AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR031065

Key Points:

- A regression tree ensemble is proposed to address the temporal autocorrelation of daily streamflow
- A Bayesian model average with a stratified sampling ensemble strategy is provided to improve the simulation accuracy for medium flows
- Autocorrelation of daily streamflow is particularly important for establishing reliable irrigation-discharge relationships

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

G. Huang,
huangg@uregina.ca

Citation:

Li, K., Huang, G., Wang, S., Baetz, B., & Xu, W. (2022). A stepwise clustered hydrological model for addressing the temporal autocorrelation of daily streamflows in irrigated watersheds. *Water Resources Research*, *58*, e2021WR031065. https://doi.org/10.1029/2021WR031065

Received 17 AUG 2021
Accepted 30 JAN 2022

Author Contributions:

Conceptualization: Kailong Li
Data curation: Kailong Li
Formal analysis: Kailong Li
Investigation: Kailong Li
Methodology: Kailong Li
Resources: Kailong Li
Software: Kailong Li
Validation: Kailong Li, Weihuang Xu
Visualization: Shuo Wang
Writing – original draft: Kailong Li
Writing – review & editing: Guohe Huang, Shuo Wang, Brian Baetz, Weihuang Xu

A Stepwise Clustered Hydrological Model for Addressing the Temporal Autocorrelation of Daily Streamflows in Irrigated Watersheds

Kailong Li[1], Guohe Huang[1], Shuo Wang[2], Brian Baetz[3], and Weihuang Xu[4]

[1]Faculty of Engineering, University of Regina, Regina, SK, Canada, [2]Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong, China, [3]Department of Civil Engineering, McMaster University, Hamilton, ON, Canada, [4]Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

Abstract Streamflow simulations at daily time steps are vital to water resources management, especially in arid regions. Previously, data-driven models have been used as an effective tool for daily streamflow simulation. However, the accuracy of conventional data-driven approaches is affected by the temporal autocorrelation of daily streamflow, especially in irrigated watersheds where the persistence of saturated flows dominates irrigation seasons. This study presents a Stepwise Clustered Regression Tree Ensemble (SCRTE) to address the streamflow autocorrelation. With the provision of a state-of-the-art data-driven model Stepwise Cluster Analysis (SCA), the SCRTE enables both single- and multi-output settings (i.e., model predictand can be either a scalar or a vector), which can thus address interactions among streamflow values over multiple consecutive days. The autocorrelation effect of daily streamflow is evaluated based on single- and multi-output SCA ensembles, which can then be aggregated according to their performance for various streamflow quantile ranges. To facilitate the irrigation scheduling decision-making under rigorous transboundary water regulations, the SCRTE is applied to three interconnected watersheds with mixed land use, located in a floodplain of the Yellow River basin in China. The results show that the SCRTE outperforms seven well-known benchmark models across seven evaluation metrics. Our findings reveal that the SCRTE can reflect the varying effects of autocorrelation over different streamflow quantile ranges, thereby improving the streamflow simulation. The multi-output SCA ensembles are more capable of addressing the medium flows, while the single-output one can better simulate the low and high flows.

1. Introduction

Data-driven models have received increasing attention from hydrological communities (ASCE Task Committee, 2000b; Elshorbagy et al., 2010; Kratzert, Klotz, Shalev, et al., 2019; H. Zhang et al., 2019). They are flexible enough to approximate various complex processes and interrelationships, and allow for direct mapping from meteorological and ancillary inputs (e.g., soil moisture (Schmidt et al., 2020), catchment attributes (Kratzert, Klotz, Shalev, et al., 2019) and irrigation scheduling (Mohan & Vijayalakshmi, 2009)) to streamflow or other output fluxes (Solomatine & Ostfeld, 2008). Many data-driven hydrological models have been developed (Adnan et al., 2020; ASCE Task Committee, 2000a; Y. Yang et al., 2020; H. Zhang et al., 2019). However, most of the existing modeling efforts are focused on short-term predictions, such as single- or multi-step-ahead forecasting (Adnan et al., 2019; Badrzadeh et al., 2013; Bray & Han, 2004; Campolo et al., 2003; Fleming et al., 2015; He et al., 2014; Kisi et al., 2012; Toth et al., 2000; J. Yang et al., 2013). Short-term predictions may not be useful for supporting the formulation of water-related policies that are mostly of long-term considerations (M. Cheng et al., 2020; S. Duan et al., 2020). Therefore, long-term streamflow simulation is needed (M. Cheng et al., 2020; S. Duan et al., 2020; Ren et al., 2018; Sarzaeim et al., 2017; Tongal & Booij, 2018; H. Zhang et al., 2019).

The most commonly used method for long-term simulation is the iterated (also known as serial propagated) prediction method (Bontempi et al., 2012; M. Cheng et al., 2020; Taieb et al., 2010), where the single-step-ahead prediction is iterated N times (N is the total time steps of the prediction horizon). Once the model predicts a future (streamflow) value, this value is fed back as an input to the following prediction. Hence, the model considers the predicted value as an input, instead of the actual observation (Taieb et al., 2010). The iterated method enables a model to memorize previous system conditions, and can thus reflect the temporal autocorrelation of streamflow (i.e., effects of internal water storage mechanisms such as groundwater and wetlands) (Bierkens &

Van Beek, 2009; Wada et al., 2010). However, the iterated method may suffer from poor performance in the long-term streamflow simulation (Chakraborty et al., 2021; Chang et al., 2007; S. Duan et al., 2020; J. Yang et al., 2013), since it is tuned based on a single-step-ahead criterion and cannot reflect various temporal complexities (Taieb et al., 2010). Moreover, the estimated values are iteratively used as inputs for simulating the next-step streamflow values, thereby leading to the accumulation of simulation errors (Chang et al., 2007; C. Cheng et al., 2008; M. Cheng et al., 2020; Taieb et al., 2010; J. Yang et al., 2013).

To avoid the accumulation of simulation errors over time, many long-term simulation efforts have relied on external forcing data (e.g., meteorological data), without using the streamflow data generated from previous time steps (Dibike et al., 2001; S. Duan et al., 2020; Ren et al., 2018; Sarzaeim et al., 2017; Shortridge et al., 2016; H. Zhang et al., 2019). By doing so, the streamflow value at any time step is estimated independently without the information of previous time steps (S. Duan et al., 2020), such that the autocorrelation effect can hardly be addressed adequately. To this end, modelers have been using advanced architectures and/or complicated predictors to address the streamflow autocorrelation.

In terms of modeling architecture, long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997), as a type of recurrent neural network, has proved to be a powerful approach for addressing the autocorrelation of hydrological time series (Fang & Shen, 2020). The dedicated architecture of LSTM allows for storing and regulating information over a long period of time, enabling the model to capture the long-range dependencies between inputs and outputs (Kratzert, Klotz, Shalev, et al., 2019; LeCun et al., 2015). Nevertheless, LSTM is well known for requiring a large amount of training data to achieve a satisfactory prediction (Bai et al., 2021; Ma et al., 2021). The desired length of training data varies case by case, depending on the application case, data distributions and model complexity (Goodfellow et al., 2017; Wunsch et al., 2021). For watersheds with inadequate observations, a conventional LSTM model may not necessarily outperform other data-driven models such as support vector regression (Xiang et al., 2020) and feed-forward neural networks (S. Duan et al., 2020; Wunsch et al., 2021).

Besides the model architecture, the use of multiple backward moving sums (or averages) of meteorological data (e.g., rainfall totals or averages over the past days or weeks) as additional predictors is another effective approach to address the streamflow autocorrelation (S. Duan et al., 2020; Schmidt et al., 2020; Shortridge et al., 2016). This is because autocorrelation also exists in meteorological time series (such as daily temperature and precipitation). The backward moving sums of meteorological time series may act as surrogates to characterize the initial conditions for catchment water storage, such as soil moisture, and channel and bank storages (Fleming et al., 2015). The streamflow autocorrelation can thus be addressed based on the relationships between the surrogates and the streamflow. However, such modeling efforts are mainly focused on the establishment of potential relationships between multiple inputs (i.e., predictors) and a single output (i.e., streamflow value at a given time step). The complex effects of dynamic interactions among multi-step streamflow values cannot be adequately reflected. Some multi-input multi-output models are available for addressing the interactions among streamflow values across multiple time steps (e.g., streamflow values at time t, $t + 1$ and $t + 2$) (Chang et al., 2007). Such efforts have only been successfully used for short-term predictions (e.g., a multi-input six-output model for 6-day-ahead streamflow forecasting) (Campolo et al., 2003; Toth et al., 2000). In addition, the multi-input multi-output model suffers from a drawback that its objective function is often defined as a weighted sum of squares error of multiple outputs, and thus can hardly guarantee a minimum error for each time step (Chang et al., 2007). In general, neither the multi-input single-output nor multi-input multi-output model can adequately tackle complex hydrological processes, thereby compromising the performance of the relevant long-term simulation.

In light of the above considerations, the objective of this study is to develop a Stepwise-Clustered Regression Tree Ensemble (SCRTE) for long-term daily streamflow simulation. The proposed SCRTE will improve upon the conventional data-driven hydrological models by addressing the multi-input single-output relationship and the multi-input multi-output ones. Thus, the simulated streamflow will be able to reflect the underlying interactions among the multi-step streamflow values (e.g., streamflow at time t, $t + 1$ and $t + 2$), as well as the streamflow autocorrelation. This study entails (i) application of the SCRTE to three interconnected mixed-landuse watersheds located in a floodplain of the Yellow River Basin, China; (ii) comparative assessment of simulation accuracies against seven benchmark models using seven evaluation metrics; (iii) exploration of the varying effects of autocorrelation from the perspective of flow magnitudes; (iv) establishment of irrigation-discharge relationships for supporting the irrigation scheduling under rigorous transboundary water regulations; (v) quantification of importance ranking for predictor variables to obtain better insight into the hydrological processes.

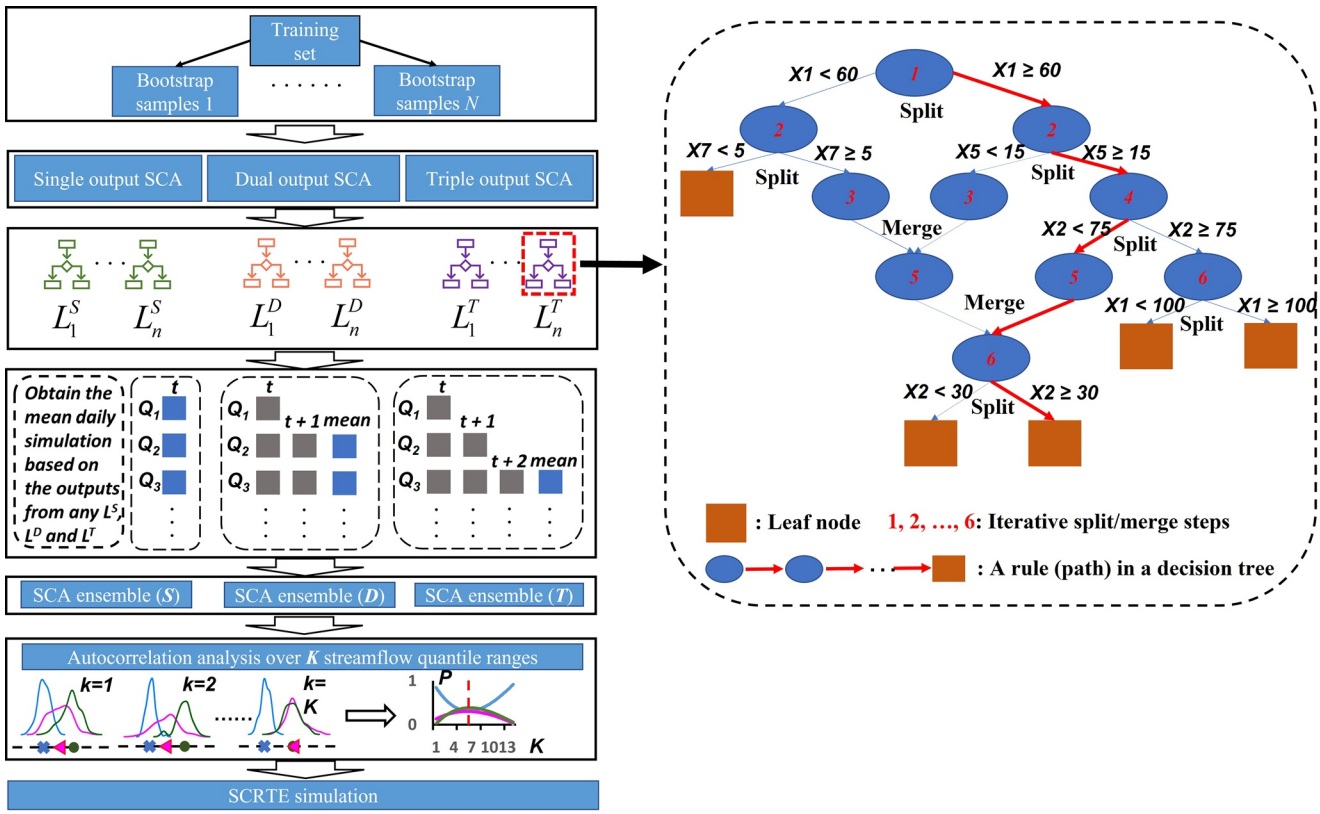

Figure 1. Flow chart of the proposed SCRTE (left) along with an example of an SCA tree (right). *Note* ***S***, ***D*** and ***T*** represent the SCA settings of single, dual and triple-output variables, respectively. ***L*** represents the regression tree of SCA, and ***N*** is the total number of bootstrap samples. For the example of an SCA tree, splitting and merging actions are iteratively performed (i.e., stages 1, 2, 4 and 6 are splitting actions while stages 3 and 5 are merging actions) until no node(s) can be split or merged. The undividable node is called the leaf node. ***X*1**, …, ***X*7** denote predictor variables used for node-splitting process.

2. Stepwise Clustered Regression Tree Ensemble

2.1. Modeling Framework and SCA Ensemble

The SCRTE is built upon a conventional random forest (RF) framework (Breiman, 2001), in which the functions of classification and regression (Breiman et al., 1984) are accomplished through stepwise cluster analysis (SCA) (Huang, 1992). SCA has been proven to be an effective approach for hydrological modeling owing to its ability to reduce overfitting and facilitate transparent inference (Fan et al., 2017; Han et al., 2016; Li et al., 2015). However, there has been no attempt to use the SCA within a regression tree ensemble (RTE) framework (James et al., 2013), even though RTE has been reported as an effective approach to improve the accuracy of a single tree (Breiman, 2001). The SCA enables both single- and multi-output settings, implying that the model predictand can be either a scalar (e.g., streamflow values at time t) or a vector (e.g., streamflow values at time t, $t+1$ and $t+2$). A multi-output SCA can help address the interactions among streamflow values over multiple consecutive days starting from a target day, which is crucial for reflecting the autocorrelation effects of streamflow time series.

The modeling process of SCRTE (Figure 1) involves the following procedures: (1) training SCA trees with single-, dual- and triple-output settings in parallel; (2) building three SCA ensembles (indicated as ***S***, ***D*** and ***T***) according to the three settings in step 1; (3) quantifying the effects of autocorrelation on the three SCA ensembles for a number of streamflow quantile ranges; and (4) aggregation of SCA ensemble simulations based on the effects of autocorrelation obtained in step 3.

Each of the SCA ensembles (i.e., ***S***, ***D*** and ***T***) is built by following the RF process. In detail, each SCA tree (from an SCA ensemble) grows in accordance with a random subset of predictors sampled without replacement and a bootstrapped version of the training set (same size to the original training set), drawn randomly from the initial training dataset with replacement. Such a bootstrap sampling process can leave about 1/3 of the training data as

out-of-bag (OOB) data, which will not be involved in training the $\boldsymbol{n}_{th}$ SCA tree and can thus be used for validating the corresponding tree. Since the validation results for each tree randomly cover 1/3 of the data over the training period, the ensemble (average) of validation results for all trees can then cover the entire dataset over the training period when $\boldsymbol{N}$ is large enough (e.g., $\boldsymbol{N} = 200$).

2.2. Training Processes of an Individual SCA Tree

The training process of an SCA tree is illustrated in Figure 2. In general, the original sample space is divided into a number of non-overlapping subspaces based on recursive binary splitting and merging operations (Huang, 1992; Wang et al., 2013). If $\boldsymbol{k}$ instances are applied for the node splitting process, a total of $\boldsymbol{k}$ sets of independent and dependent variables will be obtained in the first step (S1; Figure 2). If there are $\boldsymbol{n}$ independent variables and $\boldsymbol{m}$ dependent variables, the dataset can be presented as matrices $\boldsymbol{X} = (\boldsymbol{X}_{ip})$ and $\boldsymbol{Y} = (\boldsymbol{Y}_{iq})$, where $\boldsymbol{i} = 1, 2, \ldots, \boldsymbol{k}; \boldsymbol{p} = 1, 2, \ldots, \boldsymbol{n}; \boldsymbol{q} = 1, 2, \ldots, \boldsymbol{m}$. In the second step (S2), the values of $\boldsymbol{X}_{ip}$ are sorted in ascending order based on the $\boldsymbol{p}_{th}$ column. Therefore, all of the other columns in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are reordered accordingly. Let $\boldsymbol{X}_{\boldsymbol{jp}}$ and $\boldsymbol{Y}_{\boldsymbol{jq}}$ denote the reordered X_{ip} and Y_{iq}, respectively, where $\boldsymbol{j} = 1, 2, \ldots, \boldsymbol{k}$. Then we can go through each instance of $\boldsymbol{X}_{jp}$ from the top and examine a total of $\boldsymbol{k} - 1$ candidate split points. Any instance z ($z \in 1, 2, ..., k$) in $\boldsymbol{X}_{\boldsymbol{jp}}$ can split the predictor space into two subspaces as $\boldsymbol{X1}_{jp}$ and $\boldsymbol{X2}_{jp.}$ The response space $\boldsymbol{Y}_{jq}$ will be correspondingly divided into two subspaces as $\boldsymbol{Y1}_{jq}$ and $\boldsymbol{Y2}_{jq}$. In the third step (S3), Wilks lambda ($\boldsymbol{\Lambda}$) statistic (Wilks, 1967) is used to determine the split point of each node. In a multivariate analysis of variance, the Wilks $\boldsymbol{\Lambda}$ value can be used to assess the differences between two or more groups, and can be defined as $\boldsymbol{\Lambda} = \mathrm{Det}(W)/\mathrm{Det}(B + W)$, where $\mathrm{Det}(\cdot)$ is the determinant of a matrix, $\boldsymbol{W}$ is the within-group sum of squares (within-group variations), and $\boldsymbol{B}$ is the between-group cross-product matrix (between-group variations). The $\boldsymbol{W}$ and $\boldsymbol{B}$ for a particular split point can be calculated as:

$$W_{jq} = \sum_{j=1}^{z} \left[Y1_{jq} - \overline{Y1_q}\right]' \cdot \left[Y1_{jq} - \overline{Y1_q}\right] + \sum_{j=z+1}^{k-z} \left[Y2_{jq} - \overline{Y2_q}\right]' \cdot \left[Y2_{jq} - \overline{Y2_q}\right] \tag{1}$$

$$B_{zq} = \frac{z(k-z)}{k}\left(\overline{Y1_q}\,\overline{Y2_q}\right)' \cdot \left(\overline{Y1_q}\,\overline{Y2_q}\right) \tag{2}$$

where $\overline{Y1}_q = \frac{1}{z}\sum_{j=1}^{z} Y1_{jq}$ and $\overline{Y2}_q = \frac{1}{k-z}\sum_{j=z+1}^{k} Y2_{jq}$. For $\boldsymbol{q} = 1, 2, \ldots, \boldsymbol{m}$, $\overline{Y1}_q$ and $\overline{Y2}_q$ can be represented as vectors $\overline{Y1}_q = \left\{\overline{Y1}_1, \overline{Y1}_2, ..., \overline{Y1}_m\right\}$ and $\overline{Y2}_q = \left\{\overline{Y2}_1, \overline{Y2}_2, ..., \overline{Y2}_m\right\}$, respectively; $Y1_{jq}$ and $Y2_{jq}$ can be represented as vectors $Y1_{jq} = \{Y1_{j1}, Y1_{j2}, ..., Y1_{jm}\}$, $j \in 1, 2, ..., z$, and $Y2_{jq} = \{Y2_{j1}, Y2_{j2}, ..., Y2_{jm}\}$ $j \in z+1, z+2, ..., k$, respectively.

