Science of the Total Environment 761 (2021) 144057

Contents lists available at ScienceDirect

Science of the Total Environment

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

Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach

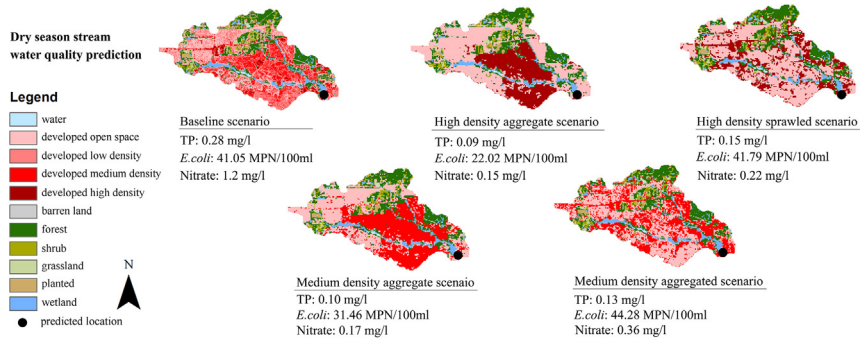
Runzi Wang [a,*], Jun-Hyun Kim [b], Ming-Han Li [b]

[a] School for Environment and Sustainability, University of Michigan, 440 Church Street, Ann Arbor, MI 48109-1041, United States of America
[b] School of Planning, Design and Construction, Michigan State University, 552 W Circle Dr, East Lansing, MI 48823, United States of America

HIGHLIGHTS

- Pattern is more important than percentage of urban area for stream water quality.
- High density aggregated development leads to much lower TP and NO_3^--N concentration.
- LPI, COHESION, SPLIT of developed areas are important predictors for water quality.
- The effect of urban development pattern has high spatial and seasonal variations.
- Machine learning is promising for predicting stream water quality.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 July 2020
Received in revised form 20 November 2020
Accepted 22 November 2020
Available online 14 December 2020

Editor: Ashantha Goonetilleke

Keywords:
Urban form
Water quality
Landscape metrics
Urban sprawl
Machine learning
Scenario planning

ABSTRACT

Urban development pattern significantly impacts stream water quality by influencing pollutant generation, build-up, and wash-off processes. It is thus necessary to understand and predict stream water quality in accordance with different urban development patterns to effectively advise urban growth planning and policies. To do so, we collected pollutant concentration data on nitrate (NO_3^--N), total phosphate (TP), and *Escherichia coli* (*E. coli*) from 1047 sampling stations in the Texas Gulf Region. We utilized a Random Forest (RF) machine learning model to predict stream water quality under four planning scenarios with different urban densities and configurations. SHapley Additive exPlanations (SHAP) was used to prove the importance of urban development pattern in influencing stream water quality. The spatial variations of the impact of these patterns were explored with Geographically Weighted Regression (GWR). SHAP results indicated that Largest Patch Index (LPI), Patch Cohesion Index (COHESION), Splitting Index (SPLIT), and Landscape Division Index (DIVISION) were the most important urban development pattern metrics affecting stream water quality. The spatial variations of such patterns were shown to impact stream water quality depending on pollutants, seasonality, climate, and urbanization level. RF prediction results suggested that high density aggregated development was more effective in reducing TP and NO_3^--N concentrations than the current sprawl development, but had the potential risk of increasing *E. coli* pollution in the wet season. The results of this study provide empirical evidence and a potential mechanistic explanation that stream water quality degradation is a consequence of urban sprawl. Lastly, machine learning is a powerful tool for scenario prediction in land use planning to forecast environmental impacts under different urban development pattern scenarios.

* Corresponding author.
E-mail addresses: runziw@umich.edu (R. Wang), junhkim@msu.edu (J.-H. Kim), minghan@msu.edu (M.-H. Li).

https://doi.org/10.1016/j.scitotenv.2020.144057
0048-9697/

1. Introduction

Human-induced land use, such as urban and industrial land use, is recognized as a dominant factor affecting stream water quality. A small increase in the percentage of urban land use has been found to exert a disproportionately large influence on generation of pollutants (Ai et al., 2015; Giri and Qiu, 2016; Oeding et al., 2018; Sun et al., 2011; Wijesiri et al., 2018). Within a similar percentage of urban developed areas, varying patterns of urban development can contribute to considerable differences in stream water quality due to different pollutant generation, build-up, and wash-off processes (Goonetilleke et al., 2005; Liu et al., 2012). Therefore, stream water quality prediction in various locations, densities, and patterns of urban development can serve as a basis for developing sound stream water quality management schemes (Fan and Shibata, 2015; Holcomb et al., 2018). However, the specific influence of urban development patterns on stream water quality remains unclear. In a large, regional spatial extent with the heterogeneous landscape, the influence of urban development patterns on stream water quality also shows great spatial and temporal non-stationarity (Chen et al., 2016; Pratt and Chang, 2012; Tu and Xia, 2008).

Urban development pattern has complex influences on stream water quality as measured by the interactions between area, shape, edge, aggregation of urban areas, and stream pollutant concentrations (Forman, 2014; Sun et al., 2014; Yu et al., 2013). Theoretically, large areas of directly connected impervious areas (DCIA) have been shown to harm downstream water bodies (Del Monaco, 2017; Jones et al., 2005; Obropta and Del Monaco, 2018; Sohn et al., 2017). However, this does not necessarily mean that urban development should be more dispersed to reduce DCIA, as it can lead to potential ecosystem fragmentation and difficulty in implementing management practices (Bu et al., 2014; Shi et al., 2017). The ambiguity regarding whether intact or fragmented urban areas cause stream water degradation is apparent in the contradictory conclusions of investigations between urban development patterns and stream water quality. Some researchers have argued that an intact urban pattern with large amounts of impervious surface can contribute to water quality deterioration (Ding et al., 2016; Li et al., 2015). However, other studies found that greater interspersion of urban areas, as indicated by high Contiguity Index and Patch Cohesion Index, significantly increased the export of pollutants due to the destruction of natural areas (Lv et al., 2015; Shi et al., 2013). More research is needed to address this question, particularly by controlling urban developed areas at the same percentage. Doing so ensures that different urban development patterns are comparable in terms of their influence on stream water quality.

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

Furthermore, quantifying the relationships between stream water quality and urban development pattern necessitates the development of predictive models that can be used to forecast stream water quality in alternative urban planning scenarios (Avila et al., 2018; Holcomb et al., 2018; Molina-Navarro et al., 2020; Sharifi et al., 2017). Machine learning algorithms, such as boosted regression tree analysis, neural networks, and self-organizing maps, have been applied to depict the complex, non-linear relationships between landscape characteristics and stream water quality with satisfactory model performance (Castrillo and García, 2020; Hameed et al., 2016; Kalteh et al., 2008; Sajedi-Hosseini et al., 2018; Xu et al., 2020). One advantage of the machine learning model is the possibility of revealing the complex, nonlinear relationships between land cover characteristics and stream water quality (Mirzaei et al., 2020). The other advantage is that after the accuracy of a machine learning model is tested on a new dataset, it could be applied to forecast stream water quality under future land use planning scenarios to support policy decision-making (Chermack and Swanson, 2008; Schreiber et al., 2019). This study enhances the existing machine learning studies in the area of stream water quality prediction by creating interpretable machine learning models that uncover the importance of urban development pattern, and facilitate scenario prediction of stream water quality with the same impervious area but different urban development patterns.

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

2. Data and methods

2.1. Study site

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

Moreover, the increasing population in the study site has resulted in problems associated with urban sprawl, which has put natural forest areas at risk and degraded stream water quality. Texas currently has a population of approximately 29 million people, with a growth rate of 1.8% every year (World Population Review, 2019). Nonpoint source pollution closely related to urban expansion contributes to 75% of stream water quality impairment in Texas (Texas Commission on Environmental Quality, 2012). The resulting high nitrogen and

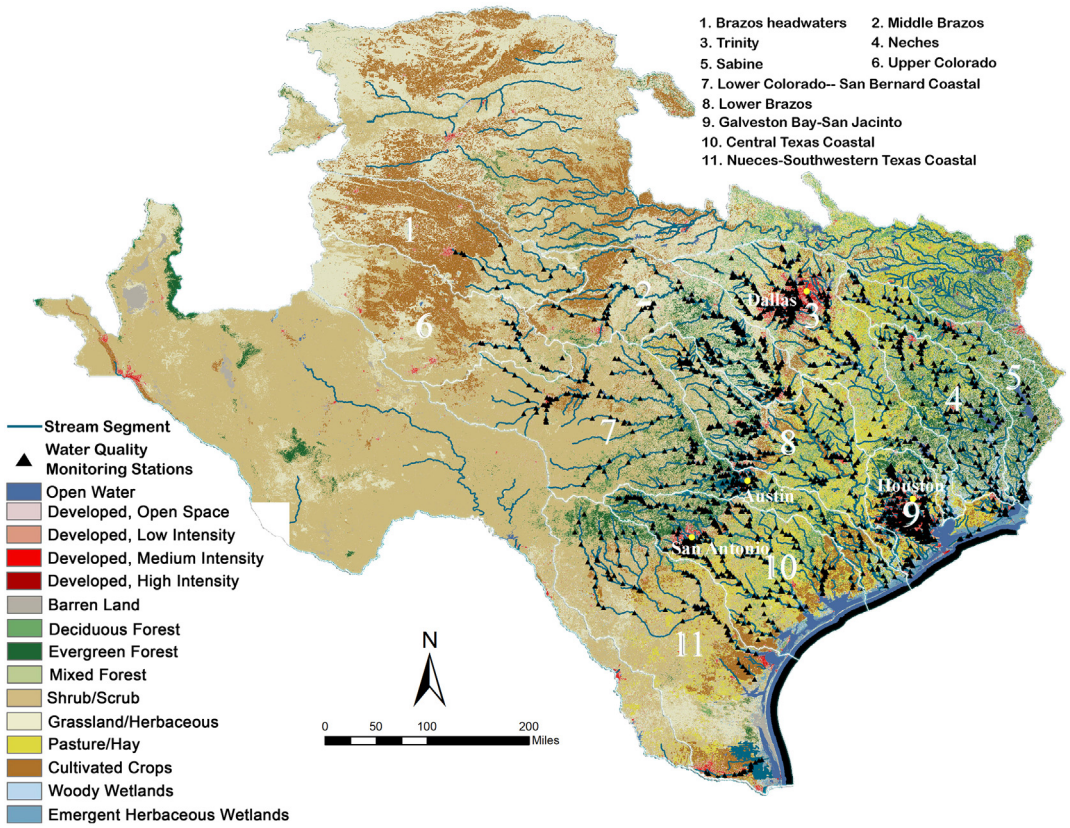


Fig. 1. Map of study site showing land cover, basins, and the locations of sampling stations.

