7-6-2020

Spatial Statistical Approaches to Water Quality Modelling

Janardan Mainali
Portland State University

Recommended Citation

Mainali, Janardan, "Spatial Statistical Approaches to Water Quality Modelling" (2020). *Dissertations and Theses.* Paper 5509.
https://doi.org/10.15760/etd.7383

Spatial Statistical Approaches to Water Quality Modelling

by

Janardan Mainali

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
in
Earth, Environment and Society

Dissertation Committee:
Heejun Chang, Chair
Jennifer Morse
Daniel Taylor- Rodriguez
Jeremy Spoon

Portland State University
2020

Abstract

This dissertation aims to advance the existing knowledge related to spatial modeling of water quality by exploring and introducing innovative approaches to different spatial conceptualizations for water quality modeling and incorporating upstream-downstream relations in geographically weighted regression. By carrying out a systematic literature review of four different classes of spatial models in Chapter One, this dissertation identifies the following major research gaps: lack of incorporation of multiscale processes, not enough emphasis on spatial weights matrices, and unavailability of upstream-downstream relationships in geographically weighted regressions. Chapters Two and Three were designed to address these gaps in the literature. In Chapter Two, different spatial conceptualizations of sampling sites were compared based on their capacity to predict dissolved oxygen and electrical conductivity utilizing geographic information system derived explanatory variables in rivers of the Setikhola watershed in central Nepal. The model strengths are better while considering graph types close to the stream network structure for dissolved oxygen. The graph types that account for neighbors in all directions are better suited for electrical conductivity modeling. In Chapter Three, this dissertation demonstrates that a successful geographically weighted regression model could be developed using an upstream distance matrix that has comparable model strength with that of standard Euclidean distance weighted geographically weighted regression. The human impacts as population density and increased sand and gravel cover can be detected impacting water quality in the study watershed. The relationships between socio-environmental factors and water quality and their spatial interrelationships identified in the second chapter shed light on the source, mobilization, and transport of dissolved oxygen and electrical conductivity and can assist the water quality management endeavor. The local insights obtained from the upstream distance weighted geographically weighted regression of the third chapter help understand fine-scale impacts of socio-environmental and biophysical factors on water quality and assist in designing locally specific water quality management efforts.

Dedication

I dedicate my dissertation to the high school in my village (Sundrawati, Dolakha, Nepal) which my parents and their generation built with their manual labor and led me to submit this dissertation.

Acknowledgments

I would like to thank adviser, Dr. Heejun Chang, for his continuous support throughout the dissertation. Thanks, are also due to my dissertation committee members Dr. Jennifer Morse, Dr. Daniel Taylor Rodriguez, and Graduate School Representative Dr. Jeremy Spoon. I also like to thank Dr. Melissa Lucash for serving on my dissertation preliminary exam committee. I thank Dr. Barbara Brower for providing review comments on some of the chapters. I also owe a gratitude to Dr. Jay ver Hoef, a brief discussion with whom led to materialize third chapter of this dissertation.

This dissertation journey was possible because of various funding agencies. The National Science Foundation Grant #1560907 offered me research assistantships to develop this dissertation. The fieldwork portion of the dissertation was supported by Rufford Small Grant Foundation grant # 26386-1, WWF Nepal grant #GM-66, and Idea Wild.

I thank Mr. Hemanta Dhakal for coordinating field visits. Field team members Kanchan Upadhyay, Birkha Raj Gurung, and Achyut Bhandari deserve a huge accolade for their enthusiastic efforts.

I express gratitude to officials of the Department of Geography, EES Ph.D. program, and Office of Graduate Studies of Portland State University for their support during this journey.

I am grateful to fellow lab mates Benjamin Fahy, Katelyn Michelson, Junjie Chen, Alexander Ross, Rebecca Talbot, and Arun Pallathadka for their friendship and support. I also like to thank my friend Muhammad Ilyas who facilitated my move to Portland.

I like to thank my parents and sisters for facilitating my study to the terminal degree.

Finally, I am indebted to my wife Anjana Dhakal for her support over my Ph. D. journey.

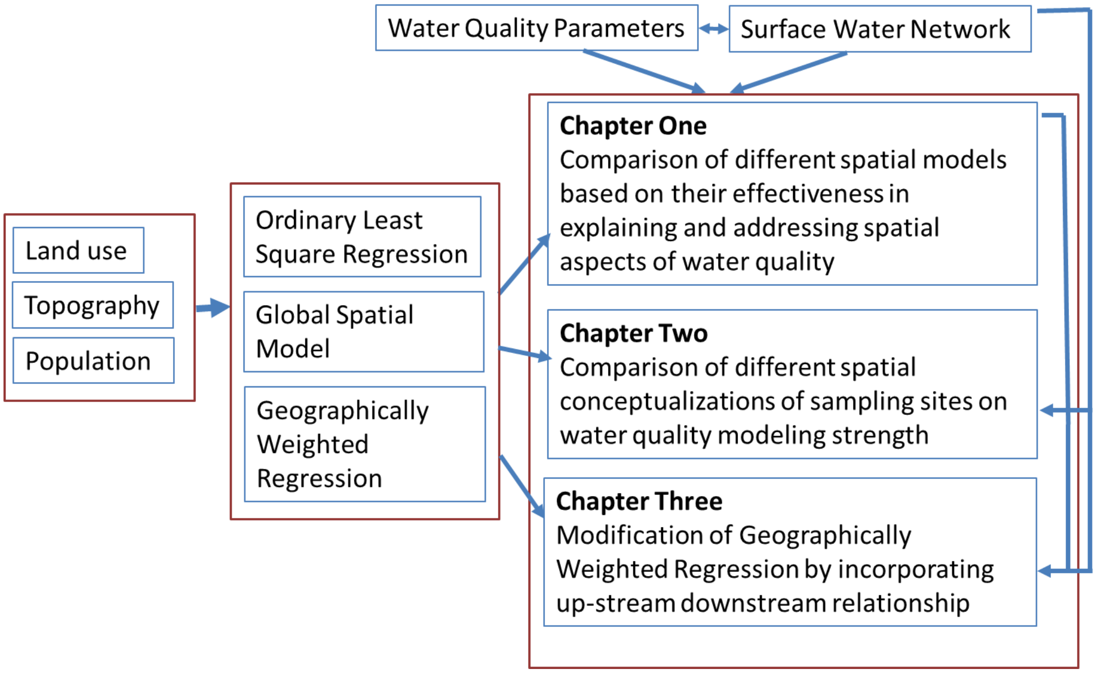
Table of Contents

List of Tables

List of Figures

Introduction

This dissertation aims to advance the existing knowledge related to spatial modeling of water quality by exploring and introducing innovative approaches to different spatial conceptualizations for water quality modeling and incorporating upstream-downstream relations in geographically weighted regression. I use these approaches to investigate the water quality of one of the Himalayan watersheds of central Nepal to understand the impacts of socio-environmental factors in surface water quality and develop models to explore and predict water quality when data are not available (Figure 0-1).



Water Quality Parameters
Surface Water Network
Land use
Topography
Population
Ordinary Least Square Regression
Global Spatial Model
Geographically Weighted Regression
Chapter One
Comparison of different spatial models based on their effectiveness in explaining and addressing spatial aspects of water quality
Chapter Two
Comparison of different spatial conceptualizations of sampling sites on water quality modeling strength
Chapter Three
Modification of Geographically Weighted Regression by incorporating up-stream downstream relationship

Figure 0-1: Conceptual framework of the dissertation

Water quality is defined as the physical, chemical, and biological characteristics of water based on the observation and measurement of various parameters like concentration of salts, nutrients, or presence of a species to determine its suitability for a particular use (USGS, 2009). Different substances enter surface water bodies by various pathways such as overland flow, in-stream flow, or atmospheric deposition (Lintern et al., 2018). Within the water bodies also these substances undergo various physical, chemical, and biological changes leading to different water quality characteristics in different sections of the surface water bodies (Lintern et al., 2018; Zhai et al., 2014). These various physical, biological, and hydrological processes are impacted by human, natural, and climatic interactions (Mainali and Chang, 2018; Mouri et al., 2011; Shen et al., 2015; Wang et al., 2013). Human interventions as increased agricultural and urban land-use and modification of river environments are the main sources of water quality deteriorations (Bu et al., 2014; Finlay et al., 2013).

The impacts of these various social, climatic, and natural factors on water quality are explored using various mathematical modeling approaches. Several processes based or statistical modelling approaches like SPARROW, QUAL, BASINS, WASP, QUASAR, MIKE, and GWR are used in water quality modelling (Brunsdon et al., 1998; Schwarz, 2006.; Wang et al., 2013). Among them, different regression modeling approaches are commonly used to establish relationships between multiple explanatory variables and a water quality response variable measured from the surface water body (Ullah et al., 2018). Different landscape characteristics like land use, land management, slope, soils, or geology as explanatory variables are used to model water quality (Lintern et al., 2018). The population density, socioeconomic status, and other social-environmental variables are also used to assess their impacts on surface water bodies

(Chen and Lu, 2014; Mainali and Chang, 2018). The traditional regression models like Ordinary Least Square Regressions (OLS) are not statistically valid when there is spatial autocorrelation of model residuals. The presence of spatial autocorrelation results in a spatially biased trend and violates the assumption of random and independent samples and un-correlated residuals of most standard parametric statistical procedures (Cliff and Ord, 1972; Legendre, 1993; Sokal and Oden, 1978). On the other hand, OLS models do not incorporate seemingly obvious spatial interrelationships between neighboring and upstream-downstream data points in stream environments. Several spatial regression approaches, which account for such spatial dependence, have been used in water quality modeling. These approaches include spatial lag model, spatial error model, geographically weighted regression (GWR), spatial eigenvector mapping, and spatial-stream-network based model (Blanchet et al., 2008; Borcard and Legendre, 2002; Brunsdon et al., 1998; Getis and Griffith, 2002; Ver Hoef et al., 2018; Ver Hoef and Peterson, 2010).

In this dissertation, I attempt to examine various spatial modeling approaches to find novel ways to incorporate spatial interrelationships, collect first-hand water quality data, extract explanatory variables, and develop models to demonstrate novel spatial statistical methods. In the first chapter, I conducted a review of recent literature to compare different statistical models based on their effectiveness in explaining and addressing spatial aspects of water quality. I, along with coauthors, specifically examine spatial autocorrelation of the water quality parameters, residual spatial autocorrelation, use of weights matrix, and incorporation of directional spatial processes in the model and attempt to identify knowledge gaps related to spatial modeling of water quality.

In the second chapter, I tackle one of the research gaps in the spatial modeling literature, which is a comparison of different spatial conceptualizations of sampling sites on water quality modeling. I compare five different spatial conceptualizations using graph theories to evaluate their effectiveness in modeling Dissolved Oxygen and Electrical Conductivity at two different spatial scales. I also explore spatial patterns of Dissolved Oxygen and Electrical Conductivity in the Setikhola Watershed of Central Nepal by collecting first-hand water quality data. I further explore how different landscape features like land cover, topography, and population density affect the water quality in the watershed.

In chapter three, I attempt to modify Geographically Weighted Regression by incorporating up-stream downstream relationships. This chapter builds upon the findings of the first chapter, which discovered that the stream network structure and up-stream down-stream relationships are not yet incorporated in geographically weighted regression. I use a spatial stream network model to extract flow connected distance matrix to run geographically weighted regression and compare the model outputs of standard and upstream distance weighted geographically weighted regression.

Introduction References

Blanchet, F.G., Legendre, P., Borcard, D., 2008. Modelling directional spatial processes in ecological data. Ecol. Model. 215, 325–336. https://doi.org/10.1016/j.ecolmodel.2008.04.001

Borcard, D., Legendre, P., 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol. Model. 153, 51–68.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. R. Stat. Soc. Ser. Stat. 47, 431–443.

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

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

Cliff, A., Ord, K., 1972. Testing for spatial autocorrelation among regression residuals. Geogr. Anal. 4, 267–284.

Finlay, J.C., Small, G.E., Stener, R.W., 2013. Human Influences on Nitrogen Removal in Lakes. Science 342, 247–250. https://doi.org/10.1126/science.1242575

Getis, A., Griffith, D.A., 2002. Comparative Spatial Filtering in Regression Analysis. Geogr. Anal. 34, 130–140. https://doi.org/10.1111/j.1538-4632.2002.tb01080.x

Legendre, P., 1993. Spatial Autocorrelation: Trouble or New Paradigm? Ecology 74, 1659–1673. https://doi.org/10.2307/1939924

Lintern, A., Webb, J.A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P., Western, A.W., 2018. Key factors influencing differences in stream water quality across space. Wiley Interdiscip. Rev. Water 5, 1–31. https://doi.org/10.1002/wat2.1260

Mainali, J., Chang, H., 2018. Landscape and anthropogenic factors affecting spatial patterns of water quality trends in a large river basin, South Korea. J. Hydrol. 564, 26–40. https://doi.org/10.1016/j.jhydrol.2018.06.074

Mouri, G., Takizawa, S., Oki, T., 2011. Spatial and temporal variation in nutrient parameters in stream water in a rural-urban catchment, Shikoku, Japan: Effects of land cover and human impact. J. Environ. Manage. 92, 1837–1848. https://doi.org/10.1016/j.jenvman.2011.03.005

Schwarz, G.E., 2006. The SPARROW Surface Water-Quality Model: Theory, Application and User Documentation--Part 1 29.

Shen, Z., Hou, X., Li, W., Aini, G., Chen, L., Gong, Y., 2015. Impact of landscape pattern at multiple spatial scales on water quality: A case study in a typical urbanised watershed in China. Ecol. Indic. 48, 417–427. https://doi.org/10.1016/j.ecolind.2014.08.019

Sokal, R.R., Oden, N.L., 1978. Spatial autocorrelation in biology: 1. Methodology. Biol. J. Linn. Soc. 10, 199–228.

Ullah, K.A., Jiang, J., Wang, P., 2018. Land use impacts on surface water quality by statistical approaches. Glob. J. Environ. Sci. Manag. 4, 231–250. https://doi.org/10.22034/gjesm.2018.04.02.010

USGS, 2009. What Does “Water Quality” Mean? [WWW Document]. Water Qual. Inf. Top. URL https://www.usgs.gov/special-topic/water-science-school/science/water-quality-information-topic (accessed 4.14.20).

Ver Hoef, J.M., Peterson, E.E., 2010. A Moving Average Approach for Spatial Statistical Models of Stream Networks. J. Am. Stat. Assoc. 105, 6–18.

Ver Hoef, J.M., Peterson, E.E., Hooten, M.B., Hanks, E.M., Fortin, M.-J., 2018. Spatial autoregressive models for statistical inference from ecological data. Ecol. Monogr. 88, 36–59. https://doi.org/10.1002/ecm.1283

Wang, Q., Li, S., Jia, P., Qi, C., Ding, F., 2013. A Review of Surface Water Quality Models. Sci. World J. 2013, 1–7. https://doi.org/10.1155/2013/231768

Wang, R., Xu, T., Yu, L., Zhu, J., Li, X., 2013. Effects of land use types on surface water quality across an anthropogenic disturbance gradient in the upper reach of the Hun River, Northeast China. Environ. Monit. Assess. 185, 4141–4151. https://doi.org/10.1007/s10661-012-2856-x

Zhai, X., Xia, J., Zhang, Y., 2014. Water quality variation in the highly disturbed Huai River Basin, China from 1994 to 2005 by multi-statistical analyses. Sci. Total Environ. 496, 594–606. https://doi.org/10.1016/j.scitotenv.2014.06.101

Chapter One: A Review of Spatial Statistical Approaches to Modeling Water Quality

Janardan Mainali, Heejun Chang, and Yongwan Chung

Progress in Physical Geography: Earth and Environment

Mainali, J., Chang, H., Chun, Y., 2019. A review of spatial statistical approaches to modeling water quality. Progress in Physical Geography: Earth and Environment 43, 801–826. https://doi.org/10.1177/0309133319852003

A Review of Spatial Statistical Approaches to Modeling Water Quality

Janardan Mainali and Heejun Chang

Abstract

We review different regression models related to water quality that incorporate spatial aspects in their model. Spatial aspects refer to the location of different sites and are usually characterized by the distance between different points and directions by which they are related to each other. We focus on spatial lag and error, spatial eigenvector-based, geographically weighted regression, and spatial stream network-based models. We evaluated different studies using these methods based on how they dealt with clustering (spatial autocorrelation) of response variables, incorporated those clustering in the error (residual spatial autocorrelation), used multiscale processes, and improved the model performance. The water quality-based regression modeling approaches are shifting from straight-line-distance-based spatial relations to upstream-downstream relations. Calculation of spatial autocorrelation and residual spatial autocorrelation was dependent upon the type of spatial regression used. The weights matrix is used as available in the software and most of the studies did not attempt to modify it. Different scale processes like certain distance from rivers vs consideration of entire watersheds are dealt separately in most of the studies. Generally, the capacity of the predictor variables to predict the response variable significantly improves when spatial regressions are used. We identify new research directions in terms of spatial considerations, weights matrix construction, inclusion of multiscale processes, and identification of predictor variables in such models.

Keyword: Water quality, hydrology, watershed, spatial statistics, spatial autocorrelation, scale

Introduction

Water quality defined as the physical, chemical, and biological characteristics of water is directly associated with the human and ecosystem health. The water quality, itself is dependent on various factors, including land cover, land use, land management, atmospheric deposition, geology and soil type, climate, topography, and catchment hydrology (Lintern et al., 2018). Water quality parameters vary across space and time because of variations in these different factors. For effective water quality management, it is crucial to understand these factors and the pathways by which they affect water quality. Understanding spatial patterns of water quality parameters and factors affecting them, therefore, is crucial in pinpointing locations of interventions for improving water quality in surface water bodies.

The most common approach of water quality research involves the statistical method, which typically process raw quantitative data using mathematical models, formula, and techniques to extract information and generate meaningful output (Nature Statistics, 2019). Regressions are most common statistical methods to understanding the relationship between water quality and watershed characteristics (Chang, 2008; Shi et al., 2016; Zhou et al., 2012). Regression approaches may or may not include spatial aspects of water quality parameters (Ullah et al., 2018). Spatial aspects refer to location and relative position to each other usually analysed using different spatial statistics. A relatively new sets of spatial statistical approaches, which typically extend from linear regression analysis, attempt to incorporate spatial processes to identify environmental and spatial determinants of water quality in surface water (Blanchet et al., 2008; Legendre, 1993).

Many studies have examined spatial aspects of water quality (e.g., spatial autocorrelation and distribution of high and low values along a river network.) using various modeling techniques to explore the effect of landscape-level variables in the water quality. These studies include several review papers that synthesized different aspects of water quality research. Giri and Qiu (2016) reviewed the current understanding of the relationship between land use and water quality, while Ullah et al. (2018) examined different statistical approaches to modeling water quality using land use types as predictor variables. Lintern et al. (2018) conducted a comprehensive review of key factors affecting the spatial patterns of water quality, while Guo et al., (2019) reviewed various factors affecting temporal patterns of water quality. Isaak et al., (2014) conducted a review of research on a group of spatial statistics, spatial stream network based models. However, there is not any comprehensive review related to the spatial aspects of water quality modeling that offers water quality researchers a way to understand the basic concept of the spatial statistics and help them choose an appropriate modeling approach.

We carry out this review to compare different statistical models based on their effectiveness in addressing spatial aspects of water quality. We specifically examine spatial autocorrelation of the water quality parameters, residual spatial autocorrelation (RSAC), use of weights matrix, and incorporation of directional spatial processes in the model. In the first section, we discuss how these methodologies have evolved, while in the later section we perform a systematic literature review to identify knowledge gaps related to spatial autocorrelation, use of multiscale processes, and directional spatial processes. We review papers related to spatial lag and error model, spatial eigenvector based models, geographically weighted regression, and spatial stream network based models. We recognize that there are other spatial modeling approaches which are not covered in this review, including spatial kriging, P-splines, and several spatial autoregressive models (e.g., McLean et al., 2019).

II Spatial Statistical Approaches in Water Quality Studies

In the watershed science, watershed, basin, or sub-basin are considered units of analysis. Extracting predictor variables that affect surface water quality mostly involves consideration of entire watershed. Several ways exist to incorporate different scales in the water quality modeling endeavor (Allan, 2004; Mainali and Chang, 2018). One of the most common involves creating a buffer of a specified distance from stream or lakes. Some studies also use a threshold distance upstream from sampling point (e.g., Shi et al., 2017). Some new methods provide higher weight to the landscape factors close by the streams based on Euclidean (straight line) distance, flow distance, or flow accumulation (Grabowski et al., 2016; King et al., 2005; Peterson et al., 2011).

Spatial variations in the watershed properties draining into the river results in the variable water quality across different parts of the river, which typically lead to a specific spatial pattern of water quality. As nearby places are more alike than distant spaces (Tobler, 1970), there might be a cluster of high or low values of water quality parameters. This phenomenon, spatial autocorrelation, is a measure of whether a data value of one location is independent of data values of other locations (Sokal and Oden, 1978). Spatial autocorrelation can be positive when similar data values are close to each other, or negative when dissimilar data values are neighbored (Legendre, 1993; Sokal and Oden, 1978). Spatial autocorrelation opens new avenues, to statistically analyze, seemingly obvious but ignored spatial pattern of water quality and its relations with the watershed attributes (Legendre, 1993).

A family of statistical tools is being used to analyze spatial autocorrelation among sampling stations. Moran's *I* is the most commonly used measure to evaluate the pattern of the attributes as clustered, dispersed, or random in space. This is a global statistics, one that offers a single set of statistics for the entire set of data. Moran's *I* has been used to analyze different water quality attributes in order to identify whether the water quality attributes show any global pattern of spatial dependence (Liu et al., 2016; Miralha and Kim, 2018; Pratt and Chang, 2012). As Moran's *I* statistics only offer information about the level of spatial autocorrelation for an entire set of data, we cannot use it to identify any local clusters. There are a few statistical approaches developed to identify local clusters in spatial data and are also being used to explore clusters of sites with degraded or not-degraded water quality. Getis-Ord's G_i and local Moran's *I* are commonly used in such local statistics (Anselin, 1995; Getis and Ord, 1992). These methods identify whether or not similar high or low values are clustered together locally and identify those clusters in geographical space. Many water quality analyses works have used these statistics to explore local relations in a sampling space (Brody et al., 2005; Mainali and Chang, 2018; Tu and Xia, 2008).

The spatial autocorrelation in any data is associated with spatial dependence among different neighboring data points, resulting in a spatially biased trend and violating the assumption of independence of most standard parametric statistical procedures (Cliff and Ord, 1972; Legendre, 1993; Sokal and Oden, 1978). In regression analysis, biases due to such neighboring data points need to be accounted for, as they can produce autocorrelated residuals (differences between actual and predicted values) and ultimately inflate Type I error, leading to wrongfully rejecting the null hypothesis (Bini et al., 2009; Cliff and Ord, 1972; Miralha and Kim, 2018). It is not possible to account for such influence only using traditional simple linear regression approaches that assume that data points are randomly distributed in the sampling space, and that model residuals are not autocorrelated. Several spatial regression approaches that account for such spatial dependence and are being used in water quality modeling, including spatial lag model, spatial error model, Geographically Weighted Regression (GWR), spatial eigenvector mapping, and spatial-stream-network based model (Blanchet et al., 2008; Borcard and Legendre, 2002; Brunsdon et al., 1998; Getis and Griffith, 2002; Ver Hoef et al., 2018; Ver Hoef and Peterson, 2010).

III Spatial Weight Matrix and Spatial Regression Models in Water Quality Studies

1 Spatial Weight Matrix

The spatial dependence between sampling points is formally expressed as a weights matrix and is a necessary element of spatial regression models (Anselin, 2001; Getis and Aldstadt, 2004). Each spatial weight refers to the relative influence of different spatial units under consideration to the candidate spatial unit. These weights matrices can be defined in several ways, according to spatial interactions among different factors under consideration, and the hypotheses of interest (Sokal and Oden, 1978). The most essential aspect of the weights matrix is defining a neighborhood set for each location. The neighborhood sets are specified for each location as the row and the neighbors as the columns in a matrix. Non-zero weight is assigned when observations are within a given number of nearest neighbors or specified distance. In the spatial statistics literature, the weight can be specified based on Euclidean distance, economic distance, number of nearest neighbors, or empirical flow matrices (Anselin, 2001). The weight matrices use several approaches to incorporate the effect of adjacent observations. Sometimes, a certain number of nearest neighbors is used, while in other cases only observations within a certain distance is used with the same weight to all the observations within that distance (Figure 1-1). Spatial regression models usually differ in terms of conceptualizing the spatial relationships usually through the weights matrix. In this section, we discuss how these different spatial regression approaches are conceptualized and used in water quality modeling endeavors (Figure 1-2).

