sustainability

MDPI

Article

A C-Vine Copula-Based Quantile Regression Method for Streamflow Forecasting in Xiangxi River Basin, China

Huawei Li [1,2], Guohe Huang [3], Yongping Li [2,3,*], Jie Sun [4] and Pangpang Gao [1]

[1] Sino-Canada Energy and Environmental Research Center, North China Electric Power University, Beijing 102206, China; hwlee121@163.com (H.L.); 17755837728@163.com (P.G.)
[2] State Key Joint Laboratory of Environmental Simulation and Pollution Control, School of Environment, Beijing Normal University, Beijing 100875, China
[3] Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, SK S4S 7H9, Canada; huang@iseis.org
[4] School of Environmental Science and Engineering, Xiamen University of Technology, Xiamen 361024, China; sunj@xmut.edu.cn
* Correspondence: yongping.li@iseis.org

Abstract: In this study, a C-vine copula-based quantile regression (CVQR) model is proposed for forecasting monthly streamflow. The CVQR model integrates techniques for vine copulas and quantile regression into a framework that can effectively establish relationships between the multidimensional response-independent variables as well as capture the upper tail or asymmetric dependence (i.e., upper extreme values). The CVQR model is applied to the Xiangxi River basin that is located in the Three Gorges Reservoir area in China for monthly streamflow forecasting. Multiple linear regression (MLR) and artificial neural network (ANN) are also compared to illustrate the applicability of CVQR. The results show that the CVQR model performs best in the calibration period for monthly streamflow prediction. The results also indicate that MLR has the worst effects in extreme quantile (flood events) and confidence interval predictions. Moreover, the performance of ANN tends to be overestimated in the process of peak prediction. Notably, CVQR is the most effective at capturing upper tail dependences among the hydrometeorological variables (i.e., floods). These findings are very helpful to decision-makers in hydrological process identification and water resource management practices.

Keywords: streamflow forecasting; C-vine copula; quantile regression; joint dependencies; water resource management

check for updates

Citation: Li, H.; Huang, G.; Li, Y.; Sun, J.; Gao, P. A C-Vine Copula-Based Quantile Regression Method for Streamflow Forecasting in Xiangxi River Basin, China. *Sustainability* **2021**, *13*, 4627. https://doi.org/10.3390/su13094627

Academic Editor: Wen-Cheng Liu

Received: 8 March 2021
Accepted: 19 April 2021
Published: 21 April 2021

Highlights

- A C-vine copula-based quantile regression (CVQR) model is developed.
- The CVQR model is applied to monthly streamflow forecasting in the Xiangxi River basin.
- It can establish relationships between multidimensional response and independent variables.
- It can also capture tail or asymmetric dependences such as extremes values.
- The results are helpful to decision-makers in water resource management practices.

1. Introduction

With continuously growing populations, water resources are becoming more and more important for urbanization and agricultural intensification, especially for developing countries [1–3]. In the process of water resource planning, streamflow forecasting plays a key role in hydrological risk assessment, reservoir operations, drought/flood prevention, and water resource allocation [4–6]. More importantly, the management efficiency of water resource systems mainly depends on the reliability and accuracy of hydrological prediction. Consequently, it is desirable to employ streamflow forecasting models for effective water resources planning and management.

Over the last few decades, great efforts have been made towards developing advanced forecasting techniques to improve hydrological prediction, including process-driven and data-driven statistical approaches [7–9]. Process-based modeling methods are based on the principle of water cycle balance coupling various physical processes, such as precipitation, evaporation, infiltration, and other processes [10,11]. These models use large amounts of data (e.g., hydrometeorology, topography, and land use/cover) and robust calibration techniques, while data-driven models can be easily built in practice without considering physical process information from hydrological models and have been extensively used [12–14]. Therefore, data-driven technology is very useful and valuable as an option for streamflow forecasting.

Previously, a variety of data-driven modeling techniques were proposed and promoted for streamflow forecasting, including autoregressive moving average, multiple linear regression (MLR), stepwise cluster analysis, artificial neural networks (ANN), genetic programming, and support vector regression (SVR) [15–17]. For example, Besaw et al. [18] employed the ANN method for streamflow forecasting in ungauged basins. The results showed that local climate measurements with time delays as the input to the model are key to improving hydrological forecasting. Guo et al. [19] coupled an SVR model with adaptive insensitive factors to predict monthly streamflow, which was proven to be effective and to have high accuracy in streamflow prediction. Terzi and Ergin [20] used autoregressive (AR) modeling, gene expression programming (GEP), and adaptive neuro-fuzzy inference system (ANFIS) to predict the monthly mean flow of a watershed in Turkey. The results indicated that the developed models had good performance. Fan et al. [21] established a stepwise cluster forecasting (SCF) model for monthly streamflow forecasting, which effectively reflected the nonlinear and discrete relationships between climatic factors and streamflow. In general, these data-driven techniques can effectively simulate hydrological elements by capturing the complex interrelationships among the multiple hydrometeorological inputs. However, these models can often be flawed when predicting outliers (such as flood events), leading to illusory relationships between the response and independent variables [22].

To overcome these limitations, in this study, the copula method is proposed to flexibly construct the joint distribution to describe the complicated dependence structure between stochastic variables. Copula functions have been extensively applied to construct multivariate models and forecasting in several areas such as flood frequency and drought analysis, rainfall and climate predictions, financial risks, and energy [23–26]. However, it is difficult to derive multivariate copulas directly. Fortunately, vines known as pair copula constructions (PCCs) can describe the correlation structures between high-dimensional response-independent variables, providing an efficient and flexible tool to analyze the dependency structures between complex coupled correlated variables [27]. Moreover, the vine copulas coupling the quantile regression provide a more complete statistical analysis of random relationships between stochastic variables, such as tail or asymmetric dependence. Specially, quantile regression (QR) was introduced by Koenker and Bassett to estimate the conditional quantiles [28]. Given the distribution of the variables, the QR method can capture the total variation, heavy tail, skewness, and kurtosis of variables and can support the calculation of confidence intervals. Moreover, the method can estimate the levels of risk in extreme cases [29,30]. Quantile regression has been successfully applied in various scientific fields, such as economics, finance, and medicine [31–33]. Therefore, this study integrates the copula and quantile regression methods to explore the complex dependence among variables. Notably, the data-driven model is often influenced by the division of training and validation data sets. In many cases, the simulation and validation effects of the model are often affected by the data inputs, especially in a changing climate environment. Therefore, in order to overcome the possible influence of different data inputs on the model and randomness errors in the simulation process, the calibration and verification data sets are divided at certain points with the five-fold cross-validation method. In this study, the predictions are repeated five times using different training and test data sets.

