https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions

Multi-step ahead daily inflow forecasting using ERA-Interim reanalysis dataset based on gradient boosting regression trees

Shengli Liao[1], Zhanwei Liu[1], Benxi Liu[1], Chuntian Cheng[1], Xinfeng Jin[1], Zhipeng Zhao[1]

[1]Institute of Hydropower System and Hydroinformatics, Dalian University of Technology, Dalian 116024, China

Correspondence to: Zhanwei Liu (337891617@qq.com)

Abstract. Inflow forecasting plays an essential role in reservoir management and operation. The impacts of climate change and human activities make accurate inflow prediction increasingly difficult, especially for longer lead times. In this study, a new hybrid inflow forecast framework with ERA-Interim reanalysis data as input, adopting gradient boosting regression trees (GBRT) and the maximum information coefficient (MIC) was developed for multi-step ahead daily inflow forecasting. Firstly, the ERA-Interim reanalysis dataset provides enough information for the framework to discover inflow for longer lead times. Secondly, MIC can identify effective feature subset from massive features that significantly affects inflow so that the framework can avoid over-fitting, distinguish key attributes with unimportant ones and provide a concise understanding of inflow. Lastly, the GBRT is a prediction model in the form of an ensemble of decision trees and has a strong ability to capture nonlinear relationships between input and output in long lead times more fully. The Xiaowan hydropower station located in Yunnan Province, China is selected as the study area. Four evaluation criteria, the mean absolute error (MAE), the root mean square error (RMSE), the Nash-Sutcliffe efficiency coefficient (NSE) and the Pearson correlation coefficient (CORR), were used to evaluate the established models using historical daily inflow data (1/1/2017-31/12/2018). Performance of the presented framework was compared to that of artificial neural networks (ANN), support vector regression (SVR) and multiple linear regression (MLR) models. The experimental results indicate that the developed method generally performs better than other models and significantly improves the accuracy of inflow forecasting at lead times of 5-10 days. The reanalysis data also enhances the accuracy of inflow forecasting except for forecasts that are one-day ahead.

Keywords Inflow forecasting, Gradient boosting, Regression trees, Maximum information coefficient, ERA-Interim

1 Introduction

Reliable and accurate inflow forecasting 1-10 days in advance is significant for efficient utilization of water resources, reservoir operation and flood control, especially in areas with concentrated rainfall. Rainfall in southern China is usually concentrated for several days at a time due to strong convective weather, such as typhoons. Low accuracy inflow predictions can easily cause the failure of power stations to make reasonable power generation plans 7-10 days ahead of disaster events and lead to unnecessary water abandonment and even substantial economic losses. Fig. 1 shows the loss of electric quantity due to discarded water (LEQDW) in Yunnan and Sichuan Provinces, China from 2011 to 2016. The total amount of LEQDW in Yunnan and Sichuan Provinces increased from 1.5 billion kWh to 47.9 billion kWh from 2011 to 2016, with an average annual growth rate of 99.9%. In recent years, due to the increased number of hydropower station and installed hydropower capacity, the problem of discarding water caused by inaccurate inflow forecasting is becoming increasingly serious, which has also had a negative impact on the development of hydropower in China.

The main challenges in inflow forecasting caused by climate change and human activities at present are low accuracy, especially for longer lead times (Badrzadeh et al., 2013; El-Shafie et al., 2007). To address the problem, a variety of models and approaches

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019


Hydrology and Earth System Sciences
Discussions

have been developed. These approaches can be divided into three categories: statistical methods (Valipour et al., 2013), physical methods (Duan et al., 1992; Wang et al., 2011; Robertson et al., 2013), and machine learning methods (Chau et al., 2005; Liu et al., 2015; Rajaee et al., 2019; Zhang et al., 2018). Each method has its own conditions and scope of application. Statistical methods are usually based on historical inflow records and mainly include use of the autoregressive model, the autoregressive moving average (ARMA) model and the autoregressive integrated moving average (ARIMA) model (Lin et al., 2006), which assume that the inflow series is stationary, and the relationship between input and output is simple. However, real inflow series are complex, nonlinear and chaotic disturbances (Dhanya and Kumar, 2011), making it difficult to obtain high-accuracy predictions using statistical models. Physical methods are implemented using theories of inflow generation and confluence, which have clear mechanisms. These methods can reflect the characteristics of catchment but are very strict with initial conditions and input data (Bennett et al., 2016). Meanwhile, these methods are used for flood forecasting have a shorter lead time and cannot be used to acquire long-term forecasting results due to input uncertainty. Machine learning methods, having a strong ability to handle the nonlinear relationship between input and output and recently shown excellent performance in inflow prediction, are widely used for medium and long-term inflow forecasts. In particular, several studies had shown that artificial neural networks (ANN) (Rasouli et al., 2012; Cheng et al., 2015; El-Shafie and Noureldin, 2011) and support vector regression (SVR) (Tongal and Booij, 2018; Luo et al., 2019) are the two powerful models for inflow predicting. However, these models still have some inherent disadvantages. For example, ANN is prone to being trapped by local minima, and both ANN and SVR suffer from over-fitting problems and reduced generalizing performance. Recent years, gradient boosting regression trees(GBRT) (Fienen et al., 2018; Friedman, 2001), a nonparametric machine learning method based on a boosting strategy and decision trees, was developed and had been used in traffic (Zhan et al., 2019) and environmental (Wei et al., 2019) field and proved to alleviate these problems mentioned above. Thus, GBRT were selected for daily inflow prediction with a lead time of 1-10 days in this paper. Compared with ANN and SVR, GBRT also has two other advantages. Firstly, GBRT can rank features according to their contribution to model scores, which is of great significance for reducing the complexity of the model. Secondly, GBRT is a white box model and can be easily interpreted. To the best of our knowledge, GBRT has not been used for daily inflow prediction with a lead time of 1-10 days before. For comparison purposes, ANN and SVR have also been employed to forecast daily inflow, and multiple linear regression (MLR) was used as a benchmark.

In addition to forecasting models, a vital reason why many approaches cannot attain higher accuracy for inflow predictions is that inflow is influenced by various factors (Yang et al., 2019), such as rainfall, temperature, humidity, pressure, dew point, etc. Thus, it is very difficult to select appropriate features for inflow forecasting. Current feature selection methods for inflow forecasting mainly include the correlation coefficient method (Badrzadeh et al., 2013; Siqueira et al., 2018; Pal et al., 2013), the stepwise selection method (Wei, 2016), and the Gamma test method (Chang and Tsai, 2016), etc. These methods have limited ability for capturing nonlinear relationships or tend to need much more computation resource. In order to select effective input factors accurately and quickly, the maximum information coefficient (MIC) (Reshef et al., 2011), was used to select input factors for inflow forecasting. MIC is a robust measure of the degree of correlation between two variables and has attracted a lot attention from academia (Zhao et al., 2013; Ge et al., 2016; Lyu et al., 2017; Sun et al., 2018). In addition, sufficient potential input factors are the prerequisite for obtaining reliable and accurate prediction results and it is not enough to use only antecedent inflow values as the input of the model. To enhance the accuracy of inflow forecasting and acquiring a longer lead time, increasing amounts of meteorological forecasting data are being used for inflow forecasting (Lima et al., 2017; Fan et al., 2015; Rasouli et al., 2012). However, with extended lead times, the errors of forecast data continuously increase because the variables obtained by numerical weather prediction (NWP) system are also affected by complex factors (Mehr et al., 2019). Moreover, with the continuous improvement of forecasting systems, it is difficult to obtain consistent, long series of forecasting data

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions

(Verkade et al., 2013). To mitigate these problems, the reanalysis data generated by ECMWF (ERA-Interim) (Dee et al., 2011), which was proved to be one of the best methods for reanalysis of data describing atmospheric circulation and elements (Kishore et al., 2011), has been used as an input. The reanalysis data are the result of assimilating observed data with forecast data, which has less error than observed data and forecast data. ERA-Interim shows the results of a global climate reanalysis from 1979 to date, which are produced by a fixed version of a NWP system. The fixed version ensures there are no spurious trends caused by an evolving NWP system. Therefore, meteorological reanalysis data satisfies the need for long sequences of consistent data and have been used for the prediction of wind speeds (Stopa and Cheung, 2014) and solar radiation (Ghimire et al., 2019; Linares-Rodr ́guez et al., 2011).

This study aims to provide a reliable inflow forecasting framework with a longer lead time for daily inflow forecasting. The framework adopts the ERA-Interim reanalysis dataset as the input which ensured ample information was supply to depict inflow. MIC was used to select appropriate features so that avoiding over-fitting and waste of computing resources caused by feature redundancy. GBRT, which is robust to outliers and has strong non-linear fitting ability, was used as the prediction model to improve inflow forecasting accuracy of longer lead times.

This paper is organized as follows: Section 2 describes a case study and collected data, Section 3 introduces the theory and process of methods used, including MIC and GBRT. Section 4 shows the results and discussion of the data, followed by the conclusions in Section 5.