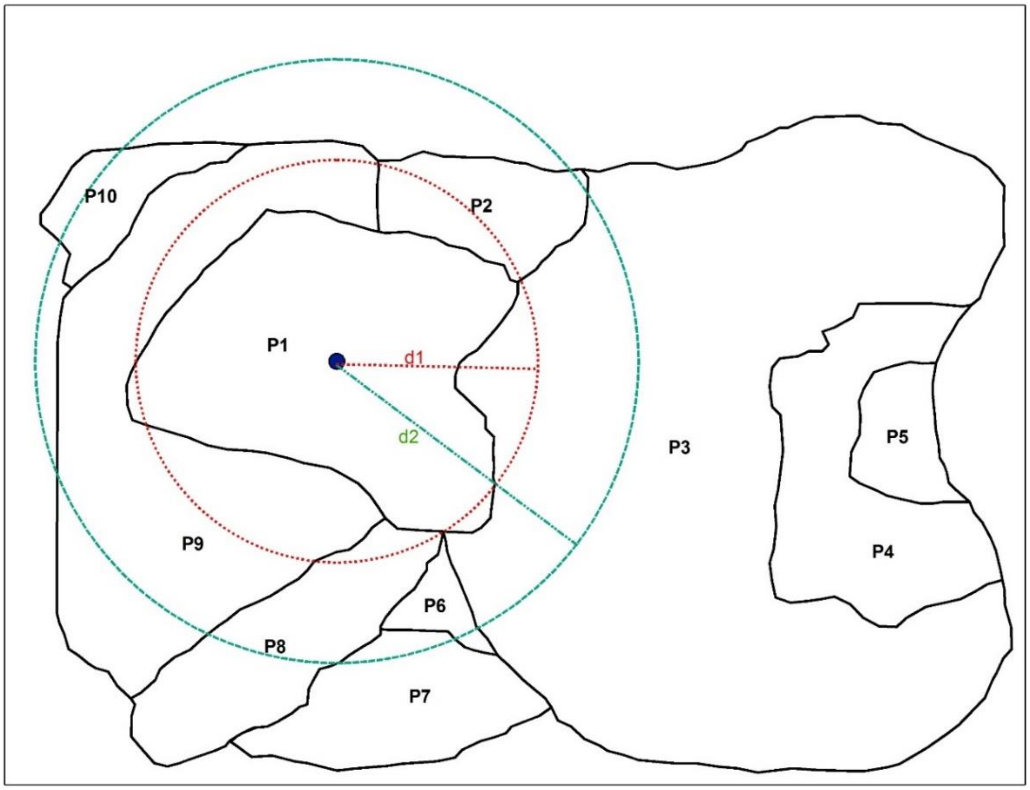


Figure 1-1: Conceptualization of different weights matrices

A spatial weights matrix is created based on whether polygons share a common boundary or not (a binary decision with 0 or 1). For example, for P1, four polygons (i.e. P2, P3, P6, P8, and P9) are considered as neighbors based on Queen's connectivity), or P6 will not be included if a zero-distance common boundary (i.e. point connectivity) does not count (Rook's connectivity). A contiguity-based spatial weights matrix can be specified with either the length of a common boundary or the area of an adjacent polygon instead of 0–1 binary values. For example, for the length of a common boundary, P9 has the longest common boundary with P1 and, thus, will have the largest weight, while P8 shares the shortest common boundary with PI and has the smallest weight. For the area of an adjacent polygon, P3 is the largest adjacent polygon of P1 and will have the largest weight, while P2, which is the smallest adjacent polygon of P1, has the smallest weight. 2) Nearest neighbor: sometimes weight can also be provided based on the numbers of neighbors for each candidate polygon (k-nearest neighbor). If we only use one closest neighbor, polygons (first order) defined in the Queen's case (P2, P3, P6, P8, and P9) are considered. If two nearest neighbors (second order) are considered, in addition to the polygons adjacent to P1, the polygons sharing

a boundary with those (P2, P3, P6, P8, and P9) are also included during the weights matrix construction for the candidate polygon (P1), which results in the inclusion of P4, P7, and P10, but not P5. Nearest neighbors (which are often called k-nearest neighbors) are specified with a fixed number of neighbors. It is often adoptively utilized for a case in which observations are not (relatively) evenly distributed. For example, one remote point (it is often specified for points) may not have any neighbor, which is a problem in spatial analysis. To avoid this problem, k-nearest neighbors can be utilized. 3) Threshold distance: spatial neighbors can be specified based on a preset distance from the centroid of a polygon. Here, with a threshold distance d1, polygons inside the circle of radius d1 are considered as spatial neighbors for polygon P1. In this case, P1 has four neighbors: P2, P3, P8, and P9. If d2 is used for a threshold distance, all polygons but P4 and P5 are neighbors of P1 and have a non-zero weight

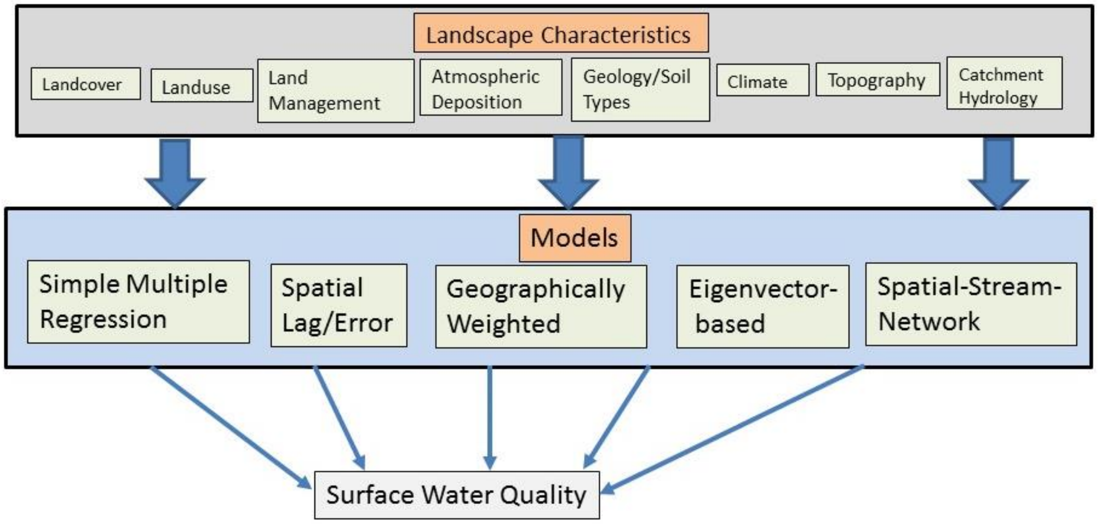


Figure 1-2: Use of different landscape characteristics (Lintern et al., 2018) in different spatial statistical models reviewed in this chapter

2 Spatial Lag and Error Model

Spatial lag models and spatial error models are the commonly used global regression models that account for spatial dependence among observations in a model specification. Global models refer to the regression models that produce a single set of model statistics for a set of data. Spatial lag model (Anselin, 1988, 2001) is applied when response variables suffer from significant spatial autocorrelation. A spatially lagged variable is created by averaging the values of the response variable at neighboring locations (Figure 1-3a). The spatial lag model includes a spatially lagged dependent variable with a weights matrix to account for the spatial autocorrelation. Such a weights matrix is often constructed without consideration of a stream network, so it tends to have more neighboring sites than one that of a stream network (Figure 1-3a). Spatial error model (Anselin 2001) is used when model residuals suffer from significant spatial autocorrelation. This is similar to the spatial lag model except that it accounts for spatial autocorrelation in the error term.

Several researchers have been using these methods to model water quality and reported a general improvement in model performance when such spatial models are used (Chang, 2008; Huang et al., 2016; Miralha and Kim, 2018). This improvement in the model performance typically relates to the degree of spatial autocorrelation and residual spatial autocorrelation (Kim et al., 2016; Kim and Shin, 2016; Miralha and Kim, 2018).

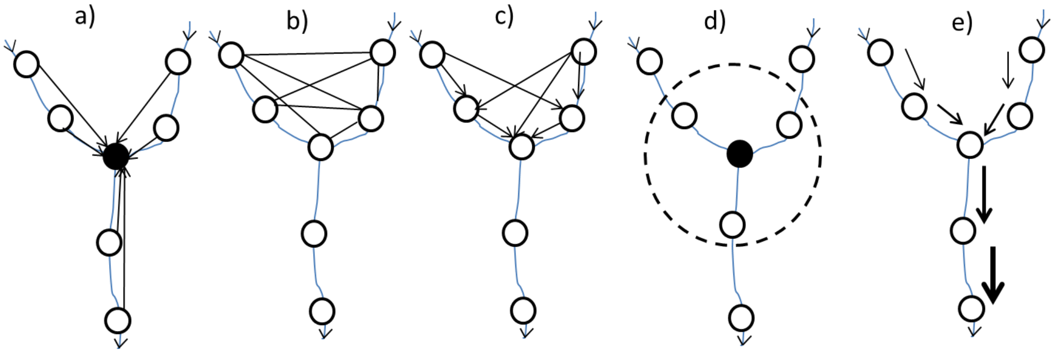


Figure 1-3: Spatial relations among sampling stations for a spatial weight matrix creation in different types of spatial modeling for surface water quality. The black arrows refer to directionality of the spatial relations and the dotted circle represents a certain bandwidth.

a) Spatial lag and/or error model – both upstream and downstream stations affect a station of interest. b) Moran's eigenvector maps – all surrounding stations are considered with no directionality between upstream and downstream stations modified from Sharma et al., (2011) , c) Asymmetrical eigenvector maps – only upstream stations are considered, but stations in different tributaries could affect each other, d) Geographically Weighted Regression – only neighbors within a threshold distance are considered with no specific upstream and downstream relationships, and e) Spatial Stream Network Based model. Arrows in spatial stream network models refer to the direction of the relation and moving average function. The width of the arrow refers to the strength of the influence for each potential neighborhood location. Spatial autocorrelation occurs when the moving average function overlaps. Modified from (Peterson and Hoef, 2010). Once a spatial weights matrix specified, all of the four modeling method can be used with the spatial weights matrix

3 Geographically Weighted Regression

Global spatial regression models, such as spatial lag models and spatial error models, are used to develop a spatially rectified global regression model by accounting for the spatial dependence of an entire dataset. They only produce a single set of statistics for the entire dataset under consideration, hence are are a member of global spatial regression models. In reality, a relationship between predictors and a response variable can vary within a catchment, and the strength of those relations might also be different across regions. In order to address this issue, Geographically Weighted Regression (GWR) can be used to allow model coefficients to vary for each observation and create a set of local models based on the location of sampling sites (Brunsdon et al., 1998). The observed data included in each local model are geographically weighted, depending on the proximity of the location and are used to estimate local R^2 and coefficients for each sample observation. The number of samples included for each data point is defined using a bandwidth function (Figure 1-3d). Although a fixed-distance band can also be used, a flexible bandwidth that adapts to the spatial pattern of the data can be more effective, particularly when data are not evenly distributed over space (Fotheringham et al., 2002). During the modelling process, the nearby data points are weighted more heavily than those from more remote locations using a kernel function. GWR is increasingly used in water quality modeling not only to estimate the model parameters but also to explore the variabilities of those relationships in different watersheds (Chen et al., 2016; Pratt and Chang, 2012; Chang and Psaris 2013; Tu, 2011).

4 Moran Eigenvector Maps and Spatial Filtering

Eigenvector-based models are spatial models in which the vectors are derived using neighborhood criteria or distance with neighbors. In these models a matrix is constructed based on the geographical distance between locations. This matrix is transformed into eigenvectors by eigenfunction decomposition (Figure 1-3b). This method was originally proposed by Borcard and Legendre (2002) as the principal component of neighborhood matrix (PCNM), also called Moran's Eigenvector Maps (MEM). This method incorporates spatial autocorrelation in modeling ecological processes. Eigenvectors corresponding to positive eigenvalues are used as spatial descriptors in regression or canonical analysis (Borcard and Legendre, 2002). Vrebos et al. (2017) modeled water quality of 75 stations in the Kleine Nete Catchment in Northern Belgium and reported that about 30 percent of variation was explained by catchment land cover while about 11 percent was explained by spatial Eigenvectors that of MEM.

There are both distance-based eigenvector maps and spatial filtering based upon a geographic connectivity matrix (Borcard and Legendre, 2002; Getis and Griffith, 2002; Griffith, 2010; Griffith and Peres-Neto, 2006). Eigenvector-based spatial filtering is used to separate spatial effects in regression modeling from model residuals so that a standard regression model can be used without suffering from spatial autocorrelation (Getis and Griffith, 2002). Similar to the eigenvector mapping approach, it also uses "eigenfunctions of spatial configuration matrices to derive the spatial eigenvectors" (Griffith and Peres-Neto, 2006). This approach has been used to model soil attributes (Kim et al., 2016), plant diversity (Kim and Shin, 2016), crime patterns (Chun, 2014), and diseases (Jacob et al., 2008). Mainali and Chang (2018) used this approach to model the water quality trends of the Han River Basin, South Korea, reporting that it significantly increased model performance and removed the residual spatial autocorrelation.

5 Asymmetrical Eigenvector Maps

All of the spatial statistical approaches discussed in the previous section assume that the relations among sampling sites are multidirectional. The spatial associations of different points along the river are usually unidirectional as the water flows downstream (Figure 1-3c). Therefore, upstream water quality affects downstream water quality but not vice-versa. Recently, new spatial statistical methods have been developed in order to account for such directionality in water quality modeling. Blanchet et al. (2008) modified MEM's approach in order to incorporate the directional process of rivers and streams as Asymmetrical Eigenvector Maps (AEMs). They propose that "gradients influencing spatial distribution can be studied via spatial variables (eigenfunctions) that represent directional spatial processes." This is also a part of the eigenfunctions-based spatial filtering framework, with the added feature that it "constructs space in an asymmetric way" by only accounting for the sites connected through the water flow. The modeling involves defining a connection diagram based on the directional spatial process and creation of sites-by-edges matrix which are transformed into spatial eigenvectors.

6 Spatial Stream Network

A river can be effectively represented as a dendritic network, and any scientific inquiries and management decisions related to river networks should acknowledge this (Peterson et al. 2013). Dendritic networks use points and lines in geographical space, and typically have a directional component (Peterson et al., 2013). The modification of the autocovariance model that incorporates the dendritic network structure of rivers is dubbed a spatial stream network (SSN) model (Ver Hoef et al., 2006, 2014). It uses a set of autoregressive functions to derive the predictor variables to be used in the regression modeling. The weight of those directional processes can be river distance, flow volume, or catchment size, or any relevant variables for the watershed of interest (Figure 1-3e). The SSN allows users to test spatial autocorrelation and develop model at various scenarios like flow-connected, flow-unconnected, and Euclidean distance (Isaak et al., 2017; Neill et al., 2018; Scown et al., 2017). It not only allows the development of models but also lets users explore the spatial properties of the data in relation to various in-stream processes (e.g., McGuire et al., 2014).

Table 1-1: Papers reviewed in different models

Spatial Models	**No of papers included**	**References**
Spatial error and lag	14	Chang, 2008; Engström et al., 2017; Fox and Alexander, 2015; Huang et al., 2016; Miralha and Kim, 2018; Sanchez et al., 2014; Snelder et al., 2018; Su et al., 2013; Vitro et al., 2017; Walters et al., 2018; Wan et al., 2015; Xu et al., 2016; Yang et al., 2017; Yang and Jin, 2010
Geographically Weighted Regression	18	Bhowmik et al., 2015; Chang and Psaris, 2013; Chen et al., 2016; Chu et al., 2018; Eccles et al., 2017; Kim et al., 2018; Pratt and Chang, 2012; Salles et al., 2018; Shrestha and Luo, 2017; Sun et al., 2014; Taghipour Javi et al., 2014; Tu, 2013; Tu and Xia, 2008; Wang and Zhang, 2018; Wilson, 2015; Xia et al., 2018; Yu et al., 2016; Zhao et al., 2015
Spatial Eigenvector Based Models	10	Brogna et al., 2017; Catherine et al., 2016; de Oliveira Marcionilio et al., 2016; Mainali and Chang, 2018a; Piorkowski et al., 2014; Pond et al., 2017; Souza-Bastos et al., 2017; Strangway et al., 2017; Vrebos et al., 2017; Zorzal-Almeida et al., 2018
SSN	12	Detenbeck et al., 2016; Falke et al., 2016; Frieden et al., 2014; Holcomb et al., 2018; Isaak et al., 2018; Marsha et al., 2018; Neill et al., 2018; Post et al., 2018; Scown et al., 2017; Steel et al., 2016; Turschwell et al., 2016
Total		**54**

IV A Systematic Review of Current Studies

We carried out a systematic review of articles related to different types of spatial regression of water quality published from 2000 to 2018, using the Web of Science database on November 9, 2018 (Table 1). The search phrases we used included “water quality” and “spatial regression”, “water quality” and “eigenvector”, “water quality and “autocorrelation”, and “water quality” and “spatial stream network.” We identified 54 articles with a water quality focus that used at least one type of spatial regression (Table 1). Notice that it may not be a comprehensive list, as we only searched for the term "water quality". The water quality information might well be published as water pollution, or in terms of individual parameter names such as temperature, pH, nitrogen, or phosphorus. These names were not included in our search term. We also removed studies that did not have spatial regression approaches. Although we mostly focused on surface water, we also included a few groundwater-quality works in this review. We focused our review on the use of spatial statistical methods to account for spatial autocorrelation and residual spatial autocorrelation, weight matrix construction, scale considerations, and improvements in model performance in different types of spatial statistical modeling. We also attempted to identify the spatial pattern of these studies to explore where such research efforts have been concentrated.

1 Geographic Distribution of Studies

The majority of study sites of research related to spatial statistical modeling of water quality are concentrated in USA and China with a few exceptions: Canada, Brazil, South

Korea, Australia, and some countries of Europe (Figure 1-4). This is likely because of the fact that these countries have relatively dense networks of monitoring stations over a large area. Only 15 nations were represented from 54 studies. Although developing countries are most vulnerable to water quality degradation (Schwarzenbach et al., 2010), very little research has been carried out there. This list may not be comprehensive, but we assume that this map represents the spatial pattern of current research related to spatial aspects of water quality.

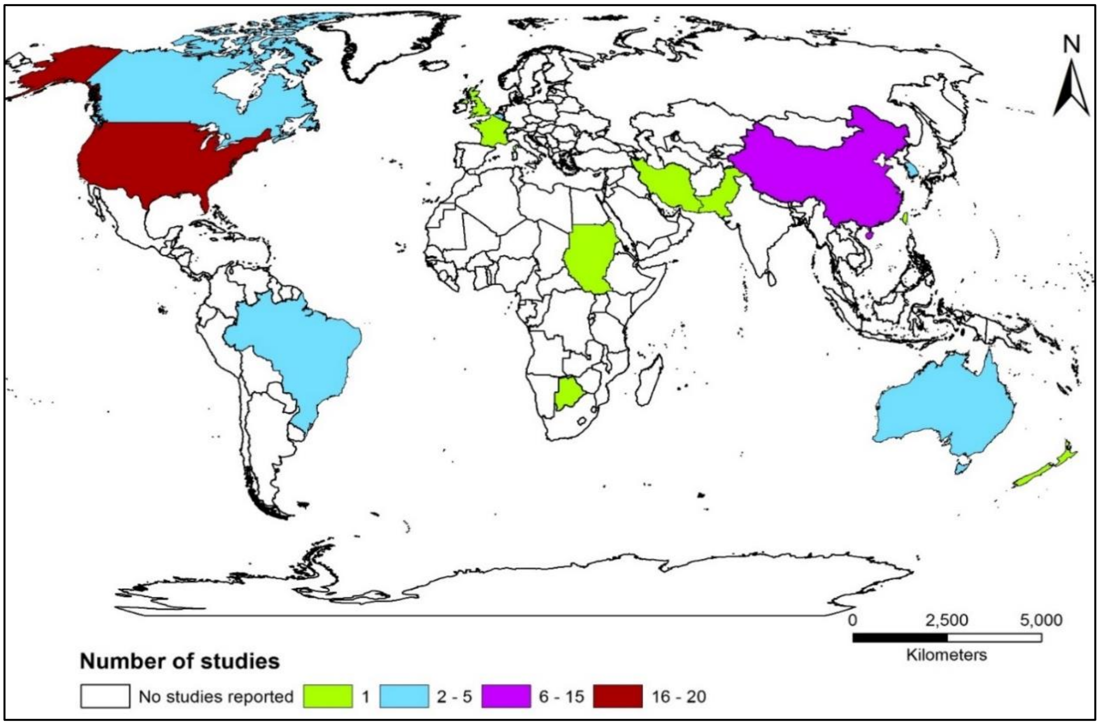


Figure 1-4: Country-wise distribution of the sites of the studies included in this review (n= 54).

2 Spatial Autocorrelation in Different Spatial Regression

Theoretically, exploring the spatial autocorrelations of the dependent variables and residual autocorrelations, and examining the significance of spatial autocorrelations, are the first steps in incorporating spatial relations into the models. Although the relationship between residual spatial autocorrelation and variation of the model (pseudo-) R^2 and coefficients is discussed in most of the studies, the relationship with the spatial autocorrelation of dependent variables is usually not taken into consideration. Many new studies have reported that spatial autocorrelation of dependent variable and residual spatial autocorrelation are usually related; the choice of covariates also affects the significance of residual spatial autocorrelation (Miralha and Kim 2018, Mainali and Chang, 2018).

We find that the use of spatial autocorrelation statistics of the dependent variable is generally associated with the type of spatial regressions used. Approximately 43 % of papers that used either a spatial lag model or a spatial error model calculated the spatial autocorrelation of the dependent variable, while only 30 % of papers using eigenvector-based model did so. Similarly, 43 % of papers using geographically weighted regression calculated spatial autocorrelation of the dependent variable, while about 75 % of SSNM papers did so. Forty eight percent of spatial-error/lag, 70% of Eigenvector-based, 61% of GWR, and 100% of SSN-model papers tested for residual spatial autocorrelation.

The analysis of spatial autocorrelation in water quality leads to a better understanding of the extent of spatial organization (clustered, dispersed or random) of water quality variables, and also helps explore the capacity of the independent variables to predict the water quality pattern (e.g., Miralha and Kim 2018). Accounting for spatial autocorrelation in regression can correct bias in parameter estimation and, hence, helps avoid an incorrect conclusion for potential factors. A higher percentage of residual spatial autocorrelation testing in more recent studies stems from the fact that the independent variables might not explain all the spatial autocorrelation, and results in residual spatial autocorrelation. That is, spatial autocorrelation in residuals is the one that should be examined. A high spatial autocorrelation in the response variable may give a hint for spatial autocorrelation in residuals, but is not necessarily a reason to use spatial regression as long as there is no significant residual spatial autocorrelation. A future suggestion in this field would be checking for residual spatial autocorrelation before performing spatial regression models, if the researchers are concerned that the regression model does not account for the spatial autocorrelation.

3 Spatial Weights Matrix

All spatial statistical modeling approaches are based on some form of spatial weights matrix. The most common type of weights matrix, distance matrix, is constructed using the distance among the sampling sites based on geographical coordinates; sites are weighted based on distance, number of neighbors, or other relevant attributes. The other attributes include Euclidean distance upstream, river distance upstream, catchment size, and river flow (Isaak et al., 2018). There are several standard distance matrices available for different types of spatial regression approaches. For example, spatial lag and spatial error methods use nearest-neighboring stations (Chang 2008, Huang et al. 2014); the spatial filtering approach uses at least one neighbor; the geographically weighted approach mostly uses adaptive bandwidth to include the desired number of sites; SSN uses river distance, flow volume, or upstream catchment area. However, spatial statisticians recommend modifying the weights matrix based on the hypothesis being tested, the scale of analysis, the spatial distribution of the sampling station, and spatial issues being addressed (Blanchet et al., 2008b; Sokal and Oden, 1978).

Based on our review, we find that most of the papers use a 'standard' weight matrix provided by the software on which model is being implemented (Table 1-2). Traditionally, spatial-lag models use observations in all directions to create a spatial lag variable. Some studies attempted to modify the existing weights matrix to incorporate hydrologic connectivity. For example, Vitro et al. (2017) modified a spatial weights matrix to incorporate the effect of only upstream stations in a spatial lag model. They provided relative weights to upstream stations based on the proximity to the candidate station being considered. Engström et al. (2017) used two different weights matrices, one with all proximate stations and the other with proximate and upstream stations. Most other studies used only a set number of nearby stations to define weights. For example, Chang (2008) and Huang et al. (2014) used four closest stations, Su et al. (2013) used ten such stations, and Yang and Jin (2010) used only adjacent stations. However, no study has tested how the study results might be sensitive to changes in weight matrices.

Geographically Weighted Regression (GWR) uses an exponential (or Gaussian) distance decay function to create spatial weights among the sampling sites included within the specified distance defined by the bandwidth. A majority of the GWR papers use flexible (or adaptive) bandwidth to derive the spatial weights to be used in the regression models. An adaptive bandwidth allows the band (or buffer) around a sampling station to vary according to the number of nearby sampling stations. The bandwidth is small for clustered data and large for scattered data, based on the distance between sampling stations. Most of these papers use a software-defined standard bandwidth approach (mostly adaptive bandwidth) available in ArcGIS. We did not find any studies that use GWR by including the effect of only upstream stations. However, Tu (2013) used sampling stations only from mutually exclusive watersheds, thereby avoiding any complexity that would be caused by upstream stations in the model. While this approach avoids the issue of upstream influence on downstream water quality, the sample size will be lowered as many spatially dependent stations are discarded for analysis. Additionally, most studies did not address the potential issues of a small sample size when GWR models were used for water quality studies. This can be a new research direction where researchers define band only towards the upstream stations and weight those values to derive the local models, which hypothetically, would better explain the local patterns. Our hypothesis is based on the general understanding of the river flow where most of the physical and chemical components flow downstream.

The research papers using MEMs and AEMs approaches also use a standard weights matrix based on the Borcard and Legendre (2002). As scale can be an issue in these kinds of weights matrices, some researchers construct eigenvectors at different scales. For instance, de Oliveira Marcionilio et al. (2016) calculated their weights matrix using eight different distance classes (50 meters to 450 meters, with an interval of 50 meters) to incorporate the effect of scale on their analysis. The SSN modeling approach was initially proposed to incorporate weights based on the stream distance, flow volume, or stream order. When flow volumes are not available, the catchment area is commonly used as a weight attribute (Ver Hoef and Peterson, 2010). But other attributes such as slope, shrieve's stream order, and Euclidean distance among stations are also used depending upon the nature of the watershed and the availability of data.

We notice from this review that a spatial weights matrix typically does not gain enough attention, in spite of its being the backbone of spatial modeling. Most previous studies rely on a weights matrix readily available in the ‘standard’ tools offered in software packages, rather than putting additional effort into generating a revised weight matrix that considers water flow along the hydrologic network. Therefore, researchers ought to be mindful of the spatial relations of water quality in the sampling space and design the weights matrix to best capture such spatial relations. We also need to be aware of the spatial relations of water quality sampling sites to source, mobilization process, delivery mechanism, and in-stream movement, and use appropriate weighting schemes to capture those processes.

4 Use of Multiscale Processes

The predictor variables for regression analysis are generally derived using a watershed because all the water flowing in the river comes from some part of the watershed, and watershed characteristics are reflected in river water quality (Allan, 2004). Researchers have worked to identify the scale at which water quality is best correlated with watershed characteristics (Figure 5). Although a majority of researchers used spatial lag/error, GWR, or MEM to extract predictor variables at different scales, they did not compare the effect of different scales in model prediction (Table 2). They rarely used different scaled data under the same regression model. The papers using SSN models, however, recognized the effects of variables at different scales and incorporated those in the models.

Some researchers have used different buffer distances from the river and/or sampling station. For example, vegetation cover within a 10m buffer is used for temperature modeling by Isaak et al. (2018), while other variables were used at the watershed scale. Turschwell et al. (2016) used 10m buffer for riparian vegetation and additionally used inverse-distance weighted effects of grazing land cover, while other variables were used as the lump attributes at the watershed scale, and reported significantly higher R^2 values when SSN models were used.

Like any other natural processes, the factors affecting water quality operate at different scales. These factors must be identified based on the understanding of the scale related to the source, mobilization, delivery, and instream processes related to these parameters (Lintern et al. 2018). This also depends on the scale at which disturbances drive water quality (Pond et al. 2017). If an "upland disturbance" is a driving factor of deteriorating water quality, using data derived only at the riparian buffer scale does not work (Pond et al. 2017). Our review also shows that the scale effects in water quality modeling using landscape characteristics are not universal, as they vary by parameters studied, location, seasons, and covariates used (Liu et al. 2017; Mainali and Chang 2018).

Isaak et al. (2018) argue that the covariates used in modeling approaches should come from a review of the literature and an understanding of a plausible mechanism that could cause a variation in a particular water-quality parameter. If the scale is not clear for the parameter, it is always safe to start with the watershed scale and incorporate other scales (e.g., Mainali and Chang 2018). In large-scale analysis, the availability of particular datasets also determines the scale at which covariates are extracted. Our review shows that the researchers should be able to provide explanations for the reasons behind choosing a particular covariate, its scale, and the need for any weights treatment in the spatial statistical modeling of water quality.

Water flowing from various parts of a watershed drains into surface water bodies via multiple pathways. Water quality along the stream network, therefore, depends on the sources of the parameter, their delivery, and instream processes occurring in the vicinity of an area where water flows (Lintern et al., 2018). To best capture such spatial variations researchers need to collect data or install the monitoring network carefully. The spatial and temporal scale of data collection and monitoring should be informed by the available geographical information of the watershed related to land use, human impact, geology, and hydrological characteristics of the stream. While increasing the spatial and temporal scale of analyses could help improve our understanding of the relationship between water quality and landscape variables, such effort requires time and resources (both human and computation resources). To make optimum use of time and resources, a selection of the data collection sites and appropriate scale should incorporate all the relevant characteristics of the range of watershed conditions (Jackson et al., 2015).

While it is beyond the scope of this paper to list all different scales at which predictor variables are extracted, here we list different statistical methods to effectively include different scale processes in water quality modeling identified in the papers we reviewed. Multi-scale data sets can be treated with principal-component analysis to reduce the dimension of the data and include the variability of different scale processes (Miralha and Kim, 2018). Redundancy analysis can identify which variables at what scale can explain variation in water quality, and use them as a predictor in the spatial regression (Strangway et al., 2017). To avoid overfitting of the data that identify the best subset of the covariates, a "Best Subset Regression" can be used (Scown et al. 2017). The Best Subset Regression uses Akaike Information Criteria (AIC) variation to identify a maximum number of covariates set by the analyst. Review of potential factors affecting water quality is of utmost importance before undertaking any water quality modeling efforts. From our review, we notice that there might be dozens of such candidate covariates. An appropriate variable reduction or selection method should be used in order to include a manageable number of water quality parameters representing different scales.