Therefore, this study aims to develop a C-vine copula-based quantile regression (CVQR) model for streamflow forecasting. The proposed CVQR model can construct a conditional copula prediction model to capture the relationship between streamflow and hydrometeorology variables. The developed method has advantages in (i) modelling the dependence among the multidimensional response-independent variables, (ii) revealing the complicated interrelationships among hydrometeorological factors, and (iii) outperforming MLR and ANN on issues related to upper tail dependence (i.e., flood events). These findings are very helpful to decision-makers in hydrological process identification and water resource management practices.

In this study, the CVQR model is applied to the Xiangxi River basin to illustrate its applicability in streamflow prediction with multiple hydrometeorological factors. Specially, the structure of this article is as follows. Firstly, the MLR, ANN, and CVQR models are introduced in Section 2. Next, the study area and database, and the method of evaluation for the various functions are depicted in Section 3. In Sections 4 and 5, relevant results from the proposed model applied in our research area, and a comparison with and discussion about the results of different models are described.

2. Model Development

In this study, multiple linear regression (MLR), artificial neural network (ANN), and the proposed C-vine copula-based quantile regression (CVQR) models are used for streamflow forecasting. In the model development section, the MLR, ANN, and CVQR models are described, which together constitute the main modules of the proposed framework shown in Figure 1. Generally, the framework of this study entails the next four steps: (1–2) fitting and standardizing the predictors (i.e., $x_1, \ldots\ x_{n-1}$) and predicted variable (x_n); (3) simulating the monthly streamflow for the calibration process using the MLR, ANN, and proposed CVQR models; and (4) performing monthly streamflow prediction during the calibration and verification periods based on the results of step 3 and comparing the results of R^2, RMSE, and NSE for each model.

2.1. Multiple Linear Regression (MLR)

The purpose of multiple linear regression (MLR) is to investigate the relationship between the independent variables and a dependent variable. Assuming that the dependent variable y is a function of n independent variables x_1, x_2, x_3, ..., x_n, then the MLR can be expressed as follows:

$$y = a + b_1 x_1 + \ldots + b_n x_n + e \tag{1}$$

where a indicates the intercept; $b_1, \ldots, b_n$ are the slope coefficients of the corresponding independent variables; e is the random error; and y represents the independent variable. For more details, please refer to Yan and Su [34]. In this study, a generalized linear regression model is used to fit the relationship between the response variable y (monthly streamflow data) and the explanatory variables x (other hydrometeorological factors), and then, the model is used to predict the streamflow (y) with the new observations (x).

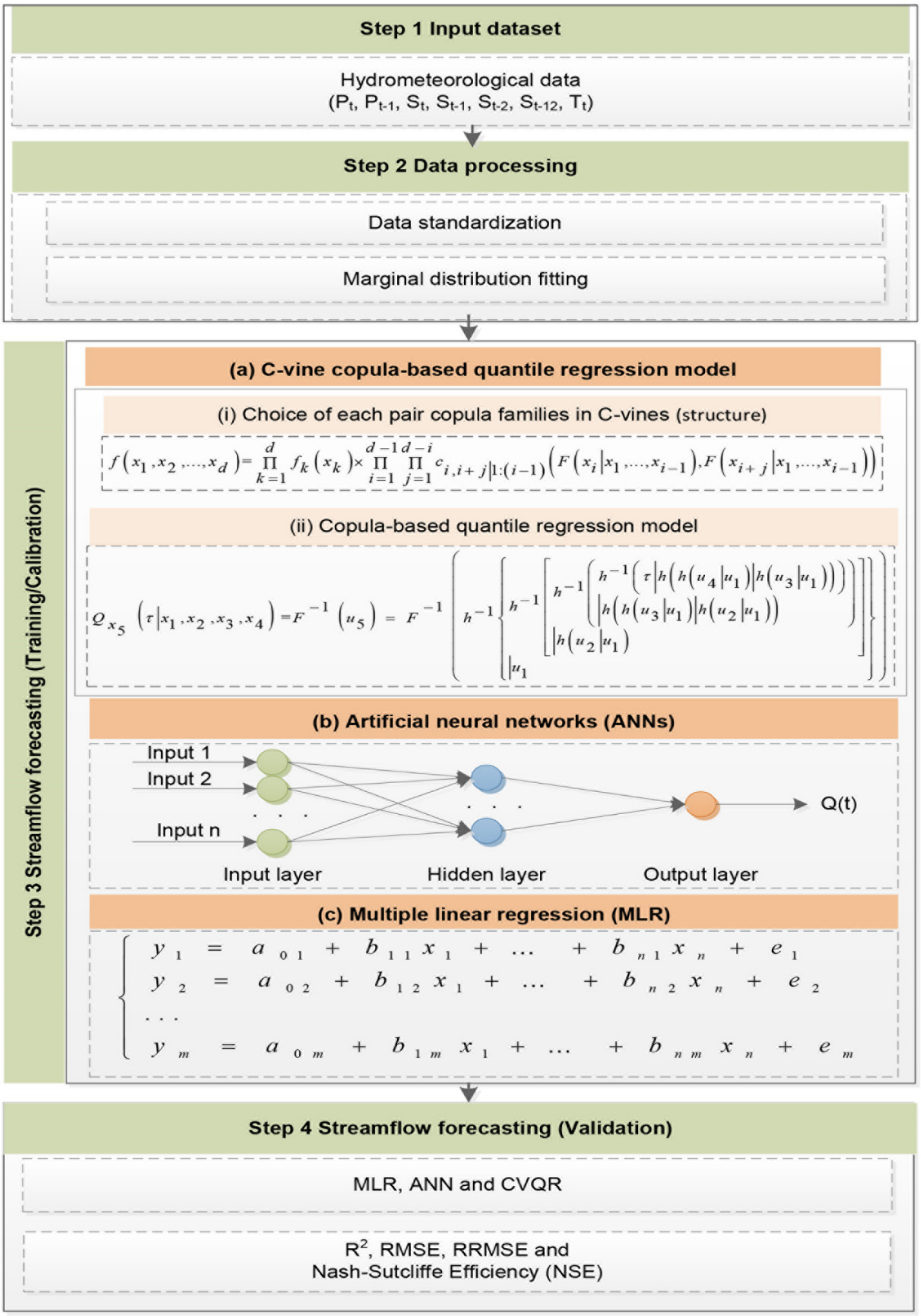
Step 1 Input dataset
Hydrometeorological data
(P_t, P_{t-1}, S_t, S_{t-1}, S_{t-2}, S_{t-12}, T_t)
Step 2 Data processing
Data standardization
Marginal distribution fitting
Step 3 Streamflow forecasting (Training/Calibration)
(a) C-vine copula-based quantile regression model
(i) Choice of each pair copula families in C-vines (structure)
$f(x_1,x_2,\dots,x_d)=\prod_{k=1}^{d} f_k(x_k)\times\prod_{i=1}^{d-1}\prod_{j=1}^{d-i} c_{i,i+j|1:(i-1)}\left(F(x_i|x_1,\dots,x_{i-1}),F(x_{i+j}|x_1,\dots,x_{i-1})\right)$
(ii) Copula-based quantile regression model
$Q_{x_5}(\tau|x_1,x_2,x_3,x_4)=F^{-1}(u_5)=F^{-1}\left(h^{-1}\left\{h^{-1}\left[h^{-1}\left(h^{-1}\left(\tau|h(h(u_4|u_1)|h(u_3|u_1))\right)|h(h(u_3|u_1)|h(u_2|u_1))\right)|h(u_2|u_1)\right]|u_1\right\}\right)$
(b) Artificial neural networks (ANNs)
Input 1
Input 2
Input n
Q(t)
Input layer
Hidden layer
Output layer
(c) Multiple linear regression (MLR)
$y_1 = a_{01} + b_{11}x_1 + \dots + b_{n1}x_n + e_1$
$y_2 = a_{02} + b_{12}x_1 + \dots + b_{n2}x_n + e_2$
...
$y_m = a_{0m} + b_{1m}x_1 + \dots + b_{nm}x_n + e_m$
Step 4 Streamflow forecasting (Validation)
MLR, ANN and CVQR
R^2, RMSE, RRMSE and
Nash-Sutcliffe Efficiency (NSE)

