




2018

## ACCOUNTING FOR SPATIAL AUTOCORRELATION IN MODELING THE DISTRIBUTION OF WATER QUALITY VARIABLES

Lorrayne Miralha

*University of Kentucky*, [lorrayne.miralha@gmail.com](mailto:lorrayne.miralha@gmail.com)

Author ORCID Identifier:

 <https://orcid.org/0000-0003-1448-9321>

Digital Object Identifier: <https://doi.org/10.13023/ETD.2018.196>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

### Recommended Citation

Miralha, Lorrayne, "ACCOUNTING FOR SPATIAL AUTOCORRELATION IN MODELING THE DISTRIBUTION OF WATER QUALITY VARIABLES" (2018). *Theses and Dissertations--Geography*. 55.

[https://uknowledge.uky.edu/geography\\_etds/55](https://uknowledge.uky.edu/geography_etds/55)

This Master's Thesis is brought to you for free and open access by the Geography at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Geography by an authorized administrator of UKnowledge. For more information, please contact [UKnowledge@lsv.uky.edu](mailto:UKnowledge@lsv.uky.edu).

## **STUDENT AGREEMENT:**

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

## **REVIEW, APPROVAL AND ACCEPTANCE**

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Lorrayne Miralha, Student

Dr. Jonathan Phillips, Major Professor

Dr. Andrew Wood, Director of Graduate Studies

ACCOUNTING FOR SPATIAL AUTOCORRELATION  
IN MODELING THE DISTRIBUTION OF WATER QUALITY VARIABLES

---

THESIS

---

A thesis submitted in partial fulfillment of the  
requirements for the degree of Master of Arts in the  
College of Arts and Sciences at the University of Kentucky

By

Lorrayne Miralha Marins da Silva

Lexington, Kentucky

Director: Dr. Jonathan Phillips, Professor of Geography

Lexington, Kentucky

2018

Copyright© Lorrayne Miralha 2018

## ABSTRACT OF THESIS

### ACCOUNTING FOR SPATIAL AUTOCORRELATION IN MODELING THE DISTRIBUTION OF WATER QUALITY VARIABLES

Several studies in hydrology have reported differences in outcomes between models in which spatial autocorrelation (SAC) is accounted for and those in which SAC is not. However, the capacity to predict the magnitude of such differences is still ambiguous. In this thesis, I hypothesized that SAC, inherently possessed by a response variable, influences spatial modeling outcomes. I selected ten watersheds in the USA and analyzed them to determine whether water quality variables with higher Moran's  $I$  values undergo greater increases in the coefficient of determination ( $R^2$ ) and greater decreases in residual SAC (rSAC) after spatial modeling. I compared non-spatial ordinary least squares to two spatial regression approaches, namely, spatial lag and error models. The predictors were the principal components of topographic, land cover, and soil group variables. The results revealed that water quality variables with higher inherent SAC showed more substantial increases in  $R^2$  and decreases in rSAC after performing spatial regressions. In this study, I found a generally linear relationship between the spatial model outcomes ( $R^2$  and rSAC) and the degree of SAC in each water quality variable. I suggest that the inherent level of SAC in response variables can predict improvements in models before spatial regression is performed. The benefits of this study go beyond modeling selection and performance, it has the potential to uncover hydrologic connectivity patterns that can serve as insights to water quality managers and policy makers.

**KEYWORDS:** spatial autocorrelation; water quality variables; spatial regression modeling; coefficient of determination; residual autocorrelation

---

Lorrayne Miralha

---

April 26<sup>th</sup>, 2018

---

ACCOUNTING FOR SPATIAL AUTOCORRELATION  
IN MODELING THE DISTRIBUTION OF WATER QUALITY VARIABLES

By

Lorrayne Miralha Marins da Silva

Dr. Jonathan Phillips

---

Director of Thesis

Dr. Andrew Wood

---

Director of Graduate Studies

April 26<sup>th</sup>, 2018

---

DEDICATION

*To my family, Deyse, Jorge, Lawson and  
my fiancée, Hisham, for their love, support, and affection.*

## ACKNOWLEDGMENTS

First, I would like to thank my adviser Dr. Daehyun Kim for all his patience and support during these 2 years at the University of Kentucky. He was always ready to help and answer my questions whenever I ran into research problems. He was also a friend that listened to my personal struggles during graduate school and offered support and guidance. I could never have found a better adviser. Dr. Kim knew the right moment to push me forward, had faith in my work, and believed in my potential. For all of it, I will always be grateful to have found you, Boss.

I would also like to thank the experts that I had the pleasure of meeting throughout my journey, most notably my committee members. Thank you to Dr. Jonathan Phillips, known as ‘the angry professor’, for sharing your knowledge, your jokes, and for accepting the position of my official adviser in this last year. Dr. Liang Liang, thank you for always keeping your door open, for having patience, and for always being willing to answer my questions. I would also like to thank Dr. Ole Wendroth and Dr. Alice Turkington for their kindness and for sharing their experience and wisdom with me. All the professors and staff (principally Lori and Jeff) I have met during this master’s program have made each second an experience I will never forget.

I also want to thank my family. My parents, Deyse and Jorge, and my brother, Lawson. I can say that I grew up the way I am because of you. You were part of every phase of my life, stood beside me throughout the difficult and happy moments, and celebrated my success. I will always be thankful for every piece of advice, for every second spent shaping me into a better person, and for all of the effort you invested to make my life easier than I could have expected.

I am also thankful to my fiancée, Hisham, who walked with me through this graduate school journey, believed in my potential, and gave me incentive to keep going. I am happy to have found you, and I am grateful for all the moments we have spent together.

I am thankful for all the friends I have made. I would like to thank Tasnuba, Marissa, Jaeyeon, and Laçin for making grad school a lighter environment and full of joy. I am also thankful for Dayna, Guetchine, and all the students in the department of geography at UK. Each one of you contributed to where I am today.

Finally, I also would like to acknowledge the financial support of National Science Foundation, National Research Foundation of South Korea, and the Department of Geography at the University of Kentucky. This support was essential in the process of developing this research and attaining my master’s degree.

## TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	iii
LIST OF TABLES.....	vi
LIST OF FIGURES.....	vii
CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: LITERATURE REVIEW	
2.1. Spatial Autocorrelation (SAC)	
2.1.1. Definition.....	5
2.1.2. Sources of SAC .....	6
2.1.3. Importance of accounting for SAC .....	8
2.1.4. Methods to measure and account for SAC	
2.1.4.1 Measuring SAC.....	11
2.1.4.2 Accounting for SAC.....	14
2.2. Water Quality	
2.2.1. Importance of water quality.....	19
2.2.2. SAC and water quality.....	19
2.2.3. Spatial approaches to water quality modeling.....	21
CHAPTER 3: METHODOLOGY	
3.1. Study areas.....	24
3.2. Dependent variables.....	28
3.3. Delineation of upstream area.....	28
3.4. Independent variables.....	30
3.5. Data processing.....	33



3.6. Testing for Spatial Autocorrelation (SAC).....	34
3.7. Statistical models.....	34
3.8. Model comparison.....	35
<b>CHAPTER 4: RESULTS</b>	
4.1. Changes in R <sup>2</sup> values.....	37
4.2. Changes in residual Spatial Autocorrelation (rSAC).....	39
4.3. Overall changes between non-spatial OLS and spatial regression models.....	40
4.4. Summary of findings.....	42
<b>CHAPTER 5: DISCUSSION.....</b>	<b>43</b>
<b>CHAPTER 6: CONCLUSIONS &amp; FUTURE WORK.....</b>	<b>46</b>
<b>APPENDICES</b>	
Appendix A: Larger maps for better visualization of water stations location.....	49
Appendix B: Model outcomes and Moran's <i>I</i> values per water quality variables on each watershed.....	59
<b>REFERENCES.....</b>	<b>70</b>
<b>VITA.....</b>	<b>77</b>

## LIST OF TABLES

Table 1. Description of the ten study sites investigated in this research.....	26
Table 2. Study areas, number of stations per study area, and water quality variables with the respective Moran's <i>I</i> values in parentheses.....	27
Table 3. Data sources and details of dependent and independent variables.....	32
Table 4. Summary of the mean values of spatial autocorrelation ( <i>I</i> ) in response variables, mean values of the non-spatial OLS outcomes and mean improvement in $R^2$ and reduction rSAC after spatial regression per state. Additionally, linear regression model coefficients, $R^2$ , and <i>p</i> -value of the Changes in $R^2$ and rSAC per state.....	41

## LIST OF FIGURES

Figure 1. Conceptualization of the main ideas (PC, principal components; OLS, ordinary least squares; SAC, spatial autocorrelation).....	4
Figure 2. Land cover characteristics of each state and watershed shape.....	25
Figure 3. Upstream area delineation and their respective buffer zones in tones of gray...	29
Figure 4. Evaluation of the hypothesis.....	36
Figure 5. Relationship between the spatial autocorrelation (SAC) of each water quality variable and $R^2$ .....	38
Figure 6. Relationship between the spatial autocorrelation (SAC) of each water quality variable (represented by Moran's $I$ values) and the SAC of model residuals (rSAC).....	40
Figure 7. Linear regression models demonstrating that the magnitude of improvement of model performance after spatial lag and error modeling is significantly and linearly explained by the SAC inherently possessed by water quality variables.....	42

## **CHAPTER 1: INTRODUCTION**

Water is an element crucial for life on Earth and is closely linked to the well-being of societies as well as the sustainability of aquatic ecosystems. A combination of natural and anthropogenic factors can adversely impact water quality. Human impacts involve general land use practices (e.g., agriculture, irrigation practices, urbanization, and deforestation), while natural factors include slope, elevation, vegetation cover, soil type, precipitation, and streamflow (Calow & Petts, 1992; Pratt & Chang, 2012; Yu et al., 2013). River characteristics are generally dependent upon land use and geomorphological features of the watershed. In addition, water use patterns associated with the location of a region and its interactions with neighboring regions influence the quality of water bodies (Franczyk & Chang, 2009). These factors are responsible for the spatial variability of water quality and are often treated as predictor variables in many hydrologic models (Vreboš et al., 2017). To provide better insights to future watershed management policies, understanding spatial trends associated with water quality variables is of extreme importance.

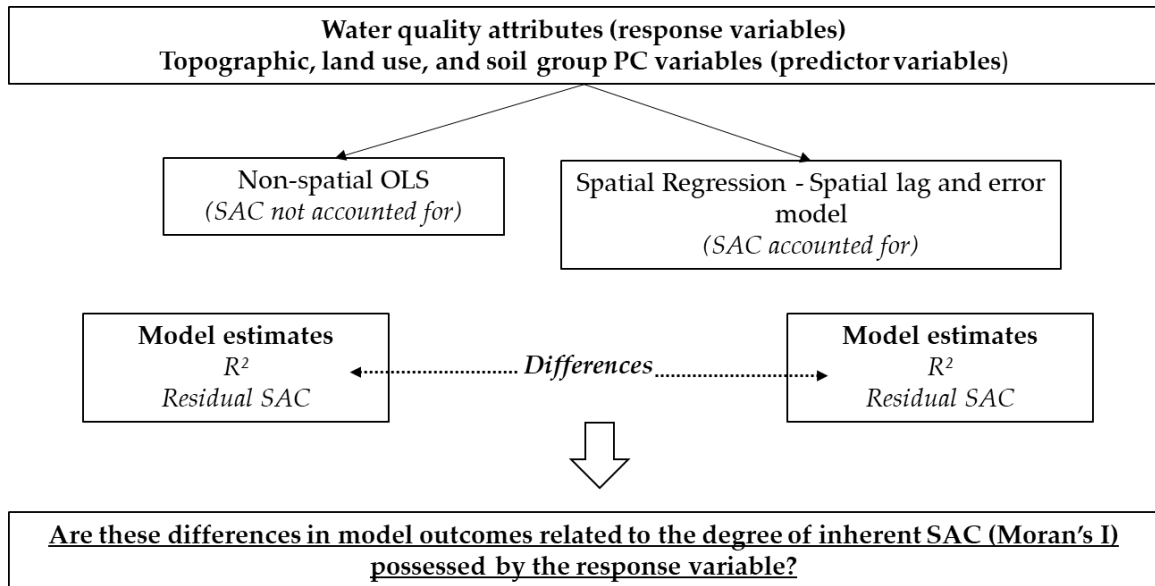
Space serves a vital role in structuring hydrological systems. Spatial autocorrelation (SAC) is an inherent property of spatial features such as streams and rivers (Legendre & Fortin, 1989). Legendre loosely defined the concept of SAC as “the property of random variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation), or less similar (negative autocorrelation) than expected for randomly associated pairs of observations” (p. 1659) (Legendre, 1993). For example, causes of positive autocorrelation in stream water quality could be associated with similarities in local habitats or turbulent mixing and water chemistries of stream flows. In

contrast, specific local built structures, such as beaver dams, fallen trees in stream channels, and territorial fishes, could be causes of negative SAC (Isaak et al., 2014). Given these interactions over space (i.e., water flow from upstream to downstream areas, local biota, and water use patterns), it is necessary to consider the presence and potential effects of SAC in water quality modeling.

Numerous studies in ecology, geography, and hydrology have noted the importance of accounting for SAC (Miller et al., 2007; Chang, 2008; Tu, 2011; Kim, 2013). These studies show that ignoring SAC can bias model outcomes and parameter estimates, leading to poor statistical inference and violation of the independence assumption of conventional regression approaches (Cliff & Ord, 1972; Dormann, 2007; Beale et al., 2010; Isaak et al., 2014; Kim et al., 2016). For example, models that ignore spatial effects (e.g., ordinary least squares; OLS) are likely to produce autocorrelated residuals violating the independent errors assumption. This can inflate the Type I error rate, wrongfully rejecting a null hypothesis. Many spatial approaches have been developed in order to overcome such limitations of non-spatial counterparts. These approaches include, but are not restricted to, regression kriging, simultaneous autoregressive modeling, conditional autoregressive modeling, spatial lag modeling, spatial error modeling, spatial eigenvector mapping, and geographically weighted regression (Griffith, 2000; Lichstein et al., 2002; Hengl et al., 2004; Griffith & Peres-Neto, 2006; Ver Hoef et al., 2006; de Marco et al., 2008; Kissling & Carl, 2008; Bini et al., 2009; Miller, 2012; Václavik et al., 2012; Kim, 2013; Isaak et al., 2014).

Several water quality studies have compared outcomes between spatial and non-spatial regressions (Chang, 2008; Franczyk & Chang, 2009; Tu, 2011; Chang et al., 2012;

Pratt & Chang, 2012; Yu et al., 2013; Huang et al., 2014; Netusil et al., 2014). In general, spatial models presented significant increases in  $R^2$  values and decreases in residual SAC (rSAC), indicating that spatial model performance exhibited clear improvements over the non-spatial approach. However, according to the literature on hydrological modeling, it is still uncertain when such improvements become large or small. Assuming that each water quality variable presents a unique degree of inherent SAC, I hypothesize that this SAC (possessed by a response variable; i.e., a water quality variable) influences the outcomes of spatial modeling. This study tests if water quality variables with a higher amount of SAC would exhibit greater improvement in model outcomes than those with a lower amount of SAC (see Figure 1). I evaluate this hypothesis across divergent regions of the USA to enable a general understanding of the effect of SAC possessed by water quality variables. I examine if SAC is a consistent determinant of the magnitude of model improvements even when watershed characteristics diverge. If this is indeed the case, I can potentially determine the degree of improvement in model fit before performing a spatial regression simply by measuring the inherent SAC level of a water quality variable. This study can also serve as a useful screening technique where modelers could use a SAC metric to predict the spatial pattern in the independent variable using a spatially explicit method.



**Figure 1.** Conceptualization of the main ideas of the study (PC, principal components; OLS, ordinary least squares; SAC, spatial autocorrelation) (from Miralha & Kim, 2018).

Following this chapter, Chapter 2 will provide the necessary background information to understand the problem addressed. I discuss the definition of SAC, conceptual background, its importance, and sources, as well as techniques to measure SAC and account for it. In the second part of this chapter, I give emphasis on the importance of water quality and the sources of SAC in aquatic ecosystems. I also explain the necessity of accounting for SAC in water quality modeling and provide examples of models that are used in water research as well as studies that accounted explicitly for SAC applying spatial modeling approaches. In Chapter 3, I will discuss the study areas, dependent and independent variables, techniques, and models used to achieve the objective of this study. Chapter 4 will cover the findings of this research. In Chapter 5, I discuss the findings and limitations. Finally, Chapter 6 concludes this thesis with a discussion of potential future research.

## **CHAPTER 2: LITERATURE REVIEW**

### **2.1. Spatial Autocorrelation (SAC)**

#### **2.1.1. Definition**

Spatial autocorrelation (SAC) has become one of the major points in modeling over a variety of fields in the past decades. According to Griffith (2009), the word auto is a prefix that literally means self while correlation is a description of the nature and the degree of a relationship between a pair of quantitative variables (p.1). Therefore, if we think about a variable and connect it to the concept of autocorrelation, we can infer that it is a variable that is correlated (has a degree of relationship) to itself. SAC is a common phenomenon in environmental and ecological data, where heterogeneity tends to be a function of clusters in environmental conditions and ecological processes (Bocard et al., 1992; Miller et al., 2007). Following this idea, we can understand what exactly spatial autocorrelation means.

SAC was defined by different authors such as Hubert, Golledge and Constanza (1981): “Given a set  $S$  containing  $n$  geographical units, spatial autocorrelation refers to the relationship between some variable observed in each of the  $n$  localities and a measure of geographical proximity defined for all  $n(n-1)$  pairs chosen from  $n$  (p. 224).” Tobler in 1970, loosely defined SAC through the attempt to establish the First Law of Geography: “everything is related to everything else, but near things are more related than distant things (p.236).” However, it is important to state that a variable can be related to itself through distance, but it does not always happen through the same mechanism (Griffith, 2009). Legendre (1993) stated that SAC is “the property of random variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation)



or less similar (negative autocorrelation) than expected for randomly associated pairs of observation (p. 1659).” He pointed out that SAC can be negative or positive depending on the distance class between the observations. Legendre and Fortin (1989) explained that positive autocorrelation (when variable takes similar values) is very common in ecology for short distances among samples. In a positive SAC scenario, when distance increases, negative SAC occurs where more distant observations have a higher chance to present significantly different values. If for short distances negative SAC occurs, it can be the result of a unique/local phenomenon or the sampling interval is too large compared to the cluster size. Several ways to explain and account for SAC have been developed. Although in different ways, the association to spatial relationship, similarity between variables per distance, and spatial dependence were always left explicitly in each definition. Therefore, SAC is a way to understand and measure the observations relationship intensity through the distance they are apart. However, the source of this phenomenon can vary, thus ways to account for and explain SAC in geographical, environmental, and ecological data are still in demand.

### **2.1.2. Sources of SAC**

Several studies have pointed the sources of SAC in environmental, ecological, and hydrological data (Cliff & Ord, 1981; Diniz-Filho et al., 2003; Dormann et al., 2007; Beale et al., 2010). Fortin et al. (2002) explained that SAC results from four sources: spurious, interpolative, true, and induced autocorrelation. The first happens when hidden processes affect the spatial arrangement of the data. Interpolative autocorrelation comes from interpolated, extrapolated, or smoothed spatial surfaces. The true SAC is associated with

causal interactions among samples near to each other while the induced SAC is related to the relationship between a dependent variable and another spatially autocorrelated variable. The last two SAC sources arise from spatial processes and have been the focus of spatial studies (Cliff & Ord, 1981; Legendre & Legendre, 1998).

There are four types of spatial processes (i.e., that operate in geographic space): diffusion, dispersal, interaction, and processes involving exchange and transfer (Haining, 2003). A diffusion process happens when an attribute spreads in a population and it is possible to observe areas or individuals that have this attribute. It is also different from the dispersal process because dispersal involves the spread of a population itself. Interaction involves the outcomes in one location that influence and are influenced by outcomes in other locations. Finally, processes of exchange and transfer are highly linked to the urban and economic studies creating inter-linked economic spaces.

