average daily discharge. Medalie (2014) estimated TP and DP fluxes from 1990-2012, and TN and TSS fluxes from 1992-2012, using a “Weighted Regressions on Time, Discharge, and Season” (WRTDS) method. The WRTDS method uses average daily discharge, similar to Smeltzer et al. (2012), but instead allows for maximum flexibility in representing the temporal trends, seasonal effects, and discharge-related components of the water-quality variable of interest. It is designed to provide internally consistent estimates of the measured
concentration and fluxes, as well as histories that minimize the influence of year-to-year variations in stream flow (Hirsch, 2010). Although this method addresses seasonality, the use of average daily discharge to calculate load estimates can potentially underestimate loading estimates associated with critical time periods, such as the effect of spring snowmelt or increased frequency of large storm events (Miatke, 2015). Furthermore, both studies were limited to a total of about 14-19 water quality samples per year as part of the Lake Champlain Long-Term Water Quality and Biological Monitoring Project (VT DEC and NY DEC 2012). In another study, LaBeau et al. (2015) addressed the importance of seasonal changes in discharge for the U.S. Great Lakes Basin using total phosphorus loadings calculated with average daily discharge measurements. Collecting data for other analytes is critical for understanding the relationship between species. For example, soluble or sorbed phosphorus has been observed to react differently under varying discharges given their source; Dorioz et al., 1989 showed that phosphorus sorbed to sediment eroding from agricultural lands may be more sensitive to high-discharge events than dissolved phosphorus. Watersheds with greater seasonal discharge variability may therefore be more susceptible to sorbed phosphorus loading. Measuring total phosphorus, dissolved phosphorus, total nitrogen, and total suspended solids helps improve understanding of the complex biogeochemistry and overall nutrient loading to water bodies.

1.5 OVERVIEW OF THE THESIS AND SPECIFIC OBJECTIVES

The specific objective of this research was to develop a new framework for improving seasonal nutrient and sediment load estimates by identifying (i.e., optimally clustering) the temporal variability embedded in high resolution discharge records. We use
water quality concentration and associated 15-minute discharge data for nine watersheds in Vermont’s Lake Champlain Basin as our test bed to determine how Lake Champlain Basin nutrient/sediment load estimates vary in space (from river to river) and time (annually, and seasonally). Analysis of these nine watersheds also revealed differences in load estimates in space (from river to river), but this was not a specific focus of our work. The framework is developed to cluster monthly discharge-load regressions resulting in improved linear models that are based on a minimum number of sampled measurements, and slope-intercept coefficients of individual regression models. A specific focus of this work is to identify the optimal scale (time period) to best estimate multiple/correlated analyte loadings during this interesting hydrological regime. In this work, we hypothesize that a data-driven framework for optimizing the temporal scale that defines a “season or hydrologic regime” will provide more accurate estimates of nutrient and sediment loadings.

The method produces multiple regression models whose time span is clustered based on number of samples, the temporal order of the months, and the slope and intercept coefficients of the individual load-discharge regressions using the observed sample concentrations for a variety of constituents (i.e., total phosphorus (TP), dissolved phosphorus (DP), total nitrogen (TN), and total suspended solids (TSS)) from a larger water quality dataset and near continuous discharge measurements. The “optimized” time periods result in improved linear models. This data-driven framework leverages the recent proliferation of high-resolution sensor networks and helps identify more appropriate time scales given adequate amounts of data compared to traditional annual or meteorological four-season models. We focus on the period of spring snowmelt runoff as a case study to
demonstrate the proposed framework for estimating seasonal loadings. Furthermore, the flexibility in determining regression coefficients for site-specific watersheds or goals will help identify the complex linkages between water quality basin characteristics (e.g., land use/cover, soil type), hydrologic regimes, and temporal variability in discharge to better develop time-dependent best management practices for a changing climate.

1.5.1 Spring Study Site- Missisquoi Watershed

The Missisquoi River runs 88 miles through northern Vermont and southern Quebec, and drains to Lake Champlain in Missisquoi Bay (Figure 2.1). Missisquoi Bay drains 855 square miles of northwestern Vermont and southern Quebec with almost 60% of the drainage area in Vermont. Despite significant phosphorus-load reduction efforts in the Missisquoi River Basin, a large agricultural basin, the land-use practices over the past centuries in the watershed have led to a degradation of the water quality in the river and the bay. The spring time period (i.e., period associated with hydrologic processes involving snowmelt and runoff over frozen ground) is analyzed for the Missisquoi watershed; and increased sampling effort, specifically targeting spring water quality, occurred to better represent this critical time period.

1.6 THESIS OUTLINE

This chapter has presented the specific objective of this thesis and a brief literature review and description of the study site. The second chapter, A Framework for Improving Nutrient and Sediment Load Estimates by Leveraging Temporal Variability Embedded in Water Monitoring Data, presents the bulk of this work in manuscript format and will be
submitted to Journal of Great Lakes Research. A script of the Matlab code used to program the framework and models presented in chapter two is available in Appendix A.

1.7 REFERENCES


Newcomb D., 2007. Links between geomorphic condition, water quality, and phosphorus loading in Hungerford Brook, Vermont. Thesis presented to Faculty of the Graduate College of The University of Vermont.


Research on Adaptation to Climate Change (RACC), 2015. VT-EPSCoR, http://epscor.w3.uvm.edu/2/node/30


CHAPTER 2: MANUSCRIPT FOR JOURNAL

A Framework for Estimating Nutrient and Sediment Loads that Leverages the Temporal Variability Embedded in Water Monitoring Data

Baxter G. Miatke*, Donna M. Rizzo¹, Kristen L. Underwood¹, Andrew W. Schroth²

For submission to Journal of Great Lakes Research

Corresponding Author: Baxter Miatke, baxterm419@gmail.com

* Corresponding Author- baxterm419@gmail.com
¹ Department of Civil & Environmental Engineering, University of Vermont, 33 Colchester Ave., Burlington, VT 05405, drizzo@uvm.edu, Kristen.Underwood@uvm.edu
²Department of Geology, University of Vermont, 180 Colchester Ave., Burlington, VT 05405, aschroth@uvm.edu
2.1 ABSTRACT

Rivers often deliver most of their sediment and solute loads during particular times of the year, which drive their impact on receiving water quality. Here, we establish a data-driven framework to develop the most representative empirical equations to estimate sediment and nutrient loading that capture temporal variability embedded in hydrologic regimes and the complexities driving nutrient transport over the course of a hydrologic year. We use water quality concentration and associated 15-minute discharge data for nine watersheds in Vermont’s Lake Champlain Basin to determine how nutrient/sediment load estimates vary in space and time. The framework optimally clusters time periods for analysis using the slope and intercept coefficients from best-fit load-discharge regressions using 16 years of historical data from nine basins draining to Lake Champlain; two of the nine basins were supplemented with water quality data from our 2012-2015 field work. Optimized regression models are used to estimate annual and “spring” loadings for total phosphorus, dissolved phosphorus, total nitrogen, and total suspended sediment loads. These models performed on average ~20% better than annual regression models in terms of Nash-Sutcliffe efficiency across all four constituents, and exhibited marked improvements in both bias and error. The largest loadings (25% to 70% of the total annual estimated load) occurred in the “spring”. The framework developed here is robust and can analyze other units of time depending on the hydrologic regime of interest, and can also be configured to address a variety of site specific management applications provided that adequate monitoring data exist.
2.2 INTRODUCTION

Increases in nutrient and sediment loading have been linked to local and regional changes in land-use/land-cover (Carpenter et al., 1998), changes in climate and hydrologic processes (Betts et al., 2014; Rosenberg 2016), and seasonal patterns in hydrology (i.e., low discharge during summer; heavy spring rains and snowmelt). The extent and magnitude of riverine sediment and nutrient loading can vary considerably due to these different hydrologic processes, making the choice of an “appropriate” time scale a critical consideration when developing models to project riverine loading across the hydrologic year (Royer et al., 2006; Danz et al., 2010). Snowmelt and heavy spring rains, for example, have been suggested as critical drivers of harmful algal blooms (HABs) in many temperate systems (Daloglu et al., 2012; Stottlemeyer, 2001). In the Northeastern United States, the spring time period is specifically important for quantification of nutrient loading because of the high flows generated by melting snowpack and the high number of storm events occurring during this season (Rosenberg, 2016). The spring runoff period in temperate climates is characterized by variably frozen and saturated soils, which are known to affect surface and shallow subsurface flows, and may lead to accumulation of nutrients and sediments either within the melting snowpack or in the near-surface soils (Bayard et al., 2005; Groffman et al., 2001; Hardy et al., 2001).

Furthermore, analysis of more than fifty years of climate data showed that precipitation events are increasing in both frequency and magnitude across the Northeastern US, with the largest increases occurring in April (Guilbert et al., 2015). The Lake Champlain Basin, our selected study area, will likely experience significant climate
change during the twenty-first century as the annual snowfall and the number of days below 0°C are expected to decrease by 50% and 45 days, respectively (Guilbert et al., 2014). As a result, the onset and duration of the spring runoff season in this region will likely vary significantly in the coming years, with a predicted temporal shift in warmer winters possibly causing changes in snow magnitude and earlier onset of spring runoff (Betts et al., 2014; Crossman et al., 2013; Betts, 2011). These changing weather patterns will therefore impact the temporal and spatial drivers of springtime nutrient and sediment loading to receiving waters of the Lake Champlain. Methods for developing loading models that will be sensitive to such changes in seasonal loading dynamics are of particular importance in the rapidly changing environment of the Anthropocene.

Many previous studies have estimated loads using annual regression equations (Dolan et al., 1981; Cohn et al., 1992; Syvitski, 2000), but the temporal resolution of available data is often insufficient to capture changes in concentration-discharge relationships that occur on seasonal timescales. Despite the importance of using this temporal variability to better estimate nutrient loads (Adhikari et al., 2010), relatively few methods use near-continuous discharge measurements to produce seasonal (or shorter time-scale) regressions; and instead rely on annual regressions using average daily discharge values (Johnes, 2007; Medalie et al., 2012; LaBeau et al., 2015). When seasonality is addressed, standardized definitions of season (e.g., 3-months) tend to be the default regardless of differences in watershed sensitivity to temporally-variant biogeochemical processes. This sensitivity is particularly important in Vermont basins because loading is driven by snow melt and other high-flow events (Adhikari et al., 2010).
The importance of the seasonal variability in discharge for the U.S. Great Lakes Basin was studied by LaBeau et al. (2015) for total phosphorus loadings that use load-discharge regressions on average daily discharge measurements. Collecting data for other chemical analytes is critical for understanding the relationship between species. For instance, soluble or sediment-bound phosphorus may react differently to varied discharges given their source. One source of sediment-bound phosphorus is erosion on agricultural lands, which may be more sensitive to high-discharge events than dissolved phosphorus (Dorioz et al., 1989). Watersheds with greater seasonal discharge variability may therefore be more susceptible to sediment-bound phosphorus loading. Measuring total phosphorus, as well as dissolved phosphorus, total nitrogen, and total suspended solids helps improve the analysis of the complex biogeochemistry and overall nutrient loading to water bodies.

To ensure accuracy, regression should be based on, water quality samples obtained under a full range of flow conditions, with a strong emphasis on high-flow conditions to help improve the precision of annual mass balance loading estimates (Johnes, 2007). Samples representing a full range of conditions are necessary for temporal variability to be addressed in a seasonal framework. However, few studies in Vermont have investigated the relationship between the temporal variability associated with weather and stream flows throughout a calendar year and the complex interaction with land use/cover on the flux of river nutrient/sediment loadings. Regression estimates can further be improved through stratification of samples into groups wherein intragroup variability is minimized (Quilbe et al., 2006). Examples include separating observations by season or by rising versus falling limb of the hydrograph (Littlewood, 1995; Asselman, 2000), but
few studies have promoted clustering or grouping methods for optimizing the temporal stratification of time series data to improve accuracy of load estimation (Cohn et al., 1992; Smeltzer et al., 2009).

The specific objective of this research was to develop a new framework for improving seasonal nutrient and sediment load estimates by identifying (i.e., optimally clustering) the temporal variability embedded in high resolution discharge and water quality monitoring efforts. We use water quality concentration and associated 15-minute discharge data for nine watersheds in Vermont’s Lake Champlain Basin as our test bed to determine how Lake Champlain Basin nutrient/sediment load estimates vary in time (annually, and seasonally). Analysis of these nine watersheds also revealed differences in load estimates in space (from river to river), but this was not a specific focus of our work.

The method developed here produces multiple “seasonal” regressions whose time span is optimized by clustering data based on the number of samples, and the load-discharge linear regression coefficients, using the observed constituent concentrations collected under a full range of flow conditions and near-continuous discharge measurements. The framework proposed here can be used to leverage the recent proliferation of high-resolution data from sensor networks to identify the appropriate temporal scales for grouping data to improve loading estimates. The between-watershed differences in temporally-optimized regression models may also suggest variable influence of bio-geochemical and hydrologic processes driven, in part, by land use and soil characteristics, which in turn may help develop better “temporal” land management practices for Vermont basins.
2.3 METHODS

2.3.1 Study Area

Nine of the twenty-two Lake Champlain long-term monitoring river sites (VTDEC, 2015) were chosen for this research representing a range of drainage areas, land use and hydrologic features across Vermont (Figure 2.1 and Table 2.1). All nine watersheds drain to Lake Champlain, the 14th largest freshwater lake by volume in the United States. Lake Champlain is over 125 miles (201 km) long, 14 miles (23 km) across at its widest part, and drains north to the larger St. Lawrence River Basin that connects the Great Lakes with the Atlantic Ocean. The maximum depth is 400 feet (122 m); but with an average depth of 64 feet (19.5 m) and shallower in most other areas, the lake as whole experiences little mixing and is fragmented into multiple arms and bays (Isles et. al 2015). The land to water surface ratio is 16.8 to 1, which is unusually large and promotes water quality vulnerability to changing land use practices that enhance nutrient and/or sediment loading.

Elevations in the study basins range from 1,339 m (4,393 ft) above mean sea level at Mount Mansfield (divide between Basins 7 and 8 in Figure 2.1) to 29 m (95 ft) at the average stage of Lake Champlain. The Champlain Basin has a humid temperate climate, with mean annual precipitation ranging from over 1,270 mm (50 in.) along the north-south trending spine of the Green Mountains to a low of 813 mm (32 in.) in the Champlain Valley (Randall, 1996). Stream discharge regimes vary from small flashy systems higher in the headwaters to larger flood-prone rivers lower in the Champlain Valley. Overall variability in mean daily flow was calculated for a period from 1990-2015 as the ratio of the 95th percentile to the 5th percentile of flow (Table 2.1). In general, the more mountainous basins
have higher ratios, while the lower-relief basins and those with higher percentages of wetlands, lakes and ponds or human impoundments (e.g., Otter Creek) have lower ratios. Within a typical year, a majority of the runoff from Lake Champlain tributaries occurs between ice-out and late spring (Shanley and Denner, 1999). Analysis of 2012-2014 hydrographs from these nine watersheds shows the peak snowmelt discharge to generally occur during March-April (Miatke, 2015), consistent with Shanley and Denner (1999).

Land use in the study basins ranges from 7 - 49% agricultural and 37 - 77% forested. Urban land uses, including transportation corridors, range from 6 - 13% (Troy et al., 2007). Given the overall shallow depths and limited mixing, nutrient inputs from tributaries and land-use management in the contributing watersheds are critical drivers of lake health. This is especially evident in Missisquoi Bay, a shallow unconnected bay in the northeast portion of the lake fed by the Missisquoi River (Basin 9, Figure 2.1). Severe HABs in Missisquoi Bay have made this a targeted research area leading to deployment of additional sensors and automated water samplers in both the bay and along the Missisquoi River (Isles et al., 2015; Pearce et al., 2013; Rosenberg, 2016).

