INVESTIGATING LANDSCAPE-STREAM WATER QUALITY RELATIONSHIPS AND STREAM WATER QUALITY PRESERVATION STRATEGIES IN THE TEXAS GULF REGION USING A HYBRID OF MACHINE LEARNING AND HYDROLOGICAL MODELING APPROACH

By

Runzi Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Planning, Design and Construction—Doctor of Philosophy

2020

ABSTRACT

INVESTIGATING LANDSCAPE-STREAM WATER QUALITY RELATIONSHIPS AND STREAM WATER QUALITY PRESERVATION STRATEGIES IN THE TEXAS GULF REGION USING A HYBRID OF MACHINE LEARNING AND HYDROLOGICAL MODELING APPROACH

By

Runzi Wang

This research investigates how land use, urban development pattern, topography, soil, climate, and population influence the stream nitrate (NO_3^--N), ammonium (NH_4^+-N), orthophosphate (PO_4^{3-}-P), total phosphate (TP), and *Escherichia coli* (*E.coli*) concentrations in the Texas Gulf Region. Specifically, the study focuses on how the land-stream water relationship varies by different sample sites, basins, ecoregions, and different years between 1991 and 2011. It also examines the benefits of compact urban development and verifies the management strategies to place best management practices (BMP) in hydrologically sensitive areas (HSAs).

The 2011 cross-sectional study in the Texas Gulf Region indicates that the connectedness of developed areas and the adjacencies between developed areas and other land covers were more significant than the percentage of developed areas in their effect on stream water quality. The relationships between landscape factors and stream water quality varied by season, location, and pollutant category, with these associations generally stronger in dry seasons and in coastal suburban watersheds. Using a random forest machine learning algorithm, a predictive model demonstrated that high density aggregated urban development is the most effective in protecting stream water quality. The predicted average dry season NO_3^-N and TP concentrations were 0.17 mg/l and 0.09 mg/l in high density aggregated scenarios, compared to 1.2 mg/l and 0.28 mg/l in the current sprawled development scenario.

The longitudinal study from 1991-2011 confirms the effects of controlling developed areas and agricultural areas in improving stream water quality. With the derived annual land cover composition and longitudinal nutrient and *E.coli* concentration data, it was found that adding 1 percent of developed area led to a 6.31% increase of NO_3^--N concentration and a 3.52% increase of PO_4^{3-}-P concentration in the Texas Gulf Region. Some unobserved characteristics led to high nutrient concentrations in the Middle Colorado-Concho and the Lower Trinity basins, and high *E.coli* concentration in the San Jacinto basin. The relationships between land cover and stream water quality varied more at the local scale than basin and region scales; they did not change significantly in the 20 years between 1991 and 2011.

In the BMP siting strategy study, the effectiveness of placing BMP in HSAs was verified using a Soil & Water Assessment Tool (SWAT). The hydrological sensitivity of subbasins had a significantly nonlinear positive association with NO_3^-N concentrations. Defining HSAs as areas with the highest 2% hydrological sensitivity and designating them to be preserved as green space was the most effective in reducing NO_3^-N output. Generally, it was suggested that evidence-based ecological planning should incorporate performance evaluation with valid data-driven methods.

Overall, this research was one of the first empirical studies to demonstrate the water quality degradation consequence of urban sprawl and the advantage of compact urban development. Machine learning and big data approaches were proven to be powerful tools for scenario prediction in land use planning to forecast environmental impacts of different urban development patterns. This study also established a robust Texas regional scale longitudinal water quality modeling approach depending upon efficient data fusion techniques, which can guide multiscale land use planning and watershed management.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my major advisor, Dr. Ming-Han Li. Without Dr. Li's excellent mentorship, it is impossible for me to become an independent researcher and find a good place to continue my academic career. I am also very appreciated to my committee members at MSU for their support to develop my thesis. They are Dr. Jun-Hyun Kim, Dr. Mark Wilson, and Dr. Scott Loveridge. In addition, I feel grateful to my previous committee members at TAMU, Dr. Xiao Yu, Dr. Xinyuan Wu, and Dr. Sorin Popescu. They help me a lot in initiating this thesis proposal and passing the preliminary exam at TAMU.

I want to thank all the faculty at SPDC for inspiring my work. Dr. Yue Cui supported me financially to engage in the COPR research project. I also thank Dr. Galen Newman at TAMU for recommendations in my job search. I took lots of courses at both MSU and TAMU, and I feel grateful to all the instructors who taught me useful knowledge and skills to help me become an interdisciplinary researcher.

I appreciate all the great time I spent with my friends at both MSU and TAMU. My friends support me with knowledge and take care of me in my life. I also thank my colleagues on the SESYNC project team. My life and career will be totally different without my friends. I sincerely hope we can be life-long partners and always help each other.

Finally, I would like to deeply thank my husband Xuewen Zhang, who always supported me for more than ten years. He is the best engineer, friend, and life-long partner I can think of. I also want to thank my parents, and my parents in law for their love and understanding.

Runzi Wang

7/28/2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AI Aggregation Index

AIC Akaike information criterion

AREA_MD Median of Patch Area

BD Biomass Decrease

BI Biomass Increase

BMP Best Management Practice

CA Total (Class) Area

CIRCLE Median of Related Circumscribing Circle

COHESION Patch Cohesion Index

CONTAG Contagion

CONTIG_MD Median of Contiguity Index

CRP Clean Rivers Program

CSA Critical Source Area

CV Change Vector

DCIA Directly Connected Impervious area

DEM Digital Elevation Model

DIVISION Landscape Division Index

dNBR differenced Normalized Burn Ratio

dNDVI differenced Normalized Difference Vegetation Index

E.coli Escherichia coli

ED Edge Density

ENN_MD Median of Euclidean Nearest Neighbor Distance

FRAC_MD Median of Fractal Dimension Index

GEE Google Earth Engine

GLCM Grey Level Co-occurrence Matrix

GWR Geographically Weighted Regression

GYRATE_MD Median of Radius of Gyration

HSA Hydrologically Sensitive Area

HUC Hydrologic Unit Code

IJI Interspersion Juxtaposition Index

LASSO Least Absolute Shrinkage and Selection Operator

LID Low Impact Development

LPI Largest Patch Index

LSI Landscape Shape Index

LS factor Length and Steepness factor

MESH Effective Mesh Size

MSE Mean Square Error

MSIDI Modified Simpson's Diversity Index

MSIEI Modified Simpson's Evenness Index

NDVI Normalized Difference Vegetation Index

NH_4^+-N Ammonium

NLCD National Land Cover Database

NLS Non-linear Least Squares

NO_3^--N Nitrate

NP Number of Patches

NPS Nonpoint Source Pollution

NSE Nash–Sutcliffe efficiency

OLS Ordinary Least Squares

PAFRAC Perimeter-Area Fractal Dimension

PARA_MD Median of Perimeter-Area Ratio

PCA Principal Component Analysis

PD Patch Density

PLADJ Proportion of Like Adjacencies

PLAND Percentage of Landscape

PR Patch Richness

PRD Patch Richness Density

PO_4^{3-}-P Orthophosphate

RCVMAX Relative Change Vector MAXimum

RF Random Forest

RUSLE Revised Universal Soil Loss Equation

SHAPE_MD Median of Shape Index

SHDI Shannon's Diversity Index

SHEI Shannon's Evenness Index

SIDI Simpson's Diversity Index

SIEI Simpson's Evenness Index

SPLIT Splitting Index

SSURGO Soil Survey Geographic Database

SWAT Soil and Water Assessment Tool

SWQM Surface Water Quality Monitoring

TCEQ Texas Commission on Environmental Quality

TE Total Edge

TOPMODEL TOPography based hydrological MODEL

TP Total Phosphorous

TRI Terrain Ruggedness Index

TWI Topographic Wetness Index

VSA Variable Source Area

CHAPTER 1 INTRODUCTION

In Texas, 410 out of a total of 1214 water bodies did not meet the applicable water quality standards or were threatened for one or more designated uses according to a 2012 Texas Commission on Environmental Quality (TCEQ) integrated report. Nonpoint source (NPS) pollution contributes to 45% of stream water quality impairment and 48% of lake water quality impairment (TCEQ, 2014). NPS pollution that results from a variety of sources such as lawns, construction areas, farms, and highways is difficult to control. To address the issue, Texas has Watershed Protection Plans to protect and restore stream water quality on a watershed basis across multiple jurisdictions. Therefore, technical support is pressingly needed to meet the complex challenge of stream water quality management at the watershed scale, especially from the NPS pollution point of view.

It is a general understanding that land use practices including urbanization, agricultural intensification, and deforestation are dominant drivers in influencing stream water quality (Yu et al., 2013; Manfrin et al., 2016; Zhang et al., 2017). However, this conclusion is sometimes not well applied to the local water environment because there are considerable differences in the relationships between stream water quality and local landscape features in different regions and basins (Ding et al., 2016). In Texas, there are a few studies investigating lake and reservoir water quality, while research efforts on stream water quality are very limited (Santhi et al., 2006; Patino et al., 2014). It is necessary for new studies to provide scientific and technical support in managing stream water quality in response to changing landscapes in Texas. This kind of support will benefit the formulation and implementation of stream water quality conservation policies and practices.

Stream water quality is related to many natural and anthropogenic factors such as land use composition, landscape configuration, topography, geology, climate, hydrology, and socioeconomic factors. The interactions between these explanatory factors are also complex

depending on spatial and temporal scales. To uncover the complicated nonlinear land-water relationships accurately and explicitly, several knowledge gaps need to be addressed. Firstly, there are fewer predictive studies compared to the common interpretation studies. If stream water quality can be predicted accurately with landscape characteristics, it will inform urban planning and watershed management policy makers about stream water quality under specific planning scenarios. Additionally, although the variation in the stream water quality is well explained by landscape factors using conventional statistical models, it is not guaranteed that the derived quantitative relationship could be generalized to new planning scenarios. Secondly, most previous studies investigating land-water relationships are cross-sectional studies, which are often criticized due to their relatively weak internal validity. Some research based on data from multiple years always treats samples from different years independently. Longitudinal research is thus needed to model how the land-water relationships change with long-term urban development, considering the dependency in stream water quality data from multiple years. Thirdly, to control stream water pollution, although some research has proposed prioritized sites to place best management practices (BMP) or low impact development (LID) practices to treat contaminants before they enter the streams, there are few studies verifying the effectiveness of BMP and LID siting strategies. For example, it is suggested that LID and BMP be placed in HSAs, which is a small portion of the watershed more susceptible to producing runoff (Walter et al., 2000, Martin-Mikle et al., 2015). However, empirical studies to verify the function of HSAs as critical source areas (CSAs) of pollution is still needed.

The overall goal of this study is to understand the complex relationships between landscape characteristics and stream water quality in the Texas gulf region with advanced analytical methods;

specifically, a combination of conventional statistical models, machine learning algorithms, and hydrological models. The three objectives below will be addressed in the following three chapters:

1) Chapter 2 focuses on predicting stream water quality in the Texas gulf region with landscape characteristics, with the focus on urban developed pattern. It also interprets variations in stream water quality with the most important landscape features with the consideration of spatial variability. The importance of urban development density and urban area configuration on stream water quality is verified.

2) Chapter 3 investigates the changing relationship between land use and stream water quality in the Texas gulf region from 1991 to 2011. It discovers how the variations in the land-water relationships are partitioned spatially and temporally. It also generates annual land cover maps from 1991 to 2011 to match the temporal resolution of the stream water quality data.

3) Chapter 4 confirms that placing BMPs in HSAs is efficient in reducing nutrient loadings in streams with hydrological models. It proposes an interdisciplinary data-driven framework to make suggestions to ecological planning and design.

The significance of this study is to apply big data and cutting-edge technologies to frame a large-scale longitudinal study in the landscape architecture discipline. Several key questions related to the stream water quality in Texas were answered, including urban developed pattern impact, regional-scale spatial variations, temporal changes and causal inference, and target management practices. It serves as a comprehensive and multidimensional theoretical and technical guide to the sustainable stream water quality management in Texas.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ding, J., Jiang, Y., Liu, Q., Hou, Z., Liao, J., Fu, L., & Peng, Q. (2016). Influences of the land use pattern on water quality in low-order streams of the Dongjiang River basin, China: a multi-scale analysis. *Science of the total environment*, *551*, 205-216.Manfrin, A., Bombi, P., Traversetti, L., Larsen, S., & Scalici, M. (2016). A landscape-based predictive approach for running water quality assessment: a Mediterranean case study. *Journal for nature conservation*, *30*, 27-31.

Martin-Mikle, C. J., de Beurs, K. M., Julian, J. P., & Mayer, P. M. (2015). Identifying priority sites for low impact development (LID) in a mixed-use watershed. *Landscape and urban planning*, *140*, 29-41.

Patiño, R., Dawson, D., & VanLandeghem, M. M. (2014). Retrospective analysis of associations between water quality and toxic blooms of golden alga (Prymnesium parvum) in Texas reservoirs: Implications for understanding dispersal mechanisms and impacts of climate change. *Harmful Algae*, *33*, 1-11.

Santhi, C., Srinivasan, R., Arnold, J. G., & Williams, J. R. (2006). A modeling approach to evaluate the impacts of water quality management plans implemented in a watershed in Texas. *Environmental modelling & software*, *21*(8), 1141-1157.

Texas Commission on Environmental Quality. (2014). *Managing nonpoint source pollution in Texas, 2013 annual report*

Walter, M. T., Walter, M. F., Brooks, E. S., Steenhuis, T. S., Boll, J., & Weiler, K. (2000). Hydrologically sensitive areas: variable source area hydrology implications for water quality risk assessment. *Journal of Soil and Water Conservation*, *55*(3), 277-284.

Yu, D., Shi, P., Liu, Y., & Xun, B. (2013). Detecting land use-water quality relationships from the viewpoint of ecological restoration in an urban area. *Ecological Engineering*, *53*, 205-216.

Zhang, L., Karthikeyan, R., Bai, Z., & Srinivasan, R. (2017). Analysis of streamflow responses to climate variability and land use change in the Loess Plateau region of China. *Catena*, *154*, 1-11.

CHAPTER 2 PREDICTING STREAM WATER QUALITY UNDER DIFFERENT URBAN DEVELOPMENT PATTERN SCENARIOS WITH A MACHINE LEARNING APPROACH

2.1 Introduction

Human-induced land use, such as urban and industrial land use, is recognized as a dominant factor affecting stream water quality. For example, a small increase in the percentage of urban land use has been found to exert a disproportionately large influence on pollutant generation (Ai et al., 2015; Giri and Qiu, 2016; Oeding et al., 2018; Sun et al., 2011; Wijesiri et al., 2018). Within a similar percentage of urban developed areas, varying patterns of urban development can contribute to considerable differences in stream water quality due to different pollutant generation, built-up and wash-off processes (Goonetilleke et al., 2005; Liu et al., 2012). Therefore, stream water quality prediction in various locations, densities, and patterns of urban development can serve as a basis for developing sound stream water quality management schemes (Fan and Shibata, 2015; Holcomb et al., 2018). However, the specific influence of urban development patterns on stream water quality, as well as the influence of spatial and temporal dynamics, remains unclear.

Urban development pattern has complex influences on stream water quality as measured by the interactions between area, shape, edge, aggregation of urban areas, and stream pollutant concentrations (Forman, 2014; Sun et al., 2014; Yu et al., 2013;). Theoretically, large areas of directly connected impervious areas (DCIA) have been shown to harm downstream water bodies (Del Monaco, 2017; Jones et al., 2005; Obropta and Del Monaco; 2018 Sohn et al., 2019). However, this does not necessarily mean urban development should be more dispersed to reduce DCIA as it can lead to potential ecosystem fragmentation and difficulty in implementing management practices (Bu et al., 2014; Shi et al., 2017). The ambiguity regarding whether intact or fragmented urban areas cause stream water degradation can be seen in the contradictory conclusions of

investigations between urban development pattern and stream water quality. Some researchers have argued that intact urban patterns with large amounts of impervious surface can contribute to water quality deterioration (Ding et al., 2016; Li et al., 2009). However, other studies found that greater interspersion of urban areas, as indicated by high Contiguity Index and Patch Cohesion Index significantly increased the export of pollutants due to the destruction of natural areas (Lv et al., 2014; Shi et al., 2013). More research is needed to address this question, particularly in terms of controlling the percentage of urban developed areas at the same levels. Doing so ensures different urban development patterns are comparable in terms of their influence on stream water quality.

One of the major challenges in quantifying stream water quality in accordance with factors of urban development patterns is to understand which factors are the most important/efficient in influencing stream water quality. Some studies have found that size and number of urban areas—as quantified by Patch Density, Largest Patch Index, and Edge Density—showed higher degrees of relationships to water quality compared to the isolation and connectedness of urban areas (Carey et al., 2011; Lee et al., 2009). Others have found that shape and aggregation of urban developed areas had a higher explanatory power in predicting stream water quality variations (Li et al., 2015; Yu et al., 2013). These varying results from previous studies regarding the correlation between urban development pattern and stream water quality have been attributed to two reasons. First, many studies reported important urban development pattern metrics at the local level using a small number of catchment samples (Li et al., 2015; Lintern et al., 2017; Sun et al, 2014). Thus, few studies have investigated the importance of urban development pattern in the context of a large heterogeneous area with a large watershed sample size. Second, there is a lack of more robust methods for improving the generalization of results regarding the importance of urban

development pattern metrics. For example, stepwise regression, the most commonly used algorithm for finding variable importance in predicting stream water quality, was found to sometimes generate problematic results due to approaches intent on only local optimization at each selection step (Harrell, 2017).

Furthermore, quantifying the relationships between stream water quality and urban development pattern necessitates the development of predictive models that can be used to forecast stream water quality in alternative urban planning scenarios (Avila et al., 2018; Holcomb et al., 2018; Molina-Navarro et al., 2020; Sharifi et al., 2017). Machine learning algorithms like boosted regression tree analysis, neural networks, and self-organizing maps have been applied to depict the complex, non-linear relationships between landscape characteristics and stream water quality with satisfactory model performance (Clapcott et al., 2012; Hameed et al., 2016; Kalteh and Berndtsson, 2008; Lek, 1999; Mirzaei et al., 2019). One advantage of machine learning application in stream water quality prediction is the possibility of controlling the same percentage of urban developed area in scenario prediction to determine the partial effect these patterns have on stream water quality. The other advantage is that after the accuracy of machine learning model is tested on a new dataset, the generalizability can be ensured and it can then be applied to forecast stream water quality under future land use planning scenarios to support policy decision-making (Chermack et al., 2008; Schreiber et al., 2019). Although stream water quality prediction with different land use scenarios has been explored in such predictive studies using machine learning algorithms, to our best knowledge, very few studies focus on the impact of urban development pattern.

The goal of this study is thus to provide a comprehensive understanding of how different urban development patterns influence stream water quality, covering the aspects of important factors, spatial variations, predictive models, and potential mechanisms. Using the Texas Gulf Region as

the study site, stream water quality—represented by NO_3^--N, TP, and *E.coli* concentrations—was quantified and predicted by metrics of patterns of urban development, controlling for landscape spatial pattern, topography, soil, climate, and population. Specifically, this study has three objectives: 1) To identify the most important factors of urban development pattern that influence NO_3^--N, TP, and *E.coli* concentrations and suggest specific urban forms to protect stream water quality; 2) To uncover the seasonal and spatial non-stationary relationships between urban development pattern and stream water quality; and 3) To develop predictive models that can forecast stream water quality based on different scenarios of urban development densities and configurations as well as provide implications for land use planning.

2.2 Data and Method

2.2.1 Study Site

The study site was the Texas Gulf Region, which has an area of 471,080 km^2 (Figure 2-1). It is one of 21 water resource regions (HRU 02) in the United States, consisting of 11 subregions (HRU 04) and 23 basins (HRU 06). The climate of this region is diverse, with a maritime climate along the coast, a continental climate in the central and northern areas, and a dry and hot climate in the west. These diverse climates lead to heterogeneous landscapes across the region. From east to west, the terrain ecosystem changes from coastal swamps and piney woods to rolling plains and rugged hills. The heterogeneity of these climate and landscape factors provide ideal samples for studying their influences on stream water quality.

Moreover, the increasing population in the study site has resulted in problems associated with urban sprawl, which has put natural forest areas at risk and degraded stream water quality. Texas currently has a population of approximately 29 million, with a growth rate of 1.8% every year (World Population Review, 2019). Nonpoint source pollution closely related to urban expansion

contributes to 45% of stream water quality impairment in Texas. Bacteria, nutrients, dissolved oxygen, and organics are the major causes of stream water quality degradation (Texas Commission on Environmental Quality, 2014). I therefore selected NO_3^--N, TP, and *E.coli* concentrations as the contaminants of interest in this study. Other common pollutants such as Total suspended solid and heavy metal were not included because of the data quality and availability.

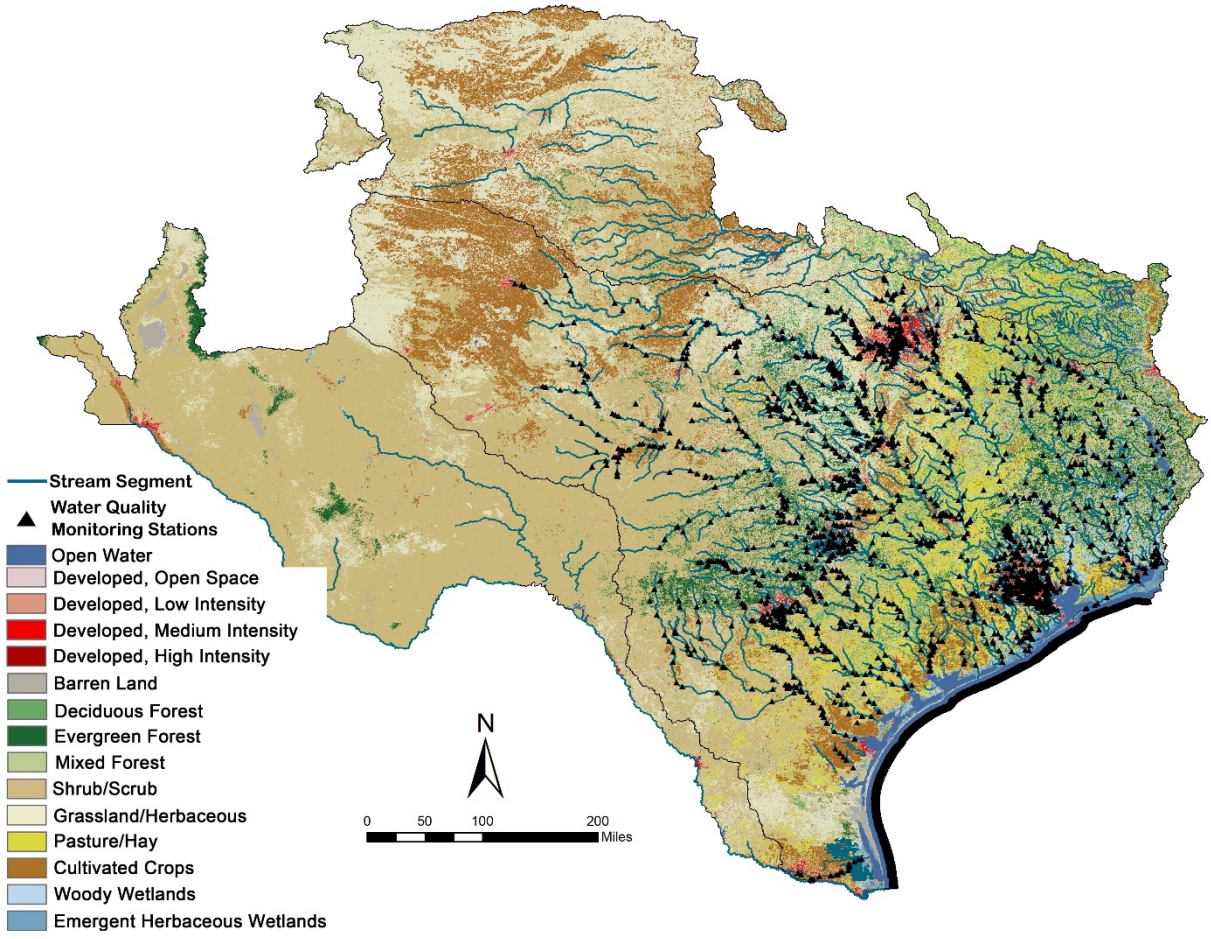


Figure 2-1. Study Site

2.2.2 Data and variables

Pollutant concentration data from 1,047 sampling stations in the Texas Gulf Region were used as predicted variables in this study. To monitor and assess stream water quality, the Texas

Commission on Environmental Quality's (TCEQ) Surface Water Quality Monitoring (SWQM) Program has installed over 3,000 active monitoring stations throughout the region. Pollutant concentration data in 2011 were obtained from the SWQM program and aggregated in dry and wet seasons by taking the average values. According to the monthly average precipitation in Texas, the dry season went from November to April and the wet season occurred the rest of year (Pratt and Chang, 2012).

Landscape metrics at both class and landscape levels, climate, soil, topography, and population were included as explanatory variables to explain variations in stream water quality. The class level metrics included land covers of developed area, developed open area, forest area, and planted area, which have been demonstrated to be major environmental drivers of changes in stream water quality (Clement et al., 2017; Glinska-Lewczuk et al., 2016; Teklu et al., 2016). Our analytical steps focused on metrics from urban development pattern and used other metrics as control variables. The definition of all land covers was in accordance with NLCD (Homer et al., 2015), and all variables in this study and their corresponding data sources are presented in Table 2-1.

Table 2-1. Data sources

Dataset	Structure	Variables	Spatial Resolution
NLCD 2011: USGS National Land Cover Database	Raster	landscape metrics	30m
Tiger census block	Shapefile	population, population density	NA
USGS National Elevation Dataset 1/3 arc-second	Raster	elevation, slope	0.33 arc seconds/ 30m
PRISM Monthly spatial climate dataset AN81m	Raster	precipitation, temperature	2.5 arc minutes/ 5km

Table 2-1 (cont'd)

SSURGO database	Shapefile	hydrological soil groups, soil storage depth	1:12000
TCEQ SWQM program	Table	stream pollutant concentration	NA

I incorporated high dimensions of landscape metrics in the machine learning models, including 76 class level metrics and 32 landscape level metrics in the categories of area, edge, shape, and contagion/interspersion (McGarigal, 1995), as presented in Table 2-2. It was assured that correlated variables would not cause multi-collinearity issues in the machine learning models and a large set of features can potentially increase predicting accuracy.

Table 2-2. Explanatory variables

Category	Subcategory	Variable
Class Level Metrics (76) [1] (including classes of developed open area, developed area, forest area, and planted area)	Area (28)	Percentage of Landscape (PLAND), Total Area (CA), Median of Patch Area (AREA_MD), Median of Radius of Gyration (GYRATE_MD), Largest Patch Index (LPI), Number of Patches (NP), Patch Density (PD)
	Edge (8)	Total Edge (TE), Edge Density (ED)
	Shape (20)	Median of Perimeter-Area Ratio (PARA_MD), Median of Shape Index (SHAPE_MD), Median of Fractal Dimension Index (FRAC_MD), Median of Related Circumscribing Circle (CIRCLE), Median of Contiguity Index (CONTIG_MD)
	Contagion/Interspersion (20)	Landscape Division Index (DIVISION), Splitting Index (SPLIT), Interspersion Juxtaposition Index (IJI), Landscape Shape Index (LSI), Patch Cohesion Index (COHESION)

Table 2-2 (cont'd)

Landscape Level Metrics (32)	Area (6)	Total Area (CA), Largest Patch Index (LPI), Median of Patch Area (AREA_MD), Median of Radius of Gyration (GYRATE_MD), Number of Patches (NP), Patch Density (PD)
	Edge (2)	Total Edge (TE), Edge Density (ED)
	Shape (6)	Perimeter-Area Fractal Dimension (PAFRAC), Median of Perimeter-Area Ratio (PARA_MD), Median of Shape Index (SHAPE_MD), Median of Fractal Dimension Index (FRAC_MD), Median of Related Circumscribing Circle (CIRCLE), Median of Contiguity Index (CONTIG_MD)
	Contagion/Interspersion (10)	Landscape Division Index (DIVISION), Splitting Index (SPLIT), Effective Mesh Size (MESH), Interspersion Juxtaposition Index (IJI), Landscape Shape Index (LSI), Patch Cohesion Index (COHESION), Contagion (CONTAG), Proportion of Like Adjacencies (PLADJ), Aggregation Index (AI), Median of Euclidean Nearest Neighbor Distance (ENN_MD)
	Diversity (8)	Patch Richness (PR), Patch Richness Density (PRD), Shannon's Diversity Index (SHDI), Simpson's Diversity Index (SIDI), Modified Simpson's Diversity Index (MSIDI), Shannon's Evenness Index (SIEI), Simpson's Evenness Index (SIEI), Modified Simpson's Evenness Index (MSIEI)
Climate (24)	Precipitation (12)	Monthly Precipitation, Seasonal Average Precipitation
	Temperature (12)	Monthly Temperature, Seasonal Average Temperature
Topography (2)		Elevation, Slope

Table 2-2 (cont'd)

Soil (6)	Soil Storage, the Presence of Hydrologic Soil Groups A, B, C, D, C/D, B/D
Population (2)	Population, Population Density

I added environmental and social control variables including precipitation, temperature, slope, elevation, soil type, soil storage depth, population, and population density to control for model bias. In terms of the climatic variables, seasonal total precipitation and mean temperature were included in the statistical models to simplify interpretation. Monthly total precipitation and mean temperature were used in the machine leaning models to facilitate higher predicting accuracy. In this study, soil type referred to hydrological soil groups (HSG). HSG A, B, C, and D have a high infiltration rate, a moderate infiltration rate, a slow infiltration rate, and a very slow infiltration rate, respectively. If a soil was placed in HSG D because of a high-water table, it might be assigned to a dual hydrologic group such as A/D, B/D, or C/D. The first letter of the pair represented the soil's group if drained and the second letter, D, represented the natural drainage condition.

2.2.3 Data analysis

As presented in Figure 2-2, I first applied LASSO regression to identify whether urban development patterns were the dominant factors in determining stream water quality among all the catchment characteristics. GWR models were then developed to understand the spatial variation of the relationships between urban development pattern and pollutant concentrations. RF regression was used to train machine learning models to predict stream water quality. After confirming test set accuracy using RF regression was satisfactory, the final model was employed to predict stream water quality under four scenarios of different urban development patterns.

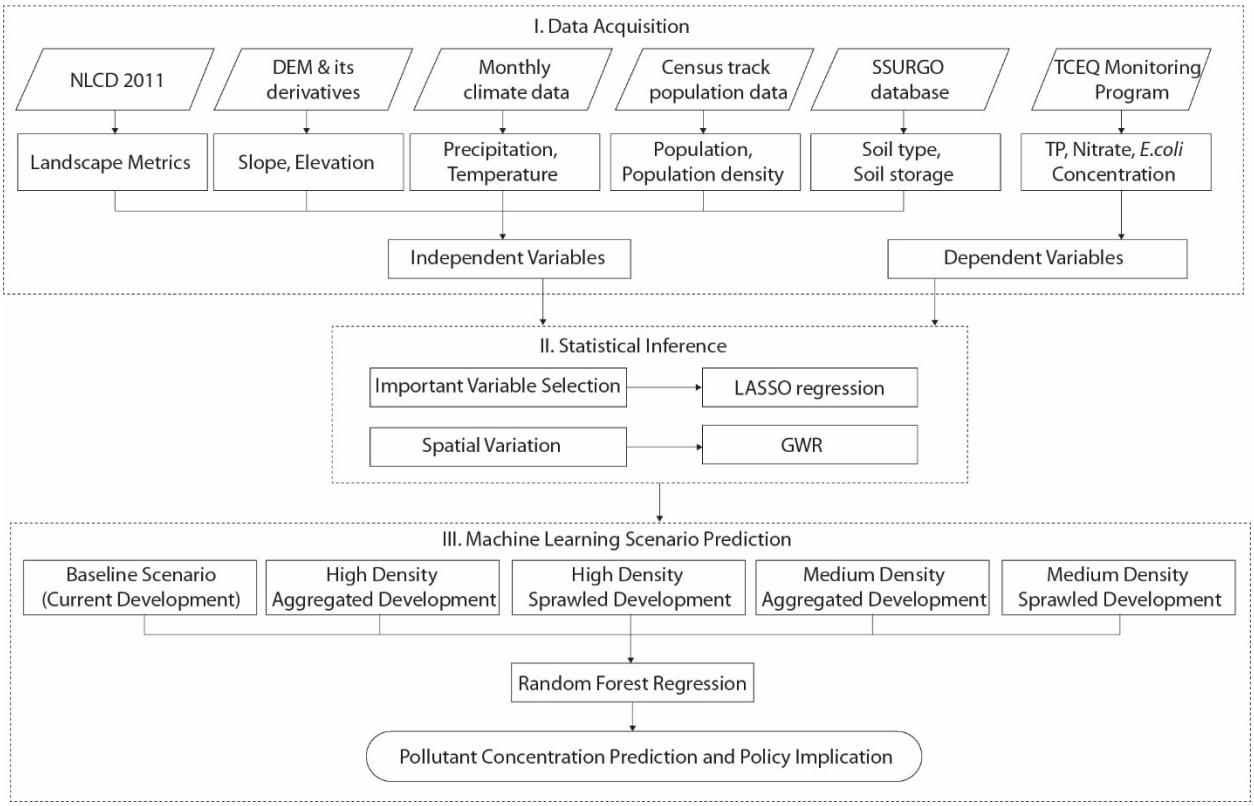


Figure 2-2. Data analysis flowchart

- **LASSO Regression**

LASSO regression was employed to select for important factors in stream water quality while minimizing prediction error. LASSO regression results identified key factors in urban development patterns that determine stream water quality. The results were also used to select other important catchment characteristics. LASSO regression is a machine learning method that performs both variable selection and regularization to improve prediction accuracy and a regression model's interpretability (Tibshirani, 1996). It selects only a subset of covariates by forcing the sum of the absolute value of regression coefficients to be less than a fixed value, which forces some variable coefficients to be set to zero. Variables with non-zero coefficients are then considered more important in predicting the outcomes. The objective of LASSO regression is to solve Equation 2-1, where y_i is the outcome and x_i is the covariate vector. The parameter t, which

determines the amount of regularization, is tuned throughout the cross-validation process. Compared to the common stepwise regression approach used widely in previous studies assessing stream water quality, LASSO regression has the advantage of reaching a global rather than local optimization to make a prediction. With the cross-validation process tuned to the hyperparameter *t*, LASSO regression also guarantees model generalization in a new dataset. I implemented LASSO regression in "scikit-learn" and "statsmodels" packages in Python 3.0.

