FDC prediction and inference: insights from the fusion of machine learning methods and basin characteristic factors

Yu Zhou[1] and Wuyi Wan[1]

[1]Zhejiang University

November 21, 2023

Abstract

This paper aims to solve the problem of accurately estimating flow duration curves (FDC) in catchments lacking diachronic flow data. Based on 645 sets of observed data in the middle and lower reaches of the Yangtze River (YZR), which include 22 basin characteristic variables, eight machine learning (ML) models (SVM, RF, BPNN, ELM, XGB, RBF, PSO-BP, GWO-BP) were integrated to predict the FDC (quantiles of flow rate corresponding to 15 exceedance probabilities were studied), after which the model most suitable for predicting was determined. Finally, the SHapley Additive exPlanation (SHAP) method was used to determine and quantify the impact of various input variables on different quantiles and the degree of that influence. Results indicate that: (1) The GWO-BP model is the best ML model for predicting FDC among the eight, having good prediction performances throughout the entire duration with determination coefficients (R2) on the testing set of 0.86 to 0.94 and Nash-Sutcliff criterion (NSE) of 0.78 to 0.94. (2) The ML model (BPNN) optimized using swarm intelligence can effectively predict FDC. (3) The predictive impact of variables on different quantiles varies, with and BFI_mean contributes significantly to predicting FDC. The former has a negative effect on the prediction result and has better contribution to predicting higher flow rate (i.e., having higher accuracy in predicting the upper tail of FDC), whereas the latter is the opposite. SHAP's explanations are consistent with the physical model, revealing local interactions between predictive factors. The results demonstrate that the method proposed in this paper can greatly improve the prediction accuracy and is innovative and valuable in model interpretation and factor selection.

FDC prediction and inference: insights from the fusion of machine learning methods and basin characteristic factors

Yu Zhou [1], Wuyi Wan [1]

[1] Dept. of Hydraulic Engineering, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China

Corresponding Author : Address: Wuyi Wan, Zhejiang University, Hangzhou 310058, China

Email: wanwuyi@zju.edu.cn

Abstract

This paper aims to solve the problem of accurately estimating flow duration curves (FDC) in catchments lacking diachronic flow data. Based on 645 sets of observed data in the middle and lower reaches of the Yangtze River (YZR), which include 22 basin characteristic variables, eight machine learning (ML) models (SVM, RF, BPNN, ELM, XGB, RBF, PSO-BP, GWO-BP) were integrated to predict the FDC (quantiles of flow rate corresponding to 15 exceedance probabilities were studied), after which the model most suitable for predicting was determined. Finally, the SHapley Additive exPlanation (SHAP) method was used to determine and quantify the impact of various input variables on different quantiles and the degree of that influence. Results indicate that: (1) The GWO-BP model is the best ML model for predicting FDC among the eight, having good prediction performances throughout the entire duration with determination coefficients (R^2) on the testing set of 0.86 to 0.94 and Nash-Sutcliff criterion (NSE) of 0.78 to 0.94. (2) The ML model (BPNN) optimized using swarm intelligence can effectively predict FDC. (3) The predictive impact of variables on different quantiles varies, with and BFI_mean contributes significantly to predicting FDC. The former has a negative effect on the prediction result and has better contribution to predicting higher flow rate (i.e., having higher accuracy in predicting the upper tail of FDC), whereas the latter is the opposite. SHAP's explanations are consistent with the physical model, revealing local interactions between predictive factors. The results demonstrate that the method proposed in this paper can greatly improve the prediction accuracy and is innovative and valuable in model interpretation and factor selection.

Keywords: Flow duration curve (FDC), Streamflow quantile, Basin characteristics, Machine learning, SHAP

INTRODUCTION

1.1Estimating flow duration curves in ungagged catchments

The flow duration curve (FDC) is defined as the relationship between a certain flow rate and the frequency greater than or corresponding to that flow rate during certain periods of time. It is essentially a cumulative distribution function (Searcy, 1959), which comprehensively describes the entire characteristics of runoff in a basin from low flow to flood, and can better reflect the precipitation and runoff conditions of the basin (Cheng et al., 2012). However, many water resource projects are often located in areas without measured runoff data which leads to difficulty to directly obtain the FDC (Veber Costa, 2020). Regional analysis methods can be used to obtain the regional FDC from the areas with measured runoff data, and convert it to areas without measured runoff data to meet the design needs of water resource projects in that area, which is of great help for water resource planning, design, and runoff prediction in areas without measured data (Manuel Almeida, 2021; Veber Costa, 2020; Mancini, 2016; Li et al., 2010; Croker K M, 2003).

There are usually two existing methods to establish FDC in unmeasured areas: process and statistical based method (Blöschl, 2013) Based on process-based method, the probability distribution of daily flow is simulated by establishing a hydrological model of the watershed, considering the characteristics of the watershed and the physical mechanisms of the hydrological process (Cheng et al., 2012; Yokoo and Sivapalan, 2011; Ceola et al., 2010; Botter et al., 2007; Doulatyari et al., 2015) Statistical based method, on the other hand, are modeling methods based on statistical models and data analysis. It infers future hydrological variables by analyzing the statistical characteristics and patterns of historical observed data, which does not require in-depth understanding of the physical mechanisms of hydrological processes and typically require a large amount of data for training and optimization (Burgan and Aksoy, 2022a; Müller et al., 2014; Atieh et al., 2017a). The advantage of process-based models is that they can analytically derive the probability density function of flow and independently simulate the impact of climate or geomorphic changes on FDC, with reliability under non-steady conditions (Ghotbi et al., 2020; Ghotbi et al., 2020) . Its disadvantage is that the assumption of spatial homogeneity in the watershed makes its applicability relatively low (Leong and Yokoo, 2021) . However, parameter estimation for process-based models is less demanding, and can be determined using information such as rainfall, climate, and geomorphic characteristics of the watershed at any location

with data (Schaefli et al., 2013; Karst et al., 2019). Based on existing analysis and research on process-driven method, the physical characteristics of the basin (such as average temperature, potential evapotranspiration, elevation, etc.) distribute precipitation to various parts of the river: groundwater recharge and base flow, surface runoff, and rainstorm flow (Rice and Emanuel, 2017; Ye et al., 2012) . Therefore, the precipitation accumulated and the features of the basin will mainly influence the shape of FDC (Luan et al., 2021).

However, due to the uncertainty in runoff and climate mechanisms, this method has limited application in areas without data (Reichl and Hack, 2017) . In addition, statistical methods often have better estimation performance on FDC than process-based methods (Engeland and Hisdal, 2009; Over et al., 2018) . It mainly include (1) Using regression methods to independently estimate quantiles through basin characteristics (Farmer and Vogel, 2016) (2) Estimating statistical moments and fitting the FDCs using appropriate distribution functions, finding the relationship between the statistical parameters of the function and the basin features (Almeida et al., 2021; Burgan and Aksoy, 2022b; Shin and Park, 2023) ; (3) Using streamflow index-based method (Atieh et al., 2017b). (4) Using geostatistical method ,etc (Goodarzi and Vazirian, 2023) . By comparing these two methods, it has been found that the statistical approach is more sensitive to spatially sparse data, while the process-based approach is more sensitive to observations that are temporally limited (Müller and Thompson, 2016). Although statistical methods often provide better FDC predictions than process-based methods, it is obvious that they typically need a significant amount of post-processing to explain the physical untrustworthiness in the results. There have been numerous studies attempting to combine process-based models and data-driven models in hopes of fully leveraging their respective advantages to improve prediction accuracy. For example, the relationship between quantiles and watershed features was studied to ensure the monotonicity of quantile estimation and explore the relationship between quantiles and their related basin features (Requena et al., 2018; Poncelet et al., 2017) .

1.2 Artificial intelligence methods in the prediction of FDCs

With the development and maturity of data science and artificial intelligence, the research focus of hydrological prediction models has gradually shifted from process-drive to data-driven models (Mohammadrezapour et al., 2019; Sharifi Garmdareh et al., 2018) . The data-driven model was based on the statistical properties of the data, without considering the physical causes of runoff, and directly calculates the correlation between the input and output of the model to obtain hydrological prediction results. Machine learning models typically exhibit a relatively complex model structure. By adjusting parameters and conducting model training, the model can continuously approach the optimal mapping relationship between the input and output, and the predicted results usually have high accuracy. However, due to the limitations of the “black box”, decision-makers cannot directly know how machine learning models calculate decision results (Cortez and Embrechts, 2013) . The ”black box” of machine learning models simplifies model input and training, which makes its prediction results lack practical physical significance, and the model is unable to explain how to obtain prediction results from the causes and mechanisms of runoff formation, resulting in low credibility in practical prediction work. But machine learning methods are widely used in hydrology (Khan et al., 2016; Khan et al., 2019) because they have unreasonable effectiveness when applied to real-world problems (Shen, 2018) . Due to the complexity of hydrological systems which cannot be easily represented by simple conceptual relationships between variables and the nonlinear relationship between watershed characteristics and hydrological characteristics, traditional methods lack sufficient ability to predict FDCs, while artificial intelligence models have some applicative potential (Nearing and Gupta, 2015) .

SVM, ANN, and nonlinear regression (NLR) were used for regression prediction using different runoff duration as output variables and six basin feature selections as input variables in a study of 33 watersheds. The results indicate that SVR is the most suitable model for estimating FDC (Vafakhah and Khosrobeigi Bozchaloei, 2020) . A multi-output neural network model was developed to predict the FDC of 9203 dataless areas in the southeastern United States over a 60-year period from 1950 to 2009, suggesting that compared with single-output neural-network models, multi-output neural networks is capable of learning monotonic relationships between adjacent quantiles and yield better predictions (Worland et al., 2019) .

Machine learning (ML) has demonstrated outstanding performance in forecasting FDC and is extensively

utilized for predicting (Ley et al., 2023; Vaheddoost et al., 2023) . Existing research has primarily concentrated on enhancing the prediction accuracy of FDC through single ML model, neglecting the impact of its influencing factors, and the prediction accuracy through traditional prediction methods is relatively low. Moreover, there are few research of using multiple machine model algorithms for comprehensive comparison, and conducting regionalization research on FDC prediction based on geographical and climatic characteristics. Explainable machine learning (eg. SHAP) is a rapidly developing subfield aimed at understanding how models use inputs for prediction and eliminating the black box problem (Kim, 2017) . Thus, the main issues studied in this paper include (see **Figure 1**):

[Insert Figure 1]

Figure 1 Framework of the prediction and inference of FDC using ML

This paper utilizes a total of 645 sets of samples, made up of 22 basin characteristic variables (including "mutable" and "immutable") in 30 years from 244 hydrometric stations located in the middle and lower reaches of the Yangtze River basin. Using typical characteristics of the basin, regional FDC model was established through machine learning methods and the performance of these methods was compared to determine the most suitable model for predicting the FDC. Firstly, the model includes 22 basin characteristics that were selected and divided into mutable and immutable variables and 15 corresponding quantiles of FDC. Secondly, basin characteristic variable-flow quantile database was established using eight typical ML models to study the nonlinear relationship between the input parameter (basin characteristics) and the fifteen flow quantiles which affect the shape of the FDC. Each quantile was predicted and Taylor plots were applied to compare different ML models to select the best one to estimate FDC. Finally, the key influencing factors of various input on the fifteen quantiles were quantified and determined using SHAP. How these important hydrological factors affect the results was also discussed.

