Accepted refereed manuscript of: Xu G, Fan H, Oliver DM, Dai Y, Li H, Shi Y, Long H, Xiong K & Zhao Z (2022) Decoding river pollution trends and their landscape determinants in an ecologically fragile karst basin using a machine learning model. *Environmental Research*, 214, Art. No.: 113843. https://doi.org/10.1016/j.envres.2022.113843

Decoding river pollution trends and their landscape determinants in an ecologically fragile karst basin using a machine learning model

Hightlight

Spatial and temporal patterns of river water quality in Wu jiang River basin (WRB) were analyzed from 2014 to 2019

Machine learning model (XGBoost) was developed to predict robust spatially-distributed continuous water quality patterns

SHAP was used as a powerful model interpreter to decode the black box of a ML model indicating the drivers of water quality deterioration

Geological and climatic vulnerabilities drive management decisions for control of pollution in these critical areas

Abstract

Karst watersheds accommodate high landscape complexity and are influenced by both human-induced and natural activity, which affects the formation and process of runoff, sediment connectivity and contaminant transport and alters natural hydrological and nutrient cycling. However, physical monitoring stations are costly and labor-intensive, which has confined the assessment of water quality impairments on spatial scale. The geographical characteristics of catchments are potential influencing factors of water quality, often overlooked in previous studies of highly heterogeneous karst landscape. To solve this problem, we developed a machining learning method and applied Extreme Gradient Boosting (XGBoost) to predict the spatial distribution of water quality in the world's most ecologically fragile karst watershed. We used the Shapley Addition interpretation (SHAP) to explain the potential determinants. Before this process, we first used the water quality damage index (WQI-DET) to evaluate the water quality impairment status and determined that COD_{Mn}, TN and TP were causing river water quality impairments in the WRB. Second, we selected 46 watershed features based on the three key processes (sources-mobilization-transport) which affect the temporal and spatial variation of river pollutants to predict water quality

in unmonitored reaches and decipher the potential determinants of river impairments. The predicting range of COD_{Mn} spanned from 1.39 mg/L to 17.40 mg/L. The predictions of TP and TN ranged from 0.02 to 1.31 mg/L and 0.25 to 5.72 mg/L, respectively. In general, the XGBoost model performs well in predicting the concentration of water quality in the WRB. SHAP explained that pollutant levels may be driven by three factors: anthropogenic sources (agricultural pollution inputs), fragile soils (low organic carbon content and high soil permeability to water flow), and pollutant transport mechanisms (TWI, carbonate rocks). Our study provides key data to support decision-making for water quality restoration projects in the WRB and information to help bridge the science:policy gap.
Keywords: Ecologically fragile karst basin; Water quality assessment; XGBoost regression; Shapley additive explanations; Determinant analysis.

Introduction

Anthropogenic interferences have dramatically hindered natural hydrological and nutrient cycles, in turn threatening river water quality in many countries and regions across the world (Mandaric et al., 2018; Ockenden et al., 2017). Controlling pollutant emissions has become the focus of many global environmental policies (Cardinale, 2011; Mandaric et al., 2018; Vorosmarty and Sahagian, 2000). Water quality can be impacted by anthropogenic factors (such as the land use and land cover changes) (Baker, 2003; Liu et al., 2018; Yan et al., 2021a). Urbanization has led to an increase in impervious surfaces, which alters hydrological flow paths and deliver pollutants to the river network more efficiently resulting in additional pressure and degradation of river water quality (Marinoni et al. 2013). Intensification of agricultural activities may result in increased nutrient loads due to fertilization and changes in surface soil properties.

Geographical factors (e.g. climate change, atmospheric deposition, geology and topography, soil types, catchment hydrology, land use/cover and land management) are summarized as three key process factors (i.e., sources, mobilisation and delivery) that define how determinants spatially and temporally affect water quality in a watershed (Alvarez-Cabria et al., 2016a; Fan and Shibata, 2015; Heckmann and Schwanghart, 2013; Lintern et al., 2018; Liu et al., 2021; Noori et al., 2012; Varanka et al., 2015). Identifying the influences of watershed geographical characteristics on river water quality is helpful to understand the evolution of river ecosystem in this region because these key factors vary widely across different geographical regions (Liu et al., 2021; Mainali and Chang, 2018).

Karst is globally distributed landscape and supports approximately 20% of the world's population (Ford and Williams, 2007; Hartmann et al., 2014). The Wu jiang River basin (WRB) is located in the world's largest continuous landscapes of karst, which is deemed as an ecological barrier for the Yangtze River Basin and also defined as one of the most ecologically fragile regions in the world (Xu et al.,2021). Land use patterns (e.g. sloped planting and overgrazing) interact with the heterogeneous karst landscape composition and configuration in complex ways (Varanka et al., 2015; Xu et al., 2019). As a result, karst landscapes are more fragile and this can influence the formation and processes of runoff, sediment connectivity and the delivery of pollutants from land to water (Ai et al., 2015; Heckmann and Schwanghart, 2013; Yan et al., 2021b). Further, hydrological processes in karst landscapes deviate from typical responses in non-karst environments and this can lead to river water quality impairments differentiating from those of the plains (Deng, 2020; Liu et al., 2020). Due to the influence of subtropical humid monsoon, rainfall in this region is seasonally unevenly distributed and heavy rainfall events are a frequent occurrence, which exacerbates the mobilization and transport of pollutants (Powers et al., 2016; Singh et al., 2005a; Sinha and Michalak, 2016). In recent years, the nutrient balance of the WRB has become a controversial issue due to the construction of cascade dams (Li and Ji, 2016; Winemiller et al., 2016). The geographical characteristics and human disturbance of WRB lead to serious water pollution and complex environmental response in karst areas.

Field assessments of water quality can support catchment managers and stakeholders in identifying spatio-temporal sensitive areas of managed landscapes and help to evaluate the benefits and risks of water management strategies in priority areas (Altenburger et al., 2015; Huang et al., 2021; Yi et al., 2017). However, most water quality assessments are limited to particular river reaches due to the costs associated with data collection; therefore, many low-order streams are not evaluated, which can limit understanding of water quality challenges in a watershed (Altenburger et al., 2015; Ding et al., 2016; Mello et al., 2018). Thus, physical process-based model simulation can complement field monitoring investigations. Models, e.g., HSPF and HYPE, SWAT, AGNPS or semi-distributed process based model SPARROW or INCA can simulate complex nonlinear interactions between nutrient transport dynamics and biogeochemical processes (Arhonditsis et al., 2007; Hashemi et al., 2016; Mayorga et al., 2010; Singh et al., 2005b). Such physical process-based models often preclude the identification of dominant processes operating within a watershed due to uncertainties associated with parameter calibration across a large watershed (Badham et al., 2019; Jakeman et al., 2006; Knoben et al., 2020). The complexity of environmental processes often results in physical process-based models being costly and labor-intensive inputs of dataset collection. Moreover, the karst zone under the thin soil layer in karst regions has high permeability and accommodates a complex subsurface hydrological system which makes the parameterization of such models difficult and hinders the transferability of mechanical process approaches to karst areas (Fiorillo et al., 2015; Hartmann et al., 2015; Li et al., 2021; Malago et al., 2016). On the contrary, data-driven machine learning (ML) models are recognized as an effective alternative method and offer advantages for modeling complex nonlinear systems over deterministic and statistical models when handling multi-source data for prediction of river water quality due to improved model interpretability, prediction accuracy, and reduced computational cost (Najah Ahmed et al., 2019; Sun and Scanlon, 2019; Wang et al., 2021b; Zou et al., 2019).

As an optimized distributed gradient lifting algorithm, Extreme Gradient Boosting (XGBoost) delivers high accuracy and fast processing time (Lundberg et al., 2020). Indeed, XGBoost outperformed several other machine learning techniques (e.g., Gradient Boosting and Deep Neural Network, Bayesian Regularized Neural Network and Random Forest algorithm) to predict probabilities, and is especially used when dealing with spatial data (Just et al., 2020; Mokoatle et al., 2019). Tree-based ML models are often regarded as unexplainable black box models (Moreira et al., 2020; Parsa et al., 2020). However, data-driven machine learning models suffer from several drawbacks. First, ML models often require a large amount of training data to obtain robust performance (Kratzert et al., 2019). Second they are still not as easily interpretable as traditionally-used physics-based conceptual hydrologic models (Höge et al., 2022). Shapley Additive Explanations (SHAP) is considered as a state-of-the-art model interpretation to decode the black-box of ML models. It can connect optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (Adadi and Berrada, 2018; Lundberg and Lee, 2017; Molnar, 2020). This helps to understand the magnitude and direction of the influence of input variables on the output variable.

The overarching aim of our study, therefore, was to investigate the spatial distribution of river water quality impairments in the WRB and decipher how watershed features, both anthropogenic and natural, impair water quality. Firstly, we compiled a complete time series trend (2014–2019) dataset of river water quality (14,845 records from 207 water quality sampled sites) to identify spatio-temporal water quality impairments and screen the key variables that contribute to water quality impairment in the WRB. We hypothesized that watershed landscape attributes are important in interpreting water quality at different regional scales in the WRB. Then a powerful ML method was developed to predict water pollution concentrations in unmonitored reaches and we used the SHAP value to determine the significant landscape covariates of water quality in the WRB.

2. Materials and methods

2.1 Study area

The WRB (25°39′13″~25°41′00″N, 105°36′30″-105°46′30″E) is located in southwest China, Guizhou province, which comprises a total area of 80300 km^2, see Fig.1. Though located in an area of subtropical humid monsoon climate, with average annual precipitation of 1300 mm, there is a serious shortage of clean drinking water for people and livestock (Qin et al., 2015). The WRB provides agricultural irrigation, urban development, river navigation and other functions for more than 35 million people in 54 counties in Guizhou Province (Xu et al., 2021b). Due to the local government promoting strict farmland protection policy, the cultivated land in the WRB have remained stable during the period of the Outline of the 12th Five-Year Plan and the 13th Five-year Plan for National Economic and Social Development of the People's Republic of China. The total use of fertilizers has increased significantly in accordance with the increase of grain demand (Li et al., 2020a; Oliver et al., 2020), together with runoff and infiltration of pollutants leading to a serious crisis of the river water quality in the WRB (Li et al., 2020a; Xu et al., 2021b). Due to the slow soil formation of carbonate rocks, large landform slopes, low vegetation cover, water and soil conservation is at risk from natural disasters and poor approaches to agricultural production have exacerbated soil erosion and rock desertification (Xu et al., 2021b). Because the original surface vegetation was mostly destroyed, the vegetation of the region is mainly a secondary forest, consisting of subtropical evergreen and deciduous broad-leafed mixed trees, mainly composed of species of ***genera Cyclobalanopsis*, *Pinus, Betula, and Cupressus***(Sheng et al., 2018).The bedrock of the WRB is mainly composed of carbonate rocks, dolomite, and limestone micaceous(Han and Liu, 2004), and the main soil types are yellow loam, paddy soil, and calcareous soil. The hydrogeological conditions are complex due to the unique geology of the region, which has contributed to mature underground rivers. The region is covered by shallow soil (<1 m), mainly composed of lime developed from dolomite (>50%), some of which is mineralized and some of which is presented in loose form (Nie et al., 2017). The soil types in the WRB are more permeable soils type A and B and may result in higher water tables and accelerated nutrient flow to the soil (Rodriguez-Galiano et al., 2014). Their texture is silty loam, sandy soils have higher porosity and therefore lower water retention, resulting in lower absorption of pollutants such as pesticides, metal ions and solutes. In addition to the aerated structure and inadequate bonding of humus to sand grains, these properties preferentially allow the infiltration of water and associated contaminants (Andry et al., 2009; Zalidis et al., 2002).