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}\left|\beta_j\right|\le t$$

Equation 2-1

- **Geographically Weighted Regression (GWR)**

In this study, GWR was applied to investigate the spatially varying associations between urban development pattern metrics selected by LASSO regression and stream water quality. GWR allows linear predictors to be a function of spatial coordinates *(u, v)*, as represented in Equation 2-2. In this equation, y is the pollutant concentration, x_j is the covariate vector, and β_j is the corresponding vector coefficient. GWR assumes that the contribution of each sample to the local regression model is weighed according to its proximity to the local sample point. A common choice of weighting function is the Gaussian curve, as shown in Equation 2-3, where d_{ij} is the distance between observation point *i* and the realization point *j*, and the bandwidth *b* is the parameter to be determined. An adaptive kernel bandwidth was employed in this study in accordance with the judgement of AIC. GWR was implemented in "spgrw" package in R.

$$y=\sum_{j=1}^{p+1}\beta_j(u,v)x_j$$

Equation 2-2

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2b^2}\right)$$

Equation 2-3

- **Random forest regression**

RF regression was used to train models to quantify the nonlinear relationships between explanatory variables and stream water quality. It was further applied to scenario predictions of pollutant concentrations in accordance with different urban development patterns. RF is an ensemble learning method that consists of a large number of individual decision trees. Random samples are taken with replacement and a random subset of features are used to generate each regression decision tree. A prediction is made by averaging the results of all regression trees (Breiman, 2001).

To guarantee the generalization of the predictive models, 90% of the samples were used to train the models and the remaining sample was used to test the models' performance metrics, including Mean Square Error (MSE) and R^2 (Wang et al., 2019). Ten-fold cross validation was employed to train the hyperparameters including the maximum depth of the tree (max_depth), the minimum number of samples required to split a node (min_sample_split), and the maximum number of features to look for the best split (max_features) using a grid search fashion. The number of regression trees were set to be 1,000. Random forest regression was also implemented in Python 3.0 "scikit-learn" package.

2.2.4 Scenario Design

To understand the effects of urban developed density and configuration on stream water quality, I created four alternative urban development scenarios in the upstream area of The Woodland, TX

and predicted their pollutant concentrations of NO_3^--N, TP, and *E.coli* in both dry and wet seasons (Figure 2-3). The Woodlands was well-known for Ian McHarg's ecological planning approach (McHarg and Sutton, 1975). The current development condition was chosen as the baseline scenario, where 33.6% of the area (24 km^2) was developed into urban areas. Low density development is the major development type in the current condition. The boundary of the scenario site is the Bear Branch-Panther Branch sub-watershed boundary with the HUC12 ID 120401020211.

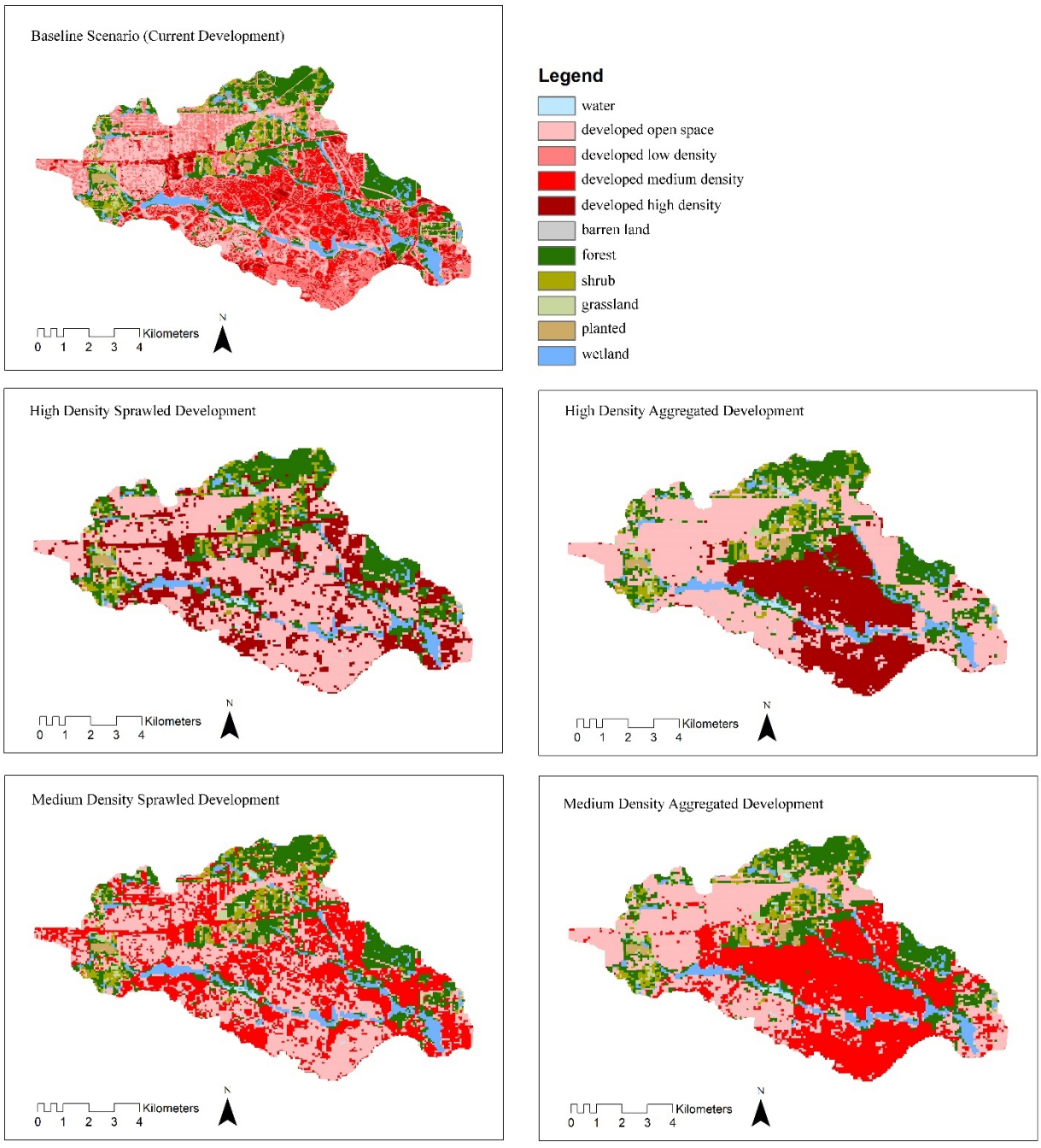


Figure 2-3. Scenario Maps

The alternative scenarios included four extreme development scenarios where developed areas were extremely scattered or aggregated: high-density aggregated development, high-density sprawl development, medium density aggregated development, and medium density sprawl

development (Figure 2-3). I applied two criteria to create the four development scenarios. First, the total impervious surface area was the same as the baseline scenario. According to the land cover description of NLCD, impervious surface accounts for 20%-49% in low-density development, 50%-79% in medium density development, and 80%-100% in high-density development. To quantify impervious surface in urban areas for each density, I used the median value of the impervious surface percentage, which were 35%, 65%, and 90% for low density, medium density, and high-density developments, respectively (Yang and Li, 2011). The impervious surface area added up to be 16.4 km2 in all scenarios. Second, all of the existing land cover types in the baseline scenario—including water, forest, grassland, planted, and wetland—stayed the same. The reduced urban areas in the four alternative scenarios were changed to forest areas that represent undeveloped conditions. To approximate the maximum degree of aggregated/sprawl development, I manually chose locations of high/medium density development that had changed to forest areas in ArcGIS 10.5.

The key difference in each scenario was urban development patterns, as presented in Table 2-3. Compared to the two sprawled scenarios, the two aggregated scenarios were characterized by higher LPI, COHESION, lower ED, LSI, and shape complexity. Therefore, developed areas were clumped into larger patches with simpler shape and were more physically connected in the two aggregated scenarios. These differences in urban development pattern metrics laid the foundation for quantifying stream water quality with different urban densities and configurations.

Table 2-3. Scenario description

	Metrics	Baseline Scenario	High density Aggregated Scenario	High Density Sprawled Scenario	Medium Density Aggregated Scenario	Medium Density Sprawled Scenario
Urban development density	Impervious area (%)	16.4	16.4	16.4	16.4	16.4
	High density developed area (%)	1.4 (1.2)[1]	18.2 (16.4)	18.2 (16.4)	0	0
	Medium density developed area (%)	12.8 (8.3)	0	0	25.2 (16.4)	25.2 (16.4)
	Low density developed area (%)	19.4 (6.8)	0	0	0	0
Urban development configuration (represented by landscape metrics of developed area[2])	LPI	41.60	15.75	3.45	32.90	9.81
	NP	494	58	124	175	277
	ED (meters per hectare)	86.88	13.07	41.55	28.11	61.45
	FRAC_MD	622470	95850	304800	206100	448290
	CIRCLE_MD	0.4907	0.4123	0.4907	0.4123	0.4907
	LSI	27.34	6.10	19.33	10.78	23.54
	IJI (percent)	33.53	40.65	42.39	42.98	36.40
	COHESION	99.48	99.02	95.43	99.42	98.02

2.3 Result

2.3.1 Important catchment characteristics selected by LASSO regression

COHESION, IJI, and LPI of developed areas were found to be important in affecting TP in both dry and wet seasons (Table 2-4). When developed areas were more interspersed with other land cover types (indicated by IJI) and became less physically connected (indicated by COHESION), TP concentration was likely to reduce. Larger patches of developed area (indicated by LPI) were positively, significantly correlated with TP concentration in dry seasons. Because urban development pattern metrics were the only landscape metrics selected by the TP LASSO

regression, urban development patterns outweighed other land use patterns in affecting TP concentration. In addition, areas with a very low infiltration rate (indicated by soil group D) significantly contributed to low TP concentration. The presence of soil group C/D was significantly associated with high TP concentration in the wet season. High temperature, low forest percentage, and low slope catchments were significantly associated with high TP concentration. The important catchment characteristics affecting TP concentration in dry and wet seasons were similar, with larger and more significant effects in the dry season.

Table 2-4. LASSO linear regression results of TP concentration

		Wet season			Dry season		
		coefficient	t value	*p* value	coefficient	t value	*p* value
	Constant	-1.783	-41.932	<0.001	-1.946	-51.951	<0.001
Developed area class level metrics	COHESION	0.074	1.453	0.147	0.077	1.764	0.078
	IJI	-0.090	-1.553	0.121	-0.067	-1.438	0.151
	LPI	0.229	1.347	0.178	0.144	2.942	0.003**
	PLAND	-0.068	-0.375	0.708	n/a	n/a	n/a
Control variables	Soil storage	0.184	2.930	0.003**	0.224	4.283	<0.001**
	The presence of soil group D	-0.120	-2.444	0.015*	-0.102	-2.492	0.013*
	The presence of soil group C/D	0.100	2.18	0.028*	0.077	1.958	0.051
	slope	-0.091	-1.356	0.175	-0.064	-1.139	0.255
	population	0.097	2.018	0.044*	0.118	2.783	0.006**
	Percentage of forest area	-0.148	-2.268	0.024**	-0.142	-2.752	0.006**
	Mean temperature	0.157	2.675	0.008**	0.390	9.278	<0.001**
	elevation	-0.250	-3.392	0.001**	n/a	n/a	n/a
	The presence of soil group B/D	0.059	1.219	0.223	n/a	n/a	n/a

Note of Table 2-4, Table 2-5, and Table 2-6: * indicates the significance level of 0.05; ** indicates the significance level of 0.01

The number of important variables associated with *E.coli* concentration were found to be larger than those of TP concentration, which indicated a more complex mechanism, particularly in wet seasons (Table 2-5). Complex shape (indicated by SHAPE) of urban developed areas was positively and significantly related to *E.coli* concentration in wet seasons. Similarly, high edge density (ED) and shape complexity (SHAPE) of planted areas was found to be significantly and positively correlated with *E.coli* concentration in both dry and wet seasons. At the landscape level, the median of CONTIG had significant positive correlation with *E.coli* concentration in dry seasons, meaning that the high spatial connectedness of land cover patches was likely to increase *E.coli* concentration. Moreover, low soil storage capacity and low infiltration rates helped to significantly reduce *E.coli* concentration. High temperature, high soil storage, and the presence of soil group D all contributed to high *E.coli* concentration.

Table 2-5. Lasso linear regression results of *E.coli* concentration

		Wet season			Dry season		
		coefficient	t value	*p* value	coefficient	t value	*p* value
	Constant	4.123	53.486	<0.001	4.534	71.632	<0.001
Developed area class level metrics	PLAND	0.328	0.747	0.455	0.643	2.580	0.010**
	IJI	-0.121	-0.956	0.339	-0.111	-1.337	0.182
	LPI	0.250	0.744	0.457	n/a	n/a	n/a
	PD	-0.093	0.868	0.386	n/a	n/a	n/a
	Median of CIRCLE	-0.157	-1.608	0.108	n/a	n/a	n/a
	Median of SHAPE	0.213	2.619	0.009**	n/a	n/a	n/a
	DIVISION	n/a	n/a	n/a	-0.170	-1.259	0.209
Planted area class level metrics	ED	-0.286	-2.533	0.012*	n/a	n/a	n/a
	NP	-0.067	-0.572	0.567	n/a	n/a	n/a
	Median of SHAPE	n/a	n/a	n/a	0.250	3.369	0.001**
	PLAND	n/a	n/a	n/a	0.181	1.911	0.056
Forest area class level metrics	PLAND	-0.146	-1.362	0.174	-0.094	0.570	0.569
	Median of FRAC	0.134	1.207	0.228	n/a	n/a	n/a
	SPLIT	-0.105	-1.249	0.212	n/a	n/a	n/a
	Median of AREA	n/a	n/a	n/a	0.106	1.510	0.131
Landscape level metrics	Median of CONTIG	0.503	1.749	0.083	0.562	6.453	<0.001**
	Median of AREA	-0.103	-0.297	0.766	n/a	n/a	n/a
	Median of FRAC	0.162	0.579	0.562	n/a	n/a	n/a
	IJI	0.078	0.548	0.584	n/a	n/a	n/a
	MESH	-0.874	-0.835	0.404	n/a	n/a	n/a
	TE	n/a	n/a	n/a	-0.144	-2.088	0.037
	AI	n/a	n/a	n/a	-0.119	-1.679	0.094
Other control variables	Soil storage	0.526	4.567	<0.001**	0.455	5.187	<0.001**
	The presence of soil group D	-0.258	-2.956	0.003**	-0.160	-2.261	0.024**
	The presence of soil group C/D	0.181	2.220	0.027*	0.105	-2.261	0.024**
	Mean temperature	0.173	1.726	0.085	0.438	5.384	<0.001**
	Population density	0.254	1.171	0.242	0.098	0.570	0.569
	Population	0.196	2.035	0.042*	n/a	n/a	n/a

All aspects of urban development patterns were importantly associated with NO_3^--N concentration, including area, cohesion, adjacency, edge, shape, and area. Meanwhile, the mechanism of planted and forest areas were relatively simple (Table 2-6). In dry season, when the proportion of developed areas increased and these areas became more connected (indicated by COHESION), NO_3^--N concentration significantly decreased. When developed area became more interspersed to other land cover patches (indicated by IJI), NO_3^--N concentration significantly decreased as well. The percentages (PLAND) and connectedness (CONTIG) of planted areas were also positively and significantly correlated with NO_3^--N concentration in dry seasons. It was found that a simple and intact shape (PARA) of forest area significantly contributed to the reduction of NO_3^--N concentration in both dry and wet seasons. At the landscape level, the more fragmented (PAFRAC) the landscape, the higher and more significant the NO_3^--N concentration in dry season. NO_3^--N concentration was negatively associated with precipitation and thus, NO_3^--N concentration in the wet season was lower than in the dry season.

Table 2-6. Lasso linear regression results of NO_3^--N concentration in wet seasons

		Wet season			Dry season		
		coefficient	t value	*p* value	coefficient	t value	*p* value
	Constant	-0.867	-9.903	<0.001	-0.436	-5.730	<0.001
Developed area class level metrics	COHESION	0.141	1.101	0.272	0.324	2.968	0.003**
	IJI	-0.169	-0.749	0.454	-0.319	-3.044	0.003**
	LPI	0.169	1.110	0.268	n/a	n/a	n/a
	ED	0.003	0.024	0.981	n/a	n/a	n/a
	Median of AREA	n/a	n/a	n/a	0.174	2.157	0.032*
Developed open area class level metrics	Median of CONTIG	-0.103	-0.489	0.625	-0.124	-1.179	0.239
	Median of CIRCLE	-0.076	-0.369	0.713	n/a	n/a	n/a
Planted area class level metrics	PLAND	n/a	n/a	n/a	0.312	3.409	0.001**
	Median of CONTIG	n/a	n/a	n/a	-0.356	3.323	0.001**
Forest area class level metrics	PLAND	-0.242	-2.030	0.043*	-0.190	-1.849	0.065
	Median of PARA	0.261	2.441	0.015*	0.439	4.595	<0.001**
Landscape level metrics	PAFRAC	n/a	n/a	n/a	0.218	2.291	0.023*
Other control variables	The presence of soil group D	-0.206	-2.148	0.032*	-0.116	-1.450	0.148
	The presence of soil group C/D	0.228	2.499	0.017*	0.183	2.244	0.025*
	Population	0.113	0.981	0.327	0.068	0.704	0.482
	Mean temperature	0.426	3.792	<0.001**	0.706	7.052	<0.001**
	Soil storage	0.160	1.447	0.032*	n/a	n/a	n/a
	Mean precipitation	-0.396	-4.098	<0.001**	n/a	n/a	n/a

2.3.2 Spatial variation of the effects of urban development pattern on stream water quality

In this study, TP GWR performed better in coastal areas such as the Neches Basin, the Lower Brazos Basin, and the Central Texas Coastal Basin; with the R^2 higher than 0.4 (Figure 2-4). However, they did not perform equally well in the Houston metropolitan area and the agricultural watersheds like the Middle Brazos Basin. The performance of the *E.coli* GWR was also better in coastal areas, including the Galveston Bay-San Jacinto Basin and the Neches Basin.

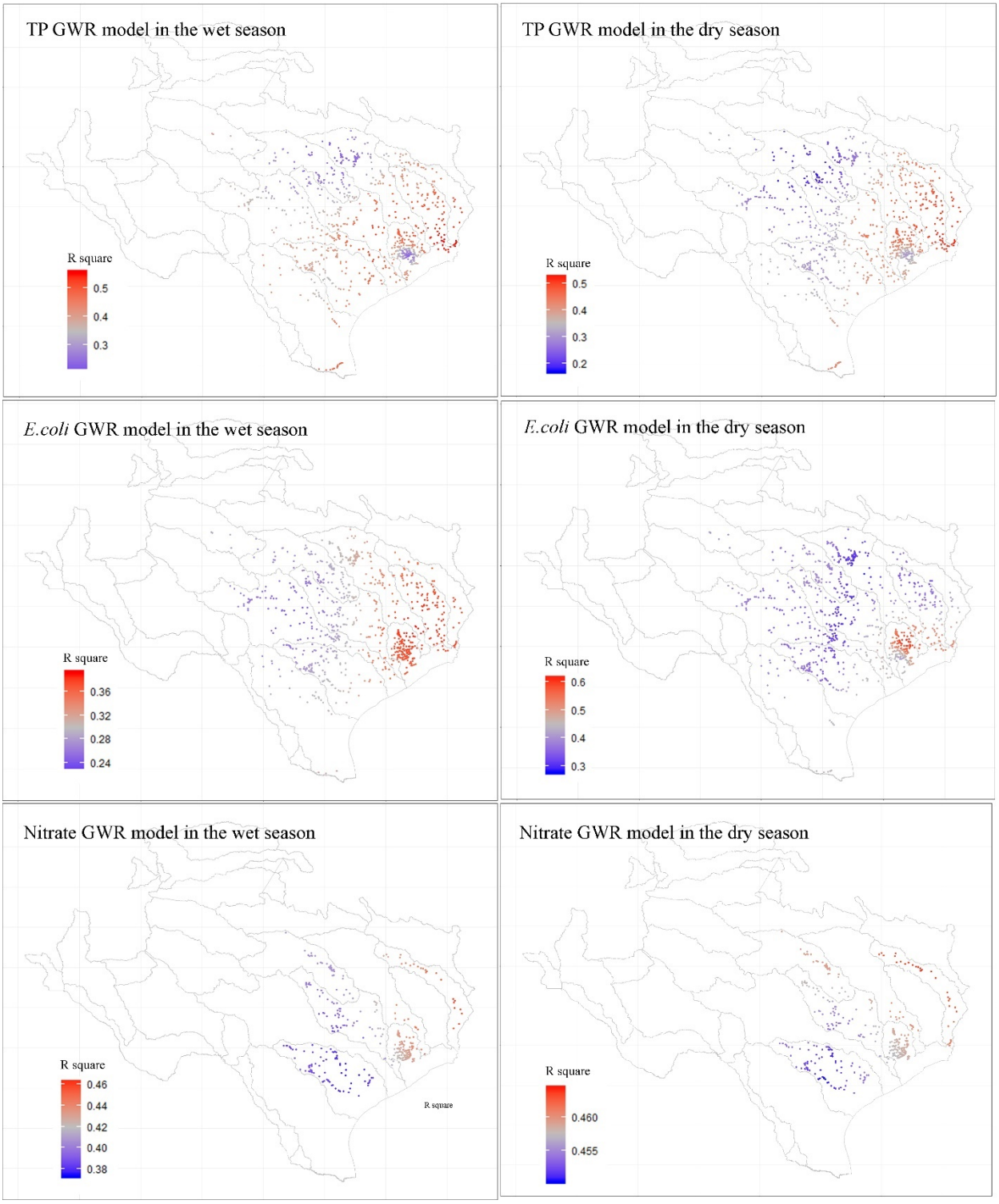


Figure 2-4. TP, *E.coli*, and NO_3^--N GWR model performance

Compared to LASSO regression, GWR performed better in predicting TP and *E.coli* concentrations, indicated by a lower AIC and higher R^2 (Table 2-7). The performance of NO_3^--N

models were similar between GWR and LASSO regressions, which was attributed to the relatively smaller sample size and spatial extent.

Table 2-7. Model performance comparison between Lasso linear regression and GWR

		N Observation	Model df	R^2		AIC	
				LASSO regression	GWR	LASSO regression	GWR
TP	wet season	804	13	0.32	0.49	2596	2394
	dry season	868	10	0.34	0.44	2645	2526
E.coli	wet season	754	22	0.29	0.32	3416	3365
	dry season	788	15	0.33	0.44	3158	3061
NO_3^--	wet season	329	14	0.42	0.42	1252	1242
N	dry season	355	13	0.45	0.46	1276	1259

Among the most important metrics of urban development, COHESION, i.e., the aggregation of urban developed areas, exerted a greater positive effect on TP concentration in the southern portion of the study area, which included the Nueces-Southwestern Texas Coastal Basin, the Central Texas Coastal Basin, the Lower Colorado-San Bernard Coastal Basin, and the Lower Brazos Basin (Figure 2-5). The effects of COHESION in the Galveston Bay-San Jacinto Basin and the Trinity Basin, in contrast, trended towards negative. When developed areas were more proportionally interspersed with other land cover types (higher IJI), TP concentration in the Central Texas Coastal Basin and the Galveston Bay-San Jacinto Basin were likely to decrease. Large patches of developed area (higher LPI) were shown to have a spatially heterogeneous effects on TP concentration. In the Central Texas Coastal Basin, the Lower Colorado-San Bernard Coastal Basin, and the Lower Brazos Basin, the effect was positive, and changed to negative in the Trinity Basin and most of the coastal areas.

Complex shape (SHAPE) and large patches (LPI) of urban developed areas had greater positive effects on *E.coli* concentration in coastal basins, including the Nueces-Southwestern Texas Coastal Basin, the Central Texas Coastal Basin, the Galveston Bay-San Jacinto Basin, the Sabine Basin, and the east parts of the Lower Colorado-San Bernard Coastal Basin and the Lower Brazos Basin.

SHAPE had negative effects on *E.coli* concentration in some agricultural basins such as the Middle Brazos Basin and west of the Lower Brazos Basin. The IJI of developed areas had a greater negative effect on *E.coli* concentration in the northwest part of the study area, including the Middle Brazos Basin and the Lower Brazos Basin. I discuss the mechanism that is potentially driving the spatial variation in the effects of urban development pattern in section 4.3.

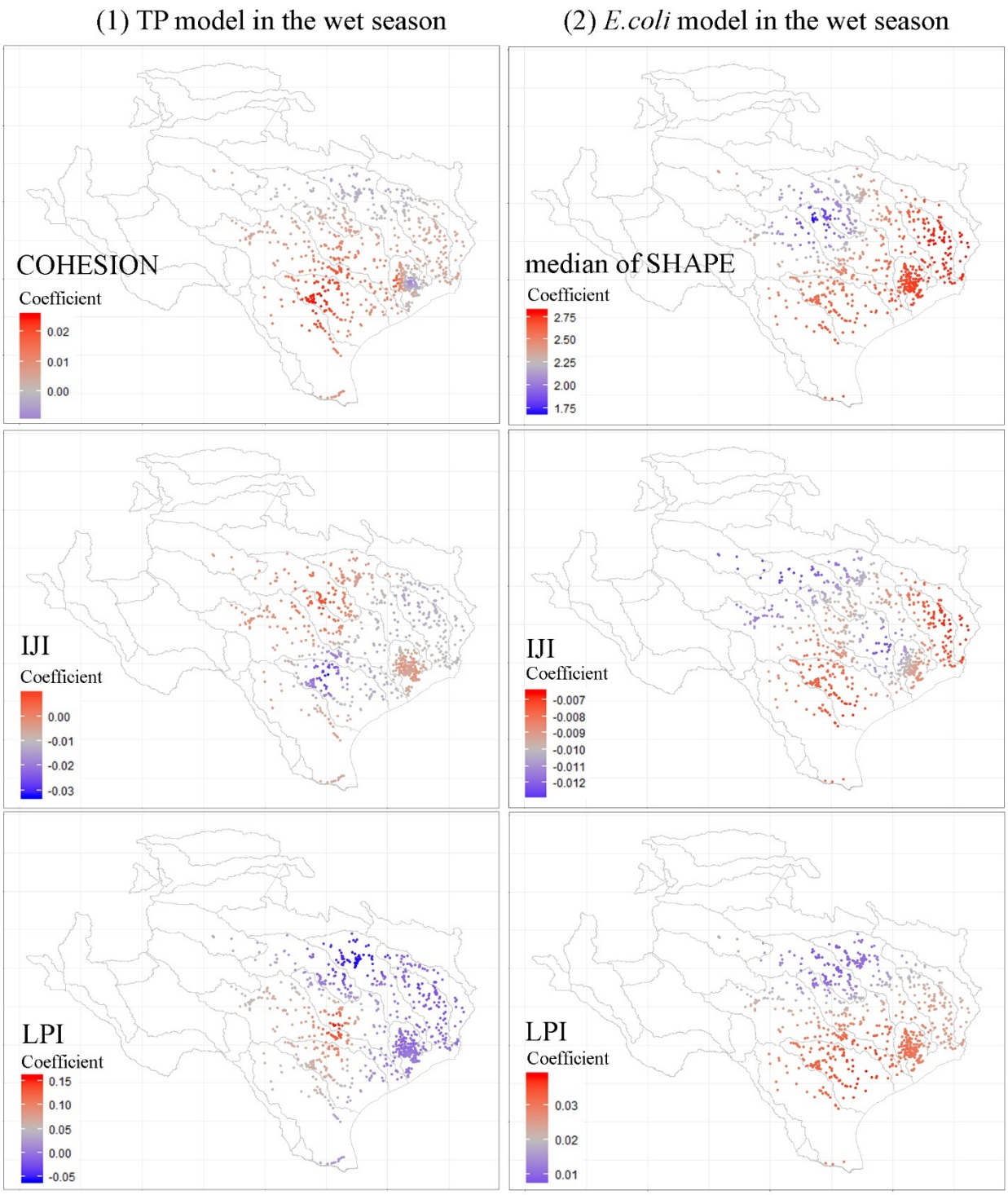


Figure 2-5. GWR model coefficients of urban development pattern effects in the wet season (TP model on the left and *E.coli* model on the right)

2.3.4 Stream water quality prediction under alternative planning scenarios

In the RF regression model, the variations and trends of all pollutant concentrations were well captured in the test set; however, the extreme values were not well predicted (Figure 2-6). The very low concentrations tended to be overestimated and the very high concentrations tended to be underestimated. In the wet season, the R^2 in the test set was 0.56, 0.45, and 0.66 for the TP, *E.coli*, and NO_3^-N RF models, respectively (Table 2-8). Similar to the GWR performance, RF predicting accuracy in the dry season were slightly higher than in the wet season.

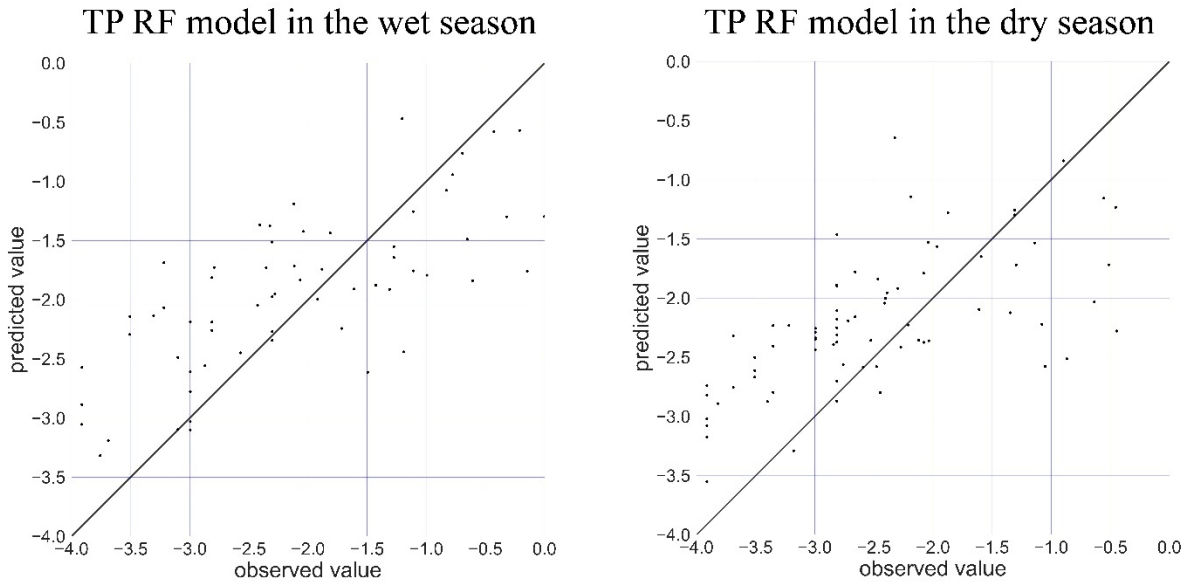


Figure 2-6. Scatter plots of predicted values against observed values of TP concentration in the test set

Table 2-8. Random forest prediction results

		Train set			Test set			Top 10 important variables
		correlation	R^2	MSE	correlation	R^2	MSE	
TP	wet season	0.98	0.96	0.16	0.74	0.55	1.23	IJI of forest area, precipitation, population, temperature, COHESION of developed area, population density, PAFRAC, soil storage
	dry season	0.98	0.96	0.15	0.75	0.56	0.85	ED of planted area, precipitation, temperature, slope, IJI of developed area, PAFRAC, COHESION of developed area, population density, PD of planted area, population, elevation, soil storage
E.coli	wet season	0.98	0.96	0.58	0.65	0.42	4.61	Temperature, percentage of planted area, ED of planted area, precipitation, COHESION of developed area, DIVISION of developed area, soil storage, population density, LPI of developed area, percentage of developed area
	dry season	0.98	0.96	0.35	0.67	0.45	2.62	Precipitation, IJI, temperature, population density, soil storage, percentage of developed area
NO_3^--N	wet season	0.97	0.94	0.35	0.80	0.64	1.75	Precipitation, population density, LPI of developed open area, temperature, soil storage, COHESION of developed area, population density, median of CIRCLE of forest area, median of GYRATE of forest area, LPI of forest area
	dry season	0.97	0.94	0.29	0.81	0.66	1.36	PAFRAC, elevation, temperature, soil storage, population, median of GYRATE of forest area, IJI of developed area, COHESION of developed area, LPI of forest area, precipitation, median of AREA of forest area, median of CONTIG of forest area

The importance of climatic factors was highlighted in the RF regression because using monthly average temperature and total precipitation data yielded much higher accuracy than that obtained through seasonal average temperature and total precipitation. It is therefore likely that climatic factors exhibited interaction effects with urban development patterns and other environmental variables on stream water quality. According to the variable importance of RF regression (Table 8), the aggregation and interspersion of developed areas (indicated by COHESION and IJI) were important in affecting TP and NO_3^--N concentrations, which aligned with the LASSO regression results. COHESION, LPI, and DIVISION of developed areas were important in affecting the concentration of *E.coli* in wet seasons. With respect to other landscape patterns, the ED of planted areas was significant for both TP and *E.coli* concentrations. Shape complexity and aggregation of forest areas were important factors in influencing NO_3^--N concentration. Landscape level metrics were found to not be as important as class level metrics.