2 DATA AND METHODS

2.1 Study area

The Yangtze River Basin (YRB) has a well-developed water system, with four main tributaries: the Yalong River, the Minjiang River, the Jialing River, and the Han River. This study was conducted in the middle and lower reaches of the YRB with the area of $1.8*10^6 km^2$, located from 90°33' to 122°25' E longitude and 24°30' to 35°45'N latitude (see **Figure 2**) (Li et al., 2021).

[Insert Figure 2]

Figure 2 Study area and location of 224 hydrometric stations

2.2 Data

The daily streamflow data of 267 hydrometric stations was downloaded from the Annual Hydrological Report of the People's Republic of China, in which 224 hydrometric stations with 30 years records from 1970 to 1990 and from 2007 to 2016 were selected (see **Figure 3**). According to the China Meteorological Data Network (https://data.cma.cn/), daily precipitation, potential evaporation and temperature data from 1961 to 2016 were downloaded from 698 evenly distributed weather stations. The observed precipitation, potential evaporation and temperature data were interpolated into the whole Yangtze River basin with the method of Thiessen polygon (Meena et al., 2013). The interpolated precipitation, potential evaporation and temperature data of the basin area corresponding to 224 hydrometric stations were averaged. From the geospatial data cloud (*http://www.gscloud.cn/*), the 30-meter digital elevation model (DEM) was downloaded. Data on 361 reservoirs located in the mid-lower reaches of the YZR were retrieved from the Global Reservoir and Dam database (GRanD) (Lehner et al., 2011) .

[Insert Figure 3]

Figure 3 Streamflow record for 224 hydrologic stations. We here used complete decadal streamflow data for year 1970–1990 and 2007-2016.

2.3 FDC and its corresponding streamflow percentiles

In order to present the characteristics of interdecadal changes, the changes of every 10 years: 1970-1979, 1980-1989, 2007-2016 were counted for total 30 years, providing more observational data for regionalization. Runoff data includes 15 percentiles corresponding to exceedance probabilities: $Q_{0.3}$, $Q_{0.5}$,Q_{1}, Q_{5},Q_{10}, Q_{20},Q_{30}, Q_{50},Q_{70}, Q_{80},Q_{90}, Q_{95},Q_{99}, $Q_{99.5}$,$Q_{99.7}$, representing 0.3%, 0.5%, 1%, 5%, 10%, 20%, 50%, 70%, 80%, 90%, 95%, 99%, 99.5%, 99.7% respectively. These 15 quantile flow values are calculated and are directly related to the return period. Each quantile represents a different part of FDC, ranging from particularly high flow ($Q_{0.3}$,$Q_{0.5}$, Q_{1} andQ_{5}) to particularly low flow (Q_{95}, Q_{99},$Q_{99.5}$ and $Q_{99.7}$). For example, taking Wuxi station as an example (see **Figure 4**) and draw its fitting curve using gamma distribution, the quantile corresponding to Q_5 is 35.1160m^3/s, and the return period of 35.1160m^3/s is a 20-year return period (1/p=1/0.05=20). It is noteworthy that each quantile was log transformed ($\log_{10}$(streamflow)) prior to modelling.

The variation characteristic of the **Figure 4** shows that although the quantiles are randomly selected, they are used to determine the overall shape of each FDC. Because low and high flow rates are usually more important for drought and flood research, which are related to many processes occurring in ecosystems, the selected exceedance probabilities are closer to both ends to explain very large or very small changes of corresponding quantiles, while the distribution of the center part of the curve is determined by exceedance probabilities which are distributed more evenly and fewer. The overall variation of the curve is significant, with both the high and low flow parts of the curve showing a downward trend, and the FDC showing an S-shape. The change at the low tail of the curve is greater than the change at the front of the curve, indicating that the low flow part has a greater change.

[Insert Figure 4]

Figure 4 Streamflow values corresponding to different probability of exceedance (Year 1980-1989_Wuxi Station). Red marks represent the point of 15 streamflow percentiles analyzed in this paper.

For a more intuitive presentation, the final dataset includes a 10-year combination of 224 sites, with each quantile normalized by the logarithmic coordinate transformation (log (quantile)). The 80% of randomly selected stations were used as training sets, while the remaining 20% were reserved for testing.

2.4 Machine learning algorithms

2.4.1 Support vector machine (SVM)

SVM has developed from the optimal classification surface in linearly separable cases, and based on statistical learning theory, it has excellent generalization ability in machine learning by replacing the empirical risk minimization principle with the structural risk minimization principle (Araghinejad, 2013) .By introducing appropriate inner product kernel functions, the samples in the input space can be mapped to high-dimensional spaces, thereby achieving linear classification or regression after a certain nonlinear transformation without increasing computational complexity (see **Figure 5** (a)). When SVM is applied to regression problems, its learning goal is to find the best hyperplane closest to all data points at a given interval. Given the dataset $\{(x_1.\ y_1), (x_2.\ y_2), \ldots, (x_i,\ y_i)\}$, the problem can be converted into an optimization problem under given objective functions and constraint which shows as followed:

Where, and b are the weight coefficients and bias coefficients of the optimal hyperplane respectively. C is the penalty factor, 、 are introduced slack variables; is the given interval (ARNARI S, 1999; Choubin et al., 2018; VAPNIK V, 1997) .

2.4.2 Random forest (RF)

Breiman improved the regression tree model based on the Bagging algorithm and proposed the random forest algorithm (BREIMAN, 2001) , which consists of sub-training sets and sub-regression models (decision trees), which extracts m multiple sample data points from the original sample set D through Bootstrap resampling method to form a sub-training sample set with the same sample size as the original one (see **Figure 5** (b)). For each sub-training sample set, a sub-regression model is constructed, which is called random forest model (Das et al., 2017; Ibarra-Berastegi et al., 2015; Nashwan and Shahid, 2019) .

2.4.3 Back propagation neural network (BPNN)

BPNN is a kind of multi-layer feedforward neural network used for nonlinear functions, which trains the weights and thresholds for many times. When the actual output does not match the observed, it enters the error backpropagation stage: it updates the weights between the hidden layer and the input layer, as well as the thresholds of the hidden layer based on the error of the output values (Maier and Dandy, 2000) . In this paper, the sigmoid (x) function serves as the activation function for input-hidden layer and hidden-output layer (see **Figure 5** (c)), and the formula is as followed:

2.4.4 Extreme learning machine (ELM)

ELM is a feedforward neural network with only single hidden layer (HUANG et al., 2004) , which differs from the BP algorithm in that it does not need to adjust the weight of the hidden layer through reverse iteration. The weight of the input feature vector is randomly assigned from the input layer to the hidden laye. (Huang et al., 2006) (see **Figure 5** (d)). Its difference with BP is that the weights between the layers all need to be iteratively solved using the gradient descent method in the BP algorithm, But for ELM, the weight of input layer and the hidden layer can be determined without iteration for the certain input feature vector.

2.4.5 Extreme gradient boosting (XGB)

The Extreme gradient boosting (XGB) algorithm reduces the error of the previous prediction step by continuously generating new regression trees, gradually narrowing the gap between the true and predicted values, and thereby improving the prediction accuracy (Chen and Guestrin, 2016) (see **Figure 5** (e)). It improves the generalization ability and computational efficiency of the model by introducing regularization terms and parallel computing techniques on the basis of the original gradient boosting decision tree (GBDT) algorithm.

2.4.6 Radial-Basis Function (RBF)

Radial-Basis Function (RBF) network can approximate any nonlinear function, deal with the difficulty to analyze regularity in the system (Majnooni et al., 2023) (see**Figure 5** (f)). Compared with BP, RBF has only one hidden layer, while BP does not limit the number of it. BP is a global approximation of nonlinear mapping, while RBF is a local approximation of nonlinear mapping, with faster training speed.

[Insert Figure 5]

Figure 5Principles of the six machine learning algorithms

2.5 Swarm intelligence optimization algorithm

Although BP neural networks can quickly adapt to various problems, they are also prone to falling into local optimum and overfitting, which will affect the prediction results, while swarm intelligence optimization algorithms demonstrates their effectiveness in solving complex optimization problems (Cai et al., 2018). Therefore, we present swarm intelligence optimization algorithms to seek out both optimal global and local solution. These algorithms generate initial populations and utilize heuristic rules to carry out searches and reduce overfitting risks by optimizing weight parameters to improve the accuracy of FDC prediction.

*2.5.1 Particle swarm optimization*and *gray wolf optimizer algorithm*

The particle swarm optimization algorithm (PSO) originates from a group of animal social interaction models that search for food (Poli et al., 2007) , in which birds are regarded as particles, and all particle information within the group is shared to find the optimal strategy (see**Figure 6** (a)). Assuming a fixed search area is limited to ad -dimensional space, where the number of particles is n . During the iteration process, particles track the "individual extremum " and "global extremum " by changing their own speed and position. The formula for particle speed and position is as follows:

Where and represent the velocities of particle at the iteration and , having both size and direction; Similarly, and are the positions of those particles; is the individual extremum and is the global extremum which is the uniformly distributed random number of the currently found optimal solution in the particle swarm. and denote acceleration coefficients; and are two random numbers between 0 and 1; is the inertia weight.

The Grey Wolf Optimizer (GWO) boasts several advantages, including simplicity in structure and parameters, and capabilities in adaptive adjustment. These features render it an effective tool in striking a balance between local and global optimization (Dehghani et al., 2019; Seifi and Soroush, 2020) . It includes three main stages: finding, surrounding, and attacking preys (Mirjalili et al., 2014) .**Figure 6** (b) shows the working structure. 、 、 and the remaining wolves dominate the social hierarchy, and the level of dominance of wolf species descends from to , which means that is the most powerful wolf category, with wolves 、 and guiding the remaining wolves to search towards their targets. GWO can bypass local optimal stagnation and enhance the global optimal convergence ability (Adnan et al., 2023) (Adnan et al., 2023). The hunting behavior of gray wolves encircling their prey is quantified using the following position update formula (Maroufpoor et al., 2020; Zhou et al., 2021) .

Where and represent the positions of the prey and the grey wolf respectively; t represents the current number of iteration; D represents the distance between the wolf and its prey; is the wolf's position at time (t +1); C and A are coefficient vectors; The above equation represents the distance between the grey wolf and its prey, and the following formula represents the the grey wolves' position updated.

[Insert Figure 6]

Figure 6 The principles of swarm intelligence optimization algorithm: PSO and GWO

2.6 Evaluation of model performance

The performance evaluation criteria including R -squared (R ^{2}), root mean squared error (RMSE), and Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) were used to evaluate the ML performance.

where and are the predicted and the observed value; and are the average values of observed and predicted; n represents the sample number.

2.7 Explanation of ML's output

SHapley Additive exPlanation (SHAP) is based on the Shapley value in game theory, which can take into account the mutual influence between all features, providing more accurate and comprehensive explanations. For each prediction sample, the model generates a corresponding prediction value, and the SHAP value represents the contribution of each feature to the model's overall prediction (Lee, 2017) . The Shapley value was interpreted as an additive feature attribution method.

3 DATABASE DESCRIPTION AND ANALYSI

3.1 Selection of input variables

Though multiple research have been carried out on the prediction of flow rate in the regions without historical flow rate data via FDC, the accuracy of prediction and applicability to regions is still unsettled. Therefore, 224*3 sets of test data were used in this paper. All basin characteristics selected in this paper (Table 1) which

were selected as input variables are divided into two categories: mutable variables and immutable variables. Characteristics that may change over time (such as Aridity index (AI), precipitation, etc.) are regarded to as "mutable" variables. Those characteristics which are reflected as quantities that are fixed over time, for example, such as elevation、drainage area、reservoir control area、average slope (dimensionless)、baseflow index (dimensionless)、topographic wetness index (dimensionless) et al., are regarded as "immutable" variables. The impact of basin characteristics on FDC seems to be empirical (Elena Ridolf et al., 2020). By assessing the value of each parameter, 15 unique corresponding flow percentile values can be obtained.

