https://doi.org/10.5194/hess-2024-355
Preprint. Discussion started: 21 January 2025

Hydrology and Earth System Sciences
Discussions

Advancing flow duration curve prediction in ungauged basins using machine learning and deep learning

Sooyeon Yi[1], Jeongin Yoon[2], Chulhee Lee[2], Seonmi Lee[2], Jungwon Ji[2], Eunkyung Lee[2], Jaeeung Yi[2]

[1] Department of Environmental Science, Policy, and Management, University of California, Berkeley, 94720, USA,
[2] Department of Civil Systems Engineering, Ajou University, 206 Worldcup-ro Yeongtong-gu, Suwon, 16499, South Korea

Correspondence to: Jaeeung Yi (jeyi@ajou.ac.kr)

Abstract. The flow duration curve (FDC) represents the distribution of streamflow, providing vital information for managing river systems. Constructing FDC is especially challenging in ungauged basins where streamflow data are lacking. This study addresses key gaps by utilizing machine learning and deep learning models to predict FDC in ungauged basins. The objectives include: (a) identifying influential hydrologic, meteorological, and topographic factors, (b) evaluating various combinations of predictor variables, (c) assessing the effects of different precipitation metrics on flow predictions, and (d) comparing ML and DL model performance. We developed and evaluated random forest (RF), deep neural network (DNN), support vector regression (SVR), and elastic net regression (ENR) models using historical data from 140 streamflow stations. Feature importance analysis revealed that watershed area and precipitation were the key factors for high discharge percentiles, whereas land use and basin characteristics gained greater importance for medium and low flows. Scenario analysis showed that combining all variables yielded the highest accuracy in predicting FDC. Different precipitation metrics had minimal impact on streamflow predictions, indicating that other factors played a more significant role. The DNN outperformed RF, SVR, and ENR in predicting low (Q_{95}), medium (Q_{50}), and high flows (Q_5), achieving an average coefficient of determination that was 8.03% higher, a root mean square error that was 227.4% lower on average, and a standard deviation that was 46.4% lower. This study demonstrates the effectiveness of advanced ML and DL approaches for predicting FDC in ungauged basins, offering a foundation for advancing hydrological prediction.

1 Introduction

Accurate prediction of streamflow in ungauged basins remains a critical challenge in hydrology, essential for effective water resource management, ecosystem protection, and sustainable development (Booker & Snelder, 2012; Castellarin et al., 2004). One widely used approach to tackle this challenge is constructing flow duration curves (FDC), which are instrumental in various water-related applications, such as hydropower generation, irrigation system design, stream-pollution management, river and reservoir sedimentation control, and fluvial erosion (Yi & Yi, 2024). However, constructing FDC in ungauged or poorly gauged basins presents significant challenges due to the lack of reliable streamflow data. In many regions globally, discharge has not been measured regularly or accurately, leading to the designation of such areas as ungauged basins. The scarcity of streamflow data is a well-recognized issue, as demonstrated by numerous studies across diverse geographic regions, such as Canada (LeBoutillier & Waylen, 1993), China (Ma et al., 2024), Greece (Mimikou & Kaemaki, 1985), Mexico (Arsenault et al., 2019), Korea (Won et al., 2023), USA (Ridolfi et al., 2020). While gauged basins allow for empirical derivation of FDC using long-term flow records, ungauged basins require alternative approaches to estimate streamflow. Recognizing the importance of predicting FDC at ungauged sites, substantial research efforts have been dedicated to addressing this challenge (Castellarin et al., 2018; Costa et al., 2014; Li et al., 2010). The International Association of Hydrological Sciences (IAHS) has also promoted initiatives like predictions in ungauged basins (PUB) to foster research in unmonitored basins (Sivapalan, 2003). As a result, prediction of FDC at ungauged sites has become a major focus within PUB, due to the widespread use of FDC for planning and managing water resources.

PUB has been a long-standing challenge in hydrology (Smakhtin et al., 1997). Common methods include regionalization techniques, where hydrological relationships are transferred from nearby gauged catchments with similar characteristics, and hydrological modeling, where conceptual or physically based models simulate flow using catchment attributes and climate inputs (Mohamoud, 2008; Razavi & Coulibaly, 2013; Shu & Ouarda, 2012). A wide range of approaches—statistical, conceptual, and physical methods—have been applied to this problem. For example, one study proposed a method for constructing FDC at gauged stations using continuous historical flow data (Vogel & Fennessey, 1994). However, continuous flow data is limited in many parts of the world due to the costs associated with installing, operating, and managing gauges (Hrachowitz et al., 2013; Mishra & Coulibaly, 2010; Sivapalan et al., 2003). To address the lack of data, numerous methods for predicting flow in ungauged basins have been widely researched (Mohamaoud, 2008; Shu & Ouarda, 2012). Regression equations are often used to estimate FDC percent exceedances and the parameters of probabilistic models that represent FDC (Mohamaoud, 2008; Pugliese et al., 2016). However, this approach requires defining hydrologically homogeneous regions and involves uncertainties due to the number of physical and climatic variables influencing the water regime of a basin (Castellarin et al., 2004). Regional hydrological models for estimating daily FDC at ungauged river basins exist (Fennessey & Vogel, 1990), but these models may perform poorly at specific gauging stations where the hydrological behavior deviates from the general characteristics of the basin (Burgan & Aksoy, 2022). Alternative approaches include multivariate statistical models (Holmes et al., 2002), geostatistical methods (Goodarzi & Vazirian, 2023; Pugliese et al., 2014)**,** kriging techniques (Castellarin, 2014)**,** linear and nonlinear mathematical equations (Ganora et al., 2009; Yaşar & Baykan, 2013)**,** and spatial nonlinear interpolation methods (Archfield & Vogel, 2010; Hughes & Smakhtin, 1996; Mohamaoud, 2008).

However, these approaches carry inherent uncertainties, as they rely on assumptions about the similarity between basins or the accuracy of model inputs and parameters (Farmer & Vogel, 2013; Gianfagna et al., 2015; Razavi & Coulibaly, 2013; Zelelew & Alfredsen, 2014). The primary limitation lies in the complexity of regional hydrological characteristics and the difficulty of accurately translating those characteristics between ungauged basins (Yi, 2024). Machine learning (ML) techniques have recently emerged as promising alternatives for estimating FDC in ungauged basins, leveraging basin attributes to predict flow characteristics. Although these models offer notable advantages, they are still limited by their dependence on the quality of training data and the potential risk of overfitting, especially in the absence of observed flow data. The lack of observed data fundamentally limits the validation of any FDC derived in ungauged basins, thus making the accurate characterization of flow variability an ongoing challenge.

To address the limitations in predicting FDC in ungauged basins, numerous machine learning algorithms have been applied, including artificial neural networks (ANN) (Atieh et al., 2017), gene expression programming (GEP) (Razaq et al., 2016), multi-output neural networks (MNN) (Worland et al., 2019), support vector machines (SVM) (Razaq et al., 2016), and long short-term memory (LSTM) models (Feng et al., 2021). Despite these advances, significant gaps persist. First, many studies only compare two ML algorithms, lacking a comprehensive evaluation of diverse models, particularly deep learning (DL) methods for streamflow prediction (Arsenault & Brissette, 2014). Second, existing models face challenges in capturing flow variability in ungauged basins due to persistent data scarcity (Feng et al., 2021). This limitation makes it uncertain whether advanced DL models like LSTM are suitable for PUB applications, where there is a dearth of sufficient training data. Third, while various hydrologic, meteorological, and topographic factors have been used to predict flow duration curves (Atieh et al., 2017), few studies rank the importance of these variables across different flow percentiles. Understanding which factors are most influential under specific flow conditions is crucial for accurately predicting streamflow variability. Fourth, most studies use average annual precipitation as the primary predictor (Arsenault & Brissette, 2014; Atieh et al., 2017; Worland et al., 2019), without exploring the impact of different precipitation metrics—such as those accumulated over various durations—on low, medium, and high flow predictions. This lack of exploration leaves gaps in understanding how different rainfall characteristics influence flow predictions. This study aims to address these four gaps through the following objectives:

(a) Determine the importance of different independent variables across various discharge percentiles, particularly examining which factors are most influential in low (Q_{95}), medium (Q_{50}), and high flows (Q_5) discharge scenarios.

(b) Evaluate different combinations of independent variables for FDC predictions, identifying the scenario that provides the most accurate prediction results.

(c) Assess the influence of various precipitation variables on streamflow predictions, particularly their impact on high and low flow conditions.

(d) Compare the performance of different ML and DL models to identify the best-performing model for FDC predictions.

The novelty of this research lies in the development of a comprehensive approach that integrates a wide range of predictor variables—including hydrologic, meteorological, and topographic factors—to predict the full range of discharge percent exceedances using ML and DL algorithms. Moreover, unlike previous studies, this work assesses the relative importance of different factors for each discharge percent exceedances along the FDC, providing valuable insights into the dominant drivers for different flow conditions. The contribution of this study is not only in predicting streamflow in ungauged basins but also in advancing our understanding of how specific physical factors influence different portions of the FDC, thereby supporting more effective water resource management and planning.

2 Study area

Approximately two-thirds of South Korea, predominantly along its eastern coast, consists of mountains. This topography directs the flow of river water westward, leading it into the Yellow Sea. Geographically, South Korea lies within the East Asian region, significantly impacted by the Asian monsoons with four distinct seasons (Fig. 1). The mean annual precipitation in South Korea is approximately 1,300 mm, about 1.6 times greater than the global average (Lee et al., 2023). A significant portion of this precipitation, approximately two-thirds of the annual total, occurs during the rainy season from June to September, frequently resulting in floods. The winter lasts from December to February, while the summer lasts from June to September. The low streamflow is usually observed between December and February. South Korea, experiences high humidity in summer (June to August) due to the influence of the North Pacific high-pressure system.

The country encompasses five major river basins: the Han, Nakdong, Geum, Seomjin, and Yeongsan River Basins (Fig. 1). The Han River Basin has the largest basin area, while the Nakdong River Basin has the longest river course (Table 1). The study area covers the entire South Korea including 140 discharge stations from Han, Nakdong, Geum, Seomjin, and Yeongsan, respectively.