The prediction results of the alternative planning scenarios suggested that high density aggregated development patterns were advantageous in reducing TP and NO_3^--N concentrations (Table 2-9). All high density and medium density compact developments had a lower than half concentration of all pollutants compared to the current development, indicating the benefits of small footprint urban areas. Aggregated development in both high and medium density scenarios had lower TP and NO_3^--N concentrations when compared to sprawl development of the same density. However, aggregated development contributes to higher *E.coli* concentrations than sprawl development of the same density in wet seasons. Specifically, for TP concentration, two sprawled development scenario result in the hypereutrophic conditions, while two aggregated development scenarios result in eutrophic conditions in lotic ecosystems (Grand River Water Management Plan, 2013). Unpolluted water generally has a NO_3^--N concentration of less than 1.0 mg/l, which can be

achieved in the four alternative high and medium density development but cannot be achieved in the low-density current development scenario.

Overall, the most recommended urban development pattern for stream water quality protection was high density aggregated development; though specific attention should be paid in areas with potential *E.coli* pollution to avoid very high density development. It was worth noting that the predicted values of TP and NO_3^--N were comparable to the measured data at the TCEQ Station #16629 , which was located close to the outlet of the basin, indicating the reliability of our prediction models.

Table 2-9. Scenario prediction results of pollutant concentration

		Current development	High-density aggregated development	High-density sprawl development	Medium-density aggregated development	Medium-density sprawl development
TP	wet season	0.28 (0.55[1])	0.10	0.14	0.11	0.18
	dry season	0.28 (0.11)	0.09	0.14	0.09	0.13
E.coli	wet season	119.72	43.68	30.28	98.58	76.59
	dry season	67.23	26.31	41.50	54.90	32.02
NO_3^--N	wet season	1.98 (2.92)	0.1	0.19	0.15	0.25
	dry season	1.2 (1.58)	0.17	0.25	0.22	0.42

Notes of Table 2-9:

1. Values in the parentheses are measured pollutant concentrations at the TCEQ Station # 16629, which is close to the outlet of this basin
2. The unit is mg/l for TP and NO_3^--N and MPN/100ml for *E.coli*

2.4 Discussion

2.4.1 Planning implication based on urban development pattern metrics

Given that interpreting urban development pattern metrics can be difficult in land use planning, in this section I discuss how to link these metrics with specific land cover maps using sample watersheds in the study region. Three pair-wise comparisons of land cover maps with similar developed percentages but different TP concentrations were given in Figure 2-7. The two watersheds—(a) and (b) as represented in Figure 2-7 (1)—produced very different TP concentrations, which was likely associated with different IJI in developed areas. The watershed #12083 (Figure 2-7-a) was identified as more aggregated development, with a relatively integral natural core in the west. The IJI of developed area in this watershed was larger because the developed area was more equally adjacent to other land patch types. The watershed #11155 watershed (Figure 2-7-b) was low density development with scatted developed open areas. The IJI in this watershed was small because the developed area was largely adjacent to the developed open area only. The higher TP concentration in watershed # 11155 was likely caused by pollutants generated from the landscape gardens in the developed open areas and the greater extent of road surface area associated with detached houses (Goonetilleke et al., 2005).

The comparisons between watersheds in Figure 2-7-c and Figure 2-7-d and watersheds in Figure 2-7-e and Figure 2-7-f showed how shape and edge complexity potentially affected pollutant concentration. Different shape and edge complexity was associated with different drainage connections and road systems that influence runoff velocity, pollutant travel distance, and time of transport (Liu et al., 2012). The watershed #20730 (Figure 2-7-c) had a lower percentage of developed area but a higher TP concentration than the watershed #16655 (Figure 2-7-d). The complex shape and sprawled development of watershed #20730 led to more interspersed

land uses and more complex drainage and road systems. Higher ED of developed area in the watershed #17406 (Figure 2-7-e) was found to be associated with higher TP concentration than watershed #11405 (Figure 2-7-f) given the similar percentage of developed area. The high ED of developed area in watershed #17406 implied a sprawled road system that degraded the structure of natural areas (Lee et al., 2009).

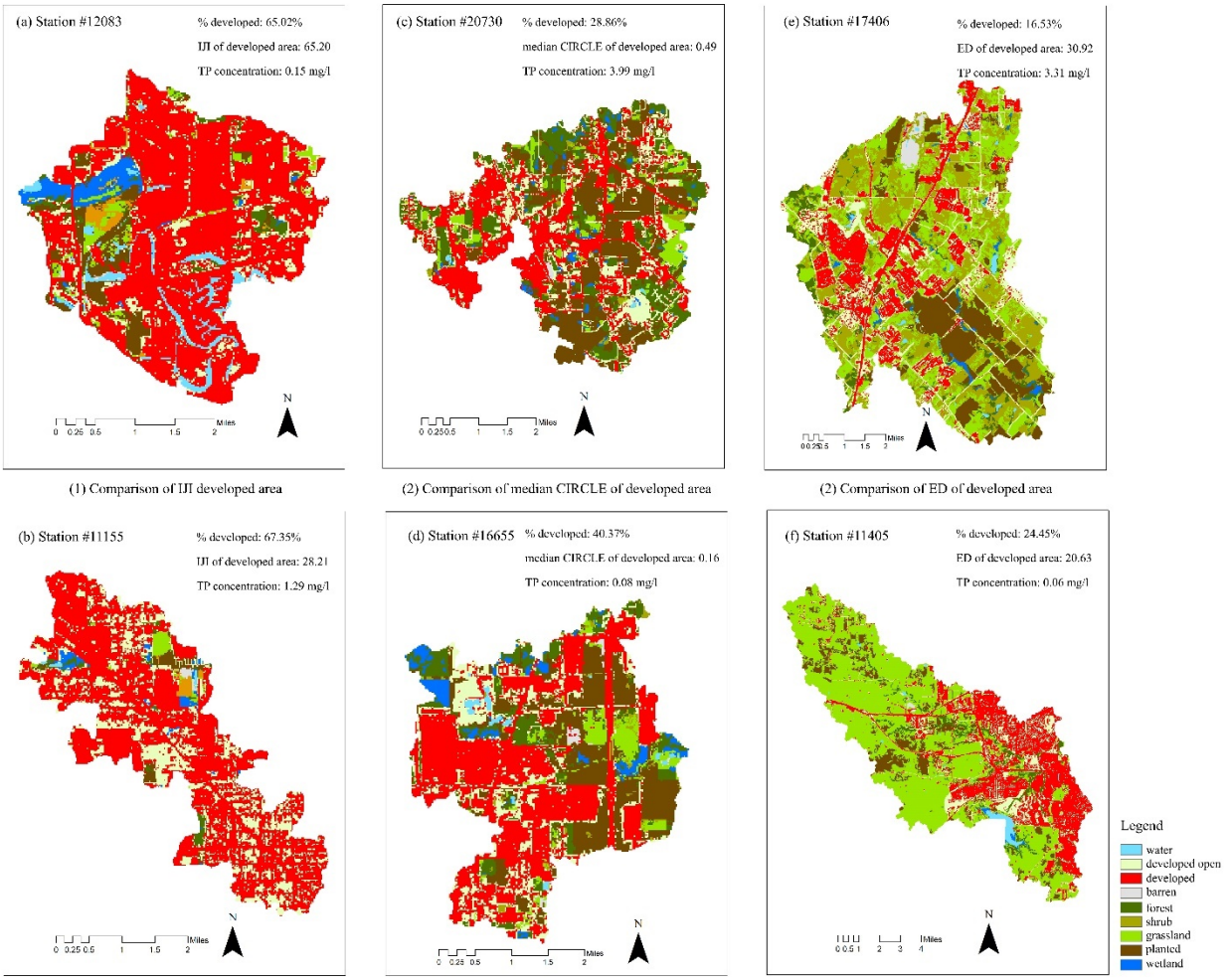


Figure 2-7. Examples of watersheds with the similar percentage of developed area but different urban development pattern metrics and TP concentration

Urban development pattern metrics are related to percentage, aggregation, patch shape, and connectivity of developed areas, and thus can represent characteristics of urban sprawl like low-

density development, leapfrog development over vacant lands, and decentralization (Riitters et al., 1995; Gordon and Richardson, 1996; Ewing, 2008; Bhatta, 2010). I argue that urban sprawl had a direct relationship to stream water quality, as it affected pollutant generation, build-up, and wash off by altering the structure of urban forms and the surrounding natural areas (Goonetilleke et al., 2005; Liu et al., 2012). To sum up, the ideal urban form for stream water quality protection should avoid (1) sprawl of low-density development with large lawn areas and complex road systems; (2) complexly shaped of urban areas that are likely to have complicated drainage and road systems; and (3) scatted patches of urban areas that destroy integral natural areas.

2.4.2 The complexity of the impact of urban development pattern on stream water quality

The results of this study indicate that both size and connectedness of urban developed areas (LPI, COHESION, and IJI of developed areas) were important in influencing stream water quality. This conclusion differs from Lee and others' argument that the dispersion and connectedness of land cover appear to be less informative in measuring the relationship between land use and water quality compared to size and number metrics (Lee et al., 2009).

Regarding the aggregation of urban areas, COHESION has showed a negative correlation with runoff and pollutant concentration in some studies (Li et al., 2015), while large and aggregated urban area, as indicated by high contiguity index (CONTIG) or contagion index (CONTAG), has been associated with poor stream water quality in others (Lee et al., 2009; Lv et al., 2015; Shi et al., 2017). Because greater interspersion and increases in the number of urban patches may accelerate soil erosion and sediment exportation (Shi et al., 2013), I argue that, although an intact urban area with large impervious surfaces can result in the deterioration of water quality (Alberti et al., 2007; Lee et al., 2009), the same area of impervious surface can lead to worse stream water quality with greater dispersion, as verified in our scenario prediction.

It is worth noting here that, without the control of developed area percentage, the effect of urban developed pattern on water quality should always be interpreted with caution due to the collinearity between urban development pattern and urban area percentage. As indicated in Figure 2-8, IJI, COHESION, and the percentage of developed areas were correlated with each other. Therefore, the effect of urban development pattern on stream water quality derived in statistical models can sometimes be caused by the percentage of urban developed area. I also wanted to note that, in Figure 2-8, the relationships between urban development patterns and percentage of urban developed area were not linear. Specifically, a low percentage of developed area does not necessarily mean low COHESION or high IJI. Thus, the percentage of urban developed area cannot replace urban development pattern metrics. This means that, in land use planning policy, IJI and COHESION should be considered together with the percent of urban developed area to evaluate the possible influence on stream water quality.

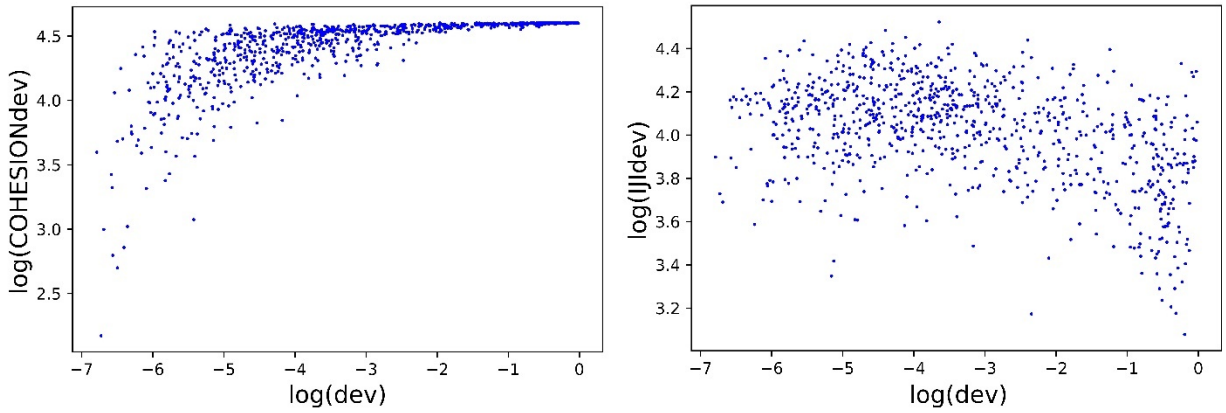


Figure 2-8. Scatter plots showing correlations between IJI, COHESION and the percentage of urban developed areas

The shape and edge complexity of developed areas were useful but not as important as aggregation/interspersion metrics in influencing stream water quality. Among some highly

colinear shape metrics, I found that the median of CIRCLE, FRAC, and SHAPE of a developed area were more efficient compared to other metrics. SHAPE was frequently applied to measure the effect of shape complexity on stream water quality and was found to be negatively associated with pollutant concentration at the catchment scale (Li et al., 2015; Yu et al., 2013; Lee et al., 2009; Shi et al., 2017). I found that FRAC and CIRCLE were also efficient metrics for measuring urban pattern shape. The importance of CIRCLE indicated that patch elongation was as important as patch compactness of urban area in evaluating stream water quality.

I furthermore found that class level landscape metrics were more effective than landscape level metrics in predicting stream water quality. The reason for this is that class level metrics had different influences on water quality depending on land cover type. For example, COHESION and IJI were important to developed areas in terms of their influence on stream water quality, but the COHESION and IJI of forest and planted areas were not as important. Researchers have argued that, at the landscape level, the landscape level SHDI and ED affect steam water quality at both watershed and reach scale (Shi et al., 2013; Sun et al., 2014,). However, I found that PAFRAC was the most significant factor affecting all pollutant concentrations instead of SHDI and ED.

2.4.3 Interpretation of the spatiotemporal non-stationary land-water relationships

In this study, the effects of IJI, LPI, and COHESION of developed areas on TP concentrations were more significant in the dry season than in the wet season. The absolute values of urban development pattern metrics' coefficients were larger in *E.coli and* NO_3^--N regressions in dry seasons, thereby indicating that urbanization had a larger effect on stream water quality in dry seasons. Precipitation had a significantly negative association with NO_3^--N concentration, indicating a potential dilution effect of prolonged precipitation in wet seasons (Chen et al., 2016). I also found that more urban development pattern metrics were selected by LASSO regression in

wet seasons, which represented more complex relationships than in dry seasons. Under future climate change conditions, urban development pattern might have a more complicated effects on stream water quality due to more precipitation in coastal areas. Future research should thus investigate the interaction effects between precipitation and the impacts of urban development pattern on stream water quality to further understand this mechanism.

Moreover, the influence that urban development pattern exerted on stream water quality had high spatial variations, which might be attributed to different pollutant sources. The LPI of developed area had a negative correlation with TP concentration in the highly urbanized areas like the Dallas metropolitan area in Texas. This finding differed from existing studies that have reported that the LPI of residential areas was a strong positive predictor of pollutant loading (Carey et al., 2011). Alternately, I argue that, in highly urbanized areas, larger LPI of developed area corresponded to aggregated development with fewer urban patch numbers, while smaller LPI of developed area was associated with smaller but more patches of impervious areas. Larger LPI of developed area in this case contributed to better water quality because of the smaller urban footprint of aggregated development. However, in the agricultural area, the relationship between LPI of developed area and TP concentration changed to significantly positive. In these watersheds, there were not many urban patches and large LPI of developed area simply implied larger urban core areas and total impervious areas, which contributed to the increasing pollutant concentration. This conclusion supports previous findings that indicate urbanization in agricultural watersheds can lead to larger increases in pollution compared to urban watersheds (Chen et al., 2016; Huang et al., 2015).

Furthermore, the IJI of developed area had a higher negative influence on TP concentration primarily in agricultural watersheds. In these agricultural watersheds, low IJI of developed area

was usually associated with low density development and high IJI was associated with medium to high density development. If developed areas were mostly adjacent to developed open areas in low density development, the watersheds typically had a low IJI of developed area and a high TP concentration. This phenomenon might be attributed to the application of phosphorus-based fertilizers on lawns in low-density residential areas (Wilson, 2015). TP concentration in highly urbanized areas, such as watersheds around Houston and Dallas, had a weak dependence on the IJI of developed area. Because of highly mixed land use in the high-density urban areas, the IJI of developed area might not be a reliable indicator of specific urban forms.

Complex shape (SHAPE) of developed area was associated with high *E.coli* concentration in all the watersheds, with stronger influences in San Jacinto Basin, the Neches Basin, and the Sabine Basin than in other basins. The similarity in these regions was higher total precipitation in wet seasons, which might be a reason for *E.coli* wash off from urban areas. It is also possible that aggregated development led to more *E.coli* pollution in wet seasons, which aligned with our scenario prediction results. Compared to TP, the effect of IJI of developed area on *E.coli* concentration had a lower spatial variation. The negative effect of urban sprawl on TP concentration was stronger than that on *E.coli* concentration, indicating that the mechanisms might differ and thus worth future investigation.

2.4.4 The advantages and limitations of applying machine learning in scenario prediction

The major advantage of the machine learning approach in this study was the successful quantification of complex, nonlinear land-water relationships. Overall, it facilitated more accurate water quality predictions under different planning scenarios. As the generalizability of machine learning is guaranteed by large sets of training samples and the train-test split method, it can be used to predict water quality under new land use plans in the Texas Gulf Region, especially in the

coastal area where the model performance was better than in the inland area. Policy makers can use this information to decide whether the resulting contaminant concentration meets regulation standards under the future land use scenario. This prediction framework can also be generalized to other watersheds and regions for the purposes of informing planning policy. As using a machine learning model alone is difficult in revealing the contribution of each catchment characteristic on stream water quality, statistical models were useful for uncovering the direction and spatial variation of each urban development pattern metrics' influence on stream water quality. I therefore suggest that combining statistics and machine learning was helpful for both predicting and interpreting water quality variations.

As mentioned in previous studies, a key gap in water quality studies has been a lack of consideration of cross effects between explanatory variables, such as the cross-correlation between land covers and the cross-correlation between land cover and climate in influencing water quality (Li et al., 2015; Hwang et al., 2016; Lintern et al., 2017). Machine learning can make use of all cross effects between variables and improve model predicting accuracy, which is an advantage over traditional statistical models.

Another advantage is that RF regression handles high dimensional data well since it works with subsets of data in each tree. It is therefore flexible and can accommodate more factors to improve water quality prediction accuracy, e.g., the inclusion of a monthly climatic variable in this study. Under climate change scenarios, climatic variables can therefore be included in machine learning models to forecast future stream water quality under extreme climate conditions. Overall, machine learning models can be used to predict water quality by taking into consideration any variables of interest in future research, the mechanism of which can be obscure and hard to model with a

physical-based model. In the predicting process specifically, it is applicable for integrating a set of planning factors to draw management implications of interest.

The major limitation of this study was that some catchment characteristics were excluded because they were not readily available. Such variables included point source pollution, animal products, wastewater treatment plants, and so on (Chen et al., 2014; Zhou et al., 2016). Future machine learning predictions of stream water quality should take these important aspects into consideration in order to obtain more unbiased models. Another limitation was the selection of appropriate variables. In this study, I conducted trials of variable selection in the RF regression using mutual info regression, which entailed dropping a specific number of variables with the lowest mutual information regarding pollutant concentration (Kraskov et al., 2011). I at last decided to keep the whole set of independent features in the prediction model because, after iterating all possible numbers of input variables, the RF regression accuracy did not significantly improve. Future studies should also try other engineering algorithms, such as recursive feature elimination.

2.5. Conclusion

Urban development patterns were found to significantly influence stream TP, NO_3^--N, and *E.coli* concentrations in the Texas Gulf Region, with the relationships among them varying according to season and location. LPI, COHESION, and IJI of developed areas were the most efficient urban development pattern metrics associated with stream water quality. Furthermore, shape complexity and edge density of urban developed areas were positively correlated with pollutant concentrations. The effect of urban development pattern on stream water quality was more stable and significant in dry seasons and more variable and complex in wet seasons. The IJI of developed area had a higher negative influence on water quality in less urbanized watersheds. The LPI of developed

area had a negative correlation with TP concentration in the highly urbanized area, but a positive correlation in the agricultural area.

It was predicted by RF regression that high density aggregated development was the most effective in reducing TP and NO_3^--N concentrations compared to medium density development and the current sprawl development. However, aggregated development contributed to *E.coli* pollution in wet seasons. To conclude, this study demonstrated the environmental consequences of urban sprawl and supported policy orientation towards compact city planning according to the machine learning predictive framework.

APPENDIX

Table 2A-2-10. Description of landscape metrics.

Category	Variable	Range	Description
Area	Percentage of Landscape (PLAND)	(0,100]	the percentage the landscape comprised of the corresponding patch type
	Total Area (CA)	(0, ∞)	the sum of the areas of all patches of the corresponding patch type
	Median of Patch Area (AREA_MD)	(0, ∞)	the median of all patches of the corresponding patch type
	Median of Radius of Gyration (GYRATE_MD)	(0, ∞)	the median of mean distance between each cell in the patch and the patch centroid.
	Largest Patch Index (LPI)	(0,100]	the area of the largest patch of the corresponding patch type divided by total landscape area
Edge	Total Edge (TE)	[0, ∞)	the sum of the lengths of all edge segments involving the corresponding patch type
	Edge Density (ED)	[0, ∞)	the sum of the lengths of all edge segments involving the corresponding patch type, divided by the total landscape area
Shape	Median of Perimeter-Area Ratio (PARA_MD)	(0, ∞)	the median of the ratio of the patch perimeter to area
	Median of Shape Index (SHAPE_MD)	(0, ∞)	the median of patch perimeter divided by the minimum perimeter possible for a maximally compact patch of the corresponding patch area.
	Median of Fractal Dimension Index (FRAC_MD)	[1, 2]	the median of 2 times the logarithm of patch perimeter divided by the logarithm of patch area
	Median of Related Circumscribing Circle (CIRCLE)	[0, 1)	the median of 1 minus patch area divided by the area of the smallest circumscribing circle
	Median of Contiguity Index (CONTIG_MD)	[0, 1]	the median of the average contiguity value for the cells in a patch minus 1 divided by the sum of the template values minus 1
Subdivision	Number of Patches (NP)	[1, ∞)	the number of patches of the corresponding patch type
	Patch Density (PD)	(0, ∞)	the number of patches of the corresponding patch type divided by total landscape area
	Landscape Division Index (DIVISION)	[0, 1)	1 minus the sum of patch area divided by total landscape area, quantity squared, summed across all patches of the corresponding patch type
	Splitting Index (SPLIT)	[1, $Ncell^2$]	the total landscape area squared divided by the sum of patch area squared, summed across all patches of the corresponding patch type
Aggregation	Interspersion Juxtaposition Index (IJI)	(0,100]	minus the sum of the length of each unique edge type involving the corresponding patch type divided by the total length of edge involving the same type, multiplied by the logarithm of the same quantity, summed over each unique edge type; divided by the logarithm of the number of patch types minus 1

Table 2A-2-10 (cont'd)

Landscape Shape Index (LSI)	[1, ∞)	the total length of edge involving the corresponding class, divided by the minimum length of class edge possible for a maximally aggregated class
Patch Cohesion Index (COHESION)	[0,100)	1 minus the sum of patch perimeter divided by the sum of patch perimeter times the square root of patch area for patches of the corresponding patch type, divided by 1 minus 1 over the square root of the total number of cells in the landscape

BIBLIOGRAPHY

BIBLIOGRAPHY

Ai, L., Shi, Z. H., Yin, W., & Huang, X. (2015). Spatial and seasonal patterns in stream water contamination across mountainous watersheds: Linkage with landscape characteristics. *Journal of Hydrology*, 523, 398–408. doi:10.1016/j.jhydrol.2015.01.082

Alberti, M., Booth, D., Hill, K., Coburn, B., Avolio, C., Coe, S., & Spirandelli, D. (2007). The impact of urban patterns on aquatic ecosystems: An empirical analysis in Puget lowland sub-basins. *Landscape and urban planning*, *80*(4), 345-361.

Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E. (2018). Evaluating statistical model performance in water quality prediction. *Journal of Environmental Management*, 206, 910–919. doi:10.1016/j.jenvman.2017.11.049

Bhatta, B. (2010). Urban Growth and Sprawl. Analysis of Urban Growth and Sprawl from Remote Sensing Data, 1–16. doi:10.1007/978-3-642-05299-6_1

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

Bu, H., Meng, W., Zhang, Y., & Wan, J. (2014). Relationships between land use patterns and water quality in the Taizi River basin, China. *Ecological Indicators*, 41, 187–197. doi:10.1016/j.ecolind.2014.02.003

Carey, R. O., Migliaccio, K. W., Li, Y., Schaffer, B., Kiker, G. A., & Brown, M. T. (2011). Land use disturbance indicators and water quality variability in the Biscayne Bay Watershed, Florida. *Ecological Indicators*, *11*(5), 1093-1104.

Chen, J., & Lu, J. (2014). Effects of Land Use, Topography and Socio-Economic Factors on River Water Quality in a Mountainous Watershed with Intensive Agricultural Production in East China. *PLoS ONE*, 9(8), e102714. doi:10.1371/journal.pone.0102714

Chen, Q., Mei, K., Dahlgren, R. A., Wang, T., Gong, J., & Zhang, M. (2016). Impacts of land use and population density on seasonal surface water quality using a modified geographically weighted regression. *Science of the total environment*, *572*, 450-466.

Chermack, T. J., & Swanson, R. A. (2008). Scenario planning: human resource development's strategic learning tool. *Advances in Developing Human Resources*, 10(2), 129-146.

Clapcott, J. E., Collier, K. J., Death, R. G., Goodwin, E. O., Harding, J. S., Kelly, D., ... & Young, R. G. (2012). Quantifying relationships between land-use gradients and structural and functional indicators of stream ecological integrity. *Freshwater Biology*, *57*(1), 74-90. doi:10.1111/j.1365-2427.2011.02696.x

Clément, F., Ruiz, J., Rodríguez, M. A., Blais, D., & Campeau, S. (2017). Landscape diversity and forest edge density regulate stream water quality in agricultural catchments. *Ecological Indicators*, 72, 627–639. doi:10.1016/j.ecolind.2016.09.001

Ding, J., Jiang, Y., Liu, Q., Hou, Z., Liao, J., Fu, L., & Peng, Q. (2016). Influences of the land use pattern on water quality in low-order streams of the Dongjiang River basin, China: A multi-scale analysis. *Science of The Total Environment*, 551-552, 205–216. doi:10.1016/j.scitotenv.2016.01.162

Del Monaco, N. (2017). Reducing directly connected stormwater infrastructure and the associated benefits (Doctoral dissertation, Rutgers University-Graduate School-New Brunswick).

Ewing, R. H. (n.d.). Characteristics, Causes, and Effects of Sprawl: A Literature Review. *Urban Ecology*, 519–535. doi:10.1007/978-0-387-73412-5_34

Fan, M., & Shibata, H. (2015). Simulation of watershed hydrology and stream water quality under land use and climate change scenarios in Teshio River watershed, northern Japan. *Ecological Indicators*, 50, 79–89. doi:10.1016/j.ecolind.2014.11.003

Forman, R. T. (2014). Land Mosaics: The ecology of landscapes and regions (1995) (p. 217). Island Press.

Goonetilleke, A., Thomas, E., Ginn, S., & Gilbert, D. (2005). Understanding the role of land use in urban stormwater quality management. *Journal of Environmental Management*, 74(1), 31–42. doi:10.1016/j.jenvman.2004.08.006

Grand River Water Management Plan. (2013). Update Water Quality Targets to Support Healthy and Resilient Aquatic Ecosystems in the Grand River Watershed

Giri, S., & Qiu, Z. (2016). Understanding the relationship of land uses and water quality in Twenty First Century: A review. *Journal of Environmental Management*, 173, 41–48. doi:10.1016/j.jenvman.2016.02.029

Glińska-Lewczuk, K., Gołaś, I., Koc, J., Gotkowska-Płachta, A., Harnisz, M., & Rochwerger, A. (2016). The impact of urban areas on the water quality gradient along a lowland river. *Environmental Monitoring and Assessment*, 188(11). doi:10.1007/s10661-016-5638-z

Gordon, P., & Richardson, H. W. (1996). Beyond Polycentricity: The Dispersed Metropolis, Los Angeles, 1970-1990. *Journal of the American Planning Association*, 62(3), 289–295. doi:10.1080/01944369608975695

Hameed, M., Sharqi, S. S., Yaseen, Z. M., Afan, H. A., Hussain, A., & Elshafie, A. (2016). Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. *Neural Computing and Applications*, 28(S1), 893–905. doi:10.1007/s00521-016-2404-7

Harrell, F. (2017). Regression modeling strategies. BIOS, 330, 2018.

Holcomb, D. A., Messier, K. P., Serre, M. L., Rowny, J. G., & Stewart, J. R. (2018). Geostatistical Prediction of Microbial Water Quality Throughout a Stream Network Using Meteorology, Land Cover, and Spatiotemporal Autocorrelation. *Environmental Science & Technology*, 52(14), 7775–7784. doi:10.1021/acs.est.8b01178

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., ... & Megown, K. (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.

Huang, Z., Han, L., Zeng, L., Xiao, W., & Tian, Y. (2015). Effects of land use patterns on stream water quality: a case study of a small-scale watershed in the Three Gorges Reservoir

Area, China. *Environmental Science and Pollution Research*, 23(4), 3943–3955. doi:10.1007/s11356-015-5874-8

Hwang, S.-A., Hwang, S.-J., Park, S.-R., & Lee, S.-W. (2016). Examining the Relationships between Watershed Urban Land Use and Stream Water Quality Using Linear and Generalized Additive Models. *Water*, 8(4), 155. doi:10.3390/w8040155

Jones, J. E., Earles, T. A., Fassman, E. A., Herricks, E. E., Urbonas, B., & Clary, J. K. (2005). Urban Storm-Water Regulations—Are Impervious Area Limits a Good Idea? *Journal of Environmental Engineering*, 131(2), 176–179. doi:10.1061/(asce)0733-9372(2005)131:2(176)

Kalteh, A. M., Hjorth, P., & Berndtsson, R. (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. *Environmental Modelling & Software*, 23(7), 835–845. doi:10.1016/j.envsoft.2007.10.001

Kraskov, A., Stögbauer, H., & Grassberger, P. (2011). Erratum: Estimating mutual information [Phys. Rev. E 69, 066138 (2004)]. Physical Review E, 83(1), 019903.

Lee, S.-W., Hwang, S.-J., Lee, S.-B., Hwang, H.-S., & Sung, H.-C. (2009). Landscape ecological approach to the relationships of land use patterns in watersheds to water quality characteristics. *Landscape and Urban Planning*, 92(2), 80–89. doi:10.1016/j.landurbplan.2009.02.008

Lek, S. (1999). Predicting stream nitrogen concentration from watershed features using neural networks. *Water Research*, 33(16), 3469–3478. doi:10.1016/s0043-1354(99)00061-5

Li, Y., Li, Y., Qureshi, S., Kappas, M., & Hubacek, K. (2015). On the relationship between landscape ecological patterns and water quality across gradient zones of rapid urbanization in coastal China. *Ecological Modelling*, 318, 100–108. doi:10.1016/j.ecolmodel.2015.01.028

Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., … Western, A. W. (2017). Key factors influencing differences in stream water quality across space. Wiley Interdisciplinary Reviews: *Water*, 5(1), e1260. doi:10.1002/wat2.1260

Liu, A., Goonetilleke, A., & Egodawatta, P. (2012). Inadequacy of Land Use and Impervious Area Fraction for Determining Urban Stormwater Quality. *Water Resources Management*, 26(8), 2259–2265. doi:10.1007/s11269-012-0014-4

Lv, H., Xu, Y., Han, L., & Zhou, F. (2014). Scale-dependence effects of landscape on seasonal water quality in Xitiaoxi catchment of Taihu Basin, China. *Water Science and Technology*, 71(1), 59–66. doi:10.2166/wst.2014.463

McGarigal, K. (1995). FRAGSTATS: spatial pattern analysis program for quantifying landscape structure (Vol. 351). US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

McHarg, I.L., Sutton, J., 1975. Ecological plumbing for the Texas coastal plain: The Woodlands New Town Experiment. *Landscape Archit*. 65 (1), 80–90.