The importance of SAC in explaining spatial processes is scale dependent (Haining, 2003), thus understanding hidden mechanisms will depend on the scale of the observed phenomenon or event. Václavik et al. (2012) explained that there are two types of recognized factors associated with scale, exogenous and endogenous. Exogenous factors are linked to broad-scale spatial trends and include underlying environmental conditions such as soil, topography, and hydrology. Endogenous factors occur at fine-scale and are usually associated with biological processes such as dispersal, vegetative reproduction, metapopulation dynamics, predation, and competition (Lichstein et al., 2002; Miller, 2012; Kim & Shin, 2016). These factors are also known in the literature as extrinsic (exogenous) and intrinsic (endogenous) causes of SAC (Koenig, 1999; Beale et al., 2010). The exogenous factors can easily be inserted in statistical models as environmental covariates.

However, endogenous factors are difficult to account for because their quantification is limited by data availability (Dormann, 2007). In practice, these two causes of SAC are expected to be found simultaneously (Diniz-Filho et al., 2003). Thus, to correctly account for these SAC sources, decisions should involve the use of appropriate spatial modeling approaches.

### **2.1.2. Importance of accounting for SAC**

Environmental and ecological processes are structured over space. Thus, understanding the contribution of the pure spatial component in these processes structure can be beneficial. Miller et al. (2007) stated that different scales of vegetation spatial distribution have the potential to explain the mutual environmental patterns and processes as well as help in large-scale biodiversity assessment and ecosystem management. For instance, biological processes operating in a spatially patterned environment generate the structured spatial distribution of species. Hawkins (2012) emphasized that any broad-scale samples or representation of nature have a spatially structured distribution, and he pointed that if spatial structure is not observed, it means that valuable information is missing to reveal key spatial patterns. In other words, everything is related to everything else in nature, thus nature is spatially autocorrelated. Therefore, it is of our interest as biogeographers, ecologists, geomorphologists, and hydrologists to understand the influence of spatial autocorrelation (SAC) as well as try to explain how it occurs and changes in the environment considering distinct scenarios in terms of scale and sampling design.

Ecological and environmental studies suggested that ignoring SAC in statistical analysis can violate the independence assumption of traditional regression approaches because of autocorrelated residuals (Cliff & Ord, 1968, 1972; Dormann, 2007; Peres-Neto & Legendre, 2010; Václavik et al., 2012; Kim, 2013). SAC would not cause inconsistency in analyses in cases where (1) the causes of spatial pattern in the dependent variable are fully explained by the measured independent variables, or when (2) the causes of SAC in a dependent variable does not exist (Beale et al., 2010). In other words, since sources of SAC are associated with environmental and ecological variables, in some cases, SAC may be fully explained if these variables are sufficiently taken in consideration. Also, if the variable of modeling interest has no spatial structure characteristic, then causes of SAC may not exist. However, these two conditions are never met simultaneously, thus errors are expected to be autocorrelated violating the basic statistics assumption. This lack of independence can lead to difficulties in the hypothesis testing causing inflation of the Type I error, when the null hypothesis is rejected while it is true (Lennon, 2000; Dormann et al., 2007; Peres-Neto & Legendre, 2010; Miller, 2012). Additionally, SAC can inflate the significance of measured relationships as well as bias the model parameters when non-spatial techniques are used to model spatially structured data (Anselin, 2002; Bini et al., 2009; Václavik et al., 2012). Lennon (2000) called this issue as the red shift problem. He argued that, with non-spatial models, these spatially dependent predictors have their magnitudes inflated and errors underestimated. In other words, ignoring SAC can lead to bias in predictions and in the interpretation of patterns. Therefore, incorporating SAC is important to clarifying the effect of explanatory variables and improving inferences.

Considering SAC will also depend upon whether spatial dependence exists in the chosen scale and sample interval in the data acquisition process. Large plot sizes reduce the ability to discriminate finer spatial variation, while small plot sizes may not represent the area adequately. Additionally, densely collected samples (i.e., narrow intervals) can present strong SAC because different processes may not be identified within a small distance while a wide sampling interval may not detect spatial dependence at all (Bellehumeur & Legendre, 1998; Miller et al., 2007). An approach to solve this scale issue is to choose an appropriate sampling unit size for a predictor variable considering the ecological scale that the response variable operates. Data availability can also limit the potential to detect SAC. For example, one of the factors that result in spatial dependence in species distribution is dispersal, but data and the knowledge to estimate dispersal processes may be unavailable (Václavik et al., 2012). In this case, evaluating SAC at various scales may be crucial to unveil these dynamic processes.

SAC is often seen as a problem or, sometimes, is totally ignored (Lennon, 2000; Hawkins, 2012). For example, studies used subsampling strategies to eliminate the effect of SAC in model outcomes (e.g. Barringer et al., 1990). However, SAC is not an issue, it is what we need to understand. Griffith (1987) explained that natural phenomenon can generate patterns, and, as a researcher, understanding these patterns may answer important concerns. He gave an example of land parcels where the spreading of fertilizer was also a result of rainfall events. He claimed that the rainfall in those lands have created a spillover effect, and this effect if understood can explain crops distribution, correct for bias in modeling, and uncover independent elements responsible for the creation of spatial patterns. Furthermore, in an ecological study, Chase and Knight (2013) argued that spatial-

scale dependent approaches may help addressing questions about the most important drivers of patterns in biodiversity. Therefore, understanding how nature is structured over space (i.e., spatial autocorrelation) is important. It can help unveiling the quality and quantity of information on spatial data as well as answer complex questions about driving processes in the environment.

Overall, answering questions about drivers of processes in nature can help in the management of natural resources and in the development of solutions to problems such as climate change and human activity impacts. Thus, SAC may be a key tool to examine these processes at multiple scales and improve spatial prediction of variables that play a vital role in structuring spatial patterns.

#### **2.1.4. Methods to measure and account for SAC**

##### **2.1.4.1 Measuring SAC**

Cliff and Ord (1968) published a study that proposed to derive SAC from statistics. They used statistical approaches from Moran (1950) and Geary (1954). Moran's  $I$  and Geary's  $c$  are SAC coefficients useful to estimate the spatial intensity and scale of adjacent or noncontiguous sampling units (Fortin et al., 2002). Moran's  $I$  computes the degree of correlation between neighboring values of a variable, and can be comparable to the Pearson's correlation coefficient. Moran's  $I$  varies from -1 (perfect negative autocorrelation) to 1 (perfect positive autocorrelation), where 0 represents the absence of SAC (Fortin et al., 2002; Diniz-Filho et al., 2003; Dormann, 2007; O'sullivan & Unwin, 2014). Moran's  $I$  is computed as:

$$I = \frac{n}{\sum_{i=1}^n (X_i - \bar{X})^2} \frac{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n \sum_{j=1}^n W_{ij}},$$

where,  $X_i$  and  $X_j$  refer to the value of a certain variable at the location  $i$  and location  $j$ , respectively.  $\bar{X}$  is the overall mean of the variable, and  $W_{ij}$  is the spatial weight matrix. Geary's  $c$  measures the degree of difference between the values of a variable in neighboring locations. Geary's  $c$  varies from 0, which indicates perfect positive autocorrelation, to around 2 for negative autocorrelation. When no SAC is detected, Geary's  $c$  is 1. We calculate Geary's  $c$  using:

$$c = \frac{\frac{\sum \sum w_{ij}(d)(x_i - x_j)^2}{W(d)}}{\frac{\sum (x_i - \bar{x})^2}{(n - 1)}},$$

where  $w_{ij}$  is the spatial weight matrix,  $d$  is the distance class,  $W(d)$  is the sum of  $w_{ij}(d)$ . As in Moran's  $I$ ,  $x_i$  and  $x_j$  indicates the value of a certain variable at the location  $i$  and  $j$ , respectively. Users of these techniques must pay attention to normalize the data set of the analyses because Moran's  $I$  and Geary's  $c$  are sensitive to extreme values as well as asymmetric data distribution (Legendre & Legendre, 1998). Overall, both measurements are used to the entire study area and produce global values (Fortin et al., 2002). However, Moran's  $I$  is the most used metric for SAC because of its direct comparison with Pearson's correlation coefficient, while the same comparison with Geary's  $c$  values would require a transformation. Therefore, they are considered global statistics and are useful when high spatial autocorrelation is possible to occur as well as when few samples are available to describe an area with distinct spatial settings.

Global statistics are not appropriate to investigate local spatial association. Thus, local statistics were developed to measure the spatial dependence in a portion or geographic subset of the study area, where the general pattern does not hold. Where the investigation of local spatial dependence can potentially reveal interesting findings, *local index of spatial association* (LISA) can be used (Anselin, 1995; Longley & Batty, 1996). Local statistics of spatial dependence can identify hot spots, cases of non-stationarity and heterogeneous data (Fotheringham et al., 2002). Anselin (1995) illustrated how global autocorrelation statistics such as Moran's I and Geary's c can also be a case of LISA and explained four more metrics for local spatial association measurement. LISA is a statistic of the form:

$$\Gamma = \sum_j w_{ij} y_{ij} ,$$

where the  $w_{ij}$  and  $y_{ij}$  are elements of matrices W and Y, and the focus is on the value of  $\Gamma$  at location i. W is the spatial association between site i and other sites j, while Y is the association of values of a random variable at the site i with values at other sites. The Y matrix is the one from which other local statistics are formed. For local Moran's statistic ( $I_i$ ), which is based on covariance, the  $y_{ij}$  are of the form of  $(x_i - \bar{x})(x_j - \bar{x})$ . For Geary's local statistics ( $c_i, K_{1i}, K_{2i}$ ), the  $y_{ij}$  has a difference structure form like  $(x_i - x_j)^2$ . For local Getis-Ord ( $G_i$  and  $G_i^*$ ), the  $y_{ij}$  takes the form of either  $x_j$  or  $(x_i + x_j)$ . These evaluate spatial association by comparing local weighted averages to global averages for 'hot spots' checking.

Bocard et al (1992) and Legendre (1993) also worked to quantify SAC. They separated the variation of the response variables into four parts: 1) unexplained variation, which represents the model error; 2) explained environmental variation (e.g. climate and



topography); 3) variation explained by both spatial and environmental variables; and 4) variation explained only by spatial structure. To identify each part usually partial regression analysis is performed (Legendre and Legendre 1998; Miller, 2007).

In sum, these metrics are useful to quantify the degree of SAC possessed by any spatial structured variables, which can be a response or an explanatory variable in the modeling procedure.

#### **2.1.4.2 Accounting for SAC**

Here, I review the main methods used to account for SAC in different scales. To overcome limitations of non-spatial techniques different modeling approaches have been used in the literature. In sum, these methods are applied according to the scale of interest in the study and, determined by sampling units and design (Miller et al., 2007; Franklin, 2010). Common methods used in ecological, species distribution, soil-landform, and hydrological modeling are regression kriging, simultaneous autoregressive modeling, conditional autoregressive modeling, spatial lag modeling, spatial error modeling, spatial eigenvector mapping, and geographically weighted regression (GWR) (Griffith, 2000; Lichstein et al., 2002; Hengl et al., 2004; Griffith & Peres-Neto, 2006; Ver Hoef et al., 2006; de Marco et al., 2008; Kissling & Carl, 2008; Bini et al., 2009; Miller, 2012; Václavík et al., 2012; Kim, 2013; Isaak et al., 2014).

Autoregressive models (AR) are linear regression models with an additional term that incorporates SAC (Anselin, 2003; Kissling & Carl, 2008). This additional term has a spatial weights matrix that requires the distance between neighbors of each location and

the weight of each neighbor, where closer neighbors receive higher weights. This information serves as input to calculate the spatial dependence of a location and produce a variance-covariance matrix (Anselin, 1988; Cressie, 1993; Kissling & Carl, 2008). These approaches can describe fine-scale spatial patterns that are associated with local factors such as dispersal, disturbance, and competition (Lichstein et al., 2002; Miller et al., 2007). AR models can be defined as:

$$y = \alpha + \rho Wy + \varepsilon$$

where  $\alpha$  is the constant,  $\rho$  is the autoregressive coefficient,  $Wy$  the spatial lag for variable  $y$  where  $W$  is the neighborhood based on distance or other topology, and  $\varepsilon$  is the error term (Anselin, 1993). This model is known as spatial lag model and it assumes that the autoregressive process occurs in the response variable  $y$ , also called inherent SAC (Anselin, 1988; Kissling & Carl, 2008). This model can be generalized with the addition of other predictor variables:

$$y = \alpha + \rho Wy + X\beta + \varepsilon$$

where  $X\beta$  are the predictor variables and coefficients. This addition can improve the predictive ability of the associated model. For cases where the spatial autoregressive process occurs in the error term, which can happen when the explanatory variables do not fully explain SAC, it is advised to use the spatial error model (Haining, 2003; Kissling & Carl, 2008). This model is represented as:

$$y = \alpha + X\beta + \lambda W\mu + \varepsilon$$

where  $\lambda$  is the autoregressive coefficient,  $W\mu$  the spatial matrix for the  $\mu$  which is the spatially dependent error term, and  $\varepsilon$  is the spatially independent error (Dormann et al., 2007). Kissling & Carl (2008) compared a non-spatial model (ordinary least square (OLS)) to three autoregressive spatial approaches (spatial lag (SARlag), spatial error (SARerr), and mixed (lagged and error together; SARmix) in species distribution data. They found that SARerr in all cases is more reliable and that OLS, SARlag and SARmix can perform poorly in terms of type I error and present unpredictable biases in parameter estimates.

Geostatistical methods such as regression kriging, one of the most widely used interpolation techniques (Cressie, 1993), model SAC explicitly through a variogram separating it from the deterministic variation and noise. These methods consider a stationary environment where the mean and the variance are constant over space (Anselin, 2002; Legendre & Fortin, 1989). Assuming stationarity, these methods describe the spatial structure through a variogram as a function of distance (Fortin et al., 2002). Distance, in kriging methods, is the most important predictor, except for co-kriging and universal kriging that allow the addition of a second variable. Co-kriging adds co-variation information between two variables to model one of them (Miller et al., 2007). The quality and amount of sample data may affect the ability of these techniques to detect local information or local spatial dependence. Therefore, geostatistical methods are more adequate to model broad-scale SAC which is compared to the true gradient defined by Legendre and Legendre (1998), where the environmental gradient coincides with the geographical gradient.

Another approach to explicitly account for broad-scale SAC that has been widely used in ecological and soil modeling is trend surface analysis (TSA; Legendre & Legendre,

2012). Lichstein et al. (2002) emphasized that while autoregressive models account for fine-scale variations, models based on trend surface polynomials account for broad-scale spatial patterns. TSA is a polynomial technique that, based on a plane, determines the broad trend in the spatial data with the objective of minimizing the error between the interpolated value at a known location and the original value (Lichstein et al., 2002; Kim, 2013). Comparing a non-spatial model with TSA model outcomes, Kim (2013) explained that TSA performed better than the non-spatial approach. TSA presented lower AIC (Akaike information criterion) and greater  $R^2$  values compared to the non-spatial method outcomes. However, compared to other spatial models such as spatial eigenvector mapping (SEVM), TSA did not have the best performance.

It is hard to separate processes that occur only in one scale in nature, thus methods that have the potential to explain multiple scale SAC are necessary. Eigenfunction spatial filtering has been introduced to deal with the multiple scale SAC. This approach is a nonparametric technique that accounts for the inherent SAC in spatial models by introducing appropriate variates called spatial filters (Legendre & Legendre, 2012). It is considered an alternative methodology to account for a specific type of SAC originated from missing variables that are spatially correlated (Getis & Griffith 2002; Griffith & Peres-Neto, 2006; Fischer & Griffith, 2008; Kim, 2013). Griffith (2000) emphasized that the eigenvector function can handle very well the conversion from spatially autocorrelated to spatially non-autocorrelated data. He also pointed it as a solution to the difficulties that autoregressive models face to deal with normalizing factors. Václavík et al. (2012) explained the issues of predicting invasive species distribution and showed how models can be improved by accounting for SAC at multiple scales. The authors used four models,

one that ignores SAC and three that incorporates SAC at different scales. They used TSA (trend surface analysis) to account for broad-scale SAC, autocovariates for local-scale, and SEVM (spatial eigenvector mapping), a filtering technique to account for SAC at multiple scales. In the results, the authors argued that accounting for SAC at multiple scales can improve our understanding of dynamic processes that drive the distribution of invasive species as well as the predictive performance of statistical techniques.

There are methods to account for SAC in a variety of fields. The manifestation of SAC can be differentiated (sometimes superficially) by the scale in which its influence is observed (Franklin, 2010; Václavik et al., 2012). The previous methods (AR, TSA, Kriging, and SEVM) describe consistently the relationships throughout a region of interest, thus representing global parameter estimates (Fortin et al., 2002; Miller et al., 2007). Thus, the decision among statistical methods that investigate SAC effects will depend on the scale of an observed phenomenon, and the patterns that the researcher is looking for to understand or explain.

## **2.2. Water Quality**

### **2.2.1. Importance of water quality**

Streams and rivers represent a considerable part of Earth's biodiversity and are responsible for crucial ecosystem services that are beneficial to the human population (e.g. drinking water, irrigation use, industrial purposes). Natural and anthropogenic factors can impact river systems. Calow (1992) notes that natural factors include soils, atmospheric precipitation, slope, elevation, vegetation cover, and river discharge while anthropogenic aspects include urbanization, agricultural practices, and deforestation. Praskievicz & Chang (2009) explain that hydrological responses are affected by processes such as urbanization because it leads to changes in the magnitude of peak flow during rainfall events (i.e., as the impervious surface increases, the entire water balance of the watershed is altered). These factors can directly or indirectly impact the hydrological, biological, and chemical processes of aquatic ecosystems as well as degrade water quality conditions (Pratt & Chang, 2012). Therefore, assessing the condition of aquatic resources has become one of the major concerns worldwide and one of the most important areas of interdisciplinary environmental research.

### **2.2.2. SAC and water quality**

Analyses of water quality are complicated by various sources of SAC in hydrological data. As previously mentioned, SAC is the likelihood of the value of a variable in one location to be similar to the measurement of the same variable in a neighboring location. Closer samples tend to be similar, thus resulting in positive spatial

autocorrelation, while far samples are usually different, representing cases of negative autocorrelation. Isaak et al. (2014) explained that positive autocorrelation in water quality may result from local habitat similarities or turbulent stream flows. However, even though cases of negative spatial autocorrelation are rare in ecological data (Dale & Fortin 2002; Beale et al., 2010), negative autocorrelation in water quality may still occur as a result of a too wide sampling interval for a specific variable within a watershed, as well as local existent structures or ecosystems (Pringle, 2001). SAC originates from exogenous (e.g. topography) and endogenous (e.g. dispersal) factors (Miller, 2012). Nonetheless, this will all depend on the scale of interest. Stream water quality patterns can present spatial homogeneity at different scales because of sink-source relationships (Valett et al., 2008). However, it can also be heterogeneous at both fine and broad scales because of channel and catchment characteristics (Pringle et al., 1988; Cooper et al., 1997).

SAC is particularly important for water quality modeling because water quality conditions are influenced by human factors (e.g. land use practices), and natural factors, such as topography, climate, and hydrological processes (Pringle, 2001; Chang et al., 2012). Some hydrological processes that influence the nearby water quality samples are rainfall intensity and channel characteristics. Also, it is well-known that physical and biological processes such as metapopulation dynamics and disturbance regimes occur in the catchment area and influence the characteristics of streams and rivers (e.g. network structure, connectivity, stream-flow direction) (Johnson & Gage, 1997; Peterson et al., 2013). These hydrological, physical, and biological processes in conjunction with the resulting stream characteristics typically create spatially structured patterns that should not be ignored. For example, the physical structure of a stream may serve as an ecological

corridor to an organism or a material, but the efficiency of this corridor will also depend on the processes that involve the organism or material in observation (Peterson et al., 2013). Accounting for SAC in water quality studies may reveal these complex spatial patterns and help water quality experts to understand what drives main changes in hydrological systems, thus enhancing water quality management laws and practices.