2 Data

2.1 Study area and collected data

The Xiaowan Hydropower Station in the lower reaches of the Lancang River was chosen as the study site (Fig. 2). The Lancang River is approximately 2000 km long and has a drainage area of 113300 km^2 above the Xiaowan Hydropower Station. The Lancang River originates in the Tibetan Plateau and runs through China, Myanmar, Laos, Thailand, Cambodia, and Vietnam and is also known as the Mekong River. The major source of water flowing into the Lancang River in China comes from melting snow on the Tibetan plateau (Commission Mekong River, 2005).

We collected EAR-Interim reanalysis dataset, observed daily inflow and rainfall data for Xiaowan for 8 years (January 2011 to December 2018). Fig. 3 depicts the daily inflow series. The data from January 2011 to December 2014 (1461 days, approximately 50% of the whole dataset), from January 2015 to December 2016 (731 days, approximately 25% of the whole dataset) and from January 2017 to December 2018 (730 days, approximately 25% of the whole dataset) were used for training, validation and testing, respectively. The reanalysis dataset can be downloaded from *https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/* and is provided every six hours on a spatial grid size of $0.75° \times 0.75°$. According to the physical meaning of the variables, the near-surface 79 variables from the reanalysis data are considered as potential selected predictors for inflow forecasting, which include the total precipitation (tp), the 2 meter temperature (t2m), the total column water (tcw), etc..

2.2 Data scaling and feature selection

Feature scaling is necessary for machine learning methods and all features were scaled to the range between 0 and 1 before taking part in the calculation, as follows:

$$x_{scale} = \frac{x_{original} - x_{min}}{x_{max} - x_{min}} \quad (1)$$

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019


Hydrology and Earth System Sciences
Discussions

where x_{scale} and $x_{original}$ indicate the scaled and original data, respectively and x_{max} and x_{min} represent the maximum and minimum of inflow series, respectively.

Reasonable selection of input variables can accelerate the calculation speed and improve the prediction accuracy of the model by removing redundant feature information and reducing the dimensions of the features. If too many features are selected, model will become very complex, which will cause trouble when adjusting parameters, resulting in over-fitting and difficult convergence. Moreover, natural patterns in the data will be blurred by noise (Zhao et al., 2013). On the other hand, if too few features are chosen, there will be not enough information for inflow forecasting. After eliminating invalid variables, MIC was employed to select inputs from 79 potential predictors from reanalysis data and observed inflow and rainfall lags are identified by autocorrelation function (ACF).

2 Methodology

3.1 Feature selection via maximal information coefficient

The calculation of MIC is based on concepts of the mutual information (MI) (Kinney and Atwal, 2014). For a random variable X, such as observed inflow, the entropy of X is defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{2}$$

where $p(x)$ is the probability density function of X = *x*. Furthermore, for another random variable Y, such as observed rainfall, the conditional entropy of X given Y may be evaluated from the following expression

$$H(X \mid Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x \mid y) \tag{3}$$

where $H(X \mid Y)$ is the uncertainty of X given knowledge, $p(x, y)$ and $p(x \mid y)$ are the joint probability density and the conditional probability of X = *x* and Y = *y*, respectively. The reduction of the original uncertainty of X, due to the knowledge of Y, is called the MI (Amorocho and Espildora, 1973; Chapman, 1986), defined by

$$MI(X,Y) = H(X) - H(X \mid Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x) p(y)} \tag{4}$$

The calculation of MIC is divided into three steps. Consider given a dataset D, including variable X and Y with a sample size *n*. Firstly, drawing scatter plots of X and Y and drawing grids for partitioning which is called an *x*-by-*y* grid. Let D|G denote the distribution of D divided by one of x-by-y grids as G. $MI^*(D, x, y) = \max MI(D \mid G)$, where $MI(D \mid G)$ is the mutual information of D|G. Secondly, characteristic matrix is defined as

$$M(D)_{x,y} = \frac{MI^*(D, X, Y)}{\log(\min(x, y))} \tag{5}$$

Lastly, MIC is introduced as the maximum value of characteristic matrix, that is, $MIC(D) = \max_{xy<B(n)} M(D)_{x,y}$, where B(*n*) is the upper bound of the grid size need to be considered.

We perform feature selection in two steps via MIC. First, compute MIC value of each reanalysis factors and observed inflow. Then, sort features based on MIC in a descending order and select the top *k* features according to the set threshold. The selected *k* features from reanalysis data are used as part of input to the model.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

3.2 Gradient boosting regression trees

Gradient boosting regression trees is an ensemble model which mainly includes two algorithms: decision tree algorithm and the boosting algorithm. The decision tree robust to outliers is used as a primitive model and boosting algorithm as integration rule is used to improve inflow forecasting accuracy.

3.3.1 The decision tree

The decision tree in this paper refers to decision tree learning used in computer science, which is one of the predictive modeling approaches used in machine learning. A decision tree consists of branch nodes (the tree structure) and leaf nodes (the tree output). Supposing a training dataset is given in a feature space with n features and each feature with N samples, $\{(X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\}$ ($X_i = (x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, N$). In the input space where the training set is located, each region is recursively divided into two subregions and the output value of each subregion is used to construct a binary decision tree. The top-down cyclic branch learning of the decision tree adopts a greedy algorithm where each branch node only cares about its own objective function. By traversing all features and all segmentation points of each feature, the best feature j and segmentation points s can be found by minimizing square loss:

$$\min_{j,s}\left[\min_{c_1}\sum_{X_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{X_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{6}$$

where

$$\begin{aligned} R_1(i,s) &= \left\{X_i \mid x^{(j)} \leqslant s, i=1,2,\cdots,N\right\} \\ R_2(i,s) &= \left\{X_i \mid x^{(j)} > s, i=1,2,\cdots,N\right\} \end{aligned} \tag{7}$$

$$c_m = \frac{1}{N_m}\sum_{X_i\in R_m(j,s)} y_i \quad (m=1,2) \tag{8}$$

y_i is the observed value and $R_1(j, s)$ and $R_2(j, s)$ are the results of partitioning. c_1 and c_2 are output values of $R_1(j, s)$ and $R_2(j, s)$, respectively. Fig. 4 shows an example of a decision tree model with a max depth and number of leaf nodes of 3 and 5, respectively. If the threshold of loss is set as the stopping condition of the decision tree, it will easily lead to over-fitting problems. Hence, we set the following parameters to alleviate the over-fitting problem of the decision tree model: the maximum depth of the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node and the number of leaf nodes. These parameters are also the ones used for optimization when using the decision tree.

3.3.2 The boosting algorithm

The idea of gradient boosting originated in the observation by Breiman (Breiman, 1997) and can be interpreted as an optimization algorithm based on a suitable cost function. Explicit regression gradient boosting algorithms were subsequently developed (Friedman, 2001; Mason et al., 2000). The boosting algorithm used is described here. Supposing a training dataset with N sample $\{(X_1,y_1),(X_2,y_2), \cdots; (X_N,y_N)\}$, a square loss function is used to train the decision tree:

$$L(y, f(X)) = \sum_{i=1}^{N}(y - f(X_i))^2 \tag{9}$$

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

The core of the GBRT algorithm is the iterative process of training the decision with a residual method. The iterative training process of GBRT with *M* decision trees is as follows:

1) Initialization $f_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c)$.

2) For *m*-th (m=1, 2, ..., *M*) decision trees:

a) Operating *i*-th (i=1, 2, ..., *N*) sample points. Using the negative gradient of the loss function to replace the residual in the current model $r_{mi} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$.

b) Fitting a regression tree with $\{(x_i, r_{mi})\}$. The *i*-th regression tree with R_{mt} (t = 1, 2, ..., T) as its corresponding leaf node region is obtained, where *t* is the number of leaf nodes of regression.

c) For each leaf region *t* = 1, 2, ..., *T*, and the best fitting value is calculated by $c_{mt} = \arg\min_c \sum_{x_i \in R_{mt}} L(y_i, f_{m-1}(x_i) + c)$.

d) The fitting results are updated by adding the obtained fitting values to the previous ones using $f_{mt}(x_i) = f_{m-1}(x_i) + \sum_{t=1}^{T} c_{mt} I \quad (x_i \in R_{mt})$.

3) Finally, a strong learning method is obtained $\hat{f}(x_i) = f_M(x_i) = \sum_{m=1}^{M}\sum_{t=1}^{T} c_{mt} I \quad (x_i \in R_{mt})$.

According to the above introduction to GBRT, the parameters of the GBRT can be divided into two categories: boosting parameters and decision tree parameters. The boosting parameters include the learning rate and the number of weak learners (*n_estimators*). The learning rate setting is used for reducing the gradient step. The learning rate influences the overall time of training, and the smaller the value is, the more iterations are required for training. There are four tree parameters: *max_leaf_nodes*, *min_samples_leaf*, *min_samples_split* and *max_depth*. Hence, GBRT has six parameters control model complexity (Fienen et al., 2018), five of which we adjusted for tuning, except for learning rate determined by trial in advance.