Table 1 Basin characteristic parameters selected for basins of hydrometric stations

[Insert Table 1]

3.2 Correlation and description of basin characteristics

The results judged by the Spearman correlation coefficients (see**Figure 7**) show that BFI_ Mean and has a significant correlation with flow data includes 15 percentiles corresponding to exceedance probabilities. Smax, AP_max, DP_ std, ATP and SWS all exhibit a negative correlation with the streamflow percentile. From **Figure 7** (c), the relationship between DA and the streamflow percentile varies from negative to positive with the rise of percentiles (decrease in the rate of flow and the return period), while the relationship between wind and flow percentile is reversed. Moreover, BFI_ Mean has a better correlation with low flow rate (bottom part of the FDC curve), while has a higher correlation with high flow rate.

[Insert Figure 7]

Figure 7 Correlation of input and output variables. (a) shows the correlation between variables of selected characteristics and Q0.3. (b) shows the correlation between all the variables. (c) shows the correlation between variables of selected characteristics and the flow rate includes 15 streamflow percentiles corresponding to exceedance probabilities.

[Insert Figure 8]

Figure 8 Distribution of the seven variables with high correlation with $Q_{0.3}$

As shown in **Figure 8** , scatter matrix are further used to display the data distribution features of these input variables. Considering the large size of the matrix, here only a small portion of the matrix was shown, selectively analyzing the seven variables (BFI_mean、Smax、ATP、DP_std、AP_max、、SWS) with high correlation with $Q_{0.3}$ in **Figure 7** to show the data situation. The frequency distribution of these input variables are represented in the diagonal line. The upper right section portrays the scatter plot of input basin characteristics. It can be observed that ATP and DP_std follow a linear relationship. Nevertheless, the variables exhibit a tendency towards nonlinear distribution, thus uncomplicated linear regression and nonlinear regression models are not able to precisely describe the distribution features of the input variables and their impacts on the output variables. The lower left section displays the probability density of the distribution of these input variables, with darker colors indicating a higher probability of data appearing. To take an instance, ATP is concentrated near the value of 500 when BFI_mean assumes the value of 0.5. In order to fully evaluate their impacts on them, SHAP was utilized to increase the interpretability of ML models and disclose the mechanism in section 4.4.

4 RESULTS AND DISCUSSIONS

4.1 Parametric analysis

[Insert Figure 9]

Figure 9

Parametric optimization of the algorithm (SVM and RF)

ML models' performance is influenced by their parameters, making it essential to optimize the parameters by pre-setting the value of MSE. There are several methods adopted to address the poor prediction accuracy and acquire the optimal parameter.

(1) There are two ways to adjust the network parameters of the BPNN: (a) Determining the minimum MSE value by comparing different numbers of hidden layers or nodes in the hidden layer to obtain the optimal parameters. (b) Algorithms of PSO and GWO were combined for parameter optimization and the optimization ability of two methods was compared in**Figure 17** .

(2) The penalty coefficient c and gamma g play an important role in the SVM model, where c determines the generalization ability and g affects the prediction accuracy. The libSVM toolbox (Chang and Lin, 2001) is adopted for parameter optimization of c and g , the process of which is shown in**Figure 9** (a).

(3) In RF model, leaves' number has an impact on the prediction accuracy and grown trees determine whether the model will be over-fitted. The process of optimization are shown in **Figure 9** (b).

(4) The optimal number of nodes of the hidden layer is needed to be determined in the ELM. Therefore, the models' optimization search is completed by pre-setting MSE.

(5) The XGB optimizes parameters through regularization, cross-validation, and so on.

4.2 ML predictive performance analysis

[Insert Figure 10]

Figure 10 Comparison of predicted and observed $Q_{0.3}$

[Insert Figure 11]

Figure 11 Comparison of predicted and observed Q_5

[Insert Figure 12]

Figure 12 Comparison of predicted and observed Q_{50}

[Insert Figure 13]

Figure 13 Comparison of predicted and observed Q_{90}

[Insert Figure 14]

Figure 14 Comparison of predicted and observed $Q_{99.7}$

[Insert Figure 15]

Figure 15 Performance of 8 ML models on the training sets

[Insert Figure 16]

Figure 16 Performance of 8 ML models on the testing sets

In this paper, a comparative analysis is conducted on the predicted and observed values of 15 streamflow percentiles corresponding to the FDCs obtained from 8 models, and the predicted and observed values show consistency. From **Figure 10** to **Figure 14** , we mainly analyze the prediction results of five key streamflow percentiles (, ,,,).

Overall, the ratio of predicted to observed values is stable around 1, and R^2 is close to 1. Neural networks have better generalization capabilities than other machine learning algorithms, as evidenced by their better predictive accuracy on the testing set. The predictive accuracy of each model for the upper tail of the FDC is higher than that for the lower tail, with Q_5 having the highest predictive accuracy and $Q_{99.7}$ having the lowest predictive accuracy. The prediction difference between observed and predicted values may be attributed to the random-like property of hydrological phenomena. Related literature (Montanari and

Koutsoyiannis, 2012) also reached similar conclusions. Among the eight ML models on the testing set, the prediction performance of ELM and RBF is worse, probably due to the simplicity of single-layer neural networks. The predictive ability of the XGB, PSO-BP and GWO-BP models is significantly better. We noticed that these three models show good predictive ability on both the training and testing set, with the R^2 for both the training and testing set being greater than 0.8 at different streamflow percentiles. The predictive performance of models shouldn't only be evaluated by a single metric. Compared to scatter plots, which can only display the relationship between individual indicators (Choubin et al., 2018) , the Taylor diagram integrates three evaluation metrics: correlation coefficient, centered root-mean-square, and standard deviation, based on the cosine relationship between the three to evaluate the predicting performance from different perspectives (**Figure 15** 、**Figure 16**). When the model prediction results are consistent with the observed values, the closer the point "model" is to the point "observed" on the x-axis, the higher the correlation between such models and observations.

[Insert Figure 17]

Figure 17 Comparison of predicted value obtained by PSO-BP, GWP-BP and observed data (Testing sets)

For training sets, the RBF and ELM models' prediction performances are poor, while XGB performs the best. For high tails, the prediction performance of the eight models can be ranked as XGB > SVM >RF > PSO-BP> BPNN > GWO-BP > RBF > ELM, while the prediction performance can be ranked as XGB > RF > BPNN > GWO-BP > SVM >PSO-BP > RBF > ELM for low-tailed data.

As for testing sets, the prediction performance of the BPNN is not as good as the other seven models, while XGB, PSO-BP, and GWO-BP all exhibit good performance. For high tails, the prediction performance can be ranked as follows: PSO-BP > GWO-BP > XGB > RF > SVM > BPNN > ELM > RBF. For low-tailed data, the prediction performance is ranked as follows: GWO-BP > XGB > PSO-BP > RF >BPNN > ELM >SVM >RBF.

It is worth noting that the prediction accuracy of the lower tail of FDC through machine learning is significantly lower than that of the upper tail, but GWO-BP and XGB perform well in predicting the lower tail. By comparing evaluation indicators, it is determined that the GWO-BP and XGB models are the best models for predicting FDC. Moreover, it can be concluded that optimizing ML model parameters using the swarm intelligence optimization algorithms can effectively and significantly enhance the model's predictive capability and generalization ability by comparing BPNN, PSO-BP, and GWO-BP (**Figure 17**).

4.3 Prediction results throughout the entire duration

[Insert Figure 18]

Figure 18 Overall evaluation (R^2) (testing sets) of the estimated quality of ML models (15 streamflow percentiles)

Multiple points of streamflow percentiles can reflect the shape of the FDC. The R^2 and NSE are usually used to assess the model prediction. We believe that an R^2greater than 0.85 indicates good predictive performance for the model. In addition to considering R^2, *std* ,*cor* , and R MSE which we have analyzed and discussed in the condition of six most typical streamflow percentiles in the previous section, models with *NSE* values less than or equal to 0.50, 0.50˜0.65, 0.65˜0.75, and greater than 0.75 are considered to represent 4 categories: bad, satisfactory, good, and excellent performance respectively (Fatehi et al., 2015) .

[Insert Figure 19]

Figure 19 Overall evaluation (NSE) (testing set) of the estimated quality of ML models (15 streamflow percentiles)

As shown in **Figure 18** and **Figure 19** , considering theR^2 and *NSE* criteria, the results show that RF, PSO-BP, and XGB all achieve very good performance, except for Q99.7 which only has satisfactory

results. And it is observed that the XGB model has less predictive power for larger and smaller streamflow percentiles than for the middle streamflow percentiles, with particularly good predictive performance for the middle part. The GWO-BP model performs well across the entire duration range in the testing set (i.e., high flow to low flow) with R^2 of 0.86 to 0.94 and *NSE* of 0.78 to 0.94. The performance of the RF, PSO-BP, and XGB models is also good throughout the entire duration, but it is lower than that of the GWO-BP model. From the perspective of the sustained range of the entire FDC, the GWO-BP is the best model to predict FDC among all the models in this paper.

Compared with the research of Vafakhah and Khosrobeigi Bozchaloei (Vafakhah and Khosrobeigi Bozchaloei, 2020) , which is believed that SVR is the optimal model for predicting FDC with relative RMSE of 9.37 to 1.45 and *NSE* of 0.54 to 0.91, the GWO-BP model we selected in this paper greatly improve the accuracy of prediction.

4.4 Feature importance analysis of the processes

[Insert Figure 20]

Figure 20 Interpretation of the predicted FDC and feature important analysis

Due to the "black box" issue, the ML models have their limitations. (Esterhuizen et al., 2022) . The "feature importance" merely reflects which feature is more important, but how it influences the prediction results is unknown. In this paper, Shapley was used to explain the results of machine learning. The advantage of SHAP values is that they not only reveal the impact of each feature in given samples but also indicate the sign of that impact (i.e., whether it is positive or negative) (Dikshit and Pradhan, 2021) .

From **Figure 20** , it can be seen that the impact of 22 variables of basin characteristics on 6 critical streamflow percentiles (,,,,,) was analyzed. The SHAP values are calculated for each sample and variable, globally demonstrating the impact of feature on the model, which quantifies the contribution of 22 environmental variables to different streamflow percentiles. Each row represents an environmental input variable, with the horizontal axis indicating the distribution of SHAP values. Each point represents a sample, with color indicating the feature value number (red for high values and blue for low values).

Comparing **Figure 20** (a) and Section 3.2, the correlation coefficients between BFI_mean, Smax, ATP, DP_-std, AP_max, , SWS and are relatively high, which are slightly different from the neural network prediction results but generally consistent. It is worth paying attention to that the zero value represents the average value of on the horizontal axis. Considering , there is a rise in the value of SHAP when decreases (changes in the color from red to blue). When reaches its maximum, the is 1 lower than its average value, while reaches its minimum, the is 2 higher than its average value. This is because a high probability of no precipitation days indicates a decrease in precipitation frequency (Cheng et al., 2012) , which will significantly lower the value of flow rate quantile.