phosphorous levels may contribute to the eutrophication of waterbody that fuels the growth of phytoplankton and periphyton, and lead to low dissolved oxygen levels for the aquatic system. The combination of high nutrient levels, high water temperature, and low dissolved oxygen levels that are frequently observed along the Texas coast is primarily responsible for fish kills (Parsons, 2019; Texas Commission on Environmental Quality, 2019). Elevated concentration of bacteria such as *E. coli* and fecal coliform, which is found in the south-central Texas streams, may indicate the health risk in contact recreational activity (Texas Commission on Environmental Quality, 2005). According to the above regional water quality issues, we selected NO_3^--N, TP, and *E. coli* concentrations as the contaminants of interest in this study.

2.2. Data and variables

Pollutant concentration data of NO_3^--N, TP, and *E. coli* from 1047 sampling stations in the Texas Gulf Region were used as predicted variables in this study. To monitor and assess stream water quality, the Texas Commission on Environmental Quality's (TCEQ) Surface Water Quality Monitoring (SWQM) Program has installed over 3000 active monitoring stations throughout the region. Pollutant concentration data in 2011 were obtained from the SWQM program and aggregated in dry and wet seasons by taking the average values (Table S2). According to the monthly average precipitation in Texas, the dry season went from November to April, and the wet season occurred the rest of year (Pratt and Chang, 2012). In this study, TP and *E. coli* samples covered most basins in the Texas Gulf Region. After seasonal aggregation, the sample size of TP and *E. coli* data in the dry season was 868 and 788 respectively. However, because of data availability, NO_3^--N concentration sample size was 329 in the dry season, covering only the coastal basins. All the pollutant concentration data were highly positively skewed (Table S1).

Landscape metrics at both class and landscape levels, climate, soil, topography, and population were included as explanatory variables to explain variations in stream water quality (Table 1). The class level metrics included land covers of developed area, developed open area, forest area, and planted area, which have been demonstrated to be major environmental drivers of changes in stream water quality (Clément et al., 2017; Glińska-Lewczuk et al., 2016; Teklu et al., 2016). Our analytical steps focused on metrics from urban development pattern and used other metrics as control variables. The definition of all land covers was in accordance with NLCD (Homer et al., 2015), and all variables in this study and their corresponding data sources are presented in Table S2.

We incorporated high dimensions of landscape metrics in the machine learning models, including 76 class level metrics and 32 landscape level metrics in the categories of area, edge, shape, and contagion/interspersion (McGarigal, 1995), as presented in Table 1. The random forest machine learning model that we used in this study generally works well with high-dimensional problems because of the random subsets of variables in the predicting process (Darst et al., 2018). A large set of features can also potentially increase predicting accuracy. The explanation and calculation of all the landscape metrics can be found in Table S3.

We added environmental and social control variables including precipitation, temperature, slope, elevation, soil type, soil storage depth, population, and population density to control for model bias (Table 1). The PRISM monthly climate dataset, which is a gridded climate dataset, was used in this study. The watershed climatic variables were calculated by taking the average values among all the grids inside the watershed boundary. To simplify interpretation, the aggregated seasonal total precipitation and mean temperature were used to find variable importance. The original monthly total precipitation and mean temperature were used in the predictive models to facilitate higher predicting accuracy. In this study, soil type referred to hydrological soil groups (HSG). HSG A, B, C, and D have a high infiltration rate, a moderate infiltration

Table 1
Explanatory variables.

Category	Subcategory	Variable
Class Level Metrics (76)[a] (including classes of developed open area, developed area, forest area, and planted area)	Area (28)	Percentage of Landscape (PLAND), Total Area (CA), Median of Patch Area (AREA_MD), Median of Radius of Gyration (GYRATE_MD), Largest Patch Index (LPI), Number of Patches (NP), Patch Density (PD)
	Edge (8)	Total Edge (TE), Edge Density (ED)
	Shape (20)	Median of Perimeter-Area Ratio (PARA_MD), Median of Shape Index (SHAPE_MD), Median of Fractal Dimension Index (FRAC_MD), Median of Related Circumscribing Circle (CIRCLE_MD), Median of Contiguity Index (CONTIG_MD)
	Contagion/Interspersion (20)	Landscape Division Index (DIVISION), Splitting Index (SPLIT), Interspersion Juxtaposition Index (IJI), Landscape Shape Index (LSI), Patch Cohesion Index (COHESION)
Landscape Level Metrics (32)	Area (6)	Total Area (CA), Largest Patch Index (LPI), Median of Patch Area (AREA_MD), Median of Radius of Gyration (GYRATE_MD), Number of Patches (NP), Patch Density (PD)
	Edge (2)	Total Edge (TE), Edge Density (ED)
	Shape (6)	Perimeter-Area Fractal Dimension (PAFRAC), Median of Perimeter-Area Ratio (PARA_MD), Median of Shape Index (SHAPE_MD), Median of Fractal Dimension Index (FRAC_MD), Median of Related Circumscribing Circle (CIRCLE_MD), Median of Contiguity Index (CONTIG_MD)
	Contagion/Interspersion (10)	Landscape Division Index (DIVISION), Splitting Index (SPLIT), Effective Mesh Size (MESH), Interspersion Juxtaposition Index (IJI), Landscape Shape Index (LSI), Patch Cohesion Index (COHESION), Contagion (CONTAG), Proportion of Like Adjacencies (PLADJ), Aggregation Index (AI), Median of Euclidean Nearest Neighbor Distance (ENN_MD)
	Diversity (8)	Patch Richness (PR), Patch Richness Density (PRD), Shannon's Diversity Index (SHDI), Simpson's Diversity Index (SIDI), Modified Simpson's Diversity Index (MSIDI), Shannon's Evenness Index (SIEI), Simpson's Evenness Index (SIEI), Modified Simpson's Evenness Index (MSIEI)
Climate (24)	Precipitation (12)	Monthly Precipitation, Seasonal Average Precipitation
	Temperature (12)	Monthly Temperature, Seasonal Average Temperature
Topography (2)		Elevation, Slope
Soil (6)		Soil Storage, the Presence of Hydrologic Soil Groups[b] A, B, C, D, C/D, B/D
Population (2)		Population, Population Density

[a] The number in the parenthesis indicates the number of variables in the group.
[b] If a soil was placed in HSG D because of a high-water table, it might be assigned to a dual hydrologic group such as A/D, B/D, or C/D. The first letter of the pair represented the soil's group if drained and the second letter, D, represented the natural drainage condition.

rate, a slow infiltration rate, and a very slow infiltration rate, respectively.

2.3. Data analysis

In this study, we applied Random forest (RF) regression to build the predictive models of stream water quality. SHapley Additive exPlanations (SHAP) feature importance from RF models was used to explore whether urban development pattern was the dominant factor in determining stream water quality among all the catchment characteristics. Geographically Weighted Regression (GWR) was then applied to understand the spatial variation of the relationships between urban development patterns and pollutant concentrations. Finally, the trained RF models were employed to predict stream water quality under four scenarios of different urban development patterns.

2.3.1. Random Forest (RF) regression and the corresponding SHapley Additive exPlanations (SHAP)

RF regression was used to train models to quantify the nonlinear relationships between explanatory variables and stream water quality. It was further applied to scenario predictions of pollutant concentrations in accordance with different urban development patterns. RF is an ensemble learning method that consists of a large number of individual decision trees. Random samples are taken with replacement, and a random subset of features are used to generate each regression decision tree. A prediction is made by averaging the results of all the regression trees (Breiman, 2001). We selected RF for this study because it handles high non-linearity between independent variables well and is robust to outliers, which is suitable for our dataset (Kho, 2018).

To guarantee the generalization of the predictive models, 90% of the samples were used to train the models, and the remaining sample was used to test the models' performance. Performance metrics included Mean Square Error (MSE) and R^2, which are commonly used metrics in water quality prediction study (Lek et al., 1999; Wang et al., 2019), and Nash-Sutcliffe model coefficient (NSE), which is used to assess the predictive performance of hydrological models (Nash and Sutcliffe, 1970). Prior to RF model training process, all the independent variables were standardized to transform data into more standard, normally distributed data. Natural log was taken to all the pollutant concentration values. Very large outliers (TP > 4 mg/l or *E. coli* > 10,000 MPN/100 ml or NO_3^--N > 17 mg/l), which were approximately three standard deviations above the mean concentration, were removed. Ten-fold cross validation was employed to tune the hyperparameters including the number of trees (n_estimators), the maximum depth of the tree (max_depth), the minimum number of samples required to split a node (min_sample_split), and the maximum number of features to look for the best split (max_features) using a grid search fashion. Random forest regression was implemented in Python 3.0 “scikit-learn” package.