Mirzaei, M., Jafari, A., Gholamalifard, M., Azadi, H., Shooshtari, S. J., Moghaddam, S. M., … Witlox, F. (2020). Mitigating environmental risks: Modeling the interaction of water quality parameters and land use cover. *Land Use Policy*, 95, 103766. doi:10.1016/j.landusepol.2018.12.014

Molina-Navarro, E., Segurado, P., Branco, P., Almeida, C., & Andersen, H. E. (2020). Predicting the ecological status of rivers and streams under different climatic and socioeconomic scenarios using Bayesian Belief Networks. *Limnologica*, 80, 125742. doi:10.1016/j.limno.2019.125742

Obropta, C. C., & Del Monaco, N. (2018). Reducing Directly Connected Impervious Areas with Green Stormwater Infrastructure. *Journal of Sustainable Water in the Built Environment*, 4(1), 05017004. doi:10.1061/jswbay.0000833

Oeding, S., Taffs, K. H., Cox, B., Reichelt-Brushett, A., & Sullivan, C. (2018). The influence of land use in a highly modified catchment: Investigating the importance of scale in riverine health assessment. *Journal of Environmental Management*, 206, 1007–1019. doi:10.1016/j.jenvman.2017.12.005

Pratt, B., & Chang, H. (2012). Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. *Journal of Hazardous Materials*, 209-210, 48–58. doi:10.1016/j.jhazmat.2011.12.068

Riitters, K. H., O'Neill, R. V., Hunsaker, C. T., Wickham, J. D., Yankee, D. H., Timmins, S. P., … Jackson, B. L. (1995). A factor analysis of landscape pattern and structure metrics. *Landscape Ecology*, 10(1), 23–39. doi:10.1007/bf00158551

Schreiber, J., Jessulat, M., & Sick, B. (2019). Generative Adversarial Networks for Operational Scenario Planning of Renewable Energy Farms: A Study on Wind and Photovoltaic. *Artificial Neural Networks and Machine Learning* – ICANN 2019: Image Processing, 550–564. doi:10.1007/978-3-030-30508-6_44

Sharifi, A., Yen, H., Boomer, K. M. B., Kalin, L., Li, X., & Weller, D. E. (2017). Using multiple watershed models to assess the water quality impacts of alternate land development scenarios for a small community. *CATENA*, 150, 87–99. doi:10.1016/j.catena.2016.11.009

Shi, Z. H., Ai, L., Li, X., Huang, X. D., Wu, G. L., & Liao, W. (2013). Partial least-squares regression for linking land-cover patterns to soil erosion and sediment yield in watersheds. *Journal of Hydrology*, 498, 165–176. doi:10.1016/j.jhydrol.2013.06.031

Shi, P., Zhang, Y., Li, Z., Li, P., & Xu, G. (2017). Influence of land use and land cover patterns on seasonal water quality at multi-spatial scales. *CATENA*, 151, 182–190. doi:10.1016/j.catena.2016.12.017

Sohn, W., Kim, J.-H., & Li, M.-H. (2017). Low-impact development for impervious surface connectivity mitigation: assessment of directly connected impervious areas (DCIAs). *Journal of Environmental Planning and Management*, 60(10), 1871–1889. doi:10.1080/09640568.2016.1264929

Sun, R., Chen, L., Chen, W., & Ji, Y. (2011). Effect of Land-Use Patterns on Total Nitrogen Concentration in the Upstream Regions of the Haihe River Basin, China. *Environmental Management*, 51(1), 45–58. doi:10.1007/s00267-011-9764-7

Sun, Y., Guo, Q., Liu, J., & Wang, R. (2014). Scale Effects on Spatially Varying Relationships Between Urban Landscape Patterns and Water Quality. *Environmental Management*, 54(2), 272–287. doi:10.1007/s00267-014-0287-x

Teklu, B. M., Hailu, A., Wiegant, D. A., Scholten, B. S., & Van den Brink, P. J. (2016). Impacts of nutrients and pesticides from small- and large-scale agriculture on the water quality of Lake Ziway, Ethiopia. *Environmental Science and Pollution Research*, 25(14), 13207–13216. doi:10.1007/s11356-016-6714-1

Texas Commission on Environmental Quality. (2014). Managing nonpoint source pollution in Texas, *2013 annual report*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

Wang, R., Zhang, X., & Li, M.-H. (2019). Predicting bioretention pollutant removal efficiency with design features: A data-driven approach. *Journal of Environmental Management*, 242, 403–414. doi:10.1016/j.jenvman.2019.04.064

Wijesiri, B., Deilami, K., & Goonetilleke, A. (2018). Evaluating the relationship between temporal changes in land use and resulting water quality. *Environmental Pollution*, 234, 480–486. doi:10.1016/j.envpol.2017.11.096

Wilson, C. O. (2015). Land use/land cover water quality nexus: quantifying anthropogenic influences on surface water quality. *Environmental Monitoring and Assessment*, 187(7). doi:10.1007/s10661-015-4666-4

World Population Review (2019). *Texas Population 2019*. Retrieved from http://worldpopulationreview.com/states/texas-population/

Yang, B., & Li, M.-H. (2011). Assessing planning approaches by watershed streamflow modeling: Case study of The Woodlands; Texas. *Landscape and Urban Planning*, 99(1), 9–22. doi:10.1016/j.landurbplan.2010.08.007

Yu, D., Shi, P., Liu, Y., & Xun, B. (2013). Detecting land use-water quality relationships from the viewpoint of ecological restoration in an urban area. *Ecological Engineering*, 53, 205–216. doi:10.1016/j.ecoleng.2012.12.045

Zhou, P., Huang, J., Pontius, R. G., & Hong, H. (2016). New insight into the correlations between land use and water quality in a coastal watershed of China: Does point source pollution weaken it? *Science of The Total Environment*, 543, 591–600. doi:10.1016/j.scitotenv.2015.11.063

CHAPTER 3 DERIVING ANNUAL LAND COVER MAPS AND MODELING THE LONGITUDINAL EFFECT OF LAND COVER CHANGE ON NUTRIENT AND BACTERIA CONCENTRATIONS

3.1 Introduction

Land cover change is an important driver of many environmental issues such as climate change, hydrological cycle alteration, nonpoint source pollution, biodiversity declines, and so on (Kalnay and Cai, 2003; Sajikumar and Remya, 2015; Newbold et al., 2016; Zhao et al., 2016; Oeding et al., 2018). Assessing the relationship between land cover and stream water quality is recognized as an imperative step to help manage nonpoint source pollution and to inform land use policies in the watershed (Ai et al., 2015; Giri and Qiu, 2016; Wijesiri et al., 2018). The significant impact of land cover change, together with climatic, geo-morphological, and socioeconomic factors on stream water quality has been highlighted in recent research (Ding et al., 2016; Manfri et al., 2016; Zhou et al., 2016; Lintern et al., 2018).

Two broad issues are associated with this area of research. Firstly, because many water quality studies are conducted at a local and cross-sectional scale with limited samples (Huang et al., 2016; Luo et al., 2017; Rodrigues et al., 2018), confidence about land cover effect on stream quality is not high; therefore there is little capacity to make generalizations regionally. Secondly, the spatial and temporal variations in the land-water relationship are difficult to quantify with a simple and robust model structure (Sun et al., 2013; Bu et al., 2014; Kibena and Gumindoga, 2014; Walsh and Webb, 2014). To concentrate on the two issues, I propose a linear mixed model structure that employs 20-year water quality data from 1991 to 2011 in the Texas Gulf Region. The abundance of data is able to capture the variation in the natural and anthropogenic drivers of water quality degradation in this region.

With respect to the first issue, one difficulty in conducting a long-term study to quantify the land cover effect on water quality is the mismatch between the temporal resolution of land cover data and the stream water quality data. Land cover data such as NLCD, measured at 2-3 year intervals, is much coarser than the measured water quality data (Seeboonruang, 2012; du Plessis et al., 2015; Homer et al., 2015; Vrebos et al., 2017). In addition, the quality of land cover maps before 2001 is not as high as more recent land cover maps (Vogelmann et al., 2001). There is a great need to generate consistent land cover maps over a long period of time that match the temporal range and resolution of stream water quality data. In addition, land cover maps should have relatively balanced accuracy among different classes, because less common land cover types can still affect environmental processes and functions significantly (Zhu et al., 2016; Heydari and Mountrakis, 2018).

Mapping large-area heterogeneous landscapes and detecting changes is always challenging (Schneider et al., 2010; Rodriguez-Galiano et al., 2012; Thakkar et al., 2017). The selection of classifier, the inclusion of auxiliary training features, and the training sample size and distribution are all critical to improve classification performance (Millard and Richardson, 2015; Zhang and Roy, 2017; Liu et al., 2018). Recent developments in long-term land cover classification and change detection methods have incorporated spectral, spatial, and temporal data, as well as the knowledge of logic processes to provide reliable outcomes (Manandhar et al., 2009; Gómez et al., 2016; Jin et al., 2017; Liu et al.,2019). One of the efficient approaches is to use the multi-threshold method to identify change groups, such as biomass increase and decrease groups with multiple spectral indices (Jin et al., 2013, Jin et al., 2017). In this study, a comprehensive investigation of training samples and classifiers, and a multi-threshold post-classification quality control process, are the two primary attempts to obtain the annual land cover data.

To address the second issue, it is important to construct a simple but credible model to quantify the spatial and temporal variations in the long-term land cover and water quality relationships to advise both local and regional planning (Wang et al., 2014; Chen et al., 2016; Shi et al., 2017). Significant associations between land cover and water quality have been found using Ordinary Least Squares (OLS) regression with the assumption that the relationship is constant across space (Rothwell et al., 2010; Carey et al., 2011; Chu et al., 2013; Jordan et al., 2018). Ordinary Least Squares (OLS) regression leads to general inferences of the land cover effect on water quality but neglects spatial autocorrelation among water quality samples (Tu, 2011). Geographically Weighted Regression (GWR) demonstrates great improvements in model performance over OLS because it assumes that the samples closer to the location of an observation have a higher impact on the local parameter estimation (Tu, 2013; Chen and Lu, 2014; Chen et al., 2016). However, if the spatial extent and the number of water quality samples become too great, there is a risk that GWR model parameters and the underlying spatial relationships become too complicated.

Linear mixed models can handle both spatial and temporal correlation structures among samples with flexible model structures (Molenberghs and Verbeke, 2000; Kuznetsova et al, 2017). These models have provided insights to predict many environmental parameters such as carbon cycles, soil productivity and forest density (Doetterl et al., 2013; Sakai et al., 2013; Zou et al., 2017). The advantage of using linear mixed models in water quality prediction is that samples can be grouped as random components to explore the unobserved characteristics in each group which are not expressed in the fixed effects. For example, water quality samples can be grouped according to the year they were taken, the site they were taken at, and the antecedent discharge, depending on which are the factors of interest (Sheldon et al., 2012; Lessel and Bishop, 2013; Bonansea et al., 2015). Compared to GWR, linear mixed models are more flexible in grouping

water quality samples at the scale of policy interest and can account for temporal variation at the same time.

The novelty of this study is two-fold. Firstly, it provides an efficient classification and change detection algorithm for generating annual land cover maps. The algorithm can be applied to obtain historical land cover data where only one-year land cover map is available. Secondly, it is one of the few water quality studies with a large regional scale and a long time range. The derived regional-scale knowledge matches the spatial scale of urban and regional planning. This study involves four research objectives: 1) To develop a robust annual land cover classification workflow implemented on the GEE platform. 2) To explore the land cover change and stream water quality change trajectory from 1990 to 2011. 3) To find the most appropriate linear model correlation structure to model the longitudinal relationships between land cover and nutrient and bacteria concentrations. 4) To provide land use and watershed management policy implications at both regional and basin levels in Texas.

3.2 Data and Method

3.2.1 Study Site

The Texas Gulf Region is one of the 21 water resource regions within the first-level hydrological units in the United States. It consists of 11 subregions and 23 basins with a total drainage area of 471,080 km^2. It covers most areas of Texas and discharges into the Gulf of Mexico. The climate of this region is quite diverse, with a maritime climate along the coast, a continental climate in the central and northern areas, and a dry and hot climate in the west. These diverse climates lead to a heterogeneous landscape across the region. From east to west, the terrain ecosystem changes from coastal swamps and piney woods to rolling plains and rugged hills. According to landscape characteristics, Texas can be divided into 10 ecoregions or natural regions, with 9 of them located

in the Texas Gulf Region, including the Piney Woods, the Gulf Prairies and Marshes, the Post Oak Savannah, the Blackland Prairies, the Cross Timbers, the South Texas Plains, the Edwards Plateau, the Rolling Plains, and the High Plains (Figure 3-1).

Texas is the second largest state in the United States with a current population of 29 million. It has an annual population growth rate of 1.8%, ranking the third in the country (World Population Review, 2019). The increasing population results in the problem of urban sprawl, which has put natural forest areas at risk and caused stream water quality degradation. In Texas, 410 out of 1214 water bodies do not meet applicable water quality standards or are threatened for one or more designated uses, among which bacteria, dissolved oxygen, nutrients, and organics are the major concerns. Nonpoint source pollution closely related to land use contributes to approximately 45% of stream water quality impairment (Texas Commission on Environmental Quality, 2014). To monitor and assess stream water quality conditions, the Texas Commission on Environmental Quality (TCEQ) Surface Water Quality Monitoring (SWQM) Program has over 3000 active monitoring stations throughout the state.

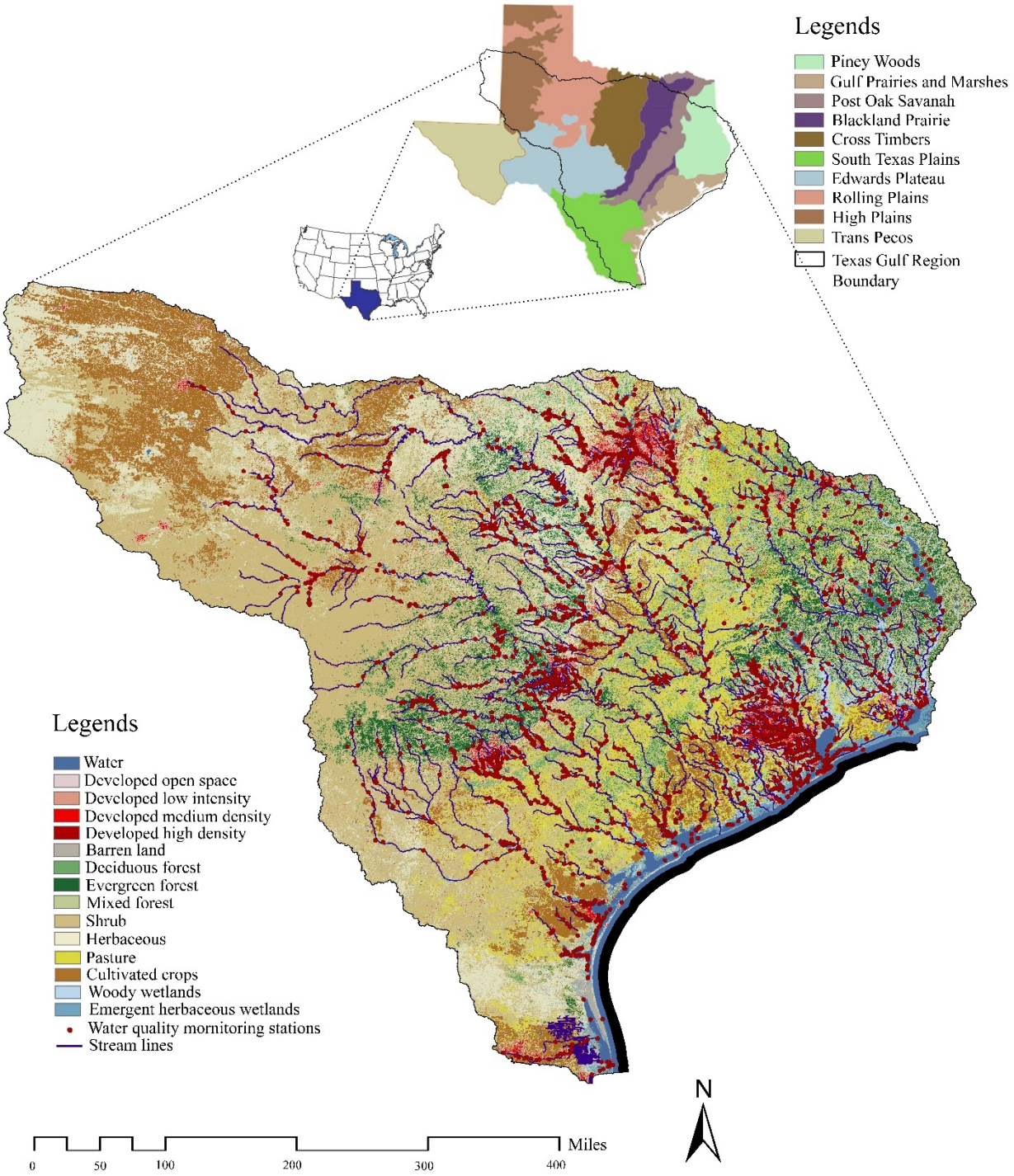


Figure 3-1. Texas Gulf Region with a base map of NLCD 2011 and the Texas ecoregions

3.2.2 Data

- **Data for image classification**

NLCD with an 89% overall accuracy at Level I is the most fundamental data to investigate the impact of land cover change on ecosystems in the United States (Tran et al., 2010; Homer et al., 2015; Wickham et al., 2017). It provides land cover data at a 30 m resolution from 2001 to 2016 at 2-3 year intervals. NLCD 1992 is also available but is not recommended for any direct comparisons with the subsequent NLCD products due to the change of legends and mapping methods (Vogelmann et al., 2001). In this study, NLCD 2011 was used to extract ground truthed land cover types for image classification, and NLCD 2006 and 2001 were used as validation maps to evaluate classification performance.

The USGS Landsat 5 Surface Reflectance Tier 1 product was used as the base map to extract the spectral training features. This dataset is the atmospherically corrected and orthorectified surface reflectance data. The USGS National Elevation Dataset with the spatial resolution of 30m resampled from the original 1/3 arcsecond was used to derive elevation, slope and other terrain features. All the classification training data was extracted from the GEE platform.

- **Data for stream water quality prediction**

The stream water quality data was acquired from the Texas Clean Rivers Program (CRP). There are 1783 water quality monitoring stations in the Texas Gulf Region in operation between 1991 and 2011, from which all the available NO_3^--N, PO_4^{3-}-P, NH_4^+-N, TP and *E.coli* concentration data were obtained. Then the 1783 contributing areas were delineated according to the 30m Digital Elevation Model (DEM) with the water quality monitoring stations as the subbasin outlets. The delineated watershed boundary was used to obtain all the independent variables of each subbasin. The annual land cover areal percentages were calculated from the classified land cover maps. The elevation and slope data was derived from the 30m USGS National Elevation Dataset. The climatic data, including monthly total precipitation and average temperature, was acquired from PRISM

Monthly Spatial Climate Dataset AN81m. All the independent variables were obtained from the GEE platform.

3.2.3 Methods

Two major steps were implemented in this study as shown in the flowchart (Figure 3-2): First, annual land cover maps from 1991 to 2011 were generated for the whole Texas Gulf Region. Local random forest classifiers were applied in each ecoregion with a combination of spectral, ancillary, seasonal, and textural training features. Then the 20-year independent classification maps were passed through the post-classification quality control algorithm to produce the final images. Second, land cover percentages in each year were calculated from the 20-year land cover maps to build longitudinal regression models together with nutrient and bacteria concentrations as dependent variables using linear mixed models.

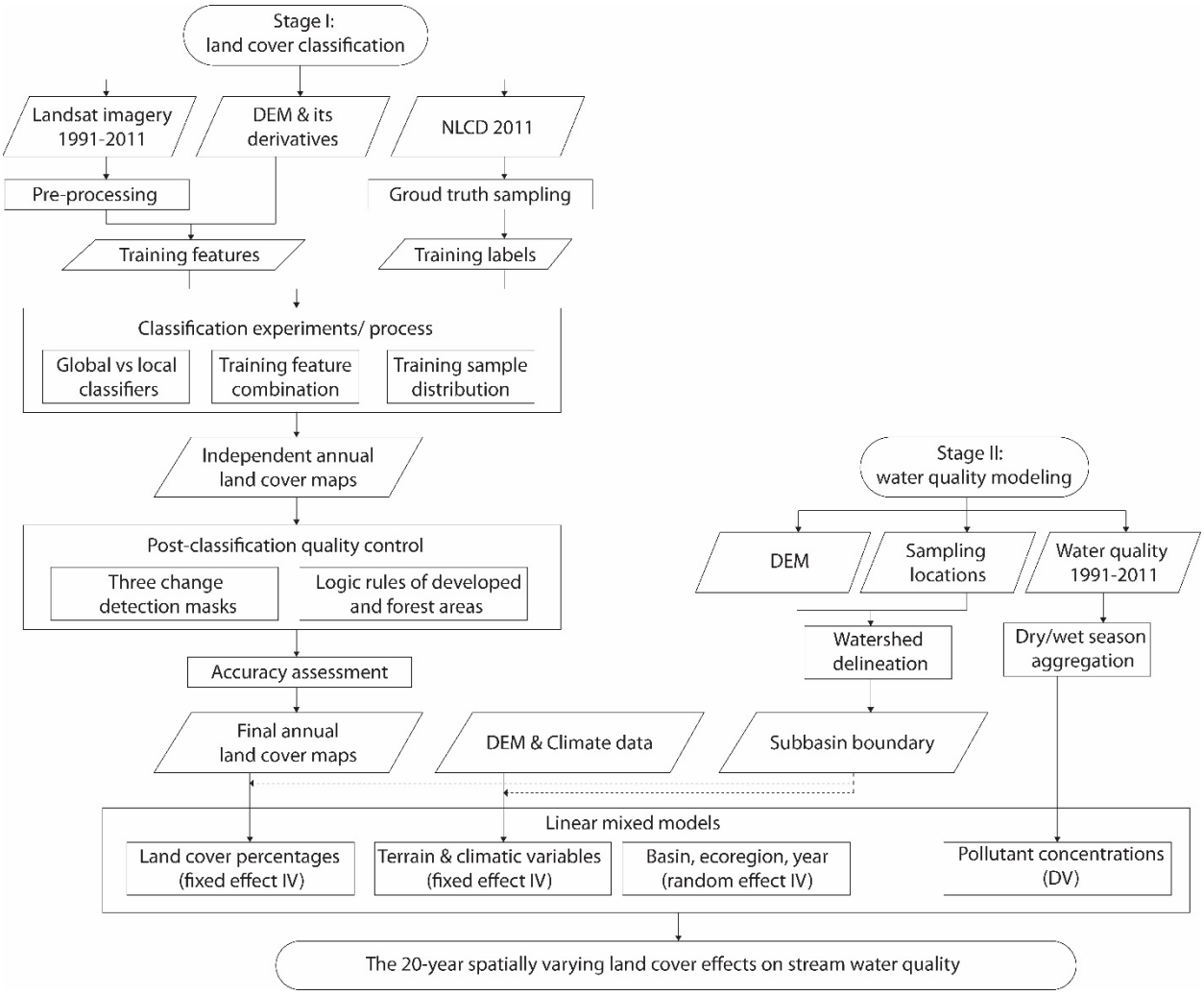


Figure 3-2. Method flowchart

- **Annual land cover classification**

Local random classifiers applied to every ecoregion were tested to outperform a single random classifier because the dominant land cover types were different among ecoregions. The Post Oak Savannah, the Blackland Prairies and the Cross Timbers ecoregions share some landscape similarities and they were merged to become one region (Post Oak and Prairie) in this study. The High Plains were excluded from the classification process because no water quality monitoring stations are in this ecoregion. Therefore, six local random forest classifiers were fitted independently in the Piney Woods, the Gulf Prairies and Marshes, the Post Oak and Prairies, the

South Texas Plains, the Edwards Plateau, and the Rolling Plains ecoregions. In each local random forest classifier, the number of trees was set to 10, the number of variables per split was set to the square root of the number of variables, and the minimum size of a terminal node was set to 1.

A pair of cloud-free Landsat images in both leaf-on and leaf-off seasons were generated every year from 1991 to 2011 to extract the training samples. Specifically, the median values of the clear and water pixels with low or median cloud confidence in the pixel quality band of Landsat 5 were selected to generate the cloud-free images. To ensure the reliability of the training samples, two control principles were implemented. 1) Only pixels with consistent land cover labels in NLCD 2001, 2006 and 2011 were included in the training sample pool. 2) A spatial filter was applied to all the pixels to filter pixels with land cover labels the same as the surrounding eight pixels. In each ecoregion, 160,000 training samples were selected as input to the local random forest classifier.

Three groups of training features were used in the classification process, which were basic spectral features, ancillary features, and texture features. The basic spectral features included band 1 to band 7 of Landsat 5 imagery. The topography-based ancillary features included elevation, slope, Terrain Ruggedness Index (TRI), Topographic Wetness Index (TWI), slope Length and Steepness factor (LS factor) (Moore et al., 1993; Riley et al., 1999; Panagos et al., 2015). The spectral-based ancillary features included the ratio of near infrared band to the red band, NDVI, Tasseled Cap wetness, and greenness and brightness index (Crist and Cicone, 1984). Texture features calculated from the Grey Level Co-occurrence Matrix (GLCM) were also included in the classification process to aid the detection of developed area and planted area (Rodriguez-Galiano et al., 2012). In this study, the kernel of size 7*7 pixels was used to derive texture features from both the Landsat 5 NIR band and the NDVI image based on the GLCM. The six most important

texture features discovered by the Principal Component Analysis (PCA) were selected, including difference entropy, cluster prominence, correlation, cluster shade, information measure of correlation, and sum average. In addition, all the spectral-based features were derived from both the leaf-on and the leaf-off images to add seasonal information. In total, 53 training features were used in the classifier training process.

The agreement between the classified map and NLCD was referred to as "accuracy" in this study. The original classification scheme was the eight Anderson Level I land cover classes, which are water, developed, barren, shrubland, herbaceous, planted/cultivated, and wetlands (Anderson et al., 1976). The developed open space, barren, and wetlands were excluded in this study and six land cover classes remained in the classified land cover maps. Barren lands are occupied by less than 15% vegetation and their effect on water quality is similar to those of the developed lands. Wetlands are composed of water and vegetation covers and they would be classified as either water or as whatever vegetation covers them.

The water, developed, forest, shrubland, herbaceous and planted land cover classes are all critical to stream nutrient and bacteria concentrations. Therefore, both the overall accuracy and the minimum accuracy of each class are important to the water quality models (Heydari and Mountrakis, 2018). Proportionally distributed training samples yield higher overall accuracy and equally distributed training samples lead to higher minimum accuracy of each class (Mellor et al., 2015; Zhu et al., 2016). In this study, a balance was sought between high overall accuracy and good accuracy within each class. Specifically, tests were conducted to find a balance between proportional samples and equal samples by increasing the sample size of the minority classes.

Logical trajectory information together with spectral characteristics was used to correct classification errors of the 20 independent land cover maps with a comprehensive post-

classification quality control approach. This quality control approach was modified based on Xian and Homer's method (Xian and Homer, 2010) and Jin and others' method (Jin et al., 2013), with an adjustment of control principles and threshold selections to adapt to the local conditions. The quality control process involved two steps. 1) The unchanged mask, the Biomass Increase (BI) mask and the Biomass Decrease (BD) mask were generated to recover some pixels' labels to those of NLCD 2011. 2) Classification maps were updated in a way that the changes of developed area and forest area were logical. Developed areas, once established, should not change to other land cover types; and if forest areas changed to other land cover types, they would not be able to change back in just 20 years.

In the first step, four spectral indices, including Change Vector (CV), the Relative Change Vector MAXimum (RCVMAX), the differenced Normalized Burn Ratio (dNBR), and the differenced Normalized Difference Vegetation Index (dNDVI), were used in the quality control process (Equation 3-1). The four indices indicate the spectral changing conditions of one image compared to another, which implies the possibilities of land cover change. In the equations, B_{1i} denotes the *i*th band of the early Landsat image and B_{2i} represents the *i*th band for the later Landsat image. CV and RCVMAX were used to generate the unchanged mask. For example, by comparing the classified image with NLCD 2011, water pixels with Z score of CV smaller than 2 or RCVMAX smaller than 1 were labeled as unchanged. The four indices were used together to generate the BI mask and the BD mask. For example, pixels with Z score of dNDVI larger than 0, dNBR larger than 1, and RCVMAX larger than 1 were designated as part of the BI mask. If the land cover changed from forest to grass, which was a biomass decrease, but the pixels were in the BI mask, they would be corrected to NLCD 2011 labels as forest. In the quality control process,

pairs of spectral indices in both leaf-on and leaf-off seasons were generated every year and combined with the "OR" principle.

$$dNBR = (B_{14} - B_{17}) / (B_{14} + B_{17}) - (B_{24} - B_{27}) / (B_{24} + B_{27})$$
$$dNDVI = (B_{14} - B_{13}) / (B_{14} + B_{13}) - (B_{24} - B_{23}) / (B_{24} + B_{23})$$
$$CV = \sum_i (B_{1i} - B_{2i})^2$$
$$RCVMAX = \sum_i \left[(B_{1i} - B_{2i}) / \max(B_{1i}, B_{2i})^2 \right]$$

Equation 3-1

The thresholds of the three change detection masks were defined with exploratory statistics and decision tree algorithms. The spectral characteristics of pixels with unchanged labels, biomass increase, and biomass decrease were carefully reviewed by comparing NLCD 2006 and NLCD 2011. Multi-threshold methods were designed using the four indices to generate the change detection masks. The quality control procedure was implemented in an iterated fashion from 1991 to 2011. Finally, the accuracy assessment was conducted by calculating the confusion matrix, the overall accuracy, and the kappa coefficient in 2006 and 2001 with NLCD as the validation data. R^2 was also calculated as the most important performance measurement, representing the agreement between the true number of land cover pixels and the classified number of land cover pixels among all the classes in all the subbasins.

- **Statistical analysis**

Land cover percentages were retrieved from the classification maps at a yearly base. The pollutant concentration data of NO_3^--N, PO_4^{3-}-P, NH_4^+-N, TP, and *E.coli* was aggregated yearly in both dry and wet seasons, as were the average temperature and total precipitation. All the pollutant

concentrations were log transformed to make them close to normal distributions. The land cover change trends as well as the nutrient and bacteria concentrations were explored.

Linear mixed models are key methods of modeling the spatial dependency and temporal dependency among the water quality samples. In a mixed model, fixed effects are assumed constant across samples while random effects vary. Random effects represent groups of samples that share the same unobserved characteristics in each group. Random intercept models were used in this study to avoid overcomplicated parameters and the over-fitting issue. For example, if basins were the only random intercepts, the underlying assumption was that except for land cover, topography, and climatic fixed effects, each basin has some unobserved factors that affect stream water quality, represented by a random intercept. In this study, the potential random effects were years, basins, regions, and sampling stations. The random effects were assumed to be independent from each other. The matrix form of a linear mixed model is as follows (Equation 3-2):

$$Y = X\beta + U\gamma + \varepsilon$$

Equation 3-2

In the above equation, Y is a known vector of observations, which is the vector of pollutant concentration of all the samples. X is the design matrix representing the fixed effect covariates of the samples, which are land cover, topography and climate. U is the design matrix of random effect covariates, which can be columns of years, regions, basins and sampling stations. β is the unknown fixed effect coefficient vector and γ is the unknown random effect coefficient vector to be estimated. In a random intercept model, the correlations among samples in the same group are assumed to be the same.

Several candidate models were compared in this study. The fixed effect covariates in all the models were percentages of water, developed, forest, shrubland, and planted land covers, temperature, precipitation, elevation, and slope. The dependent variables were yearly mean pollutant concentrations of NO_3^--N, PO_4^{3-}-P, NH_4^+-N, TP, and *E.coli*. Dry and wet season models were constructed separately. The first model was the fixed effect multiple linear regression models with no random effects. The second model had only random intercepts of years. The third model had random intercepts of years and ecoregions. The fourth model had random intercepts of years, ecoregions, and basins. The fifth model included random intercepts of years, ecoregions, basins and monitoring stations. Candidate models were compared with respect to the R^2, the Akaike Information Criterion (AIC), and the likelihood ratio test to detect significant differences between models. The selected model was used to draw the longitudinal relationships between land cover and pollutant concentrations.

