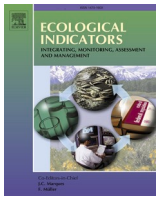
Ecological Indicators 165 (2024) 112156



Contents lists available at ScienceDirect

Ecological Indicators

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



Original Articles



Building an XGBoost model based on landscape metrics and meteorological data for nonpoint source pollution management in the Nakdong river watershed

Sun Hee Shim, Jung Hyun Choi *

Department of Environmental Science and Engineering, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, South Korea

ARTICLE INFO

Keywords:
Land use
Landscape
Machine learning
Total Maximum Daily Load
Water quality
Watershed

ABSTRACT

Purpose: To effectively operate river water quality under current fourth-phase Total Maximum Daily Load (TMDL) management system, we built a machine learning model that predicts whether water quality goals are achieved for the entire Nakdong river watershed in Korea.

Methods: First, to consider the effects of land use type on the runoff characteristics of pollutants, K-means clustering was used to classify the watershed into three areas: agricultural areas, forest areas, and urban areas. Next, we developed a machine learning model to predict the achievement of BOD, TP, and TOC water quality goals in the different rainfall seasons. At this time, the Isolated Forest and ADASYN machine learning techniques were used to preprocess the training data. Finally, SHAP was used to find the factors with the greatest effects on the achievement of water quality goals.

Results: This model's average prediction results for TP, BOD, and TOC showed accuracy ranging from 0.6 to 1.0. Meteorological factors, particularly monthly precipitation and average temperature, were found to highly influence the model predictions for all land use types. In the landscape metrics, ED showed a high level of importance in all land use types. CONTAG was the main factor in agricultural areas; ED, LPI, CONTAG, COHESION and SHDI were the main factors in forest areas; and PD, ED, SHDI, and COHESION were the main factors in urban areas.

Conclusion: The monthly precipitation and average temperature significantly affected whether the TMDL water quality goals were achieved in all sub watersheds, and the landscape metrics calculated as highly influenced factors differed depending on the land use type. Therefore, customized watershed management according to land use characteristics is necessary. These results provide valuable ideas for land use managers and landscape planners to achieve water quality goals through the management of non-point source pollution.

1. Introduction

Rivers are vital ecosystems for preserving ecological equilibrium (Wu et al., 2021). The water quality of rivers is deteriorating steadily due to the cumulative effects of human activities (Xu et al., 2022a; Wu and Lu, 2021). such that wastewater discharge has reached a limit, even if pollutant sources comply with discharge standards (Kim et al., 2023). To address that issue, comprehensive water management measures and special laws were introduced in 1988 (Cho et al., 2020). Since 2004, South Korea has implemented TMDL management system to manage water quality. This system aims to establish scientifically grounded water quality goals for individual sub watersheds and involves calculating pollutant loads under which the desired water quality can be attained and upheld. Moreover, the TMDL management system regulates and oversees the cumulative discharge of pollutant loads from each sub watershed so that the designated goals for overall water quality can be achieved (Kim et al., 2022; Shim et al., 2021; Lee et al., 2013).

With the practice of TMDL management system, Biochemical Oxygen

Abbreviations: TMDL, Total Maximum Daily Load; BOD, Biochemical Oxygen Demand; TP, Total Phosphorus; TOC, Total Organic Carbon; NRW, Nakdong River Watershed; XGBoost, Extreme Gradient Boosting; SHAP, SHapley Additive exPlanations; ASOS, Automated Synoptic Observing System; AWS, Automatic Weather System; IF, Isolation Forest; ADASYN, Adaptive Synthetic Sampling; LPI, Largest Patch Index; LSI, Landscape Shape Index; PD, Patch Density; ED, Edge Density; CONTAG, Contagion Index; COHESION, Cohesion Index; SHDI, Shannon's Diversity Index; CWA, Clean Water Act; WWTPs, Waste Water Treatment Plants.

* Corresponding author.
E-mail addresses: sunhee960@naver.com (S. Hee Shim), jchoi@ewha.ac.kr (J. Hyun Choi).

https://doi.org/10.1016/j.ecolind.2024.112156
Received 1 November 2023; Received in revised form 6 March 2024; Accepted 17 May 2024
Available online 30 May 2024
1470-160X/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Demand (BOD) was designated as a target pollutant in 2004, and Total Phosphorus (TP) was added in 2016. In the Nakdong River Watershed (NRW), the application of TMDL management system led to a significant reduction of approximately 63 % in BOD discharges, from 176 tons/day in 2004 to 65 tons/day in 2020, and TP discharges decreased from 11 tons/day in 2010 to 4 tons/day in 2020 (ME, 2022). However, downstream of the Nakdong river, in the drinking water supply area for a metropolitan area, the Total Organic Carbon (TOC) concentration increased by 19 % from 3.7 mg/L in 2011 to 4.4 mg/L in 2020, highlighting the need for TOC management (ME, 2022). Therefore, starting from the second half of 2023, TOC was selected as a pilot target for TMDL implementation. See (Table 1).

Despite TMDL implementation and management efforts, non-point source pollution contributes around 70 % of the total pollutant load, posing challenges in achieving the water quality standard (Yu et al., 2020; Joint Agency, 2012). Non-point source pollution is challenging to manage because its sources are dispersed, making it difficult to identify specific discharge points (Zhou et al., 2023; Ki et al., 2019; Korea Institute of Civil engineering and building Technology (KICT), 2000). Pollutants are carried to the rivers through surface runoff from various land uses, including agricultural areas, forests, and urban areas (Jung et al., 2012).

The land use composition (percentage of various land uses) and configuration (structure of land use) are influenced by factors such as natural site conditions, cultivation traditions, social development, and past and present economic forces (Pan et al., 2023; Rahman and Szabó, 2021; Kaim et al., 2018; Bartel, 2000). These natural conditions and anthropogenic activities can be reflected by landscape metrics. Because river water quality is increasingly influenced by the landscape, it is necessary to quantify the role of landscape metrics at the watershed scale to develop better management strategies for river water quality (Nafi'Shehab et al., 2021; Zhang et al., 2020, Xu et al., 2022a).

Many recent studies have used landscape metrics to manage river water quality. Huang (2016) analyzed the correlation between landscape metrics and water quality in the Gorges reservoir area. Their study revealed a close relationship between water quality in the reservoir and landscape metrics such as Patch Density (PD) due to fragmentation of the watershed caused by land use changes, urbanization, and agricultural activities. Zhang (2022) performed a linear regression analysis among water quality and land use changes and landscape metrics of the Jing river in northwestern China and found that the cohesion was significantly correlated with the Chemical Oxygen Demand (COD), mineralization degree, and total nitrogen (TN). Moreover, PD correlated negatively with TN. Nafi'Shehab et al. (2021) studied how landscape metrics and land use changes affected water quality in the Bentong river basin in Malaysia. As a result, they emphasized the need for buffer zones in river water quality management because tropical forests have a positive influence on water quality.

Although previous studies have used statistical methods to analyze relationship between landscape metrics and water quality, machine learning has not been widely used to predict and analyze water quality based on landscape metrics. Machine learning approaches have gained attention as methods that might overcome the limitations of physically based models such as HSPF, SWAT, and AGNPS, which can exclude essential characteristics within large watersheds through parameter calibration and uncertainty (Xu et al., 2022b; Hashemi et al., 2016).

In this study, we developed an XGBoost (Extreme Gradient Boosting) machine learning model to predict the achievement of water quality targets (BOD, TP, TOC) in the NRW. After building the machine learning model, we used SHAP (Shapley Additive exPlanations) values to identify factors that most influenced whether the water quality goals were achieved. The model considered not only landscape metrics but also weather conditions to account for the acceleration of global warming and climate change. Intense rainfall events are becoming more frequent in the region because the continuous rise in atmospheric temperature drives fluctuations in atmospheric energy and water circulation (Fowler et al., 2021). Kwon et al. (2008) found that the rainfall pattern in South Korea tend to increase in summer and fall from 1970 to 2000. Notably, this study showed significant effects near the NRW, especially along the South Sea and East Sea. generally, South Korea's weather exhibits distinct monsoonal characteristics, with significant seasonal variations in precipitation that produce substantial fluctuations in runoff carrying non-point source pollutants (Kang et al., 2010).

Our main objectives for this research are as follows: (a) Build a machine learning model to predict whether TMDL water quality goals will be achieved; (b) Identify which land use types and landscape metrics influence water quality in the NRW during both the dry and wet seasons; (c) Provide fundamental data for establishing optimal river water quality management strategies based on watershed characteristics.

Table 1
Landscape metrics that can be analyzed in FRAGSTATS4.2.

Index (Abbreviation)	Unit	Equation	Range
Largest patch index (LPI)	%	$LPI = \frac{max(a_{ij})}{A}(100) a_{ij} = area(m^2) of patch ij. A = total landscape area(m^2).$	$0 < LPI \leq 100$
Landscape shape index (LSI)	None	$LSI = \frac{.25E^*}{\sqrt{A}} E^* = total length(m) of edge in landscape; includes the entire landscape boundary and some or all background edge segments.$ $A = total landscape area(m^2).$	$1 \leq LSI < \infty$
Patch density (PD)	N/100 ha	$PD = \frac{N}{A}(10,000)(100) N = total number of patches in the landscape. A = total landscape area(m^2).$	$0 < PD < \infty$
Edge density (ED)	Meters per ha	$ED = \frac{E}{A}(10,000) E = total length(m) of edge in landscape. A = total landscape area(m^2).$	$0 \leq PD < \infty$
Contagion index (CONTAG)	%	$CONTAG = \left[1 + \frac{\sum_{i=1}^{m}\sum_{k=1}^{m}\left[P_i \bullet \frac{g_{ik}}{\sum_{k=1}^{m} g_{ik}}\right] \bullet \left[ln\left(P_i \bullet \frac{g_{ik}}{\sum_{k=1}^{m} g_{ik}}\right)\right]}{2\ln(m)}\right](100) P_i =$ $proportion of the landscape occupied by patch type(class) i. g_{ik} =$ $number of adjacencies(joins) between pixels of patch types(classes) i and k based on the double-count method. m =$ $number of patch types(classes) present in the landscape, including the landscape border if present.$	$0 < CONTAG \leq 100$
Cohesion index (COHESION)	None	$COHESION = \left[1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij}^*}{\sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij}^* \sqrt{a_{ij}^*}}\right] \bullet \left[1 - \frac{1}{\sqrt{Z}}\right]^{-1} \bullet (100) P_{ij}^* = perimeter of patch ij in terms of number of cell surfaces. a_{ij}^* =$ $area of patch ij in terms of number of cells. Z = total number of cells in the landscape.$	$0 < COHESION < 100$
Shannon's diversity index (SHDI)	Information	$SHDI = -\sum_{i=1}^{m}(P_i \bullet lnP_i) P_i = proportion of the landscape occupied by patch type(class) i.$	$0 < SHDI < \infty$

2. Methods

2.1. Study area

The Nakdong river is the longest river in South Korea, located in the southeastern part of the Korean Peninsula, with a length of 510.36 km and a drainage area of 23,790 km^2, which accounts for 25.9 % of the total area of South Korea (Cho and Kim, 2018) (Fig. 1). The study area, NRW, is characterized by numerous mountains and wide valleys and is associated with larger seasonal and regional changes than other rivers in Korea. It plays a crucial role in providing drinking water, domestic water, and agricultural water for the southeastern part of Korea, including several metropolitan areas (Kang et al., 2019).

The weather in the NRW is generally influenced by seasonal movement of the polar front, which produces cold and hot climates via continental air movement from northern Asia and tropical and humid air currents from the Pacific Ocean, respectively. Winter lasts mainly from November to March, with cold dry winds blowing from the northwest, whereas summer, from June to mid-September, is characterized by the continuous influence of tropical and humid air currents from the Pacific Ocean (Shon et al., 2010).

Since 2004, officials have been implementing TMDL management for 41 sub-watersheds in the NRW and achieved the water quality goals. BOD, TN, TOC, and flow variations have been monitored at the boundaries of cities and counties within the sub-watersheds since October 2010 to evaluate the implementation of TMDL management (Cho et al., 2020). Fig. 1 shows the NRW and the locations of the water-quality monitoring network in an illustration provided by the National Spatial Data Infrastructure Portal operated by the Ministry of Land, Infrastructure and Transport (https://www.nsdi.go.kr).