2.3.2 Water Quality Data and Sampling

The Vermont Department of Environmental Conservation (VTDEC) has sampled water quality in the nine study basins since 1990 for a wide range of parameters including total phosphorus (TP), dissolved phosphorus (DP), total nitrogen (TN), and total suspended solids (TSS). These tributary samples are obtained over a full range of flow conditions each year, but with a strong emphasis on high flow conditions to help improve annual mass balance loading estimates (VTDEC, 2015). Twelve or more samples per year were
obtained for most parameters, including TP, DP, TN, and TSS, with additional sampling (for phosphorus only) in certain tributaries. Sampling sites at or near bridge crossings are located as close as possible to the mouth of each river. It is important to note that not all VTDEC long-term monitoring sites are aligned at the confluence of the rivers and Lake Champlain, nor are they always located at the United States Geological Survey (USGS) gauging stations where discharge is measured. Watersheds were delineated to VTDEC sampling points in Figure 2.1; the ratios of watershed area to total drainage area ranged from 0.6 to 1.1 (Table 2.1). USGS gauge stations have provided near-continuous (i.e., 15-minute) discharge data since 1990 with some gaps in data collection due to ice cover or equipment malfunction. These gaps in discharge were filled by reconstructing the hydrograph using available stage (height) data and the stage-discharge relationship for that particular gauge, when possible.

This work also used additional TP, DP, TN, and TSS water quality data sampled using Teledyne™ ISCO automatic samplers at the Missisquoi River USGS gauge station at Swanton and at the Winooski River USGS gauge station at Essex Junction (Basins 9 and 7, Figure 2.1) from 2012-2015 as part of the Vermont Experimental Program to Stimulate Competitive Research (VT-EPSCoR). Since these two sites (bolded in Table 2.2), are located in close proximity to the VTDEC sample locations, we combined the two water quality datasets for this research. The ISCO samplers were programmed to collect samples during large discharge events above specified flows. Additional grab samples were obtained between storm events to monitor base flow conditions, and during the months of March and April, 2014, to capture the effect of snowmelt on constituent concentrations.
Sampling was conducted following quality assurance procedures from Worsfold (2005). Table 2.2 summarizes the period of record for all nine study sites and includes the total number of water samples collected for each constituent that had corresponding 15-minute discharge measurements and passed lab quality assurance tests. We should note there were samples from both datasets that did not have 15-minute discharge data at the time of sampling and could not be used; this highlights a limitation of using 15-minute discharge data instead of average daily discharge.

2.3.3 Load-Discharge Regression Development

Load estimates of TP, DP, TN, and TSS were calculated by multiplying observed concentrations with corresponding 15-minute USGS discharge measurements to obtain load estimates in kg/15-minute interval. The bivariate linear regression relationship between the logarithm of concentration and flow has been modified from Vogel et al. (2005) replacing concentration with load as follows:

\[ L = b_0 + b_1 \bar{Q} + \varepsilon, \]  

(1)

where \( L \) is the base-10 logarithm of load in kg/15 minutes, \( \bar{Q} \) is the base-10 logarithm of discharge in L/sec, \( b_0 \) and \( b_1 \) are the intercept and slope regression coefficients, respectively; \( \varepsilon \) is the normally distributed model error with zero mean and variance. Equation 1 is back-transformed to the following power model:

\[ L = 10^{b_0} \times Q^{b_1} \times 10^\varepsilon. \]  

(2)

Thus, an estimation of \( 10^\varepsilon \) is all that is required to account for the bias in predictions when back transforming the power model. In this research, the model residuals were shown to be normally distributed using a Shapiro-Wilk W statistic, allowing us to use
the estimate of the bias from Newman (1993), $10^\varepsilon = 10^{\frac{MSE}{2}}$, where MSE is the mean square error from the regression. If the residuals had not been normally distributed, then the smearing estimate of bias would need to be determined (see Newman, 1993; Duan, 1983). Individual model residuals are not reported due to the large number of models evaluated.

The load-discharge regression model is most commonly used with average annual observations; an example using the Winooski and Missisquoi TP data is provided in Figure 2.2. Visible inspection suggests that not all the data are best fit by a single annual regression model, and that a better model might include regressions stratified by finer time scales (Quilbe et al., 2006). Our proposed framework was motivated by the desire to optimize the time periods used to best fit linear regression models.

2.3.4 Framework Development

The framework is designed to optimize the time period over which linear regression models best fit the available load-discharge data using a hierarchical cluster analysis of the regression coefficients (i.e., intercept ($b_0$) and slope ($b_1$)). For instance, one best-fit regression model might be constructed using the long-term monitoring data clustered over the months of March, April, and May as a single unit of time, while another may cluster only data from the month of June. The data-driven framework is designed to guide users through the clustering process (Figure 2.3a), and begins (step 0) by having the user select a base unit of time. Any base unit may be used (e.g., biweekly, weekly, or even daily); in this work, we chose a time unit of one month. Next, the full time-series of flow and concentration data available for each analyte were grouped by this user-defined time
unit (e.g., all available January observations of discharge and load from the 1990-2015 time
series were grouped together, all February observations were grouped, etc.) Step 1 then
ensures that a minimum number of samples exist within each user-selected base time unit.
In this research, a minimum number (n=7) of samples was selected following the
Figure 2.3b shows an example (using TP) of the number of samples per month over the
1990-2015 monitoring period. In step 2, the user must define a starting time period, ideally
one of particular hydrologic importance for the region or process under study, and assign
an ordinal integer value for use in the subsequent cluster analysis. This step is important as
it preserves the sequential order of the time units, which helps ensure that units far apart in
time will not cluster together. In this case, the month of March was assigned an ordinal
value of “1” to target the spring snowmelt and runoff, a time period of interest based on
previous studies (Miatke, 2015; Rosenberg, 2016). Each subsequent time unit (i.e., month)
increases by a value of one, ending with “11” for the January-February time unit (i.e.,
January and February were combined due to an insufficient number of samples in step 1;
see Figure 2.3a). In step 3, linear regressions are performed on the group of observations
within each time unit to generate the slope and intercept coefficients (Figure 2.3c, 2.3d).
These regression coefficients along with the integer values assigned to preserve temporal
order (from step 2) are used as input to a hierarchical cluster analysis using Ward’s method
(step four). Ward’s method, known as the minimum variance method, finds the distance
between two clusters and the analysis of variance sum of squares between the two clusters.
Ward’s method tends to join clusters with a small number of observations and is strongly
biased toward producing clusters with approximately the same number of observations (Milligan and Cooper, 1985).

The resulting temporal clusters are visualized in a dendrogram such that the user can confirm whether the clusters appear logical and statistically significant at the 5% level for a two-tailed test. Acceptable clusters of data are used to re-create “optimized” load-discharge regressions. Alternatively, the analysis may be repeated with a different base time unit, a higher threshold for the minimum number of samples per base time unit, or a different starting point for the sequential ordering of time units. This framework also allows a user to skip the hierarchical cluster analysis and specify any number of clusters (e.g., k-means clustering), a priori, based on knowledge of dominant processes active during certain time periods. As previously noted, this framework may be more finely resolved (e.g., on the order of weeks or days) given a sufficient amount of data. We use monthly units in this proof-of-concept to ensure a sufficient number of water quality samples for each temporal regression and to help evaluate land-management practices that typically occur on monthly to seasonal time scales.

2.3.5 Global Performance Metrics

In addition to testing the statistical significance of linear regressions (step 3) and clusters (step 4), a set of global performance metrics was calculated to compare across models. Performance, in this context, is defined as the goodness of fit of the load-discharge models (Moriasi et al., 2007). Nash-Sutcliffe efficiency (NSE) indicates how well the observed measurements versus simulated (optimized) data fit a 1:1 line, and is computed as:
\[ NSE = 1 - \frac{\sum_{i=1}^{n} (Y_{i}^{obs} - Y_{i}^{sim})^2}{\sum_{i=1}^{n} (Y_{i}^{obs} - Y_{i}^{mean})^2} \]  

(4)

where \( Y_{i}^{obs} \) is the \( i \)th observation for load, \( Y_{i}^{sim} \) is the \( i \)th modeled value, \( Y_{i}^{mean} \) is the mean of the observed data, and \( n \) is the total number of data points. A NSE value of 1.0 indicates a perfect match between modeled and observed data; and NSE values >0.50 generally indicate satisfactory model performance with increasing performance as NSE approaches 1.0 (Moriasi et al., 2007). NSE values \( \leq 0.50 \) are considered unsatisfactory; and NSE values \( \leq 0.0 \) indicate that the observed mean is a better predictor than the model. The average adjusted \( R^2 \) coefficient of determination and average bias of the observed and modeled loads are also reported. Bias values were calculated using:

\[ Bias = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_{i}^{obs}}{Y_{i}^{sim}} \right) \]  

(5)

where bias values <1.0 represent a model that overestimates the observed values and a bias > 1.0 represents a model that underestimates compared to observations. All above analyses and methods were performed using Matlab R2015b software (Mathworks, 2015) and will be made available online.

2.5 RESULTS AND DISCUSSION

Application of the framework is shown for TP in the Missisquoi basin in Figure 2.3b-2.3d. In this example, the optimized number of clusters was six (see dendrogram of Figure 2.4), comprised of [March-April], [May-June], [July], [August-September], [October-November], and [December-January-February]. We refer to the final framework result as the optimized annual watershed model because estimated loadings have been aggregated over the course of an entire year. However, we remind the reader that these
estimated loadings were constructed using a number of “optimized” linear regressions on data that may have clustered into 1-, 2-, 3-, and/or 4-month time periods. The optimized time period for each best-fit regression model is presented by constituent in Table 2.3 for each watershed. The number of optimized temporal clusters across all watersheds ranged from five to seven per year, comprised of anywhere between 1 and 4 months each. In most cases, there were six total clusters across all watersheds and constituents, with December, January, and February often clustering together as a function of insufficient n values. November also clustered with these three months for TSS samples (in all but Basins 9 and 10). When time-series data for TP were analyzed, the month of March emerged as a 1-month cluster for four watersheds and clustered with April for the other five, indicating that the optimized regression coefficients (slope and intercept) were statistically different enough in March to be treated as an individual cluster. The seasonal clustering for DP followed similar trends to TP, with the exception of the Lewis and LaPlatte watersheds, which overall had a different number of total clusters. Analysis of the Lewis watershed also clustered February with March based on step seven of the framework. The optimized groupings for TN were similar to the pattern for TP in 4 out of 9 basins, identifying March as an individual cluster. The seasonal clusters for TSS were the most difficult to optimize using the proposed framework because they had the lowest number of samples compared to other constituents (see Table 2.2). As a result, four-month “winter” clusters were common for TSS, while the number of summer clusters showed more variability than for other constituents. Global performance metrics (adjusted $R^2$, RMSE, NSE, and bias) were
calculated and partitioned by constituent for both the optimized and traditional annual regression models (Table 2.3).

2.5.1 Total Phosphorus

The traditional annual models for TP had high adjusted $R^2$ values, but poor efficiency for the Lamoille, Otter, and Winooski watersheds. Whereas the optimized annual TP models had similarly high adjusted $R^2$ values, but also had higher efficiencies for all watersheds with the exception of the Lewis watershed. On average, there was an increase of $\sim$19% in efficiency, a decrease in error of $\sim$18.5%, and a decrease in bias of $\sim$12.5% across eight of the nine watershed models when using the optimized regressions compared to a traditional annual regression. While the optimized models generally performed better, two of the study watersheds (Otter and Winooski), had an NSE value below 0.50, which is considered poor. While the NSE is low for the Winooski watershed, the optimized model improved the large bias (1.91 for the traditional model) by 21.5% to achieve a value closer to 1. These metrics suggest the optimized TP regressions should provide better loading estimates compared to traditional annual regressions.

2.5.2 Dissolved Phosphorus

The relative performance of the optimized and traditional annual models of DP were similar to TP with an average increase in efficiency of 26.1%, a decrease in error of 17.8%, and a decrease in bias of 10.8% across all nine of the watersheds. All optimized DP models had NSE values greater than 0.50 and were considered good.
2.5.3 Total Nitrogen

The traditional annual TN models had highest performance metrics compared to the other constituents and despite being of high quality for use in practice, the performance metrics improved even further when clustered into the optimized model by an average 5.7% increase in efficiency, an 11% decrease in error, and a 3.5% decrease in bias across all nine watersheds.

2.5.4 Total Suspended Solids

TSS performance metrics for both the optimized and traditional annual models were the worst across all constituents. All optimized annual TSS models had efficiencies less than 70%, even though the efficiency increased on average by 26% across eight watersheds when compared to the traditional annual models. In the Lewis watershed the NSE had a negative value, indicating that the mean would be a better indicator than using the TSS optimized model. TSS optimized annual models also had the highest error and bias compared to the other constituents despite a significant decrease in average error and bias by 8.4% and 22.1%, respectively, across eight of the watersheds. Overall, the optimized load-discharge regressions produced markedly improved performance metrics over traditional annual load-discharge regressions.

2.5.5 Watershed Trends and Discussion

Overall, the optimized annual model had good global model metrics for each study site, with the exception of Otter Creek, which showed some of the lowest performance metrics across all constituents. The latter may be due to the small number of
observations across all four analytes in Otter Creek (n ranges from 157 to 291) compared to other watersheds (see Table 2.2) despite the low variability in discharge (Q95/Q5; Table 2.1). The low variability in discharge may, in part, be attributed to a 9,000-acre complex of wetlands in the floodplain of the Otter Creek between Middlebury and Rutland. The wetlands store flood waters and slowly release them over time so that Middlebury and other locations downstream (i.e. the USGS Gauge Station) experience reduced flood peaks (LCBP, 2012). Also, Otter Creek flows are somewhat regulated by a series of upstream hydroelectric dams and reservoirs (USGS, 2015). Regulated systems like these make load estimation challenging because they dampen the load-discharge relationship compared to other unimpeded rivers; and dams can serve as points of discontinuity in the longitudinal transport of sediment and sediment bound nutrients (Magilligan et al., 2003; Williams and Wolman, 1984). Other watersheds, such as Missisquoi and Winooski, had much better optimized model performance, which may be due to additional water quality samples added to the VTDEC dataset, particularly those collected during the spring snowmelt under VT-EPSCoR. From a constituent point of view, the TSS load-discharge relationships were the most challenging to model and had the poorest performance metrics. This may be due to the fact that sediment transport processes and soil types vary significantly across the nine watersheds and across seasons. Hysteresis effects at the storm event-scale also confound sediment load-discharge relationships (Williams, 1989). Due to the disproportionate impact of the spring runoff period on nutrient and sediment loading during typical hydrologic years in the Lake Champlain Basin, we will focus on insight gleaned from our regression framework for this critical period in subsequent discussions.
2.5.6 “Spring” Trends and Model Metrics

Optimized temporal clustering results for time periods coincident with the meteorological definition of spring (i.e., March, April, May) were variable across the watersheds and across constituents. When examining TP in some watersheds (i.e., Lamoille, LaPlatte, Winooski), the framework identified March as its own singular-month cluster, and April and May in a two-month cluster. While for other watersheds, March and April clustered together, and May was either on its own or clustered with June (i.e., Lewis, Little Otter, Mettawee, Missisquoi). These results indicate differential variance in the load-discharge relationships between watersheds, and may be a function of differences in hydrologic regimes and biogeochemical filtering capacities between watersheds. Snowmelt, for instance, occurs at different times across even a single watershed and mobilizes nutrient runoff differently producing different optimized $b_1$ and $b_0$ coefficients across time periods, catchments and between measured parameters. These results might also be explained by temporal shifts in spring snowmelt over the data series, given that spring snowmelt is occurring earlier (i.e., earlier in March rather than April) in Vermont (Betts, 2011). While there are other confounding drivers of these clusters, the two-month clusters in spring may therefore be indicative of snow-melt occurring (on average) over both March and April and influencing the slope and intercept of the load-discharge relationship.