To understand which variables dominate RF models and their associations with stream water quality, SHAP feature importance was calculated for each RF model. SHAP is a game theoretic approach that can interpret any machine learning model. The goal is to explain the prediction of an instance by computing the contribution of each feature to the prediction. Global feature importance is calculated by averaging the absolute Shapley values per feature across all the samples (Lundberg and Lee, 2017; Molnar, 2019). From SHAP summary plot, we know the relationship between each catchment characteristic and the prediction of pollutant concentration. SHAP results were derived in Python 3.0 “shap” package.

2.3.2. Geographically weighted regression (GWR)

GWR was applied to investigate the spatially varying associations between important urban development pattern metrics and stream water quality. GWR was used because it improves the model capacity to reveal the local cause of stream water pollution in this large study area. The independent variables included in the GWR models were land cover percentages, topography, population, soil storage, seasonal mean temperature and total precipitation, and important landscape metrics derived by SHAP of RF models. The important landscape metrics

were COHESION, DIVISION, LPI, SPLIT, IJI, ED, PD of urban development areas and planted areas, FRAC_MD, CONTIG_MD, AREA_MD, SPLIT of forest areas, and ED of developed open areas.

GWR allows linear predictors to be a function of spatial coordinates (u, v), as represented in Eq. (1). In this equation, y_i is the pollutant concentration, $x_{k,i}$ is the covariate vector, and β_k is the corresponding vector coefficient. GWR assumes that the contribution of each sample to the local regression model is weighed according to its proximity to the local sample point. A common choice of weighting function is the Gaussian curve, as shown in Eq. (2), where d_{ij} is the distance between observation point i and the realization point j, and the bandwidth b is the parameter to be determined. An adaptive kernel bandwidth was employed in this study in accordance with the judgement of AIC. GWR was implemented in "spgrw" package in R.

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k} \beta_k(u_i, v_i) x_{k,i} + \varepsilon_i \quad (1)$$

$$\omega_{i,j} = \exp\left(-\frac{{d_{ij}}^2}{2b^2}\right) \quad (2)$$

2.4. Scenario design

To understand the effects of urban developed density and configuration on stream water quality, we created four alternative urban development scenarios in the upstream area of The Woodland, TX, and predicted their pollutant concentrations of NO_3^--N, TP, and *E. coli* in both dry and wet seasons (Fig. 2). The Woodlands are well-known for Ian McHarg's ecological planning approach (McHarg and Sutton, 1975). The current development condition was chosen as the baseline scenario, where 33.6% of the area (24 km^2) was developed into urban areas. Low-density development is the major development type in the current condition. The boundary of the scenario site is the Bear Branch-Panther Branch sub-watershed boundary with the HUC12 ID 120401020211.

The alternative scenarios included four extreme development scenarios where developed areas were extremely scattered or aggregated: high-density aggregated development, high-density sprawl development, medium-density aggregated development, and medium-density sprawl development (Fig. 2). We applied two criteria to create the four development scenarios. First, the total impervious surface area was the same as the baseline scenario. To quantify impervious surface in urban areas for each density, we used the median value of the impervious surface percentage from NLCD description, which were 35%, 65%, and 90% for low density, medium density, and high-density developments, respectively (Yang and Li, 2011). The impervious surface areas added up to be 16.4 km^2 in all scenarios. Second, all of the existing land cover types in the baseline scenario—including water, forest, grassland, planted, and wetland—stayed the same. The reduced urban areas in the four alternative scenarios were changed to forest areas that represent undeveloped conditions. To approximate the maximum degree of aggregated/sprawl development, we manually chose locations of high/medium-density development that had changed to forest areas in ArcGIS 10.5. When selecting the locations of urban developed areas, we maintained the historical trend of The Woodland development, that is to develop from downstream to upstream along the Panther Creek to the north (Yang and Li, 2011). Therefore, the aggregated development was located close to the downstream area. This trend was also in accordance with McHarg's ecological planning approach to develop residential areas on land with low soil permeability (McHarg, 1996).

The key difference in each scenario was urban development pattern, as presented in Table S4. Compared to the two sprawled scenarios, developed areas were clumped into larger patches with simpler shapes, and were more physically connected in the two aggregated scenarios. The two aggregated scenarios were thus characterized by higher LPI, COHESION, lower ED, LSI, and shape complexity.

3. Results

3.1. RF model accuracy and the important catchment characteristics in influencing stream water quality

In the RF regression model, the variations and trends of all pollutant concentrations were well captured in the test set; however, the extreme values were not well predicted (Fig. S2). The very low concentrations tended to be overestimated and the very high concentrations tended to be underestimated. In the wet season, the R^2 of the test set was 0.64, 0.46, and 0.64 for the TP, *E. coli*, and NO_3^- N RF models, respectively. The NSE of the test set was 0.61, 0.45, and 0.69 for TP, *E. coli*, and NO_3^- N wet season models, respectively (Table 2).

Urban development pattern outweighed other land use patterns in affecting TP concentration. We present the ten most important variables and visualize their associations with pollutant concentrations in Fig. 3. The twenty most important variables and their directions of influence are shown in Table S5 for further reference. COHESION, DIVISION, LPI, and SPLIT of developed areas were found to be the most important urban development pattern metrics in affecting TP concentration (Fig. 3). A lower DIVISION of developed areas represented the condition where the proportion of developed areas increased, and the developed patches increased in size. TP concentration was likely to increase in this situation. When developed patches became more physically connected, as indicated by a higher COHESION, TP concentration also increased. Large patch sizes of developed areas, as indicated by a larger LPI, led to high TP concentration. The decrease of SPLIT also implied an increase in TP concentration, where developed patches increased in area and became less subdivided. When developed areas were more equally interspersed with other land cover types, as indicated by a higher IJI, TP concentration decreased in dry seasons (Table S5). The important catchment characteristics affecting TP concentration in dry and wet seasons were similar, with most of the urban development pattern effects greater in wet seasons.

The important variables influencing *E. coli* concentration were associated with more variable categories, including patterns of urban developed areas, developed open areas, planted areas, and some landscape level metrics. It indicated a potentially more complex mechanism (Table S5). Similar to TP results, higher *E. coli* concentration was associated with higher LPI, higher COHESION, lower DIVISION, and lower SPLIT of developed areas (Fig. 3). Higher proportion of developed areas with larger patch sizes and more connection with other developed patches contributed to higher *E. coli* concentration. In addition, higher ED of developed areas, developed open areas, and planted areas, which implied more complex edges of these patch types, were associated with higher *E. coli* concentration. At the landscape level, the median of GYRATE and the median of FRAC had positive associations with *E. coli* concentration, while IJI had a negative association with *E. coli* concentration in the wet season. Therefore, large patch size, complex patch shape, and uneven distribution of adjacencies among patch types led to high *E. coli* concentration. Similar to TP results, urban development pattern had larger effects on *E. coli* concentration in wet seasons than in dry seasons.

Both urban development and forest patterns were importantly associated with NO_3^--N concentration, with forest pattern more important in dry seasons than in wet seasons. Similar to TP and *E. coli* results, higher COHESION, higher LPI, lower SLPIT, and lower DIVISION of developed areas were associated with higher NO_3^--N concentration (Fig. 3). The higher connectedness and larger patch sizes of developed areas contributed to higher NO_3^--N concentration. In dry seasons, the situation where developed areas were more equally adjacent with other land covers (indicated by a higher IJI) helped reduce NO_3^--N concentration. A lower median of CONTIG, lower ED, and higher SPLIT of forest

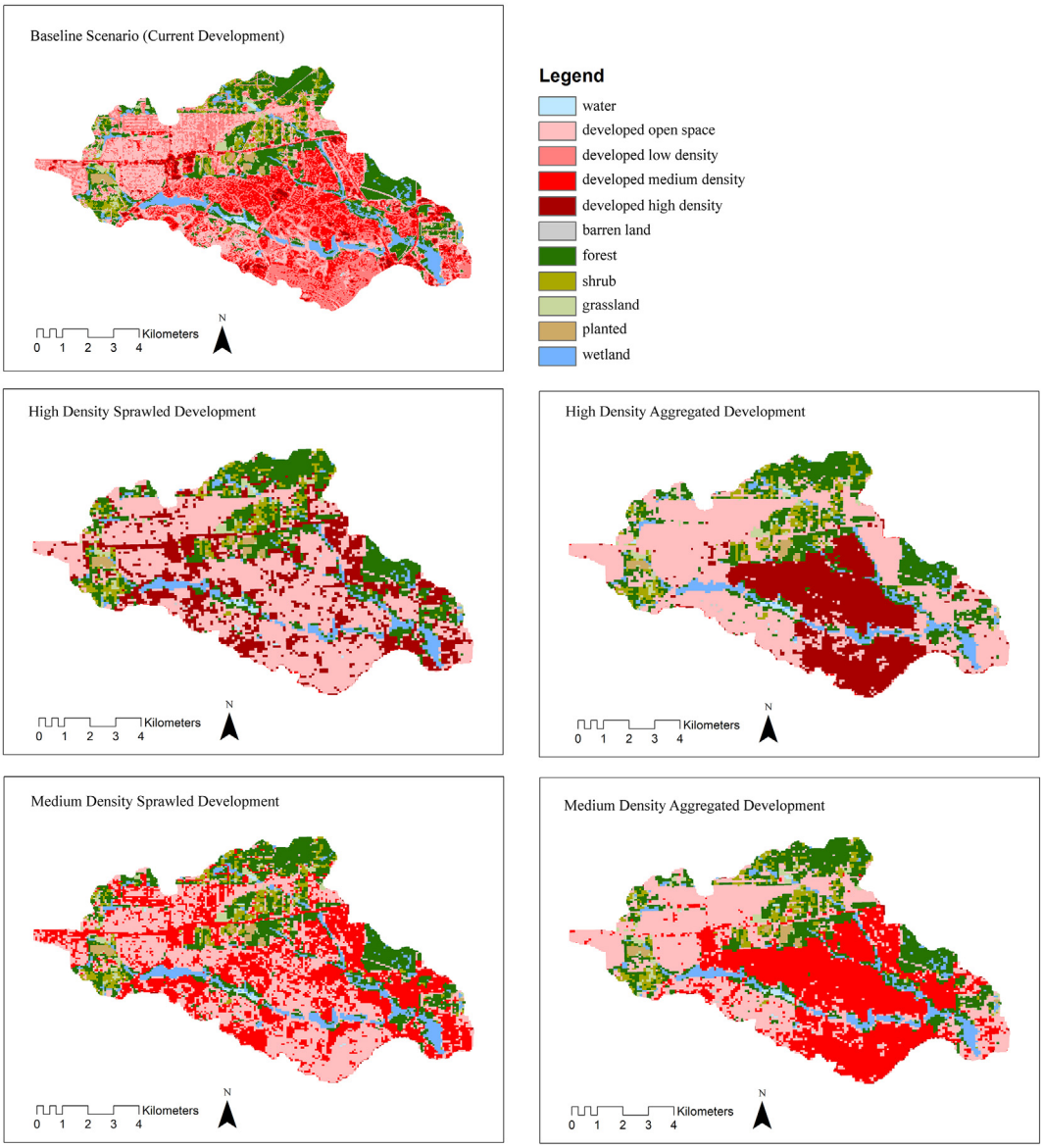


Fig. 2. Scenario maps.