Figure 1. Framework of this study.

2.2. *Artificial Neural Networks (ANNs)*

An artificial neural network is an information processing system inspired by biological neural networks (such as the brain). Artificial neural networks can model the complex relationships between the input and output by simulating human learning [35]. Neural networks can be described as simple processing nodes or neurons, which generally include inputs, weights, a sum function, an activation function, and outputs and perform the corresponding numerical operations in a specific order [36]. An ANN model is usually made up of three parts: the input layer, the hidden layer, and the output layer, each of which do not have a unique number of layers. Multilayer feedforward ANNs, also known as multilayer perceptron, are commonly used in drought and water resource management and contain one input layer, one or more hidden compute node layers, and one output layer [37]. The three-layered ANNs can be expressed as follows:

$$\left\{\begin{array}{l}
\underbrace{x_j}_{the\,input\,layer\,I} \quad \Rightarrow \\
\underbrace{H_i^{in} = \sum_{j=1}^{m} w_{ij}x_j + b_{hi}}_{iutput\,ith\,node\,for\,the\,hidden\,layer\,H} \quad \Rightarrow \underbrace{H_i^{out} = \varphi\left(\sum_{j=1}^{m} w_{ij}x_j + b_{hi}\right)}_{output\,ith\,node\,for\,the\,hidden\,layer\,H} \Rightarrow \\
\underbrace{O_k^{in} = \sum_{i=1}^{p} w_{ki}\left(H_i^{out}\right) + b_{ok}}_{input\,kth\,node\,for\,the\,output\,layer\,O} \Rightarrow y_k = \underbrace{\psi\left(\sum_{i=1}^{p} w_{ki}\left(\varphi\left(\sum_{j=1}^{m} w_{ij}x_j + b_{hi}\right)\right) + b_{ok}\right)}_{output\,kth\,node\,for\,the\,output\,layer\,O}
\end{array}\right. \tag{2}$$

where w_{ij} is the weight between node i of the hidden layer and node j of the input layer; w_{ki} is the weight between the ith hidden layer node and the kth output layer node; b_{hi} and b_{ok} are the bias weights of ith node for the hidden layer and of the kth node for the output layer; and $\varphi()$ and $\psi()$ indicate the activation functions of the hidden and output layers, respectively. In this study, the multilayer feedforward ANNs with the back-propagation algorithm are used for monthly streamflow forecasting, and the number of hidden nodes is determined as five by the trial and error method. For more details, refer to Tan et al. [38].

2.3. Development of C-Vine Copula-Based Quantile Regression (CVQR) Model

In general, vine copulas are represented using a graph called R-vine, which consists of a series of trees (undirected acyclic graphs) [39]. Specially, the hierarchical structure, called a regular vine (R-vine), contains a series of connected trees $\boldsymbol{T} := (T_1, T_2, \ldots, T_d)$ along with the series of edges $E(\boldsymbol{T}) := E_1 \cup E_2 \cup \ldots \cup E_{d-1}$ and the series of nodes $N(\boldsymbol{T}) := N_1 \cup N_2 \cup \ldots \cup N_{d-1}$. However, regular vines in terms of pair-copulas are still very general and do not have unique decomposition. Thus, the canonical vine (C-vine) and the D-vine are two most common structures of regular vines [40]. C-vine has a stellar structure in their tree sequence, while D-vine has a path structure. In hydrological field in this study, the monthly streamflow is affected by various climatic and hydrological factors. Therefore, the runoff factor that has a strong dependence on all other variables is selected as the first root for C-vine construction instead of D-vines. Here, two five-dimensional examples of possible tree sequences are shown in Figure 2.

2.3.1. Copula Function

The general expression of bivariate copulas can be written as follows:

$$H(x, y) = C\left(u_x, u_y; \theta\right) \tag{3}$$

where (x, y) are correlated random variables. θ can often be derived from Kendall's τ as a preliminary estimation, and (u_x, u_y) are the marginal cumulative distribution functions of x and y, respectively. Kendall's τ is the rank correlation coefficient proposed by Kendall [41]. Let (x_1, y_1), (x_2, y_2), ..., $(x_\mathrm{n}, y_\mathrm{n})$ be a set of observations of the joint random variables X and Y, respectively, and empirical Kendall's τ can be defined as $\tau = 2(C_n - D_n)/n(n-1)$, where C_n and D_n indicate the number of concordant and discordant pairs, respectively.