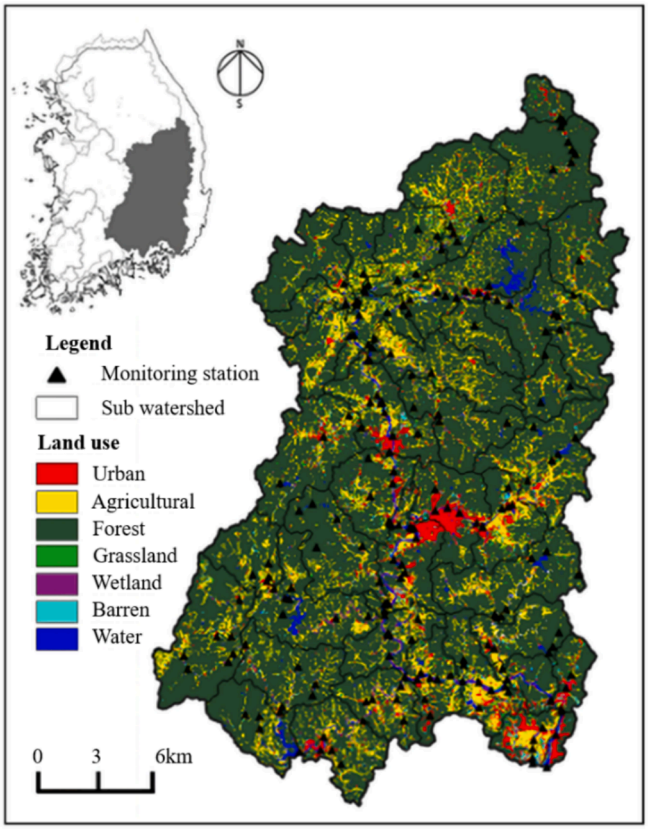


Fig. 1. Location of NRW and water quality monitoring network sites.

2.2. Data resources and preprocessing

2.2.1. Water quality and meteorological data resources

First, we developed an integrated database from the 204 water quality monitoring stations within the 40 sub-watersheds except 1 sub-watershed comprised of highly urbanized metropolitan city. Through the Water Environment Information System operated by the Ministry of the Environment (https://water.nier.go.kr/), we collected monthly TP, BOD, and TOC concentration (*unit: mg/L*) data from all 204 monitoring stations from January 2017 to December 2021. Meteorological data were obtained from the Open MET Data Portal provided by the Korea Meteorological Administration (https://data.kma.go.kr/). Monthly precipitation (*unit: mm*) and average temperature (*unit:* ℃) data from January 2017 to December 2021 were collected from the Automated Synoptic Observing System (ASOS) and Automatic Weather System (AWS) points closest to the 204 water quality monitoring stations.

After developing the water quality and meteorological database, we considered the land use type of the sub watersheds in relation to different rainfall patterns. The study period was divided into wet seasons (June, July, August) and dry seasons excluding winter (March, April, May, September, October, November).

2.2.2. Removing outliers with an isolation forest

Good data management practices require that data be thoroughly screened to identify outliers, which could potentially be incorrect values stemming from analytical or transcription errors (Robinson et al., 2005). Isolation Forest (IF) is an unsupervised learning anomaly detection algorithm based on ensemble learning widely used to identify anomalies in complex data (Song et al., 2021; Ahmed et al., 2019). Therefore, we used IF, a decision tree-based outlier detection method, to remove outliers. The advantage of IF is that outliers can be directly identified based on the concept of isolation. In IF, the term isolation means separating one instance from the other instance (Liu et al., 2012). IF can achieve high detection performance with high efficiency with only a small subsampling size than other outlier detection algorithms such as density-based and distance-based algorithms, and neither distance nor density measurements are essential for anomaly detection (Wang et al., 2024). Therefore, by isolating instances in the data, the technique effectively identifies outliers (Chen et al., 2016).

The IF does not rely on data distance or density but instead uses a machine learning model called a decision tree to create a classification model (Fig. 2). IF is similar to dichotomy as it does not require defining a mathematical model or training (Zou et al., 2019). See (Fig. 3 and Fig. 4).

Because IF is based on the assumption that abnormal instances are usually fewer than and different from normal instances in a data set, they are easier to isolate than normal instances in many binary tree structures (Song et al., 2021). The IF randomly divides the data using random splitting criteria. The tree structure is constructed from root to leaves by partitioning the data at each node using randomly selected features and thresholds (Wang et al., 2024). Normal instances (x_i) require many splits to reach the terminal node, whereas outlier instances

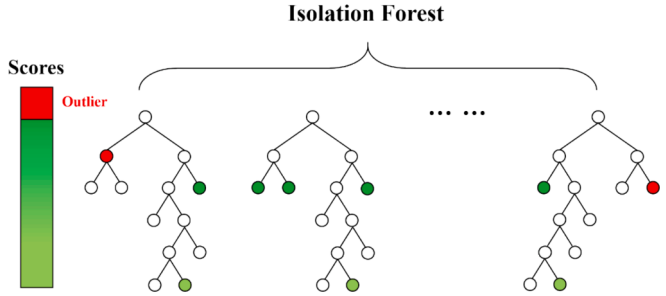


Fig. 2. Description of how the Isolation Forest detects outliers.

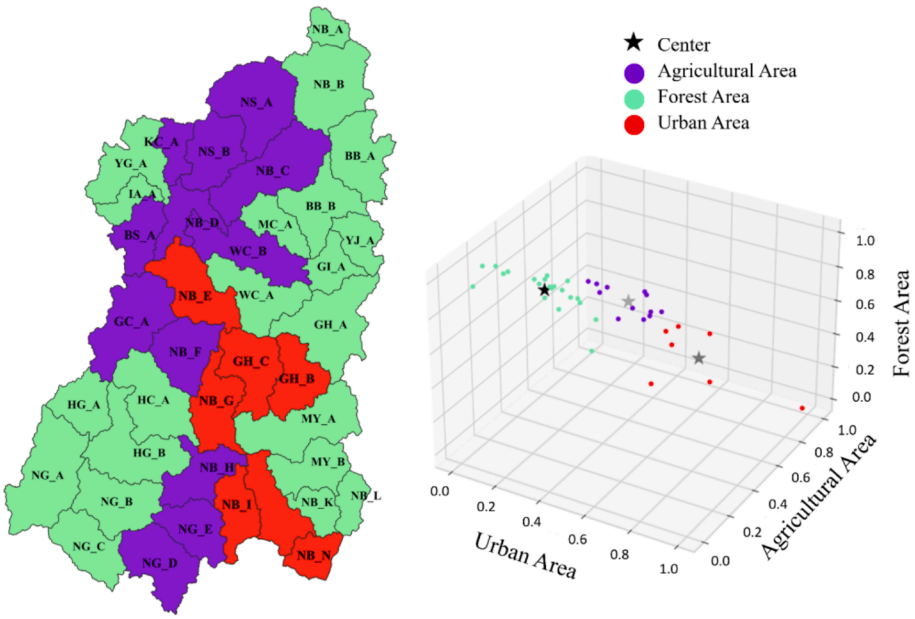


Fig. 3. K-means Clustering results and map.

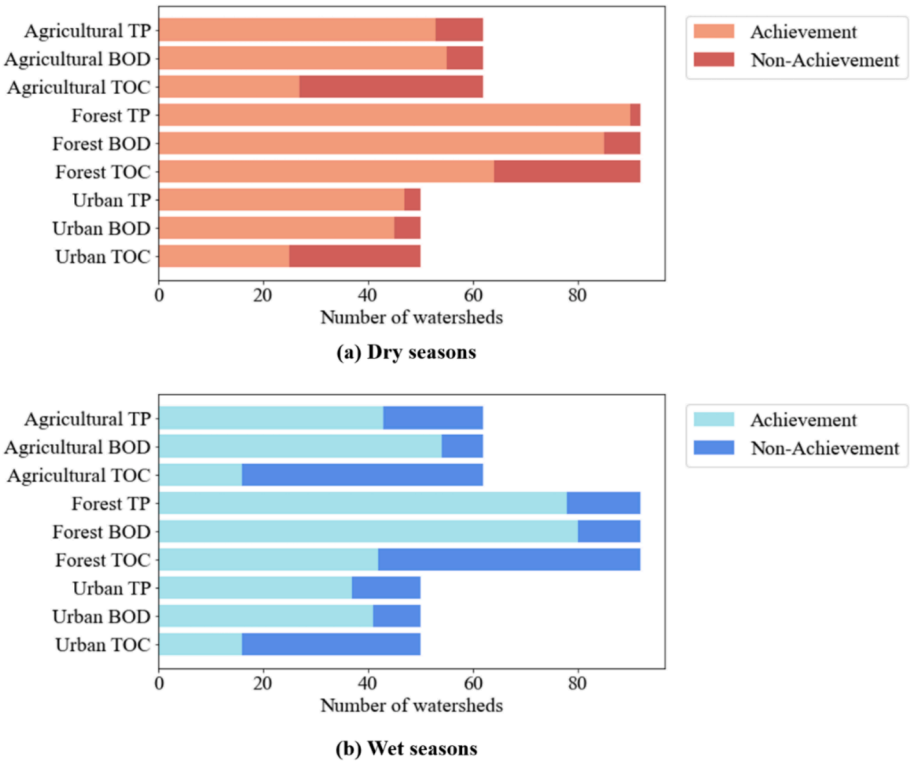


Fig. 4. Rate of achievement and non-achievement of water quality goals in the (a) Dry seasons and (b) Wet seasons.

(x_o) reach the terminal node with only a few splits, leading to their isolation (Chen et al., 2016). Overall, the average path length of an anomalous instance in a tree is much shorter than the average path length of a regular instance, and the unusual score assigned to an instance can be computed using the average path length of the instance in a binary tree (Liu et al., 2012). In other words, outliers are likely to be isolated early during tree partitioning, positioning them close to the root node of the tree, whereas normal instances tend to be located closer to the leaf nodes. The Isolation Forest thus classifies outliers and normal instances based on the lengths of the paths they take to isolation.

2.2.3. Adaptive synthetic sampling

Adaptive Synthetic Sampling (ADASYN) is a data augmentation technique used to address class imbalance in minority class data. It is an improved version of the Synthetic Minority Over-sampling Technique (SMOTE) (Gosain and Sardana, 2017). ADASYN generates an appropriate number of synthetic data samples for each minority datapoint based on the density distribution calculated using the K-nearest neighbor function (Xu et al., 2020).

In ADASYN, some degree of noise is added to the synthetic data samples to resolve the overfitting problem encountered in SMOTE,

where the generated synthetic data become too similar to the original data (Gosain and Sardana, 2017). In this study, we set the hyperparameter 'n_neighbors' in the ADASYN model, which determines the number of neighboring data points used when creating new synthetic data, to 1. This setting allows for the diverse generation of synthetic samples for minority data instances.

2.2.4. K-means clustering

The Summary Report on the Basic Plan for Total maximum daily Load Management in the Nakdong River Watershed provides land use data for each sub watershed classified into 14 categories: paddy fields, dry fields, orchards, ranch lands, forests, residential areas, industrial sites, public facilities, school grounds, transportation areas, gas station sites, sports facilities, amusement parks, and other areas. However, in order to effectively study the characteristics of the entire NRW, we deemed it necessary to simplify the land use types into three categories. Therefore, based on land use data for 2017, which is the benchmark year for implementing the 4th step of TMDL management system, paddy fields, dry fields, and orchards were grouped as agricultural areas; residential areas, industrial sites, public facilities, school grounds, transportation areas, gas station sites, sports facilities, amusement parks, and other areas were grouped as urban areas; and forests were classified as forest areas. Ranch lands were excluded from the classification because they account for less than 2.3 % of the total area in all sub-watersheds.

Clustering is an unsupervised learning technique that groups similar data into clusters. It can be broadly categorized as hierarchical clustering and nonhierarchical clustering (Lee, 2014). K-means clustering is one of the most commonly used nonhierarchical clustering algorithms. First, we identify k centroids and then assign data points to the clusters with the closet centroids based on Euclidean distance (Lee, 2012).

To determine the most appropriate number of centroids, we used the Elbow method by implementing the KElbowVisualizer function from the yellowbrick library in python3.9.1 and testing k values from 1 to 9. After $k = 3$, the inter-cluster distances did not change significantly but the training time increased. Therefore, the optimal number of clusters was 3. According to the K-means clustering results, 21 sub-watersheds were classified as forest areas, and 12 sub-watersheds were classified as agricultural areas and the remaining 7 watersheds were classified as urban areas.

