ABSTRACT

MILLER, JONATHAN WILLIAM. Assessing Natural and Anthropogenic Drivers of Regional Water Quality using Hierarchical Modeling (Under the direction of Dr. Daniel R. Obenour).

Identifying drivers of water quality degradation is important to assess current risks to aquatic systems as well as design management strategies that might improve future conditions. Water quality monitoring data from state, federal, and local agencies has become abundant in the past few decades, but statistical frameworks are needed that can leverage these large data sources to statistically infer relationships between watershed stressors, water quality, and ecological condition. Hierarchical modeling (mixed-effects modeling or multi-level modeling) is an extension of regression modeling that can account for variance that occurs at group-levels (e.g., estuary, basins, sites) concurrently with predictors. Hierarchical modeling is utilized in this dissertation to assess the effect of natural and anthropogenic watershed stressors on water quality indicators across two large environmental systems: Gulf of Mexico (GoM) estuaries and NC Piedmont wadeable streams. In the GoM, a hierarchical generalized linear modeling approach was developed to predict fish and invertebrate species presence in bottom trawl samples, using data from 33 estuaries over a nineteen-year study period. This was the first GoM estuary assessment to leverage Gulf-wide trawl data to develop species-level indicators and a quantitative index of estuary disturbance. Influential stressors [e.g., total anthropogenic (urban, crop and pasture) land use, crop land use, and the number of toxic release sites in upstream watersheds] were identified for 57 fish and invertebrate species and used to set baseline conditions for the GoM and rank estuaries based on their levels of anthropogenic stress. In general, disturbance levels were greatest in estuaries west of the Mississippi delta and in highly developed estuaries in southwest Florida. In NC Piedmont streams, hierarchical multiple linear regression models were used to identify major natural and anthropogenic drivers of water quality

indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC); and predicted their values throughout the Upper Neuse River Basin (UNRB). Impervious cover (IC) and stream canopy buffers had significant influence over all water quality variables. The BI model suggested that newer development (post-1980 IC) degrades stream biological health 30% less than older development, and canopy removal in stream buffers had 2 to 9 times the impact on BI on a per hectare basis, when compared to the addition of IC in upstream watersheds. Water quality models showed that FC and TDU are driven largely by short-term precipitation, while TN, TP, and SC variability is best explained by point source loadings. Five potential management scenarios were forecasted for the UNRB and compared to current conditions under different hydro-climatological conditions. Restoring canopy cover in stream buffers was more impactful on mean water quality conditions than mitigating the effects of IC and point source loading. The largest potential for water quality improvements came from identifying site-specific causes of water quality (i.e., site and basin random effects) not represented by other predictors in our model. Forecasts were developed probabilistically to account for model and weather uncertainty for future sampling and showed that BI, TN, and FC were the indicators with the most likely chance of observing improvements.

Assessing Natural and Anthropogenic Drivers of Regional Water Quality using Hierarchical Modeling

by
Jonathan William Miller

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

Civil Engineering

Raleigh, North Carolina
2019

APPROVED BY:

Dr. Daniel R. Obenour Committee Chair	Dr. Sankar Arumugam
Dr. Emily Z. Berglund	Dr. Brian J. Reich

Dr. Michael J. Paul
External Member

DEDICATION

I would like to dedicate my dissertation to my wife, Erika Poma, and son, Sanders Miguel Miller-Poma. They are the reason that I returned to school, and they have been extremely understanding throughout my studies. Without their love and support, I would not have been able to complete my degree. Thank you for the joy that you have brought into my life and I pray that my studies can be a blessing for us all. I love you both very much.

BIOGRAPHY

Jonathan William Miller was born May 19th, 1975 in Madison, Wisconsin and grew up in Cary, NC. He lived in a Walnut Hills, went to Cary High School, and attended Western Boulevard Presbyterian Church, all of which formed him into who he is now. He majored in Mathematics and Religious Studies at the University of North Carolina at Chapel Hill where he graduated in 1997. Though he always loved book learning, he felt that there was something missing in his education when he left college: real life experience. In an effort to understand the world around him better, he traveled around the US working as a carpenter in rural Mississippi and as an electrician in Austin, Texas. He also hiked 600 miles following the Columbia River through Washington and was amazed by its power and majesty. From the mile-long Grand Coulee Dam that created a seventy-mile long lake to the man-made jetties at its mouth that melded into the Pacific Ocean, some of man's great engineering feats were only overshadowed by the power of nature itself. This journey taught Jonathan the rewards of perseverance, the peace of nature, and the power of engineering. Jonathan then worked as a missionary in Quetzaltenango, Guatemala for five years where he taught Math and Spanish classes in three Mayan villages: San Juan Ixcoy, San Juan Ostuncalco, and La Pinada, Baja Verapaz. In Guatemala, Jonathan met his wife, Erika Poma and they were married in Raleigh in 2007. Sanders Miguel Miller-Poma, their son, was born in 2008 and quickly became the center of their lives. When Jonathan returned to the US from Guatemala, he taught high school Mathematics at East Montgomery High School in Biscoe, North Carolina and the Academy of High Point Central in High Point, NC, before deciding to return to school in order to change careers. He enrolled in engineering classes at North Carolina A & T for a year before transferring to NCSU

where he received a Master's in Civil Engineering in 2014. That fall, he enrolled for his Ph.D. program under the direction of Dr. Daniel R. Obenour where he has been working to this day.

ACKNOWLEDGMENTS

I would first like to acknowledge God's help in my studies. The opportunity to pursue this degree and the ability to finish it was only possible through His grace.

Next, I would like to thank my advisor, Dr. Daniel R. Obenour, for his guidance and support. He has shown me how to understand and model complex environmental systems and overcome problems inherit in that endeavor. His trust in me has been constant, even when I was not sure in myself; and under his direction, I have become more thorough in my work, concise in my writing, and professional in how I present my research. His guidance was instrumental at every stage of my work and he has taught me invaluable lessons about engineering and modeling. I would also like to thank Michael J. Paul, who was kind enough to not only share data essential for my research, but also collaborate with one of my projects and participate as a member of my committee. Without his help, I would have reached the end of my work as timely as I did.

Many people have helped me during my return to school. First and foremost in those, was my wife, Erika Poma, who was willing to support me as in this long pursuit as we moved three times and left Greensboro for Raleigh. I would also like to thank Vernon Matzen, the former head of graduate studies for Civil, Construction and Environmental Engineering at NCSU. His openness and excitement about helping students return to school to obtain Master's degree gave me the hope and assurance that I needed to change careers. Dr. Sameer Hamoush, the former head of Civil Engineering at North Carolina A & T, also helped by supporting me when I took engineering classes in Greensboro for a year before enrolling at NCSU. Many great professors at NC A & T, including Dr. Stephanie Luster- Teasley, Dr. Manoj Jha, and Dr. Taher Abu-Lebdeh taught me much about engineering and inspired me to continue. At NCSU, Dr. Ranji Ranjithan

and Dr. E. Downey Brill also encouraged me to continue pursuing a Ph.D. and were great teachers along the way. I would like to personally thank Dr. Sankar Arumugam for being my teaching mentor when I taught CE 383 while in the Preparing the Professoriate program. His words of praise and constant optimism were exactly the type of guidance that I needed. Finally, thank you to Dr. Emily Z. Berglund and Dr. Brian J. Reich for being on my committee.

My work was supported by both the National Oceanic Atmospheric Administration and Environmental Protection Agency. I would like to thank those organizations for their financial support as well as for all my numerous collaborators from those projects: Kristin Blackhart, Pete Esselman, Ibrahim Allameddine, Roger Von Haefen, George Von Houtven, and Melissa Kenney.

Last, but certainly not least, I would like to thank my Mom and Dad who have been instrumental in supporting my education both now and throughout my lifetime. Their generous support has made my pursuit of a Ph.D. possible. I also have such warm feelings for the NC State campus due to visits to my Dad's biochemistry laboratory in Polk Hall when I was a boy. I am proud that I am able to graduate from the university where he worked and taught for so many years.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Introduction

Aquatic water systems throughout the world are affected by anthropogenic disturbances such as hydrologic alteration (Booth 1995; Malmqvist and Rundle 2002), nutrients and eutrophication (Rabalais et al., 2002; Howarth and Marino, 2006), toxic pollution (NRC, 2000; Miller et al., 2018), and overuse (Coll et al. 2010). Assessing and identifying principal drivers of aquatic degradation and the magnitude of their impact on *in situ* water quality is an essential first step in developing plans to improve future water quality. Environmental engineers and scientists are often tasked to determine the extent of impairment in natural systems due to chemical or biological factors and design management strategies that will bring them under federal or state compliance. Initially, reducing point source loadings was used to improve water quality in the US, but many aquatic systems still remain disturbed after economically feasible point discharge reductions occurred (USEPA 2006; 2016). In addition, many smaller streams are impaired without having any upstream point sources, often the result of the "urban stream syndrome" (Schueler 1994; Paul and Meyer 2001) or extensive agricultural practices (Cuffney et al., 2010; Qian et al, 2010; USEPA 2016). Understanding the role and influence of non-point sources on downstream water quality is a scientific priority, but their true effect is often obscured by variability due to natural factors (e.g., hydro-climatological conditions, geology), variation across regions, and temporal variations in land use and development. Policy and management decisions rest on the ability to synthesize data from these highly variable systems, understand principal drivers of water quality degradation, and make forecasts that can adequately account for and quantify uncertainty (Beck 1987; Clark et al. 2001).

Water quality data collected by state, federal, and local agencies have become abundant in the past few decades, but the majority of these data are only analyzed to determine if sites exceed state or national criteria. The ability to leverage these spatially expansive multi-decadal datasets as part of broader efforts to answer questions about watershed development and its effect on water quality would help researchers better understand the principal mechanisms that universally degrade water quality as well as give insights into regional differences. Several obstacles to modeling water quality datasets of this kind persist including how to account for repeatedly sampled locations (intra-class correlation), and disentangle multi-collinear watershed predictors (i.e., land covers). Constructing models in statistical frameworks that can incorporate these aspects while identifying key watershed drivers of water quality is a current challenge for water quality modelers.

Hierarchical modeling (mixed-effects modeling or multi-level modeling) is an extension of classical modeling that includes both "random" and "fixed" effects and will be used extensively throughout this dissertation. Fixed effects are coefficients for model predictors normally used in regression modeling, while random effects represent group-level differences in the data (e.g. watershed, estuary, state). Random effects differ from categorical variables in that they are drawn from a normal distribution, called a hyperdistribution (Gelman and Hill 2006) and they do not have to be statistically significant to be included in analyses. Random effects can incorporate intra-class correlation present in collected data as well as quantify group-level differences not accounted for by predictor variables (Gelman et al. 2014). Hierarchical modeling is considered a partial pooling approach that allows the model to fit fixed (coefficients of predictors) and random effects (e.g., sites, estuaries, basins) together leveraging data across the whole study area (Gelman and Hill 2006). This allows for watershed or group level analysis

which is often the goal of research studies (Gelman and Hill 2006), and offers more predictive power than classical regressions (Wikle 2003; Qian and Shen 2007; Cressie et al. 2009; Qian et al. 2010). The flexibility of hierarchical models and the ability to extend them to generalized linear mixed-effects models (GLMM) have made these a common method for modeling ecological datasets (Clark and Gelfand 2006, Bolker et al. 2009), though they have not been used widely in water quality modeling.

The goal of this research was to identify watershed stressors that affect downstream water quality in two complex aquatic systems: Gulf of Mexico (GoM) estuaries and NC Piedmont wadeable streams. This work demonstrates how a hierarchical modeling approach is an accessible and tractable way to model these complex and data-intensive systems, and it rigorously accounts for natural and group-level (i.e., estuary, watershed) variability while identifying watershed stressors of downstream water quality and ecosystem condition. In Chapter 1, multi-stressor hierarchical generalized linear models were developed for individual species to determine dominant watershed stressors and identify benchmark conditions for the Gulf of Mexico. Hierarchical offsets were used to control for intra-class correlation and large amounts of natural variability in the region in order to screen watershed stressors. In Chapter 2, a hierarchical multiple linear regressions (HMLR) was used to model stream health (i.e., biotic index) and compare the relative importance of watershed stressors in NC Piedmont streams. Hierarchical site random effects were then used to help identify potential restoration sites. Finally, in Chapter 3, HMLRs were used to model five additional water quality indicators and forecast their values throughout the Upper Neuse River Basin to assess potential water quality improvements based on future management scenarios.

Chapter 1 presents a published manuscript in *Ecological Indicators*, "Hierarchical Assessment of the Influence of Watershed Stressors on Fish and Invertebrate Species in Gulf of Mexico Estuaries." This chapter describes the first large-scale fish-habitat assessment model for the Gulf of Mexico that identified important watershed stressors of GoM fish and invertebrate species, determined the amount of anthropogenic stress affecting estuaries as compared to benchmark conditions, and ranked GoM estuaries based on these criteria. Using trawl data from 1991-2009, hierarchical generalized linear models were constructed for 57 fish and invertebrate species to help compare 33 Gulf estuaries. The model considers both natural and anthropogenic sources of variability in species presence/absence at various spatial scales. Statistical relationships between species presence and watershed stressors were used as indicators to determine the degree of estuary habitat degradation due to anthropogenic watershed disturbance. This research establishes an effective methodology for performing across-estuary assessments and created an estuary stress index that compares GoM estuaries to regional benchmark conditions.

Chapter 2 describes a HMLR to model stream BI across the NC Piedmont. Using over 3,000 biological samples from 1982-2016, important natural and anthropogenic factors that relate to stream health across the NC Piedmont were identified. Accounting for sources of variability at multiple spatial scales, statistical inference was used to assess the relative contributions of various drivers of stream health. The results provide new insights into the temporal effects of urban development on stream health (i.e., older vs. newer impervious cover) as well as the spatial effects of stream buffers, wastewater treatment plants, and reservoirs. In addition, the study demonstrates a novel application of HMLR modeling to study a complex multi-decadal environmental dataset in an accessible and computationally tractable way. Results

illustrate how the hierarchical framework can help determine potential restoration sites, which is of particular value to those tasked with the care, management, and restoration of degraded urban aquatic systems.

Chapter 3 expands on the work in Chapter 2 by modeling five additional water quality indicators: FC, TDU, TN, TP, and SC, along with BI. The addition of short-term precipitation and wastewater nutrient loadings as candidate predictors for multi-stressor HMLR models were also included in this chapter and were key drivers of stream water quality. Current water quality conditions throughout a large stream network (UNRB) were assessed and then forecasted for future potential management scenarios. Expected water quality improvements were initially summarized based on mean expected improvements across the UNRB. Then, variability due to hydro-climatological conditions and model uncertainty were explored probabilistically to determine the amount of post-restoration monitoring needed to observe potential water quality improvements.

Chapter 1: Hierarchical modeling assessment of the influence of watershed stressors on fish and invertebrate species in Gulf of Mexico estuaries

Miller, J., Esselman, P. C., Alameddine, I., Blackhart, K., & D. R. Obenour. 2018. Hierarchical modeling assessment of the influence of watershed stressors on fish and invertebrate species in Gulf of Mexico estuaries. *Ecological Indicators*, *90*, 142-153.

This work was supported by the National Oceanic and Atmospheric Administration and results from this study are also published online as part of a 2015 report "Through a Fish's Eye." http://assessment.fishhabitat.org/

Ecological Indicators 90 (2018) 142–153

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original Articles

Hierarchical modeling assessment of the influence of watershed stressors on fish and invertebrate species in Gulf of Mexico estuaries

Jonathan Miller[a,*], Peter C. Esselman[b], Ibrahim Alameddine[c], Kristan Blackhart[d], Daniel R. Obenour[a]

[a] North Carolina State University, 208 Mann Hall, Raleigh, NC 27695, United States
[b] US Geological Survey, Great Lakes Science Center, 1451 Green Road, Ann Arbor, MI 48105, United States
[c] American University of Beirut, Bliss Street, PO Box 11-0236, Beirut, Lebanon
[d] ECS Federal, Inc. in support of National Marine Fisheries Service, 2725 Montlake Blvd East, Seattle, WA 98112, United States

ARTICLE INFO

Keywords:
Biological assessment
Species indicators
Anthropogenic stressors
Hierarchical modeling
Watershed development
Gulf of Mexico

ABSTRACT

The northern Gulf of Mexico (GoM) spans five U.S. states and encompasses estuaries that vary greatly in size, shape, upstream river input, eutrophication status, and biotic communities. Given the variability among these estuaries, assessing their biological condition relative to anthropogenic stressors is challenging, but important to regional fisheries management and habitat conservation initiatives. Here, a hierarchical generalized linear modeling approach was developed to predict species presence in bottom trawl samples, using data from 33 estuaries over a nineteen-year study period. This is the first GoM estuary assessment to leverage Gulf-wide trawl data to develop species-level indicators and a quantitative index of estuary disturbance. After controlling for sources of variability at the sampling event, estuary, state, and sampling program levels, our approach screened for statistically significant relationships between watershed-level anthropogenic stressors and fish and invertebrate species presence. Modeling results indicate species level indicators with sensitivities to landscape stressor gradients. The most influential stressors include total anthropogenic land use, crop land use, and the number of toxic release sites in upstream watersheds, as well as agriculture in the shoreline buffer, each of which was significantly related to between 21% and 39% of the 57 species studied. Averaging the effects of these influential stressors across species, we develop a quantitative estuary stress index that can be compared against benchmark conditions. In general, disturbance levels were greatest in estuaries west of the Mississippi delta and in highly developed estuaries in southwest Florida. Estuaries from the Florida panhandle to the eastern Mississippi delta had less anthropogenic stress.

1. Introduction

Fishing is central to the social and economic well-being of the northern Gulf of Mexico (GoM) region of the United States (U.S.), making sustainable management of fisheries a regional priority. The seafood industry in the Gulf States of Florida, Alabama, Mississippi, Louisiana, and Texas (FL, AL, MS, LA, TX) contributed $7.9B to the 2012 U.S. Gross Domestic Product (GDP) and provided 160,000 jobs to coastal residents (NMFS, 2014), while recreational fisheries provided an additional $7.8B to the regional GDP in 2012 as a result of the activities of 3.1 M anglers (NMFS, 2014). Some of the most valuable species in both the commercial and recreational fisheries, including shrimps (Family: Penaeidae), Gulf Menhaden (*Brevoortia patronus*), and Spotted Seatrout (*Cynoscion nebulosus*), have strong affiliations to estuary habitats for a portion of their life cycles. The estuaries of the GoM are subject to disturbance from a wide range of anthropogenic activities, potentially putting commercial and recreational fisheries at risk. Understanding the spatial patterns and causes of degradation to estuary fisheries and fish habitats are important to the conservation of these economic, recreational, and ecological resources.

Environmental degradation in some areas of the GoM is already advanced due to anthropogenic disturbances that include hydrologic alteration, eutrophication, toxic pollution, and overfishing (NRC, 2000; Rabalais et al., 2002; Yáñez-Arancibia and Day, 2004, Howarth and Marino, 2006). Altered patterns of freshwater inflow to GoM estuaries have resulted from upstream damming, river channelization, and water abstraction (Harwell, 1997; Flannery et al., 2002), which have led to changed salinity regimes, reduced dilution of estuary pollutants, and

* Corresponding author.
E-mail addresses: jwmille7@ncsu.edu (J. Miller), pesselman@usgs.gov (P.C. Esselman), ia04@aub.edu.lb (I. Alameddine), kristan.blackhart@noaa.gov (K. Blackhart), drobenour@ncsu.edu (D.R. Obenour).

https://doi.org/10.1016/j.ecolind.2018.02.040
Received 6 October 2017; Received in revised form 6 February 2018; Accepted 18 February 2018
1470-160X/

land subsidence (Day et al., 2000). Excessive runoff of nitrogen and phosphorus from fertilizers and urban activities in catchments has caused eutrophication, and in severe cases, low dissolved oxygen and fish kills (Rabalais et al., 2002; Diaz and Rosenberg, 2008; 2011). Toxic chemicals associated with industry, urban development, and agriculture are strongly concentrated in some areas and have been shown to negatively affect benthic organisms in the GoM (Brown et al., 2000). Some of these toxic releases originate from the petroleum industry, which is especially concentrated in coastal areas of LA and eastern TX (Adams et al., 2004). Estuarine fish and invertebrate species integrate habitat conditions differently over both time and space and can be helpful as biological indicators of larger trends in ecosystem degradation (e.g., Macauley et al., 1999).

A central challenge in biological assessment is to distinguish the responses of species to anthropogenic stress from the high levels of natural background variation common to ecological systems (Hawkins et al., 2010). This challenge is particularly acute in GoM estuaries where the daily and seasonal variability in temperature, salinity, and dissolved oxygen leads to annual variation in biological community composition and structure (Peterson and Ross, 1991; Akin et al., 2003; Baltz et al., 1993; Gelwick et al., 2001; Granados-Dieseldorff and Baltz, 2008). While high natural environmental variability can be stressful to many species, anthropogenic stress to estuaries has been shown to result in dominance by opportunistic habitat or trophic generalists, to the detriment of rare or specialized taxa (Felley, 1987; Chesney and Baltz, 2001; Lewis et al., 2011). Such findings suggest that some species may be particularly sensitive to human impacts, including even estuarine species adapted to high degrees of natural environmental variation. Many estuarine studies conducted to date have focused on responses by groups of species with similar life history or functional characteristics (i.e. community metrics; Macauley et al., 1999; Summers, 2001; Hughes et al., 2002; Meng et al., 2002; Jordan et al., 2010; Cabral et al., 2012). However, evidence from freshwater systems suggests that community metrics may obscure biological responses to stressor gradients that can be detected in species-specific indicators (Baker and King, 2010; King and Baker, 2010). Further, individual species indicators may also allow for more direct linkages to management by focusing on populations of economically valuable taxa or taxa with high conservation value.

Multiple approaches have been used to account for natural background variation in biological assessment. One approach is to define different biological indicators within discrete salinity zones (Coates et al., 2007; Breine et al., 2010; Cabral et al., 2012) or different types of estuaries (Harrison and Whitfield, 2006). Another approach is to screen for indicators that are sensitive to anthropogenic stress, but insensitive to natural gradients (Jordan et al., 2010). Moreover, models can be developed and used to account for the effects of natural variables before testing for indicator sensitivity (Engle et al., 1994). Multivariable models (e.g. multiple linear regression) have proven useful for predicting species or community responses to both natural and anthropogenic gradients (Lewis et al., 2007; Courrat et al., 2009; Delpech et al., 2010). By controlling for natural variation with model coefficients, biological responses to one or multiple ecological stressors can be predicted at different stressor levels which can be particularly helpful for judging whether current conditions differ from a benchmark or reference condition (Hawkins et al., 2010). Without benchmarks, little context exists for interpreting the measured value of an ecological resource, which can vary substantially with natural differences among sites.

Efforts to classify the ecological status of estuaries have occurred globally over the past two decades. In the U.S., studies have focused on characterizing estuaries based on their water quality, susceptibility to pollution based on geomorphologic and flow conditions, and watershed stressors such as land cover and point sources of pollution (Bricker et al., 2008; Greene et al., 2015). Some studies have classified U.S. estuaries based on their fish populations (Gleason et al., 2011; Hughes et al., 2014), but only a few nekton species in limited regions have been modeled to connect fish presence to estuary anthropogenic stress (Toft et al., 2015). In Europe, regional and country specific multi-metric indices have been developed based on biological communities to calculate overall ecological health (Breine et al., 2007; Coates et al., 2007; Delpech et al., 2010; Cabral et al., 2012; Harrison and Kelly, 2013), but limited progress has been made toward showing strong relationships between these indices and anthropogenic stressors (Pasquaud et al., 2013). Such analyses have been complicated by the additional effort required to "intercalibrate" the results from different studies in order to make intra-continental comparisons. In particular, setting regional benchmark conditions and comparing studies with different sampling protocols or metrics is an ongoing challenge not easily resolved (Poikane et al., 2014). Though much progress has been made in sampling and quantifying estuary ecological status, further efforts and techniques are needed to model species and biological communities across varying natural settings, link biological conditions to watershed stressors, and set benchmark conditions.

Hierarchical (or 'multi-level') modeling provides a potential methodology for linking watershed stressors and estuary biological condition, accounting for discrepancies among sampling programs and natural variability among estuaries. Hierarchical modeling accounts for variability among different groups of data at different spatial or organizational levels using regression coefficients (i.e., 'random effects') that vary by group as members of a common statistical hyperdistribution (Gelman and Hill, 2006). This extension of classical regression modeling accounts for intra-class correlation among data from common groups (i.e., estuaries, states, trawl programs), allowing for statistically valid hypothesis testing of group-level predictor variables (Gelman et al., 2014). Thus, for grouped data, hierarchical modeling is often an improvement over classical regression modeling in terms of both predictive performance and causal inference (Wilde, 2003a; Cressie et al., 2009; Qian et al., 2010), and these models have been used extensively to study environmental and ecological systems (Wilde et al., 1998; Wilde, 2003b; Clark and Gelfand, 2006; Bolker et al., 2009; Kashuba et al., 2010; Cuffney et al., 2011). In this study, hierarchical modeling is instrumental in controlling for the variability of species presence across different estuaries, states, and monitoring programs.

The aim of the current study is to identify key sources of watershed stress (i.e., stressors) that are related to species presence in GoM estuaries, and to aggregate these relationships to assess the relative intensity of watershed stress in each estuary using hierarchical generalized linear modeling. This new approach to assessing the biological health of estuaries allows us to: (1) combine nearly 70,000 trawl samples collected by separate research efforts; (2) control for natural environmental variation while identifying statistically significant estuary-level stressors affecting the presence of fish and invertebrate species; and (3) create an estuary stress index that quantifies the amount of anthropogenic stress affecting GoM estuaries as compared to benchmark conditions in the region.

2. Methods

2.1. Study area

Our study spanned 33 estuaries across the five U.S. GoM states (Fig. 1). Estuary habitats were classified to include open water and wetland classes from the Coastal Change Analysis Program (C-CAP) dataset (NOAA, 2006) and the National Wetlands Inventory (NWI; USFWS, 2012). To summarize landscape influence on each estuary, it was necessary to define their spatial extents and tributary influences. The seaward extent of estuaries was limited to the 4-m depth contour based on an examination of plots of salinity-at-depth. The salinities at the defined seaward depth did not generally drop below 32 practical salinity units (psu). The landward extent of the estuaries was drawn to include open water areas as well as estuarine emergent wetlands from

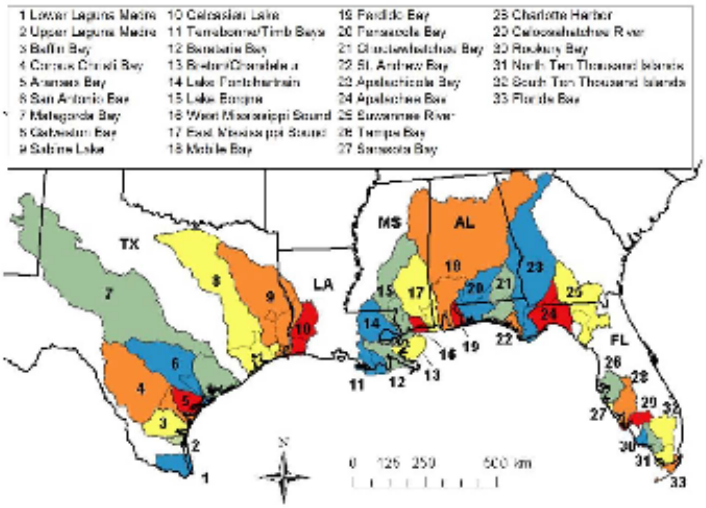


Fig. 1. Basins for the 33 estuaries and estuarine drainage areas (EDAs), representing the most downstream portion (HUC 8) of each basin, are delineated using gray lines while state lines are shown in black. For some smaller drainage basins, the total basin and EDA are equivalent (no gray line within the colored basin). The GoM includes five states: Texas (TX), Louisiana (LA), Mississippi (MS), Alabama (AL), and Florida (FL). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

C-CAP. Atchafalaya Bay, LA was excluded because it is a statistical outlier in terms of its freshwater inflow and nitrogen load (from the greater Mississippi River Basin), and thus difficult to compare to other GoM estuaries. Additional coastal areas were omitted from this analysis because they did not meet the conventional definition of an estuary (i.e., a partially enclosed coastal body of brackish water with one or more rivers or streams flowing into it and with a free connection to the open sea; Pritchard, 1967) or they lacked sufficient watershed level stressor data needed for analysis. One estuary, Rookery Bay, was included in the final analysis because it had watershed level stressor data even though it did not possess any biological data (i.e., trawl data).

Data were collected at two different spatio-temporal scales. Throughout this paper, "event-level data" refer to the data recorded at the specific times and locations of individual bottom trawl samples. "Estuary-level data" refer to average conditions within estuaries or in the watersheds connecting to an estuary. Physical features include both event-level data (temperature, salinity and distance-to-shore) and estuary-level data such as estuary volume, estuary area, percent of estuary open to the sea, and average freshwater inflows. All anthropogenic stressors included in this analysis, such as toxic releases and land covers, are estuary-level data.

2.2. Trawl events

Sampling events, performed by state and federal programs, consisted of towing otter trawls through estuary bottom waters for set distances or periods of time to determine the presence and abundance of different fish and invertebrates living in those areas. This study analyzed 69,570 bottom trawl samples from 1991 through 2009 (Table 1) gathered by the five GoM states (FL, AL, MS, LA, TX) and two federal research programs run by the U.S. Environmental Protection Agency (EPA): the Environmental Monitoring and Assessment Program (EMAP) and National Coastal Assessment (NCA). Sampling gears and protocols varied across programs (Table 1). For instance, programs used nets with varying dimensions and cod-end mesh sizes, and towed for different distances and at different speeds. Therefore, a measure of effort (hectares (ha) sampled) was derived for each sampling event, and the cod-end mesh size was recorded for each sampling program. Variation that existed in sampling protocols was accounted for by a hierarchical grouping of trawl data that allowed the model intercept to vary by sampling program (see section 2.6).

Table 1
Summary of fish and invertebrate trawl data for Florida (FL), Alabama (AL), Mississippi (MS), Louisiana (LA), Texas (TX), Environmental Monitoring and Assessment Program (EMAP), and National Coastal Assessment (NCA).

Program	# of trawls	# of species recorded	Mesh size (mm)	Trawl effort (ha)	Time period
FL	9580	213	3.2	5.6	1991–2005
AL	2620	102	9.5	4.2	1991–2006
MS	708	25	6.4	4.2	1991–2005
LA	23,580	209	6.4	3.2	1991–2007
TX	31,870	488	38	2.8	1991–2009
EMAP	418	157	25	5.0	1991–1994
NCA	795	244	38	5.0	2000–2004

For each bottom trawl event, fish and invertebrate species were identified and enumerated along with basic environmental data such as location, temperature, and salinity. Some trawl events yielded no data because no species were collected, while most events contained records for multiple species. Data on over 500,000 fish and invertebrate individuals collected from the trawl events were used in this study, but only species that were sampled by at least six of the seven monitoring programs and caught in a minimum of 120 trawls were used for final results (see section 2.7). This limited our study to 57 fish and invertebrate species (Table 2) whose ranges were widely distributed throughout the GoM and had sufficient data to adequately parameterize logistic models (Peluzzi et al., 1996; Allison, 2012).

2.3. Event-level variables

Predictor variables were compiled at both the event and estuary levels. Preliminary analyses suggested that temperature, salinity and distance-to-shore were important event-level variables for the majority of species while average estuary depth was not. Temperature and salinity can determine important characteristics of fish habitat due to their effect on primary production and fish metabolism (Fry, 1971; Brett and Groves, 1979), while distance-to-shore can act as a proxy for near-shore habitats and changes in water depth. Reliable event-level depth estimates were missing for a substantial portion of the trawl data and as such this variable had to be excluded from the analysis, despite

Table 2
Modeled fish and invertebrate species along with overall percent presence in trawl samples (T) and in estuaries (E).

Common/Scientific species name	% pres.		Common/Scientific species name	% pres.	
	T	E		T	E
Atl. Croaker/*Micropogonias undulatus*	52	85	Southern Flounder/*Paralichthys lethostigma*	5	73
Bay Anchovy/*Anchoa mitchilli*	49	85	Striped Mullet/*Mugil cephalus*	5	61
Blue Crab/*Callinectes sapidus*	42	94	Black Drum/*Pogonias cromis*	5	64
Spot/*Leiostomus xanthurus*	39	82	Striped Anchovy/*Anchoa hepsetus*	4	76
White Shrimp/*Litopenaeus setiferus*	35	70	Silver Jenny/*Eucinostomus gula*	3	82
Brown Shrimp/*Farfantepenaeus aztecus*	34	70	Lookdown/*Selene vomer*	3	73
Hardhead Catfish/*Ariopsis felis*	30	85	Star Drum/*Stellifer lanceolatus*	3	52
Sand Seatrout/*Cynoscion arenarius*	28	85	Gulf Toadfish/*Opsanus beta*	3	76
Pinfish/*Lagodon rhomboides*	27	94	Lined Sole/*Achirus lineatus*	3	73
Gulf Menhaden/*Brevoortia patronus*	20	79	Atl. Moonfish/*Selene setapinnis*	3	55
Silver Perch/*Bairdiella chrysoura*	20	82	Chain Pipefish/*Syngnathus louisianae*	2	76
Pink Shrimp/*Farfantepenaeus duorarum*	13	85	Crevalle Jack/*Caranx hippos*	2	52
Least Puffer/*Sphoeroides parvus*	12	70	Silver Seatrout/*Cynoscion nothus*	2	70
Bay Whiff/*Citharichthys spilopterus*	11	67	Scaled Sardine/*Harengula jaguana*	2	55
Gafftopsail Catfish/*Bagre marinus*	10	79	Blue Catfish/*Ictalurus furcatus*	2	67
Black Tonguefish/*Symphurus plagiusa*	9	76	Gizzard Shad/*Dorosoma cepedianum*	2	64
Fringed Flounder/*Etropus crossotus*	9	79	Spotfin Mojarra/*Eucinostomus argenteus*	2	52
Atl. stingray/*Dasyatis sabina*	8	73	Gulf Pipefish/*Syngnathus scovelli*	2	79
Inshore Lizardfish/*Synodus foetens*	8	97	Naked Goby/*Gobiosoma bosc*	1	73
Hog Choker/*Trinectes maculatus*	7	85	Lane Snapper/*Lutjanus synagris*	1	58
Bighead Searobin/*Prionotus tribulus*	7	85	Red Drum/*Sciaenops ocellatus*	1	64
Atl. Bumper/*Chloroscombrus chrysurus*	7	76	Atl. Threadfin Herring/*Opisthonema oglinum*	1	55
Atl. Spadefish/*Chaetodipterus faber*	7	73	White Mullet/*Mugil curema*	1	48
Ground Mullet/*Menticirrhus americanus*	7	82	Bluntnose Jack/*Hemicaranx amblyrhynchus*	0.9	58
Pigfish/*Orthopristis chrysoptera*	7	82	Spanish Mackerel/*Scomberomorus maculatus*	0.8	61
Gulf Butterfish/*Peprilus burti*	6	70	Ladyfish/*Elops saurus*	0.8	64
Spotted Seatrout/*Cynoscion nebulosus*	6	79	Lined Seahorse/*Hippocampus erectus*	0.6	48
Threadfin Shad/*Dorosoma petenense*	6	61	Sea Bass/*Centropristis philadelphica*	0.3	52
Atl. Oulassfish/*Trichiurus lepturus*	5	85			

Table 3
Summary of event-level data for trawl events.

	Temperature (°C)	Salinity (psu)	Distance-to-shore (km)
Mean	23.1	18.0	2.7
Standard deviation	6.1	11.0	3.3
2.5% quantile	11	0	0.04
97.5% quantile	31	38	14.1

its obvious importance for fish habitat selection and trawl catch efficiency. Temperature, salinity, and distance-to-shore exhibited substantial variability as seen in Table 3. The variations in the event-level predictors reflect, to a large extent, the natural heterogeneity of estuarine conditions across the study area. While hydrologic alteration in the GoM can affect freshwater inflows and estuary salinity levels (Orlando et al., 1993; Day et al., 2000), these large-scale anthropogenic modifications were not considered directly in this study because they could not be explicitly quantified, though they are likely related to some of the available watershed stressor variables.

All data were checked for potential outliers. Events that reported temperature values over 35 °C or below 6 °C were considered suspect and removed. Similarly, all non-winter records with temperatures below 10 °C were excluded. In total, 285 trawl events were removed for having temperature values outside of these ranges, representing 0.4% of the original dataset. Given that salinity values above 40 psu are common in hyper-saline estuaries such as Baffin Bay, TX, no salinity values were excluded from this analysis (maximum salinity = 66 psu). Sample months were aggregated into seasons: spring (March, April, May), summer (June, July, August), fall (September, October, November), and winter (December, January, February), which were considered as a categorical variable within the models.

2.4. Estuary-level variables

Land cover data were compiled and analyzed at three different spatial scales: basin, estuarine drainage area (EDA; Fig. 1) and shoreline buffer. Basin data covers were summarized from the entire drainage area of the estuary to the topographic divide with adjacent river basins with outlets to the sea, while EDAs were limited only to the lowest most proximate 8-digit hydrologic unit code (HUC) watershed that drains to the coast (NOAA/NOS, 1985). The shoreline buffer was defined as the 500-meter buffer inland from each estuary boundary. EDA sizes ranged from 246 km^2 to 13,690 km^2, while basin areas ranged from 330 km^2 to 121,762 km^2. For some estuaries with entirely coastal drainage areas, basin and EDA values were the same (e.g., Sarasota Bay, FL, Perdido Bay, AL, and Baffin Bay, TX; Fig. 1). Land cover data from 1992, 2001 and 2006 were highly correlated ($r > .98$) such that all estuary-level land cover data was calculated using the 2001 National Land Cover Database (Homer et al., 2007) representing the middle of the study period and supplemented with wetland classifications from C-CAP (NOAA, 2006) and the National Wetlands Inventory (USFWS, 2012). We grouped land cover categories to reduce potential collinearity and to generate combinations of land cover types expected to have similar impacts on estuary habitat quality. The "Agriculture" class was formed by grouping crop and pasture, both of which are expected to yield excess nutrients to waterways while the "Anthropogenic" class aggregated urban, crop, and pasture into one class to represent land use predominantly affected by humans. The land cover classes adopted in the model were normalized by the total land area (A_L) of the given spatial unit (basin, EDA, or shoreline buffer), creating percent land cover values for each spatial scale. Other approaches to normalizing land cover are described below.

Watershed land cover, population, number of toxic release sites, and number of National Pollutant Discharge Elimination System (NPDES) sites (USEPA, 2015) can be interpreted as proxies for pollution, both at the basin and EDA levels. For example, crop areas may export nitrogen,

Table 4
Estuary-level predictor variables considered in the modeling analysis using various normalization options (1 = no normalization; A_E = estuary area (km^2); Q = flow (m^3/day); A_L = total land area). Note: I(0:3) indicates an integer variable (0, 1, 2, or 3).

Variable	Unit	1	A_E	Q	A_L
Watershed					
Shoreline Urban	km^2		X		X
Shoreline Crop	km^2		X		X
Shoreline Agriculture	km^2		X		X
Shoreline Anthropogenic	km^2		X		X
Shoreline Wetlands	km^2		X		X
EDA Urban	km^2		X	X	X
EDA Crop	km^2		X	X	X
EDA Agriculture	km^2		X	X	X
EDA Anthropogenic	km^2		X	X	X
Basin Urban	km^2		X	X	X
Basin Crop	km^2		X	X	X
Basin Agriculture	km^2		X	X	X
Basin Anthropogenic	km^2		X	X	X
EDA toxic releases	#		X	X	
EDA NPDES	#		X	X	
EDA population	#		X	X	
Basin population	#		X	X	
Estuary					
Mean estuary salinity	psu	X			
Estuary percent open to sea	%	X			
Hypoxic condition	I(0:3)	X			
Toxic algal condition	I(0:3)	X			
Eutrophic condition	I(1:5)	X			

phosphorous, pesticides, and various other agricultural chemicals downstream to estuaries. We would have preferred to use more direct predictors like nitrogen, phosphorus, or other pollutant loads, but reliable loading data were unavailable for many estuaries. Land-based pollution loads are often attenuated through biogeochemical processing, settling, and flushing (Chapra, 2008). In the current study, it was infeasible to estimate reaction rates or settling velocities for the diverse range of potentially important pollutants. In an effort to indirectly account for settling and flushing, many estuary-level variables were normalized by estuary area (A_E) and flow (Q), in addition to A_L, creating several stressors per variable (e.g., EDA Crop/A_E, EDA Crop/Q, and EDA Crop/A_L; Table 4). Shoreline buffers are not large enough to produce significant loads, but may affect nearshore habitat conditions, and were thus normalized by only A_E. Five estuarine variables from the 2007 National Estuarine Eutrophication Assessment (Bricker et al., 2008), including estuary average salinity, estuary percent open to sea (%), hypoxic condition, and toxic algal condition (scaled from 0 = no problem to 3 = high), and eutrophication condition (scaled from 1 = low to 5 = high), were also included without normalization. In total, 47 candidate estuary-level variables were considered (Table 4). Basic estuary-level data can be found in Supplementary Material, SM 1 and SM 2.

2.5. Hierarchical generalized linear model

Species presence and absence in trawl samples was modeled as a binary response, where trawls that contained a given species were assigned a value of one and trawls that did not collect that species were assigned a value of zero. A common approach for modeling binary data is to use a generalized linear model that extends the framework of linear regression modeling to non-normally distributed variables by the use of a link function. These types of models are used extensively in many fields, including ecology (Guisan and Zimmermann, 2000; Guisan et al., 2002; Gelfand et al., 2005; Gelman and Hill, 2006; Latimer et al., 2006, Bolker, 2008).

$$P(y_i) = \text{logit}^{-1}(X_i\beta) \quad (1)$$

where $P(y_i)$ is the probability of fish presence in trawl i, X_i is a matrix of predictor variables, and β is a vector of corresponding model coefficients. The inverse-logit function is used to back transform the response from the continuous modeling domain ($-\infty$ to $+\infty$) to probability space values between 0 and 1: $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$.

In addition to accounting for binary responses, our model is constructed hierarchically in order to capture different levels of organization within the data. Hierarchical models use "random effects" to account for data that are similar and can be grouped as members of a common statistical hyperdistribution (Gelman and Hill, 2006). This approach helps account for intra-class correlation among grouped data allowing for statistically valid hypothesis testing of group-level (i.e., estuary-level) predictor variables (Gelman et al., 2014). Random effects were used in this study to account for variation at the estuary, program, and state level that was not explained by other predictors in the model. Without this added model flexibility, prescreening would not have been possible without much noise from spurious correlations between predictors and watershed stressors (see Section 4.4).

2.6. Single-stressor models

For each fish and invertebrate species, we initially screened all 47 possible estuary-level predictor variables (x_{pred}) separately while controlling for event-level variables. Logistic hierarchical models were fit using the "lme4" R package with parallel processing provided by the "doparallel" R package (R Development Core Team, 2011; Bates et al., 2015; Calaway et al., 2015). The adopted model structure is shown below.

$$P(y) = \text{logit}^{-1}(\beta_0 + \beta_{season} + \alpha_{estuary} + \alpha_{state} + \alpha_{program} + \beta_{temp} * x_{temp} + \beta_{sal} * x_{sal} + \beta_{salsq} * x_{salsq} + \beta_{dist} * x_{dist} + \beta_{pred} * x_{pred}) \quad (2)$$

where P(y) is the probability of occurrence of a given species, β_0 is the overall y-intercept for the model, and β_{season} is a categorical variable used to account for seasonal differences. Random effects, $\alpha_{estuary}$, α_{state}, and $\alpha_{program}$, were assumed to follow normally distributed hyperdistributions centered around zero. For each species modeled, 32 different values of $\alpha_{estuary}$ were returned (Rookery Bay is excluded due to no trawl events), five values for α_{state} and seven values for $\alpha_{program}$. Slope coefficients β_{temp}, β_{sal}, β_{salsq}, and β_{dist} were determined for the event-level effects of temperature (x_{temp}), salinity (x_{sal}), salinity^2 (x_{salsq}), and distance-to-shore (x_{dist}), respectively. The square of salinity (x_{salsq}) was included because preliminary analysis indicated a potential non-linear (quadratic) relationship between salinity and probability of fish or invertebrate presence. β_{pred} is the slope coefficient of the estuary-level anthropogenic variable being screened. These single-stressor models were used to test the significance of each estuary-level predictor (Table 4) on the occurrence of every fish and invertebrate species (Table 2). Predictor variables were considered statistically significant when their corresponding β_{pred} coefficients were significantly different from zero at a 95% confidence level ($p < 0.05$).

2.7. Multi-stressor models

Candidate estuary-level variables for multi-stressor modeling were selected based on the single stressor modeling results. Favoring parsimonious multi-stressor models, we selected only estuary-level variables that were found to be significant for over 20% of the species. Furthermore, variables were tested for multicollinearity, and when one or more variables were found to be correlated ($|r| \geq 0.65$; Supplementary Material, SM 3), only the most significant variable was included. This process reflects two assumptions: 1) stressor variables with truly mechanistic underpinnings should affect a broad range of species across the GoM and 2) variables that represented similar mechanisms (e.g., Basin Crop/A_L is a subset of Basin Agriculture/A_L) should only enter the model once. Performing variable selection on a subset of probable stressors instead of the entire 47 dependent variables

Table 5
Summary of highly significant estuary-level stressor variables (statistically related to over 20% of species), as determined from the single-stressor models. Negative relationships indicate the percent of significant relationships that are negative. Candidate variables are those tested in the multi-stressor models.

Stressor	Species affected (%)	Negative relationships (%)	Candidate variables
Basin Anthropogenic/A_L	53	77	X
Mean estuary salinity	49	61	X
Basin Agriculture/A_L	47	70	
Basin Crop/A_L	44	72	
Basin Crop/Q	39	73	X
Basin Anthropogenic/Q	37	67	
Shoreline Agriculture/A_L	35	15	X
Basin Agriculture/Q	35	70	
EDA toxic releases/A_E	23	85	X
Basin Urban/Q	21	58	

(Table 4) avoids the inclusion of spurious or "noisy" predictors and thus leads to more robust models (Cohen, 1990; Derksen and Keselman, 1992).

For each species, all event-level variables (temperature, salinity, salinity^2, and distance to shore) and the selected candidate estuary-level variables were included together in a backward model selection procedure using the "LMERConvenienceFunctions" R package (Tremblay and Ransijn, 2015). Backward selection eliminates variables in a step-wise fashion based on statistical significance, until the final model contains only predictors significant at the 95% confidence level. Note that since multi-stressor models could have up to nine predictors (four natural variables and five anthropogenic ones; Eq. (2) and Table 5) and three levels of random effects based on preliminary modeling results, these models were only developed for species caught in at least 120 trawl events. This insures that we have at least 10 observations for each possible predictor and level of random effect (Peduzzi et al., 1996; Allison, 2012). Model selection using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) was also explored (Sakamoto et al., 1986; Schwarz, 1978), but both of these criteria added insignificant predictors as well as eliminated significant predictors making these methods not desirable for the overall project goals.