Table 1-2: Consideration of weights matrix, spatial autocorrelation, and residual spatial autocorrelation

Model type	**Scale**	**Spatial autocorrelation (SAC)**	**Weights matrix**	**Residual Spatial Autocorrelation (RSAC)**
Spatial Lag/Spatial Error	Predictor variables extracted at multiple scales. Entire catchments (Yang and Jin, 2010), a buffer of a certain distance (Chang, 2008), circular upstream buffer, multiscale (Su et al., 2013, Chang, 2008)	About 60 percent of the papers evaluate SAC of response variable before pursuing these models.	Most of the papers use weights matrix based on the Euclidean distance between neighboring stations while some modify it to test a different hypothesis (e.g., Engström et al., 2017; Vitro et al., 2017).	Most of the papers do not evaluate whether RSAC has been an issue or not. Only a couple papers used it (Miralha and Kim, 2018, Engström et al., 2017)
Eigenvector-based (MEM/AEM/ Spatial filters)	Some papers only used watershed while the majority used different scales (Strangway et al., 2017Mainali and Chang 2018). Scale information derived from eigenvectors are also used (Vrebos et al., 2017)	Only about a quarter of papers appeared in our list explored global or local SAC.	Mostly used standard weights matrix derived using a binary coded sites-by-edges table and distance between the sites. Some modify it based on the distance classes (de Oliveira Marcionilio et al., 2016).	Majority of the papers report RSAC except Strangway et al. (2017) of the model. RSACs are removed when this modeling approach is used.
Geographically Weighted Regression (GWR)	Although the majority of papers only use watershed or some distance from the sampling station, some of the papers used different scales (Pratt and Chang, 2012).	The autocorrelation of the response variable is tested scantly.	Mostly adaptive or fixed bandwidth approach is used as available in the software. Shrestha and Luo (2017) tried to make sure that there are certain numbers of stations (119) nearest neighbors in each local models.	As there is an inbuilt function to test RSAC in ArcGIS interface of GWR, most of the papers mention it in their model.
Spatial Stream Network (SSN) Model	Most of the papers using SSN use a multi-scale approach where relevant covariates are extracted from either whole watershed, or buffer, or using distance weighted approaches	Semivariogram and Torgegrams are used to explore SAC almost exclusively although some papers do not	Different attributes are used as weights like river distance, discharge, and catchment size with different spatial connectivity considerations like flow connected, not	RSAC of the models are tested almost exclusively and SSN models have found to remove it.

5 Comparison of Model Performance

As expected, the spatial regression models typically explain the variation of the dependent variable better than their aspatial counterparts (Table 1-3). Studies using spatial-lag and error models generally reported improved model performances from an aspatial linear regression model. An increase in R^2 and a decrease in AIC indicate the improved model performance of these models over an aspatial one (Chang, 2008; Engström et al. 2017; Huang et al. 2014; Yang and Jin, 2010). While using eigenvector-based spatial filtering approach, Mainali and Chang (2018) reported that the model strengths (R^2) significantly increase when an aspatial model suffered from residual spatial autocorrelation. However, most of eigenvector-based spatial statistical models we reviewed did not make an explicit comparison between aspatial and spatial models, as they used landscape characteristics and eigenvectors in the same model and used redundancy analysis to parse out the effect of ‘environmental’ and ‘spatial’ predictors (Souza-Bastos et al., 2017; Vrebos et al., 2017). Geographically weighted regression (GWR)-based models consistently showed higher model strengths than linear regression. Chu et al. (2018) reported that GWR performed better than linear regression, which was superseded by geographically and temporally weighted regression. Similarly, Tu (2013) reported that the model performance increased by up to 10-fold when GWR was used against linear regression models. Tu and Xia (2008) also found some “dramatic” increases in R^2 when GWR models were used. Most other papers using GWR for water quality modeling also reported a significant increase in model performance (Kim et al., 2018; Pratt and Chang, 2012; Shrestha and Luo, 2017; Sun et al., 2014; Yu et al., 2013). The spatial stream network (SSN) based models have shown to produce high R^2 values in modeling water quality parameters. An R^2 value of higher than 0.9 was reported for modeling summer temperature using SSN (Isaak et al., 2018). Turschwell et al. (2016) found SSN performing strongest among different models used. However, in some cases, SSN-based models did not significantly improve model performance (e.g., Frieden et al., 2014). These varying results appear to be associated with the choice of water quality parameters, landscape variables, the scale of analysis, sample size, and watershed conditions.

Table 1-3: Improvement of model performance using spatial statistical models.

Author	WQ parameters	Predictor(s)	Range of R^2 change
		Spatial Lag and Error Model	
(Yang et al., 2017)	TN	Land use types and hydrological soil groups	Increase in R2 values ranged from 0.06 to 0.12
Miralha and Kim, (2018)	pH, T, SC, DO, TDS, TN, DIN, KjN, TP, Tur, Br, Cl, Mg, Na, Ca, SiO2, Fe, K, CO_2, Mn, Alk, SO_4^{--} F, T, Csu, Chla, TOC, DOC, As, Cd, Zn, PO_4^{---}, NO_3, Al,	Land cover, elevation, slope, hydrological soil groups	Increase in R^2 values ranged from 0.03 to 0.29
(Vitro et al., 2017)	Fecal coliform	Demographic, sewer, sine, landcover, policy dummies	Model performance increased from 0.44 to 0.46662 to 0.4665
(Engström et al., 2017)	Microbiological contamination	Distance to informal settlement, share of informal settlement, different land use, distance to marshland, etc.	Reduction in model AIC from 158.31 to 153.2
(Sanchez et al., 2014)	Different components of biological integrity	Race, income, education, housing, and population size, household size. etc	DIC decreased in spatial model against the spatial model (2131 vs 2064, 1848.7 vs 1673.8, 2428 vs 2270, 1252 vs 1143) .
(Huang et al., 2014)	NH_4, NO^3 , COD, SRP, Cl, Na, K, and Mg^{++}	Landscape composition, pattern, topography, geology, population, GDP	Increase in R^2 ranged from 0.003 to 0.2

(Su et al., 2013)	DO, NH_3, and TP	Population, GDP, soil, land use,	R^2 values not compared only spatial regressions run
(Yang and Jin, 2010)	NO_3, NO_2-N	Landuse/cover, soil, slope, and area of watershed	Increase in R^2 values ranged from 0.04 to 0.1.
(Chang, 2008)	T, TN, TP, pH, COD, BOD, SS, DO	Land use, topography, soil	R^2 values generally increased up to 0.3
(Fox and Alexander, 2015)	E. Coli, TSS, DO, Cond, Temp	Land use, Floodplain, wildlife, elephant-specific fecal count, wildlife species	Quantitative change in R^2 is not reported. But spatial models performed better
Walters et al. 2018	TP	Land use composition and pattern, area, precipitation	Result of spatial regressions only reported.
Snelder et al. 2017	TN, NO_3, TP, and DRP	Climate, topography, geology and land cover	No comparisons were made
Xu et al. 2016	Nitrogen Loss	Morphometric variables and soil drainage of each land cover type	No comparison only spatial lag model
(Souza-Bastos et al., 2017)	Hematocrit, Plasma Osmolality, sodium, chloride, Mg, K, Cortisol, Glucose, etc.	Different water quality parameters	Spatial factors accounted for about 2% variation of dependent variables.
(Wan et al., 2015)	Macroinvertebrates	Different water quality parameters	Spatial factors (eigenvectors) more important than the environmental factors. Overland distance worked

			better (6.7 to 9.5, and 10.2 to 10.7 percent).
(Brogna et al., 2017)	DO, DOC, TP, NH_4, NO_2, NO_3, pH, Cl, SO_4	Forest cover	Variability explained by forest covers when elevation is included accounts for 9.3 percent of variation in water quality which would be 33.8 if elevation was not included
(Vrebos et al., 2017)	T, pH, O, NO3, NO2, NH4, TP, CL, Co2, BSi, Ca, Fe, K, Mg, Na, SiO2, Zn, COD, SS, Chl-a, Cond	Land use and soil	Space (Euclidean distance based MEM) explained for both analyses circa 22% of variance. But non of the AEMs were significant predictors
(Strangway et al., 2017)	TP, OP, E.Coli, KjN, DOC, pH, Cond, various metals, NO_3, DO, dissolved Br, Ca, Mg, and SO4, F, Hg, Sb, As, B, Se, Si, Tellurium etc.	Land use, road density	River network based model explained the greater variations.
(Catherine et al., 2016)	Phytoplankton species	Water quality parameters, land use, rainfall, water temperature, altitude, etc.	No significant effect of MEMs were reported in the model performance
Mainali and Chang 2018	TN. TP, COD, SS	land use topography, soil, population	Increase in R^2 ranged from -0.16 to 0.31

(de Oliveira Marcionilio et al., 2016)	Chl-a	Water quality parameters, depth, vegetation cover	Addition of spatial factors at eigenvector slightly increased the model performance (39 vs 28 %)
Zorzal-Almeida et al. 2018	Trans., CO_2, DO, Cond., pH, NH4, NO3, TN, PO_4, TP, Chla, TOC, TN, TP, C/N, δ13C, and δ15N	Land use index	AEM R^2s are higher from 0.13 to 0.24 over MEM. Only environmental
Piorkowski et al. 2013	E. coli	Organic carbon and water velocity	MEMs explain 26.9 % of the population variance during baseflow and 31.7% post stream flow.
(Xia et al., 2018)	Cu, Zn, Pb, Cr and Cd	Land use	GWR didn't always increase R^2 values. R^2 change ranged from -0.029 to 0.663
(Kim et al., 2018)	Cyanobacteria	band 2, 4, and 5 of RapidEye imagery	R2 was increased to 0.719 from 0.615, and AICc was also reduced from 1735 to 1710
(Salles et al., 2018)	Amplitude of the water table variation	Soil water, soil types, drainage network, slope etc.	0.22 in OLS vs 0.9 in GWR
(Wang and Zhang, 2018)	Water Quality Index (12 different parameters)	Landscape pattern matrix	Global R^2 of GWR models were not reported but increase in R2

			in GWR models can be inferred from the results.
(Chu et al., 2018)	TB, which refers to the haziness of fluid caused by suspended solids in flowing water	Red, green, and blue reflectances	R^2 values of LR, GWR and GTWR are 0.37, 0.44, and 0.87 respectively.
(Shrestha and Luo, 2017)	Groundwater Nitrate	Fertilizer, manure, crop, permeability, precipitation, slope, DO, Clay, Iron, and Mg	GWR regression increased by 0.05.
(Eccles et al., 2017)	Total Coliform, E. coli	Aquifer depth, hydraulic connectivity, flood hazard types, land cover data, abandoned well, population and dwelling density, number of farms, and hectares of farmland	R^2 increased from 0.013 to 0.11, 0.099 to 0.155
(Chen et al., 2016)	TN, TP, DO, COD	Different Land use types, census	Corresponding GWR models had adjusted R^2 values an average of 59.2% higher than the optimal OLS models
(Chang and Psaris, 2013)	Temperature related matrix	Base flow, precipitation, stream oreder, distance to coast, topography, and land cover	R^2 values increased from 0 to 0.08
(Zhao et al., 2015)	COD, BOD, NH_3, TP, Hg	Land use change intensity	R^2 change not compared as no OLS were run
(Sun et al., 2014)	Temp, pH, DO, chla, Sal, Cond, TOC, TN, TP	Land use composition, and matrix, topography	Global value of GWR R^2 was not reported.

Yu et al. 2013	T, pH, DO, PP, BOD, NH_3, TP, TN, Faecal choliform, anionic surfactant dissolved oxygen	Land use composition and matrix (mostly matrix)	About 59% of GWR models have significantly higher explanatory power for water quality than the corresponding OLS models
Tu 2013	SC, DO, OC, TN, KjN, NO3, NO2	Land use data in Year 2005	R^2 values sometimes increased by 10 folds
(Pratt and Chang, 2012)	Cond, DO, NO_3, pH, TP, TS, T	land cover, topography, built structure	R2 values increased from 0.04 to 0.44
(Tu and Xia, 2008)	SC, NH3-N, NO2-N, KN, NO3-N, P, Ca, Mg, Na, K, Cl, SO4, DS	Land use and population	A dramatic improvement in R^2 of GWR over OLS is observed for every pair of models
(Taghipour Javi et al., 2014)	Groundwater level changes and groundwater withdrawal differences (GWD)	Land use/cover	Increase in R2 ranged from 0.11 to 0.48
Bhowmik et al. 2015	As, Cd, Cr, Cu, Fe, Mn, Hg, Ni, Pb, Zn	Land use, soil, elevation	Not compared
Wilson 2015	TSS, TP	Different water quality parameters, land use, negativity, rainfall, water temperature, altitude, etc.	Only temporal changes of GWR models are presented not compared with aspatial model
(Neill et al., 2018)	E. coli	Land use, soil, Anthropogenic Impact Index	R^2 values increased from 0 to 0.2. R^2 value neared

			one when random effects were included.
(Marsha et al., 2018)	Temperature	Elevation	Quantitative comparisons not made. But linear model and SSN had mixed effects in different kind of matrices.
(Isaak et al., 2018)	Temperature	Elevation, slope, lake percentage, glacier, ppt, northing, base flow index, drainage area, riparian canopy, air temperature, discharge, tailwater	No comparisons made but overall model performance of SSN was more than 90 %
(Scown et al., 2017)	TP	Area, stream category, slope, soil area, clermont area, land use, septic systems, NPDES permit address, total P released, average tp concentration	AIC value slightly reduced (134.98 to 133.76).
Steel et al 2016	Temperature	elevation, mean annual discharge, and percent commercial area	Explicit comparisons not made
Frieden et al. 2014	Macroinvertebrates	Air temperature, catchment area, soil, direction, land use	Spatial models did not substantially increase model performance over the non-spatial models
Turschwell et al. 2016	Different temperature matrices	Elevation, air temperature, riparian vegetation within 100 m buffer, IDW-HA of grazed land, solar radiation	SSNM, RF, and Nonspatial R^2s are 0.825, 0.81, and 0.824 respectively

(Shi et al., 2016)	DO, NH_3, COD, TP	Land cover and topography (slope, and elevation)	Only aspatial multiple regressions were run
Detenbeck et al. 2018	Temperature	Land cover, air temperature, slope, drainage, imperviousness etc.	Yes compared against non-spatial model
Falke et al. 2015	Temperature	No predictors	No comparisons made
Holcomb et al. 2018	Microbial Water Quality	Landuse, rainfall	The OLS model and the three spatial models performed similarly, with the OLS model faring slightly worse by all three metrics and the Euclidean space-only model performed slightly better by AIC
Post et al. 2018	DO, Temperature, and Salinity.	Space-time predictors	Spatial and non-spatial model R^{2}s worked similarly.

(SC: specific conductance; DO: dissolved oxygen; TDS: total dissolved solids; TSS: total suspended 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; FC: fecal coliform; DOC: dissolved organic carbon; Pb: lead; Zn: zinc; Cd: cadmium; CO2: carbon dioxide; SiO2: silicon dioxide; PO4: phosphate; As: arsenic; PP: potassium permanganate; BOD: biochemical oxygen demand; dissolved reactive phosphorus; DRP Cr: chromium; Cu: copper; Fe: iron; Mn: manganese; Hg: mercury; Ni: nickel; cond: conductivity; C/N: carbon-to-nitrogen ratio; Sal: salinity;

SO4 sulphate; NO3: nitrate; E. coli: Escherichia coli; NO2: nitrite-nitrogen; GDP: gross domestic product; OLS: ordinary least square regression; AIC: Akaike information criteria; DIC: deviance information criteria; MEM: Moran’s eigenvector maps; AEM: asymmetrical eigenvector maps; GTWR: geographically and temporally weighted regression; NPDES: National Pollutant Discharge Elimination System; SSNM: Spatial Stream Network Model.)

V Conclusions

Spatial modeling of water quality is gaining increased attention, and researchers have been using novel and creative ways to incorporate spatial aspects into surface water quality modeling. Our review identifies a few aspects of these modeling that stood out.

- Research in this field is dominated by resource-rich countries like the US and China. This may be associated with the availability of data over a large geographical area.
- There is still insufficient emphasis on spatial autocorrelation and residual spatial autocorrelation, which deserve more attention as these techniques can help understand unidirectional, multidirectional, and river network-based spatial attributes of the dependent variable and overall models of surface water quality. A suggestion based on this review would be to check for residual spatial autocorrelation before performing spatial regression models if the researchers are concerned with the regression model not being able to account for the spatial autocorrelation.
- Weight matrices have great potential in informing spatial autocorrelation of dependent variables at different scales, and in helping test several hypotheses of spatial eco-socio-hydrological processes in relation to surface water. Thus, testing the model’s sensitivity to different weight matrices needs further investigation.

However, no study considered in our review has tested the sensitivity of a model against the changes in weight metrics.

- Our reviews show that the modification of a weights matrix should be informed by spatial organization of water quality data points, understanding of the source, mobilization, and delivery of a particular water quality parameter, the hypothesis being tested, and the scale of analysis.
- In most regression models except SSNs, predictor variables extracted from different scales are used differently to compare the model strength. A fusion of predictor variables extracted from different scales, such as in a multiscale model, might be better suited to predict water quality, as different processes occur at several different scales simultaneously.
- A thorough review of source, mobilization, delivery, and instream flow mechanism of the water quality parameters under consideration might be necessary in order to include suitable predictor variables, multiscale processes, and identify appropriate weight matrix in the model. This should be accompanied by proper variable reduction statistics, like brute-force reduction, in order to include manageable and meaningful predictors.
- Although most of the spatial models are recognizing and incorporating the directional aspect of water flow, we did not find any papers using GWR doing so. Researchers can attempt to modify GWR to incorporate directional process and river network structures.

- Researchers should also explore different spatial representations of the landscape matrix (e.g. composition, patterns, distance weighting, and hydrological weighting) in order to identify an appropriate approach to use them in spatial modeling of water quality.

VI. References

Allan JD (2004) Landscapes and Riverscapes: The Influence of Land Use on Stream Ecosystems. *Annual Review of Ecology, Evolution, and Systematics* 35(1): 257–284. DOI: 10.1146/annurev.ecolsys.35.120202.110122.

Anselin L (1988) *Spatial Econometrics: Methods and Models*. Kluwer Academic, Dordrecht.

Anselin L (1995) Local indicators of spatial association—LISA. *Geographical analysis* 27(2): 93–115.

Anselin L (2001) Spatial econometrics. *A companion to theoretical econometrics* 310330.

Bini LM, Diniz-Filho JAF, Rangel TFLVB, et al. (2009) Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. *Ecography* 32(2): 193–204. DOI: 10.1111/j.1600-0587.2009.05717.x.

Blanchet FG, Legendre P and Borcard D (2008) Modelling directional spatial processes in ecological data. *Ecological Modelling* 215(4): 325–336. DOI: 10.1016/j.ecolmodel.2008.04.001.

Borcard D and Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* 153(1): 51–68.

Brody SD, Highfield W and Peck BM (2005) Exploring the mosaic of perceptions for water quality across watersheds in San Antonio, Texas. *Landscape and Urban Planning* 73(2–3): 200–214. DOI: 10.1016/j.landurbplan.2004.11.010.

Brunsdon C, Fotheringham S and Charlton M (1998) Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(3): 431–443.

Chang H (2008) Spatial analysis of water quality trends in the Han River basin, South Korea. *Water Research* 42(13): 3285–3304. DOI: 10.1016/j.watres.2008.04.006.

Chen Q, Mei K, Dahlgren RA, et al. (2016) Impacts of land use and population density on seasonal surface water quality using a modified geographically weighted regression. *Science of The Total Environment* 572: 450–466. DOI: 10.1016/j.scitotenv.2016.08.052.

Chun Y (2014) Analyzing Space–Time Crime Incidents Using Eigenvector Spatial Filtering: An Application to Vehicle Burglary. *Geographical Analysis* 46(2): 165–184. DOI: 10.1111/gean.12034.

Cliff A and Ord K (1972) Testing for spatial autocorrelation among regression residuals. *Geographical analysis* 4(3): 267–284.

de Oliveira Marcionilio SML, Machado KB, Carneiro FM, et al. (2016) Environmental factors affecting chlorophyll-a concentration in tropical floodplain lakes, Central

Brazil. *Environmental Monitoring and Assessment* 188(11). DOI: 10.1007/s10661-016-5622-7.

Forman RTT and Godron M (1981) Patches and Structural Components for a Landscape Ecology. *BioScience* 31(10): 733–740. DOI: 10.2307/1308780.

Fotheringham AS, Brunsdon C and Charlton M (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, England ; Hoboken, NJ, USA: Wiley.

Getis A and Aldstadt J (2004) Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Analysis* 36(2): 90–104. DOI: 10.1111/j.1538-4632.2004.tb01127.x.

Getis A and Griffith DA (2002) Comparative Spatial Filtering in Regression Analysis. *Geographical Analysis* 34(2): 130–140. DOI: 10.1111/j.1538-4632.2002.tb01080.x.

Getis A and Ord JK (1992) The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 24(3): 189–206. DOI: 10.1111/j.1538-4632.1992.tb00261.x.

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

Grabowski ZJ, Watson E and Chang H (2016) Using spatially explicit indicators to investigate watershed characteristics and stream temperature relationships.

Science of The Total Environment 551–552: 376–386. DOI: 10.1016/j.scitotenv.2016.02.042.

Griffith DA (2010) Spatial Filtering. In: *Handbook of Applied Spatial Analysis*. Springer, Berlin, Heidelberg, pp. 301–318. DOI: 10.1007/978-3-642-03647-7_16.

Griffith DA and Peres-Neto PR (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* 87(10): 2603–2613.

Guo D, Lintern A, Webb JA, et al. (2019) Key Factors Affecting Temporal Variability in Stream Water Quality. *Water Resources Research*. DOI: 10.1029/2018WR023370.

Huang J, Huang Y and 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. Chapman M (Gee) G (ed.) *PLoS ONE* 9(3): e91528. DOI: 10.1371/journal.pone.0091528.

Huang Z, Han L, Zeng L, et al. (2016) Effects of land use patterns on stream water quality: a case study of a small-scale watershed in the Three Gorges Reservoir Area, China. *Environmental Science and Pollution Research* 23(4): 3943–3955. DOI: 10.1007/s11356-015-5874-8.

Isaak DJ, Peterson EE, Ver Hoef JM, et al. (2014) Applications of spatial statistical network models to stream data: Spatial statistical network models for stream data. *Wiley Interdisciplinary Reviews: Water* 1(3): 277–294. DOI: 10.1002/wat2.1023.

Isaak DJ, Ver Hoef JM, Peterson EE, et al. (2017) Scalable population estimates using spatial-stream-network (SSN) models, fish density surveys, and national geospatial database frameworks for streams. *Canadian Journal of Fisheries and Aquatic Sciences* 74(2): 147–156. DOI: 10.1139/cjfas-2016-0247.

Isaak DJ, Wenger SJ, Peterson EE, et al. (2018) The NorWeST Summer Stream Temperature Model and Scenarios for the Western U.S.: A Crowd-Sourced Database and New Geospatial Tools Foster a User-Community and Predict Broad Climate Warming of Rivers and Streams. *Water Resources Research*: n/a-n/a. DOI: 10.1002/2017WR020969.

Jackson FL, Malcolm IA and Hannah DM (2015) A novel approach for designing large-scale river temperature monitoring networks. *Hydrology Research*: nh2015106. DOI: 10.2166/nh.2015.106.

Jacob BG, Muturi EJ, Caamano EX, et al. (2008) Hydrological modeling of geophysical parameters of arboviral and protozoan disease vectors in Internally Displaced People camps in Gulu, Uganda. *International Journal of Health Geographics* 7(1): 11. DOI: 10.1186/1476-072X-7-11.

Kearns FR, Kelly NM, Carter JL, et al. (2005) A method for the use of landscape metrics in freshwater research and management. *Landscape Ecology* 20(1): 113–125. DOI: 10.1007/s10980-004-2261-0.

Kim D and Shin YH (2016) Spatial autocorrelation potentially indicates the degree of changes in the predictive power of environmental factors for plant diversity. *Ecological Indicators* 60: 1130–1141. DOI: 10.1016/j.ecolind.2015.09.021.

Kim D, Hirmas DR, McEwan RW, et al. (2016) Predicting the Influence of Multi-Scale Spatial Autocorrelation on Soil–Landform Modeling. *Soil Science Society of America Journal* 80(2): 409. DOI: 10.2136/sssaj2015.10.0370.

King RS, Baker ME, Whigham DF, et al. (2005) Spatial Considerations for Linking Watershed Land Cover to Ecological Indicators in Streams. *Ecological Applications* 15(1): 137–153. DOI: 10.1890/04-0481.

Lausch A, Blaschke T, Haase D, et al. (2015) Understanding and quantifying landscape structure – A review on relevant process characteristics, data models and landscape metrics. *Ecological Modelling* 295: 31–41. DOI: 10.1016/j.ecolmodel.2014.08.018.

Legendre P (1993) Spatial Autocorrelation: Trouble or New Paradigm? *Ecology* 74(6): 1659–1673. DOI: 10.2307/1939924.

Lintern A, Webb J a., Ryu D, et al. (2018) Key factors influencing differences in stream water quality across space. *Wiley Interdisciplinary Reviews: Water* 5(1): n/a-n/a. DOI: 10.1002/wat2.1260.

Liu J, Zhang X, Xia J, et al. (2016) Characterizing and explaining spatio-temporal variation of water quality in a highly disturbed river by multi-statistical techniques. *SpringerPlus* 5(1): 1171.

Mainali J and Chang H (2018) Landscape and anthropogenic factors affecting spatial patterns of water quality trends in a large river basin, South Korea. *Journal of Hydrology* 564: 26–40. DOI: 10.1016/j.jhydrol.2018.06.074.

McGuire KJ, Torgersen CE, Likens GE, et al. (2014) Network analysis reveals multiscale controls on streamwater chemistry. *Proceedings of the National Academy of Sciences* 111(19): 7030–7035. DOI: 10.1073/pnas.1404820111.

McLean MI, Evers L, Bowman AW, et al. (2019) Statistical modelling of groundwater contamination monitoring data: A comparison of spatial and spatiotemporal methods. *Science of The Total Environment* 652: 1339–1346. DOI: 10.1016/j.scitotenv.2018.10.231.

Miralha L and 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. DOI: 10.3390/ijgi7020064.

Nature Statistics (2019) Statistical methods - Latest research and news | Nature. Available at: https://www.nature.com/subjects/statistical-methods (accessed 28 February 2019).

Neill AJ, Tetzlaff D, Strachan NJC, et al. (2018) Using spatial-stream-network models and long-term data to understand and predict dynamics of faecal contamination in a mixed land-use catchment. *Science of The Total Environment* 612: 840–852. DOI: 10.1016/j.scitotenv.2017.08.151.

Peterson EE and Hoef JMV (2010) A mixed-model moving-average approach to geostatistical modeling in stream networks. *Ecology* 91(3): 644–651.

Peterson EE, Sheldon F, Darnell R, et al. (2011) A comparison of spatially explicit landscape representation methods and their relationship to stream condition. *Freshwater Biology* 56(3): 590–610. DOI: 10.1111/j.1365-2427.2010.02507.x.

Peterson EE, Ver Hoef JM, Isaak DJ, et al. (2013) Modelling dendritic ecological networks in space: an integrated network perspective. Blasius B (ed.) *Ecology Letters* 16(5): 707–719. DOI: 10.1111/ele.12084.

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

Sharma S, Legendre P, De Cáceres M, et al. (2011) The role of environmental and spatial processes in structuring native and non-native fish communities across thousands of lakes. *Ecography* 34(5): 762–771. DOI: 10.1111/j.1600-0587.2010.06811.x.