3.3 Result

3.3.1 Land cover change in the Texas Gulf Region

- **Land cover classification accuracy**

There was strong agreement between the classified land cover maps and NLCD in both 2006 and 2001. The classified maps achieved 96.19% and 94.69% overall accuracy; and the kappa coefficients were 0.94 and 0.92 in 2001 and 2006 respectively. Table 3-1 and Table 3-2 show that the classification performed particularly well in mapping water and shrubland areas, with a recall of 99.01% and 97.31% in 2006, and a precision of 97.70% and 98.13% in 2001. The precision of developed areas was 87.91% in 2006 and 82.97% in 2001, where some developed areas were misclassified as planted areas. The recall of herbaceous areas was 90.13% in 2006 and 91.78% in 2001, where some shrublands and planted were misclassified as herbaceous. The R^2 of the true

land cover areas versus the classified land cover areas was 0.98 in both 2006 and 2001, which was calculated among all the 1783 subbasins. The R^2 of forest, shrubland and herbaceous were particularly high of 0.97, 0.99 and 0.97 in both 2006 and 2001.

Table 3-1. Confusion Matrix of the classification agreement compared with NLCD 2006.

2006 (OA=96.19%, kappa=0.94, r2=0.98)							
	water	developed	forest	shrubland	herbaceous	planted	precision
water	15199[a]	2	31	49	31	39	99.01%
developed	32	10386	210	577	159	450	87.91%
forest	6	119	67758	1295	648	76	96.93%
shrub	170	84	2435	303126	3380	2296	97.31%
herbaceous	52	88	391	2631	49065	2011	90.46%
planted	97	77	206	1219	1154	60827	95.67%
recall	97.70%	96.56%	95.39%	98.13%	90.13%	92.59%	
R^2	0.99	0.94	0.97	0.99	0.97	0.97	

a. The units of pixel numbers are 1000 pixels for all the land cover types.

Table 3-2. Confusion matrix of the classification agreement compared with NLCD 2001

2001 (OA=94.69%, kappa=0.92, R^2=0.98)							
	water	developed	forest	shrubland	herbaceous	planted	precision
water	17103[a]	3	30	94	51	48	98.70%
developed	66	13231	471	626	538	1015	82.97%
forest	27	95	69841	1693	1091	59	95.93%
shrub	194	57	3131	150444	2842	2008	94.81%
herbaceous	102	111	699	2497	66049	1425	93.18%
planted	396	83	788	1473	1395	95524	95.85%
recall	95.61%	97.43%	93.17%	95.93%	91.78%	95.45%	
R^2	0.94	0.92	0.97	0.99	0.97	0.96	

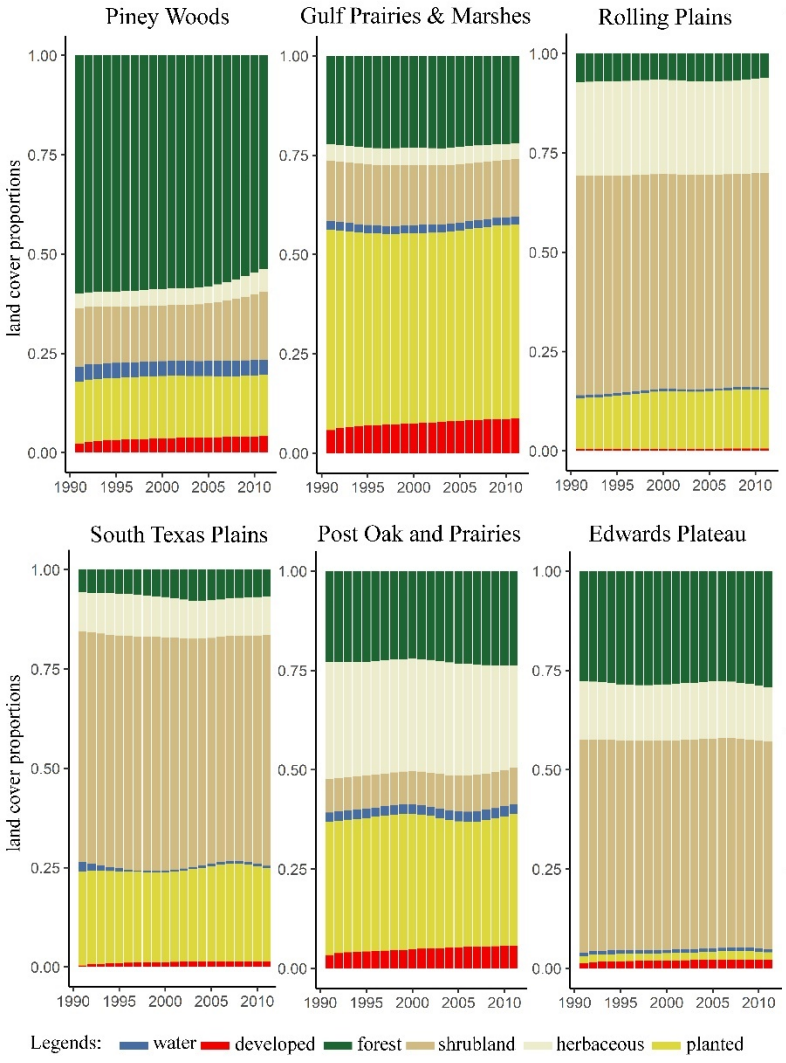
a. The units of pixel numbers are 1000 pixels for all the land cover types.

- **Land cover proportions and changes**

The land cover areal percentages from 1991 to 2011 were smoothed and presented in Figure 3-3, together with conversion tables of the six ecoregions (Figure 3-3). An obvious deforestation trend was found in the Piney Woods ecoregion. This region had the largest proportion of forest area, but more than 4000 km^2 of forest changed to shrubland or herbaceous land. The forest degradation

trend was particularly rapid from 2005 to 2011. The Gulf Prairies and Marshes ecoregion has the largest percentages of urban area and planted area. Around 1000 km^2 forest in this ecoregion changed to planted or developed areas from 1991 to 2011, but the deforestation trend has recently slowed.

The Post Oak and Prairies ecoregion was occupied by balanced proportions of forest, herbaceous, and planted areas. There seemed to be a forest restoration in this ecoregion after 2000. The South Texas Plains, the Rolling Plains and the Edwards Plateau ecoregions were primarily occupied by shrubland. Water area has decreased in the South Texas Plains, with more than 1000 km^2 water changing to planted or shrubland areas. In the Rolling Plains ecoregion, most of the land cover was relatively stable, with slight forest degradation.



Piney Woods

1991 \ 2011	water	developed	forest	shrubland	herbaceous	planted
water	2271	10	34	49	34	20
developed	0	1172	0	0	0	0
forest	29	466	27351	2861	1344	543
shrubland	21	245	939	6948	281	400
herbaceous	6	55	76	64	1516	166
planted	9	101	248	164	93	6034

Gulf Prairies & Marshes

1991 \ 2011	water	developed	forest	shrubland	herbaceous	planted
water	9985	28	7	14	47	180
developed	0	2677	0	0	0	0
forest	7	379	4904	108	56	775
shrubland	4	119	206	3869	167	510
herbaceous	18	73	27	35	1946	292
planted	102	676	507	573	708	23199

Rolling Plains

1991 \ 2011	water	developed	forest	shrubland	herbaceous	planted
water	437	3	5	49	32	2
developed	0	399	0	0	0	0
forest	9	10	4464	785	917	96
shrubland	12	44	10	31568	598	1032
herbaceous	9	23	9	252	18794	708
planted	3	9	1	597	370	9060

South Texas Plains

1991 \ 2011	water	developed	forest	shrubland	herbaceous	planted
water	278	159	30	726	155	396
developed	0	389	0	0	0	0
forest	1	9	1926	163	26	86
shrubland	8	153	121	29218	215	594
herbaceous	2	42	11	259	8504	288
planted	12	483	89	1477	1208	15620

Post Oak and Prairies

1991 \ 2011	water	developed	forest	shrubland	herbaceous	planted
water	2323	34	28	12	18	27
developed	0	3061	0	0	0	0
forest	24	318	21041	534	637	1390
shrubland	3	65	96	7439	69	250
herbaceous	34	1096	382	308	22818	2476
planted	37	712	380	964	1135	32879

Edwards Plateau

1991 \ 2011	water	developed	forest	shrubland	herbaceous	planted
water	42	1	0	1	1	0
developed	0	89	0	0	0	0
forest	1	24	1498	66	28	3
shrubland	0	12	13	4301	45	25
herbaceous	0	20	11	23	810	19
planted	0	2	1	32	13	148

Figure 3-3. Land cover proportions and conversions of the six ecoregions from 1991 to 2011.

3.3.2 The spatial and temporal distributions of nutrient and bacteria concentrations

There were large variations in both spatial and temporal distributions of nutrient and bacteria concentrations. Trend plots in different ecoregions of the yearly mean concentrations of NO_3^--N, PO_4^{3-}-P, NH_4^+-N, TP, and *E.coli* from 1991 to 2001 are present in Figure 3-4. The South Texas Plains, the Gulf Prairie and Marshes, the Piney Woods and the Post Oak and Prairies all faced the issue of the increasing NO_3^--N pollution, while the Rolling Plains had a decreasing trend of NO_3^--

N concentration. The increasing trend of NO_3^--N was particularly significant in the Gulf Prairie and Marshes ecoregion with the average concentration in 2011 rising to higher than 5.0 mg/l. The PO_4^{3-}-P concentration in the Gulf Prairie and Marshes and the South Texas Plains were higher than the other ecoregions. The average concentration in both regions were higher than 5 mg/l after 2005. There were also increasing trends of PO_4^{3-}-P in the Rolling Plains and Piney Woods. In the Rolling Plains and the Gulf Prairie and Marshes, there was an increasing trend of TP after 2000. The TP concentration in the Gulf Prairie and Marshes reached around 1mg/l in 2011, compared to around 0.5 mg/l in 2000. The high *E.coli* concentration in the Piney Woods and the Gulf Prairie and Marshes in 2001 was well controlled and had started to decrease since then. In 2011, the *E.coli* concentration in all the ecoregions was lower than 2000 MPN/100ml. However, the *E.coli* concentration slightly increased after 2009 in the Post Oak and Prairies, the Rolling Plains and the Edwards Plateau ecoregions.

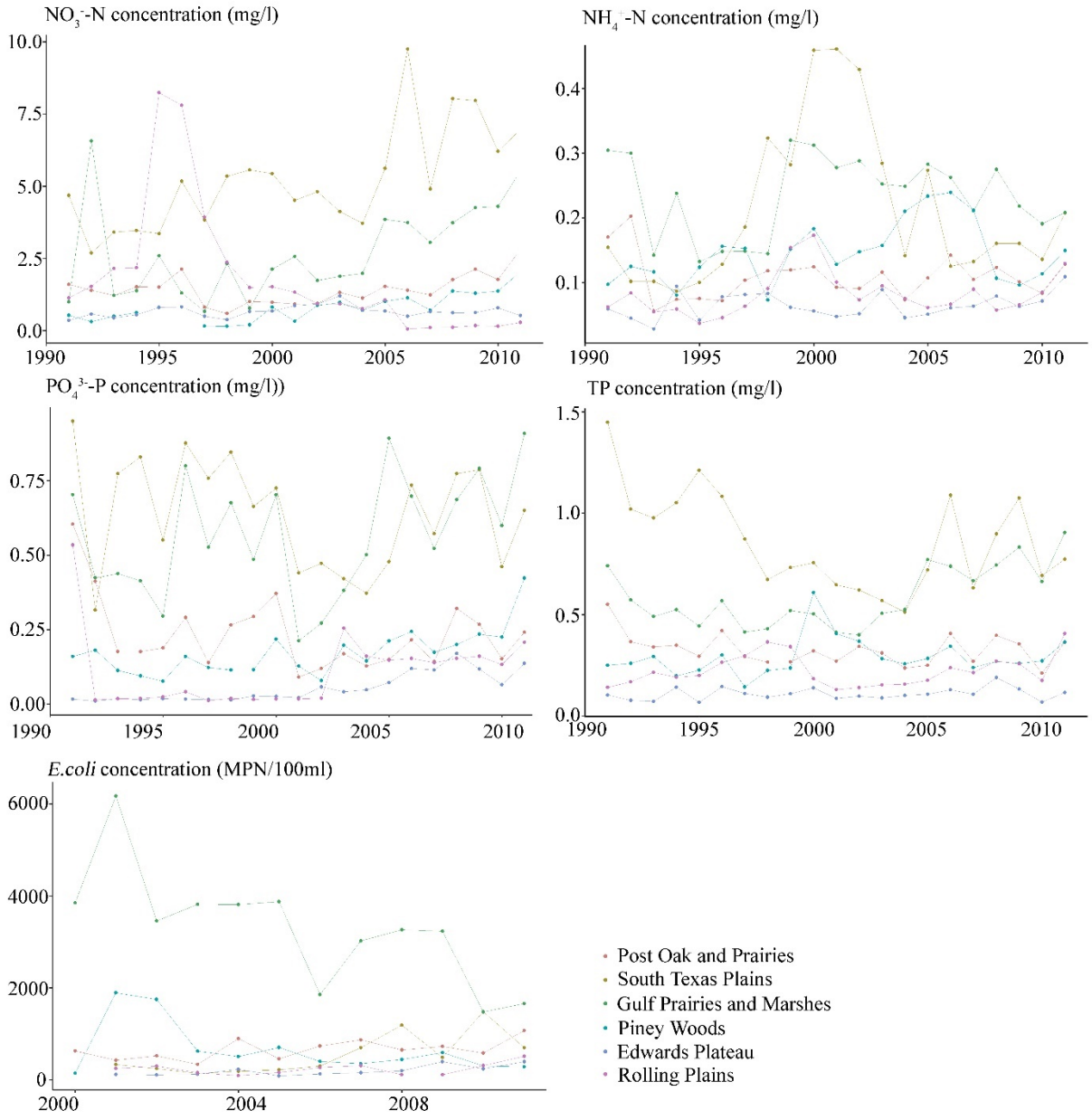
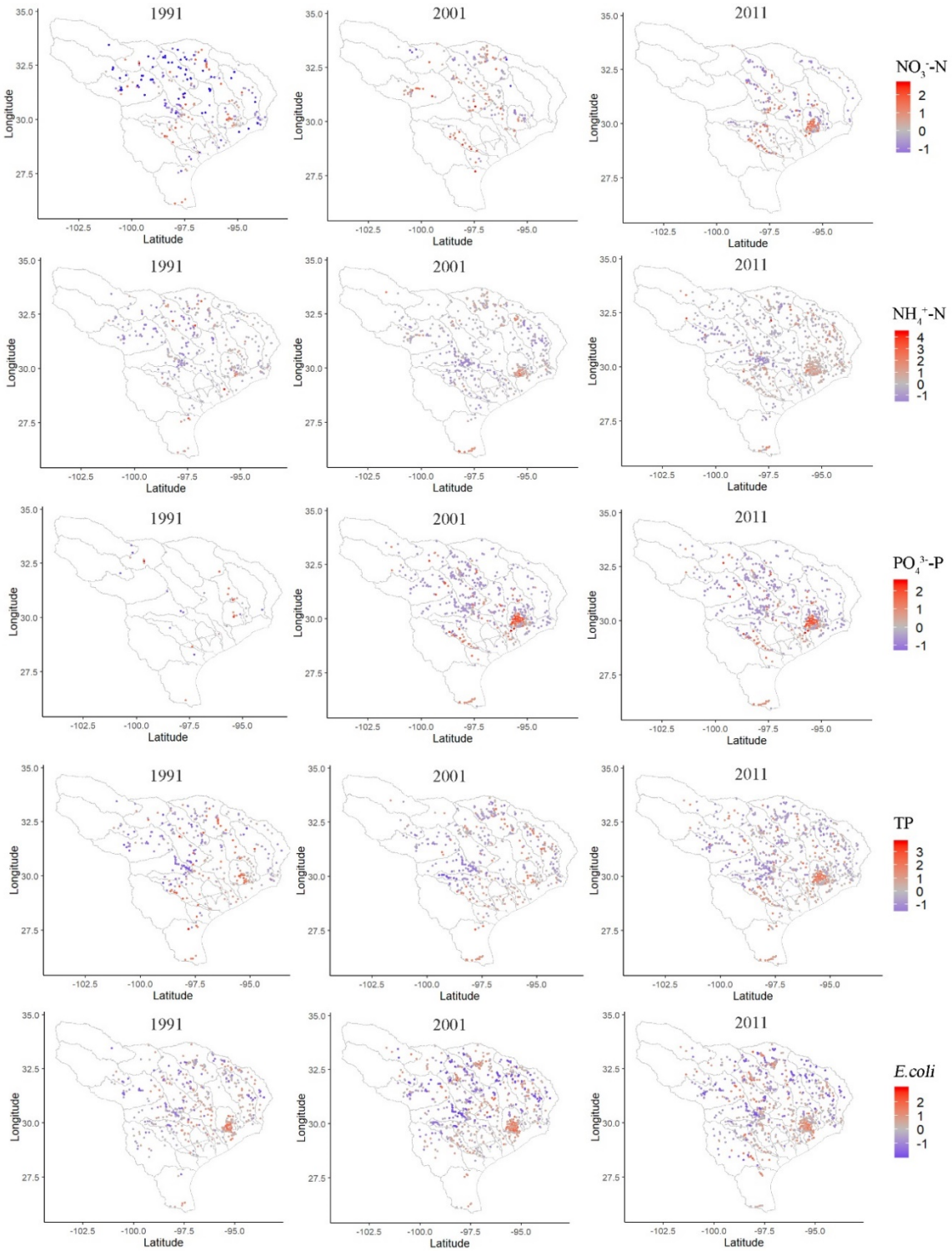


Figure 3-4. Change of nutrients and bacteria concentrations in the six ecoregions

The spatial distribution of nutrient and *E.coli* concentrations in 1991, 2001 and 2011 are present in Figure 3-5. After the log transformation and standardization, the positive range of pollutant concentrations was close to but larger than the negative range, indicating that there were some

extremely high concentration values for all the pollutants, represented by the dark red points in Figure 3-5.

The NO_3^--N concentration increased dramatically in the Middle Brazos, the Lower Brazos and the San Jacinto basins after 2001. The San Jacinto and the San Antonio basins faced the most severe NO_3^--N pollution, with the average concentration of 3.36 mg/l and 4.20 mg/l. NH_4^+-N concentration remained relatively stable from 1991 to 2011, with a slight increase in the Neches and the San Jacinto basins. The PO_4^{3-}-P had very few measurements in 2001, but some hotspots could still be found in the upper and lower Trinity basins and the San Jacinto basin. In 2011, the highest PO_4^{3-}-P concentration appeared in the San Jacinto, the Southwestern Texas Coastal and the San Antonio basins. TP concentrations increased significantly in the Neches basin and the San Jacinto basins from 2001 to 2011, with the most polluted areas along the coastal line. The highest average PO_4^{3-}-P concentration were in the San Jacinto basin and the Southwestern Texas Coastal basin, with the mean concentration of 0.74 mg/l and 0.64 mg/l. The *E.coli* concentration generally became lower after 1991. Areas with high *E.coli* concentrations were in the San Jacinto basin and the Southwestern Texas Coastal basin, with the mean concentration of 4343 MPN/100ml and 1517 MPN/100ml respectively.



Notes: The value in Fig. 5 is the standardized log-transformed yearly mean pollutant concentration. In the standardization process, the original values were subtracted by the mean and divided by the standard deviation, calculated based on all the 1991, 2001 and 2011 samples. The base map is the HUC 6 (basin) maps of the Texas Gulf Region.

Figure 3-5. The spatial distributions of nutrient and *E.coli* concentrations in 1991, 2001, and 2011.

3.3.3 The longitudinal relationship between land cover and water quality

- **The longitudinal model selection**

Comparison among the five models in predicting NO_3^--N concentration in wet seasons is present in Table 3-3. Model 1 contained only fixed effects and did not specify correlations among samples. The R^2 of this model was 0.31; and the coefficient of shrubland was significantly positive, which was not reasonable in reality. It proved that models with independent assumptions among samples might lead to wrong inference. After adding a random intercept of years in Model 2, R^2 increased to 0.35. The variance explained by the sampled year was 4%. The random intercept of ecoregions was added in Model 3 and R^2 increased to 0.4. In this model, 16% of the variance was partitioned to the ecoregions and only 2% was partitioned to years, indicating that the spatial variation of pollutant concentration was much larger than the temporal variation. Model 4 with random intercepts of years, ecoregions and basins had R^2 of 0.55. The coefficient of shrubland in this model changed to significantly negative, showing a reasonable result that shrubland had a positive impact on mitigating NO_3^--N pollution. In this model, 25% of the variance was explained by the basin intercept, 12% of the variance was explained by the ecoregion intercept, and only 2% of the variance was explained by year. The likelihood ratio tests were conducted to compare the five models and it was found that every model was significantly different from the previous one.

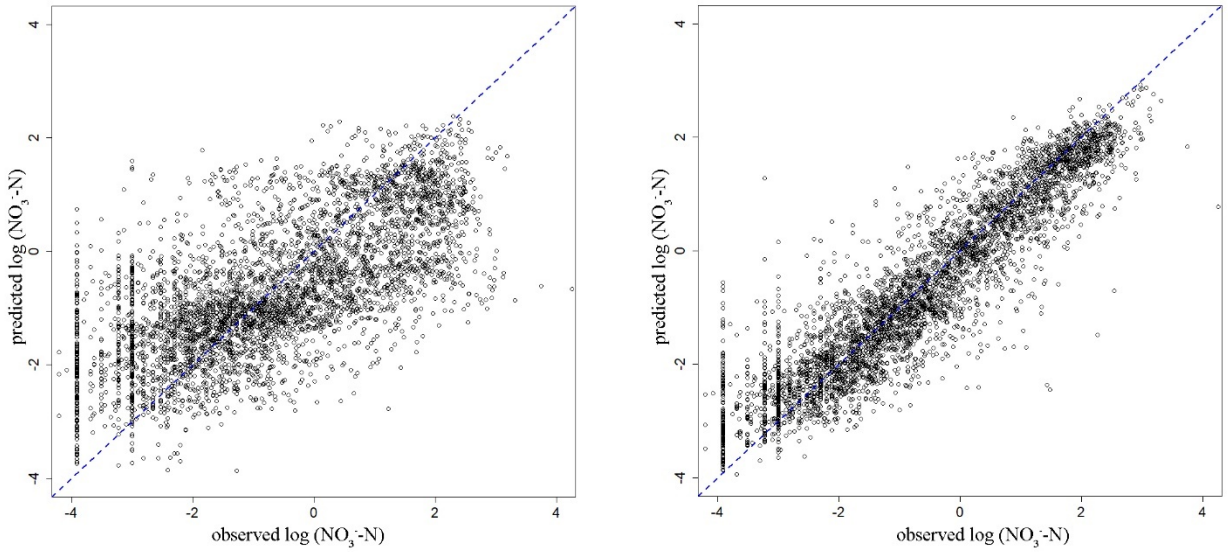
The R^2 of Model 4 and Model 5 were 0.54 and 0.82 respectively. Model 4 had a moderate prediction power while Model 5 performed the best in predicting NO_3^--N concentration (Figure 3-6). However, there was a generalizability issue with Model 5. The model variance explained by residuals was only 24% and 51% of the variance was explained by the location of monitoring

stations. Therefore, model 5 was more suited to explain location-based stream water quality, but not the general land cover effect on water quality. Considering that Model 4 had a balance of R^2 and generalization capacity, this model structure was used to draw inferences regarding the land cover effect on all the pollutants in the next step. Figure 3-6 also indicates that both models can predict NO_3^--N concentration better if the observed values are larger than 0.01 mg/l. Although the very small values were not predicted accurately, these values were not as important as the normal and high concentration values in reality.

Table 3-3. Candidate models to predict log (NO3-N) concentration in wet seasons and their comparison

		Model 1	Model 2	Model 3	Model 4	Model 5
		fixed effect model	model with random intercepts of year	model with random intercepts of year and ecoregion	model with random intercepts of year, ecoregion and basin	model with random intercepts of year, ecoregion, basin and monitoring station
model coefficients and significance	%forest	-0.89**[a]	-0.82**	-0.61**	-0.76**	-0.56
	%developed	2.48**	2.19**	2.53**	1.99**	2.12**
	%planted	2.02**	1.84**	1.86**	2.68**	2.65**
	%shrubland	0.94**	0.69**	0.12	-0.85**	-0.69
	%water	-2.30**	-2.34**	-2.29**	-2.27**	-1.51**
	year after 1991	0.0015	NA	NA	NA	NA
	slope	-0.026	-0.039	-0.13**	-0.013	-0.039
	elevation	0.0002**	0.0016**	0.0029**	0.0016**	0.0018**
	precipitation	0.0011	0.0023	0.0019*[b]	0.0015	0.0021**
	temperature	0.40**	0.60**	0.45**	0.25**	0.15**
model performance	R^2	0.31	0.35	0.40	0.55	0.82
	AIC	15656	15595	15393	14725	12060
variance partitions	residual variance proportion	100%	96%	82%	57%	24%
	year variance proportion	NA	4%	2%	2%	2%
	ecoregion variance proportion	NA	NA	16%	10%	7%
	basin variance proportion	NA	NA	NA	31%	16%
	station variance proportion	NA	NA	NA	NA	51%

a. ** indicates $p < 0.01$
b. * indicates $0.01 < p < 0.05$



Notes: In Fig. 6, the left figure is the scatter plot of Model 4 and the right figure is the scatter plot of Model 5. The unit of the original value is mg/l

Figure 3-6. The scatter plots of predicted values vs observed values of Model 4 and Model 5.

- **The longitudinal model inference**

The relationship between land cover and NO_3^--N in wet seasons was the strongest among all the pollutants, represented by an R^2 of 0.55. The R^2 of PO_4^{3-}-P, TP, and *E.coli* models in wet seasons were 0.44, 0.39, and 0.41 respectively. The relationship between land cover and NH_4^+-N was relatively weak, as indicated by an R^2 of 0.3 in wet seasons. The land cover effect on stream water quality were generally stronger in wet seasons than in dry seasons (Table 3-4).

The positive impact of forest was significant in reducing all the nutrient concentrations. The impact of forest was particularly strong in wet seasons in mitigating NO_3^--N, TP, and NH_4^+-N pollution. After some calculation, it was found that if adding 1 percent of forest area, the NO_3^--N concentration in wet seasons was expected to drop 1.14%, and the TP concentration in wet seasons was expected to drop 1.36%. Developed land cover was significantly positively associated with all the pollutant concentrations. The impact was strong in both dry and wet seasons. For example,

a 1% addition of developed area caused a 5.23% increase of *E.coli* concentration and a 6.31% increase of NO_3^--N concentration in wet seasons. The significantly positive impact of planted area on NO_3^--N concentration was very strong. Adding 1% of planted area led to a 13.59% increase of NO_3^--N in wet seasons. Planted area was also significantly positively associated with PO_4^{3-}-P, TP, NH_4^+-N, and *E.coli* concentrations. Water area significantly reduced NO_3^--N, PO_4^{3-}-P, TP, and *E.coli* concentrations. Adding 1% of water area led to an 8.6% decrease in NO_3^--N concentration and a 6.7% decrease in TP concentration. Water had the most significant influence on reducing *E.coli* concentration, and the contribution might be attributed to some wetland areas. Shrubland area had a significantly negative association with NO_3^--N, TP, and *E.coli* concentrations, with the impact strongest on NO_3^--N. Adding 1% of shrubland area caused a 1.3% decrease of NO_3^--N concentration. Slope generally had a negative impact on pollutant concentrations. In summary, the most influential land covers were developed and planted areas, with a negative impact, and water/wetland areas, with a positive impact in the study area.

Table 3-4. Mixed model results to predict pollutant concentrations

	log (NO_3^--N)		log (PO_4^{3-}-P)		log (TP)		log (NH_4^+-N)		log (*E.coli*)	
season	wet	dry	wet	dry	wet	dry	wet	dry	wet	dry
%forest	-0.76**[a]	-0.39	-0.34*	-0.34*	-0.74**	-0.27**	-0.86**	-0.48**	-0.12	-0.39
%developed	1.99**	2.01**	1.51**	1.59**	0.42**	0.97**	0.39**	0.57**	1.83**	2.55**
%planted	2.68**	2.29**	0.91**	0.85**	0.39**	1.03**	-0.002	0.39**	0.49**	0.19
%shrubland	-0.85**	-0.59*	0.02	0.18	-0.34**	-0.58**	-0.41**	-0.095	0.16	-0.33
%water	-2.27**	-3.84**	-1.35**	-1.15**	-1.92**	-2.12**	-0.46**	-0.03	-8.34**	-8.54**
slope	-0.013	-0.041	-0.019	-0.039**	-0.069**	-0.098**	-0.0023	-0.0079	-0.13**	-0.16**
elevation	0.0016**	0.0016**	0.00021	0.00051	-0.0012**	-0.00069**	0.00027	0.00037*	0.0012**	0.0021**
precipitation	0.0015	0.0024*[b]	-0.00007	0.00021	-0.0018**	0.00079	-0.00068*	-0.00081	0.0033**	0.0025*
temperature	0.25**	0.024**	-0.025	0.0071	-0.076	-0.079**	0.037	0.014	-0.069	0.0059
R^2	0.55	0.49	0.44	0.39	0.39	0.36	0.3	0.21	0.41	0.37

a. ** indicates $p < 0.01$
b. * indicates $0.01 < p < 0.05$

The basin random intercepts of all the wet-season mixed models are present in Figure 3-7, which represented the unobserved basin characteristics that adjusted the stream water quality prediction. The Middle Colorado-Concho basin and the Middle Brazos-Clear Fork basin had some characteristics leading to high NO_3^--N concentration. After some calculations, it was found that 6.7 mg/l and 4.6 mg/l should be added besides the fixed effects when estimating NO_3^--N concentrations in the above two basins. The Lower Trinity basin and the Lower Colorado basin were likely to have higher PO_4^{3-}-P concentration. A 2.48 mg/l and a 1.89 mg/l should be added when estimating PO_4^{3-}-P concentration in the two basins. The Middle Brazos-Clear Fork basin was likely to have a higher TP concentration, where a 2.77 mg/l should be added to the predicted results.

According to the random intercepts of ecoregions, The South Texas Plains and the Gulf Prairie and Marshes ecoregions had positive random intercepts for all the pollutants, while the Rolling Plains and Edwards Plateau ecoregions had negative random intercepts for all the pollutants. The Piney Woods ecoregion had some positive characteristics that led to higher NO_3^--N and *E.coli* concentrations. The Post Oak and Prairies ecoregion also had some factors that caused higher *E.coli* concentration. When considering random intercepts of years, pollutant concentration after 2006 was likely to be higher than in earlier years under a fixed land use scenario.

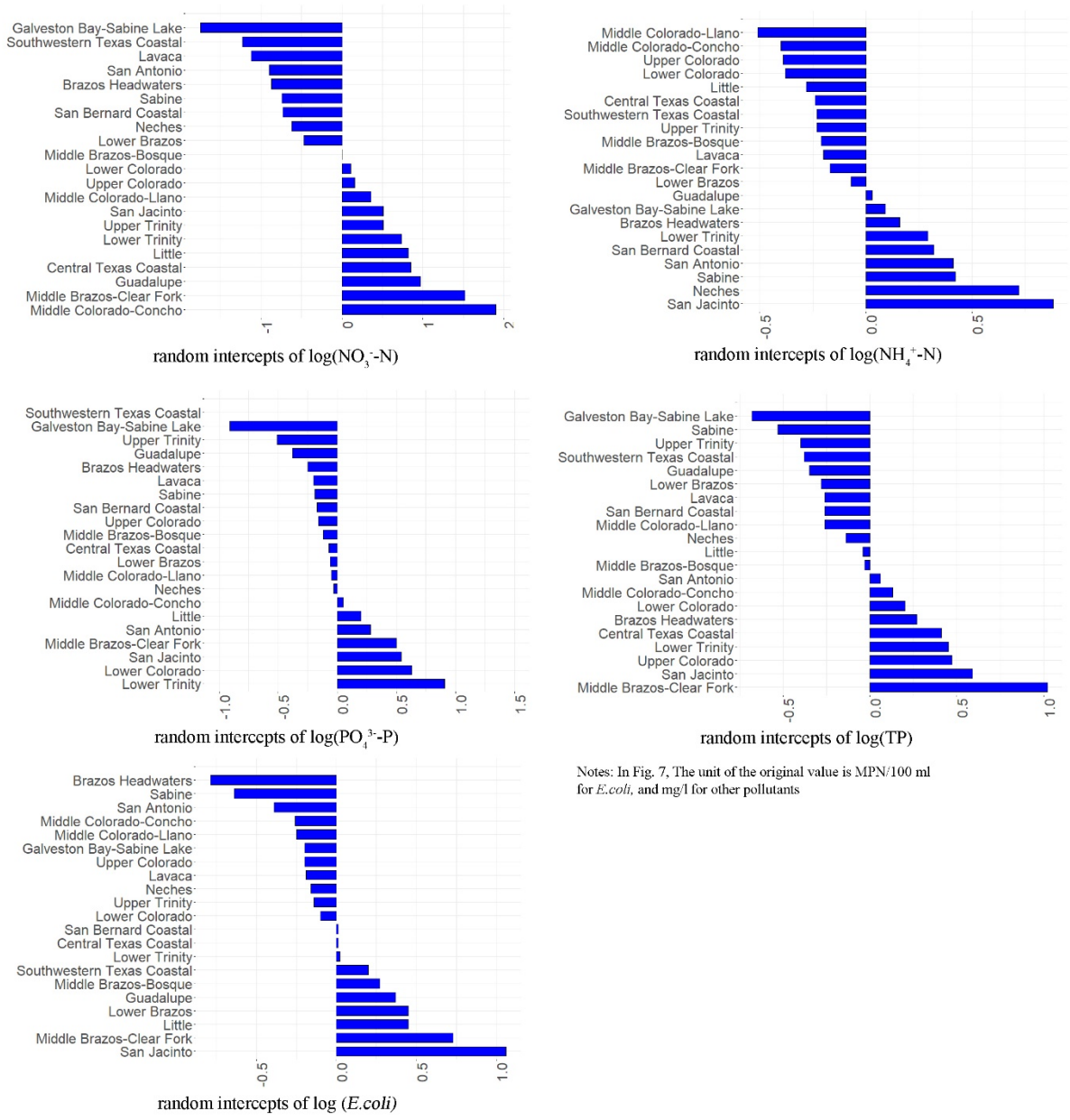


Figure 3-7. Bar charts of random intercepts of basin

3.4 Discussion

3.4.1 The impact factors on stream water quality in the Texas Gulf Region

With an abundance of water quality data provided by TCEQ, water quality study in Texas was still very limited (Santhi et al., 2006; Gelca et al., 2016). The land cover effect on nutrients and bacteria obtained from this study was qualitatively consistent with existing research in other regions. Quantified land cover effect was modified for the Texas Gulf Region, with the effect primarily focused on the regional scale.

