



Stream water quality prediction using boosted regression tree and random forest models

Ali O. Alnahit² · Ashok K. Mishra¹ · Abdul A. Khan¹

Accepted: 1 December 2021 / Published online: 20 January 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Reliable water quality prediction can improve environmental flow monitoring and the sustainability of the stream ecosystem. In this study, we compared two machine learning methods to predict water quality parameters, such as total nitrogen (TN), total phosphorus (TP), and turbidity (TUR), for 97 watersheds located in the Southeast Atlantic region of the USA. The modeling framework incorporates multiple climate and watershed variables (characteristics) that often control the water quality indicators in different landscapes. Three techniques, such as stepwise regression (SR), Least Absolute Shrinkage and Selection Operator (LASSO), and genetic algorithm (GA), are implemented to identify appropriate predictors out of 28 climate and catchment-related variables. The selected predictors were then used to develop the Random Forest (RF) and Boosted regression tree (BRT) models for water quality predictions in selected watersheds. The results highlighted that while both algorithms provided reasonable results (based on statistical metrics), the RF algorithm was easier to train and robust to model overfitting. Partial dependence plots highlighted the complex and nonlinear relationships between the individual predictors and the water quality indicators. The thresholds obtained from partial dependence plots showed that the median values of total nitrogen (TN) and total phosphorus (TP) in streams increase significantly when the percentage of urban and agricultural lands is above 40% and 43% of the watershed area, respectively. Furthermore, when soil hydraulic conductivity increases, the reduction in runoff results in decreased Turbidity levels in streams. Therefore, identifying the key watershed characteristics and their critical thresholds can help watershed managers create appropriate regulations for managing and sustaining healthy stream ecosystems. Besides, the forecasting models can improve water quality predictions in ungauged watersheds.

Keywords Water quality · Machine learning algorithms · Random forests · Boosted regression trees

1 Introduction

The rapid change in land use and agricultural practices can alter the stream water quality (Castela et al., 2008; Mishra et al., 2020; Walsh et al. 2005; Wang et al., 2021a, b). Maintaining healthy streams poses a challenge, mainly because of the many pollutant sources and the complex interaction between different watershed characteristics (Waite et al., 2010; Walsh & Webb, 2016; Yu et al., 2014).

The increase in total nitrogen (TN) and total phosphorus (TP) concentrations in rivers is often linked to the high percentages of urban lands within a watershed (Bucak et al., 2018; Castela et al., 2008; Johnson & Ringler, 2014; Mattsson et al., 2005; Walsh et al. 2005). This high level of pollutants in streams can lead to eutrophication and water quality degradation (Correll, 1999; Hecky & Kilham, 1988; Paerl, 1988).

Previous studies highlighted the significant correlations between the anthropogenic variables (e.g., urbanization and agricultural activities) and the concentration of TN and TP in a watershed (Allan, 2004; Giri & Qiu, 2016; Lintern et al., 2018). Watershed characteristics and climatic variables (e.g., topography, soil, climatic data) can also influence stream water quality (Alnahit et al., 2020; Lintern et al., 2018; Tramblay et al., 2010). For example, a steep

✉ Ashok K. Mishra
ashokm@g.clemson.edu

¹ Glenn Department of Civil Engineering, Clemson University, Clemson, SC, USA

² Department of Civil Engineering, King Saud University, Riyadh, Saudi Arabia

slope may influence stream water quality by mobilizing pollutants into streams, leading to water quality degradation (Alnahit et al., 2020; Kang et al., 2010; Lintern et al., 2018). Similarly, soil properties can also affect water quality (Alnahit et al., 2020; Lintern et al., 2018; Varanka et al., 2015). For instance, watersheds dominated by parent rock showed low values of dissolved ions; on the other hand, soft sedimentary rocks showed high values of dissolved ions (Young et al., 2005). Furthermore, a high phosphorus level in rivers was noticed in a watershed with high values of sediment depositions (Dillon & Kirchner, 1975). Different watershed characteristics can potentially influence water quality since they influence the mobilization process and the delivery of indicators into rivers (Granger et al., 2010; Lintern et al., 2018).

Overall, there are two commonly modeling strategies for predicting stream water quality in ungauged watersheds, (1) deterministic physically-based models (e.g., distributed hydrologic and water quality models) and (2) statistical and machine learning methods (e.g., decision tree models). This study uses machine learning methods to estimate the long-term median stream water quality indicators using several climate and watershed characteristics. Linear regression models are commonly used to explore the relationship between water quality and different land-use variables (Seber & Lee, 2012; Tong & Chen, 2002; Zampella et al., 2007). However, the effects of watershed characteristics on water quality indicators are often complex and nonlinear. Recent machine learning algorithms can handle nonlinear relationships associated with complex watershed processes (Alpaydin, 2020; Konapala & Mishra, 2020; Shen et al., 2020). Moreover, these algorithms determine the relationship between response variables (e.g., water quality indicators) and predictors (e.g., land-use variables) instead of a priori assumption, improving the model prediction accuracy. Several studies have applied techniques adapted from machine learning models to understand the relationships between water quality and land use variables (e.g., Bui et al., 2020; Castrillo & García, 2020; Fatehi et al., 2015; Ko et al., 2015; Puissant et al., 2014; L. Q. Shen et al., 2020; Singh et al., 2017; Tu & Xia, 2008; R. Wang et al., 2021a, b). These studies highlighted that these algorithms are more suitable than linear models such as Bayesian linear regression, stepwise linear regression, and partial least squares regression, especially when human/landscape interactions are complex (Giri et al., 2019; Mouazen et al., 2010).

Among the previously used machine learning algorithms, the boosted regression tree (BRT) algorithm and the random forest (RF) algorithm recently gained a lot of attention (Chen et al., 2020; Fang et al., 2021; Knierim et al., 2020; Konapala & Mishra, 2020; Shen et al. 2020; Veetil & Mishra, 2020). BRT and RF have fewer parameters, and both can investigate and provide estimates

related to the hierarchy of variables in the classification (Everingham et al., 2016). Additionally, RF and BRT algorithms (1) have less user-defined parameters; (2) are flexible in handling nonlinear relationships, missing values, and outliers; (3) can limit model overfitting; (4) are capable of incorporating qualitative and quantitative variables; and (5) have been applied successfully in different areas (Giri et al., 2019; Konapala & Mishra, 2020; Veetil & Mishra, 2020; Yang et al. 2016; Shen et al. 2020).

Many recent studies highlighted the use of machine learning algorithms to study the potential influence of human activities on water quality parameters (e.g., Giri et al., 2019; Jeung et al., 2019; Onderka et al., 2012; Trambly et al., 2010; Tung & Yaseen, 2021; Wang et al., 2021a, b). However, prior studies have used a limited number of watersheds and associated variables. Additionally, no prior studies performed a comprehensive analysis using RF and BRT algorithms to predict water quality indicators (TN, TP, TUR) for a large number (97 nos) of watersheds based on a combination of climate, watershed, and morphological variables in the southeast USA.

This study will complement previous studies that used only a limited number of watersheds and associated variables. The median values of water quality indicators are selected for individual watersheds, and corresponding 28 variables associated with watershed, climate, and topographic and soil characteristics are used for the model development. The selected watersheds represent various land use, climate, watershed characteristics with different watershed areas to improve our understanding of the predictive power of two selected machine learning algorithms that can capture the linkage between climate-watershed characteristics and water quality indicators. The RF and BRT algorithms use an ensemble of many simple tree models to optimize predictive performance instead of a single tree model used in the traditional simple regression. The water quality indicators investigated in this study are TN, TP, and TUR, while the predictors (independent variables) represent a combination of the climatic and watershed characteristics.

Overall, this study aims to address the following research questions: (1) to compare and identify the best machine learning algorithms based on the classification and decision tree approach for water quality (TN, TP, and TUR) prediction in streams; and (2) to investigate the functional relationships and interactions among dominant variables influencing stream water quality based on the interpretive machine learning techniques (i.e., partial dependence analysis). The remainder of the manuscript is organized as follows: Sect. 2 introduces the study area and data used in the study. The methods employed in the study are discussed in Sect. 3. Section 4 presents the results, while the discussion is provided in Sect. 5. The conclusions drawn from this study are summarized in Sect. 6.

2 Study area and data

2.1 Study area

This study includes 97 watersheds located in North Carolina, South Carolina, and Georgia (Fig. 1a). These watersheds are located in three main physiographic regions, including coastal plain, blue Ridge, and Piedmonts (Turner & Ruscher, 1988). There are more than 250 watersheds with water quality monitoring stations in the region; however, only 97 watersheds were selected based on the following criteria: (1) nested watersheds were not included to avoid pollutant transfer from other watersheds; (2) watersheds with reservoirs covering more than 25% of the watershed were excluded, and (3) water quality stations located less than 50 km downstream of a reservoir outlet were eliminated.

The watersheds were delineated using a 10 m Digital Elevation Model (DEM). The latitude and longitude of each watershed outlet were located, and then the Soil and

Water Assessment Tool (SWAT) was used to generate the watershed boundary (Arnold et al., 2012). The selected watersheds vary in size from 72 to 5786 km². In addition, the selected watersheds experience different degrees of human activities (urbanization and agricultural activities) (Fig. 1b). The primary urbanization form is expanding low-density residential areas, medium-density residential areas, and high-density residential areas. Such changes in land use have altered watersheds hydrology and the environmental conditions of streams in the study area.

The study area climate is characterized by a humid subtropical climate, with hot summers and mild winters. The mean annual temperature is 20 °C, while the mean annual evapotranspiration is 635 mm/year (SCDHEC, 2016). The study area runs from the north to the south, with elevation ranging from 2035 to 0 m above sea level (Fig. 1b). Land use is dominated by forest (approximately 55%, mainly located in the northern side of the study area, Fig. 1b).

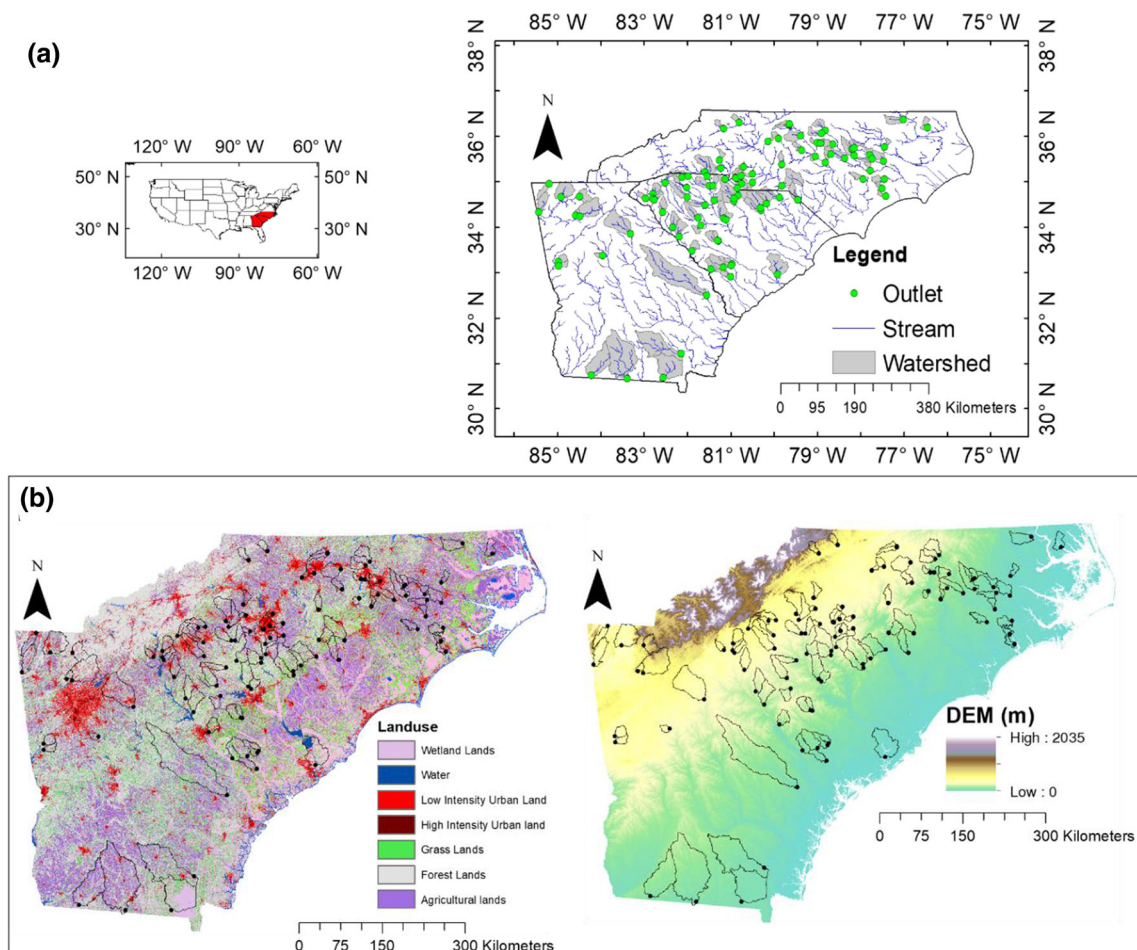


Fig. 1 a Selected watersheds located within the Southeastern part of the USA. b Examples of watershed characteristics: land use/land cover and the digital elevation model (DEM) over the selected watersheds