3.3 Evaluation criteria of the models

It is critical to carefully define the meaning of performance and to evaluate the performance on the basis of the forecasting and fitted values of the model compared with historical data. The root mean square error (RMSE) and mean absolute error (MAE) are the most commonly used criteria to assess model performance and are calculated using Eq. (10) and Eq. (11), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Q}_i - Q_i)^2} \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Q}_i - Q_i\right| \tag{11}$$

where $\hat{Q}_i$ and Q_i are the inflow estimation and observed value at time *i*, respectively and *n* is the number of samples. The RMSE is more sensitive to extremes in sample sets and thus it was used to evaluate the model's ability to simulate flood peaks.

The Nash-Sutcliffe efficiency coefficient (NSE) (Nash and Sutcliffe, 1970) is commonly for evaluating the performance of hydrological models and it is one of the best performance metrics for reflecting the overall fit of a hydrograph. The NSE is calculated using Eq. (12)

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

$$NSE = 1 - \frac{\sum_{i=1}^{n}(\hat{Q}_i - Q_i)^2}{\sum_{i=1}^{n}(Q_i - \bar{Q})^2} \tag{12}$$

where $\bar{Q}$ is the mean of the observed values. An NSE = 1 means the model is perfect and an NSE < 0 reflects that the model forecasts are a worse estimation than the mean value of observed values.

The Pearson correlation coefficient (CORR) is a good measurement of the average error which is calculated according to Eq. (13).

$$CORR = \frac{\sum_{i=1}^{n}(Q_i - \bar{Q})(\hat{Q}_i - \bar{\hat{Q}})}{\sqrt{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}\sqrt{\sum_{i=1}^{n}(\hat{Q}_i - \bar{\hat{Q}})^2}} \tag{13}$$

where $\bar{\hat{Q}}$ is the mean of the estimation values. The range of the CORR is between 0 and 1 and values close to 1 demonstrate a perfect estimation result.

3.4 Overview of framework

Fig.5 illustrates the overall structure of framework presented. This structure consists of two major models: GBRT and GBRT-MIC.

In GBRT, we measure the relevance of different lags observed inflow and rainfall with observed inflow at the time of forecast via autocorrelation function (ACF) (Badrzadeh et al., 2013) and select appropriate lags as predictors of model by the set threshold. Then, data preprocessing and scaling were carried out for selected predictors. Next, dividing the dataset into train set, validation set, and test set according to the length of each data set specified in advance (in Section 2.2). A grid search algorithm was guided to optimization model parameters by evaluation of validation set for each lead time (Chicco and Davide, 2017). Lastly, prediction results are evaluated based on test set. Compared with GBRT, GBRT-MIC adds reanalysis data which were selected via MIC (in Section 3.1) as the input of the model. Moreover, GBRT-MIC also calculates the importance of features according to the prediction results and ranks the features. The model structures of GBRT and GBRT-MIC are as follows:

$$\hat{Q}^{I}_{t+T} = f(\vec{\theta}^{I}_{t}; Q_t, Q_{t-1}, ..., Q_{t+1-p}, R_t, R_{t-1}, ..., R_{t+1-q}) \ (T = 1, 2, ..., 10) \tag{14}$$

$$\hat{Q}^{II}_{t+T} = f(\vec{\theta}^{II}_{t}; Q_t, Q_{t-1}, ..., Q_{t+1-p}, R_t, R_{t-1}, ..., R_{t+1-q}, E^1_t, E^2_t, ..., E^k_t) \ (T = 1, 2, ..., 10) \tag{15}$$

where $\hat{Q}^{I}_{t+T}$ and $\hat{Q}^{II}_{t+T}$ are the forecasted value of GBRT and GBRT-MIC at the lead time T of current time t, respectively. $\vec{\theta}^{I}_{t}$ and $\vec{\theta}^{II}_{t}$ are parameters of GBRT and GBRT-MIC at the lead time T of current time t, respectively. p and q are lags of observed inflow and rainfall determined via ACF, respectively. E_t is the features from reanalysis data at the current time t and k is the number of features from reanalysis data determined via MIC.

4. Experimental results and discussion

In order to compare with GBRT-MIC, the ANN-MIC, SVR-MIC and MLR-MIC, obtained by replacing GBRT in the framework with ANN, SVR and MLR, respectively, were also employed for inflow forecasting with lead times of 1-10 days. As mentioned

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019


Hydrology and Earth System Sciences
Discussions

earlier, the RMSE, MAE, NSE and CORR of each model were calculated to compare with the performance of models based on the test set. We also explored the feature importance based on the GBRT-MIC model.

4.1 Feature selection

Table 1 shows the ACF of a lag of 1-10 days of inflow and rainfall. The results show the ACF decreases as the lag order increases. Considering the problems of over-fitting and computational time caused by large amounts of input data, thresholds of inflow and rainfall are set as 0.931, 0.226, respectively and threshold of MIC is set as 0.625. And thus a lag of one day of 10 predictors from the reanalysis data and a lag of 1-3 days of inflow and rainfall were selected as the model inputs, that is k=10, p=3 and q=3. Simultaneously, according to the causes of inflow, the total precipitation was also selected as a predictor.

Finally, a total of 17 variables including 6 observed variables and 11 reanalysis variables were selected as the model inputs (Table 2). No. 7 to 17 are reanalysis variables and the range of MIC of the reanalysis variables is 0.625 to 0.853 except for the total precipitation. No. 7 to 10, No. 15 and No. 16 are variables related to solar radiation or temperature that are used to represent the melting of snow. Albedo (No. 7) is a measure of the reflectivity of the Earth's surface. Typically, the highest Albedo is found on snow and ice, followed by land and the lowest in the ocean. Albedo has the highest correlation (negative correlation) with inflow in terms of MIC. Soil temperature level 3 (No. 8) is the temperature of the soil in layer 3 (28-100 cm, the surface is at 0 cm). The temperature of the snow layer (No. 15) gives the temperature of the snow layer from the ground to the snow-air interface. Surface thermal radiation downwards (No. 16) is the amount of thermal (also known as longwave or terrestrial) radiation emitted by the atmosphere and clouds that reaches the Earth's surface. Total column water (No. 11) is the sum of water vapor, liquid water, cloud ice, rain and snow in a column extending from the surface of the Earth to the top of the atmosphere. The total column water vapor (No. 12) is only the total amount of water vapor, which is a fraction of the total column water. Runoff (No. 13) is a measure of the availability of water in the soil, including at the surface (surface runoff) and under the ground (subsurface runoff). Volumetric soil water layer 3 (No. 14) is the volume of water in soil layer 3. Total precipitation (No. 17) is the accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface. In summary, all the selected predictors are interpretable and have a good physical connection with inflow.

4.2 Hyperparameter optimization

For machine learning methods, hyperparameters are parameters that are set before starting the learning process, rather than parameters obtained through training. In order to improve the learning performance, it is imperative to tune the hyperparameters of models. Grid searching was employed to tune the hyperparameters of GBRT, GBRT-MIC, ANN-MIC and Bayesian optimization (Snoek et al., 2012) was employed to tune the hyperparameters of SVR-MIC. All computations of this paper are performed in the version 3.7.10 of Python, using the scikit-learn package (Pedregosa et al., 2011).

Besides the training algorithms and the number of hidden layers of the ANN, there are also two parameters controlling the model structure which need to be adjusted. A range of 2-20 neurons and four activation functions (Table 3) are selected by grid searching. To alleviate the influence of random initialization of weights, 50 ANN-MIC models were trained for each combination of activation and the number of hidden neurons. We obtained activation functions and the number of hidden neurons by selecting the optimal median of the MAE of the validation set for each lead time. The results of the trials show identity is most robust activation function (Fig. 6) and less number of nodes is inclined to obtain lower error for except for relu. The optimal combinations of number of hidden layer and activation function for each lead time are listed in Table 4. It is found that the optimal number of neurons is 2 or 3 and the optimal activation function is either logistic or tanh.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019


Hydrology and Earth System Sciences
Discussions

For SVR, there are three parameters that need to be optimized, the regularization parameter C, the loss function tolerance threshold ϵ and the parameter of RBF γ. The search space of ϵ, C and γ are [0, 2], [e^{-5}, e^{5}], [e^{-13}, e^{-1}], respectively. The tuning parameters for each lead time are listed in Table 4 and the range of C, ϵ and γ are 0.51 to 144.48, 0.0004 to 0.0085 and 0.0012 to 0.2150, respectively.

According to trial and error, the learning rate of GBRT and GBRT-MIC are all decided to be 0.01. For parameters *max_depth*, *max_leaf_nodes*, *min_samples_leaf* values between 3 and 9 with an increment of 2 were evaluated, for *min_samples_split* values between 10 and 50 with an increment of 10 were evaluated, and *n_estimators* between 200 and 1200 with an increment of 50 were evaluated. We developed 13440 models for each lead time and Table 5 lists the parameters selected for GBRT and GBRT-MIC.