area also led to higher NO_3^--N concentration in dry seasons, which suggests that intact forest areas with complex edges could potentially reduce NO_3^--N concentration. In wet seasons, percentages of both urban developed areas and developed open areas were positively associated with NO_3^--N concentration. A higher ED of developed open areas also contributed to higher NO_3^--N concentration. It was also discovered that urban development pattern had a larger effect NO_3^--N concentration in wet seasons than in dry seasons.

In summary, urban development pattern was more important than the percentage of urban developed areas and other land use patterns in determining the pollutant concentration of TP, *E. coli*, and NO_3^--N. The most important aspects were patch size and connectedness of urban developed areas. With respect to other catchment characteristics, high temperature was important in driving the increase of pollutant concentration. Areas with high soil storage levels were associated with TP and *E. coli* pollution.

Table 2
Predicting accuracy of RF regression.

		Train set				Test set			
		Correlation	R^2	MSE	NSE	Correlation	R^2	MSE	NSE
TP	Wet season	0.98	0.96	0.16	0.91	0.80	0.64	0.79	0.61
	Dry season	0.98	0.96	0.15	0.91	0.76	0.58	0.85	0.56
E. coli	Wet season	0.98	0.96	0.58	0.90	0.68	0.46	3.04	0.45
	Dry season	0.98	0.96	0.35	0.90	0.71	0.50	2.06	0.49
NO_3^--N	Wet season	0.97	0.94	0.35	0.91	0.83	0.69	1.37	0.64
	Dry season	0.97	0.94	0.29	0.91	0.82	0.68	1.56	0.64

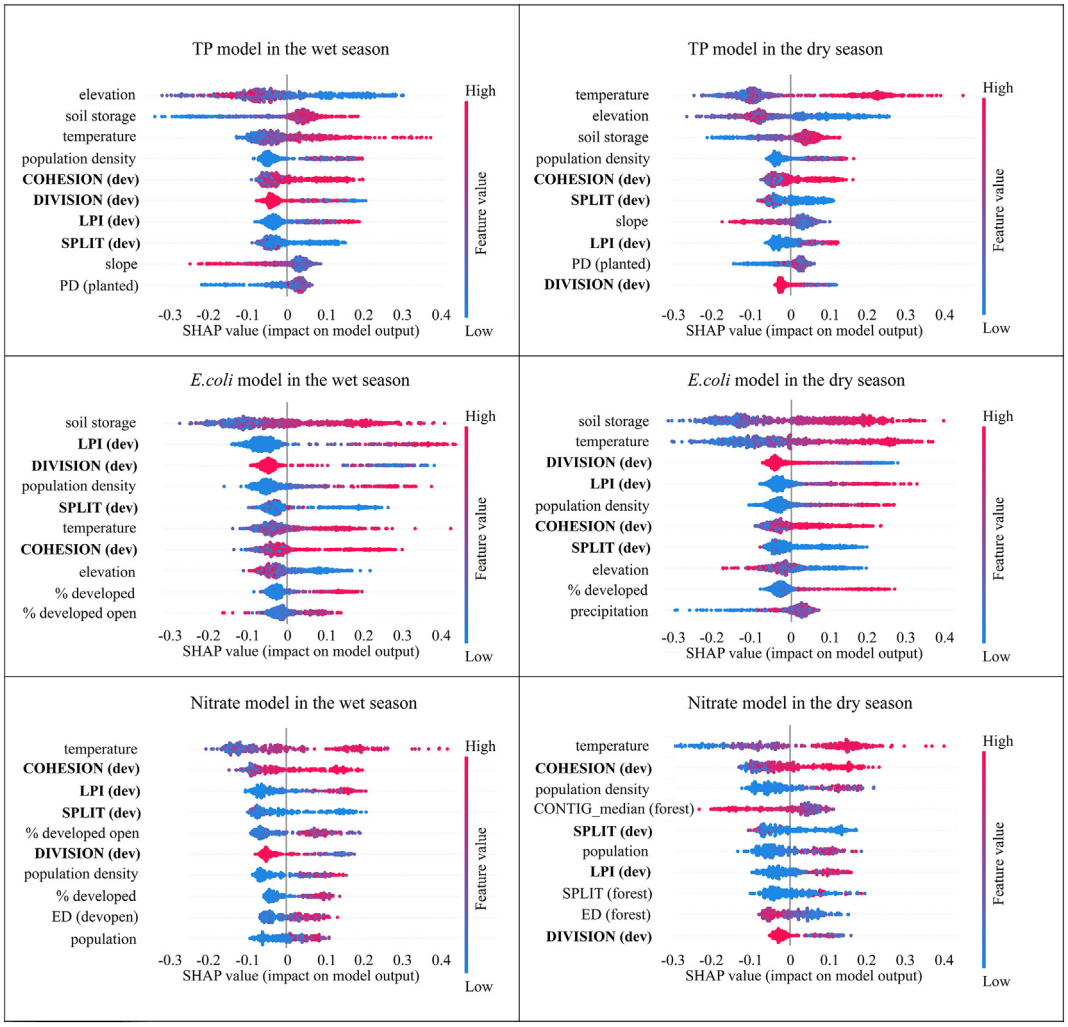


Fig. 3. SHAP results of feature importance from RF regression.

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

In this study, TP GWR performed better in coastal areas such as the Neches Basin and the Galveston Bay-San Jacinto Basin, with the R^2 higher than 0.45 (Fig. S1). The performance of the *E. coli* GWR was also better in coastal areas. The wet season *E. coli* model had R^2 above 0.4 in the coastal areas, while the dry season *E. coli* model had R^2 above 0.31 with a lower variation. Because of the smaller spatial extent, the variation of NO_3^--N GWR models was much smaller than that of TP and *E. coli* models. The NO_3^--N GWR model performed better in the west basins (the Nueces-Southwestern Texas Coastal and the Lower Colorado-San Bernard Coastal Basin) than the east basins (the Galveston Bay-San Jacinto Basin, the Trinity Basin, and the Sabine Basin). Lastly, the NO_3^--N GWR model performed better in the dry season than in the wet season, with the global R^2 0.54 and 0.48, respectively.

We selected COHESION, IJI, and LPI of urban developed areas to explore their spatially varying relationship with pollutant concentrations in Fig. 4. The three metrics were all important as indicated by the SHAP results, and distinct in their conceptual meanings. Among the most important urban development metrics, COHESION exerted a greater positive effect on TP concentration in the southern portion of the study area, which included the Nueces-Southwestern Texas Coastal Basin, the Central Texas Coastal Basin, the Lower Colorado-San Bernard Coastal Basin, and the Lower Brazos Basin (Fig. 4). When developed areas were more proportionally interspersed with other land cover types (indicated by a higher IJI), TP concentration in the east portion of the study area (the Trinity Basin, the Neches Basin, and the Galveston Bay-San Jacinto Basin) decreased more than the west portion. Large patch sizes of developed areas (indicated by a higher LPI) were shown to have a larger effect on TP concentration in the south and west parts of the study area. These areas were mainly crop, pasture, and forest areas.

Similar to TP models, highly aggregated urban developed areas (high COHESION of developed areas) had larger effects on *E. coli* concentration in the southern part of the study area. In the very northern part like the Trinity Basin, COHESION had a negative influence on *E. coli* concentration. Large patch sizes of urban developed areas had greater positive effects on *E. coli* concentration in coastal basins, including the Central Texas Coastal Basin, the Galveston Bay-San Jacinto Basin, the Nueces-Southwestern Texas Coastal Basin, and the east part of the Lower Colorado-San Bernard Coastal Basin. IJI of developed areas had a negative effect on *E. coli* concentration in the southeastern coastal areas. The effects trended towards positive in the northwest part of the study area.