2.2 Datasets

For each watershed, the water quality monitoring data from 2000 to 2019, including TN, TP, and TUR, were downloaded using data retrieval tools from R software package “dataRetrieval” (<https://github.com/USGS-R/EflowStats>). The water quality monitoring data was expressed as a concentration (mg/l) (or in NTU in the case of TUR). Since the stationarity of the time series is crucial, the stationarity was checked at each site using two methods. Specifically, each time series was split into four sections, and the mean and variance were computed for each section. The augmented Dickey-Fuller (ADF) unit root test and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test were utilized (Vazifekhhah et al., 2019). Most of the time series passed the stationarity tests. We performed first-order differencing for the time series that did not pass the stationarity test to generate stationary time series (Mishra & Desai, 2005). Furthermore, a t-test at a 95% confidence interval was performed to exclude outliers for each time series.

The watershed characteristics selected in this study were land use, topography, geology, and climatic data (Table 1). The land use data were obtained from the National Land Cover Dataset (NLCD) for the year 2011. The land use data of 2011 was used to represent the whole period (2000 to 2019) to capture the broad impacts of land use on water quality. The Soil data was downloaded from the Soil Survey Geographic (SSURGO) database (SSURGO, 2018). The climate data (precipitation and temperature) data from 2000 to 2019 over the study area was downloaded from Parameter-elevation Relationships on Independent Slopes Model (PRISM) (Daly et al., 2008). PRISM was developed employing ground rain gauge data, DEM, and interpolation schemes (Daly et al., 2008). The precipitation and temperature data were averaged over each watershed (areal average) using the Zonal Statistics tools in ArcMap (Esri, 2014). The topographic data for each watershed (e.g., mainstream length–width ratio, watershed slope, and watershed elevation) was extracted from a 10 m DEM using SWAT model. Twenty-eight different watershed/climatic characteristics were obtained from these datasets (Table 1). Following previous research, these characteristics were selected to identify the essential predictors influencing the water quality indicators (Alnahit et al., 2020; Lintern et al., 2018; Mainali & Chang, 2018; Varanka & Luoto, 2012). Based on EPA criteria, the concentration for TN and TP should be about 0.90 mg/l and 0.04 mg/l, respectively (US EPA, 2002; Ice & Binkley, 2003). The water quality indicators and land use vary within the selected watersheds (Fig. 2). For example, the median TN based on the 97 watersheds ranged from 0.54 to 1.9 mg/l, while the overall median for all watersheds is

about 0.9 mg/l (Fig. 2a). FRST land has the highest percentage among different types of land use, followed by URBAN, AGRL, GRAS, HAY, and WTLN (Fig. 2b).

3 Model development

The Classification and Regression Tree (CART) (Breiman, 2001; Friedman & Meulman, 2003; Golden et al., 2016; Yang et al. 2016) is a flexible and nonparametric method implemented in this study. The CART method can handle outliers, missing values, multicollinearity, and heteroscedasticity in the datasets. CART method is commonly used to investigate complex datasets with numeric and/or categorical variables (predictor variables) that interact with each other nonlinearly (De’ath and Fabricius 2000). Both RF and BRT belong to the CART family, which has been implemented in different disciplines, such as species distributions (Shabani et al., 2017), groundwater mapping (Naghbi et al., 2016), water quality (Golden et al., 2016; Povak et al., 2014), aquatic ecosystems (Elith et al., 2008; Smucker et al., 2013; Tonkin et al., 2014), and environmental modeling (Giri et al., 2019; Strobl et al., 2008).

Watershed characteristics and climatic variables (total of 28 characteristics) were chosen as predictor variables (independent variables), while the water quality indicators (TN, TP, and TUR) were chosen as dependent variables. The median values of temporal variations of TN, TP, and TUR at each watershed outlet were calculated and used as the dependent variables. The one-way variance test indicated significant differences in water quality indicators’ median values among the watersheds [confidence interval of 95%; $\alpha = 0.05$; $n = 97$]. The overall modeling framework is shown in Fig. 3, which are discussed in the following sections.

3.1 Variables selection

Three different approaches were used to select the predictor variables (Fig. 3b). In addition to using all the 28 predictor variables, a stepwise linear regression (SR) was used to select the smallest number of relevant variables that provide the best linear combination (Lima et al., 2016; Wang et al., 2018). However, SR may have statistical deficiencies, such as bias estimates, standard error, and size of p-values (Harrell, 2001; Mo et al., 2016); therefore, the Least Absolute Shrinkage and Selection Operator (LASSO) was also used for variable selection (Bardsley et al., 2015; Tibshirani, 1996). LASSO uses a cross-validation technique to find a set of significant variables with the optimal performance; LASSO shrinks regression coefficients to zero if there is a strong correlation with another

Table 1 Definitions of the selected independent variables to quantify relationships between watershed characteristics and climatic variables on the mean water quality indicators

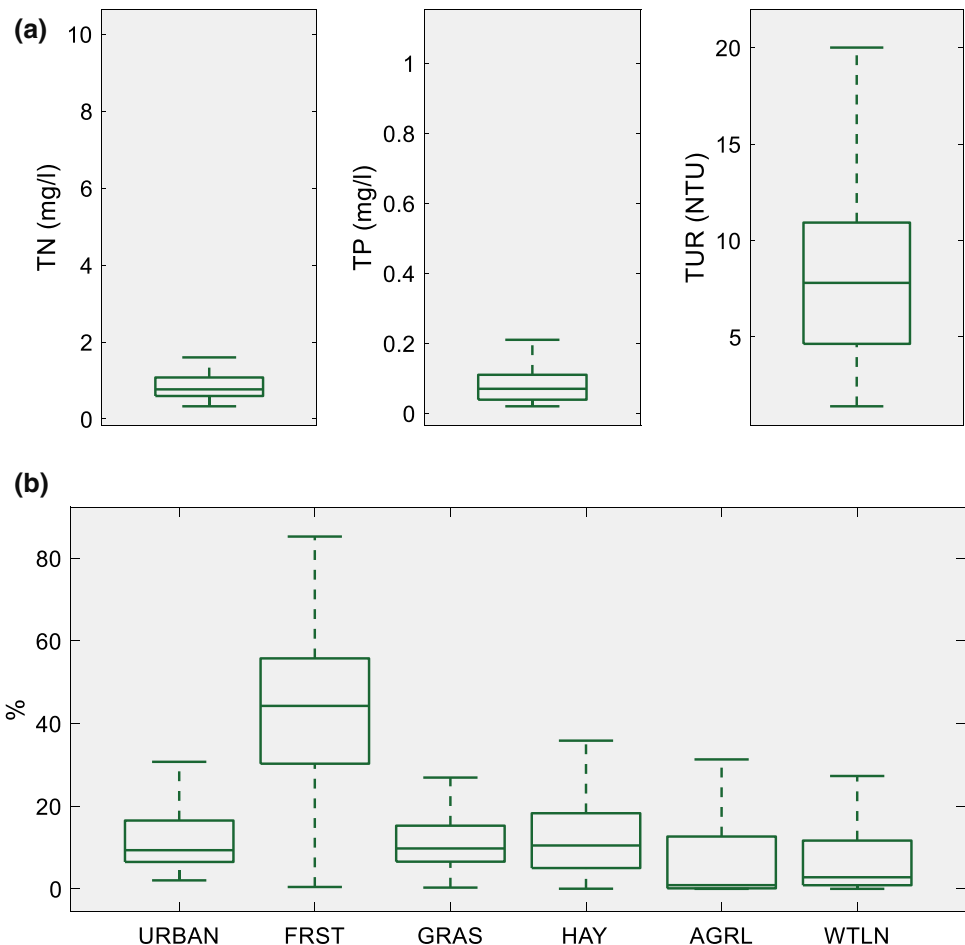
Name	Definition	Unit
(a) Land use variables		
URBAN	Cumulative percent of low density, medium density, and high-density urban land use	%
FRST	Cumulative percent of forest land use (e.g., deciduous, evergreen, and mixed)	%
GRAS	Cumulative percent grassland use (Range-Brush and Range-Grasses)	%
HAY	Cumulative percent of Hay/pasture lands	%
AGRL	Cumulative percent of cultivated crops	%
WTLN	Cumulative percent of wetland land use (Wetlands-forested and Wetlands-non-forested)	%
(b) Soil variables		
SOL_K	Saturated hydraulic conductivity (areal average)	mm/hr
SOL_F	Soil erodibility (areal average)	–
SOL_OM	Organic matter content in the soil (areal average)	%wt
CLAY	Clay content (areal average)	%wt
SAND	Sand content (areal average)	%wt
SILT	Silt content (areal average)	%wt
SOL_AWC	Available water capacity of the soil layer (areal average)	mm/mm
SOL_pH	Soil pH value (areal average)	–
(c) Topography variables		
AREA	Area of the watershed	Km ²
WT_S	Mean Slope of the watershed	%
CH_L	The longest flow path in the watershed	m
CH_W	The mean width of the main channel in the watershed	m
CH_D	The mean depth of the main channel in the watershed	m
CH_S	The mean slope of the main channel	m/m
CH_WDR	The ratio of the main channel (width to depth)	–
Elev	Elevation of the watershed (weighted average)	m
(d) Climatic variables		
MeanRain	Mean total rainfall (areal average)	mm/year
WetMRain	Mean rainfall of the wettest month (areal average)	mm
DryMRain	Mean rainfall of the driest month (areal average)	mm
MeanTemp	Mean annual temperature (areal average)	°C
ColdMTemp	Mean temperature of the coldest month (areal average)	°C
HotMTemp	Mean temperature of the hottest month (areal average)	°C

variable (Bardsley et al., 2015). Furthermore, a non-linear method (genetic algorithm, GA) was included to choose the most significant climatic/watershed characteristics (Huang et al. 2016; Taghizadeh-Mehrjardi et al., 2016). GA is an adaptive optimization search method that mimics Darwinian natural selection theory to find optimal values of a function (Huang et al., 2016; Taghizadeh-Mehrjardi et al., 2016). Three standard parameter settings were defined for the GA, population size of 50, crossover rate of 0.80, and mutation rate of 0.1 based on the recommendation of (Welikala et al., 2015). The relevant variables based on the four different datasets were used to develop predictive models based on RF and BRT algorithms.