4.3 Inputs comparison

Fig. 7 illustrates the relationship between performance indices and lead times of GBRT and GBRT-MIC in the test set (2017-2018). It is obvious that the use of reanalysis data selected by MIC as model inputs makes a great improvement on the GBRT forecasting. For Fig. 7(a), the MAE of GBRT-MIC decreases from 171 to 167, a decrease of 2.4% for two-day-ahead forecasting and decreases from 270 to 241, a decrease increasing to 10.7% for ten-day-ahead forecasting compared with GBRT. For Fig. 7(b), the RMSE of GBRT-MIC achieves 4.4% and 9.6% reduction for two and ten-day-ahead forecasting, respectively, compared with GBRT. For Figs. 7(c) and 7(d), the NSE and CORR of GBRT-MIC increase by 1.1%, 1.1% for two-day-ahead forecasting and 5.2%, 2.3% for ten-day-ahead forecasting, respectively. Thus, the above analysis indicates the improvements increase as the lead time increases and reanalysis data significantly improves the accuracy of inflow prediction for longer lead times in terms of all four different evaluation measures in the test set.

Fig. 8(a) shows the ten-day-ahead forecasted inflow of GBRT-MIC and GBRT versus the observed inflow in the test set. The slope of fitting curve of GBRT-MIC and GBRT are 0.78 and 0.72, respectively, which demonstrates that GBRT-MIC can obtain more accuracy inflow forecasting than GBRT. Fig. 8(b) illustrates the distribution of the forecast errors of GBRT and GBRT-MIC. The results show the prediction error of two models approximate to normal distribution. It demonstrates that the prediction error contains less information that is not extracted by the model and more errors of forecasted inflow concentrate at 0 around by GBRT-MIC than GBRT. Fig. 8(c) provides the ten-day-ahead forecasted results of GBRT-MIC and GBRT in the test set. It can be seen that forecast error of peak value of GBRT-MIC significantly decreases than GBRT. This reveals that the problem of inaccurate peak flow prediction arisen in in areas with concentrated rainfall for the GBRT model could be mitigated by incorporating the reanalysis data identified by MIC as model inputs.

4.4 Model comparison

GBRT-MIC, SVR-MIC, ANN-MIC with obtained optimal model parameters were employed for inflow forecasting of one to ten-day-ahead. Summarized results for train, validation and test set are presented in Table 6, Table 7 and Table 8, respectively. Notably, the ANN-MIC models were trained 50 times for each lead time, and the median of the performance indices are listed in these tables. It is clear from Table 6 that the GBRT-MIC are more efficient in the train set than other models at lead times of 1-10 days, and machine learning models obtain better simulation results than MLR-MIC. However, GBRT-MIC did not always perform best for the test set at lead times of 1-10 days. From Table 8, we notice that SVR-MIC outperforms the other models for a lead time of one day. Furthermore, ANN-MIC performs best in terms of RMSE and NSE at lead times of 2-4 days, and GBRT-MIC performs best for longer lead times (5-10 days). At a lead time of ten days, the NSE of GBRT-MIC even reached 0.8084.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019


Hydrology and Earth System Sciences
Discussions

The relationship between performance indices and lead times of these four models in the test set (2017-2018) is presented in Fig. 9. The results indicate the performance of these four models decreases (higher MAE and RMSE, and lower CORR and NSE) as the lead time increases and the four models perform equally well for one- to four-day-ahead forecasting, whereas significant differences among their performances are found as the forecasting lead time exceeds four days. It clearly indicates that the GBRT produce much higher CORR and NSE, and lower MAE and RMSE than the other three models for five to ten-day-ahead forecasting except that the SVR perform best in terms of MAE for ten-day-ahead forecasting. As mentioned earlier (Section 4.3), the prediction error of these four models approximate to normal distribution and according to (Chai and R., 2014), the RMSE is more appropriate for representing model performance than MAE when the error distribution is expected to be Gaussian. Thus, GBRT-MIC preforms best for five to ten-day-ahead forecasting.

4.5 Feature importance

A benefit of using gradient boosting is that after the boosted trees are constructed, relative importance scores for each feature can be acquired to estimate the contribution of each feature to inflow forecasting. Fig. 10 shows the feature importance based on GBRT-MIC for lead times of one and ten days. The lagged observed values (e.g., Q_{t-1} and Q_{t-2}) are more important for shorter lead times (Fig. 10(a)), which demonstrates that the observed values are essential for inflow forecasting of shorter lead times.

The features (e.g., $stl3_{t-10}$ and $d2m_{t-10}$) from the reanalysis data have a high relative importance for longer lead times (Fig. 10(b)). Based on the analysis of the concepts of $stl3_{t-10}$ and $d2m_{t-10}$ (Section 4.1), we infer that the temperature near the ground effects the inflow by affecting the melting of snow which is consistent with the fact that the Lancang River is a snow-melt river. Meanwhile, it is found that the reanalysis data provides important information for inflow forecasting at longer lead times.

5. Conclusion

In this study, GBRT-MIC are employed to make inflow forecasts for lead times of 1-10 days and ANN-MIC, SVR-MIC and MLR-MIC are developed to compare with GBRT-MIC. The reanalysis data selected by MIC, the antecedent inflow and the rainfall records selected by ACF are used as predictors to drive the models. These models are compared using four evaluation criteria, the RMSE, MAE, NSE and CORR. It is shown that GBRT-MIC, ANN-MIC and SVR-MIC outperform MLR-MIC at lead times of 1-10 days, and GBRT-MIC performs best at lead times of 5-10 days.

According to compare the forecasted results of GBRT and GBRT-MIC, we concluded that GBRT-MIC could be used for more accurate and reliable inflow forecasting for 1-10 day lead times, and the use of reanalysis data selected by MIC as model inputs makes a great improvement on the GBRT forecasting, especially for 5-10 day lead times. In addition, the feature importance achieved by GBRT-MIC demonstrates that soil temperature and dewpoint temperature near the ground significantly affects inflow for longer lead times.

In summary, the developed framework that integrates reanalysis data, MIC and GBRT can well perform inflow forecasting for lead times of 1-10 days. The results of this study are of significance to assist power stations in making power generation plans 7-10 days in advance in order to reduce LEQDW and flood disasters.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Acknowledgements

This research was supported by National Natural Science Foundation of China (No. 51979023, No. U1765103) and the Liaoning province Natural Science Foundation of China (No. 20180550354). We are grateful for reanalysis data provided by European Centre for Medium-Range Weather Forecasts.

Code/Data availability

Request for materials should be addressed to Zhanwei Liu.

Author contributions

S.L. carried out the study design, the analysis and interpretation of data, and drafted the manuscript. Z.L. and B.L. participated in the study design, data collection, analysis of data, and preparation of the manuscript. C.C. and Z.Z. carried out the experimental work and the data collection and interpretation. X.J. participated in the design and coordination of experimental work, and acquisition of data. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no conflict of interest.

References

Amorocho, J., and Espildora, B.: Entropy in the assessment of uncertainty in hydrologic systems and models, Water Resour. Res., 9, 1511-1522, 10.1029/WR009i006p01511, 1973.

Badrzadeh, H., Sarukkalige, R., and Jayawardena, A.: Impact of multi-resolution analysis of artificial intelligence models inputs on multi-step ahead river flow forecasting, J. Hydro., 507, 75-85, 10.1016/j.jhydrol.2013.10.017, 2013.

Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, Water Resour. Res., 52, 8238–8259, 10.1002/2016WR019193, 2016.

Breiman, L.: Arcing the edge, Statistics Department, University of California,Technical Report 486, 1997.

Chai, T., and R., D. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, 1247-1250, 10.5194/gmd-7-1247-2014, 2014.

Chang, F. J., and Tsai, M.-J.: A nonlinear spatio-temporal lumping of radar rainfall for modelling multi-step-ahead inflow forecasts by data-driven techniques, J. Hydro., 535, 256-269, 10.1016/j.jhydrol.2016.01.056, 2016.

Chapman, T. G.: Entropy as a Measure of Hydrologic Data Uncertainty and Model Performance, J. Hydro., 85, 111-126, 10.1016/0022-1694(86)90079-X, 1986.

Chau, K., Wu, C., and Li, Y.: Comparison of several flood forecasting models in Yangtze River, J. Hydro. Eng., 10, 485-491, 10.1061/(ASCE)1084-0699(2005)10:6(485), 2005.

Cheng, C. T., Zhong-Kai, F., Wen-Jing, N., and Sheng-Li, L.: Heuristic Methods for Reservoir Monthly Inflow Forecasting: A Case Study of Xinfengjiang Reservoir in Pearl River, China, Water-Sui, 7, 4477-4495, 10.3390/w7084477, 2015.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Chicco, and Davide: Ten quick tips for machine learning in computational biology, Biodata Min., 10, 35, 10.1186/s13040-017-0155-3, 2017.

Commission Mekong River: Overview of the Hydrology of the Mekong Basin, Mekong River Commission, Vientiane, 82, 73, 2005.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., and Bauer, d. P.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., 137, 553-597, 10.1002/qj.828, 2011.

Dhanya, C. T., and Kumar, D. N.: Predictive uncertainty of chaotic daily streamflow using ensemble wavelet networks approach, Water Resour. Res., 47, W06507, 10.1029/2010WR010173, 2011.

Duan, Q., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28, 1015-1031, 10.1029/91WR02985, 1992.

El-Shafie, A., Taha, M. R., and Noureldin, A.: A neuro-fuzzy model for inflow forecasting of the Nile river at Aswan high dam, Water Resour. Manag., 21, 533-556, 10.1007/s11269-006-9027-1, 2007.