Based on the $\boldsymbol{\Lambda}$ definition, if $\mathrm{Det}(B)$ is larger than $\mathrm{Det}(W)$, then Λ value will become smaller, implying a larger difference between $Y1_{jq}$ and $Y2_{jq}$. To determine whether there exists a significant difference between $Y1_{jq}$ and $Y2_{jq}$ for the purpose of splitting, $\boldsymbol{\Lambda}$ is approximated using Rao's $\boldsymbol{F}$-approximation ($\boldsymbol{R}$-statistic) to perform the F-test. The $\boldsymbol{R}$-statistic is defined as (Huang, 1992):

$$R = \frac{1 - \Lambda^{1/S}}{\Lambda^{1/S}} \cdot \frac{Z \cdot S - m \cdot (r-1)/2 + 1}{m \cdot (r-1)} \tag{3}$$

$$Z = k - 1 - (m + r)/2 \tag{4}$$

$$S = \frac{m^2 \cdot (r-1)^2 - 4}{m^2 + (r-1)^2 - 5} \tag{5}$$

where the $\boldsymbol{R}$-statistic is distributed approximately as an $\boldsymbol{F}$-variate with $n_1 = m \cdot (r-1)$ and $n_2 = m \cdot (r-1)/2 + 1$ degrees of freedom; r is the number of groups. Since the number of groups is two (i.e., $Y1_{jq}$ and $Y2_{jq}$), an $\boldsymbol{F}$-test can be performed based on the following Wilks Λ criterion:

$$F(m, k-m-1) = \frac{1-\Lambda}{\Lambda} \cdot \frac{k-m-1}{m} \tag{6}$$

The two subspaces ($Y1_{jq}$ and $Y2_{jq}$) can be compared against each other for examining the significance of their differences through the $\boldsymbol{F}$-test. The null hypothesis would be H_0: $\boldsymbol{\mu}(\boldsymbol{Y1}) = \boldsymbol{\mu}(\boldsymbol{Y2})$ versus the alternative hypothesis H_1: $\boldsymbol{\mu}(\boldsymbol{Y1}) \neq \boldsymbol{\mu}(\boldsymbol{Y2})$, where $\boldsymbol{\mu}(\boldsymbol{Y1})$ and $\boldsymbol{\mu}(\boldsymbol{Y2})$ are the means of $\boldsymbol{Y1}$ and $\boldsymbol{Y2}$, respectively. Let the significance level be

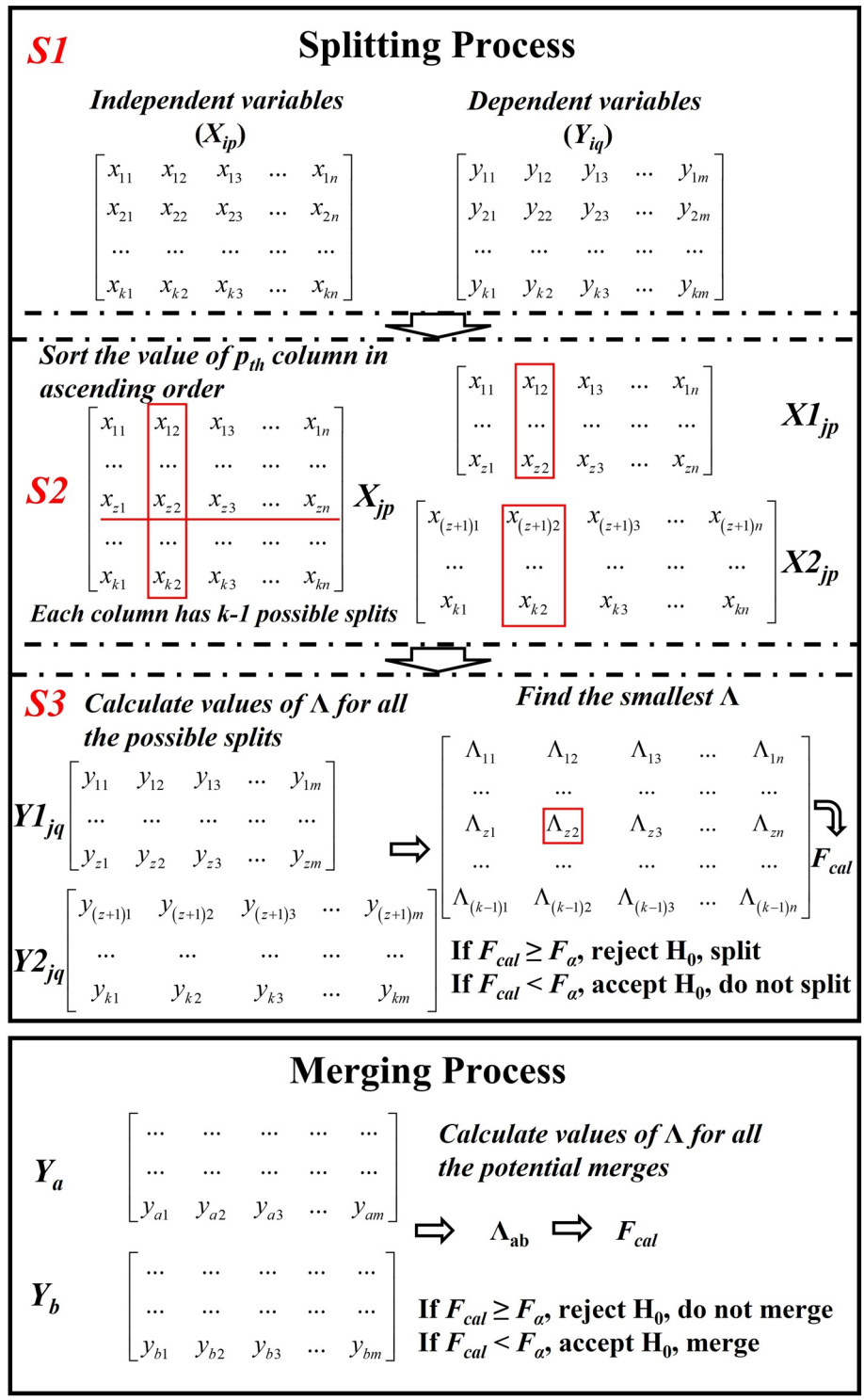

Figure 2. Deductive reasoning for the SCA inference tree.

$\boldsymbol{\alpha}$ (which was set as 0.05 in this study), the split criterion would be: $F_{cal} \geq F_{\alpha}$ (i.e., H_0 is false), which implies that the difference between the two subspaces is significant and thus the sample can be split. As shown in the third step (S3) of Figure 2, the Wilks $\boldsymbol{\Lambda}$ value can be evaluated for each candidate split point. The matrix of $\boldsymbol{\Lambda}$ can be represented as $\boldsymbol{\Lambda}_{ip}$, where $\boldsymbol{i} = 1, 2, \ldots, \boldsymbol{k} - 1; \boldsymbol{p} = 1, 2, \ldots, \boldsymbol{n}$. The potential split point (with the most significant

difference between $Y1_{jq}$ and $Y2_{jq}$) is always found at the point with the smallest value in $\mathbf{\Lambda}_{ip}$. If the H_0 is false for the smallest value of $\mathbf{\Lambda}_{ip}$, the node should be split; otherwise, the node should not be split.

Once all nodes at the current stage have been examined in the splitting process, the merging process will be followed (Figure 1). The merging process is to compare any given pair of nodes based on the Wilks Λ value to test whether the pair can be merged. Assume that $\boldsymbol{a}$ and $\boldsymbol{b}$ are two nodes among the existing $\boldsymbol{H}$ nodes ($\boldsymbol{H}$ is the total number of nodes at the current stage), the Λ value can be calculated using the dependent variables $\boldsymbol{Y}_a$ and $\boldsymbol{Y}_b$ in nodes $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively (Figure 2). According to Equation 6, if $F_{\text{cal}} < F_{\alpha}$ (i.e., H_0 is true), nodes $\boldsymbol{a}$ and $\boldsymbol{b}$ do not have a significant difference and thus can be merged.

Such splitting and merging processes are iteratively performed until no node(s) can be further split or merged. The mean value of dependent variables for those undividable nodes (leaf nodes) can be used for making predictions. When a new sample set ($\boldsymbol{X}_{\text{new}}$) is provided, the value of $\boldsymbol{X}_{\text{new}}$ will be compared with $\boldsymbol{X}_{zp}$ (the splitting point of $\boldsymbol{X}_{jp}$) at each split node. If the value of $\boldsymbol{X}_{\text{new}}$ is greater than the value of $\boldsymbol{X}_{zp}$, the sample set will fall into the right child node; otherwise, it will fall into the left one. The sample can eventually enter a leaf node.

2.3. Aggregation of SCA Ensembles Based on Streamflow Autocorrelation

Streamflow autocorrelation can be addressed through the mapping from the predictor values to the Wilks Λ ones. Variations in streamflow over multiple consecutive time steps can be directly reflected in the outputs of multi-output SCAs. Using a dual-output SCA as an example (Figure 1), SCA outputs include the streamflow simulations for time t ($\boldsymbol{Q}_t = \{\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_n\}$, $t \in 1, 2, ..., n$, and n is the number of time steps to be simulated) and those for $t + 1$ ($\boldsymbol{Q}'_t = \{\boldsymbol{Q}'_2, \boldsymbol{Q}'_3, \ldots, \boldsymbol{Q}'_{n+1}\}$, $t \in 2, 3, ..., n + 1$). For the above expressions, $\boldsymbol{Q}_2$ and $\boldsymbol{Q}'_2$ are streamflow values of the same time ($t = 2$). In detail, $\boldsymbol{Q}_2$ carries the information of potential interactions with $\boldsymbol{Q}'_3$, while $\boldsymbol{Q}'_2$ represents the potential interactions with $\boldsymbol{Q}_1$. The expected dual-output SCA results ($\boldsymbol{E}_t$) for such a single tree can then be expressed as $\boldsymbol{E}_t = \boldsymbol{Q}_t$ when $t = 1$, and $\boldsymbol{E}_t = (\boldsymbol{Q}_t + \boldsymbol{Q}'_t)/2$ when $t \in 2, 3, ..., n$. Thus, the $\boldsymbol{E}_t$ can carry the information of potential interactions among streamflow values across multiple time steps ($t = n - 1$, n and $n + 1$). The dual-output SCA results obtained for all trees (i.e., dual-output SCA ensemble) can be expressed as $D_t = \frac{1}{N} \sum_{\text{Tree}=1}^{N} E_t^{\text{Tree}}$.

Since streamflow autocorrelation varies in time and space, a multi-output SCA ensemble may not always outperform a single-output one. Therefore, the three SCA ensembles are aggregated through the Bayesian model averaging with a stratified sampling (BMASS) approach. Different from conventional BMA (Raftery et al., 2005), the BMASS applies the BMA to various quantile ranges (Duan et al., 2007; Zhu et al., 2015). Generally, the BMASS involves two steps: (1) estimation of stratified BMA weights through the expectation and maximization algorithm (Dempster et al., 1977), and (2) aggregation of three SCA ensemble simulations based on the BMA weights. In the first step, the streamflow observations over the training period are categorized into $\boldsymbol{K}$ streamflow classes (i.e., intervals) based on a given set of quantile ranges (e.g., 0-tenth, 10-twentieth, …, 90–100th). For each instance (i), the streamflow observations and the corresponding validation results for the three SCA ensembles belong to a particular streamflow class $C_i^{\text{Obs(training)}} \in [1, 2, ..., K]$. Therefore, each class contains a number of observed streamflow values and the corresponding validation results. BMA is then performed for each streamflow class to obtain the BMA weights (Duan et al., 2007). In the second step, daily means of the three SCA ensemble simulations over the testing period are used for categorizing the $\boldsymbol{K}$ streamflow classes ($C_i^{\text{Sim(testing)}} \in [1, 2, ..., K]$), based on the same set of quantile ranges in the first step. The BMA simulation for a particular streamflow class is generated from the weighted average of the three SCA ensembles based on the corresponding BMA weights. Lastly, the resulting SCRTE is the combined BMA simulations for all quantile ranges.

3. Application

3.1. The Ancient Yellow River Irrigation System

Three interconnected watersheds, located at the upper reach of China's Yellow River, were investigated using the proposed SCRTE. The landscape of the study area (Figure 3) has been largely modified by farming and urbanization since B.C. 215 of Qin Dynasty (Zhang & Deng, 1987). In the irrigation system, the ground is lower than the river surface, allowing a diversion dam to channel water into irrigation canals. The total length of irrigation canals

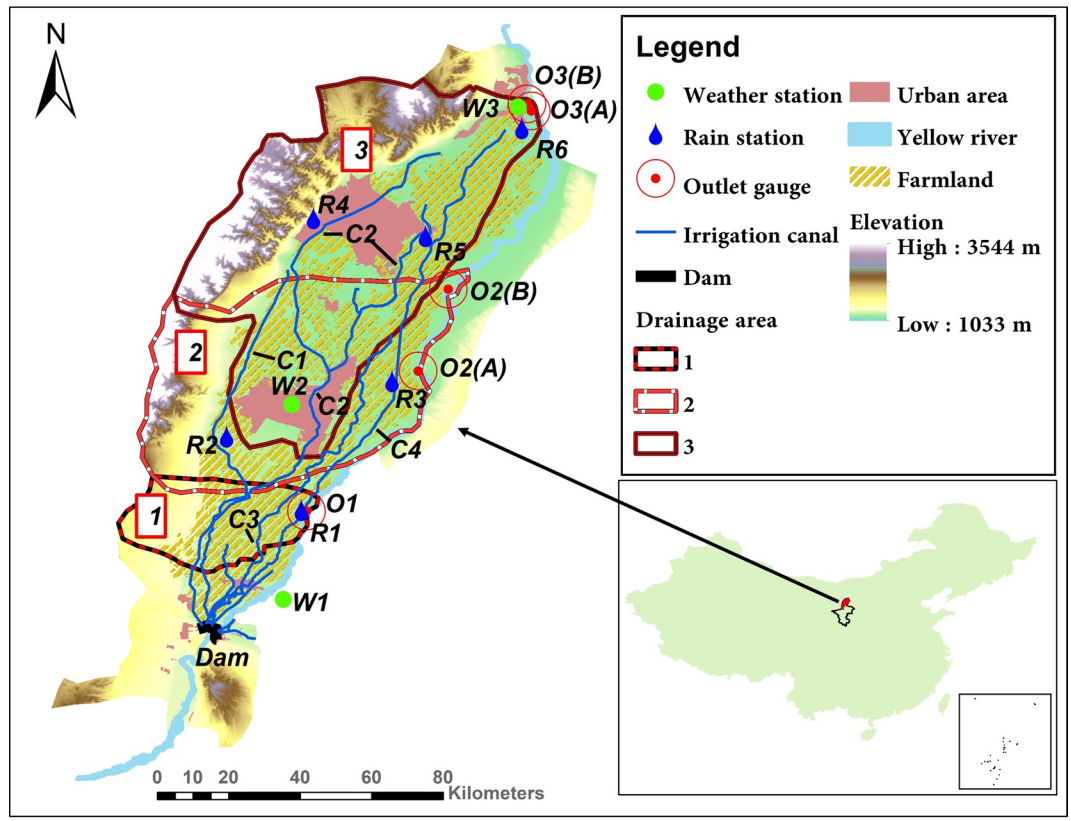

Figure 3. Map of the study area. Note that ***W*** denotes weather stations, ***R*** denotes rain stations, ***G*** denotes groundwater level gauges, ***C*** denotes irrigation canals and ***O*** denotes drainage gauges. Both the second and the third drainage areas contain two interwoven drainages with strong hydrological connections.

is over 1,200 km, which has been acknowledged as the oldest and largest irrigation system in China's northwest region (International Commission on Irrigation and Drainage, 2017).

Most of the lands have been relying on flood irrigation to take advantage of the nutrient-abundant sediment from the Yellow River (Dong et al., 2015). The irrigation can be divided into two periods: Spring irrigation (SI) and Winter irrigation (WI). The SI spans the entire growing season (April to September). The WI, which starts at the end of October and lasts till the end of November, freezes the topsoil for preserving moisture and facilitating the plowing for the next growing season. Due to the scarcity of precipitation (180–200 mm per year) (Yang et al., 2015), the high water demands of crops (e.g., rice, wheat and corn), and the serious water-related conflict between Ningxia and downstream provinces, the local government has been facing great challenges of water shortage, especially in SI period.

Following the strict water regulations in each province (State Council of the People's Republic of China, 2012), the Ningxia Water Resources Conservancy (NWRC) must secure the quantity of transboundary water (for flowing to the adjacent province). The transboundary water mainly consists of the streamflow discharged from the dam and the returned surface/subsurface runoffs after irrigation and precipitation (i.e., return flows). The highly porous soils in most of the irrigated regions (Table S1 in Supporting Information S2) enable a rapid and deep percolation, facilitating irrigation water to recharge the nearby drainages quickly. According to the Ningxia Water Conservancy (2003–2015), during the months with intensive irrigation (May and June), up to 65% of the transboundary water comes from return flows. Therefore, an appropriate estimation of transboundary water has become a primary challenge to the NWRC.