Schwarzenbach RP, Egli T, Hofstetter TB, et al. (2010) Global Water Pollution and Human Health. *Annual Review of Environment and Resources* 35(1): 109–136. DOI: 10.1146/annurev-environ-100809-125342.

Scown MW, McManus MG, Carson JH, et al. (2017) Improving Predictive Models of In-Stream Phosphorus Concentration Based on Nationally-Available Spatial Data Coverages. *JAWRA Journal of the American Water Resources Association* 53(4): 944–960. DOI: 10.1111/1752-1688.12543.

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

Shi W, Xia J and Zhang X (2016) Influences of anthropogenic activities and topography on water quality in the highly regulated Huai River basin, China. *Environmental*

Science and Pollution Research 23(21): 21460–21474. DOI: 10.1007/s11356-016-7368-8.

Sokal RR and Oden NL (1978a) Spatial autocorrelation in biology: 1. Methodology. *Biological journal of the Linnean Society* 10(2): 199–228.

Sokal RR and Oden NL (1978b) Spatial autocorrelation in biology: 1. Methodology. *Biological journal of the Linnean Society* 10(2): 199–228.

Strangway C, Bowman MF and Kirkwood AE (2017) Assessing landscape and contaminant point-sources as spatial determinants of water quality in the Vermilion River System, Ontario, Canada. *Environmental Science and Pollution Research* 24(28): 22587–22601. DOI: 10.1007/s11356-017-9933-1.

Tobler WR (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46: 234. DOI: 10.2307/143141.

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

Tu J and Xia Z (2008) Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation. *Science of The Total Environment* 407(1): 358–378. DOI: 10.1016/j.scitotenv.2008.09.031.

Turschwell MP, Peterson EE, Balcombe SR, et al. (2016) To aggregate or not? Capturing the spatio-temporal complexity of the thermal regime. *Ecological Indicators* 67: 39–48. DOI: 10.1016/j.ecolind.2016.02.014.

Ullah KA, Jiang J and Wang P (2018) Land use impacts on surface water quality by statistical approaches. *Global Journal of Environment Science and Management* 4(2): 231–250. DOI: 10.22034/gjesm.2018.04.02.010.

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

Ver Hoef JM and Peterson EE (2010) A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association* 105(489): 6–18.

Ver Hoef JM, Peterson E and Theobald D (2006) Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics* 13(4): 449–464. DOI: 10.1007/s10651-006-0022-8.

Ver Hoef JM, Peterson EE, Hooten MB, et al. (2018) Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs* 88(1): 36–59. DOI: 10.1002/ecm.1283.

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

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

Xiao R, Wang G, Zhang Q, et al. (2016) Multi-scale analysis of relationship between landscape pattern and urban river water quality in different seasons. *Scientific Reports* 6: 25250. DOI: 10.1038/srep25250.

Yang X and Jin W (2010) GIS-based spatial regression and prediction of water quality in river networks: A case study in Iowa. *Journal of Environmental Management* 91(10): 1943–1951. DOI: 10.1016/j.jenvman.2010.04.011.

Zhou T, Wu J and Peng S (2012) Assessing the effects of landscape pattern on river water quality at multiple scales: A case study of the Dongjiang River watershed, China. *Ecological Indicators* 23: 166–175. DOI: 10.1016/j.ecolind.2012.03.013.

Chapter Two: Environmental and Spatial Factors Affecting Surface Water Quality in a Himalayan Watershed, Central Nepal

Janardan Mainali and Heejun Chang

Environmental and Spatial Factors Affecting Surface Water Quality in a Himalayan Watershed, Central Nepal

Janardan Mainali and Heejun Chang

Abstract

Various spatial interrelationships among sampling stations are not well explored in the spatial modeling of water quality literature. This research explores the relationship between water quality and various social, demographic, and topographic factors in an urbanizing watershed of Nepal with a comparison of different connectivity matrices to conceptualize spatial interrelationships. We collected electric conductivity and dissolved oxygen (DO) data from surface water bodies using a handheld probe, and used the data to establish relationships with land use, topography, and population density-based explanatory variables at both watershed and 100-meter buffer scales. The linear regression model was compared with different eigenvector-based spatial filtering models. These spatial filtering models were constructed using five different spatial conceptualizations based on different graph types generated from the geographic coordinates of the sampling sites. Population density, elevation, and percentage sand in the watershed and riparian regions are most important in explaining DO concentration and electric conductivity. A human signature as population density and increased sand and gravel cover can be detected in this watershed impacting water quality. Among different graph types compared, the relative graph type provided the highest model strength signifying stronger upstream-downstream relationship to DO, while k-mean graph types with four neighbors provided the strongest model performance, indicating the impact of local factors in electric conductivity. The relationships between socio-environmental factors and water quality and their spatial interrelationships identified in this work shed light on the source, mobilization, and transport of DO and conductivity and can assist the water quality management endeavor.

1. Background

1.1. Landscape characteristics and water quality

A stream's water quality is a result of a complex interaction of natural and anthropogenic processes in the watershed. Land-use change, population density, geology, and topography affect water quality in rivers (Baker, 2003; Lintern et al., 2018a). Human-modified land use is generally associated with degraded water quality and undermines ecosystem sustainability, including degradation of the freshwater ecosystem (Allan, 2004; Foley et al., 2005; Zampella et al., 2007). The anthropogenic impacts on surface water quality are not always straightforward, as complex interactions among various social, environmental, climatic, and political factors determine the consequences of these changes (Baker, 2003; Turner and Rabalais, 2003). These impacts are usually manifested as increased water temperature, increased nutrients (e.g., nitrogen and phosphorus), salt compounds, reduction in oxygen availability, and increased conductivity (Lintern et al., 2018a). The high concentration of nutrients and increased water temperature typically results in reduced oxygen levels in the water, as increased temperature reduces the solubility of oxygen, and remaining dissolved oxygen is also consumed rapidly by aquatic organisms, signifying eutrophication and deteriorated water quality (Cox, 2003).

Researchers have been using watershed characteristics at different scales to understand the spatial patterns of different water quality parameters across the stream network (Allan, 2004; King et al., 2005). Different landscape characteristics such as landcover types, topography, and other relevant explanatory features are extracted at scales including the entire watershed, riparian buffer, or some intermediate scales. The scale effects are not universal, as some factors are likely to affect water quality at the riparian scale, while others tend to do that at a watershed scale (Mainali et al., 2019). These relationships are different among different sites, seasons, and parameters studied as well. For example, Uriarte et al., (2011) reported that turbidity and DO responded to land-use change at a larger watershed scale while nitrogen was affected at the riparian buffer scale. While Mainali and Chang (2018) found a generally stronger influence on water quality at the stream buffer scale, the impact of scale in their model performance varied according to the parameters studied and seasons at which water quality data were collected. Some studies like Pratt and Chang (2012), Sliva and Williams (2001), and Zampella et al., (2007) reported a more significant influence of the whole watershed than a 100m buffer in their analyses.

Regression modeling approaches are commonly used to explore landscape factors affecting water quality at different scales. As water quality information is tied to location, regression modeling approaches are expected to incorporate spatial interrelationships among different locations from which water quality information is collected. If spatial relationships are not considered, regression modeling might violate the assumption of independence of the residuals of such models. There are several spatial regression models that overcome the limitation of ordinary least square (OLS) models in analyzing the relationship between water quality and landscape variables. These models include spatial lag and error models (Anselin, 1988), spatial eigenvector-based models (Borcard and Legendre, 2002; Tiefelsdorf and Griffith, 2007), geographically weighted regression (GWR) models (Brunsdon et al., 1998), and spatial stream network-based models

(Peterson and Hoef, 2010; Ver Hoef et al., 2006). In this work, we use an eigenvector-based spatial filtering-based regression method to explore the relationships between water quality and landscape matrices. We use eigenvector-based spatial filters to capture the spatial heterogeneity in the data and remove any clustering of residuals, which might lead to residual spatial autocorrelation (Getis and Griffith, 2002). Spatial filtering techniques generate a new set of explanatory variables representing the response variable's spatial structure. A selected set of those eigenvectors are then used as spatial predictors along with other predictor variables in the regression models. This approach has been recently used to model average and trends in water quality (Mainali and Chang, 2020, 2018).

1.2. Spatial Filtering and Different Graph Types

In the water quality modeling literature, different spatial conceptualizations of sampling sites, and their role in model outputs are not adequately explored (Mainali et al., 2019). Most studies use the spatial filtering approach with standard neighborhood criteria and weight matrix parameters without any attempt to modify them. In this work, we aim to explore how spatial conceptualizations of sampling sites rendered as different graph types in spatial-filtering affect the model output of DO and conductivity. We generate spatial eigenvector-based filters using five different graph types -- Delaunay, Gabriel, Relative, Minimum Spanning Tree, and k-mean—and use respectively fitted spatial filters in the regression model to compare their effectiveness in modeling dissolved oxygen and conductivity against the simple linear regression models.

1.3. Water Quality in Nepal

This work uses the Setikhola watershed in central Nepal as a case study to explore the relationships between water quality and landscape features in the Nepal Himalaya. In Nepalese Himalaya, different water quality parameters respond to the differences in land use, land management, natural vegetation, and atmospheric deposition that are usually directly affected by elevation (Jenkins et al., 1995). As in most of the other parts of the world, nutrient loss from forested lands is lower than non-forested lands in the Himalayan region (Pandey et al., 1983). Collins and Jenkins (1996) reported that although the agriculture catchments showed higher ammonium content during the wet season, they were unlikely to damage aquatic biota in Nepal’s mostly non-commercial agriculture practices. However, fertilizer input per hectare has since substantially increased, from 31 kg in 1995 to 131 kg in 2015. As a result, surface water pollution due to agricultural runoff has also increased, especially in the mid-hill and lowland Terai region of Nepal (Bista et al., 2016; Sharma et al., 2005). Urbanization has also significantly increased in Nepal. In the study watershed, the urbanized area more than doubled from 1990 to 2013 (Rimal et al., 2015). The impact of urbanization on water quality is sparsely studied in Nepal and is mostly focused in the capital city of Kathmandu (Kannel et al., 2007a, 2007b; S. Hammoud et al., 2018; Vaidya and Labh, 2017). The spatially explicit information related to water quality and the role of different landscape characteristics were not explored in the study watershed.

1.4. Dissolved Oxygen and Conductivity

We assessed the spatial patterns of DO and conductivity using the data collected from the field in December 2018 and January 2019. DO and electrical conductivity were chosen because they are important indicators of water pollution and the ecological integrity of surface water bodies (Cox, 2003; Lintern et al., 2018a). Data related to conductivity provide us information about the ability of water to pass electrical current, a measure of the availability of anions usually sourced from various chemicals, including alkali, chlorides, sulfides, and carbonate compounds. Conductivity is also related to temperature, as a warmer temperature tends to have higher conductivity (US EPA, 2013). Conductivity values are important indicators of biological integrity, as changes in conductivity usually indicate that pollution from discharge or other sources is entering the water bodies. The survival of aquatic organisms like fishes, algae, and macrophytes is directly related to oxygen availability in water. DO provides information about the human impacts in the water bodies, as increased temperature from anthropogenic activities leads to the reduction of dissolved oxygen. Polluted water has lower DO concentration because aquatic plants and bacteria in the polluted water consume oxygen, as does the decay of organic materials, which leads to eutrophic conditions (USGS DO, 2006).

1.5. Objectives and Research Questions

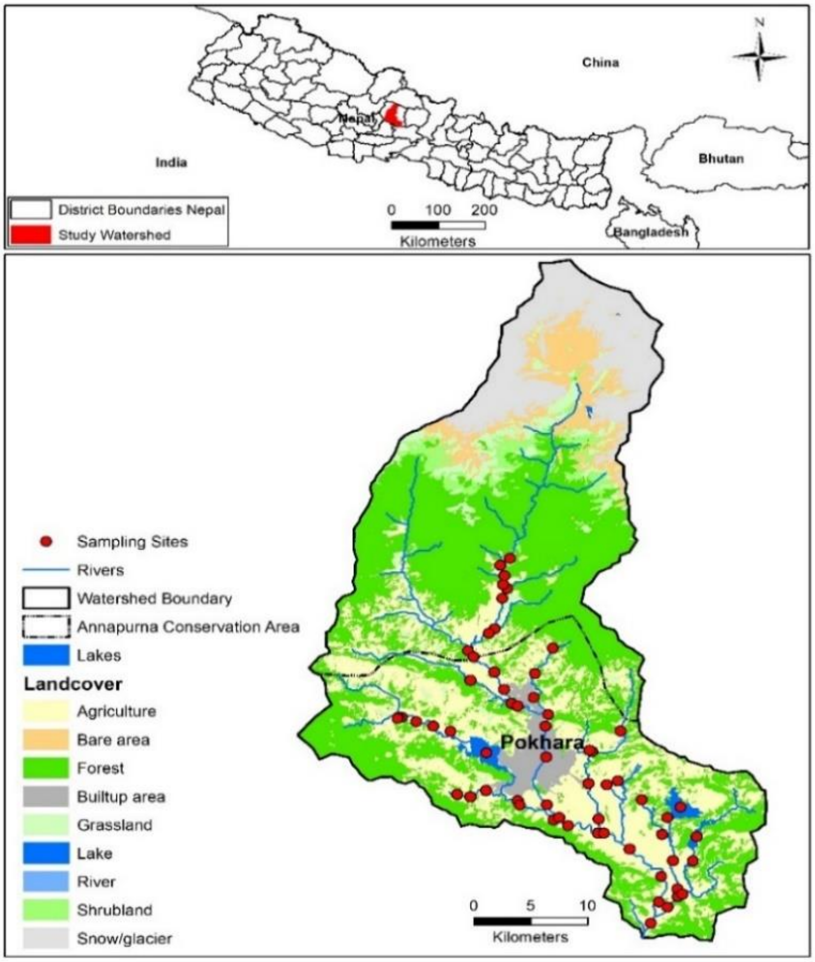
A recent review by Mainali et al. (2019) reported that different spatial conceptualizations of the sampling sites to incorporate the neighborhood impacts on water quality remain unexplored in water quality modeling literature. In this work, we compare various spatial conceptualizations of sampling sites by leveraging the graph theory literature and statistical packages available in R software. We attempt to answer the following research questions:

(1) How do DO and conductivity spatially vary in this watershed? (2) How different landscape features like the land cover, topography, and population density affect the water quality in the study watershed? and (3) How do different spatial conceptualizations of the sampling sites affect model results in this watershed?

2. Methods

2.1. Study area

Our study area is the Setikhola watershed which includes the Pokhara valley and adjoining hills and mountains (Figure 2-1). It provides an example of an urbanization gradient in Nepal (Rimal et al., 2015). The city of Pokhara is one of the biggest cities in Nepal and a famous tourist destination, and gateway to the popular Annapurna Conservation Area. The valley floor is a metropolis with a population greater than 500,000, while the hills are dominated by subsistence agriculture. The high elevation regions are mostly near-wilderness with forests, prairies, and snow-covered mountains, protected as a part of the Annapurna Conservation Area (ACAP, 2017). The area of this watershed is about 990 km^2 and includes 381 kilometers of the river; three major lakes cover approximately 9 km^2 (Baral Gauli et al., 2016)



.

Figure 2-1: Map of the study area with sampling sites

The elevation of the watershed ranges from 700 meters to more than 8000 meters above sea level. This watershed is located in one of the wettest regions of Nepal, with a total annual rainfall of about 4000 to 5400 mm, most of which falls during the monsoon season, June-August (CBS, 2013). The flow of rivers and the volume of lakes respond to the cyclic pattern of rainfall. The flow rate of the river was recorded at 40 ± 37 m^3/s during June and July of 2012 (Pokharel et al., 2018). The lake system of the valley floor was recently added to the list of important wetlands as a Ramsar site (Baral Gauli et al., 2016). The water bodies of the proposed study area are home to dozens of waterbird species, native fishes, endangered otters, and amphibians (Bhandari and GC, 2008; Husen and Sherpa, 2017; Kafle et al., 2008). Many endangered raptors, including the slender-billed vulture, also inhabit this area and depend on the water resources directly and indirectly.

Most of the recent biodiversity-related studies in this region only focused on terrestrial systems like forests and rangelands, typically overlooking aquatic biodiversity (Thapa et al., 2015). The water system is an important habitat for different aquatic organisms, provides ecosystem services to people living around it, and is also a major economic driver in this valley, including the tourist attractions in lakes and rivers, and fishery activities in the lakes (Gurung et al., 2005; Husen and Sherpa, 2017). Understanding the factors affecting the quality of surface water, therefore, is of paramount importance for both people and the ecosystem in this watershed.

2.2. Data Collection

2.2.1. Water quality data

We sampled 93 data points from rivers and lakes of the watershed. These data points were aggregated to 61 points after combining duplicate sampling in the river and different locations in the lake (Table 2-1). We collected pH, conductivity, DO, and temperature data using the YSI probe (Professional Plus #603190). We also collected several other ancillary data such as land-use, depth, and width of the stream, pollution signs, and the pictures of the waterbodies we sampled. The field data were collected during December 2018 and January 2019. This dry winter period was chosen to minimize the effect of meteorological factors on water quality. In this work, we only use conductivity and DO data because they were stable across the different times of the day in the watershed, thereby allowing spatial pattern analysis.

2.2.2. Landcover Data

A landcover classification of a Landsat 8 image was performed using the Google Earth Engine (Google Earth Engine, 2020). A cloud-free image was selected for the year 2017 as there was not any cloud-free image available for the year of 2018 or early 2019 when sampling was performed. We used the Classification and Regression Tree (CART) classification method to classify land cover into seven different classes (Urban Light, Urban Dense, Agriculture, Forest, Sand, Bare, Snow & Glaciers). The overall accuracy of the landcover map was about 82 percentage. The accuracy was measured by creating an error-matrix with a total of 115 polygons. Based on landcover information collected in the field, a set of known landcover type polygons were created, covering the entire watershed. The landcover category of those polygons was compared with the classified image by creating a confusion matrix (Lewis and Brown, 2001). The confusion matrix provides us information about the percentage of pixels correctly classified in different landcover types. The confusion matrix was used to calculate the user's accuracy and the producer's accuracy, which were averaged to derive an overall accuracy.

2.2.3. Population Data

The latest population estimate based on WorldPop data was used (WorldPop Nepal, 2015). This is a 100-meter resolution population estimate for the year of 2015. The population raster was clipped with a watershed boundary shapefile.

2.2.4. Topographic Data

We used the Department of Survey, Government of Nepal's 20-meter contour data as our elevation dataset. This dataset was interpolated using the topo-to-raster the interpolation technique with ArcGIS (ArcGIS 10.5.1, 2020). The elevation surface was converted into a slope raster using the surface analysis tool of ArcGIS 10.5.1. The interpolated elevation surface was also used to delineate the watershed boundary for each sampling station. The watershed polygons were used to extract the percentage of different landcover types, human population density, and an average of elevation and slope.

Table 2-1: List of different types of data used in the analysis

Data Name	**Type**	**Resolution**	**Source**
Water Quality	Point	Point data	Field sampling, 2019
Land Use Types	Raster	30-meter raster	Classified from Landsat 8, 2017
Elevation and Slope	Contour layer converted to raster	30-meter raster	Department of Survey, Nepal, 1986
Population	Raster	100 m, resampled to 30 m	WorldPop Nepal, 2015

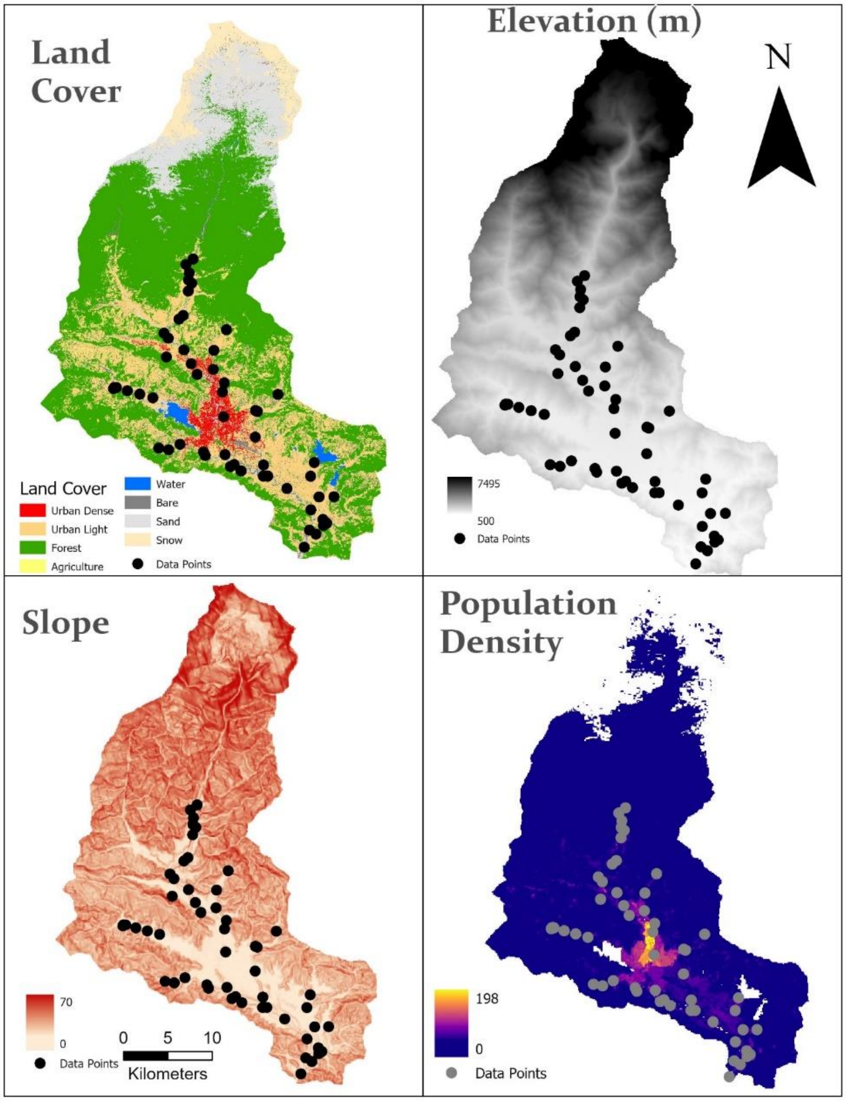


Figure 2-2: Spatial patterns of different explanatory variables used in the analysis

2.3. Data Processing and Analysis

2.3.1. Watershed delineation and predictor variables extraction

The watershed and subwatershed boundaries of the study area were delineated for each sampling point using the watershed hydrology tool of ArcGIS, which involved calculating flow direction, flow accumulation, and delineation of watershed boundary based on the user-defined outlet. We used the zonal statistics tool to calculate an average and standard deviation of elevation, slope, and population. The zonal histogram tool was used to calculate the number of pixels of each landcover type for each watershed draining to the sampling points. That value was converted to the percentage of each landcover type. A buffer of 100 m from the center of the stream was calculated using the buffer tool in ArcGIS. Those buffer polygons were clipped for each watershed. Predictor variables were extracted for the buffer of each watershed draining into the sampling point.

2.3.2. Exploratory Data Analysis

We mapped the spatial patterns of different water quality parameters and compared the differences between rivers and lakes. To test whether there is significant spatial clustering, we carried out spatial cluster and outlier analysis (Anselin Local Moran’s I) statistics using ArcGIS. This clustering was used to map high and low-value clusters of the water quality parameters in the watershed.

2.3.3. Regression analysis

After all the explanatory data sets were extracted for each sampling point, we used R version 3.6.1 software to analyze the data (Bivand, 2019; R Core Team, 2019). Only stream data points were used during regression analysis to remove any noise from the lakes. The response data sets were evaluated for their distribution using the Shapiro-Wilk test. We found that DO concentration was normally distributed while conductivity was not. Therefore, water conductivity was log-transformed before the regression modeling. The variation inflation factor (VIF) statistics were run to identify the predictor variables that were not autocorrelated. We chose predictor variables having VIF less than 10.

Using the predictor variables, regression analysis was run for dissolved oxygen and conductivity both at the watershed and buffer scale.

2.3.4. Spatial Regression Models and Different Graph Types

In this work, different spatial interrelationships among sampling sites were explored using graph theory. Graph theory uses the simple mathematical concept of nodes connected by the edges that have weights and directions. These edges connected by nodes can be used to decipher the processes and mechanisms of the underlying spatial phenomenon being studied (Dale and Fortin, 2010). There are several graph types being used in graph theory literature. These different graph types have different levels of connectivity and result in different adjacency matrix (Yan et al., 2019). We hypothesize that using different connectivity matrices resulted from these graph types allows us to examine the spatial relation among sampling stations to better understand the underlying process and mechanism of water quality parameters. A default spatial graph type of spatial filtering algorithm is the Delaunay graph type, a 6-node degrees graph type (each node connects to 6 other nodes). The other graph types used are the subgraphs of the Delaunay that have different node degrees: Gabriel- 4, Minimum Spanning Tree- 2, k-nearest neighbor- 2, and relative - 3 (Dale and Fortin, 2010). All the graph types used in this analysis are undirected maps where edges link two vertices symmetrically (Figure 2-3). Some of the graph types, like Relative and Minimum Spanning Tree, mimic the stream network to a certain extent.

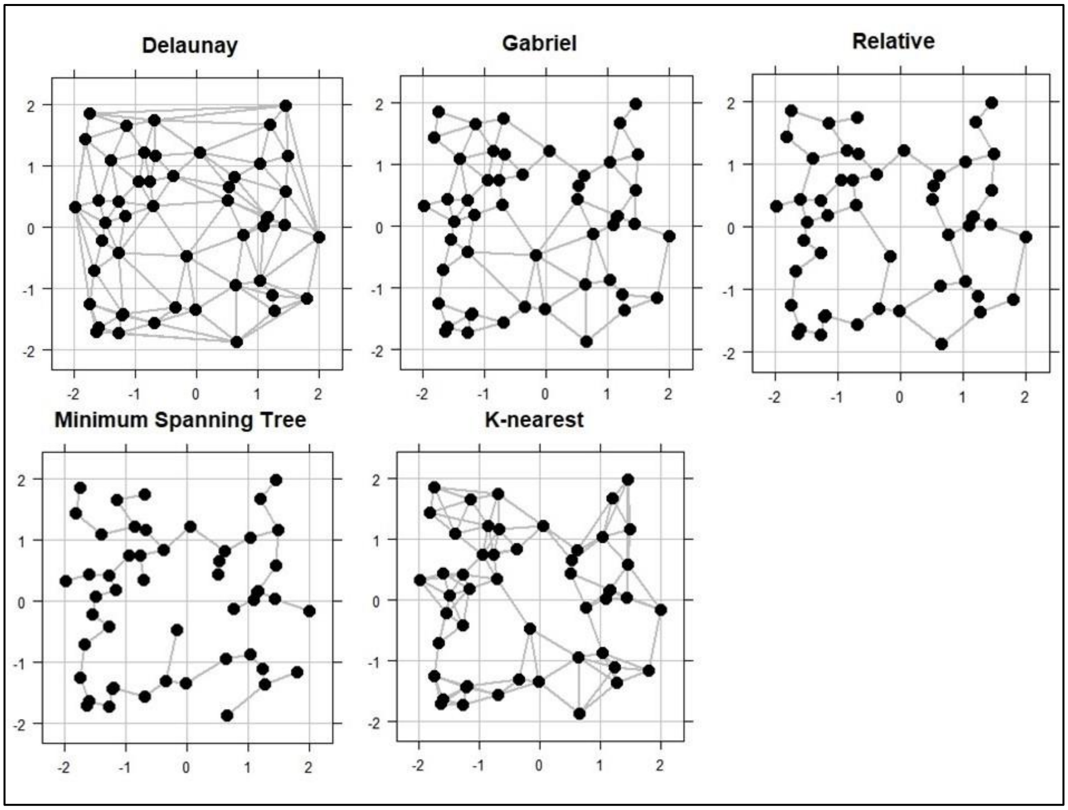


Figure 2-3: Schematic representation of spatial patterns of the data points based on different graph types (Data points are created randomly using R software version 3.6.1)

Spatial-filtering algorithms were implemented using the *spatialreg* package in R version 3.6.1 (Bivand, 2019; R Core Team, 2019). The first step of this process involved creating a weight matrix based on neighborhood criteria using different graph types (Figure 3). Each weight matrix was then decomposed and transformed using a set of mathematical functions to create eigenvalues and corresponding n-1 eigenvectors (Chun et al., 2016; Tiefelsdorf and Griffith, 2007). A set of fitted spatial filters that mimics the spatial structure of the response variable and can reduce the residual spatial autocorrelation was then selected to use as predictor variables along with other environmental variables in spatial regression for each graph type (Tiefelsdorf and Griffith 2007).