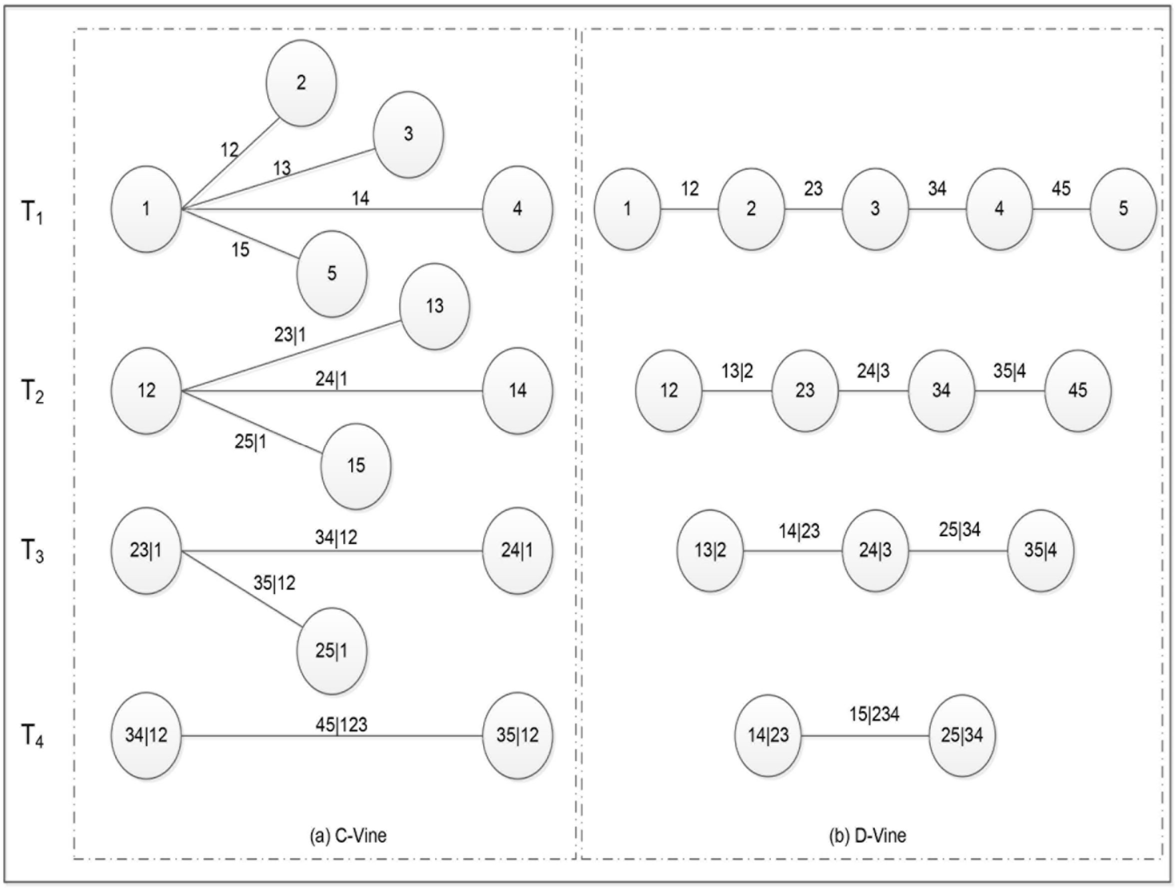

Figure 2. Examples of 5-dimensional C-vine (**a**) and D-vine (**b**).

A d-dimensional copula C: [0, 1] d → [0, 1] with uniformly distributed marginals U (0, 1) on the interval [0, 1] was introduced by Sklar [42]. According to Sklar's theorem, every joint cumulative distribution function (CDF) F on R^d with marginals $F_1(x_1), F_2(x_2), \ldots,$ $F_d(x_d)$ can be written as follows:

$$F(x_1, x_2, \ldots, x_d) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)), \ \forall \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in R^d \qquad (4)$$

Similarly, the multivariate density $f(x_1, x_2, \ldots, x_d)$ with marginal densities $f_1(x_1), f_2(x_2), \ldots, f_d(x_d)$ and join probability density of copula c (u_1, u_2, … , u_d) can be written as follows:

$$f(x_1, x_2, \ldots, x_d) = \left[\prod_{i=1}^{d} f_i(x_i)\right] c(u_1, u_2, \ldots, u_d), \ \forall \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in R^d \qquad (5)$$

and vice versa:

$$C(u_1, u_2, \ldots, u_d) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_d^{-1}(u_d)\right), \ \forall \boldsymbol{u} = (u_1, u_2, \ldots, u_d) \in (0, 1) \qquad (6)$$

where $u_i = F_i(x_i)$, $(i = 1, 2, \ldots, d)$, and $F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_d^{-1}(u_d)$ are the inverse distribution functions of the marginals.

2.3.2. Vine Copulas

For actual statistical inference, a d-dimensional copula density c can be decomposed into a product of d (d−1)/2 so-called pair-copula constructions (PCCs) based on bivariate (conditional) copulas [43]. The PCCs involve marginal conditional distributions of the form $F(x\,|\boldsymbol{\omega}\,)$. Joe [44] showed that, for every j,

$$h(x\,|\boldsymbol{\omega}\,) := F(x\,|\boldsymbol{\omega}\,) = \frac{\partial C_{x,\omega_j|\boldsymbol{\omega}_{-j}}\left(F\left(x|\boldsymbol{\omega}_{-j}\right),F\left(\omega_j|\boldsymbol{\omega}_{-j}\right)\right)}{\partial F\left(\omega_j|\boldsymbol{\omega}_{-j}\right)} \tag{7}$$

where $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_j,\ldots,\omega_n)$ is a n-dimensional vector, $\boldsymbol{\omega}_j$ is an arbitrarily selected component of the vector $\boldsymbol{\omega}$, and $\boldsymbol{\omega}_{-j}$ is a vector of $\boldsymbol{\omega}$ without the jth component; $h(x\,|\boldsymbol{\omega}\,)$ is the conditional distribution function given the k-dimensional vector $\boldsymbol{\omega}$ (i.e., h-function) [43].

Then, the C-vines with one node connected to all others is the focus of this study (as shown in a). The density of the d-dimensional C-vine can be factorized as follows [45]:

$$f(x_1,x_2,\ldots,x_d) = \prod_{k=1}^{d} f_k(x_k) \times \prod_{i=1}^{d-1}\prod_{j=1}^{d-i} c_{i,i+j|1:(i-1)}\left(F(x_i|x_1,\ldots,x_{i-1}),F\left(x_{i+j}|x_1,\ldots,x_{i-1}\right)\right) \tag{8}$$

where $c_{i,i+j|1:(i-1)}$ are the bivariate (conditional) copula densities, index j indicates the trees, while i runs over the edges in each tree.