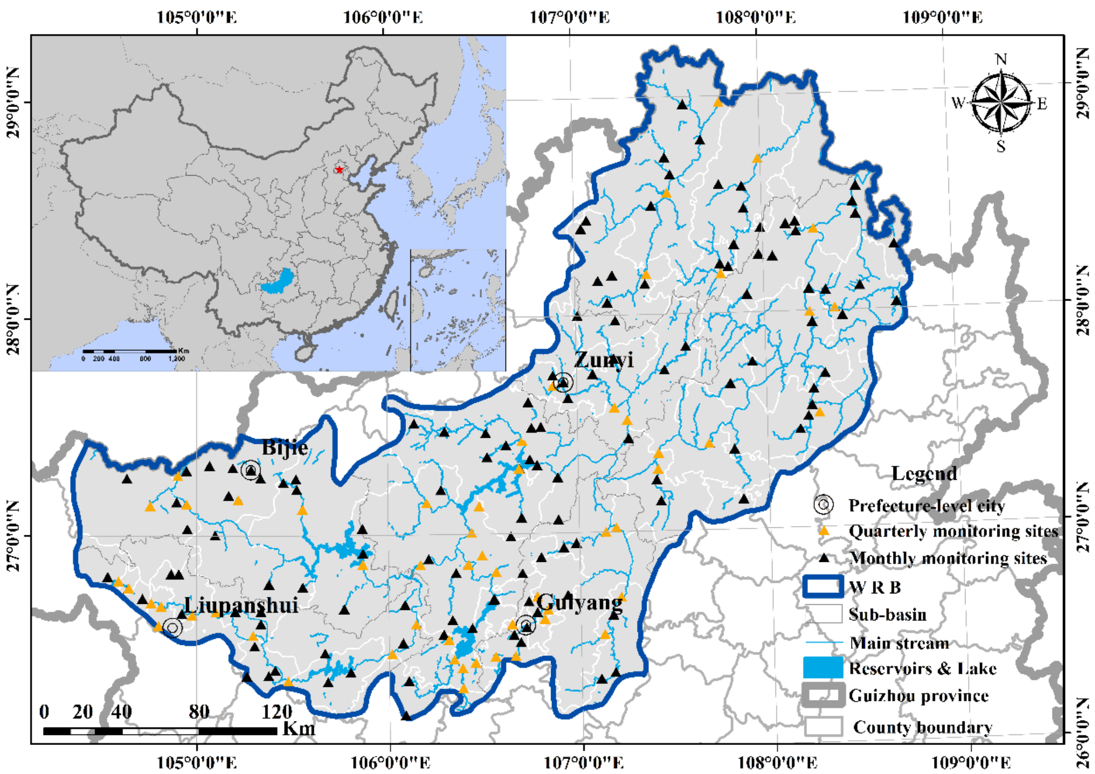


Fig.1. The monitoring river networks and overview of WRB

2.2 Data resources and pre-processing

2.2.1 Water quality data resources

We applied a complete time series water dataset from 207 surface water quality monitoring sections sampled by the Bureau of Water Resources Department and Environmental Protection Department in Guizhou province respectively. A monthly sampling frequency is used for national-controlled and provincial-controlled water quality sections are once per month, while that of water functional areas is once per quarter. The time scales were across from 2014 to 2019. All the indicators were collected by mixed samples and tested in laboratory. The Sample pretreatment and pollutant concentration determination methods are mainly based on "Environmental Quality Standard for Surface Water" (GB3838-2002), and specific detection methods are presented in the supplementary material ST3. Therefore, the complete time series pollution indexes in the WRB were selected including Ammonia nitrogen (NH_4^+-N), Total phosphorus (TP), Five-day biochemical oxygen demand (BOD_5), Dissolved oxygen (DO), Anionic surfactant (AS), Temperature (T), Hydrogen ion concentration index (pH), Electrical conductivity (EC), total nitrogen (TN), sulfide (SO_4^{2-}), Potassium permanganate index (COD_{Mn}). These 11 indexes were used as water environment dataset in our study. We set the measured value below the detection limit as the detection limit (Farnham et al., 2002).

2.2.2 Meteorological and landscape data sources and data pre-processing

We developed an integrated database including 46 watershed landscape characteristics contributing to the spatial variability of river pollution according to three key driving processes (sources, mobilization and delivery) put forward by multiple studies (Granger et al., 2010; Hrachowitz et al., 2016; Lintern et al., 2018). More details are shown in supplementary material Table.S1. Then we used the hydrological tool "Burn-in" method in ArcGIS 10.2 software with a watershed pixel threshold of 15,000 to delineate 792 reaches and sub-basins of the WRB, and their distribution was adjusted to be more correct according to satellite images of the basin. The topographic wetness index (TWI) was calculated by using the 8-flow method proposed by Quinn (Gruber and Peckham, 2009; Quinn et al., 1991). The grid soil permeability data was computed by the Python-ROSSETA model according to soil texture (Zhang et al., 2018). The average annual streamflow of the reaches is modeled based on the water balance Budyko model (Zhang et al., 2004) through annual predication and potential evapotranspiration, and all the parameters are participated in calculation referred to (Dai et al., 2021). The landscape index is calculated from Fragstats 4.3. The distance between landslide geological disaster points and water body is analyzed by nearest neighbor analysis ArcGIS 10.2. Industrial Point sources emission data came from fifteen thousands sewage draining outlets of Guizhou Provincial Environmental Protection Bureau. Nighttime light data was provided by (Li et al., 2020b). A detailed description of the data sources and processing steps are provided in supplementary materials, see Fig.S1-27 and Table.ST2.

2.3. Modeling and database processing

First, we assessed water quality conditions and identified key variables deteriorating water quality calculating by the Water Quality Index (WQI-DET) proposed by (Huang et al., 2019). Second, we used Zonal statistics in ArcGIS 10.2 to extract the watershed characteristic data to pair with geographical location of the water quality sampled sites. The integrated subbasin units (SUs) database matrix containing water quality data (as model inputs) and corresponding watershed characteristic data (as model output) prior to developing the prediction model are shown in Fig S1, Table S1 and S2. We first detrended the water quality for use in modeling (Schwarz et al., 2006). A large uncertainty would exist in time-averaged water quality on account of the temporal variability within water quality datasets. Prior to performing ML models, we used the car package in R 4.03 to perform the Box-Cox transformation for the site-level average mean concentrations of each water quality index (Fox et al., 2012; Guo et al., 2019). The Box-Cox parameter λ was estimated individually and presented in supplementary material Table S3 and all the transformed water quality variables were normally distributed based on the Shapiro-Wilk's test (Box and Cox, 1964; Liu et al., 2021; Wang et al., 2021a). We only selected data from 151 water quality sites for ML modeling on account of the completeness of the dataset for total nitrogen. In order to improve the precision and computational efficiency of the ML model, through feature selection, we first removed redundant and irrelevant features according to Spearman correlation analysis, see section 3.2. Spearman correlation coefficient between each pair of features were performed and Mantel text was used to test the relationship between environmental factors and water quality variables(Legendre et al., 2015), see Fig.5. Due to the special geological conditions, debris flow, landslide and other geological disasters often occur in some parts of the WRB. We added the distance between the landslide damage points to the center of water body in the ML model. Although this metric is not filtered into the four models, we still included this index into the inputs of the ML model to verify if it has an impact to the water quality impairment. All the prediction models in this study were performed on the Jupiter notebook platform using the open sources libraries in Python3.7 (Scikit-learn, Hyperopt, XGBoost). The visualization and calculation of SHAP value applied the SHAPforxgboost by Liu (2019) in R 4.03 https://liuyanguu.github.io/post/2019/07/18/Visualization-of-shap-for-xgboost/. ArcGIS®10.2 was used for process and analysis of all watershed attributes and visualization of the results.

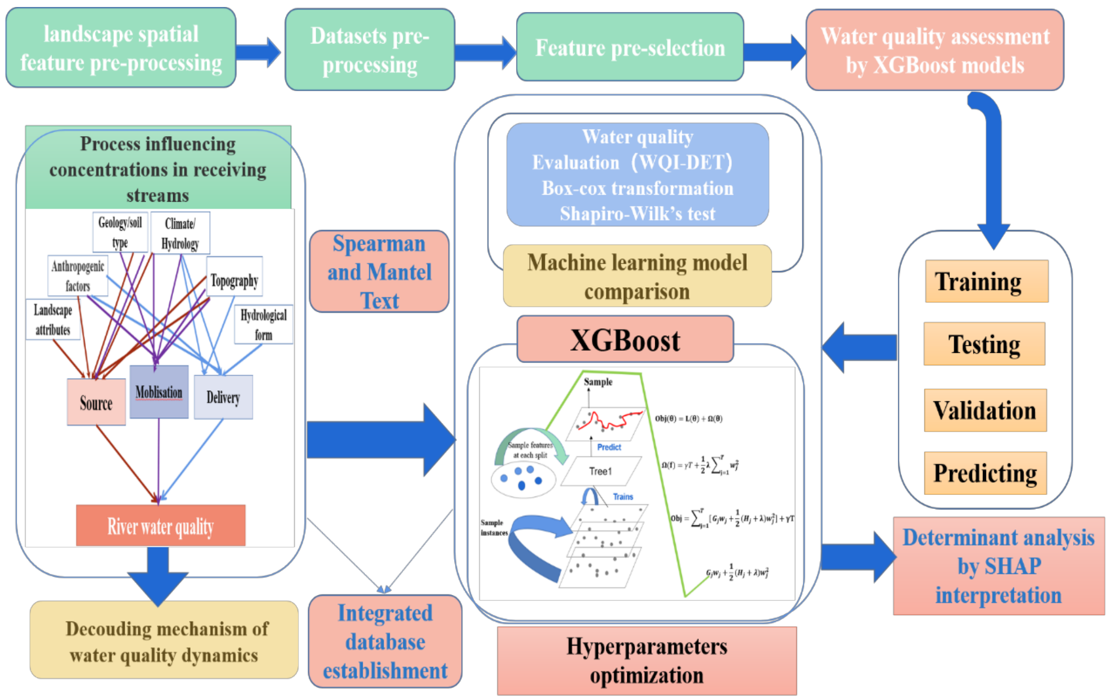


Fig.2. The schematic framework of overall methods used in this study

2.3.1 Water quality impairments evaluation

Water quality can be expressed in terms of scores calculated through integrating complex data into a mathematical expression (Nazeer et al., 2014). Water quality index (WQI) (dimensionless value) based on multiple water quality indicators has been widely used to characterize the degradation degree of surface and groundwater water quality (Lumb et al., 2011; Sutadian et al., 2016; Wu et al., 2018). We referred to the algorithm of modification water quality index (WQI-DET) to determine key variables leading to deterioration of water quality put forward by(Huang et al., 2019). The most sensitive indicators of river water quality impairments were evaluated according to the relative frequency of a variable leading to negative WQI-DET values in the WRB during 2014-2019. Indices of WQL-DET indicate extremely poor water quality (score of-oo) through to good water quality (score of 100). We can calculate the value of a WQI-DET of a single water sample by equation (1), from which monthly WQI-DET was calculated by averaging all the values within a given month. Eeleven (11) water quality variables were used to calculate WQI-DET, i.e., n = 11, and their concentrations were evaluated against the corresponding surface water quality classes.

$$WQI_{DET}^{j} = min(WQI_{DET_1}^{j},\ldots,WQI_{DET_i}^{j},\ldots,WQI_{DET_n}^{j}) \qquad (1)$$

$$WQI_{DET\_i}^{j} = 100 - max\left(0, \frac{C_{ij} - C_i^I}{C_i^V - C_i^I} \times 100\right) \quad (2)$$

(WQI_{DET}^{j}) is the WQI-DET value for the variable i of the water sample j; C_{ij} is the concentration of the environmental variable i of the water sample j; C_i^V and C_i^I are the concentration of the variable i at class I and V according to (GB3838-2002), respectively.

2.3.2 Machine learning prediction method

Boosting regression Tree (Boosting) is a machine learning technique commonly used for regression and classification problems. It generates prediction models in the form of collections of weak prediction models (usually decision trees) and modelling complex phenomena(Friedman, 2001; Strobl et al., 2009). The XGBoost package is an optimized distributed gradient enhancement library that reduces the gradient of the loss function (Chen et al., 2015). The component trees using recursive binary partitioning of predictive variables are chosen to minimize the variance of residuals and segment of all predictive variables, which is considered robust to outliers (Chen and Guestrin, 2016). To be self-contained, we just provide a brief description of the XGBoosting model here, and the detailed equation can be referred to the literature elsewhere(Chen and Guestrin, 2016). Eq (1) describes the training loss and regularization which consists of the two parts of XGBoost's objective function:

$$\mathrm{Obj}(\theta) = \mathrm{L}(\theta) + \Omega(\theta) \quad (3)$$

where $\mathrm{L}(\theta)$ is the training loss function employing to evaluate the model simulated performance for training data and $\Omega(\theta)$ is the regularization term aiming to control the overfitting of model (Gao et al., 2018). In addition, the complexity of each tree is often computed as the following Eq. (2):

$$\Omega(\mathrm{f}) = \gamma\mathrm{T} + \frac{1}{2}\lambda\sum_{j=1}^{T}\mathrm{w}_j^2 \quad (4)$$

In Eq (4) w_j is represented by the vector of scores on leaves while T represented by leaves respectively. The Eq.(3) is defined as the objective function of the structure score of XGBoost.

$$\mathrm{Obj} = \sum_{j=1}^{T}[\,\mathrm{G_j w_j} + \frac{1}{2}(\mathrm{H_j} + \lambda)\mathrm{w_j^2}] + \Gamma\mathrm{t} \quad (5)$$

The form $G_j w_j + \frac{1}{2}\left(H_j + \lambda\right)w_j^2$ (6) is quadratic and the best w_j to a given structure q(x). In each distinct round of cross-validation we tuned the hyperparameters of the XGBoost model. Grid Search CV was applied to automate the tuning of hyperparameters to determine the optimal value of the given model to satisfy the model generalizability (Moriasi et al., 2007). 70% of the randomly selected data sets was used as the training set and 30% as the test set. The prediction model is only established by using the data from the training sets, and the training sets are randomly divided into five parts by non-repeated sampling. Four of them were used to train the model each time, and the remaining ones was used to verify the accuracy of the four trained models. The step was repeated five times until each subset had a chance to be used as the validation set, and the remaining subset was used as the training set. The average of the five test results was calculated as an estimate of model accuracy and as a model performance indicator of the model under the implementation of the five-fold cross-validation. Finally, we validated with the remaining 30% of the test sets. We evaluated average model predictions performance based on coefficient of determination (R^2), root mean square error (RMSE), and Nash-Sutcliffe coefficient (NSE) (Nash and Sutcliffe, 1970).

