Environmental Earth Sciences (2021) 80:747
https://doi.org/10.1007/s12665-021-10054-5

ORIGINAL ARTICLE

Random forest and extreme gradient boosting algorithms for streamflow modeling using vessel features and tree-rings

Hossein Sahour[1] · Vahid Gholami[2,3] · Javad Torkaman[4] · Mehdi Vazifedan[5] · Sirwe Saeedi[6]

Received: 7 January 2021 / Accepted: 11 October 2021 / Published online: 1 November 2021

Abstract

Monitoring temporal variation of streamflow is necessary for many water resources management plans, yet, such practices are constrained by the absence or paucity of data in many rivers around the world. Using a permanent river in the north of Iran as a test site, a machine learning framework was proposed to model the streamflow data in the three periods of growing seasons based on tree-rings and vessel features of the *Zelkova carpinifolia* species. First, full-disc samples were taken from 30 trees near the river, and the samples went through preprocessing, cross-dating, standardization, and time series analysis. Two machine learning algorithms, namely random forest (RF) and extreme gradient boosting (XGB), were used to model the relationships between dendrochronology variables (tree-rings and vessel features in the three periods of growing seasons) and the corresponding streamflow rates. The performance of each model was evaluated using statistical coefficients [coefficient of determination (*R*-squared), Nash–Sutcliffe efficiency (NSE), and root-mean-square error (NRMSE)]. Findings demonstrate that consideration should be given to the XGB model in streamflow modeling given its apparent enhanced performance (*R*-squared: 0.87; NSE: 0.81; and NRMSE: 0.43) over the RF model (*R*-squared: 0.82; NSE: 0.71; and NRMSE: 0.52). Furthermore, the results showed that the models perform better in modeling the normal and low flows compared to extremely high flows. Finally, the tested models were used to reconstruct the temporal streamflow during the past decades (1970–1981).

Keywords Dendrochronology · Machine learning · RF · XGB · Growing season · Streamflow

Introduction

Temporal streamflow records are essential for any long-term water resource plans, including optimal design of hydraulic structures, controlling extreme events, and determining ecological water budgets for aquatic ecosystems (Hirsch and Costa 2004). Streamflow is typically measured using automated stream gauges mounted in a stream. Such practices are constrained by the absence or paucity of gauging stations, temporal gaps in the measured time-series data, and data quality issues in many parts of the world (Zhang and Post 2018). Even in the areas with available streamflow records, the data are limited to recent decades, making it challenging to investigate the long-term variation of streamflow. Therefore, alternative models and methods have been used to fill gaps or reconstruct the past streamflow time series.

Hydrologists have made use of statistical applications to predict the streamflow using the observed relationship between precipitation and runoff (Khan and See 2006). However, the development of advanced machine learning algorithms and artificial neural networks (ANNs) in the past decade has prompted extensive research into advanced data-driven models (Alshehri et al. 2020; Sahour et al. 2020a). These models can predict the streamflow by establishing linear or nonlinear relationships between streamflow and a set of explanatory variables (Tongal and Booij 2018).

✉ Vahid Gholami
Gholami.vahid@guilan.ac.ir

[1] Marine Science Institute, The University of Texas at Austin, Port Aransas, TX 78373, USA

[2] Department of Range and Watershed Management, Faculty of Natural Resources, University of Guilan, Sowmeh Sara, Guilan, Iran

[3] Department of Water Engineering and Environment, University of Guilan, Rasht, Guilan, Iran

[4] Department of Forestry, Faculty of Natural Resources, University of Guilan, Sowmeh Sara, Iran

[5] Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, USA

[6] Department of Computer Sciences, Western Michigan University, Kalamazoo, MI 49008, USA

These models have proven to be powerful tools for streamflow modeling (Wang et al. 2019; Zhang et al. 2020). For instance, Adnan et al. (2019) successfully implemented an optimally pruned extreme learning machine (OP-ELM) model to predict daily streamflow using hydro-climatic data as inputs. They also found that including local hydroclimate data significantly improves the accuracy of the model. Meng et al. (2019) proposed a modified empirical mode decomposition support vector machine (M-EMDSVM) to predict the streamflow in the Wei River Basin of China. Their comparative analysis showed that the M-EMDSVM was superior to ANN and a single support vector machine (SVM) model to predict strong non-stationary streamflow. In all the abovementioned studies, streamflow prediction was carried out using hydroclimate variables as the models' inputs. The application of those approaches is constrained by the availability of hydroclimate data such as gauge-based precipitation. Moreover, in many stations, the available data may suffer from gaps in the time series. One plausible solution could be the use of modeled or downscaled data; however, the modeled data are typically associated with high levels of uncertainties (Qi et al. 2020).

Dendrochronological records yield valuable information about past hydroclimate variability through the response of tree growth to variations of precipitation and streamflow (Khaleghi 2018; Liu et al. 2018; Wu et al. 2020). Therefore, tree-rings and vessel features can be used as a proxy of hydroclimate variables such as precipitation and environment moisture for the streamflow modeling (Gholami et al. 2015, 2017; Liu et al. 2017). The advantage of using tree-rings and vessel features is to reconstruct the streamflow for past centuries, a practice that is not achievable using gauge-based precipitation records since these data are typically limited to the past few decades in most stations around the world. The relationship between tree-rings and time series of streamflow has been previously used to reconstruct the flow in several rivers around the world (Meko et al. 2012). For example, Akkemik et al. (2008) reconstructed the 350 years of streamflow for the Filyos river basin in Turkey using tree-ring records. Therrell and Bialecki (2015) identified 39 flood-ring years from 1770 to 2009 in the Lower Mississippi River using dendrochronology records. A network of multispecies tree-ring records was used to reconstruct the Suwannee River flow in Florida from 1555 to 2005 CE (Harley et al. 2017). Similar studies using tree-rings for the reconstruction of streamflow have been carried out in Canada (Case and MacDonald 2003), Chile (Urrutia et al. 2011), China (Gou et al. 2010), and Sudan (Mokria et al. 2018).

Tree-rings width depends on temperature and environmental moisture. Therefore, tree-ring chronologies can provide information about past hydrologic conditions of environments (Allen et al. 2015; Ferrero et al. 2015; Kames et al. 2016; Wu et al. 2020). Similar relationships exist between vessel features (vessel diameter, vessel area, and vessel perimeter) and the presence of moisture. Therefore, these parameters can also be used to predict hydrological data during growing seasons (Campelo et al. 2010; Fonti and Garcia Gonzalez 2004; Gholami et al. 2019).

Previous studies have used a combination of various dendrochronology data to reconstruct hydroclimate variabilities. For example, Allen et al. (2015) used wood properties in addition to tree-rings to reconstruct historical December–January inflow and streamflow in southeastern Australia.

In this research, vessel features (vessel area, diameter, and perimeters) were also incorporated into the streamflow modeling process in addition to tree-ring chronologies in the three time periods of a growing season. Their modeling results rely heavily on new tree-ring chronologies based on properties such as tracheid radial diameter, density, and cell wall thickness, underscoring the importance of these different types of chronologies in reconstructions. Second, applying state-of-art machine learning methods to predict the streamflow from dendrochronology data, a practice that has been typically performed by mathematical data-driven methods in previous studies (Anderson et al. 2019; Chen et al. 2019; Li et al. 2019; Sahour et al. 2021). Additionally, to investigate the relationships (e.g., linear, nonlinear, monotonic) between dendrochronology data (tree-rings, vessel features) and streamflow variation using variable importance (VI) and partial dependence plots. Third, the southern coastal plain of the Caspian Sea is one of the most densely populated areas in Iran, where being an agricultural and industrial pole places a higher demand on water resources. Using one of the rivers (Khalkaee river) within this region as a test site, the study aims to provide a feasible conventional approach to reconstruct the streamflow using dendrochronology data in the north of Iran. For this purpose, *Zelkova carpinifolia,* a ring-porous species, was selected for dendrochronology studies. *Zelkova carpinifolia* is a mesophytic deciduous tree that prefers to grow in the riversides, mixed lowlands, and ravine forests. This species has been previously used in dendrochronology studies (Grissino-Mayer 1993; Davis et al. 2012). For example, it was used to reconstruct precipitation and water table fluctuations on the southern coast of the Caspian Sea in Iran (Gholami et al. 2017, 2019).