The NWRC currently uses a rule-of-thumb approach to estimate return flows as a fixed proportion of irrigation ones as gauged at the diversion dam. However, such an approach can hardly support effective water allocation under the increasingly rigorous water supply-demand conflicts. Due to the local water accounting regulations,

Table 1
Datasets Used for Each Drainage Basin

Drainage ID	Predictors	Response
First	$C1_i$, $C2_i$, $C3_i$, $R1_j$, $WR1_j$, $WT1_k$, $C(1,2,3)_{15}$, $C(1,2,3)_{30}$	$O1$
Second	$C1_i$, $C2_i$, $C3_i$, $C4_i$, $R2_j$, $R3_j$, $R5_j$, $WR2_j$, $WT2_k$, $C(1,2,3,4)_{15}$, $C(1,2,3,4)_{30}$	$O2(A) + O2(B)$
Third	$C1_i$, $C2_i$, $C4_i$, $R4_j$, $R5_j$, $R6_j$, $WR2_j$, $WT2_k$, $WT3_k$, $C(1,2,4)_{15}$, $C(1,2,4)_{30}$	$O3(A) + O3(B)$

Note. $WR1$ and $WT1$ denote the precipitation and air temperature monitored by the weather station $W1$, respectively. Subscripts i, j and k denote the moving windows for irrigation, precipitation and temperature, respectively; where $i \in [1, 3, 5, 7]$, $j \in [1, 3, 5]$ and $k \in [3, 5]$. For example, $C1_3 = C1(t) + C1(t-1) + C1(t-2)$; where $C1(t)$, $C1(t-1)$ and $C1(t-2)$ are the daily irrigation flows for the $C1$ canal original, previous-day and previous-2-day time series, respectively. The 15-day and 30-day moving sum irrigations (e.g., $C(1,2,3)_{15}$) are derived from the average irrigation of all canals considered in the corresponding drainage basin. The time series of $R6$ is identical to $WR3$ due to their close distance (Figure 3). Therefore, only $R6$ was used for modeling the third drainage basin.

the return flows are not allowed to be reused arbitrarily. Nevertheless, an accurate estimation of return flows is still challenging because they vary dramatically with climate factors, crop types, irrigation techniques, land slopes, soil properties, and land uses (Dewandel et al., 2008; Kim et al., 2009; Yalcin, 2019). For instance, complicated irrigation schedules (especially for the allocation of insufficient irrigation water and the control of flood; Li et al., 2019), compounded with intensive precipitation during July to September (Yang et al., 2015), pose a substantial challenge to return flow estimation. Under such complexities, the estimation of return flows should take into account climatic factors (e.g., precipitation and temperature), irrigation schedule, and surface/subsurface responses to irrigational and climatic events.

The study area can be delineated into three drainage basins (Figure 3). Each of these basins is individually modeled. The spatial connections among these basins can be implicitly reflected by overlapped modeling inputs. The datasets of daily time series are obtained from the NWRC (2003–2015). In the modeling process, the cropping pattern and irrigation method generally remain stable. Therefore, it is reasonable to assume that there is a relatively stationary relationship between the total water received (i.e., the sum of irrigation and precipitation) and the total runoff generated. There are five drainage gauges for monitoring return flows to the three drainage basins, four irrigation canal gauges for measuring the daily irrigation flows at the canal inlets, six rain stations and three weather stations for monitoring precipitation and air temperature (Figure 3). Previous studies have successfully used the simulated crop water demands as predictors to predict the hydrological fluxes in irrigated watersheds (Mohan & Vijayalakshmi, 2009; Sahoo et al., 2017). However, the simulated crop water demands in previous studies are oversimplified spatially and temporally, thereby affecting further modeling efforts. In this study, daily irrigation flows gauged at irrigation canals (Figure S1 in Supporting Information S1) are used as an improved approximation to the basin-wide crop water demand because the flows are scheduled based on the crop water demands collected from irrigators who usually have accurate judgments on their crops. In detail, the irrigator association initially collects the most recent water demand from irrigators and then reports the water demand to the NWRC. Based on such information, NWRC can then schedule the daily irrigation flow.

Backward sums with multiple moving windows are used to generate time series information (of each drainage basin) to reflect the initial catchment conditions. Table 1 shows the predictors (with backward sums) and predictands for each drainage basin.

3.2. Evaluation Metrics

In order to assess how the flow characteristics can be addressed by the SCRTE, seven error metrics were employed in this study, including the coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), Kling-Gupta efficiency (KGE) (Gupta et al., 2009), Nash–Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970), Log Nash-Sutcliffe efficiency (LogNSE) (Krause et al., 2005), and Volumetric efficiency (VE) (Criss & Winston, 2008).

The R^2 measures how well the simulated values and observations can be fitted into a linear regression model, and can be expressed as:

$$R^2 = \left[\sum_{n=1}^{N} (y_n - \overline{y})(y_n^* - \overline{y^*}) \Big/ \sqrt{\sum_{n=1}^{N} (y_n - \overline{y})^2} \sqrt{\sum_{n=1}^{N} (y_n^* - \overline{y^*})^2} \right]^2 \quad (9)$$

where y_n and y_n^* are the nth observed and simulated daily streamflow values, respectively; $\overline{y}$ is the mean of observed daily streamflow time series; $\overline{y^*}$ is the mean of simulated daily streamflow time series; N is the number of observations.

The MAE is a measure of difference between observations and simulations, and can be expressed as:

$$\text{MAE} = \frac{1}{n} \sum_{n=1}^{N} |y_n - y_n^*| \quad (10)$$

The RMSE is the square root of the average squared difference between observations and simulations, and is expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{n=1}^{N} (y_n - y_n^*)^2} \quad (11)$$

The KGE is a combined measure of correlation, standard deviation and mean squared error, which is sensitive to the flow variability.

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (12)$$

where r equals to the square root of R^2, α denotes the measure of relative variability in the simulated and observed values ($\alpha = \sigma_{y^*}/\sigma_y$, in which σ denotes the variance of samples), and β represents the ratio of simulated and observed flow means ($\beta = \overline{y^*}/\overline{y}$).

The NSE is the most commonly used metric to assess the predictive power of hydrological models. It ranges from $-\infty$ to 1, and is defined as:

$$\text{NSE} = 1 - \sum_{n=1}^{N} (y_n - y_n^*)^2 \Big/ \sum_{n=1}^{N} (y_n - \overline{y})^2 \quad (13)$$

The LogNSE is the NSE with natural logarithm transformed streamflow, and is sensitive to the low flows of the hydrograph (Krause et al., 2005).

$$\log\ \text{NSE} = 1 - \sum_{n=1}^{N} (\ln\ y_n - \ln\ y_n^*)^2 \Big/ \sum_{n=1}^{N} (\ln\ y_n - \ln \overline{y})^2 \quad (14)$$

The VE is an index of how well the simulated volume matches the observed one over a given time interval (Criss & Winston, 2008), and it ranges from 0 to 1.

$$\text{VE} = 1 - \left| \sum_{n=1}^{N} (y_n - y_n^*) \right| \Big/ \sum_{n=1}^{N} y_n \quad (15)$$

3.3. Training, Validation and Testing Procedures

In this study, SCRTE was compared against seven well-known data-driven models, including random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM), multilayer perceptron neural network (MLP), long short-term memory (LSTM), sparse regression model (SLM), and generalized linear model (GLM). The inputs of daily time series were divided into two subsets, including one from 01/01/2003 to 31/12/2011 for model training, and the other from 01/01/2012 to 31/12/2015 for the purpose of testing. All of the models were independently trained for the SI (April to September) and WI (October to March) periods in the three drainage

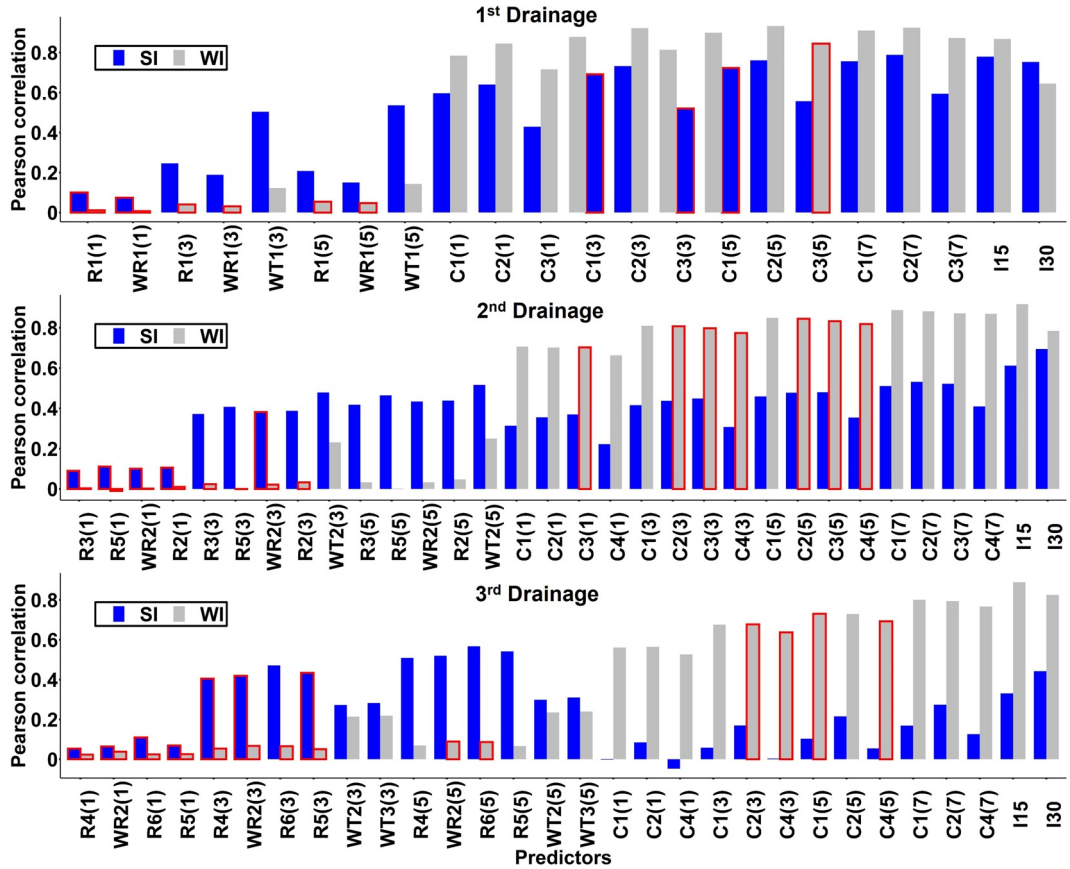

Figure 4. Pearson correlations between predictors and streamflow. Note that the bars in red outline represent the eliminated predictors (for ***S***, ***D*** or ***T***) during the RFE process.

basins. The 10-fold cross-validation scheme (Stone, 1974) was applied to the models (except for SCRTE and RF). It uses the data over the training period to generate 10 dataset groups, each of which contains 90% of the (randomly sampled) dataset for training and 10% for validation. Through modeling with these groups, desired model configuration (based on the validation dataset) can be identified for further testing. SCRTE and RF were associated with the out-of-bag (OOB) datasets (as mentioned in Section 2) for the validation purpose. In addition, these two models used the recursive feature elimination (RFE) strategy (Guyon et al., 2002) to identify the desired model configuration. The RFE strategy in this study followed the work of Schmidt et al. (2020), which started from training a model with all predictors, and then the three least relevant predictors that made little improvement on the RMSE (over the OOB datasets) were removed at each iteration. The model with the lowest RMSE (among iterations) was used for further testing. The other six error metrics were calculated to provide additional information on the modeling performance. The Pearson correlation analysis was conducted to examine the correlation between predictors and streamflow (Figure 4). Higher correlations between irrigation water and streamflow can be found in the WI periods than in the SI ones. In contrast, the correlations between rainfall and streamflow are higher in SI than those in WI. The bars with a red outline represent the eliminated predictors (for ***S***, ***D*** or ***T***) during the RFE process.

Five hyperparameters need to be determined for training the SCRTE, including the significance level (α) in F-test during the node splitting process, the generated flow quantile ranges (*Qr*), the number of bootstrap replicates (*Nrep*), the minimum number of samples in a node for splitting action (*Nmin*), and the number/ratio of features in a subspace (*Mtry*). In this study, the α value was set to 0.05 as suggested by Huang(1992). The *Qr* was designated as the 0–25th quantile range for baseflow; 25–50th for low flow; 50, 55, …, 95th for mid and high flows; 99–99.5th and 99.5–100th for peak flows. The *Nrep* was set as 200, after which no further improvement in accuracy can be achieved. The *Nmin* was set as 5 to balance overfitting and robustness. The *Mtry* was set as 50% as suggested by Barandiaran (1998), indicating that a half of the predictors were selected for each tree.

RF and XGB are regression tree ensemble approaches. The regression trees in RF are aggregated using average simulated values, whereas the XGB sequentially aggregates the trees based on the errors learned from the previous aggregation. The default parameters, as described in the R statistical computing software (hereinafter referred to as R software) (R Development Core Team, 2019) with the package "randomForest" (Liaw & Wiener, 2002), were used to train the RF. The XGB was trained using the R package "XGBoost" (Chen et al., 2015), and two modeling parameters (i.e., *eta* and *maximum depth*) were optimized using the grid search method (Gilli et al., 2019).

SVM is a classification and regression approach that uses multiple sets of decision boundaries in the predictor space to divide observations into different classes. The selection of appropriate kernel functions is critical for generating desired modeling performance (Deka, 2014). In this study, the SVM model was trained using the R package "e1071" (Meyer et al., 2019), and its Kernel function parameters were optimized using the grid search method.

MLP is a well-established artificial neural network that consists of a network of nodes and links between predictors and predictions. Each link in the network represents a function that maps the input nodes into the output one(s) (Ripley, 2007). MLP was trained using the stochastic gradient descent backpropagation as a learning function with the R package "RSNNS" (Bergmeir & Benítez Sánchez, 2012). The training data were fed into the MLP model with a maximum of 1,000 times (i.e., epochs = 1,000). Two hyperparameters, including the learning rate and the number of units in the hidden layer(s), were optimized using the grid search method.

The LSTM is a type of recurrent neural network and has been successfully used for hydrological simulations (Kratzert et al., 2018). Each LSTM layer uses a memory cell and three gates to control information in the hidden neuron. In this study, the LSTM model was evaluated using Keras 2.2.5 (Allaire & Chollet, 2019) and TensorFlow 2.0.0 (Allaire & Tang, 2019) with the R software. The model comprised two stacked LSTM layers with a dropout rate of 0.5 and a neuron size of 16. The model was established with a sequence-to-value mode to predict a daily streamflow value based on the meteorological and irrigation data of the past 30 days, which made the input structure different from those of other benchmark models. The training dataset was fed into the LSTM with 1,000 iterations (i.e., epochs = 1,000) using the loss function of MSE, and the LSTM model configuration with the lowest MSE was used for further testing.

SLM and GLM are the two main types of linear regression. Compared to GLM, SLM selects a subset of predictors and then evaluates the least-square fit of all possible sub-models (i.e., models trained with a subset of predictors) in order to identify the optimal one that provides the best fit. SLM and GLM were fitted using the R package "scalreg" (Sun, 2019) and a basic statistical package in the R software, respectively.

3.4. Results

3.4.1. Evaluation of Modeling Performance

The training, validation and testing results are shown in Tables 2–4. In general, all of the models achieved reasonable accuracy over the testing period for all drainage basins. Notably, SCRTE achieved the overall best performance for six out of seven error metrics when compared with the other benchmark models (except the KGE, for which SCRTE and GLM have the same level of accuracy). The combined use of error metrics demonstrates that SCRTE can address various aspects of flow characteristics (e.g., flow variations and volumes). However, the benchmark models are focused on specific aspects of flow characteristics. For instance, the overall KGE value for the GLM is as high as that for SCRTE, but the LogNSE value for GLM is the second lowest (Table 4). Table 3 shows that, even though the overall validation performance of LSTM (with the convergence test in Figure S2 in Supporting Information S1) is comparable to that of SCRTE, the overall testing performance of LSTM is much lower than that of SCRTE (Table 4). In fact, recent studies have indicated that deep learning methods such as LSTM may require a tremendous amount of training data for the long-term streamflow simulation (Bai et al., 2021; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019). For the cases with limited training data, LSTM may not necessarily outperform other data-driven models (Duan et al., 2020; Wunsch et al., 2021; Xiang et al., 2020).

The overall performance of RF tends to be the best among the benchmark models based on the metrics of MAE, RMSE, NSE, VE and LogNSE. These findings agree with the previous studies of Shortridge et al. (2016), Schmidt

Table 2
Modeling Performances for the Training Dataset

Model		1st		2nd		3rd		
		SI	WI	SI	WI	SI	WI	Overall
R^2	SCRTE	0.93 ± 0.1%	0.97 ± 0.0%	0.94 ± 0.1%	0.97 ± 0.0%	0.94 ± 0.1%	0.96 ± 0.0%	0.95 ± 0.0%
	RF	**0.98 ± 0.0%**	**0.99 ± 0.0%**	**0.98 ± 0.0%**	**0.99 ± 0.0%**	**0.98 ± 0.0%**	**0.99 ± 0.0%**	**0.98 ± 0.0%**
	XGB	0.89 ± 0.1%	0.98 ± 0.0%	0.88 ± 0.1%	0.96 ± 0.0%	0.86 ± 0.1%	0.95 ± 0.1%	0.92 ± 0.1%
	SVM	0.83	0.94	0.78	0.93	0.72	0.90	0.85
	MLP	0.76 ± 0.3%	0.91 ± 0.2%	0.68 ± 0.6%	0.90 ± 0.4%	0.61 ± 0.7%	0.88 ± 0.3%	0.79 ± 0.4%
	SLM	0.74	0.89	0.63	0.86	0.48	0.83	0.74
	GLM	0.74	0.90	0.65	0.90	0.51	0.87	0.76
	LSTM	0.87 ± 0.4%	0.95 ± 0.2%	0.87 ± 0.9%	0.94 ± 0.5%	0.83 ± 1.2%	0.94 ± 0.5%	0.90 ± 0.6%
MAE	SCRTE	1.20 ± 0.4%	0.54 ± 0.4%	1.36 ± 0.4%	0.66 ± 0.4%	1.55 ± 0.7%	0.87 ± 0.4%	1.03 ± 0.5%
	RF	**0.64 ± 0.2%**	**0.38 ± 0.2%**	**0.70 ± 0.1%**	**0.41 ± 0.2%**	**0.81 ± 0.3%**	**0.53 ± 0.2%**	**0.58 ± 0.2%**
	XGB	1.53 ± 0.3%	0.50 ± 0.2%	1.84 ± 1.2%	0.75 ± 0.3%	2.10 ± 0.9%	1.01 ± 0.4%	1.29 ± 0.6%
	SVM	1.71	0.69	2.18	0.84	2.60	1.25	1.55
	MLP	2.16 ± 3.4%	0.86 ± 0.6%	2.86 ± 5.3%	1.05 ± 1.7%	3.42 ± 2.6%	1.51 ± 3.8%	1.98 ± 2.9%
	SLM	2.32	1.65	3.17	2.58	4.08	2.22	2.67
	GLM	2.30	0.87	3.02	1.05	3.94	1.60	2.13
	LSTM	1.68 ± 4.3%	0.77 ± 0.9%	1.97 ± 7.5%	0.92 ± 2.2%	2.33 ± 3.2%	1.29 ± 4.4%	1.49 ± 3.8%
RMSE	SCRTE	1.70 ± 0.8%	0.87 ± 0.6%	1.84 ± 0.6%	0.89 ± 0.6%	2.08 ± 1.0%	1.29 ± 0.4%	1.44 ± 0.7%
	RF	**0.97 ± 0.6%**	**0.60 ± 0.5%**	**0.99 ± 0.4%**	**0.58 ± 0.3%**	**1.17 ± 0.6%**	**0.82 ± 0.4%**	**0.85 ± 0.5%**
	XGB	2.10 ± 0.5%	0.72 ± 0.3%	2.44 ± 1.3%	1.01 ± 0.5%	2.77 ± 1.1%	1.44 ± 0.9%	1.75 ± 0.8%
	SVM	2.67	1.32	3.23	1.27	3.89	2.27	2.44
	MLP	3.09 ± 1.6%	1.58 ± 2.2%	3.89 ± 4.2%	1.52 ± 2.9%	4.57 ± 2.8%	2.33 ± 1.8%	2.83 ± 2.6%
	SLM	3.19	2.09	4.18	2.94	5.27	3.06	3.46
	GLM	3.18	1.62	4.08	1.51	5.08	2.45	2.99
	LSTM	2.33 ± 1.5%	1.24 ± 3.9%	2.54 ± 5.0%	1.25 ± 3.0%	3.02 ± 3.3%	1.77 ± 2.0%	2.02 ± 3.1%
NSE	SCRTE	0.93 ± 0.1%	0.97 ± 0.0%	0.93 ± 0.0%	0.97 ± 0.0%	0.92 ± 0.1%	0.96 ± 0.0%	0.95 ± 0.1%
	RF	**0.98 ± 0.0%**	**0.99 ± 0.0%**	**0.98 ± 0.0%**	**0.99 ± 0.0%**	**0.97 ± 0.0%**	**0.98 ± 0.0%**	**0.98 ± 0.0%**
	XGB	0.89 ± 0.0%	0.98 ± 0.0%	0.87 ± 0.1%	0.96 ± 0.0%	0.86 ± 0.1%	0.95 ± 0.1%	0.92 ± 0.1%
	SVM	0.82	0.94	0.78	0.93	0.72	0.89	0.85
	MLP	0.76 ± 0.3%	0.91 ± 0.2%	0.68 ± 0.9%	0.90 ± 0.4%	0.62 ± 1.0%	0.88 ± 0.3%	0.79 ± 0.5%
	SLM	0.74	0.84	0.64	0.63	0.48	0.80	0.69
	GLM	0.74	0.90	0.65	0.91	0.52	0.87	0.76
	LSTM	0.86 ± 0.2%	0.95 ± 0.3%	0.86 ± 1.2%	0.94 ± 0.5%	0.83 ± 1.5%	0.94 ± 0.4%	0.90 ± 0.7%
KGE	SCRTE	0.89 ± 0.1%	0.96 ± 0.1%	0.87 ± 0.1%	0.95 ± 0.1%	0.84 ± 0.2%	0.95 ± 0.1%	0.91 ± 0.1%
	RF	**0.93 ± 0.1%**	**0.97 ± 0.1%**	**0.93 ± 0.1%**	**0.97 ± 0.1%**	**0.91 ± 0.1%**	**0.96 ± 0.1%**	**0.95 ± 0.1%**
	XGB	0.89 ± 0.1%	0.98 ± 0.0%	0.86 ± 0.1%	0.96 ± 0.0%	0.84 ± 0.1%	0.95 ± 0.1%	0.91 ± 0.1%
	SVM	0.81	0.89	0.78	0.93	0.71	0.84	0.83
	MLP	0.79 ± 1.7%	0.91 ± 1.8%	0.75 ± 2.1%	0.91 ± 2.4%	0.68 ± 2.0%	0.90 ± 1.7%	0.82 ± 1.9%
	SLM	0.81	0.71	0.75	0.60	0.62	0.79	0.72
	GLM	0.80	0.93	0.73	0.93	0.60	0.90	0.82
	LSTM	0.89 ± 2.3%	0.92 ± 2.8%	0.87 ± 3.2%	0.90 ± 3.7%	0.82 ± 3.6%	0.90 ± 1.9%	0.89 ± 2.9%

