International Journal of Geo-Information

MDPI

Article

Accounting for and Predicting the Influence of Spatial Autocorrelation in Water Quality Modeling

Lorrayne Miralha [1] and Daehyun Kim [2,*]

[1] Department of Geography, University of Kentucky, Lexington, KY 40508, USA; lorraynemiralha@uky.edu
[2] Department of Geography, Seoul National University, Seoul 08826, Korea
* Correspondence: biogeokim@snu.ac.kr; Tel.: +82-2-880-4059; Fax: +82-2-876-9498

Received: 28 November 2017; Accepted: 17 February 2018; Published: 19 February 2018

Abstract: Several studies in the hydrology field have reported differences in outcomes between models in which spatial autocorrelation (SAC) is accounted for and those in which SAC is not. However, the capacity to predict the magnitude of such differences is still ambiguous. In this study, we hypothesized that SAC, inherently possessed by a response variable, influences spatial modeling outcomes. We selected ten watersheds in the USA and analyzed if water quality variables with higher Moran's *I* values undergo greater increases in the coefficient of determination (R^2) and greater decreases in residual SAC (rSAC). We compared non-spatial ordinary least squares to two spatial regression approaches, namely, spatial lag and error models. The predictors were the principal components of topographic, land cover, and soil group variables. The results revealed that water quality variables with higher inherent SAC showed more substantial increases in R^2 and decreases in rSAC after performing spatial regressions. In this study, we found a generally linear relationship between the spatial model outcomes (R^2 and rSAC) and the degree of SAC in each water quality variable. We suggest that the inherent level of SAC in response variables can predict improvements in models before spatial regression is performed.

Keywords: spatial autocorrelation; water quality; spatial modeling; coefficient of determination; residual autocorrelation

1. Introduction

Water is an element crucial for life on Earth and is closely linked to the well-being of societies as well as the sustainability of aquatic ecosystems. A combination of natural and anthropogenic factors can adversely impact water quality. Human impacts involve general land use practices (e.g., agriculture, irrigation practices, urbanization, and deforestation), while natural factors include slope, elevation, vegetation cover, soil type, precipitation, and streamflow [1–3]. River characteristics are generally dependent upon land use and geomorphological features of the watershed under study. In addition, water use patterns associated with the location of a region and its interactions with neighboring regions influence the quality of water bodies [4]. These factors are responsible for the spatial variability of water quality, and are often treated as predictor variables in many hydrologic models [5]. To provide better insights to future watershed management policies, understanding spatial processes associated with water quality variables is of extreme importance.

Space serves a vital role in structuring hydrological systems. Spatial autocorrelation (SAC) is an inherent property of spatial features such as streams and rivers [6]. Legendre defined the concept of SAC as "the property of random variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation), or less similar (negative autocorrelation) than expected for randomly associated pairs of observations" (p. 1659) [7]. For example, causes of positive autocorrelation in stream water quality could be associated with similarities in local habitats or

turbulent mixing and water chemistries of stream flows. In contrast, specific local built structures, such as beaver dams, fallen trees in stream channels, and territorial fishes, could be causes of negative SAC [8]. Given these interactions over space (i.e., water flow from upstream to downstream areas, local biota, and water use patterns), it is necessary to consider the presence and potential effects of SAC in water quality modeling.

Numerous studies in ecology, geography, and hydrology have noted the importance of accounting for SAC [9–12]. These studies show that ignoring SAC can bias model outcomes and parameter estimates, leading to poor statistical inference and violation of the independence assumption of conventional regression approaches [8,13–16]. For example, models that ignore spatial effects (e.g., ordinary least squares; OLS) are likely to produce autocorrelated residuals violating the independent errors assumption. This can inflate the Type I error rate, wrongfully rejecting a null hypothesis. Many spatial approaches have been developed in order to overcome such limitations of non-spatial counterparts. These approaches include, but are not restricted to, regression kriging, simultaneous autoregressive modeling, conditional autoregressive modeling, spatial lag modeling, spatial error modeling, spatial eigenvector mapping, and geographically weighted regression [8,9,17–26].

Several water quality studies have compared outcomes between spatial and non-spatial regressions [2–4,10,11,27–29]. In general, spatial models presented significant increases in R^2 values and decreases in residual SAC (rSAC), indicating that spatial model performance exhibited clear improvements over the non-spatial approach. However, as per the literature on hydrological modeling, it is still uncertain when such improvements become large or small. Assuming that each water quality variable presents a unique degree of inherent SAC, we hypothesize that this SAC (possessed by a response variable; i.e., a water quality variable) influences the outcomes of spatial modeling. We test if water quality variables with a higher amount of SAC would exhibit greater improvement in model outcomes than those with a lower amount of SAC (see Figure 1). We evaluate this hypothesis across divergent regions of the USA to enable a general understanding of the effect of SAC possessed by water quality variables. We examine if SAC is a consistent determinant of the magnitude of model improvements even when watershed characteristics diverge. If this is indeed the case, we can potentially determine the degree of improvement in model fit before performing a spatial regression simply by measuring the inherent SAC level of a water quality variable. This study can also serve as a useful screening technique where modelers could use Moran's *I* to predict the spatial pattern in the independent variable using a spatially explicit method.

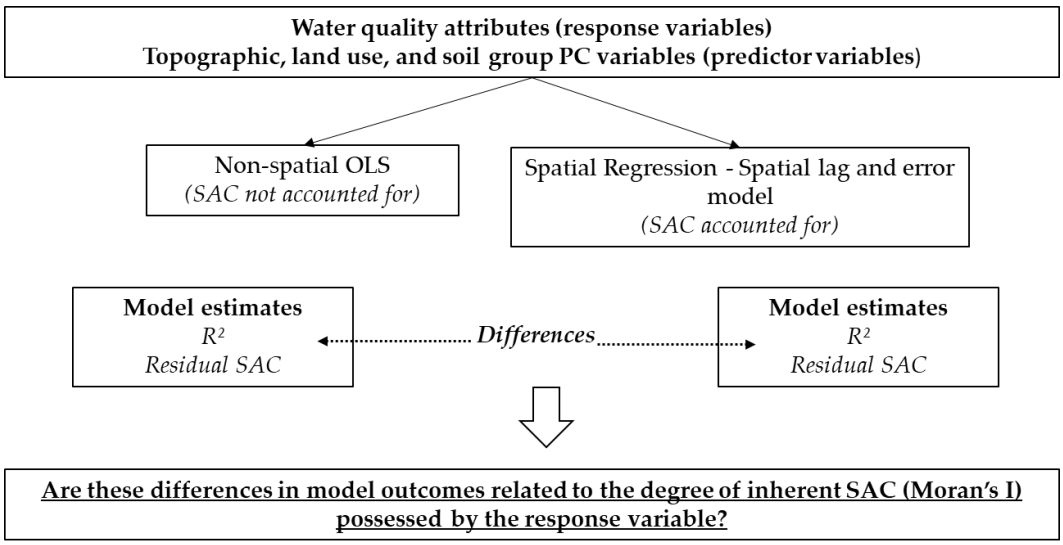


Figure 1. Conceptualization of the main ideas of the study (PC, principal components; OLS, ordinary least squares; SAC, spatial autocorrelation).

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

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

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

Study Areas										
State	**LA**	**AZ**	**KS**	**VA**	**CA**	**CO**	**DE**	**ID**	**IA**	**KY**
Watershed	**Bayou Louis/ Lake Louis**	**Cherry Creek**	**Indian Creek**	**Difficult River**	**Headwaters Tuolumne River**	**Upper San Miguel River**	**Clay, Mill, Bradywine Creek, and Cristina River**	**Lower South Fork Coeur d'Alene River**	**Iowa River**	**Beaver Creek**
Area (km^2)	288.58	586.26	193.8	150.84	553.66	763.71	352.24	308.49	193.96	407.07
Stations	29	31	33	33	31	32	36	32	32	54
Water quality parameter (Moran's *I*)	pH (0.13) T (0.15) SC (0.20) DO (0.28) TDS (0.53)	DO (−0.08) * pH (−0.07) * T (0.54) SC (0.59)	TN (0.013) SC (0.022) DIN (0.07) KjN (0.10) TP (0.15) T (0.20) Tur 0.25) DO (0.44) pH (0.72)	Tur (−0.28) * TDS (−0.26) * SC (0.06) Br (0.09) Cl (0.12) Mg (0.15) Na (0.15) DO (0.16) Ca (0.17) SiO_2 (0.19) Fe (0.21) K (0.25) CO_2 (0.34) Mn (0.34) pH (0.39) Alk (0.40) TP (0.42) SO_4^{2-} (0.45) F (0.54) T (0.69)	Csu (−0.20) * T (0.30) Mg (0.42) K (0.46) Ca (0.55) Cl (0.58) Na (0.59) SiO_2 (0.62) SO_4^{2-} (0.65) TDS (0.73) Alk (0.80) pH (0.82)	DO (0.39) SC (0.36) pH (0.37) T (0.67)	SC (−0.05) * T (−0.006) * Chla (0.02) TN (0.03) Nin (0.05) Alk (0.08) TP (0.12) DO (0.15) pH (0.16) Cl (0.23) TOC (0.32) DOC (0.32)	Pb (0.11) T (0.15) Zn (0.24) pH (0.31) Cd (0.35) As (0.47) SC (0.56)	DO (0.18) pH (0.34) NO_3^- (0.36) T (0.49) PO_4^{3-} (0.66) Cl (0.67)	Al (0.005) Ba (0.06) Alk (0.11) Na (0.14) Cl (0.23) K (0.26) Nin (0.29) TDS (0.32) SO_4^{2-} (0.38) Fe (0.40) KjN (0.43) Mg (0.47) Ca (0.55) Mn (0.58)