Only a few dendrochronology-based reconstructions of streamflow have been carried out in the Middle East (Akkemik et al. 2008). Considering the increasing population of the region, scarcity of water resources, and the need for reliable data to investigate the impact of climate change on water resources, the methodology could potentially be adopted for other rivers across the region to fill the temporal gaps in streamflow records and reconstruct the hydrological conditions of rivers during past decades. This study aimed to predict the historical stream flow using dendrohydrology and machine learning models.

Study area

The study was conducted on a permanent river in the north of Iran, namely Khalkaee, located between 48° 50′ to 49° 20′ E and 37° 17′ to 37° 28′ N (Fig. 1). The river's length is almost 71 km, originating from Alborz Mountains and drains into the Caspian Sea. The mean discharge of the Khalkaee river is 3.9 cubic meters per second during the trees growing season (April–September). The study area has a humid climate with an average annual precipitation of 1000 mm. The primary type of precipitation in the study area is rainfall. Precipitation mainly occurs during the fall and winter, and it decreases during the growing season as temperature increases. The average annual temperature in the region is 15 °C. The average temperature during the growing season is 20.8 °C, with a minimum and a maximum of 13.5 and 31 °C, respectively.

The major types of land use are forest lands, paddy lands, and residential areas. The geologic formation consists of alluvial sediment. The altitude of the site is between 30 and 40 m above the mean sea levels. One stream gauging station (Taskoh station) over the Khalkaee river has been recording the streamflow from 1982 to the present (2020). The selected trees for records dendrochronology studies were located downstream of the river (Fig. 1).

Materials and methods

In this section, the steps toward providing the dendrochronology data and machine learning techniques for streamflow modeling are described.

Dendrochronology data

The samples were taken from *Zelkova carpinifolia* trees, also known as Caucasian elm. Caucasian elm is a native to the Caucasia and Alborz mountains (Andrews 1993). Thirty sampling trees were selected from a single site with an area of 4 km^2. The sampling trees were near the Khalkaee river, because our goal was to reconstruct the streamflow of this river (Fig. 1). All sampling trees were young, healthy, almost the same age, and with no sign of damages.

First, the full-disc samples were taken from a cross-section of the stems perpendicular to the growing axis. The samples underwent preprocessing (sanding and polishing) to enhance the visibility of the tree-rings. Samples were cut into small cubes with width of 2 cm^2. The cubes were boiled in hot water for 4 h to soften the wooden tissues. The thin cross-sectional slices were provided using a microtome. Colors were used to increase the visibility of the slices under the microscope. High-resolution pictures were taken from the tree-rings, and vessel features on a radial path using a digital camera mounted on the stereomicroscope (Fig. 2). The width of the tree-rings and vessel features (vessel diameter, vessel area, and vessel perimeter) were measured using Digimizer image analysis software with an accuracy of 0.001 mm. Cross-dating of the vessels and tree-rings was performed simultaneously according to their cambial age. The cross-dating process was evaluated in two vertical directions. We used TSAP software to perform time series analysis and cross-dating.

Non-hydrologic trends were removed from the tree-ring and vessel features time series, the program Auto Regressive STANdardization (ARSTAN) was used (Cook 1985). The generated time series was standardized using a spline function with a 50% frequency response of 50 years. Non-hydrologic growth trends were excluded by dividing the original data by the fitted curve. Finally, four tree samples were excluded from the dataset due to the inconsistency with the other tree-rings samples. Those samples produced a high standard deviation which could not represent the chronology of the sampling trees.

During the growing season, the cambium produces several large cells with thin walls that form the earlywood, also known as springwood, identified by their light color rings (due to the larger size of the constituent elements and more favorable environmental conditions) under a microscope. Toward the end of summer, when the growth rate slows, summerwood is formed. Summerwood can be identified by its small-sized and darker color cells (Zhang et al. 2008). Moreover, the width of the tree-ring is an indicator of climate variability. For example, an increase in streamflow causes wider tree-rings and larger vessel sizes, because the trees are more likely to receive adequate moisture from soil and air. Therefore, different periods of the growing season can be identified by evaluating the change in wood texture, wood color, and vessels size. Furthermore, identifying earlywood (springwood) and latewood (summer wood) is a helpful indicator in this process. Finally, the mean tree-rings width and the mean vessels' features in the three time periods of growing seasons [early spring (April), the early summer (June), and the end of summer (August) from 1970 to 2018] was measured. The mean values represent the average from the several samples. Both tree-ring width and vessel features data were provided in two types include the tree-ring width and vessel features in the three desired time period and the cumulative tree-rings and vessel features. The tree-ring width and vessel features in the desired time period shows the mean measurement of tree-ring width and vessel features values for a particular period of a growing season (early

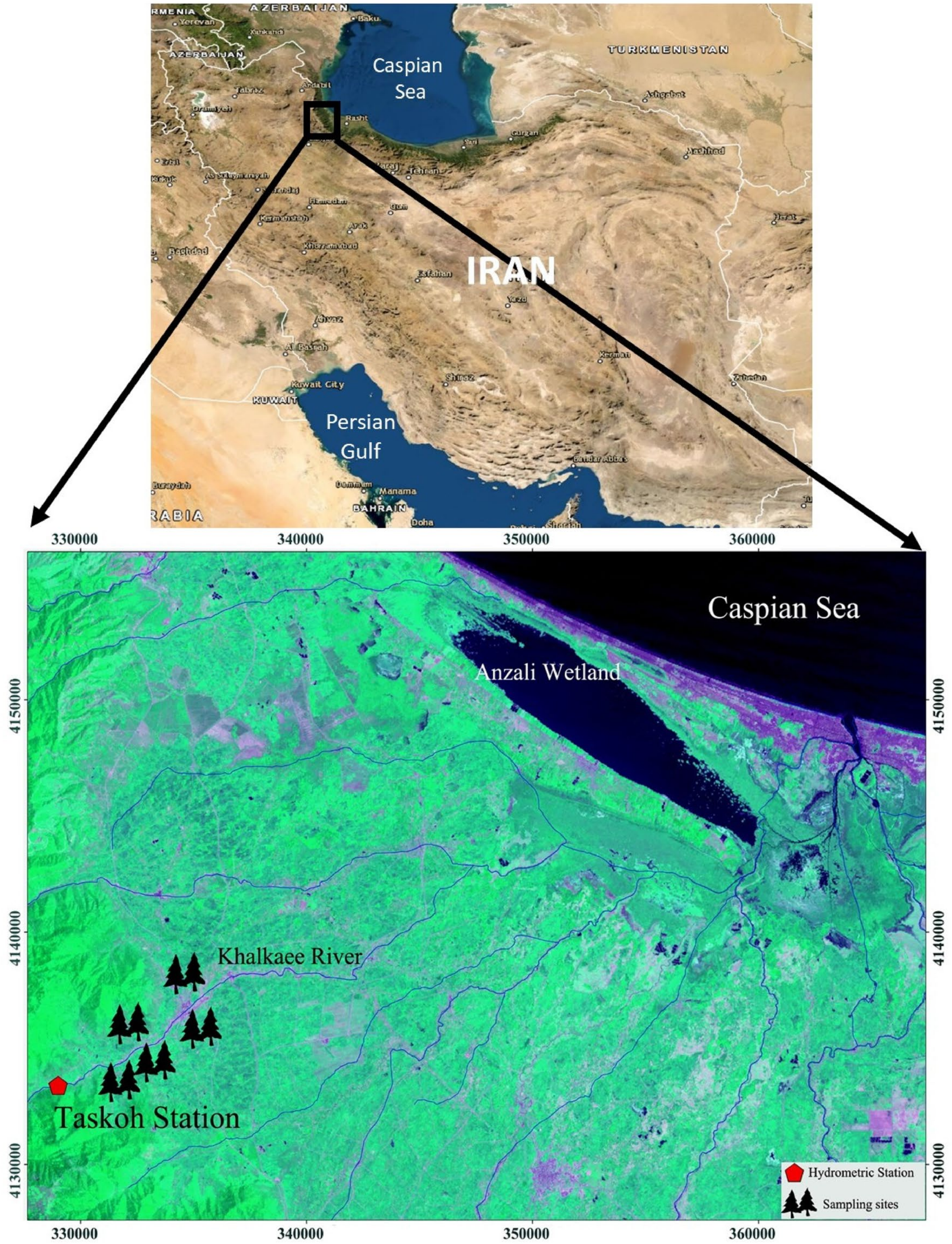


Fig. 1 Location of the study site (Khalkaee river) on the southern coast of the Caspian Sea

spring, early summer, and end of summer). The cumulative value of a particular period shows the mean tree-ring widths or vessels features from the beginning of the growing season to the desired time period.