2.3. Modeling and database processing

2.3.1. FRAGSTATS

Landscape refers to the spatial and temporal patterns that result from interactions between nature and human activities. *Landscape change* refers to alterations in the shape and location of landscape elements or their replacement by various external factors such as natural disturbances, disturbances caused by human activities, and others (Heo et al., 2007). To quantitatively assess landscape changes, ecologists have developed landscape metrics, which represent the structure, function, and change patterns of landscape ecosystems in numerical terms (Baker and Cai, 1992).

Among the various software programs available for calculating landscape metrics, FRAGSTATS computes a wide variety of landscape metrics for categorical map patterns. The original software was released in the public domain in 1995 in association with the publication of a USDA Forest Service General Technical Report (McGarigal and Marks, 1995). FRAGSTATS facilitates the quantitative evaluation of landscape structures, shapes, functions, and changes within a study area and can consider the most diverse range of landscape metrics among the currently used programs (Kang et al., 2018). Indeed, numerous studies have provided compelling evidence that landscape metrics play a crucial role in elucidating the relationship between land uses and water quality and affect hydrological processes (Nafi'Shehab et al., 2021).

Therefore, we assessed how landscape metrics influence river water quality in reference to various previous studies. Specifically, we calculated the Largest Patch Index (LPI), Landscape Shape Index (LSI), Patch Density (PD), Edge Density (ED), Contagion Index (CONTAG), Cohesion Index (COHESION), and Shannon's Diversity Index (SHDI) at the landscape scale (Xu et al., 2023; Yao et al., 2023; Xu et al., 2022a; Zhang et al., 2022; Huang et al., 2016).

Largest patch index (LPI) is the area of the largest patch in a landscape (m^2) divided by the total landscape area (m^2) and converted to a percentage. LPI shows the dominance of the largest patch over the entire landscape area and is a measure of dominant land use characteristics. Landscape shape index (LSI) indicates the shape and complexity of landscape patches through edge length and density compared to landscape area. If the landscape shape consists of a single square patch, the value is 1, and if the landscape shape is irregular, the value increases infinitely as the edge length within the landscape increases.

Shannon's diversity index (SHDI) is a value obtained by adding up the ratio of patch types multiplied by the natural logarithm of the ratio for all patch types and adding a minus sign. It has a value of 0 when there is only one landscape patch, that is, when there is no diversity. Patch density (PD) is the number of patches in the landscape divided by the total landscape area (m2) and then multiplied by 10,000 and 100. The total landscape area (A) includes the internal landscape. Edge density (ED) is the sum of all edge lengths (m) in the landscape divided by the total landscape area (m2) multiplied by 10,000. The larger the value, the more edges there are in the landscape, meaning it is more fragmented.

Lastly, Contagion index (CONTAG) is an index that measures continuity between adjacent patches and reflects all patch types in the landscape. And the cohesion index (COHESION) is an index that measures the physical connectivity of each patch in the landscape, and the higher the physical connectivity of each patch in the landscape, the higher the value.

We calculated those landscape metrics based on the land cover classification system provided by the Environmental Geographic Information Service, which includes major categories such as urban dry area, agricultural area, forests, grassland, wetland, barren land, and water areas. Calculations within a 1 km radius of the water quality monitoring network within the NRW were performed using FRAGSTATS 4.2 version.

2.3.2. XGBoost

XGBoost, developed by Tianqi Chen in 2016, is a versatile machine-learning method suitable for classification, regression, and ranking tasks. Its superior predictive performance in classification, compared with other machine-learning models, has been widely recognized (Chung et al., 2021; Lee and Sun, 2020; Chen and Guestrin, 2016). As a Gradient Boosting Machine (GBM), XGBoost uses gradient descent with boosting techniques to iteratively create new predictive models and update weights in the direction of improved accuracy (Oh et al., 2019).

Unlike conventional GBM, which updates weights sequentially during training, XGBoost achieves faster computation through parallel processing. Additionally, it incorporates cross-validation and regularization functions to address overfitting and enhance model robustness (Sim et al., 2022; Chen and Guestrin, 2016). The structure score of XGBoost is defined by its objective function, which explains training loss and regularization and contributes to its effective performance.

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{1}$$

L(θ) is used to evaluate the model performance on the training data, and Ω(θ) is used for regularization to control overfitting of the model. Additionally, the complexity of the tree is computed as follows:

$$\Omega(f) = \gamma T + 0.5\,\lambda \sum_{j=1}^{T} w_j^2 \tag{2-1}$$

$$Obj = \sum_{j=1}^{T} \left[G_j w_j + 0.5\left(H_j + \lambda\right) w_j^2 \right] + \Gamma t \tag{2-2}$$

- T represents the number of terminal nodes in the tree.
- w_j is the weight associated with the j-th terminal node.
- γ and λ are hyperparameters that control the regularization strength.

This formulation combines both the number of terminal nodes (to control the depth of the tree) and the sum of squared weights (to control the complexity of the individual nodes) for effective regularization. The equation $G_j w_j + 0.5(H_j + \lambda) w_j^2$ is a quadratic function used to find the optimal w_j, where w_j denotes the vector of scores on leaves, and T signifies the individual leaves.

Subsequently, we used GridSearchCV to optimize the hyperparameters of the XGBoost model. XGBoost offers superior predictive performance, compared with other machine-learning techniques, because it supports more than 86 hyperparameters (Hwang et al., 2022). The selection of appropriate hyperparameters during model training significantly influences prediction accuracy, so identifying the most suitable ones for training data is crucial (Park et al., 2020).

In this study, we optimized four key hyperparameters, n_estimators, max_depth, colsample_bytree, and learning_rate, in GridSearchCV with 4-fold cross-validation. Specifically, we considered integer values ranging from 5 to 15 for n_estimators and 2 to 10 for max_depth. For colsample_bytree, we varied the values from 0.5 to 0.9 in increments of 0.1. To explore the learning_rate, we tested values of 0.005, 0.01, 0.05, and 0.1 to identify the most optimal set of hyperparameters for the XGBoost model. Throughout this process, we randomly selected 70 % of our data for the training set, leaving the remaining 30 % for use as the test set. This approach enabled us to fine-tune the XGBoost model and enhance its prediction accuracy, thereby providing a robust foundation for our research analyses.

2.3.3. SHAP (SHapley additive exPlanations) values

Machine-learning models have a challenging black-box nature, rendering the internal decision-making process difficult to comprehend. To address that issue, SHAP has proved to be a valuable tool for explaining model outputs and visualizing complex causal relationships between the driving forces and the prediction target (Li et al., 2018).

As an additive explanation model, SHAP draws inspiration from the theoretically optimal Shapley value of cooperative game theory, which treats all characteristics as contributors (Štrumbelj and Kononenko, 2014). The determination of Shapley values adheres to several axioms, ensuring fair allocation of contributions within a group N comprising N features (Lundberg et al., 2017). The formulation of a linear function of binary features $g(z')$ is realized through the additive feature attribution method represented by Equation 3, where z' takes the value 1 when a feature is observed and 0 otherwise, and M denotes the number of input features.

$$\varnothing_i = \sum S \subseteq F\{i\} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{Su\{i\}m.}\left(x_{Su\{i\}}\right) - f_S(x_s) \right] \quad (3\text{-}1)$$

$$g(z') = \varnothing_0 + \sum_{i=1}^{M} \varnothing_i Z_i^1 \quad (3\text{-}2)$$

To accurately compute TreeSHAP values in tree integration models, we used the TreeExplainer proposed by Lundberg and Lee (2017). The SHAP values for a given prediction variable and observation value reflect variations in the outputs, particularly the values associated with the predicted water quality indicator. This variability determines the suitability of using the prediction variable after each observation. The mean absolute SHAP values of all observations effectively synthesize the significance of global features and can be interpreted via scatterplots of individual predictive variables (Just et al., 2020).

3. Results and discussion

3.1. Achievement and non-achievement of water quality TMDL targets in the NRW

TMDLs determine the overall quantity of pollutants that a waterbody can integrate while still adhering to established water quality standards (Kwon et al., 2023). Section 303(d)(1)(C) of the USA Clean Water Act (CWA) requires that TMDLs be developed and implemented for specific pollutants in impaired waters. To achieve national-level TMDL goals, it is necessary to identify and concentrate on managing water quality pollutants that are problematic in each watershed. In addition, section 303(d) indicates that land use controls could potentially be an effective method for mitigating non-point source pollution (Lebowitz, 2001).

Therefore, it is essential to identify watersheds that do not achieve their TMDL water quality goals, examine factors such as land use composition and configurations that influence water quality in those watersheds, and manage those factors intensively (Kim and Yi. 2011). We built an XGBoost model using landscape metrics and meteorological data to investigate whether the TP, BOD, and TOC water quality targets were achieved.

Prior to model development, we compared the database with the water quality goals specified in the TMDL standards for the NRW. The database was classified as “achievement” or “non-achievement” to indicate that the water quality goal was achieved or not. In addition, the database was divided into ‘dry season’ and ‘wet season’ to examine the effects of rainfall patterns.

From the initial data distribution in the dry seasons, the non-achievement ratios of the TMDL water quality targets were as follows: agricultural areas (14.52 %) > urban areas (6.00 %) > forest areas (2.17 %) for TP; agricultural areas (11.29 %) > urban areas (10.00 %) > forest areas (7.61 %) for BOD; agricultural areas (56.45 %) > urban areas (50.00 %) > forest areas (30.43 %) for TOC. Agricultural areas exhibited the highest non– achievement rates, followed by urban areas.

The initial data distribution in the wet seasons showed the following non-achievement ratios: agricultural areas (30.65 %) > urban areas (26.00 %) > forest areas (15.22 %) for TP; urban areas (18.00 %) > forest areas (13.04 %) > agricultural areas (10.00 %) for BOD; agricultural areas (74.19 %) > urban areas (68.00 %) > forest areas (54.35 %) for TOC. Agricultural areas again showed the highest non-achievement rates in the wet seasons, just as in the dry seasons.

Surface runoff is mainly affected by climate and land use type in any region (Li et al., 2018). Several researchers have reported that grass and forest areas generally tend to have a positive influence on the water quality of river (Lee et al., 2009, Walker et al., 2009), while agricultural and residential areas have a negative impact on the water quality of river (Bu et al., 2014, Namugize et al., 2018). In this study, the results were relatively consistent with those of previous researches, and forest area usually has better water quality and plays a crucial role in improving water quality in various watersheds worldwide (de Oliveira et al., 2016). These results reinforce the hypothesis that grass and forest areas have a fixation and absorption impact on pollutants in river systems (Bu et al., 2018). In addition, forest acts as a filter or a sieve that screens and reduces pollutants and sediments conveyed with surface runoff (Ding et al., 2013). Furthermore, grass acts as a retention agent that lessens the amount and effect of surface runoff (Ouyang et al., 2010).

On the contrary, several studies have illustrated that agricultural and urban areas tend to have most adverse effects on the water quality (Huang et al., 2016). The non-achievement in agricultural areas could be influenced by source factors such as the soil and its management. The movement of soil is accompanied by the fertilizers and pesticides used in agricultural areas and is facilitated by the topography and movement of water over or through the soil profile. Elevated ratios of non-achievement were observed during the wet seasons, especially in agricultural areas. The cultivation patterns in agricultural areas could be a source factor for the movement of pollutants to the river through the

baseflow condition. In addition to the spatial effect on water quality, there were obvious temporal effects associated with seasonal changes. This study also confirmed that there was higher rate of non-achievement in the water quality goals during wet seasons. Huang et al. (2016) stated that seasonal fluctuations in hydrological processes, such as precipitation, runoff, and interception, have a significant impact on river flow, resulting in different levels of sediments and nutrients from season to season.

3.2. The results of water quality prediction by XGBoost

We built an XGBoost model using landscape metrics and meteorological data to investigate whether the TP, BOD, and TOC water quality targets were achieved. XGBoost results can be quantified according to their accuracy, precision, recall, and F1-score. To estimate those metrics, the following four parameters need to be determined: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The mathematical equations below are used to evaluate performance (Parsa et al., 2020).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4\text{-}1)$$

$$Precision = \frac{TP}{TP + FP} \quad (4\text{-}2)$$

$$Recall = \frac{TP}{TP + FP} \quad (4\text{-}3)$$

$$F1 - score = \frac{2{*}precision \times recall}{precision + recall} \quad (4\text{-}4)$$

Fig. 5 shows the average accuracy, precision, recall, and F1-score values, with the size of each circle representing the number of datapoints used in the XGBoost model. As shown in Fig. 5, the model most accurately predicted all water quality parameters except TOC in the forest areas. In the dry seasons, the average accuracy for forest areas was 91.08 %, followed by agricultural areas (78.74 %) and urban areas (76.53 %).