2.8. Assessment of estuary level anthropogenic stress

Once the multi-stressor models were developed, they were used to create an estuary stress index which quantifies the amount of anthropogenic disturbance in estuaries by comparing sampled conditions to those predicted under two different benchmarks of anthropogenic disturbance. Least disturbed conditions (LDC) were defined as the minimum value of each stressor observed at the regional scale (SM 1), and minimally disturbed conditions (MDC) were predicted by setting stressor values in the models to zero. LDC and MDC are benchmarks of ecological condition frequently used to assess the degree to which the current conditions of a particular system deviate from a more desirable state (Stoddard et al., 2006; Hawkins et al., 2010). We used the region-wide stressor minima to specify LDC as an example of one benchmark to assess estuaries in the GoM, but sub-regional LDCs might also be developed and used within this same framework. To judge the amount of anthropogenic watershed stress on each estuarine system, the absolute value of each stressor coefficient was averaged across all species to determine the mean absolute rate of change in probability of species presence (on the log-odds scale) associated with each stressor. Mean absolute coefficients were then multiplied by the actual estuary-level conditions and compared to both LDC and MDC to create the estuary stress index.

3. Results

3.1. Single-stressor models

All estuary-level stressor variables (Table 4) were screened in single-stressor models (Eq. (2)) that controlled for natural variability at the event-level (e.g., temperature, salinity) as well as for intra-class correlation and unknown variability at group levels (i.e., estuary, state, program). Variables were then ranked according to the number of species with which they had significant relationships (Table 5). Basin Anthropogenic/A_L, Basin Crop/A_L, and Basin Agriculture/A_L were found to be statistically significant for over 40% of the assessed fish and invertebrate species with more than 70% of those significant relationships being negative. Of the other anthropogenic candidate stressors, Basin Crop/Q and EDA toxic releases/A_E were also negatively related to the majority of the species, while Shoreline Agriculture/A_L had a predominantly positive relationship. Mean estuary salinity was found to be a significant predictor for 49% of modeled species.

3.2. Multi-stressor models

Multi-stressor models for individual fish and invertebrate species were developed by performing backward variable selection on event-level variables (temp, sal, salsq, and dist) and the five candidate estuary-level stressors identified in Table 5 (see Methods section for selection criteria). This resulted in a unique multi-stressor model for each fish and invertebrate species (SM 4). For example, the model for Silver Perch (*Bairdiella chrysoura*) was:

$$\begin{aligned} P(y) = \text{logit}^{-1}(&-2.92 + 0.02*[\text{temp}] + 0.04*[\text{sal}] - .001*[\text{salsq}] \\ &-0.05*[\text{dist}] - 0.40*[\text{EDA toxic releases}/A_E] - 0.22*[\text{Basin Crop}/Q] \\ &+ 0.29*[\text{Mean estuary salinity}] + \beta_{\text{season}} + \alpha_{\text{estuary}} + \alpha_{\text{state}} \\ &+ \alpha_{\text{program}}) \end{aligned} \quad (3)$$

The model indicates that Silver Perch presence increases with higher temperatures and in areas closer to the shore. The relationship between salinity and Silver Perch is quadratic with an optimal value at 20 psu, which is consistent with the fact that the Silver Perch prefer higher saline estuaries. Silver Perch was negatively related to EDA toxic releases/A_E and Basin Crop/Q. Note, coefficients for event-level variables are calculated in their natural units (i.e., °C) while coefficients for estuary-level variables (e.g., EDA toxic releases/A_E, estuary salinity) are based on their normalized values (with mean of zero and standard deviation of one). Therefore, a change of 1 °C at the sampling spot has an effect of 0.02 on the log odds of Silver Perch presence while the change of one standard deviation of EDA toxic releases/A_E (6.3×10^{-4} releases/km^2) represents a change of 0.40 (SM 1,4). Multi-stressor models can be used to predict changes in species presence within estuarine systems due to changes in estuary-level stressors (SM 5).

The fraction of anthropogenic (agricultural plus urban) land cover in an estuarine basin (Basin Anthropogenic/A_L) was found to be the most significant anthropogenic stressor in multi-stressor models affecting 39% of species, of which 77% showed a negative relationship (Table 6). Basin Crop/Q was related to 37% of species, with 86% of those relationships being negative. Shoreline Agriculture/A_L was related to 23% of the species with most relationships being positive, and EDA toxic releases/A_E was related to 21% of the species with most relationships being negative. Average estuary salinity was related to 25% of species as well. Mean absolute values of stressor coefficients in log odds showed that Basin Anthropogenic/A_L and Basin Crop/Q exerted the largest influence on species presence throughout the GoM (Table 6).

Table 6
Percent of species significantly related to candidate estuary-level stressors in multi-stressor models and percent of those relationships that are negative. Mean absolute change is calculated by averaging the absolute value of the slope coefficients for all 57 species.

Stressor	Species affected (%)	Negative relationships (%)	Mean absolute change (log odds)
Basin Anthropogenic/A_L	39	77	.27
Basin Crop/Q	37	86	.24
Shoreline Agriculture/A_L	23	8	.15
EDA toxic releases/A_E	21	83	.14
Mean estuary salinity	25	36	.16

3.3. Assessment of estuary level anthropogenic stress

The aggregated intensity of anthropogenic stress for each GoM estuary was combined to form an estuary stress index based on the mean absolute change in log odds of species presence, comparing current to benchmark (LDC or MDC) conditions (Fig. 2). We used the mean absolute value of both positive and negative coefficients, reflecting that increases in stress-tolerant species can indicate estuary disturbance in the same way that declining sensitive species do (Pelley, 1987; Caddy, 2000; de Leiva Moreno et al., 2000; Chesney and Baltz, 2001; Lewis et al., 2011). The estuary stress index provides a concise and quantitative way to estimate the amount of anthropogenic stress affecting each estuary, and since values are additive in log odds, the contributions of individual stressors can be observed separately (Fig. 2). Basin Anthropogenic/A_L had the largest effect on species presence in the GoM, averaging nearly 70% of total deviations, and was most prevalent in the highly urbanized southwest FL estuaries (Sarasota Bay, Charlotte Harbor, Caloosahatchee River, and Tampa Bay) along with several estuaries in the western GoM (Baffin Bay, Upper and Lower Laguna Madre, and Aransas Bay, TX and Lake Borgne and Lake Pontchartrain, LA). Basin Crop/Q and EDA toxic releases/A_E both had large effects in relatively few estuaries where those stressors had elevated values (Upper Laguna Madre, Baffin Bay, Galveston Bay, and Sabine Lake, TX and Calcasieu Lake, LA) while Shoreline Agriculture/A_L was prevalent in several TX and LA estuaries.

Anthropogenic influence was also split between positive and negative effects (Fig. 3) in order to assess in which direction watershed stressors were impacting GoM estuaries. Both positive and negative components exist for the same stressor because some of the 57 species modeled are positively related to the stressor while other species are negatively related to the same stressor (Table 6; SM 4). The magnitude of negative impacts are significantly larger for Basin Anthropogenic/A_L, Basin Crop/Q, and EDA toxic releases/A_E, while the opposite is true for Shoreline Agriculture/A_L (Fig. 3).

3.4. Indicator species

Modeling individual species allowed for the identification of fish and invertebrate species whose presence had significant negative or positive relationships with anthropogenic stressors. Species with the largest negative coefficients for Basin Anthropogenic/A_L were Atlantic Moonfish (*Selene setapinnis*), Gulf Butterfish (*Peprilus burti*), Bay Whiff (*Citharichthys spilopterus*), White Shrimp (*Litopenaeus setiferus*), Blue Catfish (*Ictalurus furcatus*), Bluntnose Jack (*Hemicaranx amblyrhynchus*), Sea Bass (*Centropristis philadelphica*) and Pringed Flounder (*Etropus crossotus*) (SM 4). Species with large negative relationships with EDA toxic releases/A_E were Gulf and Chain Pipefish (*Syngnathus scovelli*, *Syngnathus louisianae*), Pink Shrimp (*Farfantepenaeus duorarum*), Atlantic Threadfin Herring (*Opisthonema oglinum*), Atlantic Stingray (*Dasyatis sabina*), Blackcheek Tonguefish (*Symphurus plagiusa*), Silver Perch, and Blue Crab (*Callinectes sapidus*). Several species had consistently positive responses to anthropogenic stressors including: Black Drum (*Pogonias cromis*), Gafftopsail Catfish (*Bagre marinus*), Hardhead Catfish (*Ariopsis felis*), White and Striped Mullet (*Mugil curema*, *Mugil cephalus*), and Lined Sole (*Achirus lineatus*).

3.5. Random effects and seasonality

State random effects were included to account for the potential impact of state-level fishing regulations on species presence, as well as large-scale natural variations in habitat that may occur at the state level. In addition, program random effects were included to account for variation among different trawl programs (efficiency and effort). State and program random effects were averaged across all species (Fig. 4). Mean state effects ranged from −0.3 (FL) to 0.4 (TX) and program effects vary from −0.8 (MS) to 0.6 (AL) suggesting substantial variability in the effects of different program sampling methods on the probability of species presence in trawl events. Seasonal effects were included as categorical variables. Fall had the highest mean effect while mean summer, spring, and winter offsets were 0.38, 0.36, and 0.38 lower (on log odds scale), respectively. It is important to note that event-level temperature and salinity measurements were also included as predictors in the model, so that these seasonal effects represent variability that exists in addition to seasonal patterns in temperature and salinity.

The magnitudes of estuary-level random effects were determined by averaging across all species (Fig. 5). The relatively high random effects

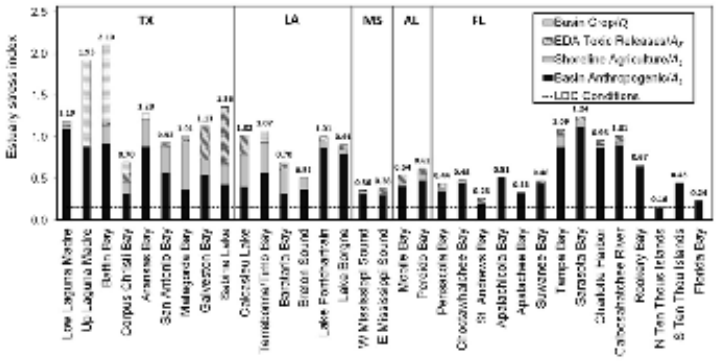


Fig. 2. Estuary stress index is the mean anthropogenic effect in logs odds based on the absolute value of model coefficients. The x-axis represents minimally disturbed conditions (MDC; all anthropogenic stressors = 0) and the dotted line represents least disturbed conditions (LDC) for the GoM (SM 1).

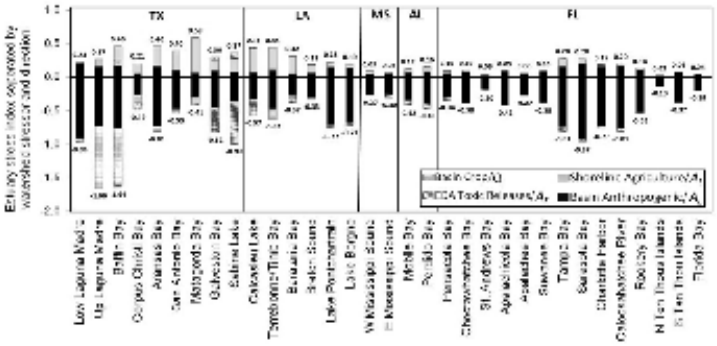


Fig. 3. Mean anthropogenic effect by estuary separated by positive and negative effects. The sum of the mean positive and mean negative effects equal the total absolute effect shown in Fig. 2.

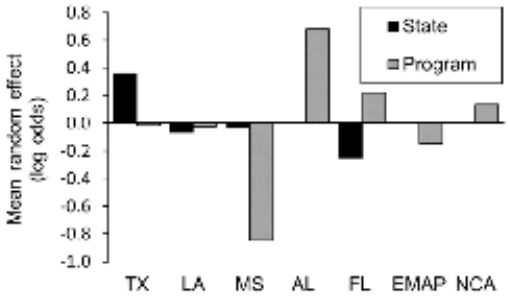


Fig. 4. Mean state and program random effects across all fish and invertebrate species.

observed in estuaries such as Aransas Bay, Corpus Christi, TX, Terrebonne/Timbalier Bays, LA, and Apalachicola Bay, FL point to estuaries where species presence was higher than would be expected based on the predictors and other random effects included in the models. Similarly, estuaries such as Baffin Bay, TX and Breton Sound, LA had a lower probability of presence than would be expected. In such estuaries, other biophysical or watershed factors that were unaccounted for by this analysis may be affecting habitat quality.

4. Discussion

4.1. Anthropogenic stressors

Our study contrasts with previous biological assessment efforts that sought to characterize biological condition, but which were not designed to quantitatively relate that condition to watershed anthropogenic stressors (Deegan et al., 1997; Macauley et al., 1999; Summers, 2001; Hughes et al., 2002; Meng et al., 2002; Harrison and Whitfield, 2004, 2006; Breine et al., 2007; Coates et al., 2007; Uriarte and Borja, 2009; Breine et al., 2010; Delpech et al., 2010; Jordan et al., 2010; Cabral et al., 2012). Previous work in the U.S. (e.g., Greene et al., 2015) has characterized the magnitudes of anthropogenic stressors in estuarine watersheds, but there has been limited progress toward quantitatively linking these stressors to the biological condition of estuaries. In Europe, Teichart et al. (2016) used ecology quality ratios (EQR) scores from seven European countries to determine water quality stressors that affect estuarine biological condition. Though this was a step toward linking biological condition to watershed stressors, there are still questions about how comparable EQR ratios are among countries (Poikane et al., 2014) and basin-level stressors were not explored. In this study, we demonstrate a linkage between watershed stressors at three spatial scales (basin, EDA, and shoreline) and estuarine biological condition using regression models that relate watershed stressors to the log odds of species presence while accounting for natural variation and other potentially confounding effects (e.g., systematic biases associated with different sampling programs). We also develop a quantitative

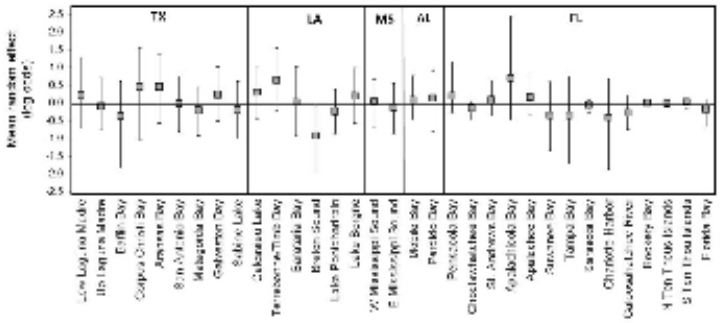


Fig. 5. Estuary random effects. Grey square indicates the mean (across species) and error bars show the 0.1 and 0.9 quantiles. Positive random effects imply species presence was under predicted by the other components of the model. Negative random effects imply the opposite. Rookery Bay does not have an estuary random effect because it had no trawl samples.

index of estuary stress using mean stressor coefficients (averaged across species) and the variability in stressors across estuarine watersheds (Figs. 2, 3). We found widespread and largely negative influences of landscape stressors on both fish and invertebrate species. Our results suggest that estuaries east of the Mississippi delta to the Florida panhandle are, in general, less impacted by watershed stress than other systems in the GoM. Our indicator screening process also provides new insight into species-level responses to estuary stressors, revealing species indicators on both the tolerant and sensitive sides of the 'stress-tolerance gradient' (Whittier et al., 2007).

Our results indicate that human activities in upstream watersheds have statistically significant and largely negative relationships on a substantial portion of estuarine fish and invertebrate species. The percent of anthropogenic land cover in river basins (Basin Anthropogenic/A_L) has the strongest relationship, affecting 22 species (17 negatively) in the multi-stressor models (Table 6) including five of the ten most prevalent species in the GoM (SM 4). It accounts for nearly 70% of all absolute deviations from MDC (Fig. 3). The "Anthropogenic" land category is a composite of urban, crop, and pasture, and it generally outperformed the individual land categories in terms of predictive performance (Table 5). Nonetheless, it is interesting to consider whether one of the three land cover classes (urban, crop and pasture) plays a dominant role. The high ranking of crop and agriculture in single stressor models (Table 5) coupled with the fact that agriculture comprises 69% of anthropogenic land use within our study area, suggests that it is perhaps the dominant form of land cover leading to estuary disturbance. To further explore this, we repeated the single-stressor modeling three times, omitting estuaries in FL, TX, and the central Gulf states (LA, MS, and AL), in turn, as a simple sensitivity analysis. Basin Anthropogenic/A_L remained the top stressor when both the central state estuaries and the TX estuaries were omitted, but Basin Crop/A_L became the top stressor when FL estuaries were omitted. The increased importance of Basin Crop/A_L when Florida estuaries are excluded implies that crop production might be more influential in the western GoM. Of note, the categorical variables representing hypoxic, toxic algal, and eutrophic condition estimates (Bricker et al., 2008) were not found to be significant predictors in single-stressor models, suggesting these metrics, as calculated and used in this study, are not primary drivers of species presence in GoM estuaries.

Basin Crop/Q, reflecting crop areas normalized by average flow rate, is significantly related to 21 species (18 negatively). This stressor is likely a proxy for the concentration of crop-related pollutants entering a system. Basin Crop/Q was highest in Baffin Bay and Upper Laguna Madre, two hyper-saline southern TX estuaries, and elevated in Corpus Christi Bay, TX and Terrebonne/Timbalier Bays, LA as well (SM 1). These results imply that as estuary flushing decreases, estuaries become more susceptible to pollution. Hydrologic alteration as well as consumptive water use by urban and agricultural sources can reduce water inflow to estuaries and the dilution of contaminants in surface water (Harwell, 1997; Day et al., 2000; Flannery et al., 2002).

The number of toxic release sites in the estuarine drainage area (EDA) normalized by the estuary area is significantly related to 12 species (10 negatively). Five of the negative relationships are with species associated with bottom environments (e.g., Atlantic Stingray, Blue Crab, Pink Shrimp, Blackcheek Tonguefish, and Hog Choker (*Trinectes maculatus*); SM 4). Galveston Bay and Sabine Lake, TX and Calcasieu Lake, LA have the highest historical levels of toxic releases, potentially due to upstream petrochemical plants (USEPA, 2015; Blackburn, 2015). Our findings are consistent with previous research reporting measurable negative effects of toxic contamination and releases on estuarine organisms, particularly in the benthos (Pearson and Rosenberg, 1978; Boesch and Rosenberg, 1981; Malins et al., 1985; Engle et al., 1994; Macauley et al., 1999; Brown et al., 2000).

The percent of the estuarine shoreline buffer comprised of agricultural land cover is the only anthropogenic stressor identified to be predominantly associated with an increase in species presence. It is significantly related to 13 species; 12 of these positively. Estuaries with shoreline agriculture greater than 10% are generally found in TX and LA (SM 1). One possible explanation for this finding is a positive response to incipient stress as observed in other estuary studies (Jordan and Vaas, 2000; Jordan and Smith, 2005). A possible mechanism for a positive response to stress could be greater productivity and diversity due to nutrient enrichment (i.e., a release from oligotrophy; Nixon and Buckley, 2002), or by a combination of stressors that has not reached a threshold for causing negative effects (Jordan et al., 2010). Alternatively, long-term effects of nutrient enrichment (and overfishing) have been shown to lead to the dominance of pelagic over benthic/demersal species (Caddy, 2000; de Leiva Moreno et al., 2000). Nine of the 12 species positively linked to shoreline percent agriculture are pelagic species (e.g., Gulf Menhaden, Atlantic Bumper (*Chloroscombrus chrysurus*), Ground Mullet (*Mentcirrhus americanus*), and White Mullet; SM 4) and two more are catfish species which thrive in high nutrient waters (*Ariopsis felis*, *Bagre marinus*). More investigation into the positive relationship between shoreline agriculture and the presence of several species is warranted.

4.2. Assessment of estuary level anthropogenic stress

We adopted both MDC and LDC as possible benchmarks against which to compare estuary disturbance in order to demonstrate how our index of estuary stress could be used with both; however, each of these potential benchmarks have limitations. MDC requires a somewhat greater degree of model extrapolation, while LDC is not referenced to an absolute ecological state making it susceptible to shifting baselines as watersheds continue to develop over time (Pauly, 1995). LDC conditions, as defined in this study, do not deviate greatly from MDC (0.16 change in log odds; Fig. 2). LDC for both Shoreline Agriculture/AL and EDA toxic releases/A_E were zero (i.e., equal to MDC) while LDC for Basin Crop/Q is essentially the same as MDC. The minimum LDC value for Basin Anthropogenic/A_L was 8.9% and this stressor explains the majority of the difference between MDC and LDC. The application of LDC across such a large system as the GoM might be debatable (Stoddard et al., 2006); however, comparing the MDC and LDC benchmarks in Fig. 2 suggest the overall disturbance pattern indicated by our results should be fairly robust to small changes in benchmark assumptions. Estuary rankings essentially remain the same whether they are based on deviation from LDC or MDC conditions since both benchmarks are constant across the GoM and would only change if sub-regional benchmarks are developed on the assumption that baseline conditions vary by region. Another approach that might lead to different estuary ranks would be if the estuary stress index was defined as only the negative (i.e., negative component in Fig. 3) or the net effect of watershed level stressors (i.e., the difference between the positive and negative components in Fig. 3) for specific subsets of species of interest. If system productivity (i.e., more fish for anglers) is of interest, then an estuary stress index could be created specifically for species of interest for commercial and recreational fisherman. However, both of these alternatives ignore the value of stress-tolerant species as potential indicators of anthropogenic influence (see below; Whittier et al., 2007).

4.3. Indicator species

The estuary stress index proposed here is based on the responses of widely distributed fish and invertebrate species to different landscape anthropogenic stressors that provide insights into important regional indicator species. Forty-eight of the 57 species studied are significantly related to at least one anthropogenic stressor in the multi-stressor models (SM 4). While the overwhelming direction of response is negative, responses of species with apparent stress tolerance (i.e., those that responded positively to increasing stressor levels) are also present (see Section 3.4). Tolerant species tend to have greater behavioral and diet plasticity, greater niche breadth and may benefit from the absence

of more sensitive competitors (Vázquez and Simberloff, 2002; Swihart et al., 2003; Devictor et al., 2008; Wilson et al., 2008; Segurado et al., 2011).

4.4. Hierarchical modeling

The ability to model variability at multiple scales (i.e., estuaries, states, programs) is an important advantage of hierarchical modeling in this application. Critically, it allowed us to perform statistically valid hypothesis testing concerning which watershed anthropogenic stressors are significantly related to species presence. The hierarchical approach can be contrasted with more traditional "no pooling" and "complete pooling" approaches to regression modeling (Gelman and Hill, 2006; Qian et al., 2010; Cuffney et al., 2011). If we had developed independent models (or parameter estimates) for each estuary (no pooling), then it would be infeasible to statistically assess the significance of stressors acting across multiple estuaries. On the other hand, if all data were combined into a single model, but random effects were omitted, then we would be treating each trawl sample as statistically independent, failing to address the substantial intraclass correlation that exists among samples from the same estuary, state, and program (Gelman et al., 2014). As an example, single-stressor models (Eq. (2)) were run omitting the estuary random effect ($\alpha_{estuary}$) and 40 of the 47 predictor variables listed in Table 4 were significant for over 80% of species; an implausible result related to the failure to account for intraclass correlation.

In addition, random effects allow us to characterize variability over different scales (i.e., sample, estuary, state, and programs) (Wilde, 2003a; Cressie et al., 2009). For example, program-level random effects allow us to assess the variability associated with differences in sampling protocols. The MS program only recorded 23 of the 57 modeled species, resulting in a much lower mean program random effect when compared to other state and federal programs (Fig. 4). State random effects are also of interest, with TX having the highest state random effect and FL having the lowest, though the variation across states was less than the variation across programs (Fig. 4). Exploring the causes of state-level variation could be a subject for future research, as this variation could plausibly be related to either state fishing regulations, biogeography, or other factors that vary over a large spatial scale. The ability to discern between state and program level random effects is critical, as state random effects reflect actual differences in species prevalence, whereas program random effects reflect sampling efficiency. Distinguishing between state and program random effects within the model was only possible because of the two federal interstate programs (EMAP and NCA), and this is an important reason to continue these federal trawl programs in the future.

4.5. Summary and future directions

This study identifies watershed anthropogenic stressors that are most strongly associated with fish and invertebrate species presence, and then used the modeled stressor relationships to develop an index of watershed stress for estuaries across the GoM. The hierarchical generalized linear model used here provides a tool to control for variability that occurs among programs, estuaries, and states, as well as among individual trawl samples. A large percentage of GoM species exhibited negative associations with anthropogenic stressors, but some positive associations were present as well. The percent anthropogenic land cover (urban, crop, and pasture) within an estuary's drainage basin was consistently ranked as a leading predictor of estuary disturbance. When determining deviations from benchmark conditions (Fig. 2), we implicitly assumed that the statistically significant predictor variables identified through hierarchical modeling are, in fact, drivers of species presence/absence. However, the empirical relationships developed in the single and multi-stressor models presented here do not directly demonstrate causal relationships, as it is well known that correlation is not necessarily causation. The results do, however, provide a data-supported line of evidence relevant to large-scale estuarine assessments and fisheries management in the GoM. Past studies have found negative effects of nitrogen loading, eutrophication, and urbanization on estuary species (Dauer et al., 2000, Diaz and Rosenberg, 2008; Coll et al., 2010), but our results are perhaps the most compelling evidence to date for widespread population-level effects on fish and invertebrate habitats in GoM estuaries by watershed stressors, given the comprehensiveness of the datasets used and the spatial extent of the study. We expect this study to motivate and focus future research efforts to understand the mechanisms by which watershed stressors affect estuary habitats. Our findings may also be useful to help guide and prioritize watershed restoration activities.

While the current study focuses on species-level responses, other studies have focused on the responses of biological assemblages (i.e., metrics of community composition, structure, and function) to regional stressor gradients (Lewis et al., 2007; Piazza and La Peyre, 2009). Acknowledging the potential advantages to species-specific indicators (see Section 1), our approach could be extended to consider the presence/absence of species functional groups that share common habitat, life history, or feeding strategies. Modeling functional groups may help corroborate the patterns documented here and provide further insights into the mechanisms by which landscape stressors act on biological communities. Further, modeling functional groups could facilitate comparison of results to previous studies that have relied on community metrics for estuarine biological assessment (Engle et al., 1994; Summers, 2001; Hughes et al., 2002; Jenkins, 2004; Jordan et al., 2010).

This hierarchical approach could be extended to consider species abundance, which is related (imperfectly) to the number of individuals of a given species collected in a trawl event. An assessment of how species abundance responds to anthropogenic stressors could also be valuable, particularly for species of social or economic importance. Many of the species found to have sensitivities to stressors in this study are also targeted by the regional commercial fishery (including shrimps, Gulf Menhaden, and Blue Crab), and species that are heavily targeted in recreational fisheries (Spotted Seatrout, Sand Seatrout (*Cynoscion arenarius*), Silver Seatrout (*Cynoscion nothus*), Ground Mullet, and Atlantic Croaker (*Micropogonias undulatus*); NMFS, 2014). A natural extension of this research is to focus on the responses of economically valuable species in terms of both their probabilities of occurrence and abundances.

Finally, the hierarchical modeling approach developed in this study could be enhanced to consider additional spatial and temporal variability. Our approach does consider event-level variables like temperature and salinity that explain some intra-estuary variability, but which cannot be directly related to stressors. Like other studies before ours (Jordan et al., 2010), our assessment is not capable of estimating variation in biological condition at a sub-estuary scale. If trawl samples could be organized into different estuarine subsections (perhaps based on estuary geomorphology), then species presence could be related to more localized stressors associated with shoreline development. Shoreline hardening associated with sea walls has been shown to reduce both richness and abundance of species in near shore areas (Peterson and Lowe, 2009; Gittman et al., 2016; Dethier et al., 2016), but precise estuary-level estimates of shoreline hardening were not available for this study (Gittman et al., 2015). Regarding temporal variability, future analyses could explore how species presence changes over time, with alterations in watershed development and basin flows as potential predictor variables. A better understanding of temporal trends in estuary condition could further prioritize watershed management and restoration activities.

Acknowledgements

This work would not have been possible without the contributions

of Correigh Greene, Steve Brown, Harmon Brown (NOAA NMFS), Tom Minello (NOAA retired), Kirsten Larsen (NOAA NESDIS), and Moe Nelson (NOAA NOS). The Southeast Aquatic Resources Partnership Science and Data Committee provided important guidance. Hiroo Imaki, Amy Drohan, Bethany Craig, Joshua Simms, and Michelle Zapp-Sluis were instrumental in dataset assembly and creation, and data were generously provided by Florida Fish and Wildlife Conservation Commission, Alabama Department of Conservation and Natural Resources, Louisiana Department of Wildlife and Fisheries, Mississippi Department of Marine Resources, and Texas Parks & Wildlife Department. Justin Davenport and Alexey Katin of NCSU also provided important technical assistance with results processing and figures. We also thank Song Qian and Yoon Kyung Cha for their methodological input during the early phases of this work. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Funding

This work was supported by the National Oceanic and Atmospheric Administration [Grant Nos. NA14OAR4170299, NA15OAR4170215] and North Carolina State University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolind.2018.02.040.

References

Adams, C.M., Hernandez, E., Cato, J.M., 2004. The economic significance of the Gulf of Mexico related to population, income, employment, minerals, fisheries and shipping. Ocean Coast. Manag. 47 (11), 565–580.
Akin, S., Winemiller, K.O., Gelwick, F.P., 2003. Seasonal and spatial variations in fish and macrocrustacean assemblage structure in Mad Island Marsh estuary, Texas. Estuar. Coast. Shelf Sci. 57 (1), 269–282.
Allison, P.D., 2012. Logistic Regression Using SAS: Theory and Application. SAS Institute.
Baltz, D., Rakocinski, C., Fleeger, J., 1993. Microhabitat use by marsh-edge fishes in a Louisiana estuary. Environ. Biol. Fishes 36 (2), 109–126.
Baker, M.E., King, R.S., 2010. A new method for detecting and interpreting biodiversity and ecological community thresholds. Methods Ecol. Evol. 1 (1), 25–37.
Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effect models using lme4. J. Stat. Softw. 67, 1–48. http://dx.doi.org/10.18637/jss.v067.i01.
Blackburn, J.R., 2015. The Book of Texas Bays. Texas A&M University Press.
Bolker, B.M., 2008. Ecological models and data in R. Princeton University Press.
Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24 (3), 127–135.
Boesch, D.F., Rosenberg, R., 1981. Response to stress in marine benthic communities. In: Barrett, G.W., Rosenberg, R. (Eds.), Stress Effects on Natural Ecosystems. John Wiley & Sons, New York, pp. 179–200.
Breine, J.J., Maes, J., Quataert, P., Van den Bergh, E., Simoens, I., Van Thuyne, G., Belpaire, C., 2007. A fish-based assessment tool for the ecological quality of the brackish Schelde estuary in Flanders (Belgium). Hydrobiologia 575 (1), 141–159.
Breine, J., Quataert, P., Stevens, M., Ollevier, F., Volckaert, F.A., Van den Bergh, E., Maes, J., 2010. A zone-specific fish-based biotic index as a management tool for the Zeeschelde estuary (Belgium). Mar. Pollut. Bull. 60 (7), 1099–1112.
Brett, J.R., Groves, T.D.D., 1979. Physiological energetics. Fish Physiol. 8, 279–352.
Bricker, S.B., Longstaff, B., Dennison, W., Jones, A., Boicourt, K., Wicks, C., Woerner, J., 2008. Effects of nutrient enrichment in the nation's estuaries: a decade of change. Harmful Algae 8 (1), 21–32.
Brown, S.S., Gaston, G.R., Rakocinski, C.F., Heard, R.W., 2000. Effects of sediment contaminants and environmental gradients on microbenthic community trophic structure in Gulf of Mexico estuaries. Estuaries 23 (3), 411–424.
Cabral, H.N., Fonseca, V.F., Gamito, R., Gonçalves, C.I., Costa, J.L., Erzini, K., Gonçalves, J., Martins, J., Leite, L., Andrade, J.P., Ramos, S., 2012. Ecological quality assessment of transitional waters based on fish assemblages in Portuguese estuaries: the Estuarine Fish Assessment Index (EFAI). Ecol. Ind. 19, 144–153.
Caddy, J.F., 2000. Marine catchment basin effects versus impacts of fisheries on semi-enclosed seas. ICES J. Marine Sci. 57 (3), 628–640.
Calaway, R., Revolution Analytics, S. Weston, and D. Tenenbaum, 2015. 'Foreach' Parallel Adaptor for the 'parallel' Package. CRAN. https://cran.r-project.org/web/packages/doParallel/index.html.
Chapra, S.C., 2008. Surface Water-quality Modeling. Waveland press.
Chesney, E.J., Baltz, D.M., 2001. The effects of hypoxia on the northern Gulf of Mexico coastal ecosystem: a fisheries perspective. In: Rabalais, N.N., Turner, R.E. (Eds.), Coastal Hypoxia: Consequences for Living Resources and Ecosystems. American Geophysical Union, Washington, pp. 321–354.
Clark, J.S., Gelfand, A.E., 2006. A future for models and data in environmental science. Trends Ecol. Evol. 21 (7), 375–380.
Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M.V., Wikle, C.K., 2009. Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. Ecol. Appl. 19 (3), 553–570.
Coates, S., Waugh, A., Anwar, A., Robson, M., 2007. Efficacy of a multi-metric fish index as an analysis tool for the transitional fish component of the Water Framework Directive. Mar. Pollut. Bull. 55 (1), 225–240.
Cohen, J., 1990. Things I have learned (so far). Am. Psychol. 45 (12), 1304.
Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Lasram, F.B.R., Aguzzi, J., Ballesteros, E., 2010. The biodiversity of the Mediterranean Sea: estimates, patterns, and threats. PloS one 5 (8), e11842.
Courrat, A., Lobry, J., Nicolas, D., Laffargue, P., Amara, R., Lepage, M., et al., 2009. Anthropogenic disturbance on nursery function of estuarine areas for marine species. Estuar. Coast. Shelf Sci. 81 (2), 179–190.
Cuffney, T., Kashuba, R., Qian, S., Alameddine, I., Cha, Y., Lee, B., Coles, J., McMahon, G., 2011. Multilevel regression models describing regional patterns of invertebrate and algal responses to urbanization across the USA. J. North Am. Benthol. Soc. 30 (3), 797–819.
Dauer, D.M., Ranasinghe, J.A., Weisberg, S.B., 2000. Relationships between benthic community condition, water quality, sediment quality, nutrient loads, and land use patterns in Chesapeake Bay. Estuaries 23 (1), 80–96.
Day, J.W., Britsch, L.D., Hawes, S.R., Shaffer, G.P., Reed, D.J., Cahoon, D., 2000. Pattern and process of land loss in the Mississippi Delta: a spatial and temporal analysis of wetland habitat change. Estuaries 23 (4), 425–438.
de Leiva Moreno, J.I., Agostini, V.N., Caddy, J.F., Carocci, F., 2000. Is the pelagic-demersal ratio from fishery landings a useful proxy for nutrient availability? A preliminary data exploration for the semi-enclosed seas around Europe. ICES J. Marine Sci. 57 (4), 1091–1102.
Deegan, L.A., Finn, J.T., Ayvazian, S.G., Ryder-Kieffer, C.A., Buonaccorsi, J., 1997. Development and validation of an estuarine biotic integrity index. Estuaries 20 (3), 601–617.
Delpech, C., Courrat, A., Pasquaud, S., Lobry, J., Le Pape, O., Nicolas, D., Boët, P., Girardin, M., Lepage, M., 2010. Development of a fish-based index to assess the ecological quality of transitional waters: the case of French estuaries. Mar. Pollut. Bull. 60 (6), 908–918.
Derksen, S., Keselman, H.J., 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br. J. Math. Stat. Psychol. 45 (2), 265–282.
Dethier, M.N., Raymond, W.W., McBride, A.N., Toft, J.D., Cordell, J.R., Ogston, A.S., Heerhartz, S.M., Berry, H.D., 2016. Multiscale impacts of armoring on Salish Sea shorelines: evidence for cumulative and threshold effects. Estuar. Coast. Shelf Sci. 175, 106–117.
Devictor, V., Julliard, R., Jiguet, F., 2008. Distribution of specialist and generalist species along spatial gradients of habitat disturbance and fragmentation. Oikos 117, 507–514.
Diaz, R.J., Rosenberg, R., 2008. Spreading dead zones and consequences for marine ecosystems. Science 321 (5891), 926–929.
Diaz, R.J., Rosenberg, R., 2011. Introduction to environmental and economic consequences of hypoxia. Int. J. Water Res. Dev. 27 (1), 71–82.
Engle, V.D., Summers, J.K., Gaston, G.R., 1994. A benthic index of environmental condition of Gulf of Mexico estuaries. Estuaries 17 (2), 372–384.
Felley, J.D., 1987. Nekton assemblages of three tributaries to the Calcasieu Estuary, Louisiana. Estuaries 10 (4), 321–329.
Flannery, M.S., Peebles, E.B., Montgomery, R.T., 2002. A percent-of-flow approach for managing reductions of freshwater inflows from unimpounded rivers to southwest Florida estuaries. Estuaries 25 (6), 1318–1332.
Fry, F.E.J., 1971. The effect of environmental factors on the physiology of fish. Fish Physiol. 6, 1–98.
Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J.A., Latimer, A., Rebelo, A.G., 2005. Modelling species diversity through species level hierarchical modelling. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 54 (1), 1–20.
Gelman, A., Hill, J., 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2014. Bayesian data analysis Vol. 2 Chapman & Hall/CRC, Boca Raton.
Gelwick, F.P., Akin, S., Arrington, D.A., Winemiller, K.O., 2001. Fish assemblage structure in relation to environmental variation in a Texas Gulf coastal wetland. Estuaries 24, 285–296.
Gittman, R.K., Fodrie, F.J., Popowich, A.M., Keller, D.A., Bruno, J.F., Currin, C.A., Peterson, C.H., Piehler, M.F., 2015. Engineering away our natural defenses: an analysis of shoreline hardening in the US. Front. Ecol. Environ. 13 (6), 301–307.
Gittman, R.K., Scyphers, S.B., Smith, C.S., Neylan, I.P., Grabowski, J.H., 2016. Ecological consequences of shoreline hardening: a meta-analysis. BioScience, biw091.
Gleason, M.G., Newkirk, S., Merrifield, M.S., Howard, J., Cox, R., Webb, M., et al., 2011. A Conservation Assessment of West Coast (USA) Estuaries. The Nature Conservancy.
Gonzalez-Dieseldorff, P., Baltz, D.M., 2008. Habitat use by nekton along a stream-order gradient in a Louisiana estuary. Estuaries Coasts 31, 572–583.
Greene, C.M., Blackhart, K., Nohner, J., Candelmo, A., Nelson, D.M., 2015. A national assessment of stressors to estuarine fish habitats in the contiguous USA. Estuaries Coasts 38 (3), 782–799.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2), 147–186.
Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive

models in studies of species distributions: Setting the scene. Ecol. Model. 157 (2), 89–100.
Harrison, T.D., Whitfield, A.K., 2004. A multi-metric fish index to assess the environmental condition of estuaries. J. Fish Biol. 65 (3), 683–710.
Harrison, T.D., Whitfield, A.K., 2006. Application of a multimetric fish index to assess the environmental condition of South African estuaries. Estuarine Coasts 29 (6), 1108–1120.
Harrison, T.D., Kelly, F.L., 2013. Development of an estuarine multi-metric fish index and its application to Irish transitional waters. Ecol. Indicators 34, 494–506.
Harwell, M.A., 1997. Ecosystem management of south Florida. Bioscience 47, 499–512.
Hawkins, C.P., Olson, J.R., Hill, R.A., 2010. The reference condition: predicting benchmarks for ecological and water-quality assessments. J. North Am. Benthol. Soc. 29 (1), 312–343.
Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J.N., Wickham, J., 2007. Completion of the 2001 national land cover database for the conterminous United States. Photogramm. Eng. Remote Sens. 73 (4), 337–341.
Howarth, R.W., Marino, R., 2006. Nitrogen as the limiting nutrient for eutrophication in coastal marine ecosystems: evolving views over three decades. Limnol. Oceanogr. 51 (1), 364–376.
Hughes, B.B., Levey, M.D., Brown, J.A., Fountain, M.C., Carlisle, A.B., Litvin, S.Y., et al., 2014. Nursery functions of US West Coast estuaries: the state of knowledge for juveniles of focal invertebrate and fish species. Nature Conservancy, Arlington, Virginia.
Hughes, J., Deegan, L., Weaver, M., Costa, J., 2002. Regional application of an index of estuarine biotic integrity based on fish communities. Estuaries Coasts 25, 250–263.
Jenkins, J.A., 2004. Fish bioindicators of ecosystem condition at the Calcasieu Estuary. Louisiana. USGS Open-File Report 2004 1132.
Jordan, S.J., Vaas, P.A., 2000. An index of ecosystem integrity for Northern Chesapeake Bay. Environ. Sci. Policy 3, 59–68.
Jordan, S.J., Smith, L.M., 2005. Indicators of ecosystem integrity for estuaries. CRC Press, Estuarine Indicators. Boca Raton.
Jordan, S.J., Lewis, M.A., Harwell, L.M., Goodman, L.R., 2010. Summer fish communities in northern Gulf of Mexico estuaries: Indices of ecological condition. Ecol. Ind. 10, 504–515.
Kashiba, R., Cha, Y., Alameddine, I., Lee, B., Cuffney, T.F., 2010. Multilevel hierarchical modeling of benthic macroinvertebrate responses to urbanization in nine metropolitan regions across the conterminous United States. USGS Scientific Investigations Report 2009–5243.
King, R.S., Baker, M.E., 2010. Considerations for analyzing ecological community thresholds in response to anthropogenic environmental gradients. Journal of the North American Benthological Society 29 (3), 998–1008.
Latimer, A.M., Wu, S., Gelfand, A.E., Silander, J.A., 2006. Building statistical models to analyze species distributions. Ecological applications 16 (1), 33–50.
Lewis, M., Jordan, S., Chaney, C., Harwell, L., Goodman, L., Quarles, R., 2007. Summer fish community of the coastal northern Gulf of Mexico: Characterization of a large-scale trawl survey. Trans. Am. Fish. Soc. 136 (3), 829–845.
Lewis, M.A., Goodman, L.R., Chaney, C.A., Jordan, S.J., 2011. Fish assemblages in three northwest Florida urbanized bayous before and after two hurricanes. J. Coastal Res. 27 (1), 35–45.
Macauley, J.M., Summers, J.K., Engle, V.D., 1999. Estimating the ecological condition of the estuaries of the Gulf of Mexico. Environ. Monit. Assess. 57, 59–83.
Malins, D.C., Krahn, M.M., Myers, M.S., Rhodes, L.D., Brown, D.W., Krone, C.A., McCain, B.B., Chan, S.L., 1985. Toxic chemicals in sediments and biota from a creosote-polluted harbor: relationships with hepatic neoplasms and other hepatic lesions in English sole (Parophrys vetulus). Carcinogenesis 6 (10), 1463–1469.
Meng, L., Orphanides, C.D., Powell, J.C., 2002. Use of a fish index to assess habitat quality in Narragansett Bay, Rhode Island. Trans. Am. Fish. Soc. 131 (4), 731–742.
Nixon, S.W., Buckley, B.A., 2002. "A striking rich zone": Nutrient enrichment and secondary production in coastal marine ecosystems. Estuaries 25, 782–796.
NMFS (National Marine Fisheries Service), 2014. Fisheries Economics of the United States, 2012. U.S. Department of Commerce NOAA Technical Memorandum NMFS-F/SPO-137.
NRC (Natural Resource Council)., 2000. Clean Coastal Waters: Understanding and Reducing the Effects of Nutrient Pollution. National Academies Press, Washington.
NOAA (National Oceanic and Atmospheric Administration), 2006. Costal Change Analysis Program regional land cover [GIS files]. http://coast.noaa.gov/czapatlas/.
NOAA/NOS (National Oceanic and Atmospheric Administration/National Ocean Service), 1985. National Estuarine Inventory Data Atlas, Volume 1: Physical and Hydrologic Characteristics. Ocean Assessments Division, National Ocean Service, National Oceanic and Atmospheric Administration, Rockville, Maryland pp. 103.
Pasquaud, S., Courrat, A., Fonseca, V.F., Gamito, R., Gonçalves, C.I., Inbry, J., et al., 2013. Strength and time lag of relationships between human pressures and fish-based metrics used to assess ecological quality of estuarine systems. Estuar. Coast. Shelf Sci. 134, 119–127.
Pauly, D., 1995. Anecdotes and the shifting baseline syndrome of fisheries. Trends Ecol. Evol. 10 (10), 430.
Pearson, T.H., Rosenberg, R., 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanogr. Marine Biol. Assoc. Rev. 16, 229–311.
Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R., 1996. A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 49 (12), 1373–1379.
Peterson, M.S., Ross, S.T., 1991. Dynamics of littoral fishes and decapods along a coastal river-estuarine gradient. Estuar. Coast. Shelf Sci. 33, 467–483.
Peterson, M.S., Lowe, M.R., 2009. Implications of cumulative impacts to estuarine and marine habitat quality for fish and invertebrate resources. Rev. Fish. Sci. 17 (4), 505–523.
Piazza, B.P., La Peyre, M.K., 2009. The effect of Hurricane Katrina on nekton communities in the tidal freshwater marshes of Breton Sound, Louisiana, USA. Estuar. Coast. Shelf Sci. 83 (1), 97–104.
Pollutane, S., Zampoukas, N., Borja, A., Davies, S.P., van de Bund, W., Birk, S., 2014. Intercalibration of aquatic ecological assessment methods in the European Union: lessons learned and way forward. Environ. Sci. Policy 44, 237–246.
Pritchard, D.W., 1967. What is an estuary: physical viewpoint. In: Lauff, G.H. (Ed.), Estuaries. AAAS Publications, Washington, pp. 3–5.
Qian, S.S., Cuffney, T.F., Alameddine, I., McMahon, G., Reckhow, K.H., 2010. On the application of multilevel modeling in environmental and ecological studies. Ecology 91 (2), 355–361.
R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna [online] http://www.R-project.org.
Rabalais, N.N., Turner, R.E., Dortch, Q., Justic, D., Bierman Jr, V.J., Wiseman, W.J., 2002. Nutrient-enhanced productivity in the northern Gulf of Mexico: Past, present and future. Hydrobiologia 475 (1), 39–63.
Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1986. Akaike information criterion statistics.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Segurado, P., Santoy, J.M., Pont, D., Melcher, A.H., Jakob, D.G., Hughes, R.M., Ferreira, M.T., 2011. Estimating species tolerance to human perturbation: expert judgement versus empirical approaches. Ecol. Ind. 11, 1623–1635.
Stoddard, J.L., Larsen, D.P., Hawkins, C.P., Johnson, R.K., Norris, R.H., 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. Ecol. Appl. 16 (4), 1267–1276.
Summers, J.K., 2001. Ecological condition of the estuaries of the Atlantic and Gulf coasts of the United States. Environ. Toxicol. Chem. 20, 99–105.
Swihart, R.K., Gehrig, T.M., Kolozsvary, M.B., Nupp, T.E., 2003. Responses of 'resistant' vertebrates to habitat loss and fragmentation: the importance of niche breadth and range boundaries. Divers. Distrib. 9, 1–18.
Toft, Jason D, S.H. Munsch, J. Cordell, K. Sillett, V.C Hare, B. Holycross, L.A. Dehnychare, C.M. Greene, 2015. Nursery Functions of West Coast Estuaries: Data Assessment for Juveniles of 15 Focal Fish and Crustacean Species. Final report to the Pacific Marine and Estuarine Fish Habitat Partnership. http://www.pacificfishhabitat.org/assessment-reports/. **Last Accessed January 2018.**
Trumbley, A., Rensijn, J., 2015. LMERConvenienceFunctions: Model selection and post-hoc analysis for (G) LMER models. R Package Version 2.
Uriarte, A., Borja, A., 2009. Assessing fish quality status in transitional waters, within the European Water Framework Directive: setting boundary classes and responding to anthropogenic pressures. Estuar. Coast. Shelf Sci. 82 (2), 214–224.
USEPA (United States Environmental Protection Agency), 2015. Geospatial data download service. [GIS files] http://www2.epa.gov/enviro/geospatial-data-download-service.
USFWS (U.S. Fish and Wildlife Service), 2012. National wetlands inventory [GIS files]. http://www.fws.gov/wetlands/.
Vázquez, D.P., Simberloff, D., 2002. Ecological specialization and susceptibility to disturbance: conjectures and refutations. Am. Nat. 159, 606–623.
Whittier, T.R., Hughes, R.M., Lomnicky, G.A., Peck, D.V., 2007. Fish and amphibian tolerance values and an assemblage tolerance index for streams and rivers in the western USA. Trans. Am. Fish. Soc. 136 (1), 254–271.
Wikle, C.K., 2003a. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84 (6), 1382–1394.
Wikle, C.K., 2003b. Hierarchical models in environmental science. Int. Stat. Rev. 71 (2), 181–199.
Wikle, C.K., Berliner, L.M., Cressie, N., 1998. Hierarchical Bayesian space-time models. Environ. Ecol. Stat. 5 (2), 117–154.
Wilson, S.K., Burgess, S.C., Cheal, A.J., Emslie, M., Fisher, R., Miller, I., Polunin, N.P., Sweatman, H.P., 2008. Habitat utilization by coral reef fish: implications for specialists vs. generalists in a changing environment. J. Anim. Ecol. 77, 220–228.
Yáñez-Arancibia, A., Day, J.W., 2004. The Gulf of Mexico: Towards an integration of coastal management with large marine ecosystem management. Ocean Coast. Manag. 47, 537–563.