2.3.3. SHAP analysis

Shapley Additive Explanations (SHAP) is a unified approach to create interpretable machine learning models. It helps to explain the output of any ML model and to visualize and describe the complex causal relationship between driving forces and the prediction target (Li et al., 2018). SHAP, an additive explanation model, is inspired by the theoretically optimal Shapley value of cooperative game theory, with all the characteristics treated as "contributors" (Lundberg and Lee, 2017; Strumbelj and Kononenko, 2014). Shapely values are determined according to several axioms to help allocate the contribution fairly for a group N (with *N* features) . (Lundberg et al., 2020; Lundberg et al., 2018). A linear function of binary features g is defined based on the following additive feature attribution method in equation (6):

$$\emptyset_i = \sum S \subseteq F\{i\} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{Su\{i\}m,}(x_{Su\{i\}}) - f_S(x_s)] \qquad (6)$$

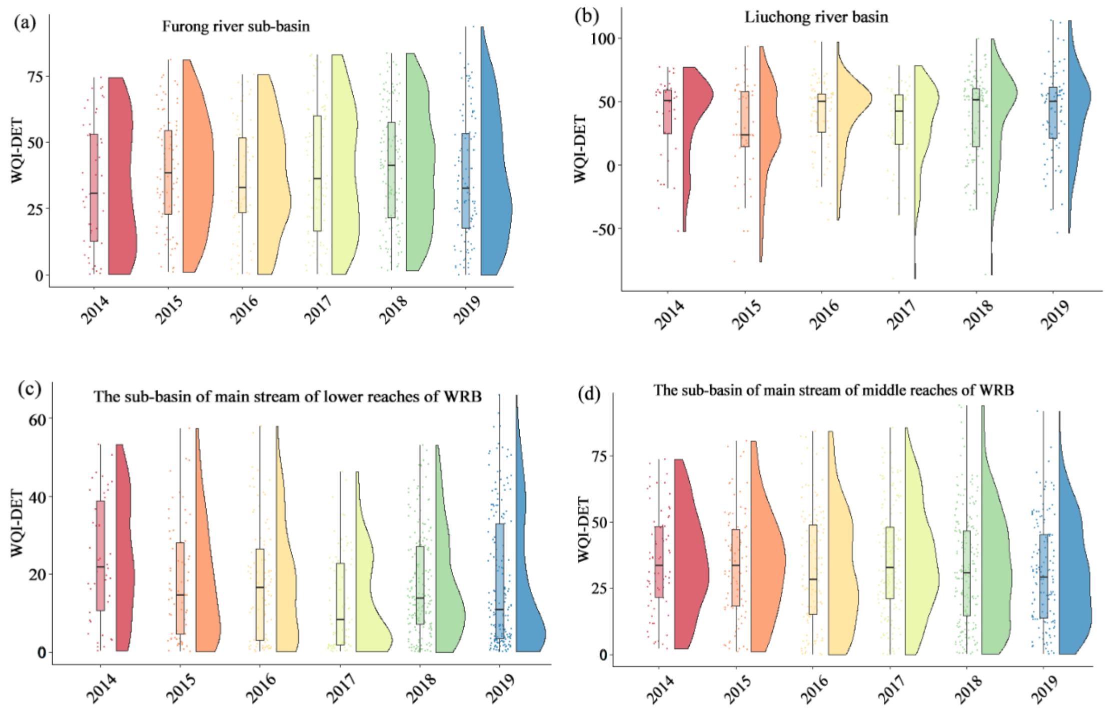
$$\mathrm{g}(\mathrm{z}') = \emptyset_0 + \sum_{\mathrm{i}=1}^{\mathrm{M}} \emptyset_\mathrm{i} \mathrm{Z}_\mathrm{i}^1 \qquad (7)$$

where z', equals to 1 when a feature is observed, otherwise it equals to 0, and M is the number of input features. In this study, we apply TreeExplainer proposed by Lundberg and Lee (2017) to accurately calculate TreeSHAP values of the tree integration models. Hyperparameters tuning of the XGBoosting model are performed separately in each round of cross-validation, and the overall RMSE was calculated based on the out-of-sample prediction after cross-validation. The SHAP values of a given prediction variable and observation value exist differences in the outputs, i.e. the predicted water quality indicator, up to whether the model is suitable for using or not using the prediction variable after performing each observation. The mean absolute SHAP values of all observed values summarize the importance of global features, and can be interpreted by more local models through scatter plots of individual predictive variables and their SHAP values (Just et al., 2020).

3. Results

3.1 Assessment of water quality impairments in WRB

According to the method of section 2.3.1, river water quality in the WRB shows temporal and spatial variation. Water quality is roughly consistent with the distribution of population density and the geographical line of terrain (decreasing from southwest to northeast along the elevation), presenting a complicated characteristic of fractal phenomenon (Xu et al., 2021). To be specific, from 2014 to 2019, 37.2% of monitoring sections in Furong River basin in the northeast of the WRB showed good water quality. While only 25.7 % and 17.1 % of monitoring sections in the middle and southwest part of the WRB showed good water quality. Furong River Basin has the best water quality and the mean value of median of WQI-DET is 62.15. While the Sancha River Basin (SCRB), Liuchong River Basin (LCRB) and Qingshui River Basin (QRB) had the worst water quality, with the mean values of median of WQI-DET from 2014 to 2019 being 42.75, 44.15 and 49.81, respectively. The mean value of the median of WQI-DET values of the Xiangjiang River basin (XJRB) and the middle reaches of main stream of the WRB (MRMS of WRB) and lower reaches of main stream of the WRB (LRMM of WRB) were 44.71, 51.65 and 46.65, respectively. The worst water quality monitored section in the WRB was found mainly in the SCRB and LCRB, among which 74.1% (23) of 31 sampled sites and 76.4.4% (26) of 34 sampled points are worse than Class V on the grounds of water quality standards (GB3838-2002). The water quality of 18 sampling sites (71.76%) in the LCRB was very poor, which was significantly higher than that in central regions (XJRB and LRWRB) and northeast regions (FRB). The sampled sites with extremely poor water quality accounted for 46.1%, 36.4% and 18.19%, respectively. During the year of 2014-2019, the median trend line of WQI-DET (Figure 4) of the FRB, middle reaches and lower reaches of the WRB illustrated a positive slope (k) ($P < 0.01$). It showed a small improvement of water quality in general, increasing 0.52, 0.13 and 0.34 of WQI-DET per year. However, the water quality of the SRB and XRB showed a negative slope, and the overall water quality decreased slightly. The K value of WQI-DET decreased by -1.71 and -1.14 per year respectively. It is worth noting that the river reaches in the rural areas around the cities, are being seriously polluted. The WQI-DET of water quality decreased from upstream to downstream reaches of the WRB. The water quality in the middle reaches of the WRB was slightly better than that in the lower reaches of WRB, which may be due to the large discharge and the cumulative effect of pollutants from upstream to downstream. The absolute number of WQI-DET showed a slight downward trend from 2014 to 2019, but it began to increase after 2017, mainly due to the increase of water quality sampled sites and samples. In all sampled sites of the whole WRB, 61.3% (127/207) of water quality conditions were seriously impaired. We used the relative frequency of negative WQI-DET value caused by each variable to determine the most sensitive indices of river water quality impairment in the WRB during 2014-2019, and these were COD_{Mn}, TN, and TP. In particular, the contribution of COD_{Mn} (reflecting organic pollutants) and TP to water quality impairment increased, especially in the river reaches around densely populated urban and rural residential areas.



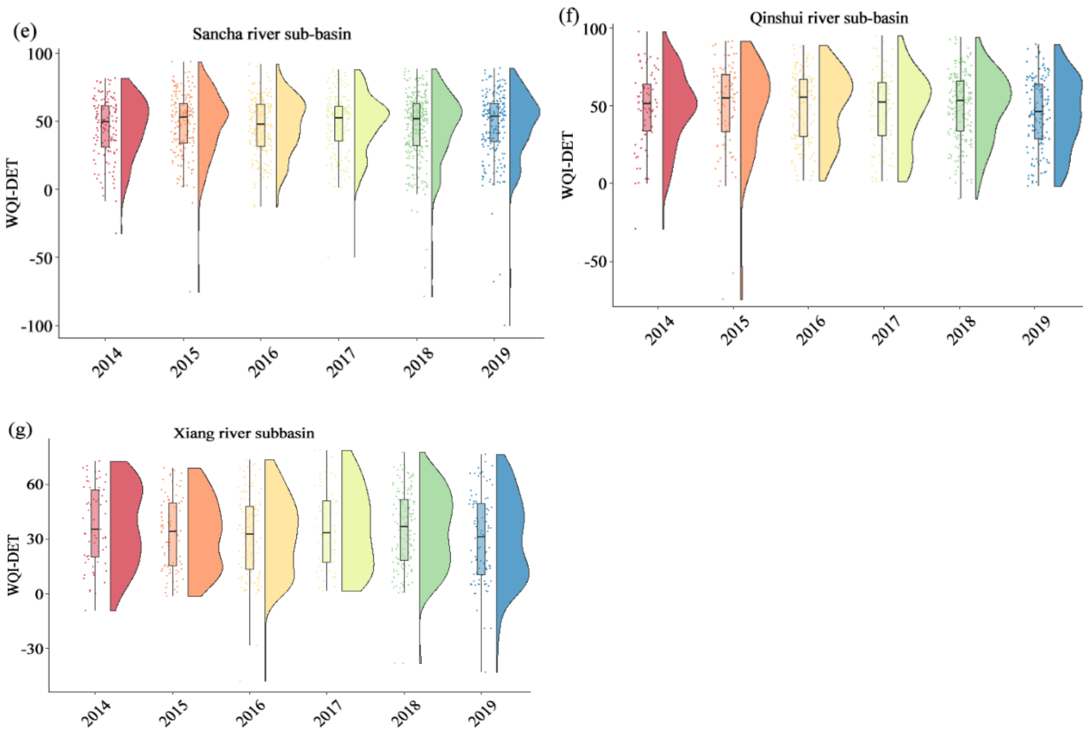


Fig.4. Inter-annual variability of the WQI-DET values for seven sub-basins of WRB during the 2014–2019 period. The cloud and rain map represents the WQI-DET distribution in the seven sub-basins of WRB. X-axis 1-7 represents Furong River Basin (coral pink), Liuchong River Basin (orange), Qingshui River Basin (straw orange) and the low reaches of main stream of WRB (apple green), and Xiang River Basin(blue) the middle reaches of main stream of the WRB (olivine). N is the number of sampling data. Note：To better visualize Fig.4, WQI-DET<-100 was omitted

3.2 The results of water quality prediction based on XGBoost

The Pearson correlation coefficients between each pair of watershed features were calculated initially, and only those with a Spearman correlation larger than 0.5 were kept (Fig.5). The Mantel test of mutual information (Mantel text) is a nonlinear correlation metric for pairs of geographic characteristics or environmental factors(Legendre et al., 2015). About 50% of features with low correlation (for COD_{Mn}, TN and TP and $p< 0.05$) were further discarded. Meanwhile, a Partial Mantel test can eliminate the interference of autocorrelation between environmental factors. The larger the correlation coefficient of Mantel test, the smaller the P value is. It indicates the greater the impact of geographical landscape factors on a water quality index(Zeller et al., 2016).