South Korea's major river basins—Han, Nakdong, Geum, Seonjin, and Yeongsan—each contribute uniquely to the country's water resources (Table 1). These basins vary in size, length, and precipitation, reflecting the diverse hydrological patterns across the nation. The Han and Nakdong basins are the largest and serve as critical water sources for the densely populated and industrialized regions, while the smaller Geum, Seonjin, and Yeongsan basins play vital roles in supporting agriculture and local ecosystems. The variation in precipitation among these basins underscores the importance of tailored water management strategies to ensure sustainable use and preservation of water resources across the country.

Table 1 Characteristics of five river basins in South Korea.

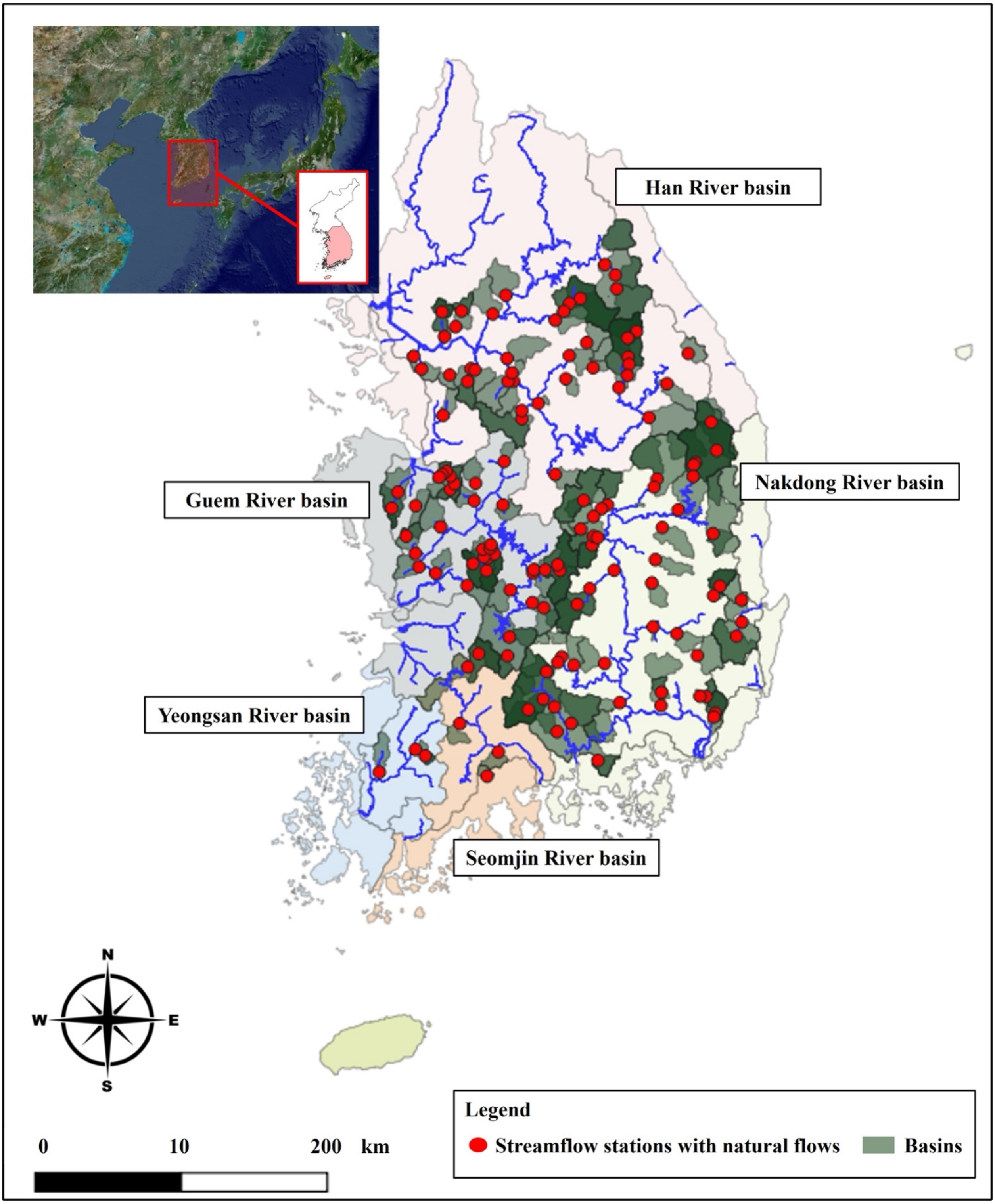
River Basin	Han	Nakdong	Geum	Seonjin	Yeongsan
Area (km^2)	34,428	23,690	9,914	4,914	3,469
River Length (km)	483.0	511.0	388.0	222.0	135.0
Average precipitation	1,261.1	1,163.4	1,225.6	1,415.5	1,310.4

Table 2 Statistics (mean, max, min, and standard deviation (SD) of the hydrologic, meteorological, and physical variables for 140 stream gauge stations.

Variable	Unit	Mean	Max	Min	SD
Watershed Area	km^2	276.9	1590	28	260
Avg basin elevation	m	364.4	914.4	42.9	225
Avg basin slope	%	11.7	19.3	2.8	3.7
Urban area	%	4.9	46.5	0.01	8.8
Forest/Mountain area	%	67.7	97.6	3.1	30.1

Agriculture area	%	26.6	95.3	1.6	28.8
Water area	%	0.8	6.1	0	0.8
P_{annual}	mm	1,231	2,810	854	209

https://doi.org/10.5194/hess-2024-355
Preprint. Discussion started: 21 January 2025



120

Figure 1 Map of South Korea showing the five major river basins: Nakdong, Han, Geum, Seomjin, and Yeongsan River Basins. The red points indicate the locations of 140 streamflow stations with natural flows, and the green areas represent the corresponding basins. © Google Maps (2024).

3 Materials and methods

3.1 Workflow

Fig. 2 presents the methodology of this study. The initial step involves gathering input data, such as discharge, Digital Elevation Model (DEM), Land Use Land Cover (LULC) product, and precipitation. The second step focuses on data preprocessing. Here, discharge data is utilized to estimate the different discharge percentile (Q_5, Q_{10}, Q_{20}, Q_{30}, Q_{40}, Q_{50}, Q_{60}, Q_{70}, Q_{80}, Q_{90}, Q_{95}), while DEM is employed to determine basin area, slope, and elevation. The LULC data assists in calculating the percentages of agriculture, forested, water, and urban areas. Additionally, precipitation data is used to compute the cumulative precipitation over 120 days for each basin. The third step involves developing ML and DL models, including RF, SVR and ENR for ML, and DNN for DL. In the fourth step, the performance of these ML and DL models is evaluated using metrics such as coefficient of determination (R^2), root mean square error (RMSE), standard deviation (SD), and Taylor diagram. The final step entails applying the optimized model to forecast future low flow scenarios using new input data that was not included in the model training and testing phases.

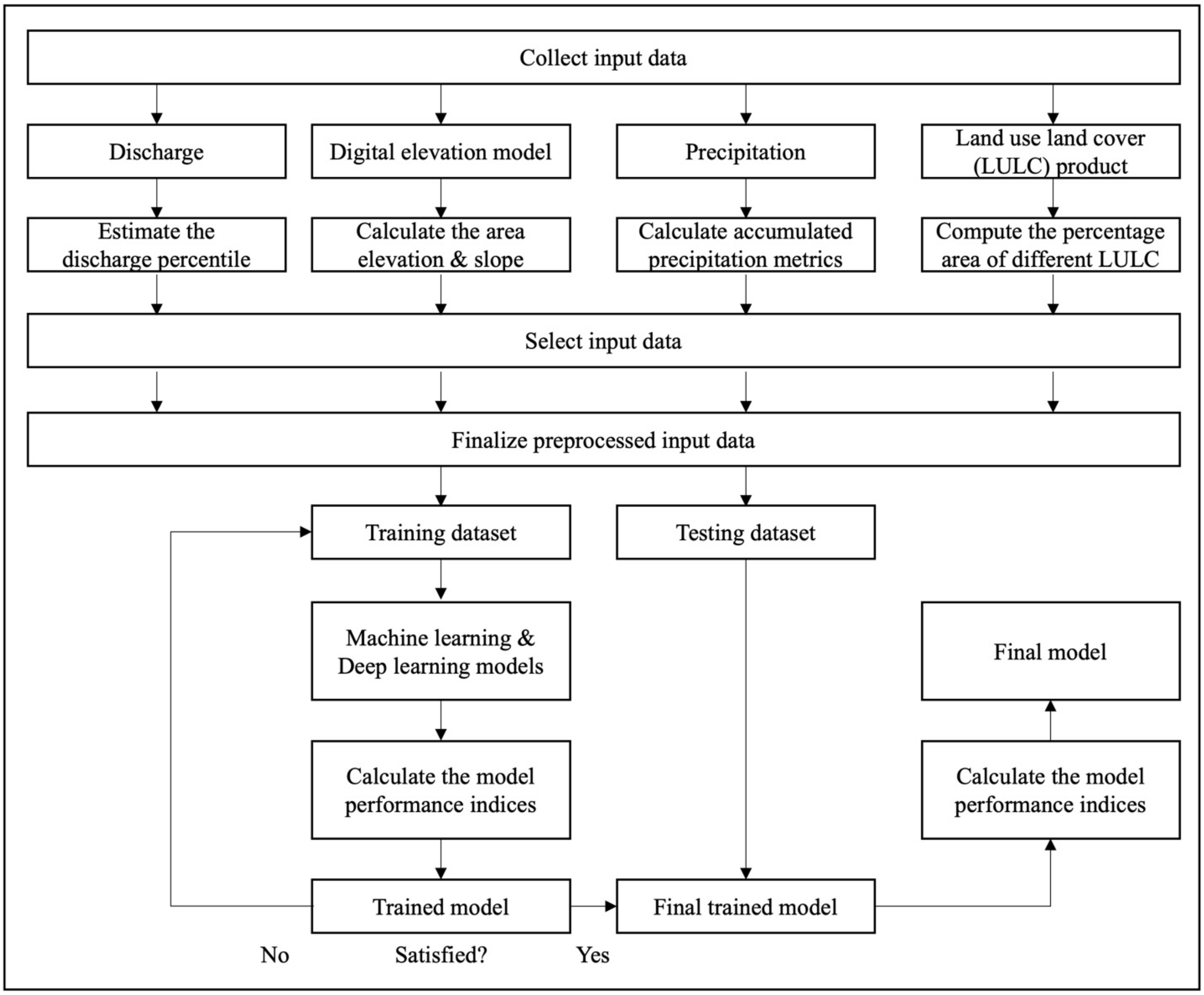

Collect input data
Discharge
Digital elevation model
Precipitation
Land use land cover (LULC) product
Estimate the discharge percentile
Calculate the area elevation & slope
Calculate accumulated precipitation metrics
Compute the percentage area of different LULC
Select input data
Finalize preprocessed input data
Training dataset
Testing dataset
Machine learning & Deep learning models
Final model
Calculate the model performance indices
Calculate the model performance indices
Trained model
Final trained model
No
Satisfied?
Yes

Figure 2 A workflow illustrating the sequence starting from the collection of input data and culminating in the selection of the final model.