It also can be seen that the impact of environmental variables on different streamflow quantiles varies noticeably. However, the two main influencing factors for streamflow quantiles remain nearly unchanged, with and BFI_mean playing the key roles. High values of the will reduce the flow rate of streamflow quantiles, exerting a negative impact, while high values of the BFI_mean feature will increase the flow rate of streamflow quantiles, exerting a positive impact. with higher SHAP values results in lower streamflow percentiles, exerting a negative impact to the output values, while BFI_mean with higher SHAP values results in higher streamflow percentiles, having positive impacts on the output values. Additionally, it can be observed that contributes more to the prediction of high flow rate values such as ,, which means it will predict the upper tail of FDC more accurately, while BFI_mean has a greater impact on the prediction of low flow rate values such as , which means it will predict the lower tail of FDC more accurately.

The main influencing factors obtained through SHAP in this paper are consistent with the physical controls of the gamma distribution fitting parameters in the same region, proving the accuracy of the model in this paper (Yu Zhou, 2023). The influence of annual average precipitation and maximum precipitation on

the prediction results of flow quantiles is not significant, while the prediction results of flow quantiles are closely related to , indicating that high flow may be driven by short-term precipitation events, which are closely related to the frequency of precipitation occurrence, and these events cannot captured by annual average precipitation. The frequency of precipitation has a significant impact on the prediction results of FDC, mainly because it directly affects the runoff generation mechanism and water balance of the watershed (Butcher et al., 2021). High precipitation frequency means that there will be more precipitation in a shorter period of time, which will lead to faster collection of surface runoff, the increase in saturation degree of soil and reduction of the infiltration capacity of soil (Crow et al., 2018). Soil moisture content is high and the evaporation amount will decrease accordingly, resulting in reduced water consumption. Thus, the negative impact of precipitation frequency on the streamflow corresponding to percentiles may be mainly due to the fact that frequent precipitation can lead to faster collection of surface runoff, resulting in slower increases or faster decreases in the streamflow. The BFI is largely influenced by the water storage capacity of the aquifer and human activities. It indicates the importance of the aquifer's water storage capacity in predicting low flow parts (Mazvimavi et al., 2004) . Basins with low BFI cannot maintain good water flow mobility. This may result in a shorter duration of flow in high flow areas, while basins with high BFI can better maintain water flow mobility, thus maintaining high flow conditions for a longer period of time, which can explain why BFI_mean exerts positive impacts to the output values and has greater impacts on the prediction of low flow rate values.

Like other statistical-based methods (Burgan and Aksoy, 2022c) , this paper also has shortcomings of the subjectivity and uncertainty in variables selection (Veber Costa, 2020) . In future research, except for exploring more watershed characteristics that influence FDCs and incorporating them into the model for more precise prediction, larger datasets and scales (e.g., global scale) are needed to be considered and examined to enhance the applicability of the model before it can be applied to various watersheds with more diverse climate and landscape conditions.

For most data-driven models, such as neural networks, only the correlation between inputs and outputs is utilized, and the impact mechanism of influencing factors is unknown (Atieh et al., 2017c; Bozchaloei and Vafakhah, 2015) (Atieh et al., 2017c; Bozchaloei and Vafakhah, 2015). Scholars pointed out that machine learning (ML) can help hydrology make progress in many ways, including (1) incorporating physics into ML models; and (2) improving the explanatory ability of ML models (Shen, 2018) (Shen, 2018). From these two perspectives, the findings of this paper can provide new methods and insights for more accurately data-driven FDC curve prediction and analysis, which will help provide scientific basis for water resource management and hydrological forecasting and reveal the undelying physical processes.

5 CONCLUSION

This paper proposed the different ML methods to estimate FDCs. Based on a total of 645 sets of samples, made up of 22 basin characteristic variables (including "mutable" and "immutable"), eight ML models are integrated to predict the FDC (flow quantiles corresponding to 15 exceedance probabilities). Moreover, the SHAP analysis was used to identify the main input variables that affect the prediction results of different streamflow quantiles and the degree of that influence. The optimal model for predicting under this environmental condition was found. The main conclusions can be drawn in the following:

1. With the high prediction accuracy and good generalization ability, GWO-BP and XGB are the best models for predicting FDC. Moreover, optimizing ML model parameters using the swarm intelligence optimization algorithms can significantly enhance the model's predictive capability and generalization ability of the original BPNN.
2. From the perspective of the sustained range of the entire FDC, GWO-BP is the best predictive model among the eight with R^2 of 0.86 to 0.94 andNSE of 0.78 to 0.94 in the testing set. It significantly improved the prediction accuracy of existing research, which is believed that SVR is the optimal model for predicting FDC with RMSE of 9.37 to 1.45 and NSE of 0.54 to 0.91 (Vafakhah and Khosrobeigi Bozchaloei, 2020) .

3. The predictive impact of variables on different quantiles varies, with and BFI_mean contributes the most significantly to predicting FDC. The has negative effects on the prediction result and has better contribution to predicting higher flow rate, which is mainly due to the fact that frequent precipitation can lead to faster collection of surface runoff, resulting in slower increases or faster decreases in the streamflow. Basins with low BFI cannot maintain good water flow mobility, which may result in a shorter duration of flow in high flow areas, while basins with high BFI can better maintain water flow mobility, thus maintaining high flow conditions for a longer period of time. Therefore, BFI_mean exerts positive impacts to the output values and has a greater impact on the prediction of low flow rate values.

ACKNOWLEDGEMENTS

The authors are deeply grateful for the support provided by the National Natural Science Foundation of China (No. 52379080 and No. 52079122)

DATA AVAILABILITY

Daily precipitation and temperature data (1961-2016) were from the China Meteorological Data Network (https://data.cma.cn/). The 30-meter digital elevation model (DEM) was from the geospatial data cloud (http://www.gscloud.cn/). Data on 361 reservoirs located in the mid-lower reaches of the YZR were retrieved from the Global Reservoir and Dam database (GRanD) (Lehner et al., 2011) .

REFERENCES

References:

Adnan, R.M. et al., 2023. Improved prediction of monthly streamflow in a mountainous region by Metaheuristic-Enhanced deep learning and machine learning models using hydroclimatic data. Theoretical and applied climatology.

Almeida, M., Pombo, S., Rebelo, R. and Coelho, P., 2021. The probability distribution of daily streamflow in perennial rivers of Angola. Journal of Hydrology, 603: 126869.

Araghinejad, S., 2013. Data-Driven Modeling: Using MATLAB in Water Resources and Environmental Engineering, 67. Springer Nature, Dordrecht.

ARNARI S, W.S., 1999. Improving support vector machine classifiers by modfying kernel functions. Neural Networks, 12(6): 783-789.

Atieh, M., Taylor, G., M. A. Sattar, A. and Gharabaghi, B., 2017a. Prediction of flow duration curves for ungauged basins. Journal of Hydrology, 545: 383-394.

Atieh, M., Taylor, G., M. A. Sattar, A. and Gharabaghi, B., 2017b. Prediction of flow duration curves for ungauged basins. Journal of Hydrology, 545: 383-394.

Atieh, M., Taylor, G., M. A. Sattar, A. and Gharabaghi, B., 2017c. Prediction of flow duration curves for ungauged basins. Journal of Hydrology, 545: 383-394.

Blöschl, G., 2013. Runoff prediction in ungauged basins: synthesis across processes, places and scales. Cambridge University Press.

Botter, G., Peratoner, F., Porporato, A., Rodriguez Iturbe, I. and Rinaldo, A., 2007. Signatures of large-scale soil moisture dynamics on streamflow statistics across U.S. climate regimes. Water Resources Research, 43(11).

Bozchaloei, S.K. and Vafakhah, M., 2015. Regional Analysis of Flow Duration Curves Using Adaptive Neuro-Fuzzy Inference System. Journal of Hydrologic Engineering, 20(12): 06015008.

BREIMAN, L., 2001. Random forests. Machine Learning, 45(1): 5-32.

Burgan, H.I. and Aksoy, H., 2022a. Daily flow duration curve model for ungauged intermittent subbasins of gauged rivers. Journal of Hydrology, 604: 127249.

Burgan, H.I. and Aksoy, H., 2022b. Daily flow duration curve model for ungauged intermittent subbasins of gauged rivers. Journal of Hydrology, 604: 127249.

Burgan, H.I. and Aksoy, H., 2022c. Daily flow duration curve model for ungauged intermittent subbasins of gauged rivers. Journal of Hydrology, 604: 127249.

Butcher, J.B. et al., 2021. An Efficient Statistical Approach to Develop Intensity-Duration-Frequency Curves for Precipitation and Runoff under Future Climate. Clim Change, 164(1-2): 1-3.

Ceola, S. et al., 2010. Comparative study of ecohydrological streamflow probability distributions. Water Resources Research, 46(9).

Chang, C.C. and Lin, C.J., 2001. Training nu-support vector classifiers: theory and algorithms. Neural Comput, 13(9): 2119-47.

Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. ACM, Ithaca, pp. 785-794.

Cheng, L. et al., 2012. Exploring the physical controls of regional patterns of flow duration curves – Part 1: Insights from statistical analyses. Hydrology and Earth System Sciences, 16(11): 4435-4446.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F. and Klove, B., 2018. River suspended sediment modelling using the CART model: A comparative study of machine learning techniques. Science of The Total Environment, 615: 272-281.

Cortez, P. and Embrechts, M.J., 2013. Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225: 1-17.

Croker K M, Y.A.R.Z., 2003. Flow duration curve estimation in ephemeral catchments in Portugal. Hydrological sciences journal, 48(3): 427-439.

Crow, W.T., Chen, F., Reichle, R.H., Xia, Y. and Liu, Q., 2018. Exploiting Soil Moisture, Precipitation, and Streamflow Observations to Evaluate Soil Moisture/Runoff Coupling in Land Surface Models. Geophysical Research Letters, 45(10): 4869-4878.

Das, S., Chakraborty, R. and Maitra, A., 2017. A random forest algorithm for nowcasting of intense precipitation events. Advances in Space Research, 60(6): 1271-1282.

Dehghani, M., Seifi, A. and Riahi-Madvar, H., 2019. Novel forecasting models for immediate-short-term to long-term influent flow prediction by combining ANFIS and grey wolf optimization. Journal of Hydrology, 576: 698-725.

Dikshit, A. and Pradhan, B., 2021. Interpretable and explainable AI (XAI) model for spatial drought prediction. Science of The Total Environment, 801: 149797.

Doulatyari, B. et al., 2015. Predicting streamflow distributions and flow duration curves from landscape and climate. Advances in Water Resources, 83: 285-298.

Engeland, K. and Hisdal, H., 2009. A Comparison of Low Flow Estimates in Ungauged Catchments Using Regional Regression and the HBV-Model. Water Resources Management, 23(12): 2567-2586.

Esterhuizen, J.A., Goldsmith, B.R., Linic, S. and Univ. Of Michigan, A.A.M.U., 2022. Interpretable machine learning for knowledge generation in heterogeneous catalysis. Nature catalysis, 5(3): 175-184.

Farmer, W.H. and Vogel, R.M., 2016. On the deterministic and stochastic use of hydrologic models. Water Resources Research, 52(7): 5619-5633.

Fatehi, I., Amiri, B.J., Alizadeh, A. and Adamowski, J., 2015. Modeling the Relationship between Catchment Attributes and In-stream Water Quality. Water resources management, 29(14): 5055-5072.

Ghotbi, S., Wang, D., Singh, A., Bloschl, G. and Sivapalan, M., 2020. A New Framework for Exploring Process Controls of Flow Duration Curves. Water Resources Research, 56(1).