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

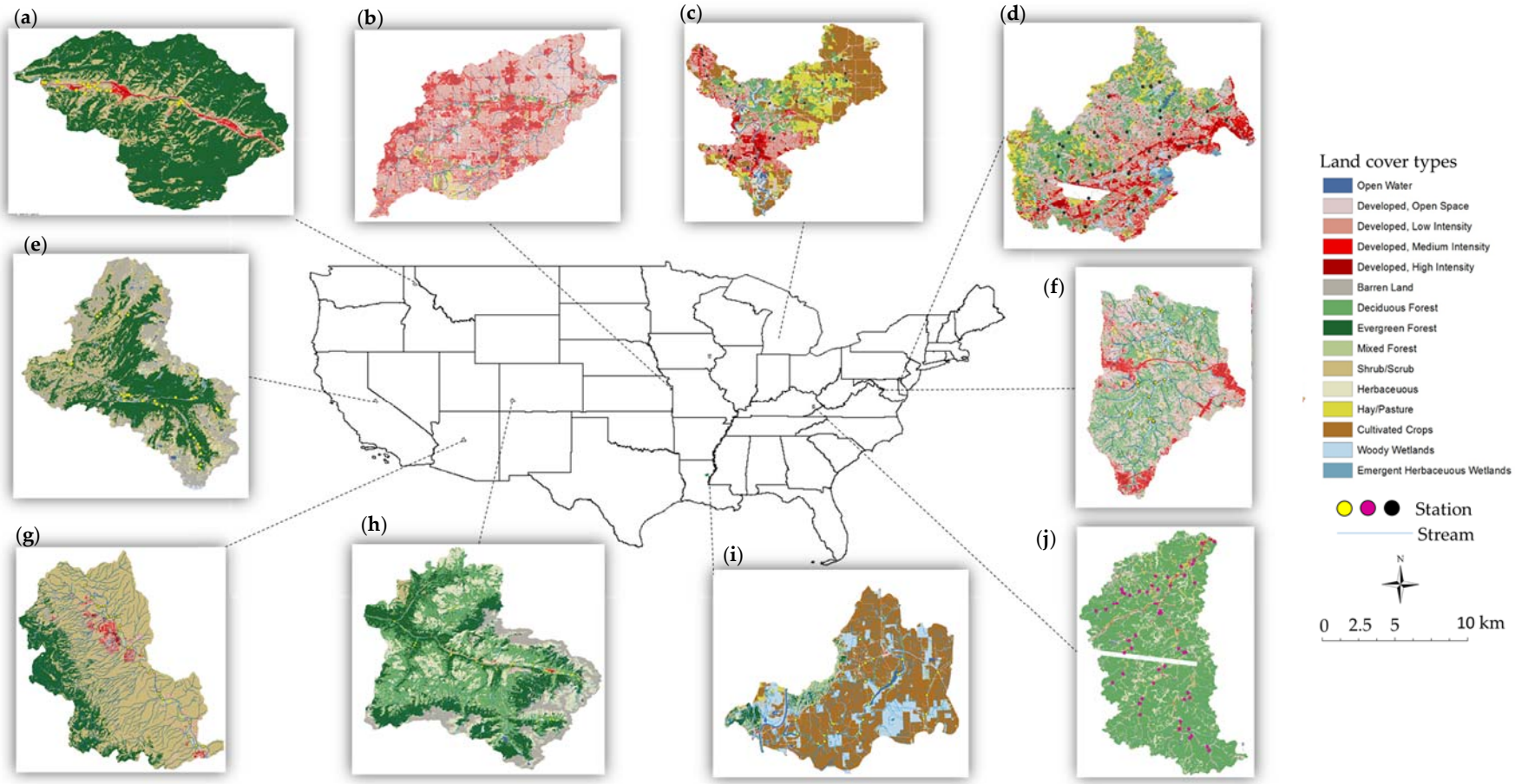


Figure 2. Land cover characteristics of each state and watershed shape. Idaho (**a**); Kansas (**b**); Iowa (**c**); Delaware (**d**); California (**e**); Virginia (**f**); Arizona (**g**); Colorado (**h**); Louisiana (**i**); and Kentucky (**j**). To better visualize the water quality stations spatial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

2. Materials and Methods

2.1. Study Areas

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

2.2. Dependent Variables

Water quality data from 2011 to 2017 were obtained online from the national Water Quality Portal (WQP) [30]. The WQP integrates publicly available water quality data from three very important and widely used sources for research in the US: the USGS National Water Information System (NWIS), the EPA STOrage and RETrieval (STORET) Data Warehouse, and the United States Department of Agriculture (USDA) Sustaining the Earth's Watersheds Agricultural Research Data System (STEWARDS) through the Water Quality eXchange (WQX).

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

2.3. Delineation of Upstream Area

Characteristics of the sub-watershed area upstream of sampling stations affect water quality variables [11]. Thus, sub-watershed boundaries were delimited using the 'ArcHydro' package tool of ArcGIS 10.3 (Environmental Systems Research Institute, Redlands, CA, USA). We downloaded spatial stream data from the 2016 US Geological Survey (USGS) National Hydrography Dataset [31]. The distance between stations varied, as did the size of each upstream area delineated. Land use characteristics as well as topography and soil far from the stream channel might contribute less to changes in water quality across space [3]. Thus, we used the upstream area to separate the stream network specific to each station, and delineated the riparian zone around the stream. Many studies have conducted analyses at the riparian area scale, mainly by considering a buffer area on each side of the stream. Overall, there was no specific buffering distance recommended [3,11,32]. In this study, we used a buffer zone of 50 m each side of the stream (i.e., a 100 m buffer in total) as the area that can contribute the maximum to water quality changes (Figure 3). We performed these analyses for all watersheds in this study.

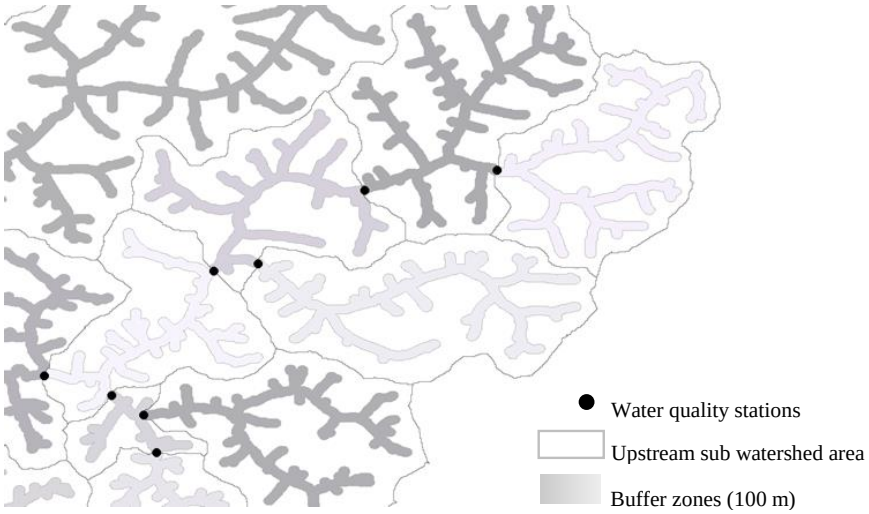


Figure 3. Upstream area delineation and their respective buffer zones in tones of gray. The solid circles are water quality stations (sites).

2.4. Independent Variables

Using the buffer zones of the upstream area, we extracted the land use, topography, and soil types associated with each sampling station. These variables were treated as independent variables in the subsequent modeling. The summary of these variables is shown in Table 3. We downloaded the land use raster with 30 m resolution from USGS The National Map—2011 National Land Cover Database (USGS TNM-NLCD) [33]. In this study, we considered the percentage of four major land use types surrounding stream networks: urban, agriculture, forest, and wetland. To extract this information, we used the 'Zonal Statistics' toolset in ArcGIS 10.3. The percentage of urban area in each upstream buffer zone was calculated using the sum of the low-, medium-, and high-intensity urbanization, and open space values in the land use raster. The sum of values for pasture and cultivated crop was used to calculate the percentage of agricultural land in the area. The values for deciduous forest, evergreen forest, and mixed forest were used to arrive at the percentage of forest, while the values for woody wetlands and emergent herbaceous wetland were combined to calculate wetland percentage. For the topographic variables, we used 10 m resolution digital elevation models (DEMs) downloaded from USGS the National Map Elevation Products (USGS TNM 3DEP) [34]. Using the same upstream area and zonal statistic toolset, we extracted the mean and standard deviation of the elevation and slope respectively for each station's upstream area. These variables were used to account for topographic complexity.

We downloaded the hydrological soil groups (HSGs) from the Natural Resources Conservation Service's (2017 NRCS) Soil Survey Geographic (SSURGO) database [35]. We extracted the percentages of A, B, C, D, A/D, B/D, and C/D categories of soil for each site. The HSGs are categorized by the hydraulic conductivity level of a soil and how much runoff it produces. This is usually associated with the percentage of sediment grain sizes a soil presents. Typically, group A soils have a low runoff capacity because the water transmissivity through the soil profile is very high. Thus, group A soils are composed of a high percentage of sediments with large grain size, such as sand or gravel. Group B soils have a moderate runoff capacity. Nevertheless, water flows freely through the soil profile and the percentage of large-sized grains is high. In this case, however, small grain size sediments such as clay can reach up to 20 percent of the total. Group C soils have a moderately high runoff capacity and have a higher clay percent, with less than 50 percent of sand. Group D soils are characterized as having the highest percentage of fine grains such as clay and silt. The dual HSGs (A/D, B/D, and C/D) are wet soils where the water table is within 60 cm below the surface but can still be drained adequately. The first letter indicates the drained condition, and the second, the undrained [36].