The sequence of accuracy in the wet seasons was consistent with that in the dry seasons: forest areas (79.22 %), followed by agricultural areas (78.54 %) and urban areas (71.87 %). It is generally known that machine-learning models require a sufficient amount of data to achieve high accuracy (Jung et al., 2022). In this study, forest areas had an average of 161 datapoints, followed by agricultural areas with an average of 93 datapoints and urban areas with an average of 78 datapoints. Thus, a higher number of datapoints correlates with higher prediction accuracy.

The TOC prediction accuracy in forest areas was lower compared to other water quality parameters (Fig. 5 (b), (e)). This can be explained by the distribution of TOC concentration in forest areas compared to the other water quality parameters in other areas. Fig. 6 showed that the concentration distribution of water quality parameters by land use characteristics using a box plot. A box plot is a method to graphically represent groups of numerical data through quartiles, making it easy to organize large amounts of data and visualize outliers (Horvat and Job, 2020). As a result of the box plot, it was confirmed that TOC concentration in forest areas had greater variability compared to agricultural and urban areas (Fig. 6). The forest areas have large spatial scales across multiple watersheds and temporal variability driven by hydroclimatic conditions such as streamflow, rainfall, and temperature. In general, high variability in water quality across both space and time reduces the ability to accurately assess the status of water quality and develop effective management strategies. Especially, it can be expected that there will be substantial difference in predicting the water quality with high concentration values. Therefore, the spatiotemporal variability in water quality hinder the sound understanding of the relationships with between specific landscape characteristics and water quality (Guo et al., 2020).

The TP prediction accuracy in urban areas during the dry seasons showed the highest accuracy at 100 % (Fig. 5 (c)). The outstanding prediction accuracy for TP might be due to TP discharge from Waste Water Treatment Plants (WWTPs) in the dry seasons. Because large scale WWTPs have a continuous and persistent impact on downstream river water quality through consistent loading characteristics (Kim and Chung, 2022), TP concentration in urban areas during the dry seasons mainly reflects WWTP discharge, showing high prediction accuracy.

Because significant regional variations in the concentrations of water quality parameters occur during wet seasons (Kwon et al., 2020), higher accuracy was expected for the results of the dry season simulations. Except for TOC in agricultural areas, shown in Fig. 5 (a), and BOD and TOC in urban areas, shown in Fig. 5 (c), the model generally displayed superior predictive performance in dry seasons compared with wet seasons.

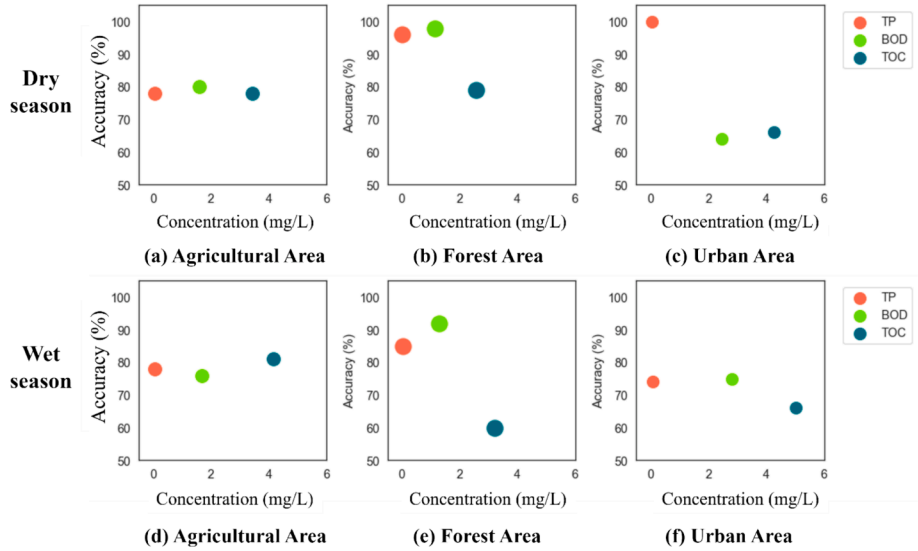


Fig. 5. Concentration and Predictive Accuracy of Targets in the Dry and Wet Seasons (The size of each circle is the number of data used in the XGBoost model).

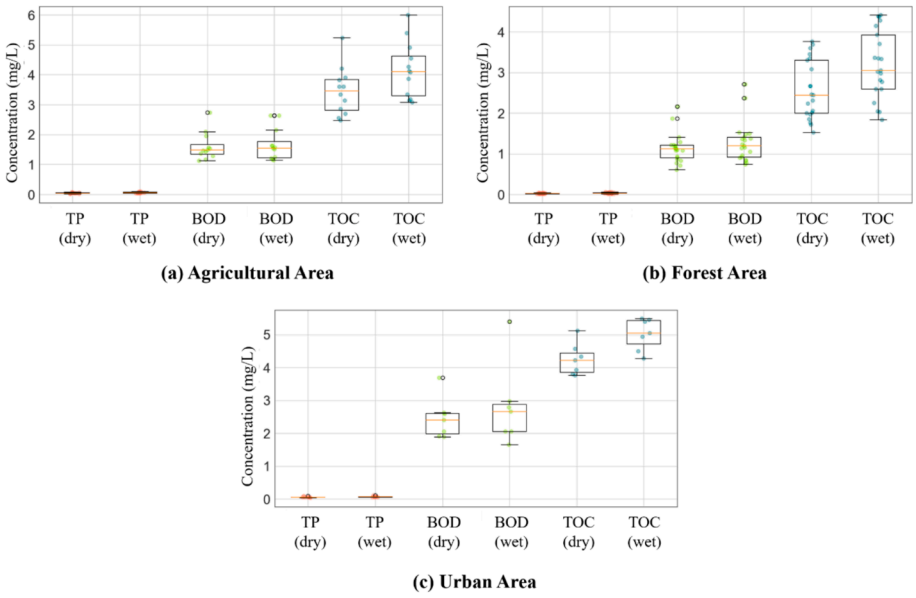


Fig. 6. The concentration distribution of water quality variables by land use types using a box plot. (a) Agricultural area, (b) forest area, and (c) urban area.

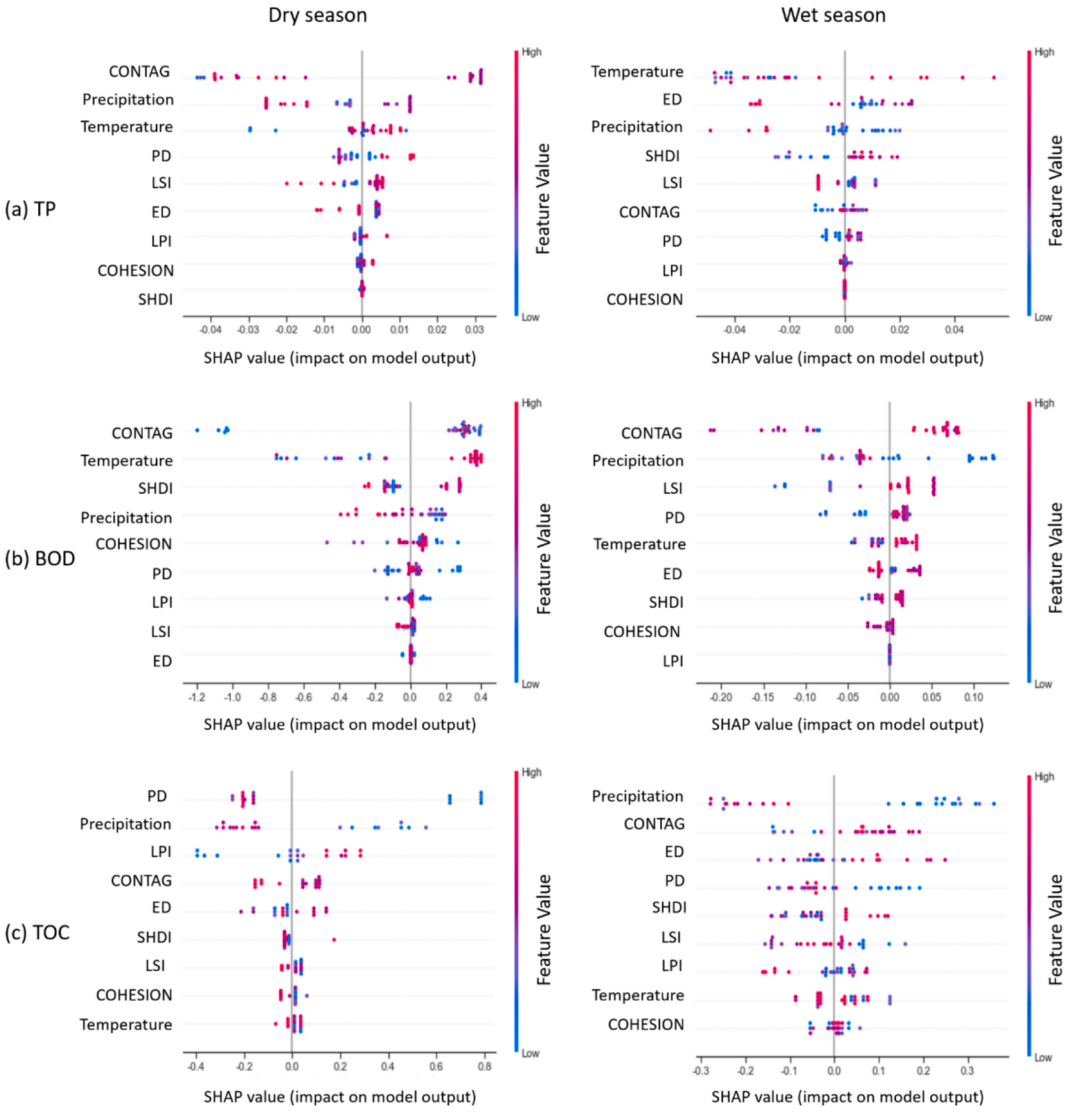


Fig. 7. SHAP summary plots that had the most impact in predicting TMDL water quality goal achievement in agricultural areas.

3.3. Analysis of water quality determinants by SHAP values

3.3.1. Agricultural areas

Non-point source pollution in agricultural areas mainly depends on the land use pattern and hydrological regimes caused by soil texture, topography, and cultivation traditions (Kim et al., 2007). Because it is difficult to investigate the type, amount, toxicity, and effects of pollutants in agricultural areas, we looked for factors associated with land use types and landscape metrics that influenced water quality in the Nakdong River during both the dry and wet seasons through a SHAP analysis of our XGBoost model.

Fig. 7 shows a SHAP summary plot that orders features based on their importance in detecting whether the TMDL water quality targets were achieved. CONTAG, a factor related to land use patterns, was a significant factor in predicting whether the target was achieved for all the water quality parameters (TP, BOD, and TOC). CONTAG is a landscape metric that reflects the ratio and degree of dispersion among various land use types within a specific area. The higher the degree of dispersion, the lower the CONTAG value. In general, the correlation between CONTAG and water quality variables is negative, suggesting that water quality of surrounding rivers tends to worsen with highly fragmented and low dispersive landscapes (Nafi'Shehab et al., 2021; Bu et al., 2014; Uuemaa et al., 2005).

Considering the important role of CONTAG in achieving water quality goals in agricultural areas, we investigated the land use types of the study area which do not achieve their TMDL water quality goals. The average CONTAG value of the agricultural watersheds was 52.24, the minimum was 0.00, and the maximum was 92.51. Especially, the Changnyeongcheon watershed, which showed a high CONTAG value of 86.07, did not achieve the TMDL water quality goals for all the water quality parameters of TP, BOD, and TOC. On the other way, the Naeseongcheon 4 waterhed, which showed a low CONTAG value of 35.12, satisfied the water quality goals for all the water quality parameters.

Fig. 8 shows a land use types within a radius of 1.2 km centered on the water quality measurement point of the Changnyeongcheon watershed and the Naeseongcheon 4 waterhed. The land use type used in this study was LEVEL-2 data from the LULC (Land Use and Land Cover) data officially provided by the Korean government. This is a more detailed classification than LEVEL-1, which is classified into 7 levels, including cities or buildings, agriculture, forests, grasslands, wetlands, wastelands, and water bodies, and is divided into a total of 22 types. The Changnyeongcheon watershed with high CONTAG value showed the fragmented land use types compared with the Naeseongcheon 4 waterhed. In the Changnyeongcheon watershed, the boundaries between the land use types are unclear and the land use types with different characteristics are mixed and intermingled. This also resulted in the overall landscape becoming more fragmented and overall connectivity lowered.