3.2 Data collection and preprocessing

We used discharge as the dependent variable, and hydrologic, meteorological, and topographic factors as the independent variables to analyze their influence on streamflow. The independent variables consist of basin area, average basin elevation, average basin slope, precipitation within each basin, and percentages of different LULC types, including urban, forest, agriculture, and water. We chose these variables to effectively capture the diverse characteristics impacting streamflow and build robust prediction models.

3.2.1 Hydrologic data

The study area covers the entire South Korea including 140 discharge stations from Han, Nakdong, Geum, Seomjin, and Yeongsan, respectively. The Korea Water Resources Corporation provides daily inflow data from observation gaging stations (http://www.water.or.kr). We selected discharge stations that have at least five years of data and no upstream water infrastructure, such as reservoirs. We used the U.S. Army Corps of Engineers, Hydrologic Engineering Center's Statistical Software Package to estimate the discharge. Q_5 refers to the discharge where 5% of the entire daily flow is greater than this value. Q_{95} refers to the discharge where 95% of the entire daily flow is greater than this value. Other discharge factors follow the same principle.

3.2.2 Meteorological data

The Korea Meteorological Administration offers open access to meteorological data for the entire South Korea (https://data.kma.go.kr/resources/html/en/aowdp.html). Precipitation data were obtained from two to eight Automatic Weather Stations within each basin. The Thiessen polygon method was utilized to approximate the precipitation data across these basins.

Precipitation data were processed into P_{annual}, P_{10}, P_{30}, P_{120}, P_{150}, and P_{180}, and used as independent variables. P_{annual} represents the annual average precipitation in each watershed, and P_{10} and P_{30} represent the maximum sustained precipitation for 10 and 30 days, respectively. P_{120}, P_{150}, and P_{180} represent the minimum sustained precipitation for 120 days, 150 days, and 180 days.

3.2.3 Topographic and land use data

The Ministry of Land, Infrastructure and Transport provides an open access to the Digital Elevation Model (DEM) (90 m) for the South Korea (http://data.nsdi.go.kr/dataset). Using QGIS, we calculate the watershed area, basin elevation, and average basin slope with the latest DEM from 2020.

The Ministry of Environment provides an open access LULC map for the entire South Korea (https://egis.me.go.kr/req/list.do). LULC maps are categorized based on resolution into three distinct types: the broad category with a 30 m resolution, the medium category at 5 m resolution, and the detailed category with a 1 m resolution. Approximately 91.5% is analyzed using the broad category map, while the remaining area is assessed through a combined use of medium and detailed category maps. For a more focused analysis, the study area is reclassified into four distinct categories: urban, forest/mountain, agriculture, and water.

https://doi.org/10.5194/hess-2024-355
Preprint. Discussion started: 21 January 2025

3.3 Machine learning models

3.3.1 Random forest

RF is an ensemble learning method used for classification and regression. It generates multiple decision trees, trains each tree on different subsets of the dataset, and aggregates several prediction results (Breiman, 2001). This algorithm is a representative ML of the bagging technique, where random samples are repeatedly drawn from the original data to train individual models, and their results are combined. By randomly selecting variables during the tree-building process, RF reduces the correlation between decision trees, thereby enhancing predictive performance and increasing efficiency.

RF reduces the risk of overfitting by randomly sampling multiple times and aggregating the diverse results (Ali et al., 2012). It is less sensitive to outlier data and has the advantage of easy parameter tuning. RF automatically assesses the significance of variables, using measures such as Gini importance, which simplifies the evaluation process. However, increasing the number of trees to enhance predictive performance can significantly increase computational load, resulting in longer training times.

The key parameters used in Random Forest include n_estimators, max_depth, min_samples_split, and min_samples_leaf (Kelkar & Bakal, 2020). n_estimators represent the number of trees in the forest, max_depth denotes the maximum depth of each tree. Min_samples_split signifies the minimum number of samples required to split an internal node, while min_samples_leaf represents the minimum number of samples required to be at a leaf node.

3.3.2 Deep neural network

DNN has become a fundamental tool in machine learning, particularly due to their ability to learn hierarchical representations from data (Yi et al., 2024). Unlike traditional models, DNNs employ multiple hidden layers that allow them to extract progressively complex features from raw inputs. This hierarchical approach has enabled DNN to outperforms many other machine learning methods, especially in tasks such as image classification, speech recognition, and language translation, where high-dimensional data and complex patterns are involved. Each layer in a DNN refines the information passed from the previous layer, allowing for more abstract and informative feature representations (Schmidhuber, 2015).

One of the key advantages of DNN is their ability to model non-linear relationships, thanks to the use of non-linear activation functions such as Rectified Linear Unit (ReLU). These functions introduce non-linearity into the system, enabling the network to solve complex problems that linear models cannot handle effectively. Additionally, DNN leverages large amounts of data and computational resources to fine-tune these models, resulting in highly accurate predictions. This adaptability makes DNN particularly useful in fields like computer vision and natural language processing, where non-linear patterns dominate (LeCun et al., 2015).

Despite their powerful capabilities, DNN also comes with challenges, particularly in the realm of training. The networks can be prone to overfitting, especially when dealing with smaller datasets. To mitigate this, techniques such as regularization, dropout, and batch normalization are often employed. These methods help to generalize the model and prevent it from memorizing the training data. Furthermore, the computational demand of DNN can be a limitation, as training deep models requires significant processing power and memory. However, advancements in hardware, such as the use of GPUs and TPUs, have helped alleviate some of these challenges (Goodfellow et al., 2016).

3.3.3 Support vector regression

SVR, a subset of SVM algorithms, is designed for regression analysis (Smola & Schölkopf, 2004). SVR aims to approximate the relationship between input variables and a continuous target variable by finding a function that minimizes prediction errors. Unlike SVMs for classification, SVR establishes a hyperplane in a continuous space to best fit the data points (Awad & Khanna, 2015). This involves mapping input variables into a high-dimensional feature space to maximize the margin—the distance between the hyperplane and the nearest data points—while minimizing error. SVR accommodates non-linear relationships using a kernel function, which maps data into a higher-dimensional space, enhancing its capability for handling complex variable interactions (Yi et al., 2022). The model employs a hyperplane and a margin within an ε-insensitive tube, which allows some deviations from the hyperplane without penalizing them as errors, providing a robust approach to regression tasks.

SVR is particularly effective in high-dimensional spaces and handles scenarios where the number of dimensions exceeds the number of samples efficiently (Drucker et al., 1997). It performs well when there is a clear margin of separation between classes, making it memory efficient and accurate in such conditions. However, SVR is less suitable for very large datasets as both its computational and memory requirements can become prohibitive. It also struggles with noisy datasets where target classes overlap and underperforms when the feature count per data point greatly exceeds the number of training samples.

C and Gamma are the parameters for a nonlinear SVR with a Gaussian radial basis function kernel (Jiang et al., 2009). *C* controls error tolerance and a low *C* makes the decision surface smooth, while a high *C* aims at classifying all training examples correctly. *Gamma* defines how much influence a single training example has. The larger *gamma* is, the closer other examples must be to be affected.

3.3.4 Elastic net regression

ENR is a versatile machine learning algorithm that integrates the strengths of both Lasso and Ridge Regression. ENR is a type of regularized linear regression model that adds appropriate constraints to the linear regression coefficients, helping to prevent model overfitting. ENR is a versatile machine learning algorithm that integrates the strengths of both Lasso and Ridge Regression. ENR is a compromise between the Ridge model, which improves model stability and solves multicollinearity problems, and the Lasso model, which selects only useful variables and reduces the model size. It utilizes both the L2 norm used in Ridge regression and the L1 norm used in Lasso regression, incorporating the sum of the absolute values and the sum of squares of the regression coefficients as constraints (Kelly et al., 2012).

ENR can effectively discard unimportant variables even in the presence of high correlations between variables, while selecting the most important ones and applying appropriate weights based on importance and correlation. It can also address multicollinearity issues between input variables.

The main hyperparameters of ENR include Alpha and Lambda. Alpha represents the mix ratio between Lasso regression (alpha = 1) and Ridge regression (alpha = 0) (Friedman et al., 2010). This means that during training, the model will test 10 values evenly spaced between 0 (pure Ridge) and 1 (pure Lasso), covering different combinations of Ridge and Lasso penalties. Lambda refers to the penalty strength parameter, which defines a range of 10 values from 10^{-4} to 10^{-1}, meaning that lambda values are tested between 0.0001 and 0.1.

3.4 Scenarios

We analyze streamflow predictions across four scenarios involving various meteorological and topographic variables (Table 3). Scenario 1 uses precipitation and basin area as independent variables. Scenario 2 includes precipitation, basin area, and LULC. Scenario 3 substitutes LULC with basin slope and elevation. Scenario 4 incorporates all these variables. Each scenario tests different accumulated precipitation values to forecast streamflow. In total, we simulate 128 cases incorporating different scenarios and precipitation values for 11 discharge levels that make up the FDC.

For predicting large stream discharge (e.g., Q_5, Q_{10}, Q_{20}, Q_{30}), we use the P_{annual}, P_{10}, and P_{30}, which represent the accumulated maximum precipitation over a short duration to capture short and intense storms (Table 4). In four scenarios for each of the four stream discharges, we simulate four different precipitation variables along with other independent variables, resulting in a total of 64 combinations.

For predicting median stream discharge (e.g., Q_{40}, Q_{50}, Q_{60}, Q_{70}), we use the P_{annual}, P_{120}, and P_{150}, which represent the accumulated maximum precipitation over a medium duration to capture medium storms. In four scenarios for each of the four stream discharges, we simulate three different precipitation variables along with other independent variables, resulting in a total of 48 combinations.

For predicting small stream discharge (e.g., Q_{90}, Q_{95}), we use the annual total precipitation P_{annual} and P_{180}, which represent the accumulated maximum precipitation over a long duration to capture rainfall events. In four scenarios for each of the two stream discharges, we simulate two different precipitation variables along with other independent variables, resulting in a total of 16 combinations.