The correlation between the chronology data was analyzed using the Pearson correlation coefficient. The dataset was tested to identify the multicollinearity among the variables using variance inflation factor (VIF), considering a threshold of 10 (VIF > 11) for individual variables as the

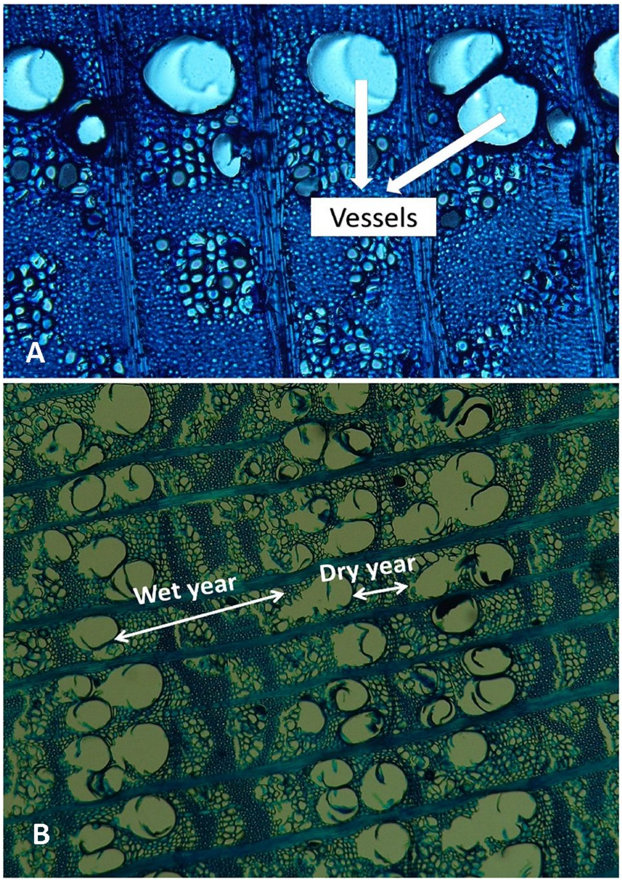


Fig. 2 Pictures taken from the samples under a microscope showing **A** vessels, and **B** different widths of the tree-rings during wet and dry years

presence of multicollinearity. High VIF values (> 11) were observed between vessel perimeter, vessel area, and vessel diameter. Therefore, the vessel perimeter and vessel area were excluded from sets of input variables.

Streamflow modeling

The streamflow records from 1982 to 2018 were obtained from the river's gauging station (Tashkoh hydrometry station). The chronology and streamflow data were randomly divided into two subsets of training (75% of the total data) and testing (25% of the total data). Two machine learning algorithms, namely random forest (RF) and extreme gradient boosting (XGB) were employed to establish a relationship between dendrochronology inputs (tree-rings and vessels features in the desired time periods, and cumulative tree-rings and vessel diameter) and streamflow during the growing season within the study period (1982–2018). Additionally, we compared the results with the traditional multivariate regression (MLR) model. The MLR derives patterns in the data and establishes the best fitting linear relationships between two or more dependent variables and the target (stream flow). In an MLR model, every value of the input variable X is associated with a value of the target variable Y.

The combination of the tree-ring widths and vessel features was used for modeling the streamflow. The training data were used for developing the models and the testing set for the evaluation of the models. The tested model and dendrochronology inputs were later used for reconstruction of the past streamflow. The streamflow data were available from 1982 to 2018, while the chronology data was available from 1970 to 2018. Therefore, the reconstruction of the streamflow was carried out for the years 1970 to 1982. Below is a detailed explanation of the RF and XGB models.

Random forest (RF)

The RF algorithm was implemented to predict the interannual streamflow using chronology data as inputs. RF is an improved version of a decision tree algorithm that combines the base principles of bagging with random feature selection to add additional diversity to the decision tree models (Breiman 2001). Decision tree learners are robust predictive modeling approaches that utilize a tree structure to establish relationships among the features and the outcomes. A tree structure mirrors how a tree begins at a wide trunk and splits into narrower branches as it is followed upward. Similarly, a decision tree learner uses a structure of branching decisions that channel examples into a final predicted class value. A decision tree is built on an entire dataset, using all the features of interest, whereas an RF randomly selects observations and specific features to build multiple decision trees and then averages the results to make predictions. In the RF model, the Gini Coefficient is used. Gini coefficient (Eq. 1) indicates how nodes on a decision tree branch. Gini is calculated as follows:

$$\text{Gini} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\overline{x}}, \tag{1}$$

where n represents the number of observations. Likewise, entropy (Eq. 2) is another indicator that determines how nodes branch in a decision tree and is calculated as

$$E_{\text{split}} = \frac{N_1}{N}E_1 + \frac{N_2}{N}E_2, \tag{2}$$

where N_1 and N_2 are the number of items of each set after the split, and E_1 and E_2 are their corresponding entropies. Random forests offer some advantages over other machine learning algorithms. For instance, it only selects the essential features and can be used on data with an extremely large number of features. A schematic diagram of the RF model is shown in Fig. 3.

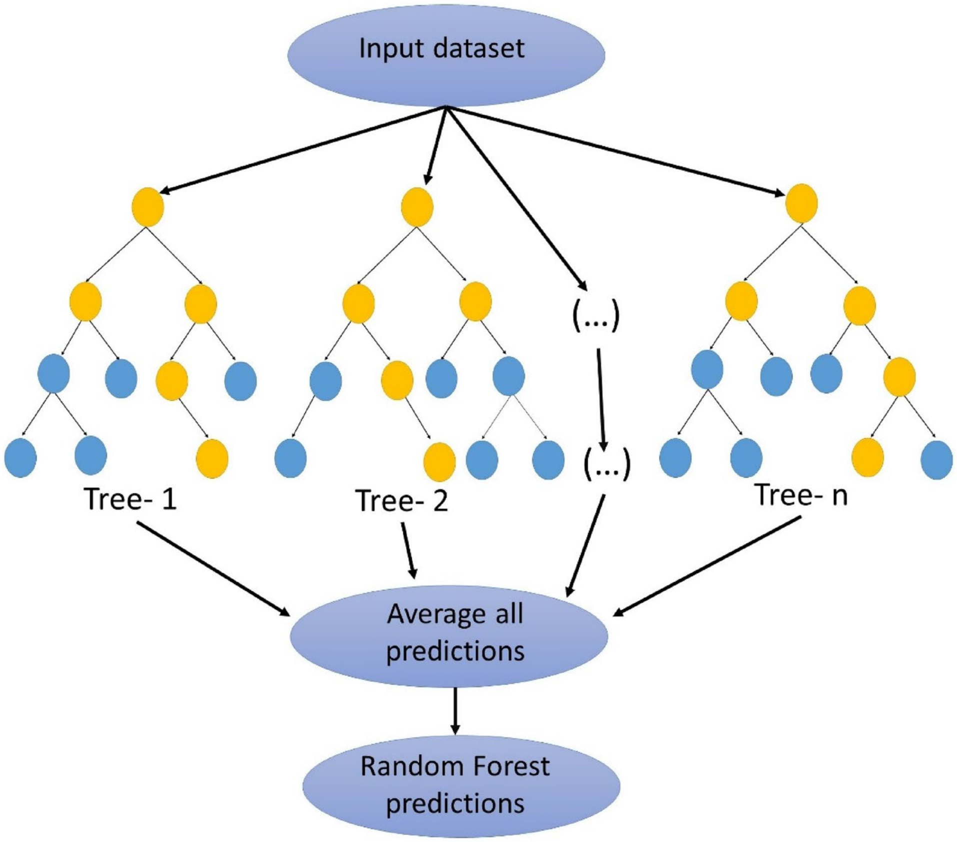


Fig. 3 Schematic diagram of the random forest algorithm

The RF model developed for this study consisted of 250 regression trees in which each tree was grown on a bootstrap sample drawn from the training dataset that contained the streamflow observations and the chronology data as input variables. One-third of the input variables (remaining in each step of variable selection) were randomly selected to grow the tree to reduce the correlation between the prediction errors of a pair of decision trees. The final predicted value in the RF model was calculated by averaging the predictions from all the individual trees. The *randomForest* package written the R programming language was used for this purpose.