We investigated the land use status of the study area, NRW. Most of Korea's agricultural areas are rice paddies and fields. According to the Korea Rural Economic Institute (Korea Rural Economic Institute (KREI), 2015), food crops were mainly grown in rice paddies. In 2014, 82.4 % of all rice paddies were used to grow food crops, of which rice accounted for 94 %. On the other hand, in fields, the order of crops was open-field vegetables (24.8 %), open-field fruit trees (20.8 %), open-field food crops (20.7 %), open-field other (15.0 %), open-field special uses (11.6 %), and facility crops (7.1 %) (Korea Rural Economic Institute (KREI), 2015).

According to previous studies, rice paddies characteristically absorb more nutrients than other land use types and have the ability to retain pollutants, which has the effect of purifying the water in rivers nearby (Cho et al., 2000; Kim et al., 2005). In this way, rice paddies are intensively farmed and have the ability to retain pollutants, so they are expected to have a relatively small effect on the water quality of rivers near paddies. In this study, the Naeseongcheon 4 watershed, which satisfied the TMDL water quality goals for all the water quality parameters, was mainly composed of paddy field, otherwise the Changnyeongcheon watershed, which did not achieve the water quality goals, was composed of protected cultivation referred to the LEVEL-2 data. Compared with paddies, fields practices are less intensive. Around 70 % of fields in South Korea are distributed on slopes. The NRW originates in Taebaek, Gangwon Province, and features numerous sloping areas. Consequently, the need for pollution source management in the upper reaches of the NRW, due to soil loss caused by upland cultivation, has been highlighted consistently (Joo et al., 2007). Therefore, to achieve the TMDL water quality goals in the NRW, it is necessary to particularly manage the outflow and movement of non-point pollution sources around fields.

3.3.2. Forest areas

Generally, areas with a lot of forests have better water quality than

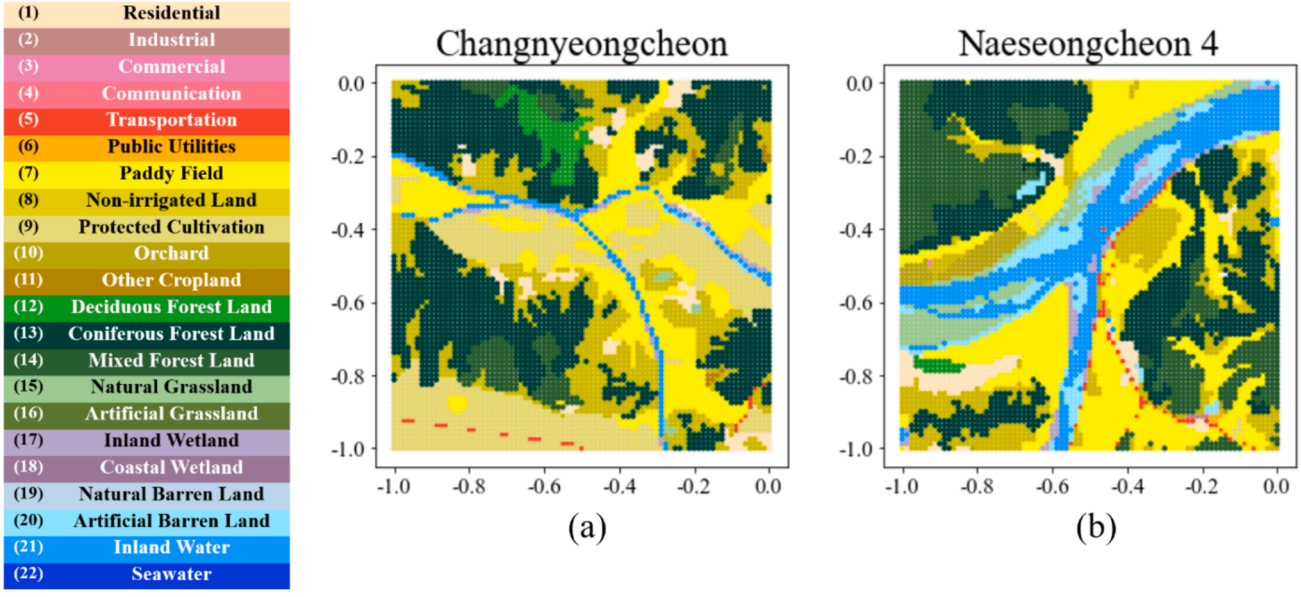


Fig. 8. Comparison of land use types of agricultural areas within a radius of 1.2 km centered on the water quality measurement point: (a) the Changnyeongcheon watershed, (b) the Naeseongcheon 4 waterhed.

watersheds with other land use types, and forest areas play an important role in filtering pollutants and sediments from surface runoff; thus, optimizing the proportion and spatial location of forests area is an important factor in sustainable watershed management (Zhang et al., 2022).

In this study, we found the factors that most influenced the model predictions about whether TMDL water quality goals in forest areas had been achieved. For TP, ED and average temperature had a high influence in the dry seasons, and precipitation and ED had a high influence in the wet seasons (Fig. 9). In the case of BOD, precipitation and CONTAG had a high influence in the dry seasons, and average temperature and LPI had a high influence in the wet seasons (Fig. 9). For TOC, precipitation and COHESION had a high influence in the dry seasons, and SHDI and average temperature had a high influence in the wet seasons (Fig. 9).

We thus confirmed that meteorological data significantly affected all the water quality parameters in both day and wet seasons. Additionally, among the landscape metrics, ED, LPI, CONTAG, COHESION, and SHDI significantly affected the results in forest areas (Fig. 9). Based on the results of SHAP analysis, it was examined the value distribution of the landscape metrics such as ED, LPI, SHDI, CONTAG, COHESION. Table 2 shows the minimum, maximum and average values of each landscape metrics for all watershed and the average value of the landscape metrics for watersheds that did not achieve the TMDL water quality goals.

ED is an indicator of landscape unevenness and fragmentation (Fichera et al., 2012). LPI reflects the largest land cover or use within a particular area (Ding et al., 2016). Similar to LPI, SHDI reflects the

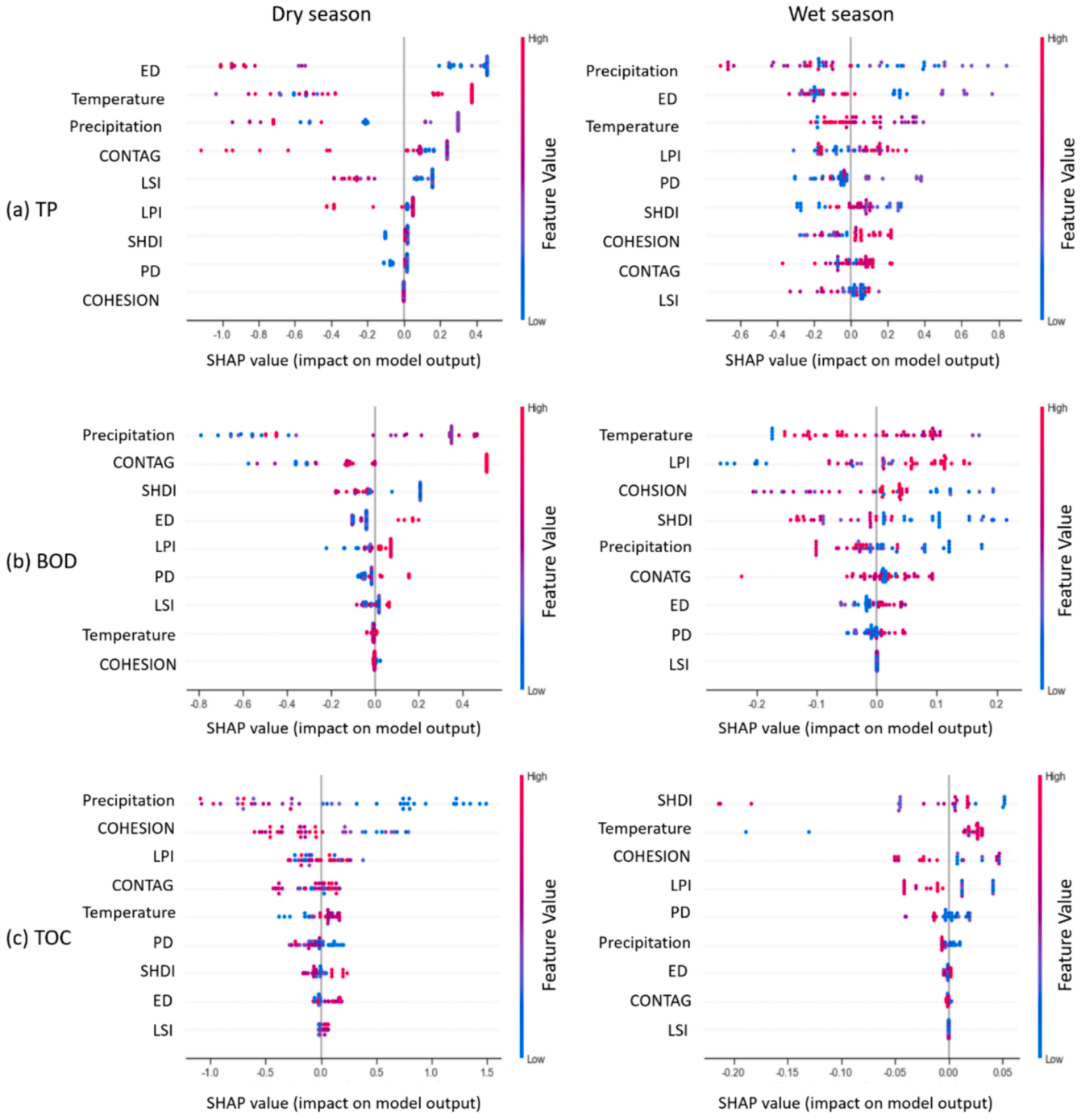


Fig. 9. SHAP summary plots that had the most impact in predicting TMDL water quality goal achievement in forest areas.

Table 2
Minimum, maximum, and average values of landscape metrics.

Values	ED	LPI	SHDI	CONTAG	COHESION
Min	0.00	30.11	0.00	0.00	90.05
Max	213.72	100.00	1.49	99.79	100.00
Average	73.56	72.35	0.58	49.73	96.89
Average of watersheds that do not achieve water quality standards	103.24	66.11	0.78	56.58	95.82

number or proportion of land uses in a region and the proportion of changes in a landscape, measuring the diversity of landscape fragments (Amaral et al., 2019), and it correlates positively with biodiversity and increases with increasing number of different types of land use. (Lee et al., 2009; Concepción et al., 2008). In this study, it was confirmed that the higher the ED and SHDI, the worse the water quality. In addition, the lower the LPI, the worse the water quality. This is consistent with previous research showing that water quality is likely to deteriorate as different land use types become largely dispersed and the number of land use types increases (Lee et al., 2009).

CONTAG percentages reflect the aggregated proportion or variance of land use patterns present in a certain area, which tends to be high when land use variance is low (Bu et al., 2014). COHESION reflects the physical and structural connectivity and connectedness of a particular land cover or use within an area and consequently the effect of landscape spread or dispersion on the variables of interest (Ding et al., 2016). In the case of forest areas, it was confirmed that CONTAG was higher in the watersheds that did not achieve the TMDL water quality goals, similar to the results of the agricultural areas. However, for the case of COHESION, there no clear difference between the watersheds that did not achieve the water quality goals and the entire watersheds.

Fig. 10 shows a land use patterns within a radius of 1.2 km centered on the water quality measurement point of the Sacheoncheon watershed and the Yonggyecheon waterhed. In the forest area, there were no watershed that did not achieve the TMDL water quality goals for all 3 water quality parameters (TP, BOD, and TOC). Only 4 watersheds did not achieve the TMDL water quality goals for 2 water quality parameters. Among them, Sacheoncheon composed of 14 land use types did not achieve the BOD and TOC goals. On the other hand, Yonggyecheon, satisfied all water quality standards and had 11 land use types, 3 land use types smaller than those of the Sacheoncheon. Through this, it was confirmed that different land uses were further mixed and that areas where the connectivity of each land use type was relatively low did not achieve the water quality standard.

We examined land use changes from the 1980 s to the 1990 s, a period of rapid industrialization and urbanization in Korea. We found that the forest area, which accounted for 73.2 % of the total area of the NRW in the 1980 s, decreased to 67 % of the total area in the 1990 s, a decrease of about 1500 km^2. In addition, as urbanization progressed, the impervious layer area increased by approximately 2.5 times, from 1.6 % to 4.25 % of the total area (Park et al., 2005).