Table 2
Continued

Model		1st		2nd		3rd		
		SI	WI	SI	WI	SI	WI	Overall
VE	SCRTE	0.88 ± 0.0%	0.85 ± 0.1%	0.90 ± 0.0%	0.89 ± 0.1%	0.87 ± 0.1%	0.85 ± 0.1%	0.87 ± 0.1%
	RF	**0.94 ± 0.0%**	**0.90 ± 0.0%**	**0.95 ± 0.0%**	**0.93 ± 0.0%**	**0.93 ± 0.0%**	**0.91 ± 0.0%**	**0.93 ± 0.0%**
	XGB	0.84 ± 0.0%	0.87 ± 0.1%	0.86 ± 0.1%	0.87 ± 0.0%	0.83 ± 0.1%	0.83 ± 0.1%	0.85 ± 0.1%
	SVM	0.83	0.81	0.84	0.85	0.79	0.79	0.82
	MLP	0.78 ± 0.3%	0.77 ± 0.2%	0.79 ± 0.4%	0.82 ± 0.3%	0.72 ± 0.3%	0.75 ± 0.7%	0.77 ± 0.3%
	SLM	0.77	0.56	0.77	0.55	0.66	0.63	0.66
	GLM	0.77	0.77	0.78	0.82	0.68	0.73	0.76
	LSTM	0.83 ± 0.2%	0.81 ± 0.3%	0.85 ± 0.5%	0.85 ± 0.2%	0.81 ± 0.7%	0.81 ± 0.3%	0.83 ± 0.4%
LogNSE	SCRTE	0.95 ± 0.0%	0.91 ± 0.1%	0.93 ± 0.0%	0.86 ± 0.3%	0.84 ± 0.3%	0.73 ± 0.2%	0.87 ± 0.2%
	RF	**0.99 ± 0.0%**	**0.95 ± 0.0%**	**0.98 ± 0.0%**	**0.93 ± 0.1%**	**0.94 ± 0.1%**	**0.81 ± 0.2%**	**0.93 ± 0.1%**
	XGB	0.91 ± 0.1%	0.90 ± 0.0%	0.86 ± 0.1%	0.84 ± 0.1%	0.74 ± 0.2%	0.70 ± 0.1%	0.83 ± 0.1%
	SVM	0.91	0.89	0.82	0.83	0.60	0.67	0.79
	MLP	0.84 ± 0.5%	0.85 ± 0.3%	0.73 ± 0.8%	0.76 ± 0.7%	0.46 ± 1.3%	0.59 ± 1.4%	0.70 ± 0.8%
	SLM	0.48	−6.01	0.58	−9.67	0.19	−0.93	−2.56
	GLM	0.65	0.75	0.69	0.77	0.31	0.59	0.63
	LSTM	0.91 ± 0.4%	0.85 ± 0.4%	0.84 ± 0.9%	0.80 ± 1.2%	0.69 ± 2.1%	0.73 ± 1.9%	0.80 ± 1.2%

Note. The bold number represents the best performance of the evaluation metric for each drainage basin and in each season. Models with random effects (i.e., the bootstrap sampling for SCRTE, RF, XGB and random weights initialization for MLP and LSTM) ran for 10 times to obtain the error bar (i.e., standard deviation) of each metric.

et al. (2020) and Erdal and Karakurt (2013) who reported that RF had the best overall performance compared with many other data-driven models. Figure 5 illustrates the training, validation and testing performances of SCRTE as compared with RF (i.e., the best benchmark model). Even though the overall performance of SCRTE is no better than that of RF for both training and validation datasets, the accuracy of SCRTE is better than that of RF over the testing period. In fact, SCRTE also outperformed RF for each of the individual cases over the testing period (Table 4). These results infer that the statistical-test-based tree deduction process as embedded in SCRTE could be a better alternative to the regression-based process as embedded in RF.

3.4.2. Analysis of Hydrological Extremes

The simulated hydrographs over the SCRTE testing, training and validation periods are depicted in Figure 6, Figures S3 and S4 in Supporting Information S1 respectively. For the testing period, SCRTE can well capture the streamflow dynamics in all drainage basins. However, the results for the first and third drainage basins show an underestimation of peak flows over the testing period. This is mainly caused by the limited training data and the algorithmic limitations inherent in the tree-structured models. Specifically, each simulation effort can be regarded as a process of seeking the outcomes from SCA trees based on associated rules (e.g., $(x_1 > 60) \wedge (150 < x_2 < 200) \wedge \ldots \wedge (x_m > 30)$) and then averaging those outcomes. Each rule (as indicated in Figure 1) can be regarded as a combination of events (e.g., upstream rainfall exceeds 60 mm/day compounded with a daily average irrigation flow greater than 150 m^3/s and smaller than 200 m^3/s). The ideal case is that the rules identified in SCA trees can represent all potential causes of flood events. However, due to the data-coverage limitations and large uncertainties of hydrological extremes, the limited number of rules can hardly fully represent the complexity of future hydrological extremes. Namely, future hydrological extremes probably fall outside the domain of the knowledge learned through previous experiences (i.e., training datasets). In the first drainage basin, the flood events that exceed 30 m^3/s over the training period (i.e., 22/05/2007 and 18/06/2007) were caused by the compound events of precipitation (with the maximum 24-hr rainfall exceeding 2- and 2.5-year return periods during the two flood events, respectively) and irrigation (with the maximum daily irrigation flows exceeding 63rd and 66th quantile during the two flood events, respectively). Nevertheless, the flood extreme during the

Table 3
Modeling Performances for the Validation Dataset

Model		1st		2nd		3rd		
		SI	WI	SI	WI	SI	WI	Overall
R^2	SCRTE	0.86 ± 0.1%	0.93 ± 0.1%	0.86 ± 0.1%	0.94 ± 0.1%	0.85 ± 0.1%	0.92 ± 0.1%	**0.90 ± 0.1%**
	RF	**0.87 ± 0.1%**	0.94 ± 0.1%	**0.88 ± 0.1%**	**0.95 ± 0.0%**	**0.86 ± 0.2%**	**0.93 ± 0.1%**	**0.90 ± 0.1%**
	XGB	0.81 ± 0.3%	0.93 ± 0.3%	0.80 ± 0.3%	0.92 ± 0.1%	0.76 ± 0.2%	0.90 ± 0.2%	0.86 ± 0.2%
	SVM	0.83	0.95	0.72	0.88	0.71	0.87	0.83
	MLP	0.77 ± 3.8%	0.91 ± 2.7%	0.64 ± 5.8%	0.90 ± 3.0%	0.51 ± 10.2%	0.88 ± 3.2%	0.77 ± 4.8%
	SLM	0.76	0.92	0.63	0.86	0.51	0.82	0.75
	GLM	0.76	0.92	0.62	0.88	0.50	0.83	0.75
	LSTM	0.83 ± 4.2%	**0.95 ± 3.6%**	0.87 ± 8.7%	**0.95 ± 4.9%**	0.83 ± 13.2%	**0.93 ± 7.4%**	0.89 ± 7.0%
MAE	SCRTE	1.66 ± 0.5%	0.72 ± 0.5%	1.92 ± 0.7%	0.83 ± 0.6%	2.17 ± 1.1%	1.18 ± 0.6%	1.41 ± 0.7%
	RF	**1.55 ± 0.4%**	**0.67 ± 0.3%**	**1.73 ± 0.6%**	**0.72 ± 0.3%**	**1.99 ± 1.0%**	**1.07 ± 0.3%**	**1.29 ± 0.5%**
	XGB	1.93 ± 0.9%	0.71 ± 0.7%	2.31 ± 1.4%	0.92 ± 0.6%	2.65 ± 1.2%	1.29 ± 0.8%	1.64 ± 0.9%
	SVM	1.60	0.54	2.27	0.85	2.50	1.14	1.48
	MLP	1.94 ± 11.8%	0.70 ± 3.3%	2.64 ± 9.1%	0.95 ± 5.3%	3.12 ± 11.0%	1.33 ± 6.0%	1.78 ± 7.8%
	SLM	2.10	1.41	2.87	2.41	3.79	2.08	2.44
	GLM	2.11	0.68	2.75	0.89	3.74	1.42	1.93
	LSTM	1.79 ± 14.8%	1.05 ± 4.7%	2.00 ± 12.6%	0.91 ± 6.6%	2.50 ± 17.3%	1.32 ± 9.0%	1.60 ± 10.8%
RMSE	SCRTE	2.40 ± 1.0%	1.30 ± 0.9%	2.63 ± 1.1%	1.17 ± 0.8%	2.95 ± 1.5%	1.84 ± 0.8%	2.05 ± 1.0%
	RF	**2.30 ± 0.6%**	1.28 ± 1.0%	**2.44 ± 0.8%**	**1.11 ± 0.4%**	**2.82 ± 1.5%**	**1.78 ± 0.6%**	**1.96 ± 0.8%**
	XGB	2.71 ± 1.9%	1.31 ± 2.9%	3.10 ± 2.4%	1.33 ± 1.0%	3.54 ± 1.7%	2.03 ± 2.2%	2.34 ± 2.0%
	SVM	2.29	0.84	3.10	1.25	3.42	1.66	2.09
	MLP	2.69 ± 16.7%	1.22 ± 11.0%	3.53 ± 13.0%	1.30 ± 6.5%	4.10 ± 10.3%	1.85 ± 10.0%	2.45 ± 11.3%
	SLM	2.73	1.58	3.60	2.71	4.75	2.59	2.99
	GLM	2.74	1.06	3.51	1.25	4.65	1.90	2.52
	LSTM	2.42 ± 20.0%	**1.16 ± 16.3%**	2.53 ± 16.0%	1.17 ± 9.0%	3.15 ± 15.6%	1.81 ± 10.1%	2.04 ± 14.5%
NSE	SCRTE	0.93 ± 0.1%	0.97 ± 0.0%	0.93 ± 0.0%	0.97 ± 0.0%	0.92 ± 0.1%	0.96 ± 0.0%	0.95 ± 0.1%
	RF	**0.98 ± 0.0%**	**0.99 ± 0.0%**	**0.98 ± 0.0%**	**0.99 ± 0.0%**	**0.97 ± 0.0%**	**0.98 ± 0.0%**	**0.98 ± 0.0%**
	XGB	0.89 ± 0.0%	0.98 ± 0.0%	0.87 ± 0.1%	0.96 ± 0.0%	0.86 ± 0.1%	0.95 ± 0.1%	0.92 ± 0.1%
	SVM	0.82	0.94	0.78	0.93	0.72	0.89	0.85
	MLP	0.76 ± 0.3%	0.91 ± 0.2%	0.68 ± 0.9%	0.90 ± 0.4%	0.62 ± 1.0%	0.88 ± 0.3%	0.79 ± 0.5%
	SLM	0.74	0.84	0.64	0.63	0.48	0.80	0.69
	GLM	0.74	0.90	0.65	0.91	0.52	0.87	0.76
	LSTM	0.83 ± 4.6%	0.95 ± 3.1%	0.86 ± 7.8%	0.95 ± 2.5%	0.83 ± 9.3%	0.93 ± 3.4%	0.89 ± 5.1%
KGE	SCRTE	0.89 ± 0.1%	0.96 ± 0.1%	0.87 ± 0.1%	0.95 ± 0.1%	0.84 ± 0.2%	0.95 ± 0.1%	0.91 ± 0.1%
	RF	**0.93 ± 0.1%**	**0.97 ± 0.1%**	**0.93 ± 0.1%**	**0.97 ± 0.1%**	**0.91 ± 0.1%**	**0.96 ± 0.1%**	**0.95 ± 0.1%**
	XGB	0.89 ± 0.1%	0.98 ± 0.0%	0.86 ± 0.1%	0.96 ± 0.0%	0.84 ± 0.1%	0.95 ± 0.1%	0.91 ± 0.1%
	SVM	0.81	0.89	0.78	0.93	0.71	0.84	0.83
	MLP	0.79 ± 1.7%	0.91 ± 1.8%	0.75 ± 2.1%	0.91 ± 2.4%	0.68 ± 2.0%	0.90 ± 1.7%	0.82 ± 1.9%
	SLM	0.81	0.71	0.75	0.60	0.62	0.79	0.72
	GLM	0.80	0.93	0.73	0.93	0.60	0.90	0.82
	LSTM	0.84 ± 5.0%	0.91 ± 3.3%	0.88 ± 4.4%	0.91 ± 2.8%	0.84 ± 5.9%	0.90 ± 3.7%	0.88 ± 4.2%

Table 3
Continued

Model		1st		2nd		3rd		
		SI	WI	SI	WI	SI	WI	Overall
VE	SCRTE	0.83 ± 0.1%	0.81 ± 0.1%	0.86 ± 0.1%	0.86 ± 0.1%	0.82 ± 0.1%	0.80 ± 0.1%	0.83 ± 0.1%
	RF	**0.84 ± 0.0%**	**0.82 ± 0.1%**	**0.87 ± 0.0%**	**0.87 ± 0.0%**	**0.84 ± 0.1%**	**0.82 ± 0.1%**	**0.84 ± 0.1%**
	XGB	0.80 ± 0.1%	0.81 ± 0.2%	0.83 ± 0.1%	0.84 ± 0.1%	0.78 ± 0.1%	0.78 ± 0.1%	0.81 ± 0.1%
	SVM	0.83	0.84	0.83	0.84	0.79	0.79	0.82
	MLP	0.80 ± 0.9%	0.78 ± 2.0%	0.80 ± 0.8%	0.83 ± 0.8%	0.74 ± 1.4%	0.76 ± 2.2%	0.78 ± 1.3%
	SLM	0.77	0.57	0.78	0.57	0.68	0.62	0.66
	GLM	0.77	0.80	0.79	0.83	0.68	0.74	0.77
	LSTM	0.82 ± 0.5%	0.80 ± 1.0%	0.85 ± 0.2%	0.86 ± 0.5%	0.80 ± 0.6%	0.81 ± 2.0%	0.82 ± 0.8%
LogNSE	SCRTE	0.92 ± 0.1%	0.88 ± 0.1%	0.87 ± 0.1%	0.83 ± 0.3%	0.74 ± 0.4%	0.68 ± 0.2%	0.82 ± 0.2%
	RF	**0.93 ± 0.1%**	**0.89 ± 0.0%**	**0.89 ± 0.1%**	**0.87 ± 0.1%**	**0.77 ± 0.3%**	**0.72 ± 0.2%**	**0.85 ± 0.1%**
	XGB	0.87 ± 0.1%	0.88 ± 0.1%	0.81 ± 0.3%	0.81 ± 0.1%	0.65 ± 0.5%	0.67 ± 0.1%	0.78 ± 0.2%
	SVM	0.90	0.87	0.83	0.77	0.71	0.73	0.80
	MLP	0.86 ± 1.3%	0.82 ± 2.9%	0.76 ± 3.1%	0.74 ± 4.6%	0.48 ± 7.0%	0.59 ± 5.2%	0.71 ± 4.0%
	SLM	0.45	−7.43	0.56	−11.55	0.31	−1.98	−3.27
	GLM	0.72	0.65	0.70	0.76	0.39	0.64	0.64
	LSTM	0.88 ± 0.8%	0.83 ± 1.7%	0.83 ± 3.5%	0.82 ± 5.0%	0.71 ± 8.3%	0.71 ± 7.9%	0.80 ± 4.5%

testing period (i.e., 30/07/2012) was out of the intensive irrigation season (with the maximum daily irrigation flows exceeding the 32nd quantile) and was mainly caused by precipitation (with a 6-year return period). Due to these complexities, the streamflow simulation on 30/07/2012 did not follow the rules of extreme streamflow simulations over the training period, causing an underestimation of peak flow on 30/07/2012.