The relative performance of each model relies heavily on the biogeochemical processes operating to influence the load-discharge relationship at any given temporal unit being analyzed. The specific focus for demonstrating this particular proof-of-concept was
to improve loading estimates under the influence and timing of snowmelt occurring in each watershed. Thus, three “spring” time models were compared for performance. We refer to these as: annual spring, 3-month spring, and optimized spring. The annual spring model used the traditional annual regression model (generated with all of the data) and then estimated loads for the meteorologically defined spring (March, April and May) based on these annual regression parameters. The 3-month spring model, on the other hand, used a regression model generated using load-discharge observations from only March, April, and May; and applied the resulting slope and intercept coefficients to 15-minute discharge observations for those same months. The optimized spring model used the optimized (clustered) regressions (Table 2.3) to make estimates for March, April and May. An example of the three “spring” models for Missisquoi using TSS and TP (Figure 2.5) suggests that the annual regression alone does not capture “spring” effectively. The 3-month and optimized spring regressions, which have significantly different slope and intercept coefficients, may better capture specific hydrologic processes during that time period. The performance metrics for all three models are displayed in Table 2.4 for all nine watersheds and all four constituents. Performance for the three spring models varied greatly between watersheds and across constituents. Optimized spring regressions did not always perform better than the annual regression (see Lewis, Otter, and Mettawee watersheds). In the two watersheds (Missiquoi and Winooski) with additional spring sampling, global performance metrics for the optimized models improved only for TP and DP. In general, the watersheds with higher number of samples in March, April, or May had better model performance using the optimized regressions than watersheds with small numbers of data.
The TP data for Missisquoi is a good example; the optimized model had similar adjusted R\textsuperscript{2} values across all three models, but exhibited lower error and bias, and higher efficiency. In this case, the optimized spring model would be expected to produce more reliable loading estimates compared with the annual or 3-month models. One clear exception was the Otter Creek watershed, which had poor spring model performance across all constituents similar to the annual models. The Mettawee optimized spring model for TP was another example of a model that performed poorly in terms of NSE compared to the 3-month spring model, despite having a relatively high R\textsuperscript{2} value and virtually no bias. This may be due in part to the low gauged-to-sampled ratio in these two watersheds, where the USGS gauge is located further upstream from the sampling station and may not be as representative of flow conditions at the sampled station. Thus, it is important to examine all of the global metrics before estimating loads in order to describe different models in terms of efficiency, bias, and error. These performance metrics would be expected to improve with more water quality samples and proximal discharge measurements. The bias performance metric, suggests that the annual spring models tend to underestimate loads, whereas the optimized annual and spring models showed less bias (Table 2.3, Table 2.4). When true loadings are underestimated, the target reductions actually needed to meet total maximum daily load (TMDL) limits will vary significantly, which directly affects stakeholders and land management plans. Results from this framework can help stakeholders optimize their management strategies at specific times of year and save costs associated with low-impact land management practices that overlook the importance or
sensitivity of the watershed’s nutrient loading to temporal/hydrologic variability embedded in certain seasons.

2.5.7 Framework Estimation of Annual Loads

To demonstrate another application of this framework, annual loadings were estimated using the optimized annual models. The Missisquoi watershed was selected for proof-of-concept because of the additional sampling efforts in this watershed and the focused research interest in loading effects on HAB’s in the Missisquoi Bay (RACC, 2015). Since the global performance metrics were the best for the TP models, we generated cumulative loading estimates over the course of a year (2014) to demonstrate differences between using the annual and optimized models. The optimized model shows a much larger cumulative annual loading (~ 40%) compared to the annual model (Figure 2.6). When superimposed on the respective hydrograph, one can see that snowmelt and large flow events during storm events throughout the year are largely responsible for the difference between the two models. Data from the historic Mt. Mansfield snow stake in Stowe, VT (UVM-EcoInfo 2015) suggests that snowmelt occurred generally in late March/early April coinciding with the optimized spring regressions and steeper slopes. Larger slope coefficients during the spring snowmelt may be explained by consistently higher flows concurrent with exposed bare ground and flow paths that connect biogeochemical hot spots that have accumulated under snowpack to the river that uniquely impact the relation between concentration and discharge (and thus load discharge) (Bayard et al., 2005). This manifests differently across basins and constituents due to differences in land use/land
cover and river configurations, as well as the (similar or different) pathways and processes that govern constituent mobility.

Sediment and nutrient loadings to Lake Champlain from these study basins have been estimated by others as part of the long-term monitoring in the context of the Lake Champlain TMDL for phosphorus (Semeltzer et al. 2012; Medalie, 2014). While our methods for estimating loads differed considerably (i.e., were not flow-normalized, were based on near-continuous time series rather than daily mean flows, and were based on temporally-optimized regressions), it is instructional to compare results. Annual load estimates using our framework for TP, DP, TN, and TSS were calculated for water years 2000-2015 using 15-minute discharge data and the optimized annual models. Results for the Missisquoi watershed are provided in Figure 2.7. Nutrient loadings were produced by Smeltzer et al. (2012) and Medalie (2014) for 18 rivers and streams in the Lake Champlain Basin with a focus on phosphorus and nitrogen. Smeltzer et al. (2012) generated total phosphorus load estimates using annual regression relationships to predict concentrations and calculate daily load estimates from 1991-2008 using average daily discharge. Medalie (2014) estimated total phosphorus and dissolved phosphorus fluxes for 1990-2012 and total nitrogen and total suspended sediment fluxes for 1992-2012, but used a “Weighted Regressions on Time, Discharge, and Season” (WRTDS) method. The WRTDS method uses average daily discharge, similar to Smeltzer et al. (2012), but instead allows for maximum flexibility in representing the temporal trends, seasonal effects, and discharge-related components of the water-quality variable of interest. This method is designed to provide internally consistent estimates of the measured concentration and fluxes, as well
as histories that eliminate the influence of year-to-year variations in stream flow (Hirsch et al., 2010). Although this method addresses seasonality and is useful for evaluating long-term, flow-normalized trends, the use of average daily discharge to calculate load estimates can potentially underestimate critical time periods, such as the effect of spring snowmelt or increased frequency of extreme flow events (Miatke, 2015). Furthermore, this method dilutes the importance of the slope and intercept regression coefficients and potential insights gleaned from analysis of temporal variability embedded in those parameters across seasons or between river systems.

The magnitude and range of annual flux for each constituent using our optimized framework were within similar ranges to Medalie et al. (2014). Annual load estimates generated from our framework were slightly lower in certain cases; this was particularly noticeable with TSS, which differed by ~50,000 metric tons per year for Missisquoi. Methods of Smeltzer (et al., 2012) and Medalie (2014) adjusted their load estimates to reflect loading from the total watershed area to Lake Champlain, while our methods estimate loads from the portion of the basin upstream of the USGS sampling station, which may contribute to this difference between our estimates and theirs. In Missisquoi, the gaged drainage area (2205 km2) is ~98% of the total basin area (2,240 km2). The largest annual flux across all nutrients occurred in water year 2011, which reflects the impact on nutrient loading from large storm events, like Tropical Storm Irene (August) and multiple flooding events in April and May of that year, when peak flows exceeded Tropical Storm Irene. While the annual loadings fluctuate from year to year, there is also an increasing trend in DP flux over the 15-year period. The magnitude of this DP increase, however, is
low compared to the magnitude of the TSS loading. These annual load estimates have optimized the estimation of temporally-variable nutrient and sediment flux by aggregating loads computed from regressions on data sets from clustered time units. The largest nutrient/sediment load in any given year generally occurred during the “spring” time period (i.e., March and April) in the Missisquoi TP model; and it was, on average, more than 40% of the annual estimated load for the respective water year. In certain years, (e.g., 2014) the load estimates for March and April alone were ~45% of the annual water year load. When March-April-May loadings are combined, the total comprises more than 50% of the annual load, which is comparable to results from Rosenberg (2016), who found the spring load to be as high as 70% using a different method. Given these results, the spring snowmelt period is an important period of focus (i.e., a hot moment (McClain et al., 2003)) from an environmental stewardship perspective, as management practices to reduce loads during this specific time period might constitute a more cost-effective and perhaps less labor-intensive strategy. Vermont presently bans manure spreading, for instance, from Dec 15 through April 1 but perhaps the ending date should be more variable in light of sensitivity of that snowmelt period. While significant strides have been taken to reduce agricultural nonpoint source pollution (e.g., the implementation of soil, manure, and fertilizer management practices), agriculture remains one of the most significant sources of nonpoint source pollution (Edwards & Hooda, 2008; Hively et al., 2005). Insights gleaned from utilizing this framework, which allows one to identify periods and locations of disproportionate importance in driving nutrient loading to receiving waters, could help to
maximize the potential utility of limited resources toward suppression of pollutant loading across a wide range of applications and issues.

2.6 CONCLUSIONS AND RECOMMENDATIONS

Intra-annual variability in hydrologic regimes affects the rate of nutrient and sediment loading as a function of discharge. A new framework was developed to capture this variability using the slope and intercept coefficients from individual load-discharge regressions that were best fit by 16 years of historical data from nine basins draining to Lake Champlain; two of the nine basins were supplemented with water quality data from our 2012-2015 field work. The framework introduced here optimally clusters time periods for analysis that better fit linear models (i.e., provide individual watershed slope and intercept regression coefficients for each clustered time unit) to improve loading estimates. These coefficients were analyzed with a hierarchical cluster analysis that preserved the temporal order of the data to optimally cluster load-discharge relationships that may subsequently be used in an optimized annual model of loading. We introduced ordinal integer values to the time unit of interest (individual months in this case) to preserve the temporal sequence of the user-selected base time units. This is important for the cluster analysis because these data act as a surrogate for water/air temperature data, which were not available at our discharge measurement locations. Future work should consider replacing these ordinal integer data with continuous temperature or similar weather data to preserve a similar temporal sequence to the data. Optimized annual models appear to outperform traditional annual regression models for use in nutrient and sediment load estimates. Results suggest that the optimized annual regression models performed on
average ~20% better than traditional annual regression models using Nash-Sutcliffe efficiency as a global performance metric, and resulted in ~50% cumulative higher load estimates in a given year. The largest nutrient and sediment load estimates occurred during the “spring” unit of time and were typically more than 40% of the total annual estimated load in a given year. The optimized cluster of time representing “spring” varied for each watershed, suggesting that the spring snowmelt and runoff period may be defined differently across watersheds in the Lake Champlain Basin. “Spring” is a critical time in Vermont for nutrient and sediment loading and needs to be addressed as such for land management plans. Watersheds throughout the Lake Champlain Basin could be prioritized for management practices targeted toward reductions in loads based on which watersheds may be most vulnerable to climate change based on the shifting temporal clusters.

This work was motivated, in part, by the recent proliferation of high-resolution sensor data; and the application of this framework will become increasingly useful as these high-resolution “big data” sets come online and become publically available. The current framework is flexible, and can be used with daily mean flow data if continuous flow data are not available. It may also be used to create more targeted and cost-effective management strategies for improved aquatic health in rivers and lakes by analyzing the varying sensitivity of different months and determining time periods of importance to focus management plans, such as cover crops, buffers, manure application, etc. The coefficients from the optimized regressions also seem to reflect underlying watershed characteristics and future work should also explore a more in-depth analysis that leverages these relationships with similar cluster analysis methods over space and time.
2.7 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under VT-EPSCoR Grant No. EPS-1101317 on Research on Adaptation to Climate Change (RACC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank all researchers involved with RACC VT-EPSCoR grant and the UVM College of Engineering and Mathematical Sciences, with specific thanks to Braden Rosenberg for assistance with sampling and analysis. We also acknowledge the VTDEC for the long term water quality data set and USGS for continuous flow records.
2.8 REFERENCES


Research on Adaptation to Climate Change (RACC), 2015. VT-EPSCoR, http://epscor.w3.uvm.edu/2/node/30


Figure 2.1- Location of Watersheds Delineated to VT DEC Sampling Points. Figure created using data from VCGI: Vermont Center for Geographic Information. Available from http://vcgi.vermont.gov/opendata. *Basins excluded from this analysis due to limited water quality and discharge data.
Figure 2.2 - Annual linear regressions for log(Load Estimate) vs. log(15-Minute Discharge) for Missisquoi TP and Winooski TP. Examples show adjusted $R^2$, root mean square error (RMSE), and slope ($b_1$) and intercept ($b_0$) coefficients reported for each regression model.
Figure 2.3—(a) Framework for Selecting Optimal Combinations using Hierarchical Cluster Analysis with Missisquoi TP Example Data: (b) Number of Samples in Each Month 1990-2015 (c) Slope Coefficients for Each Unit of Time, and (d) Absolute Value of Intercept Coefficients for Each Unit of Time
Figure 2.4- Dendrogram of Monthly Clusters for Missisquoi TP Example from Figure 2.2
Figure 2.5- Traditional Annual Regressions, Traditional Spring Regressions, and Optimal Spring Regressions for log(Load Estimate) vs. log(Discharge) for Missisquoi TSS and TP Example with Adjusted $R^2$, RMSE, and slope and intercept coefficients reported for each regression.
Figure 2.6- Missisquoi TP Cumulative Loading in 2014 for Optimized vs. Simple Annual Models vs. 15-minute hydrograph
Figure 2.7- Missisquoi TP, DP, TN, and TSS annual flux from water years 2000-2015 using optimal seasonal models
Table 2.1 - Watershed Characteristics

<table>
<thead>
<tr>
<th>Watershed</th>
<th>Number of USGS Gage No.</th>
<th>Drainage Area at Gage (km²)¹</th>
<th>Drainage Area at Mouth (km²)¹</th>
<th>Q median (cfs)²</th>
<th>Q max (cfs)²</th>
<th>(Q95/Q5)²</th>
<th>Fraction of land use in 2001 (%)³</th>
<th>Urb.</th>
<th>Agr.</th>
<th>For.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lamoille</td>
<td>8</td>
<td>04292500</td>
<td>1,776</td>
<td>1,872</td>
<td>933</td>
<td>28,800</td>
<td>19</td>
<td>6</td>
<td>16</td>
<td>74</td>
</tr>
<tr>
<td>LaPlatte</td>
<td>6</td>
<td>04282795</td>
<td>115</td>
<td>137</td>
<td>21</td>
<td>2,520</td>
<td>82</td>
<td>12</td>
<td>7</td>
<td>77</td>
</tr>
<tr>
<td>Lewis</td>
<td>5</td>
<td>04282780</td>
<td>200</td>
<td>210</td>
<td>65</td>
<td>4,690</td>
<td>26</td>
<td>13</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td>Little Otter</td>
<td>4</td>
<td>04282550</td>
<td>148</td>
<td>189</td>
<td>24</td>
<td>1,750</td>
<td>30</td>
<td>8</td>
<td>23</td>
<td>65</td>
</tr>
<tr>
<td>Mettawee</td>
<td>1</td>
<td>04280450</td>
<td>482</td>
<td>1,097</td>
<td>185</td>
<td>16,800</td>
<td>34</td>
<td>8</td>
<td>16</td>
<td>66</td>
</tr>
<tr>
<td>Missisquoi</td>
<td>9</td>
<td>04294000</td>
<td>2,201</td>
<td>2,240</td>
<td>1150</td>
<td>32,200</td>
<td>95</td>
<td>8</td>
<td>45</td>
<td>42</td>
</tr>
<tr>
<td>Otter</td>
<td>3</td>
<td>04282500</td>
<td>1,626</td>
<td>2,444</td>
<td>970</td>
<td>6,570</td>
<td>14</td>
<td>10</td>
<td>49</td>
<td>37</td>
</tr>
<tr>
<td>Poultney</td>
<td>2</td>
<td>04280000</td>
<td>484</td>
<td>681</td>
<td>164</td>
<td>12,800</td>
<td>62</td>
<td>9</td>
<td>24</td>
<td>59</td>
</tr>
<tr>
<td>Winooski</td>
<td>7</td>
<td>04290500</td>
<td>2,703</td>
<td>2,759</td>
<td>1350</td>
<td>35,000</td>
<td>20</td>
<td>8</td>
<td>12</td>
<td>76</td>
</tr>
</tbody>
</table>

¹ Troy et al. (2007)
² Calculated from 15 Minute Discharge USGS (1990-2015)
³ Q95 is average daily discharge corresponding to exceedance probability, USGS (1990-2015)
⁴ UVM Spatial Analysis Lab (2001)

Table 2.2 - Period of record and source for TP, DP, TN, and TSS for all nine study watersheds

<table>
<thead>
<tr>
<th>Watershed</th>
<th>Number of Samples</th>
<th>Source</th>
<th>Period of Record</th>
</tr>
</thead>
</table>
### Table 2.3: Optimal clusters indicated by number in parentheses, shading, and vertical dividers from January to December (J to D) for the nine study watersheds partitioned by constituent TP, DP, TN, and TSS with optimal and simple annual performance metrics Adjusted $R^2$, RMSE, NSE, and Bias.