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

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

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

2.5. Data Preprocessing

We tested the normality of each dependent and independent variable using IBM SPSS Statistics for Windows Version 23.0 (Armonk, NY, USA). In this study, the independent variables are likely to present a high level of correlations due to their nature. For example, agriculture and urban zones are land use types that might express a negative relationship because, as the area under agricultural use increases, the urbanized areas will tend to decrease. Thus, to account for the multicollinearity in the subsequent modeling, we applied principal component analysis (PCA), a multivariate technique. This technique reduces the dimensionality of a multivariate dataset where variables are significantly interrelated. This reduction results in principal components (PCs), which are considered uncorrelated variables [37,38]. PCA is useful because it simplifies the description of the independent variables and the modeling procedure. We divided the independent variables into three main groups: land use, topography, and soil. Land use considered the percentage of urban, agriculture, wetland, and forest areas. The topographic group encompassed the mean and standard deviation values of slope and elevation. The soil groups represented the percentage of A, B, C, D, A/D, B/D, and C/D soil types (Table 3). Overall, we had three main PC groups used as the predictors in the models. Each variable category presents one to three PCs, depending on how significant the variables in the group are to the area of study. This means that a model can have three to nine principal components as independent variables.

2.6. Testing for Spatial Autocorrelation (SAC)

We quantified the inherent degree of SAC for each water quality parameter using Moran's I function (Equation (1)):

$$I = \frac{n}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \frac{\sum_{i=1}^{n}\sum_{j}^{n} W_{ij}\left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)}{\sum_{i=1}^{n}\sum_{j}^{n} W_{ij}} \quad (1)$$

where, X_i and X_j refer to the water quality at station i and station j, respectively. $\overline{X}$ is the overall mean water quality, and W_{ij} is the weight matrix. Moran's I values vary between -1 to 1 for maximum negative and positive autocorrelation, respectively. No-zero values of Moran's I indicate that values at a certain geographical Euclidian distance are more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for randomly assigned values [39,40]. We used the geographic coordinate system based on angular values (longitude and latitude) considering the North American 1983 as the datum for the distance calculation. We acknowledge that we did not perform projection in this study, which would have been a serious issue if we were concerned with region-scale modeling crossing multiple states. Instead, the current study examined the water quality of several stations within local watersheds (<ca. 764 km^2). Therefore, using the Euclidean distance should not be a critical problem.

2.7. Statistical Models

GeoDa version 1.8 (Chicago, IL, USA) was used to run three models in this paper. First, OLS, representing the non-spatial model, is a multiple linear regression approach (Equation (2)), where the response variable is the water quality parameter and the independent variables are the PCs of the topographic, land cover, and soil groups:

$$Y_\mathrm{i} = \beta_o + \beta_1 X_1 + \cdots + \beta_\mathrm{i} X_\mathrm{i} + \varepsilon_\mathrm{i} \quad (2)$$

where Y_i is the response variable, β_o is the constant in a linear model, β_i are coefficients associated with the independent variables, and ε_i is the error term. Notably, the same independent and dependent

variables were used as in the spatial modeling approaches. The second model was a spatial lag model (Equation (3)):

$$Y_i = X_i\beta_i + \rho WY_j + \varepsilon \quad (3)$$

where Y_i and Y_j are the dependent variables at locations i and j, respectively, X_i is the independent variable at i, β_i is the regression coefficient, ρ is the spatial autoregressive coefficient, WY_j is the spatially lagged dependent variable, and ε is the error term. This model accounts for the fact that the dependent variable is affected by the independent variables in adjacent places, and, thus, the dependent variable is spatially lagged as a predictor. The third model used was the spatial error model (Equation (4)):

$$W_i = X_i\beta_i + \varepsilon\ \varepsilon = \lambda W\varepsilon + \varepsilon \quad (4)$$

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

2.8. Model Comparison

After measuring the inherent degree of SAC for each water quality variable, we compared the outcomes of non-spatial OLS and spatial regression approaches in terms of R^2 and rSAC. To quantify rSAC, we estimated Moran's I for residuals. After the modeling procedure, we evaluated our hypothesis by plotting Moran's I values of the water quality variables against the R^2 and rSAC values for each water quality variable (Figure 4). A few water quality variables presented negative inherent SAC values and were treated as positive in this graph. This is because we intended to concentrate on the magnitude of SAC.

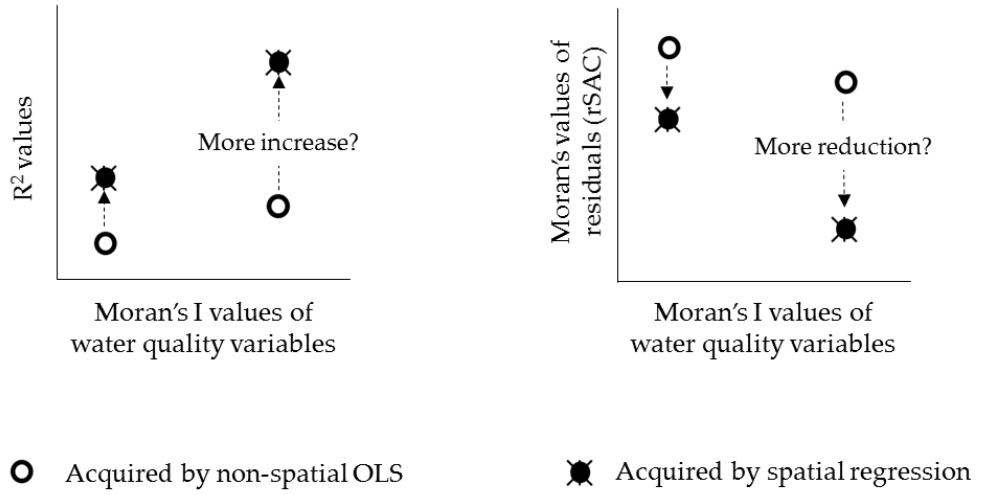


Figure 4. Evaluation of the hypothesis—Moran's I values of the water quality variables appear on the x-axis, and the model outcomes, R^2 and residual SAC, appear on the y-axis. After spatial regression, water quality variables with a higher amount of spatial autocorrelation (SAC) were hypothesized to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

3. Results

3.1. Changes in R^2 Values

Overall, Moran's I values pertaining to water quality variables varied widely, from 0.01 to 0.82, across all watershed sites (Figure 5). The relationships shown in Figure 5 indicate that the

improvements in R^2 were proportional to the degree of inherent SAC in water quality variables (i.e., the hypothesis predicting increases in R^2 as a function of the degree of SAC is supported). Whether we treated each state separately or combined them as a whole, strongly autocorrelated water quality parameters over space (i.e., having higher Moran's I values) exhibited greater increases in R^2 values after spatial regression compared to weakly autocorrelated variables (i.e., having lower Moran's I values). This pattern seemed to be less clear when water quality variables within a state had a relatively narrow range of Moran's I (e.g., Delaware).