NO_3^--N concentration was more positively affected by the connection (COHESION) and large patch size (LPI) of urban developed area in the west coast of the study area, including the Central Texas Coastal Basin and the west part of the Lower Brazos Basin. NO_3^--N concentration was more dependent on IJI of developed areas in the east coast, which

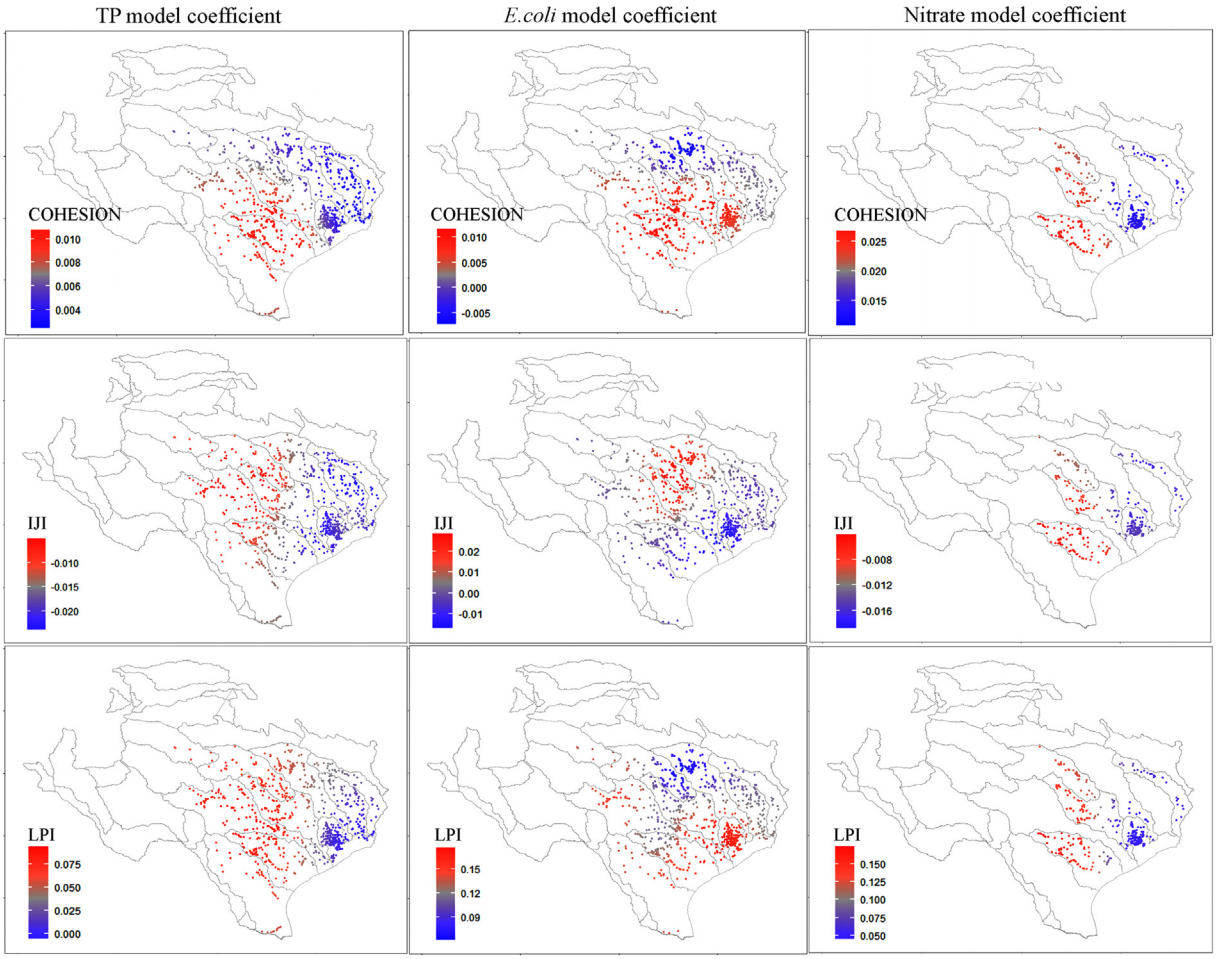


Fig. 4. GWR model coefficients of urban development pattern effects in the wet season.

encompassed the Galveston Bay-San Jacinto Basin, the south part of the Lower Colorado-San Bernard Coastal Basin, and the Sabine Basin. In these areas, when developed areas were more proportionally interspersed with other land cover types (indicated by a higher IJI), NO_3^--N concentration was more likely to decrease. We discuss the mechanism that can potentially drive the spatial variation in the effects of urban development pattern in Section 4.2.

3.3. Stream water quality prediction under alternative planning scenarios

The prediction results of the alternative planning scenarios suggest that a high-density aggregated development pattern was advantageous in reducing TP and NO_3^--N concentrations (Table 3). All high-density and medium-density compact developments had less than half the concentration of all pollutants compared to the current development, indicating the benefits of small footprint urban areas. The benefit of small footprint urban areas was most significant in reducing NO_3^--N concentrations.

Aggregated development in both high and medium density scenarios had lower TP and NO_3^--N concentrations when compared to sprawl development of the same density. However, aggregated development contributes to higher *E. coli* concentrations than sprawl development of the same density in wet seasons. Overall, the most recommended urban development pattern for stream water quality protection was high-density aggregated development; though specific attention should be paid in areas with potential *E. coli* pollution to avoid very high density development. It was worth noting that the predicted values of NO_3^--N were comparable to the measured data at the TCEQ Station #16629, which was located close to the outlet of the basin, indicating the reliability of our predictive models.

Regarding ecological implication, NO_3^--N concentration higher than 1.5 mg/l, which was presented in the current sprawl development (baseline scenario), was considered as the hypereutrophic level of lakes. All the alternative compact urban development scenarios generated NO_3^--N < 0.5 mg/l, which were oligotrophic or mesotrophic trophic levels (*Nitrogen fact sheet*, 2016). Regarding TP concentration, scenarios other than high-density aggregated and medium-density aggregated development were all associated with the hypereutrophic level, which is 0.1 mg/l (Yang et al., 2008). Therefore, aggregated development was proven to be beneficial to lotic ecosystems by avoiding potential eutrophication.

4. Discussion

4.1. The complexity of urban development pattern impact on stream water quality

DIVISION and SPLIT of developed areas were proven to be important in influencing all the pollutant concentrations by the SHAP results. They both represent patch subdivision calculated by the relationship between the single patch area and the total landscape area (Table S3). They are potentially useful in characterizing residential urban form given how they are subdivided into blocks (Bach et al., 2013). The difference between the two metrics is that DIVISION departs from the maximum limit slowly when patches become more scattered, and approaches the minimum limit quickly if very large patch areas are present and they become less subdivided. Therefore, low DIVISION of

Table 3
Scenario prediction results of pollutant concentration.

		Current development	High-density aggregated development	High-density sprawl development	Medium-density aggregated development	Medium-density sprawl development
TP	Wet season	0.22 (0.55[a])	0.10[b]	0.13	0.11	0.14
	Dry season	0.28 (0.11)	0.09	0.15	0.10	0.13
E. coli	Wet season	90.18	33.40	17.52	85.71	67.49
	Dry season	41.05	22.02	41.79	31.46	44.28
NO_3^--N	Wet season	1.95 (2.92)	0.10	0.18	0.19	0.27
	Dry season	1.2 (1.03)	0.15	0.22	0.17	0.36

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

urban developed areas is more likely to lead to higher pollutant concentration compared to low SPLIT of urban developed areas. On the contrary, a high SPLIT value is more likely to associated with clean water compared to a high DIVISION value of urban developed areas (Fig. 3).

COHESION is sensitive to the aggregation and physical connectedness of the focal class. Large and aggregated urban areas, as indicated by a high contiguity index (CONTIG) or contagion index (CONTAG), were associated with poor stream water quality in some studies (Lee et al., 2009; Lv et al., 2015; Shi et al., 2017). However, since greater interspersion and increases in the number of urban patches may accelerate soil erosion and sediment exportation (Shi et al., 2013), we argue that although an intact urban area with large impervious surfaces can result in the deterioration of water quality (Alberti et al., 2007; Lee et al., 2009), the same area of impervious surface can lead to worse stream water quality with greater dispersion, as verified in our scenario prediction results. The SHAP results also indicated that high COHESION did not necessarily lead to high pollutant concentration (Fig. 3).

It is worth noting here that the effect of the urban development pattern on water quality should always be interpreted with caution due to the collinearity between urban development pattern metrics and the urban area percentage. Therefore, the effect of urban development pattern on stream water quality derived by some statistical models can sometimes be caused by the percentage of urban developed areas. The conceptually similar urban development metrics cannot replace each other, nor can they be replaced by the percentage of urban developed area. As indicated in Fig. S3, a low percentage of developed area does not necessarily mean low COHESION or high IJI. Similarly, a high percentage of developed area does not necessarily mean low DIVISION. For example, high percentages of developed areas can be very scattered, and the COHESION of urban developed areas can be low and the DIVISION can be high in this situation.

We discovered that soil storage is a strong factor that positively affected pollutant concentration according to SHAP feature importance results. This is because soil, which acts as a primary sink and source of terrestrial contaminants, can affect stream water quality through subsurface and soil water (Liu et al., 2017; Taka et al., 2016). Therefore, soil with a deep storage layer might be associated with a larger terrestrial source and a higher risk or holding contaminants to influence surface water quality, the mechanism of which is worth future investigation. Air temperature is also important in affecting water temperature, and thus affects water chemistry. Higher temperatures were found to be correlated with higher nutrient concentrations, which could be a more serious issue as the climate warms (Baron et al., 2009). Therefore, warmer areas with larger soil storage depth could be potential hydrologically sensitive areas that generate more pollution, and should be cautiously planned with urban development.

4.2. Interpretation of spatiotemporal non-stationary land-water relationships

In this study, the effects of urban development pattern on stream water quality were greater in the wet season than in the dry season according to the SHAP results. This conclusion agrees with existing literature that the effects of composition and configuration of land cover are more evident during the rainy season (Bu et al., 2014; Shi et al., 2017). This might be because in urban areas, the flushing effect during the wet season outweighs the dilution effect in the study region (Liu et al., 2017). Under future climate change conditions, urban development pattern might have a more complicated effect on stream water quality with changing precipitation. Precipitation predictions associated with climate change vary by location and often include a lot of uncertainty, up to the point that the sign of change is uncertain (Wuebbles and Hayhoe, 2004).