Table 3 Four scenarios have different combinations of hydrologic, meteorological, and topographic factors.

Scenario	Discharge	Precipitation	Area	LULC	Slope & Elevation
1	O	O	O	X	X
2	O	O	O	O	X
3	O	O	O	X	O
4	O	O	O	O	O

Table 4 Different precipitation values to predict streamflow percent exceedances (Q_5 to Q_{95}).

Discharge	Precipitation		
Q_5	P_{annual}	P_{10}	P_{30}

Q_{10}	P_{10}	P_{30}			
Q_{20}	P_{10}	P_{30}			
Q_{30}	P_{10}	P_{30}			
Q_{40}			P_{120}	P_{150}	
Q_{50}			P_{120}	P_{150}	
Q_{60}			P_{120}	P_{150}	
Q_{70}			P_{120}	P_{150}	
Q_{80}			P_{120}	P_{150}	
Q_{90}			P_{120}	P_{150}	P_{180}
Q_{95}			P_{120}	P_{150}	P_{180}

3.5 Model performance metrics

The three model performance metrics are the R^2, RMSE, and SD. R^2 is a goodness of fit that ranges from 0 to 1 that measures the preciseness of the model outcome. Equation 1 presents the formula for calculating the R where the sum of squares of residuals (SSR) is the differences between the observed and predicted values and the total sum of squares (SST) is the sum of the squared differences between the observed values and the mean of observed values.

$$R^2 = 1 - \frac{Sum\ of\ Squares\ of\ Residuals\ (SSR)}{Total\ Sum\ of\ Square\ (SST)}, \qquad (1)$$

RMSE measures the average difference between a statistical observed values and predicted values. It is the standard deviation of the residuals. Equation 2 shows the calculation for RMSE where y_i is the actual value, $\hat{y}_i$ is the predicted value, and *n* is the total number of observations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \qquad (2)$$

To calculate the standard deviation, which measures the amount of variation or dispersion of a set of values, you can use a similar approach to the RMSE formula you provided. The standard deviation (σ) for a set of observed values (y_i) is calculated as follows, where $\underline{y}$ is the mean of the observed values and n is the total number of observations:

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \underline{y}\right)^2}, \qquad (3)$$

4 Results

4.1 Influence of hydrologic, meteorological, and topographic factors on flow dynamics

We aimed to compare prediction results of four scenarios using different combinations of multiple independent variables. We assessed the relative importance of eight independent variables to construct scenarios based on their combinations. Using percentile flow as the dependent variable and watershed area, average slope, precipitation, among others, as independent variables, we developed a multiple linear regression model. We analyzed the beta coefficients to indicate the relative importance of each independent variable. The beta coefficient represents the slope of each independent variable in the estimated regression equation. A larger beta coefficient suggests that the independent variable has a greater impact on changes in the dependent variable, implying higher relative importance of that variable in the scenario being analyzed. This approach, which involves multiple linear regression and beta coefficients, is widely used to evaluate the relative importance of independent variables. By standardizing all variables in the analysis, we controlled for the impact of variable units on the slope. As beta coefficients represent standardized slopes, they provide a straightforward interpretation: a larger beta coefficient implies a stronger influence of the corresponding independent variable on the dependent variable. We then established the rankings of the importance of eight independent variables for all discharge scenarios.

Table 5 shows the ranking of the independent variables by discharge percentiles. The importance of the eight independent variables was evaluated separately for low, medium, and high discharge percentiles. For high discharge percentiles (above Q_{30}), watershed area and precipitation exhibited the highest importance, while basin elevation and basin slope, representing watershed topographical characteristics, also ranked highly in importance. Water, urban, agriculture, and forest/mountain area (LULC) were given lower priority. In contrast, for medium and low discharge scenarios (below Q_{40}), the importance of LULC was found to be high, followed by watershed area.

Table 5 Summary table of eight independent variables (in beta coefficients) ranked by discharge percentiles. The number indicates the order of importance. In dependent variables include watershed area (Area), precipitation (Pre), basin slope, basin elevation (Elv), water area, urban area, agriculture (Agr) area, forest/mountain (Forest) area.

Discharge	Area	Pre	Slope	Elv	Water	Urban	Agr	Forest
Q_5	0.731	0.014	0.071	0.022	-0.149	-0.576	-1.26	-1.26
	(1)	(4)	(2)	(3)	(5)	(6)	(7)	(8)
Q_{10}	0.965	0.05	0.080	0.072	-0.237	-1.201	-2.59	-2.62
	(1)	(4)	(2)	(3)	(5)	(6)	(7)	(8)
Q_{20}	1.01	0.097	0.014	0.131	-0.124	-0.512	-1.25	-1.28
	(1)	(3)	(4)	(2)	(5)	(6)	(7)	(8)
Q_{30}	0.988	0.133	-0.029	0.158	-0.084	-0.368	-0.953	-0.979
	(1)	(3)	(4)	(2)	(5)	(6)	(7)	(8)
	0.920	0.145	-0.053	0.15	0.147	1.351	2.54	2.53

Q_{40}	(4)	(7)	(8)	(5)	(6)	(3)	(1)	(2)
Q_{50}	0.904	0.135	-0.075	0.156	0.526	4.24	8.36	8.40
	(4)	(7)	(8)	(6)	(5)	(3)	(2)	(1)
Q_{60}	0.903	0.111	-0.087	0.158	0.969	7.622	15.15	15.24
	(5)	(7)	(8)	(6)	(4)	(3)	(2)	(1)
Q_{70}	0.909	0.092	-0.105	0.161	0.867	7.001	13.9	13.9
	(4)	(7)	(8)	(6)	(5)	(3)	(2)	(1)
Q_{80}	0.896	0.057	-0.116	0.149	0.776	6.48	12.7	12.8
	(4)	(7)	(8)	(6)	(5)	(3)	(2)	(1)
Q_{90}	0.759	-0.016	-0.167	0.14	1.162	9.505	18.8	19.0
	(5)	(7)	(8)	(6)	(4)	(3)	(2)	(1)
Q_{95}	0.608	-0.047	-0.164	0.11	1.19	9.79	19.3	19.6
	(5)	(7)	(8)	(6)	(4)	(3)	(2)	(1)

4.2 Scenario evaluation for flow duration curve prediction

We developed a FDC prediction model using four scenarios based on combinations of independent variables. We compared the prediction results of the four scenarios for annual precipitation (P_{annual}), with RF and DNN evaluated at the 70th percentile discharge (Q_{70}), and SVR and ENR evaluated at the 30th percentile discharge (Q_{30}), to determine the scenario with the best prediction results. Additionally, we compared the prediction results of the four scenarios for other discharge percentiles and precipitation metrics. Scenario 4, which included all independent variables, generally showed the most superior prediction results.

In Fig. 3, graphs of results for each scenario are displayed. These graphs visually depict the similarity between a series of observed and predicted data, indicating correlation. Fig. 3 illustrates the results for four scenarios per machine learning technique: (a) corresponds to the results of RF for the 70th percentile discharge (Q_{70}) and annual precipitation (P_{annual}); (b) represents the results of DNN; (c) shows the results of SVR, and (d) displays the results of ENR. Overall, Scenario 4, which included all independent variables, demonstrated superior performance.

Table 6 provides the exact values of the indicators representing the scenario results, along with the correlation coefficient (R^2) shown in Fig. 3. In Fig. 3a, Scenario 4 outperforms the other scenarios by an average of 10%. In Fig. 3b, Scenario 4 outperforms the other scenarios by an average of 5%, and in Fig. 3d, by an average of 3%.

Table 6 Summary table of four scenario models with their R^2, RMSE and SD. For RF is Q_{70} and P_{annual}; DNN is Q_{70} Pannual; SVR is Q_{30} and Pannual; ENR is Q_{30} and P_{annual}.

ML model	Index	Scenario 1	Scenario 2	Scenario 3	Scenario 4
RF	R^2	0.766	0.834	0.854	0.895

	RMSE	0.498	0.419	0.394	0.334
	SD	1.18	1.15	1.10	1.09
	R^2	0.808	0.864	0.828	0.878
DNN	RMSE	0.601	0.506	0.571	0.480
	SD	1.21	1.25	1.32	1.36
	R^2	0.694	0.612	0.736	0.640
SVR	RMSE	2.540	2.90	2.36	2.75
	SD	2.41	2.80	2.25	2.70
	R^2	0.803	0.822	0.822	0.834
ENR	RMSE	2.04	1.94	1.93	1.87
	SD	2.06	1.95	1.93	1.89

https://doi.org/10.5194/hess-2024-355
Preprint. Discussion started: 21 January 2025