Ghotbi, S., Wang, D., Singh, A., Mayo, T. and Sivapalan, M., 2020. Climate and Landscape Controls of Regional Patterns of Flow Duration Curves Across the Continental United States: Statistical Approach. Water Resources Research, 56(11).

Goodarzi, M.R. and Vazirian, M., 2023. A geostatistical approach to estimate flow duration curve parameters in ungauged basins. Applied Water Science, 13(9).

HUANG, G., ZHU, Q. and SIEW, C., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks. IEEE, Piscataway NJ, pp. 985-990 vol.2.

Huang, G., Zhu, Q. and Siew, C., 2006. Extreme learning machine: Theory and applications. Neurocomputing, 70(1-3): 489-501.

Ibarra-Berastegi, G., Saenz, J., Esnaola, G., Ezcurra, A. and Ulazia, A., 2015. Short-term forecasting of the wave energy flux: Analogues, random forests, and physics-based models. Ocean Engineering, 104: 530-539.

Karst, N., Dralle, D. and Muller, M.F., 2019. On the Effect of Nonlinear Recessions on Low Flow Variability: Diagnostic of an Analytical Model for Annual Flow Duration Curves. Water Resources Research, 55(7): 6125-6137.

Khan, M.Y.A., Hasan, F., Panwar, S. and Chakrapani, G.J., 2016. Neural network model for discharge and water-level prediction for Ramganga River catchment of Ganga Basin, India. Hydrological sciences journal, 61(11): 2084-2095.

Khan, M.Y.A., Tian, F., Hasan, F. and Chakrapani, G.J., 2019. Artificial neural network simulation for prediction of suspended sediment concentration in the River Ramganga, Ganges Basin, India. International Journal of Sediment Research, 34(2): 95-107.

Kim, F.D.A.B., 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv, 2017.

Lee, S.L.S., 2017. A Unified Approach to Interpreting Model Predictions. NIPS.

Lehner, B. et al., 2011. High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. Frontiers in ecology and the environment, 9(9): 494-502.

Leong, C. and Yokoo, Y., 2021. A step toward global-scale applicability and transferability of flow duration curve studies: A flow duration curve review (2000–2020). Journal of Hydrology, 603: 126984.

Ley, A., Bormann, H. and Casper, M., 2023. Intercomparing LSTM and RNN to a Conceptual Hydrological Model for a Low-Land River with a Focus on the Flow Duration Curve. Water, 15(3): 505.

Li, M., Shao, Q., Zhang, L. and 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.

Luan, J., Liu, D., Lin, M. and Huang, Q., 2021. The construction of the flow duration curve and the regionalization parameters analysis in the northwest of China. Journal of Water and Climate Change, 12(6): 2639–2653.

Maier, H.R. and Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling & Software, 15(1): 101-124.

Majnooni, S. et al., 2023. Long-term precipitation prediction in different climate divisions of California using remotely sensed data and machine learning. Hydrological sciences journal: 1-25.

Mancini, L.B.G.R., 2016. Regionalization of Flow-Duration Curves through Catchment Classification with Streamflow Signatures and Physiographic – Climate Indices. Journal of Hydrologic Engineering, 21(3):

05015027.

Manuel Almeida, S.P.R.R., 2021. The probability distribution of daily streamflow in perennial rivers of Angola. Journal of Hydrology, 603: 126869.

Maroufpoor, S., Bozorg-Haddad, O. and Maroufpoor, E., 2020. Reference evapotranspiration estimating based on optimal input combination and hybrid artificial intelligent model: Hybridization of artificial neural network with grey wolf optimizer algorithm. Journal of Hydrology, 588: 125060.

Mazvimavi, D., Meijerink, A.M.J. and Stein, A., 2004. Prediction of base flows from basin characteristics: a case study from Zimbabwe / Prevision de debits de base a partir de caracteristiques du bassin: une etude de cas au Zimbabwe. Hydrological sciences journal, 49(4): 715-715.

Mirjalili, S., Mirjalili, S.M. and Lewis, A., 2014. Grey Wolf Optimizer. Advances in Engineering Software, 69: 46-61.

Mohammadrezapour, O., Piri, J. and Kisi, O., 2019. Comparison of SVM, ANFIS and GEP in modeling monthly potential evapotranspiration in an arid region (Case study: Sistan and Baluchestan Province, Iran). Water Supply, 19(2): 392-403.

Muller, M.F. and Thompson, S.E., 2016. Comparing statistical and process-based flow duration curve models in ungauged basins and changing rain regimes. Hydrology and Earth System Sciences, 20(2): 669-683.

Muller, M.F., Dralle, D.N. and Thompson, S.E., 2014. Analytical model for flow duration curves in seasonally dry climates. Water Resources Research, 50(7): 5510-5531.

Nash, J.E. and Sutcliffe, J.V., 1970. RIVER FLOW FORECASTING THROLIGH CONCEPTUAL MODELS PART I - A DISCLISSION OF PRINCIPLES. ECOLOGICAL MODELLING, 10(3): 0-290.

Nashwan, M.S. and Shahid, S., 2019. Symmetrical uncertainty and random forest for the evaluation of gridded precipitation and temperature data. Atmospheric Research, 230: 104632.

Nearing, G.S. and Gupta, H.V., 2015. The quantity and quality of information in hydrologic models. Water Resources Research, 51(1): 524-538.

Over, T.M., Farmer, W.H. and Russell, A.M., 2018. Refinement of a regression-based method for prediction of flow-duration curves of daily streamflow in the conterminous United States. Scientific Investigations Report.

Poli, R., Kennedy, J. and Blackwell, T., 2007. Particle swarm optimization. Swarm Intelligence, 1(1): 33-57.

Poncelet, C., Andreassian, V., Oudin, L. and Perrin, C., 2017. The Quantile Solidarity approach for the parsimonious regionalization of flow duration curves. Hydrological sciences journal, 62(9): 1364-1380.

Reichl, F. and Hack, J., 2017. Derivation of Flow Duration Curves to Estimate Hydropower Generation Potential in Data-Scarce Regions. Water, 9(8): 572.

Requena, A.I., Ouarda, T.B.M.J. and Chebana, F., 2018. Low-flow frequency analysis at ungauged sites based on regionally estimated streamflows. Journal of hydrology (Amsterdam), 563: 523-532.

Rice, J.S. and Emanuel, R.E., 2017. How are streamflow responses to the ElNino Southern Oscillation affected by watershed characteristics? Water Resources Research, 53(5): 4393-4406.

Schaefli, B., Rinaldo, A. and Botter, G., 2013. Analytic probability distributions for snow-dominated streamflow. Water Resources Research, 49(5): 2701-2713.

Searcy, J.K., 1959. Flow-duration curves, Manual of hydrology. U.S. Geological Survey.

Seifi, A. and Soroush, F., 2020. Pan evaporation estimation and derivation of explicit optimized equations by novel hybrid meta-heuristic ANN based methods in different climates of Iran. Computers and Electronics in Agriculture, 173: 105418.

Sharifi Garmdareh, E., Vafakhah, M. and Eslamian, S.S., 2018. Regional flood frequency analysis using support vector regression in arid and semi-arid regions of Iran. Hydrological sciences journal, 63(3): 426-440.

Shen, C., 2018. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. Water Resources Research, 54(11): 8558-8593.

Shin, Y. and Park, J., 2023. Modeling climate extremes using the four-parameter kappa distribution for r-largest order statistics. Weather and Climate Extremes, 39: 100533.

Vafakhah, M. and Khosrobeigi Bozchaloei, S., 2020. Regional Analysis of Flow Duration Curves through Support Vector Regression. Water resources management, 34(1): 283-294.

Vaheddoost, B., Yilmaz, M.U. and Safari, M.J.S., 2023. Estimation of flow duration and mass flow curves in ungauged tributary streams. Journal of Cleaner Production, 409: 137246.

VAPNIK V, G.S.E.S., 1997. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing, Advances in Neura

l lnformation Processing Systems, pp. 281-287.

Veber Costa, W.F.A.S., 2020. Identifying Regional Models for Flow Duration Curves with Evolutionary Polynomial Regression: Application for Intermittent Streams. Journal of Hydrologic Engineering, 25(1): 04019059.

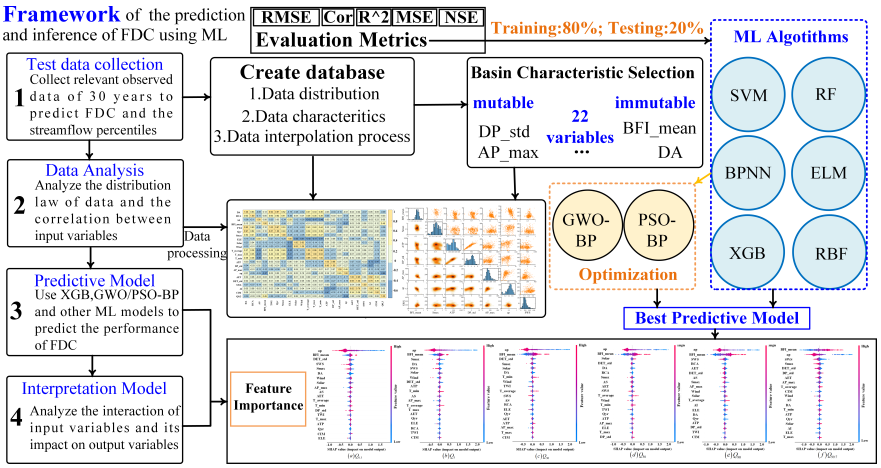
Worland, S.C., Steinschneider, S., Asquith, W., Knight, R. and Wieczorek, M., 2019. Prediction and Inference of Flow Duration Curves Using Multioutput Neural Networks. Water Resources Research, 55(8): 6850-6868.

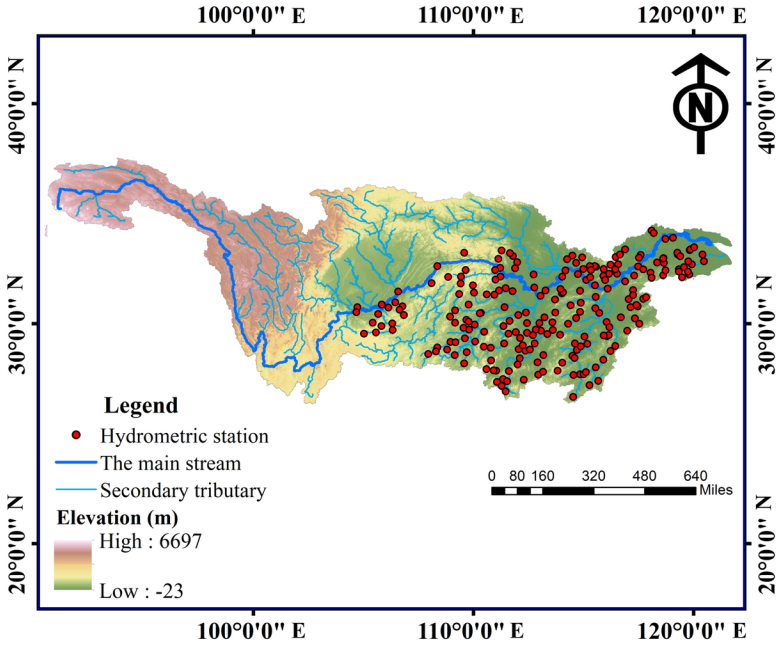
Ye, S., Yaeger, M., Coopersmith, E., Cheng, L. and Sivapalan, M., 2012. Exploring the physical controls of regional patterns of flow duration curves – Part 2: Role of seasonality, the regime curve, and associated process controls. Hydrology and Earth System Sciences, 16(11): 4447-4465.