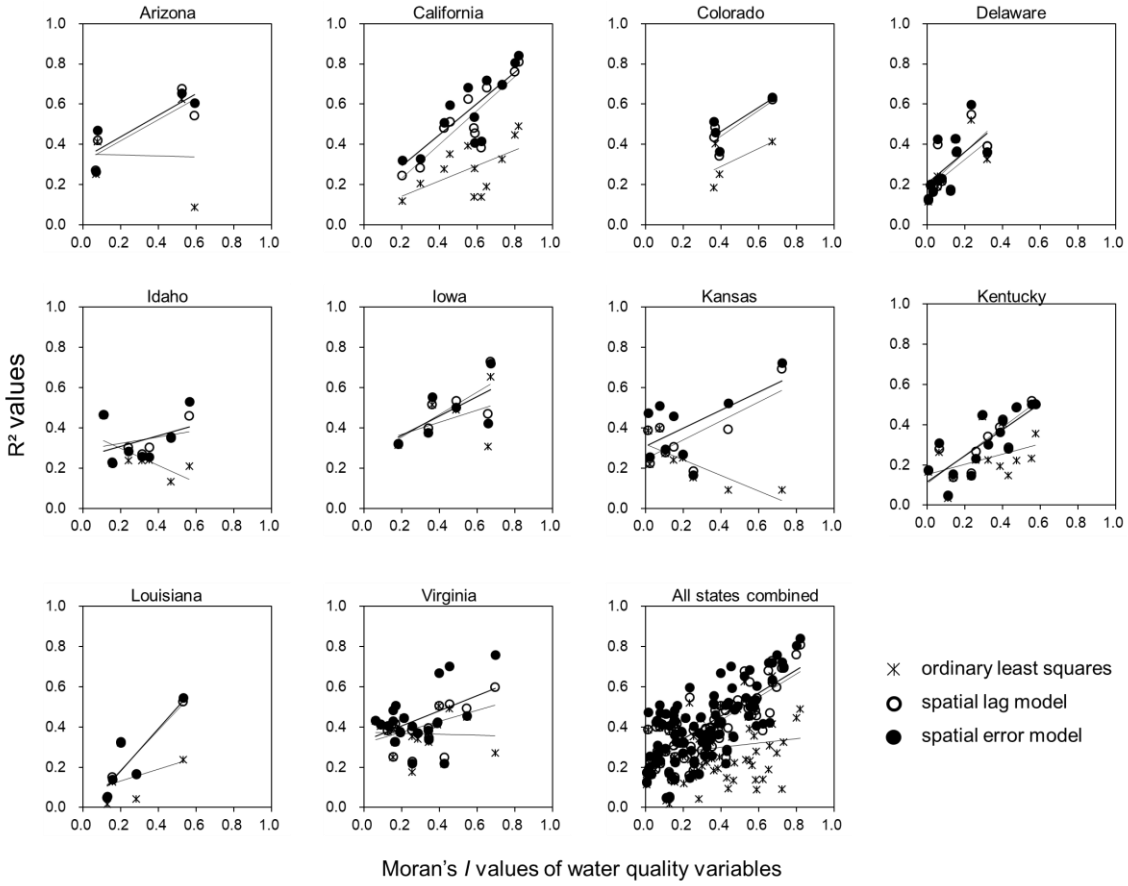


Figure 5. Relationship between the spatial autocorrelation (SAC) of each water quality variable (represented by Moran's I values) and the R^2 indicating the amount of variance in each water quality variable, explained by topographic, land use, soil groups, and spatial terms.

3.2. Changes in rSAC

The values of Moran's I indicating rSAC produced by non-spatial OLS presented a wider range than those from spatial regression (i.e., rSAC for non-spatial OLS from 0.01 to 0.72, while spatial lag rSAC ranged from 0.00 to 0.44, and spatial error, from 0.00 to 0.07). We found a positive correlation between the degree of SAC in water quality variables and rSAC from non-spatial OLS. Conversely, as expected, rSAC values acquired by spatial regressions were in general near zero. Therefore, the larger the Moran's I values possessed by water quality variables, the greater the reduction in rSAC after running models that consider spatial dependence (Figure 6; the hypothesis predicting greater decreases in rSAC, proportional to the degree of SAC in water quality variables, is supported). All states presented significant reduction in rSAC except Delaware, showing a narrow range of Moran's I values of water quality variables.

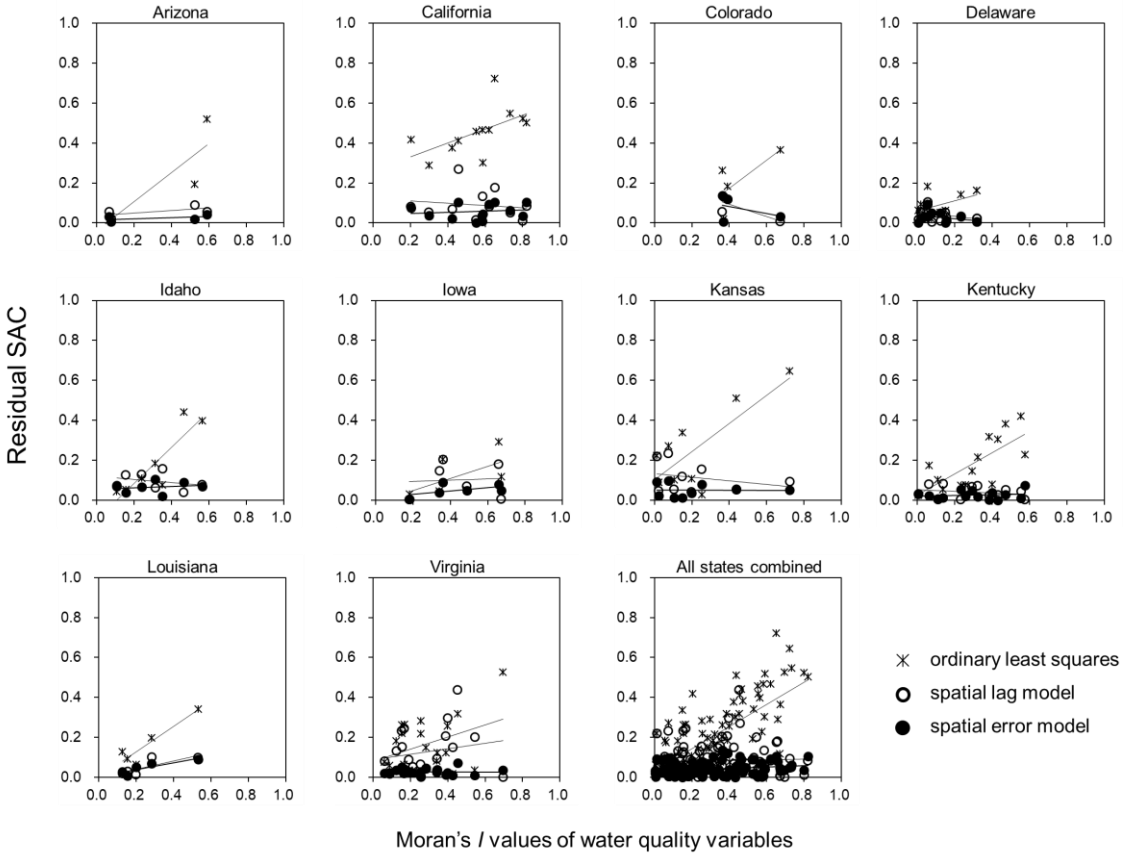


Figure 6. Relationship between the spatial autocorrelation (SAC) of each water quality variable (represented by Moran's *I* values) and the SAC of model residuals (also represented by Moran's *I* values). "All states combined" showed a general reduction in residual SAC after accounting for spatial autocorrelation in the models of each water quality variable.

3.3. Overall Changes between Non-Spatial OLS and Spatial Regression Models

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

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

			California	Colorado	Delaware	Idaho	Iowa	Kentucky	Arizona	Kansas	Louisiana	Virginia	All States Combined
	OLS	*Samples*	12	4	12	7	6	14	4	9	5	20	93
		I	0.56	0.45	0.13	0.31	0.45	0.30	0.31	0.22	0.26	0.29	0.32
		R^2	0.28	0.31	0.27	0.25	0.44	0.23	0.34	0.23	0.15	0.37	0.29
		rSAC	0.39	0.21	0.09	0.19	0.12	0.19	0.19	0.26	0.16	0.17	0.21
After spatial regression	Improvement in R^2	*lag-ols*	0.26	0.16	0.03	0.09	0.05	0.09	0.13	0.11	0.09	0.03	0.10
		error-ols	0.29	0.18	0.04	0.09	0.04	0.08	0.15	0.17	0.10	0.07	0.12
	Reduction in rSAC	*ols-lag*	0.37	0.13	0.05	0.09	0.02	0.15	0.13	0.14	0.11	0.04	0.12
		ols-error	0.40	0.13	0.07	0.12	0.07	0.16	0.17	0.21	0.12	0.14	0.17
Linear regression models for the Change in R^2 vs. *I*	Model fit Spatial Lag	R^2	0.55	0.12	0.07	0.85	0.68	0.61	0.51	0.91	0.94	0.46	0.58
		β_o	0.00	0.07	0.01	−0.09	−0.07	−0.04	−0.04	−0.07	−0.09	−0.05	−0.15
		$\beta 1$	0.46	0.19	0.11	0.58	0.26	0.44	0.55	0.86	0.70	0.30	0.74
		p-value	<0.001 *	0.10 *	0.60	0.10 *	0.53	0.08 *	0.39	0.09 *	0.38	0.28	<0.001 *
	Model fit Spatial Error	R^2	0.40	0.03	0.00	0.77	0.64	0.55	0.42	0.77	0.93	0.29	0.36
		β_o	0.07	0.11	0.03	−0.13	−0.04	−0.04	−0.02	−0.01	−0.10	−0.04	−0.04
		β_1	0.39	0.15	0.02	0.68	0.19	0.40	0.56	0.83	0.75	0.40	0.60
		p-value	<0.001 *	0.06 *	0.52	0.15	0.62	0.10 *	0.33	0.02 *	0.39	0.06 *	<0.001 *
Linear regression models for the Change in rSAC vs. *I*	Model fit Spatial Lag	R^2	0.33	0.56	0.58	0.66	0.42	0.60	0.67	0.67	0.80	0.03	0.31
		β_o	0.14	−0.32	0.01	−0.21	−0.10	−0.03	−0.07	−0.03	0.00	−0.01	−0.05
		β_1	0.41	1.01	0.36	0.98	0.27	0.57	0.66	0.80	0.42	0.17	0.18
		p-value	<0.001 *	0.18	0.01 *	0.20	0.71	<0.001 *	0.34	0.08 *	0.09 *	0.32	<0.001 *
	Model fit Spatial Error	R^2	0.32	0.87	0.42	0.84	0.28	0.45	0.77	0.60	0.74	0.17	0.39
		β_o	0.22	−0.26	0.02	−0.15	−0.03	0.01	−0.05	0.05	0.00	0.05	−0.03
		β_1	0.32	0.88	0.33	0.87	0.22	0.51	0.70	0.71	0.46	0.30	0.17
		p-value	<0.001 *	0.17	0.00 *	0.11	0.15	<0.001 *	0.25	0.02 *	0.08 *	<0.001 *	<0.001 *