Moreover, the influence that urban development pattern exerted on stream water quality had high spatial variations according to the GWR results, which might be attributed to different pollutant sources from different basins. The relationship between LPI of developed areas and TP concentration was weaker in the coastal urban area than the inland agricultural area. We argue that in highly urbanized areas, a larger LPI of developed areas corresponds to aggregated development with fewer urban patches, and in this situation, it is less likely to cause pollution because of the smaller urban footprint of the aggregated development. However, in the agricultural area, the relationship between the LPI of developed areas and the TP concentration changed to highly positive (Fig. 4). In these watersheds, there were not many urban patches, and a large LPI of developed areas simply implied larger urban core areas and larger total impervious areas, which would contribute to the increasing pollutant concentration.

COHESION of developed areas has shown the most significantly positive association with *E. coli* concentration in the Galveston Bay-San Jacinto Basin, which is a highly urbanized area. This might be because aggregated development led to more *E. coli* pollution, which aligned with our scenario prediction results. However, TP concentration showed relatively weak dependency on the COHESION of developed areas in the same highly urbanized area. Since the impact of large patch size (indicated by LPI) and more aggregation (indicated by COHESION) of developed areas on TP concentration was not large in this highly urbanized area, we argue that the negative effect of urban sprawl might play a more important role in these areas. The potentially different mechanisms on how urban sprawl affects TP and *E. coli* pollution warrant future investigation.

Furthermore, the IJI of developed area had a higher negative influence on TP and *E. coli* concentration primarily in the coastal urban watersheds (Fig. 4). In these watersheds, low IJI of developed areas was usually associated with low density development, which was the most common development type in this region. If developed areas were mostly adjacent to developed open areas in low-density development, the watersheds typically had a low IJI of developed area and high TP and *E. coli* concentrations. This phenomenon might be attributed to the application of phosphorus-based fertilizers on lawns in low-density residential areas (Wilson, 2015).

4.3. Planning implications based on urban development pattern metrics

In this section, we discuss urban planning implications based on urban development pattern metrics using sample watersheds in the

study region. Two pair-wise comparisons of land cover maps with similar percentages of developed areas but different TP and *E. coli* concentrations were shown in Fig. 5. The watershed #12083 (Fig. 5-a) was identified as having more aggregated development, with a relatively integral natural core in the west. The IJI of developed area in this watershed was larger because the developed area was more equally adjacent to other land patch types. The watershed #11155 (Fig. 5-b) had low density development with scattered developed open areas. The IJI in this watershed was small because the developed area was largely adjacent to the developed open area only. Greatly interspersed land uses accelerated soil erosions and caused the increase in pollutants (Shi et al., 2013; Sun et al., 2013). Specifically, for watershed # 11155, high TP and *E. coli* concentrations could have resulted from landscape gardens in the developed open areas and the greater extent of road surface area associated with detached houses (Goonetilleke et al., 2005).

The comparisons between watersheds in Fig. 5-c and -d showed how edge complexity potentially affected pollutant concentration. Different edge complexity was associated with different drainage connections and road systems that influence runoff velocity, pollutant travel distance, and time of transport (Liu et al., 2012). A higher ED of developed areas in the watershed #17406 (Fig. 5-c) was found to be associated with higher TP and *E. coli* concentration than in watershed #11405 (Fig. 5-d), given the similar percentage of developed areas. The complex shape and sprawled development of watershed #17406 led to more interspersed land uses and higher road density that might generate more nonpoint source pollutants. It also degraded the structure of natural systems that were important for filtering pollutants (Lee et al., 2009).

Urban development pattern metrics are related to percentage, aggregation, patch shape, and connectivity of developed areas, and thus can represent characteristics of urban sprawl like low-density development, leapfrog development over vacant lands, and decentralization (Riitters et al., 1995; Gordon and Richardson, 1996; Ewing, n.d.; Bhatta, 2010). We argue that urban sprawl has a direct relationship to stream water quality, as it affects pollutant generation, build-up, and wash off by altering the structure of urban forms and the surrounding natural areas (Goonetilleke et al., 2005; Liu et al., 2012). We suggest that urban form for stream water quality protection should avoid:

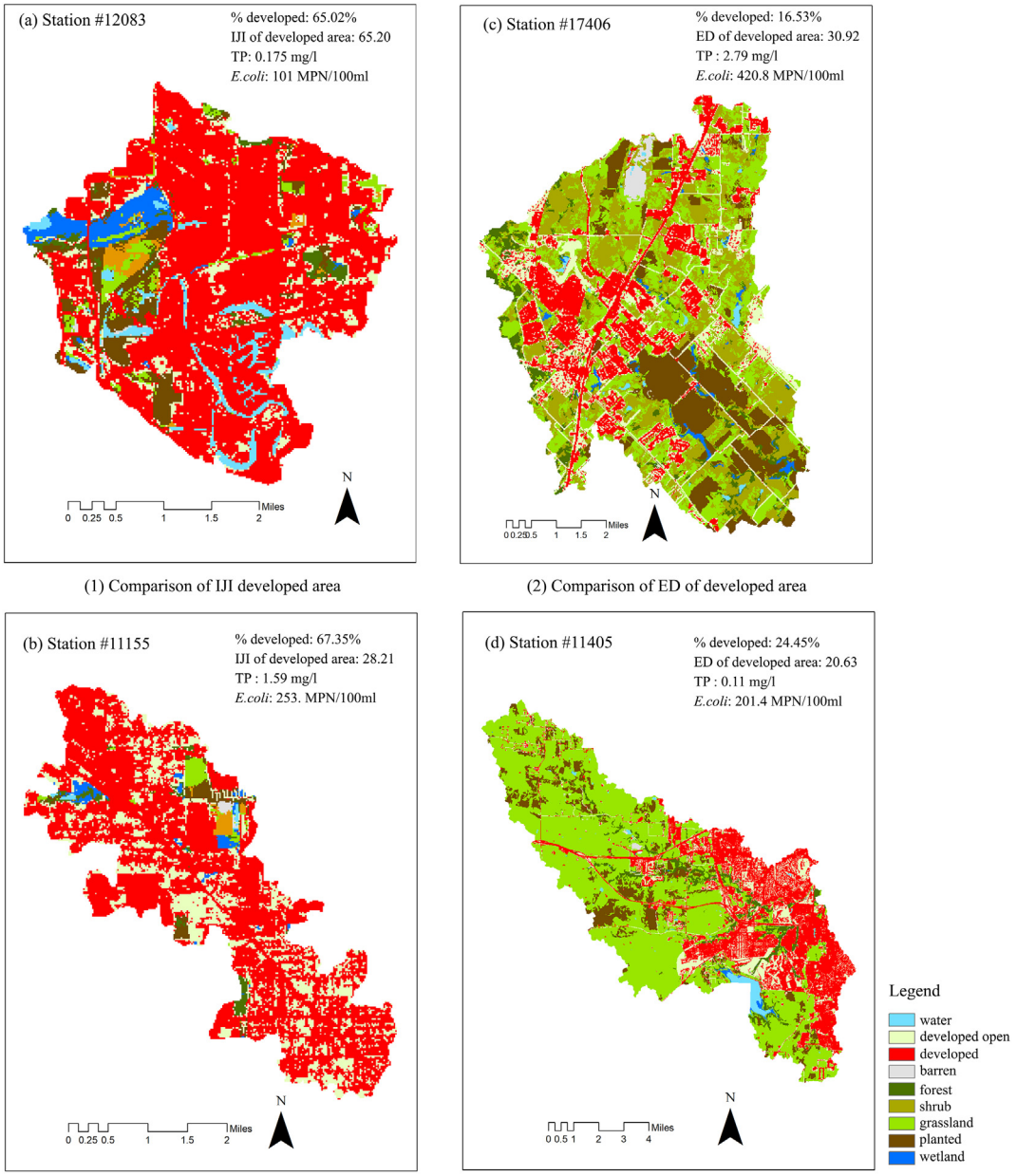


Fig. 5. Examples of watersheds with the similar percentage of developed areas but different urban development pattern metrics and pollutant concentrations.

(1) sprawl of low-density development with large lawn areas and complex road systems (Fig. 5-b) and (2) complexly shaped and scattered patches of urban areas that are likely to have complicated road systems and destroy integral natural areas (Fig. 5-d). To further justify this recommendation, future studies should measure more planning indicators related to urban sprawl, such as the layout of residential areas and the fragmentation of natural areas, to investigate how they affect stream water quality.

4.4. The advantages and limitations of applying machine learning in stream water quality scenario prediction

We integrated urban developed pattern factors and achieved the state-of-the-art stream water quality predicting accuracy. In this study, test set correlation between actual values and predicted values was 0.83 and 0.82 for NO_3^--N in wet and dry seasons, respectively. This result is comparable with Lek and others' research where test set model correlation was 0.845 and 0.832 for total nitrogen and inorganic nitrogen using an artificial neuron network (ANN) (Lek et al., 1999). In a more recent research, NO_3^--N concentration was modeled with ANN, and the test set R^2 was 0.60 (Mirzaei et al., 2020), while our RF model test set R^2 was 0.69 and 0.68 for NO_3^--N in the wet and dry season, respectively. Our test set model NSE was 0.61, 0.45, and 0.64 for TP, *E. coli*, and NO_3^--N in wet seasons. In a SWAT review study, the average NSE for monthly TP and NO_3^--N model in the calibration stage was 0.68 and 0.54, respectively (Gassman et al., 2007). However, given the noisy nature of this large-scale dataset, there was still risk that the models did not fit the data well, which might undermine the results of the scenario prediction. If we could obtain a larger sample size or apply transfer learning, this issue could potentially be addressed. In addition, adding more influential factors could also improve model fitting. For example, there might be other factors affecting *E. coli* concentration besides climate and land cover characteristics (Chelsea Nagy et al., 2012). As the advancement of machine learning algorithms, it is promising for future study to predict contaminant concentration under the future land use planning scenarios if the machine learning model fits data well.