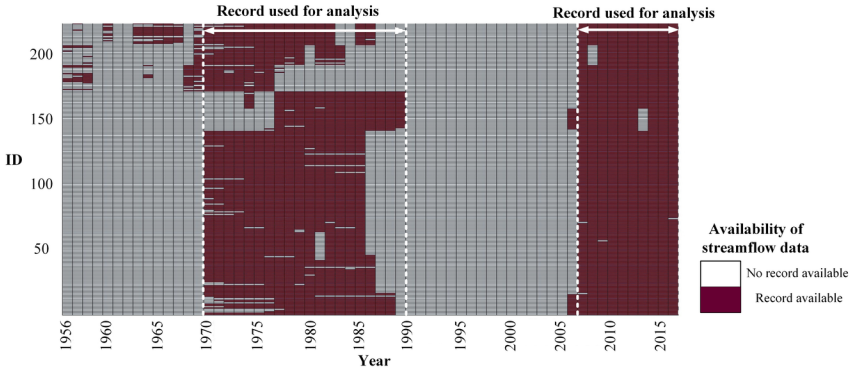
Yokoo, Y. and Sivapalan, M., 2011. Towards reconstruction of the flow duration curve: development of a conceptual framework with a physical basis. Hydrology and Earth System Sciences, 15(9): 2805-2819.

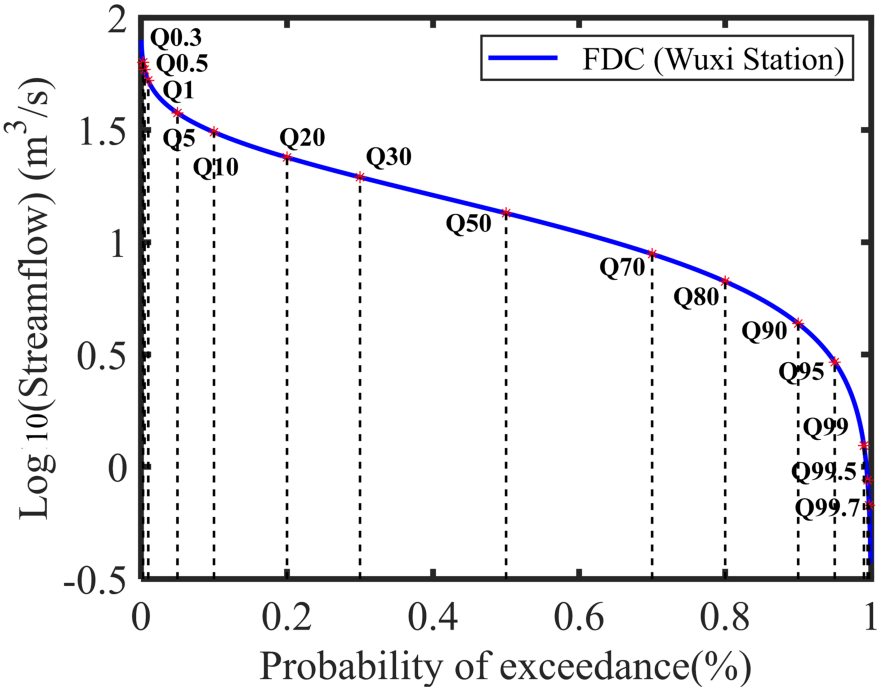
Yu Zhou, Y.Z., 2023. Physical controls of regional distribution patterns of precipitation and flow duration curves in the middle and lower reaches of the Yangtze River. Authorea.

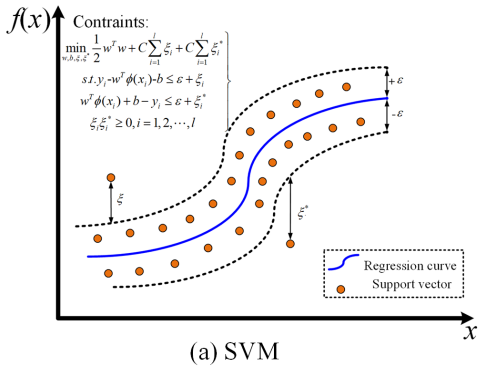
Zhou, J. et al., 2021. Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. Engineering Applications of Artificial Intelligence, 97: 104015.