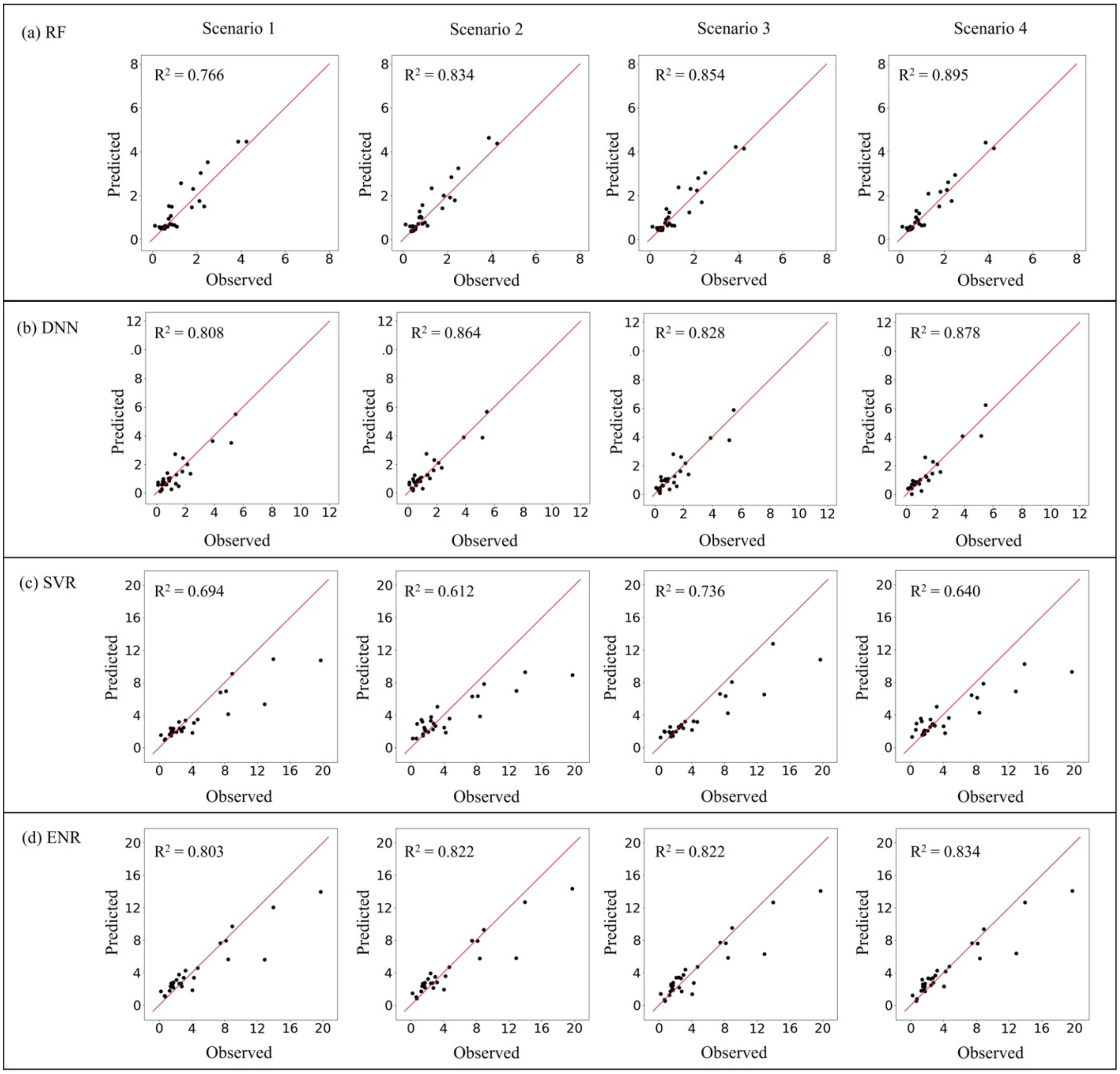


Figure 3 Graphs of results for four scenarios per machine learning technique. Black points closer to the red dashed line represent better results.

4.3 Impact of precipitation variables on streamflow prediction

We used various precipitation variables to predict accurate flow for each flow rate. For Q_5 to Q_{30}, P_{10}, P_{30}, P_{180} and P_{annual} were used, for Q_{40} to Q_{80}, P_{10}, P_{120}, P_{150}, and P_{annual} were used and for Q_{90} and Q_{95}, P_{10}, P_{120}, P_{150}, P_{180} and P_{annual} were used. To compare the results among precipitation variables, analysis was conducted on Scenario 4, which exhibited the highest prediction accuracy. Specifically, analysis was performed on Q_5, representing high flow, and Q_{95}, representing low flow. The analysis of prediction based on precipitation variable P revealed that the predicted flow was not significantly related to the type of precipitation variable P. Because the precipitation variable P has a low level of importance among the independent variables.

Fig. 4 illustrates the result of prediction based on precipitation variables. (a) represents the results of Random Forest for Q_5; (b) illustrates the results of DNN for Q_5. Both models showed similar prediction accuracy for precipitation variables. (c) covers the results of SVR for Q_{95}; (d) corresponds the results of ENR for Q_{95}. Also, both models showed similar prediction accuracy for precipitation variables. The results can be explained to the examination of the importance of independent variables, where the importance of P has low importance.

https://doi.org/10.5194/hess-2024-355
Preprint. Discussion started: 21 January 2025

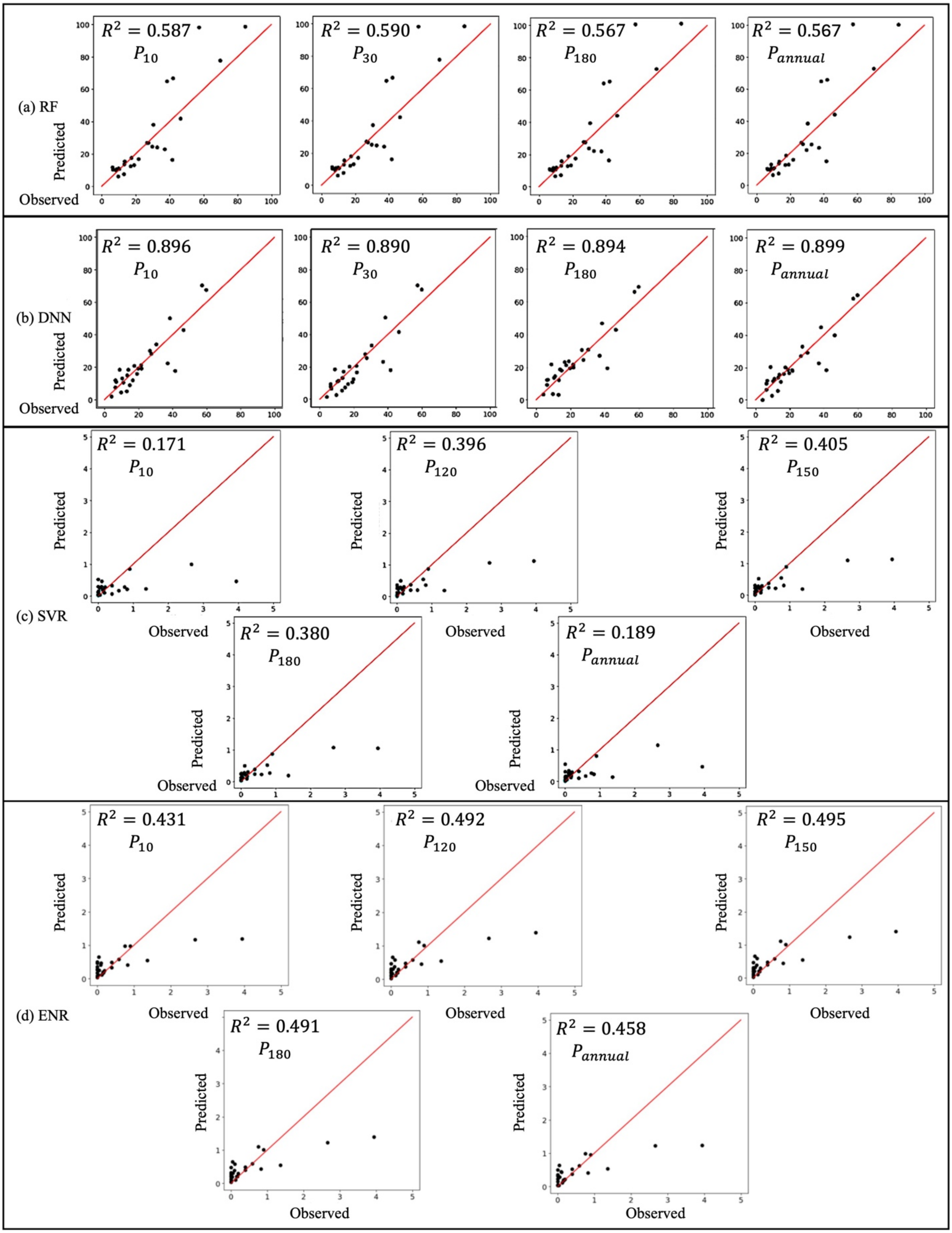


Figure 4 Graphs showing results for different precipitation variables (P) in (a) RF, (b) DNN, (c) SVR, and (d) ENR. Black points closer to the red dashed line indicate better performance.

4.4 Impact of precipitation variables on streamflow prediction

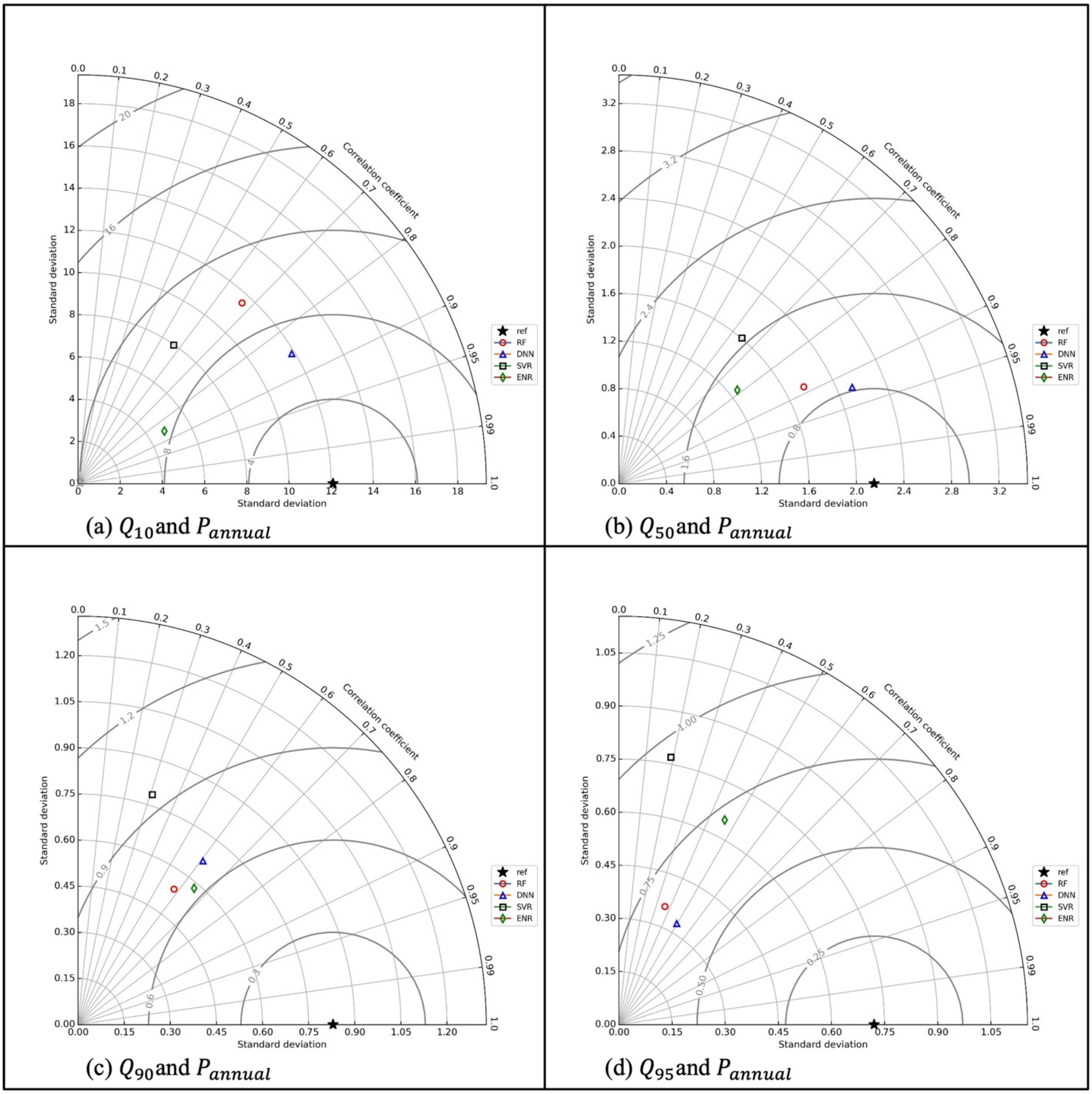
We developed FDC prediction models using RF, DNN, SVR, ENR. Scenario 4 showed the best performance among the four scenarios. Performance varied across different precipitation values at different discharge percentiles. We identified the best-performing ML algorithms by selecting low, medium, and two high discharge percentiles along with the highest-performing precipitation values.