El-Shafie, A., and Noureldin, A.: Generalized versus non-generalized neural network model for multi-lead inflow forecasting at Aswan High Dam, Hydrol. Earth Syst. Sc., 15, 841–858, 10.5194/hess-15-841-2011, 2011.

Fan, F. M., Schwanenberg, D., Collischonn, W., and Weerts, A.: Verification of inflow into hydropower reservoirs using ensemble forecasts of the TIGGE database for large scale basins in Brazil, J. Hydro., 4, 196-227, 10.1016/j.ejrh.2015.05.012, 2015.

Fienen, M. N., Nolan, B. T., Kauffman, L. J., and Feinstein, D. T.: Metamodeling for groundwater age forecasting in the Lake Michigan Basin, Water Resour. Res., 54, 4750-4766, 10.1029/2017WR022387, 2018.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Ann. Stat., 29, 1189-1232, 10.2307/2699986, 2001.

Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., Wang, G., and Zhou, F.: McTwo: a two-step feature selection algorithm based on maximal information coefficient, Bmc Bioinformatics, 17, 142, 10.1186/s12859-016-0990-0, 2016.

Ghimire, S., Deo, R. C., Downs, N. J., and Raj, N.: Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cites of queensland Australia, J. Clean. Prod., 216, 288-310, 10.1016/j.jclepro.2019.01.158, 2019.

Kinney, J. B., and Atwal, G. S.: Equitability, mutual information, and the maximal information coefficient, P. Natl. Acad. Sci. USA, 111, 3354-3359, 10.1073/pnas.1309933111, 2014.

Kishore, P., Ratnam, M. V., Namboothiri, S. P., Velicogna, I., Basha, G., Jiang, J. H., Igarashi, K., Rao, S. V. B., and Sivakumar, V.: Global (50 °S–50 °N) distribution of water vapor observed by COSMIC GPS RO: Comparison with GPS radiosonde, NCEP, ERA-Interim, and JRA-25 reanalysis data sets, J. Atmos. Sol.-Terr. Phy., 73, 1849-1860, 10.1016/j.jastp.2011.04.017, 2011.

Lima, A. R., Hsieh, W. W., and Cannon, A. J.: Variable complexity online sequential extreme learning machine, with applications to streamflow prediction, J. Hydro., 555, 983-994, 10.1016/j.jhydrol.2017.10.037, 2017.

Lin, J. Y., Cheng, C.-T., and Chau, K.-W.: Using support vector machines for long-term discharge prediction, Hydrology SCI J, 51, 599-612, 10.1623/hysj.51.4.599, 2006.

Linares-Rodr ́guez, A., Ruiz-Arias, J. A., Pozo-V ́azquez, D., and Tovar-Pescador, J.: Generation of synthetic daily global solar radiation data based on ERA-Interim reanalysis and artificial neural networks, Energy, 36, 5356-5365, 10.1016/j.energy.2011.06.044, 2011.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Liu, Z., Zhou, P., and Zhang, Y.: A Probabilistic Wavelet–Support Vector Regression Model for Streamflow Forecasting with Rainfall and Climate Information Input*, J. Hydrometeorol, 16, 2209-2229, 10.1175/JHM-D-14-0210.1, 2015.

Luo, X., Yuan, X., Zhu, S., Xu, Z., Meng, L., and Peng, J.: A hybrid support vector regression framework for streamflow forecast, J. Hydro., 568, 184-193, 10.1016/j.jhydrol.2018.10.064, 2019.

Lyu, H., Wan, M., Han, J., Liu, R., and Cheng, W.: A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining, Comput. Biol. Med., 89, 264-274, 10.1016/j.compbiomed.2017.08.021, 2017.

Mason, L., Baxter, J., Bartlett, P. L., and Frean, M. R.: Boosting algorithms as gradient descent, Adv. Neur. in., 2000, 512-518,

Mehr, A. D., Jabarnejad, M., and Nourani, V.: Pareto-optimal MPSA-MGGP: A new gene-annealing model for monthly rainfall forecasting, J. Hydro., 571, 406-415, 10.1016/j.jhydrol.2019.02.003, 2019.

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

Pal, I., Lall, U., Robertson, A. W., Cane, M. A., and Bansal, R.: Predictability of Western Himalayan river flow: melt seasonal inflow into Bhakra Reservoir in northern India, Hydrol. Earth Syst. Sc., 17, 2131-2146, 10.5194/hess-17-2131-2013, 2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825-2830, 2011.

Rajaee, T., Ebrahimi, H., and Nourani, V.: A review of the artificial intelligence methods in groundwater level modeling, J. Hydro., 572, 336-351, 10.1016/j.jhydrol.2018.12.037, 2019.

Rasouli, K., Hsieh, W. W., and Cannon, A. J.: Daily streamflow forecasting by machine learning methods with weather and climate inputs, J. Hydro., 414, 284-293, 10.1016/j.jhydrol.2011.10.039, 2012.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C.: Detecting novel associations in large data sets, science, 334, 1518-1524, 10.1126/science.1205438, 2011.

Robertson, D., Pokhrel, P., and Wang, Q.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrol. Earth Syst. Sc., 17, 579-593, 10.5194/hess-17-579-2013, 2013.

Siqueira, H., Boccato, L., Luna, I., Attux, R., and Lyra, C.: Performance analysis of unorganized machines in streamflow forecasting of Brazilian plants, Appl. Soft. Comput., 68, 494-506, 10.1016/j.asoc.2018.04.007, 2018.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical bayesian optimization of machine learning algorithms, Adv. Neur. in., 2012, 2951-2959,

Stopa, J. E., and Cheung, K. F.: Intercomparison of wind and wave data from the ECMWF Reanalysis Interim and the NCEP Climate Forecast System Reanalysis, Ocean Model., 75, 65-83, 10.1016/j.ocemod.2013.12.006, 2014.

Sun, G., Li, J., Dai, J., Song, Z., and Lang, F.: Feature selection for IoT based on maximal information coefficient, Future Gener. Comp. Sy., 89, 606-616, 10.1016/j.future.2018.05.060, 2018.

Tongal, H., and Booij, M. J.: Simulation and forecasting of streamflows using machine learning models coupled with base flow separation, J. Hydro., 564, 266-282, 10.1016/j.jhydrol.2018.07.004, 2018.

Valipour, M., Banihabib, M. E., and Behbahani, S. M. R.: Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir, J. Hydro., 476, 433-441, 10.1016/j.jhydrol.2012.11.017, 2013.

Verkade, J., Brown, J., Reggiani, P., and Weerts, A.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydro., 501, 73-91, 10.1016/j.jhydrol.2013.07.039, 2013.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Wang, E., Zhang, Y., Luo, J., Chiew, F. H., and Wang, Q.: Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data, Water Resour. Res., 47, W05516, 10.1029/2010WR009922, 2011.

Wei, C.-C.: Comparing single-and two-segment statistical models with a conceptual rainfall-runoff model for river streamflow prediction during typhoons, Environ. Modell. Softw., 85, 112-128, 10.1016/j.envsoft.2016.08.013, 2016.

Wei, Z., Meng, Y., Zhang, W., Peng, J., and Meng, L.: Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau, Remote Sens. Environ., 225, 30-44, 10.1016/j.rse.2019.02.022, 2019.

Yang, Q., Zhang, H., Wang, G., Luo, S., Chen, D., Peng, W., and Shao, J.: Dynamic runoff simulation in a changing environment: A data stream approach, Environ. Modell. Softw., 112, 157-165, 10.1016/j.envsoft.2018.11.007, 2019.

Zhan, X., Zhang, S., Szeto, W. Y., and Chen, X.: Multi-step-ahead traffic speed forecasting using multi-output gradient boosting regression tree, J. Intell. Transport. S., 1547-2442, 10.1080/15472450.2019.1582950, 2019.

Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., Liu, X., and Zhuang, J.: Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm, J. Hydro., 565, 720-736, 10.1016/j.jhydrol.2018.08.050, 2018.