SM1: Estuary-level data for anthropogenic stressors (* denotes estuary with LDC in the GoM)

					Anthropogenic stressors			
Estuary	**State**	**Area (km^2)**	**Volume (km^3)**	**Mean estuary salinity**	**% Basin Anthropogenic**	**% Shoreline Agriculture**	**EDA Toxic Releases /Estuary Area (# permits /km^2)**	**Basin Crop / Flow (km^2*d//m^3))**
Florida Bay	FL	1663	1.03	29	13.5	**0.0***	**0.0***	2.2E-04
South Ten Thousand Islands	FL	227	0.144	21	26.9	**0.0***	**0.0***	2.0E-04
North Ten Thousand Islands	FL	390	0.284	27	**8.9***	**0.0***	**0.0***	5.7E-05
Rookery Bay	FL	35	.017	29	38.5	1.3	**0.0***	2.8E-04
Caloosahatchee River	FL	67	0.126	13	55.0	1.8	31.3	1.8E-04
Charlotte Harbor	FL	502	0.819	22	53.3	1.5	23.3	2.4E-04
Sarasota Bay	FL	124	0.272	29	68.4	4.2	18.5	3.6E-05
Tampa Bay	FL	902	2.70	25	53.7	6.4	34.9	1.5E-04
Suwannee River	FL	165	0.193	17	26.4	0.8	7.9	1.3E-04
Apalachee Bay	FL	1773	3.40	29	19.1	0.2	2.6	1.3E-04
Apalachicola Bay	FL	593	1.08	19	30.4	0.1	1.2	1.0E-04
St. Andrew Bay	FL	252	0.714	20	11.3	0.7	21.4	**3.7E-06***
Choctawhatchee Bay	FL	340	1.29	14	27.0	1.3	3.2	8.0E-05
Pensacola Bay	FL	477	1.44	18	20.9	1.8	25.6	4.3E-05
Perdido Bay	AL	129	0.198	13	28.9	3.7	31.1	9.3E-05
Mobile Bay	AL	1079	2.06	15	24.4	2.1	43.1	3.4E-05
East Mississippi Sound	MS	654	1.53	14	18.3	0.8	24.6	1.4E-05
West Mississippi Sound	MS	1581	3.84	15	18.6	1.8	3.7	1.3E-05
Lake Borgne	LA	744	1.34	13	48.5	2.5	24.8	2.4E-04
Lake Pontchartrain	LA	1879	6.58	11	52.6	4.1	24.3	4.0E-04
Breton/Chandeleur Sound	LA	4301	11.8	21	21.9	6.4	1.19	5.5E-04
Barataria Bay	LA	852	0.362	10	18.8	14.2	33.3	9.6E-04
Terrebonne/Timbalier Bays	LA	1262	0.854	13	34.2	18.5	1.1	6.1E-03
Calcasieu Lake	LA	260	0.311	13	23.9	19.5	103.1	9.4E-05
Sabine Lake	TX	265	0.661	12	26.0	11.5	323.1	2.5E-05
Galveston Bay	TX	1456	2.25	15	33.0	8.8	186.7	1.9E-04
Matagorda Bay	TX	1115	1.57	16	21.6	29.2	10.9	1.5E-03
San Antonio Bay	TX	587	0.347	15	34.0	16.3	16.7	2.0E-04
Aransas Bay	TX	524	0.515	17	53.8	16.6	1.3	2.5E-03
Corpus Christi Bay	TX	571	1.53	28	18.9	7.1	49.4	5.9E-03
Baffin Bay	TX	239	0.137	29	56.0	10.9	10.9	4.3E-02
Upper Laguna Madre	TX	591	0.202	29	44.7	**0.0***	0.3	4.8E-02
Lower Laguna Madre	TX	1308	0.999	29	66.9	1.4	4.7	2.6E-03
Average				**19**	**33.6**	**5.9**	**32.3**	**3.5E-03**
Standard Deviation				**7**	**16.7**	**7.3**	**63.4**	**1.1E-02**

SM2: Estuary-level data used for preliminary analyses and prescreening. Note: I(0:3) indicates an integer variable (0, 1, 2, or 3).

Estuary	State	Average Depth (m)	Percent open to sea (%)	Mean flow (m^3/d)	Hypoxic condition I(0:3)	Toxic algal Condition I(0:3)	Eutrophic Condition I(0:5)
Florida Bay	FL	0.6	7	5.0E05	3	2	3
South Ten Thousand Islands	FL	0.6	8	3.3E06	3	0	4
North Ten Thousand Islands	FL	0.7	7	4.2E06	2	0	4
Rookery Bay	FL	0.5	8	5.9E04	1	0	4
Caloosahatchee River	FL	1.9	2	4.2E06	1	0	3
Charlotte Harbor	FL	1.6	3	5.1E06	3	2	3
Sarasota Bay	FL	2.2	1	4.9E05	2	3	4
Tampa Bay	FL	3.0	1	3.5E06	1	2	4
Suwannee River	FL	1.2	12	2.6E07	1	1	1
Apalachee Bay	FL	1.9	16	9.6E06	1	2	2
Apalachicola Bay	FL	1.8	3	6.4E07	2	2	3
St. Andrew Bay	FL	2.8	0	1.8E06	2	1	3
Choctawhatchee Bay	FL	3.8	0	1.9E07	2	3	1
Pensacola Bay	FL	3.0	0	2.6E07	2	0	2
Perdido Bay	AL	1.5	0	3.9E06	2	1	5
Mobile Bay	AL	1.9	1	1.6E08	2	2	2
East Mississippi Sound	MS	2.3	3	3.4E07	2	1	1
West Mississippi Sound	MS	2.4	2	3.1E06	1	1	1
Lake Borgne	LA	1.8	3	3.1E07	0	2	1
Lake Pontchartrain	LA	3.5	0	1.2E07	1	2	3
Breton/Chandeleur Sound	LA	2.7	3	2.4E05	1	0	3
Barataria Bay	LA	0.4	0	6.0E05	2	1	4
Terrebonne/Timbalier Bays	LA	0.7	3	6.5E04	1	3	3
Calcasieu Lake	LA	1.2	0	8.4E06	2	3	5
Sabine Lake	TX	2.5	0	4.4E07	1	1	1
Galveston Bay	TX	1.5	1	5.6E07	1	1	2
Matagorda Bay	TX	1.4	1	9.9E06	1	1	3
San Antonio Bay	TX	0.6	0	6.6E06	1	1	3
Aransas Bay	TX	1.0	0	5.9E05	1	1	3
Corpus Christi Bay	TX	2.7	0	4.9E05	1	3	3
Baffin Bay	TX	0.6	3	7.2E04	1	3	5
Upper Laguna Madre	TX	0.3	1	7.5E04	2	3	3
Lower Laguna Madre	TX	0.8	0	2.1E06	1	3	5

SM3: Correlation matrix of highly significant estuary-level stressor variables

	Basin Anthropogenic/ A_L	**Estuary salinity**	Basin Agriculture/ A_L	Basin Crop/ A_L	**Basin Crop/ Q**	Basin Anthropogenic/ Q	**Shoreline Agriculture/ A_L**	Basin Agriculture/ Q	**EDA toxic releases/ A_E**	Basin Urban/ Q
Basin Anthropogenic/ A_L	1.00									
Estuary salinity	0.21	1.00								
Basin Agriculture/ A_L	0.78	0.01	1.00							
Basin Crop/ A_L	0.65	0.10	0.89	1.00						
Basin Crop/ Q	0.33	0.38	0.52	0.51	1.00					
Basin Anthropogenic/ Q	0.33	0.40	0.51	0.49	0.996	1.00				
Shoreline Agriculture/ A_L	-0.01	-0.35	0.08	-0.03	0.03	0.04	1.00			
Basin Agriculture/ Q	0.33	0.39	0.51	0.49	0.998	1.00	0.04	1.00		
EDA toxic releases/ A_E	-0.09	-0.33	-0.07	-0.19	-0.12	-0.13	0.24	-0.12	1.00	
Basin Urban/ Q	0.33	0.42	0.41	0.41	0.89	0.92	0.12	0.90	-0.16	1.00

SM4: Multi-stressor model coefficients for individual fish and invertebrate species. Event-level coefficient units = $\frac{\Delta\, log\ odds}{\Delta\, unit\ of\ predictor}$. Estuary-level coefficients are presented for normalized values, units = $\frac{\Delta\, log\ odds}{\Delta\, standard\ deviation\ of predictor}$. Estuary coefficients are shaded dark grey for negative relationships and light grey for positive relationships. Species with stars (*) denote pelagic species.

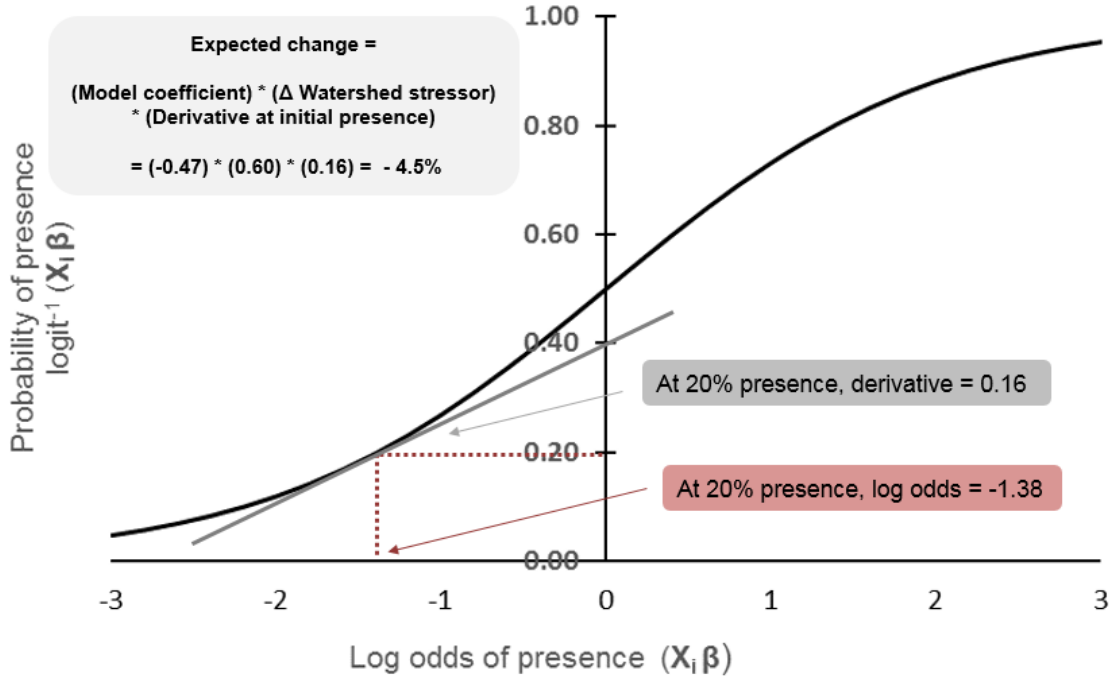
			Event-Level Coefficients				Estuary-Level Coefficients				
Species	Scientific name	% present In trawls	Temperature	Salinity	Salinity Squared	Distance to Shore	Estuary Salinity	% Basin Anthropogenic	% Shoreline Agriculture	EDA toxic releases / Area	Basin Crop / Flow
Atlantic Croaker	*Micropogonias undulatus*	52%	0.04	-0.06	0.0009	0.07	-0.49	-0.47	0.00	0.00	0.00
Bay Anchovy*	*Anchoa mitchilli*	49%	-0.02	0.02	-0.0008	0.05	-0.27	0.00	0.00	0.00	0.00
Blue Crab	*CallinectessSapidus*	42%	0.01	-0.01	-0.0007	0.00	0.00	0.20	0.00	-0.16	0.00
Spot	*Leiostomus xanthurus*	39%	0.02	0.02	-0.0008	0.02	0.00	-0.41	0.00	0.00	0.00
White Shrimp	*Litopenaeus setiferus*	35%	-0.02	-0.01	-0.0006	0.06	0.00	-0.86	0.00	0.00	0.00
Brown Shrimp	*Farfantepenaeus aztecus*	34%	0.06	0.05	-0.0014	0.06	0.00	-0.97	0.00	0.00	0.65
Hardhead Catfish	*Ariopsis felis*	30%	0.09	0.04	-0.0011	0.01	0.00	0.00	0.29	0.00	0.00
Sand Seatrout	*Cynoscion arenarius*	28%	0.07	0.05	-0.0023	0.06	0.00	0.00	0.00	0.00	-0.69
Pinfish*	*Lagodon rhomboides*	27%	0.03	0.05	-0.0003	-0.06	0.47	0.00	0.00	0.00	0.00
Gulf Menhaden*	*Brevoortia patronus*	20%	-0.01	-0.06	0.0000	0.00	0.00	-0.53	0.42	0.00	0.00
Silver Perch	*Bairdiella chrysoura*	20%	0.02	0.04	-0.0010	-0.05	0.29	0.00	0.00	-0.40	-0.22
Pink Shrimp	*Farfantepenaeus duorarum*	13%	-0.06	0.12	-0.0029	0.06	0.00	0.00	0.00	-1.28	0.00
Least Puffer	*Sphoeroides parvus*	12%	-0.02	0.15	-0.0040	0.03	0.00	0.00	0.00	0.00	0.00
Bay Whiff	*Citharichthys spilopterus*	11%	0.04	0.00	0.0000	0.00	0.00	-1.07	0.00	0.00	0.00
Gafftop Catfish	*Bagre marinus*	10%	0.17	0.05	-0.0017	0.00	0.00	0.00	0.62	0.00	-0.48
Blackcheek Tonguefish	*Symphurus plagiusa*	9%	-0.02	0.10	-0.0023	0.11	0.00	0.00	0.00	-0.58	-0.46
Fringed Flounder	*Etropus crossotus*	9%	-0.03	0.18	-0.0030	0.12	0.00	-0.63	0.00	0.00	-0.82
Atlantic Stingray	*Dasyatis sabina*	8%	-0.02	0.04	-0.0008	-0.02	0.00	0.00	0.00	-0.77	-0.27
Inshore Lizardfish	*Synodus foetens*	8%	0.03	0.17	-0.0027	0.02	0.00	0.00	0.00	0.00	-0.39
Hog Choker	*Trinectes maculatus*	7%	0.05	-0.08	0.0004	0.00	0.00	0.50	0.00	-0.25	0.00
Bighead Searobin	*Prionotus tribulus*	7%	-0.06	0.18	-0.0047	0.11	-0.40	0.00	0.00	0.00	0.00

SM4 (continued)

Atlantic Bumper*	*Chloroscombrus chrysurus*	7%	0.07	0.23	-0.0043	0.04	0.00	-0.68	0.00	0.00	-1.17
Ground Mullet	*Menticirrhus americanus*	7%	0.00	0.25	-0.0064	0.00	0.00	0.00	0.00	0.00	0.00
Atlantic Spadefish*	*Chetodipterus faber*	7%	0.09	0.16	-0.0035	0.00	0.00	0.00	0.37	0.34	-0.55
Pigfish	*Orthopristis chrysoptera*	7%	0.04	0.21	-0.0026	-0.05	0.85	0.00	0.00	0.00	-0.68
Gulf Butterfish*	*Peprilus burti*	6%	-0.05	0.25	-0.0048	0.18	0.00	-0.94	0.82	0.00	0.00
Spotted Seatrout	*Cynoscion nebulosus*	6%	-0.09	-0.04	0.0004	0.00	0.00	0.34	0.00	-0.18	0.33
Atlantic Cutlassfish*	*Trichiurus lepturus*	6%	0.05	0.15	-0.0032	0.13	0.00	0.00	0.48	0.00	-0.45
Threadfin Shad*	*Petenense*	5%	0.00	0.06	-0.0028	0.00	-0.77	-0.80	0.00	0.00	0.00
Southern Flounder	*Paralichthys lethostigma*	5%	0.03	-0.07	0.0008	-0.03	0.00	-0.60	0.00	0.00	0.00
Striped Mullet*	*Mugil cephalus*	5%	-0.09	-0.07	0.0007	-0.18	0.00	0.00	0.39	0.00	0.00
Black Drum*	*Pogonias cromis*	5%	-0.05	-0.09	0.0012	-0.17	0.51	0.00	0.79	0.61	0.50
Striped Anchovy*	*Anchoa hepsetus*	4%	0.12	0.08	-0.0015	0.03	0.00	-0.45	0.00	0.00	-0.95
Silver Jenny	*Eucinostomus gula*	3%	0.03	0.19	-0.0026	0.00	0.00	0.00	0.00	0.00	0.00
Lookdown*	*Selene vomer*	3%	0.02	0.16	-0.0031	0.00	0.00	-0.39	0.00	0.00	-0.82
Star Drum*	*Stellifer lanceolatus*	3%	0.00	0.10	-0.0017	0.18	0.00	0.00	1.14	0.00	0.00
Gulf Toadfish	*Opsanus beta*	3%	0.07	0.05	-0.0009	0.00	0.88	0.00	0.00	0.00	0.00
Lined Sole	*Achirus lineatus*	3%	0.03	0.04	-0.0003	0.06	0.00	1.20	0.00	0.00	0.00
Atlantic Moonfish*	*Selene setapinnis*	3%	0.07	0.24	-0.0047	0.14	0.00	-1.19	0.00	0.00	0.00
Chain Pipefish	*Syngnathus lousianae*	2%	0.00	0.12	-0.0011	0.00	0.00	0.00	0.00	-0.86	0.00
Silver Seatrout	*Cynoscion nothus*	2%	0.02	0.03	-0.0007	0.30	0.00	0.00	0.00	0.00	0.00
Crevalle Jack*	*Caranx hippos*	2%	0.06	0.00	-0.0005	-0.04	0.00	0.00	0.51	0.00	-0.61
Blue Catfish	*Ictalurus furcatus*	1.9%	0.00	-0.45	0.0060	0.00	0.00	0.00	0.00	0.00	0.00
Scaled Sardine*	*Harengula jaguana*	1.8%	0.11	0.10	-0.0015	0.08	0.00	0.00	0.71	0.00	-0.54
Gizzard Shad*	*Dorosoma cepedianum*	1.7%	0.00	-0.05	-0.0015	-0.11	0.00	0.00	0.00	0.00	0.00
Spotfin Mojarra	*Eucinostomus argenteus*	1.5%	0.00	0.08	0.0000	0.00	0.00	0.00	0.00	0.00	-0.81
Gulf Pipefish	*Syngnathus scovelli*	1.5%	0.00	0.04	-0.0007	-0.06	0.64	0.00	-0.90	-1.50	0.00
Naked Goby	*Gobiosoma bosc*	1.3%	-0.04	-0.12	0.0014	0.00	0.63	0.50	0.00	0.00	0.00
Lane Snapper	*Lutjanus synagris*	1.1%	0.09	0.25	-0.0028	0.00	0.00	0.00	0.00	0.00	-1.06
Red Drum	*Sciaenops ocellatus*	1.1%	0.00	-0.08	0.0000	-0.17	0.71	0.00	0.00	0.00	0.00

SM4 (continued)

Atlantic Threadfin Herring	*Opisthonema oglinum*	1.0%	0.00	0.07	-0.0016	0.20	-1.56	0.00	0.00	-0.86	0.00
White Mullet*	*Mugil curema*	1.0%	0.03	0.11	-0.0027	-0.21	0.00	0.00	0.92	0.00	0.00
Bluntnose Jack	*Hemicaranx amblyrhynchus*	0.9%	0.07	0.24	-0.0048	0.09	0.00	-0.75	0.00	0.00	0.00
Spanish Mackerel	*Scomberomorus maculatus*	0.8%	0.11	0.00	-0.0006	0.04	0.00	-0.80	0.00	0.00	0.00
Ladyfish	*Elops saurus*	0.8%	0.08	-0.02	0.0000	0.00	0.00	0.00	0.00	0.00	0.00
Lined Seahorse	*Hippocampus erectus*	0.6%	-0.07	0.28	-0.0035	0.00	0.53	0.00	0.00	0.00	-1.16
Sea Bass	*Centropristis philadelphica*	0.3%	0.00	0.10	0.0000	0.08	0.00	-1.16	0.00	0.00	0.00



SM5. Predictions of species presence for changing watershed stressor levels using multi-stressor models.

Multi-stressor species models can be used to estimate changes in the probability of species presence due to the effect of shifting watershed stressor values. This is done by multiplying the slope of the logistic function that corresponds to the current species presence with the species model coefficient and the magnitude of the watershed stressor change. For example, suppose Atlantic Croaker (*Micropogonias undulates*) is currently present 20% of the time in a given estuary where anthropogenic land use (Basin Anthropogenic/ A_L) is expected to increase from 30% to 40% (a difference of 10%). The probability of species presence (20%) is first converted to a log odds value of -1.38 (ln $(\frac{\%\ present)}{(1-\%\ present)})$) and its derivative is 0.16 $(\frac{e^{log\ odds}}{(1+2*e^{log\ odds}+e^{2*log\ odds})})$. The 10% increase in Basin Anthropogenic/ A_L is then normalized by its standard deviation across estuaries (Online Resource 3) to a value of 0.60 $(\frac{10\%}{16.7\%})$. To determine the effect on species presence, the derivative (0.16) is multiplied by the model

coefficient for Atlantic Croaker (-.47; Online Resource 4) and the normalized change in the stressor (0.60), yielding a value of -0.045. Therefore, for a 10% increase in Basin Anthropogenic/ A_L, the probability of presence for Atlantic Croaker is expected to be reduced from 20% down to 15.5%.

Chapter 2: Assessing Natural and Anthropogenic Drivers of Ecological Health in Piedmont Streams using Hierarchical Modeling

Abstract

Urban streams consistently have poorer ecological condition than natural streams due to a myriad of anthropogenic impacts that alter hydrology and increase pollutant concentrations. Though monitoring of urban streams is widespread, viable management options for improving stream health are not well defined. Understanding more precisely the factors that influence biological condition, as well as being able to identify sites that deviate from expected conditions, can help managers develop more efficient stream restoration strategies. A hierarchical (multilevel) framework is used here to model over 3,000 macroinvertebrate samples from the North Carolina Piedmont, identify important natural and anthropogenic factors that relate to stream health, and demonstrate how hierarchical random effects can help identify potential restoration sites. In addition, the spatial (e.g., watershed versus stream buffer) and temporal (e.g., age of construction) aspects of land cover development are explored within the model. Results confirm that the best predictor for biotic index (BI) is watershed impervious cover (IC). Additional factors significantly related to BI include the age of watershed IC, canopy loss in stream buffers, reservoirs, wastewater treatment plants, antecedent precipitation, and geologic soil conditions. Synthesizing these factors in a hierarchical multiple linear regression, we explain 76% of the variability in the BI (R^2) relative to 65% using watershed IC only. Remaining spatial variability is largely accounted for by site random effects (16%), which characterize the deviation between expected and actual biological condition. Of note, model results suggest that newer development (post-1980) degrades stream health 30% less than older development, and

canopy removal in stream buffers has 2 to 9 times the impact on BI on a per hectare basis, when compared to the addition of IC in upstream watersheds.

Keywords: biological assessment, anthropogenic stressors, watershed development, stream restoration, temporal land cover variability, spatial scales, macroinvertebrates

1. Introduction

Water quality in the United States (US) is degraded by various anthropogenic factors and over 40% of the nation's streams are considered to be in poor ecological condition (USEPA 2006, 2009, 2016). Urban streams, in particular, are degraded at even higher rates due to the "urban stream syndrome", which refers to the myriad of negative effects that urban settings have on stream networks from altered hydrology to increased pollutant loads (Schueler 1994, Paul and Meyer 2001, Walsh et al. 2005). The increase of impervious surfaces in urban settings alters stream health by rapidly transporting rainfall and contaminants through stormwater conveyance systems to streams (Makepeace et al. 1995, Deletic 1998, Vietz et al. 2014). Percent imperviousness and other proxies representing urbanization (e.g., road and population density) have been well established as variables that negatively correlate to stream health (Schueler 1994, Walsh et al. 2005, Bressler et al. 2009, Kashuba et al. 2010, Cuffney et al. 2010), but many scientific and methodological questions remain such as 1) how to identify multiple covarying stressors and their relative impacts (Wenger et al. 2009), 2) how to determine the effectiveness of construction and land development best management practices (BMPs) (Parr et al. 2016, Utz et al. 2016), and 3) how research and modeling efforts can characterize ecosystem health in ways that improve communication between scientists, managers, and stakeholders (Wenger et al. 2009).

The primary physical change in urban streams, elevated and more frequent peak flows, results from conversion of natural land covers to impervious cover (IC; Leopold 1968, Seaburn 1969, Booth and Jackson 1997, Boggs and Sun 2011). Though the effect of urbanization on altered hydrology can vary regionally due to natural conditions such as slope and soil permeability (Hopkins et al. 2015), development within stream basins of 10% total IC or of 2-3%

IC that is directly connected to stormwater conveyance systems can severely alter ecologically important stream attributes like physical habitat and geomorphology (Schueler 1994, Booth and Jackson 1997, Vietz et al. 2014). Sediment loads from construction activities can blanket stream bottoms and destroy habitat for biota that live in hyporheic zones (Cordone and Kelley 1961, Finkenbine et al. 2000, Miltner et al. 2004), while elevated peak flows erode and scour stream beds leading to incised channels and high banks where riparian vegetation can become hydraulically disconnected from stream baseflows (Groffman et al. 2003, Bernhardt and Palmer 2007). Elevated water temperatures also occur in urban streams due to removal of canopy cover, reduced baseflows (Klein 1979, Rose and Peters 2001, Hardison et al. 2009), and runoff from paved surfaces during summer months (Poole and Berman 2001, Nelson and Palmer 2007). Hydrologic variations, as well as elevated temperatures, have been shown to negatively impact macroinvertebrate and fish communities, common ecological indicators of stream health (Roy et al. 2005, Dunbar et al. 2010, Buchwalter et al. 2015, Elbrecht et al. 2016).

Biochemical changes in streams are caused by pollutants (i.e., nutrients, hydrocarbons, metals, pesticides, herbicides, etc.) entering from non-point sources such as urban or agricultural land (Makepeace et al. 1995, Paul and Meyer 2001, Walsh et al. 2005, Bernhardt and Palmer 2007), or point sources like wastewater treatment plants (WWTP; Dennehy et al. 1998, Paul and Meyer 2001). Increased conductivity in streams (i.e., secondary salinization; Cañedo-Argüelles et al. 2013) has also been shown to affect sensitive macroinvertebrate taxa in both mesocosms and natural studies (Kefford et al. 2011, Clements and Kotalik 2016, Boehme et al. 2016). Copper, lead, and zinc, common in road runoff (Legret and Pagotto 1999, Huber et al. 2016), settle into stream sediments where they negatively affect macroinvertebrate species (Phipps et al. 1995, Beasley and Kneale 2002, Hirst et al. 2002). In both urban and agricultural areas,

herbicides such as atrazine are widely used (Grube et al. 2011, Yun and Qian 2015) and negatively affect macroinvertebrate communities (Graymore et al. 2001). In addition to the urban stream syndrome, previous studies have also shown that both current agricultural (crop and pasture lands), as well as historical agricultural use, have been related to poor ecological stream health (Richards et al. 1996, USEPA 2006, 2009, 2016, Kashuba et al. 2010, Qian et al. 2010, Cuffney et al. 2010, 2011a, Hill et al. 2017).

Empirical modeling has shown a consistently negative relationship between macroinvertebrate community metrics and IC or urbanization (Schueler et al. 2009, Paul et al. 2009, Cuffney et al. 2011a). Regression approaches have included simple linear regressions (SLR; Schueler 1994, Booth et al. 2005), multiple linear regressions (MLR; Strayer et al. 2003, Bellucci et al. 2013), quantile regression (Paul et al. 2009), and hierarchical SLRs (Kashuba et al. 2010, Qian et al. 2010, Cuffney et al. 2011a). In all SLRs and MLRs, predictor variables are assigned (slope) coefficients, which can be used to understand the magnitude of their effect on response variables. Quantile regression, a process where a linear regression is derived through a certain quantile of the data (e.g., 95%), was used to determine an upper limit on biological communities that can be expected for urban streams (Paul et al. 2009). Hierarchical models can extend both SLR and MLR by allowing coefficients to vary by group (e.g., ecoregion, sampling location; Gelman and Hill 2006). There has been debate whether linear responses are indicative of the relationship between biological responses and urbanization and if linear responses are consistent throughout the whole range of urbanization (Allan 2004, Walsh et al. 2005, King et al. 2010, Cuffney et al. 2011b). Linear regressions are able to evaluate both linear and non-linear relationships between predictors and response, where nonlinear relationships are typically specified through variable transformation (e.g., log transformation; Faraway 2014). More

recently, machine learning techniques (e.g. random forest; Breiman 2001) have been used to predict stream health in the conterminous United States using over 190 predictor variables (Hill et al. 2017). Random forests can rank the relative importance of highly correlated predictors and have good predictive ability (Evans and Cushman 2009, Evans et al. 2011), but their unconstrained (non-parametric) form makes interpretation of the relationship between predictors and responses difficult.

Hierarchical modeling (mixed-effects modeling or multi-level modeling) is an extension of conventional regression modeling (e.g., SLR, MLR) that includes both “random” and “fixed” effects (Gelman and Hill 2006). Fixed effects refer to model coefficients fit similarly to MLRs (e.g., intercept, slope, and categorical coefficients; Faraway 2014), while random effects represent group-level differences in the data (e.g., sampling locations, ecoregions, etc.) accounting for spatial, temporal, or intraclass correlation not resolved by the available predictor variables (Gelman and Hill 2006, Gelman et al. 2014). Random effects differ from categorical variables in that random effect coefficients are drawn from a common hyperdistribution (Gelman and Hill 2006). In addition to more robust parameter estimation (accounting for intraclass correlation), hierarchical models typically have better predictive ability than conventional regressions for grouped data (Wikle 2003, Qian and Shen 2007, Cressie et al. 2009, Qian et al. 2010), and are a common framework for modeling ecological data (Clark and Gelfand 2006, Bolker et al. 2009). Hierarchical modeling has been used to compare regional differences in ecological responses to anthropogenic stressors (Qian et al. 2010, Cuffney et al. 2011a), to enhance models of watershed pollutant loading and transport (Qian et al. 2005, Strickling and Obenour 2018), and to link the biological condition of receiving waterbodies to upstream watershed stressors (Miller et al. 2018). Although not as flexible as machine learning approaches

(Breiman 2001, Olden et al. 2008, Evans et al. 2011), hierarchical modeling allows us to explore spatial and temporal variability using relationships that are readily interpretable and accessible based on their parametric form and statistical significance (Gelman and Hill 2006, Gelman et al. 2014).

The goal of this study is to address remaining urban stream research questions including how to quantify the relative impacts of co-varying stressors, test the effectiveness of improved development practices on stream health, and demonstrate how to connect and communicate research efforts with managers and stakeholders (Wenger et al. 2009, Parr et al. 2016, Utz et al. 2016). Using a hierarchical multiple linear regression (i.e., hierarchical MLR), we identify major natural and anthropogenic factors that alter the ecological health of streams using over 3000 biotic index (BI) samples collected over a 30-year period in the North Carolina (NC) Piedmont. A wide range of possible watershed predictors including the age of IC in watersheds, three spatial scales of land covers, temporally-varying hydro-climatological conditions, seasonality, geologic soil regions, and upstream reservoirs and WWTPs are tested using various statistical measures of model fit. The inferred impacts of stressors on BI are quantified and compared to determine the range of their effects on stream health. Finally, this study explores how a hierarchical framework can help inform management decisions by identifying potential restoration sites.

2. Methods

This study used hierarchical MLRs to model the BI of wadeable streams in the NC Piedmont. Data collected by the North Carolina Department of Environmental Quality (NC DEQ) and 5 major NC cities were compiled for model construction (Sections 2.1-2.2). Initially, a baseline model was created by comparing the amount of variance explained by different

representations of IC (Section 2.4.1). Then, auxiliary stressor models tested the importance of other candidate predictor variables (Section 2.3) using an information criterion approach (Section 2.4.2). Finally, a multi-stressor model was assembled to assess the combined effects of multiple natural and anthropogenic factors on BI (Section 2.4.3).

2.1. Study area

All sampling locations were located in the NC Piedmont region, which is comprised of 6 hydrologic units at the sub-basin level (Hydrological Unit Code 8): Neuse, Cape Fear, Tar-Pamlico, Roanoke, Yadkin-Pee Dee, and Santee Rivers (Fig 1), and 8 major geologic regions (Raleigh Belt, Carolina Slate Belt, Kings Mountain Belt, Inner Piedmont, Milton Belt, Charlotte Belt, Sauratown Mountains Anticlinorium, and the Triassic Basin; NC Geological Survey 2009; Fig S1, Table S1) as well as several minor ones. Triassic Basin soils exist in 3 non-adjacent sub-regions of the NC Piedmont: Durham-Sanford, Wadesboro and the Dan River subbasins. Piedmont borders were set consistent with definitions established by the NC State Climate Office (SCO; http://sco-web1.meas.ncsu.edu/climate/drought/historical, accessed June 2017), which intersects 34 counties. Elevations for sampling locations in the dataset ranged from 34-546 m above sea level. The Lower Pee-Dee River basin (i.e., Sandhills region) was excluded from this study because it is geologically part of the Coastal Plain with high-permeability soils that lead to different flow and ecological conditions compared to the rest of the study area (NC Geological Survey 2009; NRCS 2012).

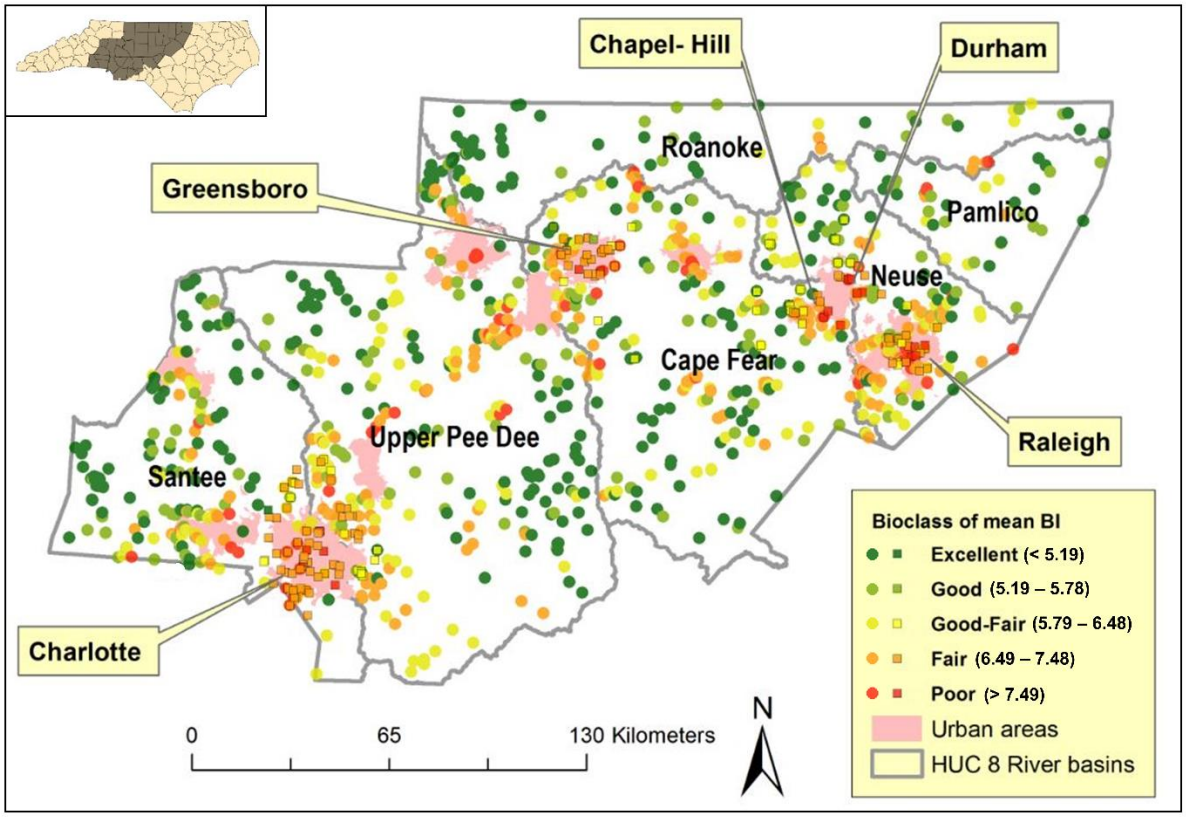


Figure 1: NC Piedmont with mean Biotic Index (BI) score for each sampling site. City sites are represented as squares while NC Department of Environmental Quality (NC DEQ) sites are circular. Bioclassification ranges in the legend are based on full scale sampling methods (NC DEQ 2016). Upper left insert shows the study area (gray) relative to the state of NC and its counties.

2.2. Macroinvertebrate data and monitoring locations

Macroinvertebrate data were obtained from the Biological Assessment Branch of NC DEQ (personal communication, Michael Walters, 2016) and 5 NC cities (Table 1). City of Raleigh data were provided by its Stormwater Management Division (personal communication, Lory Willard, 2017), while other city data were compiled by Tetra Tech (Paul 2015). NC DEQ sampling dates back to 1982 while city data were all collected after 2000 (Table 1). Only 2 years of Raleigh data were available for this study because consistent BI scores were not available prior to 2013. All macroinvertebrate samples include a BI score that is a weighted-average

tolerance score based on the abundance and pollution tolerance of the taxa collected (Hilsenhoff 1988, Mulholland and Lenat 1992, NC DEQ 2016). A total of 3127 samples collected from 1132 sites draining less than 500 km^2 (92% of all sampling sites; Fig S2) were included in the analysis. The 500 km^2 cutoff approximates 5th order streams and below, which are generally considered to be wadeable (USEPA 2006, Pierson et al. 2008).

NC DEQ uses different stream sampling methods depending on study objectives and watershed size. In this study, we utilized data based on 3 NC DEQ methods: full scale, EPT, and Qual 4 (NC DEQ 2016). The most common method for collection by NC DEQ, full scale (69% of samples), uses multiple techniques to collect all macroinvertebrate species present (i.e., 2 riffle kicks, 3 sweeps, etc.; NC DEQ 2016). The second most common sampling method, EPT (24% of samples), is used for rapid assessments. In the EPT method, fewer collection techniques are performed (1 riffle kick and 1 sweep), and only pollution sensitive taxa (i.e., Ephemeroptora, Plecoptera, and Trichoptera) are recorded and used for calculating the BI (NC DEQ 2016). The Qual 4 method (7% of samples) is performed for rapid assessments of smaller watersheds (< 7.7 km^2) and is a hybrid of the full scale and EPT method. Collection techniques (1 riffle kick, 1 sweep) are similar to the EPT method, but because lower numbers of EPT species are expected in smaller watersheds (NC DEQ 2016), all macroinvertebrate are collected for BI calculations similar to the full scale method. All city samples were collected using the full scale method.

Upstream watersheds for sampling sites were delineated using 30-m digital elevation models downloaded from the U.S. Geological Survey (USGS) National Map (Accessed 20 July 2017 at http://viewer.nationalmap.gov/viewer/). Processing was performed using Spatial Analyst tools within ArcGIS 10.3 (Esri 2015). Two additional spatial scales for each site were also processed: stream buffers and “near buffers”. Stream buffers were defined as land within 30 m of

all National Hydrograph Dataset Plus (NHD+, Moore and Dewald 2016) flowlines within the upstream watershed of each sampling site. Thirty meters has been suggested as the minimum distance needed to protect the physical, chemical, and biological health of streams (Sweeney and Newbold 2014). Near buffers were calculated as the 30 m stream buffers located within 500 m upstream of each site. The resolution of land cover data (60 m) did not allow for discerning buffer characteristics at smaller scales. Land cover variables (described below) were determined for all 3 spatial scales.

Table 1: Summary of macroinvertebrate data from different sampling programs, including the number of sites, number of samples, date range, and mean and standard deviations (SD) of watershed size, biotic index (BI), and percent of impervious cover (IC).
3

Sampling program	# of sites	# of samples	Data range	Drainage areas (km^2)		BI		IC (%)	
				Mean	SD	Mean	SD	Mean	SD
Chapel Hill	13	51	2000-2014	24.8	23.8	6.17	0.40	4.7	2.5
Charlotte	72	478	2000-2013	43.0	45.1	6.92	0.61	18.9	9.6
Durham	20	149	2002-2014	88.5	132.9	6.97	0.94	14.6	9.4
Greensboro	24	140	2000-2014	39.0	53.3	6.57	0.87	17.3	12.0
NC DEQ	984	2272	1982-2015	91.4	108.7	5.65	1.21	6.1	7.8
Raleigh	19	37	2013-2014	24.1	48.2	7.05	0.57	25.2	9.1

2.3. Candidate predictor variables

Both natural and anthropogenic factors were used as candidate predictor variables in model construction (Table 2). Each variable class (e.g., agriculture, upstream WWTP, etc.) included several candidate variables in order to test different representations of the various predictors. For instance, land cover data were compiled at different spatial scales and as both linear and log (for saturating effects) predictors. In addition, upstream reservoirs were recorded at different distances from sampling locations, and antecedent hydro-climatological conditions were summarized at 5 different temporal scales (1, 3, 6, 12, and 24 months). All log

transformations were performed using the natural logarithm. For each class, we determined the candidate predictor most closely related to BI (Section 2.4.2).

2.3.1. Land cover data

Four land cover-related variables were used in model construction: IC, canopy loss, agriculture, and wetlands because all are expected to impact stream health (Table 3; Schueler 1994, Richards et al. 1996, Schueler et al. 2009, Cuffney et al. 2010, 2011a, Qian et al. 2010, Bellucci et al. 2013, Hill et al. 2017). Land cover variables were derived from the U.S. Conterminous Wall-to-wall Anthropogenic Land use Trends (NWALT) dataset for the years 1974, 1982, 1992, 2002, and 2012 (Falcone 2015), and the National Land Cover Dataset (NLCD) for years 2001 and 2011 (Homer et al. 2007, Homer et al. 2015). The NWALT dataset included temporally consistent land cover records for: 1) low and high density residential areas, 2) transportation, 3) industrial and commercial development, 4) semi-developed (forested land adjacent to development and partially used for that purpose), 5) agriculture (crop and pasture lands), 6) low use (forests), and 7) wetlands. Land cover data were linearly interpolated in time to obtain year specific land cover values between 1974 and 2012, while land covers for 2012-2015 were linearly extrapolated using land cover trends from 2002 to 2012. Therefore, each sample was associated with a land cover estimate specific to the year it was collected.

IC and canopy cover were not available in NWALT and instead were obtained from NLCD for 2001 and 2011. To impute IC for other years, a MLR model was constructed using NLCD IC values (for all sampling site watersheds in 2001 and 2011) as the response variable and NWALT land classifications for corresponding years as the predictor variables (residential areas, transportation, wetlands, etc.). The impervious MLR model was constructed

Table 2: 16 variable classes and their associated candidate predictor variables. At most, 1 candidate predictor variable for each variable class was chosen for inclusion in the hierarchical multiple linear regression (Section 2.4).