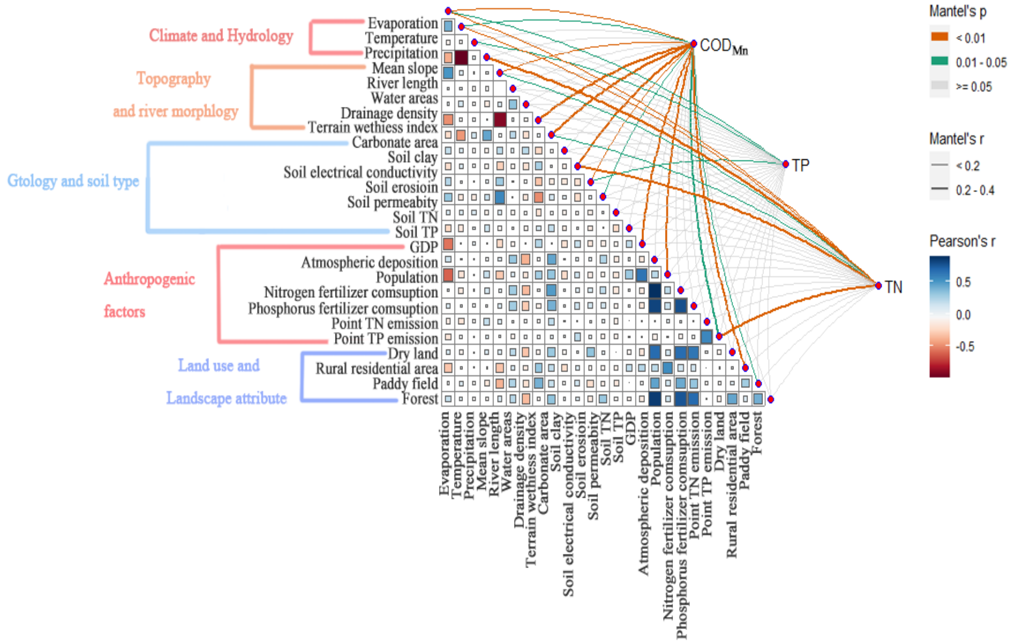


Fig.5. Pearson correlation between environmental factors is shown in the lower right corner, and Mantel correlation between watershed geographical factors and water environmental factors is shown in the upper right corner

We used XGBoost to model the relationship between 23 watershed landscape variables and three water quality indices (Fig.5). Four other popular machine learning techniques were implemented prior to this work, with adoption of XGBoost, the best predictor, to improve the performance of machine learning-based water quality predictions. Results are reported in supplementary material Table S4. According to the Nash-Sutcliffe efficiency coefficient (ranging from 0.15 to 0.77), the models for COD_{Mn}, and TN were significantly improved after feature selection, whereas the performance of the TP model was slightly better before feature selection (see supplementary material Table.S5 for more details). Among the three models, the training datasets of TP and COD_{Mn} were well fitted in the cross-validation, indicating that these two models have the highest accuracy. R^2 of CV were 0.68 and 0.57, RMSE were up to 0.79 and 0.94. These two models were applied to the TP and COD_{Mn} test datasets, with R^2 of CV of 0.71 and 0.65, and RMSE of 0.79 and 0.94, respectively.

However, the results of training the model for TN data fell short of expectations, the R^2 of CV training datasets of TN are 0.37, RMSE were 1.48. The R^2 of the test datasets of TN were 0.39, and RMSE was 1.37. Through hyperparameter optimization, the TP model was slightly improved while the COD_{Mn} and TN were greatly improved. The R^2 of COD_{Mn} training dataset and test dataset were increased to 0.67 and 0.73, RMSE was reduced to 0.79 mg/L and 0.65mg/L. R^2 of TP model training datasets and training datasets were 0.79 and 0.81, RMSE of which were 0.54 mg/L and 0.66 mg/L. Nash-Sutcliffe efficiency coefficients of those three models were improved after feature selection and parameter optimization ranging from 0.54 to 0.77, more details are provided in Table 1 and supplementary material (Table. S5 and Fig.S3).

Table 1. Performance of XGBoost models before and after hyperparameter optimization

	Water quality index	R^2		RMSE		NSE	
		Train	Test	Train	Test	Train	Test
Default Parameters	COD_{Mn}	0.57	0.65	1.96	0. 94	0.37	0.45
	TN	0.37	0.39	1.48	1.37	0.41	0.23
	TP	0.68	0.71	0.93	0.79	0.46	0.51
Optimized Parameters	COD_{Mn}	0.67	0.73	0.79	0.65	0.64	0.77
	TN	0.57	0.61	0.81	0.78	0.54	0.61
	TP	0.79	0.81	0.54	0.66	0.77	0.74

Note: WQI = water quality index, R^2 = coefficient of determination, NSE =Nash-Sutcliffe efficiency coefficient, RMSE= root mean square error

The concentrations of COD_{Mn}, TN and TP of 792 SUBs were predicted by the XGBoost model. The predictions show that COD_{Mn} concentration ranges from 0.2 to 17.31 mg/L, with an average concentration of 15.84 mg/L, See Fig. 6 (a) to (c). The reaches with higher COD_{Mn} concentration were distributed in densely populated urban reaches of QSRB, XJRB and the MRMS. The TN and TP concentrations in the WRB ranged from 0.25 to 5.72 mg/L and 0.02 to 1.31 mg/L, with a mean concentration of 3.83 mg/L and 0.56 mg/L respectively. The central and southeast portions of WRB are the most contaminated, with significant amounts of TN and TP, which is consistent with the spatial distribution of agricultural non-point source losses documented in this watershed (Dai et al., 2021; Xu et al., 2021a).

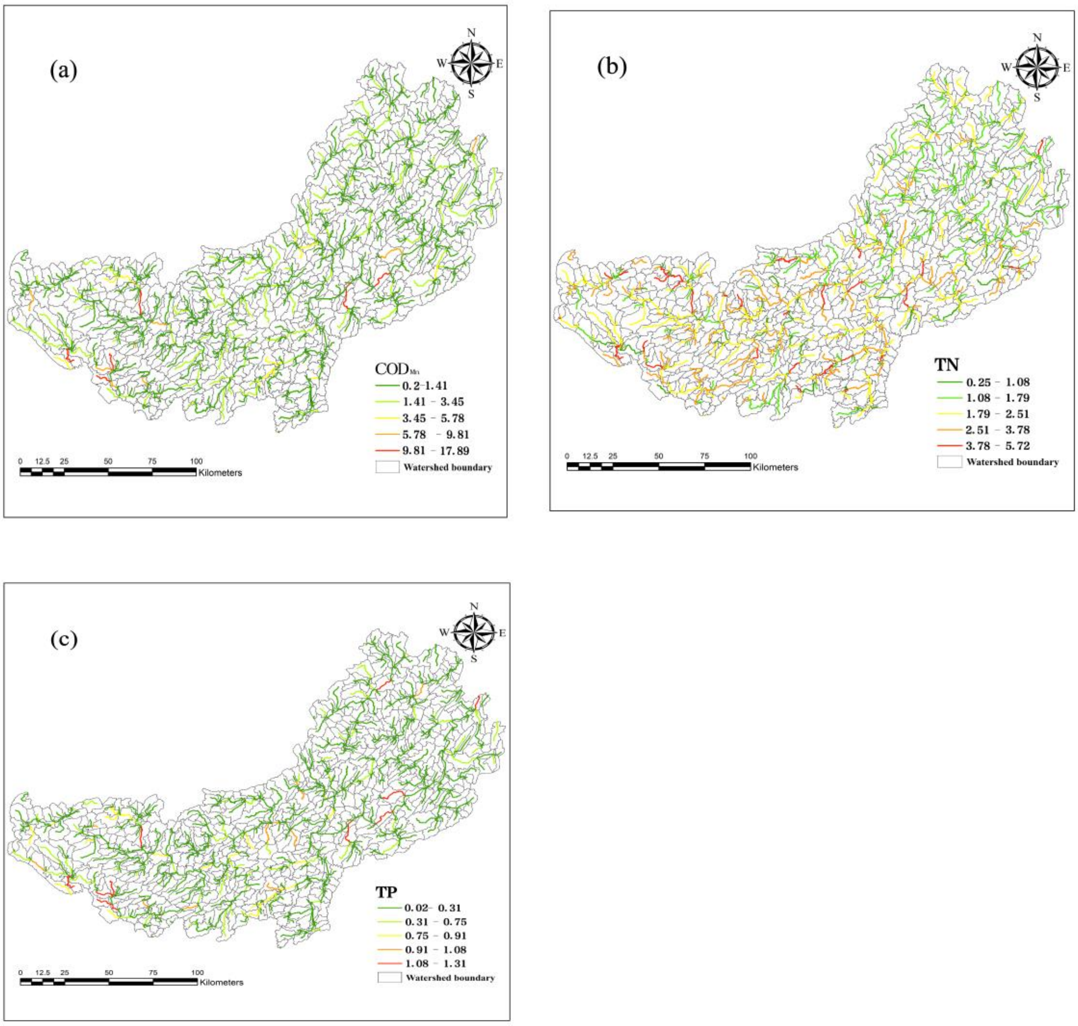


Fig.6. The XGBOOST models projected the following four water quality parameters: (a)COD$_{Mn}$, (b)TN, and (c)TP

3.3 Analysis of determinants of water quality

We used a 5-fold cross-validation split to evaluate the average absolute SHAP value as a measure of global feature importance. By performing each round of cross-validation, a recursive stepwise procedure was employed to order and remove features by increasing importance. Variable selection was run in three water quality databases using feature importance from SHAP values. Eight key features leading to water quality deterioration were selected to draw SHAP force plots according to six clusters of sub-groups for COD$_{Mn}$, TN and TP models, see our supplementary material (Fig.S4). The SHAP values of COD$_{Mn}$ ranged from 0.189 to 1.014, The SHAP values of TN and TP ranged from 0.006 to 0.054 and 0.011 to 0.041. We then pooled the features from the previous ranking and sorted the importance of the features from lowest to highest according to the average absolute SHAP value of all the features in the model. Repeating this step, least important features in each step were discarded. After plotting the overall RMSE predicted by cross-validation based on the chosen features, we finally selected the model with the lowest RMSE for each of the three water qualities. The SHAP Force plot (Fig.7) essentially superimposed these SHAP values for each observation and shows how the final output is obtained as a sum of the attributes of each predictive variable. The X-axis is set to -1 to 1 to facilitate the comparison of the three models. The Y-axis shows the order of the average absolute value of all observations (Fig.7). The eigenvalue is of absolute SHAP value is higher, the influence of the eigenvalue on the model output is greater. We used a bee swarm plot to illustrate and rank the watershed factors driving water quality in the average absolute SHAP value. COD_{Mn} was driven by anthropogenic factors, and the average absolute SHAP values were: (1.014) of nitrogen fertilizer consumption and (0.524) of night light intensity, and (0.395) of point source nitrogen emission. The land use types such as the paddy land, dry land, grass land and rural residential areas are sorted by CODMn, TP, and the mean absolute SHAP value ranged from 0.036 to 0.307. Descriptive scatter plots representing watershed characteristics and their SHAP scores are provided in supplementary material (Fig.S5), approximating their contribution (local feature importance) to the prediction of the Y-axis (three water quality characteristics). As important meteorological and hydrological factors, rainfall, evaporation and runoff drive the variation of TP, CODMn and TN, the average absolute of SHAP values were within the range of 0.028 to 0.427. River morphology factors such as river length, drainage density and water area play a pivotal role in influencing river water quality. Lithologic features (e.g., carbonate rocks) and soil property (soil erosion, Terrain wetness index (TWI), soil permeability and soil electrical conductivity, soil organic carbon) are also important determinants that can influence the deterioration of river water quality, the range of the mean absolute SHAP values were from 0.028 to 0.323 (Fig.8). The descriptive scatter plots representing watershed characteristics and their SHAP scores are provided in supplementary material (Fig.S4), approximating their contributions (local feature importance) to the prediction of the Y-axis (three water quality characteristics).