### **2.2.3. Spatial approaches to water quality modeling**

Hydrological modeling is an important tool in the investigation of processes that drive changes in water resources. Modeling water quality is difficult because water quality conditions significantly depend on complex characteristics, such as basin hydrology and vegetation dynamics, that would require their own models (Praskievicz & Chang, 2009). These characteristics can be considered the causes of SAC and, thus, ignoring SAC may prevent researchers from acquiring valuable information about stream attributes and decrease the accuracy and validity of statistical inferences.

To account for the sources of SAC existent in hydrological ecosystems, spatial techniques are necessary. Several studies in hydrology have attempted to consider SAC to understand the influence and the selection of scale and key predictors such as land use and climate in water quality. Therefore, models that explicitly accounts for SAC such as autoregressive regression (AR), geostatistical approaches, and spatial filtering techniques are necessary and have been used to investigate the causes and patterns in water quality studies. For instance, Cooper et al. (1997) pointed out that geomorphological characteristics resultant from meanders or pool-riffle spacing can be associated with spatial



relationships in streams, and to detect these spatial patterns spatial techniques such as autoregressive functions may be efficient.

Vrebos et al. (2017) attempted to account for SAC to examine if spatial processes have significant impacts on predicting water quality trends. As river systems are hierarchically organized, and with directional nature, Vrebos et al. (2017) applied a spatial filtering technique (AEM - Asymmetric Eigenvector Mapping) to evaluate the influence of land use and spatial scales in water quality changes from up to downstream within a catchment in Belgium. They compared this technique with MEM (Moran's Eigenvector Mapping), which performed better than AEM. Using MEM, they found that land use and a variety of spatial predictors of different scales were significantly impacting the water quality conditions in the region. They also pointed that human activities affected the entire chemistry balance supporting the complex characteristics of the catchment. Therefore, even though they did not identify unidirectional changes of water chemistry in the selected catchment, meaning that different directions can be affecting the trends of water quality, spatial structure proved to be significant.

Huang et al. (2014) examined the effects of natural and anthropogenic factors in the spatial variation of water quality conditions within a coastal watershed. To choose the best model that can identify significant explanatory variables for each water quality variable (response variable), they compared model outcomes ( $R^2$ , AIC and Moran's  $I$  values) of a non-spatial technique (OLS) to two spatial approaches (i.e., spatial lag and error models). They found that the spatial techniques had greater  $R^2$  results, and lower AIC values compared to OLS. Huang et al. (2014) also pointed that the spatial error model presented a slightly better performance than the spatial lag technique.

Pratt & Chang (2012) compared OLS and GWR model outcomes to observe the relationship between land cover and stream water quality, considering scale and seasonality. They concluded that scale and seasonality can impact model results. Additionally, they pointed out GWR presented greater predictive power and account for more local water quality sources of variation than OLS.

Chang (2008) applied autoregressive regression (both spatial error and lag models) to understand the complex relationships between landscape and water quality, addressing spatial and temporal trends, as well as anthropogenic and scale effects. The results showed different trends for each water quality variable. Land cover was an important predictor in explaining the spatial and temporal variation in water quality. Spatial models explained the variation of water quality better than the OLS model. Overall, Chang addressed that to understand the complex and dynamic behavior of water quality variables, the integration of landscape analysis and spatially intensive monitoring is of vital importance.

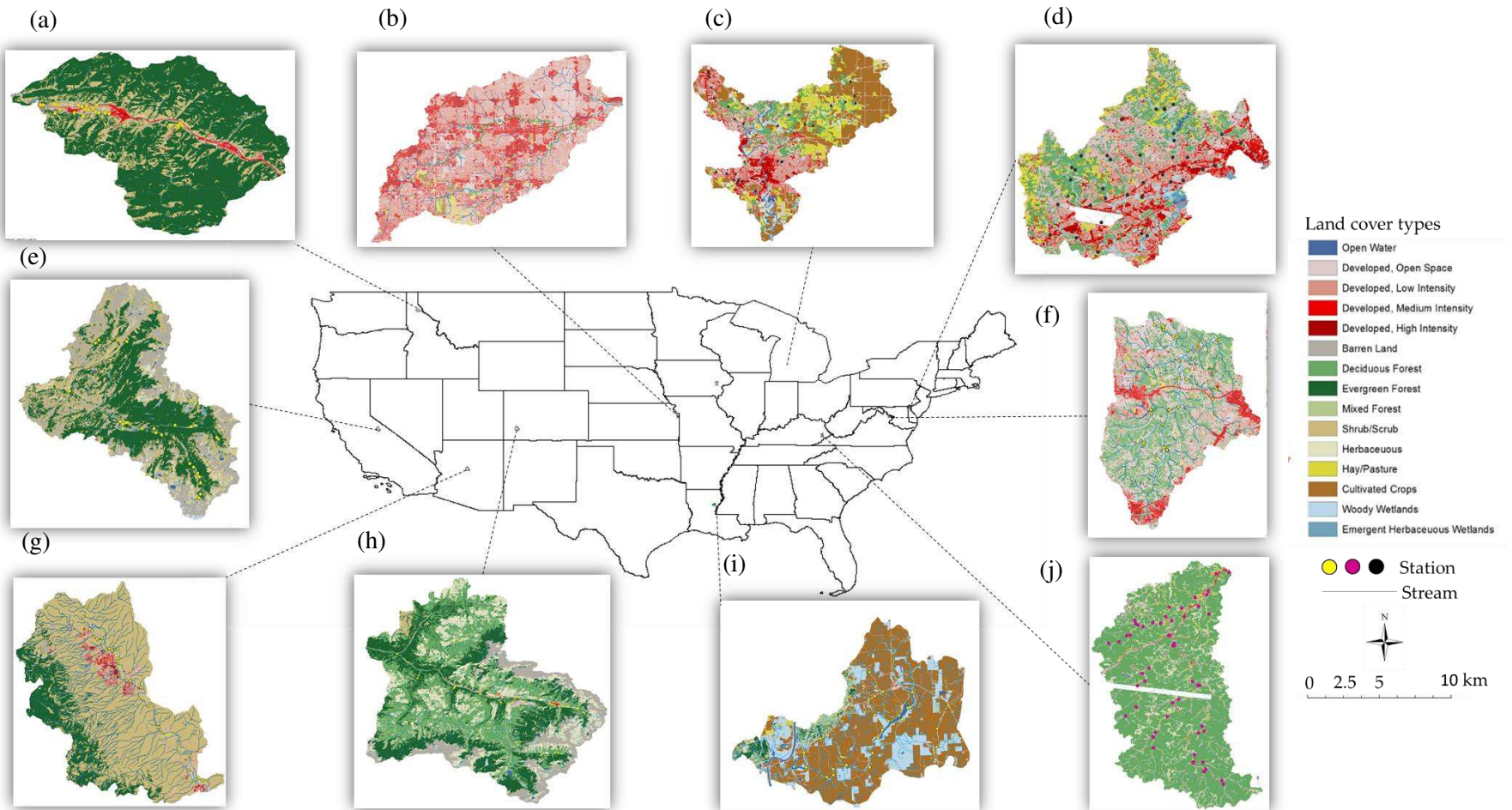
## **CHAPTER 3: METHODOLOGY<sup>1</sup>**

### **3.1. Study Areas**

The study areas are basins located in 10 states of the USA. I analyzed water quality variables in watershed and sub-watershed segments in Arizona (AZ), California (CA), Colorado (CO), Delaware (DE), Idaho (ID), Iowa (IA), Kansas (KS), Kentucky (KY), Louisiana (LA), and Virginia (VA). The basins were delineated by the U.S. Geological Survey (USGS), which states that as per the fifth and sixth levels of classification, these basins are smaller scale hydrologic units. Overall, their areas ranged from 150 to 764 km<sup>2</sup>. The climate and geology of the regions vary significantly due to their differences in latitude, longitude, and altitude. Tables 1 and 2 briefly present the climatological and geological characteristics of each state, and specific site characteristics in terms of area and water quality variables, respectively. Figure 2 illustrates the watershed shapes and their land cover characteristics.

---

<sup>1</sup> The main content of this section has been published in Miralha & Kim (2018).



**Figure 2.** Land cover characteristics of each state and watershed shape. Idaho (a); Kansas (b); Iowa (c); Delaware (d); California (e); Virginia (f); Arizona (g); Colorado (h); Louisiana (i); and Kentucky (j). To better visualize the water quality stations spatial organization, refer to Appendix A.

**Table 1.** Description of the ten study sites investigated in this research.

Region	Coordinates <sup>a</sup>	Land Cover	Biogeographic Region <sup>b</sup>	Geology	Climate <sup>b</sup>	Soil <sup>c</sup>	Surficial Lithology <sup>b</sup>
Arizona	34°40'54" N, 112°00'47" W	Herbaceous, low-intensity urbanization, and evergreen forest	North American Warm Desert	Late and middle Pleistocene surficial deposits and Pliocene to middle Miocene deposits	Cold semi-arid (BSk)	Alfisols/ Inceptisols	Non-Carbonate and Silicic Residual Material; Alluvium and Fine-textured Coastal Zone Sediment
California	38°00'00" N, 119°21'33" W	Evergreen Forest, Barren Land, and Shrubs	Mediterranean California	Mesozoic granitic rocks, unit 3 (Sierra Nevada, Death Valley area, Northern Mojave Desert, and Transverse Ranges)	Temperate Mediterranean (Csb)	Rock outcrop/ Entisols	Silicic Residual Material
Colorado	37°56'58" N, 107°56'10" W	Predominantly Evergreen and Deciduous Forest	Rocky Mountain	Mancos Shale; Pre-ash-flow andesitic lavas, breccias, tuffs, and conglomerates; Morrison, Wanakah, and Entrada Fms	Warm-summer humid continental (Dfb)	Rock outcrop/ Mollisols	Non-Carbonate and Silicic Residual Material
Delaware	39°43'36" N, 75°40'07" W	High-, medium-, and low-intensity urbanization with some deciduous forest and pasture	Gulf and Atlantic Coastal Plain	Wissahickon Schist	Humid Subtropical (Cfa)	Ultisols	Non-Carbonate and Silicic Residual Material; Alluvium and Fine-textured Coastal Zone Sediment
Idaho	47°31'01" N, 116°04'27" W	Evergreen forest, shrub, and some medium-intensity urbanization	Rocky Mountain	Siltite, argillite, dolostone, and quartzite; Middle Proterozoic Wallace Formation	Temperate Mediterranean (Csb)/Warm, dry-summer continental (Dsb)	Andisols	Non-Carbonate Residual Material
Iowa	41°37'38" N, 91°29'31" W	High and medium urbanization level with crops and pasture	Eastern Great Plains	Cedar Valley Limestone	Humid Continental (Dfa)	Mollisols	Glacial Till, Loamy; Glacial Outwash and Glacial Lake Sediment, Coarse-textured; Alluvium and Fine-textured Coastal Zone Sediment
Kansas	38°55'00" N, 94°41'14" W	Predominantly high-, medium-, and low-intensity urbanization	Eastern Great Plains	Limestone—Kansas City and Lansing Group	Humid Subtropical (Cfa)	Mollisols	Non-Carbonate Residual Material
Kentucky	37°25'01" N, 82°49'04" W	Predominantly Deciduous Forest	Central Interior and Appalachian	Middle part of Breathitt Group	Humid Subtropical (Cfa)	Inceptisols	Colluvial Sediment
Louisiana	31°48'17" N, 91°42'21" W	Predominantly cultivated crops	Gulf and Atlantic Coastal Plain	Sub/supra-glacial sediment	Humid Subtropical (Cfa)	Vertisols	Alluvium and Fine-textured Coastal Zone Sediment
Virginia	38°55'51" N, 77°18'25" W	Deciduous Forest and developed open space	Central Interior and Appalachian	Schist	Humid Subtropical (Cfa)	Alfisols/ Inceptisols	Non-Carbonate Residual Material

\*Study site names are given in the next table.

<sup>a</sup> The coordinates indicate the central point of the watershed in study.

<sup>b</sup> Biogeographic regions, Climate, and Lithology are according to Sayre, 1984.

<sup>c</sup> Soil information is according to US soil taxonomy at soil order level.

**Table 2.** Study areas (10 watersheds each in one state of the USA and their areas), number of stations per study area, and water quality parameters (response variables) with the respective Moran's *I* values in parentheses.

State	Study Areas									
	LA	AZ	KS	VA	CA	CO	DE	ID	IA	KY
Watershed	Bayou Louis/ Lake Louis	Cherry Creek	Indian Creek	Difficult River	Headwaters Tuolumne River	Upper San Miguel River	Clay, Mill, Bradywine Creek, and Cristina River	Lower South Fork Coeur d'Alene River	Iowa River	Beaver Creek
Area (km <sup>2</sup> )	288.58	586.26	193.8	150.84	553.66	763.71	352.24	308.49	193.96	407.07
Stations	29	31	33	33	31	32	36	32	32	54
Water quality parameter (Moran's <i>I</i> )	pH (0.13) T (0.15) SC (0.20) DO (0.28) TDS (0.53)	DO (-0.08) * pH (-0.07) * T (0.54) SC (0.59)	TN (0.013) SC (0.022) DIN (0.07) KjN (0.10) TP (0.15) T (0.20) Tur 0.25) DO (0.44) pH (0.72)	Tur (-0.28) * TDS (-0.26) * SC (0.06) Br (0.09) Cl (0.12) Mg (0.15) Na (0.15) DO (0.16) Ca (0.17) SiO <sub>2</sub> (0.19) Fe (0.21) K (0.25) CO <sub>2</sub> (0.34) Mn (0.34) pH (0.39) Alk (0.40) TP (0.42) SO <sub>4</sub> <sup>2-</sup> (0.45) F (0.54) T (0.69)	Csu (-0.20) * T (0.30) Mg (0.42) K (0.46) Ca (0.55) Cl (0.58) Na (0.59) SiO <sub>2</sub> (0.62) SO <sub>4</sub> <sup>2-</sup> (0.65) TDS (0.73) Alk (0.80) pH (0.82)	DO (0.39) SC (0.36) pH (0.37) T (0.67)	SC (-0.05) * T (-0.006) * Chla (0.02) TN (0.03) Nin (0.05) Alk (0.08) TP (0.12) DO (0.15) pH (0.16) Cl (0.23) TOC (0.32) DOC (0.32)	Pb (0.11) T (0.15) Zn (0.24) pH (0.31) Cd (0.35) As (0.47) SC (0.56)	DO (0.18) pH (0.34) NO <sub>3</sub> <sup>-</sup> (0.36) T (0.49) PO <sub>4</sub> <sup>3-</sup> (0.66) Cl (0.67)	Al (0.005) Ba (0.06) Alk (0.11) Na (0.14) Cl (0.23) K (0.26) Nin (0.29) TDS (0.32) SO <sub>4</sub> <sup>2-</sup> (0.38) Fe (0.40) KjN (0.43) Mg (0.47) Ca (0.55) Mn (0.58)

\* Moran's *I* values treated as absolute values. Note: Specific conductance (SC), dissolved oxygen (DO), total dissolved solids (TDS), total nitrogen (TN), dissolved nitrogen (DIN), total ammonia plus organic nitrogen (also known as Kjeldahl nitrogen, KjN), total phosphorus (TP), turbidity (Tur), alkalinity (Alk), suspended carbon (Csu), chlorophyll (Chla), inorganic nitrogen (Nin), total organic carbon (TOC), dissolved organic carbon (DOC), dissolved lead (Pb), dissolved zinc (Zn), dissolved cadmium (Cd), and dissolved arsenic (As).

### **3.2. Dependent Variables**

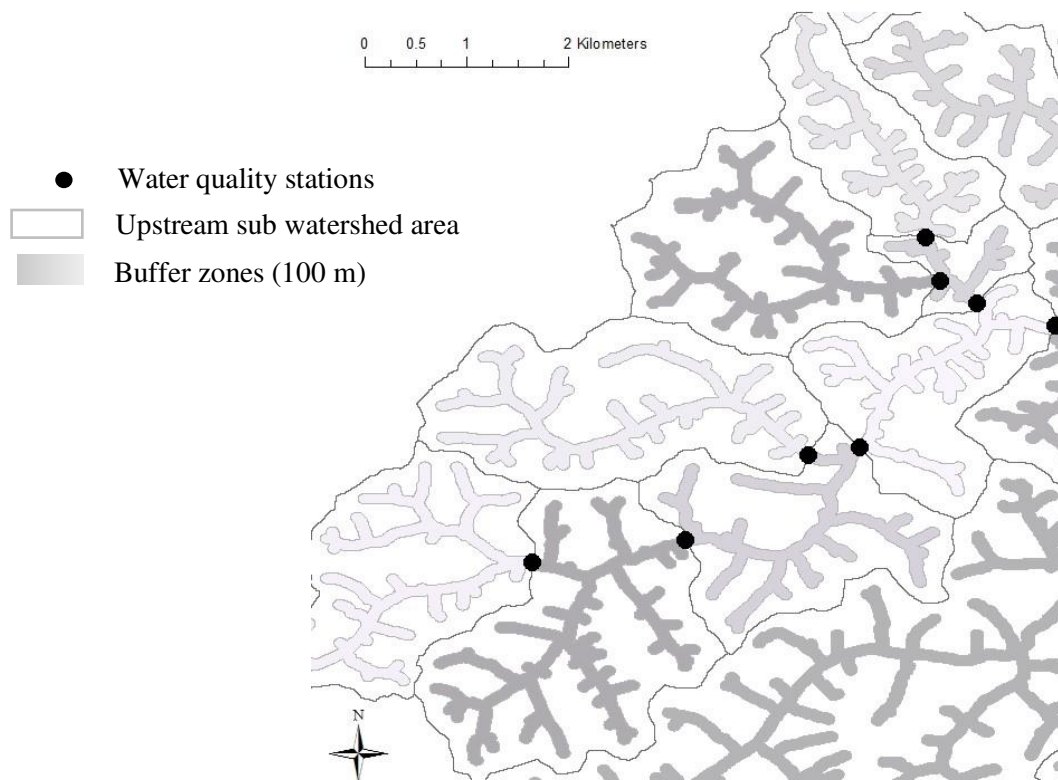
Water quality data from 2011 to 2017 were obtained online from the national Water Quality Portal (WQP) (NWQMC, accessed throughout the year 2017). The WQP integrates publicly available water quality data from three very important and widely used sources for research in the US: the USGS National Water Information System (NWIS), the EPA STORage and RETrieval (STORET) Data Warehouse, and the United States Department of Agriculture (USDA) Sustaining the Earth's Watersheds Agricultural Research Data System (STEWARDS) through the Water Quality eXchange (WQX).

Based on the data availability and site locations, 29–54 sampling stations were selected from each study watershed (Table 2). Accounting for the temporal variability in each watershed, the data were selected within the same week, month, or season. Therefore, no seasonality effect was considered in this study. Because I collected water quality data from different sources as explained above, the number and type of variables varied across watersheds (Table 2). These water quality variables were treated as dependent variables in this research.

### **3.3. Delineation of Upstream Area**

Characteristics of the sub-watershed area upstream of sampling stations should affect water quality variables downstream (Chang, 2008). Thus, sub-watershed boundaries were delimited using the 'ArcHydro' package tool of ArcGIS 10.3 (Environmental Systems Research Institute, Redlands, CA, USA). I downloaded spatial stream data from the 2016 US Geological Survey (USGS) National Hydrography Dataset (USGS, accessed

on 2017). The distance between stations varied, as did the size of each upstream area delineated. Land use characteristics as well as topography and soil far from the stream channel might contribute less to changes in water quality across space (Pratt & Chang, 2012). Thus, I used the upstream area to separate the stream network specific to each station and delineated the riparian zone around the stream. Many studies have conducted analyses at the riparian area scale, mainly by considering a buffer zone on each side of the stream. Overall, there was no specific buffering distance recommended (Chang, 2008; Li et al., 2009; Pratt & Chang, 2012). In this study, I used a buffer zone of 50 m each side of the stream (i.e., a 100 m buffer in total) as the area that can contribute the maximum to water quality changes (Figure 3). I performed these analyses for all watersheds in this study.