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>RMSE</th>
<th>NSE</th>
<th>Bias</th>
<th></th>
<th>RMSE</th>
<th>NSE</th>
<th>Bias</th>
<th></th>
<th>RMSE</th>
<th>NSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lamorille (6)</td>
<td>J F</td>
<td>M A M J J A S O N D</td>
<td>0.90</td>
<td>0.91</td>
<td>0.20</td>
<td>0.23</td>
<td>0.56</td>
<td>0.39</td>
<td>1.09</td>
<td>1.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LaPlatte (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lewis (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little Otter (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mattawa (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missisquoi (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otter (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poultney (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winnoski (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamorille (6)</td>
<td>J F</td>
<td>M A M J J A S O N D</td>
<td>0.89</td>
<td>0.90</td>
<td>0.17</td>
<td>0.20</td>
<td>0.84</td>
<td>0.70</td>
<td>1.04</td>
<td>1.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LaPlatte (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lewis (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little Otter (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mattawa (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missisquoi (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otter (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poultney (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winnoski (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamorille (6)</td>
<td>J F</td>
<td>M A M J J A S O N D</td>
<td>0.93</td>
<td>0.98</td>
<td>0.09</td>
<td>0.09</td>
<td>0.90</td>
<td>0.87</td>
<td>1.01</td>
<td>1.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LaPlatte (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lewis (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little Otter (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mattawa (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missisquoi (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otter (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poultney (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winnoski (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamorille (6)</td>
<td>J F</td>
<td>M A M J J A S O N D</td>
<td>0.84</td>
<td>0.92</td>
<td>0.28</td>
<td>0.30</td>
<td>0.66</td>
<td>0.37</td>
<td>1.15</td>
<td>1.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LaPlatte (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lewis (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little Otter (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mattawa (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missisquoi (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otter (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poultney (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winnoski (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winnoski (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.4- Simple, traditional, and optimal model performance metrics for spring months March, April and May; Adjusted $R^2$, RMSE, NSE, and Bias, partitioned by constituent TP, DP, TN, and TSS for each of the nine study watersheds

<table>
<thead>
<tr>
<th>TP</th>
<th>Adjusted $R^2$</th>
<th>RMSE</th>
<th>NSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
</tr>
<tr>
<td>Lamolii</td>
<td>0.91 0.88 0.86</td>
<td>0.23 0.20 0.19</td>
<td>0.40 0.59 0.59</td>
<td>1.03 1.06 1.06</td>
</tr>
<tr>
<td>LaPlatte</td>
<td>0.89 0.86 0.82</td>
<td>0.30 0.25 0.26</td>
<td>0.56 0.71 0.71</td>
<td>0.96 1.18 1.17</td>
</tr>
<tr>
<td>Lewis</td>
<td>0.91 0.87 0.89</td>
<td>0.25 0.24 0.23</td>
<td>0.81 0.67 0.63</td>
<td>1.12 1.10 1.07</td>
</tr>
<tr>
<td>Little Otter</td>
<td>0.90 0.88 0.91</td>
<td>0.24 0.19 0.17</td>
<td>0.71 0.71 0.72</td>
<td>0.84 1.06 1.04</td>
</tr>
<tr>
<td>Mettawee</td>
<td>0.86 0.83 0.84</td>
<td>0.25 0.22 0.21</td>
<td>0.49 0.64 0.52</td>
<td>1.00 1.08 1.05</td>
</tr>
<tr>
<td>Missisquoi</td>
<td>0.88 0.88 0.87</td>
<td>0.29 0.23 0.22</td>
<td>0.46 0.68 0.75</td>
<td>1.20 1.11 1.13</td>
</tr>
<tr>
<td>Otter</td>
<td>0.75 0.46 0.45</td>
<td>0.27 0.27 0.26</td>
<td>0.23 0.22 0.27</td>
<td>1.02 1.13 1.08</td>
</tr>
<tr>
<td>Poultney</td>
<td>0.90 0.82 0.83</td>
<td>0.27 0.25 0.24</td>
<td>0.47 0.56 0.57</td>
<td>1.07 1.11 1.09</td>
</tr>
<tr>
<td>Winooski</td>
<td>0.80 0.85 0.86</td>
<td>0.39 0.25 0.24</td>
<td>0.45 0.57 0.59</td>
<td>1.19 1.12 1.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DP</th>
<th>Adjusted $R^2$</th>
<th>RMSE</th>
<th>NSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
</tr>
<tr>
<td>Lamolii</td>
<td>0.90 0.88 0.84</td>
<td>0.20 0.14 0.12</td>
<td>0.86 0.90 0.90</td>
<td>0.89 1.03 1.03</td>
</tr>
<tr>
<td>LaPlatte</td>
<td>0.85 0.69 0.64</td>
<td>0.33 0.30 0.30</td>
<td>0.46 0.51 0.49</td>
<td>0.91 1.28 1.26</td>
</tr>
<tr>
<td>Lewis</td>
<td>0.89 0.88 0.87</td>
<td>0.21 0.15 0.15</td>
<td>0.96 0.97 0.97</td>
<td>0.86 1.03 1.02</td>
</tr>
<tr>
<td>Little Otter</td>
<td>0.90 0.89 0.90</td>
<td>0.25 0.15 0.14</td>
<td>0.74 0.79 0.82</td>
<td>0.77 1.04 1.03</td>
</tr>
<tr>
<td>Mettawee</td>
<td>0.87 0.82 0.80</td>
<td>0.23 0.19 0.18</td>
<td>0.66 0.70 0.69</td>
<td>0.89 1.07 1.06</td>
</tr>
<tr>
<td>Missisquoi</td>
<td>0.87 0.77 0.79</td>
<td>0.26 0.20 0.20</td>
<td>0.58 0.58 0.63</td>
<td>0.97 1.09 1.07</td>
</tr>
<tr>
<td>Otter</td>
<td>0.69 0.26 0.40</td>
<td>0.29 0.24 0.24</td>
<td>-0.20 0.08 0.12</td>
<td>0.93 1.11 1.07</td>
</tr>
<tr>
<td>Poultney</td>
<td>0.92 0.76 0.76</td>
<td>0.21 0.20 0.19</td>
<td>0.39 0.40 0.40</td>
<td>0.94 1.10 1.10</td>
</tr>
<tr>
<td>Winooski</td>
<td>0.80 0.84 0.84</td>
<td>0.27 0.15 0.14</td>
<td>0.75 0.80 0.81</td>
<td>1.08 1.04 1.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TN</th>
<th>Adjusted $R^2$</th>
<th>RMSE</th>
<th>NSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
</tr>
<tr>
<td>Lamolii</td>
<td>0.98 0.91 0.85</td>
<td>0.09 0.09 0.08</td>
<td>0.83 0.85 0.87</td>
<td>1.04 1.01 1.01</td>
</tr>
<tr>
<td>LaPlatte</td>
<td>0.97 0.90 0.92</td>
<td>0.17 0.15 0.16</td>
<td>0.74 0.74 0.76</td>
<td>0.91 1.04 1.02</td>
</tr>
<tr>
<td>Lewis</td>
<td>0.96 0.89 0.90</td>
<td>0.14 0.15 0.13</td>
<td>0.93 0.92 0.90</td>
<td>0.96 1.03 1.03</td>
</tr>
<tr>
<td>Little Otter</td>
<td>0.95 0.90 0.88</td>
<td>0.18 0.16 0.16</td>
<td>0.73 0.69 0.72</td>
<td>0.91 1.04 1.04</td>
</tr>
<tr>
<td>Mettawee</td>
<td>0.96 0.92 0.88</td>
<td>0.13 0.10 0.11</td>
<td>0.82 0.82 0.81</td>
<td>0.94 1.02 1.01</td>
</tr>
<tr>
<td>Missisquoi</td>
<td>0.95 0.91 0.81</td>
<td>0.13 0.12 0.13</td>
<td>0.83 0.83 0.82</td>
<td>0.98 1.02 1.02</td>
</tr>
<tr>
<td>Otter</td>
<td>0.88 0.38 0.47</td>
<td>0.15 0.18 0.17</td>
<td>0.08 0.20 0.19</td>
<td>1.04 1.05 1.04</td>
</tr>
<tr>
<td>Poultney</td>
<td>0.95 0.79 0.76</td>
<td>0.17 0.17 0.18</td>
<td>0.69 0.66 0.74</td>
<td>0.98 1.04 1.03</td>
</tr>
<tr>
<td>Winooski</td>
<td>0.79 0.84 0.78</td>
<td>0.20 0.11 0.13</td>
<td>0.70 0.81 0.81</td>
<td>1.00 1.02 1.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TSS</th>
<th>Adjusted $R^2$</th>
<th>RMSE</th>
<th>NSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
<td>Simple Traditional Optimal</td>
</tr>
<tr>
<td>Lamolii</td>
<td>0.92 0.81 0.65</td>
<td>0.30 0.29 0.32</td>
<td>0.36 0.70 0.70</td>
<td>1.24 1.14 1.14</td>
</tr>
<tr>
<td>LaPlatte</td>
<td>0.92 0.87 0.88</td>
<td>0.35 0.31 0.32</td>
<td>0.24 0.34 0.35</td>
<td>1.38 1.18 1.15</td>
</tr>
<tr>
<td>Lewis</td>
<td>0.90 0.81 0.66</td>
<td>0.39 0.37 0.39</td>
<td>0.89 -0.40 -2.50</td>
<td>1.63 1.24 1.16</td>
</tr>
<tr>
<td>Little Otter</td>
<td>0.83 0.73 0.76</td>
<td>0.39 0.35 0.32</td>
<td>0.26 0.38 0.40</td>
<td>1.39 1.24 1.23</td>
</tr>
<tr>
<td>Mettawee</td>
<td>0.87 0.85 0.79</td>
<td>0.36 0.29 0.31</td>
<td>0.36 0.19 0.35</td>
<td>1.23 1.17 1.14</td>
</tr>
<tr>
<td>Missisquoi</td>
<td>0.87 0.90 0.85</td>
<td>0.36 0.25 0.28</td>
<td>0.45 0.72 0.83</td>
<td>1.27 1.12 1.10</td>
</tr>
<tr>
<td>Otter</td>
<td>0.79 0.49 0.46</td>
<td>0.35 0.36 0.37</td>
<td>0.20 0.15 0.16</td>
<td>1.33 1.23 1.21</td>
</tr>
<tr>
<td>Poultney</td>
<td>0.88 0.76 0.72</td>
<td>0.38 0.35 0.34</td>
<td>0.33 0.35 0.36</td>
<td>1.33 1.22 1.22</td>
</tr>
<tr>
<td>Winooski</td>
<td>0.88 0.81 0.77</td>
<td>0.39 0.31 0.31</td>
<td>0.67 0.65 0.64</td>
<td>0.91 1.17 1.12</td>
</tr>
</tbody>
</table>

60
CHAPTER 3: THESIS CONCLUSIONS AND FUTURE WORK

A new framework was developed to capture temporal variability using the slope and intercept coefficients from individual load-discharge regressions that were best fit by 16 years of historical data from nine basins draining to Lake Champlain, and supplemented, for two of the nine basins, with water quality samples from our 2012-2015 field work. The importance of the slope and intercept coefficients should not be overlooked as they were critical for performing a hierarchical cluster analysis and clustering different time units. These coefficients may be further important for characterizing spatial differences across watersheds as well since they magnitudes differed significantly across our nine different watersheds. It is recommended for future work to analyze how changes in slope and intercept coefficients relate to changes in land use, soil characteristics, and climate change across these different watersheds. Other studies have considered the importance of land cover/land use, as well as temporal variability, but few have used regression coefficients to explain underlying watershed characteristics. The advantage of this framework is that it allows for accurate load estimates, similar to other studies mentioned previously, but also provides different slope and intercept coefficients for each time unit and watershed that can be further analyzed. Furthermore, the temporal order integer variable was created in this framework to preserve the temporal sequence of the user-selected base time units to optimally cluster similar load-discharge relationships. It is recommended in future work that the temporal order variable may be replaced with continuous temperature or similar weather data to maintain similar temporal sequence to the data. This will allow a climate based cluster analysis to be truly considered. Further applications of this framework
become increasingly useful with proliferation of high resolution data and “big” data sets. With increasing data sets, the varying sensitivity of time units can be better understood and help make informed decisions for practitioners. For instance, current data resolution shows sensitivity of month time units, which can help inform practitioners of particular times when loading is high/low.

Combined with knowledge of land cover/land use as well as existing land management applications, this framework can save large amounts of spending by better targeting “hot” moments in time for management plans. Future work can provide cost estimates for the amount saved when spending money on management plans for more significant time periods versus less significant ones and the impact cost of overall nutrient loading from poorly timed management plans. This work currently suggests that management plans in the months of April and May need to be re-examined as they are having the largest impact on waterways in Vermont. As higher resolution data continues to increase from more sensor networks, the already useful framework developed here will enable Vermont and other areas to more accurately estimate loads and create more fiscal and protective management plans for rivers and streams that will ultimately lead to improved water quality in receiving water bodies.
Appendix A: Matlab Script

% Lake Champlain Tributary Seasonal Load Estimation Code
% Baxter Miatke
% Master's Thesis Research Project July 2016
% Environmental Engineering, University of Vermont
% Advisor: Dr. Donna Rizzo

%% DESCRIPTION
% Historical (1990-present) water quality data (TP, DP, TN, and TSS) from
% VT-DEC is combined with RACC water quality data from Winooski and
% Missisquoi River. 15 minute continuous discharge data is obtained for all
% the Vermont tributary sites in the Lake Champlain Basin. Load Estimate
% vs. 15 minute Discharge regressions are created based on seasonal relationships.
% Regression coefficients are then used to calculate load estimates for each
% river. Performance metrics are calculated to evaluate each model.