Variables class	**Candidate predictor variables**	**Source**
Age of construction	IC before: 1974,1980,1990, and 2000	Falcone, 2015 Homer et al., 2007, 2015
Agriculture	Percent of watershed, stream buffer and near buffer (linear and log transformed)	Falcone, 2015
Antecedent hydro-climatological conditions	Precipitation (m; previous 1,3,6,12, and 24 months) Drought indices (SPI; previous 1,3,6,12,24 months) (linear and quadratic)	SCO, 2017
Canopy loss	Percent of watershed, stream buffer and near buffer (linear and log transformed)	Falcone, 2015 Homer et al., 2015
Collecting agencies	DEQ and 5 city agencies that collected samples	NC DEQ ; Paul, 2015
Drainage area	Size of entire upstream watershed (km^2) Sites that drain less than 7.7 km^2	USGS, 2017
Elevation (m)	Elevation at sampling location	USGS, 2017
IC	Percent of watershed, stream buffer and near buffer (linear and log transformed)	Falcone, 2015 Homer et al., 2007, 2015
Recent construction	Change in percent of IC in previous 2, 5 yrs. for watershed, stream buffer and near buffer (linear and log transformed)	Falcone, 2015 Homer et al., 2007, 2015
Sampling method	3 sampling methods (full scale, EPT, and Qual 4)	NC DEQ, 2016
Season	Season of macroinvertebrate sample	NC DEQ, 2016; Paul, 2015
Soil conditions	10 geological regions of NC Piedmont Triassic soils that drain less than 428 km^2	NC Geological Survey, 2009
Upstream reservoirs	Reservoirs within 1 and 5 km of sampling locations	Moore and Dewald, 2016
Upstream WWTPs	Major and minor WWTPs within 1, 5 10, and 20 km of sampling locations Wastewater ($\frac{m^3}{s}$) normalized by site drainage area (km^2)	NC DEQ, 2016
Wetlands	Percent of watershed, stream buffer and near buffer (linear and log transformed)	Falcone, 2015
Year	Year of the macroinvertebrate sample	NC DEQ, 2016; Paul 2015

with no intercept such that coefficients would correspond to the actual percent of NWALT land covers expected to be impervious (i.e., on average, 17% of residential areas were impervious, 35% of transportation, and 57% of industrial/commercial). Land cover coefficients that were not significant at the 95% level were parsed out of the MLR model, which explained 95% of the

variation in IC. Canopy cover values were also imputed using a MLR constructed in a similar way, with the exception that only 2011 NLCD data were used. 2001 NLCD canopy cover values were excluded because they were substantially biased compared to 2011 values (-11% on average; Coulston et al. 2012, Homer et al. 2015). The canopy cover MLR model explained 89% of the variability in canopy cover and implied on average 92% of low use, 88% of semi-developed, 48% of residential, and 12% of industrial/commercial development have canopy cover. Canopy cover was then converted to canopy loss for inclusion as a predictor variable by subtracting mean canopy cover values from the maximum possible value in the dataset (i.e., 92% for low use areas).

Table 3: Summary of percent land covers for 3 spatial extents of upstream watersheds. Summaries are for 1974 and 2012, with mean (M) and 10th and 90th percentiles shown.

Land cover	Watershed				Stream buffer				Near buffer			
	1974		2012		1974		2012		1974		2012	
	M	(10,90)	M	(10,90)	M	(10,90)	M	(10,90)	M	(10,90)	M	(10,90)
Agriculture	22.3	(2,45)	18.1	(1,41)	12.4	(0,27)	9.3	(0,22)	11.0	(0,42)	8.0	(0,30)
Wetlands	0.5	(0,1)	0.5	(0,1)	2.2	(0,6)	2.2	(0,6)	8.1	(0,33)	8.1	(0,33)
Impervious	6.1	(0,19)	8.9	(0,24)	4.1	(0,14)	6.1	(0,18)	3.9	(0,14)	5.7	(0,19)
Canopy loss	33.7	(12,52)	35.7	(15,53)	23.9	(8,41)	25.8	(10,44)	27.9	(0,68)	29.6	(0,69)

Temporal variability in IC was explored by considering the age of construction as well as recent construction in upstream watersheds. Variables for the log of IC constructed before given dates (Table 2) were calculated by multiplying the linear or log transformed estimate of IC by the ratio of IC at the given date to IC at the time of sample collection. We tested variables based on several break points (e.g., IC before 1974, IC before 2000) because state regulations and enforcement have come into effect incrementally over the duration of this study period. Break points beyond 2000 were not used due to the paucity of BI sampling in watersheds with more recently constructed IC. In addition, “recent construction” was tested by calculating impervious land cover change in the 2 or 5 years preceding sample collection.

2.3.2. Natural factors

Several variable classes were comprised of natural factors that might affect macroinvertebrate communities, including: geologic soil type, elevation, season of the sampling event, and upstream watershed drainage area. A categorical variable was created for the season of each sample: spring (March-May), summer (June- Sept), fall (Oct.-Nov), and winter (Dec.-Feb.). The summer season was extended to include September to be consistent with NC monitoring protocols (NC DEQ 2016). All sites that drained watersheds less than 7.7 km^2 and samples in Triassic basins that drained less than 428 km^2 were also tested as categorical variables, because both are expected to affect BI (NC DEQ 2016).

Antecedent hydro-climatological conditions are represented using drought indices [i.e., Standardized Precipitation Index (SPI; McKee et al. 1993)] and monthly precipitation estimates obtained from the NC State Climate Office (SCO, personal communication 2017). SPI represents a statistical z-score which is easy to interpret and less variable than other drought indices (McKee et al. 1993, Guttman 1998). Estimates were derived for 3 sub-regions within the Piedmont: Northern, Central, and Southern Piedmont and aggregated over different time scales (e.g., previous 1, 3, 6, 12, and 24 month periods). Both drought and precipitation estimates were represented using a linear term as well as a second order term (i.e., quadratic relationship) to test for non-linear effects between precipitation and stream health that might occur because both droughts and flooding have been shown to negatively affect macroinvertebrate communities (Fisher et al. 1982, Boulton 2003, Sponseller et al. 2010).

2.3.3. Upstream point source anthropogenic factors

Numerous sampling locations were located downstream of reservoirs in our dataset. NHD+ flowlines and reservoir polygons were used to determine 2 downstream conditions represented as categorical variables: 1) sampling sites within 1 km of any size reservoir and 2) sites within 5 km of large reservoirs that intercept more than 50% of the upstream watershed. There were 89 samples taken within 1 km of reservoirs and 208 within 5 km of large reservoirs in the dataset. Upstream WWTPs were also identified at different distances from sampling locations (1, 5, 10, or 20 km), and represented with categorical variables. Major WWTPs that discharged more than 4,546 m^3/day (1 million gallons per day) were separated from minor WWTPs that discharged below that threshold. Based on preliminary inspection of the data, NC DEQ regularly sampled 3 to 4 km downstream of major WWTPs and sporadically sampled directly below outfalls of both major and minor WWTPs. Finally, a continuous variable defined by the amount of wastewater discharged upstream of each sampling site divided by upstream drainage area was also tested to see if the ratio of wastewater discharge to watershed size affects BI. In this dataset, there were 57 sites within 1 km of major WWTPs, 204 within 5 km, 246 within 10 km, and 278 within 20 km. There were also 60 sites within 1 km of minor WWTP, 224 within 5 km, 322 within 10 km, and 354 within 20 km.

2.4. Model construction

2.4.1. Baseline impervious model

Weaker statistical relationships can be obscured or reversed by stronger relationships (Samuels 1993) and this motivated us to create a working baseline model using IC before screening other candidate predictors. All 3 spatial scales of IC, both linear and log transformed,

were tested to determine which representation of IC (x_i) explains the most variation (i.e., coefficient of determination, R^2) in BI. Natural log transformations of percent IC were performed after adding 1 to avoid non-real transformations. A categorical variable for the sampling method (α_{sm}; i.e., full scale, EPT, or Qual 4) along with site random effects (α_{st}) were also included in the baseline model. We expected different sampling techniques to potentially bias BI scores, while site random effects were necessary to account for intraclass correlation of samples collected at the same location (Gelman et al. 2014). Overall, the baseline model formulation is:

$$BI = \beta_0 + \alpha_{sm} + \alpha_{st} + x_i * \beta_i + \varepsilon \qquad \text{(Eq.1)}$$

$$\alpha_{st} \sim N(0, \sigma^2_{st}) \quad \varepsilon \sim N(0, \sigma^2_{BI})$$

where β_0 is the model intercept, β_i is the coefficient for the IC representation, and ε represents the residual error which is normally distributed with variance σ^2_{BI}. The site random effects come from a normal distribution centered around zero with variance σ^2_{st}. All hierarchical models were fit in R using the "lme4" package (Bates et al. 2015, R Core Team 2017). Of note, maximum likelihood (ML) estimation was used to fit hierarchical models [instead of restricted maximum likelihood (REML)] in order to better compare information criterion values across models (Faraway 2006, Zuur et al. 2009). Model sensitivity to ML vs. REML was low and did not result in any substantial changes to overall results.

2.4.2. Auxiliary stressor models

Once a baseline model was established, all candidate predictor variables (Table 2) were tested to assess if they improved model performance, using an information criterion approach.

Variables were added to the baseline model in 3 ways depending on expected plausible relationships: (1) adding a single linear or log-transformed variable (x_a; Eq. 2a), (2) adding the variable and its 2nd order (quadratic) term (Eq. 2b) for precipitation and drought indices, and (3) adding a categorical variable or random effect (α_a; Eq. 2c).

$$BI = \text{Baseline Model} + x_a * \beta_a \quad \text{(Eq.2a)}$$

$$BI = \text{Baseline Model} + x_a * \beta_a + x^2_a * \beta_{asq} \quad \text{(Eq.2b)}$$

$$BI = \text{Baseline Model} + \alpha_a \quad \text{(Eq.2c)}$$

All candidate predictor variables from each variable class (Table 2) were run separately in the auxiliary stressor models and compared to the baseline model (Eq.1) based on the Bayesian Information Criterion (BIC). BIC is an information criteria method that judges model fit based on log likelihood while penalizing for extra parameters (Schwarz 1978). A lower BIC scores implies better predictive performance.

2.4.3. Multi-stressor model

A multi-stressor model was then created by combining the best auxiliary stressor from each variable class (Table 2) together with the baseline model (Eq. 1). To be conservative in adding variables, only auxiliary stressors that lowered BIC by a minimum of 6 in auxiliary stressor models were considered for inclusion (Raftery 1995). If multiple candidate predictors for a given variable class lowered BIC beyond this threshold, the one that reduced BIC the most was considered for the multi-stressor model. Finally, backwards variable selection was performed using the “LMERConvenienceFunctions” R package (Tremblay and Ransihn 2015) to prune any variables with insignificant coefficients from the multi-stressor model (Faraway 2006). Variables with large ranges were centered on their mean in order to not artificially inflate the intercept (i.e.,

year -1999; elevation - 268 m). Variance inflation factors (VIF) were also recorded to test for excessive multi-collinearity between predictors (Cohen et al. 2013).

The relative importance of each predictor in the multi-stressor model was quantified by the magnitude of the standard deviation (SD) of its contribution to BI. That is, for each predictor in the multi-stressor model, influence on BI ($x^* \beta$) was calculated for all samples and the SD of this distribution was used to rank the predictor's relative contribution. In the case of precipitation, the contribution of both the linear and quadratic terms were evaluated together. This process of comparing the relative contribution of predictors is similar to comparing the magnitude of coefficients determined from normalized predictor variables, but more flexible and able to evaluate quadratic relationships (e.g., precipitation) as well as categorical variables (e.g., site random effect, upstream reservoirs).

A 3-fold cross-validation was performed to determine the consistency of predictor coefficients across subsets of the data, and to better assess how the multi-stressor model will perform when predicting at new locations or under future conditions (Elsner and Schmertmann 1994, Chatfield 2006). In each fold, 2/3 of the data were used to calibrate the full model and then predictions were made for the 1/3 of the data that was excluded. In this way, predictions were made across the entire dataset using models calibrated from other subsets of the data. The R^2 and mean absolute error (MAE) of calibration and validation results were compared using predictions without site random effects included, mirroring how predictions would be performed at unsampled locations. In addition, the model coefficients of each fold were recorded and compared to assess robustness.

Finally, the multi-stressor model was used to predict BI scores for all sampling locations in order to demonstrate how model outputs could be used to identify potential stream restoration

sites. Predictions were made for 2012 land covers, the last year of reported data (NWALT; Falcone 2015), and used mean conditions for antecedent hydro-climatological conditions (0.30 m for 3-month precipitation). Predictions were then divided into a deterministic portion which corresponded to known predictors of BI (e.g., IC, canopy loss, upstream WWTP, geologic soil) and site random effects which represented unaccounted for variation in BI across sites. The deterministic predictions and site random effects were then plotted together to show how both offer insights into the current biological condition and the potential for restoration at each location.

3. Results

3.1. Baseline impervious model

Different baseline models (Eq.1) were tested using IC at 3 spatial scales (watershed, stream buffer, and near buffer) as both linear and log transformed predictors. The form of IC that explained the most variance in BI was the log of watershed IC (65%; Table 4), followed closely by the log of stream buffer IC (64%). Watershed IC as a linear predictor explained 60% of the variation in BI, a 5 percent drop from the log transformed value. Log transformation of IC implied a saturating relationship with stream health, such that as total watershed IC increased, further additions had a reduced effect on BI. For example, an increase from 0 to 5% IC corresponded to a change in BI of 1.16, while an increase from 10% to 15% resulted in a change of just 0.25. Of note, IC in the near buffer had both a smaller effect on BI and explained less variance than both the watershed and stream buffer variables.

Table 4: Candidate baseline models for BI using different representations of impervious cover (IC) ranked by their coefficient of determination (R^2) and mean absolute error (MAE). Variables for each spatial scale were either linear or log transformed. β_i is the coefficient for the representation of IC, shown with standard error (se), and β_{ni} is the normalized coefficient. Units for β_i are BI per unit of predictor and units for β_{ni} are BI per 1 standard deviation of the predictor.

Representation of IC	R^2	MAE	β_i (se)	β_{ni}
Watershed (log)	0.65	0.55	0.63 (0.02)	0.71
Stream buffer (log)	0.64	0.56	0.62 (0.02)	0.70
Watershed (linear)	0.60	0.59	0.06 (0.002)	0.65
Stream buffer (linear)	0.56	0.62	0.07 (0.003)	0.61
Near buffer (log)	0.42	0.72	0.35 (0.02)	0.44
Near buffer (linear)	0.38	0.77	0.03(0.002)	0.42

3.2. Auxiliary stressor models

Candidate predictor variables (Table 2) were individually tested in auxiliary stressor models (Eq. 2a, b, c) and compared to the baseline model (Eq. 1) using BIC. The anthropogenic variable that improved the baseline model the most was the log of canopy loss in stream buffers, while other anthropogenic variables of note were: IC before 1980, agriculture in upstream watersheds, wetlands in stream buffers, major WWTP within 1 km of sites, and upstream reservoirs within 5 km of sites (Table 5). Though only 1 predictor from each variable class was included in Table 5; often, several auxiliary predictors within the same class improved the baseline model and offered insights into variation within that class of predictors. For example, adding age of construction improved (i.e., reduced) BIC for all break points (i.e., 1974, 1980, 1990, 2000; Table S2), and coefficients implied that the effect of post IC was gradually decreasing through time. Upstream major WWTPs had negative effects on streams (positive effects on BI) at all distances upstream of sampling locations (1, 5, 10, and 20 km), but these

effects reduced substantially as the distance upstream increased from 1 to 5 km (+0.85 to +0.48 BI; Table S2).

The natural auxiliary stressor that improved (i.e., reduced) BIC the most was underlying geologic soil conditions (as a random effect; Table 5). A seasonal categorical variable and a quadratic relationship with antecedent precipitation (in previous 3-months) also improved the baseline model implying that optimum rainfall for stream health is 0.37 m (i.e., 80th percentile).

Table 5: Best performing candidate predictor from each variable class, as determined by auxiliary stressor models, along with corresponding model coefficients. ΔBIC (Bayesian Information Criterion) represents the difference in BIC achieved by adding the auxiliary stressor to the baseline model (lower is better). The number in parentheses for hydro-climatological conditions is the second order coefficient in the quadratic relationship. Coefficients for the seasonal effect include winter, spring, and fall, as compared to the baseline of summer. Units for coefficients are *Δ* BI/unit of predictor; categorical variables are marked with an asterisk (*). The mean and standard deviation of the normal hyperdistribution (N(μ,σ)) are presented for the soil random effects. The last column includes coefficients for the multi-stressor model, with some variables being parsed out due to non-significant relationships (p>0.05).

Variable class	Best candidate predictor variable	Auxiliary model coefficient	ΔBIC	Multi-stressor coefficient
Age of construction	Pre-1980 IC (log percent)	0.27	-33.3	0.13
Agriculture	Percent of watershed (log)	0.0084	-18.8	-
Hydro-climatological conditions	3-Month precipitation (m)	-3.4 (4.6)	-33.9	-3.2 (4.1)
Canopy loss	Percent of stream buffer (log)	0.43	-106.6	0.44
Drainage area	Area less than 7.7 km^2*	-0.22	-6.1	-
Elevation	Elevation (m)	-0.0008	-25.9	-
Season	4 Seasons*	-0.11,-0.17,0.15	-32.5	-0.24,-0.24,0.05
Soil conditions	10 Soil regions*	N(0,0.45)	-152.6	N(0,0.41)
Upstream reservoirs	Reservoir within 5 km*	0.33	-10.3	0.36
Upstream WWTP	Major WWTP within 1 km*	0.85	-21.3	0.65
Wetlands	Percent of stream buffer (log)	0.13	-12.5	-
Year	Year of sample (yr)	-0.013	-80.1	-0.010

Low levels of precipitation (0.16 m; 5th percentile) would raise BI scores 0.20 on average while elevated levels of precipitation (0.49 m; 95th) would increase BI by 0.07. Finally, a long-term temporal trend (-0.010 BI/year) increased model fit implying that stream health has been

improving slightly throughout the study period. Variable classes with no candidate predictors that improved the baseline model are omitted from Table 5 (i.e., recent construction, collecting agency, etc.). Auxiliary stressor model results for all candidate predictor variables are presented in Table S2.

3.3. Multi-stressor model

The best performing candidate predictors (i.e., auxiliary stressors) from each variable class were combined with the baseline model to create a multi-stressor model. Backwards variable selection was performed to parse out insignificant coefficients, leaving 8 of the 12 candidate auxiliary predictors (Table 5). The multi-stressor model explained 76% of the variation (R^2) in BI, an 11% increase from the baseline model. If site random effects were included, 92% of the variation in BI was explained, though site random effects would be unknown if predicting at new sites. Coefficients in the multi-stressor models were consistent in sign with the baseline and auxiliary stressor models, but some coefficients were reduced in magnitude (Table 5). In particular, the coefficient for IC lowered from 0.65 in the baseline model to 0.31 in the multi-stressor model as other predictors accounted for more of the variability in BI (Table 4, S3). The coefficient for pre-1980 IC is understood in combination with the coefficient for the log of watershed IC (at time of sample). Both being selected in the model implied that the effect of pre-1980 construction was the combination of the two coefficients (i.e. 0.31 + 0.13 = 0.44), while the effect for post-1980 IC was only related to the coefficient for the log IC (0.31). On average, this implied that post-1980 IC affected stream health 30% less than pre-1980 IC.

Different sampling methods (full scale, EPT, and Qual 4) and seasonal offsets were included as categorical variables in the multi-stressor model. The EPT sampling method resulted

in significantly lower BI scores (-0.90 BI; Table S3) than the full scale method, while Qual 4 scores were slightly lower (-0.05; Table S3) but not statistically significantly different from full scale method. Seasonal offsets were compared to the summer sampling season in NC for macroinvertebrates (June – Sept) as a baseline. Stream health based on BI improved in both winter and spring (-0.24 for both) while it declined in fall (+0.05), though this coefficient was not statistically different than summer. 10 different geologic regions were modeled by the soil random effects and their values ranged from -0.59 (Smith River Allochthon) to +0.70 (Triassic soils; Table S1). The y-intercept for the multi-stressor model was 4.48.

The influence of each predictor and random effect in the multi-stressor model was calculated as the magnitude of the SD of its contribution to BI (see Methods 2.4.3; Fig 2). The two largest deterministic predictors of variation in BI, log of watershed IC (0.33 BI) and the sample method (0.38 BI) were both in the baseline model (Eq. 1). Other anthropogenic sources of variation in BI came from pre-1980 IC, canopy loss in stream buffers, major WWTPs, and reservoirs. Some predictors (e.g., WWTP within 1 km and reservoirs within 5 km) contributed smaller amounts to BI because they did not affect many sites in the dataset. However, the magnitude of WWTP and reservoir coefficients (+0.65, +0.36, respectively; Table 5), indicated the substantial negative influence they can exert when present. The largest natural factor affecting BI was the geologic soil condition followed by seasonal adjustments and antecedent precipitation. Site random effects also explained a substantial portion of BI variation (0.39 BI) implying there are other site specific factors not accounted for by predictors used in this model. Model residuals, which largely explain temporal variability at site locations, also contributed a substantial portion to BI (0.39 BI).

The effect of IC and canopy loss were both calculated on a per hectare basis to compare their relative impacts on stream health. The average watershed size in this study was 5500 hectares and the average size of stream buffers was 700 hectares. Using multi-stressor model coefficients, the expected Δ BI based on hectares of impervious cover (IC) added to upstream watersheds or canopy removed from stream buffers was calculated (Fig 3). On a per hectare basis, canopy loss in stream buffers is 2 to 9 times more harmful to stream health than adding IC to the overall watershed, depending on how much development has already occurred.

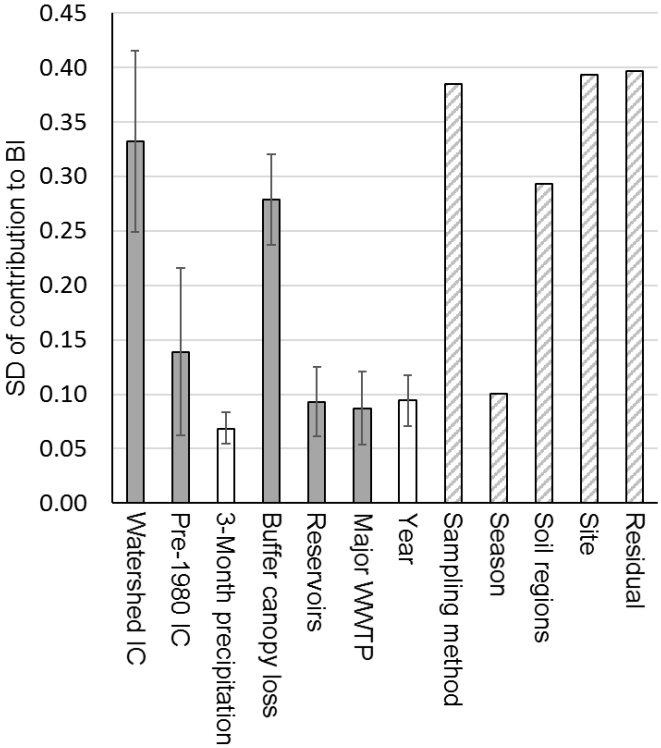


Figure 2: Relative importance of predictor variables, site random effects, and model residuals based on the standard deviation (SD) of their contribution to the biotic index (BI) across all observations. 95% error bars reflect the uncertainty in model coefficients, where applicable. Hollow represents variables that have negative relationships with BI (i.e., better stream health), while grey represents the opposite. Striped bars reflect variables with three or more categories, which are not exclusively positive or negative.

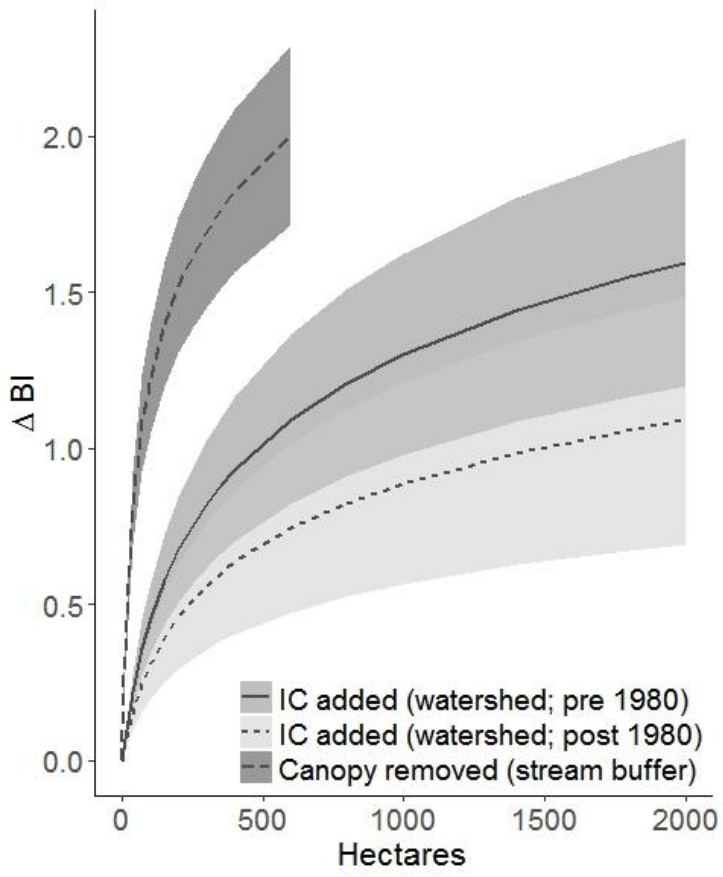


Figure 3: Expected biotic index (BI) change based on hectares of impervious cover (IC) added to upstream watersheds or canopy removed from stream buffers. For the watersheds in this study, the mean stream buffer area was 700 hectares, while the mean upstream watershed was 5,500 hectares. 95% confidence intervals are presented for each variable.

Predicted BI scores for the region were split into a deterministic portion (Fig 4A) and site random effects (Fig 4B). The deterministic portion represents the expected ecological health at a location based on known upstream watershed characteristics, while site random effects account for other sources of spatial variation not accounted for in our model. In general, a low (i.e., negative) site random effect implies that the site is in better ecological condition than would be expected based on the known predictors, while a high site random effects suggests worse-than-expected ecological health.

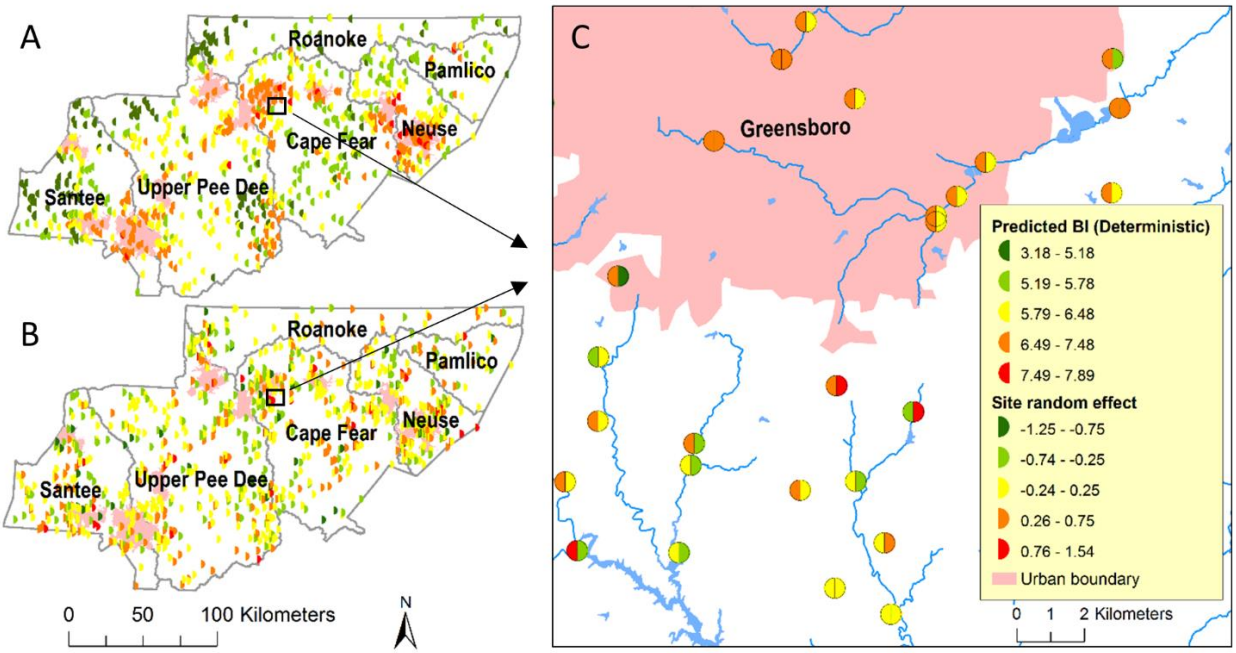


Figure 4: Predicted Biotic index (BI) for sites based on deterministic parts of the multi-stressor model (A) and site random effects (B). A zoomed in view of south Greensboro NC (C) combines predicted BI (left semicircle) and site random effects (right semicircle). Positive random effects imply ecological health is worse than expected based on deterministic predictions, while negative random effects suggests the opposite.

3.4. Model diagnostics

A 3-fold cross validation was applied to test the robustness of the multi-stressor BI model. The R^2 for validation predictions (0.653) had a moderate reduction from the multi-stressor model calibrated to the full dataset (0.756), while MAE increased slightly (0.468 to 0.557; Table S3). Model coefficients did not vary greatly or change sign across the three folds, indicating robust performance and parameterization. 3-month precipitation coefficients showed the most variation in cross validation, maybe due to the lack of spatial precision in precipitation estimates. VIF was also calculated and no predictor variables had a value above 10, implying that multi-collinearity was mild and should not strongly affect model parameterization (Table S3; Kutner et al. 2005).

4. Discussion

The goal of this study was to identify natural and anthropogenic factors that relate to stream health while rigorously quantifying other sources of variability (e.g., sampling procedures and site variation) using hierarchical MLRs. Initially, a baseline model was constructed using the best representation of IC (log of watershed IC) along with site random effects and a categorical variable for the sampling method (Table 4). Then, additional candidate predictors were tested as auxiliary stressors (Table S2) and combined to create a multi-stressor model for BI (Table 5; S3). The multi-stressor model allowed us to compare the impact of predictor variables on BI as well as quantify differences in sampling procedures, geologic soil regions, and seasonal offsets. Various empirical approaches have been used to model macroinvertebrate communities before including: regression modeling (Schueler 1994, Strayer et al. 2003, Paul et al. 2009, Cuffney et al. 2011a, Kushaba et al. 2011, Bellucci et al. 2013), multivariate modeling (Wright 1995, Hawkins et al. 2000, Clarke et al. 2003), and machine learning techniques (Hill et al. 2017). These studies focused on the relationship between IC and stream health (Schueler 1994, Schueler et al. 2009), predicting stream health throughout large regions (Bellucci et al. 2013, Hill et al. 2017), and determining standards to inform management of urban streams (Paul et al. 2009). Our work, however, is the first regional study to model macroinvertebrates using a hierarchical MLR approach to quantify and compare the magnitude of natural and anthropogenic stressor's contributions to stream health. Using temporally specific predictor data allowed us to compare the effect of older vs. newer IC and hydro-climatological conditions on BI as well. Finally, we demonstrated how hierarchical models can identify potential restoration sites based on site-specific deterministic predictions and site random effects (Fig 4).

4.1. Natural and anthropogenic predictors of ecological health

4.1.1. Land covers

Based on the baseline and multi-stressor models, IC in upstream watersheds and canopy loss in stream buffers were the two most important landscape variables associated with stream health. Baseline models (Table 4) indicated that IC has a saturating (log-transformed) relationship with BI, and that watershed-scale impervious cover was more influential than impervious cover in the (30 m) stream buffer alone. These findings support previous research showing small amounts of imperviousness can severely affect stream health (Cuffney et al. 2010, Vietz et al. 2014). In addition, this saturating relationship with IC generally supports the theory of Walsh et al. (2005), positing that streams degrade along an urbanization gradient until they reach a lower limit (i.e. maximum degradation). The log of canopy loss in stream buffers was the second most important land cover in the multi-stressor model and is associated with both urban and agricultural lands, as shown in the canopy cover MLR model (Section 2.3.1). Even small amounts of riparian removal can affect stream substrate, sedimentation levels (Richards et al. 1996, Sutherland et al. 2002, Allan 2004), stream temperature (Pluhowski 1972, Beschta and Taylor 1988, LeBlanc et al. 1997), and negatively correlate to ecological stream health (Steedman 1988, Meador and Goldstein 2003). Our results indicate that the preservation of buffers in the entire upstream network may be an effective management strategy, especially in newly developing watersheds, as removing canopy in stream buffers impacts BI 2 to 9 times more per hectare than adding IC in watersheds (Fig 3).

Land covers not selected in auxiliary and multi-stressor models are also of note. No land cover variables representing the “near buffer” improved model performance based on BIC. This

finding is consistent with previous research that has shown that riparian buffers at smaller spatial scales (e.g., reach or restoration sites) have not been consistently shown to correlate with increased biological health (Wahl et al. 2013) if large-scale watershed restoration is not also performed (Selvakumar et al. 2009, Matthews et al. 2010, Palmer et al. 2014). Also, though agriculture in upstream watersheds and wetlands in stream buffers were chosen in the auxiliary stressor models, both were parsed out of the multi-stressor model, implying these land covers have weaker relationships with BI in this region. Agricultural lands (crop and pasture) have been shown to degrade stream quality (Richards et al. 1996, EPA 2006, 2009, Cuffney et al. 2010, Hill et al. 2017), but ecological degradation has been more closely associated with highly elevated levels of croplands (Fitzpatrick et al. 2001, Allan 2004) rather than pasture lands which have less influence on stream health (Meador and Goldstein 2003, Strayer et al. 2003). Agriculture represented roughly 1/5 of land covers in this study (Table 3) and the majority (> 90%) was devoted to pasture rather than cropland. Wetlands, which represented a small fraction of land in upstream watersheds (< 1 % of watersheds and < 6% in stream buffers; Table 3) were identified in auxiliary stressor models as being negatively related to stream health, which is contrary to past studies (Table S2; Strayer et al. 2003, Bellucci et al. 2013). However, when all candidate predictor variables were included in the multi-stressor, the coefficient for wetlands switched signs. Though wetlands was eventually parsed from the multi-stressor model during backward variable selection for not being statistically significant, this sign change implies that wetlands might have a positive effect on stream health when other factors are accounted for.

4.1.2. Natural factors

Variation in geologic soil condition was the largest natural contributor to BI in our study (Fig 2). Triassic soils were the most negatively related to stream health (+0.70, Table S1), while stream health in several soil groups (e.g., Sauratown Mountains. -0.57) was generally better. Sites that drained pre-dominantly Triassic soils also improved model fit as an auxiliary stressor (+0.42; Table 2, S2) but it did not improve the baseline model as much as including all 10 soil regions. This implies that though sites in the Triassic regions appear to have a negative relationship with stream health, as recognized by state monitoring protocols (NC DEQ 2016), this soil region is not an extreme outlier as compared to other regions in the NC Piedmont. In general, soil random effects reflect soil properties that may influence sedimentation and hydrologic properties of streams, but they might also capture some spatial variability in BI attributable to east-west environmental gradients in the NC Piedmont, such as temperature, slope, or elevation. Soil regions farther west generally had more negative coefficients, implying better stream health. Various antecedent hydro-climatological conditions (i.e., precipitation and drought indices), especially those calculated on time scales of 1 to 6 months, improved model performance based on BIC (Table S2). A quadratic relationship best fit the data implying that both low as well as high flows negatively affect stream health, consistent with past research (Fisher et al. 1982, Bunn and Arthington 2002, Boulton 2003, Lake 2003, Sponseller et al. 2010). Though the magnitudes of extreme low and high conditions (i.e. 5% and 95% percentile) were not large (+0.20, +0.07, respectively), inter-annual variability in precipitation is increasing over time in the Southeast United States (Laseter et al. 2012, Burt et al. 2018), implying that the negative impact of antecedent hydro-climatological conditions on stream ecology might increase in the future.

4.1.3. Upstream point sources

Our multi-stressor model (Table 5) implied that both upstream reservoirs and major WWTPs negatively impact streams for localized areas (+0.36, +0.65, respectively; Table 5). Reservoirs are associated with altered stream ecosystems, probably attributable to altering natural flow regimes and reducing base flows through increased evaporation (Munn and Brusven 1991, Monk et al. 2008, Dunbar et al. 2010, Lobera et al. 2017). However, positive relationships between reservoirs and stream health have been noted in certain US regions (Hill et al. 2017). Localized effects of WWTPs, possibly associated with chlorine residuals (Brungs 1973, Krasner et al. 2009) and/or nutrient loadings, have been shown to alter bacterial communities in streams (Wakelin et al. 2008) which consequently alter food webs and macroinvertebrate communities (Wallace et al. 1997, Webster et al. 1999). All spatial scales for major WWTPs tested as auxiliary stressors (presence of WWTP within 1, 5, 10, and 20 km upstream) had statistically significant negative relationships with stream health (+0.85, +0.48, +0.40, +0.38, respectively, Table S2). The declining trend in coefficient values shows the ability of streams to naturally mitigate the effects of WWTP discharges at scales less than those implied in previous studies and literature (Bartsch and Ingram 1966, Chapra 2008). Of note, coefficients for minor WWTPs were smaller in magnitude, and did not improve model performance based on BIC (Table S2). Similarly, the amount of wastewater per drainage area did not improve model performance (Table S2), implying that spatial proximity to large WWTPs might be more important to stream health than the percent of wastewater in streamflow. Both reservoirs and WWTPs are expected to have strong effects on downstream ecosystems (Mason 2002), but are often avoided and/or excluded as predictors in research by omitting data collected downstream of them (Cuffney et al. 2010, Bellucci et al. 2013). In addition, when reservoirs and WWTP point sources have been

included as predictors in studies, they have not always been significantly related to downstream ecosystem health (Strayer et al. 2003, Miller et al. 2018), possibly due to the spatial resolution of data collection and their distance to sampling locations.

4.2. Temporal variation of IC

Our approach contrasts with previous research focused on stream data collected over relatively short time periods (Richards et al. 1996, Cuffney et al. 2010, Belucci et al. 2013, Hill et al. 2017). Here, we use consistent land cover data (NWALT; Falcone 2015) to model BI samples collected over a 30-year study period. Our multi-stressor model implied that pre-1980 development affects streams 30% less than development before that date (Section 3.3). The reduced impact of newer (post-1980) IC on stream health may be attributable to state regulations focused on sediment control during construction, and the installations of BMPs to control runoff from developed land after the initial construction phase (e.g., detention ponds). Sediment control in NC began in the early 1970s with the Sediment Pollution Control Act of 1973 (NC DEQ 2017a) and steadily improved throughout the 1980s (Howells 1990). Post construction rules (e.g., requiring permanent BMPs) began in the late 1990s on a voluntary basis and became mandatory in most NC cities by 2007 (USEPA 2005, NC DEQ, 2017b). Therefore the change in 1980 is potentially attributable to construction sediment controls. However, it is also possible that the effects of urbanization on streams can occur over several decades (Leopold et al. 2005), such that the reduction in stream health for more recent construction is yet to be fully realized. One limitation encountered in this study was that most sites had watersheds dominated by older construction, with very little IC constructed after 2000. More sampling in watersheds with substantial amounts of post-2000 construction are needed to empirically assess the impact of more recent development practices on BI.

4.3. Multi-stressor model and management implications

A hierarchical framework allowed us to account for the intraclass correlation among samples collected at the same location (Gelman et al. 2014). Site random effects also offer a way to quantify variability across a large number of sites (1132 in this study), that may be related to factors for which we do not have reliable data, such as illicit dumping (Duda et al. 1982), leaking sewage infrastructure and septic systems (Paul and Meyer 2001, Walsh and Kunapo 2009), or excessive sedimentation during construction (Finkenbine et al. 2000, personal communication with Boyd Devane 2018). Alternatives to using a hierarchical framework could include modeling the mean or maximum BI value at each site as the response variable. These approaches, however, give equal weight to all sampling locations regardless of how many samples were collected, and would eliminate the ability to test for drivers of temporal variability in BI (e.g., age of construction and antecedent hydro-climatological conditions).

Hierarchical modeling also creates interpretable location-specific outputs (i.e., site random effects, soil random effects) that may help address management questions (Fig 4). Practically, water quality managers are faced with the need to improve the ecological health of streams without knowing what the best possible conditions are for degraded urban areas (Paul et al. 2009, Paul 2015). Model predictions based on deterministic factors (i.e., IC, canopy loss in the stream buffer, reservoirs; etc. Fig 4A) represent the expected conditions for a site based on watershed characteristics, and these predictions can identify sites that fall within a desired range for restoration activities. Site random effects then indicate whether those sites are in better or worse ecological condition than expected based on the deterministic factors (Fig 4B). In this way, a manager can prioritize sites based not just on their expected condition, but also how much they currently deviate from that condition. Only sites where sampling data was available have

site random effects in this study, but a pseudo site random effect could be obtained for new sites by sampling them and comparing that value to their deterministic prediction. If sampled once, the difference between the sampled BI and the predicted BI would represent the site random effect and the residual together. If sampled repeatedly, the average difference essentially reflects the site random effect. Restoration could proceed by first addressing degradation that comes from deterministic factors (e.g., increasing riparian buffers and disconnecting or otherwise mitigating IC). Then, more site-specific examinations can be performed for sites with highly positive site random effects to identify additional factors that might also be addressed. Sites with deterministic components that cannot be mitigated and that have highly negative site random effects (i.e., better stream health than expected) would likely be poor candidates for restoration, as there may not be a realistic expectation of improvement. In addition, these sites with large negative random effects can be studied to better understand the reasons for the better than expected stream health. Within and around municipal areas, the ability to identify sites near the state criterion (i.e., BI = 6.49 for NC) that can be improved either through mitigating fixed effects (e.g., IC, canopy loss) or identifying and mitigating factors associated with site random effects may be beneficial for city managers. In this study area, 27% of sites are predicted to be within 0.50 BI of the state criterion, and of those sites, 24% have site random effects above 0.25, 12% above 0.50, and 7% above 0.75.

Model coefficients can also be analyzed to help inform classification protocols for BI samples. Several predictors (e.g., sites smaller than 7.7 km^2, sites in Triassic soils) were created and tested specifically because they are special designations used by NC DEQ to modify bioclassification calculations. The multi-stressor model indicates that EPT sampling methods produce much lower BI scores (-0.90; Table S3) than full scale sampling, while Qual 4

techniques did not lead to statistically different scores. This suggests it may be reasonable to treat full scale and Qual 4 samples as equivalent when performing bioclassification of streams. As discussed earlier (see section 4.1.2), geologic soil groups had large effects on BI (Table 5, S1). At present, NC DEQ does not assign bioclassifications to sites in predominantly Triassic regions, but perhaps its soil random effect (+0.70; Table S1), or the Triassic soil auxiliary stressor coefficient (+0.42, Table S2) can be starting points for providing meaningful classifications in these areas. More generally, all Piedmont sites (not in the Triassic basins) are currently classified without consideration of their geologic condition. The multi-stressor model can predict the expected BI of reference conditions (i.e., minimally disturbed conditions; Stoddard et al. 2006) for each geologic region. To estimate the summer BI for a full scale sample, all anthropogenic values and site random effects can be set to zero leaving the intercept, previous precipitation, year (2018), and the soil random effect. For a mean value of precipitation, (0.30 m in the past 3 months), the predicted BI would be 3.71 + α_{soil} (Table S1). Sampling of reference sites in each geologic region could help confirm the random effects found in this study, and might lead to different bioclassification standards for each geologic region.

4.4. Summary and future directions

This study identified major natural and anthropogenic stressors related to the ecological health of wadeable streams in the NC Piedmont using hierarchical MLRs. The log percent of IC was able to explain 65% of the variation in BI as a baseline model, while a multi-stressor model explained 76% of the variation and provided insights into other factors controlling BI. Model results imply that watershed IC and canopy removal in stream buffers are the two largest anthropogenic sources of stream degradation, and recent construction (post-1980) impacts stream health 30% less than older development (pre-1980; Fig 2, 3). In addition, a slight but significant

overall improvement in stream health (-0.010/year) over the last 30 years was observed, independent of the temporal variability in other predictors. Also, upstream reservoirs and major WWTPs had large localized effects on stream health, while soil conditions, seasonal offsets, and antecedent precipitation were also significant predictors. Though our study focused on the NC Piedmont, this hierarchal MLR approach demonstrates an accessible and computationally tractable way to model multi-decadal regional datasets within a rigorous statistical framework that accounts for intra-class correlation (i.e., repetitive sampling of data at sites) and temporal changes in land covers. Because sites were located coarsely throughout the study area, spatial correlation among random effects was not expected or observed to be strong (e.g., Fig 4B), though the potential benefits of using more complex and computationally intensive approaches to modeling spatial correlation could be considered in future research (Hoef et al. 2014, Isaak et al. 2014). State and federal agencies have collected large amounts of macroinvertebrate and water quality data over the past few decades, and the ability to efficiently leverage these datasets to identify region-specific natural and anthropogenic stressors could help inform regional management decisions.

Site random effects explained 16% of the variance in BI in the multi-stressor model implying there are additional spatial drivers of stream health beyond the predictor variables available here. Downscaling this study to focus on a particular city or region where more detailed watershed data (e.g., sewage infrastructure, BMP installations) are available, might help to further refine our understanding of factors influencing BI variation. Alternatively, water quality variables (e.g., specific conductance, dissolved oxygen, total nitrogen), which have been related to ecological health in previous assessments (USEPA 2006, 2016, Cuffney et al. 2010), might also be useful predictors within the hierarchical MLR. This study did not consider water quality

sampling variables as they were unavailable for the majority of BI observations. More detailed and comprehensive predictor datasets could help shift statistical inference from proximate causes of BI degradation (e.g., IC, buffers, WWTPs) to more direct causes (e.g., temperature, sedimentation, toxicity).

Acknowledgements

We would like to thank Erik Fleek (NC DEQ), Michael Walters (NC DEQ), and Lory Willard (Raleigh Stormwater Services) for their support and cooperation with this project. Boyd Devane and Annette Lucas (NC DEQ) as well as Jay Sauber (NC DEQ retired) were helpful in discussing past regulations and practices in NC. In addition, Erickson Saunders provided technical assistance with processing geospatial datasets. Thanks also to Alexey Katin and Shiqi Fang for technical and graphical assistance.

Funding

This work was supported by the Environmental Protection Agency (Grant #R836165) and North Carolina State University.

References

Allan, J. D. 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. *Annu.Rev. Ecol. Evol. Syst.*, *35*: 257-284.

Bartsch, A. F., and W.M. Ingram. 1966. Biological analysis of water pollution in North America: With 1 table in the text. *Internationale Vereinigung für theoretische und angewandte Limnologie:Verhandlungen*, *16*(2), 786-800.

Bates, D., Mächler, M., Bolker, B., and S. Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beasley, G., and P. Kneale. 2002. Reviewing the impact of metals and PAHs on macroinvertebrates in urban watercourses. *Progress in Physical Geography*, *26*(2): 236-270

Bellucci, C. J., Becker, M. E., Beauchene, M., and L. Dunbar. 2013. Classifying the health of Connecticut streams using benthic macroinvertebrates with implications for water management. *Environmental management*, *51*(6): 1274-1283.