**Figure 3.** Upstream area delineation example of the Beaver Creek watershed in Kentucky, and their respective buffer zones in tones of gray. The solid circles are water quality stations (sites).



### **3.4. Independent Variables**

Using the buffer zones of the upstream area, I extracted the land use, topography, and soil types associated with each sampling station. These variables were treated as independent variables in the subsequent modeling. The summary of these variables is shown in Table 3. I downloaded the land use raster with 30 m resolution from USGS The National Map—2011 National Land Cover Database (USGS TNM-NLCD) (USGS, accessed in 2017). In this study, I considered the percentage of four major land use types surrounding stream networks: urban, agriculture, forest, and wetland. To extract this information, I used the ‘Zonal Statistics’ toolset in ArcGIS 10.3. The percentage of urban area in each upstream buffer zone was calculated using the sum of the low-, medium-, and high-intensity urbanization, and open space values in the land use raster. The sum of the values for pasture and cultivated crop was used to calculate the percentage of agricultural land in the area. The values for deciduous forest, evergreen forest, and mixed forest were used to estimate the percentage of forest, while the values for woody wetlands and emergent herbaceous wetland were combined to calculate wetland percentage. For the topographic variables, I used 10 m resolution digital elevation models (DEMs) downloaded from USGS The National Map Elevation Products (USGS TNM 3DEP) (USGS, accessed in 2017). Using the same upstream area and zonal statistic toolset, I extracted the mean and standard deviation of the elevation and slope respectively for each station’s upstream area. These variables were used to account for topographic complexity.

I downloaded the hydrological soil groups (HSGs) from the Natural Resources Conservation Service’s (2017 NRCS) Soil Survey Geographic (SSURGO) database (Soil

Survey Staff, accessed in 2017). I extracted the percentages of A, B, C, D, A/D, B/D, and C/D categories of soil for each site. The HSGs are categorized by the hydraulic conductivity level of a soil and how much runoff it produces. This is usually associated with the percentage of sediment grain sizes a soil presents. Typically, group A soils have a low runoff capacity because the water transmissivity through the soil profile is very high. Thus, group A soils are composed of a high percentage of sediments with large grain size, such as sand or gravel. Group B soils have a moderate runoff capacity. Nevertheless, water flows freely through the soil profile and the percentage of large-sized grains is high. In this case, however, small grain size sediments such as clay can reach up to 20 percent of the total. Group C soils have a moderately high runoff capacity and have a higher clay percent, with less than 50 percent of sand. Group D soils are characterized as having the highest percentage of fine grains such as clay and silt. The dual HSGs (A/D, B/D, and C/D) are wet soils where water table is within 60 cm below the surface but can still be drained adequately. The first letter indicates well-draining conditions, and the second, represents poorly drained conditions (USDA-NRCS, 2009).

**Table 3.** Data sources and details of dependent and independent variables.

<b>Agency Source</b>	<b>Variable</b>	<b>Year/Data</b>	<b>PC Group</b>	<b>Derived Variable</b>	<b>Original Data</b>
WQP	Dependent	2011 to 2017—Water quality parameters		-	Physical water quality data
USGS	Independent	2017—National Elevation dataset (10 m)	Topographic	Mean elevation	Elevation
				Elevation standard deviation	
				Mean slope	
				Slope standard deviation	
USGS	Independent	2011—National Land Cover dataset (30 m)	Land use	Agriculture	Pasture, cultivated crops
				Forest	Deciduous forest, evergreen forest, mixed forest
				Urban	Low-, medium-, high-intensity urbanized areas, open space
				Wetland	Woody wetland, emergent herbaceous wetland
USDA, NRCS	Independent	2017—Hydrologic Soil Groups	Soil	A, B, C, D, A/D, B/D, C/D	Soil Survey Geographic (SSURGO) database

Note: PC (Principal Component); WQP (Water Quality Portal); USGS (United States Geological Survey); USDA, NRCS (United States Department of Agriculture, Natural Resources Conservation Service).

### **3.5 Data Preprocessing**

I tested the normality of each dependent and independent variable using IBM SPSS Statistics for Windows Version 23.0 (Armonk, NY, USA). In this study, the independent variables are likely to present a high level of correlation due to their nature. For example, agriculture and urban zones are land use types that might express a significant negative relationship because, as the area under agricultural use increases, the urbanized areas will tend to decrease. Thus, to account for the multicollinearity in the subsequent modeling, I applied principal component analysis (PCA). This technique reduces the dimensionality of a multivariate dataset where variables are significantly interrelated. This reduction results in principal components (PCs), which are considered uncorrelated variables (Jolliffe, 1986; Abdi & Williams, 2010). PCA is useful because it simplifies the description of the independent variables and the modeling procedure. I divided the independent variables into three main groups: land use, topography, and soil. Land use considered the percentage of urban, agriculture, wetland, and forest areas. The topographic group encompassed the mean and standard deviation values of slope and elevation. The soil groups represented the percentage of A, B, C, D, A/D, B/D, and C/D soil types (Table 3). Overall, I had three main PC groups used as the predictors in the models. Each variable category presents one to three PCs, depending on how significantly the variables in each group are correlated. This means that a model can have three to nine principal components as independent variables.

### 3.6. Testing for Spatial Autocorrelation (SAC)

Moran's  $I$  is the most used metric to measure SAC in spatial studies because of its similarity to the Pearson's correlation coefficient, which facilitates interpretation. Thus, I quantified the inherent degree of SAC for each water quality variable using Moran's  $I$ . I used the geographic coordinate system based on angular values (longitude and latitude) considering the North American 1983 as the datum for the distance calculation. I did not perform a projection in this study, which would have been a serious issue if I had been concerned with region-scale modeling crossing multiple states. Instead, the current study examined the water quality of several stations within local watersheds (< ca. 764 km<sup>2</sup>). Therefore, using the Euclidean distance is appropriate.

### 3.7. Statistical Models

GeoDa version 1.8 (Chicago, IL, USA) was used to run three models in this paper. First, OLS, representing the non-spatial model, is a multiple linear regression approach (Equation (1)), where the response variable is the water quality variable and the independent variables are the PCs of the topographic, land cover, and soil groups:

$$Y_i = \beta_0 + \beta_1 X_1 + \dots + \beta_i X_i + \varepsilon_i, \quad (1)$$

where  $Y_i$  is the response variable,  $\beta_0$  is the constant in a linear model,  $\beta_i$  are coefficients associated with the independent variables, and  $\varepsilon_i$  is the error term. Notably, the same independent and dependent variables were used as in the spatial modeling approaches.

The second model was a spatial lag model (Equation (2)):

$$Y_i = X_i \beta_i + \rho W Y_j + \varepsilon, \quad (2)$$

where  $Y_i$  and  $Y_j$  are the dependent variables at locations  $i$  and  $j$ , respectively,  $X_i$  is the independent variable at  $i$ ,  $\beta_i$  is the regression coefficient,  $\rho$  is the spatial autoregressive coefficient,  $WY_j$  is the spatially lagged dependent variable, and  $\varepsilon$  is the error term. This model accounts for the fact that the dependent variable is affected by the independent variables in adjacent places, and, thus, the dependent variable is spatially lagged as a predictor. The third model used was the spatial error model (Equation (3)):

$$Y_i = X_i\beta_i + \varepsilon; \varepsilon = \lambda W\varepsilon + \varepsilon, \quad (3)$$

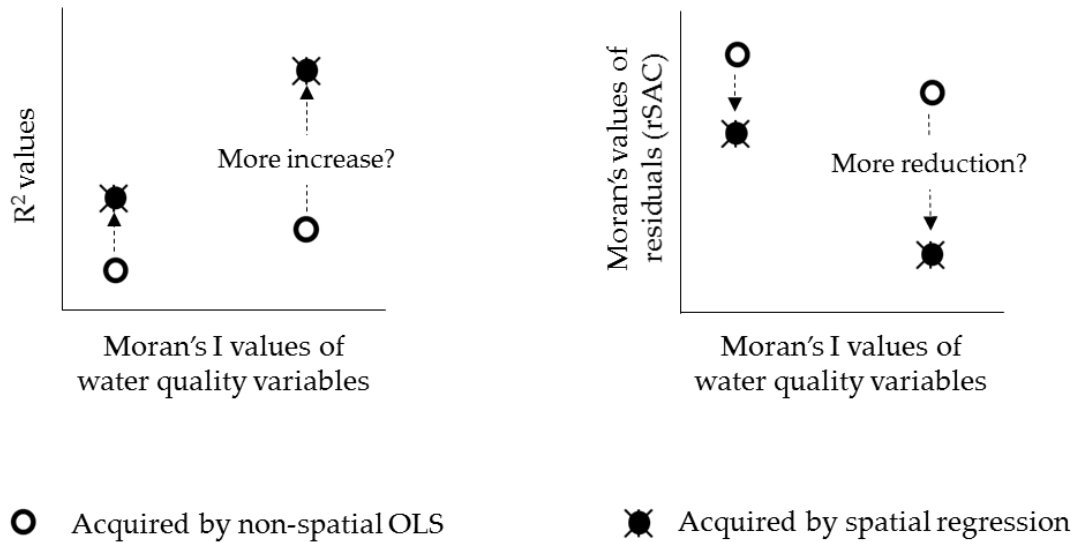
where  $Y_i$  is the dependent variable at location  $i$ ,  $X_i$  is the independent variable,  $\beta_i$  is the regression coefficient,  $\varepsilon$  is the error term,  $\lambda$  is the autoregressive coefficient,  $W\varepsilon$  is the spatially lagged error term, and  $\varepsilon$  is the homoscedastic and independent error term. This model accounts for the error terms that are correlated across different spatial units.

Spatial lag and error models are based on spatial weights matrix construction as presented in the literature (Dorman et al., 2007; Chang, 2008; Kissling & Carl, 2008). Among several methods used to account for SAC, I chose these spatial techniques because of their great flexibility in describing spatial organization in cases where the sampling sites are neighboring points separated by non-equal distance intervals (Anselin, 2002; Dray et al., 2012), which is the case of this study. In addition, these methods are the most common used spatial techniques in a variety of fields, including water quality modeling.

### **3.8. Model Comparison**

After measuring the inherent degree of SAC for each water quality variable, I compared the outcomes of non-spatial OLS and spatial regression approaches in terms of  $R^2$  and rSAC. To quantify rSAC, I estimated Moran's  $I$  for residuals. After the

modeling procedure, I evaluated the hypothesis by plotting Moran's  $I$  values of the water quality variables against the  $R^2$  and rSAC values for each water quality variable (Figure 4). A few water quality variables presented negative inherent SAC values and were treated as positive in this graph. This is because I intended to concentrate on the *magnitude* of SAC.



**Figure 4.** Evaluation of the hypothesis—Moran's  $I$  values of the water quality variables appear on the x-axis, and the model outcomes,  $R^2$  and residual SAC, appear on the y-axis. After spatial regression, water quality variables with a higher amount of spatial autocorrelation (SAC) were hypothesized to exhibit improved hydrologic modeling (i.e., more increases in  $R^2$  and more decreases in residual SAC) than those with lower SAC.

## **CHAPTER 4: RESULTS**

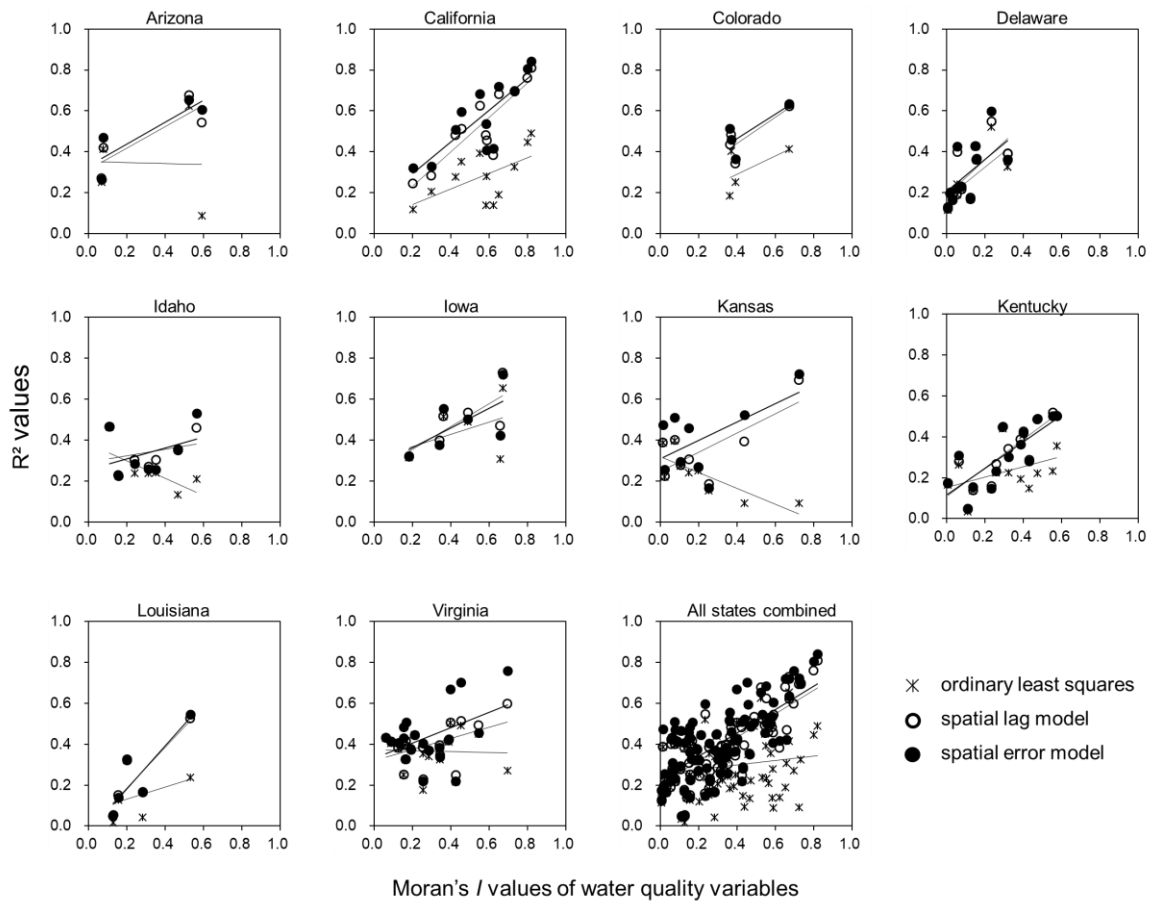
### **4.1. Changes in R<sup>2</sup> Values**

Overall, Moran's *I* values pertaining to water quality variables varied widely, from 0.01 to 0.82, across all watershed sites (Figure 5). The relationships shown in Figure 5 indicate that the improvements in R<sup>2</sup> were proportional to the degree of inherent SAC in water quality variables (i.e., the hypothesis predicting increases in R<sup>2</sup> as a function of the degree of SAC is supported). Whether I treated each state separately or combined them as a whole, strongly autocorrelated water quality variables over space (i.e., having higher Moran's *I* values) exhibited greater increases in R<sup>2</sup> values after spatial regression compared to weakly autocorrelated variables (i.e., having lower Moran's *I* values). For example, suspended carbon (Csu; *I* = 0.20) presented the lowest degree of SAC in the California study area, and pH had the highest (*I* = 0.82). For Csu, non-spatial OLS resulted in a R<sup>2</sup> of 0.12 while the spatial lag and error models resulted in R<sup>2</sup> values of 0.24 and 0.32, respectively. The OLS performance for pH improved (R<sup>2</sup> = 0.49), but spatial regression showed better results (i.e., spatial lag - R<sup>2</sup> = 0.81 and spatial error - R<sup>2</sup> = 0.72). This pattern seemed to be less clear when water quality variables within a watershed had a relatively narrow range of Moran's *I* (e.g., Delaware). A detailed example in the Kentucky study site is the variable aluminum (Al; *I* = 0.01) that presented no significant changes in R<sup>2</sup> values among non-spatial OLS, spatial lag and spatial error. R<sup>2</sup> values were equal to 0.17, 0.17 and 0.18 for OLS, spatial lag and spatial error models, respectively.

Overall, as the degree of inherent SAC increased the performance of non-spatial OLS worsened compared to spatial lag and error model R<sup>2</sup> results. Examples of this evidence are also specific conductance (SC) in the Arizona basin (*I* = 0.59; OLS - R<sup>2</sup> =



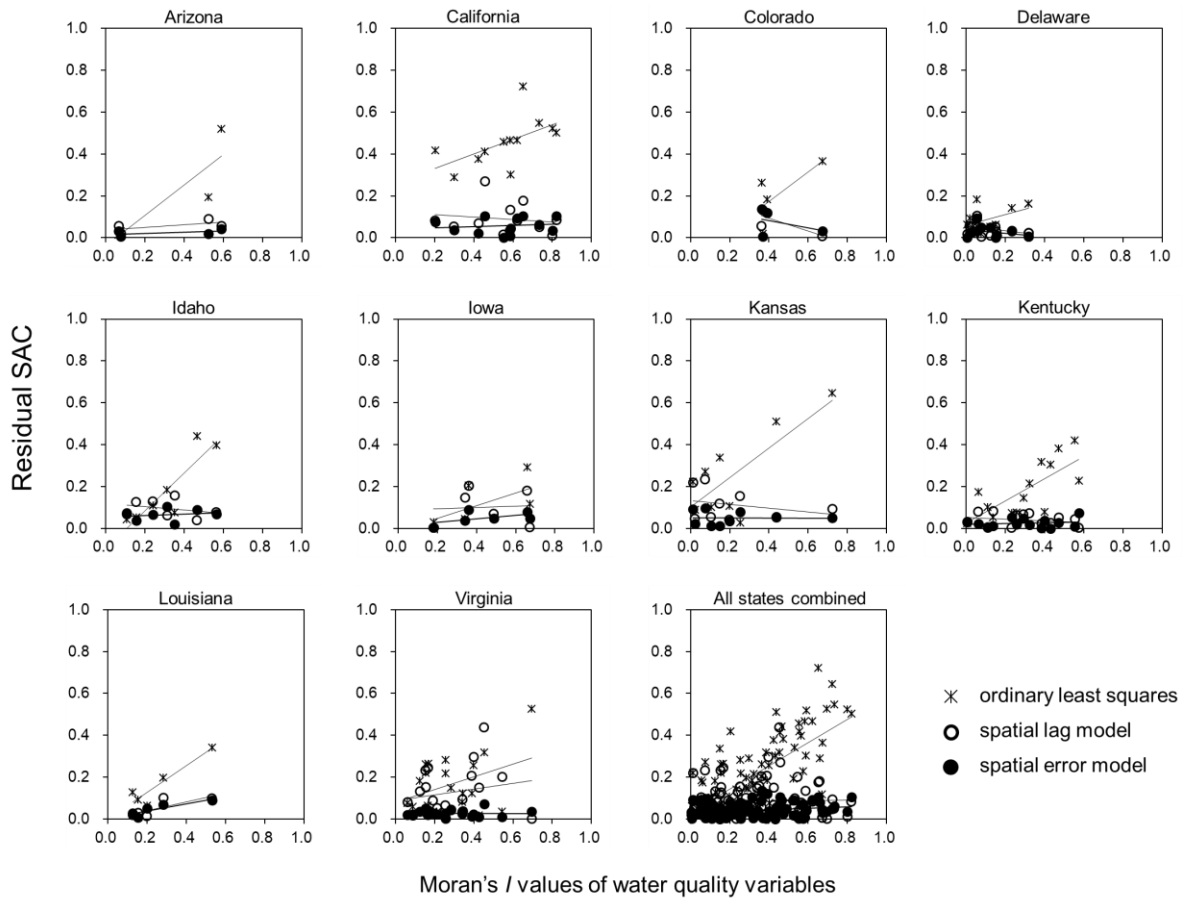
0.09, spatial lag -  $R^2= 0.54$ , spatial error  $R^2 = 0.61$ ), and temperature (T) in the Virginia site ( $I = 0.70$ ; OLS -  $R^2 = 0.27$ , spatial lag -  $R^2= 0.60$ , spatial error  $R^2 = 0.76$ ). These results also illustrate the inability of non-spatial OLS model to handle the degree of SAC inherent in water quality variables.