As mentioned in previous studies, a key gap in water quality studies has been a lack of consideration of cross effects between explanatory variables, such as the cross-correlation between land covers and the cross-correlation between land cover and climate in influencing stream water quality (Li et al., 2015; Hwang et al., 2016; Lintern et al., 2017). Machine learning can make use of all cross effects between variables and improve model predicting accuracy, which is an advantage over traditional statistical models. For example, it is likely that climatic factors exhibited interaction effects with urban development pattern and other environmental variables on stream water quality, and the predicting accuracy can thus be improved. Another advantage is that RF regression can accommodate high-dimensional factors to improve water quality prediction accuracy, e.g., the inclusion of a monthly climatic variable in this study. It is applicable for future research to integrate a set of planning factors and/or extreme climate conditions to draw management implications of interest.

The major limitation of this study was that some catchment characteristics were excluded because they were not readily available. Such variables included point source pollution, animal products, wastewater treatment plants, and so on (Chen and Lu, 2014; Zhou et al., 2016). These data were excluded because they had much lower resolution than the variables included in this study. Future machine learning predictions of stream water quality should take these important aspects into consideration, if applicable, in order to obtain more unbiased models. Another limitation was the selection of appropriate variables. In this study, we conducted trials of variable selection in the RF regression using mutual info regression (Kraskov et al., 2011), but the RF regression accuracy did not significantly improve. Future studies should also try other feature engineering algorithms, such as recursive feature elimination. Additionally, we only included one year of data because land cover did not change drastically in this region, nor did the urban development pattern. Future studies should adopt a longitudinal perspective when the change of the urban development pattern is of interest. Lastly, this study focused primarily on urban development pattern. It would be meaningful to compare urban and agricultural areas with respect to how landscape-level pattern influences stream water quality.

5. Conclusion

Urban development pattern was found to significantly influence stream TP, NO_3^--N, and *E. coli* concentrations in the Texas Gulf Region, with the relationships among them varying according to season and location. Largest Patch Index (LPI), Patch Cohesion Index (COHESION), Splitting Index (SPLIT), and Landscape Division Index (DIVISION) of developed areas were the most efficient urban development pattern metrics associated with stream water quality. Interspersion and Juxtaposition Index (IJI) and Edge Density (ED) of developed areas were also important for specific pollutants and seasons. The influence of urban development pattern on stream water quality was larger in wet seasons than in dry seasons. According to the GWR results, the effects of urban development pattern were different according to geographical locations and pollutant categories because of the different pollutant sources and transportation processes.

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

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.144057.

CRediT authorship contribution statement

Runzi Wang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jun-Hyun Kim:** Methodology, Resources, Writing – review & editing. **Ming-Han Li:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Ai, L., Shi, Z.H., Yin, W., Huang, X., 2015. Spatial and seasonal patterns in stream water contamination across mountainous watersheds: linkage with landscape characteristics. J. Hydrol. 523, 398–408. https://doi.org/10.1016/j.jhydrol.2015.01.082.

Alberti, M., Booth, D., Hill, K., Coburn, B., Avolio, C., Coe, S., Spirandelli, D., 2007. The impact of urban patterns on aquatic ecosystems: an empirical analysis in Puget lowland subbasins. Landsc. Urban Plan. 80 (4), 345–361.

Avila, R., Horn, B., Moriarty, E., Hodson, R., Moltchanova, E., 2018. Evaluating statistical model performance in water quality prediction. J. Environ. Manag. 206, 910–919. https://doi.org/10.1016/j.jenvman.2017.11.049.

Bach, P.M., Deletic, A., Urich, C., Sitzenfrei, R., Kleidorfer, M., Rauch, W., McCarthy, D.T., 2013. Modelling interactions between lot-scale decentralised water infrastructure and urban form–a case study on infiltration systems. Water Resour. Manag. 27 (14), 4845–4863.