The complete conversion of forest patches to other land use types can cause forest fragmentation and a massive loss of edge density in forest patches, leading to a decrease in LPI and ED (Mengist et al., 2022). Our results in this study show that LPI and ED were factors with a strong effect on the achievement of water quality targets. Also, fragmentation negatively affects the diversity of species living within a forest because it changes the intensity of wind and light flowing in from outside the forest (Zebisch et al., 2004). Thus, SHDI was also a factor with a strong effect on the achievement of water quality goals, and COHESION and CONTAG, which represent forest connectivity and aggregated proportion, also showed the same results.

In particular, LPI had a significant effect in predicting the achievement of the BOD water quality target. Because BOD is sensitive to the size, edge area, and shape complexity of a forest, when simpler forest areas within the watershed were divided into smaller pieces, creating more edge areas, high BOD concentrations became more likely (Lee et al., 2009). These results suggest that large, unfragmented forests in the watershed could reduce BOD concentrations.

Our results in this study show that both landscape metrics and meteorological factors play an important role in predicting the achievement of water quality targets in forest areas. Among the landscape metrics, ED, LPI, CONTAG, COHESION, and SHDI, excluding LSI and PD, had a great influence. Therefore, considering the effects of meteorological factors caused by climate change, preserving forest areas as large patches without fragmentation will positively affect river water quality (Nafi'Shehab et al., 2021).

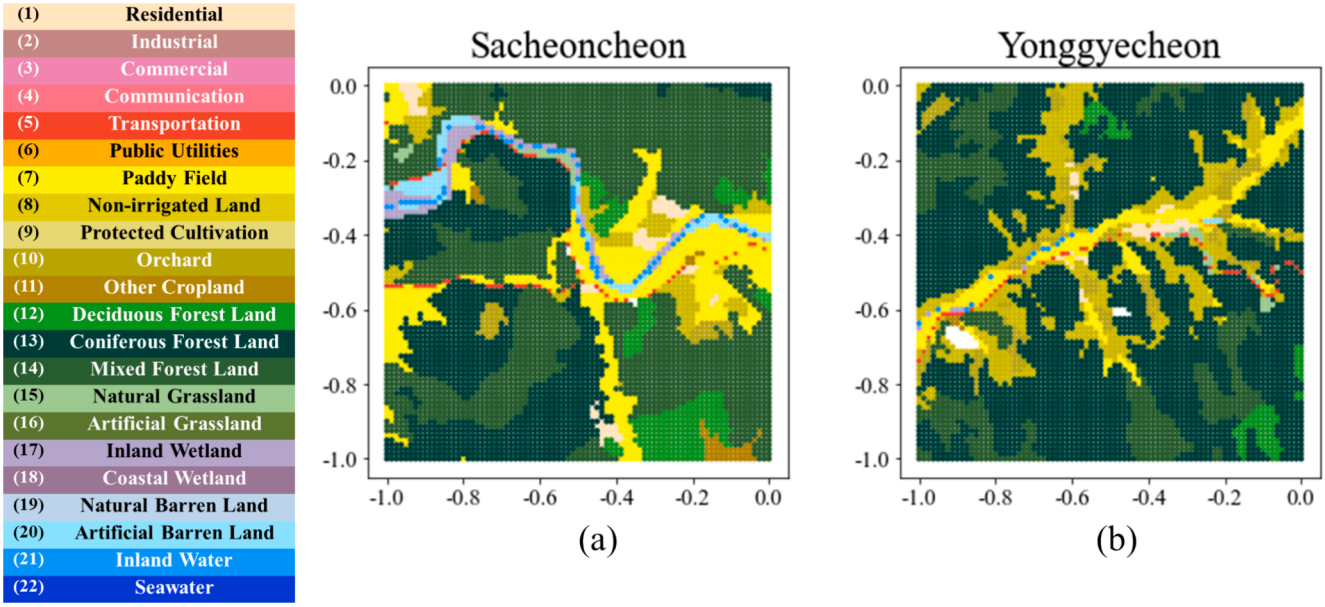


Fig. 10. Comparison of land use types of forest areas within a radius of 1.2 km centered on the water quality measurement point: (a) the Sacheoncheon watershed, (b) the Yonggyecheon watershed.

3.3.3. Urban areas

In urban areas during the wet season, precipitation ranked first and second among the main factors influencing the achievement of the TP and BOD targets, respectively. For TOC, precipitation was the third factor (Fig. 11 (a), Fig. 11 (b)). In urban areas, the loss of natural vegetation reduces the soil's ability to absorb and filter pollutants during heavy summer rainfall. Therefore, significant amounts of pollutants are discharged rapidly from urban systems into rivers (Huang et al., 2016). According to previous studies, higher runoff coefficient values generally indicate that more pollutants are being discharged from non-point pollution sources. In urban areas, the proportion of impervious surfaces is particularly high, so the runoff coefficient is high, and a significant portion of precipitation flows directly into rivers (Hwang, 2009).

The SHAP analysis of TP in urban areas showed that PD and ED were the secondary factors (Fig. 11 (a)). In general, TP concentrations are sensitive to the size and edge areas of specific land use types within a watershed, and a previous study of Korean reservoir watersheds showed that PD and ED correlated with TP in urban areas (Lee et al., 2009). PD measures the number of land use patches within a watershed, but it does not provide information about the scale or spatial distribution of land use (McGarigal and Marks, 1995). ED is a metric of landscape unevenness and fragmentation (Fichera et al., 2012). Both PD and ED are related to the number of land use patches, so many fragmented urbanized areas could reduce the quality of nearby water (Lee et al., 2009). COHESION was the secondary highly influenced factor in wet season's BOD and the first factor in dry season's TOC. COHESION reflects the

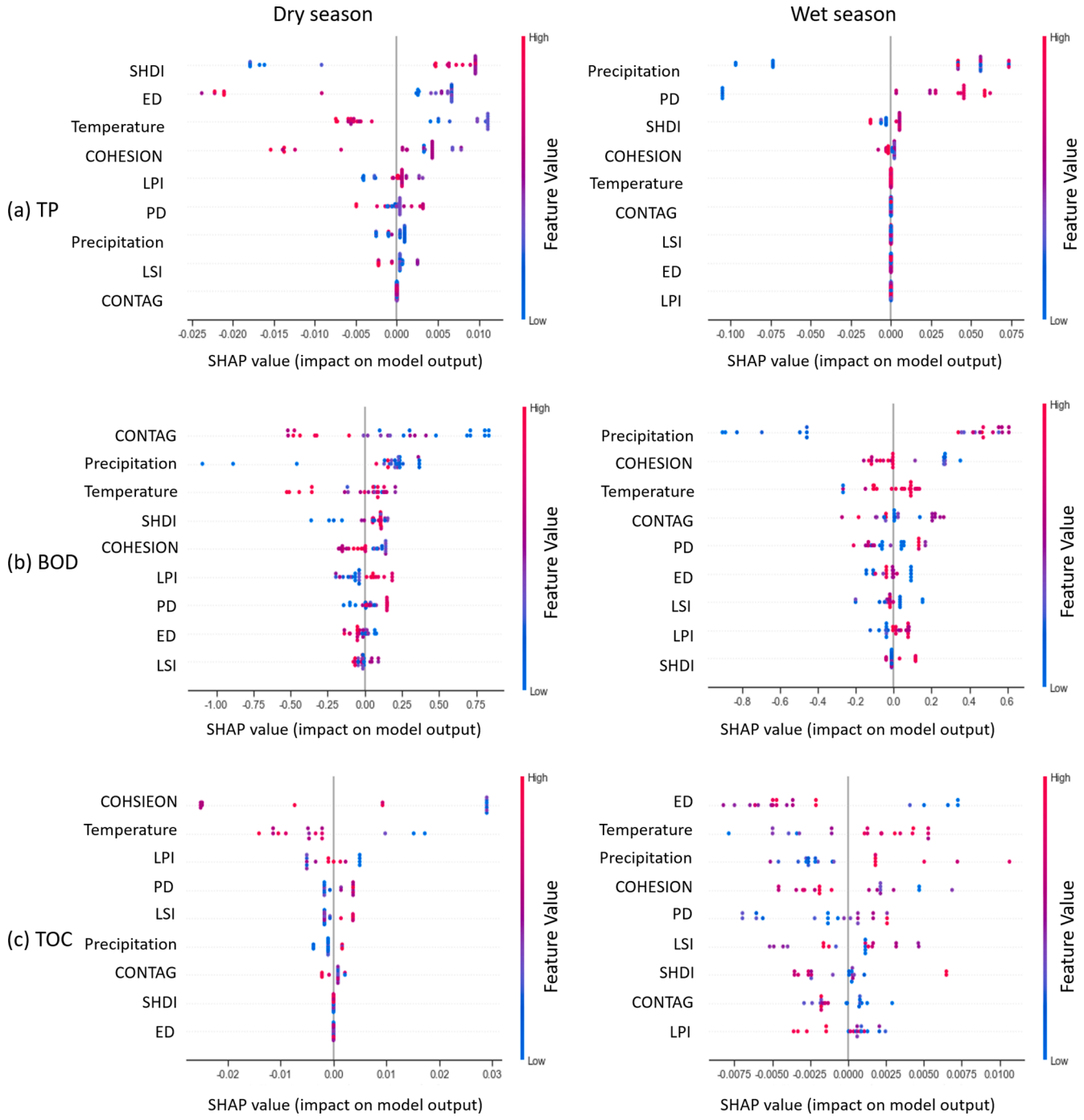


Fig. 11. SHAP summary plots that had the most impact in predicting TMDL water quality goal achievement in urban areas.

physical and structural connectivity of a specific land use distribution and tends to be negatively correlated with river water quality (Nafi'Shehab et al., 2021). In addition, the results of the SHAP analysis showed that COHESION was negatively correlated with TOC and BOD. However, from the result of this study, there was no clear correlation between landscape metrics and water quality in urban area like agricultural area and forest area. By comparing the Geumho River 8 watershed, which did not achieve the water quality goals, and Geumho River 5 watershed, which did achieve the water quality goals, it was confirmed that Geumho River 8 had more developed roads and a dense industrial complex around the river (Fig. 12). On the other hand, Geumho River 5, which is mainly populated with residential areas and has no developed roads around the river, satisfied the TMDL water quality goals.

Therefore, our analysis of factors influencing water quality goals in urban areas highlights that there is an urgent need to manage the surface runoff of pollutants from impervious surfaces during rainfall. Additionally, human land use within watersheds typically results in the fragmentation of natural areas, which has numerous negative ecological and environmental impacts (Collinge, 1996). As such, urbanization can have both direct (e.g., creating point pollution sources and nonpoint pollution sources) and indirect impacts by degrading the structure of natural areas that are important for filtering pollutants (Lee et al., 2009). Therefore, improving physical connectivity to land uses in urban areas within dispersed watersheds is essential to achieving water quality goals with the TMDL management system.

4. Conclusion

To effectively operate under the current fourth phase TMDL management system, we built an XGBoost model to predict whether the TMDL water quality goals for the entire NRW in Korea are being achieved. The factors that have the most significant impact on the achievement of water quality goals were analyzed using SHAP to find the factors that most influenced the predictions. The water resources of the NRW, the study area, play an important role as drinking water, domestic water, and agricultural water in the Gyeongsang-do region. Accordingly, the NRW has implemented the TMDLs since 2004, and water quality has been steadily managed. As a result, point pollution sources are decreasing greatly due to the installation of basic environmental facilities, and now nonpoint pollution sources account for 70 % of all pollution, making it increasingly difficult to achieve water quality goals. Nonpoint pollution sources are affected by topography, land use, climate, etc., and are generated from unspecified activities, making it difficult to identify outflow paths and making pollutants discontinuous and difficult to manage. In addition, climate change is altering the outflow characteristics of the watershed by affecting rainfall patterns, such as an increase in the frequency of heavy rainfall in summer, making it difficult to manage river basins. Therefore, we developed an XGBoost model to predict the achievement of TP, BOD, and TOC goals and confirmed that its prediction accuracy was relatively high (71.9 %). We then used SHAP to derive the major factors affecting the model predictions. We found that meteorological factors, particularly the monthly precipitation and average temperature, highly influenced the model predictions for all land use types. In the landscape metrics, ED showed a high level of importance in all land use types. CONTAG was the main factor in agricultural area; ED, LPI, CONTAG, COHESION, and SHDI were the main factors in forest area; and PD, ED, SHDI, and COHESION were the main factors in urban area. Through this study, we derived relationships that allow us to interpret and explain how various factors, including spatial patterns, rainfall timing, and landscape patterns, affect water quality variables. Understanding these relationships provides valuable data about water quality conditions and will help us outline appropriate water management practices. Therefore, our results provide valuable ideas for land use managers and landscape planners to achieve water quality goals through the management of non-point source pollution.