Bernhardt, E. S., and M.A. Palmer. 2007. Restoring streams in an urbanizing world. *Freshwater Biology*, *52*(4): 738-751.

Beschta, R. L., and R. L. Taylor. 1988. Stream Temperature Increases and Land Use in a Forested Oregon Watershed. *JAWRA Journal of the American Water Resources Association*, *24*(1): 19-25.

Boehme, E. A., Zipper, C. E., Schoenholtz, S. H., Soucek, D. J., and A. J. Timpano. 2016. Temporal dynamics of benthic macroinvertebrate communities and their response to elevated specific conductance in Appalachian coalfield headwater streams. *Ecological Indicators*, *64*: 171-180.

Boggs, J. L., and G. Sun. 2011. Urbanization alters watershed hydrology in the Piedmont of

North Carolina. *Ecohydrology*, *4*(2): 256-264.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and J. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, *24*(3): 127-135.

Booth, D. B., and C. R. Jackson. 1997. Urbanization of aquatic systems: degradation thresholds, stormwater detection, and the limits of mitigation. *JAWRA Journal of the American Water Resources Association*, *33*(5): 1077-1090.

Booth, D. B. 2005. Challenges and prospects for restoring urban streams: a perspective from the Pacific Northwest of North America. *Journal of the North American Benthological Society*, *24*(3): 724-737.

Boulton, A. J. 2003. Parallels and contrasts in the effects of drought on stream macroinvertebrate assemblages. *Freshwater Biology*, *48*(7): 1173-1185.

Breiman, L. 2001. Random forests. *Machine learning*, *45*(1), 5-32.

Bressler, D. W., Paul, M. J., Purcell, A. H., Barbour, M. T., Rankin, E. T., and V.H. Resh. 2009. Assessment tools for urban catchments: developing stressor gradients. *JAWRA Journal of the American Water Resources Association*, *45*(2): 291-305.

Brungs, W. A. 1973. Effects of residual chlorine on aquatic life. *Journal (Water Pollution Control Federation)*, 2180-2193.

Buchwalter, D., Camp, A., & Hawkins, C. (2015). *Experimental Approaches to Understanding Temperature Responses of Select North Carolina Macroinvertebrates*. Water Resources Research Institute of the University of North Carolina.

Burt, T. P., Ford Miniat, C., Laseter, S. H., and W. T. Swank. 2018. Changing patterns of daily precipitation totals at the Coweeta Hydrologic Laboratory, North Carolina,

USA. *International Journal of Climatology*, *38*(1): 94-104.

Cañedo-Argüelles, M., Kefford, B. J., Piscart, C., Prat, N., Schäfer, R. B., and C. J. Schulz. 2013. Salinisation of rivers: an urgent ecological issue. *Environmental pollution*, *173*: 157-167.

Chapra, S. C. (2008). *Surface water-quality modeling*. Waveland press.

Chatfield, C. (2006). *Model uncertainty*. *Encyclopedia of Environmetrics*, John Wiley, Hoboken, N.J.

Clark, J. S., & Gelfand, A. E. (2006). A future for models and data in environmental science. *Trends in Ecology & evolution*, *21*(7), 375-380.

Clarke, R. T., Wright, J. F., and M. T. Furse. 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling*, *160*(3): 219-233.

Clements, W. H., and C. Kotalik. 2016. Effects of major ions on natural benthic communities: an experimental assessment of the US Environmental Protection Agency aquatic life benchmark for conductivity. *Freshwater Science*, *35*(1): 126-138.

Cohen, J., Cohen, P., West, S. G., and L. S. Aiken. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Coulston, J. W., Moisen, G. G., Wilson, B. T., Finco, M. V., Cohen, W. B., and C. K. Brewer. 2012. Modeling percent tree canopy cover: a pilot study. *Photogrammetric Engineering & Remote Sensing 78 (7): 715–727*.

Cordone, A. J., and D. W. Kelley. 1961. *The influences of inorganic sediment on the aquatic life of streams*. California Department of Fish and Game.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical

modeling. *Ecological Applications*, *19*(3): 553-570.

Cuffney, T. F., Brightbill, R. A., May, J. T., and I. R. Waite. 2010. Responses of benthic macroinvertebrates to environmental changes associated with urbanization in nine metropolitan areas. *Ecological Applications*, *20*(5): 1384-1401.

Cuffney, T. F., Kashuba, R., Qian, S. S., Alameddine, I., Cha, Y. K., Lee, B., ... and G. McMahon. 2011a. Multilevel regression models describing regional patterns of invertebrate and algal responses to urbanization across the USA. *Journal of the North American Benthological Society*, *30*(3): 797-819.

Cuffney, T. F., Qian, S. S., Brightbill, R. A., May, J. T., and I. R. Waite. 2011b. Response to King and Baker: limitations on threshold detection and characterization of community thresholds. *Ecological Applications*, *21*(7): 2840-2845.

Deletic, A. 1998. The first flush load of urban surface runoff. *Water research*, *32*(8): 2462-2470.

Dennehy, K. F. 1998. *Water quality in the South Platte River basin, Colorado, Nebraska, and Wyoming, 1992-95* (Vol. 1167). US Department of the Interior, US Geological Survey.

Duda, A. M., Lenat, D. R., and D. L. Penrose. 1982. Water quality in urban streams: what we can expect. *Journal (Water Pollution Control Federation)*, 1139-1147.

Dunbar, M. J., Warren, M., Extence, C., Baker, L., Cadman, D., Mould, D. J., ... and R. Chadd. 2010. Interaction between macroinvertebrates, discharge and physical habitat in upland rivers. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *20*(S1): S31-S44.

Elbrecht, V., Beermann, A. J., Goessler, G., Neumann, J., Tollrian, R., Wagner, R., ... and F. Leese. 2016. Multiple-stressor effects on stream invertebrates: a mesocosm experiment manipulating nutrients, fine sediment and flow velocity. *Freshwater biology*, *61*(4): 362-375.

Elsner, J. B., and C. P. Schmertmann. 1994. Assessing forecast skill through cross validation. *Weather and Forecasting*, *9*(4): 619-624.

Evans, J. S., and S.A. Cushman. 2009. Gradient modeling of conifer species using random forests. *Landscape ecology*, *24*(5): 673-683.

Evans, J. S., Murphy, M. A., Holden, Z. A., and S. A. Cushman. (2011). Modeling species distribution and change using random forest. In *Predictive species and habitat modeling in landscape ecology*, pp. 139-159. Springer, New York, NY.

Falcone, J. A. 2015. US conterminous wall-to-wall anthropogenic land use trends (NWALT), 1974–2012 (No. 948). US Geological Survey.

Faraway, J. J. (2006). Extending the linear model with R: Generalized linear. *Mixed effects and nonparametric regression models*, Chapman & Hall/CRC, Boca Raton, FL.

Faraway, J. J. (2014). *Linear models with R*. Chapman & Hall/CRC press. Boca Raton, FL

Finkenbine, J. K., Atwater, J. W., and J. S. Mavinic. 2000. Stream health after urbanization. *JAWRA Journal of the American Water Resources Association*, *36*(5): 1149-1160.

Fisher, S. G., Gray, L. J., Grimm, N. B., and D. E. Busch. 1982. Temporal succession in a desert stream ecosystem following flash flooding. *Ecological monographs*, *52*(1): 93-110.

Fitzpatrick, F. A., Scudder, B. C., Lenz, B. N., and D. J. Sullivan. 2001. Effects of Multi-Scale Environmental Characteristics on Agricultural Stream Biota in Eastern Wisconsin. *JAWRA Journal of the American Water Resources Association*, *37*(6): 1489-1507.

Gelman, A., and J. Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Carlin, J. B., Stern, H. S., and D. B. Rubin. 2014. *Bayesian data analysis* (Vol. 2).

Chapman & Hall/CRC. Boca Raton, FL, USA.

Graymore, M., Stagnitti, F., and G. Allinson. 2001. Impacts of atrazine in aquatic ecosystems. *Environment international*, *26*(7): 483-495.

Groffman, P. M., Bain, D. J., Band, L. E., Belt, K. T., Brush, G. S., Grove, J. M., Pouyat, R.V., Yesilonis, I.C, and W.C. Zipperer. 2003. Down by the riverside: urban riparian ecology. *Frontiers in Ecology and the Environment*, *1*(6): 315-321.

Grube, A., D. Donaldson, T. Kiely, and L. Wu. 2011. Pesticides Industry Sales and Usage: 2006 and 2007. Market Estimates. Technical Report, U.S. Environmental Protection Agency, Washington, D.C. Guttman, N. B. 1998. Comparing the Palmer drought index and the standardized precipitation index. *JAWRA Journal of the American Water Resources Association*, *34*(1): 113-121.

Hardison, E. C., O'Driscoll, M. A., DeLoatch, J. P., Howard, R. J., and M. M. Brinson. 2009. Urban Land Use, Channel Incision, and Water Table Decline Along Coastal Plain Streams, North Carolina. *JAWRA Journal of the American Water Resources Association*, *45*(4): 1032-1046.

Hawkins, C. P., Norris, R. H., Hogue, J. N., and J. W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications*, *10*(5): 1456-1477.

Hill, R. A., Fox, E. W., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., and M. H. Weber. 2017. Predictive mapping of the biotic condition of conterminous US rivers and streams. *Ecological Applications*, *27*(8): 2397-2415.

Hilsenhoff, W. L. 1988. Rapid field assessment of organic pollution with a family-level biotic index. *Journal of the North American Benthological Society*, *7*(1): 65-68.

Hirst, H., Jüttner, I., and S. J. Ormerod. 2002. Comparing the responses of diatoms and macro-invertebrates to metals in upland streams of Wales and Cornwall. *Freshwater Biology*, *47*(9): 1752-1765.

Hoef, J. V., Peterson, E., Clifford, D., and R. Shah. 2014. SSN: An R package for spatial statistical modeling on stream networks. *Journal of Statistical Software*, *56*(3): 1-45.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., ... and J. Wickham. 2007. Completion of the 2001 national land cover database for the conterminous United States. *Photogrammetric engineering and remote sensing*, *73*(4): 337.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., ... and K. Megown. 2015. Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*, *81*(5): 345-354.

Hopkins, K. G., Morse, N. B., Bain, D. J., Bettez, N. D., Grimm, N. B., Morse, J. L., ... and A. K. Suchy. 2015. Assessment of regional variation in streamflow responses to urbanization and the persistence of physiography. *Environmental science & technology*, *49*(5): 2724-2732.

Howells, D. H. 1990. Quest for clean streams in North Carolina: an historical account of stream pollution control in North Carolina. *Water Resources Research Institute of the University of North Carolina (USA)*.

Huber, M., Welker, A., and B. Helmreich. 2016. Critical review of heavy metal pollution of traffic area runoff: Occurrence, influencing factors, and partitioning. *Science of the Total Environment*, *541*: 895- 919.

Isaak, D. J., Peterson, E. E., Ver Hoef, J. M., Wenger, S. J., Falke, J. A., Torgersen, C. E., ... and A. S. Ruesch. 2014. Applications of spatial statistical network models to stream data. *Wiley*

Interdisciplinary Reviews: Water, *1*(3): 277-294.

Kashuba, R., Cha, Y., Alameddine, I., Lee, B., and T. F. Cuffney. 2010. *Multilevel hierarchical modeling of benthic macroinvertebrate responses to urbanization in nine metropolitan regions across the conterminous United States* (No. 2009-5243). US Geological Survey.

Kefford, B. J., Marchant, R., Schäfer, R. B., Metzeling, L., Dunlop, J. E., Choy, S. C., and P. Goonan. 2011. The definition of species richness used by species sensitivity distributions approximates observed effects of salinity on stream macroinvertebrates. *Environmental Pollution*, *159*(1): 302-310.

King, R. S., and M. E. Baker. 2010. Considerations for analyzing ecological community thresholds in response to anthropogenic environmental gradients. *Journal of the North American Benthological Society*, *29*(3): 998-1008.

Klein, R. D. 1979. Urbanization and stream quality impairment 1. *JAWRA Journal of the American Water Resources Association*, *15*(4): 948-963.

Krasner, S. W., Westerhoff, P., Chen, B., Rittmann, B. E., and G. Amy.2009. Occurrence of disinfection byproducts in United States wastewater treatment plant effluents. *Environmental Science & Technology*, *43*(21): 8320-8325.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models (Vol. 4). McGraw-Hill.

Lake, P. S. 2003. Ecological effects of perturbation by drought in flowing waters. *Freshwater biology*, *48*(7): 1161-1172.

Laseter, S. H., Ford, C. R., Vose, J. M., and L. W. Swift. 2012. Long-term temperature and precipitation trends at the Coweeta Hydrologic Laboratory, Otto, North Carolina, USA. *Hydrology Research*, *43*(6): 890-901.

LeBlanc, R. T., Brown, R. D., and J. E. FitzGibbon. 1997. Modeling the effects of land use change on the water temperature in unregulated urban streams. *Journal of Environmental Management*, *49*(4): 445- 469.

Legret, M., and C. Pagotto. 1999. Evaluation of pollutant loadings in the runoff waters from a major rural highway. *Science of the Total Environment*, *235*(1): 143-150.

Leopold, L. B. 1968. Hydrology for urban land planning: A guidebook on the hydrologic effects of urban land use. Reston, VA: US Geological Survey. Circular #554.

Leopold, L. B., Huppman, R., and A. Miller. 2005. Geomorphic effects of urbanization in forty-one years of observation. *Proceedings of the American Philosophical Society*, 349-371.

Lobera, G., Muñoz, I., López-Tarazón, J. A., Vericat, D., and R. J. Batalla. 2017. Effects of flow regulation on river bed dynamics and invertebrate communities in a Mediterranean river. *Hydrobiologia*, *784*(1): 283-304.

Makepeace, D. K., Smith, D. W., and S. J. Stanley. 1995. Urban stormwater quality: summary of contaminant data. *Critical Reviews in Environmental Science and Technology*, *25*(2): 93-139.

Matthews, J., Reeze, B., Feld, C. K., and A. J. Hendriks. 2010. Lessons from practice: assessing early progress and success in river rehabilitation. *Hydrobiologia*, 655(1): 1-14.

McKee, T. B., Doesken, N. J., and J. Kleist. 1993 (January). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology* 17(22): 179-183. Boston, MA: American Meteorological Society.

Meador, M. R., and R. M. Goldstein. 2003. Assessing water quality at large geographic scales: relations among land use, water physicochemistry, riparian condition, and fish community structure. *Environmental Management*, *31*(4): 0504-0517.

Meyer, J. L., Paul, M. J., and W. K. Taulbee. 2005. Stream ecosystem function in urbanizing landscapes. *Journal of the North American Benthological Society*, *24*(3): 602-612.

Miller, J., Esselman, P. C., Alameddine, I., Blackhart, K., and D. R. Obenour. 2018. Hierarchical modeling assessment of the influence of watershed stressors on fish and invertebrate species in Gulf of Mexico estuaries. *Ecological Indicators*, *90*: 142-153.

Miltner, R. J., White, D., and C. Yoder. 2004. The biotic integrity of streams in urban and suburbanizing landscapes. *Landscape and urban planning*, *69*(1): 87-100.

Monk, W. A., Wood, P. J., Hannah, D. M., and D. A. Wilson. 2008. Macroinvertebrate community response to inter-annual and regional river flow regime dynamics. *River Research and Applications*, *24*(7): 988-1001.

Moore, R. B., and T. G. Dewald. 2016. The Road to NHDPlus—Advancements in Digital Stream Networks and Associated Catchments. *JAWRA Journal of the American Water Resources Association*, *52*(4): 890-900.

Mulholland, P. J., and D. R. Lenat. 1992. Streams of the southeastern Piedmont, Atlantic drainage. *Biodiversity of the Southeastern United States—Aquatic Communities*, 193-232.

Munn, M. D., and M. A. Brusven. 1991. Benthic macroinvertebrate communities in nonregulated and regulated waters of the Clearwater River, Idaho, USA. *Regulated Rivers: Research & Management*, *6*(1): 1-11.

Nelson, K. C., M. A. Palmer. 2007. Stream temperature surges under urbanization and climate change: data, models, and responses. *JAWRA Journal of the American Water Resources Association*, *43*(2): 440-452.

NC DEQ. 2016. Standard Operating Procedures for the Collection and Analysis of Benthic Macroinvertebrates. Division of Water Resources. Raleigh, North Carolina. February 2016.

NC DEQ. 2017a. Erosion and Sediment Control Planning and Design Manual. https://deq.nc.gov/about/divisions/energy-mineral-land-resources/energy-mineral-land-permit-guidance/erosion-sediment-control-planning-design-manual. Accessed Dec. 29th, 2017.

NC DEQ. 2017b. NPDES MS4 Permitting. https://deq.nc.gov/about/divisions/energy-mineral-land-resources/energy-mineral-land-permits/stormwater-permits/npdes-ms4. Accessed June 27th, 2017.

NC Geological Survey. 2009. NC DEQ, Division of Land Resources. Accessed from NC One Map January 2017. http://data.nconemap.gov/downloads/vector/geol.zip

NRCS. 2012. NC Soil Survey Geographic (SSURGO) database and Map Service. http://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm. Accessed January 20th, 2017.

Olden, J. D., Lawler, J. J., and N. L. Poff. 2008. Machine learning methods without tears: a primer for ecologists. *The Quarterly review of biology*, *83*(2): 171-193.

Palmer, M. A., Hondula, K. L., and B. J. Koch. 2014. Ecological restoration of streams and rivers: shifting strategies and shifting goals. *Annual Review of Ecology, Evolution, and Systematics*, *45*: 247- 269.

Parr, T. B., Smucker, N. J., Bentsen, C. N., and M. W. Neale. 2016. Potential roles of past, present, and future urbanization characteristics in producing varied stream responses. *Freshwater Science*, *35*(1): 436-443.

Paul, M. J., and J. L. Meyer. 2001. Streams in the urban landscape. *Annual review of Ecology and Systematics*, *32*(1): 333-365.

Paul, M. J., Bressler, D. W., Purcell, A. H., Barbour, M. T., Rankin, E. T., and V. H. Resh. 2009. Assessment tools for urban catchments: defining observable biological potential. *JAWRA*

Journal of the American Water Resources Association, *45*(2): 320-330.

Paul, M. 2015. *A Review of Biological Condition Ratings in Urban Settings in North Carolina for the North Carolina Water Resources Research Institute*. Water Resources Research Institute, North Carolina State University.

Phipps, G. L., Mattson, V. R., and G. T. Ankley. 1995. Relative sensitivity of three freshwater benthic macroinvertebrates to ten contaminants. *Archives of Environmental Contamination and Toxicology*, *28*(3): 281-286.

Pierson, S. M., Rosenbaum, B. J., McKay, L. D., and T. G. Dewald. 2008. Strahler stream order and Strahler calculator values in NHDPLUS. *SOSC technical paper, September*, *30*, 2008.

Pluhowski, E. J. 1972. Clear-cutting and its effect on the water temperature of a small stream in northern Virginia. *US Geological Survey Professional Papers C*, *800*, C257-C262.

Poole, G. C., and C. H. Berman. 2001. An ecological perspective on in-stream temperature: natural heat dynamics and mechanisms of human-caused thermal degradation. *Environmental management*, *27*(6): 787-802.

Qian, S. S., Reckhow, K. H., Zhai, J., and G. McMahon. 2005. Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach. *Water Resources Research*, *41*(7).

Qian, S. S., and Z. Shen. 2007. Ecological applications of multilevel analysis of variance. *Ecology*, *88*(10): 2489-2495.

Qian, S. S., Cuffney, T. F., Alameddine, I., McMahon, G., and K. H. Reckhow. 2010. On the application of multilevel modeling in environmental and ecological studies. *Ecology*, *91*(2): 355-361.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raftery, A. E. 1995. Bayesian model selection in social research. *Sociological methodology*, 111-163.

Richards, C., Johnson, L. B., and G. E. Host. 1996. Landscape-scale influences on stream habitats and biota. *Canadian Journal of Fisheries and aquatic sciences*, *53*(S1): 295-311.

Rose, S., and N. E. Peters. 2001. Effects of urbanization on streamflow in the Atlanta area (Georgia, USA): a comparative hydrological approach. *Hydrological Processes*, *15*(8): 1441-1457.

Roy, A. H., Freeman, M. C., Freeman, B. J., Wenger, S. J., Ensign, W. E., and J. L. Meyer. 2005. Investigating hydrologic alteration as a mechanism of fish assemblage shifts in urbanizing streams. *Journal of the North American Benthological Society*, *24*(3): 656-678.

Samuels, M. L. 1993. Simpson's paradox and related phenomena. *Journal of the American Statistical Association*, *88*(421): 81-88.

Seaburn, G. E. 1969. *Effects of urban development on direct runoff to East Meadow Brook, Nassau County, Long Island, New York*. US Government Printing Office.

Schueler, T. R. 1994. The importance of imperviousness. *Watershed protection techniques*, *1*(3): 100-111.

Schueler, T. R., Fraley-McNeal, L., and K. Cappiella. 2009. Is impervious cover still important? Review of recent research. *Journal of Hydrologic Engineering*, *14*(4): 309-315.

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

Selvakumar, A., O'Connor, T. P., and S. D. Struck. 2009. Role of stream restoration on improving benthic macroinvertebrates and in-stream water quality in an urban watershed: case study. *Journal of Environmental Engineering*, *136*(1): 127-139.

Sponseller, R. A., Grimm, N. B., Boulton, A. J., and J. L. Sabo. 2010. Responses of

macroinvertebrate communities to long-term flow variability in a Sonoran Desert stream. *Global Change Biology*, *16*(10): 2891-2900.

Steedman, R. J. 1988. Modification and assessment of an index of biotic integrity to quantify stream quality in southern Ontario. *Canadian Journal of Fisheries and Aquatic Sciences*, *45*(3): 492-501.

Stoddard, J. L., Larsen, D. P., Hawkins, C. P., Johnson, R. K., and R. H. Norris. 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. *Ecological Applications*, *16*(4): 1267-1276.

Strayer, D. L., Beighley, R. E., Thompson, L. C., Brooks, S., Nilsson, C., Pinay, G., and R. J. Naiman. 2003. Effects of land cover on stream ecosystems: roles of empirical models and scaling issues. *Ecosystems*, *6*(5): 407-423.

Strickling, H. L., and D. R. Obenour. 2018. Leveraging Spatial and Temporal Variability to Probabilistically Characterize Nutrient Sources and Export Rates in a Developing Watershed. *Water Resources Research*.

Sutherland, A. B., Meyer, J. L., and E. P. Gardiner. 2002. Effects of land cover on sediment regime and fish assemblage structure in four southern Appalachian streams. *Freshwater Biology*, *47*(9): 1791-1805.

Sweeney, B. W., and J. D. Newbold. 2014. Streamside forest buffer width needed to protect stream water quality, habitat, and organisms: a literature review. *JAWRA Journal of the American Water Resources Association*, *50*(3): 560-584.

Tremblay, A., and J. Ransijn. 2015. LMERConvenienceFunctions: Model selection and post-hoc analysis for (G) LMER models. R Package Version 2.

USEPA (United States Environmental Protection Agency). 2005. Stormwater Phase II Final

Rule. Office of Research and Development. Office of Water. EPA 833-F_00_001.

USEPA (United States Environmental Protection Agency). 2006. Wadeable Stream Assessment. Office of Research and Development. Office of Water. EPA 841-B-06_002.

USEPA (United States Environmental Protection Agency). 2009. National Water Quality Inventory: Report to Congress, 2004 Reporting Cycle. Report # EPA-841-R-08-001. Office of Water. District of Columbia, U.S.A. 52 pp.

USEPA (United States Environmental Protection Agency). 2016. Office of Water and Office of Research and Development. National Rivers and Streams Assessment 2008-2009: A Collaborative Survey (EPA/841/R-16/007). Washington, DC.

Utz, R. M., Hopkins, K. G., Beesley, L., Booth, D. B., Hawley, R. J., Baker, M. E., ... K. L. Jones. 2016. Ecological resistance in urban streams: the role of natural and legacy attributes. *Freshwater Science*, *35*(1): 380-397.

Vietz, G. J., Sammonds, M. J., Walsh, C. J., Fletcher, T. D., Rutherfurd, I. D., and M. J. Stewardson. 2014. Ecologically relevant geomorphic attributes of streams are impaired by even low levels of watershed effective imperviousness. *Geomorphology*, *206*: 67-78.

Wahl, C. M., Neils, A., and D. Hooper. 2013. Impacts of land use at the catchment scale constrain the habitat benefits of stream riparian buffers. *Freshwater Biology*, *58*(11): 2310-2324.

Wakelin, S. A., Colloff, M. J., and R. S. Kookana. 2008. Effect of wastewater treatment plant effluent on microbial function and community structure in the sediment of a freshwater stream with variable seasonal flow. *Applied and environmental microbiology*, *74*(9): 2659-2668.

Wallace, J. B., Eggert, S. L., Meyer, J. L., and J. R. Webster. 1997. Multiple trophic levels of a

forest stream linked to terrestrial litter inputs. *Science*, *277*(5322): 102-104.

Walsh, C. J., Roy, A. H., Feminella, J. W., Cottingham, P. D., Groffman, P. M., and R. P. Morgan II. 2005. The urban stream syndrome: current knowledge and the search for a cure. *Journal of the North American Benthological Society*, *24*(3): 706-723.

Walsh, C. J., and J. Kunapo. 2009. The importance of upland flow paths in determining urban effects on stream ecosystems. *Journal of the North American Benthological Society*, *28*(4): 977-990.

Webster, J. R., Benfield, E. F., Ehrman, T. P., Schaeffer, M. A., Tank, J. L., Hutchens, J. J., and D. J. D'angelo. 1999. What happens to allochthonous material that falls into streams? A synthesis of new and published information from Coweeta. *Freshwater Biology*, *41*(4): 687-705.

Wenger, S. J., Roy, A. H., Jackson, C. R., Bernhardt, E. S., Carter, T. L., Filoso, S., ... and J. L. Meyer. 2009. Twenty-six key research questions in urban stream ecology: an assessment of the state of the science. *Journal of the North American Benthological Society*, *28*(4): 1080-1098.

Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, *84*(6): 1382-1394.

Wright, J. F. 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. *Austral Ecology*, *20*(1): 181-197.

Yun, J., and S. S. Qian. 2015. A Hierarchical Model for Estimating Long-Term Trend of Atrazine Concentration in the Surface Water of the Contiguous US. *JAWRA Journal of the American Water Resources Association*, *51*(4): 1128-1137.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and G. M. Smith. 2009. Mixed effects

models and extensions in ecology with R. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors. *Spring Science and Business Media*. New York, NY.

Supplementary Materials

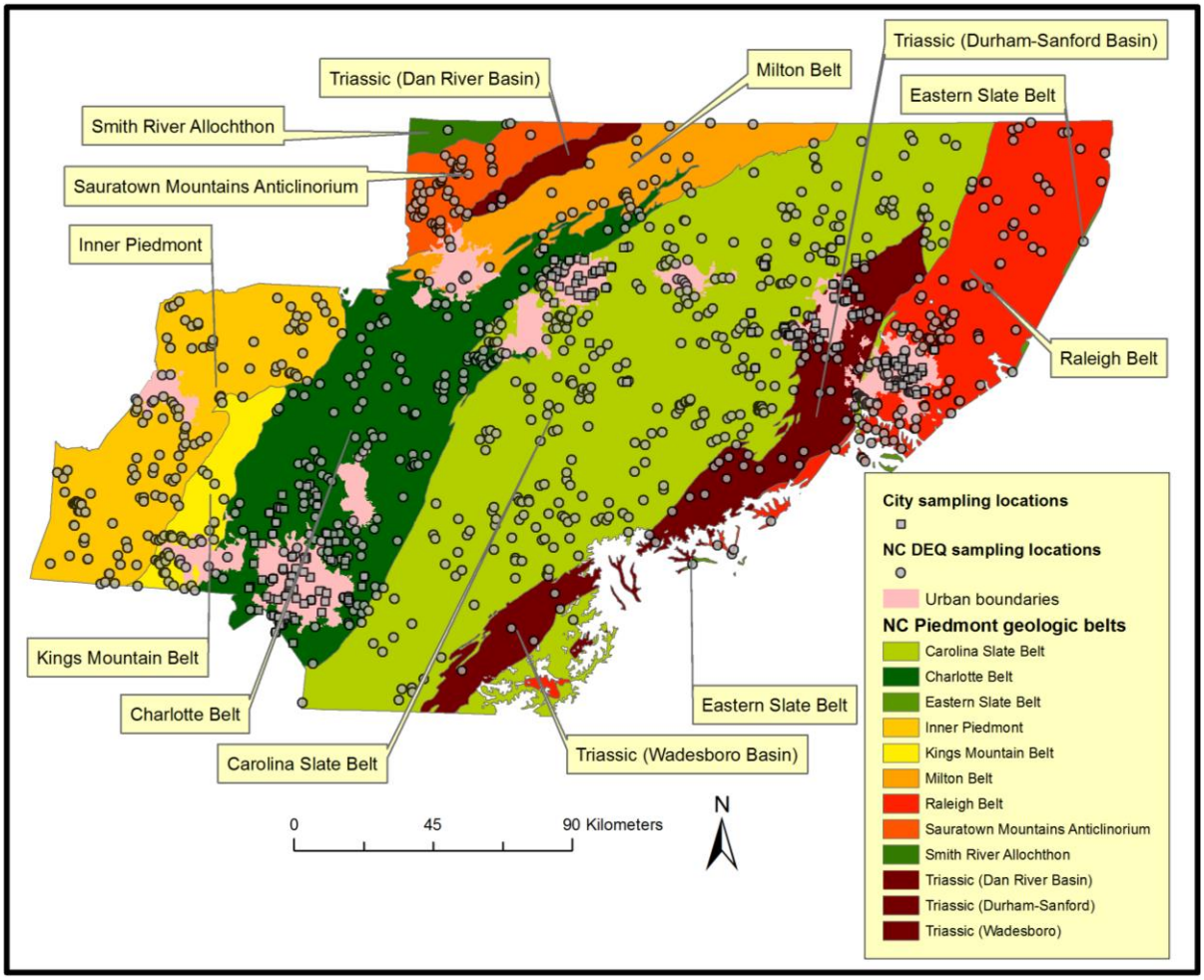


Figure S1: NC Piedmont with major geologic regions and BI sampling locations.

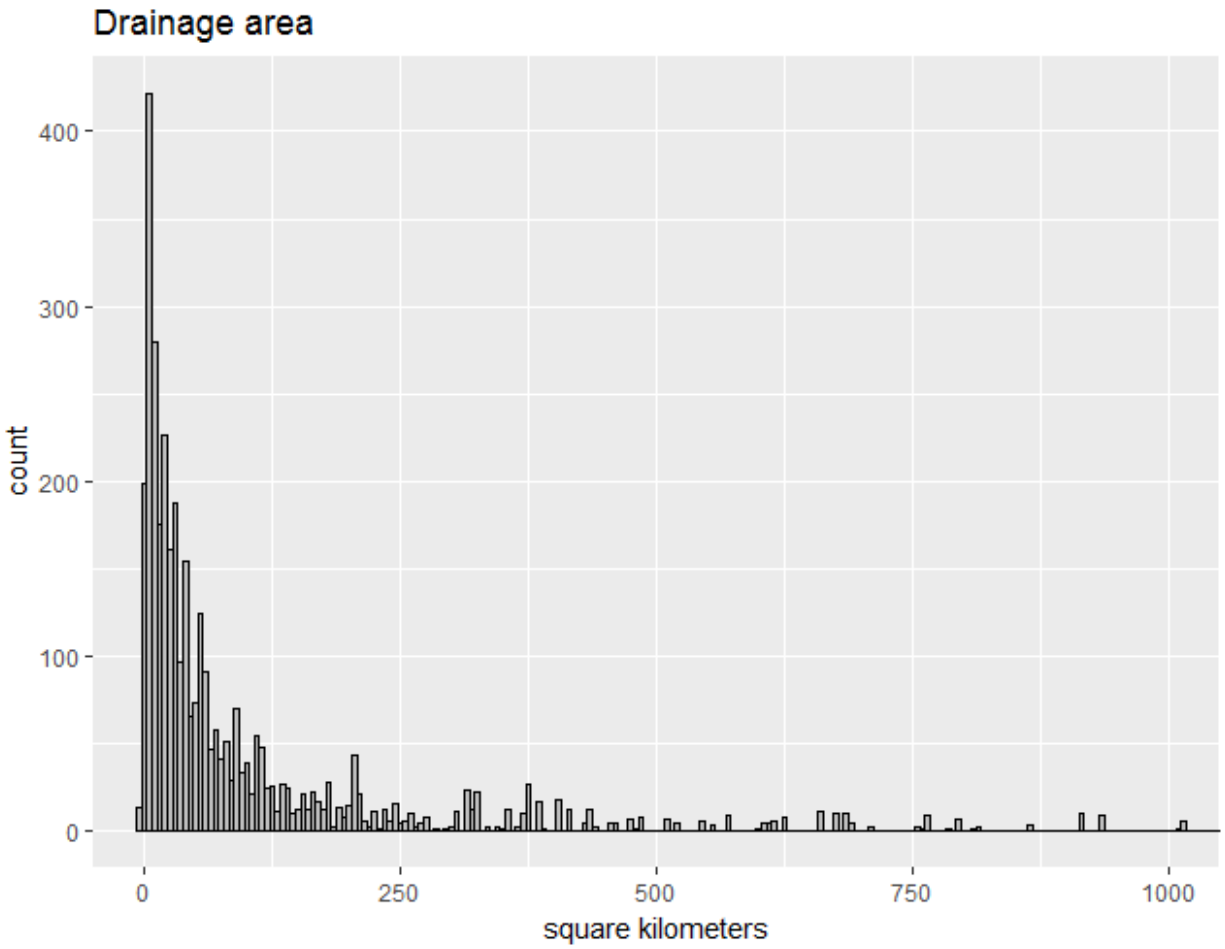


Figure S2: Histogram of drainage areas for macroinvertebrate sampling locations. 1,132 sampling locations with drainage areas less than 500 km^2 were used for model construction. 71 locations above 500 km^2 (296 samples) were removed.

Table S1: Summary statistics for 10 geologic regions including: number of samples, mean elevation, impervious cover (IC), and biotic index (BI). Soil random effects from the multi-stressor model are also presented. The NC Piedmont is shown by region below where circles represent NC DEQ sampling locations and squares are city sampling locations.

Geologic Soil	# of samples	Mean			Random effect multi-stressor model
		Elevation (m)	Impervious (percent)	BI	
Carolina Slate Belt	1069	44.8	.03	5.78	0.28
Charlotte Belt	826	58.2	.11	6.62	0.26
Triassic Basins	272	25.4	.08	6.79	0.70
Eastern Slate Belt	16	20.7	.02	5.05	-0.01
Inner Piedmont	299	75.3	.02	4.98	-0.39
Kings Mountain Belt	102	64.1	.04	5.80	0.07
Milton Belt	76	60.3	.05	5.78	0.00
Raleigh Belt	343	23.2	.06	6.07	0.23
Sauratown Mountains Anticlinorium	112	83.3	.01	4.31	-0.57
Smith River Allochthon	12	80.4	.004	3.95	-0.59

Table S2: All candidate predictor variables and their corresponding auxiliary stressor model coefficients. Starred coefficients (*) are not significant at the 95% confidence level. The best predictor variable from each variable class was included as candidate for the multi-stressor model, provided it lowered BIC by at least 6. "**X**" denotes predictors that met the BIC criteria and remained in the multi-stressor model (Table 5; Figure 2), while "**x**" denotes predictors that met this criteria but were then parsed out of the multi-stressor model due to not being significant at the 95% level.

Variable Predictor Class	Candidate predictor variable	Linear coeff.	Quad. coeff.	Δ BIC from baseline model	Candidates for multi-stressor model
Age of Construction	Impervious cover (pre 1974)	0.24		-28.8	
	Impervious cover (pre 1980)	0.27		-33.3	**X**
	Impervious cover (pre 1990)	0.31		-30.3	
	Impervious cover (pre 2000)	0.40		-12.3	
Agriculture	% Agriculture watershed	0.008		-18.8	**x**
	% Agriculture stream buffer	0.012		-17.1	
	% Agriculture near buffer	0.001*		7.5	
	% Agriculture watershed (log)	0.09		-8.2	
	% Agriculture stream buffer (log)	0.10		-11.3	
	% Agriculture near buffer (log)	0.01*		7.6	
Antecedent Hydro-Climatological Conditions	Precipitation (previous month)	-0.89		-13.7	
	Precipitation (previous 3 months)	-0.39		-8.0	
	Precipitation (previous 6 months)	-0.12*		4.8	
	Precipitation (previous 12 months)	-0.12		1.1	
	Precipitation (previous 24 months)	-0.04*		7.3	
	Drought index (previous month)	-0.007*		7.4	
	Drought index (previous 3 months)	-0.04		-15.5	
	Drought index (previous 6 months)	-0.03		-12.4	
	Drought index (previous 12 months)	-0.02		-0.3	
	Drought index (previous 24 months)	-0.01		6.4	
	Precipitation (previous month; quad.)	-2.9	8.0	-13.5	
	Precipitation (previous 3 months; quad.)	-3.4	4.6	-33.9	**X**
	Precipitation (previous 6 months; quad.)	-1.8	1.3	-5.1	
	Precipitation (previous 12 months; quad.)	-1.9	0.74	-14.2	
	Precipitation (previous 24 months; quad.)	-1.3	0.28	10.5	
	Drought index (previous month; quad.)	-0.01*	-0.01	13.9	
	Drought index (previous 3 months; quad.)	-0.04	0.01	-16.3	
	Drought index (previous 6 months; quad.)	-0.05	0.02	-27.0	
	Drought index (previous 12 months; quad.)	-0.02	0.01	-1.6	
	Drought index (previous 24 months; quad.)	-0.004*	0.01	11.5	
Canopy Loss	% Canopy loss watershed	0.01		-25.2	
	% Canopy loss stream buffer	0.02		-55.4	
	% Canopy loss near buffer	0.00		-17.8	
	% Canopy loss watershed (log)	0.35		-66.2	
	% Canopy loss stream buffer (log)	0.42		-104.5	**X**
	% Canopy loss near buffer (log)	0.10		-30.3	
Collecting Agency	Random effect for 6 agencies	-		8.8	
	Categorical variable for 6 agencies	-		28.8	

Table S2 (continued)

Drainage area	Drainage	0.0006		3.0	
	Less than 7.7 km^2	-0.22		-6.1	**x**
Elevation	Elevation	-0.0080		-25.9	**x**
Recent Construction	% Impervious watershed (in last 2 yrs, linear)	-0.18*		7.7	
	% Impervious 30m buffer (in last 2 yrs, linear)	-0.20*		7.6	
	% Impervious near buffer (in last 2 yrs, linear)	0.17*		4.9	
	% Impervious watershed (in last 5 yrs, linear)	-0.12*		7.4	
	% Impervious 30m buffer(in last 5 yrs, linear)	-0.12*		7.4	
	% Impervious near buffer(in last 5 yrs, linear)	0.15*		4.6	
	% Impervious watershed (in last 2 yrs; log)	-0.15*		7.7	
	% Impervious 30m buffer (in last 2 yrs; log)	-0.16*		7.6	
	% Impervious near buffer (in last 2 yrs; log)	0.12*		4.4	
	% Impervious watershed (in last 5 yrs; log)	-0.08*		7.5	
	% Impervious 30m buffer(in last 5 yrs; log)	-0.09*		7.5	
	% Impervious near buffer(in last 5 yrs; log)	0.11		3.7	
Season	Categorical variable for 4 seasons (w,sp,f)			-32.5	**X**
Soil conditions	Categorical variable for 10 geologic regions	-		-122.0	
	Random effect for 10 geologic regions	N(0,0.45)		-152.6	**X**
	Triassic soils draining < 428 km^2	0.42		-17.5	
Upstream Reservoirs	Reservoir (within 1 km)	0.31		-0.2	
	Reservoir (within 5 km)	0.33		-10.3	**X**
Upstream WWTPs	Major WWTP (within 1 km)	0.85		-21.3	**X**
	Major WWTP (within 5 km)	0.48		-16.8	
	Major WWTP (within 10 km)	0.40		-13.1	
	Major WWTP (within 20 km)	0.38		-12.7	
	Minor WWTP (within 1 km)	0.33		1.7	
	Minor WWTP (within 5 km)	0.13*		3.8	
	Minor WWTP (within 10 km)	0.04*		6.9	
	Minor WWTP (within 20 km)	0.06*		6.3	
	Wastewater/Drainage Area (m^3/day*km^2)	0.03		-0.4	
Wetlands	% Wetlands watershed	0.12		-7.5	
	% Wetlands stream buffer	0.02		-3.1	
	% Wetlands near buffer	0.00		-8.1	
	% Wetlands watershed (log)	0.27		-10.6	
	% Wetlands stream buffer (log)	0.13		-12.5	**x**
	% Wetlands near buffer (log)	0.06		-5.4	
Year	Year of Macroinvertebrate sample	-0.013		-80.1	**X**

Table S3: Cross validation results for the multi-stressor model. Coefficients for the parameterization of each fold are shown on the left (along with the mean of all 3 folds). On the right, the coefficients for the multi-stressor model calibrated with the full dataset is shown with their VIFs. R^2 and mean absolute error (MAE) are shown for the deterministic parts (i.e., without site random effects) of each fold.

Predictor	Fold 1	Fold 2	Fold 3	Mean	Full dataset	VIF
Impervious cover (pre 1980)	0.11	0.15	0.12	0.13	0.13	5.3
Impervious cover (watershed)	0.33	0.27	0.33	0.31	0.31	7.0
Lake 5 km	0.32	0.39	0.35	0.35	0.36	1.1
WWTP 1 km	0.69	0.62	0.63	0.65	0.65	1.0
Canopy Loss (buffer)	0.41	0.51	0.39	0.44	0.44	2.1
Year	-0.01	-0.01	-0.01	-0.01	-0.01	1.3
Antecedent Climate (3 month Prec.)	-4.19	-2.47	-2.76	-3.14	-3.10	1.1
Antecedent Climate (squared)	5.58	3.02	3.61	4.07	4.10	-
EPT sampling	-0.96	-0.89	-0.91	-0.92	-0.90	1.5
Qual 4 sampling	-0.15	0.06	-0.02	-0.04	-0.03	1.5
Fall	0.03	0.05	0.08	0.05	0.05	1.4
Spring	-0.21	-0.23	-0.25	-0.23	-0.24	1.4
Winter	-0.26	-0.25	-0.21	-0.24	-0.24	1.4
R^2	0.651	0.644	0.667	0.654	0.754	
MAE	0.559	0.564	0.550	0.558	0.467	

Chapter 3: Forecasting water quality in urban streams under future management scenarios through hierarchical modeling

Abstract

Wadeable streams are an integral part of our environment and their water quality is reflected by a myriad of indicators that fluctuate due to various natural and anthropogenic factors. Though state, federal, and local monitoring data has become abundant in the recent decades, statistical frameworks are needed that can leverage these large data sources to statistically infer relationships between upstream watershed stressors and in-situ water quality in order to assess current risks to aquatic systems as well as design management strategies that might improve future conditions. Here, we use hierarchical multiple linear regression (HMLR) models to identify major natural and anthropogenic drivers of water quality indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC); and predict their values throughout the Upper Neuse River Basin (UNRB). HMLR is a flexible modeling approach that can control for variability that occurs within grouped data (i.e., sites or basins), allowing the inclusion of data with spatial and intra-class correlations. Short and long-term hydro-climatological conditions and spatially intensive point source data (i.e., wastewater loadings) were important predictors of water quality indicators, as well as canopy loss in buffers, crop, and impervious cover (IC) in watersheds. Temporally specific data allowed us to compare the effect of recent construction as well as older vs. newer IC on water quality indicators. BI is mostly affected by upstream land covers (IC and canopy loss in stream buffers), FC and TDU are driven by natural factors (e.g., short-term precipitation), and TN, TP, and SC are best explained by point sources in the UNRB. Five potential management scenarios were evaluated and compared to current conditions under

different hydro-climatological conditions. Restoring canopy cover in stream buffers improved water quality more than mitigating IC or reducing WWTPs for all indicators except TDU. Identifying and mitigating causes of water quality variation across basins and sites (i.e., site and basin random effects) was the most impactful scenario. Uncertainty in water quality forecasts for the UNRB showed that BI, TN and FC would be the indicators most likely to show observable differences post-restoration.

1. Introduction

Water quality in streams and rivers throughout the world is degraded by various anthropogenic sources including increased nutrients, agricultural and urban land covers, water chemistry alterations, habitat changes and water allocation (Malmqvist and Rundle, 2002; Hering et al. 2006; Dudgeon et al. 2006) In the US, over 45% of streams in 2009 were in poor biological condition, more than 40% exceeded federal nutrient levels, 15% have elevated streambed sediments, and over 20% exceeded federal bacteria levels posing a risk to human health (USEPA, 2016). Though some improvements in stream quality have been observed compared to the early 2000s (e.g. less riparian disturbance), many water quality indicators have degraded further (e.g. 9% fewer streams in good biological condition, 14% more streams identified for elevated nutrients; USEPA, 2006; 2009; 2016).

Stream biological condition, as evaluated by their macroinvertebrate community, is often considered the most integrative indicator of stream health (EPA 2009) and is generally degraded by stressors from two sources: urbanization or impervious cover (IC; Schueler, 1994; Schueler et al., 2010; Vietz et al., 2014) and agricultural land uses (Richards et al., 1996, EPA 2006, 2016; Cuffney et al., 2011). The degradation of stream biological conditions due to watershed stressors occurs in small stream networks (Paul and Meyer, 2001; Walsh et al., 2005) up to regional estuary systems (Miller et al., 2018). Other water quality indicators commonly monitored in wadeable streams include nutrients, turbidity (or sediments), and bacteria levels. Elevated nutrients levels have been correlated with net anthropogenic nitrogen inputs and variations in precipitation across multiple scales (Howarth et al., 2012; Sinha and Michalak, 2016; Strickling and Obenour, 2018). Elevated nitrogen and phosphorous export associated with non-point sources such as crop, pasture, and urban lands (Beaulac and Reckhow, 1982; Alexander et al.,

2002; Alexander et al., 2007) can elevate nutrient levels as to impair streams (USEPA, 2006; EPA, 2016). Point sources such as NPDES dischargers and livestock facilities can also contribute substantially to nutrient loading (Strickling and Obenour, 2018; Tchobanoglous et al. 2003). Stream sedimentation is associated with new construction activities (Finkenbine et al., 2000), reduction of riparian buffers (Richards et al., 1996; Sutherland et al., 2002; Allan 2004), and elevated flows associated with urbanization (O Driscoll et al., 2010). Excessive stream sedimentation can blanket the hyporheic zone and destroy habitat for macroinvertebrate species and stream fauna. Elevated concentrations of bacteria, such as fecal coliform (FC), have been found near urban areas and attributed to runoff and septic leakage into streams (Kelsey et al., 2002; Walsh and Kunapo, 2009; Gentry-Shields et al., 2012). In addition to anthropogenic sources, antecedent and extreme precipitation events affect nutrients, turbidity (or sediments), and bacteria levels as well (Herlihy and Sifneos, 2008; Sinha and Michalak, 2016).