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

3.4. Summary of Findings

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

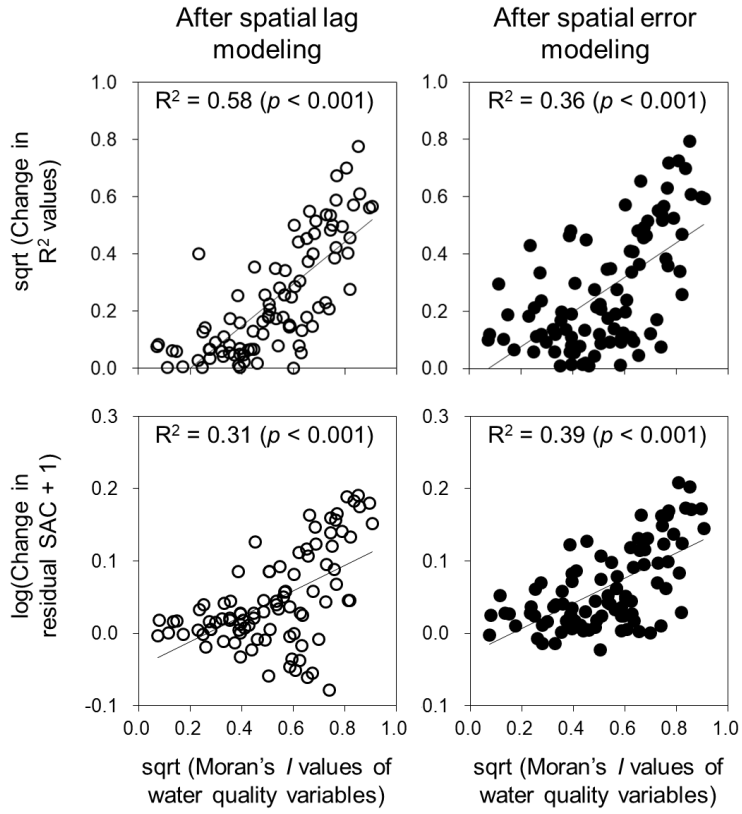


Figure 7. Linear regression models demonstrating that the magnitude of improvement of model performance after spatial lag and error modeling is significantly and linearly explained by the SAC inherently possessed by water quality variables. The Moran's I (x-axis) and Change in R^2 (y-axis) values were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

4. Discussion

The results support our hypothesis and offer insights into the field of water quality modeling. Most importantly, the level of SAC in water quality variables has the potential to indicate how much improvement a non-spatial model would experience if SAC was appropriately considered (i.e., increases in R^2 values and decreases in rSAC). We have demonstrated across divergent watersheds in the USA that the higher the SAC in a water quality variable, the greater the improvements to the model after accounting for SAC. Water quality studies, as previously mentioned, presented better results when considering spatial modeling approaches that account for SAC [2–5]. However, these studies have not considered the magnitude of SAC in the response variable as the main driver of model improvements. Furthermore, we observed that variables with lower degree of inherent SAC (lower Moran's I values) underwent smaller changes in model outcomes compared to those that presented larger Moran's I values. In this sense, higher Moran's I values imply more spatial organization (e.g., strong connection among water quality stations through the stream network) than smaller

Moran's I values. This indicates that the need for (and potentially the benefit from) accounting for SAC in water quality modeling increases as the degree of SAC increases.

In this study, we investigated water quality variables from 10 watersheds, each distinct in geology, land use, soil, and topography. We analyzed a total of 93 water quality variables that also differed among the watersheds. Despite such strong peculiarities among the watersheds, this study reveals a consistent and linear relationship between the SAC of water quality parameters and changes in the model outcomes (R^2 and rSAC). This finding perfectly accords with the study of Kim et al. [13] who evaluated the effect of SAC in soil–landform modeling to find that the degree of SAC in soil variables (i.e., dependent variables) influenced model improvements after the SAC was properly accounted for.

Our findings suggest that future water quality modeling studies should account for SAC to improve the performance of non-spatial approaches, principally when the predictors in the model cannot sufficiently account for all SAC in the model [6,7,9,12,13,15]. Overall, the improvements include increasing R^2 and decreasing rSAC. The most important point is that the degrees of these increases and decreases showed to be linearly correlated with the level of SAC in water quality variables. Therefore, water quality studies should not only focus on accounting for spatial autocorrelation, but also on understanding the magnitude of SAC inherent in water quality variables. Doing so, we could point out the degree of connectivity within water quality variables, as well as the improvement in model outcomes of a non-spatial approach before performing a spatial regression.

Adequate information on the degree of hydrologic connectivity among water quality variables is needed in watershed management and policy decisions [41,42]. The level of SAC inherent in a variable can allow managers to reveal the complex spatial relationship of water quality as well as its changes from up to downstream. It can uncover dissimilarity patterns among water quality parameters throughout the stream network in study and help in the implementation of policies that are ecologically beneficial to the aquatic ecosystem. Therefore, we highlight that the investigation of SAC in water quality modeling is not only beneficial in the model results, but also in the process of watershed management.

Streams can be considered spatially structured ecological networks, where patterns are usually associated with the in-stream flow and habitat, or even the physical structure of the network. The understanding of these patterns can be limited when only using Euclidean distance [43]. For example, two sites that are near to each other can be considered neighbors due to distance in the Euclidean technique, but they can present distinct water quality measures simply due to the water quality origins from vastly different drainage areas. Therefore, we highlight that this is a limitation in this study and further studies should focus on applying spatial network distance techniques to better understand the SAC influence.

5. Conclusions

Spatial autocorrelation (SAC) is a property possessed by any ecological or environmental variable. Consequently, its incorporation and impacts on modeling results have been studied in much detail in a variety of scientific fields. Our study demonstrates that analyzing SAC in water quality modeling provides benefits beyond just improvements in model outcomes (R^2 and rSAC): it can potentially lead to a better understanding of the extent of spatial organization of water quality variables, as well as serve as a useful screening technique to anticipate the predictability of the spatial pattern in the independent variable used in a spatially explicit model. We also highlight the benefits of understanding the level of SAC possessed by a water quality variable in the process of watershed management and point that network distances techniques could better account for the spatial pattern existent in spatially structured ecological networks such as streams.

Acknowledgments: This research was supported by (1) the National Science Foundation (#1560907) of the USA, (2) the National Research Foundation of South Korea (NRF-2017R1C1B5076922), (3) the Research Resettlement Fund for the new faculty of Seoul National University, and (4) the 4-Zero Land Space Creation of the Ministry of

Education and the NRF (#1345258304). We appreciate the constructive comments of Heejun Chang and Yongwan Chun on the earlier version of this paper.

Author Contributions: Lorrayne Miralha collected the data online, performed the data analyses, and wrote the paper. Daehyun Kim devised the project and conceptual main ideas. Daehyun Kim encouraged Lorrayne Miralha to investigate the idea in the hydrology field and supervised the findings of this work. Both authors discussed the results and contributed to the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Larger Maps of the Study Areas for Better Visualization of the Water Quality Stations