A Taylor diagram visually quantifies the similarity between a set of observations and a reference dataset, summarizing the match in terms of correlation, root-mean-square difference, and standard deviations. Fig. 5 illustrates the Taylor diagram: (a) Q_{10} and P_{annual}, (b) Q_{50} and P_{annual}, (c) Q_{90} and P_{annual}, and (d) Q_{95} and P_{annual}. Points closer to the reference point (start) represent better results. Overall, DNN (blue triangle) outperforms RF (red circle), SVR (black rectangle), and ENR (green diamond) as indicated by higher R^2 values and lower RMSEs.

Based on the Fig. 5, DNN model consistently demonstrates superior performance compared to the other models evaluated. Across all figures, DNN is closest to the reference point in terms of both standard deviation and correlation coefficient, indicating that it best captures the variability in the data while maintaining a strong correlation with the reference observations. This consistent proximity to the reference point suggests that DNN has the lowest centered RMSE across different scenarios, reinforcing its robustness and reliability. In contrast, while the other models (RF, SVR, and ENR) show varying degrees of performance, none consistently match the DNN in both key metrics. Therefore, DNN emerges as the most accurate model for predicting the dataset in question, making it the preferred choice for this analysis.

Table 7 Summary table of four ML models with their R^2 and SD.

ML model	**RF**		**DNN**		**SVR**		**ENR**	
Index	R^2	SD	R^2	SD	R^2	SD	R^2	SD
Q_{10}, P_{annual}	0.673	11.6	0.845	11.9	0.567	7.94	0.854	4.79
Q_{50}, P_{annual}	0.886	1.76	0.924	2.12	0.645	1.60	0.784	1.27
Q_{90}, P_{annual}	0.578	0.540	0.606	0.670	0.307	0.786	0.648	0.582
Q_{95}, P_{annual}	0.362	0.358	0.495	0.329	0.190	0.77	0.459	0.651



(a) $Q_{10}$and P_{annual}

(b) $Q_{50}$and P_{annual}

(c) $Q_{90}$and P_{annual}

(d) $Q_{95}$and P_{annual}

Figure 5 Taylor diagram for four different cases under Scenario 4 (a) Q_{10} and P_{annual}, (b) Q_{50} and P_{annual}, (c) Q_{90} and P_{annual}, and (d) Q_{95} and P_{annual}. Points closer to the reference point (start) represent better results.

5 Discussion

5.1 Validation

To validate the model, we used DNN model based on Scenario 4, which consistently outperformed others ML models across different scenarios. We validated our model using data from an additional station called Cheongamgyo in the Nakdong River (Table 8). This station (2002634) had six years of data. The station has no upstream water infrastructure and maintains natural flow conditions making it ideal for validating our model. We predicted the flow rates at the validation station using P_{annual} as an independent variable since it showed minimal influence on the outcomes (Fig. 6). We found that our FDC prediction models are highly influenced by the watershed area. Upon examining the data, we noticed that the discharge at the validation station was smaller compared to other basins with similar watershed areas.

The watershed area of the validation station is 473.3 square kilometers, with discharges for each percentile from Q_5 to Q_{95} as follows: 23.58 cubic meters per second (CMS) (Q_5), 11.61 CMS (Q_{10}), 5.23 CMS (Q_{20}), 2.89 CMS (Q_{30}), 1.75 CMS (Q_{40}), 1.22 CMS (Q_{50}), 0.85 CMS (Q_{60}), 0.58 CMS (Q_{70}), 0.43 CMS (Q_{80}), 0.30 CMS (Q_{90}), and 0.22 CMS (Q_{95}). To further analyze the discharge percentile, we reviewed data from seven other sites (with basin codes 2018665, 1007605, 21010625, 2010650, 2010690, 2018635, and 1003620) with watershed areas ranging from 400 to 500 square kilometers. The information on the basin codes is available at http://wamis.go.kr/ENG/. The average flow discharge at these sites for each percentile was 38.73 CMS (Q_5), 16.98 CMS (Q_{10}), 7.83 CMS (Q_{20}), 5.16 CMS (Q_{30}), 3.88 CMS (Q_{40}), 3.04 CMS (Q_{50}), 2.36 CMS (Q_{60}), 1.84 CMS (Q_{70}), 1.30 CMS (Q_{80}), 0.71 CMS (Q_{90}), and 0.42 CMS (Q_{95}), which were all higher than the validation station.

Table 8 Independent variables for the additional station (Cheongamgyo). In dependent variables include watershed area (Area), precipitation (Pre), basin slope, basin elevation (Elv), water area, urban area, agriculture (Agr) area, forest/mountain (Forest) area.

Area	Pre	Slope	Elv	Water	Urban	Agr	Forest
(km^2)	(mm)	(%)	(m)	(%)	(%)	(%)	(%)
473.3	916.9	13.3	468.1	0.3	1.4	6.9	91.4

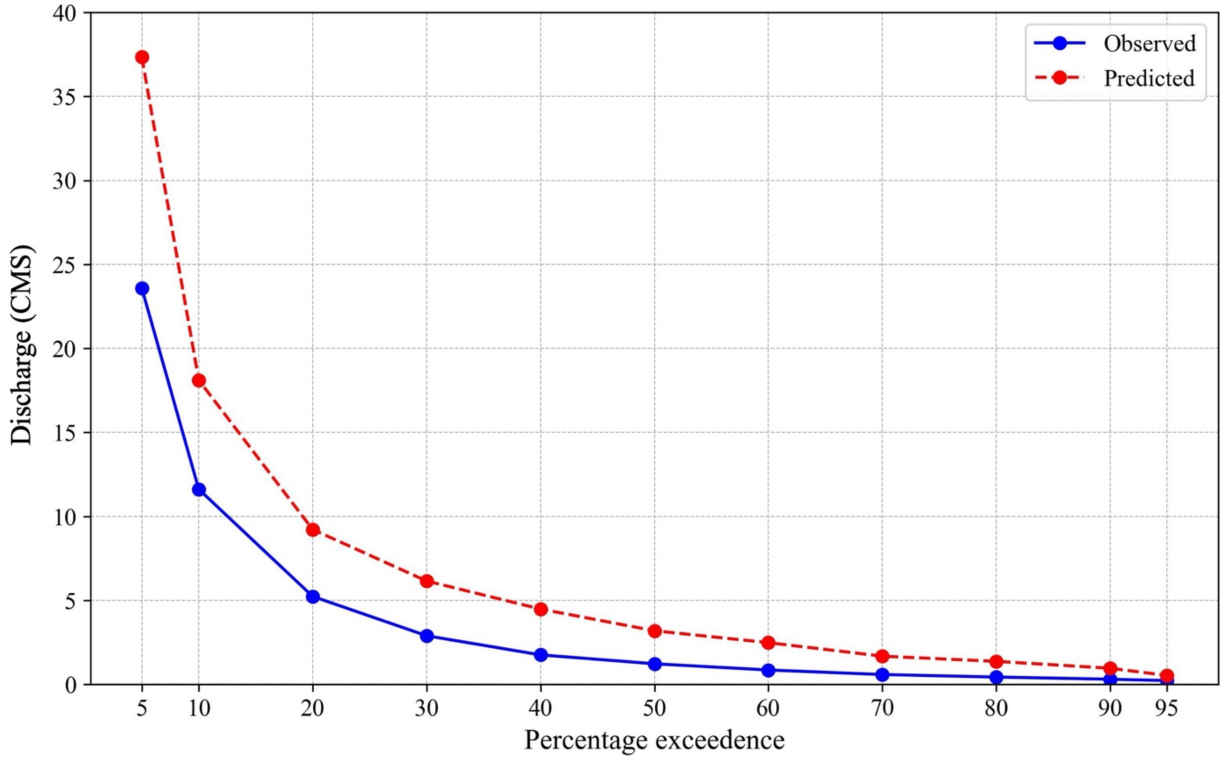


Figure 6 FDC for observed and predicted values at the validation station (Cheongamgyo).

5.2 Comparison of machine learning and deep learning models

Table 9 presents the percentage performance differences of the DNN compared to RF, SVR, and ENR across four scenarios, with positive values indicating that DNN performed better and negative values indicating otherwise. On average, DNN consistently outperformed RF, SVR, and ENR in terms of R² and RMSE. Specifically, DNN achieved an average improvement of 0.90% in R² compared to RF, 20.37% compared to SVR, and 2.80% compared to ENR, indicating better predictive accuracy. For RMSE, DNN showed substantial improvement over the other models, with average reductions in error of 23.94% compared to RF, 395.49% compared to SVR, and 262.61% compared to ENR, demonstrating its efficiency in reducing prediction error. However, for SD, DNN showed mixed results, with a lower performance compared to RF (-11.75%) but better average performance compared to SVR (98.04%) and ENR (52.86%), highlighting variability in capturing flow fluctuations. Overall, DNN performed better in most metrics, proving to be a more reliable approach for predicting FDC.

Table 9 Percentage performance differences of DNN compared to RF, SVR, and ENR across scenarios.