3.2 Random forests (RF) model

The RF algorithm approach uses an ensemble of regression (or classification) tree models (Breiman, 2001). Specifically, a series of individual trees are build based on random subsamples from the original data. Each subsample provides a decision tree, and each decision tree is used to predict the response variable (or a class). In the end, an ensemble average of all individual trees is computed. The inclusion of several trees increases the probability of deriving an effective prediction model (Breiman, 2001; Strobl et al., 2008). The accuracy of the random forests algorithm relies mainly on the strength of the individual

Fig. 2 Box plots showing the range of **a** water quality constituents (TN, TP, and TUR) and **b** land-use types. Definitions of land-use variables are shown in Table 1



tree classifiers and the dependency between the classifiers (Amit & Geman, 1997). Therefore, key parameters for RF models are the number of trees and predictor variables used to determine the split at each node (Vorpahl et al., 2012). Figure 3d illustrates the steps used to develop the RF prediction model for each watershed's median water quality indicators. The RF modeling requires two parameters: the number of trees (ntree) and the number of variables at each tree node (mtry). To optimize the two parameters, a grid search was performed using different combinations of ntree and mtry. The range of the number of ntree was set between 100 and 2000 with an increment of 50. The number of selected independent variables (mtry) ranged from 1 to 28 (or the total number of significant variables based on SR, LASSO, and GA) with an increment of 1 (Rodriguez-Galiano et al., 2015). The data was split into 10-folds for cross-validations, and the error rates for each of the 10 cross-validation partitions were aggregated into a mean percentage error. Three replicates of the ten-fold cross-validation were performed, and the process was repeated 50 times to evaluate the reliability of the predicted model (Fig. 3d).

The relative importance of each variable was calculated based on the mean decrease in accuracy (%IncMSE), as suggested by Genuer et al., (2010). The mean decrease in accuracy was calculated as a percentage of mean square error (MSE) increment when removing that variable from the prediction set. A higher value of %IncMSE for a variable indicates that the predictor has higher relative importance than other predictors. Partial dependence plots in RF model were also calculated for each independent variable.

3.3 Boosted regression trees (BRT) model

The Boosted regression trees (BRT) technique is an improvement of the regression trees model. BRT uses a boosting technique to combine decisions from a sequence of base models to enhance the accuracy of the final model (Elith et al., 2008; Naghibi et al., 2016; Yang et al. 2016). BRT is a forward and stagewise procedure, where a sub-sample of the original data is randomly selected to fit new tree models to minimize a loss function (Golden et al., 2016). The final fitted model is a linear function of the sum

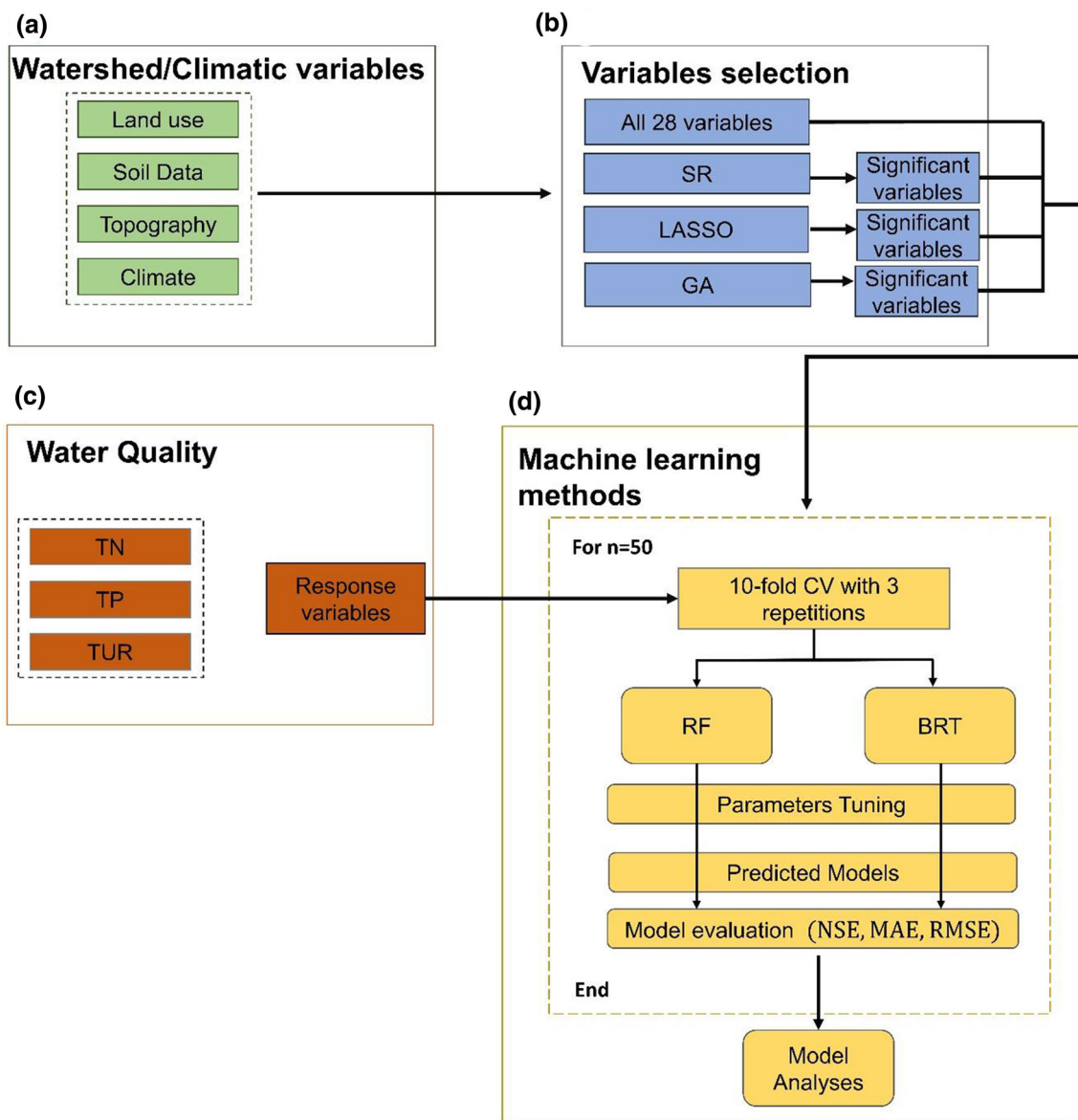


Fig. 3 The modeling framework to model the median water quality constituents in streams

of all trees multiplied by the contribution of each tree used to build the model (Elith et al., 2008). The bag fraction (BF) in BRT is the proportion of the training set used for each model fit, learning rate (LR) is the contribution of each tree to the model development, and tree complexity (TC) is the number of nodes in a tree. The number of trees (NT) required for the best model prediction is calculated based on LR and TC (Elith et al., 2008).

In BRT modeling, four parameters (LR, T, NT, and BF) need to be defined, and to optimize these parameters, several experiments were conducted using different combinations of LR, TC, and NT. The values of LR varied from 0.001 to 0.03 at 0.002 increments; the values of TC were varied from 1 to 7 with an increment of 1; the NT

values varied from 100 to 2000 at an increment of 100. These combinations generated an optimal BRT model using three repetitions of tenfold cross-validation. As in the case of RF model, the process was repeated 50 times (Fig. 3d). The variable of importance was found by the number of times a variable appeared in all trees. The mean of the relative importance of each variable from various trees was calculated. This mean was used to build a hierarchy of overall relative importance (Elith et al., 2008; Friedman & Meulman, 2003; Golden et al., 2016; Yang et al. 2016). The partial dependence plots were generated to determine the effect of the individual independent variables on the fitted function.

Both BRT and RF algorithms use several decision trees to enhance the predictive performance. BRT and RF use different techniques (boosting in the case of BRT and bagging method in the case of RF) that may lead to different results. Specifically, the boosting method is built-in subsequent trees, while the bagging approach is built-in parallel (independently). In addition, boosting is an iterative process, where tree models are built to improve the weak learners in each tree to enhance the overall model prediction accuracy (Elith et al., 2008). In the case of boosting method, the fitted values in the final model are the sum of all trees multiplied by the contribution of each tree (Elith et al., 2008). On the other hand, trees are grown independently in the bagging method, which means that each event would have an equal probability of being selected in subsequent samples. Each tree is given equal weight for final decision-making instead of higher weight for a better performing tree during training in the boosting method (Breiman, 2001; Yang et al. 2016).

3.4 Partial dependence

The concept of partial dependence aims to quantify the functional relationship between dominant predictors and the water quality indicators in streams. Partial dependence is evaluated by integrating the effects of all the predictors beside the covariate of interest (Breiman, 2001). Partial dependence of a variable x_k is computed by averaging it over the input predictors $\{X_i, i = 1, \dots, n\}$ with fixed x_k as

$$\tilde{f}_k(x_k) = \frac{1}{n} \sum_{i=1}^n \tilde{f}_k(x_{i,C_k}, x_k) \quad (1)$$

where \hat{f} is the output based on the RF and BRT models. This partial dependence estimate is usually constructed to understand the functional relationship between the variables (x_k) and their potential influence on the water quality indicators. Here, we assessed partial dependence for a subset of dominated predictors for each model (RF and BRT) to visualize the effects of a given single predictor on the outcomes of classification (RF and BRT). For a given value of the predictor, the prediction is quantified by averaging the predictions over all other predictors in the dataset.

3.5 Model validation

BRT and RF models were evaluated using a tenfold cross-validation method. The final models for each of the water quality indicators were evaluated using three statistical measures: Nash–Sutcliffe efficiency (NSE), mean absolute error (MAE), and root mean square error (RMSE) (shown in Eqs. 2–4, respectively).

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |O_i - P_i| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2} \quad (4)$$

where n is the number of watersheds, O_i is the observed water quality variable at the watershed i , \bar{O} is the mean of the observed data, \bar{P} is the mean of the predicted data, and P_i is the predicted water quality constituent at the watershed i .

NSE represents the observed and predicted data in 1:1 line, and the prediction becomes optimal as NSE approaches to 1.0. MAE indicates how close the prediction to the observation, while RMSE is the standard deviation of the residuals. MAE and RMSE are computed and reported in the same units as the variable being evaluated (Moriassi et al., 2015). Empirical relationships were categorized by R^2 values as weak ($R^2 \leq 0.25$), moderate ($0.25 < R^2 < 0.75$), and strong ($R^2 \geq 0.75$) correlation following the recommendation of Hair et al., (2013).