In the third drainage basin, the flood extremes that occurred on 16/07/2006 (in the training period) were caused by basin-wide precipitation events, with the maximum 24-hr rainfall exceeding a 26-year return period over upstream basins and a 20-year return period over downstream basins. Therefore, the rules for peak-flow simulations might be dominated by the precipitation of either upstream, downstream, or the entire basin. However, the spatial distribution of precipitation for the flood extreme on 02/08/2012 (in the testing period) was different from that of the training period, featuring a rainfall event with an 85-year return period over the upstream and a rainfall event with a 4-year return period over the downstream. Therefore, the rules associated with the simulations on 02/08/2012 might cause an underestimation if the downstream precipitation was considered to be the dominant factor during the training period. Consequently, the results on 02/08/2012 were underestimated because the model aggregated the simulations derived from the rules in all SCA trees, including those resulting in the underestimations.

The overestimations of streamflow at the second drainage basin were due to human activities. Because of the urban expansion and the application of water-saving technologies in this drainage basin (Mi et al., 2020), both irrigation area and irrigated water were reduced. This led to a decreasing trend of daily irrigation flows in two irrigation canals (i.e., ***C2*** and ***C3***) based on the two-sided Mann-Kendall test (Kendall, 1948; Mann, 1945) (with $\boldsymbol{p} = 0.01$ for ***C2***, and $\boldsymbol{p} \leq 0.01$ for ***C3***). Due to the extremely flat surface and highly permeable soils (Table S1 in Supporting Information S2) in the study area, the reduced irrigation flows increased the average groundwater depth from 1.77 m in 2007 to 2.63 m in 2015 (Table S2 in Supporting Information S2). The declining water table further led to a decreased thickness of saturated zones, and then diminished saturated flow. Moreover, the increasing trend of groundwater extraction in urban areas (Table S2 in Supporting Information S2) intensified the groundwater depletion, and further reduced the surface runoff. These trends were beyond the consideration of most data-driven models and might thus cause an overestimation of streamflow simulation in the second drainage basin.

Table 4
Modeling Performances for the Testing Dataset

Model		1st		2nd		3rd		
		SI	WI	SI	WI	SI	WI	Overall
R^2	SCRTE	**0.81 ± 0.1%**	0.92 ± 0.1%	**0.77 ± 0.1%**	0.66 ± 0.2%	**0.60 ± 0.2%**	**0.79 ± 0.3%**	**0.76 ± 0.2%**
	RF	0.80 ± 0.2%	0.91 ± 0.1%	0.74 ± 0.2%	0.66 ± 0.2%	0.54 ± 0.4%	0.78 ± 0.2%	0.74 ± 0.2%
	XGB	0.79 ± 0.3%	0.91 ± 0.2%	0.73 ± 0.5%	0.67 ± 0.2%	0.53 ± 0.6%	0.76 ± 0.4%	0.73 ± 0.4%
	SVM	0.78	**0.93**	0.71	0.64	0.51	0.76	0.72
	MLP	0.77 ± 0.3%	**0.93 ± 0.3%**	0.74 ± 0.5%	0.67 ± 0.4%	0.52 ± 0.5%	0.75 ± 0.9%	0.73 ± 0.5%
	SLM	0.74	0.91	0.72	**0.68**	0.55	0.73	0.72
	GLM	0.72	0.94	0.70	0.67	0.49	0.74	0.71
	LSTM	0.74 ± 0.3%	0.89 ± 0.2%	0.56 ± 0.9%	0.64 ± 0.5%	0.59 ± 0.5%	0.73 ± 0.7%	0.69 ± 0.5%
MAE	SCRTE	**1.74 ± 0.5%**	**0.67 ± 0.3%**	2.56 ± 0.7%	**1.47 ± 0.4%**	**4.14 ± 0.7%**	**1.90 ± 0.8%**	**2.08 ± 0.6%**
	RF	1.78 ± 0.4%	0.70 ± 0.2%	2.69 ± 1.4%	1.51 ± 0.5%	4.21 ± 1.2%	1.94 ± 0.7%	2.14 ± 0.7%
	XGB	1.87 ± 1.5%	0.70 ± 0.4%	2.70 ± 3.1%	1.51 ± 0.7%	4.30 ± 2.4%	2.01 ± 1.5%	2.18 ± 1.6%
	SVM	1.95	0.65	2.39	1.59	4.16	1.98	2.12
	MLP	1.90 ± 4.4%	0.70 ± 1.8%	**2.17 ± 12.1%**	1.75 ± 5.6%	4.51 ± 9.8%	2.35 ± 7.9%	2.23 ± 6.9%
	SLM	2.12	1.68	2.55	2.98	4.22	3.01	2.76
	GLM	2.17	0.64	2.47	1.61	4.33	2.18	2.23
	LSTM	2.31 ± 5.3%	1.02 ± 1.4%	3.25 ± 13.5%	1.66 ± 6.8%	4.33 ± 10.2%	2.34 ± 7.6%	2.49 ± 7.5%
RMSE	SCRTE	**2.48 ± 0.8%**	**1.27 ± 0.7%**	3.34 ± 0.8%	2.29 ± 0.9%	**6.01 ± 0.9%**	**3.08 ± 2.1%**	**3.08 ± 1.0%**
	RF	2.58 ± 0.8%	1.32 ± 0.6%	3.50 ± 1.8%	2.31 ± 1.4%	6.23 ± 2.8%	3.08 ± 0.9%	3.17 ± 1.4%
	XGB	2.64 ± 1.7%	1.39 ± 1.4%	3.61 ± 3.7%	2.28 ± 1.2%	6.30 ± 3.4%	3.33 ± 2.2%	3.26 ± 2.3%
	SVM	2.80	1.18	3.17	2.42	6.51	3.24	3.22
	MLP	2.72 ± 4.6%	1.24 ± 3.5%	**2.88 ± 15.3%**	**2.21 ± 6.1%**	6.80 ± 15.6%	3.69 ± 11.3%	3.26 ± 9.4%
	SLM	2.81	1.89	3.36	3.38	6.31	4.06	3.64
	GLM	2.90	**1.12**	3.27	2.43	6.63	3.41	3.29
	LSTM	3.30 ± 5.4%	2.13 ± 2.2%	4.60 ± 17.2%	2.37 ± 7.0%	6.01 ± 16.0%	3.79 ± 7.3%	3.70 ± 9.2%
NSE	SCRTE	**0.80 ± 0.1%**	0.92 ± 0.1%	0.56 ± 0.2%	0.53 ± 0.4%	**0.55 ± 0.1%**	**0.78 ± 0.3%**	**0.69 ± 0.2%**
	RF	0.78 ± 0.1%	0.91 ± 0.1%	0.52 ± 0.5%	0.52 ± 0.6%	0.51 ± 0.4%	**0.78 ± 0.1%**	0.67 ± 0.3%
	XGB	0.77 ± 0.3%	0.90 ± 0.2%	0.49 ± 1.0%	0.53 ± 0.5%	0.50 ± 0.5%	0.74 ± 0.3%	0.66 ± 0.5%
	SVM	0.74	0.93	0.61	0.48	0.47	0.75	0.66
	MLP	0.76 ± 0.8%	0.92 ± 0.4%	**0.67 ± 3.5%**	**0.56 ± 2.4%**	0.42 ± 2.7%	0.68 ± 2.0%	0.67 ± 2.0%
	SLM	0.74	0.82	0.56	−0.03	0.50	0.61	0.54
	GLM	0.72	**0.94**	0.58	0.47	0.45	0.73	0.65
	LSTM	0.66 ± 0.8%	0.80 ± 0.5%	0.17 ± 3.7%	0.54 ± 2.7%	0.55 ± 3.2%	0.69 ± 2.6%	0.56 ± 2.3%
KGE	SCRTE	0.77 ± 0.2%	**0.96 ± 0.1%**	0.75 ± 0.1%	0.75 ± 0.2%	0.53 ± 0.1%	0.80 ± 0.2%	**0.76 ± 0.2%**
	RF	0.76 ± 0.2%	0.95 ± 0.1%	0.72 ± 0.3%	0.73 ± 0.4%	0.52 ± 0.4%	0.79 ± 0.2%	0.75 ± 0.2%
	XGB	0.75 ± 0.4%	0.94 ± 0.1%	0.71 ± 0.4%	0.74 ± 0.4%	0.53 ± 0.6%	0.74 ± 0.3%	0.74 ± 0.4%
	SVM	0.72	0.95	0.80	0.70	0.48	0.77	0.74
	MLP	**0.85 ± 0.5%**	0.86 ± 1.9%	0.73 ± 3.1%	0.66 ± 3.2%	0.45 ± 2.0%	0.64 ± 2.9%	0.70 ± 2.3%
	SLM	0.82	0.65	0.75	0.37	0.54	0.68	0.63
	GLM	0.80	**0.96**	**0.78**	0.68	0.46	**0.85**	**0.76**
	LSTM	0.65 ± 0.7%	0.74 ± 0.4%	0.63 ± 4.2%	**0.76 ± 3.6%**	**0.57 ± 3.5%**	0.84 ± 3.0%	0.70 ± 2.57%

Table 4
Continued

Model		1st		2nd		3rd		
		SI	WI	SI	WI	SI	WI	Overall
VE	SCRTE	**0.80 ± 0.1%**	0.82 ± 0.1%	0.77 ± 0.1%	0.70 ± 0.1%	**0.70 ± 0.1%**	**0.71 ± 0.1%**	**0.75 ± 0.1%**
	RF	**0.80 ± 0.0%**	0.81 ± 0.0%	0.76 ± 0.1%	0.69 ± 0.1%	**0.70 ± 0.1%**	0.70 ± 0.1%	0.74 ± 0.1%
	XGB	0.79 ± 0.2%	0.81 ± 0.1%	0.76 ± 0.3%	0.69 ± 0.1%	0.69 ± 0.2%	0.69 ± 0.2%	0.74 ± 0.2%
	SVM	0.78	0.82	0.79	0.68	**0.70**	0.70	0.74
	MLP	0.79 ± 0.5%	0.81 ± 0.5%	**0.81 ± 1.1%**	0.64 ± 1.1%	0.67 ± 0.7%	0.64 ± 1.2%	0.73 ± 0.9%
	SLM	0.76	0.55	0.77	0.40	**0.70**	0.54	0.62
	GLM	0.76	**0.83**	0.78	0.67	0.69	0.67	0.73
	LSTM	0.75 ± 0.7%	0.74 ± 0.6%	0.71 ± 1.4%	**0.72 ± 1.3%**	0.69 ± 0.9%	0.67 ± 1.4%	0.71 ± 1.1%
LogNSE	SCRTE	**0.88 ± 0.1%**	**0.87 ± 0.1%**	0.72 ± 0.2%	0.42 ± 0.2%	**0.63 ± 0.2%**	0.55 ± 0.1%	**0.68 ± 0.1%**
	RF	**0.88 ± 0.1%**	0.86 ± 0.0%	0.71 ± 0.2%	0.43 ± 0.1%	0.61 ± 0.2%	0.56 ± 0.1%	0.67 ± 0.1%
	XGB	0.86 ± 0.3%	**0.87 ± 0.1%**	0.71 ± 0.8%	**0.44 ± 0.2%**	0.58 ± 0.6%	0.55 ± 0.2%	0.67 ± 0.4%
	SVM	0.85	0.85	0.74	0.41	0.62	0.56	0.67
	MLP	0.85 ± 0.5%	0.84 ± 0.7%	**0.78 ± 2.0%**	0.37 ± 2.4%	0.56 ± 1.8%	0.51 ± 1.5%	0.65 ± 1.5%
	SLM	0.03	−8.95	0.64	−9.04	0.55	−1.32	−3.02
	GLM	0.13	0.82	0.69	0.38	0.56	0.57	0.53
	LSTM	0.84 ± 0.7%	0.84 ± 1.1%	0.55 ± 3.2%	0.39 ± 2.6%	0.59 ± 1.7%	**0.77 ± 2.9%**	0.66 ± 2.0%

3.4.3. Temporal Autocorrelation of Daily Streamflow

The BMA weights of SCRTE ensemble members for different streamflow quantile ranges, as shown in Figure 7, represent how the autocorrelation of daily streamflows affects the SCRTE performance. In the SI periods, the BMA weights for the single-output SCA ensemble in each drainage basin experienced a drop from the low-to-median flow, and a rise from the median-to-high flow. The multi-output SCA ensembles in each drainage basin showed an opposite trend of BMA weights against the single-output SCA ensemble. The increased dominance of multi-output SCA ensembles indicates that the autocorrelation effects of daily streamflows were more significant under median flows than those under the low and high flows. The results also suggest that the single-output SCA ensemble was superior in simulating low and high flows. In contrast, the multi-output SCA ensembles were effective in addressing the median flows. However, similar patterns of BMA weights were not observed for the WI periods. This was due to the relatively simple irrigation plans and the absence of effective precipitation in WI periods. Consequently, all of the SCA ensembles could reflect the autocorrelation of daily streamflows with a marginal difference, leading to irregular patterns for BMA weights.

To further investigate the autocorrelation effects of daily streamflows, we introduced the lag-1 relative autocorrelative errors/residuals (hereinafter referred to as RAE). The lag-1 autocorrelative error at time $\boldsymbol{t}$ is given as $\varepsilon_t = y_t - \beta_1 \cdot y_{t-1} - \beta_0$, where y_t and y_{t-1} are the time series at time $\boldsymbol{t}$ and $\boldsymbol{t-1}$, respectively; β_1 and β_0 are the regression coefficients. Thus, the lag-1 RAE at time $\boldsymbol{t}$ is calculated as $RAE_t = |\varepsilon_t / y_t| \times 100\%$, which reflects the effect of lag-1 autocorrelation at time $\boldsymbol{t}$. A smaller lag-1 RAE value indicates a more substantial autocorrelation effect and vice versa. As depicted in Figure 7, the median flow has the lowest RAE values in the SI periods, which agrees with the trend for the BMA weights. Compared with the modeling efforts using multi-input single-output relationships (e.g., RF and single-output SCA ensemble), SCRTE can thus effectively improve the performance of simulating median flows while maintaining the performance for low and high flows.

The proposed SCRTE was further compared with its iterated prediction mode. The results indicate that SCRTE and its iterated mode produce identical results for the first and third drainage basins (Figure 8), while the iterated mode significantly underperforms its counterpart for the second basin. The hydrograph, as highlighted in Figure 8, indicates that the streamflows simulated by iterated mode (blue lines) are smoother than those by non-iterated ones. The results imply that the iterated mode can memorize the antecedent state of streamflows and

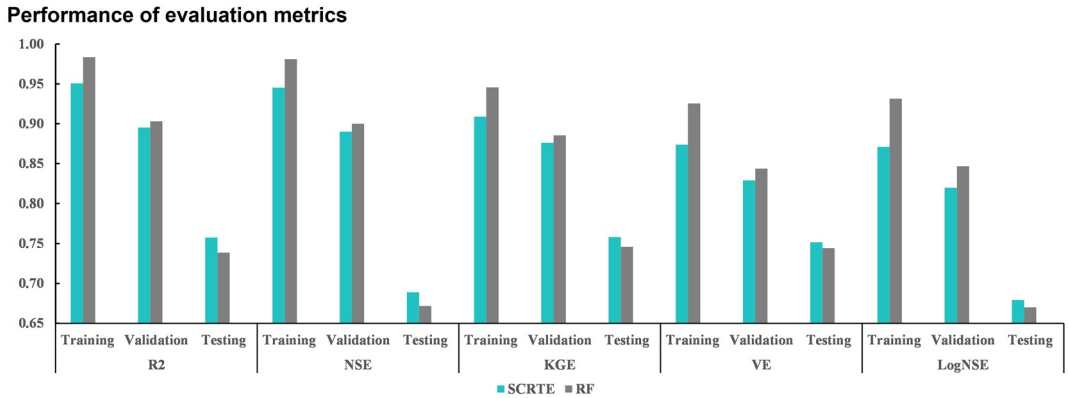

Figure 5. The overall performance of SCRTE and RF for training, validation and testing datasets.

thereby address the streamflow autocorrelation. However, the iterated mode may accumulate simulation errors during the iterative process.

3.4.4. Establishment of Irrigation-Discharge Relationships

An important advantage of the proposed SCRTE is to facilitate the irrigation decision support for securing the quantity of transboundary water. The irrigation-discharge relationship is defined as the relationship between irrigation increment (%) and discharge increment (%). With reliable irrigation-discharge relationships, the

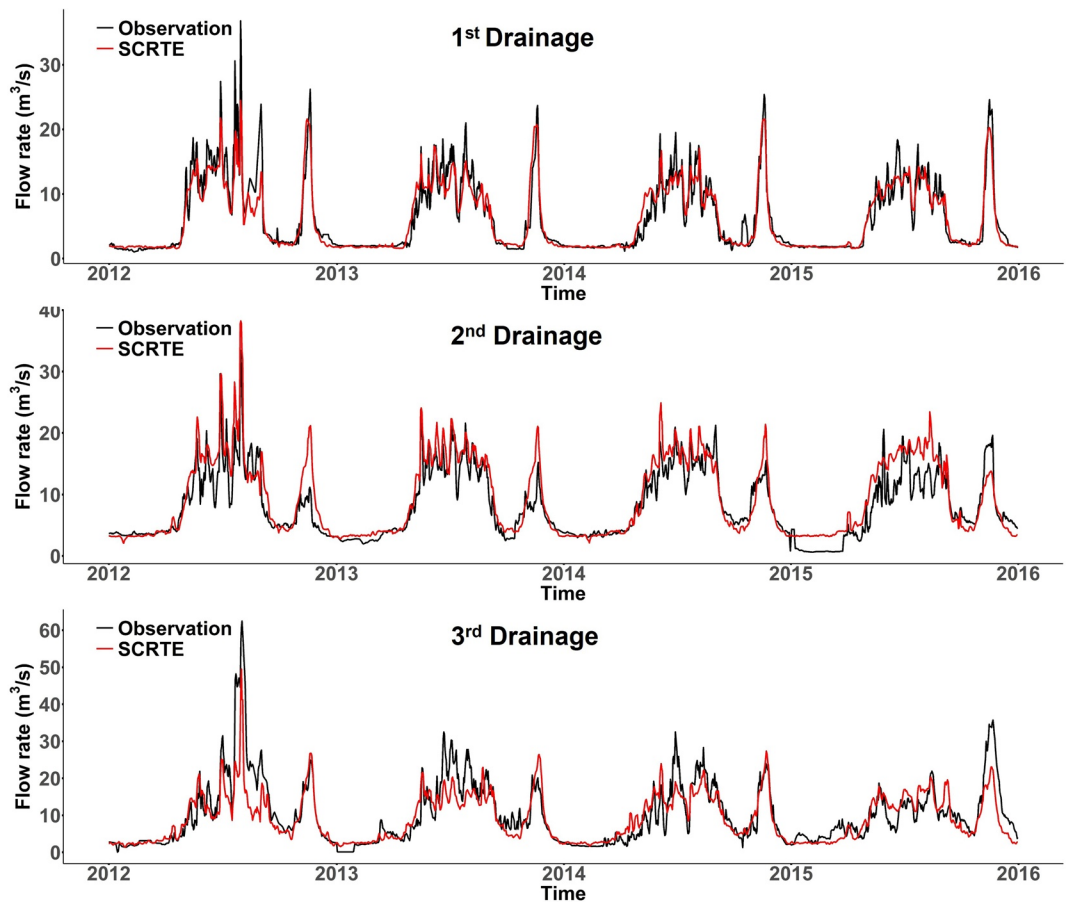

Figure 6. Hydrographs of the SCRTE over the testing period. Note plots from top to bottom show the hydrographs for the first, second, and third drainage basins.