ML model	Index	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Average
RF	R^2	5.20%	3.47%	-3.14%	-1.94%	0.90%
	RMSE	17.1%	17.2%	31.0%	30.4%	23.9%

	SD	-2.48%	-8.00%	-16.67%	-19.85%	-11.75%
	R^2	14.1%	29.2%	11.1%	27.1%	20.4%
SVR	RMSE	322.6%	473.1%	313.3%	472.9%	395.5%
	SD	99.2%	124.0%	70.5%	98.5%	98.0%
	R^2	0.62%	4.86%	0.72%	5.01%	2.80%
ENR	RMSE	239.4%	283.4%	238.0%	289.6%	262.6%
	SD	70.3%	56.0%	46.2%	39.0%	52.9%

5.3 Uncertainty analysis

For the high-flow percent exceedances (Q_5), the analysis shows higher SD, indicating greater uncertainty compared to other percent exceedances such as Q_{30} and beyond (Fig. 7). Large discharges are often influenced by unpredictable factors such as intense precipitation and human interventions, which make reliable model predictions more challenging. The variability of extreme flows contributes to the increased standard deviation, indicating that model predictions are less certain under these conditions.

On the other hand, low-flow conditions represented by Q_{95} also show higher uncertainty, though the factors are different. Low flows are typically driven by groundwater contributions, baseflow, and dry-weather conditions, which can be highly variable based on local hydrogeology, seasonal patterns, and anthropogenic influences. The difficulty in capturing these dynamics, combined with the relative scarcity of low flow events, leads to higher standard deviation and increased uncertainty for these predictions. In contrast, intermediate percent exceedances such as Q_{30} tend to have more stable and frequent flow conditions, leading to more accurate and consistent model predictions with lower SD.

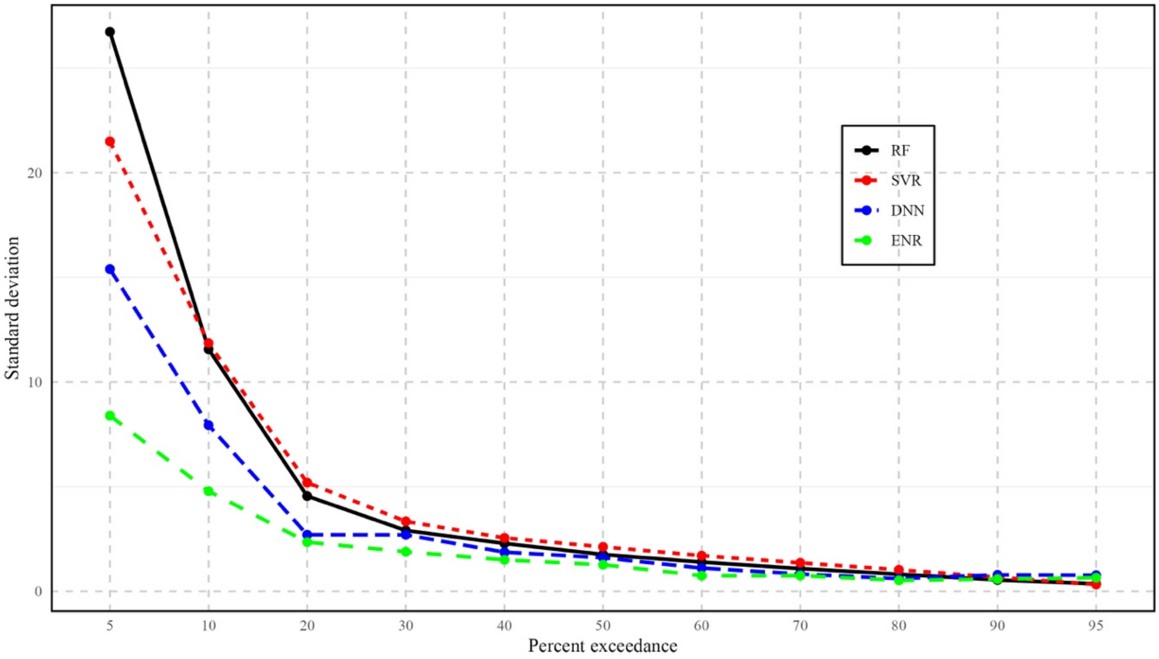


Figure 7 Uncertainty analysis for RF, DNN, SVR, and ENR models by percent exceedance.

5.4 Limitations and future studies

To predict the flow of natural rivers, we selected locations without upstream flow-regulating structures like dams and focused on sites with sufficient data (more than five years), ultimately utilizing 140 locations for model development and validation. However, the success of machine learning, particularly deep learning models, heavily depends on large amounts of high-quality data, which the selected dataset lacked, resulting in suboptimal model performance. DL models are especially effective with large datasets, but struggle to generalize with limited data, which is why simpler models can sometimes outperform more complex ones under such conditions. Despite having 897 streamflow gauging stations in Korea, constraints related to selecting sites without upstream flow-regulating structures, and the need for high-quality data with minimal missing values over five years, reduced the number of suitable locations to just 140. This limited sample size highlights the need for more data to enhance model accuracy and reliability.

For future studies, a potential solution to address the limitation of having insufficient flow observation stations is to increase the available data. We constructed a single dataset using the complete set of data (over five years) from one observation station. To increase the number of data points, if *n* years of data can be collected from one observation station, it is possible to construct a dataset for each year, resulting in n datasets, with each dataset representing one year. By employing this method, since we have collected flow observation stations with more than 5 years of data, the number of datasets can increase by at least fivefold.

With this approach, the number of data points can range from a minimum of around 1,000 to several thousand, making it more suitable for building ML models compared to the current study.

6 Conclusion

This study concludes that ML and DL models are powerful tools for predicting FDC in ungauged basins, offering key insights into hydrologic behavior in the absence of direct measurements. By evaluating the influence of hydrologic, meteorological, and topographic factors, assessing combinations of predictor variables, examining the impact of different precipitation metrics, and comparing ML and DL model performances, this research underscores the potential of advanced modeling techniques for supporting effective water resource management. Below are the key findings that emerged from this study:

a) Identifying influential factors: The importance of independent variables varies significantly across discharge percentiles. High discharges (above Q_{30}) are most influenced by watershed area and precipitation, while topographical characteristics also play a role. Conversely, for low and medium discharges (below Q_{40}), LULC becomes more important, followed by watershed area. This indicates that different variables influence streamflow at different discharge levels.

b) Assessing prediction influence: Scenario 4, which included all independent variables, consistently produced the most accurate predictions for FDC across different machine learning models and discharge percentiles. It outperformed other scenarios, achieving higher prediction accuracy as demonstrated by improved R^2.

c) Assessing prediction influence: Different precipitation variables (P) had no significant influence on streamflow prediction across various flow rates. Even in Scenario 4, which showed the highest prediction accuracy, predictions for high (Q_5) and low flows (Q_{95}) were not significantly impacted by the type of precipitation variable, indicating that precipitation variables have a relatively low importance compared to other factors.

d) Comparing ML and DL model performance: The DNN outperformed other models (RF, SVR, and ENR) in predicting FDC. The DNN showed higher correlation and lower RMSE values, demonstrating robust performance, particularly for intermediate flow conditions (e.g., Q_{30}), while exhibiting higher uncertainty at extreme high (Q_5) and low (Q_{95}) flow scenarios.

This study highlights the potential of ML and DL techniques to improve the prediction of FDC in ungauged basins, while also identifying specific limitations related to data availability and quality. By exploring the importance of various independent variables and comparing different modeling approaches, we have gained insights into factors that influence streamflow prediction across different discharge levels. Addressing these limitations through expanded datasets and incorporating more advanced algorithms could further enhance prediction reliability, ultimately contributing to more effective water resource management and planning.

Code and data availability

All data is publicly primarily available in Korean and some in English. The Korea Water Resources Corporation provides daily inflow data from observation gaging stations (http://www.water.or.kr). The Ministry of Land, Infrastructure and Transport provides an open access to the Digital Elevation Model (DEM) (90 m) for the South Korea (http://data.nsdi.go.kr/dataset). The Korea Meteorological Administration offers open access to meteorological data for the entire South Korea (https://data.kma.go.kr/resources/html/en/aowdp.html). The Ministry of Environment provides an open access LULC map for the entire South Korea (https://egis.me.go.kr/req/list.do). The input data are available online (Yi, 2024a).

Author contributions

SY: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. JY: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. CL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. SL: Data curation, Resources. JJ: Resources. EL: Resources. JY: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Competing interests

The authors declare that they have no conflict of interest.

Financial support

This work was supported by Korea Environment Industry & Technology Institute (KEITI) through the Water Management Project for Drought, funded by Korea Ministry of Environment (MOE) (2022003610004).

References

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *IJCSI International Journal of Computer Science Issues*, *9*(5).

Archfield, S. A., & Vogel, R. M. (2010). Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments. *Water Resources Research*, *46*(10). https://doi.org/10.1029/2009WR008481

Arsenault, R., Breton-Dufour, M., Poulin, A., Dallaire, G., & Romero-Lopez, R. (2019). Streamflow prediction in ungauged basins: analysis of regionalization methods in a hydrologically heterogeneous region of Mexico. *Hydrological Sciences Journal*, *64*(11), 1297–1311. https://doi.org/10.1080/02626667.2019.1639716

Arsenault, R., & Brissette, F. P. (2014). Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research*, *50*(7), 6135–6153. https://doi.org/10.1002/2013WR014898

Atieh, M., Taylor, G., M.A. Sattar, A., & Gharabaghi, B. (2017). Prediction of flow duration curves for ungauged basins. *Journal of Hydrology*, *545*, 383–394. https://doi.org/10.1016/j.jhydrol.2016.12.048

Awad, M., & Khanna, R. (2015). Support Vector Regression. In *Efficient Learning Machines* (pp. 67–80). Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

Booker, D. J., & Snelder, T. H. (2012). Comparing methods for estimating flow duration curves at ungauged sites. *Journal of Hydrology*, *434–435*, 78–94. https://doi.org/10.1016/j.jhydrol.2012.02.031

Breiman, L. (2001). Random Forest. *Machine Learning*, *45*(1), 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Burgan, H. I., & Aksoy, H. (2022). Daily flow duration curve model for ungauged intermittent subbasins of gauged rivers. *Journal of Hydrology*, *604*, 127249. https://doi.org/10.1016/j.jhydrol.2021.127249

Castellarin, A. (2014). Regional prediction of flow-duration curves using a three-dimensional kriging. *Journal of Hydrology*, *513*, 179–191. https://doi.org/10.1016/j.jhydrol.2014.03.050

Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., & Brath, A. (2004). Regional flow-duration curves: reliability for ungauged basins. *Advances in Water Resources*, *27*(10), 953–965. https://doi.org/10.1016/j.advwatres.2004.08.005

Castellarin, A., Persiano, S., Pugliese, A., Aloe, A., Skøien, J. O., & Pistocchi, A. (2018). Prediction of streamflow regimes over large geographical areas: interpolated flow–duration curves for the Danube region. *Hydrological Sciences Journal*, *63*(6), 845–861. https://doi.org/10.1080/02626667.2018.1445855

Costa, V., Fernandes, W., & Naghettini, M. (2014). Regional models of flow-duration curves of perennial and intermittent streams and their use for calibrating the parameters of a rainfall–runoff model. *Hydrological Sciences Journal*, *59*(2), 262–277. https://doi.org/10.1080/02626667.2013.802093

Drucker, H., Surges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *28*(7), 779–784.