Extreme gradient boosting (XGB)

XGB was introduced by Chen and Guestrin in 2016. Since its introduction, XGB has become one of the most popular machine learning techniques (Ni et al. 2020; Sahour et al. 2020b). The main idea of boosting is to add new models to the ensemble subsequently. In essence, boosting advances the bias-variance-tradeoff by starting with a weak model and sequentially boosts its performance by continuing to build new trees, where each new tree in the sequence tries to fix up where the previous one made the most significant errors. XGB is a significant improvement in Gradient Boosting. Figure 4 shows the schematic diagram of the XGB model.

Gradient boosting begins with a set of predictors $(X_1, ..., X_n)$ to predict a set of corresponding target values$(Y_1, ..., Y_n)$. We fit a model F(X) → Y and minimize the sum of the loss function $J = \sum_{i=1}^{n} L(Y_i, F(X_i))$ by improving the model $F(X)$. Here, L is a differentiable convex loss function that measures the difference between the prediction $F(X)$ and the target Y. Then, the following iterations are performed: first, we calculate the negative gradients of J with respect to $F(Xi)$,$-\frac{\partial J}{\partial F(X_i)}$. Then, fit a regression tree h, to negative gradients $-\frac{\partial J}{\partial F(X_i)}$ and finally, update $F(Xi)$ with $F(Xi) + \gamma h$, where γ is the step size to reach the estimated minimum of J. This iteration is performed until the desired accuracy is reached. In XGB, the loss function (Eq. 3) is

$$J = \sum_{i=1}^{n} L(Y_i, F\left(X_i\right) + h) + \Omega(h), \tag{3}$$

where $\Omega(h) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$. Here, T is the number of leaves in the tree, and ω is the leaf weights.

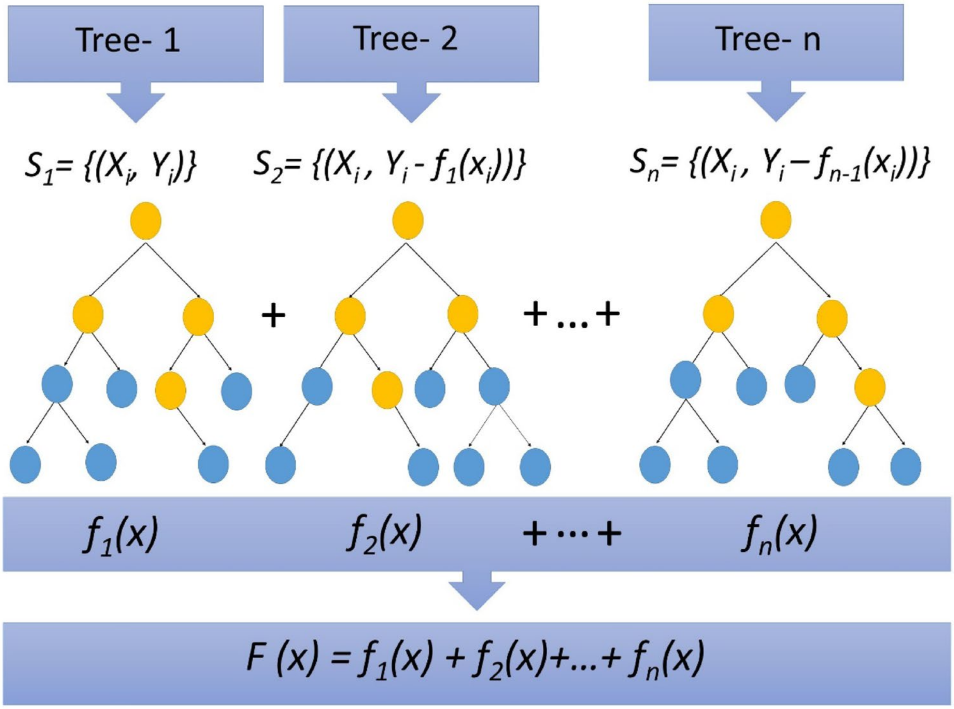

Tree- 1
Tree- 2
Tree- n
$S_1 = \{(X_i, Y_i)\}$
$S_2 = \{(X_i, Y_i - f_1(x_i))\}$
$S_n = \{(X_i, Y_i - f_{n-1}(x_i))\}$
+
+...+
$f_1(x)$
$f_2(x)$
+⋯+
$f_n(x)$
$F(x) = f_1(x) + f_2(x) + \ldots + f_n(x)$

Fig. 4 Schematic diagram of the extreme gradient boosting algorithm

Variable importance (VI)

VI represents the statistical significance of each variable in the data with respect to its effect on the generated model. VI is each predictor's ranking based on its contribution to the model. Variable importance is calculated by the sum of the decrease in error when split by a variable. Then, the relative importance is the variable importance divided by the highest variable importance value so that values are bounded between 0 and 1. The measure of VI in tree-based regressions (RF and XGB) is based on how many times a given model selects a variable for splitting and how much it is improved because of the splitting (Friedman and Meulman 2003).

Performance evaluation of the models

The adopted models (RF, XGB, and MLR) were evaluated by comparison between predicted and recorded streamflow using three statistical coefficients, namely, coefficient of determination (R-squared), Nash–Sutcliffe efficiency (NSE), and normalized root mean squared error (NRMSE). The values of these statistical coefficients range between 0 and 1. The higher values for NSE and R-squared, as well as lower values for NRMSE, indicate better prediction. The coefficients are calculated as follows:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{\left(Y_o - \hat{Y}_p\right)^2}{n}}, \tag{4}$$

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\overline{Y}_{oi}}, \tag{5}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n} \left(\hat{Y}_p - Y_o\right)^2}{\sum_{i=1}^{n} \left(Y_o - \overline{Y}_{oi}\right)^2}, \tag{6}$$

where Y_o is the recorded streamflow, $\hat{Y}_p$ is the predicted streamflow, n is the number of observations, and $\overline{Y}_{oi}$ is the average of the recorded streamflow.

Results

The streamflow during the growing season varies from 0.48 to 10.9 cubic meters per second. The mean streamflow during growing seasons was 3.9 cubic meters per second. Streamflow at the hydrometric station has been measured by hydrometric instruments. This is a relatively good flow rate for a permanent river. The mean vessel diameters in the three time periods ranged between 47 and 232 µm. The cumulative tree-ring widths ranged between 1.4 and 6.3 mm which decrease from spring to late summer. The vessel areas and perimeters ranged from 1700 to 4250 μm^2 and 140 to 1730 μm, respectively. Vessel feature decrease from spring to late summer. Moreover, expressed population signal (EPS) values were estimated to be 0.88 in the study site. Table 1 shows the correlation coefficients of tree-rings and vessel features with streamflow during the growing season. The highest correlation between streamflow and chronology parameters was associated with vessel diameter, followed by cumulative vessel diameter, cumulative tree-rings, and tree-rings. The variation in streamflow and individual chronology parameters (tree-ring and vessel diameter in the desired times and cumulative tree-ring widths and vessel diameter) during the growing season of the modeling period (1982–2018) is presented in Fig. 5.