4 Result

4.1 Variables selection using three methods

We performed a preliminary analysis based on Spearman's correlation matrix (31×31) for water-quality indicators (TN, TP, and TUR) and watershed/climatic characteristics (28 variables) (Fig. 4). A cell with a white color indicates that the correlation is statistically insignificant ($p > 0.05$). A positive correlation between TN and URBAN and a positive correlation between TN and SOL_AWC was observed. There is also a negative correlation between TN and FRST, a negative correlation between TN and WT_S, and weak correlation between TN and other watersheds/climatic characteristics. Similarly, a positive correlation between TP and both URBAN and SOL_AWC, negative correlations between TP and both FRST and GRAS, and a weak correlation between TP and other watersheds/climatic characteristics was observed. There are positive correlations between TUR and URBAN, FRST, and HAY across each watershed. Additionally, the proportion of clay and silt and the SOL_pH in a watershed are positively correlated with the median TUR values. On the other hand, TUR shows a negative correlation with GRAS, AGRL, WTLN lands at each watershed. Similarly, the proportion of sand and SOL_OM and SOL_K show a

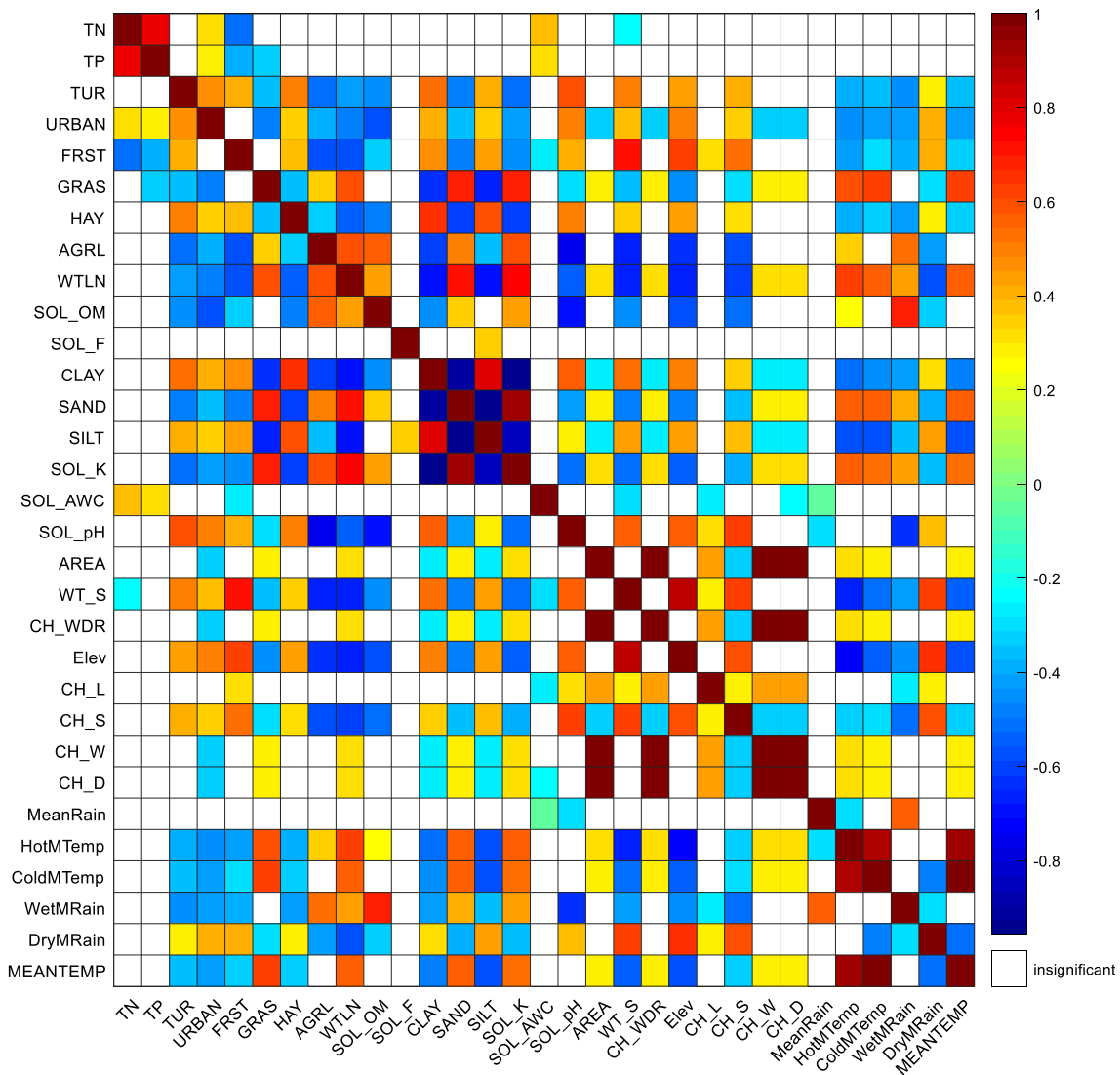


Fig. 4 The correlation matrix showing Spearman’s correlation analysis between the median water quality constituents and watershed characteristics/climatic variables. [Note: A cell with a white color

indicates that the correlation is statistically insignificant ($p > 0.05$). The definition of predictors is shown in Table 1

negative correlation with TUR. The Elev and CH_S show a positive correlation with TUR. All climatic variables (except MeanRain and DryMRain) exhibit a negative correlation with TUR values in streams.

Interestingly, while the concentrations of TN and TP are negatively correlated with FRST according to the Spearman correlation (Fig. 4), FRST is positively correlated with TUR. This is likely due to the spatial correlation between FRST lands and climatic and topographic watershed characteristics. Specifically, watersheds with higher elevation and steep slope are dominated by FRST (positively correlated with elevation and mean steep channel). Hence, FRST under these conditions may lead to more sediments and particulates being transported into receiving

streams, resulting in higher TUR values (Lintern et al., 2018; Alnahit et al., 2020).

Figure 5 shows the significant predictors for each water quality indicators (TN, TP, and TUR) selected based on SR, LASSO, and GA methods. Overall, based on the SR approach, three significant predictors are found for TN, four significant predictors are found for TP, and five significant predictors are found for TUR. On the other hand, the LASSO approach suggests that eleven predictors are significant for TN, ten predictors are significant for TP. In contrast, only eight predictors are found to be significant for TUR. A higher number of predictors are selected based on the GA approach; for example, sixteen significant predictors are selected for TN and TP, and nine predictors for TUR (Fig. 5). Specifically, URBAN, and AGRL are

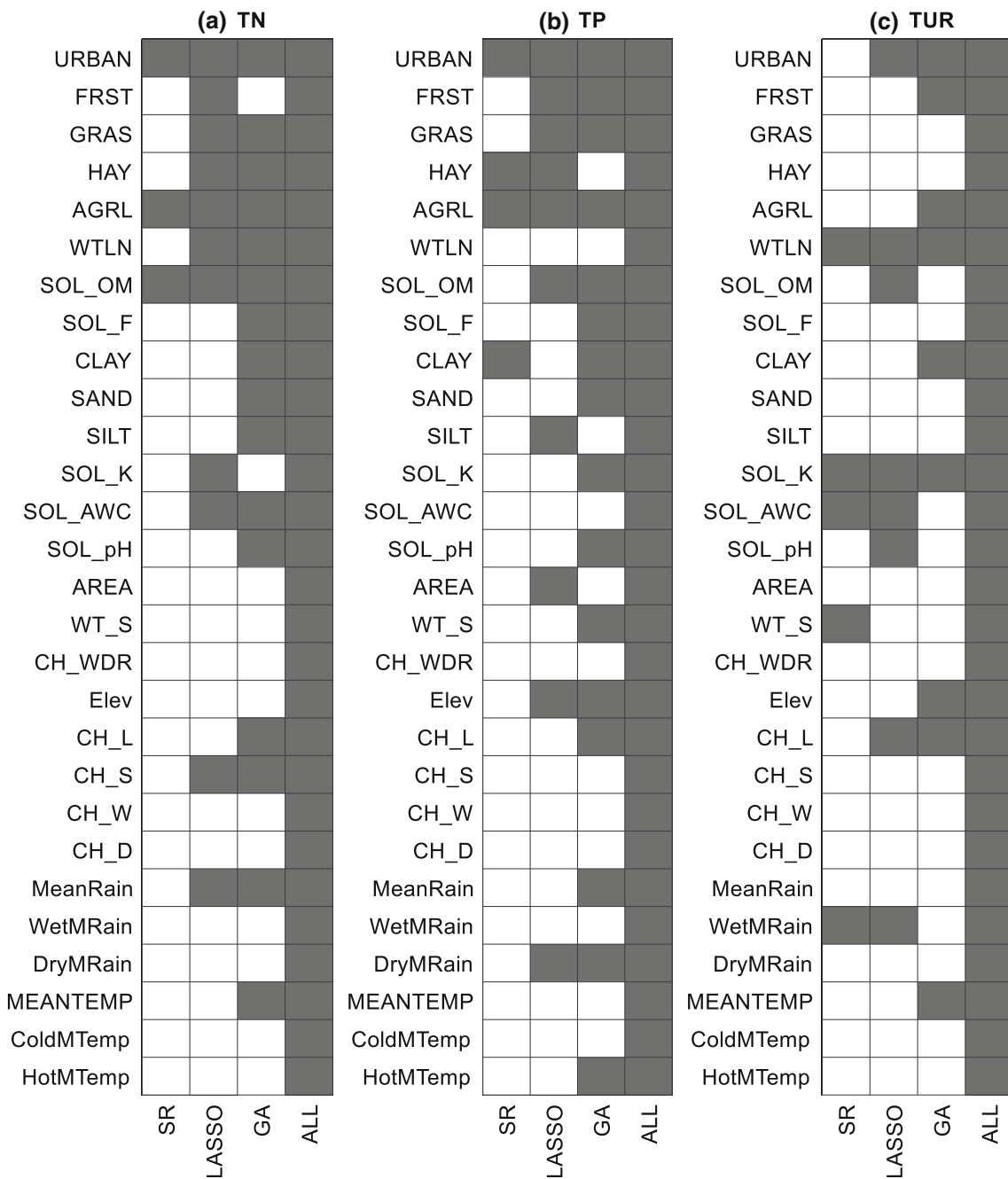


Fig. 5 Variable selection based on stepwise regression (SR), Least absolute shrinkage and selection operator (LASSO), GA (genetic algorithm), and ALL (All 28 predictor variables). [Note: A cell with a gray color indicates that the variable is selected by the variable selection method]

identified for TN by all methods. Other predictors, such as GRAS, HAY, and WTLN, were selected by all methods (except SR method) for TN.

The soil parameter such as mean SOL_OM selected by all the methods was a important predictor for TN and picked by two methods for TP. Similarly, MeanRain and mean channel slope (CH_S) are significant variables based on LASSO and GA methods (Fig. 5a). Overall, URBAN and AGRL are selected by all methods for TP, while FRST,

GRAS, Elev, and DryMRain are significant predictors based on the LASSO and GA methods (Fig. 5b). For TUR, all the three methods identified WTLN and the mean SOL_K as significant predictors, while all methods select URBAN and CH_L except for SR method (Fig. 5c). This discussion highlighted the choice of predictors can vary based on the methods (SR, LASSO, and GA), therefore it is important to evaluate the performance for the predictors for

water quality prediction, as discussed in the following section.

4.2 Evaluation of RF and BRT models

We evaluated the performance of the selected climate and watershed variables for water quality prediction over 97 watersheds. The input (predictor) variables for RF and BRT models are selected based on all 28 predictor variables (ALL), and are those identified based on the SR, LASSO, and GA methods. These four types (ALL, SR, LASSO, and GA) of input variables are selected for the individual watersheds, and the median values of water quality indicators for the same watershed is considered as an output of the model. For each water quality constituent, eight models are evaluated (four models using RF and four models using BRT models). The models are named as RF_selection method (BRT_selection method). For example, RF_LASSO represents a random forest model developed based on the variables selected by LASSO. The model performances are quantified based on the three goodness-of-fit statistics (NSE, MSE, and RMSE). The box plots of goodness-of-fit statistics developed based on selected watersheds are shown in Fig. 6.

Figure 6 shows that all models (except SR models) predicted the TN, TP, and TUR concentrations moderately well based on the median values of NSE, MAE, and RMSE. Additionally, the models selected by LASSO, GA, and the ALL models show similar levels of prediction accuracy based on the median values of NSE. The selected climatic and watershed characteristics as predictors explained at least 48% of TN, TP, and TUR variation in streams are (as indicated by NSE values). Specifically, the median NSE values explain approximately 53% of the variability in the TN, 55% of the variability in the TP, and 48% of the variability in the TUR in streams for both RF and BRT algorithms. Additionally, the random forest model algorithm performed slightly better compared to the boosted regression models for TN, TP, and TUR models (Fig. 6). For example, when using predictors selected by the GA method for TN, the model of RF_GA has higher median values of NSE (0.56) with lower median values of MAE (0.022) and RMSE (0.061) compared to BRT_GA model (NSE = 0.53, MAE = 0.024, and RMSE = 0.061).