In order to understand the decomposition of C-vine structures, only 5-dimensional C-vine structure is taken as an example to show the pair-copulas of vine structure decomposition in Figure 2a, that is, the joint density of C-vine copula can be decomposed into the following:

$$\begin{aligned} f_{12345}(x_1,x_2,x_3,x_4,x_5) &= f_1(x_1)\cdot f_2(x_2)\cdot f_3(x_3)\cdot f_4(x_4)\cdot f_5(x_5) \\ &\cdot c_{12}(F_1(x_1),F_2(x_2))\cdot c_{13}(F_1(x_1),F_3(x_3))\cdot c_{14}(F_1(x_1),F_4(x_4))\cdot c_{15}(F_1(x_1),F_5(x_5)) \\ &\cdot c_{23|1}\left(F_{2|1}(x_2|x_1),F_{3|1}(x_3|x_1)\right)\cdot c_{24|1}\left(F_{2|1}(x_2|x_1),F_{4|1}(x_4|x_1)\right)\cdot c_{25|1}\left(F_{2|1}(x_2|x_1),F_{5|1}(x_5|x_1)\right) \\ &\cdot c_{34|12}\left(F_{3|12}(x_3|x_1,x_2),F_{4|12}(x_4|x_1,x_2)\right)\cdot c_{35|12}\left(F_{3|12}(x_3|x_1,x_2),F_{5|12}(x_5|x_1,x_2)\right) \\ &\cdot c_{45|123}\left(F_{4|123}(x_4|x_1,x_2,x_3),F_{5|123}(x_5|x_1,x_2,x_3)\right) \end{aligned} \tag{9}$$

where $c_{12}(F_1(x_1), F_2(x_2))$, denoted as c_{12}, represents the density function of pair-copula with marginal distributions $F_1(x_1)$ and $F_2(x_2)$.

According to the joint density of a C-vine copula presented in Equation (9), a C-vine copula with a certain order for given data can be fitted using all of the pair-copulas (conditional bivariate copulas). Then, the conditional distribution function $C_{34\,|\,12}$ and $C_{35\,|\,12}$ from tree 3, with edges $F_{3\,|\,12}(x_3\,|\,x_1, x_2)$, $F_{4\,|\,12}(x_4\,|\,x_1, x_2)$, and $F_{5\,|\,12}(x_5\,|\,x_1, x_2)$, can be obtained using Equation (7) along with $C_{3\,|\,12}$, $C_{4\,|\,12}$, $C_{5\,|\,12}$ and C_{12}, C_{13}, C_{14}, C_{15} from the first two trees. In general, the whole inferences for the conditional distribution function of predicted variable x_5 given x_1, x_2, x_3, and x_4 can be decomposed recursively from the bivariate copulas as follows:

$$\left\{\begin{array}{l} \left.\begin{array}{l} F_{2|1}(x_2|x_1) = h_{2|1}(F_2(x_2)|F_1(x_1)) \\ F_{3|1}(x_3|x_1) = h_{3|1}(F_3(x_3)|F_1(x_1)) \\ F_{4|1}(x_4|x_1) = h_{4|1}(F_4(x_4)|F_1(x_1)) \\ F_{5|1}(x_5|x_1) = h_{5|1}(F_5(x_5)|F_1(x_1)) \end{array}\right\} \mathit{For\ Tree\ 2} \\ \left.\begin{array}{l} F_{3|12}(x_3|x_1,x_2) = h_{3|12}\left(F_{3|1}(x_3|x_1)\middle|F_{2|1}(x_2|x_1)\right) = h_{3|12}\left(h_{3|1}(F_3(x_3)|F_1(x_1))\middle|h_{2|1}(F_2(x_2)|F_1(x_1))\right) \\ F_{4|12}(x_4|x_1,x_2) = h_{4|12}\left(F_{4|1}(x_4|x_1)\middle|F_{2|1}(x_2|x_1)\right) = h_{4|12}\left(h_{4|1}(F_4(x_4)|F_1(x_1))\middle|h_{2|1}(F_2(x_2)|F_1(x_1))\right) \\ F_{5|12}(x_5|x_1,x_2) = h_{5|12}\left(F_{5|1}(x_5|x_1)\middle|F_{2|1}(x_2|x_1)\right) = h_{5|12}\left(h_{5|1}(F_5(x_5)|F_1(x_1))\middle|h_{2|1}(F_2(x_2)|F_1(x_1))\right) \end{array}\right\} \mathit{For\ Tree\ 3} \\ \ldots \\ \Rightarrow F(x_5|x_1,x_2,x_3,x_4) = h(h(T_{25,1}|T_{23,1})|h(T_{24,1}|T_{23,1})) \end{array}\right. \tag{10}$$

where $T_{ij,1} = h\left(h\left(u_j|u_1\right)|h(u_i|u_1)\right),\ 2 \le i < j \le 5.$

2.3.3. CVQR Model

Generally, taking the bivariate copula as an example, the condition distribution function of Y under the condition of X = x, i.e., $F_{Y\,|\,X}(y\,|\,x)$ can be expressed as follows:

$$F_{Y|X}(y|x) = C_1(F_X(x), F_Y(y)) = \partial C(u,v)/\partial u \tag{11}$$

where $u = F_X(x)$, $v = F_Y(y)$ are the cumulative distribution function of y and x, respectively.

For any probabilities $\tau \in (0,1)$ (e.g., τ = 0.05, 0.1, . . . , 0.95), the τth quantile function of Y given X = x from $C_1(F_X(x), F_Y(y))$ can be derived from the h-function:

$$\tau = F_{Y|X}(y|x) \equiv C_1(F_X(x), F_Y(y)) \tag{12}$$

$$Q_Y(\tau|X = x) = F_Y^{-1}\left(h^{-1}(\tau|u)\right) \tag{13}$$

where $h^{-1}()$ indicates the inverse conditional distribution function (inverse h-function) of a given parametric bivariate copula.