The most important land cover affecting phosphorous concentration was found to be agricultural land in some research (Nielsen et al., 2012; Varanka and Luoto, 2012; Zhang et al., 2018). Urban area was also proved to have a disproportionately large influence on nutrient generation (Ai et al., 2015; Huang et al., 2016; Wijesiri et al., 2018). In the Texas Gulf Region, planted, water and developed areas were comparably important to predict NO_3^--N concentration. The percentage of water area was the most important land cover to predict TP concentration, while the percentages of developed and planted areas were the secondary important predictors. Water was also the most important factor to mitigate *E.coli* pollution. The reason why water area was highlighted in this study might be that parts of wetlands were classified as water under this classification scheme; and wetlands could keep nutrients and sediments from entering the lakes and streams (Galgraith and Burns, 2007). The significance of shrubland was not mentioned much in existing literature (Meneses et al., 2015), but deserves attention in the Texas Gulf Region because shrubland occupied the largest proportion of land. The positive impact of shrubland on water quality was about similar to that of forest on NO_3^--N concentration and weaker on other nutrients and *E.coli* concentrations in the study area.

The results in this study suggested that stream water quality was generally better explained by landscape attributes in wet seasons than in dry seasons, which were consistent with existing literature (Sheldon et al., 2012; Lv et al., 2015; Shi et al., 2017). Slope was found to be negatively associated with all the pollutants, with significant effects on PO_4^{3-}-P, TP, and *E.coli* concentrations. This conclusion agreed with some literature that water quality was generally better in high slope sub-catchment because pollutants tend to decrease when water flows faster (Lv et al., 2015; Shi et al., 2016). However, some researchers claimed that a gentle slope could slow down water movement and provide a longer time to decompose pollutants (Pratt and Chang, 2012; Bu et al., 2014). Temperature had a significantly positive association with NO_3^--N concentration. Precipitation had a significantly negative association with PO_4^{3-}-P and NO_3^--N concentrations, which agreed with pervious findings of the dilution and degradation effects of rainfall on phosphates (Rothwell et al., 2010; Varanka and Luoto, 2012).

3.4.2 The performance of the model system

- **The performance of the land cover classification algorithm**

Before quality control, the classification accuracy in this study was improved from 80% to 89% in the test set after adding all the ancillary features. The inclusion of the terrain-based ancillary features and the multi-seasonal information improved the classification accuracy of vegetation classes substantially, as was shown in other studies (Lu and Weng, 2007; Sluiter et al., 2010; Eisavi et al., 2015; Yang et al., 2017). The application of texture features was helpful in discriminating urban classification, but its importance was not as great as terrain-based ancillary features and seasonal information (Ghimire et al., 2010; Gomariz-Castillo et al., 2017).

It was difficult to use a single classifier to capture all the local spectral heterogeneity in a large scale image classification (Millard and Richardson, 2015; Zhu et al., 2016; Zhang and Roy, 2017).

In this study, the local random classifiers implemented in each ecoregion improved the overall accuracy of the single random classifier from 79% to 89% in the test set before quality control. Ecoregion division was the most efficient approach in this study because the land cover percentages distribution was similar elsewhere within the same ecoregion.

The quality control process significantly improved the classification performance. After quality control, the classification R^2 was improved from 0.94 to 0.98. The multi-threshold method of identifying land cover change groups was adopted in this study and combined with knowledge-based rules to update land cover maps every year (Griffiths et al., 2014; Kim et al., 2014; Yu et al., 2016; Jin et al., 2017;). The decision tree algorithms demonstrated a high efficiency in generating thresholds in the quality control process using the label changing and spectral information learned from NLCD 2001 and 2006 (Yang et al., 2017; Wang et al., 2019).

- **The performance of the linear mixed models**

The advantages of the methodology in this study were a large spatial extent, a long time range and a large sample size, which could be used to draw more general and credible conclusions. In the existing literature, most site areas ranged from 1000 km^2 to 5000 km^2 (Fatehi et al., 2015; Grabowski et al., 2016; Gu et al., 2016), where the land cover and environmental characteristics of the study area might be homogenous. The study site of this research was 471, 080 km^2 with a great deal of variation in climate and landscape. In this study, the number of sampling stations of each pollutant was around 1000, much more than in the previous literature, which always used fewer than 100 sampling stations (Amiri and Nakane, 2009; Varanka et al., 2015; Liu et al., 2017). It was revealed by previous studies that cross-sectional and longitudinal data analysis might generate different inferences about the land use effect on water quality (Wijesiri et al., 2018). This

study overcame the lack of reliability issues in the cross-sectional model by generating 20-year land cover data as the explanatory variables.

The linear mixed models with random intercepts of years, ecoregions and basins explained from 21% to 55% of the observed variance in the water quality data, which was comparable with other research using similar methods (Uriarte et al., 2011). The predicting accuracy was lower than that of GWR in other research because the local estimation of model coefficients was omitted (Yu et al., 2013; Sun et al., 2014; Huang et al., 2015). If random intercepts of the location of monitoring stations were added into the model, R^2 could be improved to around 0.8. Using this model structure, a constant estimation of regional-scale coefficients across the study area were acquired, with the basin-scale variation partitioned to the random intercepts. This approach was well-suited to prompt a regional understanding of the stream water quality in Texas.

3.4.3 The limitations of the study and future research suggestions

One limitation of this study was that the land cover classification scheme was coarse and might conceal some important information (Wan et al., 2014). Additionally, although the land cover classification demonstrated strong agreement with NLCD 2001 and 2006 with an overall accuracy higher than 94%, the accuracy was expected to be even higher to detect subtle land cover changes accurately. Because it was not reasonable for land cover type to change back in a short period of time, the land cover percentage data was smoothed with the spline fitting method to derive the final input to the linear mixed models. The smoothed percentages might also introduce some errors. Future quality control algorithms should focus on the combination of knowledge-based methods and spectral trajectory at the pixel level to design more efficient change detection algorithms.

The inference of statistical models depends highly on the variable inclusion, sample selection and model assumption. Some explanatory variables were not readily available; therefore they were

not included in this study, such as landscape configuration metrics, soil, geology and population dynamics, which might cause biased model estimations (Chen and Lu, 2014; Sheldon et al., 2012; Sangani et al., 2015; Wilson, 2015; Bostanmaneshrad et al., 2018). In addition, the cross-correlations between the explanatory variables were not investigated, such as the cross-correlation between land covers, and the cross-correlation between land cover and climate (Li et al., 2015; Hwang et al., 2016).

In this research, the samples were aggregated in dry and wet seasons separately every year, as seasonal variation affected the relationship between land use and water quality (Hwang et al., 2016; Ai et al., 2015; Oeding et al., 2018). If taking fine-resolution climatic variables into account, such as monthly climatic variables and antecedent dry period, a finer aggregation scheme such as monthly aggregation of samples or even no aggregation should be applied to keep as much variation in the data as possible (du Plessis et al., 2015; Uwimana et al., 2017; Mello et al., 2018). With respect to spatial aggregation, the entire subbasin was adopted as the spatial aggregation unit in this study, because some literature reported that the entire watershed approach explained more variations than the riparian buffer zone approach (Pratt and Chang, 2012; Bu et al., 2014). Local-scale research should still compare catchment, riparian buffer, and reach buffer approaches to investigate the scale where each land cover type had an influence on water quality (Zhang et al., 2012; Ding et al., 2016; Liu et al., 2017).

The reasons why I selected the random intercept models were to match the objective of regional water quality estimation and to avoid too overly complex model parameters. However, this model structure might oversimplify the spatial and temporal variations in the relationship between land cover and water quality. I suggest that other statistical models can be used to extend the method framework of this study. For example, feature selection techniques can be applied to identify the

most influential independent variables prior to the regression analysis. The most important features can be selected via PCA, Redundancy Analysis (RDA), Hierarchical Partitioning (HP) and so on (Zhao et al., 2015; Huang et al., 2016; Kändler et al., 2017; Bostanmaneshrad et al., 2018; Oeding et al., 2018). Cluster analysis such as hierarchical clustering and Self-Organizing Maps (SOM) can be applied to group samples into multiple clusters according to land use and pollutant levels; and regression analysis can be conducted among samples within the same groups (Ye et al., 2009; Liu et al., 2018; Mello et al., 2018; Zhang et al., 2018). In the regression analysis, Bayesian linear regression models with random effects can be used to decompose the interactions among data into a series of conditional models and infer the distribution of model parameters (Wan et al., 2014; Wijesiri et al., 2018).

3.4.4 Model applications and management implications

The land cover maps were produced in a standard workflow on the GEE platform. The classification and change detection algorithm relied only on the Landsat imagery and an accurate land cover map in any recent year. It could be readily applied to many parts of the world to obtain historical land cover data. Similarly to other land cover classification research, the GEE platform in this study demonstrated high efficiency in automating the classification process all the way from sample generation, feature derivation to classifier training and results output (Patel et al., 2015; Gorelick et al., 2017; Huang et al., 2017; Zhao and Gao, 2019).

This study provided a solution to understand the evolution of Texas land cover with a robust classification algorithm. More information can be extracted from the classified land cover maps such as land cover trends in basins, counties, and cities to inform land use policies. According to the land cover changing status, there was a considerable deforestation trend and the corresponding ecological damage to the Piney Woods ecoregion after 2000. The reforestation efforts should be

exerted in this region to avoid further habitat loss (World Wildlife Fund, 2019). More than one third of the Texas population lives in the Gulf Prairie and Marshes ecoregion, which has been impacted by many human-induced factors. There was a more than 1000 km^2 increase of developed area and a more than 500 km^2 decrease of forest area from 1991 to 2011. The quality of the remaining habitat in this region faces drastic declines with habitat fragmentation, which requires immediate restoration actions (Texas Parks and Wildlife Department, 2012).

The proposed models are helpful for the modification of multiscale land use planning. The fixed effect land cover coefficients represent the relationship between land cover and water quality at the regional scale. Under a basin-scale land use planning scenario, water quality can be forecasted by plugging the land use percentages and the corresponding control factors into the linear mixed models, and adding the random intercepts of ecoregions and basins. It can then be decided whether the resulting contaminant concentration meets the regulation standards under the given land use scenario. The model framework is also flexible for local water quality estimation by fitting a mixed-effect model with random intercepts of monitoring stations. After the Land Change Monitoring, Assessment and Projection (LCMAP) data is published, annual land cover data from 1985 to 2017 can be derived to conduct similar research in other regions using the proposed linear mixed model structures (Zhu et al., 2016).

The inference of land cover effect on stream water quality can be directly applied to modify land use and watershed management policies. For example, land use planning should be adjusted by controlling low density urban development that occupies forest and shrub areas to mitigate NO_3^--N, PO_4^{3-}-P, and *E.coli* pollution (Bateni et al., 2013; Tu, 2013). Precision agriculture and conservation tillage should be applied in areas with high nutrient concentration such as the South Texas Plains and the Gulf Prairies and Marshes ecoregions, to reduce nutrient export from the

croplands (Shi et al., 2017). The positive impact of water and wetland areas on reducing *E.coli* concentration should be considered to guide policy in areas with rising *E.coli* concentration, such as in the South Texas Plains (Boutilier et al., 2009; Croft-White et al., 2017). In addition, the information in the random components provided some baseline information of the ecoregions and basins. Research efforts should be directed to find the unobserved factors leading to NO_3^--N and TP pollution in the Middle Colorado basin, and the factors causing NO_3^--N and PO_4^{3-}-P pollution in the Lower Trinity basin.

3.5 Conclusion

I completed a regional-scale longitudinal study of stream water modelling with land cover, terrain, and climate characteristics in the Texas Gulf Region. It involved a two-step method composed of annual land cover map classification and land cover-water quality modelling. It was the first study making use of all the available stream water quality data in a 20-year time range to derive scientific knowledge and management implications for the Texas Gulf Region.

The classified land cover maps had strong agreement with NLCD 2006 and 2001, with an accuracy of 97.70%, 96.56%, 95.39%, 98.13%, 90.13%, and 92.59% for water, developed, forest, shrubland, herbaceous, and planted land covers in 2006. The overall R^2 of the classified land cover areas versus true land cover areas calculated from all the subbasins was 0.98 in both 2001 and 2006. From the land cover maps, an obvious deforestation trend was observed in the Piney Woods, the South Texas Plains and the Gulf Prairies and Marshes ecoregions after 2000.

Linear mixed models with random intercepts of multiple spatial units can provide multiscale inference of land cover impact on water quality. Random components of years, ecoregions, and basins should be included to account for the spatial and temporal variations. The land cover change together with the terrain and climate factors explained more than 50% of the variance in NO_3^--N

concentration and more than 30% of the variance in PO_4^{3-}-P, TP, NH_4^+-N, and *E.coli* concentrations in the Texas Gulf Region. The most influential land cover types, which were significantly positively correlated with all the nutrient and bacteria concentrations, were developed areas and planted areas. Increasing water areas had a strong impact on the removal of NO_3^--N and *E.coli*.

The estimation of random intercepts provided important information regarding the unobserved basin and ecoregion characteristics that affect stream water quality. The Middle Colorado-Concho, the Lower Trinity and the San Jacinto basins had some unobserved characteristics leading to high nutrient and bacteria concentrations, with most of pollution hot spots found around the Houston metropolitan area. To sum up, this research could be applied to provide insights into the knowledge of land-water interactions, to evaluate new land use scenarios, and to inform scientific regional planning and watershed management policies.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ai, L., Shi, Z. H., Yin, W., & Huang, X. (2015). Spatial and seasonal patterns in stream water contamination across mountainous watersheds: Linkage with landscape characteristics. Journal of Hydrology, 523, 398–408. doi:10.1016/j.jhydrol.2015.01.082

Amiri, B. J., & Nakane, K. (2008). Modeling the Linkage Between River Water Quality and Landscape Metrics in the Chugoku District of Japan. Water Resources Management, 23(5), 931–956. doi:10.1007/s11269-008-9307-z

Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). A land use and land cover classification system for use with remote sensor data. Professional Paper. doi:10.3133/pp964

Bateni, F., Fakheran, S., & Soffianian, A. (2013). Assessment of land cover changes & water quality changes in the Zayandehroud River Basin between 1997–2008. Environmental Monitoring and Assessment, 185(12), 10511–10519. doi:10.1007/s10661-013-3348-3

Bonansea, M., Rodriguez, M. C., Pinotti, L., & Ferrero, S. (2015). Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina). Remote Sensing of Environment, 158, 28–41. doi:10.1016/j.rse.2014.10.032

Bostanmaneshrad, F., Partani, S., Noori, R., Nachtnebel, H.-P., Berndtsson, R., & Adamowski, J. F. (2018). Relationship between water quality and macro-scale parameters (land use, erosion, geology, and population density) in the Siminehrood River Basin. Science of The Total Environment, 639, 1588–1600. doi:10.1016/j.scitotenv.2018.05.244

Boutilier, L., Jamieson, R., Gordon, R., Lake, C., & Hart, W. (2009). Adsorption, sedimentation, and inactivation of E. coli within wastewater treatment wetlands. Water Research, 43(17), 4370–4380. doi:10.1016/j.watres.2009.06.039

Bu, H., Meng, W., Zhang, Y., & Wan, J. (2014). Relationships between land use patterns and water quality in the Taizi River basin, China. Ecological Indicators, 41, 187–197. doi:10.1016/j.ecolind.2014.02.003

Carey, R. O., Migliaccio, K. W., Li, Y., Schaffer, B., Kiker, G. A., & Brown, M. T. (2011). Land use disturbance indicators and water quality variability in the Biscayne Bay Watershed, Florida. Ecological Indicators, 11(5), 1093–1104. doi:10.1016/j.ecolind.2010.12.009

Chen, J., & Lu, J. (2014). Effects of Land Use, Topography and Socio-Economic Factors on River Water Quality in a Mountainous Watershed with Intensive Agricultural Production in East China. PLoS ONE, 9(8), e102714. doi:10.1371/journal.pone.0102714

Chen, X., Zhou, W., Pickett, S., Li, W., & Han, L. (2016). Spatial-Temporal Variations of Water Quality and Its Relationship to Land Use and Land Cover in Beijing, China. International

Journal of Environmental Research and Public Health, 13(5), 449. doi:10.3390/ijerph13050449

Chu, H.-J., Liu, C.-Y., & Wang, C.-K. (2013). Identifying the Relationships between Water Quality and Land Cover Changes in the Tseng-Wen Reservoir Watershed of Taiwan. International Journal of Environmental Research and Public Health, 10(2), 478–489. doi:10.3390/ijerph10020478

Crist, E. P., & Cicone, R. C. (1984). A Physically-Based Transformation of Thematic Mapper Data---The TM Tasseled Cap. IEEE Transactions on Geoscience and Remote Sensing, GE-22(3), 256–263. doi:10.1109/tgrs.1984.350619

Croft-White, M. V., Cvetkovic, M., Rokitnicki-Wojcik, D., Midwood, J. D., & Grabas, G. P. (2017). A shoreline divided: Twelve-year water quality and land cover trends in Lake Ontario coastal wetlands. Journal of Great Lakes Research, 43(6), 1005–1015. doi:10.1016/j.jglr.2017.08.003

Ding, J., Jiang, Y., Liu, Q., Hou, Z., Liao, J., Fu, L., & Peng, Q. (2016). Influences of the land use pattern on water quality in low-order streams of the Dongjiang River basin, China: A multi-scale analysis. Science of The Total Environment, 551-552, 205–216. doi:10.1016/j.scitotenv.2016.01.162

Doetterl, S., Stevens, A., van Oost, K., Quine, T. A., & van Wesemael, B. (2013). Spatially-explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. Geoderma, 204-205, 31–42. doi:10.1016/j.geoderma.2013.04.007

Du Plessis, A., Harmse, T., & Ahmed, F. (2015). Predicting water quality associated with land cover change in the Grootdraai Dam catchment, South Africa. Water International, 40(4), 647–663. doi:10.1080/02508060.2015.1067752

Eisavi, V., Homayouni, S., Yazdi, A. M., & Alimohammadi, A. (2015). Land cover mapping based on random forest classification of multitemporal spectral and thermal images. Environmental Monitoring and Assessment, 187(5). doi:10.1007/s10661-015-4489-3

Fatehi, I., Amiri, B. J., Alizadeh, A., & Adamowski, J. (2015). Modeling the Relationship between Catchment Attributes and In-stream Water Quality. Water Resources Management, 29(14), 5055–5072. doi:10.1007/s11269-015-1103-y

Galbraith, L. M., & Burns, C. W. (2006). Linking Land-use, Water Body Type and Water Quality in Southern New Zealand. Landscape Ecology, 22(2), 231–241. doi:10.1007/s10980-006-9018-x

Gelca, R., Hayhoe, K., Scott-Fleming, I., Crow, C., Dawson, D., & Patiño, R. (2015). Climate-water quality relationships in Texas reservoirs. Hydrological Processes, 30(1), 12–29. doi:10.1002/hyp.10545

Ghimire, B., Rogan, J., & Miller, J. (2010). Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. Remote Sensing Letters, 1(1), 45–54. doi:10.1080/01431160903252327

Giri, S., & Qiu, Z. (2016). Understanding the relationship of land uses and water quality in Twenty First Century: A review. Journal of Environmental Management, 173, 41–48. doi:10.1016/j.jenvman.2016.02.029

Gomariz-Castillo, F., Alonso-Sarría, F., & Cánovas-García, F. (2017). Improving Classification Accuracy of Multi-Temporal Landsat Images by Assessing the Use of Different Algorithms, Textural and Ancillary Information for a Mediterranean Semiarid Area from 2000 to 2015. Remote Sensing, 9(10), 1058. doi:10.3390/rs9101058

Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 116, 55–72. doi:10.1016/j.isprsjprs.2016.03.008

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18–27. doi:10.1016/j.rse.2017.06.031

Grabowski, Z. J., Watson, E., & Chang, H. (2016). Using spatially explicit indicators to investigate watershed characteristics and stream temperature relationships. Science of The Total Environment, 551-552, 376–386. doi:10.1016/j.scitotenv.2016.02.042

Griffiths, P., Kuemmerle, T., Baumann, M., Radeloff, V. C., Abrudan, I. V., Lieskovsky, J., … Hostert, P. (2014). Forest disturbances, forest recovery, and changes in forest types across the Carpathian ecoregion from 1985 to 2010 based on Landsat image composites. Remote Sensing of Environment, 151, 72–88. doi:10.1016/j.rse.2013.04.022

Gu, Q., Zhang, Y., Ma, L., Li, J., Wang, K., Zheng, K., … Sheng, L. (2016). Assessment of Reservoir Water Quality Using Multivariate Statistical Techniques: A Case Study of Qiandao Lake, China. Sustainability, 8(3), 243. doi:10.3390/su8030243

Heydari, S. S., & Mountrakis, G. (2018). Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. Remote Sensing of Environment, 204, 648–658. doi:10.1016/j.rse.2017.09.035

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., ... & Megown, K. (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.

Huang, H., Chen, Y., Clinton, N., Wang, J., Wang, X., Liu, C., … Zhu, Z. (2017). Mapping major land cover dynamics in Beijing using all Landsat images in Google Earth Engine. Remote Sensing of Environment, 202, 166–176. doi:10.1016/j.rse.2017.02.021

Huang, J., Huang, Y., Pontius, R. G., & Zhang, Z. (2015). Geographically weighted regression to measure spatial variations in correlations between water pollution versus land use in a coastal watershed. Ocean & Coastal Management, 103, 14–24. doi:10.1016/j.ocecoaman.2014.10.007

Huang, Z., Han, L., Zeng, L., Xiao, W., & Tian, Y. (2015). Effects of land use patterns on stream water quality: a case study of a small-scale watershed in the Three Gorges Reservoir Area, China. Environmental Science and Pollution Research, 23(4), 3943–3955. doi:10.1007/s11356-015-5874-8

Hwang, S.-A., Hwang, S.-J., Park, S.-R., & Lee, S.-W. (2016). Examining the Relationships between Watershed Urban Land Use and Stream Water Quality Using Linear and Generalized Additive Models. Water, 8(4), 155. doi:10.3390/w8040155

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J., & Xian, G. (2013). A comprehensive change detection method for updating the National Land Cover Database to circa 2011. Remote Sensing of Environment, 132, 159–175. doi:10.1016/j.rse.2013.01.012

Jin, S., Yang, L., Zhu, Z., & Homer, C. (2017). A land cover change detection and classification protocol for updating Alaska NLCD 2001 to 2011. Remote Sensing of Environment, 195, 44–55. doi:10.1016/j.rse.2017.04.021

Jordan, T. E., Weller, D. E., & Pelc, C. E. (2017). Effects of Local Watershed Land Use on Water Quality in Mid-Atlantic Coastal Bays and Subestuaries of the Chesapeake Bay. Estuaries and Coasts, 41(S1), 38–53. doi:10.1007/s12237-017-0303-5

Kalnay, E., & Cai, M. (2003). Impact of urbanization and land-use change on climate. Nature, 423(6939), 528–531. doi:10.1038/nature01675

Kändler, M., Blechinger, K., Seidler, C., Pavlů, V., Šanda, M., Dostál, T., … Štich, M. (2017). Impact of land use on water quality in the upper Nisa catchment in the Czech Republic and in Germany. Science of The Total Environment, 586, 1316–1325. doi:10.1016/j.scitotenv.2016.10.221

Kibena, J., Nhapi, I., & Gumindoga, W. (2014). Assessing the relationship between water quality parameters and changes in landuse patterns in the Upper Manyame River, Zimbabwe. Physics and Chemistry of the Earth, Parts A/B/C, 67-69, 153–163. doi:10.1016/j.pce.2013.09.017

Kim, D.-H., Sexton, J. O., Noojipady, P., Huang, C., Anand, A., Channan, S., … Townshend, J. R. (2014). Global, Landsat-based forest-cover change from 1990 to 2000. Remote Sensing of Environment, 155, 178–193. doi:10.1016/j.rse.2014.08.017

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, 82(13). doi:10.18637/jss.v082.i13

Lessels, J. S., & Bishop, T. F. A. (2013). Estimating water quality using linear mixed models with stream discharge and turbidity. Journal of Hydrology, 498, 13–22. doi:10.1016/j.jhydrol.2013.06.006

Li, Y., Li, Y., Qureshi, S., Kappas, M., & Hubacek, K. (2015). On the relationship between landscape ecological patterns and water quality across gradient zones of rapid urbanization in coastal China. Ecological Modelling, 318, 100–108. doi:10.1016/j.ecolmodel.2015.01.028

Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., … Western, A. W. (2017). Key factors influencing differences in stream water quality across space. Wiley Interdisciplinary Reviews: Water, 5(1), e1260. doi:10.1002/wat2.1260

Liu, C., Xiong, T., Gong, P., & Qi, S. (2017). Improving large-scale moso bamboo mapping based on dense Landsat time series and auxiliary data: a case study in Fujian Province, China. Remote Sensing Letters, 9(1), 1–10. doi:10.1080/2150704x.2017.1378454

Liu, C., Zhang, Q., Luo, H., Qi, S., Tao, S., Xu, H., & Yao, Y. (2019). An efficient approach to capture continuous impervious surface dynamics using spatial-temporal rules and dense Landsat time series stacks. Remote Sensing of Environment, 229, 114–132. doi:10.1016/j.rse.2019.04.025

Liu, J., Zhang, X., Wu, B., Pan, G., Xu, J., & Wu, S. (2017). Spatial scale and seasonal dependence of land use impacts on riverine water quality in the Huai River basin, China. Environmental Science and Pollution Research, 24(26), 20995–21010. doi:10.1007/s11356-017-9733-7

Liu, J., Shen, Z., & Chen, L. (2018). Assessing how spatial variations of land use pattern affect water quality across a typical urbanized watershed in Beijing, China. Landscape and Urban Planning, 176, 51–63. doi:10.1016/j.landurbplan.2018.04.006

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing, 28(5), 823–870. doi:10.1080/01431160600746456

Luo, K., Hu, X., He, Q., Wu, Z., Cheng, H., Hu, Z., & Mazumder, A. (2017). Using multivariate techniques to assess the effects of urbanization on surface water quality: a case study in the Liangjiang New Area, China. Environmental Monitoring and Assessment, 189(4). doi:10.1007/s10661-017-5884-8

Lv, H., Xu, Y., Han, L., & Zhou, F. (2014). Scale-dependence effects of landscape on seasonal water quality in Xitiaoxi catchment of Taihu Basin, China. Water Science and Technology, 71(1), 59–66. doi:10.2166/wst.2014.463

Manandhar, R., Odeh, I., & Ancev, T. (2009). Improving the Accuracy of Land Use and Land Cover Classification of Landsat Data Using Post-Classification Enhancement. Remote Sensing, 1(3), 330–344. doi:10.3390/rs1030330

Manfrin, A., Bombi, P., Traversetti, L., Larsen, S., & Scalici, M. (2016). A landscape-based predictive approach for running water quality assessment: A Mediterranean case study. Journal for Nature Conservation, 30, 27–31. doi:10.1016/j.jnc.2016.01.002

Mello, K. de, Valente, R. A., Randhir, T. O., dos Santos, A. C. A., & Vettorazzi, C. A. (2018). Effects of land use and land cover on water quality of low-order streams in Southeastern Brazil: Watershed versus riparian zone. CATENA, 167, 130–138. doi:10.1016/j.catena.2018.04.027

Mellor, A., Boukir, S., Haywood, A., & Jones, S. (2015). Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. ISPRS Journal of Photogrammetry and Remote Sensing, 105, 155–168. doi:10.1016/j.isprsjprs.2015.03.014

Meneses, B. M., Reis, R., Vale, M. J., & Saraiva, R. (2015). Land use and land cover changes in Zêzere watershed (Portugal) — Water quality implications. Science of The Total Environment, 527-528, 439–447. doi:10.1016/j.scitotenv.2015.04.092

Millard, K., & Richardson, M. (2015). On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. Remote Sensing, 7(7), 8489–8515. doi:10.3390/rs70708489

Molenberghs, G., & Verbeke, G. (2000). Linear Mixed Models for Longitudinal Data. Springer Series in Statistics. doi:10.1007/978-1-4419-0300-6

Moore, I. D., Gessler, P. E., Nielsen, G. A., & Peterson, G. A. (1993). Soil Attribute Prediction Using Terrain Analysis. Soil Science Society of America Journal, 57(2), NP. doi:10.2136/sssaj1993.572npb

Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., De Palma, A., Ferrier, S., … Purvis, A. (2016). Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. Science, 353(6296), 288–291. doi:10.1126/science.aaf2201

Nielsen, A., Trolle, D., Søndergaard, M., Lauridsen, T. L., Bjerring, R., Olesen, J. E., & Jeppesen, E. (2012). Watershed land use effects on lake water quality in Denmark. Ecological Applications, 22(4), 1187–1200. doi:10.1890/11-1831.1

Oeding, S., Taffs, K. H., Cox, B., Reichelt-Brushett, A., & Sullivan, C. (2018). The influence of land use in a highly modified catchment: Investigating the importance of scale in riverine health assessment. Journal of Environmental Management, 206, 1007–1019. doi:10.1016/j.jenvman.2017.12.005

Panagos, P., Borrelli, P., & Meusburger, K. (2015). A New European Slope Length and Steepness Factor (LS-Factor) for Modeling Soil Erosion by Water. Geosciences, 5(2), 117–126. doi:10.3390/geosciences5020117

Patel, N. N., Angiuli, E., Gamba, P., Gaughan, A., Lisini, G., Stevens, F. R., … Trianni, G. (2015). Multitemporal settlement and population mapping from Landsat using Google Earth

Engine. International Journal of Applied Earth Observation and Geoinformation, 35, 199–208. doi:10.1016/j.jag.2014.09.005

Pratt, B., & Chang, H. (2012). Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. Journal of Hazardous Materials, 209-210, 48–58. doi:10.1016/j.jhazmat.2011.12.068

Riley, S. J., DeGloria, S. D., & Elliot, R. (1999). Index that quantifies topographic heterogeneity. intermountain Journal of sciences, 5(1-4), 23-27.