The relative importance of the top five predictors for the TN, TP, and TUR models using RF and BRT are presented in Fig. 7 and Fig. 8, respectively. The relative importance of each predictor is calculated as the mean value of the 50 runs of each model. The TN variability in streams is influenced mainly by the presence of URBAN lands, AGRL lands, and GRAS lands, as well as the mean total rainfall (MeanRain) over a watershed. URBAN lands show the highest relative importance for all TN models, followed

by AGRL lands for RF_SR, RF_LASSO, and RF_GA methods and FRST lands in the case of RF_ALL model. On the other hand, the TP variability is influenced by URBAN, AGRL, GRAS, and watershed soil properties (the proportion of CLAY/SILT within a watershed in the SR and LASSO models). URBAN lands have the highest relative importance for all TP models, followed by MeanRain in the case of RF_SR, RF_LASSO, and RF_GA models and by HAY in the case of RF_SR model. For TUR, WTLN shows the highest relative importance for all TUR models (Fig. 7). The mean watershed slope (WT_S) appeared as an important variable in TUR_SR and TUR_ALL models.

RF and BRT models identified similar top five predictors with a high relative influence on the water quality indicators (Figs. 7 and 8). For instance, the five predictors of TN models for RF_GA and BRT_GA are the same; however, the relative importance is slightly different. URBAN is the most important predictor for TN followed by AGRL, MeanRain, GRAS, and HAY in the case of RF_GA model, while for BRT_GA model, URBAN is the most important predictor for TN followed by HAY, MeanRain, GRAS, and AGRL across the selected watershed.

Overall, the results from both the RF and BRT models suggest that the top five influential predictors for TN, TP, and TUR in streams are similar; however, the relative influence of each predictor is different in each model. This is expected as each machine-learning algorithm uses different inherent model structures. Specifically, RF algorithm generates tree independently (in parallel) where each tree is assigned equal weight for the final decision. This is different from the stagewise method of tree development that coupled with higher weight for better performing *etree* in the case of BRT. Besides, the bagging method in RF algorithm aims to minimize the variance in model fitting, while the boosting algorithm in BRT focuses on improving weak classifiers at each tree. Additionally, RF algorithm is slightly better compared to BRT algorithm. This may be due to a higher overfitting issue in BRT compared to RF. This is likely because, in the boosting algorithm, trees are grown in an adaptive way to eliminate any bias, which may reduce the variance, resulting in a model overfitting.

5 Partial dependence plots

Partial dependence plots can provide the functional relationship between an individual climate/watershed variable and the predicted water quality indicators. We assessed the partial dependence of the top dominant variables on water quality indicators for both RT and BRT models (Figs. 9 and 10, respectively). The partial plots are developed based

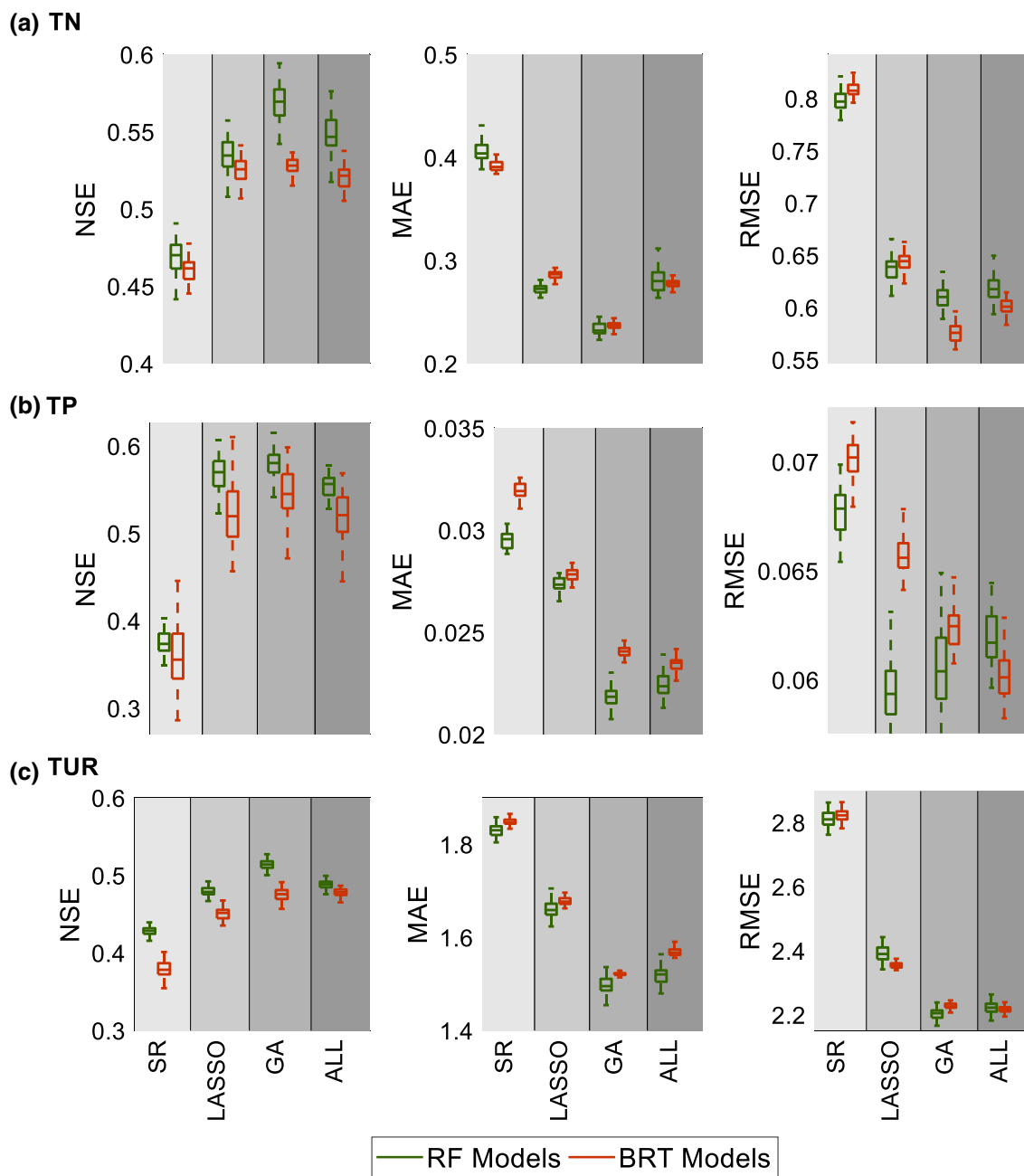


Fig. 6 The models' performance for predicting the median values of TN, TP, and TUR using random forest (RF) and boosted regression tree (BRT) with 50 runs for the different variable selection methods. The selected input variables for each method are shown in Fig. 5

on the key variables, which includes URBAN, FRST, AGRL, GRAS, WTLN, and SOL_K.

Among the most important predictors, TN and TP reveal a positive trend with the URBAN and AGRL, while they share a negative trend with the percentage of FRST and GRAS lands in the study area (Fig. 9). Specifically, for RF models, TN and TP values in streams decrease linearly as the GRAS cover increase in the watershed. On the other hand, in BRT models, the GRAS cover is nearly linearly related to TP in streams when the percentage of GRAS was

above approximately 9% of the watershed, while TN values in streams show a slight increase when the GRAS cover is around 10% of the watershed and then leveled out when the watershed is above approximately 21% of the watershed area (Fig. 10). Overall, for RF models, the partial plots suggest that TN and TP increased abruptly when the percentage of URBAN was above approximately 40% and 55% percent of the watershed, respectively and when AGRL land is above 43%.

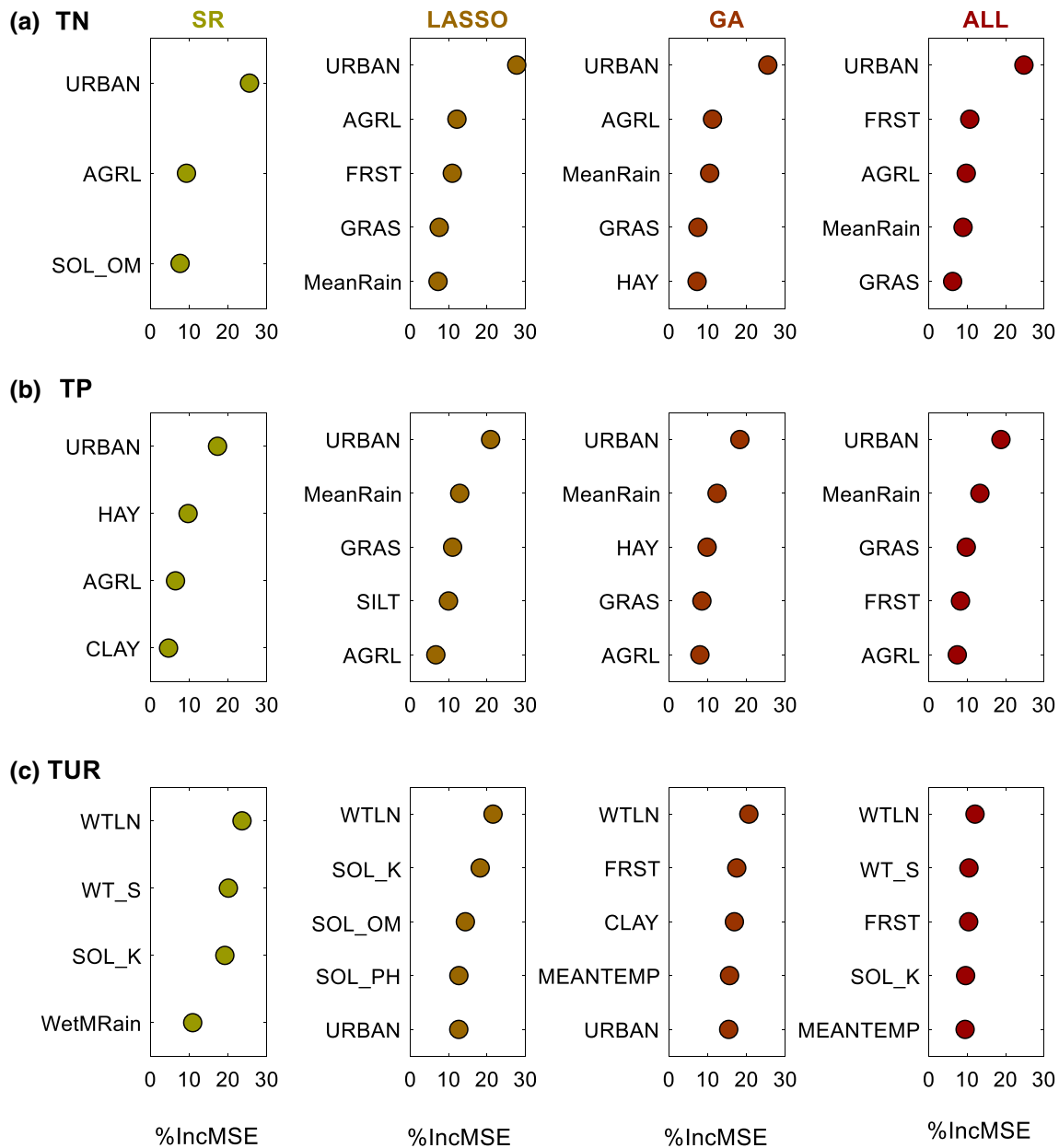


Fig. 7 The relative influence of the top 5 predictors of the median TN, TP, and TUR models based on the Random Forests (RF) algorithm

TUR shows a negative trend with the percentage of WTLN and the mean values of SOL_K, while TUR exhibits a positive trend with the percentage of URBAN and FRST. Specifically, the TUR levels in streams in both RF and BRT tend to increase as URBAN and FRST land cover increased, but only below values of about 50% of the watershed area.