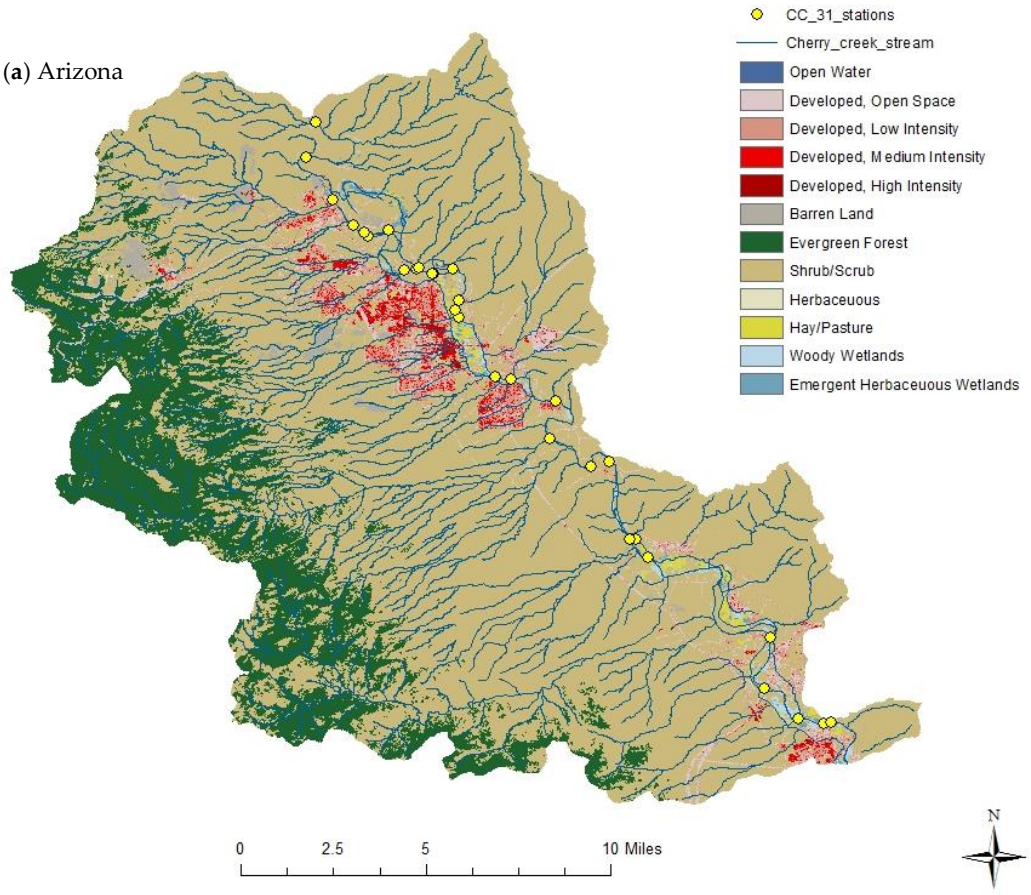


Figure A1. *Cont.*

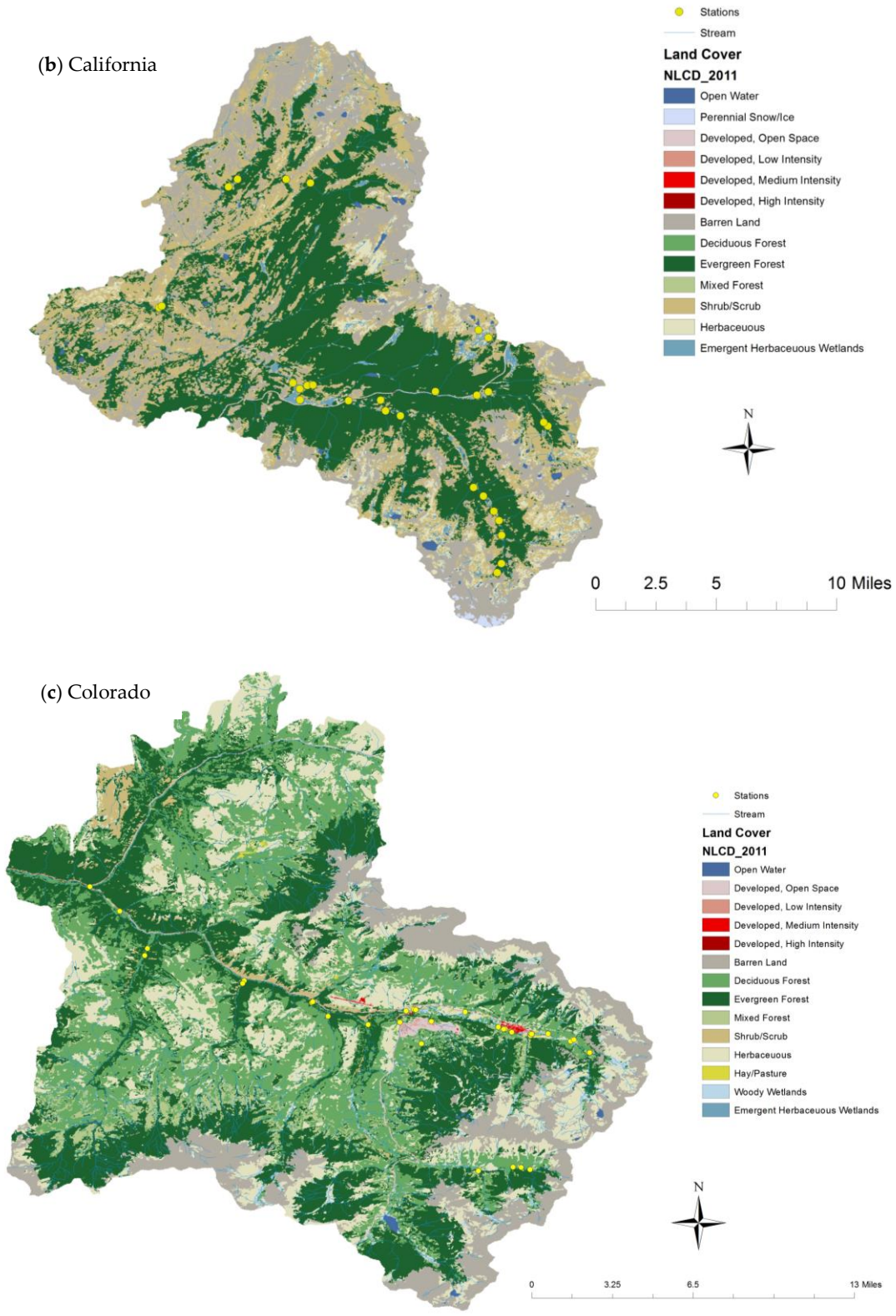


Figure A1. *Cont.*

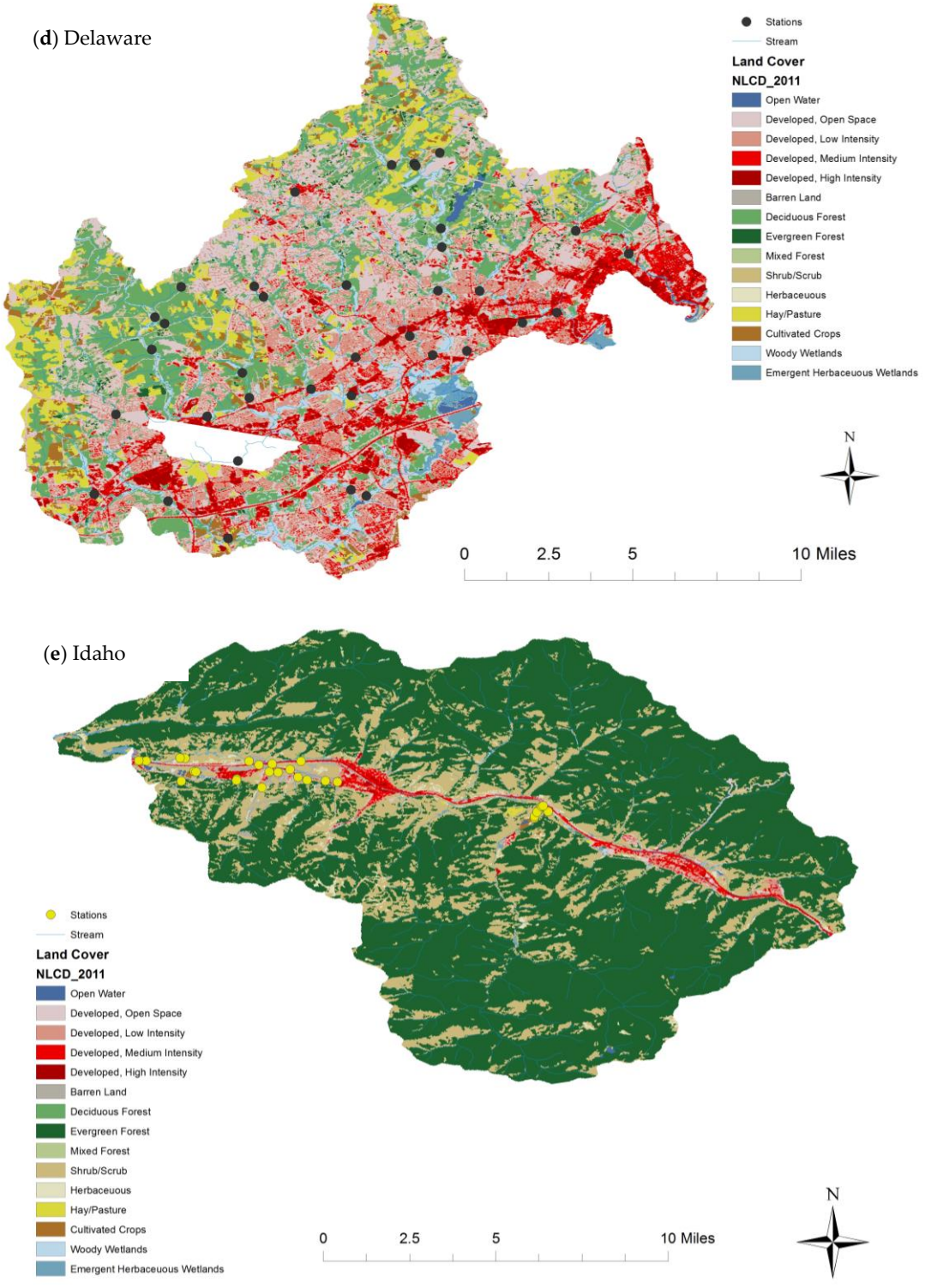


Figure A1. *Cont.*

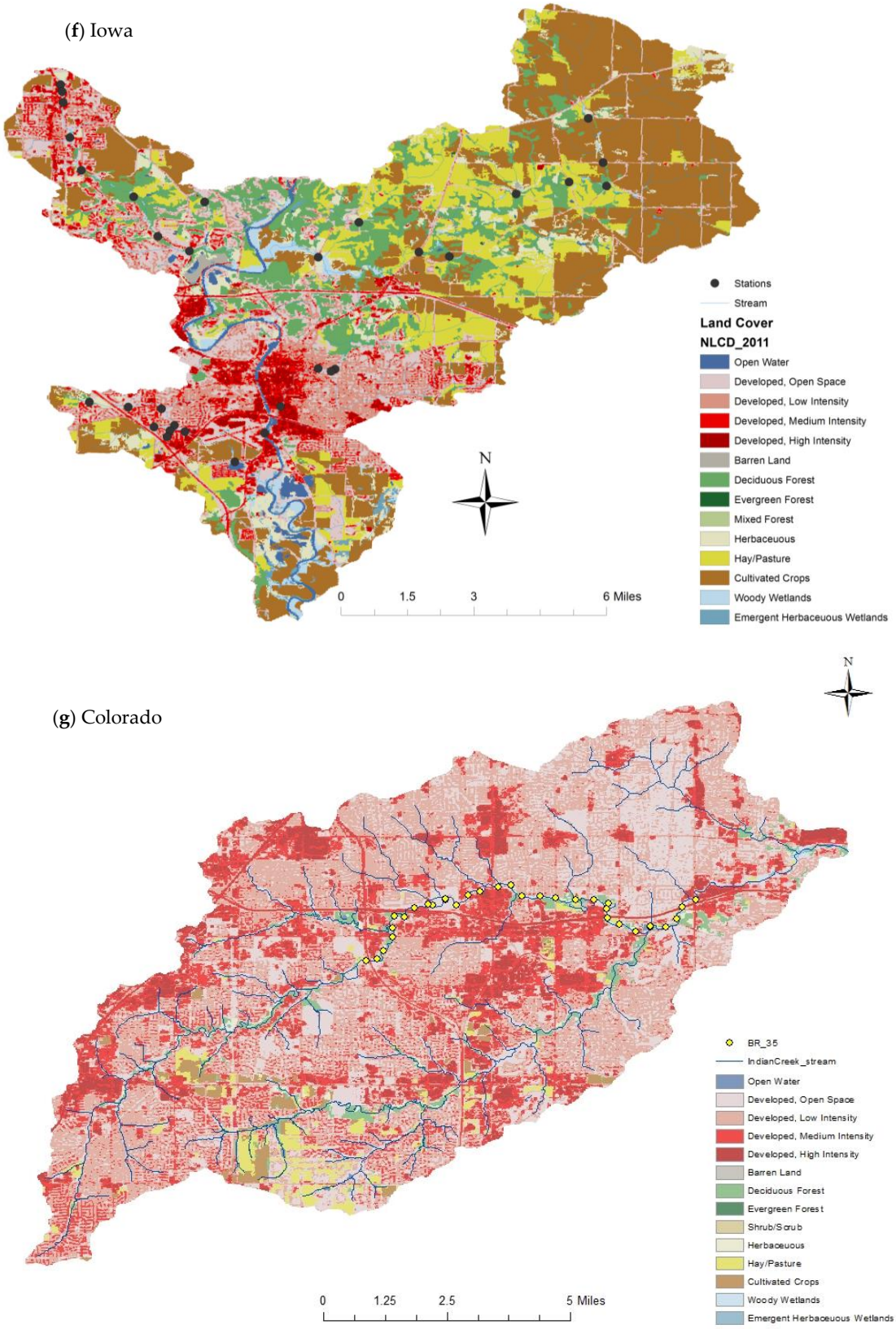
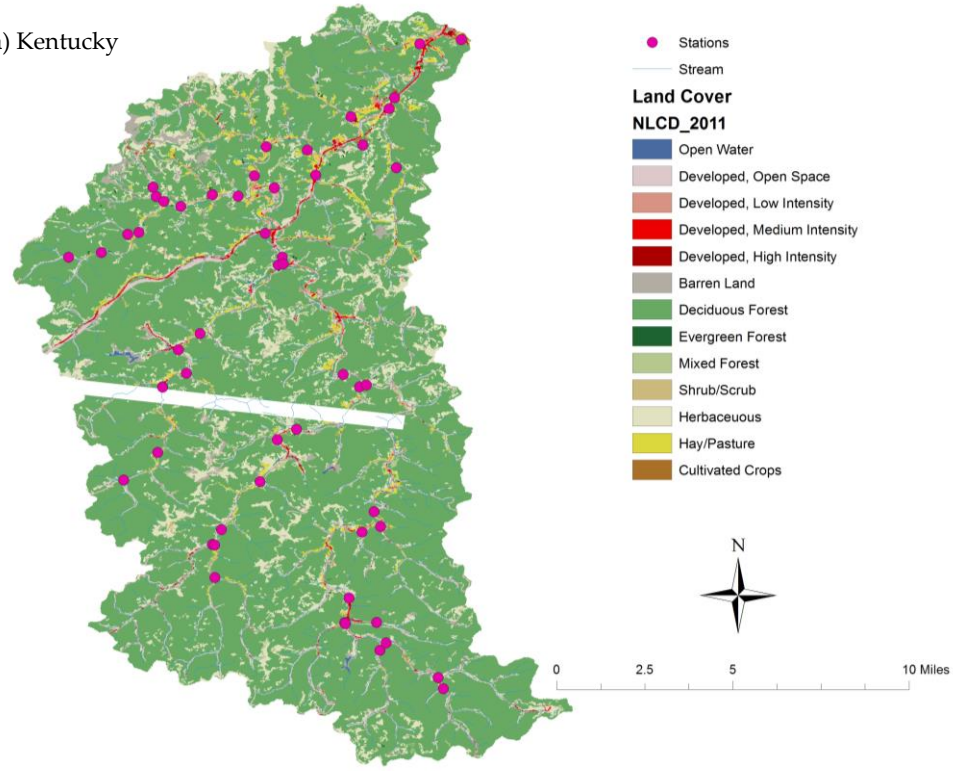


Figure A1. *Cont.*

(**h**) Kentucky



(**i**) Louisiana

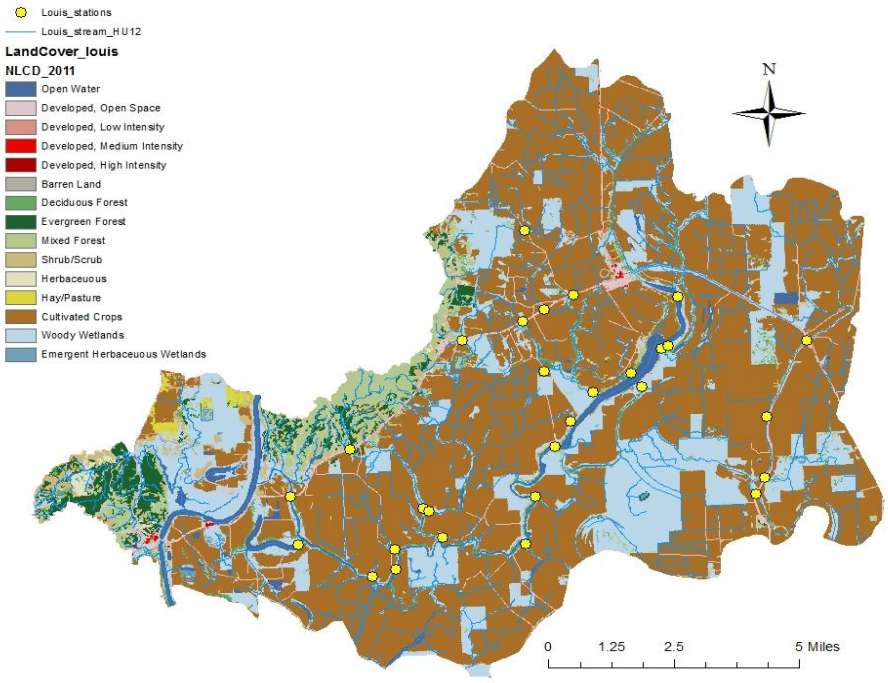