CRediT authorship contribution statement

Sun Hee Shim: Conceptualization, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jung Hyun Choi:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jung Hyun Choi reports financial support was provided by National

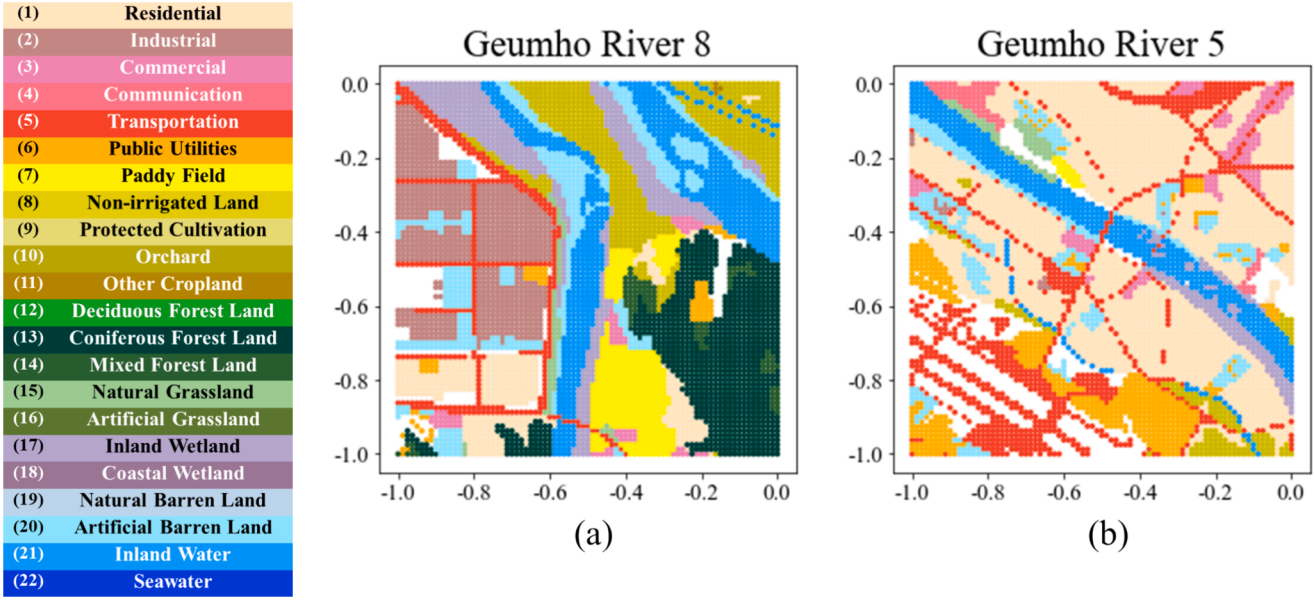


Fig. 12. Comparison of land use patterns of urban areas within a radius of 1.2 km centered on the water quality measurement point: (a) the Geumho River 8 watershed, (b) the Geumho River 5 watershed.

Research Foundation of Korea. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2018R1A6A1A08025520)

References

Agency, J., 2012. (The second) Non-point source pollution management comprehensive measures. Joint Agency, Seoul.

Ahmed, S., Lee, Y., Hyun, S.H., Koo, I., 2019. Unsupervised machine learning-based detection of covert data integrity assault in smart grid networks utilizing isolation forest. IEEE Trans. Inf. Forensics Secur. 14 (10), 2765–2777.

Amaral, Y.T., dos Santos, E.M., Ribeiro, M.C., Barreto, L., 2019. Landscape structural analysis of the Lençóis Maranhenses national park: Implications for conservation. J. Nat. Conserv. 51, 125725 https://doi.org/10.1016/j.jnc.2019.125725.

Baker, W.L., Cai, Y., 1992. The r. le programs for multiscale analysis of landscape structure using the GRASS geographical information system. Landsc. Ecol. 7, 291–302. https://doi.org/10.1007/BF00131258.

Bartel, A., 2000. Analysis of landscape pattern: towards a 'top down'indicator for evaluation of landuse. Ecol. Model. 130 (1–3), 87–94. https://doi.org/10.1016/S0304-3800(00)00214-3.

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

Chen, T., Guestrin, C., 2016. August). Xgboost: A scalable tree boosting system. In: In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Chen, W.R., Yun, Y.H., Wen, M., Lu, H.M., Zhang, Z.M., Liang, Y.Z., 2016. Representative subset selection and outlier detection via isolation forest. Anal. Methods 8 (39), 7225–7231. https://doi.org/10.1039/c6ay01574c.

Cho, J.Y., Han, K.W., Choi, J.K., Goo, J.W., Son, J.G., 2000. Runoff loading on nutrients from a paddy field during non-cropping season. Journal of the Korean Society of Agricultural Engineers 42 (2), 63–70.

Cho, H.K., Kim, S.M., 2018. Water quality correlation analysis between sewage treated water and the adjacent downstream water in Nakdong River basin. Journal of Korean Society on Water Environment 34 (2), 202–209. https://doi.org/10.15681/KSWE.2018.34.2.202.

Cho, H.K., Kwak, E.T., Kim, S.M., 2020. A study on the spatial variation of target water quality and excess rate at 41 stations in nakdong river basin after the total maximum daily loads. Journal of the Korean Society of Agricultural Engineers 62 (6), 97–109. https://doi.org/10.5389/KSAE.2020.62.6.097.

Chung, D.H., Yun, J.S., Yang, S.M., 2021. Machine learning for predicting entrepreneurial innovativeness. Asia-Pacific Journal of Business Venturing and Entrepreneurship 16 (3), 73–86.

Collinge, S.K., 1996. Ecological consequences of habitat fragmentation: implications for landscape architecture and planning. Landsc. Urban Plan. 36 (1), 59–77. https://doi.org/10.1016/S0169-2046(96)00341-6.

Concepción, E.D., Díaz, M., Baquero, R.A., 2008. Effects of landscape complexity on the ecological effectiveness of agri-environment schemes. Landsc. Ecol. 23, 135–148. https://doi.org/10.1007/s10980-007-9150-2.

de Oliveira, L.M., Maillard, P., de Andrade Pinto, É.J., 2016. Modeling the effect of land use/land cover on nitrogen, phosphorous and dissolved oxygen loads in the Velhas River using the concept of exclusive contribution area. Environ. Monit. Assess. 188, 1–19. https://doi.org/10.1007/s10661-016-5323-2.

Ding, J., Jiang, Y., Liu, Q., Hou, Z., Liao, J., Fu, L., Peng, Q., 2016. Influences of the land use pattern on water quality in low-order streams of the Dongjiang River basin, China: a multi-scale analysis. Sci. Total Environ. 551, 205–216. https://doi.org/10.1007/s10661-016-5323-2.

Ding, S., Zhang, Y., Liu, B., Kong, W., Meng, W., 2013. Effects of riparian land use on water quality and fish communities in the headwater stream of the Taizi River in China. Front. Environ. Sci. Eng. 7, 699–708. https://doi.org/10.1007/s11783-013-0528-x.

Fichera, C.R., Modica, G., Pollino, M., 2012. Land Cover classification and change-detection analysis using multi-temporal remote sensed imagery and landscape metrics. European Journal of Remote Sensing 45 (1), 1–18. https://doi.org/10.5721/EuJRS20124501.

Fowler, H.J., Lenderink, G., Prein, A.F., Westra, S., Allan, R.P., Ban, N., Zhang, X., 2021. Anthropogenic intensification of short-duration rainfall extremes. Nature Reviews Earth & Environment 2 (2), 107–122.

Gosain, A., Sardana, S. (2017, September). Handling class imbalance problem using oversampling techniques: A review. In 2017 international conference on advances in computing, communications and informatics (ICACCI) (pp. 79-85). IEEE. doi: 10.5194/hess-24-827-2020.

Guo, D., Lintern, A., Webb, J.A., Ryu, D., Bende-Michl, U., Liu, S., Western, A.W., 2020. A data-based predictive model for spatiotemporal variability in stream water quality. Hydrol. Earth Syst. Sci. 24 (2), 827–847. https://doi.org/10.5194/hess-24-827-2020.

Hashemi, F., Olesen, J.E., Dalgaard, T., Børgesen, C.D., 2016. Review of scenario analyses to reduce agricultural nitrogen and phosphorus loading to the aquatic environment. Sci. Total Environ. 573, 608–626.

Heo, S.G., Kim, K.S., Ahn, J.H., Yoon, J.S., Lim, K.J., Choi, J.D., Lyou, C.W., 2007. Landscape analysis of the forest fragmentations at Doam-Dam Watershed using the FRAGSTATS Model. Journal of the Korean Association of Geographic Information Studies 10 (1), 10–21.

Horvat, T., Job, J., 2020. The use of machine learning in sport outcome prediction: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (5), e1380.

Huang, Z., Han, L., Zeng, L., Xiao, W., Tian, Y., 2016. Effects of land use patterns on stream water quality: a case study of a small-scale watershed in the Three Gorges Reservoir Area, China. Environ. Sci. Pollut. Res. 23, 3943–3955. https://doi.org/10.1007/s11356-015-5874-8.

Hwang, B.G., 2009. Estimating unit load of non-point source pollutants for landuse types in Anseongchun Watershed. Journal of the Korea Academia-Industrial Cooperation Society 10 (5), 1026–1033. https://doi.org/10.5762/KAIS.2009.10.5.1026.

Hwang, C.G., Yoon, C.P., Kim, D.J., 2022. Indoor positioning method using WiFi signal based on Xgboost. Journal of the Korea Institute of Information and Communication Engineering 26 (1), 70–75.

Joo, J.H., Yang, J.E., Ok, Y.S., Oh, S.E., Yoo, K.Y., Yang, S.C., Jung, Y.S., 2007. Assessment of pollutant loads from alpine agricultural practices in Nakdong river basin. Korean Journal of Environmental Agriculture 26 (3), 233–238. https://doi.org/10.5338/KJEA.2007.26.3.233.

Jung, S., Eum, J., Jang, C., Choi, Y., Kim, B., 2012. Characteristics of nonpoint source pollution and relationship between land use and nutrient concentrations in the Han River watershed. Journal of Korean Society on Water Environment 28 (2), 255–268.

Jung, W., Lee, H., 2022. Data augmentation using a kernel density estimation for motion recognition applications. Journal of Korea Society of Industrial Information Systems 27 (4), 19–27. https://doi.org/10.5194/amt-13-4669-2020.

Just, A.C., Liu, Y., Sorek-Hamer, M., Rush, J., Dorman, M., Chatfield, R., Kloog, I., 2020. Gradient boosting machine learning to improve satellite-derived column water vapor measurement error. Atmos. Meas. Tech. 13 (9), 4669–4681. https://doi.org/10.5194/amt-13-4669-2020.

Kaim, A., Cord, A.F., Volk, M., 2018. A review of multi-criteria optimization techniques for agricultural land use allocation. Environ. Model. Softw. 105, 79–93.

Kang, M.A., Choi, B.W., Yu, J.J., 2010. Loading characteristics of non-point source pollutants by rainfall-case study with cherry tree plot. The Journal of Engineering Geology 20 (4), 401–407.

Kang, J.E., Choi, H.S., Hwang, H.S., Lee, S., 2018. Analysis of ecological network according to invalidation of decision on urban parks: focused on Busan. Journal of Environmental Impact Assessment 27 (6), 618–634. https://doi.org/10.14249/eia.2018.27.6.618.

Kang, D.H., So, Y.H., Park, K.D., Kim, I.K., Kim, B.W., 2019. Analyses of TOC efficiency and correlation between DO, BOD, COD and influence factors using long-term observation data in the main stream of Nakdong River. Journal of Environmental Science International 28 (5), 465–474. https://doi.org/10.5322/JESI.2019.28.5.465.

Ki, S.J., Park, H.G., Ahn, S.J., Yu, D.G., Lee, C.S., 2019. Assessing water quality characteristics and pollution loads of agricultural pumping and drainage facilities as well as drainage facilities at the lower Nakdong river. Jourral of Korean Society of Water Science and Technology 27 (6), 67–79.

Kim, S.H., Baek, S.Y., Jung, N.R., Hyun, J.M., 2022. Cost calculation of the implementation project for the management of total maximum daily loads. Ecology and Resilient Infrastructure 9 (3), 174–182.