According to the statistical analysis, sensitivity analysis, and variable importance (VI), the vessel diameter, the cumulative vessel diameters, and the cumulative tree-ring widths are the most suitable inputs for streamflow modeling. The evaluation of the adopted models on the training and testing subset is presented in Table 2. The results show that the XGB model performed better considering the higher NSE and *R*-squared and the lower NRMSE value in both the training (NSE: 0.98; *R*-squared: 0.98; NRMSE: 0.13) and testing (NSE: 0.81; *R*-squared: 0.87; NRMSE: 0.43) stage. Figure 6 shows the scatter plot of the observed and predicted streamflow in the training and testing stages. The results of the test stage showed the high performance of the XGB model in predicting the streamflow using vessels diameter and tree-ring widths during the growing season.

The variable importance for both XGB and RF models is presented in Fig. 7. The result shows that for both models, the vessel diameter is the most important parameter in the modeling process, followed by the cumulative vessel diameters, cumulative tree-ring widths, and tree-rings. However, the importance ratio of each variable varies between XGB and RF. The partial dependence plot (PDP) shows the marginal effect that input variables have on the predicted outcome of a machine learning model (Fig. 8). A partial dependence plot shows whether the relationship between the target and a predictor is linear, monotonic, or mixed. Figure 8 indicates that in the XGB model, streamflow has a linear relationship with tree-ring width, vessel diameter, and cumulative vessel diameter. The relationship between streamflow and cumulative tree-ring widths is nonlinear. The cumulative tree-rings increase by an increase in streamflow; after some point, the cumulative tree-ring widths decrease by an increase in streamflow (Fig. 8). One plausible explanation is that the vessels already reach their maximum size by providing all their water needs, and therefore, the presence of additional water does not affect the vessel size anymore. It should be noted that this analysis is only valid for growing seasons when there is a meaningful relationship between tree-rings and environmental moisture.

Table 1 The correlation coefficients and VIF values of the dendrochronology variables for streamflow modeling

Variable name	VIF	Correlation with streamflow
Tree-rings (mm)	2.2	0.68
Cumulative tree-rings (mm)	3.7	− 0.7
Vessels diameter (μm)	8.6	0.87
Cumulative vessels diameter (μm)	9.4	0.81
[a]Vessels area (μm)	14.4	0.65
[a]Vessels perimeter (μm)	12.8	0.62

[a]Variables excluded due to the presence of multicollinearity (VIF > 11)

The streamflow for the period 1970 to 1981 (the period with the absence of streamflow data, and the presence of dendrochronology records) was reconstructed using the tested RF and XGB models (Fig. 9). Results show that streamflow during the growing season between 1970 and 1981, estimated by the XGB model, ranged from 0.45 to 10.8 m^3/s. The streamflow estimated by the RF model was between 1.62 and 7.78 m^3/s.

As we mentioned before, we also used the widely used multivariate regression model to compare the results with adopted machine learning algorithms. The results showed that both RF and XGB logarithms outperform the regression model. For the regression model, the NSE, NRMSE, and *R*-squared in the test set were found to be 0.66, 0.57, and 0.73, respectively (Table 2; Fig. 9).

Discussion

This study used vessel features and tree-rings as inputs of two machine learning techniques (RF and XGBoost) to model the temporal variation of streamflow in a permanent river in the north of Iran. The analysis of the chronology inputs was conducted to select the optimal dendrochronology parameters for modeling and reconstruction of the streamflow in the growing seasons. One of the strengths of this study was to incorporate vessel features in addition to tree-rings chronology for increasing machine learning performance in streamflow modeling. We investigated the response of the vessel features through statistical analysis. VIF and correlation coefficients were used to assess the relationships between individual variables and streamflow as the target and to investigate the multicollinearity among

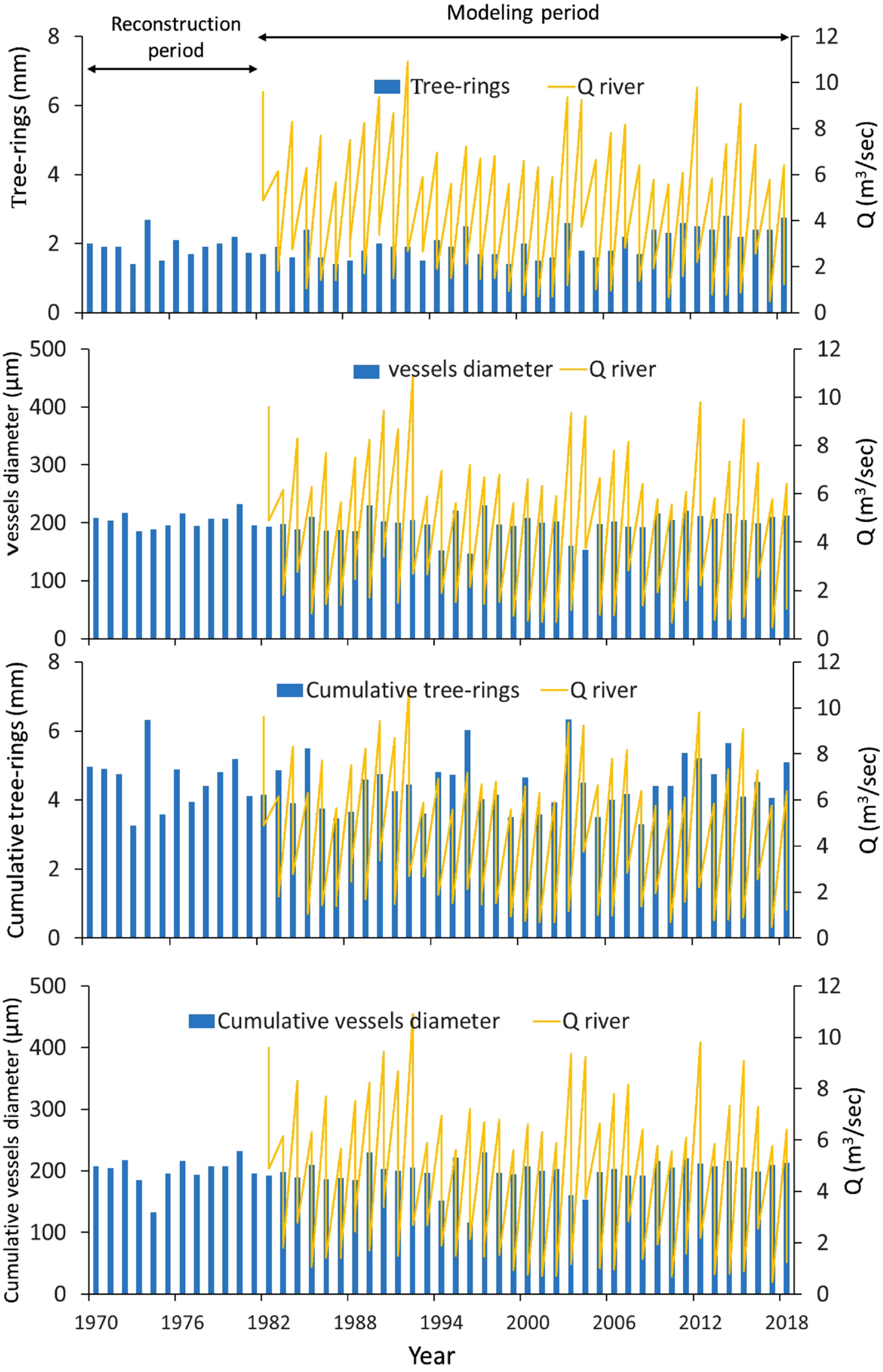


Fig. 5 Time series each chronology parameter versus reconstructed streamflow

the variables. Several vessel features, including vessel diameter, vessel area, and vessel perimeter, were extracted from the tree samples. However, due to the presence of multicollinearity among the other vessel features, only the vessel diameter and cumulative vessel diameters were selected as optimum vessel features.