6 Discussion

This study shows that urban and agricultural lands are the largest contributors to nutrient loads (TN and TP) delivered to streams. The relative importance analyses and partial dependence plots suggest that the increase in human activities (e.g., urbanization and cropping) in a watershed has led to greater TN and TP concentrations in streams. This larger proportion of urbanization in the watershed, resulting in high TN and TP in streams, may be due to the increased use of fertilizer on urban lawns, the presence of treatment plants, and stormwater discharges (Perry &

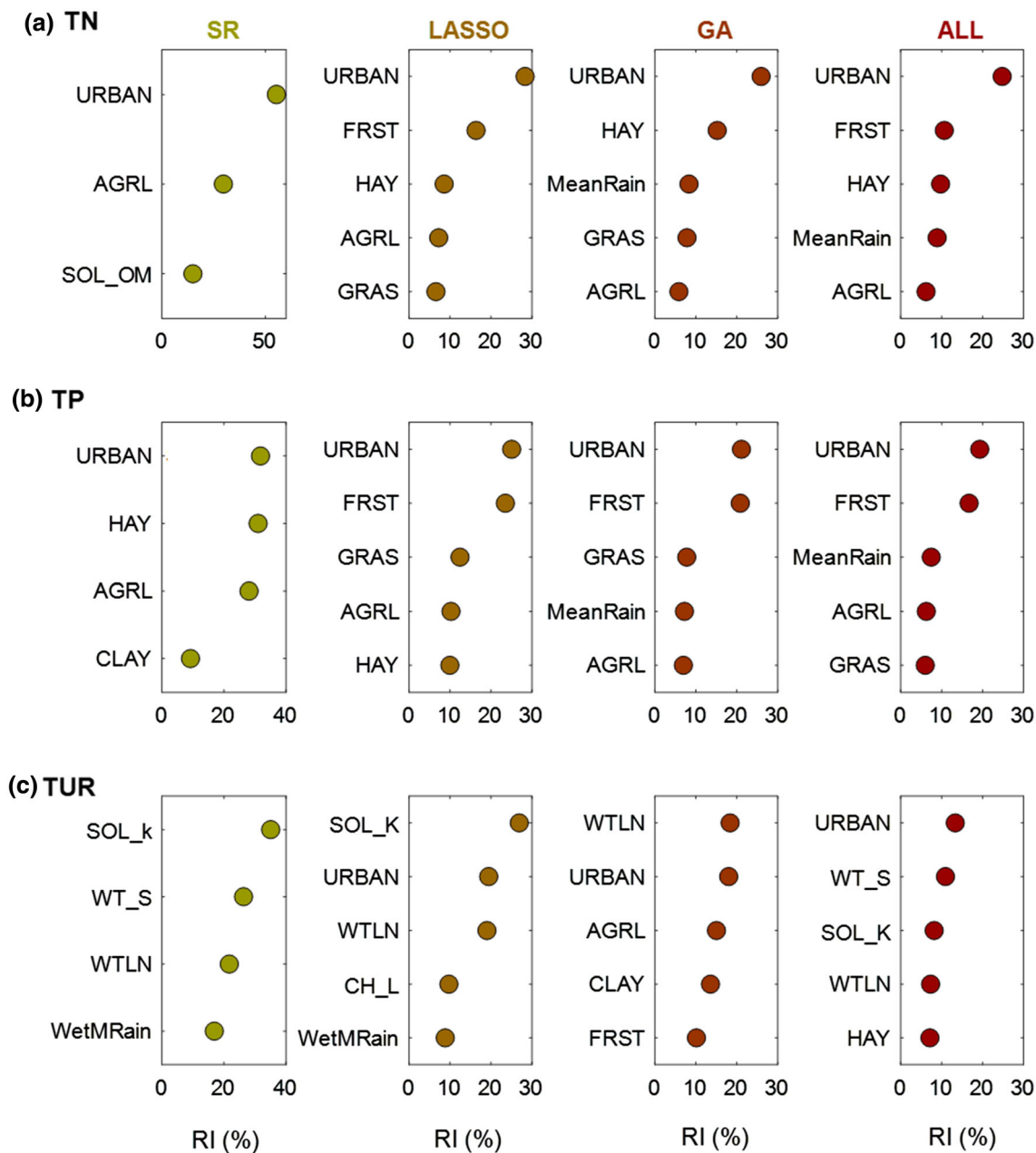


Fig. 8 The relative influence of the top 5 predictors of the median TN, TP, and TUR models based on the Boosted tree regression (BRT) algorithm

Vanderklein, 1996; Polsky et al., 2014; Tasdighi et al., 2017). These findings are expected and in agreement with previous studies that noted a positive correlation between the TN/TP and the percentage of the cropping and urbanization in the watershed (Agouridis et al., 2005; Pratt & Chang, 2012; Tasdighi et al., 2017; Wan et al., 2014; Wilson & Weng, 2010).

Additionally, WTLN lands appeared as a significant predictor for all TUR models and it has the highest relative importance value for most TUR models (Figs. 7 and 8). This is expected and it may be associated with wetlands

near streams which act as a sink for particulate matter (Cui et al., 2016; Shen et al., 2019; Suzuki et al., 2018). On the other hand, FRST and GRAS have higher relative importance for TN and TP predictions in most models (Figs. 7 and 8). On the other hand, the negative correlation between TN and TP with GRAS and FRST is expected as the GRAS, and FRST can potentially decrease nutrients in streams (Giri & Qiu, 2016; Tu & Xia, 2008).

Soil characteristics appear in all the models for TUR. For example, the proportion of clay in the watershed and the SOL_K appeared in the TUR models. When there is a

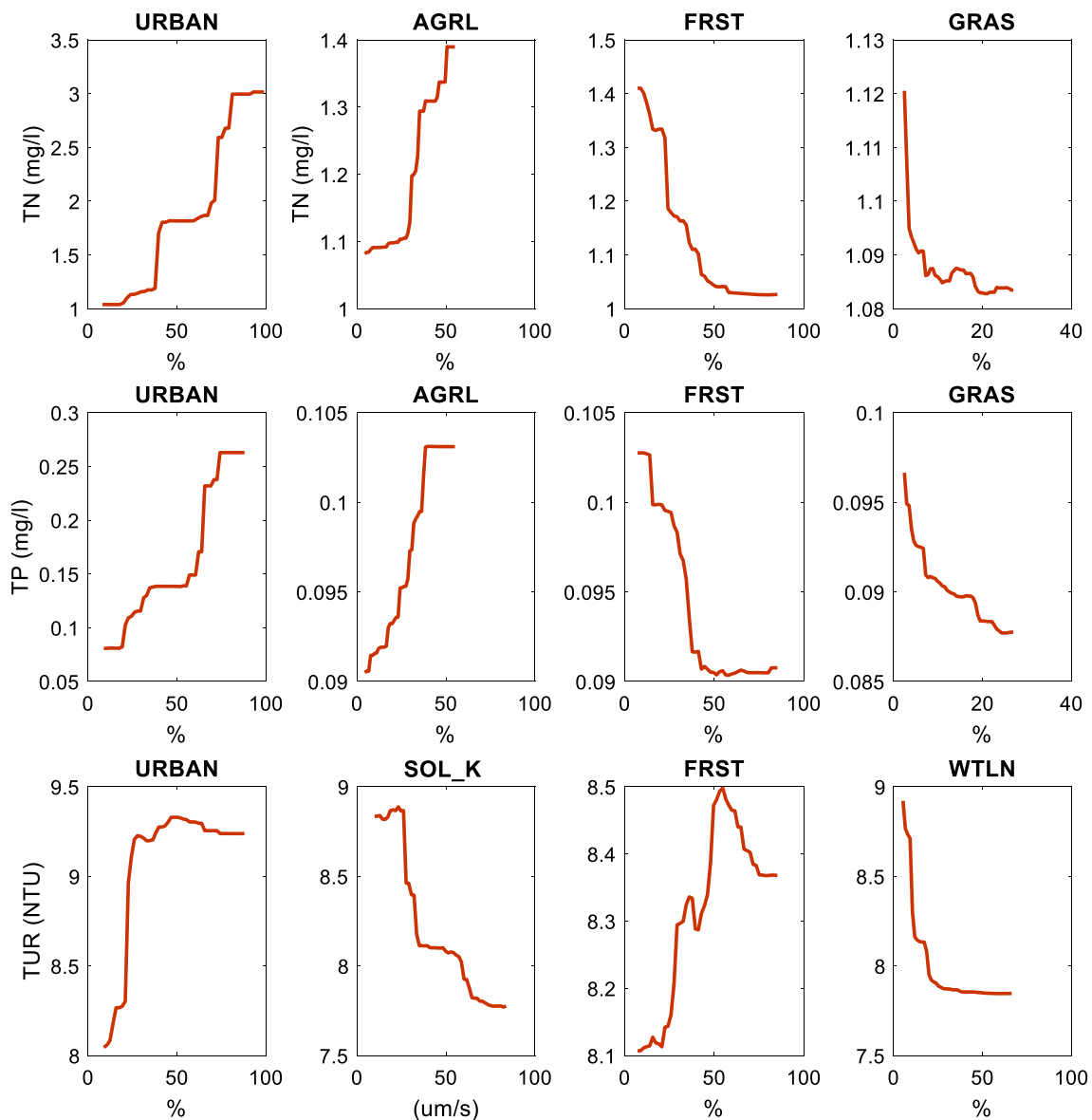


Fig. 9 Partial dependence plot based on Random Forests for TN, TP, and TUR in the streams

high percentage of clay and silt in soils, hydraulic conductivity (SOL_K) can be lower, leading to more runoff. Particulates are transported mainly from the watershed into streams by runoff. This high rate of runoff can lead to more particulates being transported over longer distances (Charlton, 2007; Wood, 1977), thus contributing to increased TUR and TP in streams. In addition, the positive relationship of TUR with the SOL_OM (organic matter) is expected, as many previous studies have indicated that organic matter can increase TUR in streams (Lenhart, 2008; Lenhart et al., 2010; Waters, 1995).

Moreover, the RF algorithm was easier to calibrate and robust to overfitting problems than BRT, which is partly associated with the bagging algorithm method that reduces the variance of the prediction model. These findings are

consistent with previous findings showing that RF performed better than BRT (Giri et al., 2019; Park & Kim, 2019; Shabani et al., 2017; Wang et al., 2018). For example, Park and Kim (2019) found that RF was slightly better than BRT in predicting landslide susceptibility mapping using different variables, such as topography and land use variables. Additionally, Shabani et al. (2017) showed that RF outperforms BRT when predicting the best location to distribute the date palm trees under different climate change scenarios. Overall, one of the advantages of using these machine learning algorithms (RF and BRT) compared to the traditional approaches (linear regression) is their ability to handle nonparametric datasets as well as nonlinear relationships (Grömping, 2009; Noi et al., 2017; Trawiński et al., 2012).