Baron, J.S., Schmidt, T.M., Hartman, M.D., 2009. Climate-induced changes in high elevation stream nitrate dynamics. Glob. Chang. Biol. 15 (7), 1777–1789.
Bhatta, B., 2010. Urban growth and sprawl. Analysis of Urban Growth and Sprawl From Remote Sensing Data, pp. 1–16 https://doi.org/10.1007/978-3-642-05299-6_1.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Bu, H., Meng, W., Zhang, Y., Wan, J., 2014. Relationships between land use patterns and water quality in the Taizi River basin, China. Ecol. Indic. 41, 187–197. https://doi.org/10.1016/j.ecolind.2014.02.003.
Carey, R.O., Migliaccio, K.W., Li, Y., Schaffer, B., Kiker, G.A., Brown, M.T., 2011. Land use disturbance indicators and water quality variability in the Biscayne Bay Watershed, Florida. Ecol. Indic. 11 (5), 1093–1104.
Castrillo, M., García, Á.L., 2020. Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. Water Res. 172, 115490.
Chelsea Nagy, R., Graeme Lockaby, B., Kalin, L., Anderson, C., 2012. Effects of urbanization on stream hydrology and water quality: the Florida Gulf Coast. Hydrol. Process. 26 (13), 2019–2030.
Chen, J., Lu, J., 2014. Effects of land use, topography and socio-economic factors on river water quality in a mountainous watershed with intensive agricultural production in East China. PLoS One 9 (8), e102714. https://doi.org/10.1371/journal.pone.0102714.
Chen, Q., Mei, K., Dahlgren, R.A., Wang, T., Gong, J., Zhang, M., 2016. Impacts of land use and population density on seasonal surface water quality using a modified geographically weighted regression. Sci. Total Environ. 572, 450–466.
Chermack, T.J., Swanson, R.A., 2008. Scenario planning: human resource development's strategic learning tool. Adv. Dev. Hum. Resour. 10 (2), 129–146.
Clément, F., Ruiz, J., Rodríguez, M.A., Blais, D., Campeau, S., 2017. Landscape diversity and forest edge density regulate stream water quality in agricultural catchments. Ecol. Indic. 72, 627–639. https://doi.org/10.1016/j.ecolind.2016.09.001.
Darst, B.F., Malecki, K.C., Engelman, C.D., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 19 (1), 65.
Del Monaco, N., 2017. Reducing Directly Connected Stormwater Infrastructure and the Associated Benefits. Doctoral dissertation, Rutgers University-Graduate School-New Brunswick.
Ding, J., Jiang, Y., Liu, Q., Hou, Z., Liao, J., Fu, L., Peng, Q., 2016. Influences of the land use pattern on water quality in low-order streams of the Dongjiang River basin, China: a multi-scale analysis. Sci. Total Environ. 551-552, 205–216. https://doi.org/10.1016/j.scitotenv.2016.01.162.
Ewing, R. H. (n.d.). Characteristics, causes, and effects of sprawl: a literature review. Urban Ecol., 519–535. doi:https://doi.org/10.1007/978-0-387-73412-5_34.
Fan, M., Shibata, H., 2015. Simulation of watershed hydrology and stream water quality under land use and climate change scenarios in Teshio River watershed, northern Japan. Ecol. Indic. 50, 79–89. https://doi.org/10.1016/j.ecolind.2014.11.003.
Forman, R.T., 2014. Land Mosaics: The Ecology of Landscapes and Regions (1995). Island Press, p. 217.
Gassman, P.W., Reyes, M.R., Green, C.H., Arnold, J.G., 2007. The soil and water assessment tool: historical development, applications, and future research directions. Trans. ASABE 50 (4), 1211–1250.
Giri, S., Qiu, Z., 2016. Understanding the relationship of land uses and water quality in Twenty First Century: a review. J. Environ. Manag. 173, 41–48. https://doi.org/10.1016/j.jenvman.2016.02.029.
Glińska-Lewczuk, K., Gołaś, I., Koc, J., Gotkowska-Płachta, A., Harnisz, M., Rochwerger, A., 2016. The impact of urban areas on the water quality gradient along a lowland river. Environ. Monit. Assess. 188 (11). https://doi.org/10.1007/s10661-016-5638-z.
Goonetilleke, A., Thomas, E., Ginn, S., Gilbert, D., 2005. Understanding the role of land use in urban stormwater quality management. J. Environ. Manag. 74 (1), 31–42. https://doi.org/10.1016/j.jenvman.2004.08.006.
Gordon, P., Richardson, H.W., 1996. Beyond polycentricity: the dispersed metropolis, Los Angeles, 1970–1990. J. Am. Plan. Assoc. 62 (3), 289–295. https://doi.org/10.1080/01944369608975695.
Hameed, M., Sharqi, S.S., Yaseen, Z.M., Afan, H.A., Hussain, A., Elshafie, A., 2016. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Comput. & Applic. 28 (S1), 893–905. https://doi.org/10.1007/s00521-016-2404-7.
Harrell, F., 2017. Regression modeling strategies. BIOS 330, 2018.
Holcomb, D.A., Messier, K.P., Serre, M.L., Rowny, J.G., Stewart, J.R., 2018. Geostatistical prediction of microbial water quality throughout a stream network using meteorology, land cover, and spatiotemporal autocorrelation. Environ. Sci. Technol. 52 (14), 7775–7784. https://doi.org/10.1021/acs.est.8b01178.
Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., ... Megown, K., 2015. Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. Photogrammetric Engineering & Remote Sensing 81 (5), 345–354.
Nitrogen fact sheet. Retrieved on 9/21/2020 from. https://www.umass.edu/mwwp/resources/factsheets.html#anchor221242.
Hwang, S.-A., Hwang, S.-J., Park, S.-R., Lee, S.-W., 2016. Examining the relationships between watershed urban land use and stream water quality using linear and generalized additive models. Water 8 (4), 155. https://doi.org/10.3390/w8040155.
Jones, J.E., Earles, T.A., Fassman, E.A., Herricks, E.E., Urbonas, B., Clary, J.K., 2005. Urban storm-water regulations—are impervious area limits a good idea? J. Environ. Eng. 131 (2), 176–179. https://doi.org/10.1061/(asce)0733-9372(2005)131:2(176).
Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. Environ. Model Softw. 23 (7), 835–845. https://doi.org/10.1016/j.envsoft.2007.10.001.
Kho, Julia, 2018. Why random forest is my favorite machine learning model—discover the real world advantages and drawbacks of the Random Forest. Retrieved on 9/21/2020 from. https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706.
Kraskov, A., Stögbauer, H., Grassberger, P., 2011. Erratum: estimating mutual information [Phys. Rev. E 69, 066138 (2004)]. Physical Review E 83 (1), 019903.
Lee, S.-W., Hwang, S.-J., Lee, S.-B., Hwang, H.-S., Sung, H.-C., 2009. Landscape ecological approach to the relationships of land use patterns in watersheds to water quality characteristics. Landsc. Urban Plan. 92 (2), 80–89. https://doi.org/10.1016/j.landurbplan.2009.02.008.
Lek, S., Guiresse, M., Giraudel, J.L., 1999. Predicting stream nitrogen concentration from watershed features using neural networks. Water Res. 33 (16), 3469–3478.
Li, Y., Li, Y., Qureshi, S., Kappas, M., Hubacek, K., 2015. On the relationship between landscape ecological patterns and water quality across gradient zones of rapid urbanization in coastal China. Ecol. Model. 318, 100–108. https://doi.org/10.1016/j.ecolmodel.2015.01.028.
Lintern, A., Webb, J.A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., ... Western, A.W., 2017. Key factors influencing differences in stream water quality across space. Wiley Interdisciplinary Reviews: Water 5 (1), e1260. https://doi.org/10.1002/wat2.1260.
Liu, A., Goonetilleke, A., Egodawatta, P., 2012. Inadequacy of land use and impervious area fraction for determining urban stormwater quality. Water Resour. Manag. 26 (8), 2259–2265. https://doi.org/10.1007/s11269-012-0014-4.
Liu, J., Zhang, X., Wu, B., Pan, G., Xu, J., Wu, S., 2017. Spatial scale and seasonal dependence of land use impacts on riverine water quality in the Huai River basin, China. Environ. Sci. Pollut. Res. 24 (26), 20995–21010.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, pp. 4765–4774.
Lv, H., Xu, Y., Han, L., Zhou, F., 2015. Scale-dependence effects of landscape on seasonal water quality in Xitiaoxi catchment of Taihu Basin, China. Water Sci. Technol. 71 (1), 59–66. https://doi.org/10.2166/wst.2014.463.
McGarigal, K., 1995. FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure. vol. 351. US Department of Agriculture, Forest Service, Pacific Northwest Research Station.
McHarg, I.L., 1996. A Quest for Life: An Autobiography. John Wiley and Sons, New York.
McHarg, I.L., Sutton, J., 1975. Ecological plumbing for the Texas coastal plain: the Woodlands New Town Experiment. Landsc. Archit. 65 (1), 80–90.
Mirzaei, M., Jafari, A., Gholamalifard, M., Azadi, H., Shooshtari, S.J., Moghaddam, S.M., ... Witlox, F., 2020. Mitigating environmental risks: Modeling the interaction of water quality parameters and land use cover. Land Use Policy 95, 103766. https://doi.org/10.1016/j.landusepol.2018.12.014.
Molina-Navarro, E., Segurado, P., Branco, P., Almeida, C., Andersen, H.E., 2020. Predicting the ecological status of rivers and streams under different climatic and socioeconomic scenarios using Bayesian Belief Networks. Limnologica 80, 125742. https://doi.org/10.1016/j.limno.2019.125742.
Molnar, C., 2019. Interpretable Machine Learning Lulu.com.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—a discussion of principles. J. Hydrol. 10 (3), 282–290.
Obropta, C.C., Del Monaco, N., 2018. Reducing directly connected impervious areas with green stormwater infrastructure. J. Sustain. Water Built Environ. 4 (1), 05017004. https://doi.org/10.1061/jswbay.0000833.
Oeding, S., Taffs, K.H., Cox, B., Reichelt-Brushett, A., Sullivan, C., 2018. The influence of land use in a highly modified catchment: investigating the importance of scale in riverine health assessment. J. Environ. Manag. 206, 1007–1019. https://doi.org/10.1016/j.jenvman.2017.12.005.
Parsons, 2019. Texas coastal waters: nutrient reduction strategies report. Retrieved on 9/21/2020:. https://www.gulfspillrestoration.noaa.gov/sites/default/files/Task%205_FNLWatershed%20Assessment_August2019_FINAL.pdf.
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.
Riitters, K.H., O'Neill, R.V., Hunsaker, C.T., Wickham, J.D., Yankee, D.H., Timmins, S.P., ... Jackson, B.L., 1995. A factor analysis of landscape pattern and structure metrics. Landscape Ecology 10 (1), 23–39. https://doi.org/10.1007/bf00158551.
Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., Pradhan, B., 2018. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci. Total Environ. 644, 954–962.
Schreiber, J., Jessulat, M., & Sick, B. (2019). Generative adversarial networks for operational scenario planning of renewable energy farms: a study on wind and photovoltaic. Artificial Neural Networks and Machine Learning – ICANN 2019: Image Processing, 550–564. doi:https://doi.org/10.1007/978-3-030-30508-6_44.
Sharifi, A., Yen, H., Boomer, K.M.B., Kalin, L., Li, X., Weller, D.E., 2017. Using multiple watershed models to assess the water quality impacts of alternate land development scenarios for a small community. CATENA 150, 87–99. https://doi.org/10.1016/j.catena.2016.11.009.
Shi, Z.H., Ai, L., Li, X., Huang, X.D., Wu, G.L., Liao, W., 2013. Partial least-squares regression for linking land-cover patterns to soil erosion and sediment yield in watersheds. J. Hydrol. 498, 165–176. https://doi.org/10.1016/j.jhydrol.2013.06.031.
Shi, P., Zhang, Y., Li, Z., Li, P., Xu, G., 2017. Influence of land use and land cover patterns on seasonal water quality at multi-spatial scales. CATENA 151, 182–190. https://doi.org/10.1016/j.catena.2016.12.017.
Sun, R., Chen, L., Chen, W., Ji, Y., 2011. Effect of land-use patterns on total nitrogen concentration in the upstream regions of the Haihe River Basin, China. Environ. Manag. 51 (1), 45–58. https://doi.org/10.1007/s00267-011-9764-7.
Sohn, W., Kim, J.-H., Li, M.-H, 2017. Low-impact development for impervious surface connectivity mitigation: assessment of directly connected impervious areas (DCIAs). J. Environ. Plan. Manag. 1871–1889.

Sun, Y., Guo, Q., Liu, J., Wang, R., 2014. Scale effects on spatially varying relationships between urban landscape patterns and water quality. Environ. Manag. 54 (2), 272–287. https://doi.org/10.1007/s00267-014-0287-x.

Taka, M., Aalto, J., Virkanen, J., Luoto, M., 2016. The direct and indirect effects of watershed land use and soil type on stream water metal concentrations. Water Resour. Res. 52 (10), 7711–7725.

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

Texas Commission on Environmental Quality, 2005. South Central Texas streams: evaluating water quality for aquatic life and recreation. Retrieved on 9/21/2020 from. https://www.tceq.texas.gov/waterquality/tmdl/31-sc_bacox_project.html#phase1.

Texas Commission on Environmental Quality, 2012. Managing nonpoint source pollution in Texas, 2011 annual report. Retrieved on 9/21/2020 from. https://texashistory.unt.edu/ark:/67531/metapth326659/m2/1/high_res_d/txca-0295.pdf.

Texas Commission on Environmental Quality, 2019. Managing nonpoint source pollution in Texas, 2018 annual report. Retrieved on 9/21/2020 from. https://www.tceq.texas.gov/assets/public/waterquality/nps/annualreports/066_18.pdf.

Tu, J., Xia, Z.G., 2008. Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation. Sci. Total Environ. 407 (1), 358–378.

Wang, R., Zhang, X., Li, M.-H., 2019. Predicting bioretention pollutant removal efficiency with design features: a data-driven approach. J. Environ. Manag. 242, 403–414. https://doi.org/10.1016/j.jenvman.2019.04.064.

Wijesiri, B., Deilami, K., Goonetilleke, A., 2018. Evaluating the relationship between temporal changes in land use and resulting water quality. Environ. Pollut. 234, 480–486. https://doi.org/10.1016/j.envpol.2017.11.096.

Wilson, C.O., 2015. Land use/land cover water quality nexus: quantifying anthropogenic influences on surface water quality. Environ. Monit. Assess. 187 (7). https://doi.org/10.1007/s10661-015-4666-4.

World Population Review, 2019. Texas population 2019. Retrieved from. 177. http://worldpopulationreview.com/states/texas-population/.

Wuebbles, D.J., Hayhoe, K., 2004. Climate change projections for the United States Midwest. Mitig. Adapt. Strateg. Glob. Chang. 9 (4), 335–363.

Xu, T., Coco, G., Neale, M., 2020. A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. Water Res. 177, 115788.

Yang, B., Li, M.-H., 2011. Assessing planning approaches by watershed streamflow modeling: case study of The Woodlands; Texas. Landsc. Urban Plan. 99 (1), 9–22. https://doi.org/10.1016/j.landurbplan.2010.08.007.

Yang, X.E., Wu, X., Hao, H.L., He, Z.L., 2008. Mechanisms and assessment of water eutrophication. J Zhejiang Univ Sci B 9 (3), 197–209.

Yu, D., Shi, P., Liu, Y., Xun, B., 2013. Detecting land use-water quality relationships from the viewpoint of ecological restoration in an urban area. Ecol. Eng. 53, 205–216. https://doi.org/10.1016/j.ecoleng.2012.12.045.

Zhou, P., Huang, J., Pontius, R.G., Hong, H., 2016. New insight into the correlations between land use and water quality in a coastal watershed of China: does point source pollution weaken it? Sci. Total Environ. 543, 591–600. https://doi.org/10.1016/j.scitotenv.2015.11.063.