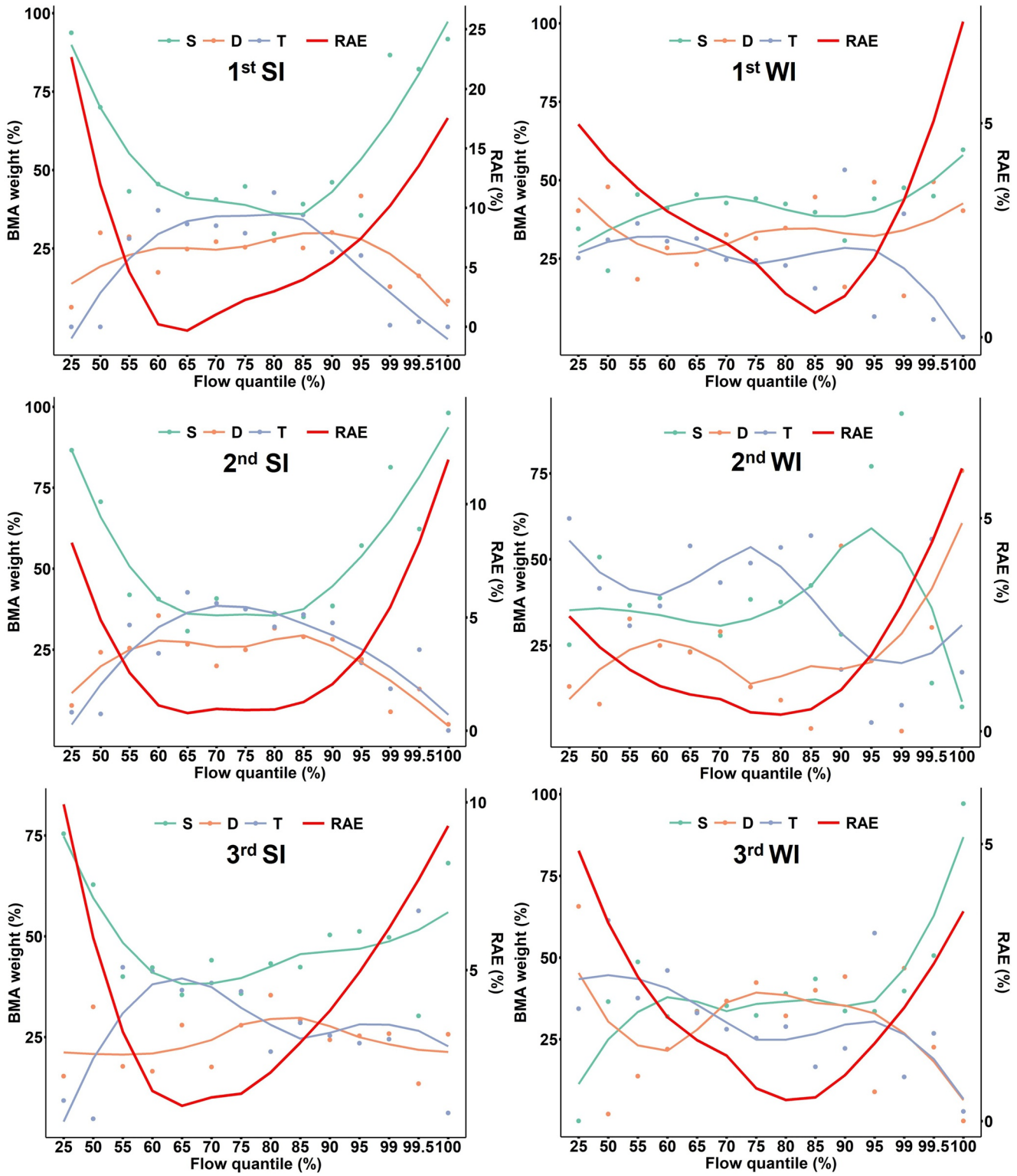

Figure 7. BMA weights of SCA ensembles. Note that the *x*-axis shows the quantile ranges of streamflow (e.g., 50 indicates 25%–50% streamflow quantile range); the sum of BMA weights for each of the quantile ranges is one; the trend lines of BMA weights for the SCA ensembles are drawn based on local weighted regression (i.e., loess) using the R package.

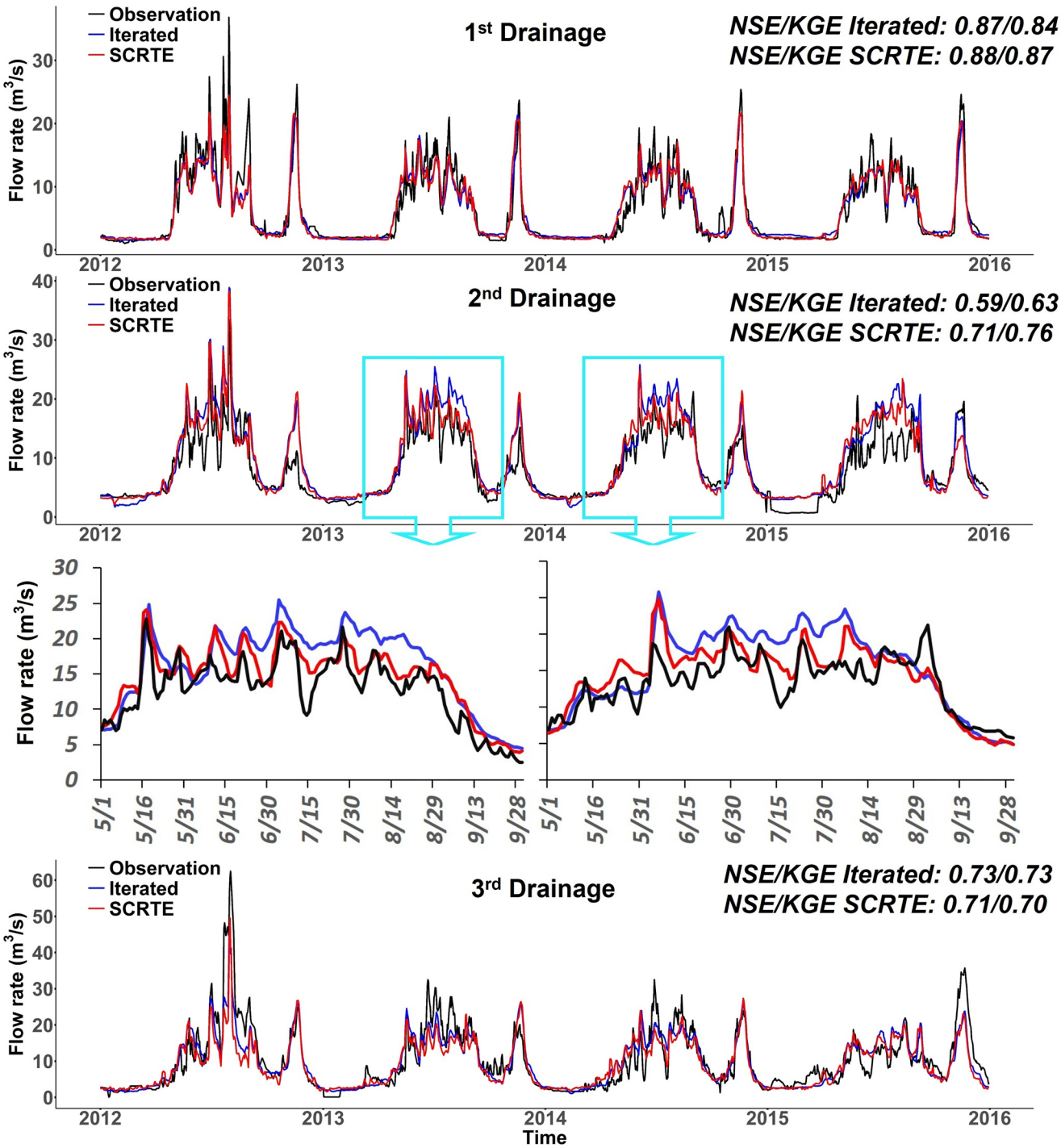

Figure 8. Hydrographs of non-iterated SCRTE (red lines) and iterated SCRTE (blue lines). Note that the input structure for training an iterated SCRTE contains all the predictors of non-iterated SCRTE and antecedent ($t-1$) streamflow observations. For the testing period, the iterated SCRTE uses simulated rather than observed streamflow records (as a predictor) at time step $t-1$ to generate streamflow values at time step $\boldsymbol{t}$. Therefore, no new streamflow observation is provided during the simulation process.

transboundary water can be precisely regulated by managing daily irrigation schedules. According to the NWRC, the most critical period is April-1 to June-10 when all crops require extensive water demand but receive a very low amount of precipitation. To establish the irrigation-discharge relations, the trained SCRTE was further evaluated using the testing dataset for the period of April-1 to June-10. In the re-evaluation process, the SCRTE

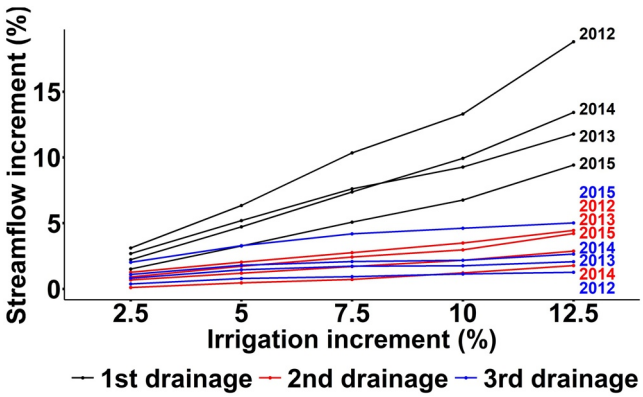

Figure 9. The irrigation-discharge relationships. Note that these relationships are based on the period between April-1 to June-10 of each year.

was fed with several scenarios of irrigation data (2.5%, 5.0%, 7.5%, 10% and 12.5% increments to the original), while the other forcing data remained unchanged. The results show that increased streamflow is expected for each basin and each year under each irrigation increment scenario (Figure 9). The hydrological characteristics of study basins are responsible for the variations in the irrigation-discharge relationships. For instance, the time of concentration for the first drainage basin is shorter than that for the second and third basins, leading to a higher runoff ratio (i.e., runoff divided by the sum of precipitation and irrigation). The interannual irrigation-discharge relationships may also vary significantly (in the same basin). For instance, when the irrigation water amount is increased by 12.5%, the streamflow increment in the first drainage basin is expected to vary by 10%–19%. The results suggest that the irrigation-discharge relationships as generated from the SCRTE are more reliable than those from the existing methods (which estimate the discharge as a fixed proportion of irrigation flow). Based on the comparison between SCRTE and RF, it is demonstrated that SCRTE is more capable of addressing the saturated subsurface flows that are responsible for the significant increment in streamflows (Figure 10). This is because SCRTE is more capable (compared to RF) of addressing median flows in the SI period, which is significantly dominated by saturated subsurface flows. Such an advantage is critical for improving the prediction accuracy and for supporting more effective irrigation management.

4. Discussion

4.1. Hydrological Inference From the Perspective of Predictor Importance

The term "predictor importance" is defined as the relative contribution of a predictor to the modeling performance. Predictor importance can advance our understanding of hydrological responses, such as streamflow persistence and flooding effect. In this study, the predictor importance for single- and multi-output SCA ensembles was investigated through the permutation feature importance (PFI) method (Molnar, 2020). The PFI method shuffles (i.e., randomly rearranges) the values of one predictor at a time while leaving the others unchanged. In this way, the ties between the predictand and the respective predictor are broken (Schmidt et al., 2020). The greater importance of a predictor, the more reduction in simulation accuracy. Once this process is repeated for all predictors, the reduction in accuracy can be used to rank the importance of predictors. The importance score of predictor j can be calculated as the reduction in accuracy of predictor j divided by the sum of reduction in accuracy of all predictors. Since the predictors used for each basin are different, the importance scores are evaluated in an aggregated manner: the importance scores characterizing various locations (e.g., the importance scores for predictors $C1_1$, $C2_1$ and $C3_1$ for the first drainage basin) are averaged. In general, all of the three SCA ensembles show similar patterns of aggregated importance scores (AISs), as shown in Figure 11. The 30-day total irrigation (I30) has the highest AISs for the SI period, which explains the persistence of streamflow as follows: the constant irrigation practice during the crop growing seasons raises the groundwater table, and thus increases the saturated subsurface flow, which then persistently feeds the streamflow in a drainage basin.

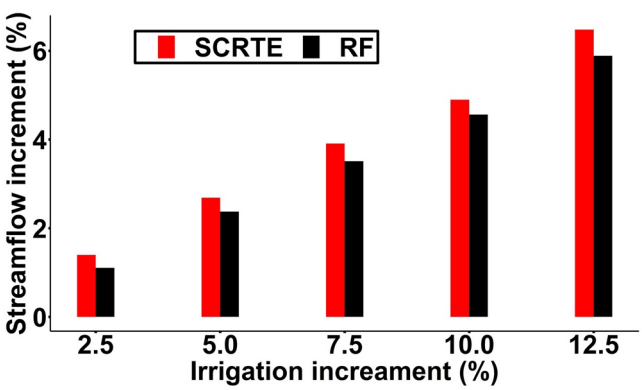

Figure 10. Comparison of the average irrigation-discharge relationships based on three basins and four years (2012–2015).

In the WI period, the highest AISs for the first, second and third drainage basins belong to I3, I7 and I15, respectively. This result can be associated with the time of concentration in these basins: the third basin (i.e., the largest one) has the highest water storage capacity in its saturated and unsaturated zones, leading to the longest retention time of irrigation water. Given the scarcity of precipitation in the WI period, I15 thus becomes the highest AIS. The situation for the first basin (i.e., the smallest one) is opposite to the third one, making I3 the highest AIS. These findings indicate that the AISs can help generate a valid inference. The results also show that the AISs for the 5-day precipitation (P5) of the third drainage basin (SI period) are 18.8%,

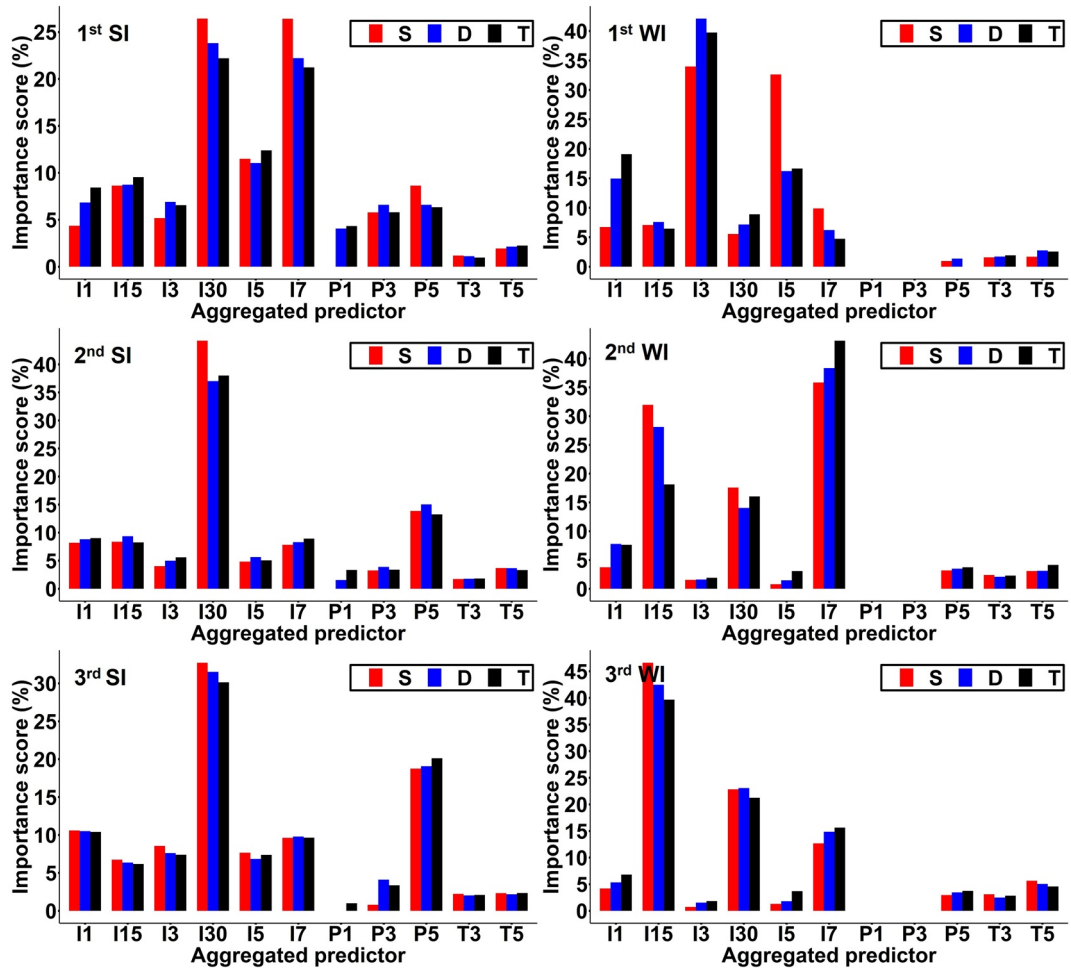

Figure 11. Importance scores for aggregated predictors. Note that plots from top to bottom represent the first, second, and third drainage basins, respectively; the sum of aggregated predictors for each SCA ensemble model (i.e., ***S***, ***D*** and ***T***) equals one.

19.1% and 20.1% for ***S***, ***D*** and ***T***, respectively, which are much higher than those for the first (8.6, 6.6 and 6.3) and the second (13.9, 15.0 and 13.2) basins. This is because the streamflow in the third basin contains a substantial proportion of mountain torrents during the rainy season. After being regulated by lakes and detention reservoirs, these flash floods are merged with irrigation return flows, becoming a substantial proportion of streamflows.

The predictors with small moving windows (e.g., I1 and P1) are more likely to have higher AISs under multi-output SCA ensembles than single-output ones. For instance, I1 for ***S***, ***D*** and ***T*** of the first drainage basin (in SI period) are 4.4%, 6.8% and 8.4%, respectively. Such an inclined pattern indicates that these predictors may have more significant relationships with streamflow variations (over multiple consecutive days) than those with streamflow values (for single time steps). The streamflow at time *t* may not be influenced by irrigation or precipitation at the same time step; instead, it may be influenced by irrigation and precipitation that occurred at a few time steps before. Thus, the multi-output SCA ensembles are able to take more advantage of these predictors than the single-output one, reflecting the streamflow dynamics over multiple consecutive days.