**Figure 5.** Relationship between the spatial autocorrelation (SAC) of each water quality variable (represented by Moran’s  $I$  values) and the  $R^2$  indicating the amount of variance in each water quality variable, explained by topographic, land use, soil groups, and spatial terms (Appendix B shows model results per water quality variable in each watershed).

## 4.2. Changes in residual Spatial Autocorrelation (rSAC)

The values of Moran's  $I$  indicating rSAC produced by non-spatial OLS presented a wider range than those from spatial regression (Figure 6; i.e., rSAC for non-spatial OLS from 0.01 to 0.72, while spatial lag rSAC ranged from 0.00 to 0.44, and spatial error, from 0.00 to 0.07). I found a positive correlation between the degree of SAC in water quality variables and rSAC from non-spatial OLS. Conversely, as expected, rSAC values acquired by spatial regressions were in general near zero. Therefore, the larger the Moran's  $I$  values possessed by water quality variables, the greater the reduction in rSAC after running models that consider spatial dependence (i.e., the hypothesis predicting greater decreases in rSAC, proportional to the degree of SAC in water quality variables, is supported). For example, in Colorado, the variable temperature (T) presented  $I = 0.67$ , and comparing rSAC values after all modeling procedure, non-spatial OLS revealed a significantly high rSAC value of 0.37 while spatial models (lag and error) reached rSAC results nearly zero (0.01 and 0.03, respectively). In Iowa, non-spatial OLS model for the variable Cl ( $I = 0.67$ ) presented rSAC = 0.12 while spatial lag rSAC was 0.01 and spatial error rSAC reached 0.05. Another example of this reduction evidence was the rSAC values after modeling the variable pH ( $I = 0.72$ ) in Kansas, spatial regression rSAC results were almost zero (spatial lag – 0.09 and spatial error – 0.05) while non-spatial OLS revealed rSAC = 0.65. Although the sites are distinct in terms of climate, geology, soil, and land use characteristics, I observed that the amount of rSAC remaining after non-spatial and spatial modeling revealed a relationship with the degree of inherent SAC in the water quality variables. Overall, all states presented a significant reduction in rSAC after spatial regression except Delaware, showing a narrow range of Moran's  $I$  values of water quality variables.



**Figure 6.** Relationship between the spatial autocorrelation (SAC) of each water quality variable (represented by Moran’s *I* values) and the SAC of model residuals (also represented by Moran’s *I* values). “All states combined” showed a general reduction in residual SAC after accounting for spatial autocorrelation in the models of each water quality variable (Appendix B shows model results per water quality variable in each watershed).

### 4.3. Overall Changes between Non-Spatial OLS and Spatial Regression Models

In general, the improvement in  $R^2$  and reduction in rSAC after spatial regression were positive, and the changes of  $R^2$  and rSAC showed to be a linear function of the degree of SAC possessed by water quality variables. I found this relationship in each study area, and the results were summarized in Table 4.

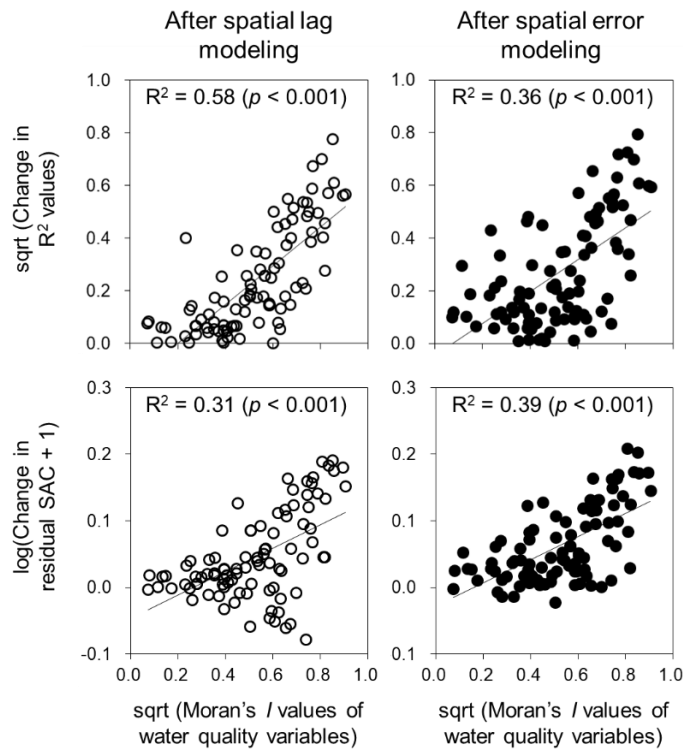
**Table 4.** Summary of mean values of spatial autocorrelation ( $I$ ) in response variables, mean values of the non-spatial OLS outcomes and mean improvement in  $R^2$  and reduction rSAC after spatial regression per state. Additionally, the linear regression model coefficients,  $R^2$ , and  $p$ -value of the Changes in  $R^2$  and rSAC per state.

			California	Colorado	Delaware	Idaho	Iowa	Kentucky	Arizona	Kansas	Louisiana	Virginia	All States Combined	
			<i>Samples</i>	12	4	12	7	6	14	4	9	5	20	93
			<i>I</i>	0.56	0.45	0.13	0.31	0.45	0.30	0.31	0.22	0.26	0.29	0.32
			$R^2$	0.28	0.31	0.27	0.25	0.44	0.23	0.34	0.23	0.15	0.37	0.29
			rSAC	0.39	0.21	0.09	0.19	0.12	0.19	0.19	0.26	0.16	0.17	0.21
After spatial regression	Improvement in $R^2$	<i>lag-ols</i>	0.26	0.16	0.03	0.09	0.05	0.09	0.13	0.11	0.09	0.03	0.10	
		<i>error-ols</i>	0.29	0.18	0.04	0.09	0.04	0.08	0.15	0.17	0.10	0.07	0.12	
	Reduction in rSAC	<i>ols-lag</i>	0.37	0.13	0.05	0.09	0.02	0.15	0.13	0.14	0.11	0.04	0.12	
		<i>ols-error</i>	0.40	0.13	0.07	0.12	0.07	0.16	0.17	0.17	0.21	0.12	0.14	0.17
Linear regression models for the Change in $R^2$ vs. $I$	Model fit Spatial Lag	$R^2$	0.55	0.12	0.07	0.85	0.68	0.61	0.51	0.91	0.94	0.46	0.58	
		$\beta_o$	0.00	0.07	0.01	-0.09	-0.07	-0.04	-0.04	-0.07	-0.09	-0.09	-0.05	-0.15
		$\beta_1$	0.46	0.19	0.11	0.58	0.26	0.44	0.55	0.86	0.70	0.30	0.30	0.74
		$p$ -value	<0.001 *	0.10 *	0.60	0.10 *	0.53	0.08 *	0.39	0.09 *	0.38	0.28	<0.001 *	
	Model fit Spatial Error	$R^2$	0.40	0.03	0.00	0.77	0.64	0.55	0.42	0.77	0.93	0.29	0.36	
		$\beta_o$	0.07	0.11	0.03	-0.13	-0.04	-0.04	-0.02	-0.01	-0.10	-0.04	-0.04	
$\beta_1$		0.39	0.15	0.02	0.68	0.19	0.40	0.56	0.83	0.75	0.40	0.60		
	$p$ -value	<0.001 *	0.06 *	0.52	0.15	0.62	0.10 *	0.33	0.02 *	0.39	0.06 *	<0.001 *		
Linear regression models for the Change in rSAC vs. $I$	Model fit Spatial Lag	$R^2$	0.33	0.56	0.58	0.66	0.42	0.60	0.67	0.67	0.80	0.03	0.31	
		$\beta_o$	0.14	-0.32	0.01	-0.21	-0.10	-0.03	-0.07	-0.03	0.00	-0.01	-0.05	
		$\beta_1$	0.41	1.01	0.36	0.98	0.27	0.57	0.66	0.80	0.42	0.17	0.18	
		$p$ -value	<0.001 *	0.18	0.01 *	0.20	0.71	<0.001 *	0.34	0.08 *	0.09 *	0.32	<0.001 *	
	Model fit Spatial Error	$R^2$	0.32	0.87	0.42	0.84	0.28	0.45	0.77	0.60	0.74	0.17	0.39	
		$\beta_o$	0.22	-0.26	0.02	-0.15	-0.03	0.01	-0.05	0.05	0.00	0.05	-0.03	
$\beta_1$		0.32	0.88	0.33	0.87	0.22	0.51	0.70	0.71	0.46	0.30	0.17		
	$p$ -value	<0.001 *	0.17	0.00 *	0.11	0.15	<0.001 *	0.25	0.02 *	0.08 *	<0.001 *	<0.001 *		

\* significant at the 0.10 level.  $I$ : Moran's  $I$  values; OLS: ordinary least squares; rSAC: residual spatial autocorrelation; lag-ols: improvement in  $R^2$  from non-spatial ols to spatial lag regression; error-ols: improvement in  $R^2$  from non-spatial ols to spatial error regression; ols-lag: reduction in rSAC from non-spatial ols to spatial lag regression; ols-error: reduction in rSAC from non-spatial ols to spatial error regression.

#### 4.4. Summary of Findings

The magnitude of model improvement (i.e., increases in  $R^2$  and decreases in rSAC), after both spatial lag and error modeling, is significantly and linearly a function of the SAC inherently possessed by water quality variables (i.e., response variables) (Figure 7). This, in turn, supported the hypothesis that water quality variables with a higher amount of SAC would exhibit greater improvement in model outcomes than those with a lower amount of SAC.



**Figure 7.** Linear regression models demonstrating that the magnitude of improvement of model performance after spatial lag and error modeling is significantly and linearly explained by the SAC inherently possessed by water quality variables. The Moran's  $I$  ( $x$ -axis) and Change in  $R^2$  ( $y$ -axis) values were transformed using square-root transformation, while the Change in rSAC ( $y$ -axis) were log-transformed.

## **CHAPTER 5: DISCUSSION**

The results support the hypothesis and offer insights into the field of water quality modeling. Most importantly, the level of SAC in water quality variables has the potential to indicate how much improvement a non-spatial model would experience if SAC was appropriately considered (i.e., increases in  $R^2$  values and decreases in rSAC). I have demonstrated across divergent watersheds in the USA that the higher the SAC in a water quality variable, the greater the improvements in the model after accounting for SAC. Water quality studies, as previously mentioned, achieved better results when considering spatial modeling approaches that account for SAC (Franczyk & Chang, 2009; Pratt & Chang, 2012; Yu et al., 2013; Vrebos et al., 2017). However, these studies have not considered the magnitude of SAC in the response variable as the main driver of model improvements. Furthermore, I observed that variables with lower degree of inherent SAC (i.e., lower Moran's  $I$  values) underwent smaller changes in model outcomes compared to those that presented larger Moran's  $I$  values. In this sense, higher Moran's  $I$  values imply more spatial organization (e.g., strong connection among water quality stations through the stream network) than smaller Moran's  $I$  values. This indicates that the need for (and potentially the benefit from) accounting for SAC in water quality modeling increases as the degree of SAC increases.

In this study, I investigated water quality variables from 10 watersheds, each distinct in geology, land use, soil, and topography. I analyzed a total of 93 water quality variables, many of which also differed among the watersheds. Despite such variations, this study reveals a consistent and linear relationship between the SAC of water quality variables and changes in the model outcomes ( $R^2$  and rSAC). This finding perfectly

accords with the study of Kim et al. (2016), who evaluated the effect of SAC in soil–landform modeling to find that the degree of SAC in soil variables (i.e., dependent variables) influenced model improvements after the SAC was properly accounted for.

These findings suggest that future water quality modeling studies should account for SAC in order to improve the performance of non-spatial approaches, principally when the predictors in the model cannot sufficiently account for all SAC in the model (Legendre & Fortin, 1989; Legendre, 1993; Dormann, 2007; Miller et al., 2007; Kim, 2013; Kim et al., 2016). Overall, the improvements include increasing  $R^2$  and decreasing rSAC. The most important point is that the degree of these increases and decreases showed to be linearly correlated with the level of SAC in water quality variables. Therefore, water quality studies should not only focus on accounting for the presence of SAC, but also on understanding the magnitude of SAC inherent in water quality variables. Doing so, we could point out the degree of connectivity within water quality variables, as well as the improvement in model outcomes of a non-spatial approach before performing a spatial regression.

Adequate information on the degree of hydrologic connectivity among water quality variables is needed in watershed management and policy decisions (Pringle, 2001, 2003). The level of SAC inherent in a variable can allow managers to reveal the complex spatial relationship of water quality as well as its changes from up to downstream. For example, pH values were available in several distinct watersheds in this study. For Cherry Creek, AZ, pH values revealed a SAC degree of 0.07, while in Headwaters Tuolumne River, CA, pH values presented an almost perfect positive SAC degree of 0.82. This differences in variable spatial distribution over distinct regions

can provide insights to the studies of hydrologic connectivity helping in the development of more efficient strategies to enhance the quality of aquatic ecosystems. It can uncover dissimilarity patterns among water quality variables throughout the stream network and help with the implementation of policies that are ecologically beneficial to the aquatic ecosystem. Therefore, I conclude that the investigation of SAC in water quality modeling is not only beneficial in the model results, but also in the process of watershed management.

Streams can be considered spatially structured ecological networks, where patterns are usually associated with the in-stream flow and habitat, or even the physical structure of the network. The understanding of these patterns can be limited when only using Euclidean distance (Peterson et al., 2013). For example, two sites that are near to each other can be considered neighbors due to the distance measured through the Euclidean technique, but they can present distinct water quality measures simply due to the water quality origins from vastly different drainage areas. It is also important to point out the directionality factor in streams that may impact the neighboring detection. Therefore, I highlight that this is a limitation in this study and further studies should focus on applying spatial network distance techniques and detecting directionality influence to better understand the SAC phenomenon.



## **CHAPTER 6: CONCLUSIONS & FUTURE WORK**

Spatial autocorrelation (SAC) is a property possessed by any ecological or environmental variable. Consequently, its incorporation and impacts on modeling results have been studied in much detail in a variety of scientific fields. This study demonstrates that analyzing SAC in water quality modeling provides benefits beyond just improving in model outcomes ( $R^2$  and rSAC): It can potentially lead to a better understanding of the extent of spatial organization of water quality variables, as well as serve as a useful screening technique to anticipate the predictability of the spatial pattern in the independent variable used in a spatially explicit model. I also highlight the benefits of understanding the level of SAC possessed by a water quality variable in the process of watershed management, and the limitation of not using network distances techniques in this study, which could better account for the spatial pattern that exist in spatially structured ecological networks such as streams.

### **Future research**

#### *Seasonality and scale*

This study aimed to compare the impact of inherent SAC in water quality variables between non-spatial and spatial modeling outcomes without explicitly taking into account seasonal and scale variability. Several water quality studies argued that seasonality and scale are significant factors that can even change the conclusions of water quality modeling. Thus, to understand if spatial modeling outcomes would generally present a linear relationship with the degree of inherent SAC in water quality, future works should consider the potential effect of seasonality and scale by acquiring intensively data at

different times of the year and scales for each watershed. These considerations would allow us to understand if the linear relationship between the spatial model outcomes and the degree of SAC holds true across broad and fine extent as well as in dry or wet conditions. Conclusions from these studies could advance water quality modeling practices as well as serve as a management action source.

### *Coefficient shift*

It is still necessary to understand the influence, source, and behavior of spatially dependent variables in river ecosystems. Studies that aimed to compare non-spatial and spatial modeling techniques argued that the coefficients of the independent variables can undergo a change in their predictive power after incorporation of SAC into the modeling procedure (Lennon, 2000; Kim, 2013). This “shift” can change our understanding of key explanatory variables in the prediction of water quality. Thus, future research should focus on understanding this shift in the predictive power of independent variables used to model water quality variables because it has the potential to change our knowledge about the causes of spatially structured water quality patterns.

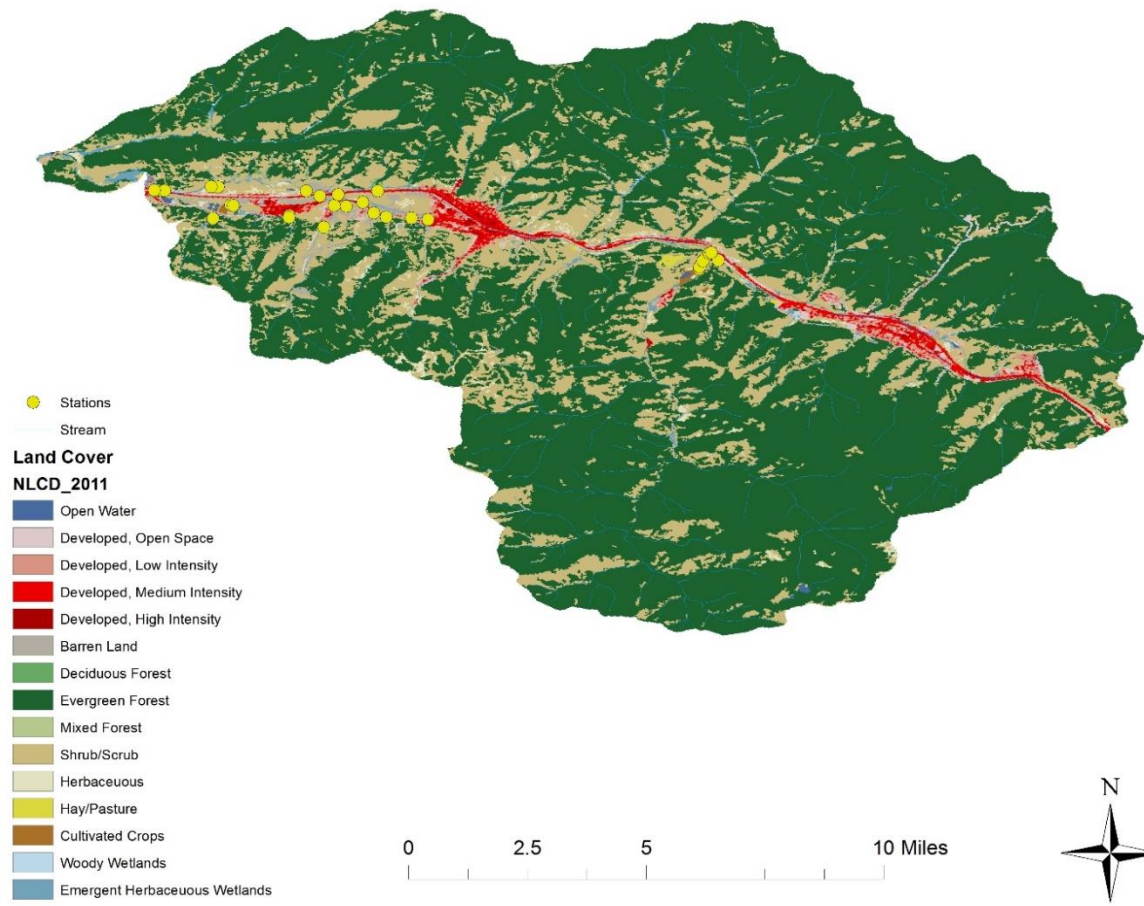
### *Spatial heterogeneity and specific water quality variable studies*

Ecologists and hydrologists are aware of main water quality variables that are important to maintaining the good condition of aquatic ecosystems. For example, specific conductance (SC), dissolved oxygen (DO), and water temperature (T) are water quality variables important in the prediction of habitat quality for fishes and other aquatic animals.

Studying one of these variables in distinct watersheds (i.e., different topography, shape, soil, and land use characteristics) may reveal the impact of spatial heterogeneity in water quality conditions as well as improve the interpretation of explanatory variables that are commonly used to model water quality. To understand the spatial heterogeneity and the importance of predictors in modeling specific water quality variables, future research should consider SAC in different scales, its influence, and its sources. Thus, this study may provide insights on best watershed practices for controlling habitat quality over divergent conditions.