In this study, the main purpose of the C-vine copula-based quantile regression (CVQR) model is to predict the quantile of a response variable Y given the outcome of some predictor variables. For the five-dimensional case, according to Equations (10)–(13), the τth conditional quantile function of x$_5$, $Q_{x_5}(\tau|x_1, x_2, x_3, x_4)$, can be derived from the recursive formulation:

$$\begin{gathered} Q_{x_5}(\tau|x_1, x_2, x_3, x_4) = F^{-1}(u_5) = \\ F^{-1}\left(h^{-1}\left\{h^{-1}\left[h^{-1}\left(h^{-1}(\tau|h(h(u_4|u_1)|h(u_3|u_1)))|h(h(u_3|u_1)|h(u_2|u_1))\right)|h(u_2|u_1)\right]|u_1\right\}\right) \end{gathered} \tag{14}$$

A C-vine copula-based quantile regression (CVQR) model is developed for monthly streamflow forecasting coupling a C-vine copula model and a quantile regression method within a general optimization framework. Specially, the CVQR model is constructed by modelling the distributions of predictors (i.e., $x_1, \ldots\ x_{n-1}$) and predicted variable (x_n) with the selected n-d C-vine (structure), i.e., unconditioned and conditioned pairs (e.g., Equation (9)); then, the predicted variable of x_n is derived from the conditional distribution function (Equations (10)–(14)). In detail, the predicted variable x_5 can be obtained from the given predictor variables x_1, x_2, x_3, and x_4. Firstly, the Monte Carlo simulation is used to generate a sample of 5000 uniformly distributed random numbers spaced [0, 1] as the quantiles τ. Secondly, the 5000 implementations of x_5 can be generated using Equation (14), with one random number generated for each quantile τ. Then, the average of these realizations is considered the general prediction.

A recommended tool for statistical inference of vine copulas is statistical software R with the VineCopula package (http://CRAN.R-project.org/, accessed on 20 January 2021). In this study, the Archimedean copula family (Frank, Clayton, and Gumbel copulas [46,47]) and Normal and Student's t copulas are employed to build the C-vine structures. The optimal bivariate copula families associated with parameter estimation are selected and calculated depending on the AIC and BIC using the maximum likelihood estimation (MLE) for the first C-vine tree. Then, based on these pair-copula families and the corresponding estimated parameters, the h-function can be used to calculate and specify the pair-copula input for the second C-vine tree. The process is iterated tree by tree until the last pair-copula is evaluated. The building steps were detailed in Brechmann and Schepsmeier [48]. Meanwhile, the goodness-of-fit test includes the λ-function and Kolmogorov–Smirnov (KS) test with p-values and statistics (Sn) to check whether the selected copula is suitable for describing the observed dependencies, where the λ function is defined as follows:

$$\lambda(v,\theta) = v - K(v,\theta) \tag{15}$$

where $K(v,\theta) = P(C(u_1, u_2|\theta) \leq v)$ is the Kendall distribution function of copula C with parameter θ, and $v \in [0,1]$, and (u_1, u_2) are the marginal cumulative distribution functions of copula C. The λ-function can be obtained by the 'BiCopLambda' function in the VineCopula package. For more descriptions, please refer to Genest and Favre [49], and Genest and Rivest [50].

In general, the main procedures of the proposed CVQR model for monthly streamflow predictions can be expresses as follows:

Step 1: Fit optimal marginal distributions, denoted as $u_i = F_i(x_i)$, $(i = 1, 2, \ldots, d)$;

Step 2: Model the joint probability distributions C $(u_1,\ u_2)$, … , C $(u_1,\ u_d)$, and then, the C-vine copula is iterated tree by tree until the last pair-copula is evaluated $F(x_1, x_2, \ldots, x_d) = C(u_1, u_2, \ldots, u_d)$;

Step 3: Calculate the conditional distribution of the predictive variable (monthly streamflow) u_d, $F(x_d|x_1, x_2, \ldots, x_{d-1})$;

Step 4: Generate uniformly distributed random numbers τ, and then, predictive variable is derived from the inverse function of the conditional distribution in Step 3, that is, $x_d = F^{-1}(\tau|x_1, x_2, \ldots, x_{d-1})$.

3. Application

3.1. Study Area and Datasets

Application of the proposed approach is proven to forecast monthly streamflow in the Xiangxi River basin, which is located in the western Hubei province and is part of the Three Gorges Reservoir region with a basin area of about 3100 km^2 (between 30°57′–31°34′ N and 110°25′–111°06′ E, shown in Figure 3) in China. The Xiangxi River, originating in the Shennongjia Mountain area, is a tributary of the Yangtze River with a main stream length of 94 km [51,52]. Due to the influence of typical subtropical continental monsoon climate characteristics, the annual precipitation in this basin is between 670 and 1700 mm [53]. The annual average temperature of this region is 15.6 °C and ranges between 12 °C and 20 °C.

The amount of streamflow is affected by many factors, a large part of which involve geographical and climatic conditions. Specifically, the climatic conditions consist of a collection of meteorological variables such as the air temperature (°C) and the precipitation (mm). Previous studies have proven that precipitation has a significant effect on both short- and long-term streamflow [54,55]. Therefore, the total monthly precipitation is used as a predictor in this study. Most importantly, the initial catchment conditions are nonnegligible factors affecting the streamflow generation and confluence. Moreover, the monthly average temperature is also applied as a predictor for streamflow forecasting [56]. It is noted that observations of hydrological processes tend to vary with time [57]. The occurrence of rainfall events is closely related to the fluctuation in streamflow, especially the distribution of a rainfall event is crucial to the influence of peak discharge (i.e., flood events). In addition, considering the climatic characteristics of the watershed, the snowmelt runoff (mainly in winter) is relatively little, so the influence of snowmelt runoff is ignored. The available hydrological (streamflow, unit: m3/s) and meteorological data (temperature and precipitation) from 1962 to 2009 were obtained from the Xingshan Hydrometric Station (located at 110°45′0″ E, 31°13′0″ N, as shown in Figure 3), which was provided by the Hydrological Bureau of Xingshan County. Considering that Xingshan Hydrometric Station is the largest hydrological control station in Xiangxi watershed (the representative station of the Three Gorges Hydrological Zone between 1000–3000 km^2), the hydrometeorological data of Xingshan Station was used for the streamflow forecasting. Moreover, as a lumped hydrological model, good results have also been achieved in the process of streamflow simulation in the earlier study of Kong [51].

In this study, considering that the current streamflow at month t and the streamflow (and precipitation) of the previous month has a certain correlation, the monthly streamflow (St) and precipitation (Pt) data sets were separated into multiple lead time factors such as Pt-1 and St-1, St-2, and St-12, where St-1, St-2, and St-12 represent streamflow at 1, 2, and 12 months ahead of forecast month t, respectively [58,59]. These factors together with the monthly average temperature (Tt) are potential prediction factors (inputs) to predict the monthly streamflow St (response variable). In the out-of-sample test of this study, the data set at a specific time point was divided into a training data set (38 years) for model calibration and a test data set (10 years) for validation of the model performance. Then, the predictions were repeated five times using different training and test data sets. The specific data set division method, namely 5-fold cross-validation models, is jointly shown in Table 1 and Figure 4.