To further investigate the reliability of modeling inference, the AISs from ***S***, ***D***, ***T***, RF and XGB are compared through the Pearson correlation test (Figure 12 and Figures S5 to S9 in Supporting Information S1). In general, most of the pairs in AISs are highly correlated, with their Pearson correlation coefficients being over 0.9. This indicates that the three tree-structured models may imply similar hydrological inferences (i.e., patterns of importance scores). These findings complement the studies of Schmidt et al. (2020), who compared three data-driven models (RF, MLP and GLM) and found that equifinality existed for these models and would hinder hydrological

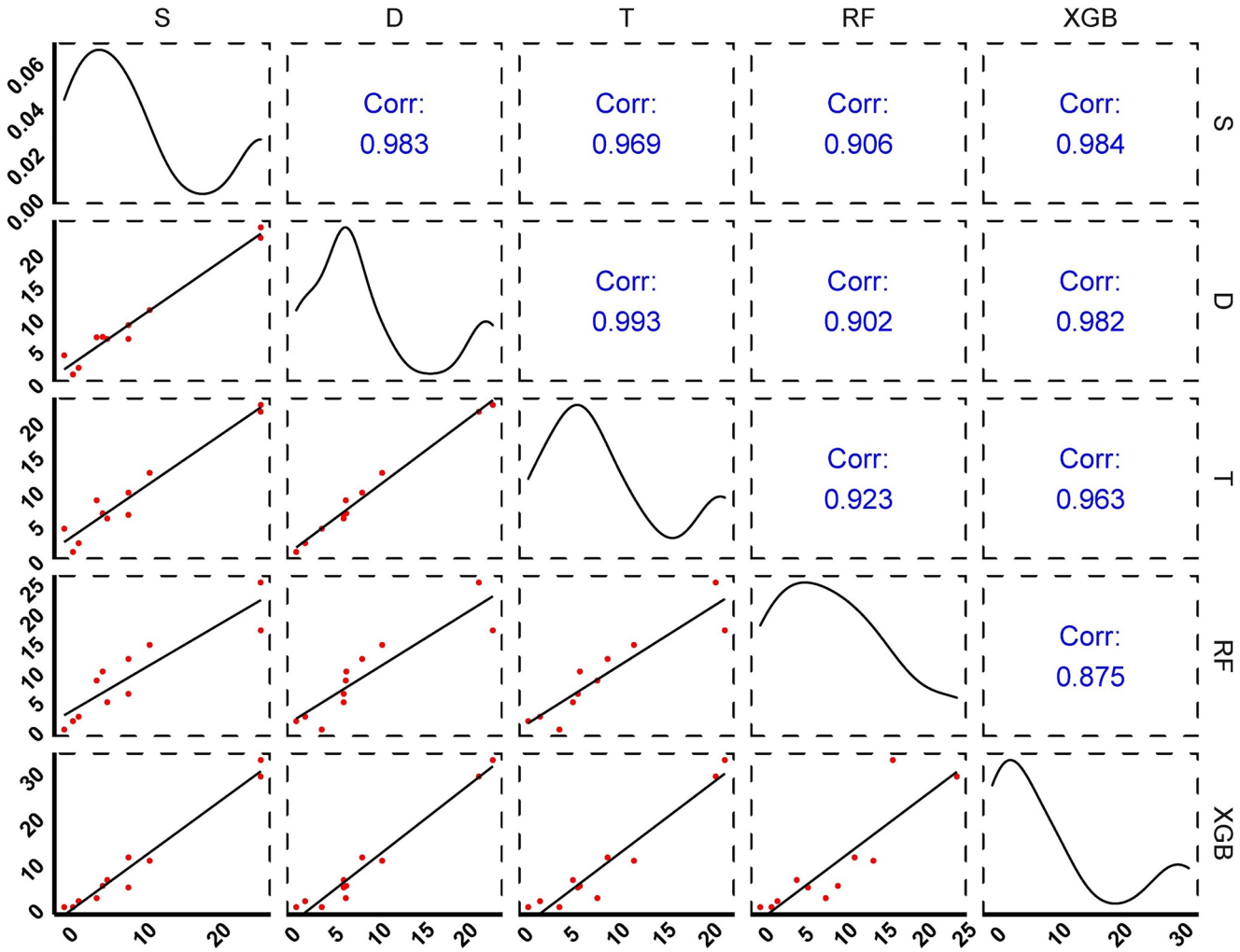

Figure 12. Pairwise comparisons of AIS for the SCA ensembles, RF and XGB over the SI period for the first drainage basin. Note that scatter plots in the lower triangular matrix represent the pairwise correlations (the ***x***-axis and ***y***-axis are AISs); plots along the diagonal represent the probability density functions of aggregated importance scores for each model (***x***-axis and ***y***-axis are AIS and probability, respectively); Pearson correlation coefficients are shown in the upper triangular matrix; the AISs for RF and XGB are measured through the PFI method, and the method embedded in XGB, respectively.

inference. Based on our results, we can infer that the data-driven models under similar algorithmic structures would not be significantly affected by equifinality. The variations in the patterns of AISs could be caused by the differences in the algorithmic structures and the interpretation approaches to quantify the relevant importance.

4.2. Robustness of the SCRTE

The convergences of SCRTE and its ensemble members are evaluated in terms of RMSE (Figure 13). In general, all of the models have converged to support reliable simulations. The multi-output SCA ensembles have higher accuracy than the single-output one in a few trees. This result implies that, for a limited number of training samples and predictors, the interactions among streamflow values, rather than individual streamflow values, serve as more significant information for the simulation efforts. This finding agrees with the previous study of Liu and Jiang (2013), in which they indicated that the multi-output training could improve the overall accuracy for MLPs for only a small number of training instances. Figure 13 also illustrates that SCRTE always outperforms at least one of its ensemble members as the number of trees increases. These results highlight the robustness of the developed SCRTE.

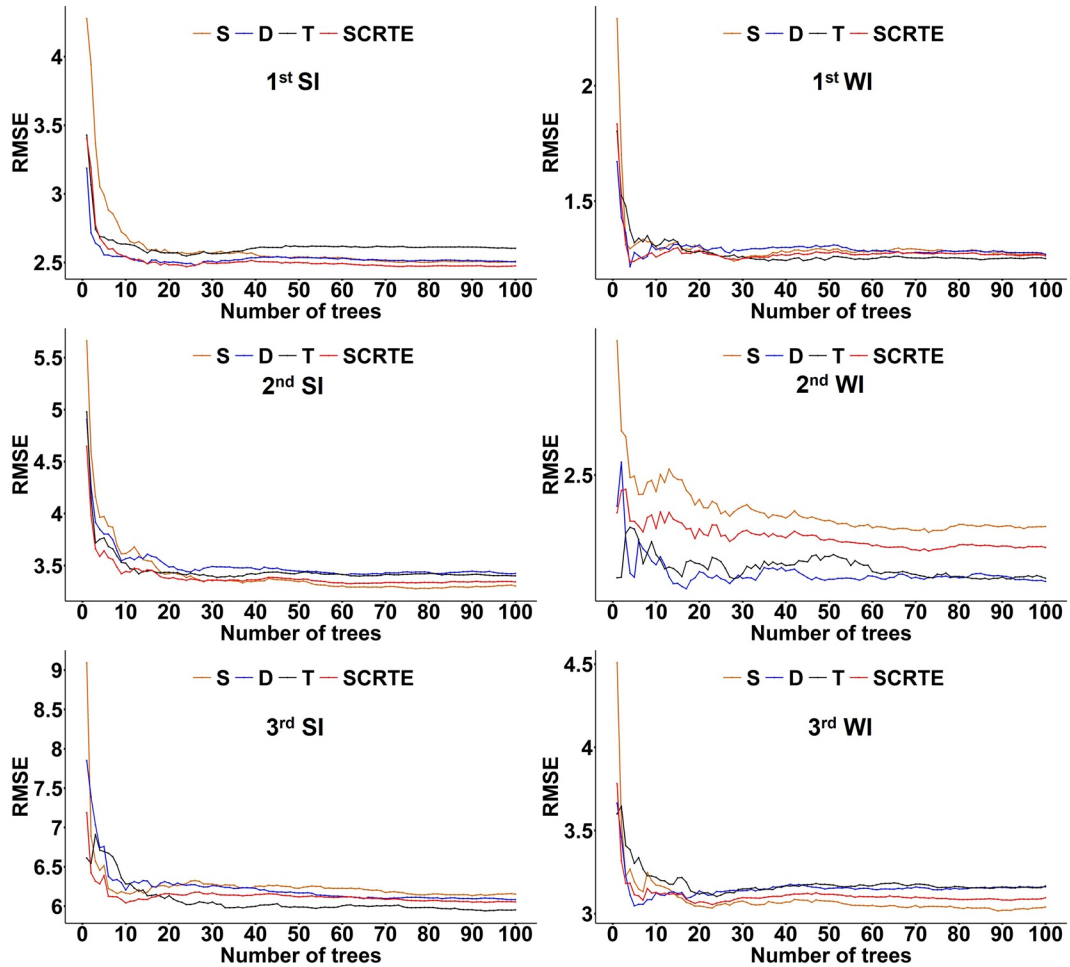

Figure 13. Convergence of the SCRTE and its ensemble members based on RMSE over the testing period. Note that since there is little change of RMSE after 100 trees, only the first 100 trees are shown in this plot.

In this study, the prediction through BMASS requires streamflow quantile ranges to be categorized using single- and multi-output SCA ensemble means. It is necessary to examine whether the ensemble-mean categorized streamflow classes ($C_i^{\mathrm{Sim(testing)}} \in [1, 2, ..., K]$, where i denotes the observation ID of the testing datasets and $\boldsymbol{K}$ denotes the number of quantile ranges) can be a valid substitute to the categorized observations ($C_i^{\mathrm{Obs(testing)}} \in [1, 2, ..., K]$). To this end, the two-sided Kolmogorov-Smirnov (KS) tests (Chakravarti & Roy, 1967) were performed to test if the two streamflow classes (i.e., $C_i^{\mathrm{Sim(testing)}}$ and $C_i^{\mathrm{Obs(testing)}}$) are from the same distribution (i.e., identical to each other). The null hypothesis is H_0: both samples come from a population with the same distribution, with a significance level of 0.05. If the $\boldsymbol{p}$ is less than 0.05, the null hypothesis is rejected (i.e., two streamflow classes are not from the same distribution); otherwise, it is accepted. Figure 14 illustrates that the simulated classes are identical to the observed ones (i.e., all pairs of cumulative densities are mostly overlapped), with all of the $\boldsymbol{p}$ significantly greater than 0.05. The difference in modeling performance using simulated and observed classes is illustrated in Table 5. The results show that the performance of simulated classes is identical to that of observed ones, which demonstrates the success of BMASS. To test whether the incorporation of other data-driven models into the BMASS ensemble scheme can improve the SCRTE performance, the RF and XGB models were considered as additional ensemble members. The results show that the modeling performance decreases when considering RF and XGB as the two additional members (Table 5). This result verifies that the single-output SCA model produces less-overfitted deduction trees than the other two models. The less overfitted trees can help produce more robust BMA weights, leading to more accurate SCRTE simulations.

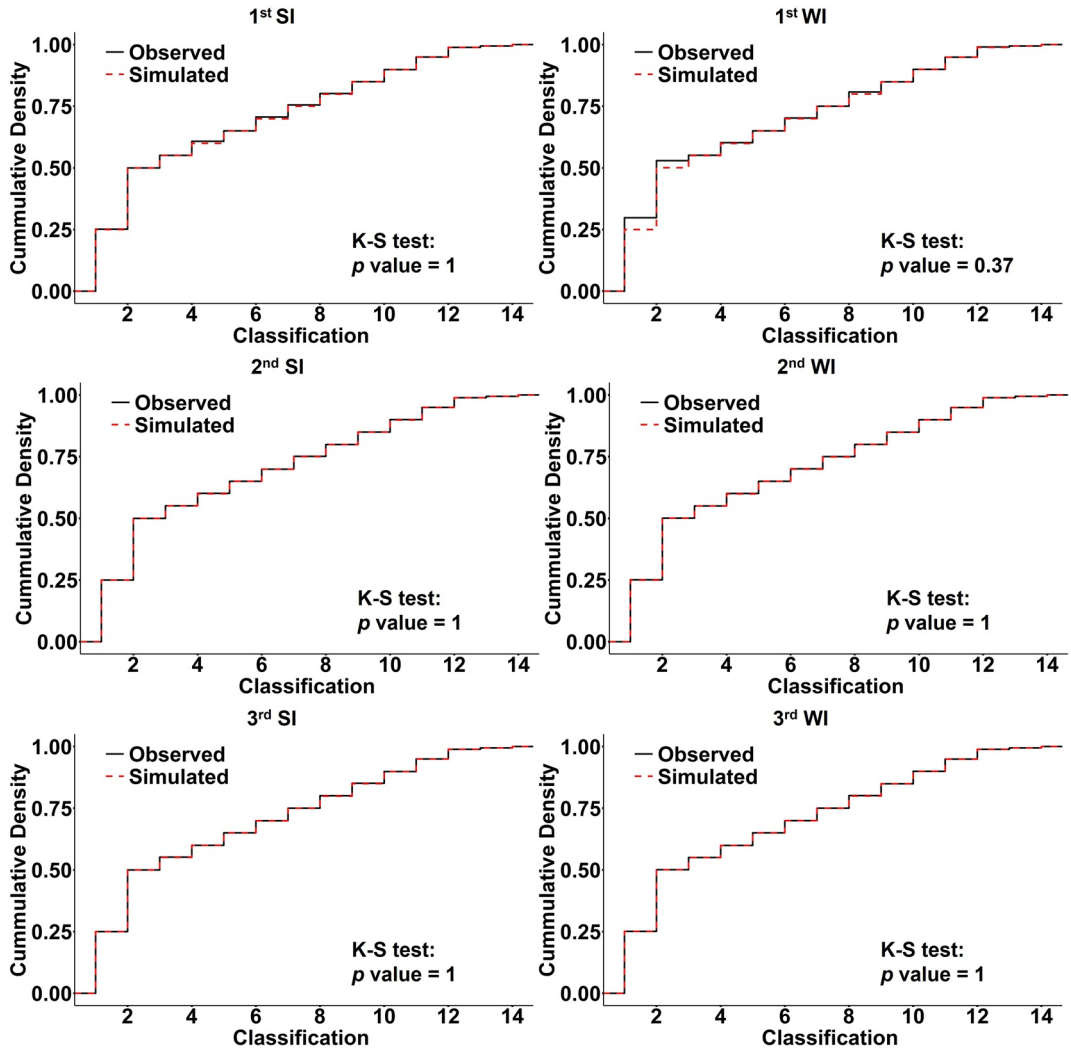

Figure 14. Cumulative density plots for two streamflow quantile classes. Note that the ***p*** value is derived from the two-sided *K-S* test.

4.3. Advantages of the Proposed SCRTE

In this study, SCRTE as a data-driven hydrological model was proposed to address the autocorrelation of daily streamflows by addressing the interactions among flow values over multiple consecutive days. Streamflow autocorrelation is critical for evaluating flow volumes, especially in dry periods. By addressing the autocorrelation, the dynamics of saturated flows can be adequately reflected, improving the median-flow simulation. In fact, median flow is critical for planning water supply, hydropower development, and irrigation activities. However, it has received less attention compared with peak flows (Kroll & Song, 2013). Thus, the proposed SCRTE can be a valuable tool to address the above-mentioned issue and to better support irrigation management in dry periods.

The performance of SCRTE and benchmark models are subject to data availability. Many recent data-driven modeling efforts have removed the barrier of watershed boundary by training one model based on big data covering hundreds of watersheds (Gauch et al., 2021; Kratzert, Klotz, Shalev, et al., 2019), and applying it to catchments of other continents (Ma et al., 2021). However, these efforts are mainly made for pristine watersheds or watersheds with little human intervention. For watersheds with intensive human activities, various complexities and non-stationarities may make such a universal model essentially inapplicable. Under such a circumstance, the proposed SCRTE would be of great value.

Table 5
Comparisons Among SCRTE Performances Using Different BMASS Schemes

Model prediction		1st		2nd		3rd		
		Spring	Winter	Spring	Winter	Spring	Winter	Overall
R^2	SCRTE	0.81	0.92	0.77	0.66	0.60	0.79	0.76
	SCRTE^O	0.81	0.92	0.78	0.65	0.60	0.79	0.76
	SCRTE*	0.82	0.91	0.76	0.67	0.57	0.78	0.75
MAE	SCRTE	1.74	0.67	2.56	1.47	4.14	1.90	2.08
	SCRTE^O	1.74	0.67	2.55	1.48	4.11	1.90	2.07
	SCRTE*	1.74	0.70	2.63	1.49	4.18	1.93	2.11
RMSE	SCRTE	2.48	1.27	3.34	2.29	6.01	3.08	3.08
	SCRTE^O	2.47	1.26	3.31	2.30	5.97	3.09	3.07
	SCRTE*	2.48	1.36	3.45	2.28	6.14	3.12	3.14
NSE	SCRTE	0.80	0.92	0.56	0.53	0.55	0.78	0.69
	SCRTE^O	0.80	0.92	0.57	0.53	0.55	0.78	0.69
	SCRTE*	0.80	0.91	0.54	0.54	0.53	0.77	0.68
KGE	SCRTE	0.77	0.96	0.75	0.75	0.53	0.80	0.76
	SCRTE^O	0.78	0.96	0.75	0.74	0.53	0.80	0.76
	SCRTE*	0.78	0.95	0.72	0.74	0.52	0.79	0.75
VE	SCRTE	0.80	0.82	0.77	0.70	0.70	0.71	0.75
	SCRTE^O	0.80	0.82	0.77	0.70	0.70	0.71	0.75
	SCRTE*	0.80	0.81	0.77	0.70	0.70	0.71	0.75
LogNSE	SCRTE	0.88	0.87	0.72	0.42	0.63	0.55	0.68
	SCRTE^O	0.88	0.87	0.72	0.42	0.64	0.56	0.68
	SCRTE*	0.88	0.87	0.72	0.44	0.62	0.56	0.68

Note. SCRTE, SCRTE^O and SCRTE* indicate SCRTE performances under simulated classes (i.e., used in this study), observed classes (i.e., streamflow quantile ranges categorized using observations) and simulated classes with additional ensemble members (i.e., RF and XGB), respectively.

Compared with neural networks, the proposed SCRTE enables the reasoning of hydrological processes because its rules for hydrological simulation are associated with compound events (e.g., rainfall and irrigation events). In this study, the cause of flood-peak underestimations in the first and third drainage basins has been explained via the rules established in the training process. However, such explanations would not be possible for neural networks. The explainable hydrological processes through the established rules can help distinguish the proposed SCRTE from "black-box" approaches.

5. Summary and Conclusions

This study proposes a state-of-the-art data-driven model SCRTE to address the effects of autocorrelation in daily streamflows. The model has been successfully applied to three watersheds with mixed land uses for supporting irrigation planning. Compared with seven well-known benchmark models based on seven evaluation metrics, the SCRTE exhibits satisfactory performances for all the drainage basins and irrigation periods. Several major flood events in the three basins have been explained via the rules established by SCRTE. The proposed method has also been used to establish the irrigation-discharge relationships for supporting irrigation scheduling under rigorous transboundary water supply-demand conditions.

There are three advantages of the proposed SCRTE. First, compared with multi-input single-output models, SCRTE can effectively improve the performance of median-flow simulation while maintaining the quality of low- and high-flow ones. Second, SCRTE does not require tremendous training data compared with deep-learning approaches, and thus it can be an effective alternative to deep learning when the training sample size is

limited. Lastly, SCRTE enables explainable processes for hydrological simulation through the established rules of deduction trees, thereby distinguishing the method from "black-box" approaches.