The evaluation of the machine learning models shows the high performance of the adopted methodology in

Table 2 Performance of the models in the training and testing stage for streamflow modeling

	NSE		NRMSE		*R*-squared	
	Training	Testing	Training	Testing	Training	Testing
Extreme gradient boosting	0.98	0.81	0.13	0.43	0.98	0.87
Random forest	0.85	0.71	0.38	0.52	0.87	0.82
Multivariate regression	0.80	0.66	0.44	0.57	0.79	0.73

NSE Nash–Sutcliffe efficiency, *NRMSE* normalized root mean squared error, *R-squared* coefficient of determination

modeling and reconstruction of streamflow. The statistical criteria (NSE, NRMSE, and *R*-squared) obtained by comparing recorded and predicted streamflow on the testing subset are very high. Moriasi et al. (2007) used the NSE and NRMSE to rank the performance of the models in hydrologic studies. According to their results, the NSE values greater than 0.75 and NRMSE less than 0.5 are indicators of the high performance of the predictive models, which was the case in this research. Using the adopted methodology, we could also analyze the importance of individual chronology parameters in streamflow modeling.

The major limitation of the study was the use of a relatively small dataset for the modeling process. Unfortunately, additional samples were not available for this study. It is well known that the reliability and the performance of the machine learning models increase in a larger dataset (Steedman et al. 2003). Therefore, providing additional observations from trees for a longer period of time could enhance the results and subsequently improve the reliability of the values for the reconstruction of streamflow during the past decades. Another limitation of streamflow modeling is the reconstruction of extremely high values. This is due to the fact that once the tree is satisfied with its water requirement, additional water is not used by the trees to enlarge the tree-rings and vessel size. Therefore, it is difficult to reconstruct the extremely high streamflow using dendrochronology inputs. However, very low streamflow values indicate the absence of adequate moisture for the trees. This affected the tree-rings and vessel growths during the growing season and was reflected in the chronology records. Therefore, the method works better for the reconstruction of low streamflow compared to extremely high flow rates. The high sensitivity of tree-rings and vessel features to low flows indicates the applicability of the adopted methodology, especially in the reconstruction of river flow during droughts. The precise streamflow modeling is an important step for many water resource management plans. Moreover, reconstruction of the minimum river flows can be used for investigating the duration and intensity of hydrologic droughts during the past decades.

Conclusion

In this study, dendrochronology parameters, including tree-ring and vessel diameter and cumulative measurements of tree-ring width and vessel diameter in the three time periods of the growing season was used to model and reconstruct the streamflow in the Khalkaee river of Iran. The statistical analysis through correlation analysis, variable importance, and partial dependence plots indicated significant relationships between chronology parameters and streamflow. Adopted machine learning methods (RF and XGB) have proven to be capable tools for streamflow modeling, given the enhanced performance of the models in the testing subset than LR method. The selection of suitable sites and tree samples are two critical parameters in dendrochronology studies. The trees were selected from *Zelkova carpinifolia* species that are highly sensitive to the presence of moisture and can better reflect the temporal variation of hydrological parameters. The sampling trees were also selected from those near the Khalkaee river. The study provides a feasible and new approach for modeling and reconstruction of the streamflow. The adopted methodology is suitable for modeling and reconstructing the streamflow. According to the results, the presented methodology is more suitable for the streams with low flows and unstable discharge regimes. Because the maximum error in the predicted streamflow was observed in modeling the maximum flows and low flows, and high fluctuations in river discharge and especially, the occurrence of droughts increase the performance of modeling using tree-rings and vessel features.

For the future, we suggest applying the presented methodology with more observation records for other rivers across the region or similar settings elsewhere.

Fig. 6 Performance of XGB, RF, and LR models on the training and test subset

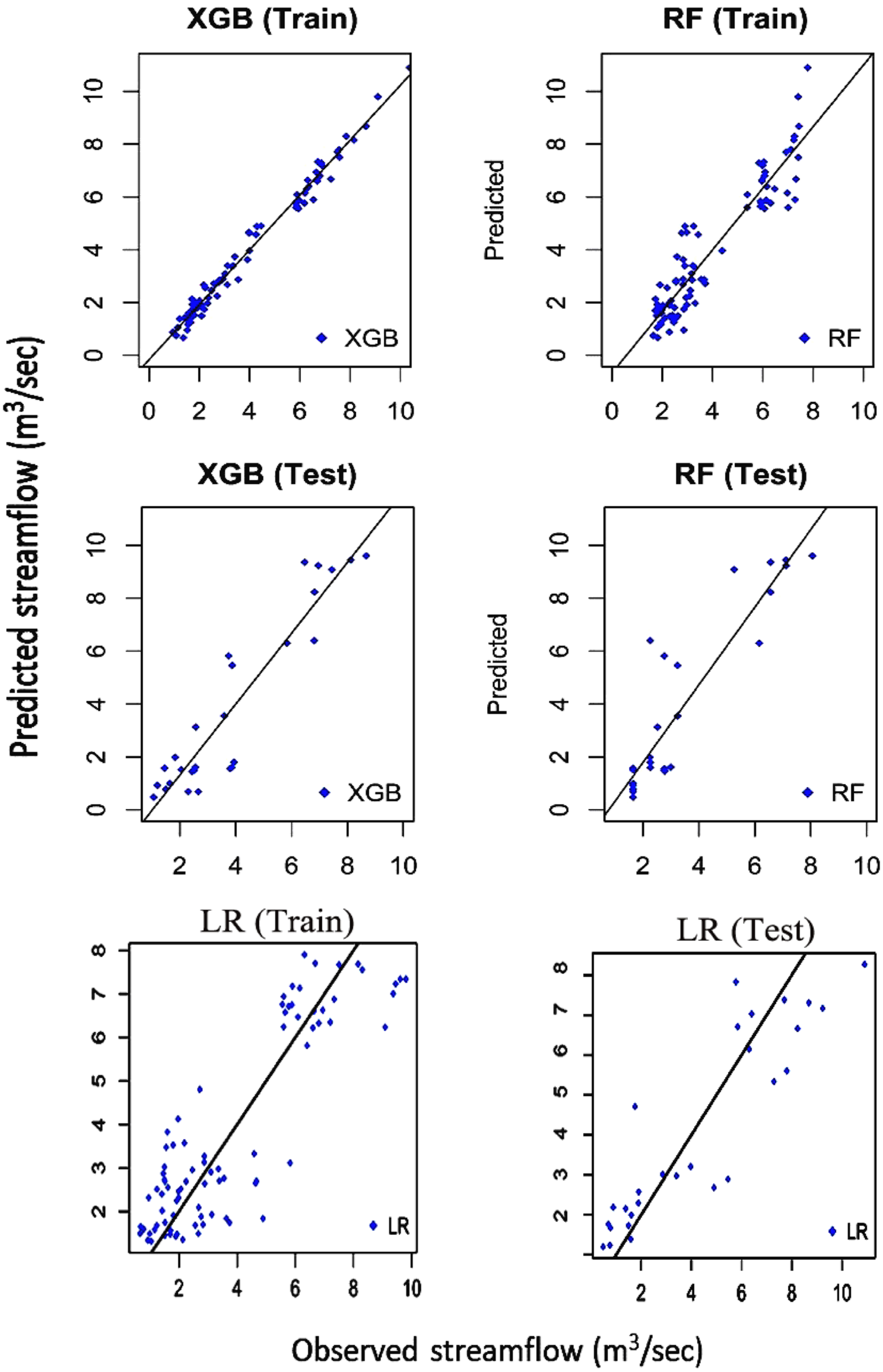
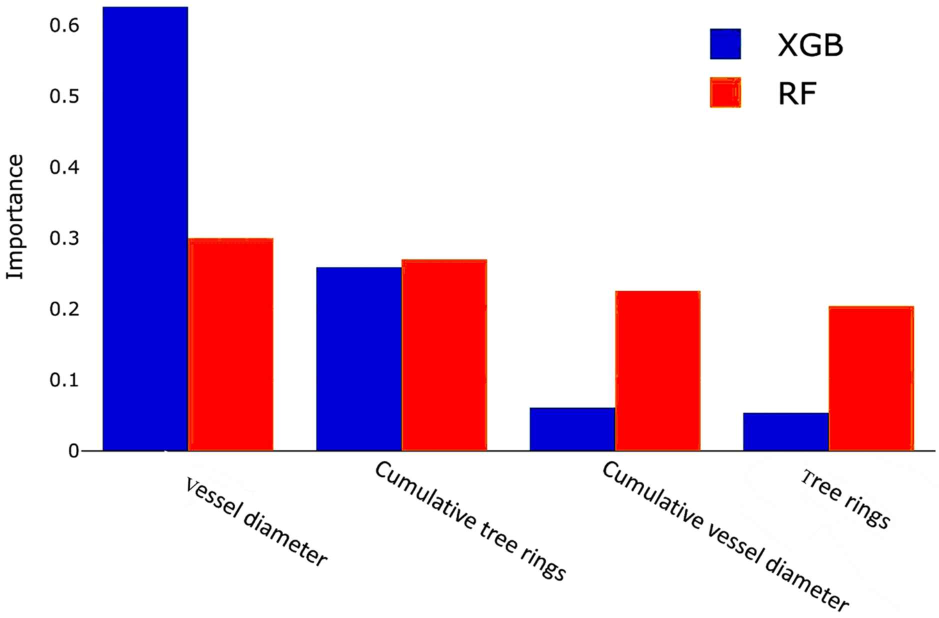