%% MAJOR CODE UPDATES/NOTES
% 9/10/15 - Initial code created to read in nutrient/discharge data,
% difficulty reading in discharge data from USGS
% 9/14/15 - Code to read in nutrients from excel file
% 9/15/15 - Set up way to separate years months and days with ymd and datetimes
% 9/21/15 - Created concentration code to plot mean years and monthly
% concentration data - Started this code update section
% 9/25/15 - Set up url read for discharge data
% 9/29/15 - url read does not work, use webread, fixed ICE entry problem, now
% have full hydrograph for all years (this is daily average though and want
% 15 minute data. 15 minute data has been moved from pre-2007 off-site from
% USGS and will need to be obtained different way. Using average values for now...
% 10/6/15 - Started automating process and creating larger loops to analyze all
% nutrients and site locations
% 10/8/15 - Created Pre-Process code separate from main code to pre-process
% and read in all other nutrient spreadsheets from other locations
% 10/14/15 - Created Get_Dischage Function separate from main code to
% get discharge data and load it in for each location (still daily average)
% 10/28/15 - Regression code to plot conc vs. discharge and create fit through
% all the data. Still need to figure out how to separate by month/year.
% Began saving old versions of code here.
% 10/29/15 - Added discharge into concentration code that finds annual and
%montly discharge values - moving towards monthly and seasonal regressions
%11/3/15- Changed get_discharge to get_discharge15 for 15-minute data, only
%available from 2007 to present, historical data is moved off-site to
%IDA, added site numbers section. Getting 15 minute discharge data
%points is
%taking long amounts of time (15min*4*24hr*365day*8yr=4,204,800)
%11/5/15- Main code was updated to deal with 15 minute discharge data
%since the time now matters. The sample times from VT-DEC are in wrong format
%1030 or 900, wrote loop to correct to datetime format 10:30 or 09:00.
%Fixed min(y) to y(1) as there is more than 1 minimum year now with new
%dataset. Need to figure out how to round sample times to 15 minute
%discharge times...
%11/6/15- Created Input and Output folders to deal with all data being read
%in and saved
%11/9/15- Updated Get Discharge with historical discharge from IDA,
%optimized code for memory issue with large dataset
%11/10/15- fixed rounding issue-> created ifelse statement to change
%minutes to either 0,15,30,or 45, fixed zeros issue with matching discharge
%to sampledate, everything matches and have TP vs. Discharge graph, created
%plots and sent update to advisors
%11/16/15- Created new Pre-process code for Missisquoi specifically to
%combine VT-DEC and RACC data. Had to change code to deal with expanded
%number of samples
%11/23/15- All pre-processing complete- units fixed micrograms to mg
%11/29/15- Successfully ran load estimation to match results with yearly
%regressions. Created monthly regressions and grouped together. Still need
%to run load estimates with monthly regressions: issue with referencing
%datetimes
%12/11/15- Minor updates and set code up for easier running when return
%in January
%1/21/16- Preliminary running of code to refresh results and needs, some
%updates in comments
%1/25/16
%1/28/16- Huge update to code - re-ordered annual and monthly sections
%to make more sense. Added load variables into loops so coefficients could be
%determined for different permutations of months (Labeau 2015). Able to get
%coefficients for each different seasonal combination- need to decided
%which ones are significant to use and then do annual load estimates with those
%coefficients.
%2/3/16- Trying to code statistical significance for b0 and b1 coefficients
to define seasonality...difficulty with using the chi-square similar
equation from Lebeau 2015
%2/5/16- Prepped output runs with new output folder names. Finished chi
square test on coefficients. All significant- look at r^2, adjr^2, CI for
significance- noted difference in gof variable - fixed structures to
cells
%and indexing to proper statistics.
%2/12/16- Added 4 month combinations, graphing GOF statistics, plotting in
%creatFit function turned off to speed up code (too many combinations
being plotted
%2/22/16- Estimating loads using seasonal regressions, defined user input
%for seasonal combinations, is solving for every month 1990-2015 right
now
%Updates as of March 24th
%ADDED COMMENTS AND INSTRUCTIONS FOR USE
%Corrected unit conversion issue between mg/L and micrograms/L
%Corrected time period when summing seasonal flux periods (check clear
%function to clear proper variables)
%Write code to export tables and store things in easy output folders
%Plot annual and seasonal flux estimates- decided conc vs. load estimate
%regression (flux and flux2)
%Incomplete 15 minute discharge is a constraint- need to reconstruct?

% Updates as of April 7th
%15 minute Discharge recorded reconstructed for Missisquoi based on stage
%measurements
%Back Transform from log was changed to Newman 1993 method because
%residuals of each seasonal model were shown to be normally distributed -
%Duan Smearing is no longer used
%Code now calculates rmse, NSE, bias, and displays all results in cleaner
$table and output. Months size changed based on rmse instead of sse.
%Running other nutrients and sites to see how seasonality changes for each

%Updates as of April 11th
%Only one set of coefficients were actually being used to estimate the
%whole load variable. Corrected so each month corresponds to its proper
%coefficient when estimating loads. Started running other sites and
%parameters and realized need a flow chart for choosing best combinations
%so it is done consistently and avoid user bias.
%Updated code to deal with parameters/sites that have less than 2 samples
% in one month
% Dates are read in slightly different in winooski data, commented
% out for writing raw conc. data for time being

% May 5th, 2016 - Code ran for all sites successfully and exported figures
and
% data to excel for further analysis

% June 5th, 2016 - Checked natural log vs. log10 transformations. Same
results
% with loading. Reported b0 and b1 coefficients in log10 units for
% comparison to published literature

% June 13th, 2016 - Cumulative Load Estimates added for 2014 to see
% difference between optimized and simple models. March and August
% individual monthly metrics added to compare time scales i.e annual,
% spring, and march

% June 27th-30th, 2016 - Cluster analysis step added. Months are first
grouped
% into enough n samples for regression analysis. Hierarchical clustering
is
% done and user selects clusters of months based on dendrogram. Those
% clusters are used for load estimation instead of previous framework.

% July, 2016 Efficiency Added:
% The calculation of log10(x) takes about 6 times longer than the
calculation
% of log(x). Even stranger, the calculation of log10(x) takes about 4
times
% longer than log(x)/log(10), even though log10(x)=log(x)/log(10). In
short,
% we decided to trade increased computational time for better results
for numbers of the form 10^k.

clc, clear all, close all

%% PRE_PROCESS

% Read nutrient data from VT-ANR website, data needs to be pre-processed
and
% saved as excel file since url is not unique to read from web
% https://anrweb.vermont.gov/dec/dec/LongTermMonitoringTributary.aspx

% Use Pre_process code separately to update nutrient data as needed or
if missing
% Needs to be saved in current folder

% RACC data was added to Pre_Process_Missisquoi separately to deal with
% different formats, will need to be done for adding other data in
addition to VTDEC as well
% Location and Site Numbers
% Use these location and site names so files are read in properly (automated
% below)
% Missisquoi-04294000
% Winooski-04290500
% Lamoille-04292500
% LaPlatte-04282795
% Lewis-04282780
% LittleOtter-04282650
% Otter-04282500
% Rock-04294140
% Pike-04294300
% Mettawee-04280450
% Poultney-04280000

% USER SELECTS SITE FROM MENU OPTIONS
iprob=menu('Select a Site To Analyze', 'Missisquoi', 'Winooski', 'Lamoille', 'LaPlatte', 'Lewis', 'Little Otter', 'Otter', 'Rock', 'Pike', 'Mettawee', 'Poultney');

% DEFINE SITE NUMBER AND TITLE BASED ON SELECTION FROM MENU
if iprob==1;
    site_no='04294000';
    location='Missisquoi';
elseif iprob==2;
    site_no='04290500';
    location='Winooski';
elseif iprob==3;
    site_no='04292500';
    location='Lamoille';
elseif iprob==4;
    site_no='04282795';
    location='LaPlatte';
elseif iprob==5;
    site_no='04282780';
    location='Lewis';
elseif iprob==6;
    site_no='04282650';
    location='LittleOtter';
elseif iprob==7;
    site_no='04282500';
    location='Otter';
elseif iprob==8;
    site_no='04294140';
    location='Rock';
disp('Rock River does not have enough valid samples <100 and short discharge record 2010-2016, cannot use, will cause error when trying to do seasons')
elseif iprob==9;
    site_no='04294300';
    location='Pike';
elseif iprob==10;
    site_no='04280450';
    location='Mettawee';
elseif iprob==11;
    site_no='04280000';
    location='Poultney';
end

%% LOAD DISCHARGE DATA IN
%RUN Get_Discharge15 code seperately to get 15 minute discharge
%All 15 minute discharge should be already saved as Q_15 files from
%Get_Discharge15 code, this can be run again to get discharge up to the
%current date, but takes some time to re-run which is why it’s done
%seperately
D=load([pwd '/Input/',site_no,'Q_15']); %loads 15 minute discharge data
D=struct2cell(D);
date_q=D{1};
q=D{2};

f=1; %figure indexing to keep track of figure numbers throughout
figure(f)
f=f+1; %next figure would be f+1

plot(date_q,q) %Plots the hydrograph for the entire record for the site
title([site_no, 'Discharge (Continuous 15 Minute)'])
xlabel('Date')
ylabel('Discharge cfs')
saveas(gcf,[pwd '/Output/',location,'/', site_no,'Q_15.fig']); %save
hydrograph in output folder

%% Select Nutrient to Be Analyzed
%User selects nutrient to be analyzed
iprob=menu('Select Nutrient', 'TP','DP','TN','TSS');
TP='TP';
DP='DP';
TN='TN';
TSS='TSS';

if iprob==1
    para={TP}; %use cell {} notation to store as text to properly save all
    figure in correct folder
elseif iprob==2
    para={DP};
elseif iprob==3
    para={TN};
elseif iprob==4;
para={'TSS'};
end

%% Load Concentration Data In and Remove NAN and Negative Values

%for k=1:4 %Used to loop, replaced with menu options to Run all 4 nutrient analysis for each location and extract parameters from filename
k=1;%keep k as placeholder %clearing variables were only important when in the loop above clear year_Result avg_conc Month_result qsamp Conc qsamp_raw SampDateTime %Clear everything for next loop (next nutrient to be analyzed)

filename=[pwd '/Input/',location,'_',para{k}]; %defines current filename loop k: TP, DP, TN, TSS
F=load(filename,'VisitDate','Time','Test','Result'); %loads the correct nutrient file for current loop
F=struct2cell(F); %converts structure to cell
VisitDate=F{1}; %date of sample (use {} for cells)
Time=(F{2}); %time of sample - THE VT_DEC is not in datetime format, but 1030, 1045, etc.
Test=F{3}; %the test being done (TP, DP, TN, TSS)
Conc=F{4}; %the concentration measured on the date from F{1}

%Remove NAN data - this messes up indexing for each analyte.
%Use logistics to remove Nan and concentrations less than zero
exclude = Conc <= 0; %Remove concentrations less than or equal to zero, and associated times to keep matrices same size for indexing
Conc(exclude) = [];
Time(exclude) = [];
VisitDate(exclude) = [];

exclude_nan=isnan(Conc);
Conc(exclude_nan) = [];
Time(exclude_nan) = [];
VisitDate(exclude_nan) = [];

%% Correct DateTimes and Discharges to Proper Format
if strcmp(location,'Missisquoi') %string compare for Missisquoi and Winooski only
    tf=strcmp(location,'Winooski');
else
    %Other sites need to fix the time issue with sample times 1030->10:30
end

if tf==1||tf2==1 %if tf or tf2 = missisquoi or winooski location %nothing, this is corrected in pre-process code already to %combine RACC samples
else
    %Other sites need to fix the time issue with sample times 1030->10:30
end
for i=1:length(Time) %This loop fixes the 1030 time issue to 10:30
and stores it new samptime
    temp=num2str(Time(i)); %first converts to string
    if length(temp)<4
        temp=['0',temp(1),temp(2), temp(3)]; %This adds a 0 to the
times before 12:00 -> 900 becomes 09:00
    end
    temp2=[temp(1),temp(2),':',temp(3),temp(4)]; %adds : in the string
    samptime(i)=datetime(temp2,'Format','HH:mm'); %converts string to
datetime
end
    Time=samptime;
end

[y,m,d] = ymd(VisitDate); %returns the year, month, and day numbers of
the datetime values (sample dates)
[hr,minute,sec]=hms(Time);%return hour, minute, second of the sample
time
[y2,m2,d2]=ymd(date_q);%same ymd result, but for all discharge dates

%Define samptime sec to 00,15,30,or 45 so can be compared to 15 minute
USGS discharge
minute2=zeros(1,length(minute));
for i=1:length(minute)
    if 0<=minute(i) && minute(i)<=7
        minute2(i)=0;
    elseif 8<=minute(i) && minute(i)<=22
        minute2(i)=15;
    elseif 23<=minute(i) && minute(i)<=37
        minute2(i)=30;
    elseif 38<=minute(i) && minute(i)<=52
        minute2(i)=45;
    else
        minute2(i)=0;
    end
end

%Redefine samptime with correct rounded 15 minute values
for i=1:length(minute2)
    hrtemp=num2str(hr(i));
    mintemp=num2str(minute2(i));
    newminute=[hrtemp,':',mintemp];
    samptime2(i)=datetime(newminute,'Format','HH:mm');
end

samptime2=samptime2';
CnvtDT = @(VisitDate,samptime2) datetime([VisitDate.Year
VisitDate.Month VisitDate.Day samptime2.Hour samptime2.Minute 0],
'Format', 'yyyy-MM-dd-HH-mm');
%This converts the dates and times and adds them into one variable
samppdatetime to match the USGS discharge format

%Pre-allocating sampdatetime causes error due to datetime format
for i=1:length(VisitDate)
    SampDateTime(i) = CnvtDT(VisitDate(i),samptime2(i)); %Define all Sample DateTimes
end

%% Match Sample Dates to Discharge Values and Plot Samples on Hydrograph
for i=1:length(SampDateTime)
    index=find(date_q==SampDateTime(i)); %find associated discharges with the sample times
    if index>=1; %only store the discharges that exist, empty matrix error
        qsamp_raw(i)=q(index); %the discharge at each sample
    else
        qsamp_raw(i)=0;
    end
end

%Plot raw samples over hydrograph
figure(f)
f=f+1;
plot(date_q,q)
hold on
plot(SampDateTime,qsamp_raw,'o')
title([location, Test(1)])
xlabel('Date')
ylabel('Discharge cfs')
legend('15-minute Hydrograph', 'Water Quality Sample Dates')
saveas(gcf,[pwd '/Output/',location,'/',para{k},'/','Sample_Hydrograph',location,para{k},'..fig']);

%This shows that not all water quality samples have an associated discharge
%value due to equipment malfunction or error. If red circles do not fall on
%the hydrograph then they do not match discharge records. This is an error
%since not using average daily discharge, but "continuous" 15 minute discharge

%% Cumulative discharge to see patterns - removed

%Cumulative discharge took too long to plot and not useful for determining
%significant change in discharge from season to season - commented out

figure(f)
% f=f+1;
% yr_dis=q(1);
    % sum_date=date_q(1);
%
    % %FOLLOWING SUM LOOP TAKES A LONG TIME- need to separate or not run every %
    % time % for i=y2(1):1:max(y2) %Use min and max as year span changes between parameters
    % for j=2:length(y2)
    %   if y2(j-1)==i %Find all the parameters associated with one year
    %     yr_dis(j)=q(j-1)+yr_dis(j-1); %this sums the previous discharge, q(1) is defined earlier
    %     sum_date(j)=date_q(j);
    %   end
    % end
    % end
    % plot(sum_date,yr_dis); %plots cummulative discharge over time
%
    %% Match water samples with corrected discharge values and use annual regression to estimate loads
%
    %Match actual corrected discharge with samples to get final values
    count=1;
clear qsamp_final conc_final l_est_final month_samp SampDateTime_final
    for i=1:length(SampDateTime)
        index=find(date_q==SampDateTime(i)); %find where the dates of discharges for each sample
        if index>=1; %only store the discharges that exist, empty matrix error
            qsamp_raw(i)=q(index);
            if qsamp_raw(i)>0 %Ice entries appear as 0 so only store discharge greater than 0
                qsamp_final(count)=qsamp_raw(i);
                conc_final(count)=Conc(i);
                l_est_final(count)=(conc_final(count)*qsamp_final(count)*28.3168466*15*60)*(10^-6);
                SampDateTime_final(count)=SampDateTime(i);
                [y,m,d] = ymd(SampDateTime(i)); %find the month of the sample being stored
                month_samp(count)=m; %store the month of the sample for monthly analysis later
                count=count+1;
            end
        end
    end
    [c,gofc]=createFit2Plot(qsamp_final,conc_final);
f=f+1;

r2_annual=gofc.rsquare(1);
title(['location para{k}', 'Annual r^2= ', num2str(r2_annual)])
xlabel('Discharge (cfs)')
ylabel(['para{k}', ' (mg/L) '])
saveas(gcf,[pwd '/Output/',location,'/',para{k},'/',location ,para{k},' Conc vs. Discharge.fig']);
cf=coeffvalues(c);
c1=cf(1);
c2=cf(2);

%Concentration Estimates
for i=1:length(q) %For all q values
c_est(i)=c1*q(i)+c2; %concentration estimate for all q values
    if c_est(i)<0
        c_est(i)=0; %set conc to 0 if less than 0
    end
end