The eigenvector-based spatial filtering can be expressed as the following equation.

$$Y = X\beta + E_k\beta_\varepsilon + \varepsilon \qquad (1)$$

In equation 1, Y is a dependent variable, X is a matrix of independent variables. E_k denotes the selected matrix of fitted spatial-filtering based eigenvectors, β is a set of regression coefficients for predictor variables, β_ε is a set of regression coefficients for selected eigenvectors, and ε is random noise (error) (Chun et al., 2016; Mainali and Chang, 2018).

3. Results

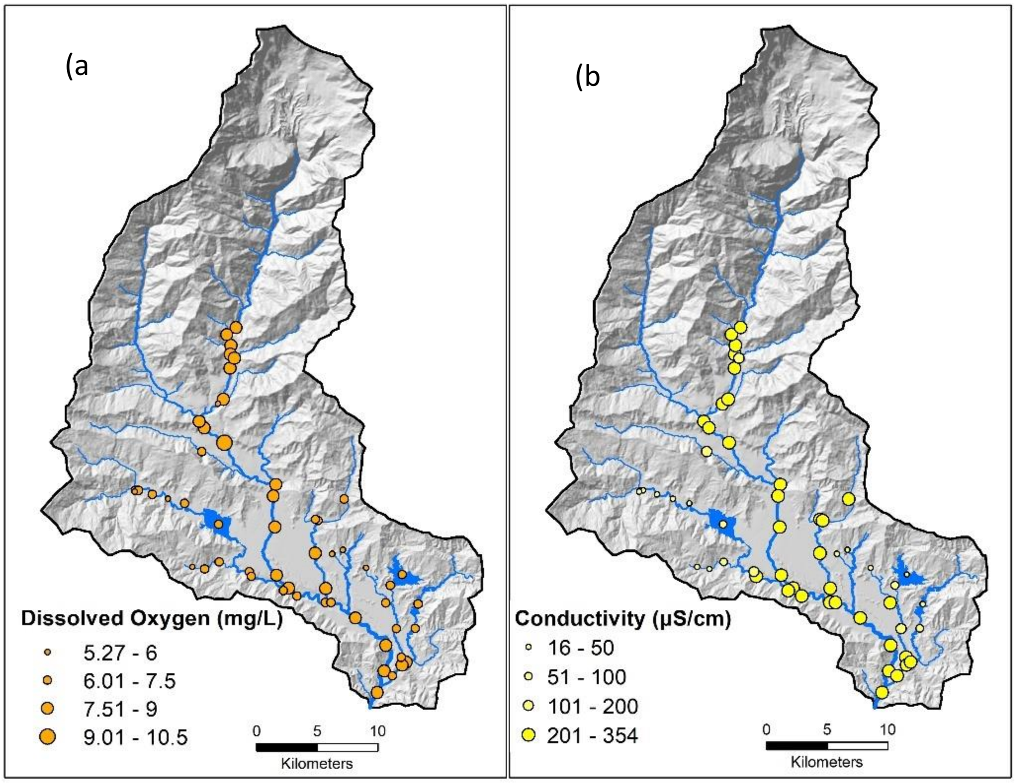


Figure 2-4: Spatial patterns of concentration of a) DO and b) Conductivity

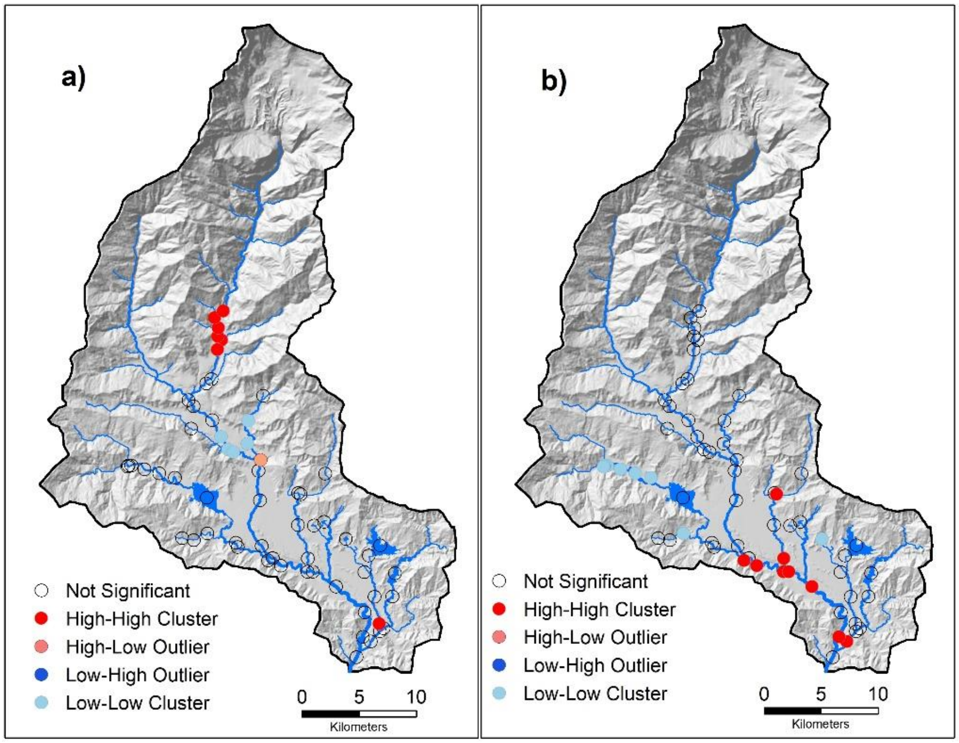


Figure 2-5: Spatial clustering of the data values a) DO and b) Conductivity

3.1. Spatial Patterns

DO

The DO values of the watershed range from 4.7 to 10.38 mg/L with an average concentration of about 7.00 mg/L. The DO concentration is highest in the main stem Seti River while they are lower in other tributaries and lakes (Figure 2-4a). There are clusters of high DO values in the high elevation regions, but no low-low clusters (Figure 2-5a). The median difference of DO is significant ($p < 0.01$, t-test) between rivers and lakes (Figure 2-6a), with higher DO in rivers than lakes. DO values along the Setikhola stem are the highest. This result shows that the main stem of Setikhola River has an excellent DO range to support aquatic life, while DO in lakes and other tributaries are lower.

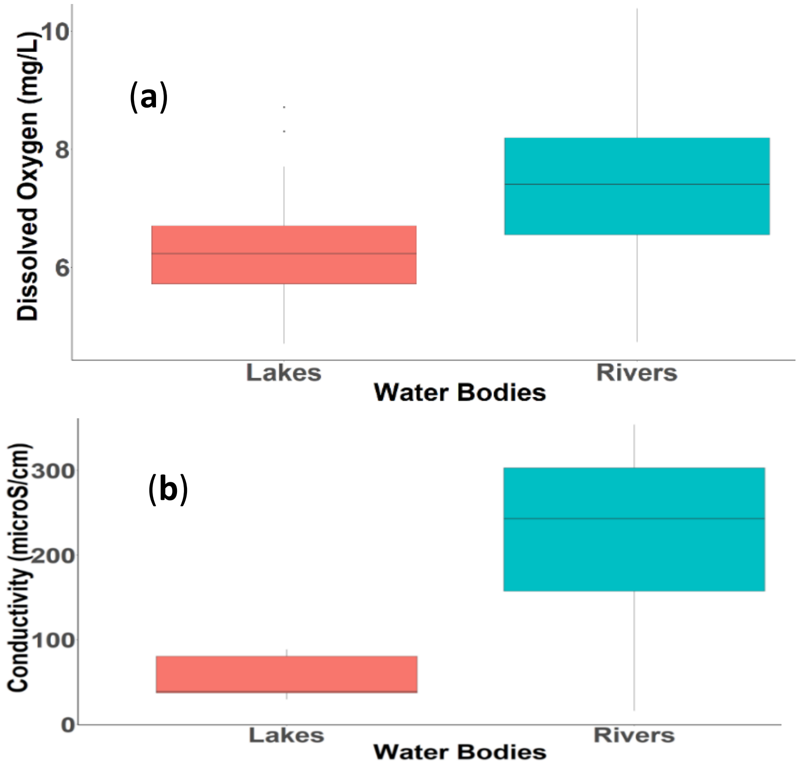


Figure 2-6: Range of DO (a) and conductivity (b) values in lakes and river

Conductivity

The conductivity of this watershed ranged from 16.1 to 354 µs/cm with a mean of about 150 µs/cm. Pokharel et al. (2018) reported an average of 166 µs/cm conductivity in the Seti-Khola River. In Figure 4b we can see that some of the western tributaries have significantly lower conductivity than the rest of the watershed (Figure 2-5a). Conductivity also substantially differed between rivers and lakes in this watershed, with significantly higher values in rivers than lakes (Figure 2-6b).

3.2. Correlation Analysis

Table 2-2: Pearson Correlation analysis (n = 54) between landscape matrices and water quality parameters at different scales. * significant at 0.05, ** significant at 0.01 level of significance

	Dissolved Oxygen		**Conductivity**	
	Buffer	**Watershed**	**Buffer**	**Watershed**
Elevation	0.48*	0.50**	0.25	0.30*
Elev Std	0.54**	0.55**	0.47**	0.52**
Slope	0.47**	-0.08	0.17	-0.04
Slope Std	0.44**	-0.18	0.41**	-0.08
Population Mean	-0.02	-0.03	0.39**	0.05
Pop Std	0.11	0.05	0.59**	0.46**
Urban Dense	0.02	0.10	0.13	0.001
Urban Light	-0.46**	-0.01	-0.26	0.115
Forest	0.43**	0.01	0.17	-0.19
Agriculture	-0.30*	-0.01	-0.29*	0.09
Sand	-0.23	-0.1	-0.48**	0.1
Bare	0.18	-0.085	0.27	0.085

The elevation standard deviation was significantly associated with both DO and conductivity at both scales, while slope was positively correlated with DO at buffer scale only (Table 2-2). But slope standard deviation was correlated significantly with DO at the buffer scale while with COND at both scales. The average population density was significant for COND at buffer scale only, while the standard deviation was significant at both scales. The forest landcover was significantly positively correlated with DO at buffer scale, while agriculture was significantly positively correlated with both DO and COND at buffer scale but not at the watershed scale. The percentage of the sand cover was significantly negatively correlated with the conductivity at the buffer scale.

3.3. Regression Results

The R^2 value of the DO model ranged from 0.25 to 0.5 while R^2 values of conductivity ranged from 0.3 to 0.85 (Table 2-3 & 2- 4). The higher R^2 values for both spatial and aspatial models were reported using the 100-meter buffer scale. Figure 2-7 displays spatial interrelationship among different sampling locations. The Relative and Minimum Spanning Tree graph types are the closest representation of the stream network, while K-nearest graph types have revealed the local clusters based on the immediate neighbors. The relative graph type yielded the highest model performance for DO, while the k-mean graph type yielded the highest model performance for conductivity (Figure 2-8).

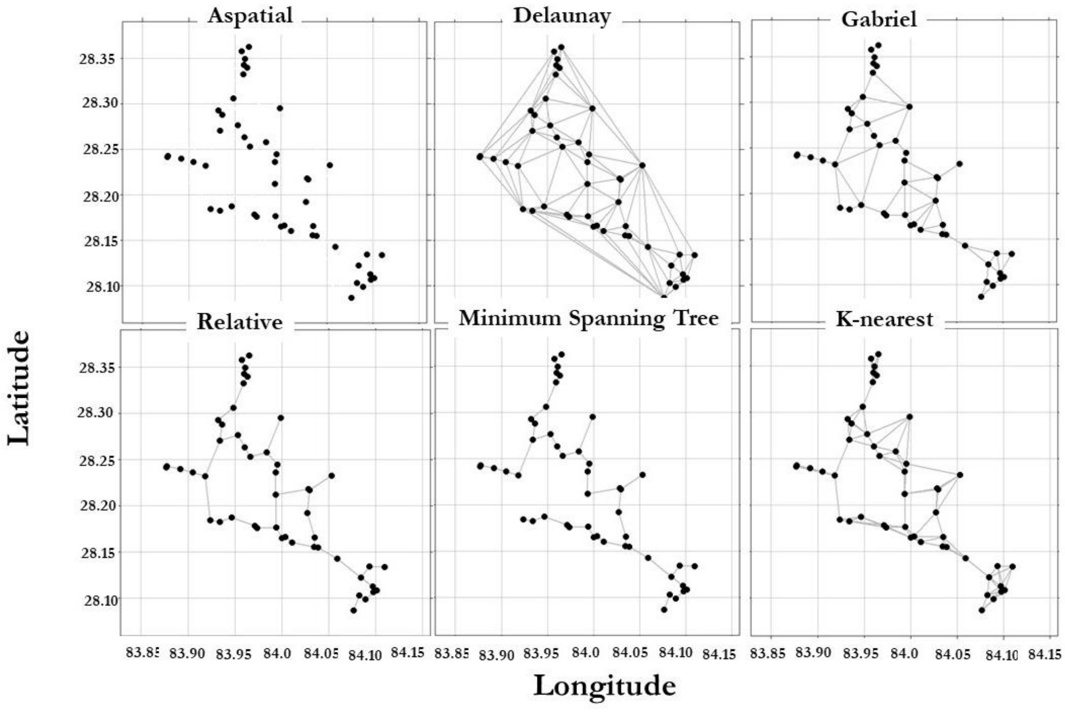


Figure 2-7: Different spatial interrelations of the study sites based on different graph types

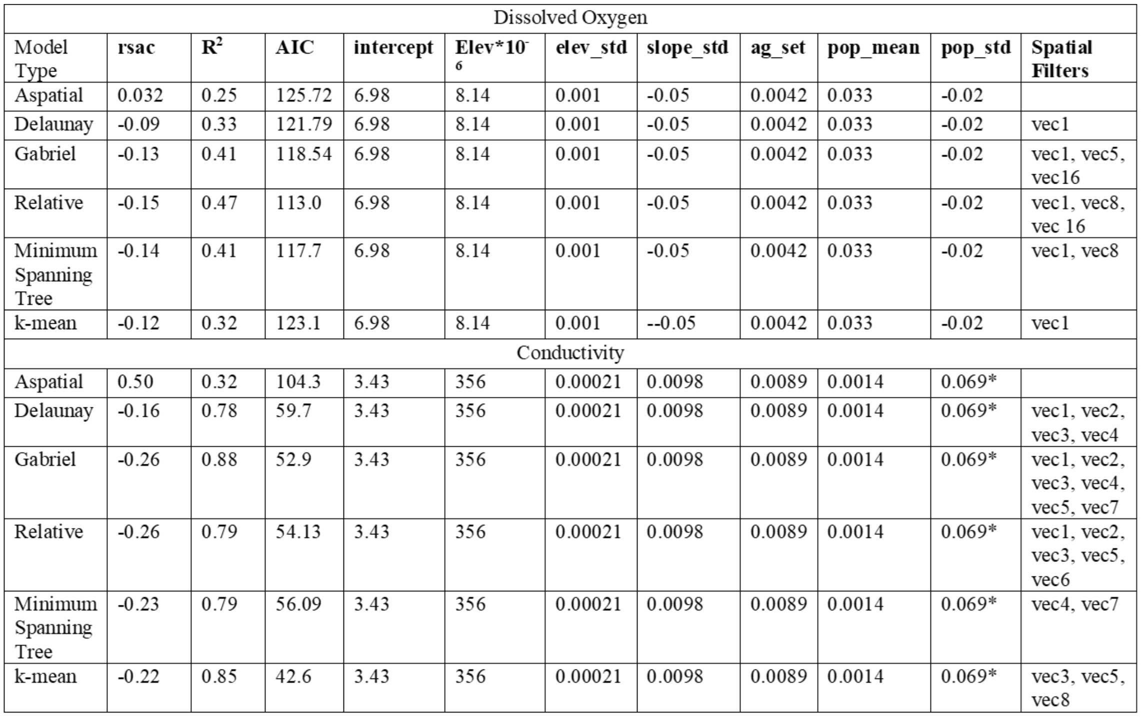
3.3.1. Dissolved Oxygen Regression Model

Different spatial conceptualizations yielded various model strengths for DO. The R^2 values with explanatory variables at the watershed scale ranged from 0.25 to 0.48, while it is generally higher at the buffer scale with values ranging from 0.35 to 0.5. All models were statistically significant with a 95 percent confidence interval ($p<=0.05$). As shown in Figure 2-8a, spatial filtering-based regression always increases model performance, but the highest model performance for DO models was achieved when the relative graph type was used in both watershed and buffer scales. Only the standard deviation of elevation was a significant predictor at a watershed level. The standard deviation of the population and percentage of sand/gravel were significant predictors at the 100-meter buffer scale (Table 2-3). The best model was derived using the relative graph spatial conceptualization at the buffer scale, with predictor variables % sand, and eigenvector number 6 and 16.

3.3.2. Conductivity Regression Model

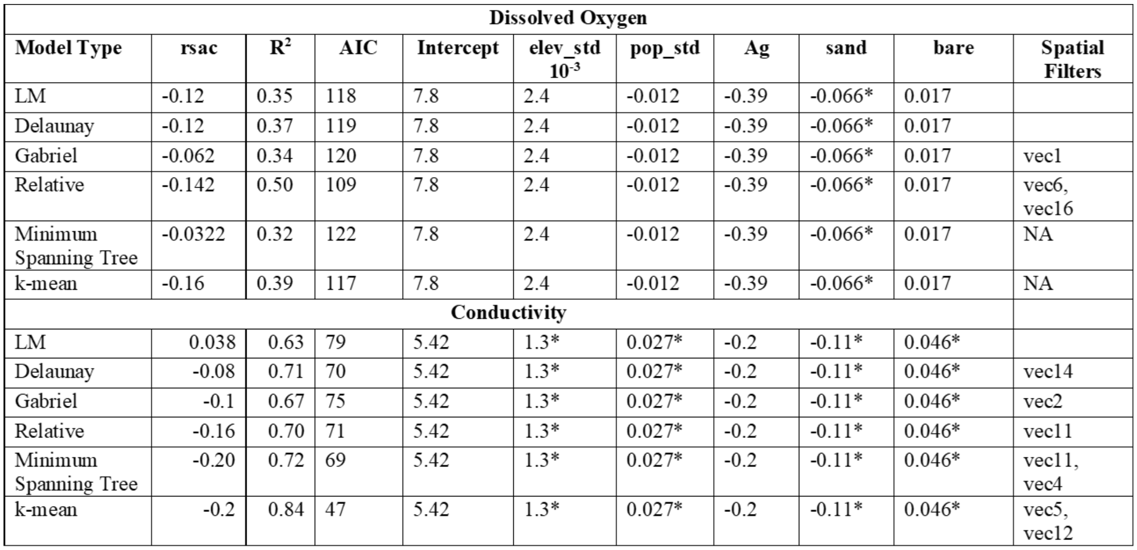
The conductivity model strengths were generally higher than DO. All models were significant at $p<=0.05$. The model strengths of conductivity also varied according to different spatial conceptualization. The R^2 values ranged from 0.3 to 0.85 at the watershed scale while the buffer scale model strength ranged from 0.62 to 0.84 R^2 values (Table2- 4, Figure 2-7b). Buffer scale models were usually weaker for conductivity models except for the aspatial linear model. The k-mean graph model strength was comparable between watershed and buffer scale models which also yielded the highest model strengths at both scales. In the regression model, the population standard deviation was always positively related to conductivity. When k-mean spatial conceptualization was used, the average elevation was also positively associated with conductivity at the watershed scale. But at the buffer scale, elevation standard deviation, population standard deviation, and percentage bare land positively explain the variation of conductivity while percentage sand predicts it negatively (Table 2-4). The k-mean graph at the watershed and the buffer scales had an R^2 value close to 0.85. However, the watershed scale model is simpler, with elevation and population standard deviation along with eigenvectors 3, 5, and 8 as the predictor variables.

Table 2-3: Watershed scale model attributes for Dissolved Oxygen and Conductivity. Full Forms: rsac: Residual Spatial Autocorrelation z value. AIC: Akaike Information Criteria. elev: average elevation, elev_std: standard deviation of elevation, slope_std: slope standard deviation



Dissolved Oxygen											
Model Type	**rsac**	**R^2**	**AIC**	**intercept**	**Elev*10^{-6}**	**elev_std**	**slope_std**	**ag_set**	**pop_mean**	**pop_std**	**Spatial Filters**
Aspatial	0.032	0.25	125.72	6.98	8.14	0.001	-0.05	0.0042	0.033	-0.02	
Delaunay	-0.09	0.33	121.79	6.98	8.14	0.001	-0.05	0.0042	0.033	-0.02	vec1
Gabriel	-0.13	0.41	118.54	6.98	8.14	0.001	-0.05	0.0042	0.033	-0.02	vec1, vec5, vec16
Relative	-0.15	0.47	113.0	6.98	8.14	0.001	-0.05	0.0042	0.033	-0.02	vec1, vec8, vec 16
Minimum Spanning Tree	-0.14	0.41	117.7	6.98	8.14	0.001	-0.05	0.0042	0.033	-0.02	vec1, vec8
k-mean	-0.12	0.32	123.1	6.98	8.14	0.001	--0.05	0.0042	0.033	-0.02	vec1
Conductivity											
Aspatial	0.50	0.32	104.3	3.43	356	0.00021	0.0098	0.0089	0.0014	0.069*	
Delaunay	-0.16	0.78	59.7	3.43	356	0.00021	0.0098	0.0089	0.0014	0.069*	vec1, vec2, vec3, vec4
Gabriel	-0.26	0.88	52.9	3.43	356	0.00021	0.0098	0.0089	0.0014	0.069*	vec1, vec2, vec3, vec4, vec5, vec7
Relative	-0.26	0.79	54.13	3.43	356	0.00021	0.0098	0.0089	0.0014	0.069*	vec1, vec2, vec3, vec5, vec6
Minimum Spanning Tree	-0.23	0.79	56.09	3.43	356	0.00021	0.0098	0.0089	0.0014	0.069*	vec4, vec7
k-mean	-0.22	0.85	42.6	3.43	356	0.00021	0.0098	0.0089	0.0014	0.069*	vec3, vec5, vec8

Table 2-4: Buffer scale model attributes for Dissolved Oxygen and Conductivity. AIC: Akaike Information Criteria. elev_std: Standard deviation of elevation, pop_std: population standard Deviation, ag: percentage agriculture land cover, sand: percentage sand cover, bare: percentage bare land cover. * refers to the coefficients significant at $p \leq 0.05$.



Dissolved Oxygen										
Model Type	**rsac**	**R^2**	**AIC**	**Intercept**	**elev_std 10^{-3}**	**pop_std**	**Ag**	**sand**	**bare**	**Spatial Filters**
LM	-0.12	0.35	118	7.8	2.4	-0.012	-0.39	-0.066*	0.017	
Delaunay	-0.12	0.37	119	7.8	2.4	-0.012	-0.39	-0.066*	0.017	
Gabriel	-0.062	0.34	120	7.8	2.4	-0.012	-0.39	-0.066*	0.017	vec1
Relative	-0.142	0.50	109	7.8	2.4	-0.012	-0.39	-0.066*	0.017	vec6, vec16
Minimum Spanning Tree	-0.0322	0.32	122	7.8	2.4	-0.012	-0.39	-0.066*	0.017	NA
k-mean	-0.16	0.39	117	7.8	2.4	-0.012	-0.39	-0.066*	0.017	NA
Conductivity										
LM	0.038	0.63	79	5.42	1.3*	0.027*	-0.2	-0.11*	0.046*	
Delaunay	-0.08	0.71	70	5.42	1.3*	0.027*	-0.2	-0.11*	0.046*	vec14
Gabriel	-0.1	0.67	75	5.42	1.3*	0.027*	-0.2	-0.11*	0.046*	vec2
Relative	-0.16	0.70	71	5.42	1.3*	0.027*	-0.2	-0.11*	0.046*	vec11
Minimum Spanning Tree	-0.20	0.72	69	5.42	1.3*	0.027*	-0.2	-0.11*	0.046*	vec11, vec4
k-mean	-0.2	0.84	47	5.42	1.3*	0.027*	-0.2	-0.11*	0.046*	vec5, vec12

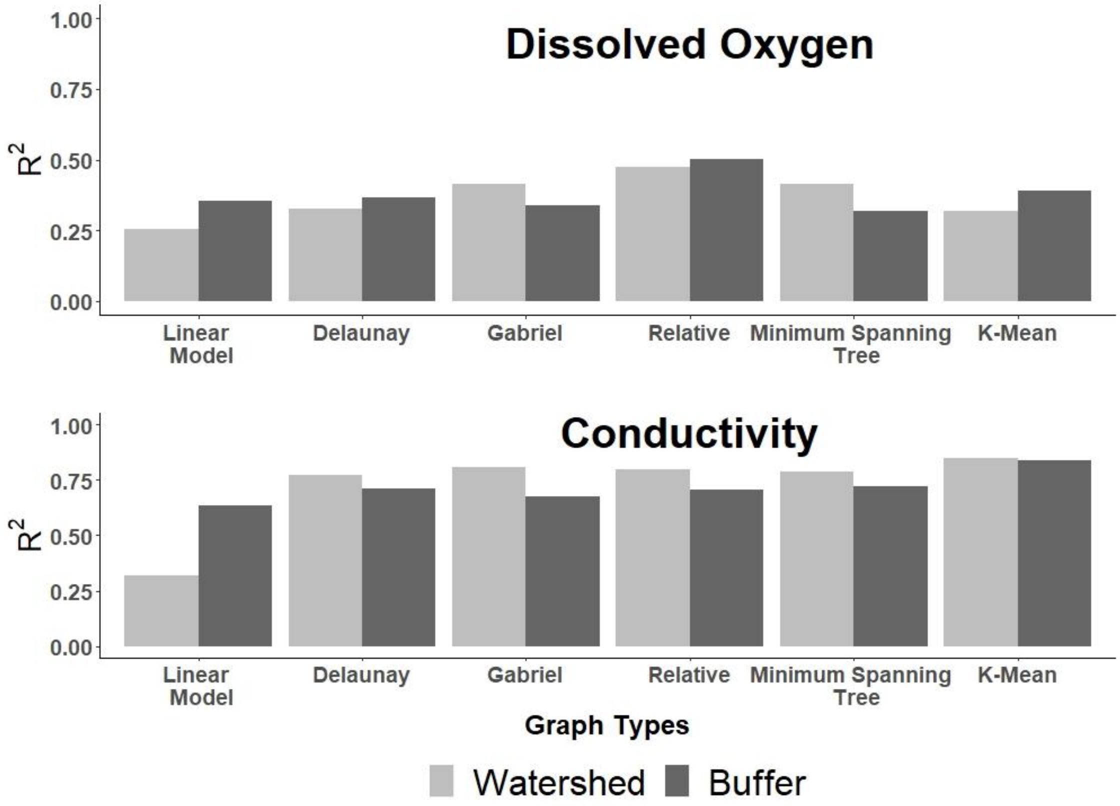


Figure 2-8: Model strengths of dissolved oxygen and conductivity at different scales and graph types

4. Discussions

4.1. Spatial Patterns of Dissolved Oxygen and Conductivity

Our DO range falls within the range reported elsewhere in Nepal and other Asian countries (Adhikari et al., 2017; Su et al., 2012; Yadav et al., 2019). Pokharel et al. (2018) reported an average of 8.0 mg/L in the Seti-Khola River from the data collected in July 2012. DO values greater than 4.0 mg/L are considered fair to support aquatic life, while higher than 6.5 is good, above 8.0 is excellent (Washington Ecology, 2002). In our study, DO is generally higher in the mainstream high-flow river, which is consistent with other studies that report increasing river flows are associated with high DO (Post et al., 2018). A relatively random spatial pattern for DO except for a high-high cluster of the high elevation result suggests that the factors affecting DO concentration are also randomly distributed in the watershed. The high-high cluster in the high elevation region might be associated with proximity to forest, cooler water temperatures coming from the snow and glaciers, the steeper slope leading, and higher turbulence resulting in rapid re-aeration. (de Mello et al., 2018; Su et al., 2013).