Zhao, X., Deng, W., and Shi, Y.: Feature selection with attributes clustering by maximal information coefficient, Procedia Comput. Sci., 17, 70-79, 10.1016/j.procs.2013.05.011, 2013.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

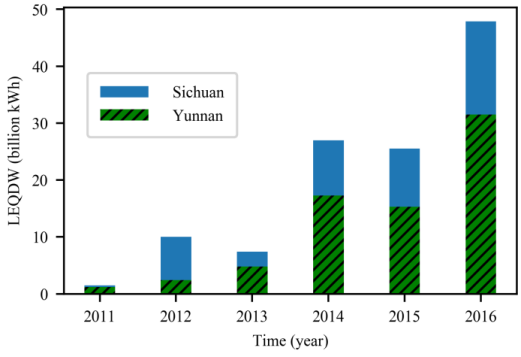
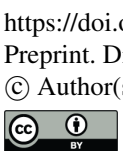
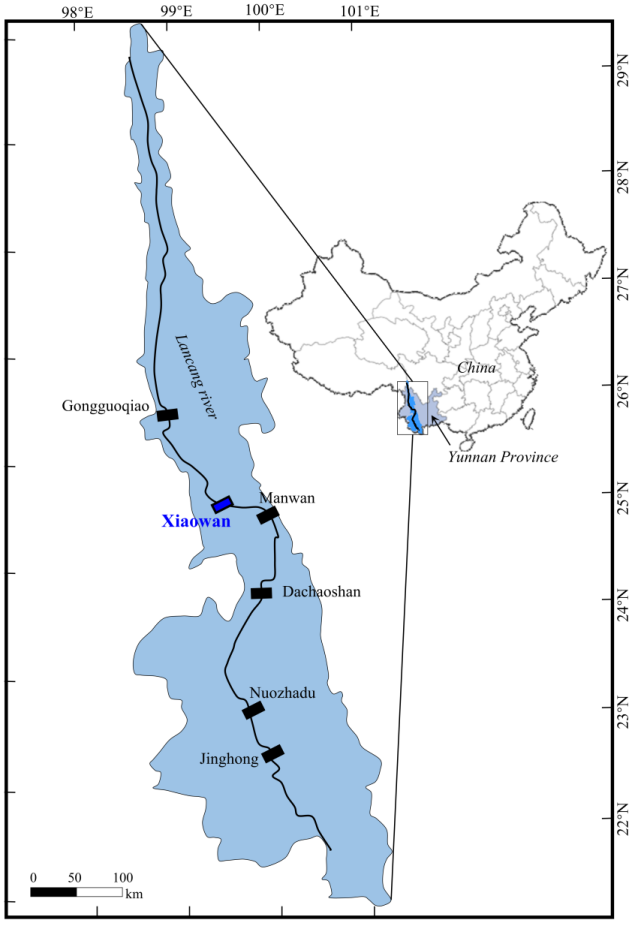


Figure 1: Loss of electric quantity due to discarded water (LEQDW).

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019



Hydrology and Earth System Sciences
Discussions
Open Access
EGU



450 **Figure 2: Location of the Xiaowan hydropower station.**

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

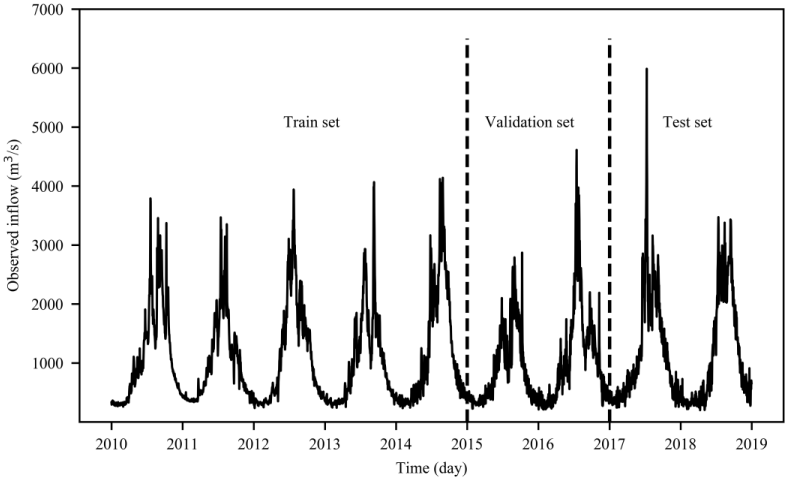
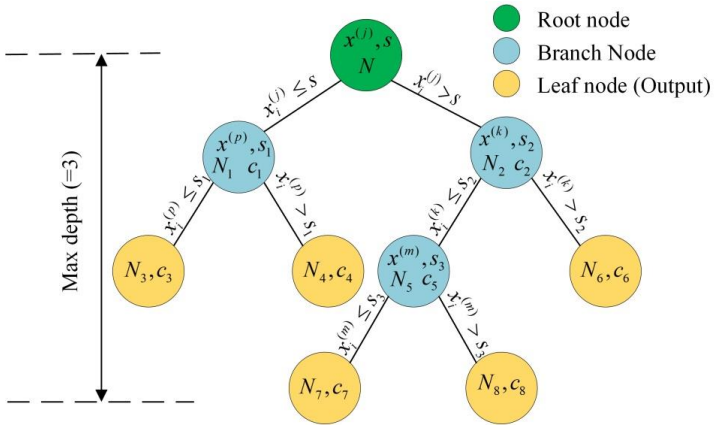


Figure 3: Daily inflow data of the Xiaowan hydropower station.



455

Figure 4: The structure of decision tree model.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

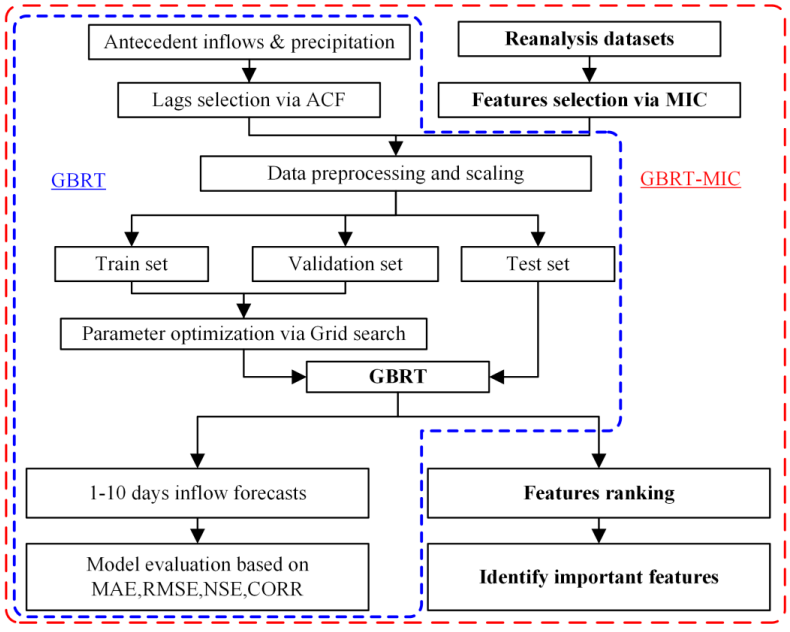
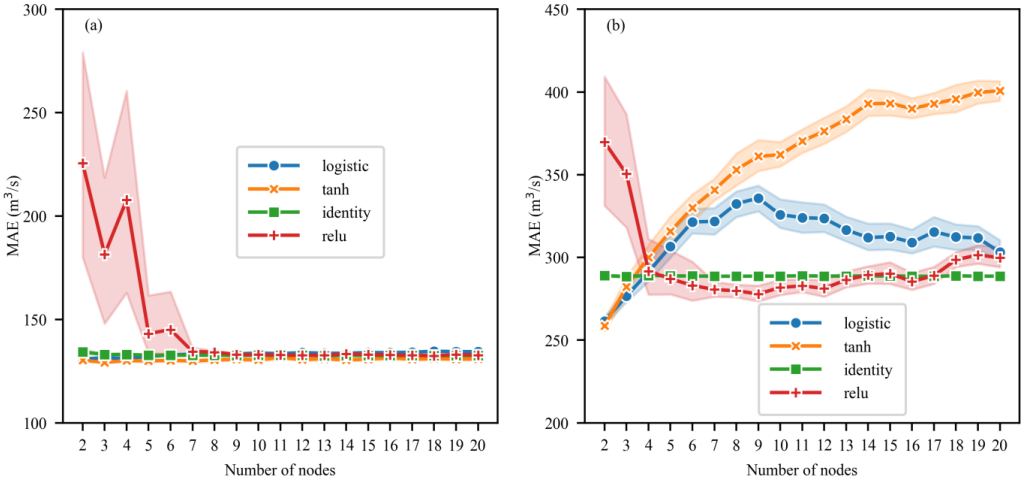


Figure 5: Overview of the framework.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU



460 **Figure 6: Sensitivity of the number of nodes and activation function in the hidden layer on the MAE of ANN-MIC, the shadow part is 95% confidence interval obtained by bootstrap of 50 trials. (a) One-day-ahead (b) Ten-day-ahead.**