Rodrigues, V., Estrany, J., Ranzini, M., de Cicco, V., Martín-Benito, J. M. T., Hedo, J., & Lucas-Borja, M. E. (2018). Effects of land use and seasonality on stream water quality in a small tropical catchment: The headwater of Córrego Água Limpa, São Paulo (Brazil). Science of The Total Environment, 622-623, 1553–1561. doi:10.1016/j.scitotenv.2017.10.028

Rodriguez-Galiano, V. F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P. M., & Jeganathan, C. (2012). Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sensing of Environment, 121, 93–107. doi:10.1016/j.rse.2011.12.003

Rothwell, J. J., Dise, N. B., Taylor, K. G., Allott, T. E. H., Scholefield, P., Davies, H., & Neal, C. (2010). Predicting river water quality across North West England using catchment characteristics. Journal of Hydrology, 395(3-4), 153–162. doi:10.1016/j.jhydrol.2010.10.015

Sajikumar, N., & Remya, R. S. (2015). Impact of land cover and land use change on runoff characteristics. Journal of Environmental Management, 161, 460–468. doi:10.1016/j.jenvman.2014.12.041

Sakai, Y., Ishizuka, S., & Takenaka, C. (2013). Predicting deadwood densities of Cryptomeria japonica and Chamaecyparis obtusa forests using a generalized linear mixed model with a national-scale dataset. Forest Ecology and Management, 295, 228–238. doi:10.1016/j.foreco.2013.01.030

Sangani, M. H., Amiri, B. J., Shabani, A. A., Sakieh, Y., & Ashrafi, S. (2015). Modeling relationships between catchment attributes and river water quality in southern catchments of the Caspian Sea. Environmental Science and Pollution Research, 22(7), 4985-5002. doi:10.1007/s11356-014-3727-5

Santhi, C., Srinivasan, R., Arnold, J. G., & Williams, J. R. (2006). A modeling approach to evaluate the impacts of water quality management plans implemented in a watershed in Texas. Environmental Modelling & Software, 21(8), 1141–1157. doi:10.1016/j.envsoft.2005.05.013

Schneider, A., Friedl, M. A., & Potere, D. (2010). Mapping global urban areas using MODIS 500-m data: New methods and datasets based on “urban ecoregions.” Remote Sensing of Environment, 114(8), 1733–1746. doi:10.1016/j.rse.2010.03.003

Seeboonruang, U. (2012). A statistical assessment of the impact of land uses on surface water quality indexes. Journal of Environmental Management, 101, 134–142. doi:10.1016/j.jenvman.2011.10.019

Sheldon, F., Peterson, E. E., Boone, E. L., Sippel, S., Bunn, S. E., & Harch, B. D. (2012). Identifying the spatial scale of land use that most strongly influences overall river ecosystem health score. Ecological Applications, 22(8), 2188–2203. doi:10.1890/11-1792.1

Shi, P., Zhang, Y., Li, Z., Li, P., & Xu, G. (2017). Influence of land use and land cover patterns on seasonal water quality at multi-spatial scales. CATENA, 151, 182–190. doi:10.1016/j.catena.2016.12.017

Shi, W., Xia, J., & Zhang, X. (2016). Influences of anthropogenic activities and topography on water quality in the highly regulated Huai River basin, China. Environmental Science and Pollution Research, 23(21), 21460–21474. doi:10.1007/s11356-016-7368-8

Shi, Z. H., Ai, L., Li, X., Huang, X. D., Wu, G. L., & Liao, W. (2013). Partial least-squares regression for linking land-cover patterns to soil erosion and sediment yield in watersheds. Journal of Hydrology, 498, 165–176. doi:10.1016/j.jhydrol.2013.06.031

Sluiter, R., & Pebesma, E. J. (2010). Comparing techniques for vegetation classification using multi- and hyperspectral images and ancillary environmental data. International Journal of Remote Sensing, 31(23), 6143–6161. doi:10.1080/01431160903401379

Sun, R., Chen, L., Chen, W., & Ji, Y. (2011). Effect of Land-Use Patterns on Total Nitrogen Concentration in the Upstream Regions of the Haihe River Basin, China. Environmental Management, 51(1), 45–58. doi:10.1007/s00267-011-9764-7

Sun, Y., Guo, Q., Liu, J., & Wang, R. (2014). Scale Effects on Spatially Varying Relationships Between Urban Landscape Patterns and Water Quality. Environmental Management, 54(2), 272–287. doi:10.1007/s00267-014-0287-x

Texas Commission on Environmental Quality. (2014). Managing nonpoint source pollution in Texas, 2013 annual report

Texas Parks and Wildlife Department. 2012. Texas Conservation Action Plan 2012 – 2016: Gulf Coast Prairies and Marshes Handbook. Editor, Wendy Connally, Texas Conservation Action Plan Coordinator. Austin, Texas

Thakkar, A. K., Desai, V. R., Patel, A., & Potdar, M. B. (2017). Post-classification corrections in improving the classification of Land Use/Land Cover of arid region using RS and GIS: The case of Arjuni watershed, Gujarat, India. The Egyptian Journal of Remote Sensing and Space Science, 20(1), 79–89. doi:10.1016/j.ejrs.2016.11.006

Tran, C. P., Bode, R. W., Smith, A. J., & Kleppel, G. S. (2010). Land-use proximity as a basis for assessing stream water quality in New York State (USA). Ecological Indicators, 10(3), 727–733. doi:10.1016/j.ecolind.2009.12.002

Tu, J. (2011). Spatially varying relationships between land use and water quality across an urbanization gradient explored by geographically weighted regression. Applied Geography, 31(1), 376–392. doi:10.1016/j.apgeog.2010.08.001

Tu, J. (2013). Spatial Variations in the Relationships between Land Use and Water Quality across an Urbanization Gradient in the Watersheds of Northern Georgia, USA. Environmental Management, 51(1), 1–17. doi:10.1007/s00267-011-9738-9

Uriarte, M., Yackulic, C. B., Lim, Y., & Arce-Nazario, J. A. (2011). Influence of land use on water quality in a tropical landscape: a multi-scale analysis. Landscape Ecology, 26(8), 1151–1164. doi:10.1007/s10980-011-9642-y

Uwimana, A., van Dam, A., Gettel, G., Bigirimana, B., & Irvine, K. (2017). Effects of River Discharge and Land Use and Land Cover (LULC) on Water Quality Dynamics in Migina Catchment, Rwanda. Environmental Management, 60(3), 496–512. doi:10.1007/s00267-017-0891-7

Varanka, S., & Luoto, M. (2011). ENVIRONMENTAL DETERMINANTS OF WATER QUALITY IN BOREAL RIVERS BASED ON PARTITIONING METHODS. River Research and Applications, 28(7), 1034–1046. doi:10.1002/rra.1502

Varanka, S., Hjort, J., & Luoto, M. (2014). Geomorphological factors predict water quality in boreal rivers. Earth Surface Processes and Landforms, 40(15), 1989–1999. doi:10.1002/esp.3601

Vogelmann, J. E., Howard, S. M., Yang, L., Larson, C. R., Wylie, B. K., & Van Driel, N. (2001). Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. Photogrammetric Engineering and Remote Sensing, 67(6).

Vrebos, D., Beauchard, O., & Meire, P. (2017). The impact of land use and spatial mediated processes on the water quality in a river system. Science of The Total Environment, 601-602, 365–373. doi:10.1016/j.scitotenv.2017.05.217

Walsh, C. J., & Webb, J. A. (2014). Spatial weighting of land use and temporal weighting of antecedent discharge improves prediction of stream condition. Landscape Ecology, 29(7), 1171–1185. doi:10.1007/s10980-014-0050-y

Wan, R., Cai, S., Li, H., Yang, G., Li, Z., & Nie, X. (2014). Inferring land use and land cover impact on stream water quality using a Bayesian hierarchical modeling approach in the Xitiaoxi River Watershed, China. Journal of Environmental Management, 133, 1–11. doi:10.1016/j.jenvman.2013.11.035

Wang, G., A, Y., Xu, Z., & Zhang, S. (2014). The influence of land use patterns on water quality at multiple spatial scales in a river system. Hydrological Processes, 28(20), 5259–5272. doi:10.1002/hyp.10017

Wang, R., Zhang, X., & Li, M.-H. (2019). Predicting bioretention pollutant removal efficiency with design features: A data-driven approach. Journal of Environmental Management, 242, 403–414. doi:10.1016/j.jenvman.2019.04.064

Wickham, J., Stehman, S. V., Gass, L., Dewitz, J. A., Sorenson, D. G., Granneman, B. J., … Baer, L. A. (2017). Thematic accuracy assessment of the 2011 National Land Cover Database (NLCD). Remote Sensing of Environment, 191, 328–341. doi:10.1016/j.rse.2016.12.026

Wijesiri, B., Deilami, K., & Goonetilleke, A. (2018). Evaluating the relationship between temporal changes in land use and resulting water quality. Environmental Pollution, 234, 480–486. doi:10.1016/j.envpol.2017.11.096

Wilson, C. O. (2015). Land use/land cover water quality nexus: quantifying anthropogenic influences on surface water quality. Environmental Monitoring and Assessment, 187(7). doi:10.1007/s10661-015-4666-4

World Population Review (2019). Texas Population 2019. Retrieved from http://worldpopulationreview.com/states/texas-population/

World Wildlife Fund (2019). Piney Woods Forest. Retrieved from https://www.worldwildlife.org/ecoregions/na0523

Xian, G., & Homer, C. (2010). Updating the 2001 National Land Cover Database Impervious Surface Products to 2006 using Landsat Imagery Change Detection Methods. Remote Sensing of Environment, 114(8), 1676–1686. doi:10.1016/j.rse.2010.02.018

Yang, C., Wu, G., Ding, K., Shi, T., Li, Q., & Wang, J. (2017). Improving Land Use/Land Cover Classification by Integrating Pixel Unmixing and Decision Tree Methods. Remote Sensing, 9(12), 1222. doi:10.3390/rs9121222

Ye, L., Cai, Q., Liu, R., & Cao, M. (2008). The influence of topography and land use on water quality of Xiangxi River in Three Gorges Reservoir region. Environmental Geology, 58(5), 937–942. doi:10.1007/s00254-008-1573-9

Ye, L., Cai, Q., Liu, R., & Cao, M. (2008). The influence of topography and land use on water quality of Xiangxi River in Three Gorges Reservoir region. Environmental Geology, 58(5), 937–942. doi:10.1007/s00254-008-1573-9

Yu, W., Zhou, W., Qian, Y., & Yan, J. (2016). A new approach for land cover classification and change analysis: Integrating backdating and an object-based method. Remote Sensing of Environment, 177, 37–47. doi:10.1016/j.rse.2016.02.030

Zhang, F., Wang, J., & Wang, X. (2018). Recognizing the Relationship between Spatial Patterns in Water Quality and Land-Use/Cover Types: A Case Study of the Jinghe Oasis in Xinjiang, China. Water, 10(5), 646. doi:10.3390/w10050646

Zhang, H. K., & Roy, D. P. (2017). Using the 500 m MODIS land cover product to derive a consistent continental scale 30 m Landsat land cover classification. Remote Sensing of Environment, 197, 15–34. doi:10.1016/j.rse.2017.05.024

Zhang, W., Li, H., Sun, D., & Zhou, L. (2012). A Statistical Assessment of the Impact of Agricultural Land Use Intensity on Regional Surface Water Quality at Multiple Scales. International Journal of Environmental Research and Public Health, 9(11), 4170–4186. doi:10.3390/ijerph9114170

Zhao, G., Gao, H., & Cuo, L. (2016). Effects of Urbanization and Climate Change on Peak Flows over the San Antonio River Basin, Texas. Journal of Hydrometeorology, 17(9), 2371–2389. doi:10.1175/jhm-d-15-0216.1

Zhao, G., & Gao, H. (2019). Estimating reservoir evaporation losses for the United States: Fusing remote sensing and modeling approaches. Remote Sensing of Environment, 226, 109–124. doi:10.1016/j.rse.2019.03.015

Zhao, J., Lin, L., Yang, K., Liu, Q., & Qian, G. (2015). Influences of land use on water quality in a reticular river network area: A case study in Shanghai, China. Landscape and Urban Planning, 137, 20–29. doi:10.1016/j.landurbplan.2014.12.010

Zhou, P., Huang, J., Pontius, R. G., & Hong, H. (2016). New insight into the correlations between land use and water quality in a coastal watershed of China: Does point source pollution weaken it? Science of The Total Environment, 543, 591–600. doi:10.1016/j.scitotenv.2015.11.063

Zhu, Z., Gallant, A. L., Woodcock, C. E., Pengra, B., Olofsson, P., Loveland, T. R., … Auch, R. F. (2016). Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative. ISPRS Journal of Photogrammetry and Remote Sensing, 122, 206–221. doi:10.1016/j.isprsjprs.2016.11.004

Zou, G., Li, Y., Huang, T., Liu, D. L., Herridge, D., & Wu, J. (2017). A Mixed-Effects Regression Modeling Approach for Evaluating Paddy Soil Productivity. Agronomy Journal, 109(5), 2302. doi:10.2134/agronj2017.02.0089

CHAPTER 4 EVALUATING THE EFFECTIVENESS OF WATERSHED PRESERVATION BASED ON THE HYDOLOGICALLY SENSITIVE AREA (HSA) SITING APPROACH—A DEMONSTRATION OF DATA-DRIVEN ECOLOGICAL PLANNING METHOD

4.1 Introduction

Landscape planning and design are decision making processes that integrate multiple domains of knowledge, including ecology, hydrology, geology, economics, history and so on (Steiner, 2011; Xiang, 2014; Wang et al., 2016). Since the 1960s and 1970s, the concept of ecological planning had brought much recognition to planners and designers. Among the most acknowledged of those planners and designers was Ian McHarg, who carried out pioneer planning projects using ecological frameworks. According to McHarg, ecological planning and design should be "an intrinsically suitable location" and included "processes with appropriate materials and forms" (McHarg, 2006, p. 123).

McHarg viewed nature as a value system by evaluating all the ecological, economic, and cultural factors as interdependent components that, together, formed a holistic social-ecological system (McHarg, 1969, p.104; Yang and Li, 2016). This fundamental theory led to the corresponding "layer-cake" model as the core method for realizing ecological planning. In the "layer-cake" model, conservation areas are delineated in terms of those that are not suitable for development according to a suitability analysis for each layer (McHarg, 1969, p.114). For example, in the early development of The Woodlands project, inventory maps including physiography, geology, soils, hydrology, vegetation, climate and resources were overlaid to determine suitability maps for proposed land uses (McHarg and Steiner, 1998; Yang et al., 2015; Yang, 2018). The "layer-cake" model has had far-reaching influence and has been widely applied in a number of

ecological planning projects over the years (Espejel et al., 1999; Sustainable Sites Initiative, 2009; Calkins, 2012).

The key step in McHarg's "layer-cake" model is the identification of critical areas that, intrinsically, have high ecological values and should thus be protected from development (Steiner et al., 2000a; Herrington, 2010). Many research efforts have aimed to expand the framework of ecological planning to become "broader" with additional layers or sublayers. An important question to consider in such work is whether each layer is "deep" enough to form a more efficient plan. We argue that one limitation of the "layer-cake" model is that the suitability analysis in each layer is a linear combination of multiple indicators. The ranking of ecological values was somewhat arbitrary due to the accuracy of environmental data and the linear overlay method (Yang et al., 2015). In fact, the non-linear behavior of ecosystems can hardly be approximated by the linear overlay approach. As such, the introduction of nonlinear interdisciplinary models in order to make each layer more physically sound has yielded promising results. For example, it is possible for the soil erosion layer to be generated by the linear overlay of hydrology, soil and topography maps (Dosskey et al., 2005). However, this approach was shown to be less accurate and efficient in comparison to the Revised Universal Soil Loss Equation (RUSLE) for mapping soil erosion (Schumacher et al., 2005). Soil erosion maps generated by RUSLE with logistic regression calibrations were tested and found to be more robust (Mueller et al., 2005), which aided in the creation of a more effective soil layer in ecological planning.

In this study, the hydrology layer in ecological planning was investigated and the hydrologically sensitive area (HSA) approach to map runoff and contaminant source areas was introduced. The hydrology layer is one of the most important components in ecological planning, as it links land use, soil, topography, and aquatic organisms to form an interactive natural process. The HSAs are

delineated according to variable source area (VSA) hydrology. VSAs are the runoff-generating areas in a watershed; they are small, variable, and predictable depending on season, climate, topography and land cover factors (Frankenberger et al., 1999; Qiu, 2003). HSAs are parts of VSAs more prone to generating runoff and are therefore susceptible to contaminant transportation (Walter et al, 2000). The spatial patterns of HSAs and their impacts on discharge and pollutant generation have been well demonstrated (Qiu, 2009; Qiu et al, 2013). In ecological planning, HSAs are the preferential locations to place best management practice (BMP) or low impact development (LID) facilities (Martin-Mikle et al., 2015).

Another limitation of the traditional ecological planning approach is that planning efficiency was often conceptually and intuitively proved, without further validation from real data. One way researchers have tried to address this issue is by evaluating the performance of ecological planning with hypothetical scenario analysis (Yang and Li, 2011, Fu et al., 2016). However, given the current, dramatically increased availability of data and advancements in hardware and software engineering, little work has yet been done to leverage these resources in ecological planning. To take advantage of various sources of publicly available environmental data, statistical analysis has been shown to be a more straightforward way to investigate the impact of landscape features on hydrology and water quality, compared to complex hydrological models (Giri and Qiu, 2016; Lintern et al., 2018). To illustrate how data-driven methods work in ecological planning, statistical verification and scenario evaluation were applied to prove the effectiveness of the HSA approach in this study.

This study utilized an interdisciplinary approach to calculate and validate the HSAs, as demonstrated in the Middle Brazos-Bosque basin in the Texas gulf region. The three objectives were: (1) to generate the HSA map in the Middle Brazos-Bosque basin. On the HSA map, areas

with high hydrological sensitivity were suggested to be prioritized as conservation areas; (2) to calculate the mean hydrological sensitivity of each subbasin in the Middle Brazos-Bosque basin and investigate if the mean hydrological sensitivity was correlated with NO_3^--N concentrations measured at the subbasin outlet; and (3) to simulate NO_3^--N outputs in scenarios where some HSAs were transformed from croplands to green infrastructures for best management practices. A threshold was suggested to delineate HSAs, which led to the most efficient scenario regarding NO_3^--N loading reduction.

4.2 Data and Method

4.2.1 Study Site

The Middle Brazos-Bosque basin (Figure 4-1) is one of the 378 hydrologic accounting units with an HUC6 number of 120602. The Brazos River is the eleventh longest river in the United States, with a total drainage area of 116,000 km^2. The main water quality concerns in the Brazos Watershed include high nutrient loadings, high bacterial, and low dissolved oxygen. The area of the Middle Brazos-Bosque basin is 19,140 km^2. It is a mixed-use watershed with the upper drainage area primarily occupied by forest and grassland. The lower drainage area is covered by planted and urban areas. A part of the city of Waco is located downstream of the Middle Brazos-Bosque basin.

According to climate data from the Waco Regional Airport Station, in the latest three decades, the annual average temperature is 19.3 °C and the annual total precipitation is 88.1 cm. The mean slope of the basin is around 21°. Hydrologic soil groups C and D are the primary soil categories in the basin, which have lower infiltration rates and higher runoff potentials. Located within the basin boundary are 89 Texas Commission on Environmental Quality (TCEQ) monitoring stations and 13 USGS monitoring stations. Because the large area of the Middle Brazos-Bosque basin would

add difficulties to the hydrological simulation process, the McGregor subbasin, with the area of 22.4 km^2, was selected as the HSA scenario analysis site. The McGregor subbasin was delineated with USGS Station 08095300 as the subbasin outlet, which is close to the city of McGregor. The primary land covers of the McGregor subbasin are herbaceous and planted.

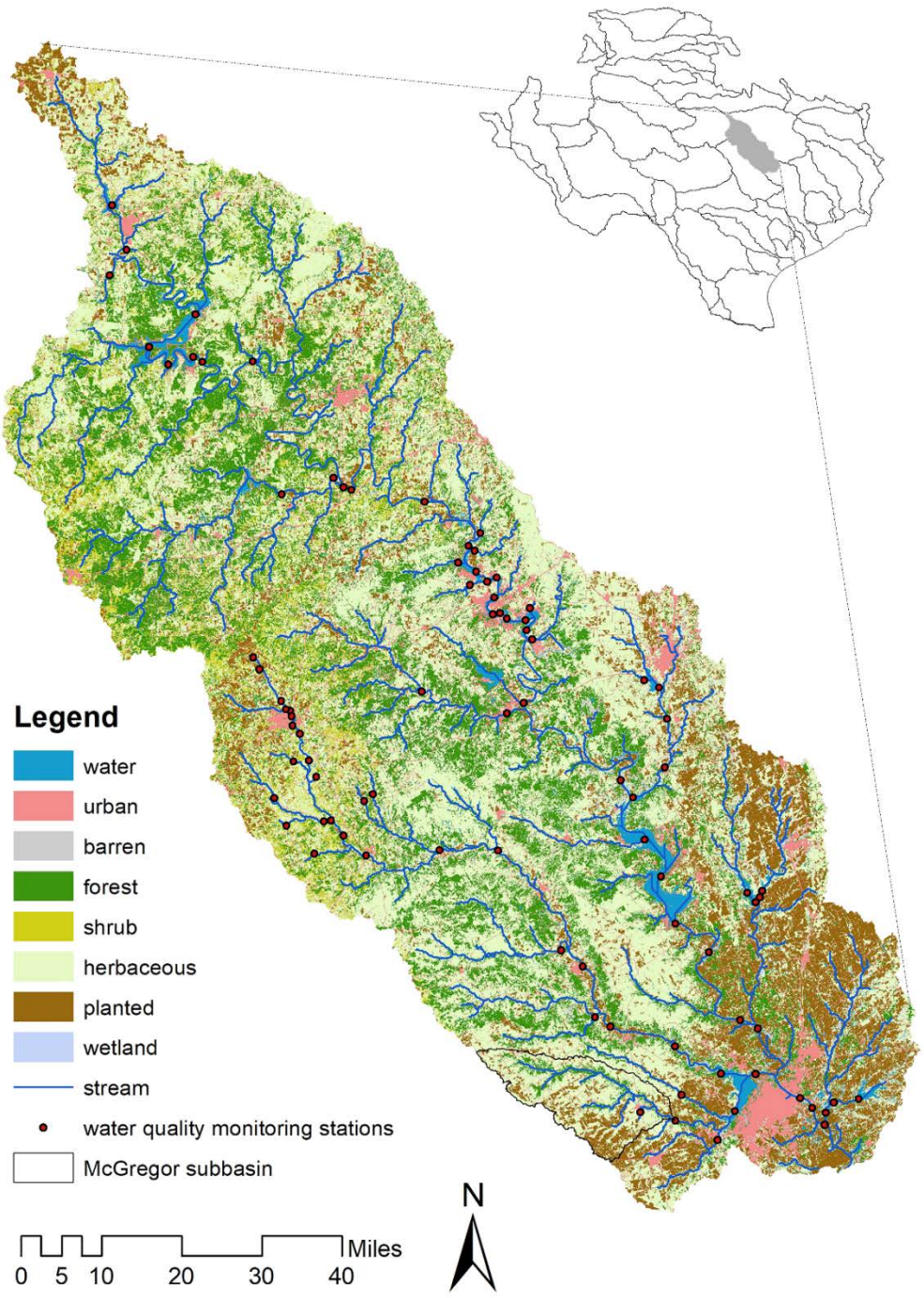


Figure 4-1. Study Site (The Middle Brazos-Bosque basin)

4.2.2 Data Acquisition

HSA mapping involved the data layers of topography, hydrology and soil. The USGS National Elevation Dataset with a spatial resolution of 30m was used to derive elevation, slope and flow accumulation data. Soil data were drawn from the Soil Survey Geographic (SSURGO) soil database. Soil conductivity and soil depth to the restrictive layer were the two parameters of interest, calculated from “component,” “corestriction” and “chorizon” tables from the SSURGO database.

Water quality data of multiple subbasins were required to perform statistical verification of the HSA approach. The locations of water quality monitoring stations were drawn from the Texas Commission on Environmental Quality (TCEQ), and subbasin boundaries were delineated accordingly. The corresponding water quality data in 2011 were obtained from the Texas Clean Rivers Program (CRP) data tool. Specifically, NO_3^--N concentration data in the wet seasons were aggregated yearly and joint with other attributes of the subbasins.

Land cover, topography and weather data were prepared for the HSA scenarios simulations. Land cover data were extracted from the 2011 National Land Cover Database (NLCD). NLCD has 16 classes of land cover at the spatial resolution of 30m. The Daymet Version 3 dataset, with a spatial resolution of 1000m, was used to aggregate daily mean temperature and daily total precipitation across the study site. Due to missing data on continuously measured pollutants, discharge data were used for model calibration as a compromising approach to simulate contaminant outputs. All the data were prepared with Google Earth Engine (GEE) and ArcMap 10.5.

4.2.3 HSA Calculation and Mapping

The gridded hydrological sensitivity maps were generated based on the TOPography based hydrological MODEL (TOPMODEL). In the TOPMODEL, the resulting topographic index from Equation 4-1 was used to model patterns of surface runoff. The larger the topographic index value, the more likely the grid is to be saturated during a rainfall event. It is therefore reasonable to keep grids with high topographic index values as conservation areas in ecological planning (Qiu, 2009).

$$\lambda = \ln(\partial / \tan\beta) - \ln(K_s D)$$

Equation. 4-1

The first part on the right side of the equation is the wetness index and the second part accounts for the soil water storage capacity (Beven and Kirkby, 1979; Walter et al., 2002). In the equation, α represents the upslope contributing area per unit contour length in meters, which is approximated by the flow accumulation value. β is the surface slope angle in decimal degrees. The term $K_s D$ is the water storage component, where K_s is the mean saturated hydraulic conductivity of the soil profile in meters per day and D is the soil depth to the restrictive layer in centimeters. The shallower the soil profile above the restrictive layers and the lower the saturated hydraulic conductivity, the higher the likelihood of runoff generation.

If there are several topsoil layers above the restrictive layer with different K_s, a compound K_s will be defined via Equation 4-2. In Equation 4-2, d is the total depth of soil above the restrictive layer, d_i is the depth of layer *i* and k_i is the corresponding saturated hydraulic conductivity of layer *i*. There are a small numbers of grids where K_s values are missing in the SSURGO database. Most

of the grids are water bodies, where green infrastructures are not suitable to build. Therefore, they are left as “no data” on the HSA map.

$$K_s = d / \sum_{1}^{n} (d_i / k_i)$$

Equation. 4-2

4.2.4 Statistical Analysis

Statistical analysis was carried out to determine the relationships between hydrological sensitivity and NO_3^--N concentration in streams. If higher hydrological sensitivity was associated with higher nutrient loadings, the effectiveness of prioritizing HSAs as conservation areas in this basin could be supported. The units of analysis were the 37 subbasins with measured NO_3^--N concentration data in the 2011 wet season. The wet season was defined as the time range from June to October. Dependent variables were the yearly averages of NO_3^--N concentrations measured at each subbasin outlet. Independent variables were the mean hydrological sensitivity of the subbasin.

The Pearson correlation analysis was performed to study the relationships between mean hydrological sensitivity and NO_3^--N concentrations. A null hypothesis was also tested to determine if any association existed between them, with a significance level of 0.05. The scatter plots of natural logarithm of NO_3^--N concentrations and mean hydrological sensitivity suggested that there might be a non-linear relationship between them. Therefore, a non-linear least squares (NLS) model with a quadratic term was fit to predict NO_3^--N concentrations with mean hydrological sensitivity.

4.2.5 SWAT modelling

The Soil and Water Assessment Tool (SWAT) was used to simulate multiple HSA scenarios. This model was selected because the hydrological response units (HRU) in SWAT integrate the components of land cover, soil and topography, which agreed conceptually with the TOPMODEL. The baseline scenario was the current land cover status of the McGregor subbasin. Two alternative scenarios were developed where HSAs were defined with 5% and 2% grids with the highest hydrological sensitivity values. The HSAs on the croplands were hypothesized to remain as forests. Discharge and NO_3^--N outputs were simulated monthly from 2008 to 2011, following a two-year warm-up period from 2006 to 2007.

Because continuously measured NO_3^--N output data was not available, only discharge was calibrated in the 2008 to 2009 period. The validation period was from 2010 to 2011. The calibrated parameters were CN value, soil evaporation compensation factor and soil available water capacity. The SWAT model efficiency was evaluated by Nash-Sutcliffe model efficiency coefficient (NSE) and R^2. The missing NO_3^--N output validation was a major drawback in the scenario simulation. The simulated NO_3^--N output in the scenario analysis was therefore only an approximation of the performance data.

4.3 Result

4.3.1 HSA Map

The distribution of the topographic index values in the Middle Brazos-Bosque basin was a right-skewed bell curve, with a mean value of 5.3, and a standard deviation of 2.8. The maximum topographic index value was 26.2. The values inside one standard deviation were from 3 to 5.2. The topographic index values in this basin had a similar range but a smaller mean than those

reported in previous studies (Qiu, 2009; Martin-Mikle et al., 2015). The reason might be that some water bodies with high hydrological sensitivities had no hydraulic conductivity data available, and we excluded them in the topographic index calculation.

Presented in Figure 4-2 is the hydrological sensitivity map of the McGregor subbasin, represented by the topographic index values. It is important to note that some HSAs with high topographic index values are located in the middle of subbasin's fields, rather than along its streams. This indicates that only protecting stream buffer areas is not sufficient for ecological planning. Grids with 5% highest topographic index values were mapped as HSAs, of which the values were larger than 11.9. The critical source areas (CSA) for nutrient generations were mapped as the HSAs on the planted area. Most of the CSAs were located in the downstream areas in the McGregor subbasin. Such CSAs were the prioritized sites to place BMP facilities.

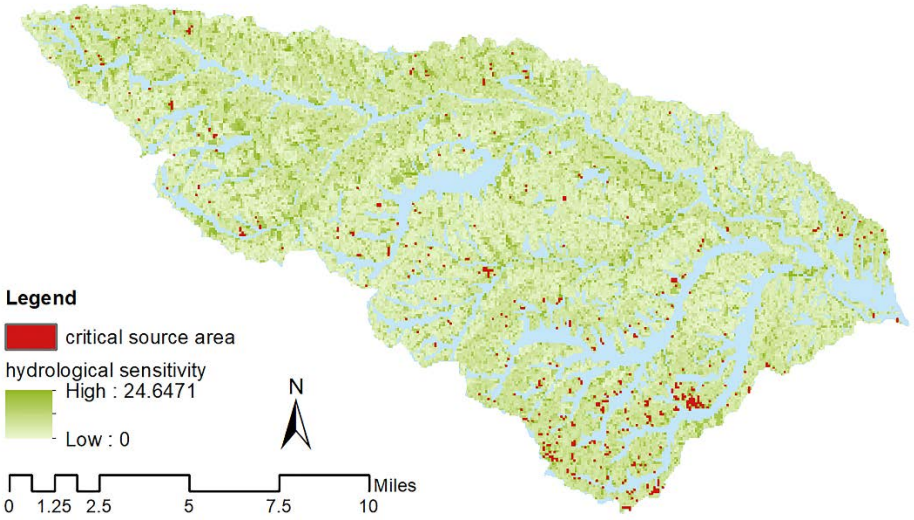


Figure 4-2. Hydrological sensitivity map and the critical source areas in the McGregor subbasin

4.3.2 The Relationships between Hydrologically Sensitivity and Water Quality

Pearson correlation results show a strong association between the mean hydrological sensitivity of the basin and the corresponding natural logarithm of NO_3^--N concentrations. The correlation between the basin's mean hydrological sensitivity and the natural logarithm of NO_3^--N concentrations was 0.4, with a p value of 0.014. The scatter plot in Figure 4-3 also indicates a positive association between hydrological sensitivity and NO_3^--N concentrations via a probable non-linear relationship. A quadratic non-linear curve fit with the NLS model is also presented in Figure 4-3. The quadratic form was significant at the 0.01 level. The results indicate that subbasins with higher hydrological sensitivity tended to have higher NO_3^--N pollutant concentrations. With increased hydrological sensitivity, its impact on NO_3^--N concentrations became stronger.

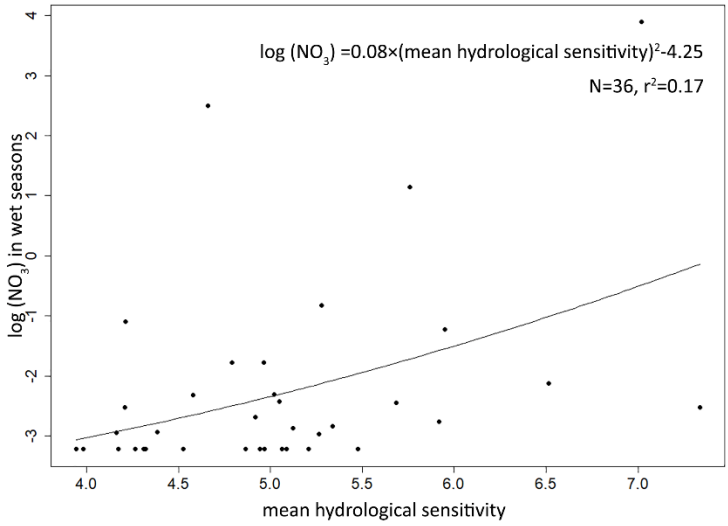


Figure 4-3. The relationship between mean hydrological sensitivity and log (NO_3^--N) in wet seasons

4.3.3 Scenario Simulation

The NO_3^--N output during the 2008 to 2011 period was approximated in SWAT under multiple HSA scenarios. In the calibration period, the R^2 and NSE of discharge were 0.93 and 0.75, respectively. In the validation period, the R^2 and NSE of discharge were 0.81 and 0.55, respectively. Table 4-1 demonstrates that scenario 2 was more efficient than scenario 1 in treating NO_3^--N pollution. In scenario 2, areas with the highest 2% hydrological sensitivity on the cropland were transformed into green space. Compared to the baseline scenario, 1.3% of croplands were transformed into green space, which only accounted for 0.25% of the total basin area and 1.3% of the total cropland area. However, 3.7% of nitrate outputs were reduced, which was disproportionately larger than the land use change.

In scenario 1, the percentages of transformed croplands and the reduction of NO_3^--N outputs were about the same; thus the efficiency of the HSA approach was not as high as that of scenario 2. The SWAT simulation results indicated that keeping areas with 2% highest hydrological sensitivity values as green infrastructure would be very efficient for NO_3^--N reduction. Increasing the percentage to 5% did not make a huge difference in further reducing NO_3^--N loadings.