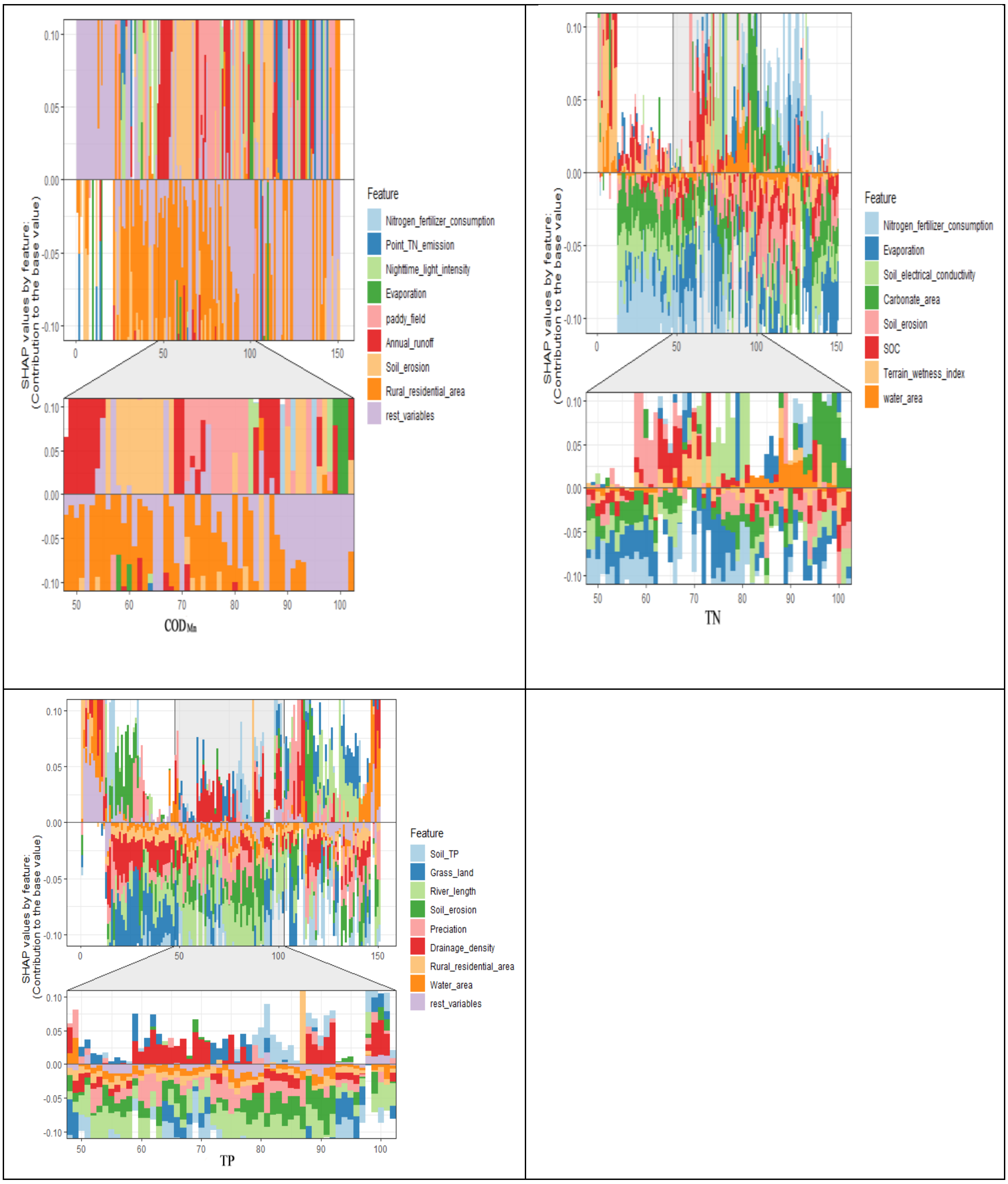
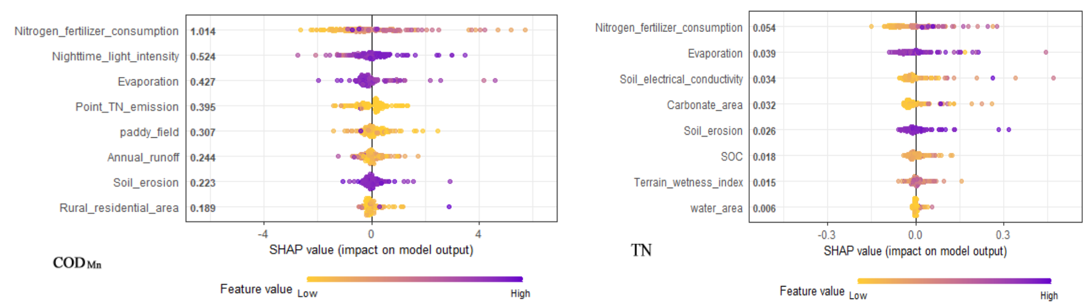


Fig.7. SHAP force plots show how the final output is the sum of the attributes of each predictive variable



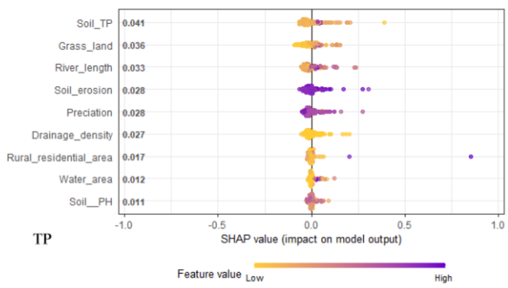


Fig.8. The Bees warm plots show SHAP values for watershed characteristics of observation using each water quality indicator. The Y-axis represents the rank of the average absolute SHAP values of the observed values (COD_{Mn}, TN, TP) of all watershed features

4 Discussion

4.1 The performance of the model to predict pollutant concentration in river water quality of WRB

The XGBOOST model developed in our study constructed a nonlinear mapping relationship between the multi-source data and the concentrations of COD_{Mn},TN and TP. This provided accurate prediction of the concentrations in unmonitored reaches of the river network. The predicted concentration of pollutants in the study area is consistent with the measured results, the average R^2 were higher than 0.78. It is worth noting that the input variables used to construct the model in this study were available to obtain. However, the model, constructed based on these input variables, can successfully predict river pollution. This provides a solution to time constraints imposed by sample collection, transportation and detection of traditional river water pollution concentrations via analysis methods, but also solves the problem that traditional monitoring methods cannot conduct rapid on-site analysis (Shuhong et al., 2019). More importantly, the predictions (Fig.S4-S7) show that the model constructed in this study can predict the concentration of pollutants in rivers better than the other 4 ML methods. Of course, the watershed characteristics used in the model will greatly affect the model performance of different water quality parameters, especially when the water quality parameters have different sources and migration/transformation processes (Alvarez-Cabria et al., 2016b). The watershed characteristics adopted by our model can explain the variability of TP and COD_{Mn} concentrations, but are inferior when predicting TN concentrations. Due the leakage of carbonate, groundwater systems can act as a net sink for dissolved Nitrogen by increasing the residence time and reducing the loads of TN through biochemical processes (Zhang et al., 2020B), thereby reducing the loads from surface water. Meanwhile, TN leaking from carbonate aquifers can be stored and converted to other nitrogen forms (e.g., by nitrification and/or N_2 to NO_3^-N by anaerobic ammonia treatment) (Dai et al., 2021; Zhang et al., 2020), However, weak statistical significance does not necessarily mean that those variables represented by watershed landscape characteristics are inherently unimportant in determining the sources, dynamics and transport of pollutants. Secondly, the accuracy of the model is not only affected by the predictors, but also impacted by the environmental behavior of the predicted targets. The determination of parameters is based not only on the statistical significance of the coefficients, but also on the overall model fitting and the physical importance of the parameters (Wang et al., 2021b). At the same time, the ML model will have better simulation performance when the sample size increases. Or it might be possible that as we do not consider the effects of seasonal variations on water quality resulting in the characteristics considered, our study could not fully explain the water quality impairments. Moreover, the diversification and spatial differences of various water quality parameters are determinants of different ecological, socio-economic and policy influences in the basin that can contribute to uncertainties in the model accuracy.

4.2 How does natural characteristics and anthropologic factors generate covariances on water quality in the WRB?

In our study, to obtain a more unbiased model, influential watershed characteristics and variables are considered as much as possible which include variables that are not readily available e.g. industrial point source pollution, fertilizer application data, sewage treatment plants, etc. River water quality is covaried with many geographical factors such as topography, land cover, biogeochemical reactivity, climate etc. In section 3.3, we apply SHAP values to decipher how these factors shape the water quality of rivers. Temperature has been identified as a key factor that directly influences riverine thermal regimes and biogeochemical processes, such as nitrification, denitrification, ammonification, and sediment diagenesis rates (Lintern et al., 2018; Sardans et al., 2008; van Vliet et al., 2013). High temperature conditions will increase the growth and degradation of algae and the capacity of sediments to adsorb phosphorus, which leads to higher COD_{Mn} concentration and higher concentrations of total phosphorus in water (Xia et al., 2015). Rainfall can profoundly modulate the flow-concentration relationship (Green et al., 2007), especially during a few short but intense precipitation events, where particulate matter and bioavailable phosphorus loads may differ by an order of magnitude between wet and dry conditions (Long et al., 2014). Environmental land use conflict caused by land use deviating from land capacity (natural use) is the root cause of accelerated deterioration of water quality, and it may lead to continuous changes in precipitation-runoff-infiltration processes, which in turn lead to extensive soil erosion and nutrient loss (Blevins et al., 1998; Pacheco et al., 2018; Suescun et al., 2017; Thomas et al., 2016; Valle Junior et al., 2014). In karst areas, as agriculture encroaches on natural lands, rapid land cover change often leads to long-term damage to soil and water conservation and other important ecosystem services (Li et al., 2021) Low natural vegetation covers owing to improper land use practices (cropping on sloping land), livestock grazing, and environmental hazards (rocky desertification) may reduce contaminant attenuation in karst area (Jiang et al., 2014).

In highly permeable carbonate karst aquifers, where there are widespread formations of fissures, fractures, and conduits, fast (e.g. conduit) and slow (e.g. fracture and matrix) flow transfer pathways will operate (Clifford and Williams, 2007). This leads to the rapid infiltration of rainwater that carries pollutants (e.g., from livestock, domestic and industrial discharge effluents) and contaminates groundwater (Wang et al., 2020; Yue et al., 2019). Geology and soil type determine the sources of sediment and natural nutrients in the catchment (Bostanmaneshrad et al., 2018; Grayson et al., 1997; Juracek and Ziegler, 2009). The erodixbility of soil and rock and the adsorption capacity of soil affect the flow of water components in a watershed. The mobilization of sediments is closely correlated to the susceptibility of the geological deposit and the soil within the catchment to erosion and weathering (Meybeck et al., 1990; Perry and Vanderklein, 2009). Lithology determines the alkalinity (pH) and conductivity of water and the concentration of different ions associated with many biogeochemical processes (Doherty et al., 2014). Soil adsorption capacity also affects nutrient mobilization in catchments. The transport of dissolved phosphorus, nitrogen and salts from catchments to recipient waters via underground flow pathways is strongly influenced by the hydrological characteristics of the soil. When the soil saturated conductivity of aquifers in watershed areas is low, the residence time of dissolved components in groundwater flow in the catchment area is increased (Lintern et al., 2018). This provides more opportunity for components to be lost from flow paths through nutrient absorption or biogeochemical processes such as denitrification (Hasani Sangani et al., 2015). The soil permeability (soil hydraulic properties) can affect soil quality and moisture, thereby altering the input of nutrients or organic matter to groundwater and river systems (Rodriguez-Blanco et al., 2015). TWI reflects topographic control of groundwater surface and soil moisture, while high TWI values indicate shallow groundwater table and high soil moisture (Rodhe and Seibert, 1999). Soil organic nitrogen is the main source of nitrates in rivers, soil moisture can promote the production of NO_3-N from the nitrification of soil organic nitrogen. This can explain that TWI entered TN model (Li et al., 2019). Soil pH plays an important role in determining the morphology of phosphate in soils because phosphate can bind to different iron when pH changes. Phosphates tend to form insoluble compounds in the presence of high concentrations of exchanged calcium. A reduction in soluble phosphorus usually occurs at higher pH (Sierra et al., 2017). As a result, high soil pH reduces water transport of phosphorus from land to rivers.

Anthropogenic activities has altered the river morphological conditions (e.g., changes in hydrological connectivity due to dam construction) and can severely impair river water quality (Maavara et al., 2020; Rodriguez-Blanco et al., 2015). In our study, (river length, drainage density, water areas) are all important covariables affecting water quality in TN, TP models. In general, higher drainage density may increase the likelihood that terrestrial pollutants carried by surface runoff will enter the water body (Alexander et al., 2002; Prasad et al., 2005). The river length determines that the transportation time of pollutants in the stream follows first-order reaction kinetics related to hydraulic residence time (Smith et al., 1997). Reservoirs play an important ecological function by hydrologically connecting upland and downstream river networks and influencing the biological cycle of nutrients. They have strong nutrient removal/interception capabilities which can be sinks of incoming nutrients or, if water quality is poor, they become sources of pollutants in downstream river reaches. Due to the need to improve engineering water shortage and flood control, the local government, most rivers in WRB are impounded (Dai et al.,2020). According to the Bureau of Hydrology and Resources of Guizhou Province, besides a dam cascade, there are 19,652 small reservoirs in the entire WRB (Dai, 2019). Dam cascade and small reservoirs have altered the hydrological regime, river morphology and lateral connectivity and increased longitudinal fragmentation of the basin (Viaroli et al., 2018), which has further amplified the instability of the biogeochemical processes and extended the range of resulting environment damage. These structural modifications have also increased regional hydraulic retention times and slowed the flow rate of rivers, in turn hindering river metabolism, amplifying nutrient transport and delivery, but also triggering eutrophication in rivers themselves (Dodds, 2006; Nizzoli et al., 2018).