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

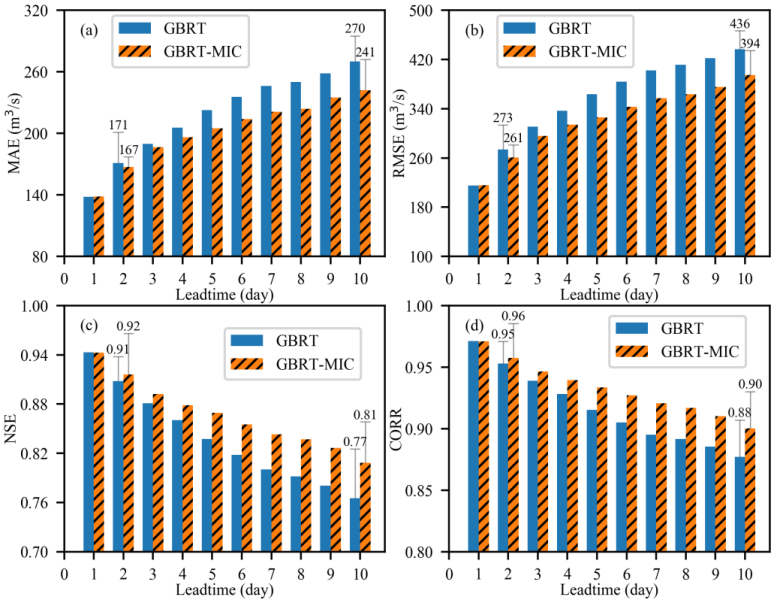
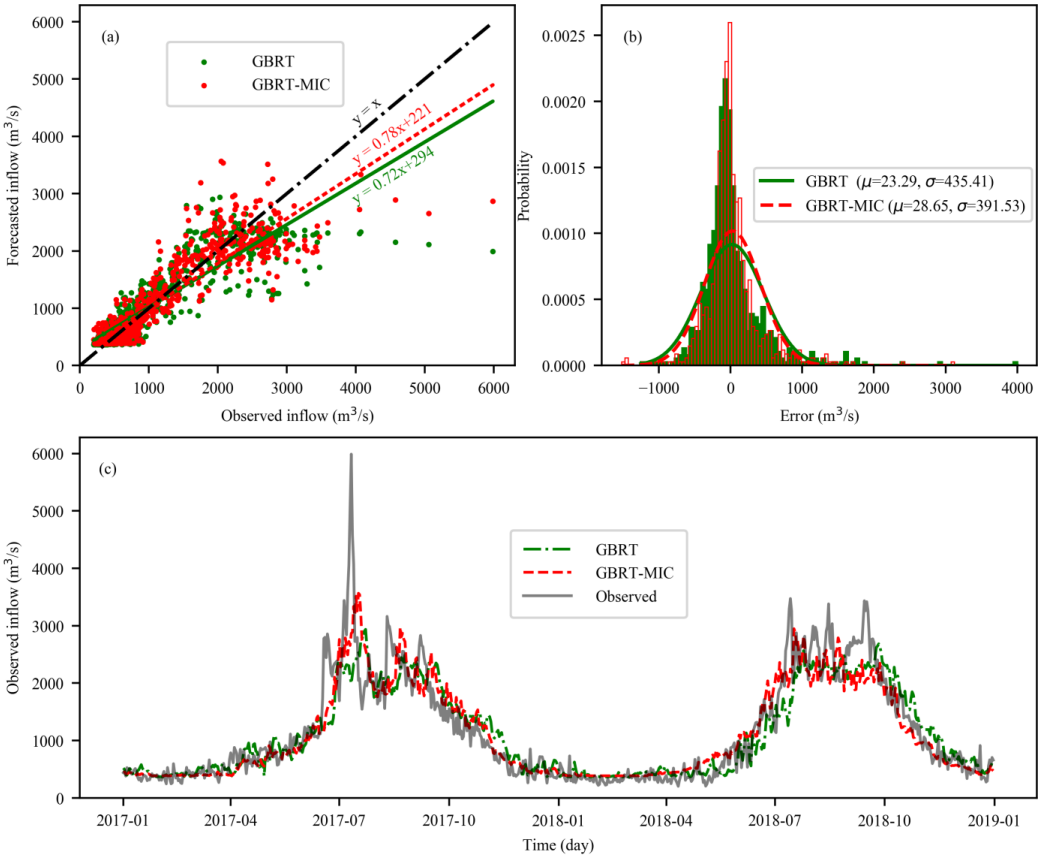


Figure 7: Performance of GBRT and GBRT-MIC for the test set (2017-2018) in term of four indices (a) MAE (b) RMSE (c) NSE (d) CORR.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU



465

Figure 8: Ten-day-ahead inflow forecasts of the GBRT and GBRT-MIC for test set (2017-2018, 730 days) (a) Observed versus forecasted inflow. (b) The histogram of predicting error of test set (c) Comparison of the observed and forecasted inflow.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

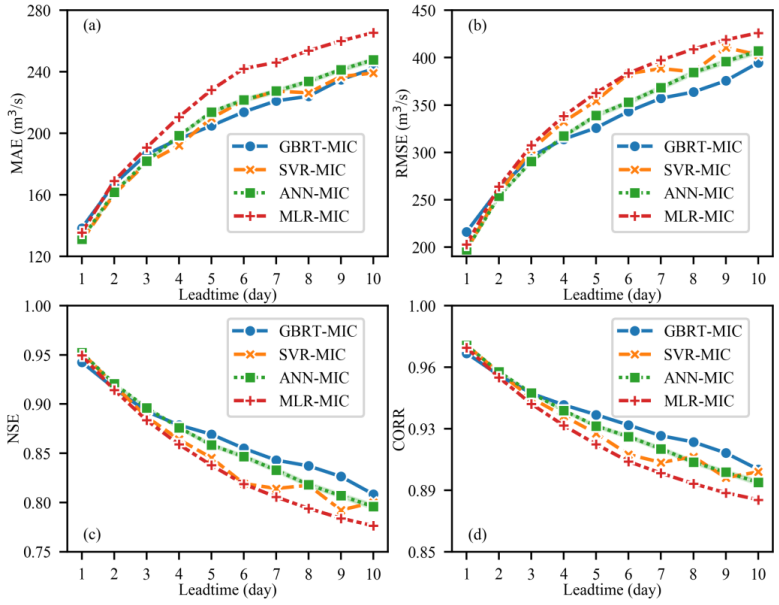


Figure 9: Performance of GBRT-MIC, SVR-MIC, ANN-MIC and MLR-MIC for the test set (2017-2018) in term of four indices. Bootstrap 95% confidence intervals of ANN-MIC for 50 trials. (a) MAE (b) RMSE (c) NSE (d) CORR.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

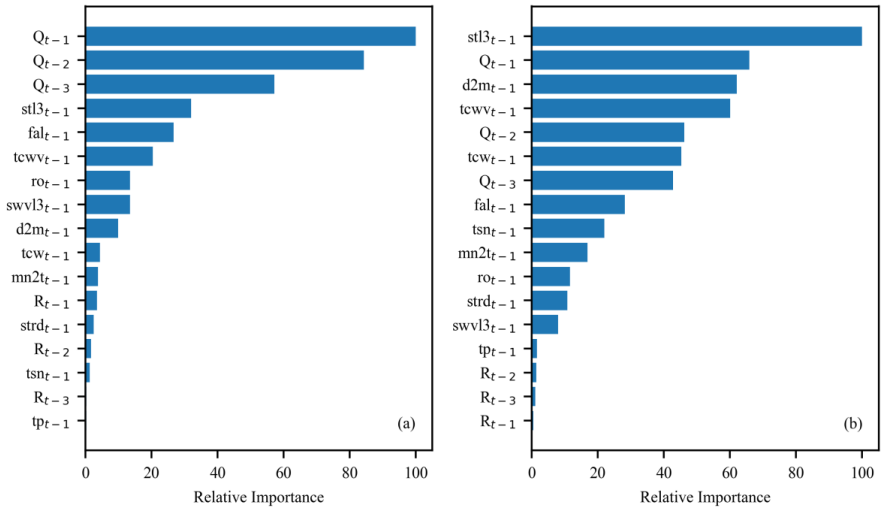


Figure 10: Feature importance for GBRT-MIC (a) One-day-ahead (b) Ten-day-ahead.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table 1

ACF of Xiaowan daily inflow and rainfall (2011-2014).

ACF	t	$t-1$	$t-2$	$t-3$	$t-4$	$t-5$	$t-6$	$t-7$	$t-8$	$t-9$
Q_t	0.974	0.950	0.931	0.914	0.899	0.883	0.870	0.858	0.847	0.838
R_t	0.504	0.273	0.226	0.233	0.195	0.174	0.153	0.170	0.171	0.174

475

Table 2

List of inputs of GBRT-MIC. There are of two types, observed and reanalysis variables. The reanalysis variables are available four time a day at 00:00 UTC, 06:00 UTC, 12:00 UTC and 18:00 UTC. The cumulative variable (e.g., Total column water) is the sum of four periods and the instantaneous variable (e.g. 2 meter dewpoint temperature) is the mean of four periods.

No.	Variable	Index	Unit	MIC	Type
1	Inflow at day t	Q_t	$m^3 \cdot s^{-1}$	-	Obs..
2	Inflow at day t − 1	Q_{t-1}	$m^3 \cdot s^{-1}$	-	Obs.
3	Inflow at day t – 2	Q_{t-2}	$m^3 \cdot s^{-1}$	-	Obs.
4	Rainfall at day t	R_t	mm	-	Obs.
5	Rainfall at day t – 1	R_{t-1}	mm	-	Obs.
6	Rainfall at day t – 2	R_{t-2}	mm	-	Obs.
7	Forecast albedo	fal_t	-	0.853	ERA-I
8	Soil temperature level 3	$stl3_t$	K	0.814	ERA-I
9	2 meter dewpoint temperature	$d2m_t$	K	0.721	ERA-I
10	Minimum temperature at 2 meters	$mn2t_t$	K	0.660	ERA-I
11	Total column water	tcw_t	$kg \cdot m^{-2}$	0.660	ERA-I
12	Total column water vapour	$tcwv_t$	$kg \cdot m^{-2}$	0.659	ERA-I
13	Runoff	ro_t	m	0.653	ERA-I
14	Volumetric soil water layer 3	$swvl3_t$	$m^3 \cdot m^{-3}$	0.637	ERA-I
15	Temperature of snow layer	tsn_t	K	0.631	ERA-I
16	Surface thermal radiation downwards	$strd_t$	$J \cdot m^{-2}$	0.625	ERA-I
17	Total precipitation	tp_t	m	0.391	ERA-I

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

480 **Table 3**

Four commonly used activation functions.