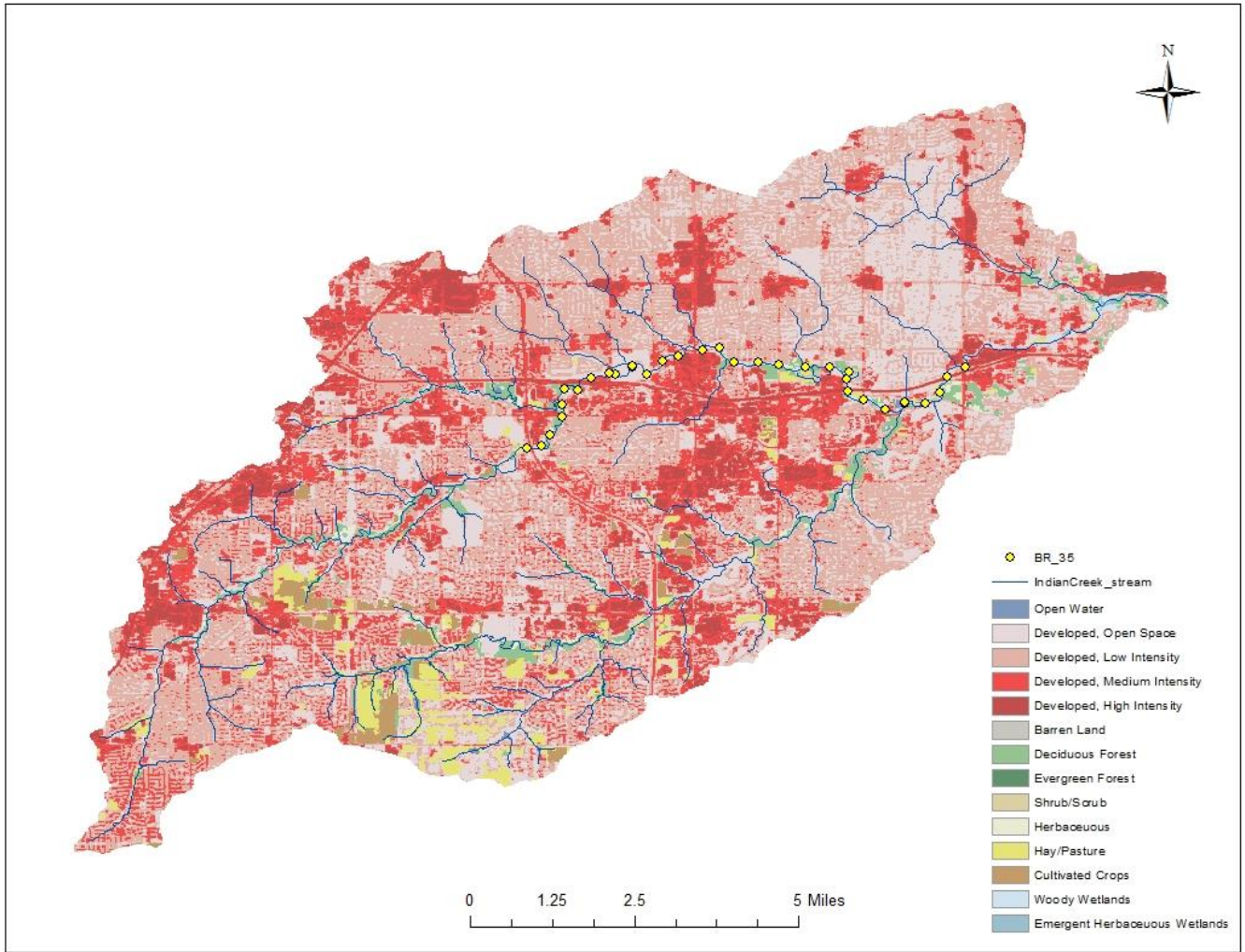
**APPENDIX A: Larger maps for better visualization of water stations location.**