Figure A1. *Cont.*

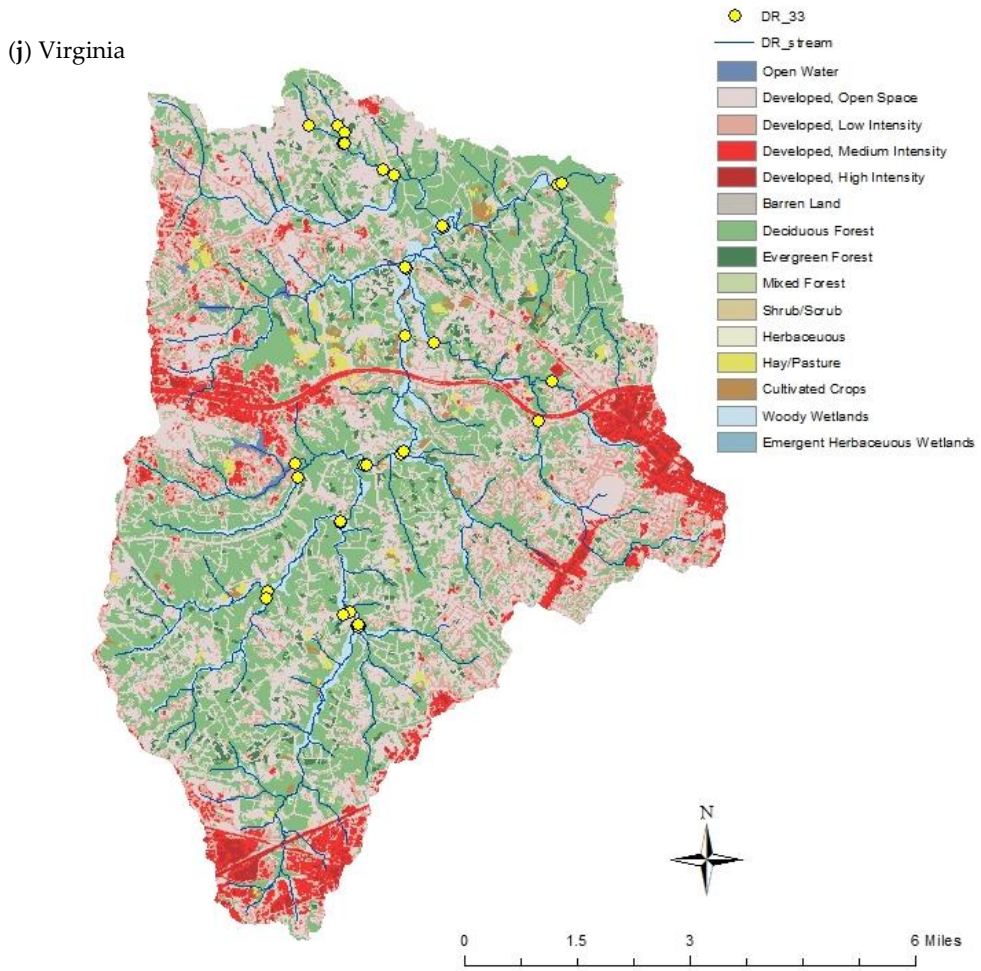


Figure A1. Water stations organization in the stream network and land cover characteristics of each study area. Idaho (**a**); Kansas (**b**); Iowa (**c**); Delaware (**d**); California (**e**); Virginia (**f**); Arizona (**g**); Colorado (**h**); Louisiana (**i**); and Kentucky (**j**).