Table 4-1. SWAT simulation results of NO_3^--N output in the period from 2008 to 2011

	baseline scenario	scenario 1	scenario 2
scenario criteria	land use of the current situation	If the grids are among the highest *5%* hydrological sensitivity with cropland land use, they are transformed into green space	If the grids are among the highest *2%* hydrological sensitivity with cropland land use, they are transformed into green space
NO_3^--N output (kg)	84662	80471	81514

	Table 4-1 (cont'd)	
the decreased NO_3^--N output compared to the baseline scenario (kg)	4191	3148
the percentage decrease of NO_3^--N output compared to the baseline scenario	5%	3.7%
the percentages of *total cropland area* that is transformed into green space compared to the baseline scenario	5.6%	1.3%
the percentages of *total basin area* that is transformed into green space compared to the baseline scenario	1%	0.25%

4.4 Discussion

4.4.1 Water Quality Management Implication

The HSA approach can be linked to land use controls, which protects scarce natural resources and mitigates the negative impacts of urbanization. Common land use controls protect water resources in steep slope areas, stream corridor areas, open space, farmland and wetlands. It was indicated that these types of land use controls could protect only around 50% of HSAs, most of which were protected by wetland conservation (Qiu et al., 2014). Based on the findings, some HSAs were located in the middle of upland fields and not along the stream corridors. These HSAs might not be effectively protected by existing land use control policies. Therefore, HSAs should be taken into consideration in land use control frameworks with additional protecting criteria.

The HSA approach can also provide a mechanistic and spatially explicit method for prioritizing LID sites. This approach ensures that LID facilities are more cost-effectively placed. In this study

site, HSAs were located dispersedly, with some patches in the stream source areas. The mapping results of HSAs were coincident with the principle of LID, which is to manage runoff at the source using a decentralized approach of controls. In addition, the scenario analysis proved the efficiency of placing LID in areas with the highest 2% hydrological sensitivity. If the measured water quality data were available, NO_3^--N outputs could be calibrated to indicate a more accurate threshold of HSA delineation, which could lead to the most effective solution regarding the removal of nutrients.

4.4.2 The Interdisciplinary Ecological Planning Approach

Typical ecological planning procedures involve planning goal initialization, inventory analysis, suitability analysis, and land use analysis (Yang and Li, 2016). In this study, the verification of ecological planning with statistical and scenario analyses was emphasized. As shown in Figure 4-4, the ecological planning workflow of a specific layer includes the steps of planning goals setup, theory formation, data acquisition, map analysis, statistical verification, and planning performance evaluation. It is a circulation process that starts with a specific goal and needs to verify whether or not the planning performance reaches this goal.

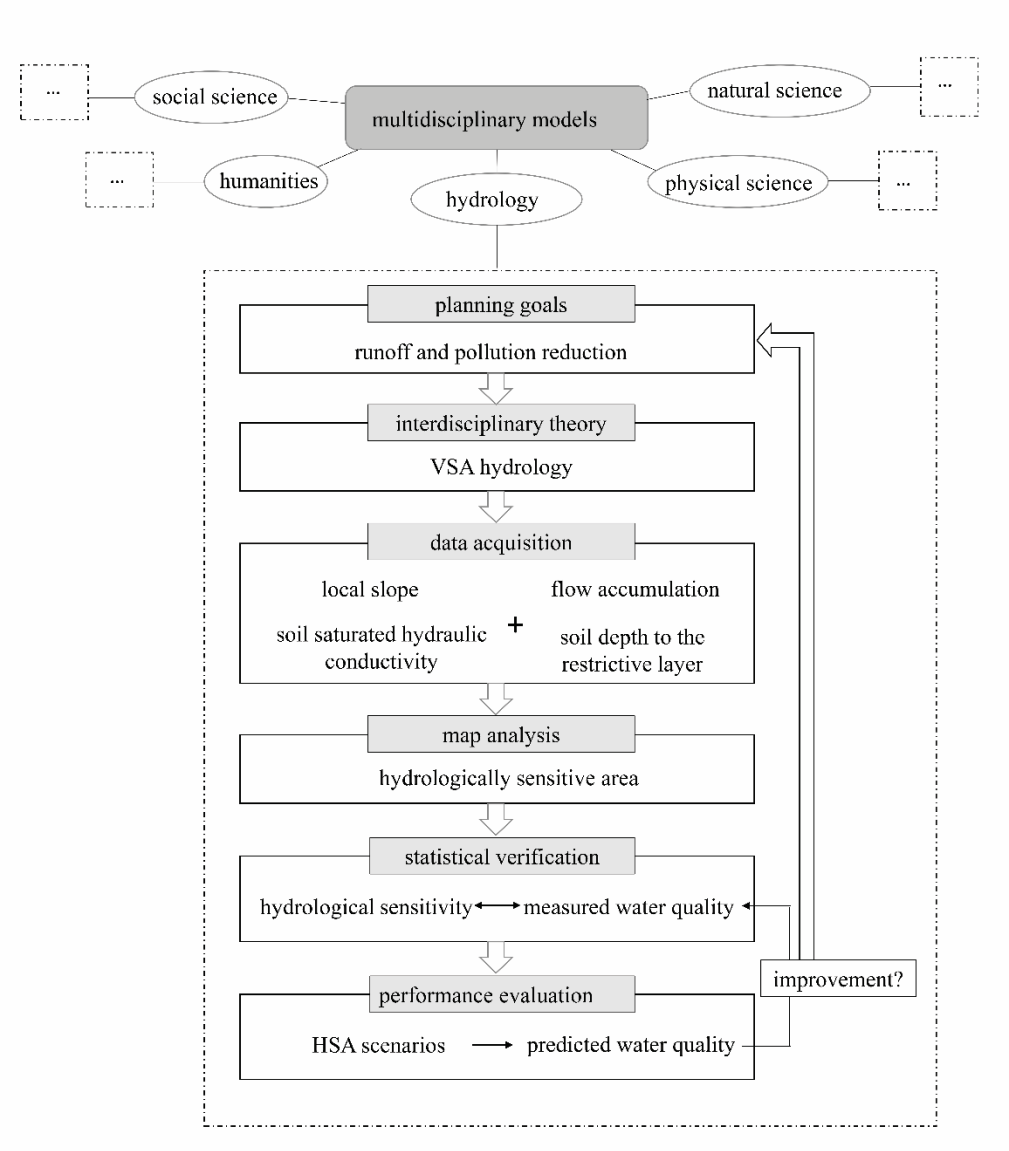


Figure 4-4. Data-driven ecological planning workflow using hydrology layer as an example

The dramatic increase of available data sources has made inventory analysis much more convenient as it is sometimes feasible to get all the in-situ data from public data sources. For example, Google Earth Engine (GEE) provides a data archive that includes more than 40 years of scientific datasets, such as climate and weather data, land cover data, geophysical data and so on (Gorelick et al., 2017). In this case study, land cover data, topography data, and climate data were

all obtained from GEE. The availability of big data greatly increases the generalization of evidence-based ecological planning, as the data are also freely available in other regions.

Using the data-driven approach to strengthen the scientific core of ecological planning is another important trend. In data-driven ecological planning, the emphasis is on identifying the most important planning factors that affect the final goals. Using the hydrology layer as an example, the most important landscape factor affecting stream water quality vary among different local contexts (Ding et al., 2016). Thus, statistical verification is needed to confirm whether or not the selected indicators in a given planning strategy have a significant impact on local stream water quality. Analytical methods such as stepwise regression, linear mixed model, geographically weighted regression (GWR), and redundancy analysis (RDA) are all helpful in analyzing the relationships between landscape factors and environmental indicators (Ragosta et al., 2010; Wang et al., 2014; Prat and Chang, 2012).

Performing scenario analysis of multiple ecological planning alternatives is important in evaluating the efficiency of different approaches, especially for multi-objective ecological planning (Yang and Li, 2011; Fu et al., 2016; Wu et al., 2016). Scenario analysis involves baseline and alternative scenario design, input data preparation, model calibration, and assessment of scenario outputs. In this study, scenario analysis was used to find an optimal threshold to define HSAs that can reduce more NO_3^--N loadings with larger areas cultivated. Some fully distributed hydrologic models have potential to simulate hydrological outcomes of ecological planning strategies with different spatial patterns. Such models include the Storm Water Management Model (SWMM), Mike SHE, Regional Hydro-Ecological Simulation System (RHESSys) and Distributed Hydrology–Soil–Vegetation Model (DHSVM) and so on (Qin et al., 2013; Trinh and Chui, 2013; Tague and Band, 2004; Cuo et al., 2008). In addition, it is helpful to measure and

document planning and design performance after a project is built, as it can be used as a reference for future ecological planning (Li et al., 2013).

Ecological planning has an interdisciplinary nature. The “layer-cake” approach requires expertise and knowledge from multiple disciplines to investigate each layer, especially in a data-driven approach. There are a number of interdisciplinary models that can be used to quantify each layer in ecological planning to support multiple goals in ecological, social and economic aspects, as shown in Figure 4-5.

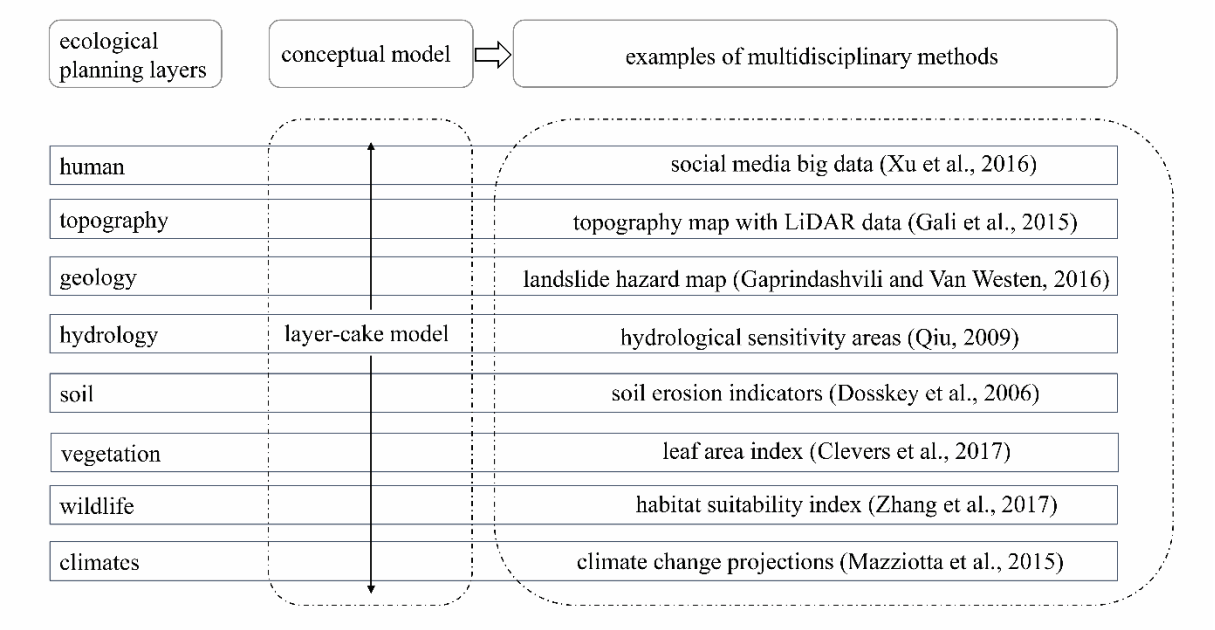


Figure 4-5. Multidisciplinary methods as extensions of the “layer-cake” model

To map the vegetation layer, leaf area index (LAI) maps which are derived from satellite imagery have been used to quantify the structure and function of forest ecosystems (Clevers et al., 2017). Areas with large LAIs represent dense forest areas and should be protected from cultivation. In the soil layer, the Water and Tillage Erosion Model (WATEM) and the Vegetative Filter Strip Model (VFSMOD) have been applied to derive soil erosion maps and designate corresponding conservation buffers (Dosskey et al., 2005; Dosskey et al., 2006). In the wildlife biology layer, a

habitat suitability index (HSI) map can be used to characterize habitat quality for selected wildlife species. For example, the HSI of marine animals was developed based on factors of sediments, water depth, water temperature, salinity, pH, dissolved oxygen and so on (Thomasma and Peterson, 1991; Chen et al., 2009; Zhang et al., 2017). In the topography layer, high resolution Light detection and ranging (LiDAR) data generate more accurate topographic maps, and perform better in mapping topographic related indexes such as power index (SPI), compound topographic index (CTI) and so on (Galzki et al., 2011; Tomer et al., 2013; Gali et al., 2015; Djodjic and Villa, 2015). In the hydrology layer, except for the HSA approach, the index method can be used to identify areas that are more sensitive to land use change, based on their contribution to the change of flow characteristics (Kalin and Hantush 2009; Noori et al., 2016).

Furthermore, ecological planning is a socio-ecological practice that incorporates social systems, such as politics, governance, economy and cultures (Xiang, 2019). Research about social impacts on planning has made some progress in quantifying the social benefits of ecological planning, such as strengthening social ties in neighborhoods, enhancing residents' mental health, increasing property values and so on (Tyrväinen and Miettinen, 2000; Francis et al., 2012; Kaźmierczak, 2013). Currently, social media data are used as a source of knowledge to measure people's attitudes and perceptions of built environment in order to inform planning and design strategies (Ciuccarelli et al., 2014; Nummi, 2019). In addition, it should be aware that the homeowner's preference, the market needs, and the public-private partnerships could all affect the implementation of an ecological planning project (Yang et al., 2015). After the formation of each layer in ecological planning, decision making models such as agent-based models (ABM), which simulate the actions and interactions of multiple entities, can be applied to find the optimal solution in complex socio-ecological systems (Matthews et al., 2007; Bruch and Atwell, 2015).

4.5 Conclusion

In this study, a data-driven approach to ecological planning by applying the HSA approach in the Middle Brazos-Bosque basin was demonstrated. Hydrological sensitivity was mapped and the most effective conservation areas for protecting a healthy watershed was designated. Correlation analysis and NLS regression results indicated that hydrological sensitivity was significantly positively correlated with NO_3^--N concentrations. Therefore, urban development and agriculture cultivation should avoid HSAs to protect stream water quality. Multiple planning scenarios were simulated in SWAT, and it was found that areas with the highest 2% hydrological sensitivity should be kept as green infrastructure in the watershed.

Given the results of this study, it was recommended that a standard data-driven approach to ecological planning should involve the steps of statistical verification and planning evaluation to test whether the proposed strategy fulfills the planning goals. There are a variety of models available from multiple disciplines for doing so, for both natural system and social systems. Data-driven approaches can offer a technical guide to realize McHarg's initial attempt at exploring a scientific and logic way of incorporating ecology in planning.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahern, J. (2011). From fail-safe to safe-to-fail: Sustainability and resilience in the new urban world. *Landscape and urban Planning*, *100*(4), 341-343.

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Journal*, *24*(1), 43-69.

Bruch, E., & Atwell, J. (2015). Agent-based models in empirical social research. *Sociological methods & research*, *44*(2), 186-221.

Calkins, M. (2012). *The sustainable sites handbook: A complete guide to the principles, strategies, and best practices for sustainable landscapes* (Vol. 39). John Wiley & Sons.

Chen, X., Li, G., Feng, B., & Tian, S. (2009). Habitat suitability index of Chub mackerel (Scomber japonicus) from July to September in the East China Sea. *Journal of oceanography*, *65*(1), 93-102.

Ciuccarelli, P., Lupi, G., & Simeone, L. (2014). *Visualizing the data city: social media as a source of knowledge for urban planning and management*. Springer Science & Business Media.

Clevers, J., Kooistra, L., & Van Den Brande, M. (2017). Using Sentinel-2 data for retrieving LAI and leaf and canopy chlorophyll content of a potato crop. *Remote Sensing*, *9*(5), 405.

Cuo, L., Lettenmaier, D. P., Mattheussen, B. V., Storck, P., & Wiley, M. (2008). Hydrologic prediction for urban watersheds with the Distributed Hydrology–Soil–Vegetation Model. *Hydrological processes*, *22*(21), 4205-4213.

Ding, J., Jiang, Y., Liu, Q., Hou, Z., Liao, J., Fu, L., & Peng, Q. (2016). Influences of the land use pattern on water quality in low-order streams of the Dongjiang River basin, China: a multi-scale analysis. *Science of the total environment*, *551*, 205-216.

Delgado, J. A., Khosla, R., & Mueller, T. (2011). Recent advances in precision (target) conservation. *Journal of Soil and Water Conservation*, *66*(6), 167A-170A.

Djodjic, F., & Villa, A. (2015). Distributed, high-resolution modelling of critical source areas for erosion and phosphorus losses. *Ambio*, *44*(2), 241-251.

Dosskey, M. G., Eisenhauer, D. E., & Helmers, M. J. (2005). Establishing conservation buffers using precision information. *Journal of Soil and Water Conservation*, *60*(6), 349-354.

Dosskey, M. G., Helmers, M. J., & Eisenhauer, D. E. (2006). An approach for using soil surveys to guide the placement of water quality buffers. *Journal of Soil and Water Conservation*, *61*(6), 344-354.

Espejel, I., Fischer, D. W., Hinojosa, A., García, C., & Leyva, C. (1999). Land-use planning for the Guadalupe Valley, Baja California, Mexico. *Landscape and Urban Planning*, *45*(4), 219-232.

Forman, R. T., & Godron, M. (1986). Landscape ecology John Wiley & Sons. *New York*, *4*, 22-28.

Francis, J., Wood, L. J., Knuiman, M., & Giles-Corti, B. (2012). Quality or quantity? Exploring the relationship between Public Open Space attributes and mental health in Perth, Western Australia. *Social science & medicine*, *74*(10), 1570-1577.

Frankenberger, J. R., Brooks, E. S., Walter, M. T., Walter, M. F., & Steenhuis, T. S. (1999). A GIS-based variable source area hydrology model. *Hydrological processes*, *13*(6), 805-822.

Fu, X., Wang, X., Schock, C., & Stuckert, T. (2016). Ecological wisdom as benchmark in planning and design. *Landscape and Urban Planning*, *155*, 79-90.

Gali, R. K., Soupir, M. L., Kaleita, A. L., & Daggupati, P. (2015). Identifying potential locations for grassed waterways using terrain attributes and precision conservation technologies. *Transactions of the ASABE*, *58*(5), 1231-1239.

Galzki, J. C., Birr, A. S., & Mulla, D. J. (2011). Identifying critical agricultural areas with three-meter LiDAR elevation data for precision conservation. *Journal of Soil and Water Conservation*, *66*(6), 423-430.

Gaprindashvili, G., & Van Westen, C. J. (2016). Generation of a national landslide hazard and risk map for the country of Georgia. *Natural hazards*, *80*(1), 69-101.

Giri, S., & Qiu, Z. (2016). Understanding the relationship of land uses and water quality in Twenty First Century: A review. *Journal of environmental management*, *173*, 41-48.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18-27.

Grove, J. M., Cadenasso, M. L., Pickett, S. T., Machlis, G. E., & Burch, W. R. (2015). *The Baltimore school of urban ecology: space, scale, and time for the study of cities*. Yale University Press.

Kalin, L., & Hantush, M. M. (2009). An auxiliary method to reduce potential adverse impacts of projected land developments: subwatershed prioritization. *Environmental management*, *43*(2), 311.

Kaźmierczak, A. (2013). The contribution of local parks to neighbourhood social ties. *Landscape and urban planning*, *109*(1), 31-44.

Kosmas, C., Ferrara, A., Briasouli, H., & Imeson, A. (1999). Methodology for mapping environmentally sensitive areas (ESAs) to desertification. *The Medalus project: Mediterranean desertification and land use. Manual on key indicators of desertification and mapping environmentally sensitive areas to desertification (Kosmas C, Kirkby M, Geeson N eds), European Union, 18882*, 31-47.

Herrington, S. (2010). The nature of Ian McHarg's science. *Landscape Journal*, *29*(1), 1-20.

Li, M. H., Dvorak, B., Luo, Y., & Baumgarten, M. (2013). Landscape performance: Quantified benefits and lessons learned from a treatment wetland system and naturalized landscapes. *Landscape Architecture Frontiers*, *1*(4), 56-68.

Liao, K. H., & Chan, J. K. H. (2016). What is ecological wisdom and how does it relate to ecological knowledge?. *Landscape and Urban Planning*, *155*, 111-113.

Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., ... & Western, A. W. (2018). Key factors influencing differences in stream water quality across space. *Wiley Interdisciplinary Reviews: Water*, *5*(1), e1260.

Martin-Mikle, C. J., de Beurs, K. M., Julian, J. P., & Mayer, P. M. (2015). Identifying priority sites for low impact development (LID) in a mixed-use watershed. *Landscape and urban planning*, *140*, 29-41.

Matthews, R. B., Gilbert, N. G., Roach, A., Polhill, J. G., & Gotts, N. M. (2007). Agent-based land-use models: a review of applications. *Landscape Ecology*, *22*(10), 1447-1459.

Mazziotta, A., Triviño, M., Tikkanen, O. P., Kouki, J., Strandman, H., & Mönkkönen, M. (2015). Applying a framework for landscape planning under climate change for the conservation of biodiversity in the Finnish boreal forest. *Global change biology*, *21*(2), 637-651.

McHarg, I. L. (1969). *Design with nature*. New York, NY: Doubleday/Natural History Press.

McHarg, I. L., & Steiner, F. R. (1998). To heal the Earth: Selected writings of Ian L.

McHarg, I. L. (2006). *The essential Ian McHarg: writings on design and nature*. Island Press.

Meerow, S. (2015). Defining urban resilience: A review, landscape and urban planning.

Mueller, T. G., Cetin, H., Fleming, R. A., Dillon, C. R., Karathanasis, A. D., & Shearer, S. A. (2005). Erosion probability maps: Calibrating precision agriculture data with soil surveys using logistic regression. *Journal of soil and water conservation*, *60*(6), 462-468.

Niemelä, J. (1999). Ecology and urban planning. *Biodiversity & Conservation*, *8*(1), 119-131.

Noori, N., Kalin, L., Sen, S., Srivastava, P., & Lebleu, C. (2016). Identifying areas sensitive to land use/land cover change for downstream flooding in a coastal Alabama watershed. *Regional environmental change*, *16*(6), 1833-1845.

Nummi, P. (2019). Social media data analysis in urban e-planning. In *Smart Cities and Smart Spaces: Concepts, Methodologies, Tools, and Applications* (pp. 636-651). IGI Global.

Palazzo, D., & Steiner, F. R. (2012). *Urban ecological design: a process for regenerative places* (Vol. 12). Island Press

Pratt, B., & Chang, H. (2012). Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. *Journal of hazardous materials*, *209*, 48-58.

Qin, H. P., Li, Z. X., & Fu, G. (2013). The effects of low impact development on urban flooding under different rainfall characteristics. *Journal of environmental management*, *129*, 577-585.

Qiu, Z. (2003). A VSA-based strategy for placing conservation buffers in agricultural watersheds. *Environmental Management*, *32*(3), 299-311.

Qiu, Z. (2009). Assessing critical source areas in watersheds for conservation buffer planning and riparian restoration. *Environmental management*, *44*(5), 968-980.

Qiu, Z., Hall, C., Drewes, D., Messinger, G., Prato, T., Hale, K., & Van Abs, D. (2013). Hydrologically sensitive areas, land use controls, and protection of healthy watersheds. *Journal of Water Resources Planning and Management*, *140*(7), 04014011.

Ragosta, G., Evensen, C., Atwill, E. R., Walker, M., Ticktin, T., Asquith, A., & Tate, K. W. (2010). Causal connections between water quality and land use in a rural tropical island watershed. *EcoHealth*, *7*(1), 105-113.

Salvati, L., Ferrara, C., & Corona, P. (2015). Indirect validation of the Environmental Sensitive Area Index using soil degradation indicators: A country-scale approach. *Ecological indicators*, *57*, 360-365.

Schumacher, J. A., Kaspar, T. C., Ritchie, J. C., Schumacher, T. E., Karlen, D. L., Venteris, E. R., ... & Fenton, T. E. (2005). Identifying spatial patterns of erosion for use in precision conservation. *Journal of soil and water conservation*, *60*(6), 355-362.

Steiner, F., McSherry, L., & Cohen, J. (2000a). Land suitability analysis for the upper Gila River watershed. *Landscape and urban planning*, *50*(4), 199-214.

Steiner, F., Blair, J., McSherry, L., Guhathakurta, S., Marruffo, J., & Holm, M. (2000b). A watershed at a watershed: the potential for environmentally sensitive area protection in the upper San Pedro Drainage Basin (Mexico and USA). *Landscape and urban planning*, *49*(3-4), 129-148.

Steiner, F. (2011). Landscape ecological urbanism: Origins and trajectories. *Landscape and urban planning*, *100*(4), 333-337.

Steiner, F. (2016). The application of ecological knowledge requires a pursuit of wisdom. *Landscape and Urban Planning*, *155*, 108-110.

Sustainable Sites Initiative. (2009). The sustainable sites initiative: guidelines and performance benchmarks 2009.

Tague, C. L., & Band, L. E. (2004). RHESSys: Regional Hydro-Ecologic Simulation System—An object-oriented approach to spatially distributed modeling of carbon, water, and nutrient cycling. *Earth interactions*, *8*(19), 1-42.

Thomasma, L. E., Drummer, T. D., & Peterson, R. O. (1991). Testing the habitat suitability index model for the fisher. *Wildlife Society Bulletin (1973-2006)*, *19*(3), 291-297.

Tomer, M. D., Crumpton, W. G., Bingner, R. L., Kostel, J. A., & James, D. E. (2013). Estimating nitrate load reductions from placing constructed wetlands in a HUC-12 watershed using LiDAR data. *Ecological Engineering*, *56*, 69-78.

Trinh, D. H., & Chui, T. F. M. (2013). Assessing the hydrologic restoration of an urbanized area via an integrated distributed hydrological model. *Hydrology and Earth System Sciences*, *17*(12), 4789-4801.

Tyrväinen, L., & Miettinen, A. (2000). Property prices and urban forest amenities. *Journal of environmental economics and management*, *39*(2), 205-223.

Walter, M. T., Walter, M. F., Brooks, E. S., Steenhuis, T. S., Boll, J., & Weiler, K. (2000). Hydrologically sensitive areas: variable source area hydrology implications for water quality risk assessment. *Journal of Soil and Water Conservation*, *55*(3), 277-284.

Walter, M. T., Steenhuis, T. S., Mehta, V. K., Thongs, D., Zion, M., & Schneiderman, E. (2002). Refined conceptualization of TOPMODEL for shallow subsurface flows. *Hydrological Processes*, *16*(10), 2041-2046.

Wang, G., Xu, Z., & Zhang, S. (2014). The influence of land use patterns on water quality at multiple spatial scales in a river system. *Hydrological processes*, *28*(20), 5259-5272.

Wang, X., Palazzo, D., & Carper, M. (2016). Ecological wisdom as an emerging field of scholarly inquiry in urban planning and design. *Landscape and Urban Planning*, *155*, 100-107.

Wu, J. (2006). Landscape ecology, cross-disciplinarity, and sustainability science. *Landscape Ecology*, *21*(1), 1-4.

Wu, J., & Wu, T. (2013). Ecological resilience as a foundation for urban design and sustainability. In *Resilience in Ecology and Urban Design* (pp. 211-229). Springer, Dordrecht.

Yang, B., & Li, M.-H. (2011). Assessing planning approaches by watershed streamflow modeling: Case study of The Woodlands; Texas. *Landscape and Urban Planning*, *99*(1), 9-22.

Yang, B., Li, M.-H., & Huang, C.-S. (2015). Ian McHarg's ecological planning in The Woodlands, Texas: Lessons learned after four decades. *Landscape Research*, 40(7), 773-794.

Yang, B., & Li, S. (2016). Design with Nature: Ian McHarg's ecological wisdom as actionable and practical knowledge. *Landscape and Urban Planning*, *155*, 21-32.

Yang, B. (2018). *Landscape Performance: Ian McHarg's ecological planning in The Woodlands, Texas*. Routledge.

Yu, K. (1996). Security patterns and surface model in landscape ecological planning. *Landscape and urban planning*, *36*(1), 1-17.

Xiang, W. N. (2014). Doing real and permanent good in landscape and urban planning: Ecological wisdom for urban sustainability. *Landscape and Urban Planning*, (121), 65-69.

Xiang, W. N. (2019). Ecopracticology: the study of socio-ecological practice. *Socio-Ecological Practice Research*, 1-8.

Xu, Z., Liu, Y., Yen, N., Mei, L., Luo, X., Wei, X., & Hu, C. (2016). Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*.

Zhang, Z., Zhou, J., Song, J., Wang, Q., Liu, H., & Tang, X. (2017). Habitat suitability index model of the sea cucumber Apostichopus japonicus (Selenka): A case study of Shandong Peninsula, China. *Marine pollution bulletin, 122*(1-2), 65-76.

CHAPTER 5 CONCLUSION AND RECOMMENDATION

Landscape-water quality nexus studies with large spatial content and a long time period require a complex research design, large data inputs, and robust analytical methods. In this research, the relationships between landscape characteristics and stream water quality in the Texas Gulf Region from 1990 to 2011 were quantified and analyzed, and the relevant management solutions were proposed. It was discovered that given the same impervious surface area, urban spatial pattern was significantly influential on stream water quality. High-density aggregated urban development led to significantly better stream water quality compared to the current sprawl development. Regarding the general land-water relationships, urban development patterns, soil, and climate were the most significant factors in determining all pollutant concentrations, but the relationships varied according to the season and location. The relationships between land cover, climate and water quality did not change significantly from 1990 to 2011. The variations of landscape and climatic factors at the local scale accounted for more than 50% of the variations in stream water quality. At the basin scale, they accounted for about 20% of the stream water quality variations. Management practice should target different regions and basins. Generally, placing BMPs in HSAs was efficient in reducing nutrient loading.

This research was novel as it combines cutting edge technologies to frame a large-scale longitudinal study in the landscape architecture discipline. Machine learning was used to find the most important factors affecting stream water quality, and to predict stream water quality given different urban spatial patterns. Linear mixed models were designed to quantify complex spatially and temporally varying landscape-water quality relationships in a simple and interpretable approach. Hydrological modeling was used to find the threshold to define HSAs that are the most efficient at reducing nutrient loadings. In addition, an annual land cover classification remote

sensing algorithm was designed to obtain the annual land cover change, which was used to explain the change in stream water quality. Overall, this dissertation determines an advanced technical workflow to study large scale water quality issues.

In the 2011 cross-sectional study, it was concluded that urban spatial patterns, soil, and climate were the most important factors in determining stream pollutant concentrations. The configuration of urban area was more important than the composition of urban area. Using a random forest predictive model, it was found that high density aggregated development contributed to the lowest level of stream pollutant concentrations. This conclusion supports the urban planning policy towards compact city planning. Methodologically, the machine learning model was flexible and robust. Thus, it could incorporate any other factors of interest, and was applicable to be generalized to other regions.

In the longitudinal study, the focus was on how the variations in the landscape-water quality relationships were explained with different spatial and temporal scales. The annual land cover classification results indicated an obvious deforestation trend in the Piney Woods, the South Texas Plains and the Gulf Prairies and Marshes ecoregions after 2000. This deforestation and urban expansion together led to water quality degradation in the Texas Gulf Region. For example, adding 1 percent of urban area led to a 6.31% increase of NO_3^--N concentration and a 3.52% increase of PO_4^{3-}-P concentration in the Texas Gulf Region. It was also discovered that some unobserved characteristics other than land cover and climate led to the high nutrient concentration in the Middle Colorado-Concho and the Lower Trinity basins, and the high *E.coli* concentration in the San Jacinto basin. Overall, the Texas Gulf Region had quite heterogenous land-water relationships, and the specific management practice should be targeted at the local level.

Finally, a basin-scale study in the Middle Brazos-Bosque basin was conducted to verify that placing BMP in HSAs was effective in reducing nutrient concentration. The HSA approach had been proposed and mapped by other studies, but there was little research effort verifying the threshold to delineate HSA. After a comparison among multiple planning scenarios simulated in SWAT, it was found that areas with the highest 2% hydrological sensitivity should be preserved as green space in the watershed to control nutrient pollutions.

Several policy recommendations were driven from this study. First, compact city should be promoted in land use planning for stream water quality protection. Regulating urban sprawl would be particularly helpful in reducing *E.coli* concentration in the Texas coastal areas. Second, stream water quality conservation should be paid greater attention to areas with higher soil storage and areas with high precipitation. Precision agriculture and conservation tillage should be applied in the north parts of the Texas Gulf Region. The reforestation efforts should be exerted in the Piney Woods ecoregion to avoid further habitat loss. Habitat restoration actions should be taken in the Gulf Prairie and Marshes ecoregion. Third, urban development and agriculture cultivation should avoid HSAs to protect stream water quality. Green infrastructure such as constructed wetland and bioretention are more appropriate to site on HSAs. Resilient redevelopment strategies such as connecting impervious surface should also be prioritized on HSAs.

Based on the findings, the recommendation is made for future research to focus on validating causal relationships between landscape factors, climatic factors, and stream water quality. It is also worth incorporating socioeconomic factors to form a comprehensive framework to understand how stream water quality responds to diverse human activities. Explicit planning policy implications and design solutions can be drawn with this full picture of land-water relationships. In such studies, a flexible combination of big data technologies, conventional statistical methods, and hydrological

modeling holds great promise in getting more interpretable and credible results. I foresee the necessity of continuous research efforts to apply cutting edge methods in water-oriented planning and design, which can contribute to the plan for a more sustainable future.