Idaho (a) – Lower South Fork Coeur d’Alene River

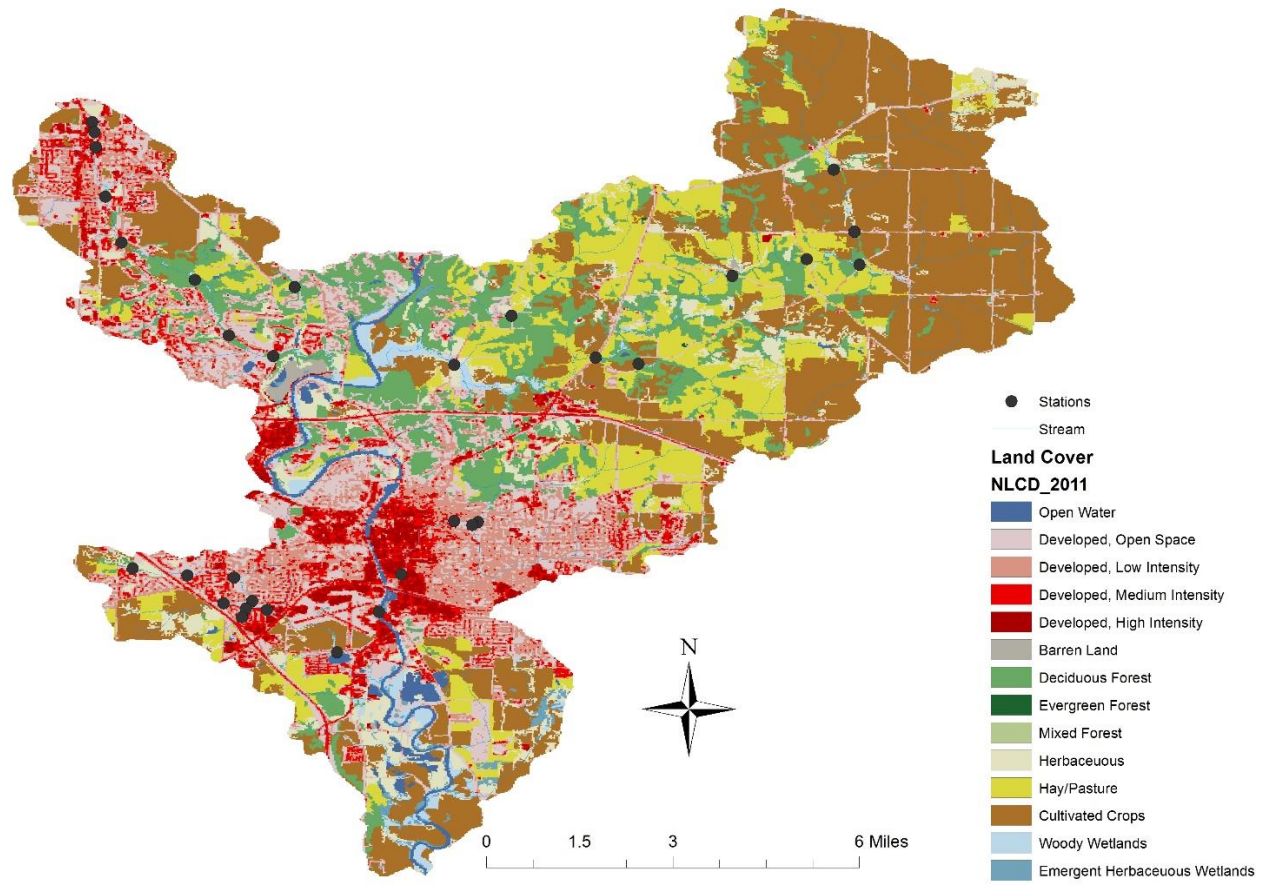


Kansas (b) – Indian Creek

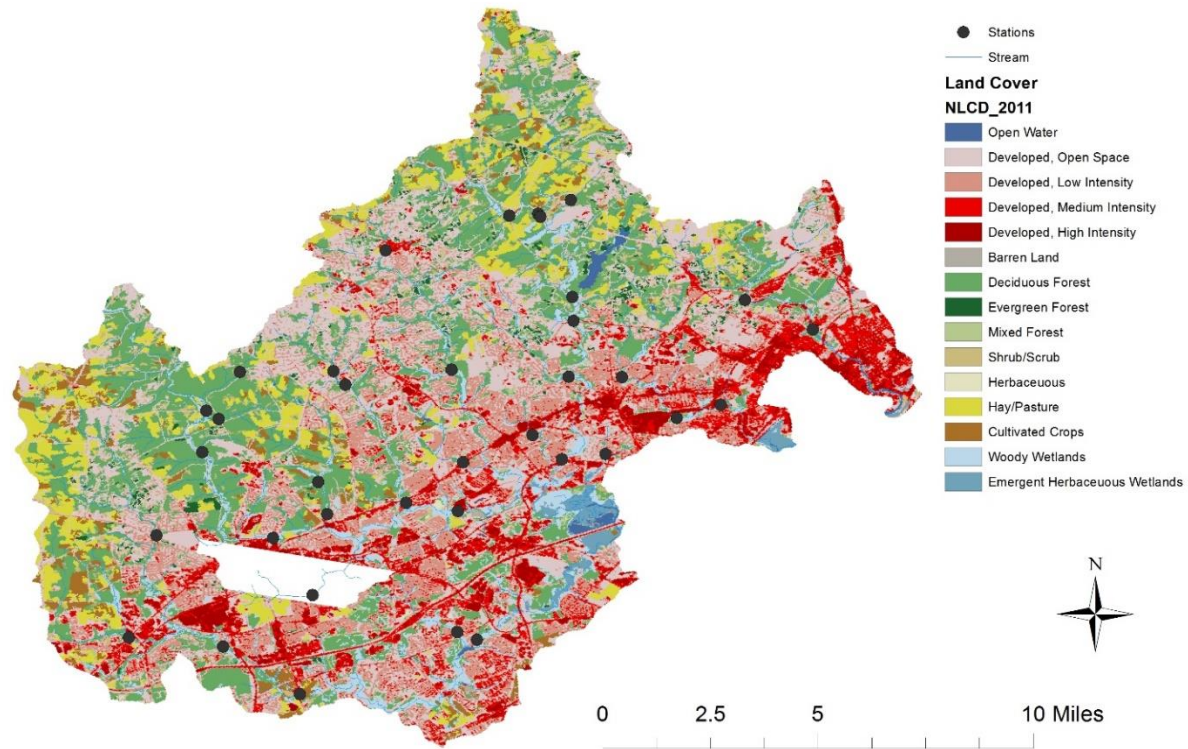
50



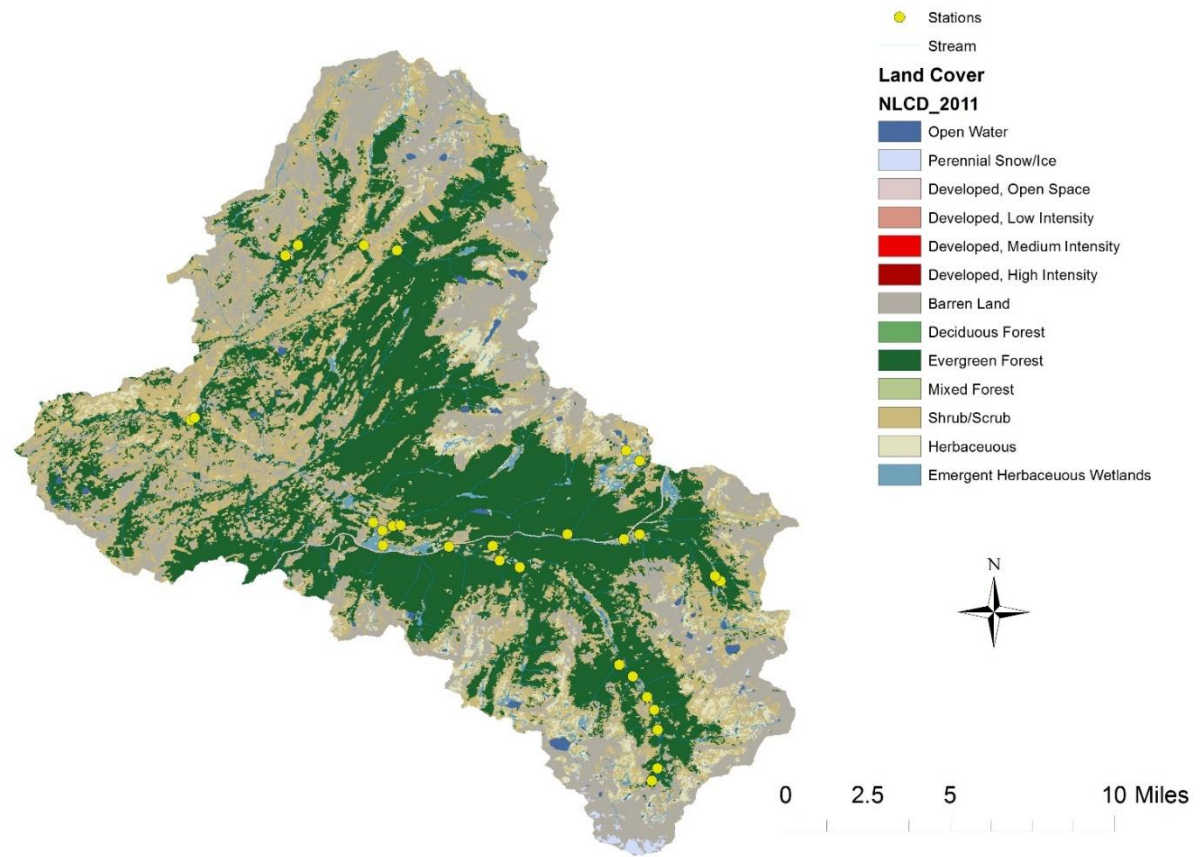
Iowa (c) – Iowa River



Delaware (d) – Clay, Mill, Bradywine Creek, and Cristina River

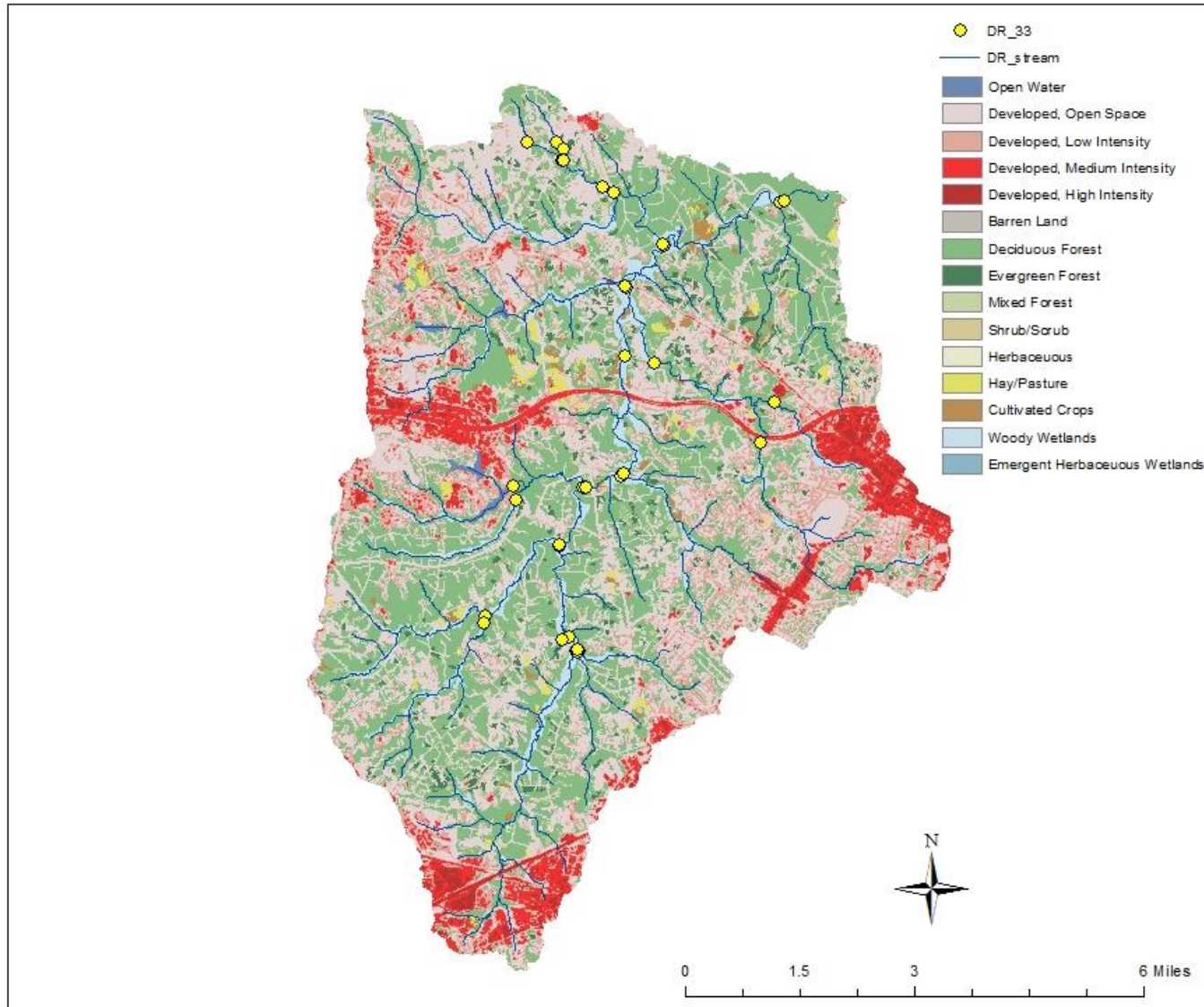


California (e) – Headwater Tuolumne River



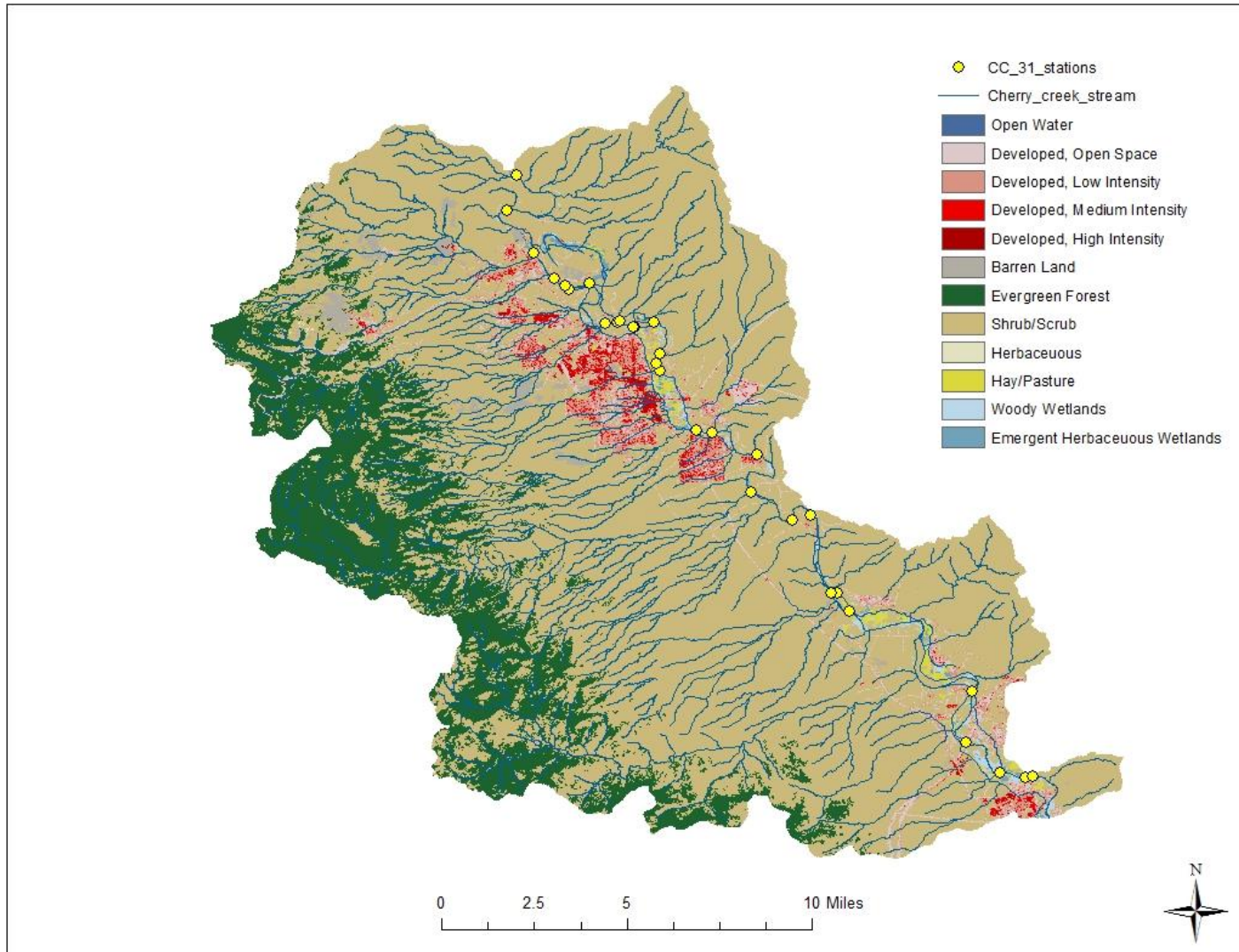


Virginia (f) – Difficult River

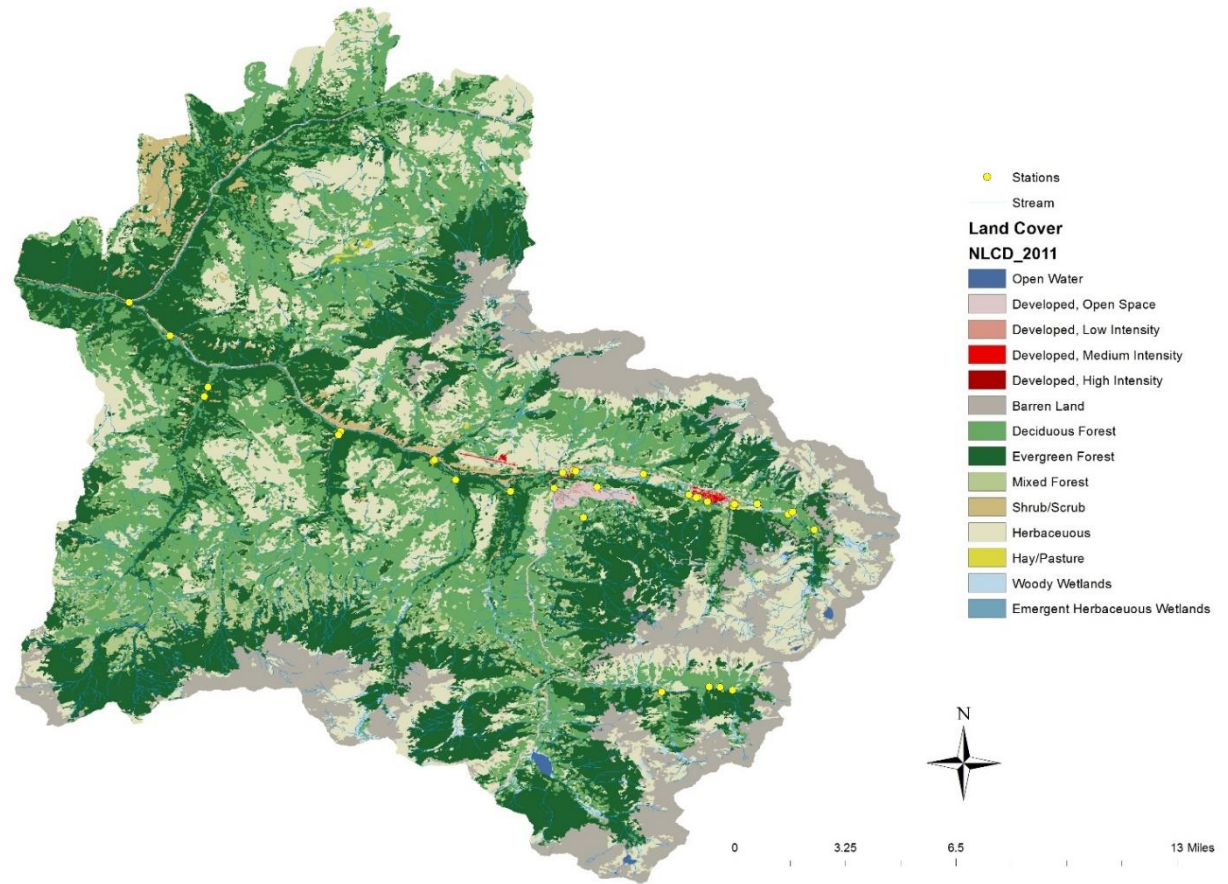


Arizona (g) – Cherry Creek

55



Colorado (h) – Upper San Miguel River

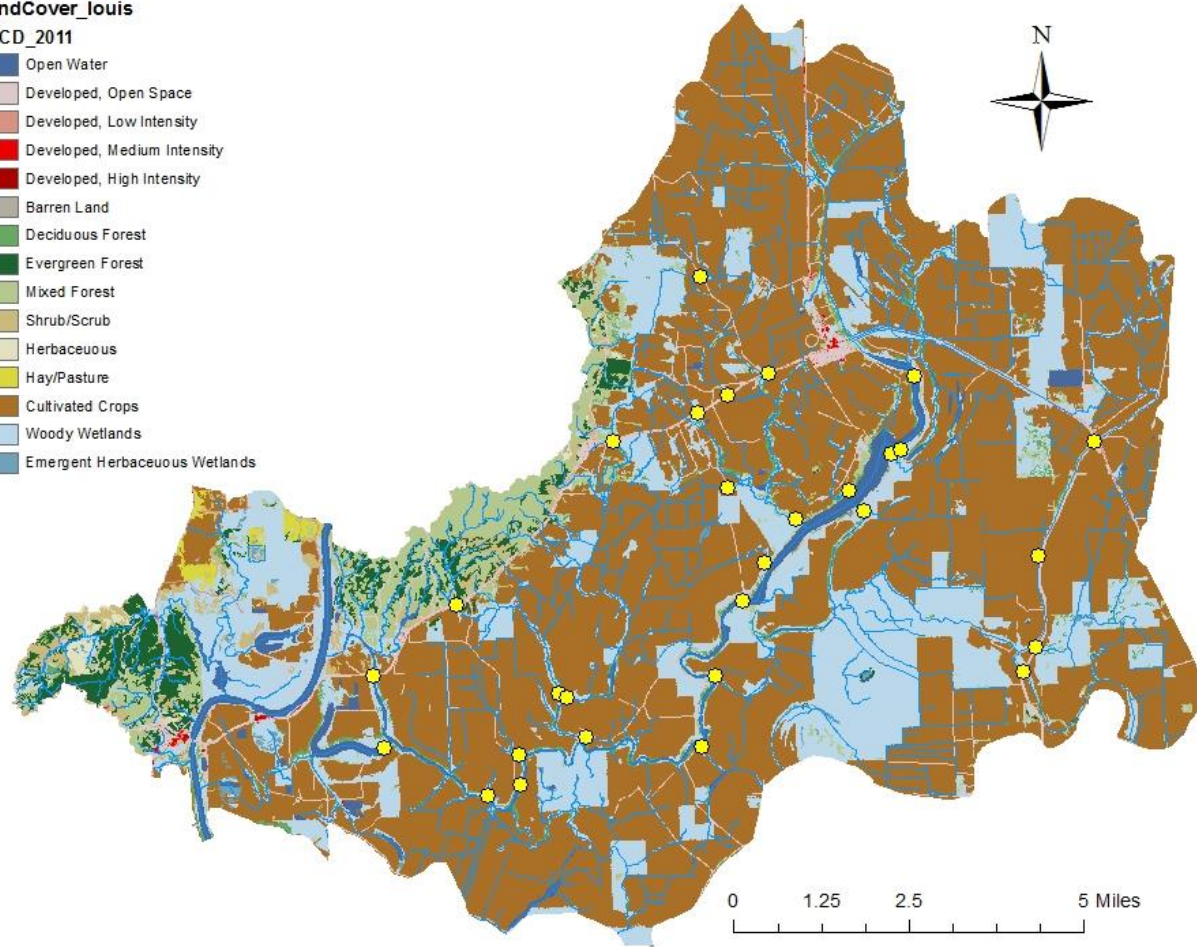


Louisiana (i) – Bayou Louis/ Lake Louis

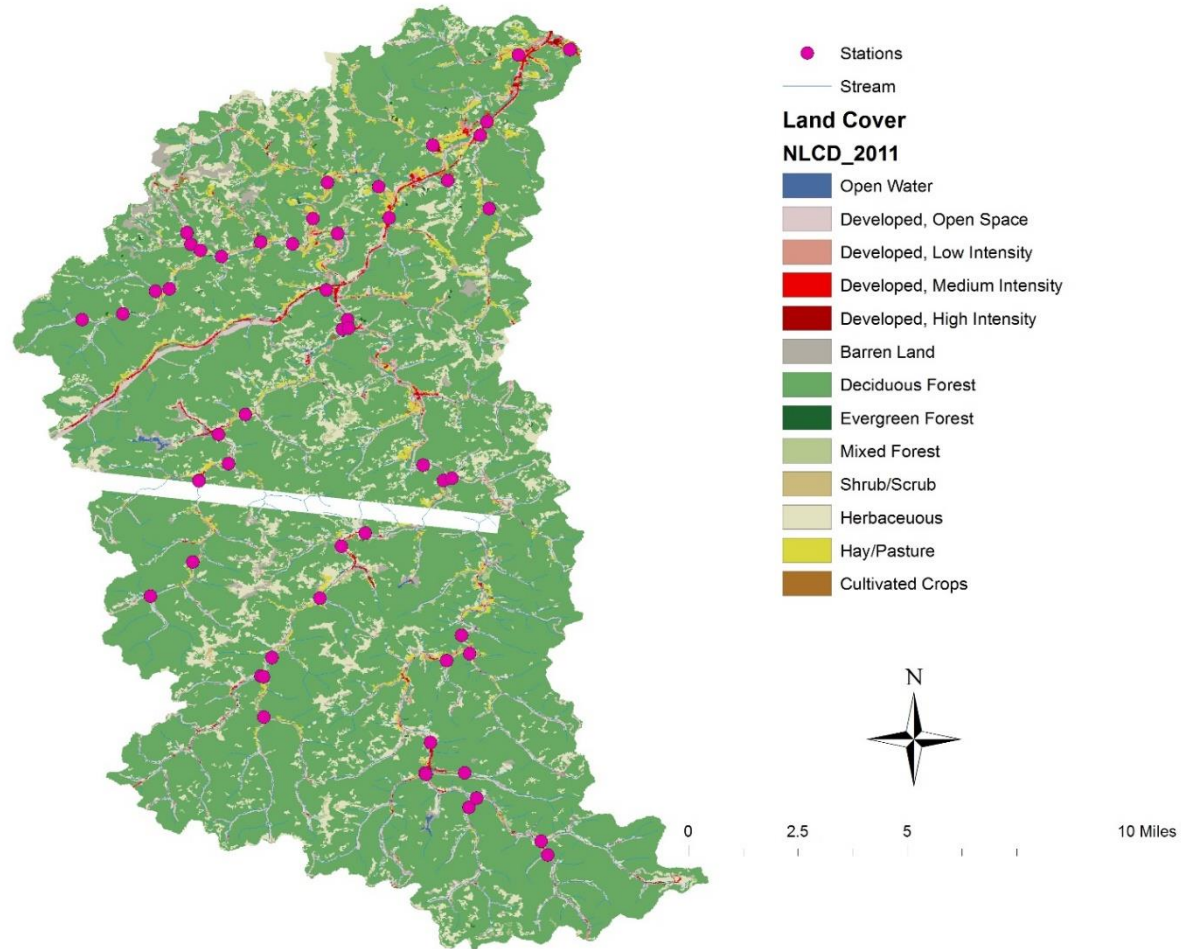
- Louis\_stations
- Louis\_stream\_HU12

**LandCover\_louis**  
**NLCD\_2011**

- Open Water
- Developed, Open Space
- Developed, Low Intensity
- Developed, Medium Intensity
- Developed, High Intensity
- Barren Land
- Deciduous Forest
- Evergreen Forest
- Mixed Forest
- Shrub/Scrub
- Herbaceous
- Hay/Pasture
- Cultivated Crops
- Woody Wetlands
- Emergent Herbaceous Wetlands



Kentucky (j) - Beaver Creek



**APPENDIX B: Model outcomes and Moran's I values per water quality variables on each watershed.**

Abbreviations	Meaning
R <sup>2</sup>	Coefficient of determination
rSAC	residual Spatial Autocorrelation
abs rSAC	absolute residual Spatial Autocorrelation
<i>I</i>	Moran's <i>I</i>
abs <i>I</i>	absolute Moran's <i>I</i>
OLS	Ordinary least squares
Lag	Spatial lag model
Error	Spatial error model
SC	Specific Conductance
DO	Dissolved oxygen
TDS	Total Dissolved Solids
TN	Total Nitrogen
DIN	Dissolved Nitrogen
KjN	Kjeldahl Nitrogen
TP	Total Phosphorus
Tur	Turbidity
Alk	Alkalinity
Csu	suspended Carbon
Chla	Chlorophyll
Nin	inorganic Nitrogen
TOC	Total Organic Carbon
DOC	Dissolved Organic Carbon
Pb	Dissolved Lead
Zn	Dissolved Zinc
Cd	Dissolved Cadmium
As	Dissolved Arsenic

Idaho (a) – Lower South Fork Coeur d’Alene River

			R <sup>2</sup>			rSAC			abs rSAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
SC	0.56	0.56	0.21	0.46	0.53	0.40	0.08	0.07	0.40	0.08	0.07
pH	0.31	0.31	0.24	0.27	0.26	0.18	0.06	0.11	0.18	0.06	0.11
T	0.15	0.15	0.22	0.23	0.23	-0.05	-0.13	-0.04	0.05	0.13	0.04
Pb	0.11	0.11	0.46	0.47	0.47	-0.04	-0.07	-0.07	0.04	0.07	0.07
Cd	0.35	0.35	0.24	0.30	0.26	0.08	-0.16	-0.02	0.08	0.16	0.02
Zn	0.24	0.24	0.24	0.30	0.28	0.11	-0.13	-0.07	0.11	0.13	0.07
As	0.47	0.47	0.13	0.36	0.35	0.44	0.04	0.09	0.44	0.04	0.09

Kansas (b) – Indian Creek

			R <sup>2</sup>			rSAC			abs rSAC		
	<i>I</i>	abs I	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
DIN	0.07	0.07	0.40	0.40	0.51	0.27	0.23	0.10	0.27	0.23	0.10
KjN	0.10	0.10	0.28	0.28	0.29	0.10	0.06	0.01	0.10	0.06	0.01
TN	0.01	0.01	0.39	0.39	0.47	0.22	0.22	0.09	0.22	0.22	0.09
DO	0.44	0.44	0.09	0.39	0.52	0.51	0.05	0.05	0.51	0.05	0.05
pH	0.72	0.72	0.09	0.69	0.72	0.65	-0.09	-0.05	0.65	0.09	0.05
TP	0.15	0.15	0.24	0.31	0.46	0.34	0.12	0.01	0.34	0.12	0.01
T	0.20	0.20	0.25	0.27	0.27	0.11	-0.04	-0.04	0.11	0.04	0.04
Tur	0.25	0.25	0.15	0.19	0.17	0.03	-0.15	-0.08	0.03	0.15	0.08
SC	0.02	0.02	0.22	0.22	0.26	0.09	0.05	-0.02	0.09	0.05	0.02



Iowa (c) – Iowa River

			R <sup>2</sup>			rSAC			abs rSAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
Cl	0.67	0.67	0.65	0.73	0.72	0.12	-0.01	0.05	0.12	0.01	0.05
DO	0.18	0.18	0.32	0.32	0.32	-0.03	0.00	0.00	0.03	0.00	0.00
PO <sub>4</sub> <sup>3-</sup>	0.66	0.66	0.31	0.47	0.42	0.29	-0.18	-0.08	0.29	0.18	0.08
NO <sub>3</sub> <sup>-</sup>	0.36	0.36	0.52	0.52	0.56	-0.20	-0.20	-0.09	0.20	0.20	0.09
pH	0.34	0.34	0.38	0.40	0.38	-0.05	-0.15	-0.04	0.05	0.15	0.04
T	0.49	0.49	0.49	0.53	0.50	0.05	-0.07	-0.05	0.05	0.07	0.05

Delaware (d) - Clay, Mill, Bradywine Creek, and Cristina River

			R <sup>2</sup>			rSAC			abs rSAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
Alk	0.08	R <sup>2</sup> +A33:L47	0.22	0.22	0.23	-0.02	0.00	0.05	0.02	0.00	0.05
Cl	0.23		0.52	0.55	0.60	-0.14	-0.03	0.03	0.14	0.03	0.03
Chla	0.02		0.19	0.20	0.20	0.09	0.06	0.03	0.09	0.06	0.03
DO	0.15		0.43	0.43	0.43	-0.06	-0.05	0.00	0.06	0.05	0.00
Nin	0.05		0.19	0.19	0.22	-0.10	-0.09	-0.03	0.10	0.09	0.03
TN	0.03		0.16	0.16	0.17	-0.05	-0.05	-0.03	0.05	0.05	0.03
TOC	0.32		0.33	0.39	0.36	0.16	-0.02	0.01	0.16	0.02	0.01
DOC	0.32		0.33	0.39	0.36	0.16	-0.02	0.01	0.16	0.02	0.01
pH	0.16		0.36	0.36	0.37	-0.03	-0.03	0.01	0.03	0.03	0.01
TP	0.12		0.17	0.18	0.17	0.05	0.01	0.05	0.05	0.01	0.05
SC	-0.05		0.24	0.40	0.43	-0.18	0.10	0.09	0.18	0.10	0.09
T	-0.01		0.12	0.12	0.13	0.06	0.02	0.00	0.06	0.02	0.00

California (e) - Headwater Tuolumne River

			R2			rSAC			abs rSAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
Alk	0.80	0.80	0.45	0.76	0.81	0.52	0.01	0.04	0.52	0.01	0.04
Ca	0.55	0.55	0.39	0.62	0.68	0.46	0.01	0.00	0.46	0.01	0.00
Csu	-0.20	0.20	0.12	0.24	0.32	-0.42	-0.08	0.08	0.42	0.08	0.08
Cl	0.58	0.58	0.14	0.48	0.54	0.47	0.03	0.01	0.47	0.03	0.01
Mg	0.42	0.42	0.28	0.48	0.51	0.38	0.07	0.02	0.38	0.07	0.02
pH	0.82	0.82	0.49	0.81	0.84	0.50	0.08	0.10	0.50	0.08	0.10
K	0.46	0.46	0.35	0.51	0.60	0.41	0.27	0.10	0.41	0.27	0.10
SiO <sub>2</sub>	0.62	0.62	0.14	0.38	0.42	0.47	-0.08	-0.09	0.47	0.08	0.09
Na	0.59	0.59	0.28	0.46	0.41	0.30	-0.13	-0.05	0.30	0.13	0.05
SO <sub>4</sub> <sup>2-</sup>	0.65	0.65	0.19	0.68	0.72	0.72	0.18	0.10	0.72	0.18	0.10
T	0.30	0.30	0.21	0.28	0.33	0.29	0.05	0.04	0.29	0.05	0.04
TDS	0.73	0.73	0.33	0.70	0.70	0.55	0.05	0.06	0.55	0.05	0.06