References

1. Calow, P.; Petts, G.E. *The Rivers Handbook*; Part 1; Blackwell Scientific: London, UK, 1992; Volume 1, ISBN 0-632-02832-7.
2. Yu, D.; Shi, P.; Liu, Y.; Xun, B. Detecting land use-water quality relationships from the viewpoint of ecological restoration in an urban area. *Ecol. Eng.* **2013**, *53*, 205–216. [CrossRef]
3. Pratt, B.; Chang, H. Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. *J. Hazard. Mater.* **2012**, *209*, 48–58. [CrossRef] [PubMed]
4. Franczyk, J.; Chang, H. Spatial analysis of water use in Oregon, USA, 1985–2005. *Water Resour. Manag.* **2009**, *23*, 755–774. [CrossRef]
5. Vrebos, D.; Beauchard, O.; Meire, P. The impact of land use and spatial mediated processes on the water quality in a river system. *Sci. Total Environ.* **2017**, *601*, 365–373. [CrossRef] [PubMed]
6. Legendre, P.; Fortin, M.J. Spatial pattern and ecological analysis. *Vegetatio* **1989**, *80*, 107–138. [CrossRef]
7. Legendre, P. Spatial autocorrelation: Trouble or new paradigm? *Ecology* **1993**, *74*, 1659–1673. [CrossRef]
8. 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.; et al. Applications of spatial statistical network models to stream data. *Wiley Interdiscip. Rev. Water* **2014**, *1*, 277–294. [CrossRef]
9. Kim, D. Incorporation of multi-scale spatial autocorrelation in soil moisture–landscape modeling. *Phys. Geogr.* **2013**, *34*, 441–455. [CrossRef]

10. Tu, J. Spatially varying relationships between land use and water quality across an urbanization gradient explored by geographically weighted regression. *Appl. Geogr.* **2011**, *31*, 376–392. [CrossRef]
11. Chang, H. Spatial analysis of water quality trends in the Han River basin, South Korea. *Water Res.* **2008**, *42*, 3285–3304. [CrossRef] [PubMed]
12. Miller, J.; Franklin, J.; Aspinall, R. Incorporating spatial dependence in predictive vegetation models. *Ecol. Model.* **2007**, *202*, 225–242. [CrossRef]
13. Kim, D.; Hirmas, D.R.; McEwan, R.W.; Mueller, T.G.; Park, S.J.; Šamonil, P.; Thompson, J.A.; Wendroth, O. Predicting the Influence of Multi-Scale Spatial Autocorrelation on Soil–Landform Modeling. *Soil Sci. Soc. Am. J.* **2016**, *80*, 409–419. [CrossRef]
14. Cliff, A.; Ord, J.K. Testing for Spatial Autocorrelation among Regression Residuals. *Geogr. Anal.* **1972**, *4*, 267–284. [CrossRef]
15. Dormann, C.F. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Glob. Ecol. Biogeogr.* **2007**, *16*, 129–138. [CrossRef]
16. Beale, C.M.; Lennon, J.J.; Yearsley, J.M.; Brewer, M.J.; Elston, D.A. Regression analysis of spatial data. *Ecol. Lett.* **2010**, *13*, 246–264. [CrossRef] [PubMed]
17. Václavík, T.; Kupfer, J.A.; Meentemeyer, R.K. Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). *J. Biogeogr.* **2012**, *39*, 42–55. [CrossRef]
18. De Marco, P.; Diniz-Filho, J.A.F.; Bini, L.M. Spatial analysis improves species distribution modelling during range expansion. *Biol. Lett.* **2008**, *4*, 577–580. [CrossRef] [PubMed]
19. Miller, J.A. Species distribution models: Spatial autocorrelation and non-stationarity. *Prog. Phys. Geog.* **2012**, *36*, 681–692. [CrossRef]
20. Bini, L.M.; Diniz-Filho, J.A.F.; Rangel, T.F.; Akre, T.S.; Albaladejo, R.G.; Albuquerque, F.S.; Aparicio, A.; Araújo, M.B.; Baselga, A.; Beck, J.; et al. Coefficient shifts in geographical ecology: An empirical evaluation of spatial and non-spatial regression. *Ecography* **2009**, *32*, 193–204. [CrossRef]
21. Kissling, W.D.; Carl, G. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Glob. Ecol. Biogeogr.* **2008**, *17*, 59–71. [CrossRef]
22. Hengl, T.; Heuvelink, G.B.; Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **2004**, *120*, 75–93. [CrossRef]
23. Griffith, D.A. A linear regression solution to the spatial autocorrelation problem. *J. Geogr. Syst.* **2000**, *2*, 141–156. [CrossRef]
24. Griffith, D.A.; Peres-Neto, P.R. Spatial modeling in ecology: The flexibility of eigenfunction spatial analyses. *Ecology* **2006**, *87*, 2603–2613. [CrossRef]
25. Ver Hoef, J.M.; Peterson, E.; Theobald, D. Spatial statistical models that use flow and stream distance. *Environ. Ecol. Stat.* **2006**, *13*, 449–464. [CrossRef]
26. Lichstein, J.W.; Simons, T.R.; Shriner, S.A.; Franzreb, K.E. Spatial autocorrelation and autoregressive models in ecology. *Ecol. Monogr.* **2002**, *72*, 445–463. [CrossRef]
27. Chang, H.; Jung, I.W.; Steele, M.; Gannett, M. Spatial patterns of March and September streamflow trends in Pacific Northwest streams, 1958–2008. *Geogr. Anal.* **2012**, *44*, 177–201. [CrossRef]
28. Huang, J.; Huang, Y.; Zhang, Z. Coupled effects of natural and anthropogenic controls on seasonal and spatial variations of river water quality during baseflow in a coastal watershed of Southeast China. *PLoS ONE* **2014**, *9*, e91528. [CrossRef] [PubMed]
29. Netusil, N.R.; Kincaid, M.; Chang, H. Valuing water quality in urban watersheds: A comparative analysis of Johnson Creek, Oregon, and Burnt Bridge Creek, Washington. *Water Resour. Res.* **2014**, *50*, 4254–4268. [CrossRef]
30. NWQMC (National Water Quality Monitoring Council). Water Quality Portal. Available online: http://www.waterqualitydata.us/ (accessed on 19 June 2017).
31. United State Geological Survey. USGS National Hydrography Dataset (NHD) Downloadable Data Collection. 2016. Available online: http://nhd.usgs.gov (accessed on 22 June 2017).
32. Li, S.; Gu, S.; Tan, X.; Zhang, Q. Water quality in the upper Han River basin, China: The impacts of land use/land cover in riparian buffer zone. *J. Hazard. Mater.* **2009**, *165*, 317–324. [CrossRef] [PubMed]
33. United State Geological Survey. The National Map, 2011, National Land Cover Database (USGS TNM-NLCD). Available online: https://viewer.nationalmap.gov (accessed on 22 June 2017).

34. United State Geological Survey. The National Map Elevation Products (USGS TNM 3DEP). 2017. Available online: https://viewer.nationalmap.gov (accessed on 22 June 2017).
35. Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database. Available online: https://sdmdataaccess.sc.egov.usda.gov (accessed on 22 June 2017).
36. United States Department of Agriculture, Natural Resources Conservation Service. Part 630 Hydrology—Hydrologic Soil Groups. In *National Engineering Handbook*; Title 210-VI [Online]; U.S. Department of Agriculture, Soil Conservation Service (SCS): Washington, DC, USA, 2009; pp. 1–7. Available online: https://directives.sc.egov.usda.gov (accessed on 08 October 2017).
37. Jolliffe, I.T. Principal component analysis. In *Principal Component Analysis*; Springer: Berlin/Heidelberg, Germany, 1986; pp. 1–9.
38. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
39. O'sullivan, D.; Unwin, D. *Geographic Information Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2014; ISBN 978-0-470-28857-3.
40. Diniz-Filho, J.A.F.; Bini, L.M.; Hawkins, B.A. Spatial autocorrelation and red herrings in geographical ecology. *Glob. Ecol. Biogeogr.* **2003**, *12*, 53–64. [CrossRef]
41. Pringle, C. What is hydrologic connectivity and why is it ecologically important? *Hydrol. Process.* **2003**, *17*, 2685–2689. [CrossRef]
42. Pringle, C.M. Hydrologic connectivity and the management of biological reserves: A global perspective. *Ecol. Appl.* **2001**, *11*, 981–998. [CrossRef]
43. 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.; et al. Modelling dendritic ecological networks in space: An integrated network perspective. *Ecol. Lett.* **2013**, *16*, 707–719. [CrossRef] [PubMed]