Efforts to protect or improve water quality in the US have focused on restoration activities for degraded streams and management activities to address future construction activities (USEPA, 2005). Over $14 B was spent on restoration efforts from 1990 to 2005 (Bernhardt et al., 2005), and the cost of retrofitting water quality controls in only one metropolitan area (Atlanta) was estimated at $1.5 B (Sample et al., 2003). Though early measures of restoration success for water quality improvements have been limited due to a lack of pre- and post-monitoring (Bernhardt et al., 2005; Brooke and Lake, 2007), more recent studies have shown some limited success (Palmer et al., 2014). One reason for little observed improvements might be the dominant effect that large-scale watershed factors have on stream quality relative to local effects (Selvakumar et al., 2009; Matthews et al., 2010; Stoll et al., 2016). Most restoration activities occur at the stream reach level, and do not address catchment-

wide efforts that would involve more resources and coordination (Walsh et al., 2016) limiting their ability to mitigate the most critical water quality issues (Bernhardt and Palmer, 2007; Palmer et al., 2010; Ernst et al., 2012).

Empirical models have been widespread in water quality modeling to determine statistical relationships between variables that are assumed to have a mechanistic connection (Mays 2010; Wilson 2016). Regression modeling has been used extensively to relate different spatial scales of upstream land covers to stream health (Strayer et al., 2003; Schueler et al. 2009), predict ecological health throughout larger stream networks (Belucci et al., 2013), and compare regional differences in the effect of land use covers on macroinvertebrate communities (Cuffney et al. 2010; Qian et al. 2010; Miller et al. In review). More recently, machine learning techniques, such as random forests (Breiman, 2001), have been used for large-scale predictive studies focused on estimating baseline water quality conditions for specific conductance and nutrient levels (Endries, 2011; Olson and Hawkins, 2012; Olson and Hawkins, 2013; Hill et al., 2017). These techniques use classification and regression trees to make predictions which can account for non-linear relationships and interactions between predictors. Though, machine learning techniques have good predictive ability (Evans and Cushman, 2009; Evans et al., 2011), they are considered “black box” methods that cannot easily answer questions about the magnitude of relationships between predictors and response variables.

Hierarchical modeling is an extension of regression modeling that offers a flexible framework for including grouped data and has become increasingly common in ecological studies (Clark and Gelfand, 2006; Bolker et al., 2009), including linking the biological conditions of streams and estuaries to upstream watershed conditions across regional datasets (Qian et al., 2010; Cuffney et al., 2010; Miller et al., 2018). Hierarchical models use “random

effects" to determine the hyperdistribution of group-level factors (e.g., sampling locations, watershed, etc.). Random effects allow for more flexibility than SLR and MLRs as these random effects can account for intra-class correlation of data allowing the inclusion of repeatedly sampled data from a location. They can also incorporate data from sites that are hydraulically connected and possibly spatially correlated by considering them to be part of the same group (Gelman and Hill 2006; Gelman et al., 2014). Previous predictive models (Hill et al., 2017) have discussed how regional models lost predictive power when joined together to create one global model for the US, which highlights the need for modeling approaches that can leverage all available data while addressing variation within and across regions (e.g., eco regions, geologic soil type, watersheds).

The aim of this study is to identify natural and anthropogenic drivers of 6 major water quality indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC). Using the NC Piedmont for inference, mean water quality conditions were then forecast probabilistically for the Upper Neuse River Basin (UNRB), a major Piedmont basin. To accomplish this, we leveraged over 90,000 water quality samples from a 17-year study period demonstrating the first use of a hierarchical multiple linear regression (HMLR) approach that can accommodate intra-class correlation of repeatedly sampling data at monitoring sites. By including temporally varying data, we rigorously tested the importance of temporal predictors such as short and long-term hydro-climatological conditions and changing land covers (IC). Mean water quality conditions for the UNRB were forecasted for potential management scenarios under different hydro-climatological conditions. Finally, uncertainty due to HMLR coefficients and errors was used to assess the probability of observing water quality improvements during post-restoration monitoring.

2. Methods

This study used HMLRs to model six water quality indicators (i.e., BI, FC, TDU, TN, TP, and SC) using data from the North Carolina (NC) Piedmont and builds upon methodology already established in Miller et al. (In review) that focused exclusively on BI. Initially, candidate predictors from 16 variable classes were evaluated separately and compared to a baseline model using an information criterion approach (see section 2.3, 2.4). Then, HMLRs for each indicator were developed using candidate predictors that improved baseline model performance. Predictor variables were then ranked based on their contribution to the variation in water quality indicators (see 2.5). Water quality predictions throughout the Upper Neuse River Basin (UNRB) were made for current conditions under different hydro-climatological conditions, and then forecast for potential management scenarios to evaluate potential improvements in water quality (see 2.6, 2.7).

2.1. Study Area

Water quality data were collected from wadeable streams throughout the NC Piedmont region (Figure 1a), which encompasses six hydrologic units at the sub-basin level (United State Geological Survey (USGS) Hydrological Unit Code 8): Neuse, Cape Fear, Tar-Pamlico, Roanoke, Yadkin-Pee Dee, and Santee Rivers (Figure 1A), and nine major geologic regions: Raleigh Belt, Carolina Slate Belt, Kings Mountain Belt, Inner Piedmont, Milton Belt, Charlotte Belt, Sauratown Mountains Anticlinorium, Smith River, and the Triassic Basin (NC Geological Survey 2009; Fig S1). “Wadeable” streams were defined as NHD stream reaches (Moore and Dewald, 2016) that drain between 2 and 500 km^2 at their outlet. Stream segments that drained less than 2 km^2 were assumed to be intermittent and excluded from the study while those over 500 km^2 approximate 5th order streams that are generally considered to be unwadeable (USEPA,

2006; Pierson et al., 2008). Elevations for water quality sampling locations in the NC Piedmont ranged from 34-546 m above sea level.

Water quality predictions were then made for wadeable streams in the non-Coastal Plains portion of the UNRB (Figure 1B; HUC 8-030201; USGS), a rapidly developing watershed in the NC Piedmont (Strickling and Obenour 2018). The northern portion of the UNRB drains into Falls Lake, a major drinking water source for Raleigh, NC. Falls Lake has been impaired due to elevated levels of chlorophyll-a since 2008, and has excessive turbidity in its northern portions (NC DEQ 2018a). Two quickly expanding urban areas are located within the UNRB, Raleigh and Durham, along with several smaller municipalities. 28% of UNRB stream reaches drain basins with over 10% IC, 20% of reaches drain basins with over 30% pasturelands, and 7% of reaches have some amount of wastewater flow. We created a prediction grid throughout UNRB streams where predictor data was compiled to make water quality forecasts (Fig S2). Each point on this grid corresponded to the midpoint of one wadeable NHD stream reach (1,423 reaches, 2,155 km of streams; Moore and Dewald, 2016). Watershed delineations and land cover processing was performed using Spatial Analyst tools within ArcGIS 10.3 (Esri 2015). The main stem of the Neuse River below Falls Lake (88 km) was the only part of the stream network excluded for being unwadeable (i.e., > 500 km^2). Though nine major geologic soil regions exist throughout the entire NC Piedmont, only three are present in the UNRB: Carolina Slate Belt, Raleigh Belt, and the Triassic Basin (Fig S1). The UNRB encompasses 2,865 km^2 and ranges in elevation from 42-221.

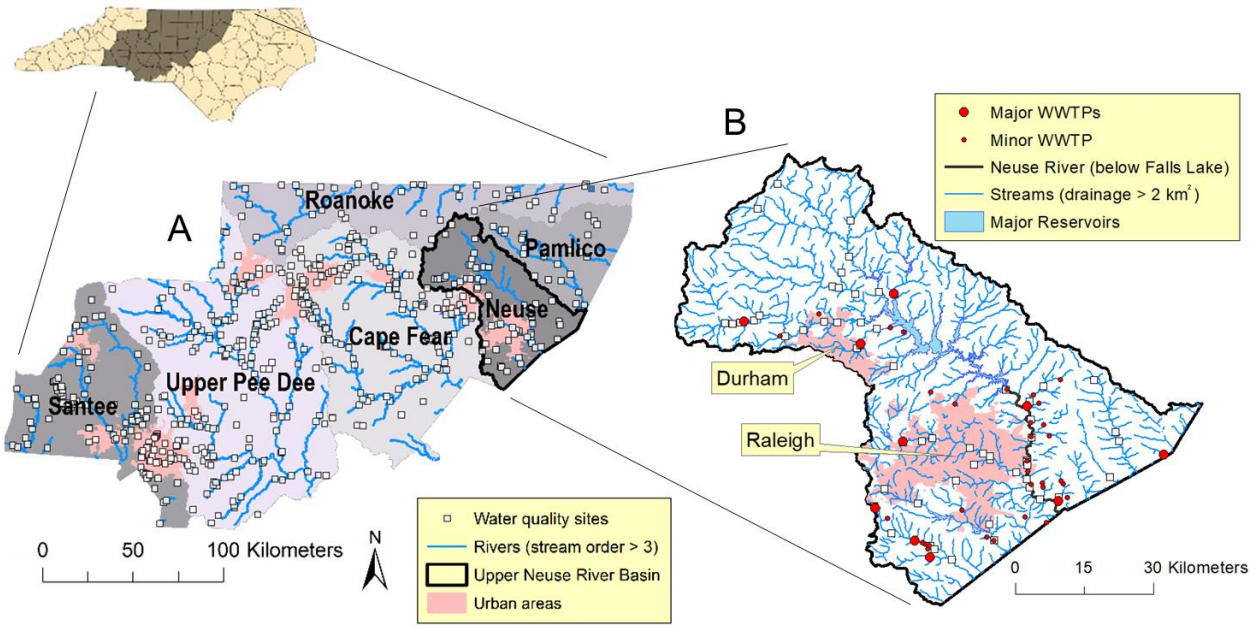


Figure 1: Location of 653 water quality stations in the NC Piedmont used for model development (A) along with the six HUC 8s where data was collected (shaded gray). The Upper Neuse River Basin, where water quality forecasts for potential management scenarios was explored, is shown on the right (B) along with wastewater treatment plants (WWTPs) and sampling locations. The main stem of the Neuse River below Falls Lake (black) was considered unwadeable, and therefore omitted from the model.

2.2. Water quality data

Water quality data for FC, TDU, TN, TP and SC (i.e., response variables for models), were collected from the Water Quality Portal (Read et al. 2017) and came from 3 main sources: USGS, NC Department of Environmental Quality (DEQ) ambient monitoring system (NC DEQ 2018b), and the NC DEQ monitoring coalition program (NC DEQ 2018c). Water quality data were collected from 653 water quality stations throughout the Piedmont, though not all water quality indicators were sampled at each site (Table 1). BI data were obtained from state and city sources as described in Miller et al. (In review) and had different locations than water quality sites. BI were sampled once a year (city sites) or every five years (NC DEQ sites), unlike data for FC, TDU, TN, TP, and SC, which were typically sampled monthly. Therefore, though BI had the

highest number of sites, it also had the fewest number of samples (Table 1). Some water quality values were removed because they had unrealistically high values, suggesting typos in the original databases. TN and TP values above 100 mg/L and 50 mg/L were removed (8 samples) as these values are significantly above values for untreated wastewater (Carey and Migliaccio, 2009).

Table 1: Summary of data for each water quality indicator: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC). All data were collected between 1994-2010.

Indicator	Units	# of sites	# of basins	# of samples	Values				Drainage size		
					5%	Median	Mean	95%	5%	Mean	95%
BI	-	1020	689	2,769	3.9	6.1	6.0	7.8	3	80	324
FC	#/100 mL	196	127	17,299	6.3	62.9	410	1435	4	156	436
TDU	NTU	192	131	21,715	1.2	4.0	7.9	24.3	6	153	436
TN	mg/L	189	138	14,635	0.5	0.8	1.4	4.2	6	154	456
TP	mg/L	158	116	8,729	0.4	0.4	0.5	0.9	3	127	436
SC	SU	257	181	26,548	20.6	55.2	72.1	181.0	3	146	432

2.3. Candidate predictor variable classes

Candidate predictor variables from 16 variable classes (Table 2) were tested as possible predictors for water quality indicators. Each variable class was compiled to represent different aspects of natural factors, upstream land covers, and point sources (e.g., WWTPs, reservoirs) that might affect in-stream water quality. Variable classes included predictors collected at different spatial or temporal scales as well as transformed (natural log and linear) to explore different potential relationships with water quality. All log transformations were performed using the natural logarithm and variables were offset by adding one to avoid non-real transformations. For each variable class (Table 2), only the predictor that improved the model performance the most was included as a candidate for the multi-stressor water quality model (see Model construction 2.4). One variable class, sampling method, was specific to the BI model, and controlled for

different macroinvertebrate collection techniques that were shown to be an important predictor as described in Miller et al. (In review).

Table 2: Sixteen variable classes and their candidate predictor variables tested for inclusion in multi-stressor water quality models.

Variable classes	Candidate predictor variables	Source
Canopy Loss	Fraction of basin and stream buffer (linear and log transformed)	Falcone, 2015 Homer et al., 2015
Crop	Fraction of basin and stream buffer (linear and log transformed)	Falcone, 2015
Drainage area	Size of entire upstream basin (km^2) Sites that drain less than 7.7 km^2	USGS 2017
Hydro-Climatology (short)	Precipitation (previous 1,3,7,14 days)	Maurer et al., 2002
Hydro-Climatology (long)	Precipitation (previous 1,3,6,12, and 24 months) Drought indices (SPI; previous 1,3,6,12,24 months) (linear and quadratic)	SCO, 2017
Elevation	Elevation at sampling location (m)	USGS 2017
IC	Fraction of basin and stream buffer (linear and log transformed)	Falcone, 2015 Homer et al., 2007,2015
IC (recent) IC (age)	Change in basin IC in previous 2, 5 yrs. IC in basin before: 1974,1980,1990, 2000, 2005 (linear and log transformed)	Falcone, 2015 Homer et al., 2007,2015
Pasture	Fraction of basin and stream buffer (linear and log transformed)	Falcone, 2015
Reservoirs	Reservoirs within 1 and 5 km of sampling locations	NHD 2007
Season	Season of macroinvertebrate sample	NC DEQ, 2017
Soil type	12 geological regions of NC Piedmont	NC Geological Survey, 2009
Wetlands	Fraction of basin and stream buffer (linear and log transformed)	Falcone, 2015
WWTP (loadings) WWTP (#, spatial proximity)	TN and TP daily loadings of upstream WWTP normalized by drainage area [kg/(day*km^2)] (linear and log transformed) Major and minor WWTPs within 1, 5 10, and 20 km Number of upstream major and minor WWTPs	NC DEQ 2017
Year	Year of the macroinvertebrate sample	NC DEQ, 2017

Natural factors included antecedent hydro-climatological conditions (i.e. precipitation and drought indices), geologic soil type, elevation, season, and the drainage area of sampling locations. Short-term precipitation data (i.e., one, three, seven, and fourteen days before samples

were taken) came from daily gridded observations from across the Piedmont at 1/8 degree spatial resolution (~ 12 km; Maurer et al., 2002), and they were matched to monitoring data based on their proximity to the centroid of upstream basins. Short-term meteorological data collected for this fine resolution grid were only available up to 2010, effectively limiting our study to that date. Long-term precipitation and drought indices were obtained from the North Carolina State Climate Office (NC SCO 2017) and matched to sites belonging to three Piedmont regions (e.g. Southern, Central, Northern; NC SCO 2018).

Anthropogenic land cover variable classes included basin IC, canopy loss, wetlands, crop, and pasture. Land cover data were obtained from the U.S. Conterminous Wall-to-wall Anthropogenic Land Use Trends (NWALT) dataset for 1974, 1982, 1992, 2002, and 2012 (Falcone 2015) and the National Land Cover Dataset (NLCD) for 2001 and 2011 (Homer et al. 2007; Homer et al. 2015). NWALT was necessary to obtain consistent land covers back to 1974 while NLCD was utilized to obtain IC and canopy cover estimates. Multiple linear regressions were then used to relate NWALT land covers designations to corresponding NLCD IC and canopy cover (Miller et al. In review). On average, 17% of residential areas, 35% of transportation, and 57% of industrial/commercial were IC, while 92% of low use, 88% of semi-developed, 48% of residential, and 12% of industrial/commercial development had canopy cover. All land cover variables were then interpolated between reported NWALT years (1982, 1992, 2002, 2012) to obtain year-specific land cover data for all sites. In addition to the entire upstream basins, buffers that included 30 m of land on either side of NHD Plus flowlines (Moore and Dewald 2016) were also calculated for each land cover variable class.

Temporal variability in IC was represented in two ways: age of IC and recent IC (Table 2). The age of IC was defined as the amount of IC constructed before a given date (e.g., pre-1980

IC) and was included to determine if the effect of IC on water quality indicators has changed in magnitude over the past several decades. Older construction has been shown to have a more negative effect on the biological health of streams (i.e., BI) than more recent construction, presumably due to more stringent standards through time (Miller et al. In review). The age of IC was defined as the product of the linear or log transformed IC of the sample year with the ratio of IC before a given date (i.e., 1974, 1980, 1990, 2000) and the IC of the sample year. Recent IC was included to test the impact that construction projects might have on water quality due to sediment runoff. Recent IC was defined as the amount of new IC added to a basin during the past two or five years. This increase in IC was then considered as a candidate predictor as both a linear and log transformed variable.

Upstream point sources were accounted for using predictors for the presence and spatial proximity of WWTPs and reservoirs, as well as WWTP nutrient loadings. Both WWTP and reservoirs were tested as categorical variables based on their spatial location upstream of monitoring sites. Distances of one, five, and ten miles were used to determine if proximity to WWTPs were important to water quality. WWTP discharge (i.e., flow) and nutrient concentrations were provided by NC DEQ (personal communication with Bridget Munger) for both major (> 4,546 m^3/day) and minor dischargers (Table S1). To calculate TN and TP loads (kg/day) specific to each WWTP, discharge and concentration records were multiplied together for corresponding months and then normalized by drainage area. Preliminary analysis showed that using median TN and TP loads for a given year were more correlated to water quality indicators than monthly TN and TP loads, probably due to missing monthly data for many WWTPs. (Table S1). Nutrient loads (TN and TP) from WWTPs were used as predictors in their respective water quality model, while TN load was also used in the additional water quality

models (BI, TDU, FC, and SC) to be a proxy for wastewater that might be in streams. Additionally, a count variable for the number of upstream dischargers was used as a proxy for water quality indicators (i.e., SC, FC and TDU) that were not routinely reported.

2.4. Model construction

Initially, baseline models for all water quality indicators were constructed to test candidate predictors from each variable class. Baseline models included site and basin random effects (α_{st} , α_{bn}; Eq.1) and up to one candidate predictor (x_b) if it explained over 50% of the variability (i.e., R^2) for the response variable (BI, FC, TDU, TN, TP, or SC). The inclusion of up to one significant predictor in baseline models was done to streamline variable selection as weaker statistical relationships can be obscured or reversed by stronger relationships (Samuels 1993). The BI model was based on previous work (Miller et al. In Review) and also included sampling method in its baseline model. Site and basin random effects identified mean deviations of monitoring sites or basins from overall mean conditions in the Piedmont. The location of each sampling site was assigned to a basin which represented sites that share more than 25% of the same drainage area. In basins where numerous sampling locations existed that met this requirement, but the two farthest sites did not, it was separated into two basins between the largest break in sampling locations. Most sampling sites were in unique basins and were thus given both a unique site and basin random effect (Table 1). All predictor candidate variables (x_p) from each variable class (Table 2) were tested individually with baseline models in three ways to determine if their addition improved model performance: (1) adding a single linear or log-transformed variable (x_p; Eq. 2a), (2) adding the variable and its 2nd order (quadratic) term (Eq. 2b) for long-term precipitation and drought indices, and (3) adding a categorical variable or random effect (α_p; Eq. 2c).

Predictor variables that improved baseline model performance (Eq. 2 a,b,c) were considered as candidates for multi-stressor HMLRs. The Bayesian Information Criterion (BIC), an information criteria that judges model fit based on log likelihood while penalizing for extra parameters (Schwarz 1978), was used to assess this improvement. A conservative level for inclusion was set to avoid over-parameterization and only candidate predictors that lowered the baseline BIC by a minimum of 6 (Raftery, 1995) were considered for inclusion in the multi-stressor model. If variable classes had several predictors that met this criterion, only the most influential from each variable class (i.e., lowest BIC) was chosen for further consideration. Variable classes with no predictor variables that met this criterion were excluded from further consideration.

The best candidate predictor from each variable class were then combined to form a multi-stressor HMLR for each water quality indicator (Eq. 3). Statistically insignificant variables were parsed out using a backward selection process using the "LMERConvenienceFunctions" R package (Tremblay and Ransihn, 2015). Variance inflation values (VIF) were also calculated for HMLRs to test for multi-collinearity in predictors. All hierarchical models were fit in R using the "lme4" package (Bates et al. 2015; R Core Team 2017). The baseline and multi-stressor model formulations are presented below:

$$WQ = \beta_0 + (x_b * \beta_b) + \alpha_b + \alpha_{st} + \varepsilon \qquad \text{(Baseline Model; Eq.1)}$$

$$WQ = \text{Baseline} + x_p * \beta_p \qquad \text{(Eq.2a)}$$

$$WQ = \text{Baseline} + x_p * \beta_p + x^2_p * \beta_{psq} \qquad \text{(Eq.2b)}$$

$$WQ = \text{Baseline} + \alpha_p \qquad \text{(Eq.2c)}$$

$$WQ = \text{Baseline} + \text{multiple candidate predictors} \qquad \text{(multi-stressor HMLR; Eq.3)}$$

where WQ is the water quality indicator, β_0 is the model intercept, β_b represents the coefficient for the baseline predictor if present, β_p is the slope coefficient for the candidate predictor being tested, and ε represents the model error [N $(0, \sigma^2_{WQ})$]. Site random effect (α_{st}) and basin random effects (α_b), come from a normal distribution centered around zero ($\alpha_{st} \sim N(0,\sigma^2_{st})$; $\alpha_b \sim N(0,\sigma^2_b)$] where σ^2_{st} and σ^2_b represent the variance of their hyperdistributions. The number of site and basin random effects varied for each water quality indicator and corresponded to the number listed in Table 1 for each water quality indicator

2.5. Model assessment

HMLR models were used to rank the importance of predictor variables for each indicator by comparing the standard deviation (SD) of the contribution of each predictor to its response variable [i.e., SD $(x * \beta)$]. For the majority of predictors, this process is similar to comparing the magnitude of coefficients determined from normalized predictor variables. Some predictors though (i.e. quadratic trends, categorical variables, or random effects) cannot be compared as normalized predictors, but this method is flexible enough to allow their contribution to the response to be compared in a similar way. Predictor contributions were then normalized by the standard deviation of their response variable, so they could be compared across HMLRs for different water quality indicators. Model skill (i.e., R^2) for multi-stressor HMLRS was calculated for deterministic factors that are always known for sites (i.e., all candidate predictors and soil type) and for the full model (deterministic plus site and basin random effects). R^2 based on averaging individual samples on a yearly basis from sites with six or more samples in a year were also calculated.

All models were subjected to a 3-fold cross validation (Elsner and Schmertmann 1994, Chatfield 2006) to check the stability of predictor coefficients and the drop of model performance for out-of-sample predictions. Multi-stressor HMLRs were recalibrated using 2/3 of the full dataset to make predictions for the 1/3 that was excluded. In this way, predictions for all data points were made using models parameterized without them included. For each water quality indicator, predictor coefficients for each fold and model fit (R^2 and MAE) were recorded.

2.6. Forecasting management scenarios in the UNRB

HMLRs were initially used to predict current conditions for water quality indicators in all wadeable stream reaches (1,423 reaches; see section 2.1; Fig 1; Fig 2) throughout the Piedmont portion of the UNRB. "Current" conditions for UNRB water quality corresponded to 2012, the last year of reported land cover data (Falcone 2015), to avoid extrapolating land cover data to the present year. Predictions were made using deterministic effects (i.e., predictors and soil random effects), which were known for all prediction sites, to obtain mean current conditions for each reach under median weather conditions (see below). Reach predictions were then weighted by their length (mean = 1.5 km; max = 7.5 km; Moore and Dewald, 2016), and averaged across the UNRB to determine a weighted mean value for each indicator across the UNRB. Weighting was done using predictions in the transformed space (i.e., in log space for FC, TDU, TN, TP, and SC) and back-transformed after the log weighted mean was calculated. Current conditions were established first as baseline conditions in the UNRB in order to judge the impact of potential management scenarios on water quality indicators.

Five potential management scenarios for improving water quality were forecasted and compared to current conditions to evaluate their possible benefits (Table 3). Scenario #1 represented a 50% decrease in the amount of canopy buffer loss throughout the UNRB (i.e.,

increase in canopy buffer), scenario #2 lowered the effect of IC 25% on streams, and scenario #3 lowered the effect of WWTPs by 25%. Scenario #4 was a combination of scenarios #1-3. For each management strategy, the impact of different candidate predictors (Table 2, 3) where lowered accordingly. Scenario #5 focused on lowering site and basin random effects, rather than reducing predictor values. Site and basin random effects represent unknown variance at these locations not explained by other predictors in our model. By lowering only positive random effects, this represented reductions to anthropogenic factors not accounted for by other candidate predictors (e.g., unaccounted for point sources, leaky infrastructure, illicit dumping) at these locations that were negatively affecting water quality. Scenario #5 assumes that some of these sources of degradation (25%) could be identified by local managers and mitigated. To forecast scenario #5, the variance of site and basin random effects hyperdistribution was randomly sampled for the 1,423 UNRB stream reaches 100 times. For each sample, positive random effects (i.e., more degraded sites), were lowered by 25% assuming mitigation occurred, while negative effects were left unchanged. Forecasts were then weighted by stream length as explained before. For each stream reach, the mean value of the 100 samples was extracted to represent the expected forecasted conditions for scenario #5 and back transformed (if needed) to compare with current conditions.

Hydro-climatologic conditions were the main driver of several indicators such that three different hydro-climatological conditions (average, dry, and wet) were explored to see their effect on water quality indicators. “Median conditions” were defined as median short and long-term hydro-climatological conditions based on historical data from 1981-2010, while “wet conditions” were set to their 90th percentile values, and “dry” conditions were represented with their 10th percentile values. All three hydro-climatological conditions were run for current

conditions as well as for scenario #4 in order to evaluate weather variability on median conditions in the UNRB and determine how varying weather conditions affected management scenarios.

Table 3: Five management scenarios and the variable classes, candidate predictors, and random effects that were adjusted to make forecasts.

	Management Scenario	**Candidate predictor variables affected**
1	Increase canopy cover in stream buffers (50%)	Canopy Loss (buffer)
2	Decrease effect of IC (25%)	IC (basin, buffer) IC (recent) IC (age)
3	Decrease effect of WWTP (25%)	WWTP (loadings; #; spatial proximity)
4	Combination of scenarios #1-3	Canopy Loss (buffer) IC (basin, buffer) IC (recent) IC (age) WWTP (loadings; #; spatial proximity)
5	Mitigate positive site and basin random effects (25%)	Site random effects Basin random effects

2.7. Uncertainty in water quality forecasts

Current conditions and scenario #4 (combination of #1, #2, and #3) were then predicted probabilistically to quantify the uncertainty in water quality forecasts due to variance associated with the multi-stressor HMLR models and weather conditions. Initially, uncertainty in fixed effects (i.e., predictor coefficients) was simulated by creating 1000 bootstrapped predictions for each reach (i.e., 1,423) using “bootmer” in “lme4” (Bates et al. 2014). Bootstrapping is a technique that relies on resampling the response variable (with replacement) to determine measures of model accuracy (Efron, 1992; Davison and Hinkley, 1997). Random model error was then added to these predictions for each water quality indicator, and these bootstrapped samples were used as the basis for further simulations. Though site and basin random effects

would not be known at a given site, it was assumed that these values would remain constant at a given reach for current and scenario #4 conditions; and therefore, not affect the uncertainty of water quality improvement.

The goal of simulations was to determine (1) if a difference in water quality indicators *should occur* based only on fixed effect uncertainty and HMLR model error, and (2) if a difference *would be observed* based on a distinct number of pre (current) and post (scenario #4) monitoring samples taken under constant and differing weather conditions. For each reach, a given number of pre and post samples (6, 12, 24, 36, 48, and 60) were randomly drawn from their (1,000) bootstrapped samples that represented uncertainty of HMLR coefficients and model error. Then, mean values of the pre and post samples (log transformed for TDU, FC, TN, TP, and SC) were compared to see if indicators were reduced in scenario #4 as compared to current conditions. To add weather variability, a random day taken from a 29-year historical weather record (1982-2010) was assigned to each pre and post sample. Short and long-term hydro-climatological data for that day were then used to adjust predictions accordingly. For each number of pre and post samples (6, 12, 24, 36, 48, and 60), this process was repeated 200 times to test the percent chance that a difference would be observed. In addition, t-test were performed on the same pre (i.e., current conditions) and post (i.e., scenario #4) samples to determine if statistically significant differences were observed. Finally, the same simulations were analyzed for subsets of the UNRB (e.g., top 10% of upstream urbanization, pasture activities), to offer insight into which anthropogenically disturbed regions might produce the most observable water quality changes.

3. Results

3.1. Water quality multi-stressor HMLR models

Baseline models for water quality indicators were created using site and basin random effects as well as candidate predictors that explained more than 50% of the variability in the response. The log of watershed IC explained 65% of the variance in BI and the log transformation of IC implied a saturating relationship such that as total watershed IC increased, further additions had a reduced effect on BI. The log of TN load (normalized by drainage area) was able to explain 50% of the variability in the TN baseline model and implied a power relationship where a 1% increase in TN load would lead to a 0.62% increase in stream TN concentrations. Other indicators (FC, TDU, TN, and SC) included only site and basin random effects in their baseline models.

Candidate predictors were then added individually to baseline models and the improvement of model performance (i.e., Δ in BIC; Table S2) was used to determine potential candidate predictors for inclusion in multi-stressor HMLRs. All six water quality models included a candidate predictor from short and long-term hydro-climatology, seasonal effect, and temporal effects of IC (recent or age) while canopy loss, IC, and WWTP loadings were chosen for inclusion in five of the six models. Multi-stressor HMLRs were formed by then joining all candidate predictors together and parsing insignificant ones. Predictors for each multi-stressor HMLR are presented to show their relative contribution to each water quality indicator (Table 4; model coefficients shown in Table S3). Predictor contributions are unitless and were derived by dividing the SD of the predictor contribution to the response variable by the SD of the response variable (see section 2.4). By comparing the magnitude of predictor contributions within models, the main driver of each water quality indicator was identified: watershed IC for BI (sampling

method represented a sampling difference, not a stressor of BI), short-term precipitation for FC and TDU, and WWTP loadings or presence for TN, TP and SC (Table 4). The contribution from seasonal, sampling methods, and random effects (α) could not be expressed as either positive or negative, but the magnitude of their contribution to the response could be judged relative to other predictors in the models to show their influence. Site random effects explained more variability in FC than other predictors and BI and SC had large soil random effects only slightly lower in magnitude than their main drivers.

Model skill for multi-stressor HMLRs was calculated for deterministic factors that could be obtained for any site in the Piedmont (i.e., all candidate predictors and soil type) and for the full model (deterministic factors plus site and basin random effects). The skill of deterministic predictions reflects the ability of HMLRs to make predictions in new locations, where site and basin random effects would be unknown. The stream biodiversity (i.e., BI) model had the highest skill using only deterministic factors (R^2= 0.76; Table 4), while TN and SC models had the next best model fits (R^2= 0.54; 0.40, respectively). Water quality driven mainly by short-term precipitation events, FC and TDU, had the lowest model skill (R^2= 0.25 and 0.28, respectively). R^2 based on yearly means for sites with six or more samples in a given year improved for FC, TN, TP, and SC models (44%, 26%, 47% and 31% increases respectively), while no change was noted for TDU. The BI model did not have a yearly R^2 calculation as BI is only sampled once a year.

The BI model was mainly driven by two anthropogenic factors: log percent of IC in upstream watersheds and the log percent of canopy loss in stream buffers (Table 4; Table S3). This model was very similar to previous work (Miller et al. In review), though it used a reduced dataset (1994 -2010) limited by the availability of other predictor data (i.e., WWTP loadings and

Table 4: Multi-stressor HMLRs for water quality indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC). The amount of variance (R^2) attributed to deterministic portions of the model (predictors, soil and seasonal effects) are shown as well as variance explained by the full HMLR (deterministic plus site and basin random effects). Deterministic variance was also calculated for averaging individual samples on a yearly basis at a site (except for BI where sampling did not occur more than once a year). Predictors (β; unitless) are represented as the standard deviation of their contribution to the response variable normalized by the standard deviation of their responses. In this way, coefficients in the table can be compared vertically (within a model) as well as horizontally (across models) to see their relative influence on water quality indicators. "Basin" represents the % of upstream watersheds covered by a given land type while "buffer" represents the % that land type in 30 m stream buffers. Categorical predictors (α) cannot be classified as either positive or negative. Of note, the BI model had a quadratic trend (q) for its long-term precipitation. Numbers in parentheses for IC (recent) and hydro-climatology (short and long) refer to the number of days, months, or years that define that predictor. For IC (age), the number represents IC constructed before a given date (e.g., 80 represents pre-1980; Table 2).

		BI	FC	TDU	TN	TP	SC
R^2	Deterministic	0.76	0.25	0.28	0.54	0.32	0.42
	Deterministic (yearly)	-	0.36	0.28	0.68	0.47	0.55
	Deterministic + site + basin	0.92	0.40	0.40	0.72	0.58	0.70
β	Canopy loss, buffer	+0.19	+0.13		+0.16	+0.07	+0.05
	Crop, buffer		-0.07				
	Crop, basin			+0.09	+0.10		-0.18
	Drainage		-0.17				
	Hydro-clim, short (days)		+0.30 (3)	+0.41 (3)	-0.06 (14)	-0.07 (7)	-0.14 (7)
	Hydro-clim, long (months)	-0.05(3; q)	+0.06 (1)	+0.12 (1)	-0.06 (6)	-0.07 (6)	-0.12 (6)
	IC, buffer				-0.10		
	IC, basin	+0.21					
	IC, age (year)	+0.10 (80)	+0.26 (00)				+0.32 (80)
	IC, recent (# of years)			+0.12 (5)		+0.20 (2)	
	Pasture, buffer		+0.19				
	Reservoirs, 5 km	+0.05					
	WWTP loadings		+0.04		+0.66	+0.47	
	# of Major WWTP						+0.36
	WWTP proximity, 1 km	+0.05					
	Wetlands, buffer				-0.06		+0.06
	Year	-0.07	-0.04		+0.04		
α	Sampling method	0.26					
	Season	0.08	0.15	0.16	0.06	0.13	0.12
	Soil type	0.19	0.16	0.14	-	-	0.26
	Site random effects	0.12	0.37	0.19	0.22	0.23	0.29
	Basin random effects	0.16	0.00	0.18	0.25	0.30	0.29
	SD of response variable	1.49	2.76	0.86	0.45	0.09	0.44

short-term precipitation; Table 2). The inclusion of both IC and pre-1980 IC imply that more recent construction (IC post-1980) impacts stream health less than older development (IC pre-1980). The BI model was the only one to have a quadratic trend in long-term precipitation implying that both low and high flows could be detrimental to stream biological health. Also, BI was the only indicator to be impacted by spatial proximity to WWTPs and reservoirs, which had large impacts within relatively small spatial distances (5 and 1 km; Table 4, Table S3).

FC and TDU were indicators whose main driver was short-term precipitation (i.e., previous 3 day precipitation; Table 4), implying that variation in these indicators at sampling locations was driven more by natural factors than anthropogenic ones (Table 4). FC did have various anthropogenic predictors though including: IC (pre-2000), pasture lands, canopy loss in buffers, and WWTP loadings. FC values were greatly elevated in the summer, compared to other seasons, and the size of drainage areas also had a negative relationship on FC. Two anthropogenic factors increased TDU, croplands and recent construction (past 5 years).

Variability in stream TN and TP concentrations was mainly explained by the log of WWTP loadings (normalized by watershed area). The influence of WWTP TN loadings on stream TN concentrations was the largest influence of any predictor on a water quality indicator throughout the Piedmont (+0.66; Table 4). The second most important anthropogenic factor for increased TN levels was canopy loss in stream buffers, which occurs mostly in agricultural (crop or pasture) and urban areas (see section 2.3). Both canopy loss and IC in stream buffers were included in the TN model and are negatively correlated with each other ($r = -0.87$). The sign and magnitude of their coefficients imply that buffer canopy loss in non-urban areas contribute to higher stream TN concentrations than in urban (high IC) areas. Crop positively contributed to elevated TN concentrations, while wetlands in stream buffers and short and long-term hydro-

climatological conditions all slightly lowered stream TN concentrations. Finally, there was a statistically significant yearly trend upward in TN concentration, implying that a TN stream concentration of 1.19 mg/L in 1994 (50th percentile for UNRB) would have increased to 1.43 mg/L by 2010. Though TP concentrations were also mainly driven by upstream WWTP loadings (TP load normalized by drainage area), model skill for deterministic factors was much lower for TP than TN (0.32 as compared to 0.54; Table 4). Additional anthropogenic factors that increased TP were recent IC added in the previous 2 years and canopy loss in stream buffers while both long and short-term precipitation decreased TP concentrations (Table S4).

The two most influential drivers of SC were both anthropogenic factors: the number of major upstream WWTPs and pre-1980 IC (Table 4). Cropland was also negatively correlated with SC, while wetland buffers had a positive relationship. Increased rainfall (short and long-term hydro-climatology) lowered SC levels as well (Table 4).

Natural factors, such as soil type and seasonal effects, had substantial effects on water quality indicators. BI values in the Triassic basin were greatly elevated (+0.65), a similar effect to a site being located within 1 km of a WWTP (Table S3). Soil type had the largest impact on SC (Table 4), which varied greatly throughout the Piedmont being highest in the Carolina Slate Belt (+0.33) and lowest in the Inner Piedmont (-0.33; Table S5). In contrast, soil type played no significant role in either TN or TP stream concentrations (Table 4). Seasonality affected all six water quality indicators and results implied that water quality is generally worse in summer for all water quality indicators (Table S4) especially for FC. Fall had statistically similar water quality to summer for four water quality indicators (BI, TN, TP, and SC) but had lower TDU and FC values as compared to summer. Finally, spring and winter had better water quality for all indicators except TDU (Table S4).

3.2. Multi-stressor HMLR model diagnostics

Multi-stressor HMLR models were all cross validated with a 3-fold approach to test for model stability and out-of-sample model performance. R^2 values for deterministic model predictions combined from all folds lowered the least for TN and TP (< 0.01), and the most for SC (0.13; Table S6). Predictor coefficients were generally stable and showed low variability with coefficient of variations across the three folds being less than 10%, except for crop (basin) and wetlands, which were both relatively uncommon in the Piedmont. Multi-stressor HMLRs were further analyzed by plotting predictions against observed values (Fig S3; S4). When only deterministic factors were used for predictions (Fig S3), model error in FC, TDU, TN, TP, and SC appeared to underpredict large spikes in the observed data. When random effects were used in the predictions (Fig S4), multi-stressor HMLR models were able to predict a broader range of observed data more accurately (both low and high), particularly for TN, TP, and TDU.

3.3. Water quality predictions across the UNRB

Water quality was predicted for 1423 stream reaches throughout the UNRB for current conditions (based on 2012 land covers) as well as forecasted for five potential management scenarios (see Section 2.5; Table 5). Reach-level predictions were aggregated as a mean (weighted by reach length) across the UNRB for each water quality indicator (log mean for FC, TDU, TN, TP, and SC; see 2.6) to compare management scenarios (Table 5). Restoring canopy in stream buffers (management scenario #1; Table 3) consistently improved water quality more than reducing the impact of IC (scenario #2) or WWTP (scenario #3) for all indicators except TDU. Improvements to water quality were nearly additive in nature such that the impact of scenario #4 (combining scenarios #1, #2, and #3), was similar in magnitude to the sum of the

individual scenarios (Table 5). Based on scenario #4, FC was the water quality indicator that showed the most potential for reduction across the UNRB while TDU showed little ability for improvement based on these scenarios. Management scenario #5, reducing the effect of positive site and basin random effects by 25%, led to the largest improvements in water quality for all indicators.

Results were summarized for three subsets of the UNRB that represented stream reaches with high levels of anthropogenic influence (i.e., urban (high IC), pasture lands, canopy loss in stream buffers). In general, urban and canopy loss areas in the UNRB had particularly elevated levels of BI, FC, and TP as compared to the whole UNRB and high pasture areas. Only slight differences were noted in TDU values for any subset, though high canopy loss areas had slightly higher TDU value. TN concentrations were nearly identical in predominantly pasture and urban areas but were highly elevated in stream reaches with excessive canopy loss. FC and SC were more elevated in high IC basins than both high pasture and canopy loss areas. For all subsets, the impact of increasing canopy cover 50% in buffers (scenario #1) was greater than mitigating 25% of the impact of IC (scenario #2), except for BI in high urban areas.

Weather events could significantly alter certain water quality indicators. To quantify the magnitude of this effect, predictions for current conditions and management scenario #4 (i.e., combination of scenarios #1, 2, and 3) were developed for two additional hydro-climatological conditions: dry conditions (10th percentile short and long-term conditions), and wet conditions (90% percentile short and long-term conditions). Mean TDU and FC values more than doubled during wet conditions (Table 5; Fig 2), while BI, TN, TP and SC all slightly improved (i.e., lowered). Conversely, during dry weather. BI, TN, TP, and SC increased, while TDU and FC were lowered. BI, TDU FC, and SC had similar water quality reductions for all management

scenarios and weather conditions (median, dry, wet), but reductions varied slightly for TN and TP where reductions for wet conditions were greater than for dry and median conditions.

Table 5: Summary of mean water quality condition in the Upper Neuse River Basin (UNRB) for current conditions and five potential management scenarios. Predictions for wet (90th percentile short and long-term precipitation) and dry conditions (10th percentile long-term and no short-term) were made for current conditions and management scenario #4. Predictions were also made for subsets of UNRB stream, defined by streams with the highest 10% impervious cover (IC), pasture lands, and stream buffer canopy loss in their basins.

	Management Scenario	BI (0-10)	TDU (NTU)	FC (#/100 mL)	TN (mg/L)	TP (mg/L)	SC (SU)
UNRB	**Current**	**5.99**	**9.5**	**76**	**1.36**	**0.20**	**108**
	Current (wet)	5.94	24.1	221	1.09	0.13	83
	Current (dry)	6.11	7.3	59	1.63	0.25	128
	#1	-5%	0%	-16%	-14%	-7%	-4%
	#2	-3%	-2%	-9%	0%	0%	-2%
	#3	0%	0%	0%	0%	0%	-1%
	#4	-7%	-2%	-23%	-13%	-8%	-6%
	#4 (wet)	-7%	-2%	-23%	-14%	-11%	-6%
	#4 (dry)	-7%	-2%	-23%	-12%	-6%	-6%
	#5	-10%	-16%	-46%	-32%	-38%	-21%
IC (top 10%)	**Current**	**7.27**	**9.3**	**230**	**1.36**	**0.28**	**130**
	Current (wet)	7.22	23.6	667	1.10	0.20	100
	Current (dry)	7.39	7.1	181	1.64	0.33	155
	#1	-4%	0%	-31%	-29%	-12%	-4%
	#2	-5%	-3%	-13%	0%	0%	-3%
	#3	0%	0%	0%	-1%	-1%	0%
	#4	-9%	-3%	-40%	-23%	-13%	-7%
	#4 (wet)	-9%	-3%	-40%	-25%	-16%	-7%
	#4 (dry)	-9%	-3%	-40%	-21%	-11%	-7%
	#5	-11%	-18%	-57%	-41%	-36%	-21%
Pasture (top 10%)	**Current**	**5.51**	**9.0**	**66**	**1.40**	**0.17**	**109**
	Current (wet)	5.46	22.9	194	1.13	0.10	84
	Current (dry)	5.63	6.9	52	1.68	0.22	130
	#1	-5%	0%	-14%	-12%	-7%	-4%
	#2	-1%	0%	-5%	0%	0%	-2%
	#3	0%	0%	0%	0%	0%	0%
	#4	-6%	0%	-18%	-12%	-7%	-6%
	#4 (wet)	-6%	0%	-18%	-13%	-11%	-6%
	#4 (dry)	-6%	0%	-18%	-11%	-6%	-6%
	#5	-9%	-16%	-42%	-33%	-43%	-20%

Table 5 (continued)

	Current	**7.19**	**10.2**	**204**	**1.67**	**0.29**	**120**
	Current (wet)	7.14	25.8	593	1.37	0.21	92
	Current (dry)	7.31	7.9	161	1.98	0.34	142
Canopy loss (top 10%)	#1	-4%	0%	-35%	-30%	-13%	-4%
	#2	-4%	-3%	-12%	0%	0%	-3%
	#3	0%	0%	0%	-1%	-1%	-1%
	#4	-9%	-3%	-43%	-25%	-14%	-7%
	#4 (wet)	-9%	-3%	-42%	-27%	-18%	-7%
	#4 (dry)	-8%	-3%	-43%	-24%	-12%	-7%
	#5	-10%	-16%	-59%	-39%	-36%	-21%

Certain management scenarios and water quality indicators showed little potential for water quality improvement. Reducing the impact of WWTPs (management scenario #3) showed no mean improvement for BI, FC, TDU, and TP, and only slight improvements for TN and SC (-2%, -1%; Table 5). This occurred, in part, because only 7% of UNRB stream reaches were located downstream of a WWTP, and most of these reaches had only minimal WWTP loadings. When reaches affected by more than a minimal WWTP load were tested [>0.1 kg TN/(day*km2); 60 reaches; Table S7], scenario #3 showed more reduction in TN, TP (-7% and -6%), but restoring buffer canopy cover still lowered nutrients more. Management scenario #2 (reducing the impact of IC) showed no ability to reduce mean TN and TP concentrations. This was expected, because the multi-stressor model for TN did not include any direct positive relationships with IC, and the TP model included only recent IC. Recent IC is not highly prevalent throughout the UNRB with a median value of 0.13% and a 95th percentile of 2.1% of total watershed area. TDU showed the least amount of improvement as a result of scenario #4 with only a 2-3% improvement. TDU had only one anthropogenic predictor in its multi-stressor HMLR (recent IC; Table 4).