GDP, industrial point emission, rural residential area and night light index were considered as key factors result in COD_{Mn} deterioration of water quality, which may also compensate for the index of population and urban development were filtered by XGBOOST. In the past decade, the WRB has made great efforts to promote the construction of municipal sewage treatment and sewage discharge standards has been strictly enforced under the background of China promoting huge investments to total environmental restoration (Xu et al.,2021). However, there is still a considerable gap in the design principle and operation performance due to treatment facilities and the sewer system lagging behind. Effluent discharge standards and sludge disposal are severely inconsistent with local conditions and environmental requirements (Lu et al., 2019; Qu et al., 2019). It might be the reason that COD_{Mn} concentrations in many urban reaches of WRB were higher than the acceptable limits. Moreover, rural residential area was an important determinant of water quality (COD_{Mn} and TP). Due to pursuing economic development of rural areas and agricultural intensification, the demand and consumption of water has been increasing and in turn runoff from fields and farms has increased in accord with the increases of discharge of domestic sewage, animal waste, leachate from manure storage facilities or green feed (Skinner et al., 1997). Moreover, the buildings in rural areas are spatially scattered, and the high construction costs are very unfavorable for the construction of public water supply and sewage treatment systems (Kupiec et al., 2021). In addition, karst rural areas not only lack the knowledge of proper manure management, but also lack proper manure storage facilities or poor technical standards (Gao et al., 2014; Norse and Ju, 2015; Oliver et al., 2020).

4.3 Management implications and future challenges

However, deterioration of water quality can be caused by many factors, such as complex geographical environment and intensive human intervention (including mining, intensive agriculture activity), inadequate sewage treatment measures (Xu et al.,2021) and poor groundwater environment (Li et al., 2020a; Zeng et al., 2020). In addition, damming and the construction of multiple small reservoirs have drastically reduced surface runoff, limiting the river's ability to dilute effluent from sewage treatment plants, see section 4.2. As mentioned above, our findings support the development of strategies by identifying key characteristics of pollutant sources and incorporating them into regional planning (e.g. changing land use, improving industrial structure and distribution). In addition, since the Chinese government has been promoting ecological rehabilitation projects to restore rocky desertification and improve local poverty, the river water quality has been neglected in the WRB. We also suggest that soil, water processes and environmental effects should be incorporated into a unified scientific management framework to best communicate the trade-offs between policy options and promotion of pollution control and ecological restoration. This will help realize ecological value and promote green development management of the WRB (Xu et al., 2021a). Our approach can not only effectively promote pollutant sources control, but also decelerate the pollutant migration and transformation process. It is imperative to adjust local economic structure and develop low-pollution water-saving industry. Water-saving irrigation schemes also appear to be a necessary measure to reduce pollutant infiltration into the soil.

And some micro-policy proposals were advocated; promoting BMPs is a good choice in this case, it will allow policy makers to mitigate non-point source pollutants and further restore river ecosystems in the agricultural areas of WRB. BMPs include improving the efficiency of fertilization, improving manure management and buffering the pollutant delivery processes between land and water (e.g. restrict livestock farming near rivers, plant more vegetation near river banks). Promoting soil remediation is also important for restoration of the water environment of WRB and requires management of vulnerable geological areas with well-drained soils, high recharge and low soil organic carbon characteristics. It is necessary to implement integrated management of surface water and groundwater to alleviate the contradiction between intensive water use and geographical environment constraints.

Although results have been achieved using ML methods to detect and evaluate water quality, we still need to consider the potential disconnects between macro-scale simulations and local social, economic, and environmental realities, as well as catchment-scale constraints for on-site water quality management. But we also need to conduct field assessments to assess the extent to which reductions in pollutants concentrations are actually achieved, based on best management practices for site-specific nutrient sources combining the landscape characteristics (Jarvie et al., 2018; Sharpley et al., 2016). In the short term, it may be unrealistic to expect pollutant concentrations to be reduced to the complianced and restricted target concentrations, especially in highly impaired karst basins with multiple complex pollutant sources and long-term legacy nutrient contributions (Jarvie et al., 2018; Sharpley et al., 2013; Xu et al., 2021a). However, our simulation to assess the nutrient limitations combined with assessment of compliance and limitation gaps, provides a basis for developing targeted approaches to nutrient water quality compliance in future work.

5 Conclusion

Understanding of the multiple forces determining river water quality and the complexity and interaction of these forces is necessary to develop successful water quality management strategies. Those knowledges can be used to develop predictive models that will help to predict river water quality. In this study, we evaluated those important factors affecting the spatio-temporal variation of water quality (COD_{Mn},TN,TP) in an ecologically fragile watershed with high landscape heterogeneity by adopting a data-driven machine learning approach. Machine learning can take advantage of all the crossover effects between variables to improve the accuracy of model predictions, which is an advantage over traditional statistical models. the Nash efficiency coefficient are ranging from 0.54 to 0.8, which indicates that our prediction is reliable and robust. Through the analysis of powerful model interpreter (SHAP), though anthropogenic factors such as land use are closely related to river pollutant concentrations, the effects of key hydroclimatic, soil types and vegetation conditions vary across different components and regions. XGBOOT can be used to identify potential water quality hot spots in unmonitored locations; this suggests that catchments with steep gradients, fragile soils or areas with widespread carbonate rocks should be sampled more frequently. Our study underlines the needs to highlight soil and water processes and integrate environmental effects into a unified scientific management framework when implementing ecological engineering restoration in karst areas. Therefore, as more land management surveys are been promoting and ongoing water quality monitoring data are available, an extended temporal or spatio-temporal modeling framework may be used to assess the success of recovery measures in the future. In the meanwhile, we should consider combining the assessment of simulated nutrient limits with the assessment of compliance and limitation gaps to provide a basis for developing a targeted approach to river water quality compliance that focuses on closing the gap between current and target concentrations.

Reference

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial

Intelligence(XAI).IeeeAccess.652138-52160.https://doi.org/10.1109/Access.2018. 2870052

Ai, L., et al., 2015. Spatial and seasonal patterns in stream water contamination across mountainous watersheds: Linkage with landscape characteristics. Journal of Hydrology. 523 398-408.https://doi.org/10.1016/j.jhydrol.2015.01.082

Alexander, R. B., et al., 2002. Estimating the sources and transport of nutrients in the Waikato River Basin, New Zealand. Water Resources Research. 38 4-1-4-23.https://doi.org/Artn1268
10.1029/2001wr000878

Altenburger, R., et al., 2015. Future water quality monitoring—adapting tools to deal with mixtures of pollutants in water resource management. Science of the total environment. 512 540-551.https://doi.org/10.1016/j.scitotenv.2014.12.057

Alvarez-Cabria, M., et al., 2016a. Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors. Science of the Total Environment. 545 152-162.https://doi.org/10.1016/j.scitotenv.2015.12.109

Alvarez-Cabria, M., et al., 2016b. Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors. Sci Total Environ. 545-546 152-62.https://doi.org/10.1016/j.scitotenv.2015.12.109

Andry, H., et al., 2009. Water retention, hydraulic conductivity of hydrophilic polymers in sandy soil as affected by temperature and water quality. Journal of Hydrology. 373 177-183.https://doi.org/10.1016/j.jhydrol.2009.04.020

Arhonditsis, G. B., et al., 2007. Eutrophication risk assessment using Bayesian calibration of process-based models: Application to a mesotrophic lake. Ecological Modelling. 208 215-229.https://doi.org/10.1016/j.ecolmodel.2007.05.020

Badham, J., et al., 2019. Effective modeling for Integrated Water Resource Management: A guide to contextual practices by phases and steps and future opportunities. Environmental Modelling & Software. 116 40-56.https://doi.org/10.1016/j.envsoft.2019.02.013

Baker, A., 2003. Land use and water quality. Hydrological Processes. 17 2499-2501.https://doi.org/10.1002/hyp.5140

Blevins, R., et al., 1998. Conservation tillage for erosion control and soil quality. Chapter in" Advances in Soil and Water Conservation.". Ann Arbor Press, Chelsea, MI. https://doi.org/10.1201/9781315136912-4

Bostanmaneshrad, F., et al., 2018. Relationship between water quality and macro-scale parameters (land use, erosion, geology, and population density) in the Siminehrood River Basin. Sci Total Environ. 639 1588-1600.https://doi.org/10.1016/j.scitotenv.2018.05.244

Box, G. E. P., Cox, D. R., 1964. An Analysis of Transformations. Journal of the Royal Statistical Society Series B-Statistical Methodology. 26 211-252.https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Cardinale, B. J., 2011. Biodiversity improves water quality through niche partitioning. Nature. 472 86-9.https://doi.org/10.1038/nature09904

Chen, T., Guestrin, C., Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp.

785-794. https://doi.org/10.1145/2939672.2939785

Chen, T., et al., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2. 1 1-4

Clifford, F., Williams, P., 2007. Karst Hydrogeology & Geomorphology.https://doi.org/ 10.1002/ 9781118684986

Dai, Y. B., et al., 2021. Modelling the sources and transport of ammonium nitrogen with the SPARROW model: A case study in a karst basin. Journal of Hydrology. 592 125763.https://doi.org/ARTN12576310.1016/j.jhydrol.2020.125763

Deng, X. J., 2020. Influence of water body area on water quality in the southern Jiangsu Plain, eastern China. Journal of Cleaner Production. 254 120136.https://doi.org/ARTN120136 10.1016/j.jclepro.2020.120136

Ding, J., et al., 2016. Influences of the land use pattern on water quality in low-order streams of the Dongjiang River basin, China: A multi-scale analysis. Science of the Total Environment. 551 205-216.https://doi.org/10.1016/j.scitotenv.2016.01.162

Dodds, W. K., 2006. Eutrophication and trophic state in rivers and streams. Limnology and Oceanography. 51 671-680.https://doi.org/10.4319/lo.2006.51.1_part_2.0671

Doherty, J. M., et al., 2014. Hydrologic Regimes Revealed Bundles and Tradeoffs Among Six Wetland Services. Ecosystems. 17 1026-1039.https://doi.org/10.1007/s10021-014-9775-3

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. Ecological Indicators. 50 79-89.https://doi.org/10.1016/j.ecolind.2014.11.003

Farnham, I. M., et al., 2002. Treatment of nondetects in multivariate analysis of groundwater geochemistry data. Chemometrics and Intelligent Laboratory Systems. 60 265-281.https://doi.org/10.1016/S0169-7439(01)00201-5

Fiorillo, F., et al., 2015. A model to simulate recharge processes of karst massifs. Hydrological Processes. 29 2301-2314.https://doi.org/10.1002/hyp.10353

Ford, D., Williams, P., Karst Hydrogeology and Geomorphology. 2007. https://doi.org/10.1002/9781118684986

Fox, J., et al., 2012. Package 'car'. Vienna: R Foundation for Statistical Computing.

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 29 1189-1232.https://doi.org/10.1214/aos/1013203451

Gao, M. F., et al., 2014. Modeling nitrogen loading from a watershed consisting of cropland and livestock farms in China using Manure-DNDC. Agriculture Ecosystems & Environment. 185 88-98.https://doi.org/10.1016/j.agee.2013.10.023

Granger, S. J., et al., Chapter 3 - Towards a Holistic Classification of Diffuse Agricultural Water Pollution from Intensively Managed Grasslands on Heavy Soils. Advances in Agronomy. Academic Press, 2010, pp. 83-115. https://doi.org/10.1016/S0065-2113(10)05003-0

Grayson, R. B., et al., 1997. Preferred states in spatial soil moisture patterns: Local and nonlocal controls. Water Resources Research. 33 2897-2908.https://doi.org/10.1029/97wr02174

Gruber, S., Peckham, S., 2009. Land-surface parameters and objects in hydrology. Developments in Soil Science. 33 171-194 https://doi.org/10.1016/S0166-

2481(08)00007-X

Guo, D., et al., 2019. Key Factors Affecting Temporal Variability in Stream Water Quality. Water Resources Research. 55 112-129.https://doi.org/10.1029/2018wr023370

Han, G. L., Liu, C. Q., 2004. Water geochemistry controlled by carbonate dissolution: a study of the river waters draining karst-dominated terrain, Guizhou Province, China. Chemical Geology. 204 1-21.https://doi.org/10.1016/j.chemgeo.2003.09.009

Hartmann, A., et al., 2015. A large-scale simulation model to assess karstic groundwater recharge over Europe and the Mediterranean. Geoscientific Model Development. 8 1729-1746.https://doi.org/10.5194/gmd-8-1729-2015

Hartmann, A., et al., 2014. Karst water resources in a changing world: Review of hydrological modeling approaches. Reviews of Geophysics. 52 218-242.https://doi.org/10.1002/2013rg000443