Kim, S., Chung, S., 2022. Causal impact analysis of enhanced phosphorus effluent standard on river water quality. J. Environ. Manage. 320, 115931 https://doi.org/10.1016/j.jenvman.2022.115931.

Kim, J.H., Kim, C.Y., Lee, S.T., Choi, C.M., Jung, G.B., Lee, J.S., Kim, W.I., 2007. Properties of water quality and land use at the rural area in the Nakdong river watershed. Korean Journal of Environmental Agriculture 26 (2), 99–106.

Kim, S.G., Oh, S.Y., Park, S.Y., Na, E.H., Kim, Y.S., 2023. Improvement and Implementation to Enhance the Effectiveness of the Total Pollution Load Control System. Journal of Korean Society on Water Quality 39 (4), 343–355.

Kim, M.K., Roh, K.A., Lee, N.J., Seo, M.C., Koh, M.H., 2005. Nutrient load balance in large-scale paddy fields during rice cultivation. Korean J. Soil Sci. Fert. 38 (3), 164–171.

Kim, Y.I., Yi, S.J., 2011. Improvement of procedures for reasonable implementation of TMDL. Journal of Korean Society of Environmental Engineers 33 (8), 617–622. https://doi.org/10.4491/KSEE.2011.33.8.617.

Korea Institute of Civil engineering and building Technology (KICT). (2000). The Optimum Management Plans and Runoff Characteristics of Nonpoint Source for PalDang Watershed.

Korea Rural Economic Institute (KREI). (2015). Analysis of Field Agricultural Production Structure and Productivity Indicators. 1-85.

Kwon, H., Jo, C., Choi, S., 2023. Diffuse pollutant load predictions in areas that implement the total maximum daily load due to climate change. Environ. Technol. Innov. 103251 https://doi.org/10.1016/j.eti.2023.103251.

Kwon, H.H., Kim, B.S., Kim, B.K., 2008. Analysis of precipitation characteristics of regional climate model for climate change impacts on water resources. KSCE Journal of Civil and Environmental Engineering Research 28 (5B), 525–533. https://doi.org/10.12652/Ksce.2008.28.5B.525.

Kwon, H., Lee, J., Lim, J., Woo, S., Kim, J., Lim, K., Hong, E., 2020. Characteristics of spatial variability in water quality on stream of lake doam watershed. Journal of the Korean Society of Agricultural Engineers 62 (6), 43–50. https://doi.org/10.5389/KSAE.2020.62.6.043.

Lebowitz, P.J., 2001. Land use, land abuse and land re-use: a framework for the implementation of TMDLs for nonpoint source polluted waterbodies. Pace Envtl. L. Rev. 19, 97.

Lee, S., 2012. Comparison of initial seeds methods for K-means clustering. Journal of Internet Computing and Services 13 (6), 1–8. https://doi.org/10.7472/jksii.2012.13.6.1.

Lee, D.J., 2014. A study on pattern making of degradation type using K-means. Journal of the Korean Institute of Electrical and Electronic Material Engineers 27 (12), 877–882. https://doi.org/10.4313/JKEM.2014.27.12.877.

Lee, A., Cho, S., Park, M.J., Kim, S., 2013. Determination of standard target water quality in the Nakdong River basin for the total maximum daily load management system in Korea. KSCE J. Civ. Eng. 17, 309–319. https://doi.org/10.1007/s12205-013-1893-5.

Lee, S.W., Hwang, S.J., Lee, S.B., Hwang, H.S., Sung, H.C., 2009. Landscape ecological approach to the relationships of land use patterns in watersheds to water quality characteristics. Landsc. Urban Plan. 92 (2), 80–89. https://doi.org/10.1016/j.landurbplan.2009.02.008.

Lee, Y., Sun, J., 2020. Predicting highway concrete pavement damage using XGBoost. Korean Journal of Construction Engineering and Management 21 (6), 46–55. https://doi.org/10.6106/KJCEM.2020.21.6.046.

Li, R., Zheng, H., Huang, B., Xu, H., Li, Y., 2018. Dynamic impacts of climate and land-use changes on surface runoff in the mountainous region of the Haihe River Basin China. Adv. Meteorol. 2018 https://doi.org/10.1155/2018/3287343.

Liu, F.T., Ting, K.M., Zhou, Z.H., 2012. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) 6 (1), 1–39.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Proces. Syst. 30 https://doi.org/10.48550/arXiv.1705.07874.

McGarigal, K., Marks, B.J. (1995). Spatial pattern analysis program for quantifying landscape structure. Gen. Tech. Rep. PNW-GTR-351. US Department of Agriculture, Forest Service, Pacific Northwest Research Station, 1-122. doi: 10.2737/PNW-GTR-351.

ME. (2022, October 17). Introduction of Total Organic Carbon Management in the Nakdong River Water System. Republic of Korea Policy Briefing. Retrieved from https://www.korea.kr/briefing/pressReleaseView.do?newsId=156531070.

Mengist, W., Soromessa, T., Feyisa, G.L., 2022. Forest fragmentation in a forest biosphere reserve: implications for the sustainability of natural habitats and forest management policy in Ethiopia. Resources, Environment and Sustainability 8, 100058. https://doi.org/10.1016/j.resenv.2022.100058.

Nafi'Shehab, Z., Jamil, N.R., Aris, A.Z., Shafie, N.S., 2021. Spatial variation impact of landscape patterns and land use on water quality across an urbanized watershed in Bentong Malaysia. Ecol. Indicators 122, 107254. https://doi.org/10.1016/j.ecolind.2020.107254.

Namugize, J.N., Jewitt, G., Graham, M., 2018. Effects of land use and land cover changes on water quality in the uMngeni river catchment, South Africa. Physics and Chemistry of the Earth, Parts A/b/c 105, 247–264. https://doi.org/10.1016/j.pce.2018.03.013.

Oh, J.Y., Ham, D.H., Lee, Y.G., Kim, G., 2019. Short-term load forecasting using XGBoost and the analysis of hyperparameters. Trans. Korean Inst. Electr. Eng 68, 1073–1078. https://doi.org/10.5370/KIEE.2019.68.9.1073.

Ouyang, W., Skidmore, A.K., Toxopeus, A.G., Hao, F., 2010. Long-term vegetation landscape pattern with non-point source nutrient pollution in upper stream of Yellow River basin. J. Hydrol. 389 (3–4), 373–380. https://doi.org/10.1016/j.jhydrol.2010.06.020.

Pan, T., Su, F., Yan, F., Lyne, V., Wang, Z., Xu, L., 2023. Optimization of multi-objective multi-functional landuse zoning using a vector-based genetic algorithm. Cities 137, 104256.

Park, K.H., Jung, S.G., Kwon, J.O., Oh, J.H., 2005. Patterns of forest landscape structure due to landcover change in the Nakdong river basin. Journal of Korean Society of Rural Planning 11 (4), 47–57.

Park, S.J., Soung, D.H., Byun, Y.C., 2020. A hybrid collaborative filtering based on online shopping patterns using XGBoost and Word2Vec. Journal of Advanced Information Technology and Convergence 18 (9), 1–8. https://doi.org/10.14801/jkiit.2020.18.9.1.

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 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.

Rahman, M.M., Szabó, G., 2021. Multi-objective urban land use optimization using spatial data: a systematic review. Sustain. Cities Soc. 74, 103214.

Robinson, R.B., Cox, C.D., Odom, K., 2005. Identifying outliers in correlated water quality data. J. Environ. Eng. 131 (4), 651–657. https://doi.org/10.1061/(ASCE)0733-9372(2005)131:4(651).

Shim, K., Kim, G., Im, T., Kim, Y., Kim, S., 2021. A study on the water quality improvement of major tributaries in Seoul, applying watershed evaluation techniques. Journal of Korean Society on Water Environment 37 (1), 32–46.

Shon, T., Lee, S., Kim, S., Shin, H., 2010. An analysis of the effect of climate change on flow in Nakdong river basin using watershed-based model. Journal of Korea Water Resources Association 43 (10), 865–881. https://doi.org/10.3741/JKWRA.2010.43.10.865.

Sim, D.C., Lee, J.Y., Jang, J., Lee, M. (2022). Prediction of chloride concentration in groundwater on Jeju Island using XGBoost regression machine learning. Journal of the Geological Society of Korea 58(2), 243-255. 10.14770/jgsk.2022.58.2.243.

Song, X., Aryal, S., Ting, K.M., Liu, Z., He, B., 2021. Spectral–spatial anomaly detection of hyperspectral data based on improved isolation forest. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41, 647–665. https://doi.org/10.1007/s10115-013-0679-x.

Uuemaa, E., Roosaare, J., Mander, Ü., 2005. Scale dependence of landscape metrics and their indicatory value for nutrient and organic matter losses from catchments. Ecol. Ind. 5 (4), 350–369.

Walker, T.R., Crittenden, P.D., Dauvalter, V.A., Jones, V., Kuhry, P., Loskutova, O., Young, S.D., 2009. Multiple indicators of human impacts on the environment in the Pechora Basin, north-eastern European Russia. Ecol. Ind. 9 (4), 765–779. https://doi.org/10.1016/j.ecolind.2008.09.008.

Wang, J., Cao, J., Liu, Z., 2024. Unsupervised machine learning-based multi-attributes fusion dim spot subtle sandstone reservoirs identification utilizing isolation forest. Geoenergy Science and Engineering 234, 212626.

Wu, J., Lu, J., 2021. Spatial scale effects of landscape metrics on stream water quality and their seasonal changes. Water Res. 191, 116811.

Wu, T., Wang, S., Su, B., Wu, H., Wang, G., 2021. Understanding the water quality change of the Yilong Lake based on comprehensive assessment methods. Ecol. Ind. 126, 107714 https://doi.org/10.1016/j.ecolind.2021.107714.

Xu, J., Bai, Y., You, H., Wang, X., Ma, Z., Zhang, H., 2022b. Water quality assessment and the influence of landscape metrics at multiple scales in Poyang Lake basin. Ecol. Ind. 141, 109096 https://doi.org/10.1016/j.ecolind.2022.109096.

Xu, T., Coco, G., Neale, M., 2020. A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. Water Res. 177, 115788 https://doi.org/10.1016/j.watres.2020.115788.

Xu, G., Fan, H., Oliver, D.M., Dai, Y., Li, H., Shi, Y., Zhao, Z., 2022a. Decoding river pollution trends and their landscape determinants in an ecologically fragile karst basin using a machine learning model. Environ. Res. 214, 113843 https://doi.org/10.1016/j.envres.2022.113843.

Xu, Q., Yan, T., Wang, C., Hua, L., Zhai, L., 2023. Managing landscape patterns at the riparian zone and sub-basin scale is equally important for water quality protection. Water Res. 229, 119280 https://doi.org/10.1016/j.watres.2022.119280.

Yao, S., Chen, C., He, M., Cui, Z., Mo, K., Pang, R., Chen, Q., 2023. Land use as an important indicator for water quality prediction in a region under rapid urbanization. Ecol. Ind. 146, 109768 https://doi.org/10.1016/j.ecolind.2022.109768.

Yu, J., Kim, Y., Sung, H.C., Lee, K.I., Choi, J.Y., Jeon, S.W., 2020. Method for calculating the pollution load amount of agricultural non-point sources using land cover map. Journal of Environmental Science International 29 (12), 1249–1260.

Zebisch, M., Wechsung, F., Kenneweg, H., 2004. Landscape response functions for biodiversity assessing the impact of land-use changes at the county level. Landsc. Urban Plan. 67 (1–4), 157–172. https://doi.org/10.1016/S0169-2046(03)00036-7.

Zhang, F., Chen, Y., Wang, W., Jim, C.Y., Zhang, Z., Tan, M.L., Rahman, H.A., 2022. Impact of land-use/land-cover and landscape pattern on seasonal in-stream water quality in small watersheds. J. Clean. Prod. 357, 131907 https://doi.org/10.1016/j.jclepro.2022.131907.

Zhang, J., Li, S., Jiang, C., 2020. Effects of land use on water quality in a River Basin (Daning) of the Three Gorges Reservoir Area, China: watershed versus riparian zone. Ecol. Ind. 113, 106226 https://doi.org/10.1016/j.ecolind.2020.106226.

Zhou, J., Liu, X., Liu, X., Wang, W., Wang, L., 2023. Assessing agricultural non-point source pollution loads in typical basins of upper Yellow River by incorporating critical impacting factors. Process Saf. Environ. Prot. 177, 17–28.

Zou, Z., Xie, Y., Huang, K., Xu, G., Feng, D., Long, D., 2019. A docker container anomaly monitoring system based on optimized isolation forest. IEEE Trans. Cloud Comput. 10 (1), 134–145.