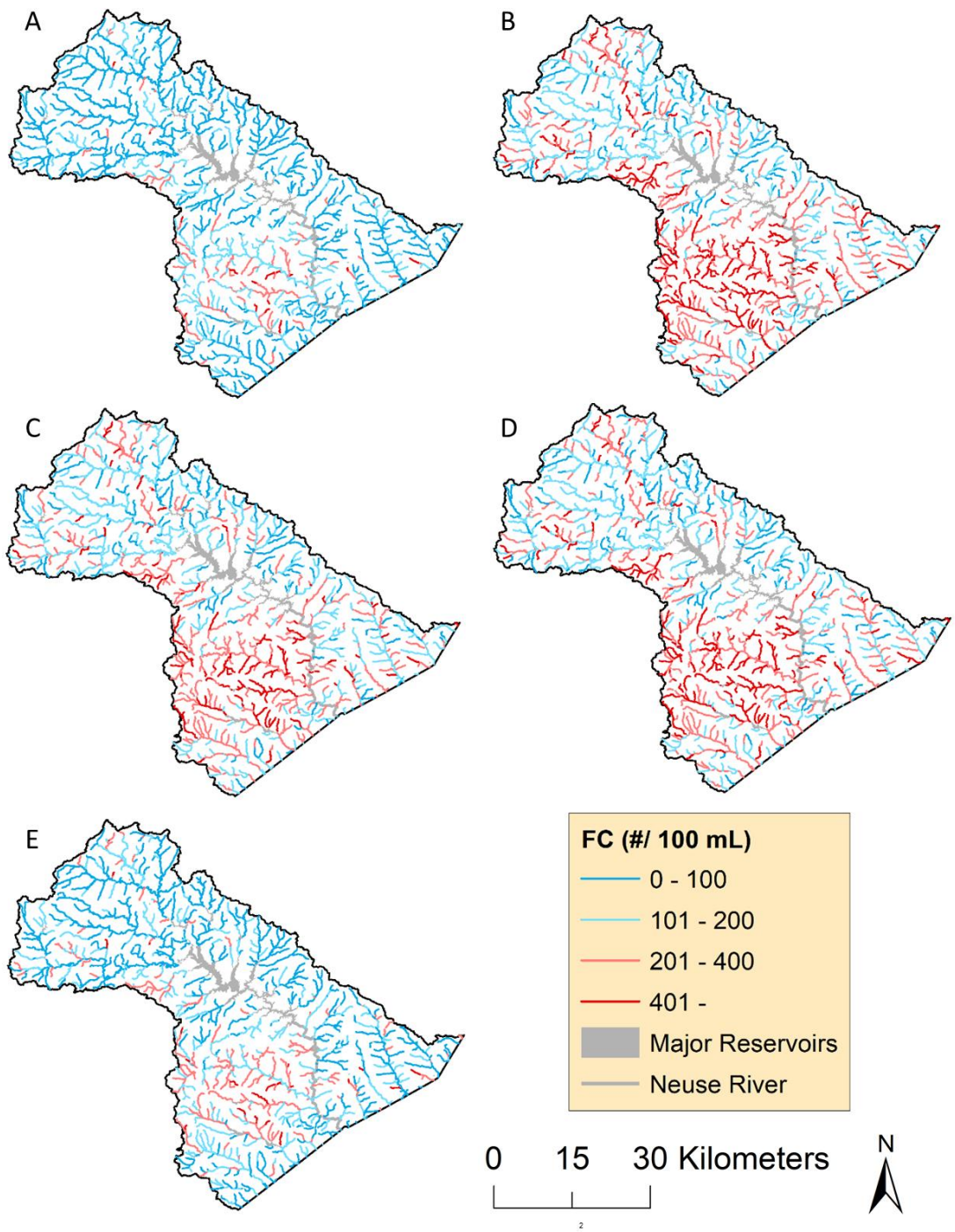


Figure 2: Mean fecal coliform (FC) predictions throughout the Upper Neuse River Basin (UNRB) for current median (A) and wet (B) hydro-climatological conditions. Management scenario #1 (C; 50% reduction in canopy loss), #2 (D; 25% reduction in the effect of impervious cover), #4 (E; combination of scenarios #1, #2, and #3 (25% reduction of WWTP discharges) are shown for wet weather conditions. The main stem of the Neuse River (gray) was considered unwadeable and excluded for predictions

The uncertainty associated with forecasted water quality improvements was explored by taking into account model uncertainty (fixed effects and model error) as well as long and short-term weather uncertainty. Based on model error and uncertainty in fixed effects, BI was the indicator where an improvement was most likely to be observed based on management scenario #4 (100% that the mean of 6 post-restoration samples was lower than the mean of 6 pre-restoration samples), followed by TN and FC (90% after 24 samples), and SC (80% after 36 samples; Fig 3). In general, increased model residual error led to less certainty in forecasted improvements. Therefore, the order of observing differences in mean pre and post-restoration sampling for indicators was similar to the order of R^2 for deterministic effects of multi-stressor HMLRs (Table 4), except for FC. Though the FC model had the lowest model skill (R^2 =0.25; Table 4), it still showed a high probability of observing improvements, probably because it had anthropogenic predictors (canopy loss, pre-2000 IC) that were of importance as well as highly prevalent in the UNRB. Other low skill models, TDU and TP, were driven by anthropogenic factors that were less common (e.g., recent IC, WWTP loadings) or land covers not affected by management scenarios (i.e., crop). TDU achieved only a 55% chance of post-restoration sampling having an improved mean value, only a slight improvement from 50%, the chance of observing a difference if no restoration efforts took place (Fig. 3). If a more stringent test of statistical difference (i.e., t-test) in pre and post-restoration sampling was used, a mean difference in BI scores was likely to be seen after 24 samples (> 90%), and it was the only water quality indicator to achieve greater than a 50% chance of having an observed improvement for mean UNRB sites (Fig 3, bottom left). Statistically significant improvements in TN and FC would be expected to be observed 50% of the time in highly urban sites, but only after 60 sampling events (i.e., 5 years of monthly sampling). Improvements to TDU, SC, and TP would have less than a

20% chance of showing a statistically significant difference even after 60 samples. When variability due to uncertain short and long-term weather conditions were added to simulations, little difference was noted, even for indicators that change greatly in wet weather (Fig S5). This probably occurred because uncertainty introduced by weather was smaller in magnitude than model uncertainty.

3.4. Basin and site random effects

Basin random effects were used to group sampling locations from the same basin that were hydrologically related, while site random effects were used to account for intraclass correlation of repeated sampling at a given location. For all water quality indicators, site and basin random effects contributed greatly to the response (Table 4). The normalized contribution of basin and site random effects were similar in magnitude in all models except FC, which had no basin effect implying that there was no significant correlation among FC values at locations within the same basin. On the other hand, TP had the largest contribution from the basin random effect of all water quality indicators (Table 4) implying that TP values were likely to be correlated at sites within the same basin.

The magnitude of basin and site random effects can be understood based on their normal hyperdistributions (Table S3). To quantify the magnitude of these hyperdistributions (i.e., effect of site and basin effects), they were used to create 95% confidence intervals for different ranges of each water quality indicator within expected predicted values [low (20th percentile), average (50th), or high (80th); Table 6]. These 95% confidence intervals were constant for differing levels of BI (low, average, high values), because BI was not log transformed. For all other indicators, these intervals increased for higher predicted values. Therefore, at highly degraded sites, more variability should be expected as a result of random effects.

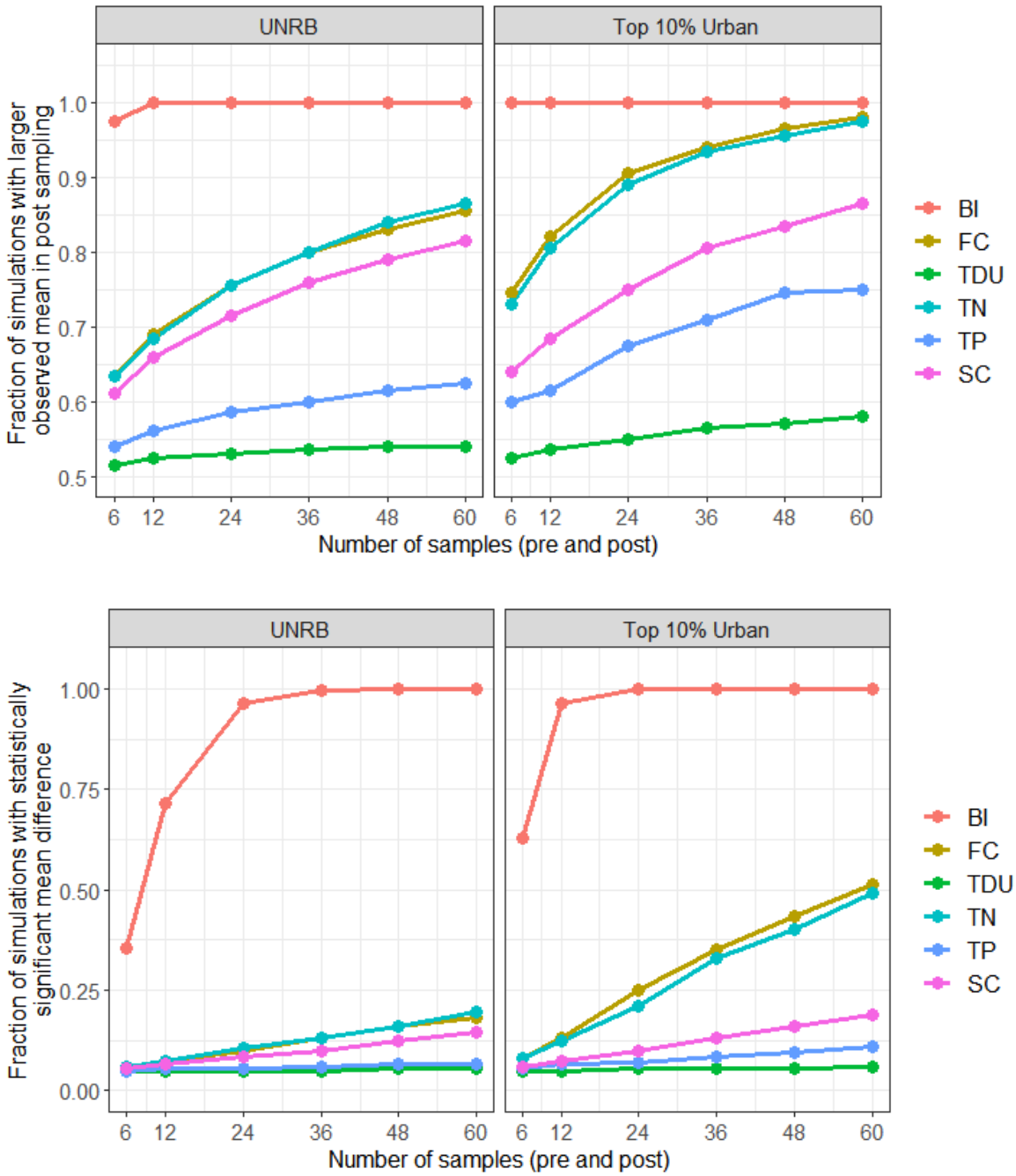


Figure 3: Chance that post-restoration sampling (management scenario #4) will have an improved (i.e., lower) mean compared to pre-restoration samples (current conditions) for all water quality indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC). Results for the entire UNRB (top left) and the top 10% urban sites (top right) are shown. The fraction of simulations that led to statistically significant differences between pre and post sampling are shown below (bottom).

Table 6: Expected variation in water quality predictions due to site and basin random effects in HMLR models. 95% confidence intervals, shown in parentheses, are based on the variance of the site and basin hyperdistributions and are shown for all water quality indicators for reaches with low (20th percentile), average (50th), or high (80th) predicted values of that indicator throughout the UNRB.

	Units	Low (20th)	Mean	High (80th)
BI	-	4.87 (3.89,5.85)	6.08 (5.10,7.06)	7.10 (6.12,8.08)
FC	#/100 mL	50 (20,124)	170 (69,420)	600 (243,1477)
TDU	NTU	4.5 (2.9,6.7)	9.8 (6.9,14.3)	23.0 (16.2,32.5)
TN	mg/L	0.61 (0.18,1.20)	1.19 (0.60,2.00)	3.82 (2.53,5.60)
TP	mg/L	0.05 (0,0.18)	0.12 (0,0.26)	0.40 (0.24,0.58)
SC	SU	93 (62,140)	149 (99,223)	260 (173,390)

4. Discussion

The goals of this research were to identify key natural and anthropogenic drivers of six water quality indicators and quantify the potential impact of future management strategies on a regional stream network (UNRB). Hierarchical multiple linear regression models were constructed using candidate predictors from 16 variable classes at distinct temporal and spatial scales. Site and basin random effects were included to account for repeatedly sampled locations and correlations among sites within the same basins. Ours is the first study to model multiple facets of water quality within a hierarchical framework including an unprecedented variety of temporal varying predictor data (e.g., land covers, hydro-climatology, WWTP loadings). Models were then used to make regional forecasts for future management scenarios. Using 17 years of data, we incorporated temporally specific data that allowed us to compare the effect of recent construction as well as older vs. newer IC in addition to short and long-term hydro-climatological conditions. Point source data (upstream reservoirs and WWTP loads) were also instrumental for model construction and forecasting water quality indicators throughout the UNRB.

This study contrasts with recent large-scale predictive studies that used machine learning techniques to estimate fish populations, baseline water quality conditions for specific conductance and nutrient levels, and current conditions for stream ecosystems across large regions (Endries, 2011; Olson and Hawkins, 2012; Olson and Hawkins, 2013; Hill et al., 2017). These approaches offer the ability to include correlated predictor data (e.g., land covers), account for non-linear relationships and interactions between predictors, and have good predictive ability (Evans and Cushman, 2009; Evans et al., 2011), but they cannot easily answer questions about the magnitude of relationships between predictors and response variables, offer confidence intervals for predictions (Mentch et al. 2016), or account for grouped data in a rigorous statistical framework as in hierarchical modeling. In addition, previous studies, focused specifically on data collected for a set time period not accounting for changes to water quality, land covers, or point sources over long time periods.

4.1. Anthropogenic and natural variable classes

Upstream land covers were found to be related to all water quality indicators in multi-stressor HMLR models. The most dominant land cover associated with degradation of water quality was some form of IC (basin, buffer, recent, age) for all indicators except TN (Table 4). IC, a common surrogate for urbanization, had its largest effect on BI, where watershed IC alone could explain 65% of the variability, consistent with previous studies (Schueler 1994; Booth et al. 1997; Schueler et al. 2010; Miller et al. In Review). Specifically, older watershed IC (pre-1980, pre-2000), had the largest effect of any land cover on BI, SC, and FC. This effect for BI might be due to development that occurred without sediment controls in place (Miller et al. In Review) while elevated SC and FC levels can possibly be attributable to leaky wastewater infrastructure (Lerner 2002; O'Driscoll et al., 2010). In addition, recent IC (previous 2 or 5

years), had the largest impact of any land cover on TDU and TP, likely because recent construction increases sediment discharge into streams (Finkenbine et al., 2000; Sutherland et al., 2002) and phosphorous is often bound to sediments (Chapra 2008). Canopy loss in 30-m stream buffers was the second most impactful land cover variable and affected all indicators except TDU (Table 4). The ability of stream buffers to filter nutrients in surface and sub-surface flows has been well-established (Daniels and Gilliam, 1996; Lowrance, 1997; Naiman and Decamp, 1997) and supports why canopy loss was the land cover variable most highly associated with increased TN concentrations. Canopy buffer loss was elevated in the UNRB in pasture and urban areas, both of which had similar mean TN values (1.40, 1.36 mg/L; Table 5). In addition, canopy loss was the second largest anthropogenic source of BI degradation, reaffirming the importance of stream buffers to downstream ecological health (Richards et al. 1996, Sutherland et al. 2002). Crop (buffer or basin) was also selected for four water quality models. It was positively associated with TDU and TN values, probably due to sediment runoff and fertilizer application (Beaulac and Reckow, 1982; Richards et al., 1996; Miller et al., 1997), while it was associated with lower FC and SC values. Lower FC values were expected in croplands due to a lack of fecal sources, but SC usually rises with increased fertilizer applications that increase major ions including calcium, magnesium, and nitrate (Johnston et al. 1990; Hamilton and Helsel, 1995). Wetlands were associated with lower TN concentrations in HMLRs, but this effect was relatively small (Table 4). Though wetlands have been shown to have the potential for denitrification (Seitzinger 1994), the capacity of wetlands to mitigate nitrogen can be limited if upstream flows pass too quickly through them (Bowden 1987). Finally, pasturelands in stream buffers were positively associated with increased FC, most likely attributable to the runoff of livestock waste.

The impact that WWTPs had on downstream water quality varied by indicator. TN and TP were highly affected by WWTP loadings, which had the two highest predictor contributions in the study (0.66, 0.47; Table 4). SC was most affected by the number of upstream major WWTPs, a count variable that did not quantify actual SC discharging from WWTPs. WWTP loading information for total dissolved solids (or SC) might have been useful for predicting SC but was largely unavailable for the study area. Model fit for SC when basin and site random effects were included, rose significantly (R^2 =0.70; Table 4), implying that large basin and site effects (not captured by the available predictors) were present. BI scores were greatly increased (i.e., negative impacted) at sites within 1 km of major WWTPs (Table S3), but this effect diminished downstream relatively quickly (Miller et al. In review). This effect might be due to altered bacterial communities found near large WWTPs (Wakelin et al. 2008), possibly associated with chlorine residuals (Brungs 1973, Krasner et al. 2009) or other pollutants, that can create shifts in food webs and macroinvertebrate communities (Wallace et al. 1997, Webster et al. 1999). FC was also associated with increased TN loadings normalized by drainage area (a proxy used to represent WWTP loadings). The inclusion of this variable implies that unsterilized effluent is being discharged from certain WWTPs. In sum, these results imply that WWTPs are major sources for increased nutrients and SC, minor sources of FC, and have large but spatially limited effects on stream ecosystems (i.e., BI).

Natural factors affecting water quality indicators were hydro-climatology, soil type, seasonal differences, and drainage area size. By including temporally varying data, we were able to determine the most influential time scales for hydro-climatological factors. Short-term precipitation (3 days) was the main driver of both TDU and FC as runoff events wash sediments and fecal matter into streams. Without spatially precise and temporally specific precipitation

data, TDU and FC model performance (i.e., R^2) would have lowered significantly by -57% and -30%, respectively. Short-term precipitation effects on TN, TP, and SC were less impactful and similar in magnitude to the impact of long-term precipitation for these indicators. BI was the only indicator not affected by short-term precipitation, a possible reason why macroinvertebrate community metrics are often considered the best way to evaluate stream water quality (USEPA 2016). Soil type had its largest effect on SC and BI, while it did not affect TN and TP, implying that natural sources of nutrients were constant throughout the Piedmont. SC levels are closely connected to underlying geology (Olson and Hawkins 2012) and the most elevated region in the Piedmont based on model results, the Carolina Slate Belt (Table S5), has also been shown to have elevated SC in its groundwater as well (Harden et al. 2009). BI is highly elevated in Triassic basins (+0.65; Table S4), and this confirms why NC does not bioclassify macroinvertebrate samples for this region (NC DQ 2016). Smith River Allochthon and Inner Piedmont soils were regions that had below average soil random effects for BI, FC, and SC, both of which are located on the western edge of the Piedmont (Fig S1), implying better background water quality in these regions. Seasonality affected all six water quality indicators implying that water quality is lower in summer for all indicators, except TDU (Table S4). Increased summer temperatures would cause elevated evapotranspiration and reduce long-term recharge to streams, which would lead to increased BI, TN, TP, and SC values (Table 4). Also, though decay rates for FC can be elevated for increased temperatures (> 30° C; de Brauwere et al., 2014), bacterial growth in contaminated riparian buffer soils and stream sediments, both of which increase FC during precipitation events, are higher during summer months (Whitman et al., 2006). Finally, larger drainage basins led to lower FC values, probably attributable to the settling of fecal matter from

upstream sources into stream sediments where they decay (Wilkinson et al. 1995; Rehmann and Soupir, 2009).

4.2. Impact and feasibility of management scenarios

Management scenarios #1-4 represented mitigation of the impact of anthropogenic land covers and WWTP point source loadings on streams. Increasing canopy cover in riparian stream buffers by 50% (scenario #1) had the most consistent improvement for all water quality indicators (Table 5). Though many factors connected to water quality are out of managers control, potential for increased canopy cover in highly urban areas has been shown to be substantial (Wilson and Lindsey 2009; McPherson et al. 2010). The importance of riparian buffers has also been shown in past studies related to BI (Richards et al. 1996; Miltner et al. 2004), increased sediments (Sutherland et al. 2002), and nutrient export (Lowrance, 1997; Naiman and Decamp, 1997).

Mitigating the effect of IC (scenario #2) led to the largest reductions for FC and smaller reductions for BI, TDU, and SC. In general, scenario #2 was less effective than restoring buffer canopy cover (scenario #1) for improving mean water quality in all subsets of the UNRB (Table 4). The impact of IC on water quality can be reduced by installing low impact development (LID) or updating stormwater control measures and these techniques have been shown to reduce runoff (20-90%), nutrient export (30-70%), and E. coli and FC (up to 90%) in small scale projects (Eckart et al. 2017). Though reductions in small-scale projects have been encouraging, basin-wide reductions for large scale projects are not expected to reach the same levels of removals (Hunt et al., 2008, Hathaway et al., 2009) mainly due to the lack of ability to retrofit all areas in a basin. IC (recent, age, basin, buffer) was a more impactful predictor for BI, FC, TDU, TP and SC (Table 4), yet restoring canopy loss led to a larger mean reduction for water quality

indicator even in stream reaches tha t drained the most urban areas in the UNRB (Table 5). This apparent contradiction occurred for two reasons: several of the IC predictors (important for TDU and TP) are not highly prevalent in the UNRB and the potential for canopy restoration was assumed to be larger (50% increase) than mitigating the effect of IC (25% reduction; Table 3). More canopy restoration was assumed to occur for several reasons: 1) a 50% reduction of buffer canopy loss is slightly less in actual restored area (mean for UNRB = 0.69 km^2) as compared to mitigating 25% of IC (mean= 0.89 km^2), 2) planting trees is cheaper than installing LID or retrofitting SCM, and 3) land near streams are more likely to be available for restoration activities as they might be in flood zones. In addition, restoring buffer canopy loss has the potential to impact both highly urban and pasture lands in the UNRB leading to its greater influence in mean predictions. Though both reductions (50% for canopy restoration, 25% for mitigating IC) might seem difficult to accomplish in the near future, management scenarios were designed to test large-scale plans that might take decades to accomplish. The true potential for restoring canopy and mitigating IC in large watersheds is unknown and might vary by region.

Management scenario #3 had almost no impact on mean UNRB water quality, partly due to the lack of WWTP effluent in the majority of UNRB streams (only 7% of stream reaches). This scenario did lower TN, TP, and SC values in the 60 stream reaches that had WWTP effluent in them (7%., 6%, and 11%, respectively; Table S7). Most major WWTPs in the NC Piedmont (Table S1) are currently discharging at mean tertiary levels for TN and TP concentrations (5.5 mg/L, 1.5 mg/L), while many minor WWTPs continue to exceed mean secondary treatment levels (20 mg/L, 5 mg/L; Tchobanoglous et al. 2003; Carey and Migliaccio, 2009). More efficient operations as well as more recent technologies such as microfiltration (Carey and Migliaccio, 2009) could lower nutrient and SC discharges, but the cost of these improvements

would have to be weighed against the influence of WWTPs on downstream water quality. Though there is more potential to upgrade minor WWTP performance in the UNRB, they had no significant effects on BI, TDU, or SC and had very limited effects on FC, TN, and TP (Table 4) due to their low discharges.

Reducing positive basin and site random effects (scenario #5), represented lowering water quality degradation not explained by deterministic predictors in the multi-stressor HMLRs. This scenario led to even higher reductions than scenario #4 (the combination of scenarios #1, 2, 3) for all water quality indicators (Table 5). This result implies that identifying and mitigating site-specific causes of water quality variation not included in the deterministic components of these models, has the potential to lower water quality more than the combination of restoring canopy, mitigating IC, and reducing WWTP impacts though additional site investigation would be required to achieve this.

4.3. Uncertainty related to water quality predictions

The ability to judge past stream restoration efforts has been limited by a lack of monitoring data which is often not available to compare pre and post-restoration conditions (Palmer et al. 2014). In order to explore uncertainty in observing water quality reductions related to potential future management scenarios, predictions for pre (current conditions) and post-monitoring (management scenario #4) samples were simulated for all water quality indicators. Reductions in post-restoration mean BI values (as compared to pre-restoration) were expected to be observed with the fewest number of samples (95% of the time for > 6 samples; Fig. 3) followed by TN, FC, and SC (80% for > 36 samples). Mean TP and TDU showed little chance for observing an improvement due to the high stochasticity of these indicators within UNRB streams (i.e., high model error). Also, the most important anthropogenic predictors for TP and

TDU that could be improved during management scenarios were not widespread in the UNRB (i.e., recent IC, TP loading). In contrast, FC had low model skill, but improvements were likely to be seen due to the strong influence of pre-2000 IC and canopy buffer loss on FC and their abundance throughout the UNRB (Table 4). If more stringent reductions (i.e., statistically significant differences in mean values of pre and post-sampling) were required, these probabilities would lower significantly for all indicators, except BI (Fig. 3 bottom).

Additional model uncertainty, not quantified in this analysis, relate to limitations inherent in using statistical inference, time lags associated with improvements in water quality, and interactions that might be occurring between predictors. Correlation is not causation and the relationship of predictors to response variables in statistical frameworks cannot be assumed to be direct causes. Even though much research has explored the mechanistic connections that occur between land cover predictors in this study (IC, buffer canopy loss, crop, wetlands) and water quality (Seitzinger 1994; Lowrance, 1997; Naiman and Decamp, 1997; Walsh et al. 2005), IC and canopy buffer loss are still proxies for urbanization and/or agricultural areas and the various forms of degradation caused by them. Simply restoring canopy cover in urban or pasturelands, might not be able to mitigate other sources of pollutants that are found concurrently near buffer canopy loss areas or reverse the effects of other factors related to the initial removal of canopy cover or addition of IC (e.g., soil impaction). Post-restoration sampling is also predicted assuming that all improvements to water quality indicators have already occurred. Though mitigating some predictors, like WWTP nutrient loadings, might immediately lower in-stream water quality values, stream ecology (i.e., BI) has been shown to return slowly, if at all, to previous conditions after restoration efforts (Spanhoff and Arle 2007), while restoring canopy cover would take decades to achieve. Finally, interactions between predictors were not

considered in this study. Buffer canopy loss and watershed IC were correlated in the UNRB (r = 0.62), and the mitigation of canopy loss in high IC areas might have a different effect than in lower IC or pasture areas. Also, mitigating IC and restoring buffer canopy might have an accelerated (or reduced) effect on water quality improvement. Though statistical models have limits and there are uncertainties inherent in forecasting water quality, our results give some guidance to how much sampling needs to be done in order to observe mean differences, and what expected post-restoration conditions might be look like if past land use effects are effectively mitigated. Changes to BI should be observable (> 95%) within 6 samples (i.e., 6 years) for mean UNRB stream, while it would take 36-48 samples (i.e., 3-4 years of monthly sampling) to have an 80% chance of observing changes to FC, TN and SC. Even after 60 samples, TP and TDU will likely show low improvement basin wide, though results for specific reaches would vary.

4.4. Site and basin random effects

A hierarchical framework was used in this study to account for intra-class correlation of repeatedly sampled data (site random effects) and to quantify correlation between sites that were hydraulically connected (basin random effects). For BI, TN, TP, SC and TDU, site variability was similar in magnitude (Table 4) to variability across basins. This implies that though there is correlation among sites within a given basin, a large amount of variability still exists among those sites. Basin variability for TN, TP, and SC might reflect variation in WWTP discharges not accounted for by mean discharge data or unaccounted for septic overflows while BI and TDU basin effects could be related to varying soil conditions in upstream watersheds that lead to varying levels of runoff and/or sediment and pollutant transport. FC was the only indicator to not have a basin random effect (Table 4; S3), implying that there is little spatial correlation of FC values in streams, probably due to particularly site-specific sources of FC along stream networks

and the quick removal of FC from the water column due to settling and other natural factors (Wilkinson et al. 1995; Rehmann and Soupir, 2009). Spatial correlation was not modeled in detail, in contrast to previous work (spatial stream networks; Hoef et al. 2010), as the large number of observations used for water quality indicators (Table 1) would have made it computationally intractable (> 1000 observations; Hoef et al. 2014). Our observed water quality data were also not temporally compatible (i.e., not collected at the same time), and spatial correlations between sampling sites were expected to be limited, because the majority of sites were not hydraulically connected (Table 1).

Though we used random effects in this study to help parameterize our model, explore regional soil differences, and quantify the accuracy of deterministic predictions; water quality managers could use them to prioritize inspections of specific basins or sites that deviate from mean conditions. In addition, analyzing deterministic predictions of water quality indicators along with their site and basin random effects can help local managers assess which sites might be ideal for restoration activities (Miller et al. In review). Variation in observing water quality differences in pre and post-restoration sites was high among stream reaches (Fig S5), which highlights the importance of choosing ideal restoration sites.

4.5. Summary

This study identified major natural and anthropogenic drivers of water quality indicators (BI, FC, TDU, TN, TP, and SC) and predicted their values throughout a regional stream network (UNRB). Water quality models show that BI is mostly affected by upstream land covers (IC and canopy loss in stream buffers), FC and TDU are driven by short-term precipitation, and TN, TP, and SC variability is best explained by WWTP loadings. Temporal variability in IC showed that

recent IC affected TDU and TP, while older IC (e.g., pre-1980) increased BI, FC, and SC. BI was the water quality indicator that was most accurately predicted using deterministic factors (R^2=0.76) and showed the highest probability of observable differences due to management scenarios. Five potential management scenarios were forecasted for the UNRB and compared to current conditions under different hydro-climatological conditions (dry, median, and wet conditions). Restoring canopy cover in stream buffers was the most impactful scenario and improved water quality for all indicators except TDU. Identifying and mitigating site-specific drivers of water quality improved water quality more than mitigating upstream land covers and point sources. Uncertainty in water quality forecasts for the UNRB was assessed and helped identify that BI, TN and FC would be the water quality indicators most likely to show observable differences after restoration occurred.

There is much value in studying long-term effects on stream ecosystems (Jackson and Fureder, 2006), but a lack of methodological tools and limited time series of predictor data (i.e., land covers) are often obstacles to analyzing long-term regional datasets. This study shows the ability of a hierarchical approach to model complex multi-decadal datasets in a computationally tractable way that offer insight into sources of water quality alteration as well as identification of potential sites or basins that are particularly elevated. Regular and intensive data collection is currently practiced by most environmental agencies, and modeling techniques that allow the inclusion of all available data (e.g., historical data, spatially correlated data) and synthesize these datasets in meaningful ways are needed in order to help inform local and state management decisions as well as bolster future research efforts.

References

Alexander, R. B., Johnes, P. J., Boyer, E. W., & Smith, R. A. 2002. A comparison of models for estimating the riverine export of nitrogen from large watersheds. *Biogeochemistry*, *57*(1), 295-339.

Alexander, R. B., Smith, R. A., Schwarz, G. E., Boyer, E. W., Nolan, J. V., & Brakebill, J. W. 2007. Differences in phosphorus and nitrogen delivery to the Gulf of Mexico from the Mississippi River Basin. *Environmental science & technology*, *42*(3), 822-830.

Allan, J. D. 2004. Landscapes and riverscapes: the influence of land use on stream ecosy stems. *Annu. Rev. Ecol. Evol. Syst.*, *35*, 257-284.

Beaulac, M. N., & Reckhow, K. H. 1982. An Examination of Land Use-Nutrient Export Relationships. *JAWRA Journal of the American Water Resources Association*, *18*(6), 1013-1024.

Bernhardt, E. S., Palmer, M. A., Allan, J. D., Alexander, G., Barnas, K., Brooks, S., ... & D. Galat 2005. Synthesizing US river restoration efforts. *Science*, *308*(5722), 636-637.

Bernhardt, E. S., & Palmer, M. A. 2007. Restoring streams in an urbanizing world. *Freshwater Biology*, *52*(4), 738-751.

Bellucci, C. J., Becker, M. E., Beauchene, M., & Dunbar, L. 2013. Classifying the health of Connecticut streams using benthic macroinvertebrates with implications for water management. *Environmental management*, *51*(6), 1274-1283.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, *24*(3), 127-135.

Booth, D. B., & Jackson, C. R. 1997. Urbanization of Aquatic Systems: Degradation Thresholds,

Stormwater Detection, and the Limits of Mitigation. *JAWRA Journal of the American Water Resources Association*, *33*(5), 1077-1090.

Bowden, W. B. 1987. The biogeochemistry of nitrogen in freshwater wetlands. *Biogeochemistry*, *4*(3), 313-348.

Breiman, L. 2001. Random forests. *Machine learning*, *45*(1), 5-32.

Carey, R. O., & Migliaccio, K. W. 2009. Contribution of wastewater treatment plant effluents to nutrient dynamics in aquatic systems: a review. *Environmental Management*, *44*(2), 205-217.

Chapra, S. C. 2008. *Surface water-quality modeling*. Waveland press.

Chatfield, C. 2006. *Model uncertainty*. *Encyclopedia of Environmetrics*, John Wiley, Hoboken, N.J.

Clark, J. S., & Gelfand, A. E. 2006. A future for models and data in environmental science. *Trends in Ecology & evolution*, *21*(7), 375-380.

Cuffney, T. F., Kashuba, R., Qian, S. S., Alameddine, I., Cha, Y. K., Lee, B., ... & McMahon, G. 2011. Multilevel regression models describing regional patterns of invertebrate and algal responses to urbanization across the USA. *Journal of the North American Benthological Society*, *30*(3), 797-819.

Daniels, R. B., & Gilliam, J. W. 1996. Sediment and chemical load reduction by grass and riparian filters. *Soil Science Society of America Journal*, *60*(1), 246-251.

Davison, A. C., Hinkley, D. V., & Young, G. A. 2003. Recent developments in bootstrap methodology. *Statistical Science*, 141-157.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., ... & Sullivan, C. A. 2006. Freshwater biodiversity: importance, threats, status and

conservation challenges. *Biological reviews*, *81*(2), 163-182.

Eckart, K., McPhee, Z., & Bolisetti, T. 2017. Performance and implementation of low impact development–A review. *Science of the Total Environment*, *607*, 413-432.

Efron, B. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569-593). Springer, New York, NY.

Elsner, J. B., and C. P. Schmertmann. 1994. Assessing forecast skill through cross validation. *Weather and Forecasting*, *9*(4): 619-624.

Evans, J. S., & Cushman, S. A. 2009. Gradient modeling of conifer species using random forests. *Landscape ecology*, *24*(5), 673-683.

Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. 2011. Modeling species distribution and change using random forest. In *Predictive species and habitat modeling in landscape ecology* (pp. 139-159). Springer, New York, NY.

Finkenbine, J. K., Atwater, J. W., & Mavinic, D. S. 2000. Stream health after urbanization. *JAWRA Journal of the American Water Resources Association*, *36*(5), 1149-1160.

Gelman, A., & Hill, J. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 2014. *Bayesian data analysis* (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Gentry-Shields, J., Rowny, J. G., & Stewart, J. R. 2012. HuBac and nifH source tracking markers display a relationship to land use but not rainfall. *Water research*, *46*(18), 6163-6174.

Hathaway, J. M., Hunt, W. F., & Jadlocki, S. 2009. Indicator bacteria removal in storm-water best management practices in Charlotte, North Carolina. *Journal of Environmental*

Engineering, *135*(12), 1275-1285.

Hamilton, P. A., & Helsel, D. R. 1995. Effects of agriculture on ground-water quality in five regions of the United States. *Groundwater*, *33*(2), 217-226.

Harden, S. L., Chapman, M. J., & Harned, D. A. 2009. *Characterization of Groundwater Quality Based on Regional Geologic Setting in the Piedmont and Blue Ridge Physiographic Provinces, North Carolina*. US Geological Survey.

Hering, D., Johnson, R. K., Kramm, S., Schmutz, S., Szoszkiewicz, K., & Verdonschot, P. F. 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. *Freshwater Biology*, *51*(9), 1757-1785.

Herlihy, A. T., & Sifneos, J. C. 2008. Developing nutrient criteria and classification schemes for wadeable streams in the conterminous US. *Journal of the North American Benthological Society*, *27*(4), 932-948.

Hill, Ryan A., Marc H. Weber, Scott G. Leibowitz, Anthony R. Olsen, and Darren J. Thornbrugh, 2016. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. Journal of the American Water Resources Association (JAWRA) 52:120-128. DOI: 10.1111/1752-1688.12372.

Hill, R. A., Fox, E. W., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. 2017. Predictive mapping of the biotic condition of conterminous US rivers and streams. *Ecological Applications*, *27*(8), 2397-2415.

Hilsenhoff, W. L. 1988. Rapid field assessment of organic pollution with a family-level biotic index. *Journal of the North American Benthological Society*, *7*(1): 65-68.

Hoef, J. M., & Peterson, E. E. 2010. A moving average approach for spatial statistical models of

stream networks. *Journal of the American Statistical Association*, *105*(489), 6-18.

Hoef, J. V., Peterson, E., Clifford, D., & Shah, R. 2014. SSN: An R package for spatial statistical modeling on stream networks. *Journal of Statistical Software*, *56*(3), 1-45.

Howarth, R., Swaney, D., Billen, G., Garnier, J., Hong, B., Humborg, C., ... & Marino, R. 2012. Nitrogen fluxes from the landscape are controlled by net anthropogenic nitrogen inputs and by climate. *Frontiers in Ecology and the Environment*, *10*(1), 37-43.

Kelsey, H., Porter, D. E., Scott, G., Neet, M., & White, D. 2004. Using geographic information systems and regression analysis to evaluate relationships between land use and fecal coliform bacterial pollution. *Journal of experimental marine biology and ecology*, *298*(2), 197-209.

Hunt, W. F., Smith, J. T., Jadlocki, S. J., Hathaway, J. M., & Eubanks, P. R. 2008. Pollutant removal and peak flow mitigation by a bioretention cell in urban Charlotte, NC. *Journal of Environmental Engineering*, *134*(5), 403-408.

Jackson, J. K., & Fuereder, L. 2006. Long-term studies of freshwater macroinvertebrates: a review of the frequency, duration and ecological significance. *Freshwater biology*, *51*(3), 591-603.

Johnston, C. A., Detenbeck, N. E., & Niemi, G. J. 1990. The cumulative effect of wetlands on stream water quality and quantity. A landscape approach. *Biogeochemistry*, *10*(2), 105-141.

Lerner, D. N. 2002. Identifying and quantifying urban recharge: a review. *Hydrogeology journal*, *10*(1), 143-152.

Lowrance, Richard, et al. 1997."Water quality functions of riparian forest buffers in Chesapeake Bay watersheds." *Environmental Management* 21.5 1997: 687-712.

Malmqvist, B., & Rundle, S. 2002. Threats to the running water ecosystems of the

world. *Environmental conservation*, *29*(2), 134-153.

Matthews, J., Reeze, B., Feld, C. K., & Hendriks, A. J. 2010. Lessons from practice: assessing early progress and success in river rehabilitation. *Hydrobiologia*, 655(1), 1-14.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., & Nijssen, B. 2002. A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *Journal of climate*, *15*(22), 3237-3251.

Mays, L. W. (2010). *Water resources engineering*. John Wiley & Sons.

McPherson, E. G., Simpson, J. R., Xiao, Q., & Wu, C. 2011. Million trees Los Angeles canopy cover and benefit assessment. *Landscape and Urban Planning*, *99*(1), 40-50.

Mentch, L., & Hooker, G. 2016. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, *17*(1), 841-881.

Miller, C. V., Denis, J. M., Ator, S. W., & Brakebill, J. W. 1997. Nutrients in Streams during Baseflow in selected Environmental Sevings of the Potomac River Basin. *JAWRA Journal of the American Water Resources Association*, *33*(6), 1155-1171.

Miller, J., Esselman, P. C., Alameddine, I., Blackhart, K., & Obenour, D. R. 2018. Hierarchical modeling assessment of the influence of watershed stressors on fish and invertebrate species in Gulf of Mexico estuaries. *Ecological Indicators*, *90*, 142-153.

Miller, J., Paul, M. J., & Obenour, D. R. (In review). Assessing Natural and Anthropogenic Drivers of Ecological Health in Piedmont Streams using Hierarchical Modeling.

Miltner, R. J., White, D., and C. Yoder. 2004. The biotic integrity of streams in urban and suburbanizing landscapes. *Landscape and urban planning*, *69*(1): 87-100.

Moore, R. B., and T. G. Dewald. 2016. The Road to NHDPlus—Advancements in Digital Stream Networks and Associated Catchments. *JAWRA Journal of the American Water*

Resources Association, *52*(4): 890-900.

Mulholland, P. J., and D. R. Lenat. 1992. Streams of the southeastern Piedmont, Atlantic drainage. *Biodiversity of the Southeastern United States—Aquatic Communities*, 193-232.

Naiman, R. J., & Decamps, H. 1997. The ecology of interfaces: riparian zones. *Annual review of Ecology and Systematics*, *28*(1), 621-658.

NC DEQ. 2016. Standard Operating Procedures for the Collection and Analysis of Benthic Macroinvertebrates. Division of Water Resources. Raleigh, North Carolina. February 2016.

NC DEQ. 2018a. Falls Lake Nutrient Strategy. https://deq.nc.gov/about/divisions/water-resources/water-planning/nonpoint-source-planning/falls-lake-nutrient-strategy#Background. Accessed November 12th, 2018.

NC DEQ. 2018b. Ambient Monitoring System (AMS). https://deq.nc.gov/about/divisions/water-resources/water-resources-data/water-sciences-home-page/ecosystems-branch/ambient-monitoring-system. Accessed July 6th, 2018.

NC DEQ. 2018c. Monitoring Coalition Program. https://deq.nc.gov/about/divisions/water-resources/water-resources-data/water-sciences-home-page/ecosystems-branch/monitoring-coalition-program. Accessed July 6th, 2018.

O'Driscoll, M., Clinton, S., Jefferson, A., Manda, A., & McMillan, S. 2010. Urbanization effects on watershed hydrology and in-stream processes in the southern United States. *Water*, *2*(3), 605-648.

Olson, J. R., & Hawkins, C. P. 2012. Predicting natural base-flow stream water chemistry in the western United States. *Water Resources Research*, *48*(2).

Olson, J. R., & Hawkins, C. P. 2013. Developing site-specific nutrient criteria from empirical models. *Freshwater Science*, *32*(3), 719-740.

Palmer, M. A., Hondula, K. L., & Koch, B. J. 2014. Ecological restoration of streams and rivers: shifting strategies and shifting goals. *Annual Review of Ecology, Evolution, and Systematics*, *45*, 247-269.

Paul, M. J., & Meyer, J. L. 2001. Streams in the urban landscape. *Annual review of Ecology and Systematics*, *32*(1), 333-365.

Qian, S. S., Cuffney, T. F., Alameddine, I., McMahon, G., & Reckhow, K. H. 2010. On the application of multilevel modeling in environmental and ecological studies. *Ecology*, *91*(2), 355-361.

Raftery, A. E. 1995. Bayesian model selection in social research. *Sociological methodology*, 111-163.

Richards, C., Johnson, L. B., & Host, G. E. 1996. Landscape-scale influences on stream habitats and biota. *Canadian Journal of Fisheries and aquatic sciences*, *53*(S1), 295-311.

Read, E. K., Carr, L., De Cicco, L., Dugan, H. A., Hanson, P. C., Hart, J. A., ... & Winslow, L. A. 2017. Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research*, *53*(2), 1735-1745.

Rehmann, C. R., & Soupir, M. L. 2009. Importance of interactions between the water column and the sediment for microbial concentrations in streams. *Water Research*, *43*(18), 4579-4589.

Sample, D. J., Heaney, J. P., Wright, L. T., Fan, C. Y., Lai, F. H., & R. Field. 2003. Costs of best management practices and associated land for urban stormwater control. *Journal of Water Resources Planning and Management*, *129*(1), 59-68.

Schueler, T. R. 1994. The importance of imperviousness. *Watershed protection techniques*, *1*(3), 100-111.

Schueler, T. R., Fraley-McNeal, L., & Cappiella, K. 2009. Is impervious cover still important? Review of recent research. *Journal of Hydrologic Engineering*, *14*(4), 309-315.

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

Seitzinger, S. P. 1994. Linkages between organic matter mineralization and denitrification in eight riparian wetlands. *Biogeochemistry*, *25*(1), 19-39.

Selvakumar, A., O'Connor, T. P., & Struck, S. D. 2009. Role of stream restoration on improving benthic macroinvertebrates and in-stream water quality in an urban watershed: case study. *Journal of Environmental Engineering*, *136*(1), 127-139.

Sinha, E., & Michalak, A. M. 2016. Precipitation dominates interannual variability of riverine nitrogen loading across the continental United States. *Environmental science & technology*, *50*(23), 12874-12884.

Spänhoff, B., & Arle, J. 2007. Setting attainable goals of stream habitat restoration from a macroinvertebrate view. *Restoration Ecology*, *15*(2), 317-320.

Stoll, S., Breyer, P., Tonkin, J. D., Früh, D., & Haase, P. 2016. Scale-dependent effects of river habitat quality on benthic invertebrate communities—implications for stream restoration practice. *Science of the Total Environment*, *553*, 495-503.

Strayer, D. L., Beighley, R. E., Thompson, L. C., Brooks, S., Nilsson, C., Pinay, G., and R. J. Naiman. 2003. Effects of land cover on stream ecosystems: roles of empirical models and scaling issues. *Ecosystems*, *6*(5): 407-423.

Strickling, H.L. & Obenour, D.R. 2018. Leveraging Spatial and Temporal Variability to Probabilistically Characterize Nutrient Sources and Export Rates in a Developing Watershed. *Water Resources Research.*

Sutherland, A. B., Meyer, J. L., & Gardiner, E. P. 2002. Effects of land cover on sediment

regime and fish assemblage structure in four southern Appalachian streams. *Freshwater Biology*, *47*(9), 1791-1805.

Tchobanoglous, G., Burton, F. L., & Stensel, H. D. 2003. *Wastewater engineering treatment and reuse* (No. 628.3 T252s). Boston, US: McGraw-Hill Higher Education.

Tremblay, A., and J. Ransijn. (2015). LMERConvenienceFunctions: Model selection and post-hoc analysis for (G) LMER models. R Package Version 2.

USEPA (United States Environmental Protection Agency). 2005. Stormwater Phase II Final Rule. Office of Water. EPA 833-F-00-002.

USEPA (United States Environmental Protection Agency). 2006. Wadeable Stream Assessment. Office of Research and Development. Office of Water. EPA 841-B-06_002.

USEPA (United States Environmental Protection Agency). 2009. National Water Quality Inventory: Report to Congress, 2004 Reporting Cycle. Report # EPA-841-R-08-001. Office of Water. District of Columbia, U.S.A. 52 pp.