The key findings from this study include: (a) the multi-output SCA ensembles are more capable of addressing median flows, whereas the single-output SCA ensemble can better simulate low and high flows. (b) SCRTE is more capable of addressing saturated subsurface flow than the single-output model (e.g., RF); such subsurface flow may lead to considerable increments of streamflows in irrigation systems. (c) The improved accuracy for low-to-median flow simulation as achieved through this study is meaningful for supporting more effective water resources management in arid and semi-arid regions. In this study, the irrigation-discharge relationships are evaluated based on irrigation information of 3-month resolution. Information with higher temporal resolutions (e.g., monthly and weekly) is desired for generating more precise relationships and thus supporting more effective irrigation management.

Appendix: Nomenclature

AISs (*Aggregated Importance Scores*) are aggregated (i.e., averaged) importance scores of the same type and moving window size.

BMA (*Bayesian model averaging*) method is an ensemble strategy used for combining the simulations from different models based on their accuracy over the training period.

BMASS (*Bayesian model averaging with stratified sampling*) method applies the BMA method to multiple quantile ranges of the target variable (e.g., streamflow), in order to achieve more accurate ensemble simulations.

GLM (*Generalized linear model*) is a linear regression model that allows the predictands to have various error distribution models rather than a normal distribution.

KGE (Kling-Gupta efficiency) is a combined measure of correlation, standard deviation and mean squared error, which is sensitive to the flow variability.

LogNSE (log Nash–Sutcliffe efficiency) is the Nash–Sutcliffe efficiency with natural logarithm transformed streamflows.

LSTM (Long short-term memory) is a type of recurrent neural network that uses a memory cell and three gates to control information in the hidden neuron.

MAE (Mean absolute error) is a measure of errors used to represent the difference between observations and simulations.

MLP (Multilayer perceptron) is a well-established artificial neural network that consists of a network of nodes and links between predictors and predictions.

NSE (Nash–Sutcliffe efficiency) is a common metric to assess the goodness of simulation of hydrological models.

OOB (Out-of-Bag) datasets are the validation datasets used for the SCRTE and the random forest model validation.

PFI (Permutation feature importance) is a technique to measure the importance of a predictor.

RAE (Lag-1 relative autocorrelative errors/residuals) is used to reflect the effect of lag-1 autocorrelation at time $\boldsymbol{t}$.

RF (Random forest) is a regression tree ensemble that consists of a weighted average of many regression trees trained in parallel.

RFE (Recursive feature elimination) is a predictor selection algorithm used to identify the core predictors associated with the model for reducing the effect of multicollinearity.

RMSE (Root mean square error) is the squared root of the average squared difference between observations and simulations.

RTE (Regression Tree Ensemble) is the simulation model composed of a weighted combination of multiple regression trees.

"S", "D" and "T" denote the SCA ensembles using single-, dual- and triple-output settings, respectively. They are also ensemble members of the SCRTE.

SCA (Stepwise Cluster Analysis) is an algorithm used to build the trees for SCRTE.

SCRTE (Stepwise-Clustered Regression Tree Ensemble) is the tree-structured data-driven model proposed in this study.

SI (Spring irrigation) period spans the entire growing season (from April to September).

SLM (Sparse linear model) selects a subset of predictors and then evaluates the least-squares fit of all possible sub-models (that trained with a subset of predictors) in order to choose the one with the best fit.

SVM (Support Vector Machine) is a classification and regression approach that uses the sets of decision boundaries in the predictor space to separate data points belonging to different classes.

VE (Volumetric efficiency) is an index of how well the simulated volume matches the observed one over a given time interval.

WI (Winter irrigation) period starts at the end of October and lasts until the end of November.

XGB (Extreme Gradient Boosting) is a regression tree ensemble that sequentially aggregates the trees based on the errors learned from the previous aggregation.

Data Availability Statement

We also appreciate the provision of climate data from China Meteorological Data Service Center (http://data.cma.cn/en). The hydrological data can be accessed from Zenodo repository (https://doi.org/10.5281/zenodo.4072259).

Acknowledgments
This research was supported by Canada Research Chair Program, Natural Science and Engineering Research Council of Canada, and MITACS. We appreciate Ningxia Water Conservancy for offering streamflow, groundwater and irrigation information and related help. We would like to express our sincere gratitude to the editor, associate editor and anonymous reviewers for their constructive comments and suggestions.

References

Adnan, R. M., Liang, Z., Heddam, S., Zounemat-Kermani, M., Kisi, O., & Li, B. (2020). Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. *Journal of Hydrology*, *586*, 124371. https://doi.org/10.1016/j.jhydrol.2019.124371

Adnan, R. M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., & Kisi, O. (2019). Daily streamflow prediction using optimally pruned extreme learning machine. *Journal of Hydrology*, *577*, 123981. https://doi.org/10.1016/j.jhydrol.2019.123981

Allaire, J., & Chollet, F. (2019). *keras: R Interface to 'Keras'. R package version 2.2.5.0*. GitHub. Retrieved from https://CRAN.R-project.org/package=keras

Allaire, J., & Tang, Y. (2019). *tensorflow: R Interface to 'TensorFlow'. R package version 2.0.0*. Retrieved from https://CRAN.R-project.org/package=tensorflow

ASCE Task Committee. (2000a). Artificial neural networks in hydrology. II: Hydrologic applications. *Journal of Hydrologic Engineering*, *5*(2), 124–137.

ASCE Task Committee. (2000b). Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering*, *5*(2), 115–123.

Badrzadeh, H., Sarukkalige, R., & Jayawardena, A. (2013). Impact of multi-resolution analysis of artificial intelligence models inputs on multi-step ahead river flow forecasting. *Journal of Hydrology*, *507*, 75–85. https://doi.org/10.1016/j.jhydrol.2013.10.017

Bai, P., Liu, X., & Xie, J. (2021). Simulating runoff under changing climatic conditions: A comparison of the long short-term memory network with two conceptual hydrologic models. *Journal of Hydrology*, *592*, 125779. https://doi.org/10.1016/j.jhydrol.2020.125779

Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 1–22.

Bergmeir, C. N., & Benítez Sánchez, J. M. (2012). Neural networks in R using the Stuttgart neural network simulator: RSNNS. *Journal of Statistical Software*, *46*(7), 1–26. https://doi.org/10.18637/jss.v046.i07

Bierkens, M., & Van Beek, L. (2009). Seasonal predictability of European discharge: NAO and hydrological response time. *Journal of Hydrometeorology*, *10*(4), 953–968. https://doi.org/10.1175/2009jhm1034.1

Bontempi, G., Taieb, S. B., & Le Borgne, Y. A. (2012). *Machine learning strategies for time series forecasting, paper presented at European business intelligence summer school*. Springer.

Bray, M., & Han, D. (2004). Identification of support vector machines for runoff modelling. *Journal of Hydroinformatics*, *6*(4), 265–280. https://doi.org/10.2166/hydro.2004.0020

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

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Campolo, M., Soldati, A., & Andreussi, P. (2003). Artificial neural network approach to flood forecasting in the River Arno. *Hydrological Sciences Journal*, *48*(3), 381–398. https://doi.org/10.1623/hysj.48.3.381.45286

Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, *170*, 114498. https://doi.org/10.1016/j.eswa.2020.114498

Chakravarti, L., & Roy, J. (1967). *Handbook of methods of applied statistics* (Vol. I, pp. 392–394). John Wiley and Sons.

Chang, F. J., Chiang, Y. M., & Chang, L. C. (2007). Multi-step-ahead neural networks for flood forecasting. *Hydrological Sciences Journal*, *52*(1), 114–130. https://doi.org/10.1623/hysj.52.1.114

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). *Xgboost: Extreme gradient boosting. R package version 0.4-2* (pp. 1–4).
Cheng, C., Xie, J., Chau, K., & Layeghifard, M. (2008). A new indirect multi-step-ahead prediction model for a long-term hydrologic prediction. *Journal of Hydrology*, *361*(1–2), 118–130. https://doi.org/10.1016/j.jhydrol.2008.07.040
Cheng, M., Fang, F., Kinouchi, T., Navon, I., & Pain, C. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. *Journal of Hydrology*, *590*, 125376. https://doi.org/10.1016/j.jhydrol.2020.125376
Criss, R. E., & Winston, W. E. (2008). Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes: International Journal*, *22*(14), 2723–2725. https://doi.org/10.1002/hyp.7072
Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, *19*, 372–386.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
Dewandel, B., Gandolfi, J. M., De Condappa, D., & Ahmed, S. (2008). An efficient methodology for estimating irrigation return flow coefficients of irrigated crops at watershed and seasonal scale. *Hydrological Processes: International Journal*, *22*(11), 1700–1712. https://doi.org/10.1002/hyp.6738
Dibike, Y. B., Velickov, S., Solomatine, D., & Abbott, M. B. (2001). Model induction with support vector machines: Introduction and applications. *Journal of Computing in Civil Engineering*, *15*(3), 208–216. https://doi.org/10.1061/(asce)0887-3801(2001)15:3(208)
Dong, L., Yu, D., Zhang, H., Zhang, M., Jin, W., Liu, Y., et al. (2015). Long-term effect of sediment laden Yellow River irrigation water on soil organic carbon stocks in Ningxia, China. *Soil and Tillage Research*, *145*, 148–156. https://doi.org/10.1016/j.still.2014.09.009
Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014
Duan, S., Ullrich, P., & Shu, L. (2020). Using convolutional neural networks for streamflow projection in California. *Frontiers in Water*, *2*, 28. https://doi.org/10.3389/frwa.2020.00028
Elshorbagy, A., Corzo, G., Srinivasulu, S., & Solomatine, D. (2010). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-Part 1: Concepts and methodology. *Hydrology and Earth System Sciences*, *14*(10), 1931–1941. https://doi.org/10.5194/hess-14-1931-2010
Erdal, H. I., & Karakurt, O. (2013). Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *Journal of Hydrology*, *477*, 119–128. https://doi.org/10.1016/j.jhydrol.2012.11.015
Fan, Y., Huang, G., Li, Y., Wang, X., Li, Z., & Jin, L. (2017). Development of PCA-based cluster quantile regression (PCA-CQR) framework for streamflow prediction: Application to the Xiangxi river watershed, China. *Applied Soft Computing*, *51*, 280–293. https://doi.org/10.1016/j.asoc.2016.11.039
Fang, K., & Shen, C. (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *Journal of Hydrometeorology*, *21*(3), 399–413. https://doi.org/10.1175/jhm-d-19-0169.1
Fleming, S. W., Bourdin, D. R., Campbell, D., Stull, R. B., & Gardner, T. (2015). Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific northwest river. *JAWRA Journal of the American Water Resources Association*, *51*(2), 502–512. https://doi.org/10.1111/jawr.12259
Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall-runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021
Gilli, M., Maringer, D., & Schumann, E. (2019). *Numerical methods and optimization in finance*. Academic Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning (adaptive computation and machine learning series)* (pp. 321–359). MIT Press.
Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422. https://doi.org/10.1023/a:1012487302797
Han, J. C., Huang, Y., Li, Z., Zhao, C., Cheng, G., & Huang, P. (2016). Groundwater level prediction using a SOM-aided stepwise cluster inference model. *Journal of Environmental Management*, *182*, 308–321. https://doi.org/10.1016/j.jenvman.2016.07.069
He, Z., Wen, X., Liu, H., & Du, J. (2014). A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology*, *509*, 379–386. https://doi.org/10.1016/j.jhydrol.2013.11.054
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Huang, G. (1992). A stepwise cluster analysis method for predicting air quality in an urban environment. *Atmospheric Environment: Part B. Urban Atmosphere*, *26*(3), 349–357. https://doi.org/10.1016/0957-1272(92)90010-p
International Commission on Irrigation and Drainage. (2017). *23rd ICID international congress on irrigation and drainage*. Retrieved from http://icidciid.org/inner_page/92
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.
Kendall, M. G. (1948), *Rank correlation methods* (4th ed.), Charles Griffin and Company.
Kim, H., Jang, T., Im, S., & Park, S. (2009). Estimation of irrigation return flow from paddy fields considering the soil moisture. *Agricultural Water Management*, *96*(5), 875–882. https://doi.org/10.1016/j.agwat.2008.11.009
Kisi, O., Shiri, J., & Nikoofar, B. (2012). Forecasting daily lake levels using artificial intelligence approaches. *Computers & Geosciences*, *41*, 169–180. https://doi.org/10.1016/j.cageo.2011.08.027
Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019wr026065
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12). https://doi.org/10.5194/hess-23-5089-2019
Krause, P., Boyle, D., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, *5*, 89–97. https://doi.org/10.5194/adgeo-5-89-2005
Kroll, C. N., & Song, P. (2013). Impact of multicollinearity on small sample hydrologic regression models. *Water Resources Research*, *49*(6), 3756–3769. https://doi.org/10.1002/wrcr.20315
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, K., Huang, G., & Wang, S. (2019). Market-based stochastic optimization of water resources systems for improving drought resilience and economic efficiency in arid regions. *Journal of Cleaner Production*, *233*, 522–537. https://doi.org/10.1016/j.jclepro.2019.05.379

Li, Z., Huang, G., Han, J., Wang, X., Fan, Y., Cheng, G., et al. (2015). Development of a stepwise-clustered hydrological inference model. *Journal of Hydrologic Engineering*, *20*(10), 04015008. https://doi.org/10.1061/(asce)he.1943-5584.0001165

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Liu, B., & Jiang, Y. (2013). A multitarget training method for artificial neural network with application to computer-aided diagnosis. *Medical Physics*, *40*(1), 011908. https://doi.org/10.1118/1.4772021

Ma, K., Feng, D., Lawson, K., Tsai, W. P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents—Leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*(5), e2020WR028600. https://doi.org/10.1029/2020wr028600

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society*, *13*, 245–259. https://doi.org/10.2307/1907187

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., et al. (2019). *Package 'e1071'. R Software package*. Retrieved from http://cran.rproject.org/web/packages/e1071/index.html

Mi, L., Tian, J., Si, J., Chen, Y., Li, Y., & Wang, X. (2020). Evolution of groundwater in Yinchuan Oasis at the upper reaches of the Yellow River after water-saving transformation and its driving factors. *International Journal of Environmental Research and Public Health*, *17*(4), 1304. https://doi.org/10.3390/ijerph17041304

Mohan, S., & Vijayalakshmi, D. (2009). Prediction of irrigation return flows through a hierarchical modeling approach. *Agricultural Water Management*, *96*(2), 233–246. https://doi.org/10.1016/j.agwat.2008.07.013

Molnar, C. (2020), *Interpretable machine learning*. Lulu Press.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Ningxia Water Conservancy. (2003–2015). *Ningxia water resources bulletin*.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/mwr2906.1

R Development Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ren, W., Yang, T., Shi, P., Xu, C.-y., Zhang, K., Zhou, X., et al. (2018). A probabilistic method for streamflow projection and associated uncertainty analysis in a data sparse alpine region. *Global and Planetary Change*, *165*, 100–113. https://doi.org/10.1016/j.gloplacha.2018.03.011

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge University Press.

Sahoo, S., Russo, T., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. *Water Resources Research*, *53*(5), 3878–3895. https://doi.org/10.1002/2016wr019933

Sarzaeim, P., Bozorg-Haddad, O., Bozorgi, A., & Loáiciga, H. A. (2017). Runoff projection under climate change conditions with data-mining methods. *Journal of Irrigation and Drainage Engineering*, *143*(8), 04017026. https://doi.org/10.1061/(asce)ir.1943-4774.0001205

Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020). Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. *Water Resources Research*, *56*(5), e2019WR025924. https://doi.org/10.1029/2019wr025924

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, *20*(7). https://doi.org/10.5194/hess-20-2611-2016

Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics*, *10*(1), 3–22. https://doi.org/10.2166/hydro.2008.015

State Council of the People's Republic of China. (2012). *Opinions of the state council on applying the strictest water resources control system*. Retrieved from http://www.gov.cn/zwgk/2012-02/16/content_2067664.htm

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Sun, T. (2019). *Package "scalreg": Scaled Sparse Linear Regression* (R package Version 1.0.1). https://CRAN.R-project.org/package=scalreg

Taieb, S. B., Sorjamaa, A., & Bontempi, G. (2010). Multiple-output modeling for multi-step-ahead time series forecasting. *Neurocomputing*, *73*(10–12), 1950–1957. https://doi.org/10.1016/j.neucom.2009.11.030

Tongal, H., & Booij, M. J. (2018). Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of Hydrology*, *564*, 266–282. https://doi.org/10.1016/j.jhydrol.2018.07.004

Toth, E., Brath, A., & Montanari, A. (2000). Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology*, *239*(1–4), 132–147. https://doi.org/10.1016/s0022-1694(00)00344-9

Wada, Y., Van Beek, L. P., Van Kempen, C. M., Reckman, J. W., Vasak, S., & Bierkens, M. F. (2010). Global depletion of groundwater resources. *Geophysical Research Letters*, *37*(20). https://doi.org/10.1029/2010gl044571

Wang, X., Huang, G., Lin, Q., Nie, X., Cheng, G., Fan, Y., et al. (2013). A stepwise cluster analysis approach for downscaled climate projection—A Canadian case study. *Environmental Modelling & Software*, *49*, 141–151. https://doi.org/10.1016/j.envsoft.2013.08.006

Wilks, S. S. (1967). *Collected papers; contributions to mathematical statistics*. Wiley.

Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: A comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, *25*(3), 1671–1687. https://doi.org/10.5194/hess-25-1671-2021

Xiang, Z., Yan, J., & Demir, I. (2020). A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resources Research*, *56*(1), e2019WR025326. https://doi.org/10.1029/2019wr025326

Yalcin, E. (2019). Estimation of irrigation return flow on monthly time resolution using SWAT model under limited data availability. *Hydrological Sciences Journal*, *64*(13), 1588–1604. https://doi.org/10.1080/02626667.2019.1662025

Yang, J., Tan, C., Wang, S., Wang, S., Yang, Y., & Chen, H. (2015). Drought adaptation in the Ningxia Hui Autonomous region, China: Actions, planning, pathways and barriers. *Sustainability*, *7*(11), 15029–15056. https://doi.org/10.3390/su71115029

Yang, J., Yu, S., & Liu, G. (2013). Multi-step-ahead predictor design for effective long-term forecast of hydrological signals using a novel wavelet-NN hybrid model. *Hydrology and Earth System Sciences Discussions*, *10*(7)

Yang, Y., Huang, T., Shi, Y., Wendroth, O., & Liu, B. (2020). Comparing the performance of an autoregressive state-space approach to the linear regression and artificial neural network for streamflow estimation. *Journal of Environmental Informatics*. https://doi.org/10.3808/jei.202000440

Zhang, H., Yang, Q., Shao, J., & Wang, G. (2019). Dynamic streamflow simulation via online gradient-boosted regression tree. *Journal of Hydrologic Engineering*, *24*(10), 04019041. https://doi.org/10.1061/(asce)he.1943-5584.0001822

Zhang, Z., & Deng, S. (1987). The development of irrigation in China. *Water International*, *12*(1–2), 46–52.
Zhu, J., Kong, F., Ran, L., & Lei, H. (2015). Bayesian model averaging with stratified sampling for probabilistic quantitative precipitation forecasting in northern China during summer 2010. *Monthly Weather Review*, *143*(9), 3628–3641. https://doi.org/10.1175/mwr-d-14-00301.1