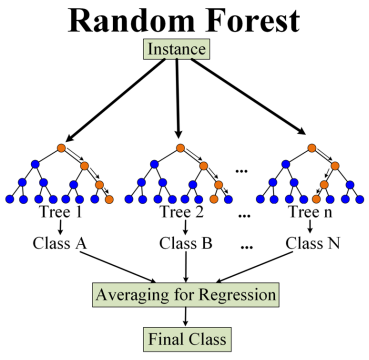




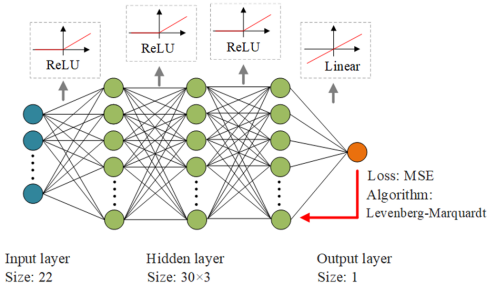




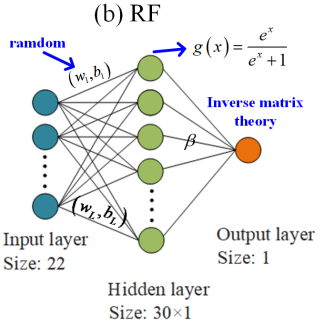
(a) SVM



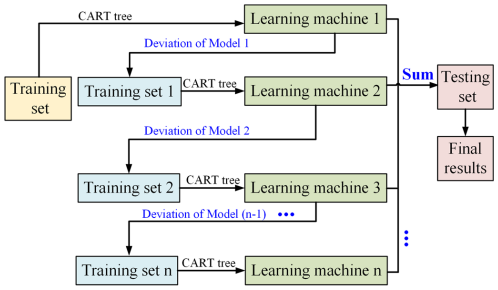
(b) RF



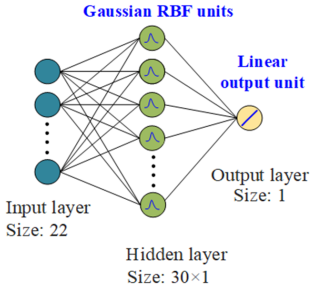
(c) BPNN



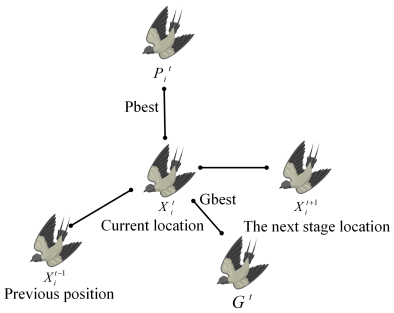
(d) ELM



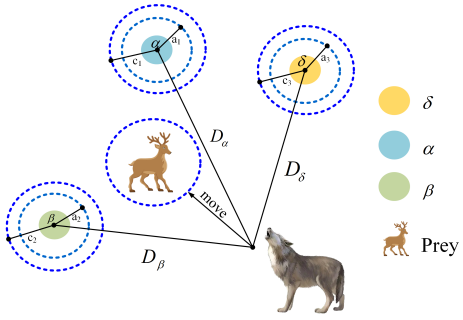
(e) XGB



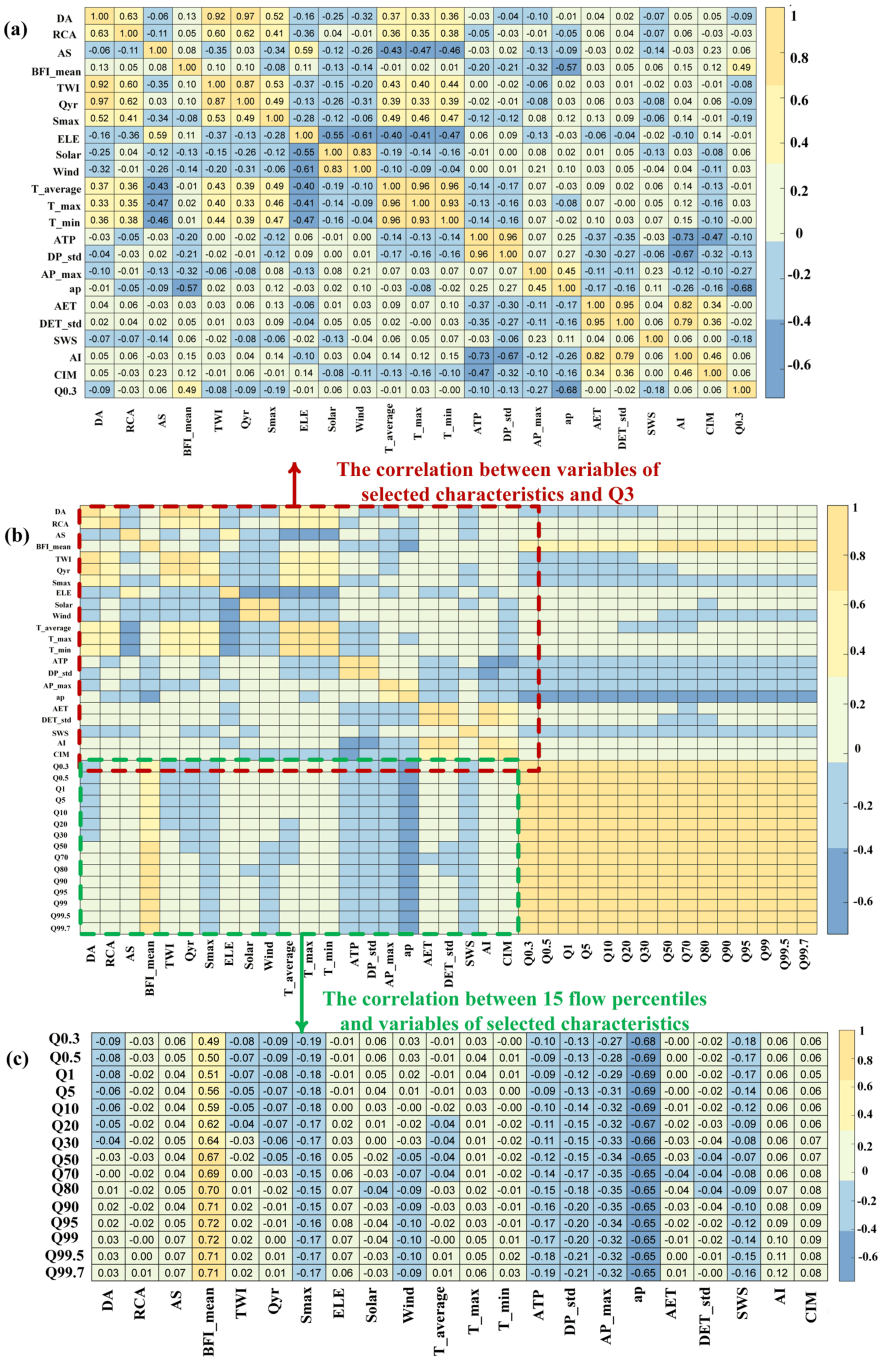
(f) RBF

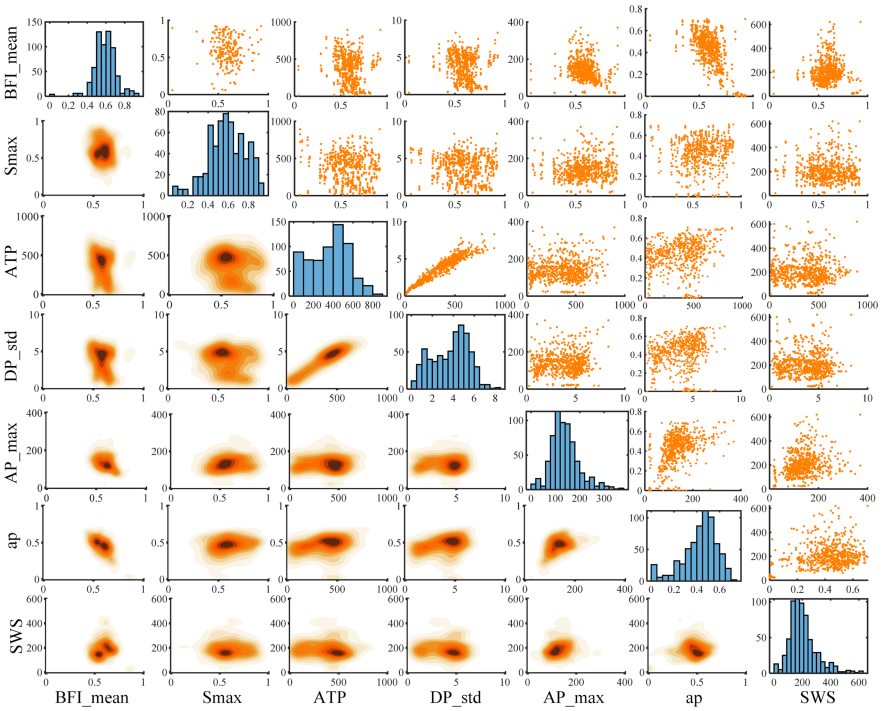


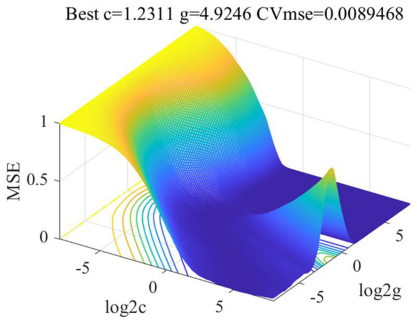
(a) PSO



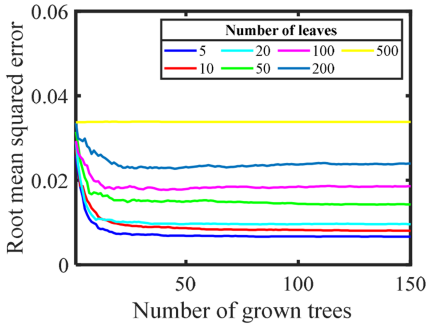
(b) GWO



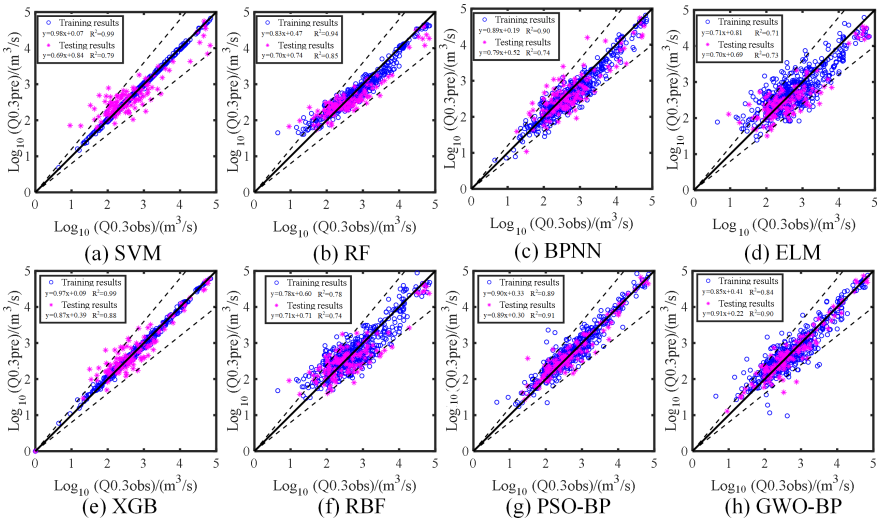


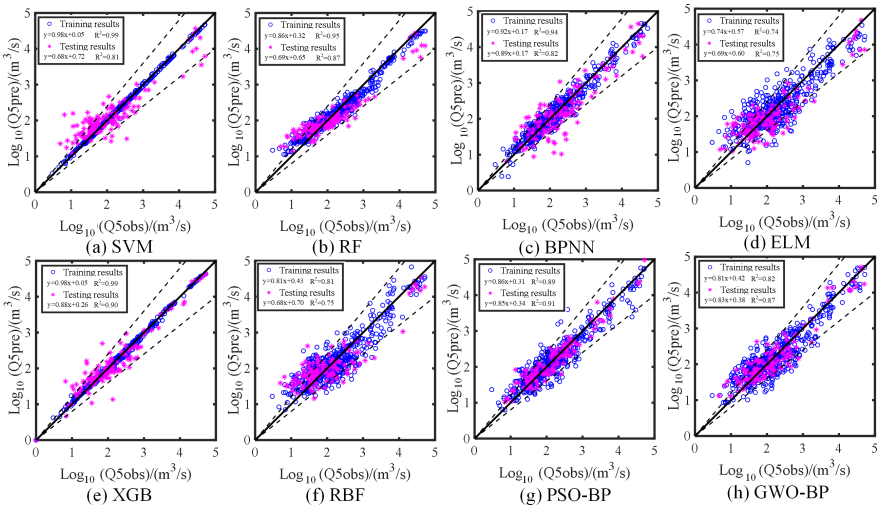


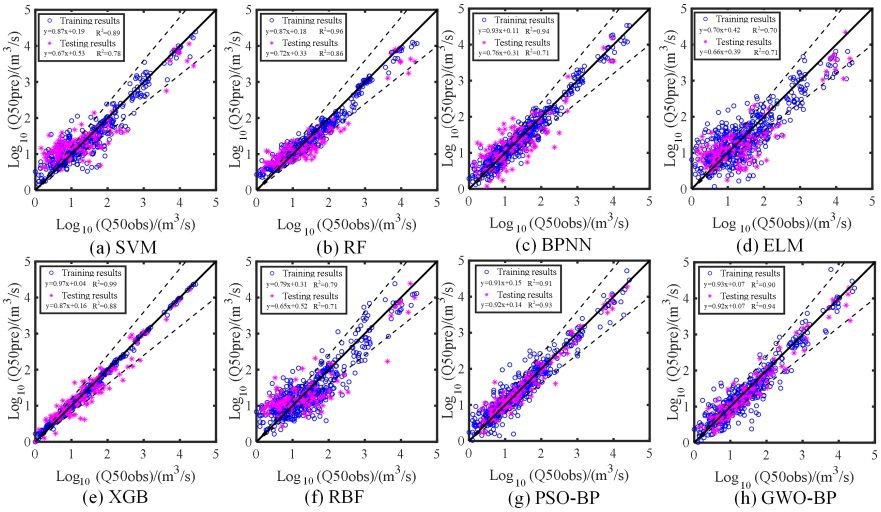
(a) SVM