The conductivity range we reported is within a standard limit (max of 1500 µs/cm) according to the Nepal government (Water Quality Standard Nepal, 2005). Our conductivity values are within the range of previous studies like Pokharel et al. (2018) who reported an average of 166 µs/cm in this watershed. The higher range of conductivity in the high-flowing river like main SetiKhola and its bigger tributaries, and lower values in the smaller tributaries and lakes, suggest that conductivity is a function of watershed size and probably in-stream activities such as the dissolution of salts from bedrock. In a larger watershed, water delivered to the surface water comes in contact with more soil surface, thereby washing more ions and increasing conductivity (Water on the Web, 2020). We also cannot rule out the possibility that the differences in conductivity in different parts of the watershed might be a consequence of the differences in underlying geology: rock types with abundant dissolvable ions tend to increase water conductivity in the stream (Water on the Web, 2020).

Water quality in the study lakes was poorer than in the fast-flowing rivers that recycle nutrients and oxygen quickly. Both DO and conductivity were lower in the lakes. Notice, however, that there were some tributaries where conductivity was lower than the lakes, probably because of their small watershed size and/or underlying geology. In many cases, lakes have different water quality conditions from rivers' because of their stagnant nature, physicochemical conditions, and responses to receiving waters that are typically affected by a combination of natural and human impacts (Low et al., 2016). Lakes hold nutrients and increase concentration over time, which can lead to eutrophication. All the lakes in this region also suffered some form of eutrophication, with such impacts more visible in small lakes (Field visit 2018/2019). According to local people, the macrophyte growths in bigger lakes are periodically removed to make room for boats. The aquaculture practices in the lakes, like fish farming in some of the lakes, and other factors like the presence of the river in the watershed, land use, geology, and climate affect the intensity of human impacts in the lake (Nielsen et al., 2012; Zang et al., 2011).

4.2. Relationship between landscape matrix and water quality

We report that the riparian forest cover is positively correlated to DO, which is in line with other studies like de Mello et al. (2018). Urban land use did not directly correlate with either DO or conductivity. It is probably because urban land use only covers a small area and is not evenly distributed across the entire watershed. The strong correlation of forest land cover with DO at the buffer scale suggests that all other human-modified landcover types are detrimental to DO, as expected according to other studies (Zhou et al., 2012). The effect of land use in DO is manifested through increasing temperature, which leads to increased biological oxygen demand and depleted oxygen in the water bodies (Schindler et al., 2017). Various other studies have also found agricultural land use affecting DO significantly, which is consistent with our finding (Yadav et al., 2019). A negative effect of the built-up area and population growth on DO are also reported in various parts of the world (Su et al., 2013).

DO in surface water measures the ability of water to support life; it can be affected by various watershed factors. Different studies have found varying levels of success in modeling DO utilizing landscape characteristics and statistical approaches. Su et al. (2013) found a maximum of 0.83 R^2 when they compared various spatial statistical models for the Qintiang river of China, while de Mello et al. (2018) reported 0.72 in the Sarapui River basin of Brazil. Chang (2008) reported R^2 values in the range of 0.7 in the study of the Han River Basin, Korea. Although lower than these studies, we were successful in deriving the model with a reasonable R^2 value of 0.5 using a combination of somewhat limited socio-environmental (population standard deviation, agriculture, sand,

and bare land cover) and spatial-filter based variables. The remaining variations might be explained by geology, soil types, and climatic variables, which are unavailable in the study region. Our result suggests that the percentage of sand coverage at the stream banks is a significant determinant of DO. This finding suggests that the sand and gravel mining rampant in the riparian area of this watershed might be reducing oxygen availability in the water bodies. Some previous studies have shown that sand and gravel mining can affect the aquatic ecosystem and also degrade overbank areas (Sreebha and Padmalal, 2011). However, the exact mechanisms by which the gravel and sand mines impact surface water quality remain to be explored.

Conductivity can be modeled with watershed characteristics better than other water quality parameters because of easier movements of soluble ions to the water, which are unique to different landscape characteristics under consideration (Lintern et al., 2018b). We found a high of 0.8 R^2 value in the current study. Conductivity can be affected by various watershed levels and in-stream factors like the concentration of phosphorus and nitrogen in the water, area of wetland surrounding water bodies, and climatic factors like precipitation (Fracz and Chow-Fraser, 2013). We also found several of these factors affecting the conductivity concentration of the river reaches. The presence of agriculture or sand cover and high population density reduces conductivity significantly in our watershed, which aligns with the study by Wenner et al. (2003) who reported that degraded streams usually had lower conductivity.

4.3. Impacts of spatial scales

Various studies have found different results in terms of the scale at which landscape matrices affect water quality. Studies have found a stronger effect of watershed characteristics than buffer scale characteristics on water quality in their models (Houlahan and Findlay, 2004; Pratt and Chang, 2012; Zhou et al., 2012). In contrast, Mainali and Chang (2018) reported a 100-meter buffer as the best scale in explaining various water quality parameters in a larger river basin in South Korea. Similarly, we found generally higher model strength at the buffer scale for DO while similar model strengths between 100-m buffer and watershed scale for conductivity. Our results also indicate that there was a higher influence of different factors at the buffer scale than the watershed scale; land use in the immediate surrounding of the river like sand and agriculture are significantly making water quality worse by reducing DO and conductivity.

4.4. Impacts of different spatial conceptualizations

We report that spatial filters significantly increase model performance, and spatial conceptualizations matter when creating spatial filters because they produce different model outputs. When spatial eigenvectors are created, the weights are provided based on the values of the neighborhood, which are different in different graph types. For DO, the highest model strengths were with Relative Graph type while it was the k-nearest for conductivity. Relative and Minimum Spanning Tree are the graph types closest to the real river network of our watershed; a difference between Relative and Minimum Spanning tree is in the connections between stations on the west side of the watershed. Relative graph type is closer to the real river network as the edges in this graph more closely follow the river network. The highest model strength in Relative Graph type suggests that DO is more directly affected by upstream-downstream relations along with the river network. Many previous studies also showed that DO concentration was predominantly governed by various upstream factors like solute concentrations (Bailey and Ahmadi, 2014) and inclusion of upstream-downstream relationships improved the model performance of DO (Money et al., 2009).

The k-mean spatial conceptualization refers to the neighbors defined around its immediate surroundings in all directions. The higher conductivity model strength using k-mean spatial conceptualization suggests that conductivity is more affected by local than upstream factors. The local clustering of conductivity could be better captured by k-mean clustering than other graph types. Previous studies also reported that the electric conductivity of the river is influenced by neighbors in all directions, or upstream values (Lintern et al., 2018b; Peterson and Hoef, 2010). It is also to be noted that the model strengths using other graph types are also significant, and only slightly lower than the k-means spatial conceptualization.

5. Conclusions

The spatial patterns of DO and conductivity, their relationships with socio-environmental factors, and various spatial and statistical interrelationships identified in this work elucidate the source, mobilization, and transport of DO and conductivity and can guide water quality management efforts. In this watershed, we report that the spatial clustering pattern of DO is affected by upstream factors, thereby revealing distinct DO concentrations in the main-stem and tributaries. Conductivity also revealed distinct spatial variations in main-stem and other tributaries and exhibited local clustering across tributaries.

The spatial regression models were successfully developed and compared using water quality data collected in the field, and various geographic information systems based on social and environmental data. Among the factors considered in the analysis, we found the population density, agricultural land cover, and sand cover negatively impact the water quality as revealed by their relationships with DO and conductivity. The inter-scale comparison revealed a generally stronger impact of a 100-m riparian scale over the entire watershed in explaining the variation of DO and conductivity.

Our work provides a novel example of using graph theory in elucidating relationships among water quality measurement sites and their affinity with landscape processes. The model strengths are usually different according to the different spatial conceptualization of interrelations among sampling stations, as demonstrated by the graph types. Among different graph types compared, the relative graph types provided the highest model strength, signifying stronger up-stream downstream relation with DO, while k-mean graph types with four neighbors provided the strongest model performance, indicating the impact of local factors in water conductivity.

References

ACAP, 2017. Annapurna Conservation Area Project [WWW Document]. URL https://www.google.com/maps/d/viewer?mid=1ujqpD_gABtG1VX-k6cxqZQ4gtc0&hl=en (accessed 1.9.18).

Adhikari, P.L., Shrestha, S., Bam, W., Xie, L., Perschbacher, P., 2017. Evaluation of Spatial-Temporal Variations of Water Quality and Plankton Assemblages and Its Relationship to Water Use in Kulekhani Multipurpose Reservoir, Nepal. J. Environ. Prot. 08, 1270–1295. https://doi.org/10.4236/jep.2017.811079

Allan, J.D., 2004. Landscapes and Riverscapes: The Influence of Land Use on Stream Ecosystems. Annu. Rev. Ecol. Evol. Syst. 35, 257–284. https://doi.org/10.1146/annurev.ecolsys.35.120202.110122

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic, Dordrecht.

ArcGIS 10.5.1, 2020. ArcGIS Desktop 10.5.1 quick start guide—ArcGIS Help | ArcGIS Desktop [WWW Document]. URL https://desktop.arcgis.com/en/arcmap/10.5/get-started/setup/arcgis-desktop-quick-start-guide.htm (accessed 2.26.20).

Bailey, R.T., Ahmadi, M., 2014. Spatial and temporal variability of in-stream water quality parameter influence on dissolved oxygen and nitrate within a regional stream network. Ecol. Model. 277, 87–96. https://doi.org/10.1016/j.ecolmodel.2014.01.015

Baker, A., 2003. Land use and water quality. Hydrol. Process. 17, 2499–2501. https://doi.org/10.1002/hyp.5140

Baral Gauli, S., Dhakal, M., Khanal, R., 2016. Lake Cluster of Pokhara Valley. Department of National Parks and Wildlife Conservation and IUCN Nepal.

Bhandari, J., GC, D.B., 2008. Preliminary Survey and Awareness for Otter Conservation in Rupa Lake, Pokhara, Nepal. J. Wetl. Ecol. 1, 2.

Bista, D.R., Dhungel, S., Adhikari, S., 2016. Status of fertilizer and seed subsidy in Nepal: review and recommendation. J. Agric. Environ. 17, 1–10.

Bivand, R., 2019. Package ‘spatialreg.’

Borcard, D., Legendre, P., 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol. Model. 153, 51–68.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. R. Stat. Soc. Ser. Stat. 47, 431–443.

CBS, 2013. Environment Statistics of Nepal 2013. Government of Nepal, National Planning Commission Secretariat, Central Bureau of Statistics.

Chang, H., 2008. Spatial analysis of water quality trends in the Han River basin, South Korea. Water Res. 42, 3285–3304. https://doi.org/10.1016/j.watres.2008.04.006

Chun, Y., Griffith, D.A., Lee, M., Sinha, P., 2016. Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. J. Geogr. Syst. 18, 67–85. https://doi.org/10.1007/s10109-015-0225-3

Collins, R., Jenkins, A., 1996. The impact of agricultural land use on stream chemistry in the Middle Hills of the Himalayas, Nepal. J. Hydrol. 185, 71–86.

Cox, B., 2003. A review of dissolved oxygen modelling techniques for lowland rivers. Sci. Total Environ. 314–316, 303–334. https://doi.org/10.1016/S0048-9697(03)00062-7

Dale, M.R.T., Fortin, M.-J., 2010. From Graphs to Spatial Graphs. Annu. Rev. Ecol. Evol. Syst. 41, 21–38. https://doi.org/10.1146/annurev-ecolsys-102209-144718

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

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., others, 2005. Global consequences of land use. science 309, 570–574.

Fracz, A., Chow-Fraser, P., 2013. Changes in water chemistry associated with beaver-impounded coastal marshes of eastern Georgian Bay. Can. J. Fish. Aquat. Sci. 70, 834–840. https://doi.org/10.1139/cjfas-2012-0431

Getis, A., Griffith, D.A., 2002. Comparative Spatial Filtering in Regression Analysis. Geogr. Anal. 34, 130–140. https://doi.org/10.1111/j.1538-4632.2002.tb01080.x

Google Earth Engine, 2020. Google Earth Engine [WWW Document]. URL https://code.earthengine.google.com/ (accessed 2.26.20).

Gurung, T.B., Wagle, S.K., Bista, J.D., Joshi, P.L., Batajoo, R., Adhikari, P., Rai, A.K., 2005. Participatory fisheries management for livelihood improvement of fishers in Phewa Lake, Pokhara, Nepal. Himal. J. Sci. 3, 47–52.

Houlahan, J.E., Findlay, C.S., 2004. Estimating the ‘critical’distance at which adjacent land-use degrades wetland water and sediment quality. Landsc. Ecol. 19, 677–690.

Husen, M.A., Sherpa, L., 2017. Native Fish Species of Begnas And Rupa Lake. Local Initiatives for Biodiversity, Research and Development (LI-BIRD).

Jenkins, A., Sloan, W.T., Cosby, B.J., 1995. Stream chemistry in the middle hills and high mountains of the Himalayas, Nepal. J. Hydrol. 166, 61–79.

Kafle, G., Cotton, M., Chaudhary, J.R., Pariyar, H., Adhikari, H., Bohora, S.B., Chaudhary, U.K., Ram, A., Regmi, B., 2008. Status of and threats to waterbirds of Rupa Lake, Pokhara, Nepal. J. Wetl. Ecol. 1, 9–12.

Kannel, P.R., Lee, S., Kanel, S.R., Khan, S.P., Lee, Y.-S., 2007a. Spatial–temporal variation and comparative assessment of water qualities of urban river system: a case study of the river Bagmati (Nepal). Environ. Monit. Assess. 129, 433–459. https://doi.org/10.1007/s10661-006-9375-6

Kannel, P.R., Lee, S., Lee, Y.-S., Kanel, S.R., Pelletier, G.J., 2007b. Application of automated QUAL2Kw for water quality modeling and management in the Bagmati River, Nepal. Ecol. Model. 202, 503–517. https://doi.org/10.1016/j.ecolmodel.2006.12.033

King, R.S., Baker, M.E., Whigham, D.F., Weller, D.E., Jordan, T.E., Kazyak, P.F., Hurd, M.K., 2005. Spatial Considerations for Linking Watershed Land Cover to Ecological Indicators in Streams. Ecol. Appl. 15, 137–153. https://doi.org/10.1890/04-0481

Lewis, H.G., Brown, M., 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. Int. J. Remote Sens. 22, 3223–3235. https://doi.org/10.1080/01431160152558332

Lintern, A., Webb, J.A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P., Western, A.W., 2018a. Key factors influencing differences in stream water quality across space. Wiley Interdiscip. Rev. Water 5, 1–31. https://doi.org/10.1002/wat2.1260

Lintern, A., Webb, J.A., Ryu, D., Liu, S., Waters, D., Leahy, P., Bende-Michl, U., Western, A.W., 2018b. What Are the Key Catchment Characteristics Affecting Spatial Differences in Riverine Water Quality? Water Resour. Res. 54, 7252–7272. https://doi.org/10.1029/2017WR022172

Low, K.H., Koki, I.B., Juahir, H., Azid, A., Behkami, S., Ikram, R., Mohammed, H.A., Zain, S.M., 2016. Evaluation of water quality variation in lakes, rivers, and ex-mining ponds in Malaysia (review). Desalination Water Treat. 57, 28215–28239. https://doi.org/10.1080/19443994.2016.1185382

Mainali, J., Chang, H., 2020. Putting space into modeling landscape and water quality relationships in the Han River basin, South Korea. Comput. Environ. Urban Syst. 81, 101461. https://doi.org/10.1016/j.compenvurbsys.2020.101461

Mainali, J., Chang, H., 2018. Landscape and anthropogenic factors affecting spatial patterns of water quality trends in a large river basin, South Korea. J. Hydrol. 564, 26–40. https://doi.org/10.1016/j.jhydrol.2018.06.074

Mainali, J., Chang, H., Chun, Y., 2019. A review of spatial statistical approaches to modeling water quality. Prog. Phys. Geogr. Earth Environ. 43, 801–826. https://doi.org/10.1177/0309133319852003

Money, E., Carter, G.P., Serre, M.L., 2009. Using river distances in the space/time estimation of dissolved oxygen along two impaired river networks in New Jersey. Water Res. 43, 1948–1958. https://doi.org/10.1016/j.watres.2009.01.034

Nielsen, A., Trolle, D., Søndergaard, M., Lauridsen, T.L., Bjerring, R., Olesen, J.E., Jeppesen, E., 2012. Watershed land use effects on lake water quality in Denmark. Ecol. Appl. 22, 1187–1200. https://doi.org/10.1890/11-1831.1

Pandey, A.N., Pathak, P.C., Singh, J.S., 1983. Water, sediment, and nutrient movement in forested and non-forested catchments in Kumaun Himalaya. For. Ecol. Manag. 7, 19–29.

Peterson, E.E., Hoef, J.M.V., 2010. A mixed-model moving-average approach to geostatistical modeling in stream networks. Ecology 91, 644–651.

Pokharel, K.K., Basnet, K.B., Majupuria, T.C., Baniya, C.B., 2018. Environmental Variables of the Seti Gandaki River Basin Pokhara, Nepal. J. Inst. Sci. Technol. 22, 129–139.

Post, C.J., Cope, M.P., Gerard, P.D., Masto, N.M., Vine, J.R., Stiglitz, R.Y., Hallstrom, J.O., Newman, J.C., Mikhailova, E.A., 2018. Monitoring spatial and temporal variation of dissolved oxygen and water temperature in the Savannah River using a sensor network. Environ. Monit. Assess. 190. https://doi.org/10.1007/s10661-018-6646-y

Pratt, B., Chang, H., 2012. Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. J. Hazard. Mater. 209–210, 48–58. https://doi.org/10.1016/j.jhazmat.2011.12.068

R Core Team, 2019. The R Project for Statistical Computing [WWW Document]. URL https://www.r-project.org/ (accessed 11.10.19).

Rimal, B., Baral, H., Stork, N., Paudyal, K., Rijal, S., 2015. Growing City and Rapid Land Use Transition: Assessing Multiple Hazards and Risks in the Pokhara Valley, Nepal. Land 4, 957–978. https://doi.org/10.3390/land4040957

S. Hammoud, A., Leung, J., Tripathi, S., P. Butler, A., N. Sule, M., R. Templeton, M., ^{1} Department of Civil and Environmental Engineering, Imperial College London, London, United Kingdom SW7 2AZ, UK, ^{2} Nepal Engineering College, Bhaktapur, Nepal, 2018. The impact of latrine contents and emptying practices on nitrogen contamination of well water in Kathmandu Valley, Nepal. AIMS Environ. Sci. 5, 143–153. https://doi.org/10.3934/environsci.2018.3.143

Schindler, D.E., Jankowski, K., A'mar, Z.T., Holtgrieve, G.W., 2017. Two-stage metabolism inferred from diel oxygen dynamics in aquatic ecosystems. Ecosphere 8, e01867. https://doi.org/10.1002/ecs2.1867

Sharma, S., Bajracharya, R.M., Sitaula, B.K., Merz, juerg, 2005. Water Quality in the Central Himalaya. Curr. Sci. 89, 774–786.

Sliva, L., Williams, D.D., 2001. Buffer zone versus whole catchment approaches to studying land use impact on river water quality. Water Res. 35, 3462–3472.

Sreebha, S., Padmalal, D., 2011. Environmental Impact Assessment of Sand Mining from the Small Catchment Rivers in the Southwestern Coast of India: A Case Study. Environ. Manage. 47, 130–140. https://doi.org/10.1007/s00267-010-9571-6

Su, S., Xiao, R., Xu, X., Zhang, Z., Mi, X., Wu, J., 2013. Multi-scale spatial determinants of dissolved oxygen and nutrients in Qiantang River, China. Reg. Environ. Change 13, 77–89. https://doi.org/10.1007/s10113-012-0313-6

Su, S., Xiao, R., Zhang, Y., 2012. Multi-scale analysis of spatially varying relationships between agricultural landscape patterns and urbanization using geographically weighted regression. Appl. Geogr. 32, 360–375. https://doi.org/10.1016/j.apgeog.2011.06.005

Thapa, G.J., Wikramanayake, E., Forrest, J., 2015. Climate-change impacts on the biodiversity of the Terai Arc Landscape and the Chitwan-Annapurna Landscape. Hariyo Ban, WWF Nepal, Kathmandu, Nepal.

Tiefelsdorf, M., Griffith, D.A., 2007. Semiparametric Filtering of Spatial Autocorrelation: The Eigenvector Approach. Environ. Plan. A 39, 1193–1221. https://doi.org/10.1068/a37378

Turner, R.E., Rabalais, N.N., 2003. Linking Landscape and Water Quality in the Mississippi River Basin for 200 Years. BioScience 53, 563. https://doi.org/10.1641/0006-3568(2003)053[0563:LLAWQI]2.0.CO;2

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

US EPA, O., 2013. Indicators: Conductivity [WWW Document]. US EPA. URL https://www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity (accessed 2.17.20).

USGS DO, 2006. Chapter A6. Section 6.2. Dissolved Oxygen. https://doi.org/10.3133/twri09A6.2

Vaidya, S.R., Labh, S.N., 2017. Determination of Physico-Chemical Parameters and Water Quality Index (WQI) for drinking water available in Kathmandu Valley, Nepal: A review.

Ver Hoef, J.M., Peterson, E., Theobald, D., 2006. Spatial statistical models that use flow and stream distance. Environ. Ecol. Stat. 13, 449–464. https://doi.org/10.1007/s10651-006-0022-8

Washington Ecology, 2002. Dissolved Oxygen and the Water Quality Standards.

Water on the Web, 2020. Water on the Web | Understanding | Water Quality | Parameters | EC [WWW Document]. URL https://www.waterontheweb.org/under/waterquality/conductivity.html (accessed 2.23.20).

Water Quality Standard Nepal, 2005. National Drinking Water Quality Standards, 2005.

Wenner, D.B., Ruhlman, M., Eggert, S., 2003. THE IMPORTANCE OF SPECIFIC CONDUCTIVITY FOR ASSESSING ENVIRONMENTALLY IMPACTED STREAMS. Presented at the Proceedings of the 2003 Georgia Water Resources Conference, p. 3.

WorldPop Nepal, 2015. Worldpop - Nepal [WWW Document]. URL http://www.worldpop.org.uk/nepal/ (accessed 9.20.18).

Yadav, S., Babel, M.S., Shrestha, S., Deb, P., 2019. Land use impact on the water quality of large tropical river: Mun River Basin, Thailand. Environ. Monit. Assess. 191. https://doi.org/10.1007/s10661-019-7779-3

Yan, X., Ai, T., Yang, M., Yin, H., 2019. A graph convolutional neural network for classification of building patterns using spatial vector data. ISPRS J. Photogramm. Remote Sens. 150, 259–273. https://doi.org/10.1016/j.isprsjprs.2019.02.010

Zampella, R.A., Procopio, N.A., Lathrop, R.G., Dow, C.L., 2007. Relationship of Land-Use/Land-Cover Patterns and Surface-Water Quality in The Mullica River Basin. J. Am. Water Resour. Assoc. 43, 594–604. https://doi.org/10.1111/j.1752-1688.2007.00045.x

Zang, C., Huang, S., Wu, M., Du, S., Scholz, M., Gao, F., Lin, C., Guo, Y., Dong, Y., 2011. Comparison of Relationships Between pH, Dissolved Oxygen and Chlorophyll a for Aquaculture and Non-aquaculture Waters. Water. Air. Soil Pollut. 219, 157–174. https://doi.org/10.1007/s11270-010-0695-3

Zhou, T., Wu, J., Peng, S., 2012. Assessing the effects of landscape pattern on river water quality at multiple scales: A case study of the Dongjiang River watershed, China. Ecol. Indic. 23, 166–175. https://doi.org/10.1016/j.ecolind.2012.03.013

Chapter Three: River Runs Downstream: Modified Geographically Weighted Regression for Stream Networks

Janardan Mainali and Heejun Chang

River Runs Downstream: Modified Geographically Weighted Regression for Stream Networks

Janardan Mainali and Heejun Chang

Abstract

The geographically weighted regression (GWR) models, which allows us to explore local variations in relationships between different factors, have been widely used to examine water quality and their relationships with the watershed structure and processes. The GWR models used in surface water quality studies to date have not paid attention to the network structure and upstream-downstream directionality of rivers and streams. We incorporate upstream-distance metrics into GWR (U-GWR) models and compare the outputs with standard GWR (S-GWR). We use Dissolved Oxygen and Conductivity data from a river and its tributaries in a mountainous watershed of central Nepal as response variables. Land use types, elevation, slope, and population density extracted at 100-m buffer and watershed-scale were used as explanatory variables. The spatial stream network-based tools were used to derive the stream network and calculate the upstream distance for each site. We compared the regression model outputs between S-GWR and U-GWR. A successful model could be developed using U-GWR having comparable model strength with that of S-GWR. The resultant model revealed different spatial patterns of model strength (R^2) as well as the relationship with explanatory variables. The U-GWR model can offer better insights into hydrological and biogeochemical relationships among different water quality measurement sites and their connections with watershed processes. These insights not only help understanding fine-scale impacts of

socio-environmental and biophysical factors to water quality but also assist in designing locally specific water quality management endeavors.

1. Background

The linear regression models like ordinary least square regression (OLS) are mostly used to elucidate the impacts of various social, environmental, and climatic factors on the surface water quality (Ullah et al., 2018). The OLS models do not account for the spatial structure of the model and are usually invalid when there is spatial autocorrelation of the residuals. The spatial modeling approach, such as Spatial lag and error model (Anselin, 2001), Spatial Filtering (Tiefelsdorf and Griffith, 2007), and Spatial Kriging (Cressie, 1988.), attempt to take into account the spatial patterns of water quality parameter being studied, help identify the watershed characteristics that impact water quality conditions, derive the spatial correlation structure among the observations, and predict water quality at unmonitored locations. (Yang and Jin, 2010, Chang 2008, Mainali et al. 2019). These regression models, however, are global as they produce only one model summary for the entire set of data. They are not useful when there is a high spatial variation in the relationships between explanatory variables and response variables. Among stream sections, the relationships between predictors and a response variable can vary, and the strength of those relations might also be dSifferent across regions (Ganio et al., 2005).