## Virginia (f) - Difficult River

			R <sup>2</sup>			rSAC			abs SAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
Alk	0.40	0.40	0.50	0.51	0.67	-0.26	-0.30	0.02	0.26	0.30	0.02
Br	0.09	0.09	0.40	0.41	0.41	-0.06	0.02	0.02	0.06	0.02	0.02
Ca	0.17	0.17	0.42	0.42	0.51	-0.26	-0.25	0.04	0.26	0.25	0.04
CO <sub>2</sub>	0.34	0.34	0.38	0.40	0.38	0.08	-0.09	-0.03	0.08	0.09	0.03
Cl	0.12	0.12	0.38	0.38	0.41	0.18	0.13	0.04	0.18	0.13	0.04
F	0.54	0.54	0.45	0.49	0.46	-0.04	-0.20	0.01	0.04	0.20	0.01
Fe	0.21	0.21	0.44	0.44	0.44	0.03	0.05	0.02	0.03	0.05	0.02
Mg	0.15	0.15	0.25	0.25	0.48	-0.26	-0.23	0.05	0.26	0.23	0.05
Mn	0.34	0.34	0.33	0.35	0.34	0.12	-0.04	0.04	0.12	0.04	0.04
DO	0.16	0.16	0.32	0.33	0.33	0.06	0.02	0.02	0.06	0.02	0.02
pH	0.39	0.39	0.41	0.42	0.42	-0.12	-0.21	-0.01	0.12	0.21	0.01
TP	0.43	0.43	0.22	0.25	0.22	-0.02	-0.15	-0.01	0.02	0.15	0.01
K	0.25	0.25	0.35	0.39	0.40	0.28	0.07	0.00	0.28	0.07	0.00
SiO <sub>2</sub>	0.19	0.19	0.37	0.38	0.37	0.04	0.09	0.03	0.04	0.09	0.03
Na	0.15	0.15	0.39	0.40	0.43	0.22	0.15	0.04	0.22	0.15	0.04
SC	0.06	0.06	0.43	0.43	0.43	0.08	0.08	0.02	0.08	0.08	0.02
SO <sub>4</sub> <sup>2-</sup>	0.45	0.45	0.49	0.51	0.70	-0.32	-0.44	0.07	0.32	0.44	0.07
T	0.70	0.70	0.27	0.60	0.76	0.53	0.00	-0.04	0.53	0.00	0.04
TDS	-0.26	0.26	0.18	0.23	0.22	-0.22	0.00	-0.03	0.22	0.00	0.03
Tur	-0.28	0.28	0.34	0.37	0.37	-0.15	-0.04	-0.05	0.15	0.04	0.05

Arizona (g) – Cherry Creek

			R <sup>2</sup>			rSAC			abs rSAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
pH	-0.07	0.07	0.25	0.27	0.26	-0.01	0.06	0.03	0.01	0.06	0.03
T	0.52	0.52	0.63	0.68	0.66	0.19	-0.09	0.02	0.19	0.09	0.02
SC	0.59	0.59	0.09	0.54	0.61	0.52	0.06	0.04	0.52	0.06	0.04
DO	-0.08	0.08	0.41	0.42	0.47	-0.03	-0.02	-0.01	0.03	0.02	0.01

Colorado (h) - Upper San Miguel River

			R <sup>2</sup>			rSAC			abs rSAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
DO	0.39	0.39	0.25	0.34	0.37	0.18	-0.12	-0.12	0.18	0.12	0.12
pH	0.37	0.37	0.40	0.48	0.46	0.02	-0.13	-0.01	0.02	0.13	0.01
T	0.67	0.67	0.41	0.62	0.64	0.37	0.01	-0.03	0.37	0.01	0.03
SC	0.36	0.36	0.18	0.44	0.51	0.26	-0.06	-0.14	0.26	0.06	0.14

Louisiana (i) - Bayou Louis/ Lake Louis

			R <sup>2</sup>			rSAC			abs rSAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
pH	0.13	0.13	0.02	0.05	0.06	0.13	-0.02	-0.03	0.13	0.02	0.03
T	0.15	0.15	0.13	0.15	0.14	0.09	-0.03	0.01	0.09	0.03	0.01
SC	0.20	0.20	0.32	0.33	0.32	0.07	0.01	0.05	0.07	0.01	0.05
DO	0.28	0.28	0.04	0.17	0.17	0.20	-0.10	-0.07	0.20	0.10	0.07
TDS	0.53	0.53	0.24	0.53	0.54	0.34	-0.10	-0.09	0.34	0.10	0.09

Kentucky (j) – Beaver Creek

			R <sup>2</sup>			rSAC			abs rSAC		
	<i>I</i>	abs <i>I</i>	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
Alk	0.11	0.11	0.04	0.05	0.05	0.10	0.00	-0.01	0.10	0.00	0.01
Al	0.01	0.01	0.17	0.17	0.18	-0.02	0.03	0.03	0.02	0.03	0.03
Ba	0.06	0.06	0.26	0.28	0.31	-0.17	-0.08	-0.02	0.17	0.08	0.02
Ca	0.55	0.55	0.23	0.52	0.50	0.42	-0.04	-0.01	0.42	0.04	0.01
Nin	0.29	0.29	0.44	0.45	0.45	0.15	0.07	0.05	0.15	0.07	0.05
Cl	0.23	0.23	0.15	0.16	0.15	0.07	0.00	0.05	0.07	0.00	0.05
Fe	0.40	0.40	0.41	0.43	0.42	0.08	-0.02	0.04	0.08	0.02	0.04
KjN	0.43	0.43	0.15	0.29	0.28	0.30	-0.03	0.00	0.30	0.03	0.00
Mg	0.47	0.47	0.22	0.49	0.49	0.38	-0.05	-0.03	0.38	0.05	0.03
Mn	0.58	0.58	0.35	0.50	0.50	0.23	0.00	0.07	0.23	0.00	0.07
K	0.26	0.26	0.23	0.27	0.23	0.08	-0.06	0.02	0.08	0.06	0.02
Na	0.14	0.14	0.14	0.14	0.16	-0.05	-0.08	-0.01	0.05	0.08	0.01
SO <sub>4</sub> <sup>2-</sup>	0.39	0.39	0.19	0.39	0.36	0.32	-0.02	0.00	0.32	0.02	0.00
TDS	0.32	0.32	0.22	0.34	0.30	0.22	-0.07	-0.02	0.22	0.07	0.02



## REFERENCES

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.
- Anselin, L. (1988). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical analysis*, 20(1), 1-17.
- Anselin, L. (1993). Discrete space autoregressive models. *GIS and Environmental Modeling*, 454-469.
- Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis*, 27(2), 93-115.
- Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural economics*, 27(3), 247-267.
- Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International regional science review*, 26(2), 153-166.
- Barringer, T., Dunn, D., Battaglin, W., & Vowinkel, E. (1990). Problems and methods involved in relating land use to ground- water quality. *JAWRA Journal of the American Water Resources Association*, 26(1), 1-9.
- Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2010). Regression analysis of spatial data. *Ecology letters*, 13(2), 246-264.
- Bellehumeur, C., & Legendre, P. (1998). Multiscale sources of variation in ecological variables: modeling spatial dispersion, elaborating sampling designs. *Landscape Ecology*, 13(1), 15-25.
- Mauricio Bini, L., Diniz-Filho, J. A. F., Rangel, T. F., Akre, T. S., Albaladejo, R. G., Albuquerque, F. S., ... & Isabel Bellocq, M. (2009). Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. *Ecography*, 32(2), 193-204.
- Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, 73(3), 1045-1055.
- Calow, P., & Petts, G. E. (Eds.). (1992). *The rivers handbook* (Vol. 1). London: Blackwell Scientific.
- Chang, H. (2008). Spatial analysis of water quality trends in the Han River basin, South Korea. *Water research*, 42(13), 3285-3304.

- Chang, H., Jung, I. W., Steele, M., & Gannett, M. (2012). Spatial patterns of March and September streamflow trends in Pacific Northwest streams, 1958–2008. *Geographical Analysis*, 44(3), 177-201.
- Chase, J. M., & Knight, T. M. (2013). Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. *Ecology letters*, 16(s1), 17-26.
- Cliff, A. D., & Ord, J. K. (1968). *The problem of spatial autocorrelation*. University of Bristol, Department of Economics and Department of Geography.
- Cliff, A., & Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical analysis*, 4(3), 267-284.
- Cliff, A. D., & Ord, J. K. (1981). *Spatial processes: models & applications*. London: Taylor & Francis.
- Cooper, S. D., Barmuta, L., Sarnelle, O., Kratz, K., & Diehl, S. (1997). Quantifying spatial heterogeneity in streams. *Journal of the North American Benthological Society*, 16(1), 174-188.
- Cressie, N. A. (1993). *Statistics for spatial data: Wiley series in probability and mathematical statistics*. New York, USA: John Wiley & Sons.
- Dale, M. R., & Fortin, M. J. (2002). Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, 9(2), 162-167.
- De Marco, P., Diniz-Filho, J. A. F., & Bini, L. M. (2008). Spatial analysis improves species distribution modelling during range expansion. *Biology Letters*, 4(5), 577-580.
- Diniz-Filho, J. A. F., Bini, L. M., & Hawkins, B. A. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global ecology and Biogeography*, 12(1), 53-64.
- Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global ecology and biogeography*, 16(2), 129-138.
- Dormann, C. F., McPherson, J., Araújo, M., Bivand, R., Bolliger, J., Carl, G., .... & Kühn, I. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5), 609-628.
- Dray, S., Péliissier, R., Coueron, P., Fortin, M. J., Legendre, P., Peres-Neto, P. R., ... & Dufour, A. B. (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs*, 82(3), 257-275.

- Fischer, M. M., & Griffith, D. A. (2008). Modeling spatial autocorrelation in spatial interaction data: an application to patent citation data in the European Union. *Journal of Regional Science*, 48(5), 969-989.
- Fortin, M. J., Dale, M. R., & Ver Hoef, J. M. (2002). Spatial analysis in ecology. *Wiley Statistics Reference Online*: <https://sites.google.com/site/jayverhoef/pubs/2002SpatialAnalysisEcology>
- Fotheringham, A. S., Brunson, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, England: John Wiley & Sons.
- Franczyk, J., & Chang, H. (2009). Spatial analysis of water use in Oregon, USA, 1985–2005. *Water Resources Management*, 23(4), 755-774.
- Franklin, J. (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.
- Getis, A., & Griffith, D. A. (2002). Comparative spatial filtering in regression analysis. *Geographical analysis*, 34(2), 130-140.
- Griffith, D. A. (1987). *Spatial autocorrelation*. Washington, D.C: Resource publications in Geography, A.A.G.
- Griffith, D. A. (2000). A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems*, 2(2), 141-156.
- Griffith, D. A., & Peres-Neto, P. R. (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology*, 87(10), 2603-2613.
- Griffith, D. A. (2009). Spatial autocorrelation. *International encyclopedia of human geography, 2009*, 308-316.
- Haining, R. P. (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.
- Hawkins, B. A. (2012). Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography*, 39(1), 1-9.
- Hengl, T., Heuvelink, G. B., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1-2), 75-93.
- Huang, J., Huang, Y., & Zhang, Z. (2014). Coupled effects of natural and anthropogenic controls on seasonal and spatial variations of river water quality during baseflow in a coastal watershed of Southeast China. *PloS one*, 9(3), e91528.

- Hubert, L. J., Golledge, R. G., & Costanzo, C. M. (1981). Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, 13(3), 224-233.
- Isaak, D. J., Peterson, E. E., Ver Hoef, J. M., Wenger, S. J., Falke, J. A., Torgersen, C. E., ... & Ruesch, A. S. (2014). Applications of spatial statistical network models to stream data. *Wiley Interdisciplinary Reviews: Water*, 1(3), 277-294.
- Johnson, L., & Gage, S. (1997). Landscape approaches to the analysis of aquatic ecosystems. *Freshwater Biology*, 37(1), 113-132.
- Jolliffe, I. T. (1986). Principal component analysis and factor analysis. In *Principal component analysis* (pp. 1-9). Springer, New York, NY.
- Kim, D. (2013). Incorporation of multi-scale spatial autocorrelation in soil moisture–landscape modeling. *Physical Geography*, 34(6), 441-455.
- Kim, D., Hirmas, D. R., McEwan, R. W., Mueller, T. G., Park, S. J., Šamonil, P., ... & Wendroth, O. (2016). Predicting the Influence of Multi-Scale Spatial Autocorrelation on Soil–Landform Modeling. *Soil Science Society of America Journal*, 80(2), 409-419.
- Kim, D., & Shin, Y. H. (2016). Spatial autocorrelation potentially indicates the degree of changes in the predictive power of environmental factors for plant diversity. *Ecological indicators*, 60, 1130-1141.
- Kissling, W. D., & Carl, G. (2008). Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography*, 17(1), 59-71.
- Koenig, W. D. (1999). Spatial autocorrelation of ecological phenomena. *Trends in Ecology & Evolution*, 14(1), 22-26.
- Legendre, P., & Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Vegetatio*, 80(2), 107-138.
- Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm?. *Ecology*, 74(6), 1659-1673.
- Legendre, P., & Legendre, L. (1998). Numerical ecology: second English edition. *Developments in environmental modelling*, 20.
- Legendre, P., & Legendre, L. F. (2012). *Numerical ecology* (Vol. 24). Elsevier.
- Lennon, J. J. (2000). Red-shifts and red herrings in geographical ecology. *Ecography*, 23(1), 101-113.

- Li, S., Gu, S., Tan, X., & Zhang, Q. (2009). Water quality in the upper Han River basin, China: the impacts of land use/land cover in riparian buffer zone. *Journal of hazardous materials*, 165(1-3), 317-324.
- Lichstein, J. W., Simons, T. R., Shriver, S. A., & Franzreb, K. E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological monographs*, 72(3), 445-463.
- Longley, P. A., & Batty, M. (Eds.). (1996). *Spatial analysis: modelling in a GIS environment*. John Wiley & Sons.
- Miralha, L., & Kim, D. (2018). Accounting for and Predicting the Influence of Spatial Autocorrelation in Water Quality Modeling. *ISPRS International Journal of Geo-Information*, 7(2), 64.
- Miller, J., Franklin, J., & Aspinall, R. (2007). Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling*, 202(3-4), 225-242.
- Miller, J. A. (2012). Species distribution models: Spatial autocorrelation and non-stationarity. *Progress in Physical Geography*, 36(5), 681-692.
- Netusil, N. R., Kincaid, M., & Chang, H. (2014). Valuing water quality in urban watersheds: A comparative analysis of Johnson Creek, Oregon, and Burnt Bridge Creek, Washington. *Water Resources Research*, 50(5), 4254-4268.
- NWQMC (National Water Quality Monitoring Council). Water Quality Portal. Available online: <http://www.waterqualitydata.us/> (accessed on 19 June 2017).
- O'sullivan, D., & Unwin, D. (2014). *Geographic information analysis*. John Wiley & Sons.
- Peres-Neto, P. R., & Legendre, P. (2010). Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography*, 19(2), 174-184.
- Peterson, E. E., Ver Hoef, J. M., Isaak, D. J., Falke, J. A., Fortin, M. J., Jordan, C. E., ... & Som, N. (2013). Modelling dendritic ecological networks in space: an integrated network perspective. *Ecology letters*, 16(5), 707-719.
- Praskievicz, S., & Chang, H. (2009). A review of hydrological modelling of basin-scale climate change and urban development impacts. *Progress in Physical Geography*, 33(5), 650-671.
- 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.

- Pringle, C. (2003). What is hydrologic connectivity and why is it ecologically important?. *Hydrological Processes*, 17(13), 2685-2689.
- Pringle, C. M. (2001). Hydrologic connectivity and the management of biological reserves: a global perspective. *Ecological Applications*, 11(4), 981-998.
- Pringle, C. M., Naiman, R. J., Bretschko, G., Karr, J. R., Oswood, M. W., Webster, J. R., ... & Winterbourn, M. J. (1988). Patch dynamics in lotic systems: the stream as a mosaic. *Journal of the North American Benthological Society*, 7(4), 503-524.
- Sayre, R. (1984). *A new map of standardized terrestrial ecosystems of the conterminous United States* (No. 1768). US Department of the Interior, US Geological Survey.
- Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database. Available online: <https://sdmdataaccess.sc.egov.usda.gov> (accessed on 22 June 2017).
- 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.
- United State Geological Survey. The National Map Elevation Products (USGS TNM 3DEP). 2017. Available online: <https://viewer.nationalmap.gov> (accessed on 22 June 2017).
- United State Geological Survey. The National Map, 2011, National Land Cover Database (USGS TNM-NLCD). Available online: <https://viewer.nationalmap.gov> (accessed on 22 June 2017).
- United State Geological Survey. USGS National Hydrography Dataset (NHD) Downloadable Data Collection. 2016. Available online: <http://nhd.usgs.gov> (accessed on 22 June 2017).
- United States Department of Agriculture, Natural Resources Conservation Service. Part 630 Hydrology—Hydrologic Soil Groups. In National Engineering Handbook; Title 210-VI [Online] U.S. Department of Agriculture, Soil Conservation Service (SCS), Washington D.C., 2009; pp. 1–7. Available online: <https://directives.sc.egov.usda.gov> (accessed on 08 October 2017).
- Václavík, T., Kupfer, J. A., & Meentemeyer, R. K. (2012). Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). *Journal of Biogeography*, 39(1), 42-55.
- Valett, H. M., Thomas, S. A., Mulholland, P. J., Webster, J. R., Dahm, C. N., Fellows, C. S., ... & Peterson, C. G. (2008). Endogenous and exogenous control of

ecosystem function: N cycling in headwater streams. *Ecology*, 89(12), 3515-3527.

Ver Hoef, J. M., Peterson, E., & Theobald, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological statistics*, 13(4), 449-464.

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, 365-373.

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.

## VITA

# Lorrayne Miralha

### Education

<b>University of Kentucky, Lexington, KY</b> M.A. in Physical Geography	<i>Expected– May 2018</i>
<b>Federal Rural University of Rio de Janeiro – RJ, Brazil</b> B.S. in Forest Engineering	<i>Mar. 2016</i>
<b>Oregon State University, Corvallis, OR</b> Exchange Program in Forest Engineering – Sponsored by CAPES	<i>Mar. 2014 – Aug. 2015</i>

### Experience

Research Assistant, NSF Project (University of Kentucky), KY, US	<i>Present</i>
Teaching Assistant, Earth's Physical Environment (University of Kentucky), KY, US	<i>Present</i>
Ad Hoc Technologies, New Maps + Committee – Graduate Primary Representative (University of Kentucky), KY, US	<i>Present</i>
Simple Day Awards Committee – Graduate Alternate Representative (University of Kentucky)	<i>Present</i>
Geography Graduate Student Union (GGSU) – Board Treasurer (University of Kentucky)	<i>2016-2017</i>
Administrative Assistant, City Hall (Rio de Janeiro – Brazil)	<i>2011-2016</i>
Intern, Reforestation Research, and Studies Laboratory (UFRRJ) – RJ/Brazil	<i>Sep 2015 - Dec 2015</i>
Research Assistant, Remote Sensing Laboratory (Oregon State University) – OR, US	<i>Mar2015 - Aug 2015</i>
Founding Member Marketing Director, Junior Company in Agrobusiness (UFRRJ) RJ/Brazil	<i>2013 - 2014</i>
English Teaching Assistant, EXCEL Language School - RJ/Brazil	<i>2006 - 2012</i>
Bilingual Secretary, Odebrecht Oil & Gas - RJ/Brazil	<i>2008 - 2010</i>
Computer Teaching Assistant, Cedaspy Informatic Course – RJ/Brazil	<i>2006 - 2007</i>



## Publication

- Miralha, L.; Kim, D. Accounting for and Predicting the Influence of Spatial Autocorrelation in Water Quality Modeling. ISPRS Int. J. Geo-Inf. 2018, 7, 64.

## Honors, Awards & Scholarships

- Geography Graduate Program - Full Tuition Scholarship – University of Kentucky (2016 – 2018)
- Honor Hall - Outstanding Academic Performance – Oregon State University – Fall 2014; Winter 2015; Spring 2015
- Exchange Program Award – Sponsored by CAPES – March 2014 – August 2015