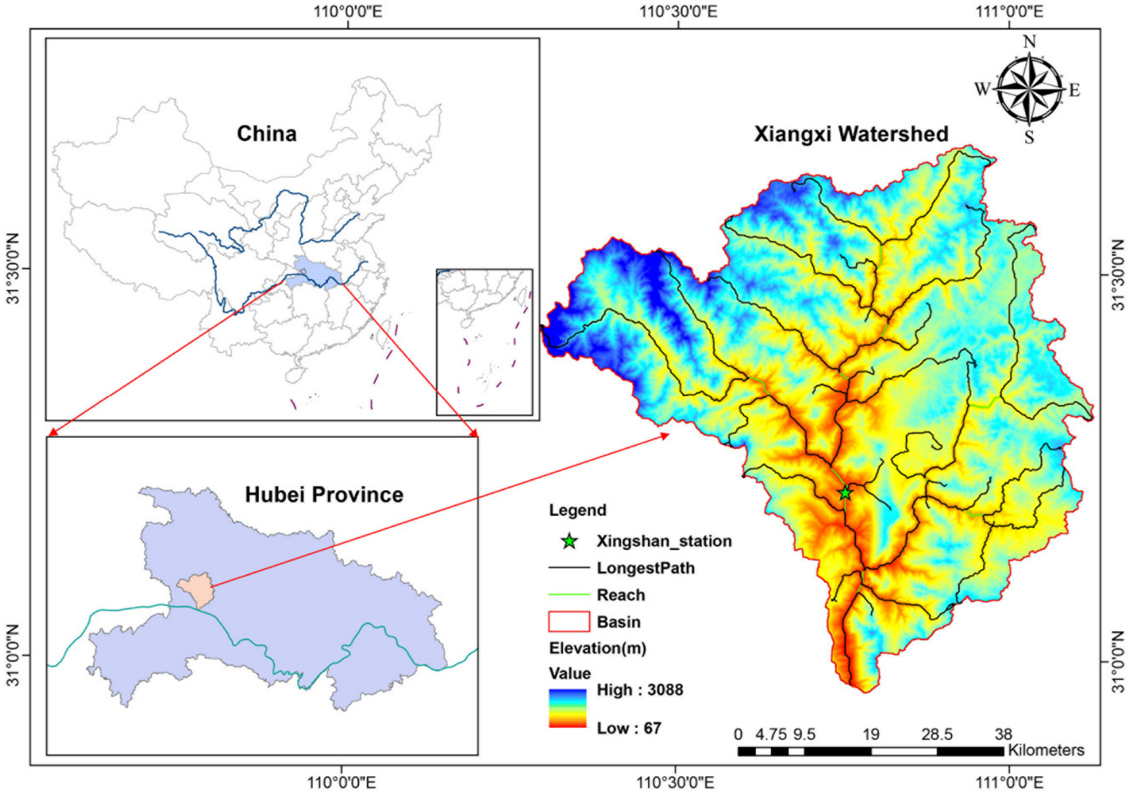

Figure 3. The study area.

Table 1. Cross-validation models with different sets of calibration and validation data.

Cross-Validation Models	Calibration Data	Validation Data
K1	1962–1971 and 1982–2009	1972–1981
K2	1962–1980 and 1991–2009	1981–1990
K3	1962–1989 and 2000–2009	1990–1999
K4	1962–1999	2000–2009
K5	1972–2009	1962–1971

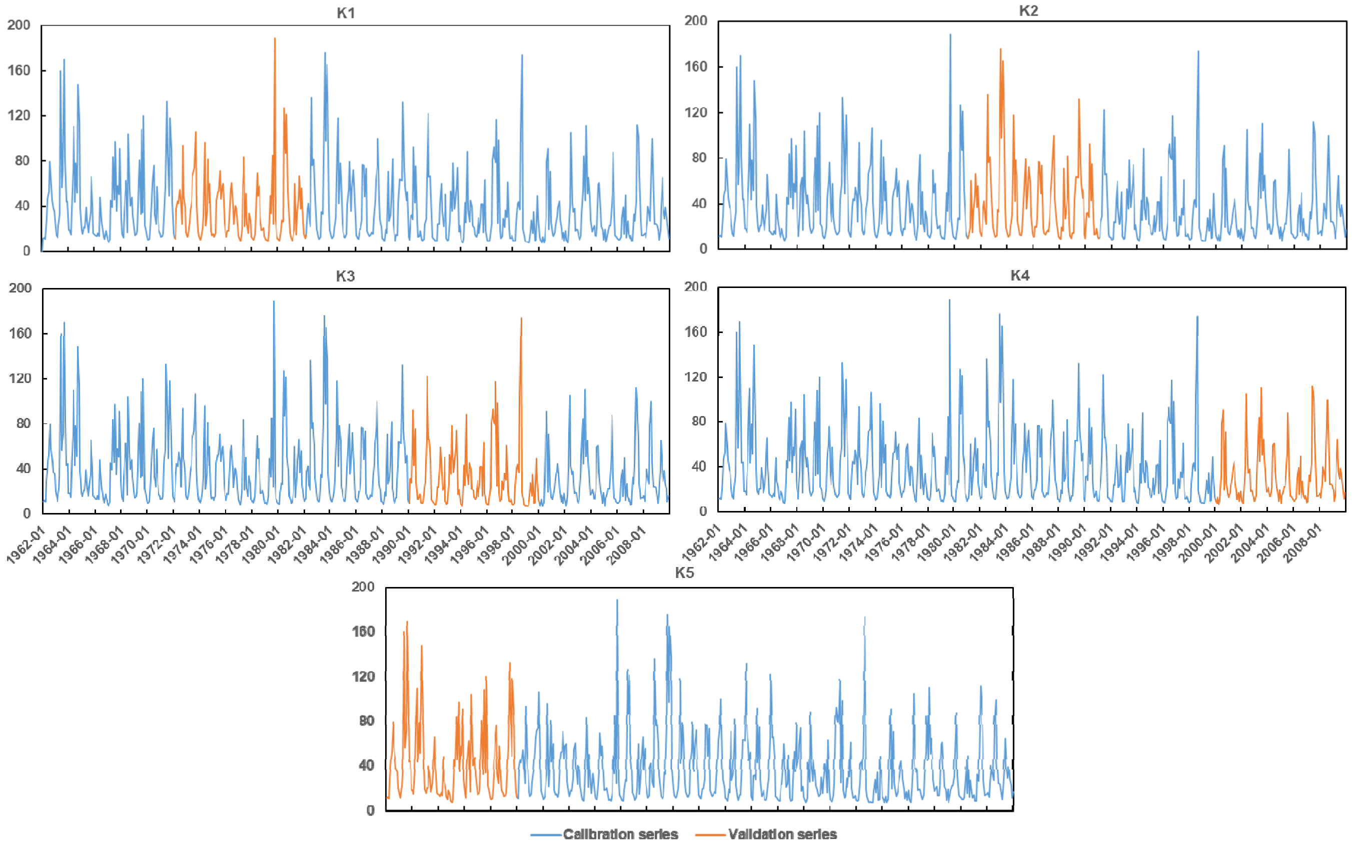

Figure 4. Selection of calibration and validation dataset with the 5-fold cross-validation method.