[c,gofc]=createFit1Plot(log10(qsamp_final), log10(l_est_final)); %annual linear regression fit of all data log=natural log
f=f+1; %create fit plots a figure

%c(x) = p1*q + p2 (conc=p1*flow+p2)
r2_annual=gofc.rsquare;
radj_annual=gofc.adjrsquare;
sse_annual=gofc.sse;
rmse_annual=gofc.rmse;
b1_annual=cf(1);
b0_annual=cf(2);
title(['location para{k}', 'Annual r^2= ', num2str(r2_annual), ', RMSE= ', num2str(rmse_annual)])
saveas(gcf,[pwd '/Output/',location,'/',para{k},'/',location ,para{k},' Log_Load vs. Discharge Fit.fig'])

%Find all load estimate values using coefficients from annual regression
for i=1:length(q) %For all q values
    l_est_yr(i)=(10^(b0_annual))*(q(i)^b1_annual);
    if l_est_yr(i)<0
        l_est_yr(i)=0; %set conc to 0 if less than 0
    end
end

%Load estimation example for 2014:
index_2014=find(y2==2014); %finds all the data for 2014
load2014_annual=l_est_yr(index_2014); %sum the l_est at the locations for 2014
load2014_annual_cum=cumsum(load2014_annual);

%ANNUAL REGRESSIONS CLOSELY MATCH RESULTS FROM PREVIOUS RESEARCH
(Miatke 2015)- checked that code works
%Create monthly regressions now that it works for annual estimates

%% MONTHLY ANALYSIS- Break data up into months
count=1; %counter
figure(f) %Set up figure for plotting outside of loop
f=f+1;
clear Month_q Month_conc Month_load %Seperate concentrations into months
for i=1:12 %12 months
clear count month_conc month_q month_load %clear variables for next loop iteration
month_conc(1)=NaN;
month_q(1)=NaN;
month_load(1)=NaN;
count=1;
for j=1:length(month_samp)
    if month_samp(j)==i %If month of sample equals current month,
then store the sample concentration from that date
    month_conc(count)=conc_final(j);
    month_q(count)=qsamp_final(j);
    month_load(count)=l_est_final(j);
    count=count+1;
end
    Month_conc{1,i}=month_conc; %store parameters from each month into cells
    Month_q{1,i}=month_q;
    Month_load{1,i}=month_load; %use later with load plot
    n=length(Month_q{1,i});
    subplot(4,3,i) %subplot for all 12 months on one figure
    plot(Month_q{1,i},Month_conc{1,i},'.') %Plot proper month concentration results
    title(["Month ", num2str(i), '(n= ', num2str(n),')]')
    hold on
end
end
end %overall subplot title
suptitle([location, Test(1)]);
saveas(gcf,[pwd '/Output/',location,'/',para(k),'','Month_Conc_',location,para(k),'.fig']);

%% FIND BEST FIT OF SEASON MONTHS (LaBeau 2014)

for i=1:12 %Perform log transform on data for each month discharge and load estimate
    Month_q_log(1,i)=log10(Month_q{i});
    Month_load_log(1,i)=log10(Month_load{i});
end
% 1 month seasons
for i=1:12 %FIND coefficients FOR EACH INDIVIDUAL MONTH DATA FIRST
    if length(Month_load_log(i))>=2;
        \[
        \text{Month}_q_{\text{log}}(1,i) = \log_{10}(\text{Month}_q(i))
        \]
        \[
        \text{Month}_\text{load}_{\text{log}}(1,i) = \log_{10}(\text{Month}_\text{load}(i))
        \]
[cp, gofc] = createFit1(Month_q_log{i}, Month_load_log{i});

% if = f+1;
% title(['Month ', num2str(i), ', R^2 = ', num2str(gofc.rsquare)])
gof1(:, i) = struct2cell(gofc); % sse, rsquare, dfe, adjrsquare, rmse
	n = length(Month_q_log{i});

cf = coeffvalues(cp);

b0_one(i) = cf(2);
b1_one(i) = cf(1);

else

gof1{1, i} = NaN;
gof1{2, i} = NaN;
gof1{3, i} = NaN;
gof1{4, i} = NaN;
gof1{5, i} = NaN;

end

end

% 2 month seasons
for i = 1:1:12 % (JAN FEB) (FEB MAR) (MAR APR) etc.

if i == 12
    Month_q2 = [Month_q_log{i} Month_q_log{1}]; % strattles 12 and 1
    Month_load2 = [Month_load_log{i} Month_load_log{1}];
else
    Month_q2 = [Month_q_log{i} Month_q_log{i+1}];
    Month_load2 = [Month_load_log{i} Month_load_log{i+1}];
end

tf = isnan(Month_load2(1));

if length(Month_load2) > 2 && tf == 0;
    [cp, gofc] = createFit1(Month_q2, Month_load2);
gof2(:, i) = struct2cell(gofc); % sse, rsquare, dfe, adjrsquare, rmse

    n = length(Month_q2);

cf = coeffvalues(cp);

b0_two(i) = cf(2);
b1_two(i) = cf(1);

else
    gof2{1, i} = NaN;
gof2{2, i} = NaN;
gof2{3, i} = NaN;
gof2{4, i} = NaN;
gof2{5, i} = NaN;

end

end

% 3 month seasons (JAN FEB MAR) (FEB MAR APR) (APR MAY JUN) etc.
for i = 1:12

if i == 11
    Month_q3 = [Month_q_log{i} Month_q_log{i+1} Month_q_log{1}];
    % strattles 11, 12 and 1

end

end
Month_load3=[Month_load_log{i} Month_load_log{12}]
Month_load_log{1}];
else
  i==12
  Month_q3=[Month_q_log{i} Month_q_log{1} Month_q_log{2}];
  %strattles 12, 1 and 2
  Month_load3=[Month_load_log{i} Month_load_log{1}]
else
  Month_q3=[Month_q_log{i} Month_q_log{i+1} Month_q_log{i+2}];
  %three months at a time
  Month_load3=[Month_load_log{i} Month_load_log{i+1}]
end
[cp,gofcp]=createFit1(Month_q3, Month_load3);

gof3(:,i)=struct2cell(gofcp); %sse, rsquare, dfe, adjrsquare, rmse
N(3,i)=length(Month_q3);
cf=coeffvalues(cp);
b0_three(i)=cf(2);
b1_three(i)=cf(1);
end

% 4 Month Seasons
for i=1:12
  if i==10
    Month_q4=[Month_q_log{i} Month_q_log{11} Month_q_log{12}]
    Month_q_log{1}]; %strattles 11, 12, 1 and 2
    Month_load4=[Month_load_log{i} Month_load_log{11}]
    Month_load_log{12}];
  elseif i==11
    Month_q4=[Month_q_log{i} Month_q_log{12} Month_q_log{1}]
    Month_q_log{2}];
    Month_load4=[Month_load_log{i} Month_load_log{12}]
    Month_load_log{1}];
  elseif i==12
    Month_q4=[Month_q_log{i} Month_q_log{1} Month_q_log{2}]
    Month_q_log{3}];
    Month_load4=[Month_load_log{i} Month_load_log{1}]
    Month_load_log{3}];
  else
    Month_q4=[Month_q_log{i} Month_q_log{i+1} Month_q_log{i+2}]
    Month_q_log{i+3}]; %four months at a time
    Month_load4=[Month_load_log{i} Month_load_log{i+1}]
    Month_load_log{i+2}];
  end
  [cp,gofcp]=createFit1(Month_q4, Month_load4);
  gof4(:,i)=struct2cell(gofcp); %sse, rsquare, dfe, adjrsquare, rmse
  N(4,i)=length(Month_q4);
cf=coeffvalues(cp);
b0_four(i)=cf(2);
b1_four(i)=cf(1);
end
save(['pwd'/Output/','/','para(k)'/','Month_load_log',location,para{k},' .mat'],['Month_load_log'])

%% Test coefficients signficance using two tail test against value of 0
matb0(1,:)=b0_one; %one month combinations
matb0(2,:)=b0_two; %two month combinations
matb0(3,:)=b0_three; %three month combinations
matb0(4,:)=b0_four; %four month combinations

matb1(1,:)=b1_one;
matb1(2,:)=b1_two;
matb1(3,:)=b1_three;
matb1(4,:)=b1_four;

save(['pwd'/Output/','/','para(k)'/,'b0_','location,para{k},' .mat'],'matb0')
save(['pwd'/Output/','/','para(k)'/,'b1_','location,para{k},' .mat'],'matb1')

%% Analyze Coefficients
figure(f)
f=f+1;
xbar=(1:1:12);
bar(xbar,b1_one)
title('b1 coefficients')
ylabel('b1')
xlabel('Month')
N1=N(1,:); %take first month combinations just to get N for each month
for i1=1:numel(N1) %label top of each bar graph
    text(xbar(i1),b1_one(i1),num2str(b1_one(i1),'%0.001f'),... 
        'HorizontalAlignment','center',... 
        'VerticalAlignment','bottom')
end

figure(f)
f=f+1;
xbar=(1:1:12);
b0_one_abs=abs(b0_one);
bar(xbar,b0_one_abs)
title('b0 coefficients')
ylabel('abs(b0)')
xlabel('Month')
for i1=1:numel(N1) %label top of each bar graph
    text(xbar(i1),b0_one_abs(i1),num2str(b0_one_abs(i1),'%0.001f'),... 
        'HorizontalAlignment','center',... 
        'VerticalAlignment','bottom')
end

%% Normalize coefficients by number of samples
Nsum=sum(N1);
Nweight=N1./Nsum;
b0_weight=b0_one_abs.*Nweight;
b1_weight=b1_one.*Nweight;

figure(f)
f=f+1;
xbar=(1:1:12);
bar(xbar,b1_weight)
title('Slope coefficients normalized by number of samples')
ylabel('b1*(N/Nmax)')
xlabel('Month')

for i=1:numel(N1) %label top of each bar graph
    text(xbar(i),b1_weight(i),num2str(b1_weight(i),'%0.001f'),...'
        'HorizontalAlignment','center',...
        'VerticalAlignment','bottom')
end

%% T-Test
for i=1:4 %1,2,3,4 month combinations
    for j=1:12 %12 coefficients in each combination
        [h(i,j),p(i,j)] = ttest(matb1(i,:),0); %null hypothesis that the pairwise difference between data vectors x and y has a mean equal to zero default at alpha=0.05
        if h(i,j)==0
            disp('h=0 does not reject null hypothesis')
        else
            disp('h=1 rejects null hypothesis')
        end
    end
end

%% Look at Goodness Of Fit parameters of each regression to decide seasons
num_month=(1:12);
GOF{1,1}=cell2mat(gof1);
GOF{1,2}=cell2mat(gof2);
GOF{1,3}=cell2mat(gof3);
GOF{1,4}=cell2mat(gof4);
%Need 12 different variables for bar graph plotting
for i=1:4 %1,2,3,4 month combinations
    sse1(i)=GOF{1,i}(1,1); %sum of squares due to error
    sse2(i)=GOF{1,i}(1,2);
    sse3(i)=GOF{1,i}(1,3);
    sse4(i)=GOF{1,i}(1,4);
    sse5(i)=GOF{1,i}(1,5);
    sse6(i)=GOF{1,i}(1,6);
    sse7(i)=GOF{1,i}(1,7);
    sse8(i)=GOF{1,i}(1,8);
    sse9(i)=GOF{1,i}(1,9);
    sse10(i)=GOF{1,i}(1,10);
    sse11(i)=GOF{1,i}(1,11);
    sse12(i)=GOF{1,i}(1,12);

    rsq1(i)=GOF{1,i}(2,1); %rsquared
    rsq2(i)=GOF{1,i}(2,2);

end
rsq3(i)=GOF{1,i}(2,3);
rsq4(i)=GOF{1,i}(2,4);
rsq5(i)=GOF{1,i}(2,5);
rsq6(i)=GOF{1,i}(2,6);
rsq7(i)=GOF{1,i}(2,7);
rsq8(i)=GOF{1,i}(2,8);
rsq9(i)=GOF{1,i}(2,9);
rsq10(i)=GOF{1,i}(2,10);
rsq11(i)=GOF{1,i}(2,11);
rsq12(i)=GOF{1,i}(2,12);

radj1(i)=GOF{1,i}(4,1);%rsquared adjusted
radj2(i)=GOF{1,i}(4,2);
radj3(i)=GOF{1,i}(4,3);
radj4(i)=GOF{1,i}(4,4);
radj5(i)=GOF{1,i}(4,5);
radj6(i)=GOF{1,i}(4,6);
radj7(i)=GOF{1,i}(4,7);
radj8(i)=GOF{1,i}(4,8);
radj9(i)=GOF{1,i}(4,9);
radj10(i)=GOF{1,i}(4,10);
radj11(i)=GOF{1,i}(4,11);
radj12(i)=GOF{1,i}(4,12);

rmse1(i)=GOF{1,i}(5,1);%rmse
rmse2(i)=GOF{1,i}(5,2);
rmse3(i)=GOF{1,i}(5,3);
rmse4(i)=GOF{1,i}(5,4);
rmse5(i)=GOF{1,i}(5,5);
rmse6(i)=GOF{1,i}(5,6);
rmse7(i)=GOF{1,i}(5,7);
rmse8(i)=GOF{1,i}(5,8);
rmse9(i)=GOF{1,i}(5,9);
rmse10(i)=GOF{1,i}(5,10);
rmse11(i)=GOF{1,i}(5,11);
rmse12(i)=GOF{1,i}(5,12);

end

%% Plot Monthly Combinations GOF Parameters to Have User Define
%% Seasons
figure(f)
% f=f+1;
% bar([sse1;sse2;sse3;sse4;sse5;sse6;sse7;sse8;sse9;sse10;sse11;sse12]);
% bar graph of 12 months with each combination
% reline(0,sse_annual)
% legend('1 Month Seasons','2 Month Seasons','3 Month Seasons','4 Month
% Seasons');
% title(['Sum of Squares Due to Error',location,Test(1)]);
% xlabel ('Start Month');
% saveas(gcf,[pwd '/Output/',location,'/',para{k},'/','Seasonal_Sum_Square_Errors_',location,para{k},'_.fig']);
% figure(f)
% f=f+1;
% bar([rsq1;rsq2;rsq3;rsq4;rsq5;rsq6;rsq7;rsq8;rsq9;rsq10;rsq11;rsq12])
% reline(0,r2_annual)
% legend('1 Month Seasons','2 Month Seasons','3 Month Seasons','4 Month Seasons')
% title(['R Squared Coefficient of Determination',location,Test(1)])
% xlabel('Start Month')
% saveas(gcf,[pwd '/Output/',location,'/',para{k},'/','Seasonal_R_Squared_',location,para{k},'_.fig']);
% figure(f)
% f=f+1;
% bar([radj1;radj2;radj3;radj4;radj5;radj6;radj7;radj8;radj9;radj10;radj11;radj12]);
% reline(0,radj_annual)
% legend('1 Month Seasons','2 Month Seasons','3 Month Seasons')
% title(['Adjusted R-Square',location,Test(1)]);
% xlabel('Start Month')
% saveas(gcf,[pwd '/Output/',location,'/',para{k},'/','Seasonal_R_Squared_Adjusted_',location,para{k},'_.fig']);
% figure(f)
% f=f+1;
% bar([rmse1;rmse2;rmse3;rmse4;rmse5;rmse6;rmse7;rmse8;rmse9;rmse10;rmse11;rmse12])
% reline(0,rmse_annual)
% legend('1 Month Seasons','2 Month Seasons','3 Month Seasons')
% title(['Root Mean Square Error (RMSE)',location,Test(1)]);
% xlabel('Start Month')
% saveas(gcf,[pwd '/Output/',location,'/',para{k},'/','Root_Mean_Square_Error',location,para{k},'_.fig'])
%

figure(f)
f=f+1;
xbar=[1:1:12];
bar(xbar,N1);
title(['Number of Samples (1990-2015) in each month', Test(1)])
xlabel('Month')
grid on
for i1=1:numel(N1) %label top of each bar graph
text(xbar(i1), N1(i1), num2str(N1(i1), '%1.0f'), 'HorizontalAlignment', 'center', 'VerticalAlignment', 'bottom')
end
%
% saveas(gcf, [pwd '/Output/', location, '/', para{k}, '/','Number_of_Samples_Per_Month_', location, para{k}, '.fig']);