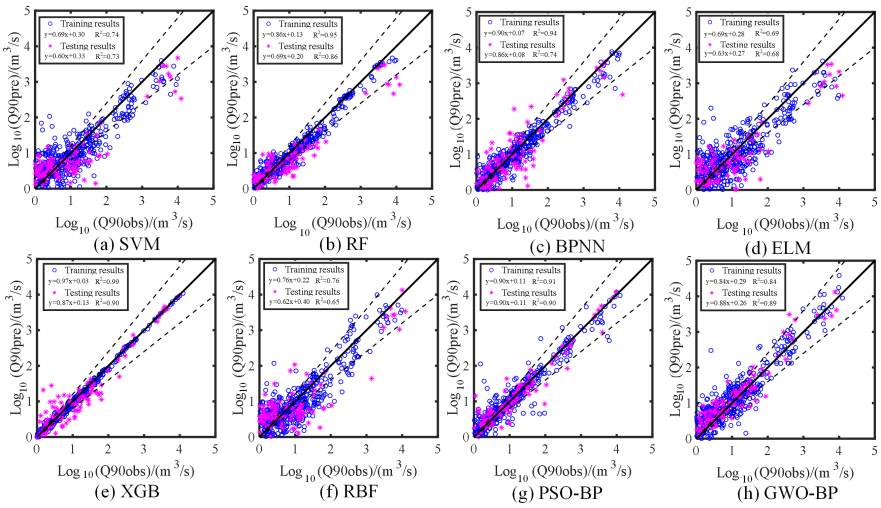


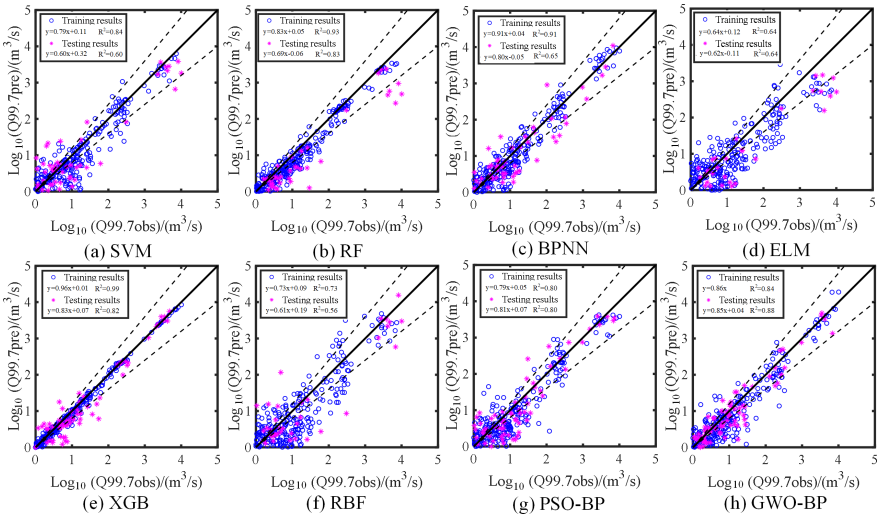
(b) RF

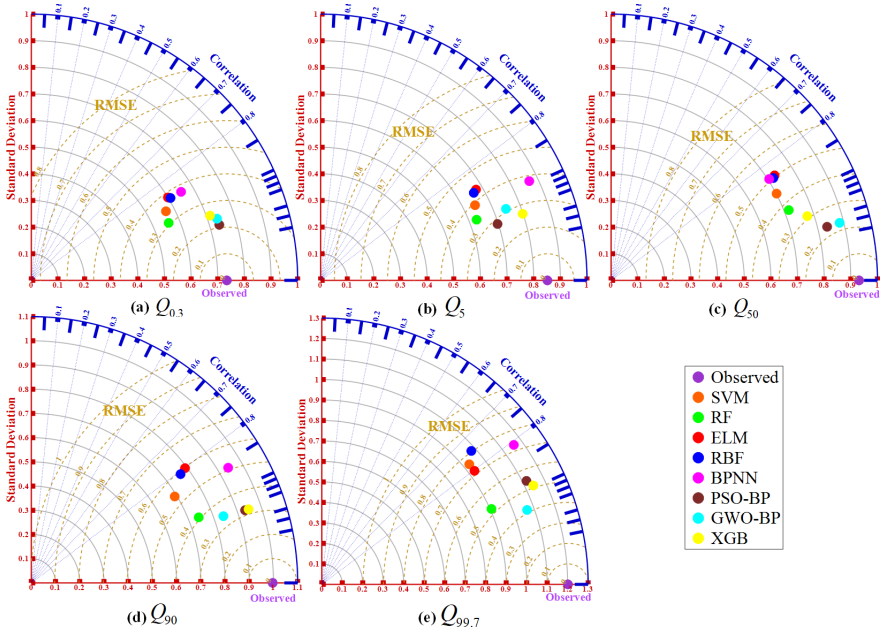


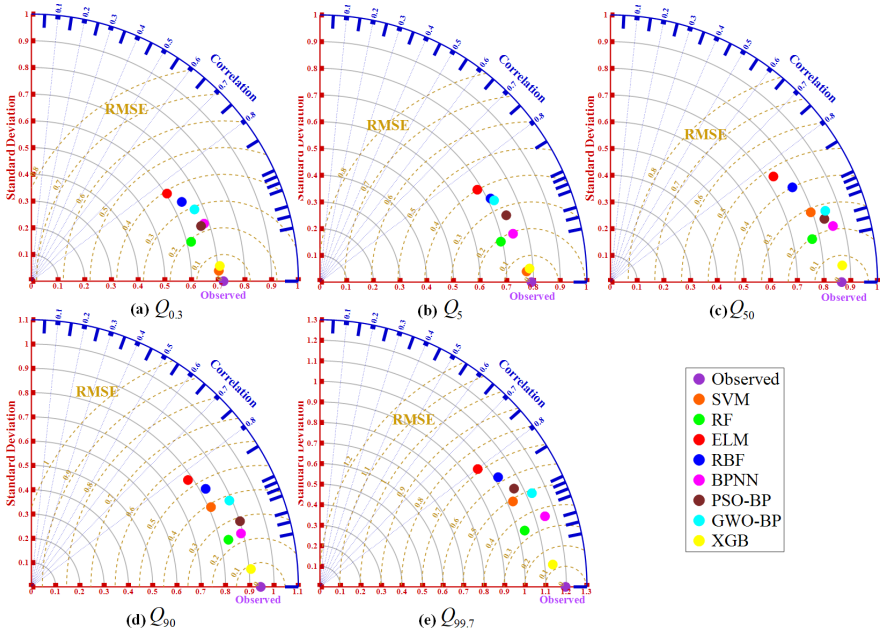


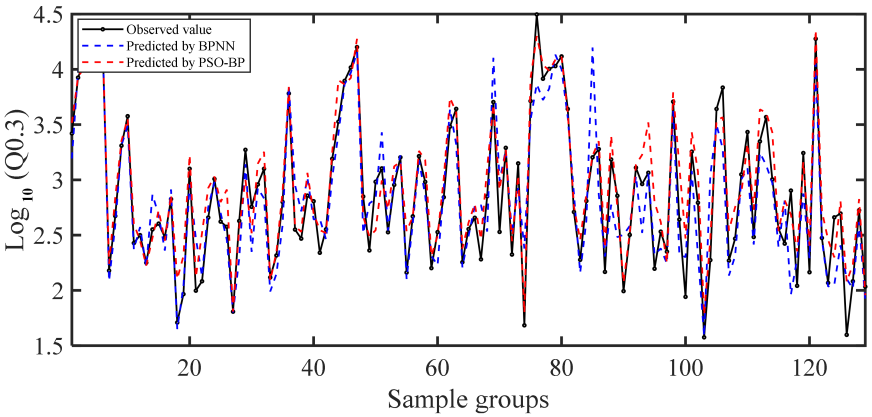




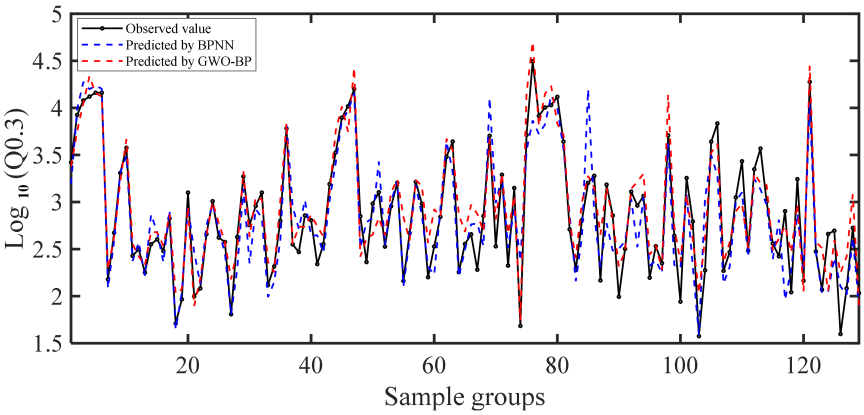




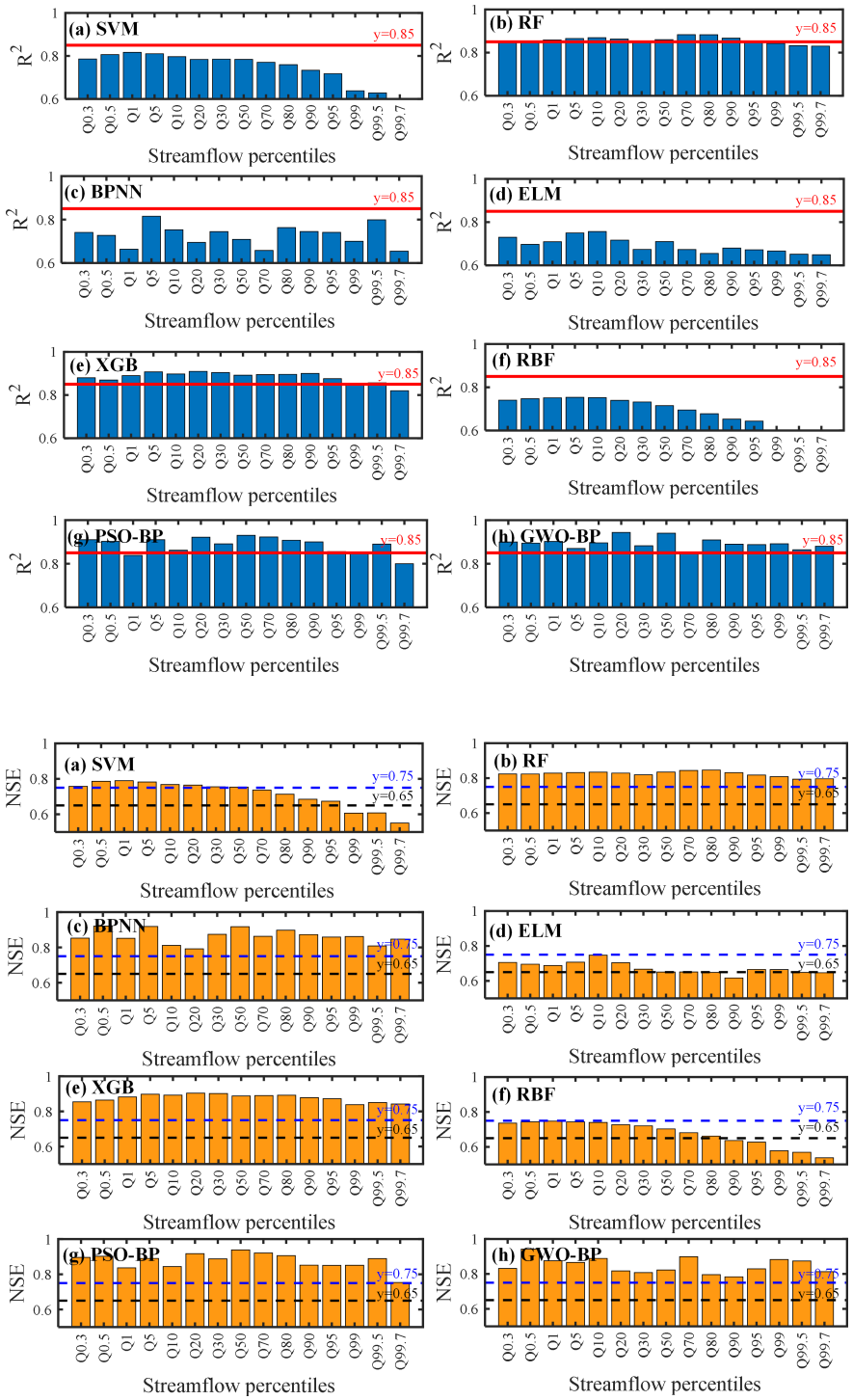


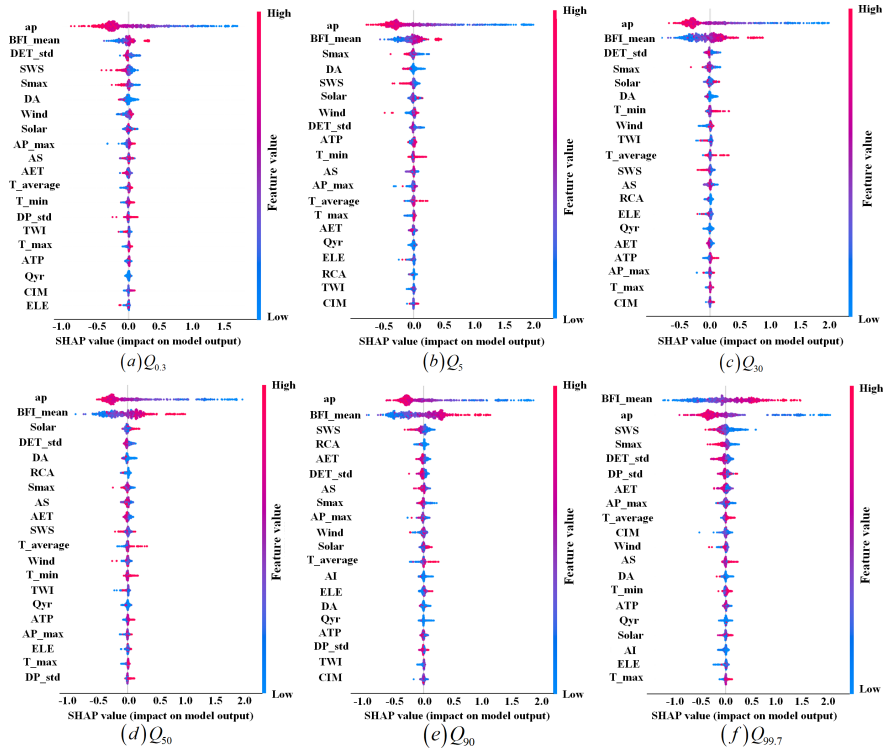


(a) Observed data and predicted value obtained by BPNN,PSO-BP



(b) Observed data and predicted value obtained by BPNN,GWO-BP





Hosted file

`table.docx` available at https://authorea.com/users/614962/articles/688423-fdc-prediction-and-inference-insights-from-the-fusion-of-machine-learning-methods-and-basin-characteristic-factors