Hasani Sangani, M., et al., 2015. Modeling relationships between catchment attributes and river water quality in southern catchments of the Caspian Sea. Environ Sci Pollut Res Int. 22 4985-5002.https://doi.org/10.1007/s11356-014-3727-5

Hashemi, F., et al., 2016. Review of scenario analyses to reduce agricultural nitrogen and phosphorus loading to the aquatic environment. Sci Total Environ. 573 608-626.https://doi.org/10.1016/j.scitotenv.2016.08.141

Heckmann, T., Schwanghart, W., 2013. Geomorphic coupling and sediment connectivity in an alpine catchment — Exploring sediment cascades using graph theory. Geomorphology. 182 89-103.https://doi.org/10.1016/j.geomorph.2012.10.033

Höge, M., et al., 2022. Improving hydrologic models for predictions and process understanding using Neural ODEs. Hydrol. Earth Syst. Sci. Discuss. 2022 1-29.https://doi.org/10.5194/hess-2022-56

Hrachowitz, M., et al., 2016. Transit times-the link between hydrology and water quality at the catchment scale. Wiley Interdisciplinary Reviews: Water. 3 629-657.https://doi.org/10.1002/wat2.1155

Huang, J., et al., 2019. How successful are the restoration efforts of China's lakes and reservoirs? Environ Int. 123 96-103.https://doi.org/10.1016/j.envint.2018.11.048

Huang, J., et al., 2021. Characterizing the river water quality in China: Recent progress and on-going challenges. Water Res. 201 117309.https://doi.org/10.1016/j.watres.2021.117309

Jakeman, A. J., et al., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software. 21 602-614.https://doi.org/10.1016/j.envsoft.2006.01.004

Jarvie, H. P., et al., 2018. Phosphorus and nitrogen limitation and impairment of headwater streams relative to rivers in Great Britain: A national perspective on eutrophication. Sci Total Environ. 621 849-862.https://doi.org/10.1016/j.scitotenv.2017.11.128

Jiang, Z. C., et al., 2014. Rocky desertification in Southwest China: Impacts, causes, and restoration. Earth-Science Reviews. 132 1-12.https://doi.org/10.1016/j.earscirev.2014.01.005

Juracek, K. E., Ziegler, A. C., 2009. Estimation of sediment sources using selected chemical tracers in the Perry lake basin, Kansas, USA. International Journal of Sediment Research. 24 108-125.https://doi.org/10.1016/S1001-6279(09)60020-2

Just, A. C., et al., 2020. Gradient boosting machine learning to improve satellite-derived column water vapor measurement error. Atmos Meas Tech. 13 4669-4681.https://doi.org/10.5194/amt-13-4669-2020

Knoben, W. J. M., et al., 2020. A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. Water Resources Research. 56 e2019WR025975.https://doi.org/10.1029/2019wr025975

Kratzert, F., et al., 2019. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research. 55 11344-11354 https://doi.org/10.1029/2019WR026065

Kupiec, J. M., et al., 2021. Assessment of the impact of land use in an agricultural catchment area on water quality of lowland rivers. PeerJ. 9 e10564.https://doi.org/10.7717/peerj.10564

Legendre, P., et al., 2015. Should the Mantel test be used in spatial analysis? Methods in Ecology and Evolution. 6 1239-1247.https://doi.org/10.1111/2041-210x.12425

Li, C., Ji, H., 2016. Chemical weathering and the role of sulfuric and nitric acids in carbonate weathering: Isotopes (13C,15N,34S, and18O) and chemical constraints. Journal of Geophysical Research: Biogeosciences. 121 1288-1305.https://doi.org/10.1002/2015jg003121

Li, C., et al., 2019. Identification of sources and transformations of nitrate in the Xijiang River using nitrate isotopes and Bayesian model. Sci Total Environ. 646 801-810.https://doi.org/10.1016/j.scitotenv.2018.07.345

Li, S. L., et al., 2021. Karst ecosystem and environment: Characteristics, evolution processes, and sustainable development. Agriculture Ecosystems & Environment. 306 107173.https://doi.org/ARTN10717310.1016/j.agee.2020.107173

Li, S. L., et al., 2020a. Effects of agricultural activities coupled with karst structures on riverine biogeochemical cycles and environmental quality in the karst region. Agriculture Ecosystems & Environment. 303 107120.https://doi.org/ARTN10712010.1016/j.agee.2020.107120

Li, X., et al., 2018. Watershed System Model: The Essentials to Model Complex Human-Nature System at the River Basin Scale. Journal of Geophysical Research-Atmospheres. 123 3019-3034.https://doi.org/10.1002/2017jd028154

Li, X., et al., 2020b. A harmonized global nighttime light dataset 1992-2018. Sci Data. 7 168.https://doi.org/10.1038/s41597-020-0510-y

Lintern, A., Webb, J., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P. and Western, A.W. (2018), Key factors influencing differences in stream water quality across space. WIREs Water, 5: e1260. https://doi.org/10.1002/wat2.1260

Liu, J., et al., 2018. Assessing how spatial variations of land use pattern affect water quality across a typical urbanized watershed in Beijing, China. Landscape and Urban Planning. 176 51-63.https://doi.org/10.1016/j.landurbplan.2018.04.006

Liu, L., et al., 2020. Insights into the long-term pollution trends and sources contributions in Lake Taihu, China using multi-statistic analyses models. Chemosphere. 242 125272.https://doi.org/10.1016/j.chemosphere.2019.125272

Liu, S., et al., 2021. A multi-model approach to assessing the impacts of catchment characteristics on spatial water quality in the Great Barrier Reef catchments. Environ Pollut. 288 117337.https://doi.org/10.1016/j.envpol.2021.117337

Long, T. Y., et al., 2014. Evaluation of stormwater and snowmelt inputs, land use and seasonality on nutrient dynamics in the watersheds of Hamilton Harbour, Ontario, Canada. Journal of Great Lakes Research. 40 964-979.https://doi.org/10.1016/j.jglr.2014.09.017

Lu, J. Y., et al., 2019. Optimizing operation of municipal wastewater treatment plants in China: The remaining barriers and future implications. Environ Int. 129 273-278. https://doi.org/10.1016/j.envint.2019.05.057

Lumb, A., et al., 2011. A Review of Genesis and Evolution of Water Quality Index (WQI) and Some Future Directions. Water Quality Exposure and Health. 3 11-24.https://doi.org/10.1007/s12403-011-0040-0

Lundberg, S. M., et al., 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2 56-67.https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., et al., 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888. https://doi.org/10.48550/arXiv.1705.07874

Lundberg, S. M, Lee, S.-I., A unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4768-4777. https://doi.org/10.48550/arXiv.1705.07874

Maavara, T., Chen, Q., Van Meter, K. et al. River dam impacts on biogeochemical cycling. Nat Rev Earth Environ 1, 103–116 (2020). https://doi.org/10.1038/s43017-019-0019-0

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

Malago, A., et al., 2016. Regional scale hydrologic modeling of a karst-dominant geomorphology: The case study of the Island of Crete. Journal of Hydrology. 540 64-81.https://doi.org/10.1016/j.jhydrol.2016.05.061

Mandaric, L., et al., 2018. Impact of urban chemical pollution on water quality in small, rural and effluent-dominated Mediterranean streams and rivers. Sci Total Environ. 613-614 763-772.https://doi.org/10.1016/j.scitotenv.2017.09.128

Mayorga, E., et al., 2010. Global Nutrient Export from WaterSheds 2 (NEWS 2): Model development and implementation. Environmental Modelling & Software. 25 837-853.https://doi.org/10.1016/j.envsoft.2010.01.007

Mello, K. d., et al., 2018. Effects of land use and land cover on water quality of low-order streams in Southeastern Brazil: Watershed versus riparian zone. Catena. 167 130-138.https://doi.org/10.1016/j.catena.2018.04.027

Meybeck, M., et al., Global freshwater quality: a first assessment. Global freshwater quality: A first assessment, 1990, pp. 306-306. https://doi.org/10.1577/1548-8659-121.1.141

Mokoatle, M., et al., Predicting road traffic accident severity using accident report data in South Africa. Proceedings of the 20th Annual International Conference on Digital Government Research, 2019, pp. 11-17. https://doi.org/10.1145/3325112.3325211

Molnar, C., 2020. 5.10 SHAP (SHapley Additive exPlanations) Interpretable machine learning. Lulu. com, 2020.

Moreira, C., et al., 2020. An Interpretable Probabilistic Approach for Demystifying Black-box Predictive Models.https://doi.org/10.48550/arXiv.2007.10668

Moriasi, D. N., et al., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the Asabe. 50 885-900.https://doi.org/Doi 10.13031/2013.23153

Najah Ahmed, A., et al., 2019. Machine learning methods for better water quality prediction. Journal of Hydrology. 578 124084.https://doi.org/10.1016/j.jhydrol.2019.124084

Nash, J. E., Sutcliffe, J. V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology. 10 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Nazeer, S., et al., 2014. Heavy metals distribution, risk assessment and water quality characterization by water quality index of the River Soan, Pakistan. Ecological Indicators. 43 262-270.https://doi.org/10.1016/j.ecolind.2014.03.010

Nie, Y., et al., 2017. Comparison of Rooting Strategies to Explore Rock Fractures for Shallow Soil-Adapted Tree Species with Contrasting Aboveground Growth Rates: A Greenhouse Microcosm Experiment. Front Plant Sci. 8 1651.https://doi.org/10.3389/fpls.2017.01651

Nizzoli, D., et al., 2018. Denitrification in a meromictic lake and its relevance to nitrogen flows within a moderately impacted forested catchment. Biogeochemistry. 137 143-161.https://doi.org/10.1007/s10533-017-0407-9

Noori, R., et al., 2012. Chemometric Analysis of Surface Water Quality Data: Case Study of the Gorganrud River Basin, Iran. Environmental Modeling & Assessment. 17 411-

420.https://doi.org/10.1007/s10666-011-9302-2

Norse, D., Ju, X., 2015. Environmental costs of China's food security. Agriculture, Ecosystems & Environment. 209 5-14. https://doi.org/10.1016/j.agee.2015.02.014

Ockenden, M. C., et al., 2017. Major agricultural changes required to mitigate phosphorus losses under climate change. Nat Commun. 8 161. https://doi.org/10.1038/s41467-017-00232-0

Oliver, D. M., et al., 2020. How does smallholder farming practice and environmental awareness vary across village communities in the karst terrain of southwest China? Agriculture, Ecosystems & Environment. 288 106715. https://doi.org/10.1016/j.agee.2019.106715

Pacheco, F. A. L., et al., 2018. An approach to validate groundwater contamination risk in rural mountainous catchments: the role of lateral groundwater flows. MethodsX. 5 1447-1455.https://doi.org/10.1016/j.mex.2018.11.002

Parsa, A. B., et al., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid Anal Prev. 136 105405.https://doi.org/10.1016/j.aap.2019.105405

Perry, J., Vanderklein, E. L., 2009. Water quality: management of a natural resource. John Wiley & Sons. https://doi.org/10.1016/S0025-326X(97)00155-0

Powers, S. M., et al., 2016. Long-term accumulation and transport of anthropogenic phosphorus in three river basins. Nature Geoscience. 9353.https://doi.org/10.1038/Ngeo2693

Prasad, V. K., et al., 2005. Exploring the relationship between hydrologic parameters and nutrient loads using digital elevation model and GIS–a case study from Sugarcreek headwaters, Ohio, USA. Environmental monitoring and assessment. 110 141-169 https://doi.org/10.1007/s10661-005-6688-9

Qin, N., et al., 2015. Impacts of climate change on regional hydrological regimes of the Wujiang River watershed in the Karst area, Southwest China. Geoenvironmental Disasters. 2 10.https://doi.org/10.1186/s40677-015-0013-x

Qu, J. H., et al., 2019. Municipal wastewater treatment in China: Development history and future perspectives. Frontiers of Environmental Science & Engineering. 13 88.https://doi.org/ARTN 8810.1007/s11783-019-1172-x

Quinn, P., et al., 1991. The Prediction of Hillslope Flow Paths for Distributed Hydrological Modeling Using Digital Terrain Models. Hydrological Processes. 5 59-79. https://doi.org/DOI 10.1002/hyp.3360050106

Rodhe, A., Seibert, J., 1999. Wetland occurrence in relation to topography: a test of topographic indices as moisture indicators. Agricultural and Forest Meteorology. 98-9 325-340.https://doi.org/10.1016/S0168-1923(99)00104-5

Rodriguez-Blanco, M. L., et al., 2015. Relating nitrogen export patterns from a mixed land use catchment in NW Spain with rainfall and streamflow. Hydrological Processes. 29 2720-2730.https://doi.org/10.1002/hyp.10388

Rodriguez-Galiano, V., et al., 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). Sci Total Environ. 476-477 189-206.https://doi.org/10.1016/j.scitotenv.2014.01.001