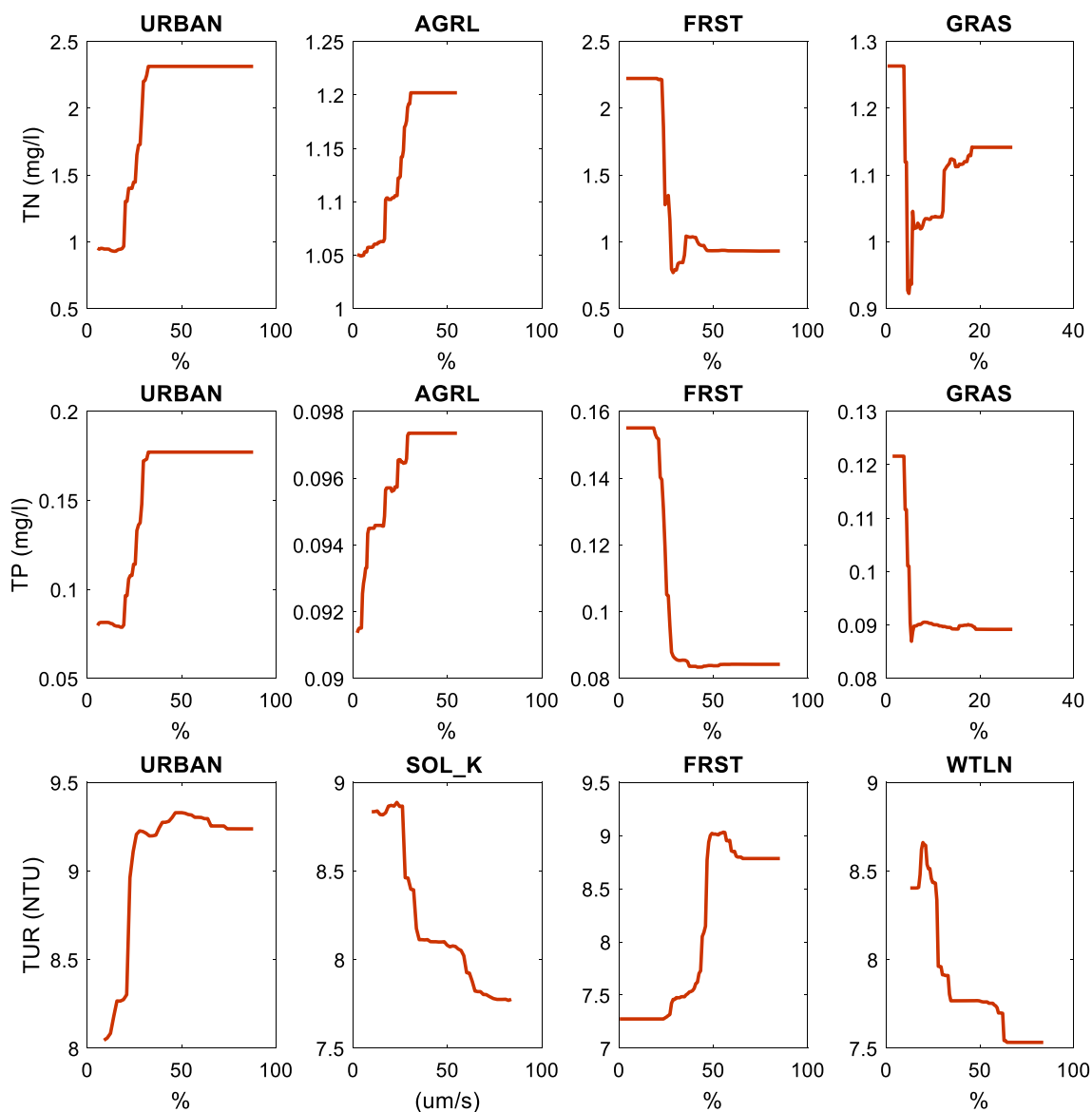


Fig. 10 Partial dependence plot based on Boosted tree regression for TN, TP, and TUR in the streams

Previous studies used stepwise regression (SR) to identify the most significant watershed characteristics influencing stream water quality (Hajigholizadeh & Melesse, 2017; Shrestha & Kazama, 2007; Wang et al., 2018). However, in this study, SR selected fewer predictors compared to LASSO and GA methods (Fig. 5) and did not perform well for RF and BRT models (Fig. 6). This may be due to the statistical deficiencies in the SR method, such as the distribution of test statistics, bias estimates, and standard error (Mo et al. 2016). Specifically, the regression error in the SR procedure follows the Gaussian distribution where the predictors and response variables are usually transformed into a Gaussian distribution. This may influence the interpretation of the regression coefficients (Hastie et al., 2017). More importantly, when solving a non-

convex optimization problem, the SR procedure often fails to find a global optimal set of variables and stays at a local optimum (Hastie et al., 2017). On the other hand, LASSO uses cross-validation to find predictors with the optimal generalization performance (Arlot & Celisse, 2010), which enhance the selection capability and providing better results compared to SR (Hammami et al., 2012). Overall, the model performance results showed that using GA models performed slightly better than LASSO and ALL models. These findings agree with Xie et al. (2015) and Wang et al. (2018), where the GA model was found to improve soil type recognition accuracy by 3–10%. These studies highlighted that SR models performed the worst, as it chooses a predictor based on the correlation's strength and ignoring interaction effects between predictors.

The models developed in this study can improve water-quality management decisions. The water-quality managers can implement the partial plots to identify the impact thresholds for different land use and watershed characteristics to formulate watershed regulation and find impaired water bodies that have not yet been assessed. Although this study focused on the Southeastern part of the United States, the methodology can be extended to other United States regions to evaluate the long-term median stream water quality.

7 Conclusion

Understanding the variability of water quality in rivers is essential to improve and predict water quality and environmental conditions in watersheds. Random forests and Boosted regression tree algorithms were evaluated to determine the most reliable model to predict the long-term median water quality indicators (TN, TP, and TUR). Different climatic and watershed characteristics across 97 watersheds located in the Southeastern of the US were used as predictor variables. The results showed that the random forests algorithm performed slightly better than boosted regression tree algorithm for predicting the median values of TN, TP, and TUR. The cross-validation results suggested that the prediction accuracy of the random forest explained 53%, 55%, 48% of variation in TN, TP, and TUR in streams, respectively. The RF algorithm was easy to train due to lesser user-defined parameters compared to BRT. Additionally, RF addressed the model overfitting issue slightly better than BRT as it uses a bagging algorithm that reduces the variance of the predictive function. Because of this, the relative importance of predictors (climatic and watershed characteristics) for the response variables (TN, TP, and TUR) was slightly different for both algorithms, leading to slight differences in model predictability.

The results also highlighted the importance of forest and grasslands within a watershed to sustain healthy streams. Identifying a threshold can help water quality watershed managers develop watershed regulations or design a restoration program based on scientific criteria. While the partial plots can be useful to identify key variables to enhance stream water quality management, additional research is needed to evaluate the different hotspots ((e.g., septic tanks, industries, biogeochemical hotspots, and the distance of pollutant sources from the streams) within the watersheds on the long term spatio-temporal water quality changes.

Acknowledgements The authors would like to thank the Ministry of Education in Saudi Arabia for providing funding through a student scholarship.

Credit Author Statement Ali O. Alnahit: Formal analysis, Investigation, Writing—original draft. Ashok Mishra: Conceptualization, Writing—review & editing. Abdul Khan: Writing—review & editing.

Declaration

Conflict of interest We have no conflict of interest to report.

References

- Agouridis CT, Workman SR, Warner RC, Jennings GD (2005) Livestock grazing management impacts on stream water quality: a review 1. *JAWRA J Am Water Resour Assoc* 41(3):591–606
- Allan JD (2004) Landscapes and riverscapes: the influence of land use on stream ecosystems. *Annu Rev Ecol Evol Syst* 35:257–284
- Alnahit AO, Mishra AK, Khan AA (2020) Quantifying climate, streamflow, and watershed control on water quality across Southeastern US watersheds. *Sci Total Environ* 139945
- Alpaydin E (2020) Introduction to machine learning. MIT Press, Cambridge
- Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. *Neural Comput* 9(7):1545–1588
- Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. *Stat Surv* 4:40–79
- Arnold JG, Moriasi DN, Gassman PW, Abbaspour KC, White MJ, Srinivasan R, Santhi C, Harmel RD, van Griensven A, van Liew MW (2012) SWAT: Model use, calibration, and validation. *Trans ASABE* 55(4):1491–1508
- Bardsley WE, Vetrova V, Liu S (2015) Toward creating simpler hydrological models: A LASSO subset selection approach. *Environ Model Softw* 72:33–43
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Bucak T, Trolle D, Tavşanoğlu ÜN, Çakıroğlu Aİ, Özen A, Jeppesen E, Beklioğlu M (2018) Modeling the effects of climatic and land use changes on phytoplankton and water quality of the largest Turkish freshwater lake: Lake Beyşehir. *Sci Total Environ* 621:802–816
- Bui DT, Khosravi K, Tiefenbacher J, Nguyen H, Kazakis N (2020) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci Total Environ* 721:137612
- Castela J, Ferreira V, Graça MAS (2008) Evaluation of stream ecological integrity using litter decomposition and benthic invertebrates. *Environ Pollut* 153(2):440–449
- Castrillo M, García ÁL (2020) Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. *Water Res* 172:115490
- Charlton R (2007) Fundamentals of fluvial geomorphology. Routledge, Milton Park
- Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, Liu F, Zuo M, Zou X, Wang J (2020) Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res* 171:115454
- Correll DL (1999) Phosphorus: a rate limiting nutrient in surface waters. *Poult Sci* 78(5):674–682
- Cui B, He Q, Gu B, Bai J, Liu X (2016) China's coastal wetlands: understanding environmental changes and human impacts for management and conservation. Springer, Berlin

- Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int J Climatol A J R Meteorol Soc* 28(15):2031–2064
- De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81(11):3178–3192
- Dillon PJ, Kirchner WB (1975) The effects of geology and land use on the export of phosphorus from watersheds. *Water Res* 9(2):135–148
- Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77(4):802–813
- Esri (2014) *ArcGIS 10.3. 1 for Desktop*. Environmental Systems Research Institute Redlands, CA USA.
- Everingham Y, Sexton J, Skocaj D, Inman-Bamber G (2016) Accurate prediction of sugarcane yield using a random forest algorithm. *Agron Sustain Dev* 36(2):27
- Fang X, Li X, Zhang Y, Zhao Y, Qian J, Hao C, Zhou J, Wu Y (2021) Random forest-based understanding and predicting of the impacts of anthropogenic nutrient inputs on the water quality of a tropical lagoon. *Environ Res Lett* 16(5):055003
- Fatehi I, Amiri BJ, Alizadeh A, Adamowski J (2015) Modeling the relationship between catchment attributes and in-stream water quality. *Water Resour Manage* 29(14):5055–5072
- Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. *Stat Med* 22(9):1365–1381
- Genuer R, Poggi J-M, Tuleau-Malot C (2010) Variable selection using random forests. *Pattern Recogn Lett* 31(14):2225–2236
- Giri S, Qiu Z (2016) Understanding the relationship of land uses and water quality in twenty first century: a review. *J Environ Manage* 173:41–48
- Giri S, Zhang Z, Krasnuk D, Lathrop RG (2019) Evaluating the impact of land uses on stream integrity using machine learning algorithms. *Sci Total Environ* 696:133858
- Golden HE, Lane CR, Prues AG, D'Amico E (2016) Boosted regression tree models to explain watershed nutrient concentrations and biological condition. *JAWRA J Am Water Resour Assoc* 52(5):1251–1274
- Granger SJ, Bol R, Anthony S, Owens PN, White SM, Haygarth PM (2010) Towards a holistic classification of diffuse agricultural water pollution from intensively managed grasslands on heavy soils. *Adv Agron* 105:83–115
- Grömping U (2009) Variable importance assessment in regression: linear regression versus random forest. *Am Stat* 63(4):308–319
- Hair JF, Ringle CM, Sarstedt M (2013) Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance. *Long Range Plan* 46(1–2):1–12
- Hajigholizadeh M, Melesse AM (2017) Assortment and spatiotemporal analysis of surface water quality using cluster and discriminant analyses. *CATENA* 151:247–258
- Hammami D, Lee TS, Ouarda TBMJ, Lee J (2012) Predictor selection for downscaling GCM data with LASSO. *J Geophys Res Atmosph* 117(D17)
- Harrell FE (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis, vol 608. Springer, New York
- Hastie T, Tibshirani R, Tibshirani R J (2017) Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *ArXiv Preprint* <https://arxiv.org/abs/1707.08692>
- Hecky RE, Kilham P (1988) Nutrient limitation of phytoplankton in freshwater and marine environments: a review of recent evidence on the effects of enrichment 1. *Limnol Oceanogr* 33(4part2):796–822
- Huang S, Huang Q, Leng G, Liu S (2016) A nonparametric multivariate standardized drought index for characterizing socioeconomic drought: A case study in the Heihe River Basin. *J Hydrol* 542:875–883
- Ice G, Binkley D (2003) Forest streamwater concentrations of nitrogen and phosphorus: A comparison with EPA's proposed water quality criteria. *J Forest* 101(1):21–28
- Jeung M, Baek S, Beom J, Cho KH, Her Y, Yoon K (2019) Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. *J Hydrol* 575:1099–1110
- Johnson SL, Ringler NH (2014) The response of fish and macroinvertebrate assemblages to multiple stressors: A comparative analysis of aquatic communities in a perturbed watershed (Onondaga Lake, NY). *Ecol Ind* 41:198–208
- Kang J-H, Lee SW, Cho KH, Ki SJ, Cha SM, Kim JH (2010) Linking land-use type and stream water quality using spatial data of fecal indicator bacteria and heavy metals in the Yeongsan river basin. *Water Res* 44(14):4143–4157
- Knierim KJ, Kingsbury JA, Haugh CJ, Ransom KM (2020) Using boosted regression tree models to predict salinity in Mississippi Embayment aquifers, central United States. *JAWRA J Am Water Resour Assoc* 56(6):1010–1029
- Ko BC, Kim HH, Nam JY (2015) Classification of potential water bodies using Landsat 8 OLI and a combination of two boosted random forest classifiers. *Sensors* 15(6):13763–13777
- Konapala G, Mishra A (2020) Quantifying climate and catchment control on hydrological drought in the continental United States. *Water Resour Res* 56(1):e2018WR024620
- Lenhart CF (2008) The influence of watershed hydrology and stream geomorphology on turbidity, sediment and nutrients in tributaries of the Blue Earth River, Minnesota, USA. University of Minnesota
- Lenhart CF, Brooks KN, Heneley D, Magner JA (2010) Spatial and temporal variation in suspended sediment, organic matter, and turbidity in a Minnesota prairie river: implications for TMDLs. *Environ Monit Assess* 165(1):435–447
- Lima AR, Cannon AJ, Hsieh WW (2016) Forecasting daily streamflow using online sequential extreme learning machines. *J Hydrol* 537:431–443
- Lintern A, Webb JA, Ryu D, Liu S, Bende-Michl U, Waters D, Leahy P, Wilson P, Western AW (2018) Key factors influencing differences in stream water quality across space. *Wiley Interdiscipl Rev Water* 5(1):e1260
- Mainali J, Chang H (2018) Landscape and anthropogenic factors affecting spatial patterns of water quality trends in a large river basin, South Korea. *J Hydrol* 564:26–40
- Mattsson T, Kortelainen P, Räsänen A (2005) Export of DOM from boreal catchments: impacts of land use cover and climate. *Biogeochemistry* 76(2):373–394
- Mishra A, Alnahit A, Campbell B (2020) Impact of land uses, drought, flood, wildfire, and cascading events on water quality and microbial communities: A review and analysis. *J Hydrol* 125707
- Mishra AK, Desai VR (2005) Drought forecasting using stochastic models. *Stoch Env Res Risk Assess* 19(5):326–339
- Moriasi DN, Gitau MW, Pai N, Daggupati P (2015) Hydrologic and water quality models: Performance measures and evaluation criteria. *Trans ASABE* 58(6):1763–1785
- Mouazen AM, Kuang B, de Baerdemaeker J, Ramon H (2010) Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma* 158(1–2):23–31
- Mo W, Wang H, Jacobs JM (2016) Understanding the influence of climate change on the embodied energy of water supply. *Water Res* 95:220–229