Name	Functional expression
logistic	$f(x)=\frac{1}{1+e^{-x}}$
tanh	$f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
identity	$f(x)=x$
relu	$f(x)=max(0,x)$

Table 4

Tuning parameters of ANN-MIC and SVR-MIC.

Model	Tuning parameter	1	2	3	4	5	6	7	8	9	10
ANN-MIC	Structure Activate function	17-3-1 tanh	17-2-1 logistic	17-2-2 logistic	17-2-3 tanh	17-2-4 logistic	17-2-5 tanh	17-2-6 logistic	17-2-7 logistic	17-2-8 logistic	17-2-9 tanh
SVR-MIC	C	8.9693	7.6987	53.1206	9.8590	44.2134	81.4581	120.2356	0.2587	144.4810	0.5105
	epsilon	0.0030	0.0007	0.0028	0.0066	0.0006	0.0015	0.0004	0.0012	0.0027	0.0085
	gamma	0.0265	0.0201	0.0037	0.0120	0.0051	0.0012	0.0088	0.2150	0.0067	0.1815

485

Table 5

Tuning parameters of GBRT and GBRT-MIC.

Tuning parameter	Tuning range	Optimal parameters (the lead times of 1-10 days)	
		GBRT	GBRT-MIC
max_leaf_nodes	[3,5,7,9]	7,7,3,5,9,3,7,9,7,3	3,3,3,7,3,7,7,9,5,9
min_samples_leaf	[3,5,7,9]	9,5,9,5,5,5,7,3,3,3	9,9,9,7,5,5,5,5,5,7
max_depth	[3,5,7,9]	3,3,9,9,9,3,3,5,9,5	9,9,7,9,9,5,7,5,7,7
min_samples_split	[10,20,30,40,50]	30,10,40,50,10,20,50,50,50,40	50,10,20,10,10,20,30,50,10,50
n_estimators	[200, 250, …, 1200]	600,600,450,450,550,500,500,550,600,600	1150,550,350,350,450,350,350,400,500,500

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table 6

Performance indices of the train set.

Indice	Model	1	2	3	4	5	6	7	8	9	10
MAE	GBRT-MIC	**70**	**98**	**126**	**141**	**150**	**167**	**173**	**172**	**170**	**169**
	SVR-MIC	86	116	138	151	161	177	176	182	192	192
	ANN-MIC	90	119	139	153	163	174	181	187	194	198
	MLR-MIC	92	124	146	162	173	185	195	205	213	219
RMSE	GBRT-MIC	**103**	**152**	**197**	**224**	**248**	**272**	**285**	**288**	**286**	**283**
	SVR-MIC	139	195	240	267	285	314	318	329	345	344
	ANN-MIC	138	190	226	248	270	287	307	319	331	337
	MLR-MIC	142	195	233	259	281	305	322	337	348	359
NSE	GBRT-MIC	**0.9844**	**0.9659**	**0.9426**	**0.9257**	**0.9090**	**0.8908**	**0.8804**	**0.8772**	**0.8791**	**0.8816**
	SVR-MIC	0.9715	0.9436	0.9149	0.8947	0.8802	0.8540	0.8508	0.8402	0.8245	0.8249
	ANN-MIC	0.9718	0.9466	0.9248	0.9091	0.8924	0.8785	0.8610	0.8495	0.8385	0.8324
	MLR-MIC	0.9703	0.9438	0.9196	0.9009	0.8833	0.8629	0.8472	0.8324	0.8206	0.8100
CORR	GBRT-MIC	**0.9923**	**0.9832**	**0.9728**	**0.9645**	**0.9548**	**0.9468**	**0.9416**	**0.9391**	**0.9397**	**0.9417**
	SVR-MIC	0.9859	0.9722	0.9592	0.9501	0.9420	0.9291	0.9257	0.9209	0.9139	0.9131
	ANN-MIC	0.9858	0.9731	0.9620	0.9538	0.9452	0.9378	0.9286	0.9225	0.9165	0.9131
	MLR-MIC	0.9851	0.9717	0.9594	0.9498	0.9406	0.9299	0.9214	0.9134	0.9070	0.9013

490 *None :* The bold numbers represent the values of performance criterion for the best fitted models.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions

Table 7

Performance indices of the validation set

Indice	Model	1	2	3	4	5	6	7	8	9	10
MAE	GBRT-MIC	**96**	**127**	**153**	**170**	**177**	**190**	**194**	**194**	**196**	**189**
	SVR-MIC	120	149	169	183	193	201	202	200	206	207
	ANN-MIC	122	154	174	188	199	206	214	215	218	216
	MLR-MIC	127	161	185	202	212	220	228	233	236	240
RMSE	GBRT-MIC	**132**	**188**	**230**	**254**	**266**	**286**	**294**	**296**	**293**	**276**
	SVR-MIC	178	240	274	294	308	326	325	326	329	330
	ANN-MIC	177	240	271	290	308	317	332	337	337	333
	MLR-MIC	184	247	281	304	319	332	342	346	349	353
NSE	GBRT-MIC	**0.9669**	**0.9324**	**0.8991**	**0.8766**	**0.8647**	**0.8440**	**0.8348**	**0.8325**	**0.8364**	**0.8544**
	SVR-MIC	0.9395	0.8898	0.8570	0.8352	0.8181	0.7969	0.7982	0.7965	0.7926	0.7922
	ANN-MIC	0.9398	0.8900	0.8594	0.8398	0.8185	0.8075	0.7893	0.7830	0.7824	0.7876
	MLR-MIC	0.9349	0.8832	0.8486	0.8236	0.8052	0.7889	0.7766	0.7708	0.7666	0.7620
CORR	GBRT-MIC	**0.9835**	**0.9661**	**0.9493**	**0.9376**	**0.9311**	**0.9203**	**0.9154**	**0.9140**	**0.9163**	**0.9260**
	SVR-MIC	0.9693	0.9434	0.9257	0.9139	0.9046	0.8927	0.8939	0.8928	0.8903	0.8906
	ANN-MIC	0.9697	0.9444	0.9289	0.9185	0.9081	0.9015	0.8925	0.8895	0.8890	0.8912
	MLR-MIC	0.9673	0.9412	0.9238	0.9112	0.9018	0.8935	0.8871	0.8839	0.8818	0.8798

None : The bold numbers represent the values of performance criterion for the best fitted models.

https://doi.org/10.5194/hess-2019-610
Preprint. Discussion started: 13 December 2019

Hydrology and Earth System Sciences
Discussions

Table 8

Performance indices of the test set.

Indice	Model	1	2	3	4	5	6	7	8	9	10
MAE	GBRT-MIC	138	167	186	196	**205**	**214**	**221**	**224**	**235**	242
	SVR-MIC	**130**	**160.5**	**180.6**	**192**	210	221	227	226	237	**239**
	ANN-MIC	131	161.4	181.4	199	214	221	228	235	242	247
	MLR-MIC	135	169	191	211	228	242	246	254	260	266
RMSE	GBRT-MIC	216	261	296	**314**	**326**	**343**	**357**	**364**	**376**	**395**
	SVR-MIC	**196.7**	256	303	332	354	383	388	385	411	403
	ANN-MIC	197.1	**253**	**290**	318	339	354	368	386	398	406
	MLR-MIC	203	264	307	338	363	383.7	397	409	419	426
NSE	GBRT-MIC	0.9426	0.9161	0.8924	**0.8787**	**0.8693**	**0.8550**	**0.8430**	**0.8372**	**0.8264**	**0.8084**
	SVR-MIC	**0.9524**	0.9190	0.8870	0.8641	0.8454	0.8193	0.8142	0.8175	0.7923	0.8001
	ANN-MIC	0.9522	**0.9211**	**0.8966**	0.8756	0.8582	0.8457	0.8337	0.8165	0.8052	0.7967
	MLR-MIC	0.9494	0.9143	0.8836	0.8590	0.8378	0.8188	0.8056	0.7940	0.7840	0.7764
CORR	GBRT-MIC	0.9710	0.9575	0.9465	**0.9395**	**0.9335**	**0.9272**	**0.9207**	**0.9169**	**0.9102**	**0.9002**
	SVR-MIC	**0.9760**	0.9591	0.9437	0.9327	0.9221	0.9093	0.9044	0.9079	0.8951	0.8987
	ANN-MIC	0.9758	**0.9598**	**0.9470**	0.9358	0.9265	0.9197	0.9132	0.9037	0.8974	0.8926
	MLR-MIC	0.9744	0.9562	0.9401	0.9269	0.9155	0.9050	0.8979	0.8915	0.8859	0.8817

None : The bold numbers represent the values of performance criterion for the best fitted models.