In order to address this issue, a widely used spatial regression model, geographically weighted regression (GWR) can be used to explore varying local relationships between predictor and response variables across different sites. GWR can be used to allow model coefficients to vary for each observation and create a set of local models based on the location of sampling sites (Brunsdon et al., 1998). The observed data included in each local model are geographically weighted with neighboring data points, depending on the proximity of the location, and are used to estimate local R^2 and coefficients for each sample observation. The number of samples included for each data point is defined using a bandwidth function. GWR is increasingly used in water-quality modeling, not only to estimate the model parameters but also to explore the variabilities of those relationships in different watersheds (Chang and Psaris, 2013; Chen et al., 2016a; Pratt and Chang, 2012; Tu, 2011; Tu and Xia, 2008). Most of these works report higher model strength of GWR over OLS. Scholars assert that the local models hence developed can facilitate site-specific water pollution mitigation efforts by accounting for local variations in pollution source, land use, and other relevant factors (An et al., 2016; Chen et al., 2016; Tu and Xia, 2008). These models, however, use the Euclidean distance approach where sites of all directions are used to derive the local models (Mainali et al. 2019). The up-stream-down-stream relations are not yet incorporated in water quality modeling using GWR.

Rivers and streams seldom behave in linear fashion (as crows fly); rather, they are better represented as a dendritic network with stems and branches (Peterson et al., 2013). The transport of energy, nutrients, sediments, and biological components like fish occur along those networks within a terrestrial landscape (Isaak et al., 2014; Ver Hoef et al., 2006). On the other hand, there is a continuum along surface water and land where water that falls on the land surface eventually ends up in surface water bodies (Vannote et al., 1980). This continuum, to some extent, is responsible for the physical, chemical, and ecological characteristics of surface water. These movements of water usually wash various non-water components from landscape to water. After these components are delivered on the river system, they undergo various chemical or physical changes like increased concentration over time, dilution due to increased water flow, and movement of those compounds downstream (e.g., Lintern et al., 2018). For spatial statistical process, the stream distance maybe a more appropriate distance metric when modeling spatial properties of the various stream and river attributes. Stream distance is defined as the shortest distance between two locations, where distance is computed only along the stream network (Ver Hoef et al., 2006). In this work, we attempt to understand the local variations of downstream movements by leveraging two major developments in spatial statistics. Building on the development of spatial stream network models and GWR, this research attempts to modify later to incorporate the unique network structure of the stream network in developing the local models for stream and river networks (Figure 3-1).

The major challenge of this approach is identifying appropriate statistical and methodological tools to define up-stream downstream relationships. Determining upstream and downstream linkage would involve calculating the distance matrix of the sampling stations based on their upstream and downstream relations and using that within a GWR framework. A recently developed spatial stream network statistical methods (SSN) can potentially be used to provide a framework to define up-stream downstream relation and derive the flow-connected distance among different sampling sites (Peterson et al., 2013; Ver Hoef and Peterson, 2010). The GIS processing toolbox STARS can be used to set up a stream network, define upstream-downstream relationships among sampling sites, and calculate the distance matrix based on those relations (Peterson and Ver Hoef, 2014).

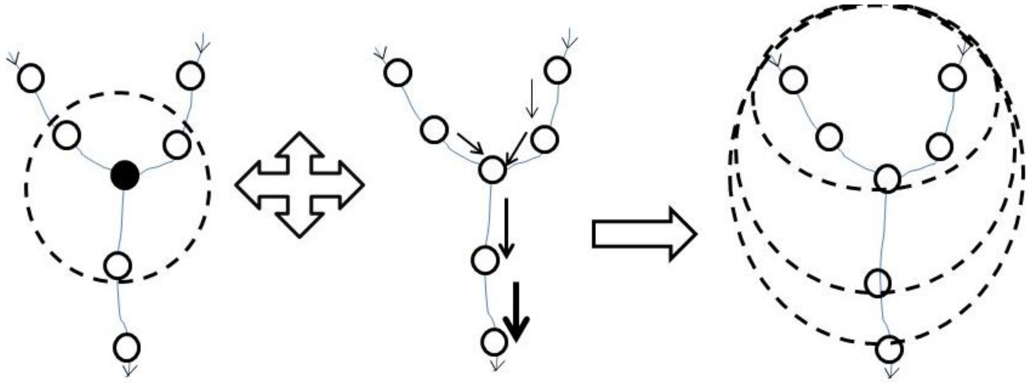


Figure 3-1: Spatial conceptualization of the project. a) A standard geographically weighted regression where a circular band is created around the site to derive the distance weight b) Upstream-downstream relations in spatial stream network models. The weights are provided based on the cumulative upstream distance or other relevant parameters like flow volume or watershed size. c) A modified geographic weighted regression where distance matrix for each site is calculated using bandwidth defined only towards upstream sites

Objectives

A general objective of this work is to compare the model outputs between standard geographically weighted regression (S-GWR) and upstream distance weighted geographically weighted regression (U-GWR). We hypothesize that the GWR models developed from upstream-distance weighted regression are stronger (having higher R^2 and lower AIC) and can capture local variability better.

2. Methods

2.1. Model Data

We use stream polyline shapefile, water quality data collected in the field and their corresponding coordinates, and raster-based explanatory variables related to land use, topography, and population density (Table 3-1). The water quality response variables are Dissolved Oxygen (DO) and Conductivity (COND). Data collection, processing, and GIS analysis processes are described in Mainali and Chang 2020, under prep).

Table 3-1: Different data types used in this work

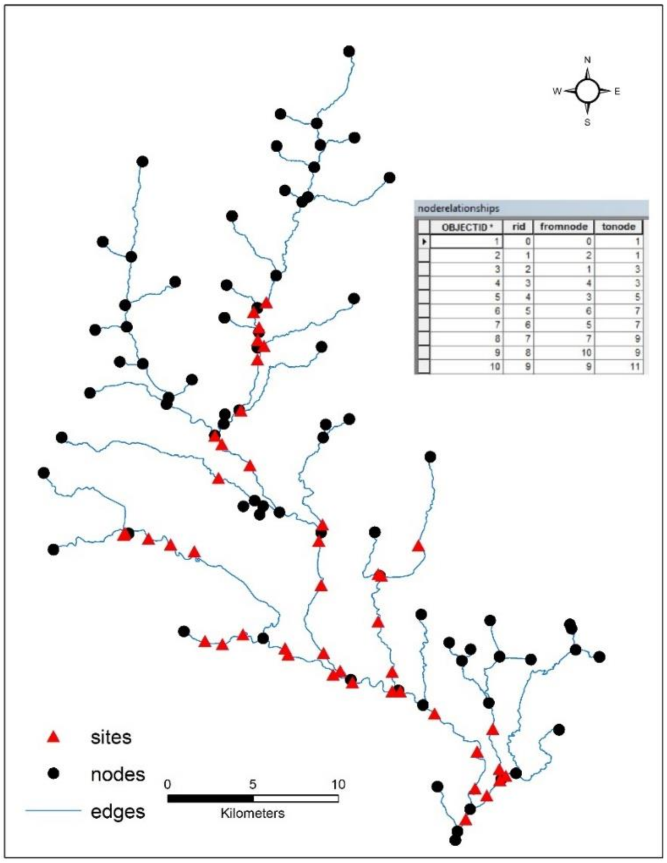
Data Name	Type	Resolution	Source
Stream Polyline Shapefile	Line Feature		Department of Survey, Nepal, 1986
Water Quality	Point	Point data	Field sampling, 2019
Land Use Types	Raster	30-meter raster	Classified from Landsat 8, 2017
Elevation and Slope	Contour layer converted to raster	30-meter raster	Department of Survey, Nepal, 1986
Population	Raster	100 m, resampled to 30 m	WorldPop Nepal, 2015

2.2. Stream Network

The first step of this work was to create topologically correct stream networks. The polyline shapefiles of the stream networks are used as the stream network. We used the Spatial Tools For The Analysis of River Systems (STARS) tool version 2.0.7 to create and analyze the stream network data (Peterson and Hoef, 2014). The STARS tool requires a carefully digitized stream network to preserve upstream to downstream direction of the river. Using the stream network shapefile and the imagery of the study area, we re-digitized a stream network for the entire watershed using ArcMap version 10.7.1 (ESRI, 2020). To make a topologically correct stream network, we digitized stream-network upstream to downstream with separate stream reaches.

2.3. *Landscape Network and Distance Matrices*

The digitized stream network was converted to a landscape network database using Polyline to Landscape Network tool available in the STARS toolset. The landscape network database consists of edges, nodes, and the relationship tables between them (Figure 3-2). The site's points were then snapped along the stream network using Snap Points to Landscape Network tool available in the STARS toolset. After the points were snapped, the upstream distance was calculated for each site using the Calculate Upstream Distance Among Sites tool. The resultant landscape matrix with upstream distance measurement for our sites was converted to a spatial stream network object. We used Create SSN Object available in the same toolset to export the resultant stream network to R software (R Project, 2020).

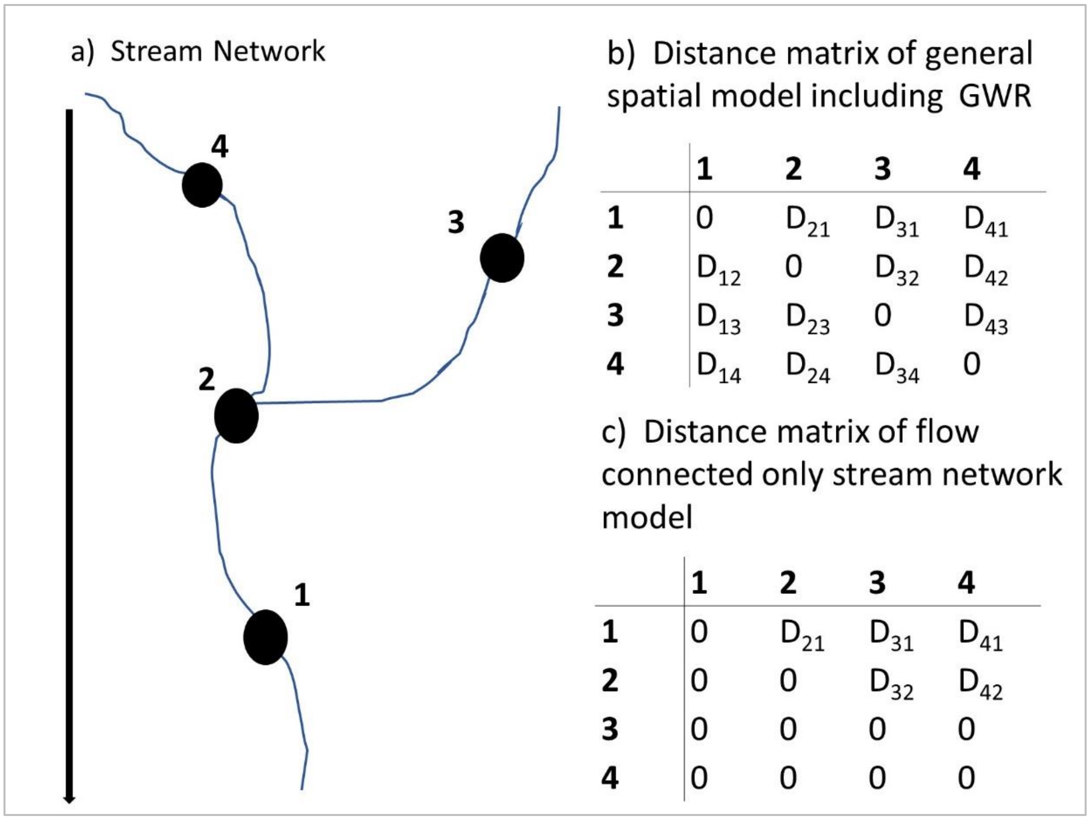


noderelationships

OBJECTID *	rid	fromnode	tonode
1	0	0	1
2	1	2	1
3	2	1	3
4	3	4	3
5	4	3	5
6	5	6	7
7	6	5	7
8	7	7	9
9	8	10	9
10	9	9	11

Figure 3-2: Landscape network with the relationship table on the inset.

After the spatial stream network object was imported to R software (version 3.6.1), we used SSN package to extract the upstream distance matrix and derive flow connected and flow unconnected semivariogram (torgegram) of our response variables DO and COND (Ver Hoef and Peterson, 2020). The distance matrices use the upstream distance calculated in the GIS environment to derive the upstream-distance matrix. The standard distance matrix has 0 on the diagonal region while having a specific distance value on the non-diagonal region (Figure 3-3). The upstream distance matrix is different as it also provides a value of 0 to the sites not connected by the flow. Using the distance matrix created, we ran torgegram for dissolved oxygen and conductivity.



b) Distance matrix of general spatial model including GWR

	1	2	3	4
1	0	D_{21}	D_{31}	D_{41}
2	D_{12}	0	D_{32}	D_{42}
3	D_{13}	D_{23}	0	D_{43}
4	D_{14}	D_{24}	D_{34}	0

c) Distance matrix of flow connected only stream network model

	1	2	3	4
1	0	D_{21}	D_{31}	D_{41}
2	0	0	D_{32}	D_{42}
3	0	0	0	0
4	0	0	0	0

Figure 3-3: Different distance matrices depicting general spatial models and spatial stream based spatial model. Notice that flow unconnected sites have 0 distance values.

2.4. Different Geographically Weighted Regression

We use R package GWmodel (Lu et al., 2019) to run S-GWR and U-GWR models. The first step of this analysis involved creating a standard distance matrix using a distance matrix function available in the package. We ran regression models for two dependent variables—DO and COND. The models are run on two different scales- watershed scale and buffer scale. The buffer scale explanatory variables capture impacts of landscape variables of the vicinity, while the watershed scale captures the entire upstream region from a site. There are, hence, four different linear models. These standard models are then compared with U-GWR.

We chose the set of explanatory variables with the lowest AICc (Akaike Information Criteria) and variation inflation factor (VIF) for each response variable and a set of explanatory variables. We used *model.selection.gwr* function available in the GWmodel package to select the best subset of the explanatory variables. That set of explanatory variables was used to run three different types of regressions- standard linear model without any spatial considerations (OLS), standard geographic weighted regression using the Euclidean distance matrix (S-GWR), and geographically weighted regression using the upstream distance matrix (U-GWR). We used *robust.gwr* function to run the GWR to come up with all relevant model outputs like local R^2, model coefficients, and significance. The model outputs from OLS, S-GWR, and U-GWR were compared based on the R^2 and AICc values. We also compared spatial patterns of R^2 and model coefficients of selected significant explanatory variables between S-GWR, and U-GWR.

3. Results

3.1. *Spatial Autocorrelation along the Network*

As shown in Figure 3-4, the spatial stream network model approach successfully derived upstream distance. The upstream distance was calculated for each stream reaches (edges), and sampling sites. The distance matrix derived from the upstream distance allows us to calculate flow-connected and flow-unconnected distance matrices where flow-unconnected reaches were excluded for creating local spatial weights.

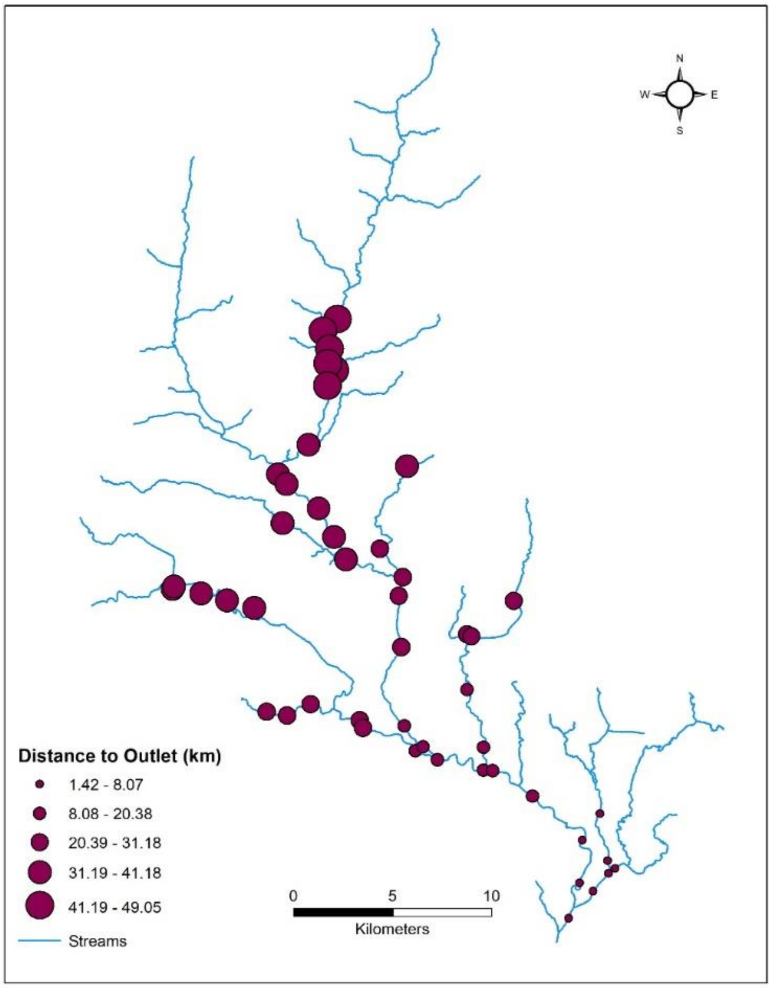


Figure 3-4: Upstream distance to the sites relative to watershed outlet defined in this study

For DO concentration, spatial autocorrelation is the highest around 15000 meters in the flow connected model, while it is around 2800 meters in the flow unconnected model, signifying the greater clustering along the stream network (Figure 3-5). Both flow connected and flow unconnected COND autocorrelations have bimodal distribution with one peak at around 10000 meters and another around 30000 meters (Figure 3-6). As we can see from the spatial distribution of COND (Chapter 2), there are distinct pockets of conductivity in different parts of the watershed with lowest in western tributary, medium ranges in eastern parts, and high values along the main stem of the river.

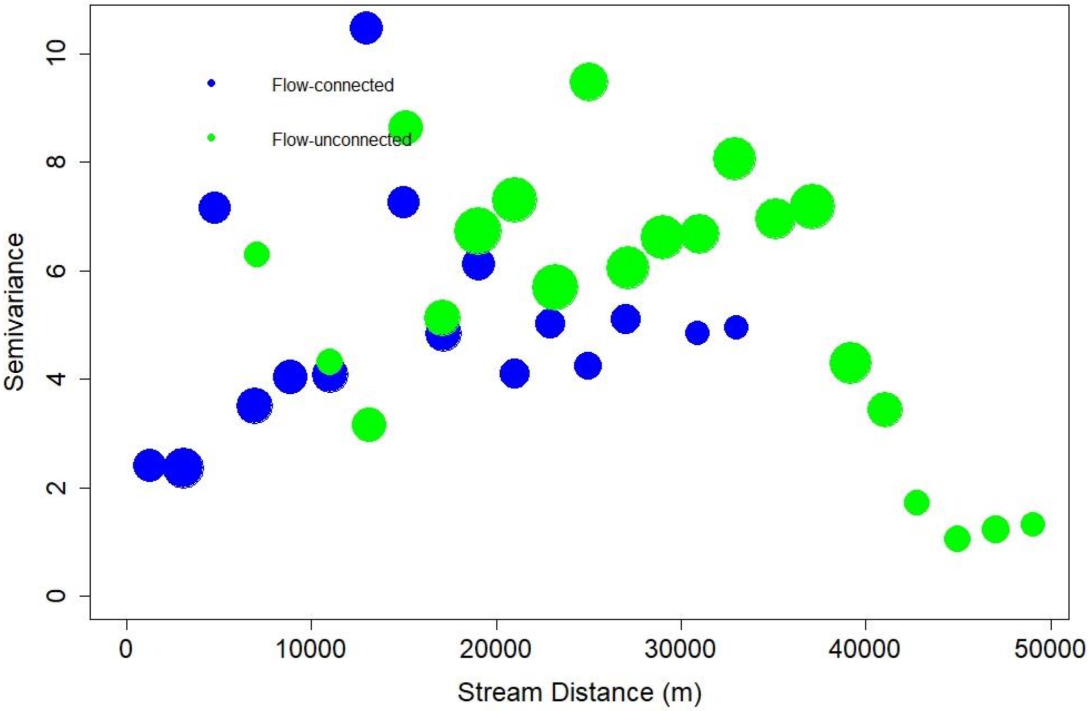


Figure 3-5: Spatial autocorrelation at different distance for Dissolved Oxygen

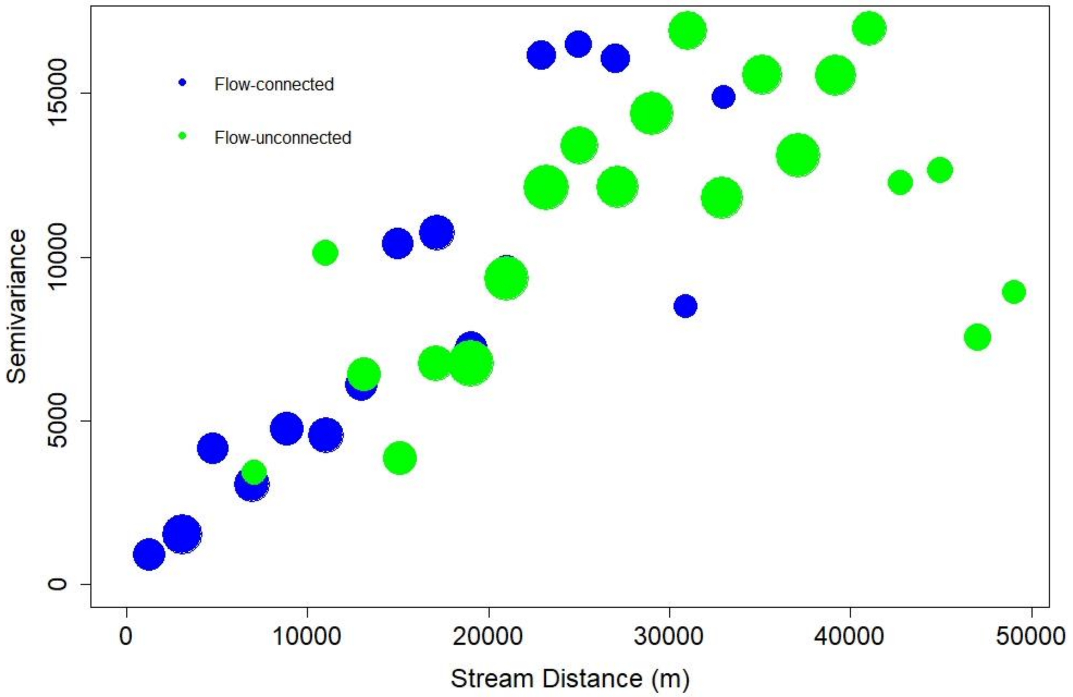


Figure 3-6: Spatial autocorrelation at different distances for Conductivity.

3.2. Comparison of different regression models

Table 3-2: Comparison of different regression models at different scales for conductivity. Notice that different sets of predictor variables were selected during model selection on the watershed and buffer scale model. Bold values are significant at p<0.05

Model Parameters	**Buffer Scale**			**Watershed Scale**		
	OLS	**S-GWR**	**U-GWR**	**OLS**	**S-GWR**	**U-GWR**
R^2	0.68	0.78	0.77	0.26	0.57	0.64
AIC	575.33	534.07	538	575.33	562.95	553.28
Intercept	219.73	208.22	175.76	552.45	512.15	474.77
Elevation	-	-	-	**-0.08**	-0.06	-0.06
Elevation Standard Deviation	**0.16**	0.18	0.05	**0.19**	0.13	0.14

Slope Standard Deviation	-	-	-	-9.58	**-16.37**	**-5.65**
Forest	-	-	-	**-2.74**	-1.93	**-2.6**
Sand	**-12.38**	-3.79	1.32	-22.56	-18.47	-14.79
Bare	2.54	-5.96	-1.32	-	-	-
Average Population	-	-	-	2.55	**2.17**	**1.82**
Population Standard Deviation	**5.43**	7.42	6.18	-	-	-

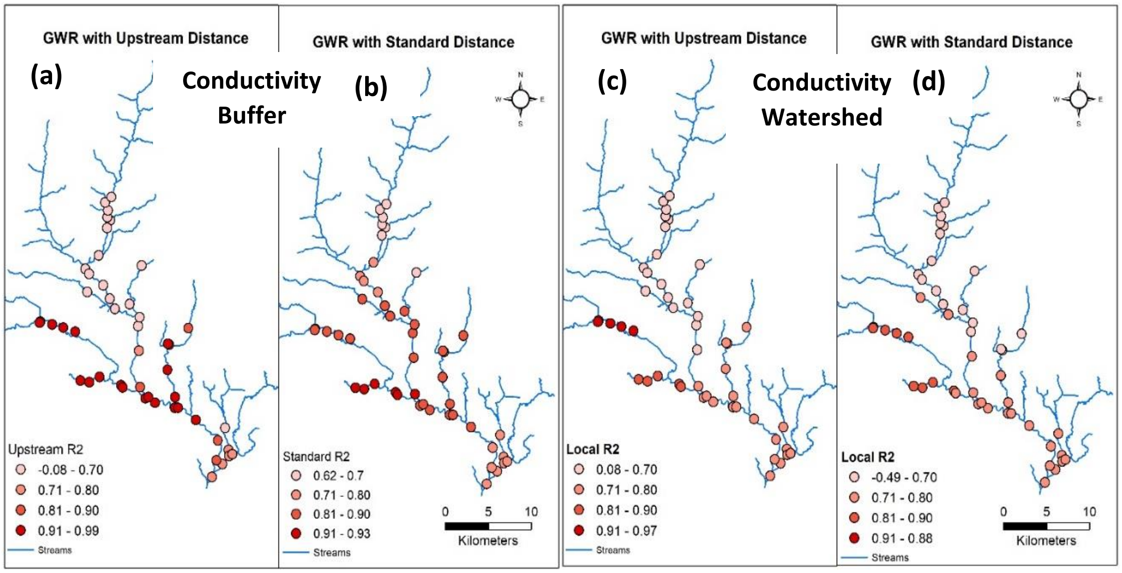


Figure 3-7: Spatial patterns of R2 values between S-GWR and U-GWR models at buffer and watershed scales. U-GWR COND model at buffer scale, b) S-GWR COND model at buffer scale, c) U-GWR COND model at a watershed scale, and d) S-GWR COND model at the watershed scale.

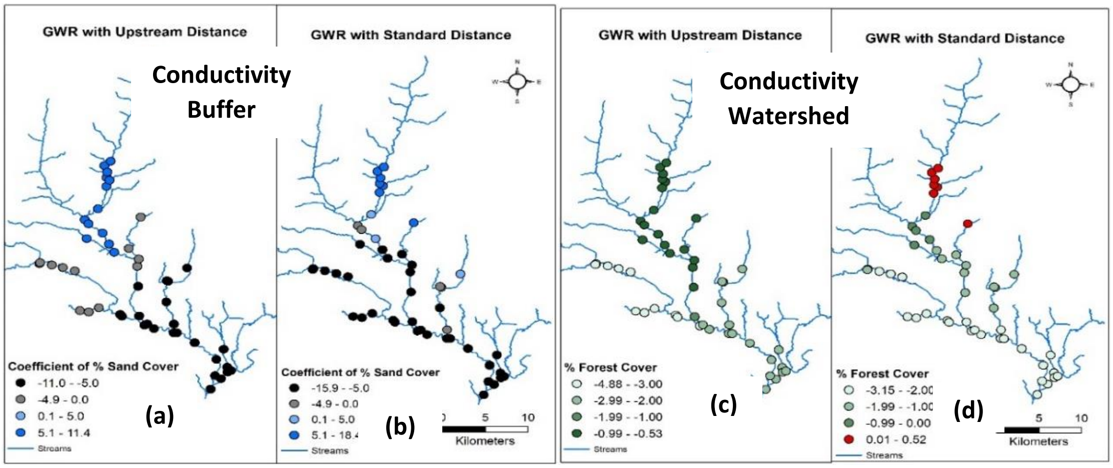


Figure 3-8: Spatial patterns of the coefficient of different explanatory variables related to S-GWR and U-GWR models at buffer and watershed scales. a) The spatial pattern of coefficients of percentage sand cover with U-GWR with buffer scale COND model b) Spatial pattern of coefficients of percentage sand cover with S-GWR COND model at buffer scale c) Spatial patterns of coefficients of percentage forest cover with U-GWR with watershed-scale COND model, and d) Spatial patterns of coefficients of percentage forest cover with S-GWR with watershed-scale COND model

3.2.1 **COND Model Details**

Overall, the model strength of the buffer scale model of conductivity is the highest with the S-GWR, although U-GWR also exhibited only slightly lower R^2 (Table 3-2). The AICc is also lowest with S-GWR, exhibiting better model output. At the buffer scale, elevation standard deviation and percentage sand are the significant explanatory variables, while none of the other variables exhibited overall significance in both GWR models. At the watershed scale, the highest R^2 and lowest AIC is with the U-GWR model. Average elevation, elevation standard deviation, and percentage forest cover are significant explanatory variables at the non-spatial linear model, while slope standard deviation and average population with S-GWR, and slope standard deviation, percentage forest, and average population in U-GWR.