- Naghibi SA, Pourghasemi HR, Dixon B (2016) GIS-based ground-water potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ Monit Assess* 188(1):1–27
- Noi PT, Degener J, Kappas M (2017) Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. *Remote Sens* 9(5):398
- Onderka M, Wrede S, Rodný M, Pfister L, Hoffmann L, Krein A (2012) Hydrogeologic and landscape controls of dissolved inorganic nitrogen (DIN) and dissolved silica (DSi) fluxes in heterogeneous catchments. *J Hydrol* 450:36–47
- Paerl HW (1988) Nuisance phytoplankton blooms in coastal, estuarine, and inland waters 1. *Limnol Oceanogr* 33(4part2):823–843
- Park S, Kim J (2019) Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. *Appl Sci* 9(5):942
- Perry JA, Vanderklein E (1996) *Water Q Natural Resour Manage*
- Polsky C, Grove JM, Knudson C, Groffman PM, Bettez N, Cavender-Bares J, Hall SJ, Heffernan JB, Hobbie SE, Larson KL (2014) Assessing the homogenization of urban land management with an application to US residential lawn care. *Proc Natl Acad Sci* 111(12):4432–4437
- Povak NA, Hessburg PF, McDonnell TC, Reynolds KM, Sullivan TJ, Salter RB, Cosby BJ (2014) Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. *Water Resour Res* 50(4):2798–2814
- Pratt B, Chang H (2012) Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. *J Hazard Mater* 209:48–58
- Puissant A, Rougier S, Stumpf A (2014) Object-oriented mapping of urban trees using Random Forest classifiers. *Int J Appl Earth Obs Geoinf* 26:235–245
- Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015) Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol Rev* 71:804–818
- Seber GAF, Lee AJ (2012) *Linear regression analysis (Vol 329)*. Wiley, New York
- Shabani F, Kumar L, Solhjoui-Fard S (2017) Variances in the projections, resulting from CLIMEX, Boosted Regression Trees and Random Forests techniques. *Theoret Appl Climatol* 129(3):801–814
- Shen G, Yang X, Jin Y, Xu B, Zhou Q (2019) Remote sensing and evaluation of the wetland ecological degradation process of the Zoige Plateau Wetland in China. *Ecol Ind* 104:48–58
- Shen LQ, Amatulli G, Sethi T, Raymond P, Domisch S (2020) Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. *Scientific Data* 7(1):1–11
- Shrestha S, Kazama F (2007) Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environ Modell Softw* 22(4):464–475
- Singh B, Sihag P, Singh K (2017) Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Model Earth Syst Environ* 3(3):999–1004
- South Carolina Department of Health and Environmental Control, *Watershed Water Quality Assessment* (2016)
- Smucker NJ, Becker M, Detenbeck NE, Morrison AC (2013) Using algal metrics and biomass to evaluate multiple ways of defining concentration-based nutrient criteria in streams and their ecological relevance. *Ecol Ind* 32:51–61
- Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. *BMC Bioinform* 9(1):1–11
- Suzuki J, Imamura M, Nakano D, Yamamoto R, Fujita M (2018) Effects of water turbidity and different temperatures on oxidative stress in caddisfly (*Stenopsyche marmorata*) larvae. *Sci Total Environ* 630:1078–1085
- Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R (2016) Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. *Geoderma* 266:98–110
- Tasdighi A, Arabi M, Osmond DL (2017) The relationship between land use and vulnerability to nitrogen and phosphorus pollution in an urban watershed. *J Environ Qual* 46(1):113–122
- Tong STY, Chen W (2002) Modeling the relationship between land use and surface water quality. *J Environ Manage* 66(4):377–393
- Tonkin JD, Stoll S, Sundermann A, Haase P (2014) Dispersal distance and the pool of taxa, but not barriers, determine the colonisation of restored river reaches by benthic invertebrates. *Freshw Biol* 59(9):1843–1855
- Tramblay Y, Ouarda TBMJ, St-Hilaire A, Poulin J (2010) Regional estimation of extreme suspended sediment concentrations using watershed characteristics. *J Hydrol* 380(3–4):305–317
- Trawiński B, Smętek M, Telec Z, Lasota T (2012) Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *Int J Appl Math Comput Sci* 22:867–881
- 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
- Tung TM, Yaseen ZM (2021) Deep learning for prediction of water quality index classification: tropical catchment environmental assessment. *Natural Resour Res*, 1–20
- Turner MG, Ruscher CL (1988) Changes in landscape patterns in Georgia, USA. *Landscape Ecol* 1(4):241–251
- US EPA (2002) *National Recommended Water Quality Criteria: 2002*. Office of Water, EPA-822-R-02-047, US Environmental Protection Agency, Washington DC. <http://www.epa.gov/water/science/standards/wqcriteria.html>.
- Varanka S, Hjort J, Luoto M (2015) Geomorphological factors predict water quality in boreal rivers. *Earth Surf Proc Land* 40(15):1989–1999
- Varanka S, Luoto M (2012) Environmental determinants of water quality in boreal rivers based on partitioning methods. *River Res Appl* 28(7):1034–1046
- Vazifehkhah S, Tosunoglu F, Kahya E (2019) Bivariate risk analysis of droughts using a nonparametric multivariate standardized drought index and copulas. *J Hydrol Eng* 24(5):05019006
- Veettil, A. V., & Mishra, A. (2020). Water security assessment for the contiguous United States using water footprint concepts. *Geophysical Research Letters*, 47(7), e2020GL087061.
- Vorpahl P, Elsenbeer H, Märker M, Schröder B (2012) How can statistical models help to determine driving factors of landslides? *Ecol Model* 239:27–39
- Waite IR, Brown LR, Kennen JG, May JT, Cuffney TF, Orlando JL, Jones KA (2010) Comparison of watershed disturbance predictive models for stream benthic macroinvertebrates for three distinct ecoregions in western US. *Ecol Ind* 10(6):1125–1136
- Walsh CJ, Roy AH, Feminella JW, Cottingham PD, Groffman PM, Morgan RP (2005) The urban stream syndrome: current knowledge and the search for a cure. *J N Am Benthol Soc* 24(3):706–723
- Walsh CJ, Webb JA (2016) Interactive effects of urban stormwater drainage, land clearance, and flow regime on stream macroinvertebrate assemblages across a large metropolitan region. *Freshwater Sci* 35(1):324–339

- Wang B, Waters C, Orgill S, Cowie A, Clark A, Li Liu D, Simpson M, McGowen I, Sides T (2018) Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. *Ecol Ind* 88:425–438
- Wang F, Wang Y, Zhang K, Hu M, Weng Q, Zhang H (2021a) Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. *Environ Res* 202:111660
- Wang R, Kim J-H, Li M-H (2021b) Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Sci Total Environ* 761:144057
- Wan N-F, Gu X-J, Ji X-Y, Jiang J-X, Wu J-H, Li B (2014) Ecological engineering of ground cover vegetation enhances the diversity and stability of peach orchard canopy arthropod communities. *Ecol Eng* 70:175–182
- Waters TF (1995) Sediment in streams: sources, biological effects, and control
- Welikala RA, Fraz MM, Dehmeshki J, Hoppe A, Tah V, Mann S, Williamson TH, Barman SA (2015) Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. *Comput Med Imaging Graph* 43:64–77
- Wilson C, Weng Q (2010) Assessing surface water quality and its relation with urban land cover changes in the Lake Calumet Area, Greater Chicago. *Environ Manage* 45(5):1096–1111
- Wood PA (1977) Controls of variation in suspended sediment concentration in the River Rother, West Sussex, England. *Sedimentology* 24(3):437–445
- Xie H, Zhao J, Wang Q, Sui Y, Wang J, Yang X, Zhang X, Liang C (2015) Soil type recognition as improved by genetic algorithm-based variable selection using near infrared spectroscopy and partial least squares discriminant analysis. *Sci Rep* 5(1):1–10
- Yang R-M, Zhang G-L, Liu F, Lu Y-Y, Yang F, Yang F, Yang M, Zhao Y-G, Li D-C (2016) Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecol Ind* 60:870–878
- Young RG, Quarterman AJ, Eyles RF, Smith RA, Bowden WB (2005) Water quality and thermal regime of the Motueka River: influences of land cover, geology and position in the catchment. *NZ J Mar Freshwat Res* 39(4):803–825
- Yu M, Li Q, Hayes MJ, Svoboda MD, Heim RR (2014) Are droughts becoming more frequent or severe in China based on the standardized precipitation evapotranspiration index: 1951–2010? *Int J Climatol* 34(3):545–558
- Zampella RA, Procopio NA, Lathrop RG, Dow CL (2007) Relationship of land-use/land-cover patterns and surface-water quality in the mullica river basin 1. *JAWRA J Am Water Resour Assoc* 43(3):594–604

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.