%% Cluster Analysis
count=0;
count2=1;
clear cluster cluster_b0 cluster_b1

while count<78 %number of seasonal periods to run through (may need to change) (sums to 78)
prompt = 'Enter the months you want grouped with enough n samples for cluster analysis: (i.e. [12,2,1]) '; %this will be done p number of times by user
index=input(prompt);
    count=(sum(index)+count);
i=length(index);
    order(count2)=mean(btv_temp(index));
if i==1
    Month_q_f{l}=Month_q_log{index(1)};
    Month_load_flog=Month_load_log{index(1)};
elseif i==2
    Month_q_f{l}=[Month_q_log{index(1)} Month_q_log{index(2)}];
    Month_load_flog=[Month_load_log{index(1)} Month_load_log{index(2)}];
    Month_load_log{index(2)};
elseif i==3
    Month_q_f{l}=[Month_q_log{index(1)} Month_q_log{index(2)} Month_q_log{index(3)}];
    Month_load_flog=[Month_load_log{index(1)} Month_load_log{index(2)} Month_load_log{index(3)}];
elseif i==4
    Month_q_f{l}=[Month_q_log{index(1)} Month_q_log{index(2)} Month_q_log{index(3)} Month_q_log{index(4)}];
    Month_load_flog=[Month_load_log{index(1)} Month_load_log{index(2)} Month_load_log{index(3)} Month_load_log{index(4)}];
end
[cl,gofcl]=createFit1(Month_q_f{l}, Month_load_flog);
cf=coeffvalues(cl);
cluster_b0(count2)=cf(2);
cluster_b1(count2)=cf(1);
count2=count2+1;
end
order=(1:length(cluster_b0));

figure(f)
f=f+1;
cluster_b0=abs(cluster_b0);
bar(order,cluster_b0)
title('b0 coefficients')
ylabel('abs(b0)')
xlabel('Month')

for il=1:numel(cluster_b0) %label top of each bar graph
text(xbar(il),cluster_b0(il),num2str(cluster_b0(il),'%0.001f'),...  
'HorizontalAlignment','center',...
'VerticalAlignment','bottom')
end

figure(f)
f=f+1;
bar(order,cluster_b1)
title('b1 coefficients')
ylabel('b1')
xlabel('Month')

for il=1:numel(cluster_b1) %label top of each bar graph
text(xbar(il),cluster_b1(il),num2str(cluster_b1(il),'%0.001f'),...  
'HorizontalAlignment','center',...
'VerticalAlignment','bottom')
end

cluster=[cluster_b0', cluster_b1',order'];
Z = linkage(cluster,'ward');
%% Dendrogram
figure(f)
f=f+1;
[H, T] = dendrogram(Z);
% get the handle of the axis
hAxis = get(H(1), 'parent');
% Get the permutation of the nodes
perm=str2num(get(hAxis,'XtickLabel'));
iprob=menu('Select groups that were clustered by n for dendrogram labels', 'Dec-Jan-Feb', 'Dec, Jan-Feb', 'Dec-Jan, Feb', 'Manually Enter Labels');
% label data
if iprob==1
    labels =
    ['Mar','April','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec-Jan-Feb'];
elseif iprob==2
    labels =
    ['Mar','April','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec','Jan-Feb'];
elseif iprob==3
    labels =
    ['Mar','April','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec-Jan','Feb'];
82
elseif iprob==4
    prompt = 'Enter All Month Labels as string: (i.e. {''Mar'',''Apr}'
    %this will be done p number of times by user
    labels=input(prompt);
end

% Create the XTickLabels
set(hAxis,'XTickLabel',labels(perm))
saveas(gcf,[pwd
'/Output/',location,'/',para(k),'','Dendrogram_','location,para(k),''.fig']);

%% User Selects Best Monthly/Seasonal Regressions Based on Dendrogram Clusters
%Organize regressions by month for conc and dis (manually selected
%off of monthly conc and discharge above
clear season_b0 season_b1 season_load_month season_year season_month
season_load_est...
    season_load2 season_load_est2 Month_load_f1 Month_conc_f Month_q_f
    Month_q_fl'
disp('Use Dendrogram to determine proper cluster of seasons')
prompt = 'How many total clusters of months do you want to use for Load Estimation = '
    p= input(prompt);
    for i=1:p %number of seasonal periods to run through
        clear season_load season_year season_month season_load2
        season_load month season_load_month 2 Month_q_f Month_q_fl Month_conc_f
        Month_load_flog Month_load_f residuals
        prompt = 'Enter a cluster of months: (i.e. [1,2,3]) ' ;
        %this will be done p number of times by user
index=input(prompt);
    ni=length(index);
    str_month{1,i}=(index);
    str{i}=[num2str(index(1)), ' to ',num2str(index(ni))];
    if ni==1
        Month_q_f=Month_q{index(1)};
        Month_conc_f=Month_conc{index(ni)};
    elseif ni==2
        Month_q_f=[Month_q{index(1)} Month_q{index(2)}];
        Month_conc_f=[Month_conc{index(1)} Month_conc{index(2)}];
    elseif ni==3
        Month_q_f=[Month_q{index(1)} Month_q{index(2)}
Month_q{index(3)}];
        Month_conc_f=[Month_conc{index(1)} Month_conc{index(2)}
Month_conc{index(3)}];
    elseif ni==4
        Month_q_f=[Month_q{index(1)} Month_q{index(2)}
Month_q{index(3)}] Month_q{index(4)}];
        Month_conc_f=[Month_conc{index(1)} Month_conc{index(2)}
Month_conc{index(3)}] Month_Conc{index(4)}];
end
[cp,gofcp]=createFit2Plot(Month_q_f, Month_conc_f);
f=f+1;
title(['Seasonal fit ', num2str(index), ' R^2= ', num2str(gofcp.rsquare), ', RMSE= ',num2str(gofcp.rmse)]);
saveas(gcf,[pwd '/Output/','location,'/','para{k}','/Seasonal Combinations/','Seasonal Fit Conc', str{i},' .fig']);

cf=coeffvalues(cp);
season_b0(i)=cf(2);
season_b1(i)=cf(1);

if ni==1
    Month_q_fl=Month_q_log{index(1)};
elseif ni==2
    Month_q_fl=[Month_q_log{index(1)} Month_q_log{index(2)}];
    Month_load_flog=[Month_load_log{index(1)} Month_load_log{index(2)}];
elseif ni==3
    Month_q_fl=[Month_q_log{index(1)} Month_q_log{index(2)} Month_q_log{index(3)}];
    Month_load_flog=[Month_load_log{index(1)} Month_load_log{index(2)} Month_load_log{index(3)}];
elseif ni==4
    Month_q_fl=[Month_q_log{index(1)} Month_q_log{index(2)} Month_q_log{index(3)} Month_q_log{index(4)}];
    Month_load_flog=[Month_load_log{index(1)} Month_load_log{index(2)} Month_load_log{index(3)} Month_load_log{index(4)}];
end

[cp,gofcp,fitoutput]=createFit1Plot(Month_q_fl, Month_load_flog);
f=f+1;
R2season(i)=gofcp.adjrsquare; %adjusted r-square
for a=1:length(index) %pull adj R2 for august
    if index(a)==8
        adjR2_aug_opt=gofcp.adjrsquare;
    end
end
rmse_opt(i)=gofcp.rmse; %root mean square error for each optimal combination
residuals=fitoutput.residuals;
title(['Seasonal Fit Load Estimate', num2str(index), ' R^2= ', num2str(gofcp.rsquare), ', RMSE= ',num2str(gofcp.rmse)]);
saveas(gcf,[pwd '/Output/','location,'/','para{k}','/Seasonal Combinations/','Seasonal Fit Load', str{i},' .fig']);

cf=coeffvalues(cp);
season_b0_log(index)=cf(2);
season_b1_log(index)=cf(1);
save([pwd '/Output/','location,'/','para{k}','/Season Coefficients'],'season_b0_log','season_b1_log');
%% Smearing estimator for backtransform log(Newman1993)
%Check if residuals are normally distributed from fit above
%rnormality(i)=ttest(residuals);  ttest- the null hypothesis that
%the data in x comes from a normal distribution with mean equal to zero
%and unknown variance at 5% level
[Hn, pValue, W]=swtest(residuals, 0.05);  
rnormality(i)=Hn;
pnormality(i)=pValue;
W(i) = W;

if rnormality(i)==0;  %rejects null
  isnormal='No';
else
  isnormal='Yes';
end

figure(f)
f=f+1;
histogram(fitoutput.residuals)
title(:,:, 'Residuals of Log-Log Fit ', num2str(index), ', Normality= ',
isnormal)

saveas(gcf,[pwd '/Output/',location,'/para{k}', '/Seasonal
Combinations/','Residuals ', str{i},'.fig']);

%Can use this method if residuals are normally distributed
mse=rmse_opt(i)^2;  %mean square error (use rmse calculated from
previous fit)
error(index)=10^(mse/2);  %bias from log transform

%% Estimate Seasonal Loads
qlog=log10(q);
for j=1:length(q)
  season_load(j)=(10^season_b0(i))+q(j)*(season_b1(i))*q(j)*28.316466*  
15*60*(10^-6);  %LoadEst=b0+(b1*q)

  if season_load(j)<0
    season_load(j)=0;  %set load to min value closest to 0
  end

  season_load2(j)=error(index(1))*(10^season_b0_log(index(1)))*(q(j)^season_b1_log(index(1)));
  if season_load2(j)<0
    season_load2(j)=0;  %set load to min value closest to 0
  end
end

%Grab seasonal coefficients for linear regressions at a particular
% combination specified by index
count=1;
count2=1;
while count2<=(length(index))  %runs for number of months input,
for r=index(count2) %the month number [1,2,3 etc.]
    for j=1:length(m2) %for all the months (the m extracted from SampDatetime)
        if m2(j)==r %if the month is equal to the month input then
            let's calc loads for that month using the corresponding coefficients
            season_load_month(count)=season_load(j);
            season_load_month2(count)=season_load2(j);
            season_year(count)=y2(j); %store years since season_load calcs all years
            season_month(count)=m2(j); %store months
            count=count+1;
        end
    end
    count2=count2+1;
end

for s=1:1:26 %26 years 1990-2015
    clear load_yr
    for w=1:length(index) %find month entered from input
        if 10<=index(w) && index(w)<=12;
            load_yr=find(season_year==(s+1988)); %finds all the data before water year (OCT-DEC)
        else
            load_yr=find(season_year==(s+1989)); %WATER YEAR (JAN-SEP)
        end
    end
    if load_yr>0
        season_load_est(i,s)=sum(season_load_month(load_yr)); %sum all data in that year
        season_load_est2(i,s)=sum(season_load_month2(load_yr));
    else
        season_load_est(i,s)=0; %Stores in matrix
        season_load_est2(i,s)=0;
    end
end
% Need final optimized load estimate that is a continuous estimate...not just seperated seasons, easier to recalc, rather than try to piece back together
for j=1:length(m2)
    num=m2(j); %find out the month you are in and use appropriate error and coefficients for that month to estimate a load
    load_final(j)=error(num)*(10^season_b0_log(num))*(q(j)^(season_b1_log(num)));
end
%USE LOAD_FINAL FOR NSE annual optimal comparison to annual simple estimates l_est_final_yr

% Cumulative Load estimation for 2014:
load2014_optimal=(season_load2(index_2014(1):index_2014(length(index_2014))));
load2014_opt_cum=cumsum(load2014_optimal);
date_2014=(date_q(index_2014(1):index_2014(length(index_2014))));
q_2014=(q(index_2014(1):index_2014(length(index_2014))));

figure(f)
f=f+1;
[ax,p1,p2]=plotyy(date_2014,load2014_opt_cum,date_2014,q_2014);
set(ax,{'
ycolor'
},{'
';;'b'
})
xlabel('Time')
ylabel(ax(1),'Cummulative Load(mt/yr)') % label left y-axis
ylabel(ax(2),'Discharge(cfs)') % label right y-axis
p1.LineStyle='--';
p1.LineWidth = 2.5;
p1.Color='r';
p2.LineStyle = ':';
p2.Color='b';

hold(ax(1))
plot(ax(1),date_2014,load2014_annual_cum,'r--','LineWidth',2.5)
set(gca,'box','off')
legend('Optimized Annual Model','Simple Annual Model','15-Minute
Hydrograph')
title([location,' Cumulative Loading'])
saveas(gcf,[pwd '/Output/',location,'/',para{k},'Cumulative_Load_Est',location,para{k},'.fig']);

%% Traditional spring (Mar, Apr, May) R^2
spring_q_fl=[Month_q_log{3} Month_q_log{4} Month_q_log{5}];
spring_load_flog=[Month_load_log{3} Month_load_log{4} Month_load_log{5}];
[sc, gofsp, spfit]=createFit1Plot(spring_q_fl, spring_load_flog);
f=f+1;
title(['Simple Spring Fit Load Estimate R^2= ',
num2str(gofsp.rsquare)])
R2spring_trad=gofsp.rsquare;
adjR2spring_trad=gofsp.adjrsquare;
sse_spring_trad=gofsp.sse;
rmse_spring_trad=gofsp.rmse;
sp=coeffvalues(sc);
b1_spring_trad=sp(1);
b0_spring_trad=sp(2);

%Can use this method if residuals are normally distributed (Newman93)
mse_spring_trad=rmse_spring_trad^2; %mean square error (use rmse calculated from previous fit)
error_spring_trad=10^(mse_spring_trad/2); %bias from log transform

count=1;
for j=1:length(q)
spring_load_trad(count)=error_spring_trad*(10^b0_spring_trad)*(q(j)^b1_spring_trad);
    count=count+1;
end

%% Do same for March and August as individual months to compare to
[fm, gofm, mfit]=createFit1Plot(Month_q_log{3},Month_load_log{3});
f=f+1;
title(['March Fit Load Estimate R^2= ', num2str(gofm.rsquare),', ', RMSE= ',num2str(gofm.rmse)])
R2march_trad=gofm.rsquare;
adjR2march_trad=gofm.adjrsquare;
sse_march_trad=gofm.sse;
rmse_march_trad=gofm.rmse;
mc=coeffvalues(fm);
b1_march_trad=mc(1);
b0_march_trad=mc(2);

%Can use this method if residuals are normally distributed (Newman1993)
mse_march_trad=rmse_march_trad^2; %mean square error (use rmse calculated from previous fit)
error_march_trad=10^(mse_march_trad/2); %bias from log transform

count=1;
for j=1:length(q)
    march_load_trad(count)=error_march_trad*(10^b0_march_trad)*(q(j)^b1_march_trad);
    count=count+1;
end

%August
[fa, gofa, afit]=createFit1Plot(Month_q_log{8},Month_load_log{8});
f=f+1;
title(['August Fit Load Estimate R^2= ', num2str(gofa.rsquare),', ', RMSE= ',num2str(gofa.rmse)])
R2aug_trad=gofa.rsquare;
adjR2aug_trad=gofa.adjrsquare;
sse_aug_trad=gofa.sse;
rmse_aug_trad=gofa.rmse;
ac=coeffvalues(fa);
b1_aug_trad=ac(1);
b0_aug_trad=ac(2);

%Can use this method if residuals are normally distributed (Newman1993)
mse_aug_trad=rmse_aug_trad^2; %mean square error (use rmse calculated from previous fit)
error_aug_trad=10^(mse_aug_trad/2); %bias from log transform

count=1;
for j=1:length(q)

...
aug_load_trad(count) = error_aug_trad*(10^b0_aug_trad)*(q(j)^b1_aug_trad);
        count = count + 1;
    end