Spatial patterns of R^2 and Coefficients

The strength and uniqueness of GWR lie in its capability to produce a model for each observation. The spatial variation in these local modeling attributes might provide us with the behavior of these water quality parameters at different locations in response to the explanatory variables under consideration. For the buffer scale model of the conductivity, the U-GWR R^2 values are significantly lower in upstream regions while those are higher in downstream. These relations are expected as the weight is higher downstream. The spatial pattern of R^2 values (Figure 3-7a) shows that at U-GWR model strength is generally weaker in the mid and upstream region and significantly stronger in the lower middle region of the watershed. Some of the up-stream sites showed very low R^2 values as low as 0.08, while it could be as high as 0.99 in some other sites in U-GWR (Figure 3-7a). The spatial pattern of R^2 with S-GWR, however, is less pronounced with R^2 values ranging from 0.62 to 0.93 (Figure 3-7b). There are still higher model strengths with the sites in the middle part of the watershed than those in the upstream or downstream region.

At the watershed scale, the general pattern of R^2 is similar between S-GWR and U-GWR with R^2 values ranging from 0.08 to 0.97 at U-GWR and -0.49 to 0.98 with S-GWR. The overall patterns of R^2 values are slightly opposite than the buffer scale model where the lower range of the value is associated with S-GWR. The spatial patterns of R^2 show that the watershed scale conductivity model exhibits higher model strength on the west side of the watershed, moderate strength on the downstream, and weaker strength on the upstream region (Figure 3-7c and 3-7d).

The spatial pattern of the coefficient in percentage sand cover is similar between U-GWR and S-GWR associated with the buffer scale COND model (Figure 3-8a and 3-8b). While the percentage of sand cover is negatively associated with conductivity in the downstream regions, which is opposite in the upstream region and is similarly distributed in both S-GWR and U-GWR. In the OLS and S-GWR, the forest cover in the watershed generally impacts water conductivity positively. When the standard distance matrix was used, the relationship changes to positive in the remote forested regions of the watershed (Figure 8c and 8d). However, the U-GWR did not result in any positive coefficients (3-8c).

Table 3-3: Comparison of different dissolved oxygen regression models. Notice that different sets of predictor variables were selected during model selection on watershed and buffer scale models. Bold values are significant at p<0.05

	Buffer Scale			Watershed Scale		
	OLS	**S-GWR**	**U-GWR**	**OLS**	**S-GWR**	**U-GWR**
R2	0.21	0.3	0.32	0.38	0.51	0.48
AIC	128.23	128.23	127.57	115.96	109.87	113
Intercept	4.9	3.85	4.96	**9.72**	**10.28**	10.26
Elevation	0.001	0.0017	**0.00001**	-	-	-
Elevation Standard Deviation	-	-	-	**0.001**	**0.001**	**0.001**
Slope Standard Deviation	0.01	0.07	0.09	**-0.18**	**-0.24**	-0.2

Forest	0.005	0.004	0.021	-	-	-
Agriculture and Settlements	0.002	0.01	0.001	-	-	-
Sand	-	-	-	**-0.34**	-0.4	-0.38
Average Population	0.09	0.1	0.11	0.006	**0.005**	0.003

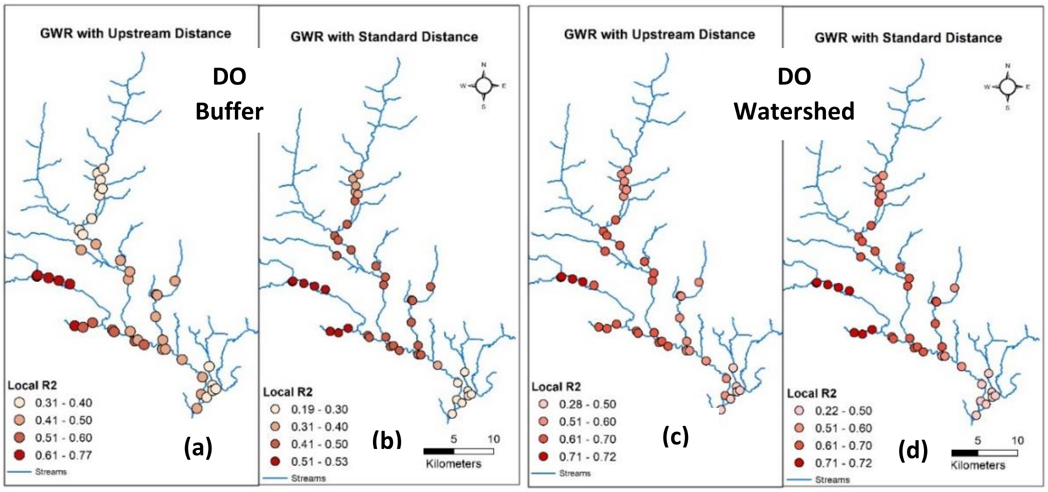


Figure 3-9: Spatial patterns of R2 values between S-GWR and U-GWR at watershed and buffer scale. U-GWR buffer scale DO model, b) S-GWR buffer scale DO model, c) U-GWR watershed-scale DO model, and d) S-GWR watershed-scale DO model

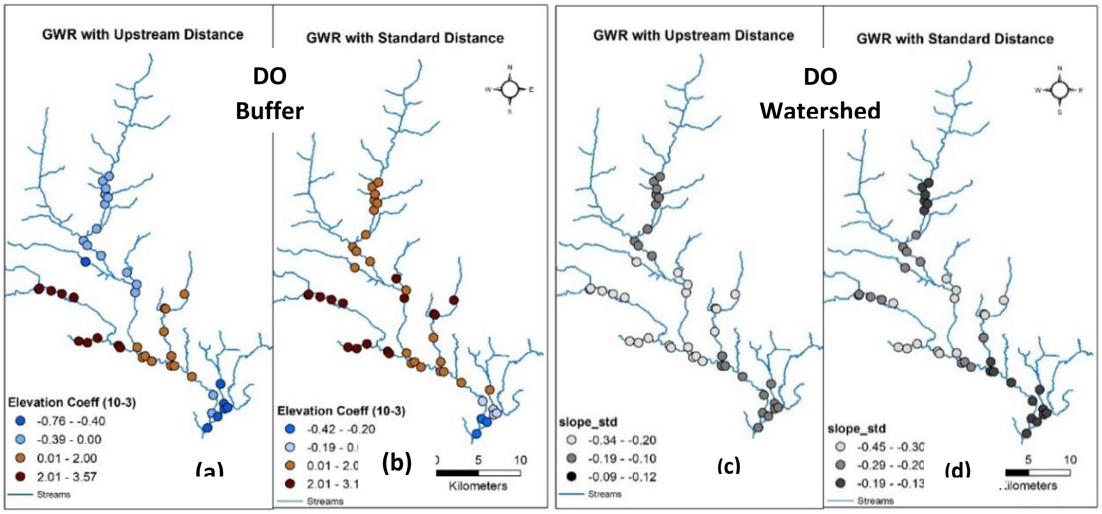


Figure 3-10: Spatial patterns of the coefficient of different predictor variables in standard and upstream distance weighted GWR models of DO. a) The spatial pattern of coefficients of average elevation with U- GWR with buffer level DO model b) Spatial pattern of coefficients of average elevation with S-GWR with buffer level DO model c) Spatial pattern of slope standard deviation with U-GWR with watershed level DO model, and d) Spatial pattern of slope standard deviation with S-GWR with watershed level DO model

3.2.2. Dissolved Oxygen

General Model Characteristic

The overall model strength of buffer scale DO is the highest in the U-GWR model with the highest R^2 and lowest AIC (Table 3-3). At the buffer scale, none of the parameters showed any significance with the OLS and S-GWR. Average elevation was a significant predictor in the U-GWR model. At the watershed scale, model strength is the highest in the S-GWR model with the highest R^2 and the lowest AIC. The elevation standard deviation, slope standard deviation, and percentage sand cover are significant variables with OLS, while elevation standard deviation, slope standard deviation, and average population density are significant on average with S-GWR. On average, the elevation standard deviation is significant with U-GWR.

Spatial Patterns of R^2 and coefficients

At the buffer scale, the spatial patterns of R^2 are similar between S-GWR and U-GWR (Figure 3-9). However, the range of R^2 values is lower in S-GWR with R^2 ranging from 0.19 to 0.53, while that of upstream-weighted R^2 values range from 0.31 to 0.77 (Figure 3-9a and 3-9b). The spatial pattern reveals that the R^2 values of both the upstream region and the downstream region are lower than the middle and west regions of the watershed with U-GWR. But, the spatial pattern of R^2 values of S-GWR shows that the R^2 values are lower in the downstream and adjacent region, while there is not much spatial variability in the rest of the watershed. At the watershed scale, both range and spatial patterns of R^2 are similar between S-GWR and U-GWR with R^2 ranging from about 0.2 to 0.72 (Figure 3-9c and 3-9d).

Elevation can be an important determinant of water quality, especially because of its impact on water temperature. This watershed also has a sharp elevation gradient with 700 to 7500 meters from sea level. The spatial patterns of the impact of elevation in DO concentration are significantly different between S-GWR and U-GWR (Figure 3-10a and 3-10b). There are generally positive impacts of buffer scale elevation in DO in the S-GWR except for a few downstream sites. These downstream sites are in the lowest elevation region of the watershed and are immediately downstream of the major urban region. The U-GWR, however, showed different spatial patterns in the majority of the upstream sites showing negative relationships with elevation. There are also a few downstream stations that showed a negative relation with DO. The negative coefficient in high elevation suggests that the DO concentration is lower in the high elevation region according to U-GWR. The slope standard deviation is a measure of the ruggedness of the landscape. The spatial patterns of the coefficients of the watershed-scale DO model are similar between S-GWR and U-GWR. Our result shows that the impact of ruggedness in DO is negative across all the regions of the watershed with strong impacts reported in upstream and downstream regions than the middle regions (Figure 3-10c and 3-10d).

4. **Discussions**

4.1. *Spatial Autocorrelation Along a Network*

The variograms that measure the variance of data in different distance intervals are used to answer several spatial problems like the size and magnitude of spatial clusters. But their use in a stream environment might be ineffectual as the spatial patterns are not governed by the straight line distance but rather by distance along the stream (Ganio et al., 2005). The variograms that are constructed based on the stream distance, therefore, provide more information about the upstream-downstream relationships of the parameters being studied and can help detect patterns along with the network independent of spatial relationships over land.

Our result also shows that the distance at which spatial autocorrelation occurs is lower for flow connected distance than unconnected distance, signifying a greater clustering pattern for DO along with the stream network. This finding is in line with our previous work (Mainali and Chang 2020, under prep) where stream-like graph type yielded stronger model performance than other standard multidirectional graph types. Other studies that examined DO along the stream networks have suggested that the DO concentration is dependent on the upstream factors like solute concentration, therefore, are modeled better when upstream-downstream relations are incorporated (Money et al., 2009). The distance at which the highest spatial autocorrelation occurred has two different peaks within both flows connected and flow unconnected Torgegram of electric conductivity. It suggests that there are various clusters of conductivity in this watershed. In our previous work (Mainali and Chang, under prep) we report that there are clusters of high to low values in different parts of the watershed with mainstem showing the highest conductivity. The two different bumps in the Torgegram might be a result of such clusters. Mainali and Chang, in prep, conclude that the local clustering of conductivity could be better captured by k-mean clustering than other graph types. Other previous studies also reported that the electric conductivity of the river is influenced by neighbors in all directions or upstream values (Lintern et al., 2018b; Peterson and Hoef, 2010).

These various levels of autocorrelation along the stream network lead to heterogeneity of the model along the stream network (Harris, 2019). This heterogeneity warrants the spatially explicit local models to account for differential relationships between stream attribute of the stream segments and various factors affecting them.

4.2. *Upstream Distance Weighted GWR*

In this work, we show that the successful local models could be developed for surface water quality parameters by incorporating stream network structures within the GWR framework. The model strengths from U-GWR yielded comparable model output as of S-GWR. The spatial patterns of model strength, as well as various coefficients from U-GWR, are crucial in elucidating local patterns as a product of upstream-downstream relation, which mimics the hydrological processes more closely than S-GWR. Although most previous studies have used straight line distance, some researchers acknowledge that the straight line based distance metrics might not reflect true spatial proximity for various social and environmental phenomena like roads and rivers (Lu et al., 2014). There are a few works that have attempted to incorporate the network structure in the GWR model. Lu et al. (2014) used network distance and travel time matrices within the road network to carry out the GWR. They also attempted to use a different set of distance matrices for different explanatory variables in their subsequent paper (Lu et al., 2017), reporting that the travel time-based distance matrix worked best to model house price data.

While network-based distance matrix generally improves the model performance of a geostatistical model, many researchers have argued that it is not always a panacea. Comber et al. (2020), for example, argue that in addition to road-network distance, modelers also need to take into account other factors like the direction of the road, congestion, and traffic lights. Although there has not been any research of local models using stream-network distance, several studies use autoregressive Spatial Stream

Network models and show significant increases in model strengths over traditional OLS models (Neill et al., 2018; Scown et al., 2017). However, other studies did not report any significant changes (Turschwell et al., 2016). We found that, although the model strength of S-GWR and U-GWR are comparable, stream network-based models did not always provide the highest model strength.

While both upstream and downstream regions have lower R^2 values, R^2 values for conductivity with both S-GWR and U-GWR are the highest in the middle section of the watershed. This might be a product of higher clustering of similar values of the response variable as well as the explanatory variables in the middle region of the watershed. We can see a differential impact of distance weighting as we go upstream with the U-GWR; as the upstream distance gets lower; the relative strength of the model seems to dilute. This seems to be an issue with GWR in general while studying a single watershed as the water quality site density becomes lower in the high elevation upstream areas.

Not many studies compared spatial patterns of model coefficient between standard GWR and non-Euclidean distance weighted GWR. Lu et al. (2014) reported that the spatial variation is larger in one of the significant coefficients when using network distance over Euclidean distance. They found that the differences are more widespread while using travel time metrics which they attributed to road network speeds used to calculate the travel time metric. We also noticed a generally different spatial pattern of some of the model coefficients. Values of those coefficients ranged from the different intensity of the same direction (e.g., positive only) between two model types to completely different directions in some sites between two model types. Especially, while comparing model coefficients the physical properties of the relationships are better captured by the S-GWR than the U-GWR. For example, the watershed scale forest impact in conductivity is positive in some upstream sites with S-GWR, while it is always negative in U-GWR. The conductivity values are generally higher along the main stem of the rivers and high elevation regions that are predominantly forested, leading to seemingly positive impacts of forest on conductivity especially on the predominantly forested area.

4.3. *Future Investigations/Limitations*

This work is a demonstration of the use of an upstream distance weighted approach in developing local models for streams and rivers. There area few things which we think could provide us with more insights and help develop stronger models. The GWR is used when there are differences in relationships between response and explanatory variables at different locations of the study sites. That could be explored by local spatial autocorrelation analysis (Harris, 2019; Ord and Getis, 1995). That would provide researchers with a better understanding of the spatial structure of data before running the model.

The current model could be improved by including additional variables. For example, geology data could further explain the spatial differences of conductivity; unfortunately, such data are currently unavailable. Another area of model improvement includes better mimicking the hydrologic features in the watershed. In our study, we simplified our stream network by removing lakes, reservoirs, and any braided structures to make the stream network amenable to STARS. The addition of new variables representing these features would add more meaning to the model and potentially increase the model performance. For example, in the travel time-related model, when additional factors like traffic light and directionality of the roads were incorporated, the models got better (Comber et al., 2020).

The spatial statistics literature takes advantage of data and model simulations. This U-GWR method could also be tested with a simulated network structure and associated datasets. Additionally, this work is an example of one watershed with only two water quality parameters with a finite set of explanatory variables. U-GWR can be tested in other watersheds that have more water quality parameters and landscape variables. Like any modeling, GWR can work better when there are more data points. More data points would increase the bandwidth for an individual site, leading to a more robust and stronger model.

5. Conclusions

We demonstrate that a successful model could be developed by combining a distance matrix derived from spatial stream network models and geographically weighted regression. The upstream distance weighted models provide a comparable model strength that of standard GWR. The spatial patterns of model strength, as well as various coefficients from the upstream distance weighted regressions, are crucial in elucidating local patterns as a product of upstream-downstream relations. The U-GWR model can offer better insights into hydrological and biogeochemical relationships among different sampling sites and their relationships with watershed processes. These insights not only help understanding these processes but also designing locally specific water quality management endeavors. Although the stream network-based models do not always provide the strongest model output, they can provide a better understanding of physical, biological, and hydrological processes occurring between land and water as well as along the upstream-downstream continuum. These local models can always be improved by incorporating more sites, using additional explanatory variables, and accounting for realistic hydrologic features.

6. References

An, K.-J., Lee, S.-W., Hwang, S.-J., Park, S.-R., Hwang, S.-A., 2016. Exploring the Non-Stationary Effects of Forests and Developed Land within Watersheds on Biological Indicators of Streams Using Geographically-Weighted Regression. Water 8, 120. https://doi.org/10.3390/w8040120

Anselin, L., 2001. Spatial Econometrics, in: A Companion to Theoretical Econometrics. Blackwell Publishing Ltd, New Jersey, USA.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician) 47, 431–443.

Chang, H., Psaris, M., 2013. Local landscape predictors of maximum stream temperature and thermal sensitivity in the Columbia River Basin, USA. Science of The Total Environment 461–462, 587–600. https://doi.org/10.1016/j.scitotenv.2013.05.033

Chen, Q., Mei, K., Dahlgren, R.A., Wang, T., Gong, J., Zhang, M., 2016a. Impacts of land use and population density on seasonal surface water quality using a modified geographically weighted regression. Science of The Total Environment 572, 450–466. https://doi.org/10.1016/j.scitotenv.2016.08.052

Comber, A., Chi, K., Huy, M.Q., Nguyen, Q., Lu, B., Phe, H.H., Harris, P., 2020. Distance metric choice can both reduce and induce collinearity in geographically weighted regression. Environment and Planning B: Urban Analytics and City Science 47, 489–507. https://doi.org/10.1177/2399808318784017

Cressie, N., 1988. Spatial prediction and ordinary kriging. Mathematical Geology 20, 17.

ESRI ArcGIS 10.7.1, 2020. ArcGIS Desktop 10.7.1 quick start guide—ArcGIS Help | Documentation [WWW Document]. URL https://desktop.arcgis.com/en/arcmap/10.7/get-started/setup/arcgis-desktop-quick-start-guide.htm (accessed 4.16.20).

Ganio, L.M., Torgersen, C.E., Gresswell, R.E., 2005. A Geostatistical Approach for Describing Spatial Pattern in Stream Networks. Frontiers in Ecology and the Environment 3, 138. https://doi.org/10.2307/3868541

Harris, P., 2019. A Simulation Study on Specifying a Regression Model for Spatial Data: Choosing between Autocorrelation and Heterogeneity Effects. Geogr Anal 51, 151–181. https://doi.org/10.1111/gean.12163

Isaak, D.J., Peterson, E.E., Ver Hoef, J.M., Wenger, S.J., Falke, J.A., Torgersen, C.E., Sowder, C., Steel, E.A., Fortin, M.-J., Jordan, C.E., Ruesch, A.S., Som, N., Monestiez, P., 2014. Applications of spatial statistical network models to stream data: Spatial statistical network models for stream data. Wiley Interdisciplinary Reviews: Water 1, 277–294. https://doi.org/10.1002/wat2.1023

Lintern, A., Webb, J.A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P., Western, A.W., 2018a. Key factors influencing differences in stream water quality across space. Wiley Interdisciplinary Reviews: Water 5, 1–31. https://doi.org/10.1002/wat2.1260

Lintern, A., Webb, J.A., Ryu, D., Liu, S., Waters, D., Leahy, P., Bende-Michl, U., Western, A.W., 2018b. What Are the Key Catchment Characteristics Affecting Spatial Differences in Riverine Water Quality? Water Resources Research 54, 7252–7272. https://doi.org/10.1029/2017WR022172

Lu, B., Brunsdon, C., Charlton, M., Harris, P., 2017. Geographically weighted regression with parameter-specific distance metrics. International Journal of Geographical Information Science 31, 982–998. https://doi.org/10.1080/13658816.2016.1263731

Lu, B., Charlton, M., Harris, P., Fotheringham, A.S., 2014. Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. International Journal of Geographical Information Science 28, 660–681. https://doi.org/10.1080/13658816.2013.865739

Lu, B., Harris, P., Charlton, M., Brunsdon, C., Nakaya, T., Murakami, D., Gollini, I., 2019. Package ‘GWmodel.’

Mainali, J., Chang, H., Chun, Y., 2019. A review of spatial statistical approaches to modeling water quality. Progress in Physical Geography: Earth and Environment 43, 801–826. https://doi.org/10.1177/0309133319852003

Money, E., Carter, G.P., Serre, M.L., 2009. Using river distances in the space/time estimation of dissolved oxygen along two impaired river networks in New Jersey. Water Research 43, 1948–1958. https://doi.org/10.1016/j.watres.2009.01.034

Neill, A.J., Tetzlaff, D., Strachan, N.J.C., Hough, R.L., Avery, L.M., Watson, H., Soulsby, C., 2018. Using spatial-stream-network models and long-term data to understand and predict dynamics of faecal contamination in a mixed land-use catchment. Science of The Total Environment 612, 840–852. https://doi.org/10.1016/j.scitotenv.2017.08.151

Ord, J.K., Getis, A., 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis 27, 286–306. https://doi.org/10.1111/j.1538-4632.1995.tb00912.x

Peterson, E., Hoef, J.V., 2014. STARS: An ArcGIS toolset used to calculate the spatial information needed to fit spatial statistical models to stream network data. Journal of Statistical Software 56, 1–17.

Peterson, E., Ver Hoef, J., 2014. STARS: An ArcGIS toolset used to calculate the spatial information needed to fit spatial statistical models to stream network data. Journal of Statistical Software 56, 1–17.

Peterson, E.E., Hoef, J.M.V., 2010. A mixed-model moving-average approach to geostatistical modeling in stream networks. Ecology 91, 644–651.

Peterson, E.E., Ver Hoef, J.M., Isaak, D.J., Falke, J.A., Fortin, M.-J., Jordan, C.E., McNyset, K., Monestiez, P., Ruesch, A.S., Sengupta, A., Som, N., Steel, E.A., Theobald, D.M., Torgersen, C.E., Wenger, S.J., 2013. Modelling dendritic ecological networks in space: an integrated network perspective. Ecology Letters 16, 707–719. https://doi.org/10.1111/ele.12084

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

R Project, 2020. R: The R Project for Statistical Computing [WWW Document]. URL https://www.r-project.org/ (accessed 4.16.20).

Scown, M.W., McManus, M.G., Carson, J.H., Nietch, C.T., 2017. Improving Predictive Models of In-Stream Phosphorus Concentration Based on Nationally-Available Spatial Data Coverages. JAWRA Journal of the American Water Resources Association 53, 944–960. https://doi.org/10.1111/1752-1688.12543

Tiefelsdorf, M., Griffith, D.A., 2007. Semiparametric Filtering of Spatial Autocorrelation: The Eigenvector Approach. Environment and Planning A 39, 1193–1221. https://doi.org/10.1068/a37378

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

Tu, J., Xia, Z., 2008. Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation. Science of The Total Environment 407, 358–378. https://doi.org/10.1016/j.scitotenv.2008.09.031

Turschwell, M.P., Peterson, E.E., Balcombe, S.R., Sheldon, F., 2016. To aggregate or not? Capturing the spatio-temporal complexity of the thermal regime. Ecological Indicators 67, 39–48. https://doi.org/10.1016/j.ecolind.2016.02.014

Ullah, K.A., Jiang, J., Wang, P., 2018. Land use impacts on surface water quality by statistical approaches. Global J. Environ. Sci. Manage 4, 231–250. https://doi.org/10.22034/gjesm.2018.04.02.010

Vannote, R.L., Minshall, G.W., Cummins, K.W., Seddel, J.R., Cushing, C.E., 1980. The River Continuum Concept. Canadian Journal of Fish and Aquatic Science 37, 130–137.

Ver Hoef, J., Peterson, E., 2020. Package 'SSN.'

Ver Hoef, J.M., Peterson, E., Theobald, D., 2006. Spatial statistical models that use flow and stream distance. Environmental and Ecological Statistics 13, 449–464. https://doi.org/10.1007/s10651-006-0022-8

Ver Hoef, J.M., Peterson, E.E., 2010. A Moving Average Approach for Spatial Statistical Models of Stream Networks. Journal of the American Statistical Association 105, 6–18.

Yang, X., Jin, W., 2010. GIS-based spatial regression and prediction of water quality in river networks: A case study in Iowa. Journal of Environmental Management 91, 1943–1951. https://doi.org/10.1016/j.jenvman.2010.04.011

Overall Conclusions

In this dissertation, I attempted to advance current knowledge in spatial modeling of surface water quality by carrying out a review of various spatial statistical approaches to water quality modeling, comparing model outputs resulted from different spatial conceptualizations of sampling sites, and demonstrating the incorporation of upstream distance while running geographically weighted regression.

This dissertation concludes that there is still an insufficient emphasis on spatial aspects of water quality measuring sites (e.g., spatial autocorrelation and residual spatial autocorrelation) while modeling water quality. Additionally, most of the current research only uses standard distance matrix and do not compare spatial conceptualizations and resultant weight matrix. Weight matrices have great potential in informing spatial autocorrelation of dependent variables at different scales, and in helping test several hypotheses of spatial eco-socio-hydrological processes in relation to surface water. Although most of the spatial models are recognizing and incorporating the directional aspect of water flow, the local model development by using geographically weighted regression models has not yet considered an up-stream distance matrix.

The second chapter provides a novel example of using graph theory in elucidating relationships among water quality measurement sites and their affinity with landscape processes. The model strengths are usually different according to the different spatial conceptualization of interrelations among sampling stations, as demonstrated by the graph types. Among different graph types compared, the relative graph types provided the highest model strength, signifying stronger up-stream downstream relation with dissolved oxygen, while k-mean graph types with four neighbors provided the strongest model performance, indicating the impact of local factors in electrical conductivity. The spatial regression models were successfully developed and compared using water quality data collected in the field, and various geographic information systems based on social and environmental data. Among the factors considered in the analysis, we found the population density, agricultural land cover, and percentage sand cover negatively impact the water quality as revealed by their relationships with DO and conductivity.

In chapter three, we demonstrated that a successful model could be developed by combining a distance matrix derived from spatial stream network models with geographically weighted regression. The upstream distance weighted models provided a comparable model strength that of standard geographically weighted regression. The spatial patterns of model strength, as well as various coefficients from the upstream distance weighted regressions, are crucial in elucidating local patterns as a product of upstream-downstream relations. The upstream distance weighted geographically weighted regression model can offer better insights into hydrological and biogeochemical relationships among different sampling sites and their relationships with watershed processes. These insights not only help in understanding these processes but also in designing locally specific water quality management endeavors.