Farmer, W. H., & Vogel, R. M. (2013). Performance-weighted methods for estimating monthly streamflow at ungauged sites. *Journal of Hydrology*, *477*, 240–250. https://doi.org/10.1016/j.jhydrol.2012.11.032

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating Prediction Error of Deep Learning Streamflow Models in Large Data-Sparse Regions With Ensemble Modeling and Soft Data. *Geophysical Research Letters*, *48*(14). https://doi.org/10.1029/2021GL092999

Fennessey, N., & Vogel, R. M. (1990). Regional Flow-Duration Curves for Ungauged Sites in Massachusetts. *Journal of Water Resources Planning and Management*, *116*(4), 530–549. https://doi.org/10.1061/(ASCE)0733-9496(1990)116:4(530)

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1). https://doi.org/10.18637/jss.v033.i01

Ganora, D., Claps, P., Laio, F., & Viglione, A. (2009). An approach to estimate nonparametric flow duration curves in ungauged basins. *Water Resources Research*, *45*(10). https://doi.org/10.1029/2008WR007472

Gianfagna, C. C., Johnson, C. E., Chandler, D. G., & Hofmann, C. (2015). Watershed area ratio accurately predicts daily streamflow in nested catchments in the Catskills, New York. *Journal of Hydrology: Regional Studies*, *4*, 583–594. https://doi.org/10.1016/j.ejrh.2015.09.002

Goodarzi, M. R., & Vazirian, M. (2023). A geostatistical approach to estimate flow duration curve parameters in ungauged basins. *Applied Water Science*, *13*(9), 178. https://doi.org/10.1007/s13201-023-01993-4

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Holmes, M. G. R., Young, A. R., Gustard, A., & Grew, R. (2002). A region of influence approach to predicting flow duration curves within ungauged catchments. *Hydrology and Earth System Sciences*, *6*(4), 721–731. https://doi.org/10.5194/hess-6-721-2002

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., … Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Hughes, D., & Smakhtin, V. (1996). Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. *Hydrological Sciences Journal*, *41*(6), 851–871. https://doi.org/10.1080/02626669609491555

Jiang, R., Tang, W., Wu, X., & Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, *10*. https://doi.org/10.1186/1471-2105-10-S1-S65

Kelkar, K. M., & Bakal, J. W. (2020). Hyper Parameter Tuning of Random Forest Algorithm for Affective Learning System. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1192–1195. https://doi.org/10.1109/ICSSIT48917.2020.9214213

Kelly, J. W., Degenhart, A. D., Siewiorek, D. P., Smailagic, A., & Wei Wang. (2012). Sparse linear regression with elastic net regularization for brain-computer interfaces. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 4275–4278. https://doi.org/10.1109/EMBC.2012.6346911

LeBoutillier, D. W., & Waylen, P. R. (1993). Regional variations in flow-duration curves for rivers in British Columbia, Canada. *Physical Geography*, *14*(4), 359–378. https://doi.org/10.1080/02723646.1993.10642485

LeCun, Y., Bengio, Y., & Hinton, G. (2015a). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Bengio, Y., & Hinton, G. (2015b). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, S., Choi, Y., Ji, J., Lee, E., Yi, S., & Yi, J. (2023). Flood Vulnerability Assessment of an Urban Area: A Case Study in Seoul, South Korea. *Water*, *15*(11), 1979. https://doi.org/10.3390/w15111979

Li, M., Shao, Q., Zhang, L., & Chiew, F. H. S. (2010). A new regionalization approach and its application to predict flow duration curve in ungauged basins. *Journal of Hydrology*, *389*(1–2), 137–145. https://doi.org/10.1016/j.jhydrol.2010.05.039

Ma, L., Liu, D., Luan, J., Ming, G., Meng, X., & Huang, Q. (2024). Connecting flow duration curve and precipitation duration curve based on the relationship deduced from machine learning in the watersheds of northern China. *Journal of Hydrology*, *635*, 131235. https://doi.org/10.1016/j.jhydrol.2024.131235

Mimikou, M., & Kaemaki, S. (1985). Regionalization of flow duration characteristics. *Journal of Hydrology*, *82*(1–2), 77–91. https://doi.org/10.1016/0022-1694(85)90048-4

Mishra, A. K., & Coulibaly, P. (2010). Hydrometric network evaluation for Canadian watersheds. *Journal of Hydrology*, *380*(3–4), 420–437. https://doi.org/10.1016/j.jhydrol.2009.11.015

Mohamaoud, Y. M. (2008). Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. *Hydrological Sciences Journal*, *53*(4), 706–724. https://doi.org/10.1623/hysj.53.4.706

Mohamoud, Y. M. (2008). Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. *Hydrological Sciences Journal*, *53*(4), 706–724. https://doi.org/10.1623/hysj.53.4.706

Pugliese, A., Castellarin, A., & Brath, A. (2014). Geostatistical prediction of flow–duration curves in an index-flow framework. *Hydrology and Earth System Sciences*, *18*(9), 3801–3816. https://doi.org/10.5194/hess-18-3801-2014

Pugliese, A., Farmer, W. H., Castellarin, A., Archfield, S. A., & Vogel, R. M. (2016). Regional flow duration curves: Geostatistical techniques versus multivariate regression. *Advances in Water Resources*, *96*, 11–22. https://doi.org/10.1016/j.advwatres.2016.06.008

Razaq, S. A., Shahid, S., Ismail, T., Chung, E.-S., Mohsenipour, M., & Wang, X. (2016). Prediction of Flow Duration Curve in Ungauged Catchments Using Genetic Expression Programming. *Procedia Engineering*, *154*, 1431–1438. https://doi.org/10.1016/j.proeng.2016.07.516

Razavi, T., & Coulibaly, P. (2013). Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. *Journal of Hydrologic Engineering*, *18*(8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

Ridolfi, E., Kumar, H., & Bárdossy, A. (2020). A methodology to estimate flow duration curves at partially ungauged basins. *Hydrology and Earth System Sciences*, *24*(4), 2043–2060. https://doi.org/10.5194/hess-24-2043-2020

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Shu, C., & Ouarda, T. B. M. J. (2012). Improved methods for daily streamflow estimates at ungauged sites. *Water Resources Research*, *48*(2). https://doi.org/10.1029/2011WR011501

Sivapalan, M. (2003). Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrological Processes*, *17*(15), 3163–3170. https://doi.org/10.1002/hyp.5155

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karamibiri, H., Lakshmi, V., Liang, X., Mcdonnel, J. J., Mendiondo, E. M., O'connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., & Zehe, E. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421

Smakhtin, V. Y., Hughes, D. A., & Creuse-Naudin, E. (1997). Regionalization of daily flow characteristics in part of the Eastern Cape, South Africa. *Hydrological Sciences Journal*, *42*(6), 919–936. https://doi.org/10.1080/02626669709492088

Vogel, R. M., & Fennessey, N. M. (1994). Flow-Duration Curves. I: New Interpretation and Confidence Intervals. *Journal of Water Resources Planning and Management*, *120*(4), 485–504. https://doi.org/10.1061/(ASCE)0733-9496(1994)120:4(485)

Won, J., Seo, J., Lee, J., Choi, J., Park, Y., Lee, O., & Kim, S. (2023). Streamflow Predictions in Ungauged Basins Using Recurrent Neural Network and Decision Tree-Based Algorithm: Application to the Southern Region of the Korean Peninsula. *Water*, *15*(13), 2485. https://doi.org/10.3390/w15132485

Worland, S. C., Steinschneider, S., Asquith, W., Knight, R., & Wieczorek, M. (2019). Prediction and Inference of Flow Duration Curves Using Multioutput Neural Networks. *Water Resources Research*, *55*(8), 6850–6868. https://doi.org/10.1029/2018WR024463

Yaşar, M., & Baykan, N. O. (2013). Prediction of Flow Duration Curves for Ungauged Basins with Quasi-Newton Method. *Journal of Water Resource and Protection*, *05*(01), 97–110. https://doi.org/10.4236/jwarp.2013.51012

Yi, S. (2024a). *Supporting data for publication (Prediction of flow duration curve for ungauged basins: Machine learning and deep learning approach)*. HydroShare. https://doi.org/http://www.hydroshare.org/resource/127d1b944fce46caaa89bac236627ed4

Yi, S. (2024b). Water Transfer Energy Efficiency Index for inter-basin water transfer projects. *Water and Environment Journal*. https://doi.org/10.1111/wej.12929

Yi, S., Kondolf, G. M., Sandoval-Solis, S., & Dale, L. (2022). Application of Machine Learning-based Energy Use Forecasting for Inter-basin Water Transfer Project. *Water Resources Management*, *36*(14), 5675–5694. https://doi.org/10.1007/s11269-022-03326-7

Yi, S., Kondolf, G. M., Sandoval Solis, S., & Dale, L. (2024). Groundwater Level Forecasting Using Machine Learning: A Case Study of the Baekje Weir in Four Major Rivers Project, South Korea. *Water Resources Research*, *60*(5). https://doi.org/10.1029/2022WR032779

Yi, S., & Yi, J. (2024). Reservoir-based flood forecasting and warning: deep learning versus machine learning. *Applied Water Science*, *14*(11), 237. https://doi.org/10.1007/s13201-024-02298-w

Zelelew, M. B., & Alfredsen, K. (2014). Use of Cokriging and Map Correlation to Study Hydrological Response Patterns and Select Reference Stream Gauges for Ungauged Catchments. *Journal of Hydrologic Engineering*, *19*(2), 388–406. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000803