    %% Save Coefficients from each month combination
    clear coeff
    coeff(1,:) = season_b0_log;
    coeff(2,:) = season_b1_log;
    coeff(1,13) = mean(season_b0_log);
    coeff(1,14) = b0_annual;
    coeff(2,13) = mean(season_b1_log);
    coeff(2,14) = b1_annual;

    % Coefficients in log10 units to compare to Labeau (2014)
    coeff_log10 = log10(exp(coeff));

    rname = {'b0'; 'b1'};
    vname = {'Jan'; 'Feb'; 'Mar'; 'Apr'; 'May'; 'Jun'; 'Jul'; 'Aug'; 'Sep'; 'Oct'; 'Nov'; 'Dec'; 'Average Seasonal'; 'Annual'};
    te = table(coeff(:,1), coeff(:,2), coeff(:,3), coeff(:,4), coeff(:,5), ...
              coeff(:,6), coeff(:,7), coeff(:,8), coeff(:,9), coeff(:,10), coeff(:,11), coeff(:,12), coeff(:,13), coeff(:,14), ...
              'VariableNames', vname, 'RowNames', rname);
    if exist([pwd '/Output/','location','/','para{k}','/Season Coefficients',location,para{k},'xlsx'], 'file') == 2
        delete([pwd '/Output/','location','/','para{k}','/Season Coefficients',location,para{k},'xlsx']);
    end
    disp(te);
    writetable(te, [pwd '/Output/','location','/','para{k}','/Season Coefficients',location,para{k},'xlsx'], 'WriteRowNames', true);

    %% Setup Seasonal Loads For Table and Percentages
    for s = 1:1:26
        season_load_est(i+1,s) = sum(season_load_est(:,s)); % add row of sum for year
        season_load_est(i+2,s) = season_load_est(i+1,s)/1000;
    end
    season_load_est2(i+1,s) = sum(season_load_est2(:,s)); % add row of sum for year
    season_load_est2(i+2,s) = season_load_est2(i+1,s)/1000;
    for i = 1:p
        season_percent(i,s) = (season_load_est(i,s)/season_load_est(p+1,s))*100; % Percentage of annual
        season_percent2(i,s) = (season_load_est2(i,s)/season_load_est2(p+1,s))*100; % Percentage of annual
    end

    % SETUP RESULTS TABLE TO DISPLAY IN MATLAB
str(i+1)="'Annual sum (kg/yr)';
str(i+2)="'Annual sum (metric tons/yr)';
variablynames=['Water_Year_1990';'Water_Year_1991';'Water_Year_1992';'
Water_Year_1993';...'

'Water_Year_1994';'Water_Year_1995';'Water_Year_1996';'Water_Year_1997'
;'
Water_Year_1998';'Water_Year_1999';...'

'Water_Year_2000';'Water_Year_2001';'Water_Year_2002';'Water_Year_2003'
;'
Water_Year_2004';'Water_Year_2005';...'

'Water_Year_2006';'Water_Year_2007';'Water_Year_2008';'Water_Year_2009'
;'
Water_Year_2010';'Water_Year_2011';...'

'Water_Year_2012';'Water_Year_2013';'Water_Year_2014';'Water_Year_2015'
};
t=table(season_load_est2(:,1),season_load_est2(:,2),season_load_est2(:,3),season_load_est2(:,4),season_load_est2(:,5),season_load_est2(:,6),sea
son_load_est2(:,7)...
,season_load_est2(:,8),season_load_est2(:,9),season_load_est2(:,10),sea
son_load_est2(:,11),season_load_est2(:,12),season_load_est2(:,13)...
,season_load_est2(:,14),season_load_est2(:,15),season_load_est2(:,16),sea
son_load_est2(:,17),season_load_est2(:,18),season_load_est2(:,19),sea
son_load_est2(:,20)...
,season_load_est2(:,21),season_load_est2(:,22),season_load_est2(:,23),sea
son_load_est2(:,24),season_load_est2(:,25),season_load_est2(:,26)...%
', 'VariableNames',variablynames, 'RowNames',str);
%
## Display, delete existing table and save new table of coefficients

 disp(t)
 % setTableTitle(t, 'Seasonal Load Estimates',location,para(k)
 if exist([pwd
 '/Output/',location,'/',para(k),'/Seasonal_Load_Estimates_',location,pa
 ra(k),'xlsx'], 'file')==2
 delete([pwd
 '/Output/',location,'/',para(k),'/Seasonal_Load_Estimates_',location,pa
 ra(k),'xlsx']);
 end
 writetable(t,[pwd
 '/Output/',location,'/',para(k),'/Seasonal_Load_Estimates_',location,pa
 ra(k),'xlsx'],'WriteRowNames',true)

%% Plot Annual Flux Estimates
% For comparison to Medalie 2014

 Year_Vec=(1990:1:2015);
 Flux=season_load_est(i+2,:);
 Flux2=season_load_est2(i+2,:);

 figure(f)
f=f+1;
 plot(Year_Vec,Flux,'o')
 title(['Annual Flux Using Conc Coefficients (metric tons/yr)',para(k)])
xlabel('Water Year')
ylabel('Annual Flux')
saveas(gcf,[pwd '/Output/','/','Annual_Flux_Cong_Est',location,para{k},'./Annual_Flux_Cong_Est.'fig']);

figure(f)
f=f+1;
plot(Year_Vec,Flux2,'o')
title(['Annual Flux Using Log Load Est Coefficients (metric
tons/yr)',['../fig'])
xlabel('Water Year')
ylabel('Annual Flux')
saveas(gcf,[pwd '/Output/','/','Annual_Flux_Load_Est',location,para{k},'./Annual_Flux_Load_Est.'fig]);

%% NSE- Nash Sutcliffe efficiency and bias
Yobs=l_est_final; %observed loads (i.e. the smaples)
Ymean=mean(l_est_final);
[y,m,d]=ymd(SampDateTime_final); %redefine year month day
count=1;
count2=1;
count3=1;
for i=1:length(SampDateTime_final)
    index2=find(SampDateTime_final(i)==date_q);
    Ysim_season(i)=load_final(index2); %all simulated loads seasonal
    Ysim_annual(i)=l_est_yr(index2); %all simulated simple annual loads
    if m(i)==3 || m(i)==4 || m(i)==5 %pull out spring (March April May) if the month equals any of these
        Ysim_spring_opt(count)=load_final(index2); %optimal spring simulated
        Ysim_spring_simple(count)=l_est_yr(index2); %simple annual spring
        Yobs_spring(count)=Yobs(i); %observed spring samples
        Ysim_spring_trad(count)=spring_load_trad(index2); %traditional spring MAM regression
        count=count+1;
    end
    if m(i)==3
        Ysim_march_opt(count2)=load_final(index2);
        Ysim_march_simple(count2)=l_est_yr(index2);
        Yobs_march(count2)=Yobs(i);
        Ysim_march_trad(count2)=march_load_trad(index2);
        count2=count2+1;
    end
    if m(i)==8
        Ysim_aug_opt(count3)=load_final(index2);
    end
Ysim_aug_simple(count3)=l_est_yr(index2);
Yobs_aug(count3)=Yobs(i);
Ysim_aug_trad(count3)=aug_load_trad(index2);
count3=count3+1;
end

%% Annual NSE
Ytop=(Yobs-Ysim_season).^2;
Ybottom=(Yobs-Ymean).^2;
NSE_season=1-(sum(Ytop)/sum(Ybottom)); %NSE seasonal optimal

Ytop2=(Yobs-Ysim_annual).^2;
NSE_annual=1-(sum(Ytop2)/sum(Ybottom)); %NSE annual simple

%% Spring NSE
% Optimal Spring - Multiple regressions used to define spring
Ytop_spring=(Yobs_spring-Ysim_spring_opt).^2;
Ybottom_spring=(Yobs_spring-mean(Yobs_spring)).^2;
NSE_spring=1-(sum(Ytop_spring)/sum(Ybottom_spring));

%Simple Annual Spring - Use annual regression to estimate spring
Ytop_spring_simple=(Yobs_spring-Ysim_spring_simple).^2;
Ybottom_spring_simple=(Yobs_spring-mean(Yobs_spring)).^2;
NSE_spring_simple=1-(sum(Ytop_spring_simple)/sum(Ybottom_spring_simple));

%Traditional Spring - Use conventional spring (MAM) observations to create spring regression %to estimate spring
Ytop_spring_trad=(Yobs_spring-Ysim_spring_trad).^2;
Ybottom_spring_trad=(Yobs_spring-mean(Yobs_spring)).^2;
NSE_spring_trad=1-(sum(Ytop_spring_trad)/sum(Ybottom_spring_trad));

%% March NSE

%March Optimal
Ytop_march=(Yobs_march-Ysim_march_opt).^2;
Ybottom_march=(Yobs_march-mean(Yobs_march)).^2;
NSE_march=1-(sum(Ytop_march)/sum(Ybottom_march));

%March Simple
Ytop_march_simple=(Yobs_march-Ysim_march_simple).^2;
Ybottom_march_simple=(Yobs_march-mean(Yobs_march)).^2;
NSE_march_simple=1-(sum(Ytop_march_simple)/sum(Ybottom_march_simple));

%March Traditional
Ytop_march_trad=(Yobs_march-Ysim_march_trad).^2;
Ybottom_march_trad=(Yobs_march-mean(Yobs_march)).^2;
NSE_march_trad=1-(sum(Ytop_march_trad)/sum(Ybottom_march_trad));

%% August NSE

%August Optimal

92
Ytop_aug=(Yobs_aug-Ysim_aug_opt).^2;
Ybottom_aug=(Yobs_aug-mean(Yobs_aug)).^2;
NSE_aug=1-(sum(Ytop_aug)/sum(Ybottom_aug));

% March Simple
Ytop_aug_simple=(Yobs_aug-Ysim_aug_simple).^2;
Ybottom_aug_simple=(Yobs_aug-mean(Yobs_aug)).^2;
NSE_aug_simple=1-(sum(Ytop_aug_simple)/sum(Ybottom_aug_simple));

% March Traditional
Ytop_aug_trad=(Yobs_aug-Ysim_aug_trad).^2;
Ybottom_aug_trad=(Yobs_aug-mean(Yobs_aug)).^2;
NSE_aug_trad=1-(sum(Ytop_aug_trad)/sum(Ybottom_aug_trad));

%% Bias
bias1=(Yobs./Ysim_season);
bias2=(Yobs./Ysim_annual);
bias_season=sum(bias1)/length(Yobs);
bias_annual=sum(bias2)/length(Yobs);

% Bias spring
bias3=(Yobs_spring./Ysim_spring_opt);
bias4=(Yobs_spring./Ysim_spring_simple);
bias5=(Yobs_spring./Ysim_spring_trad);
bias_spring=sum(bias3)/length(Yobs_spring);
bias_spring_simple=sum(bias4)/length(Yobs_spring);
bias_spring_trad=sum(bias5)/length(Yobs_spring);

% Bias March
bias6=(Yobs_march./Ysim_march_opt);
bias7=(Yobs_march./Ysim_march_simple);
bias8=(Yobs_march./Ysim_march_trad);
bias_march=sum(bias6)/length(Yobs_march);
bias_march_simple=sum(bias7)/length(Yobs_march);
bias_march_trad=sum(bias8)/length(Yobs_march);

% Bias August
bias9=(Yobs_aug./Ysim_aug_opt);
bias10=(Yobs_aug./Ysim_aug_simple);
bias11=(Yobs_aug./Ysim_aug_trad);
bias_aug=sum(bias9)/length(Yobs_aug);
bias_aug_simple=sum(bias10)/length(Yobs_aug);
bias_aug_trad=sum(bias11)/length(Yobs_aug);

%% R^2 Optimal Spring (Average R^2 if necessary for more than one
regression for MAM)
R2season_mean=mean(R2season); % Adjusted R^2 from above
rmse_mean=mean(rmse_opt);
if length(str_month{1})==3 || length(str_month{1})==4
    R2spring_opt=R2season(1);
    rmse_spring_opt=rmse_opt(1);
elseif length(str_month{1})==2 || length(str_month{1})==1 && length(str_month{2})==2
    R2spring_opt=(R2season(1)+R2season(2))/2;
    rmse_spring_opt=(rmse_opt(1)+rmse_opt(2))/2;
elseif length(str_month{1})==1
    R2spring_opt=(R2season(1)+R2season(2)+R2season(3))/3;
    rmse_spring_opt=(rmse_opt(1)+rmse_opt(2)+rmse_opt(3))/3;
end

%% Performance Metrics

l={[location,para{k}]};
vnames={"NSE_Annual_Simple","NSE_Annual_Optimal","NSE_Spring_Simple","NSE_Spring_Optimal","NSE_Spring_Traditional","NSE_Spring_Optimal"};
stats(1,1)=NSE_annual;
stats(1,2)=NSE_season;
stats(1,3)=NSE_spring_simple;%simple use annual regression to estimate only MAM
stats(1,4)=NSE_spring_trad;%traditional estimate only MAM obs linear regression
stats(1,5)=NSE_spring;%optimal
stats(1,6)=radj_annual;
stats(1,7)=R2season_mean;%optimal season
stats(1,8)=radj_annual;%This is for simple spring using annual regression to est. only MAM
stats(1,9)=adjR2spring_trad;
stats(1,10)=R2spring_opt;%optimal spring
stats(1,11)=bias_annual;
stats(1,12)=bias_season;
stats(1,13)=bias_spring_simple;
stats(1,14)=bias_spring_trad;
stats(1,15)=bias_spring;

t2=table(stats(1,1),stats(1,2),stats(1,3),stats(1,4),stats(1,5),stats(1,6),...  
stats(1,7),stats(1,8),stats(1,9),stats(1,10),stats(1,11),stats(1,12),...
.stats(1,13),stats(1,14),stats(1,15),'VariableNames',vnames,'RowNames',l);
disp(t2)
writetable(t2,[pwd  
'/Output/',location,'/',para{k},'/'Final_Model_Stats',location,para{k},  
'.xlsx'],'WriteRowNames',true)
%% spring, march, aug metrics
col1=[radj_annual,radj_annual,radj_annual]'; %simple
col2=[adjR2spring_trad,adjR2march_trad,adjR2aug_trad]'; %traditional
col3=[R2spring_opt,R2season(1),adjR2_aug_opt]'; %optimal

col4=[NSE_spring_simple,NSE_march_simple,NSE_aug_simple]'; %simple
col5=[NSE_spring_trad, NSE_march_trad,NSE_aug_trad]'; %traditional
col6=[NSE_spring,NSE_march,NSE_aug]'; %optimal

col7=[bias_spring_simple,bias_march_simple,bias_aug_simple]'; %simple
col8=[bias_spring_trad,bias_march_trad,bias_aug_trad]'; %traditional
col9=[bias_spring, bias_march, bias_aug]'; %optimal

l={'Spring (MAM)', 'March', 'August'};
vnames={'AdjR2_Simple', 'AdjR2_Traditional', 'AdjR2_Optimal', ...
    'NSE_Simple', 'NSE_Traditional', 'NSE_Optimal', ...
    'Bias_Simple', 'Bias_Traditional', 'Bias_Optimal'};
t3=table(col1,col2,col3,col4,col5,col6,col7,col8,col9,'VariableNames',vnames,'RowNames',l);
disp(t3)
write(t3,['/Output/',location,'/',para{k},'\Spring_Stats',location,para{k},'\.xlsx'])

%% RMSE
l=[location,para{k}]

vnames={'RMSE_Optimized_Annual', 'RMSE_Simple_Annual',
    'RMSE_Conventional_Spring', 'RMSE_Optimal_Spring'};
t4=table(rmse_mean,rmse_annual,rmse_spring_trad,rmse_spring_opt,'VariableNames',vnames,'RowNames',l);
disp(t4)
write(t4,['/Output/',location,'/',para{k},'\RMSE_Stats',location,para{k},'\.xlsx'])

%% DONE
disp('Analysis DONE');