Sardans, J., et al., 2008. Changes in soil enzymes related to C and N cycle and in soil C and N content under prolonged warming and drought in a Mediterranean shrubland. Applied Soil Ecology. 39 223-235.https://doi.org/10.1016/j.apsoil.2007.12.011

Schwarz, G., et al., Section 3. The SPARROW Surface Water-Quality Model—Theory, application and user documentation. Techniques and Methods, Reston, VA, 2006. e B. Hoos， RB Alexander 和 RA Smith https://doi.org/10.3133/tm6B3

Sharpley, A., et al., 2013. Phosphorus legacy: overcoming the effects of past management practices to mitigate future water quality impairment. J Environ Qual. 42 1308-26.https://doi.org/10.2134/jeq2013.03.0098

Sheng, M. Y., et al., 2018. Response of soil physical and chemical properties to Rocky desertification succession in South China Karst. Carbonates and Evaporites. 33 15-28.https://doi.org/10.1007/s13146-016-0295-4

Sierra, C. A., et al., 2017. Monitoring ecological change during rapid socio-economic and political transitions: Colombian ecosystems in the post-conflict era. Environmental Science & Policy. 76 40-49.https://doi.org/10.1016/j.envsci.2017.06.011

Singh, J., et al., 2005a. Hydrological modeling of the Iroquois river watershed using HSPF and SWAT 1. JAWRA Journal of the American Water Resources Association. 41 343-360 https://doi.org/10.1111/j.1752-1688.2005.tb03740.x

Singh, J., et al., 2005b. Hydrological Modeling of the Iroquois River Watershed Using Hspf and Swat. Journal of the American Water Resources Association. 41 343-360.https://doi.org/10.1111/j.1752-1688.2005.tb03740.x

Sinha, E., Michalak, A. M., 2016. Precipitation Dominates Interannual Variability of Riverine

Nitrogen Loading across the Continental United States. Environ Sci Technol. 50 12874-12884.https://doi.org/10.1021/acs.est.6b04455

Skinner, J. A., et al., 1997. An overview of the environmental impact of agriculture in the UK. Journal of Environmental Management. 50 111-128.https://doi.org/10.1006/jema.1996.0103

Smith, R. A., et al., 1997. Regional interpretation of water-quality monitoring data. Water resources research. 33 2781-2798 https://doi.org/10.1029/97WR02171

Strobl, C., et al., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 14 323-48.https://doi.org/10.1037/a0016973

Strumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems. 41 647-665.https://doi.org/10.1007/s10115-013-0679-x

Suescun, D., et al., 2017. Vegetation cover and rainfall seasonality impact nutrient loss via runoff and erosion in the Colombian Andes. Regional Environmental Change. 17 827-839.https://doi.org/10.1007/s10113-016-1071-7

Sun, A. Y., Scanlon, B. R., 2019. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environmental Research Letters. 14 073001.https://doi.org/ARTN07300110.1088/1748-9326/ab1b7d

Sutadian, A. D., et al., 2016. Development of river water quality indices—a review. Environmental monitoring and assessment. 188 58 https://doi.org/10.1007/s10661-

015-5050-0

Thomas, I. A., et al., 2016. Improving the identification of hydrologically sensitive areas using LiDAR DEMs for the delineation and mitigation of critical source areas of diffuse pollution. Sci Total Environ. 556 276-90.https://doi.org/10.1016/j.scitotenv.2016.02.183

Valle Junior, R. F., et al., 2014. Environmental land use conflicts: A threat to soil conservation. Land Use Policy. 41 172-185.https://doi.org/10.1016/j.landusepol.2014.05.012

van Vliet, M. T. H., et al., 2013. Global river discharge and water temperature under climate change. Global Environmental Change. 23 450-464. https://doi.org/10.1016/j.gloenvcha.2012.11.002

Varanka, S., et al., 2015. Geomorphological factors predict water quality in boreal rivers. Earth Surface Processes and Landforms. 40 1989-1999.https://doi.org/10.1002/esp.3601

Viaroli, P., et al., 2018. Space and time variations of watershed N and P budgets and their relationships with reactive N and P loadings in a heavily impacted river basin (Po river, Northern Italy). Sci Total Environ. 639 1574-1587.https://doi.org/10.1016/j.scitotenv.2018.05.233

Vorosmarty, C. J., Sahagian, D., 2000. Anthropogenic disturbance of the terrestrial water cycle. Bioscience. 50 753-765.https://doi.org/10.1641/0006-3568(2000)050[0753:Adottw2.0.Co;2

Wang, F., et al., 2021a. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environ Res. 202 111660.https://doi.org/10.1016/j.envres.2021.111660

Wang, F., et al., 2021b. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environmental Research. 202

111660.https://doi.org/https://doi.org/10.1016/j.envres.2021.111660

Wang, Z.-J., et al., 2020. Rainfall driven nitrate transport in agricultural karst surface river system: Insight from high resolution hydrochemistry and nitrate isotopes. Agriculture, Ecosystems & Environment. 291 106787.https://doi.org/10.1016/j.agee.2019.106787

Winemiller, K. O., et al., 2016. DEVELOPMENT AND ENVIRONMENT. Balancing hydropower and biodiversity in the Amazon, Congo, and Mekong. Science. 351 128-9.https://doi.org/10.1126/science.aac7082

Wu, Z., et al., 2018. Assessing river water quality using water quality index in Lake Taihu Basin, China. Sci Total Environ. 612 914-922.https://doi.org/10.1016/j.scitotenv.2017.08.293

Xia, X. H., et al., 2015. Potential Impacts of Climate Change on the Water Quality of Different Water Bodies. Journal of Environmental Informatics. 25 85-98. https://doi.org /10.3808/jei.201400263

Xu, G., et al., 2021a. Spatio-temporal characteristics and determinants of anthropogenic nitrogen and phosphorus inputs in an ecologically fragile karst basin: Environmental responses and management strategies. Ecological Indicators. 133 108453. https://doi.org/ 10.1016/j.eco lind.

Xu, G. Y., et al., 2019. Influence of Landscape Structures on Water Quality at Multiple Temporal and Spatial Scales: A Case Study of Wujiang River Watershed in Guizhou. Water. 11 159.https://doi.org/ARTN15910.3390/w11010159

Xu, G. Y., et al., 2021b. Spatio-temporal characteristics and determinants of anthropogenic nitrogen and phosphorus inputs in an ecologically fragile karst basin: Environmental responses and management strategies. Ecological Indicators. 133 108453.https://doi.o rg/ARTN 10845310.1016/j.ecolind.2021.108453

Yan, W., et al., 2021a. The effect of landscape complexity on water quality in mountainous urbanized watersheds: a case study in Chongqing, China. Landscape and Ecological Engineering. 17 165-193.https://doi.org/10.1007/s11355-021-00448-9

Yan, W. T., et al., 2021b. The effect of landscape complexity on water quality in mountainous urbanized watersheds: a case study in Chongqing, China. Landscape and Ecological Engineering. 17 165-193.https://doi.org/10.1007/s11355-021-00448-9

Yi, Q. T., et al., 2017. Tracking Nitrogen Sources, Transformation, and Transport at a Basin Scale with Complex Plain River Networks. Environmental Science & Technology. 51 5396-5403.https://doi.org/10.1021/acs.est.6b06278

Yue, F. J., et al., 2019. Land use interacts with changes in catchment hydrology to generate chronic nitrate pollution in karst waters and strong seasonality in excess nitrate export. Science of the Total Environment. 696 134062. https://doi.org/ARTN13406210.1016/j.scitotenv34062

Zalidis, G., et al., 2002. Impacts of agricultural practices on soil and water quality in the Mediterranean region and proposed assessment methodology. Agriculture Ecosystems & Environment. 88 137-146.https://doi.org/10.1016/S0167-8809(01)00249-3

Zeller, K. A., et al., 2016. Using simulations to evaluate Mantel-based methods for assessing landscape resistance to gene flow. Ecol Evol. 6 4115-28. https://doi.or g/10. 1002/ece3.2

154

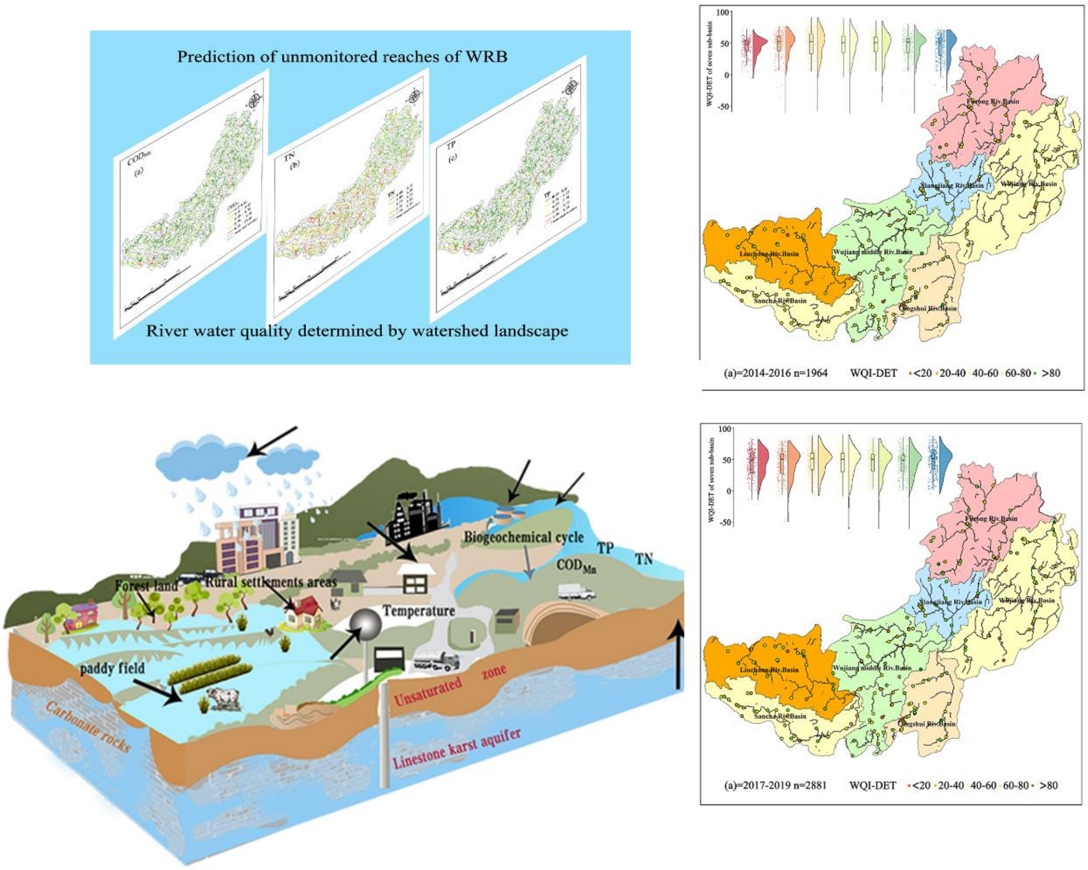
Zeng, C. N., et al., 2020. Modeling Water Allocation under Extreme Drought of South-to-North Water Diversion Project in Jiangsu Province, Eastern China. Frontiers in Earth Science. 8.https://doi.org/ARTN 54166410.3389/feart.2020.541664

Zhang, L., et al., 2004. A rational function approach for estimating mean annual evapotranspiration. Water Resources Research. 40.https://doi.org/ArtnW0250210.1029/2003wr002710

Zhang, Y. G., et al., 2018. A High-Resolution Global Map of Soil Hydraulic Properties Produced by a Hierarchical Parameterization of a Physically Based Water Retention Model. Water Resources Research. 54 9774-9790.https://doi.org/10.1029/2018wr023539

Zhang, Z., et al., 2020. Coupled hydrological and biogeochemical modelling of nitrogen transport in the karst critical zone. Sci Total Environ. 732 138902.https://doi.org/10.1016/j.sc itotenv .2020.138902

Zou, X.-Y., et al., 2019. A Novel Event Detection Model for Water Distribution Systems Based on Data-Driven Estimation and Support Vector Machine Classification. Water Resources Management. 33 4569-4581 https://doi.org/10.1007/s11269-019-02317-5


Prediction of unmonitored reaches of WRB
River water quality determined by watershed landscape
Biogeochemical cycle
TP
TN
COD$_{Mn}$
Forest land
Rural settlements areas
Temperature
paddy field
Carbonate rocks
Unsaturated zone
Linestone karst aquifer
WQI-DET of seven sub-basin
(a)=2014-2016 n=1964
WQI-DET <20 20-40 40-60 60-80 >80
(a)=2017-2019 n=2881
WQI-DET <20 20-40 40-60 60-80 >80