USEPA (United States Environmental Protection Agency). 2016. Office of Water and Office of Research and Development. National Rivers and Streams Assessment 2008-2009: A Collaborative Survey (EPA/841/R-16/007). Washington, DC.

Vietz, G. J., Sammonds, M. J., Walsh, C. J., Fletcher, T. D., Rutherfurd, I. D., and M.J. Stewardson. 2014. Ecologically relevant geomorphic attributes of streams are impaired by even low levels of watershed effective imperviousness. *Geomorphology*, *206*, 67-78.

Walsh, C. J., Roy, A. H., Feminella, J. W., Cottingham, P. D., Groffman, P. M., & Morgan II, R. P. 2005. The urban stream syndrome: current knowledge and the search for a cure. *Journal of the North American Benthological Society*, *24*(3), 706-723.

Walsh, C. J., Booth, D. B., Burns, M. J., Fletcher, T. D., Hale, R. L., Hoang, L. N., ... & A.

Wallace. 2016. Principles for urban stormwater management to protect stream ecosystems. *Freshwater Science*, *35*(1), 398-411.

Whitman, R. L., Nevers, M. B., & Byappanahalli, M. N. 2006. Examination of the watershed-wide distribution of Escherichia coli along southern Lake Michigan: an integrated approach. *Applied and Environmental Microbiology*, *72*(11), 7301-7310.

Wilkinson, J., Jenkins, A., Wyer, M., & Kay, D. 1995. Modelling faecal coliform dynamics in streams and rivers. *Water Research*, *29*(3), 847-856.

Wilson, J. S., & Lindsey, G. H. 2009. Identifying urban neighborhoods for tree canopy restoration through community participation. In *Planning and Socioeconomic Applications* (pp. 29-42). Springer, Dordrecht.

Wilson, William G. 2016. *Stormwater*. The University of Chicago Press. Chicago.

Supplementary Materials

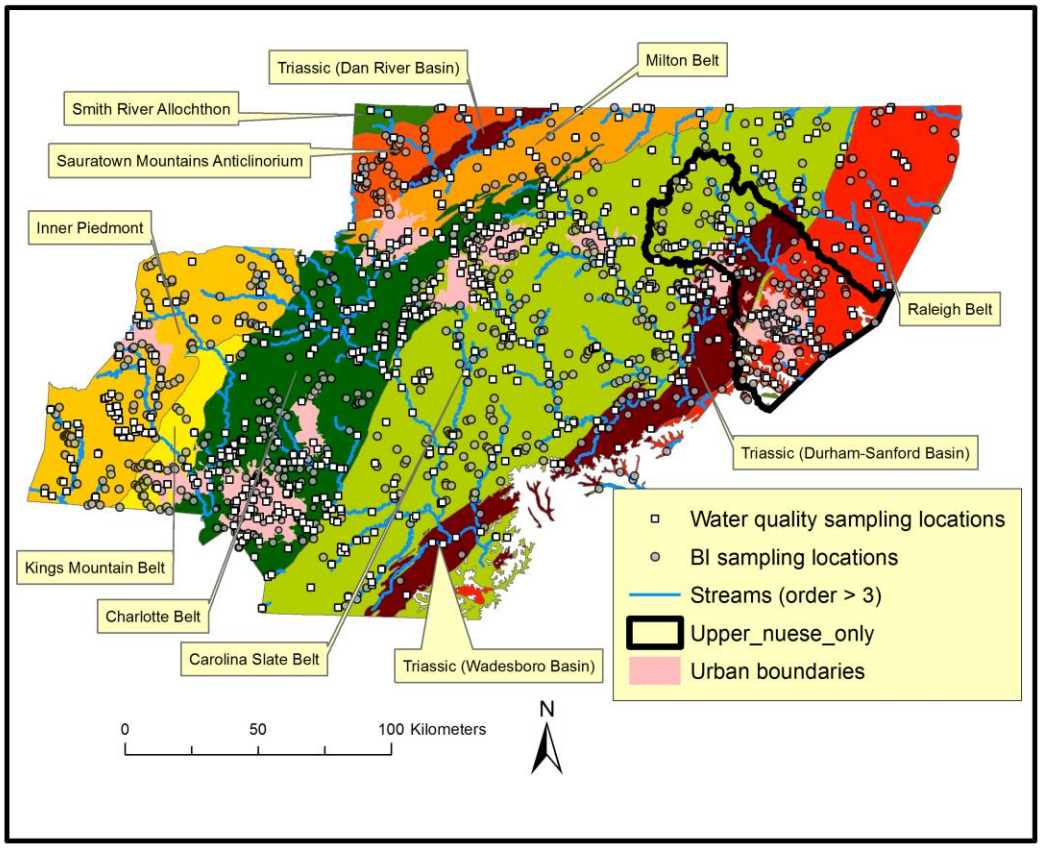


Figure S1: 9 Major geologic regions in the NC Piedmont. Water quality sampling locations had samples for either : fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and/or specific conductance (SC). Many sampling locations for the biotic index (BI) were at separate locations.

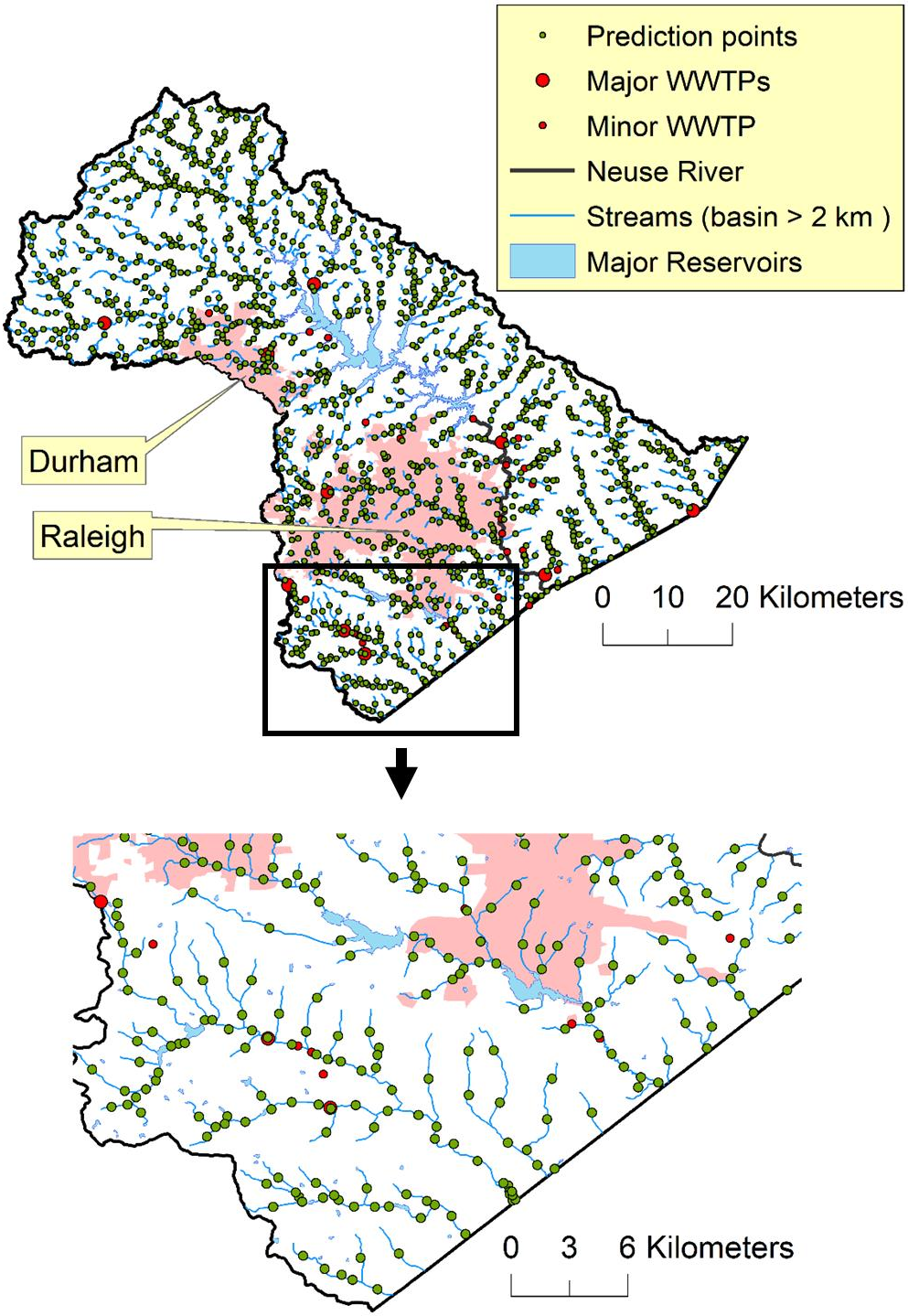


Figure S2: 1423 points used for predictions in the UNRB. Each point is the midpoint of a NHD Plus stream that was deemed to be wadeable (i.e., > 2 km^2 and < 500 km^2).

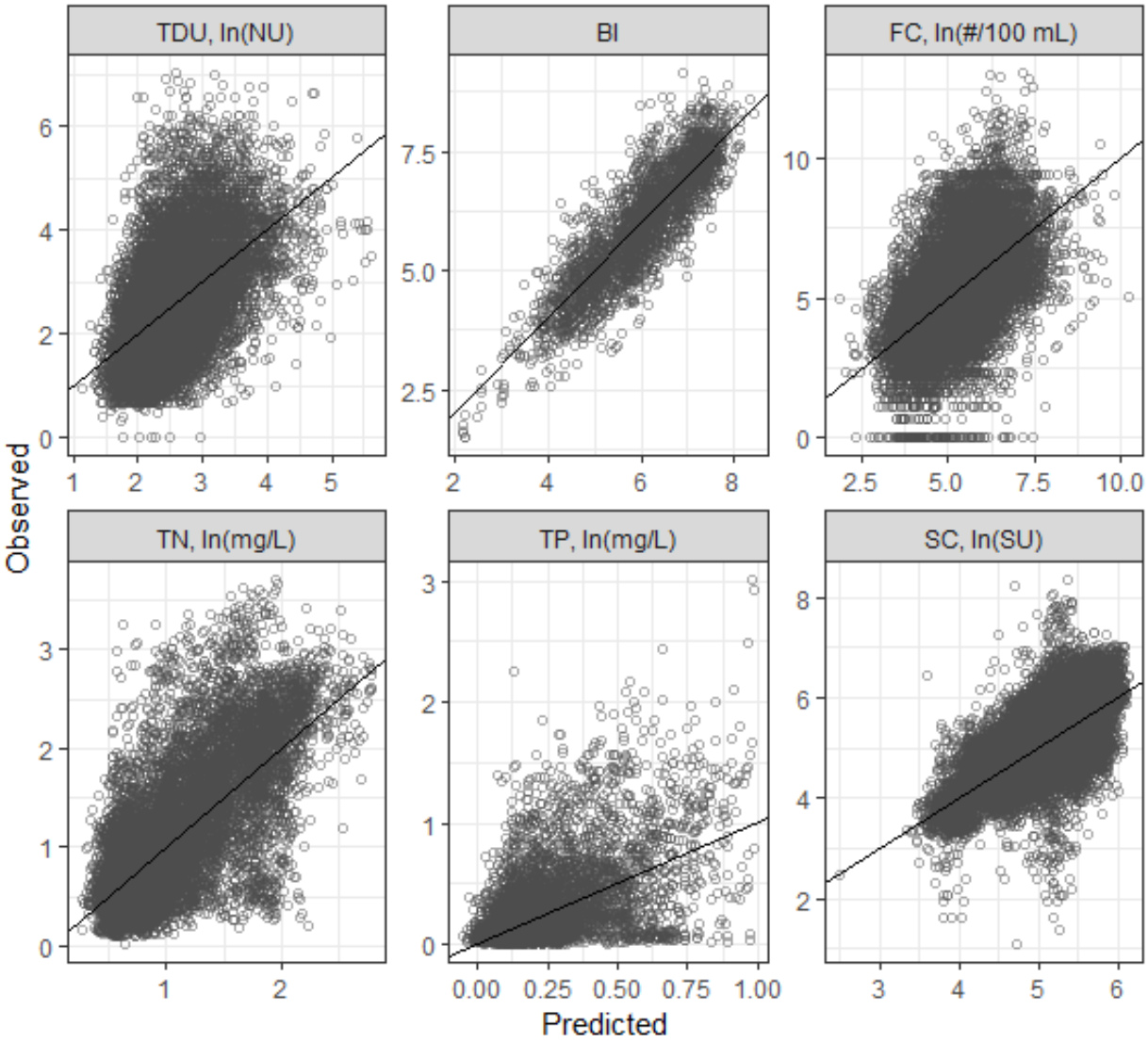


Figure S3: Observed vs. Predicted plots with only deterministic factors (candidate predictors and soil condition).

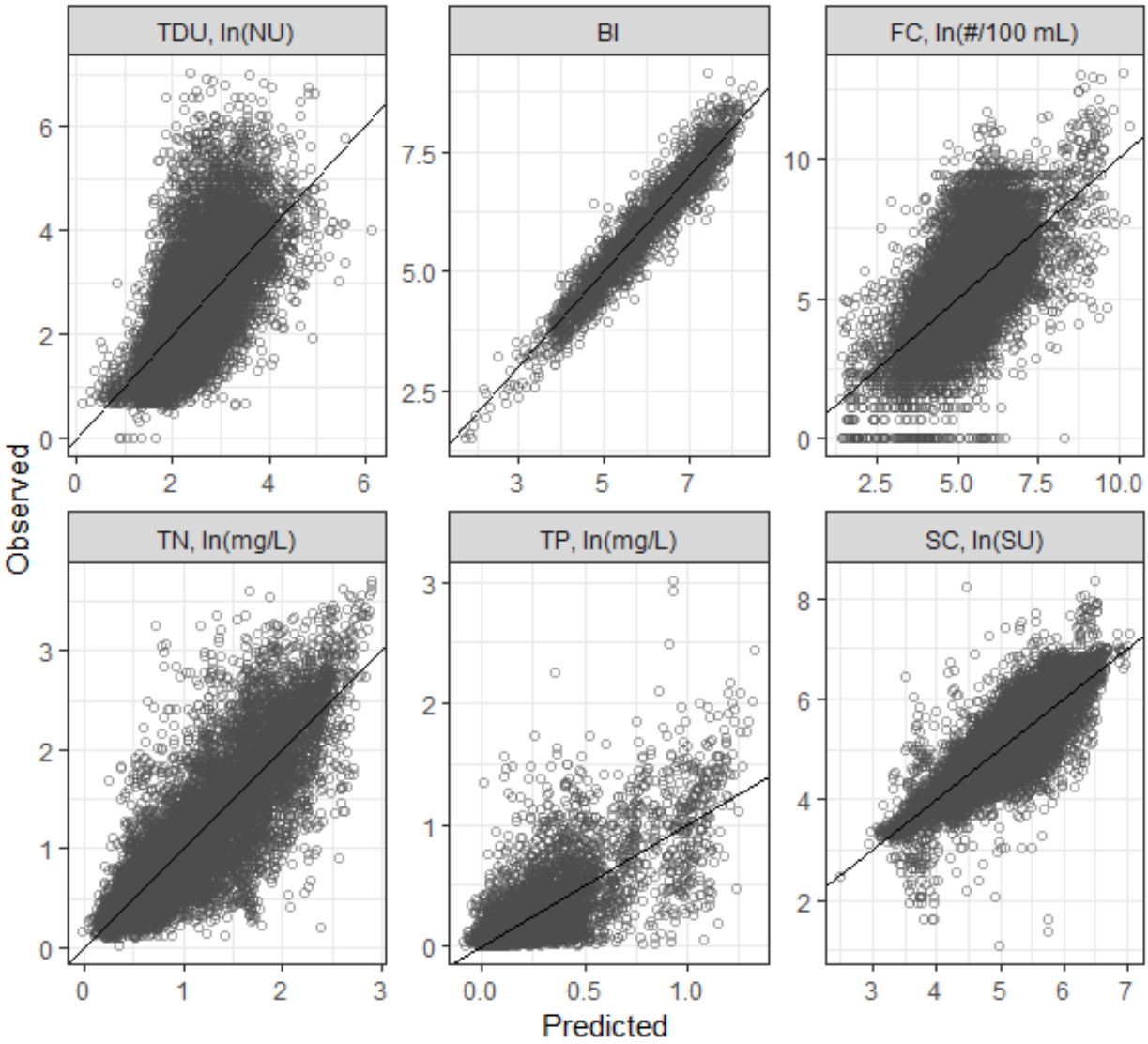


Figure S4: Observed vs. Predicted plots for full HMLRs including site and basin random effects.

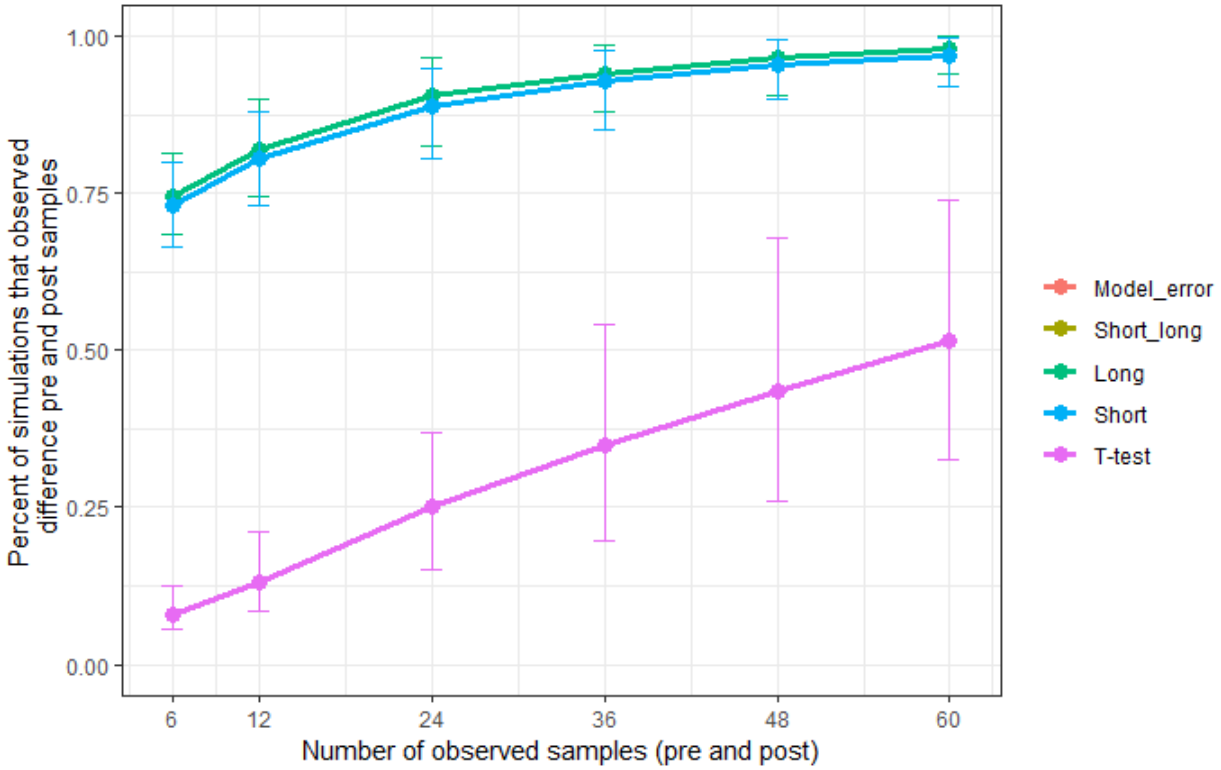


Figure S5: Chance that mean post-restoration sampling in highly urban areas will have an improved (i.e., lower) value to mean pre-restoration samples for FC. Lines represent model uncertainty (model error), model plus random short term weather (short), model plus random long term weather (long), and model plus random short and long term weather(short_long). In addition, the fraction of simulations that led to statistically significant differences between pre and post sampling are also shown (t-test). Error bars represent the 10th and 90th percentile probability of the 142 highly urban stream reaches. The line for model error does not appear on the graph because it is identical to values for "long", while "Short_long" does not appear because it is identical to "Short."

Table S1: Summary of WWTP discharges divided between major (> 1 MGD) and minor WWTPs.

Type of WWTP	Water quality parameter	Units	# of WWTP	# of monthly records	Values			
					5%	Median	Mean	95%
Major	Discharge	MGD	108	38,498	0.0	1.8	32.3	132.0
	Total Nitrogen	mg/L	102	23,624	0.7	8.8	12.9	27.4
	Total Phosphorus	mg/L	103	23,722	0.0	1.1	1.9	5.2
	Turbidity	NTU	9	732	0.2	1.5	3.6	10.9
	Specific conductance	SU	57	10,740	305.0	628.3	900.5	1920.0
Minor	Discharge	MGD	466	102,918	0.0	0.0	0.1	0.4
	Total Nitrogen	mg/L	275	25,238	0.3	12.4	17.3	48.6
	Total Phosphorus	mg/L	274	27,495	0.0	2.1	3.0	7.7
	Turbidity	NTU	99	8,120	0.0	4.5	10.0	30.5
	Specific conductance	SU	69	5,031	28.6	507.0	1211.1	2,820

Table S2: The change in BIC from baseline models when each candidate predictor variable was added to the baseline model. In each variable class, the predictor that had the lowest value (-6 at a minimum) where then selected to be candidates for the multi-stresssor HMLRs for each water quality indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC).”**X**” indicates variables that were selected and remained in multi-stressor models, while “**x**” represent variables that were parsed from the multi-stressor models for not being significant.

		BI		**FC**		**TDU**		**TN**		**TP**		**SC**	
	Candidate predictor variable	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR
Canopy Loss	% Canopy watershed	-34.4		-18.7		11.3		0.9		7.1		-33	
	% Canopy stream buffer	-64.2		-150.8	**X**	-1		-54.9	**X**	-17.7	**X**	0.3	
	% Canopy watershed (log)	-82.0		-22		6.5		0.3		5.5		-47.1	**X**
	% Canopy stream buffer (log)	-116.7	**X**	-89.3		3.2		-37.7		-13		-0.7	
Crop	% Crop basin	-15.2		-42.6		6.7		-86.6	**X**	2.5		-56.1	**X**
	% Crop buffer	-29.4		-50.8		9.1		-29.9		7.6		-0.9	
	% Crop basin (log)	-80.6	**x**	-52.7		-7.9	X	-38.6		0.4		-37.7	
	% Crop buffer (log)	-77.9		-55.5	**X**	8.6		-19.8		7.8		-0.1	
Drainage area	Drainage	1.8		-25.9	**X**	11.6		-1.1		7		-0.3	
Elevation	Elevation	-26.0	**x**	-33.7	**x**	10.1		0.9		7.2		-0.2	
Hydro-climatology (short)	Prec, 1 day	-28.4		-2518.9		-4515.4		-106.1		7.3		-954.4	
	Prec, 3 day	-29.4		-2364.7		-5188.1		-193.7		-14.9		-1566.6	
	Prec, 7 day	-33.5		-1513.1		-4266.9		-222		-62		-1874.6	
	Prec, 14 day	-36.9		-1007		-2909.4		-215.3		-62.7		-1528.6	
	Prec, 1 day (log)	-28.6		-2546.6		-4859.9		-118.4		7		-1017.8	
	Prec, 3 day (log)	-29.3		-2774.3	**X**	-5653.6	**X**	-217		-17.4		-1675.3	
	Prec, 7 day (log)	-34.5		-1625.1		-4720.9		-258.5		-68.7	**X**	-2015.1	**X**
	Prec, 14 day (log)	-38.3	**x**	-1078.5		-3326.7		-260.5	**X**	-66.2		-1681.4	

Table S2 (continued)

	Candidate predictor variable	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR
Hydro-climatology (long)	Prec, 1 month	-14.4		-110.1		-208		-80.6		-23		-570.8	
	Prec, 3 month	-6.3		-253.3		-436.8		-117.8		-18.4		-841.4	
	Prec, 6 month	5.5		-95.1		-189		-133.3		-75.8		-1046.8	
	Prec, 12 month	0.6		-154		-336.5		-167.6		-121.2		-1171	
	Prec, 24 month	7.3		-111.1		-152.4		-98.6		-91		-724.2	
	SPI, 1 month	5.4		-448.9	**X**	-1022.7	**X**	-87.6		-5.4		-747.1	
	SPI, 3 month	-18.0		-167.8		-684.7		-227.1		-71.6		-1611.2	
	SPI, 6 month	-13.9		-131.5		-771.2		-289.6	**X**	-145.1	**X**	-1997.5	**X**
	SPI, 12 month	-3.3		-267.3		-714.5		-224.4		-143.6		-1568.7	
	SPI, 24 month	6.8		-214		-341.2		-126.6		-98.3		-1029.4	
	Prec, 1 month (quad.)	-10.6		-104.5		-244.4		-82.1		-15.9		-562.4	
	Prec, 3 month (quad.)	-22.7	**X**	-251.5		-468.2		-108.7		-10		-848.5	
	Prec, 6 month (quad.)	-2.7		-90.4		-179.1		-128.3		-66.8		-1049.9	
	Prec, 12 month (quad.)	-15.9		-148.8		-400.4		-158		-124		-1188.8	
	Prec, 24 month (quad.)	11.4		-101.4		-150.1		-89.3		-89.6		-765.1	
	SPI, 1 month (quad.)	9.9		-439.1		-1013.8		-83		2.3		-750.4	
	SPI, 3 month (quad.)	-15.4		-158.7		-676		-217.6		-65.8		-1622.5	
	SPI, 6 month (quad.)	-21.1		-130.9		-762.7		-283.1		-144.1		-1992.1	
	SPI, 12 month (quad.)	-4.3		-284		-758.5		-214.8		-145.0		-1565	
	SPI, 24 month (quad.)	12.3		-204.5		-333.4		-118.5		-106		-1043.2	
IC	% IC basin	na		-15.4		11.6		-1.5		7.8		-24.4	**x**
	% IC stream buffer	2.7		-157.2	**x**	0.3		-31.8	**X**	-9.7	**x**	-0.5	
	% IC basin (log)	na	**baseline**	-18.5		6.9		0.6		5.8		-70	
	% IC stream buffer (log)	-1.7		-71.6		10.5		-6.7		-1.1		-0.9	
IC (recent) IC (age)	IC basin, previous 2 yrs	5.4		-26		0.7		-2.2		-45.5	**X**	-24.9	
	IC basin, previous 5 yrs	4.3		-19.5		-15.4		0.8		-33.4		-64	
	IC basin, previous 2 yrs (log)	1.8		-23.6		-5.6		0.6		-29		-32.4	
	IC basin, previous 5 yrs (log)	2.4		-19.4		-28.5	**X**	0.8		-17.5		-33.8	
	IC, pre 1974			-48.2		11.4		-7.3		7.1		-18.7	
	IC, pre 1980			-48.7		11.6		-7.1		7.1		-20.2	
	IC, pre 1990			-51.4		11.6		-6.9		6.6		-22.4	
	IC, pre 2000			-59.1	**X**	11		-6.7		4.1		-19	
	IC, pre 1974 (log)	-28.8		-49.4		10.6		-7.2		5.9		-60.4	
	IC, pre 1980 (log)	-32.3	**X**	-48.8		9.9		-7.2		6		-61.8	**X**
	IC, pre 1990 (log)	-25.8		-54.9		8.5		-7.5		4		-59.7	
	IC, pre 2000 (log)	-0.3		-55.6		6		-7.6	**x**	2.2		-54.4	

Table S2 (continued)

	Candidate predictor variable	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR	Δ BIC from baseline model	Selected for multi-stressor HMLR
Pasture	Pasture basin	-13.2	**x**	-10.4		11.6		0.2		7.3		0.1	
	Pasture stream buffer	-10.7		-45.7		7.1		-0.1		6.1		-3.8	
	Pasture basin (log)	-5.8		-15.2		10.7		0.9		7.4		-2.2	
	Pasture stream buffer (log)	-5.8		-68.2	**X**	3.1		-3.9		6.4		-7.8	
Reservoirs	Reservoir (within 1 km)	2.6		-12		10.5		-0.6		3.7		-0.8	
	Reservoir (within 5 km)	-6.9	**X**	-13.9		11.2		-1.4		7.3		0.3	
	Reservoir (within 10 km)	-3.5		-14.9	**x**	8.5		-1.1		6.2		-0.6	
Sampling method	BI sampling methods	na	**baseline**	na		na		na		na		na	
Season	Seasonal categorical variable		**X**		**X**		**X**		**X**		**X**		**X**
Wetlands	Wetlands basin	-6.8		-22.8		5.9		-1.2		5.7		-59.2	
	Wetlands stream buffer	-1.8		-10		3.3		-14.3	**X**	6.6		0.3	
	Wetlands basin (log)	-10.0		-27.8	**x**	2.1		-2.9		5.6		-34	**X**
	Wetlands stream buffer (log)	-10.0	**X**	-16.3		5.2		-9.7		-2.5		0.3	
WWTPs	Major WWTP (within 1 km)	-19.7	**X**										
	Major WWTP (within 5 km)	-17.0											
	Major WWTP (within 10 km)	-13.9											
	# of Major WWTP			-12.7		10.8		-0.2		-16.8		-35.8	**X**
	# of Minor WWTP			-9.8		5.1		-0.4		7.2		-2.3	
	TN (or TP) load/drainage	1.9		-9.8		10.1		na		-257.1		-17.7	
	TN (or TP) load/drainage (log)	-11.1		-19.6	**X**	1.7		baseline	**X**	-331.5	**X**	-10.7	
Year	Year of sample	-75.0	**X**	-110.8	**X**	11.6		-12.4	**x**	-41.6	**x**	0.3	

Table S3: Model coefficients for six water quality indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC). Hyperdistributions of all random effects are presented in their transformed units. All coefficients with a "*" mean that the predictor variable was log transformed.

	BI	**FC**	**TDU**	**TN**	**TP**	**SC**
Intercept	4.37	3.05	2.20	0.57	0.06	4.12
Canopy loss (buffer)	0.48*	0.02		0.008	0.001	0.07*
Crop (buffer)		-0.47*				
Crop (watershed)			0.17*	0.05		-0.05
Drainage		-0.002				
Hydro-Clim, short (days)		4.23* (3)	3.28* (3)	-0.19* (14)	-0.10* (7)	-0.54 (7)
Hydro-Clim, long (mos.)	-3.16 (4.06)	0.09 (1)	.11 (1)	-0.04 (6)	-0.02 (6)	-0.07 (6)
IC, buffer				-0.008		
IC, basin	0.30*					
IC, age (year)	0.15*	0.33* (00)				0.23*(80)
IC, recent (years)			0.30* (5)		0.12* (2)	-0.14* (5)
Pasture (buffer)		0.37*				
Reservoirs, 5 km	0.35					
Samp. Method (ept,qual4)	-0.92,-0.01					
WWTP loadings		0.16* (TN)		0.63* (TN)	0.74* (TP)	
# of WWTP upstream						
WWTP proximity	0.62					
Wetlands (buffer)				-0.01		0.06*
Year	-0.01			0.006		
Soil random effect	N(0,0.41)	N(0,0.28)	N(0,0.14)	-	-	N(0,0.24)
Site random effects	N(0,0.31)	N(0,0.90)	N(0,0.24)	N(0,0.23)	N(0,0.08)	N(0,0.28)
Basin random effect	N(0,0.39)	-	N(0.23)	N(0,0.21)	N(0,0.09)	N(0,0.29)
Season, winter	-0.24	-0.64	0.08	-0.07	-0.08	-0.14
Season, spring	-0.25	-0.54	0.00	-0.10	-0.08	-0.16
Season, fall	0.06	-0.28	-0.30	-0.01	-0.01	-0.01

Table S4: Seasonal offsets for the multi-stressor models for six water quality indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC). All values are in the units of the response variable (i.e., log transformed except for the BI), and standard errors are shown in parenthesis.

	Seasonal offsets (from Summer)		
	Fall	Winter	Spring
BI	0.06 (0.04)	-0.24 (0.03)	-0.21 (0.04)
FC	-0.28 (0.03)	-0.64 (0.03)	-0.54 (0.03)
TDU	-0.31 (0.01)	0.08 (0.01)	-0.01 (0.01)
TN	-0.01 (0.01)	-0.07 (0.01)	-0.01 (0.01)
TP	-0.01 (0.01)	-0.08 (0.01)	-0.08 (0.01)
SC	-0.01 (0.01)	-0.14 (0.01)	-0.16 (0.01)

Table S5: Soil random effects for the multi-stressor models for six water quality indicators: biotic index (BI), fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC). All values are in the units of the response variable (i.e., log transformed except for the BI). Missing values (-) for TN and TP showed there were no random effects for those models while missing values for Eastern Slate Belt and Sauratown Mountians imply there were no water quality samples taken in those regions.

	Soil random effects					
	BI	FC	TDU	TN	TP	SC
Carolina Slate Belt	0.26	-0.08	-0.32	-	-	0.33
Charlotte Belt	0.25	0.19	0.36	-	-	0.03
Triassic	0.65	0.16	-0.25	-	-	0.19
Eastern Slate Belt	0.01	-	-	-	-	-
Inner Piedmont	-0.41	-0.05	0.23	-	-	-0.33
Kings Mountain Belt	-0.03	0.07	0.05	-	-	-0.03
Milton Belt	0.03	-0.10	-0.09	-	-	0.10
Raleigh Belt	0.21	-0.15	-0.04	-	-	0.10
Sauratown Mountains	-0.49	-	-	-	-	-0.04
Smith River Allochthon	-0.47	-0.07	-	-	-	-0.12

Table S6: Cross validation of multi-stressor HMLR models

	BI	FC	TDU	TN	TP	SC
R^2 (multi-stressor HMLR)	**0.76**	**0.25**	**0.28**	**0.54**	**0.32**	**0.42**
R^2 (CV, fold #1)	0.69	0.23	0.22	0.54	0.30	0.30
R^2 (CV, fold #2)	0.73	0.22	0.25	0.54	0.33	0.27
R^2 (CV, fold #3)	0.69	0.22	0.24	0.54	0.32	0.29
R^2 (CV, total dataset)	*0.69*	*0.22*	*0.24*	*0.54*	*0.32*	*0.29*

Table S7: Summary of mean water quality condition in the Upper Neuse River Basin (UNRB) for stream reaches with above a minimal amount of WWTP loading [>0.1 kg TN/(day*km2)]. Median current conditions and five potential management scenarios are shown along with predictions for wet (90th percentile) and dry conditions (10th percentile short and long-term precipitation) made for current conditions and management scenario #4.

	Management Scenario	BI (0-10)	TDU (NTU)	FC (#/100 mL)	TN (mg/L)	TP (mg/L)	SC (SU)
Wastewater affected reaches (60 sites)	**Current**	**6.40**	**10.3**	**93**	**2.11**	**0.28**	**180**
	Current (wet)	6.35	26.1	272	1.76	0.21	139
	Current (dry)	6.52	8.0	73	2.47	0.33	214
	#1	-5%	0%	-19%	-14%	-7%	-4%
	#2	-3%	-3%	-11%	0%	0%	-2%
	#3	0%	0%	-1%	-7%	-6%	-11%
	#4	-8%	-3%	-28%	-19%	-12%	-17%
	#4 (wet)	-8%	-3%	-28%	-21%	-16%	-17%
	#4 (dry)	-8%	-3%	-28%	-18%	-11%	-17%
	#5	-10%	-18%	-48%	-37%	-35%	-29%

Conclusion

Hierarchical modeling was used throughout this dissertation to draw statistical inferences from over 70,000 fish trawls in 33 Gulf of Mexico (GoM) estuaries, and more than 90,000 samples for six water quality indicators across NC Piedmont wadeable streams. In all three chapters, the flexibility of a hierarchical framework accounted for group-level variation (e.g., estuaries, sites, and basins), allowing for rigorous screening of watershed stressors that were linked to *in situ* biological and chemical water quality. Hierarchical model outputs (i.e., random effects) were also helpful in answering different scientific and management questions in all three chapters. Though all chapters were joined by a similar methodology, model construction (e.g., variable selection, predictor data), analysis of random effects (e.g., comparing regional differences, identification of restoration sites; Fig. 2.4), and the ultimate end use of multi-stressor models (e.g., determining benchmark conditions, comparing relative importance of predictors, forecasting future water quality) varied by chapter. All chapters demonstrated that a hierarchical framework is a computationally tractable and useful modeling approach for modeling large regional aquatic systems. The following paragraphs identify the key findings for each study followed by a final section dealing with possible areas for future work.

Chapter 1 presents the first regional study to link the biological condition of estuaries to basin-level watershed stressors. We created a methodology for ranking estuaries based on their deviation from benchmark conditions: least disturbed (LDC) or minimally disturbed (MDC). By identifying dominant watershed stressors that affected large number of species throughout the GoM [total anthropogenic (urban, crop and pasture) land use, crop land use, the number of toxic release sites in upstream watersheds, and agriculture (crop and pasture) in shoreline buffers; Table 1.5], estuaries could be ranked based on their current amount of upstream watershed stress

(Fig 1.2; 1.3). Variable selection for GoM watershed stressors was based on statistical significance (95% confidence interval) and the number on species affected throughout the GoM while hierarchical offsets helped compare state sampling efficiency and distinguish estuaries that might have better or worse conditions than model predictors suggested (Fig 1.4; 1.5).

Chapter 2 presents the first use of a hierarchical multiple linear regression (HMLR) approach to model macroinvertebrates [i.e., biotic index (BI)] throughout a regional aquatic system. This study builds on previous research that used hierarchical modeling to compare individual stressors that varied across nine US regions (Cuffney et al. 2011; Qian et al. 2010). We compared the impact of co-varying predictor variables on BI and then ranked their relative impact on stream biotic indices (BI; Fig 2.2). Our results imply that watershed IC and canopy removal in stream buffers are the two largest anthropogenic sources of degradation for the biological health of streams, and canopy removal has two to 9 times the impact on BI on a per hectare basis, as compared to the addition of watershed IC (Fig 2.3). Recent construction (post-1980) also impacted stream BI 30% less than older development (pre-1980; Table 2.5; Fig 2.3), potentially attributable to NC regulatory sediment controls for construction implemented after 1980 (NC DEQ 2017; Howells 1990). Variable selection was performed by testing the improvement to baseline models by adding auxiliary stressors using the Bayesian Information Criteria (BIC). In contrast to the GoM, where change in land covers was considered constant throughout the study period, temporal changes in impervious cover were incorporated as candidate predictors in chapter 2 and allowed comparisons between the effects of older vs. newer impervious cover on BI. Water quality managers are often faced with the need to improve the ecological health of streams without knowing what the best possible conditions are for degraded urban areas (Paul et al. 2009; Paul 2015). In chapter 2, I showed how multi-stressor HMLR

predictions along with site random effects could be used to identify potential restoration sites with greater recovery potential (Fig 2.4).

Chapter 3 built upon the model construction and variable selection of chapter 2. Multi-stressor HMLR models were constructed for five additional water quality indicators: fecal coliform (FC), turbidity (TDU), total nitrogen (TN), total phosphorus (TP), and specific conductance (SC; Table 3.4); and then mean water quality values were forecasted throughout the Upper Neuse River Basin (UNRB) for potential future management scenarios (Table 3.5). Restoring canopy cover improved mean stream health throughout the UNRB more than mitigating IC or reducing wastewater treatment plant loadings. There was a large amount of site and basin variation for indicators that implied the need and importance for identifying additional site-specific sources of degradation. Considering model and weather uncertainty, BI, TN and FC were most likely to have observed water quality changes in response to management actions. Several additional predictors were considered in this chapter, temporally specific short-term precipitation data and WWTP nutrient loadings, which were instrumental for creating working models for FC, TDU, TN, TP, and SC. Also, basin random effects were added to site random effects and offered insight into spatial correlations of water quality indicators.

Future directions that logically follow from this work include the downscaling of projects, adding more complexity to random effects (i.e., random slopes), and comparing this approach to other methodologies (i.e., machine learning techniques) that have been used to model regional water quality. Though downscaling projects would eliminate the advantage of leveraging data across groups (i.e., estuaries, watersheds, sites), identifying local factors (not accounted for in our models) that are affecting water quality could help our understanding of the substantial local variance occurring with groups. Research could address the variation of the

biological condition within different estuarine subsections or focus on a particular city where more detailed watershed data (e.g., sewage infrastructure, BMP installations) are available. At smaller scales, it might also be feasible to add water quality variables (e.g., specific conductance, dissolved oxygen, total nitrogen), which have been related to ecological health in previous assessments (USEPA 2006, 2016, Cuffney et al. 2010). As more construction occurs after recent and more stringent regulations (i.e., post-2005; EPA 2005), smaller study areas are probably necessary to help determine the expected ecological benefits of these rules. Random slopes could be explored to address questions about whether the relationship between watershed stressors and water quality (i.e., model coefficients) change across groups (e.g., estuaries, city). This could test time trends in different GoM estuaries or quantify the effect of IC on BI in different cities. Finally, variable selection and the modeling approach I employed contrast with machine learning models (i.e., random forests) that have been used recently for modeling regional aquatic systems (Olson and Hawkins, 2012; Olson and Hawkins, 2013; Hill et al. 2017). Though random forest modeling has advantages including good predictive ability (Evans and Cushman 2009; Evans et al. 2011), they do not relate predictors to response variables with coefficients that quantify uncertainty and often are not conducive to answering research questions or dealing with grouped data. Comparing this type of modeling approach to hierarchical modeling to check for consistency in variable selection and water quality predictions could give insights into the advantages and limitations of each approach.

References (for Introduction and Conclusion)

Beck, M. B. 1987. Water quality modeling: a review of the analysis of uncertainty. *Water Resources Research*, *23*(8), 1393-1442.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and J. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, *24*(3): 127-135.

Booth, D. B., and Jackson, C. R. 1997. Urbanization of Aquatic Systems: Degradation Thresholds, Stormwater Detection, and the Limits of Mitigation. *JAWRA Journal of the American Water Resources Association*, *33*(5), 1077-1090.

Clark, J. S., Carpenter, S. R., Barber, M., Collins, S., Dobson, A., Foley, J. A., ... & Pringle, C. 2001. Ecological forecasts: an emerging imperative. *science*, *293*(5530), 657-660.

Clark, J. S., & Gelfand, A. E. 2006. A future for models and data in environmental science. *Trends in Ecology & evolution*, *21*(7), 375-380.

Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Lasram, F. B. R., Aguzzi, J. and R. Danovaro. 2010. The biodiversity of the Mediterranean Sea: estimates, patterns, and threats. *PloS one*, *5*(8), e11842.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, *19*(3): 553-570.

Cuffney, T. F., Kashuba, R., Qian, S. S., Alameddine, I., Cha, Y. K., Lee, B., ... and G. McMahon. 2011. Multilevel regression models describing regional patterns of invertebrate and algal responses to urbanization across the USA. *Journal of the North American Benthological Society*, *30*(3): 797-819.

De Brauwere, A., Ouattara, N. K., & Servais, P. 2014. Modeling fecal indicator bacteria concentrations in natural surface waters: a review. *Critical reviews in environmental science and technology*, *44*(21), 2380-2453.

Gelman, A., and J. Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Carlin, J. B., Stern, H. S., and D. B. Rubin. 2014. *Bayesian data analysis* (Vol. 2). Chapman & Hall/CRC. Boca Raton, FL, USA.

Hill, R. A., Fox, E. W., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., and M. H. Weber. 2017. Predictive mapping of the biotic condition of conterminous US rivers and streams. *Ecological Applications*, *27*(8): 2397-2415.

Howarth, R.W., and R. Marino. 2006. Nitrogen as the limiting nutrient for eutrophication in coastal marine ecosystems: Evolving views over three decades. *Limnology and Oceanography* 51(1): 364-376.

Howells, D. H. 1990. Quest for clean streams in North Carolina: an historical account of stream pollution control in North Carolina. *Water Resources Research Institute of the University of North Carolina (USA)*.

Leopold, L. B., Huppman, R., and A. Miller. 2005. Geomorphic effects of urbanization in forty-one years of observation. *Proceedings of the American Philosophical Society*, 349-371.

Malmqvist, B., and Rundle, S. 2002. Threats to the running water ecosystems of the world. *Environmental conservation*, *29*(2), 134-153.

Miller, J., Esselman, P. C., Alameddine, I., Blackhart, K., and D.R. Obenour. 2018. Hierarchical modeling assessment of the influence of watershed stressors on fish and invertebrate species in Gulf of Mexico estuaries. *Ecological Indicators*, *90*, 142-153.

NC DEQ. 2017. Erosion and Sediment Control Planning and Design Manual. https://deq.nc.gov/about/divisions/energy-mineral-land-resources/energy-mineral-land-permit-guidance/erosion-sediment-control-planning-design-manual. Accessed Dec. 29th, 2017.

NRC (Natural Resource Council). 2000. Clean coastal waters: Understanding and reducing the effects of nutrient pollution. Washington: National Academies Press.

Olson, J. R., & Hawkins, C. P. 2012. Predicting natural base-flow stream water chemistry in the western United States. *Water Resources Research*, *48*(2).

Olson, J. R., & Hawkins, C. P. 2013. Developing site-specific nutrient criteria from empirical models. *Freshwater Science*, *32*(3), 719-740.

Paul, M. J., and J. L. Meyer. 2001. Streams in the urban landscape. *Annual review of Ecology and Systematics*, *32*(1): 333-365.

Paul, M. J., Bressler, D. W., Purcell, A. H., Barbour, M. T., Rankin, E. T., and V. H. Resh. 2009. Assessment tools for urban catchments: defining observable biological potential. *JAWRA Journal of the American Water Resources Association*, *45*(2): 320-330.

Paul, M. 2015. *A Review of Biological Condition Ratings in Urban Settings in North Carolina for the North Carolina Water Resources Research Institute*. Water Resources Research Institute, North Carolina State University.

Qian, S. S., and Z. Shen. 2007. Ecological applications of multilevel analysis of variance. *Ecology*, *88*(10): 2489-2495.

Qian, S. S., Cuffney, T. F., Alameddine, I., McMahon, G., and K. H. Reckhow. 2010. On the application of multilevel modeling in environmental and ecological studies. *Ecology*, *91*(2): 355-361.

Rabalais, N.N., R.E. Turner, Q. Dortch, D. Justic, V.J. Bierman Jr, and W.J. Wiseman. 2002. Nutrient-enhanced productivity in the northern Gulf of Mexico: Past, present and future. *Hydrobiologia* 475(1): 39-63.

Schueler, T. R. 1994. The importance of imperviousness. *Watershed protection techniques*, *1*(3): 100-111.

USEPA (United States Environmental Protection Agency). 2006. Wadeable Stream Assessment. Office of Research and Development. Office of Water. EPA 841-B-06_002.

USEPA (United States Environmental Protection Agency). 2016. Office of Water and Office of Research and Development. National Rivers and Streams Assessment 2008-2009: A Collaborative Survey (EPA/841/R-16/007). Washington, DC.

Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* 84(6): 1382-1394.