Fig. 7 Variable importance (VI) for the extreme gradient boosting and random forest



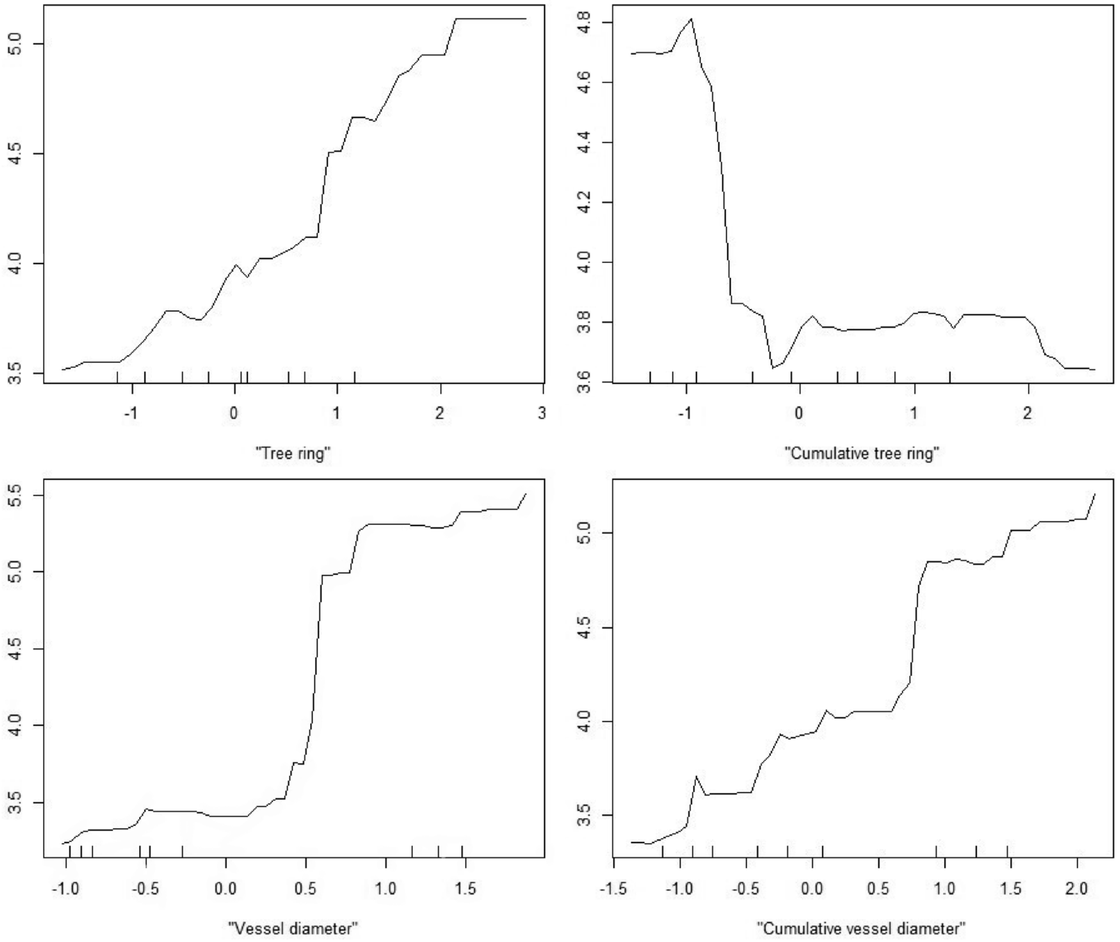


Fig. 8 Partial dependence of the chronology data in extreme gradient boosting model

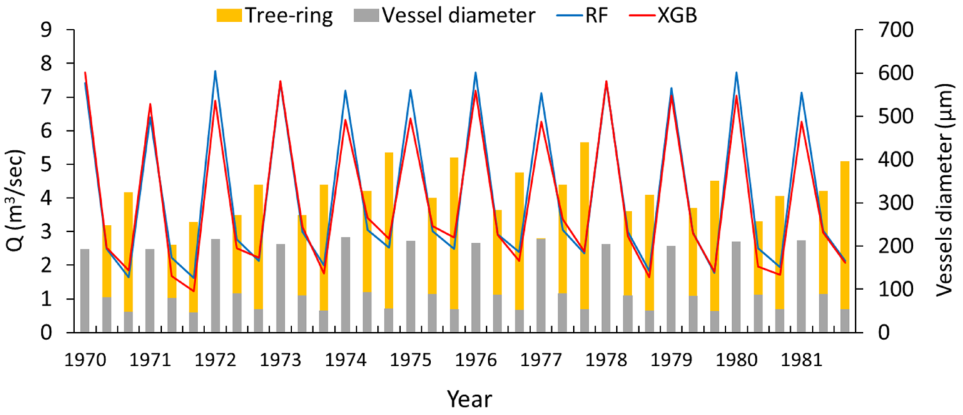


Fig. 9 Reconstruction of past streamflow using RF and XGB models. Values are plotted against the temporal changes of vessel diameter and tree-rings

Acknowledgements We thank the Regional Water Company of Guilan for providing the streamflow data and for helping us with the data preprocessing.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

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

Akkemik Ü, D'Arrigo R, Cherubini P, Köse N, Jacoby GC (2008) Tree-ring reconstructions of precipitation and streamflow for north western Turkey. Int J Climatol 28(2):173–183

Allen KJ, Nichols SC, Evans R, Cook ER, Allie S, Carson G, Ling F, Baker PJ (2015) Preliminary December–January inflow and streamflow reconstructions from tree-rings for western Tasmania, southeastern Australia. Water Resour Res 51(7):5487–5503

Alshehri F, Sultan M, Karki S, Alwagdani E, Alsefry S, Alharbi H, Sahour H, Sturchio N (2020) Mapping the distribution of shallow groundwater occurrences using remote sensing-based statistical modeling over southwest Saudi Arabia. Remote Sens 12(9):1361. https://doi.org/10.3390/rs12091361

Anderson S, Ogle R, Tootle G, Oubeidillah A (2019) Tree-ring reconstructions of streamflow for the tennessee valley. J Hydrol 6(2):34. https://doi.org/10.3390/hydrology6020034

Andrews S (1993) Tree of the year: Zelkova. Int Dendrol Soc Yearb, vol 1, pp 11–30

Breiman L (2001) Random forests. Mach Learn 2001(45):5–32. https://doi.org/10.1023/A:1010933404324

Campelo F, Nabais C, Gutiérrez E, Freitas H, García-González I (2010) Vessel features of *Quercus ilex* L. growing under Mediterranean climate have a better climatic signal than tree-ring width. Trees 24(3):463–470. https://doi.org/10.1007/s00468-010-0414-0

Case RA, MacDonald GM (2003) Tree-ring reconstructions of streamflow for three Canadian prairie rivers 1. J Am Resour Assoc 39(3):703–716

Chen F, Shang H, Panyushkina IP, Meko DM, Yu S, Yuan Y, Chen F (2019) Tree-ring reconstruction of Lhasa River streamflow reveals 472 years of hydrologic change on southern Tibetan Plateau. J Hydrol 572:169–178

Chen T, Guestrin C (2016) August. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 85–794

Cook ER (1985) A time series analysis approach to tree-ring standardization. Ph.D. thesis, The University of Arizona

Davis EL, Laroque CP, Van Rees K (2012) Evaluating the suitability of nine shelterbelt species for dendrochronological purposes in the Canadian Prairies. Agroforest Syst 87(3):713–727

Ferrero ME, Villalba R, De Membiela MD, Hidalgo LF, Luckman BH (2015) Tree-ring based reconstruction of Río Bermejo stream flow in subtropical South America. J Hydrol 525:572–584

Fonti P, Garcia Gonzalez I (2004) Suitability of chestnut earlywood vessel chronologies for ecological studies. New Phytol 163(1):77–86. https://doi.org/10.1111/j.1469-8137.2004.01089.x

Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. Stat Med 22(9):1365–1381

Gholami V, Chau KW, Fadaee F, Torkaman J, Ghaffari A (2015) Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. J Hydrol 529:1060–1069

Gholami V, Torkaman J, Khaleghi MR (2017) Dendrohydrogeology in paleohydrogeologic studies. Adv Water Resour 110:19–28

Gholami V, Torkaman J, Dalir P (2019) Simulation of precipitation time series using tree-rings, earlywood vessel features, and artificial neural network. Theor Appl Climatol 137(3–4):1939–1948

Gou X, Deng Y, Chen F, Yang M, Fang K, Gao L, Yang T, Zhang F (2010) Tree-ring based streamflow reconstruction for the Upper Yellow River over the past 1234 years. Chin Sci Bull 55(36):4179–4186

Grissino-Mayer, (1993) An updated list of species used in tree-ring research. Tree-Ring Bull 53:17–43

Harley GL, Maxwell JT, Larson E, Grissino-Mayer HD, Henderson J, Huffman J (2017) Suwannee River flow variability 1550–2005 CE reconstructed from a multispecies tree-ring network. J Hydrol 544:438–451. https://doi.org/10.1016/j.jhydrol.2016.11.020

Hirsch RM, Costa JE (2004) US stream flow measurement and data dissemination improve. EOS Trans Am Geophys Union 85(20):197–203. https://doi.org/10.1029/2004EO200002

Kames S, Tardif JC, Bergeron Y (2016) Continuous earlywood vessels chronologies in floodplain ring-porous species can improve dendrohydrological reconstructions of spring high flows and flood depths. J Hydrol 534:377–389. https://doi.org/10.1016/j.jhydrol.2016.01.002

Khaleghi MR (2018) Application of dendroclimatology in evaluation of climatic changes. J for Sci 64(3):139–147. https://doi.org/10.17221/79/2017-JFS

Khan SA, See L (2006) September. Rainfall-runoff modelling using data driven and statistical methods. In: 2006 international conference on advances in space technologies. IEEE, pp 16–20. https://doi.org/10.1109/ICAST.2006.313789

Li J, Wang Z, Lai C, Zhang Z (2019) Tree-ring-width based streamflow reconstruction based on the random forest algorithm for the source region of the Yangtze River, China. CATENA 183:104216. https://doi.org/10.1016/j.catena.2019.104216

Liu Y, Liu H, Song H, Li Q, Burr GS, Wang L, Hu S (2017) A monsoon-related 174-year relative humidity record from tree-ring δ18O in the Yaoshan region, eastern central China. Sci Total Environ 593:523–534

Liu Y, Ta W, Cherubini P, Liu R, Wang Y, Sun C (2018) Elements content in tree-rings from Xi'an, China and environmental variations in the past 30 years. Sci Total Environ 619:120–126

Meko DM, Woodhouse CA, Morino K (2012) Dendrochronology and links to streamflow. J Hydrol 412:200–209. https://doi.org/10.1016/j.jhydrol.2010.11.041

Meng E, Huang S, Huang Q, Fang W, Wu L, Wang L (2019) A robust method for non-stationary streamflow prediction based on improved EMD-SVM model. J Hydrol 568:462–478. https://doi.org/10.1016/j.jhydrol.2018.11.015

Mokria M, Gebrekirstos A, Abiyu A, Bräuning A (2018) Upper Nile River flow reconstructed to AD 1784 from tree-rings for a long-term perspective on hydrologic-extremes and effective water resource management. Quatern Sci Rev 199:126–143. https://doi.org/10.1016/j.quascirev.2018.09.011

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans ASABE 50:885–900

Ni L, Wang D, Wu J, Wang Y, Tao Y, Zhang J, Liu J (2020) Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. J Hydrol. https://doi.org/10.1016/j.jhydrol.2020.124901

Qi W, Liu J, Yang H, Zhu X, Tian Y, Jiang X, Huang X, Feng L (2020) Large uncertainties in runoff estimations of GLDAS versions 2.0 and 2.1 in China. Earth Space Sci 7(1):e2019EA000829. https://doi.org/10.1029/2019EA000829

Sahour H, Gholami V, Vazifedan M (2020a) A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. J Hydrol. https://doi.org/10.1016/j.jhydrol.2020.125321

Sahour H, Sultan M, Vazifedan M, Abdelmohsen K, Karki S, Yellich JA, Gebremichael E, Alshehri F, Elbayoumi TM (2020b) Statistical applications to downscale GRACE-derived terrestrial water storage data and to fill temporal gaps. Remote Sens 12(3):533. https://doi.org/10.3390/rs12030533

Sahour H, Gholami V, Vazifedan M, Saeedi S (2021) Machine learning applications for water-induced soil erosion modeling and mapping. Soil Tillage Res 211:105032. https://doi.org/10.1016/j.still.2021.105032

Steedman M, Osborne M, Sarkar A, Clark S, Hwa R, Hockenmaier J, Ruhlen P, Baker S, Crim J (2003) Bootstrapping statistical parsers from small datasets. In: 10th conference of the European chapter of the association for computational linguistics, pp 331–338. https://doi.org/10.3115/1067807.1067851

Therrell MD, Bialecki MB (2015) A multi-century tree-ring record of spring flooding on the Mississippi River. J Hydrol 529:490–498

Tongal H, Booij MJ (2018) Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. J Hydrol 564:266–282

Urrutia RB, Lara A, Villalba R, Christie DA, Le Quesne C, Cuq A (2011) Multicentury tree-ring reconstruction of annual streamflow for the Maule River watershed in south central Chile. Water Resour Res. https://doi.org/10.1029/2010WR009562

Wang L, Li X, Ma C, Bai Y (2019) Improving the prediction accuracy of monthly streamflow using a data-driven model based on a double-processing strategy. J Hydrol 573:733–745. https://doi.org/10.1016/j.jhydrol.2019.03.101

Wu Y, Gan TY, She Y, Xu C, Yan H (2020) Five centuries of reconstructed streamflow in Athabasca River Basin, Canada: non-stationarity and teleconnection to climate patterns. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2020.141330

Zhang Y, Post D (2018) How good are hydrological models for gap-filling streamflow data? Hydrol Earth Syst Sci 22(8):4593–4604

Zhang C, Huang B, Piper JD, Luo R (2008) Biomonitoring of atmospheric particulate matter using magnetic properties of Salix matsudana tree-ring cores. Sci Total Environ 393(1):177–190

Zhang JL, Wang X, Sun WN, Li YP, Liu ZR, Liu YR, Huang GH (2020) Application of fiducial method for streamflow prediction under small sample cases in Xiangxihe watershed, China. J Hydrol. https://doi.org/10.1016/j.jhydrol.2020.124866

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.