3.2. Evaluation Measures

In order to evaluate the performance of the developed models, in this study, four commonly used statistical evaluation methods are selected for model evaluation, including the coefficient of determination (R^2), the root mean square error (RMSE), and the Nash–Sutcliffe efficiency coefficient (NSE) and Mean Absolute Error (MAE). Then, the formulae for R^2, RMSE, NSE, and MAE can be written as follows:

$$R^2 = \frac{1}{K}\sum_{j=1}^{K}\left[\left(\frac{\sum_{i=1}^{n}\left(Q_i - Q_{avg}\right)\left(P_i - P_{avg}\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_i - Q_{avg}\right)^2}\sqrt{\sum_{i=1}^{n}\left(P_i - P_{avg}\right)^2}}\right)^2\right] \tag{16}$$

$$RMSE = \frac{1}{K}\sum_{j=1}^{K}\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i - P_i)^2}\right] \tag{17}$$

$$MAE = \frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{N}\sum_{i=1}^{N}|(P_i - Q_i)|\right] \tag{18}$$

$$NSE = \frac{1}{K}\sum_{j=1}^{K}\left[1 - \frac{\sum_{i=1}^{n}(Q_i - P_i)^2}{\sum_{i=1}^{n}(Q_i - Q_{avg})^2}\right] \tag{19}$$

where n indicates the total number of observations (or predictions), K is the number of repeated forecasting periods (K = 5), Q_i and P_i are the observed and simulated values; Q_{avg} and P_{avg} are the averages of all of the observed and simulated values, respectively.

The 90% confidential interval containing ratio (CR90) and its dispersion index (DI) are also used to evaluated the reliability and sharpness of the probabilistic predictions, respectively. CR90 is the ratio of observations covered by the 90% prediction interval. The range is between 0 and 1, and the best effect is 0.90. DI is the ratio of the average width of the 90% prediction interval to the observed value, with the lower the value, the better the prediction [60].

$$\begin{cases} CR90 = \frac{\sum\limits_{i=1}^{N} k_i}{N},\ k = \begin{cases} 1,\ s_l(i) \le o_i \le s_u(i) \\ 0,\ o_i < s_l(i)\, or\, o_i > s_u(i) \end{cases} \\ DI = \frac{1}{N}\sum\limits_{i=1}^{N} \frac{s_u(i)-s_l(i)}{o_i} \end{cases} \tag{20}$$

where k_i indicates the ith observation o_i in the 90% confidence interval with the bound $[s_l(i), s_u(i)]$ and N is the number of observations. Notably, from the perspective of flood forecasting, A high CR90 is still insufficient to illustrate a good prediction, and a high corresponding DI indicates an overestimation of uncertain boundaries.

To further illustrate the applicability of the CVQR model in streamflow forecasting, the relative estimated root mean square error (RRMSE) and relative mean absolute error (RMAE) are used to evaluate the comparison between the CVQR, ANN, and MLR models at different quantiles [61]:

$$\begin{cases} RRMSE = \frac{RMSE^{model}}{RMSE^{CVQR}} \\ RMAE = \frac{MAE^{model}}{MAE^{CVQR}} \end{cases} \tag{21}$$

in which the RMSE and MAE of the three models are acquired from Equations (17) and (18); RMAE and RRMSE stand for the relative performances of the proposed model (CVQR), for which values greater than one suggest a worse relative performance compared to the proposed model.

4. Result and Discussion

4.1. Marginal Probability Distribution Functions of C-Vine Model Variables

A two-step approach that separately evaluates the dependence function and the marginals is of great advantage in stochastic modeling of multivariate data, since many manageable distribution models are available for simulating the marginal distributions. In this study, in order to build the CVQR model, firstly, after standardization, the data are fitted with some parametric distribution functions, including the gamma, lognormal, general extreme value (GEV), and Pearson type-III (P-III) distributions, which are commonly used parameter distributions to quantify the probability distribution characteristics of hydrometeorological variables in the hydrological process [62–64]. The expressions for the gamma, GEV, lognormal, P-III, and the associated parameter values for probability functions (PDFs) are shown in Table 2. The parameters of the above distributions were obtained through the Maximum Likelihood Estimation (MLE) method.

Table 2. Parameters of optimal marginal distribution functions.

Name	Probability Density Function		Parameters						
			$\mathbf{S_{t-1}}$	$\mathbf{P_{t-1}}$	$\mathbf{S_{t-2}}$	$\mathbf{S_{t-12}}$	$\mathbf{T_t}$	$\mathbf{P_t}$	$\mathbf{S_t}$
P-III	$f(x) = \frac{\beta^\alpha (x-a_0)^{\alpha-1} e^{-\beta(x-a_0)}}{\Gamma(\alpha)}$ ***	a_0	1.88	32.12 *	1.86	2.35	Nan	32.34 *	1.83
		α	1.33	2.70	1.33	1.32	Nan	2.71	1.33
		β	0.04	0.02	0.04	0.04	Nan	0.02	0.04
Lognormal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x-\mu)^2/2\sigma^2}$	μ	3.37	3.92	3.37	3.38	2.70	3.91	3.37
		σ	0.77	1.23	0.77	0.76	0.56	1.23	0.77
GEV	$f(x) = \frac{1}{\sigma}(m)^{1+\xi}\exp(-m)$ **	ξ	0.65	0.30	0.65	0.64	-0.53	0.30	0.66
		μ	20.06	44.54	20.08	20.45	15.30	44.48	20.00
		σ	13.34	41.81	13.37	13.52	8.60	41.85	13.31
Gamma	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ ***	α	1.84	1.14	1.84	1.87	3.84	1.14	1.84
		β	0.05	0.01	0.05	0.05	0.22	0.01	0.05

Note: 32.12 *, 32.34 * indicate for −32.12 and −32.34, respectively; $** \ m = \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}; *** \ \Gamma(\alpha) = \int\limits_0^\infty u^{\alpha-1}e^{-u}du.$

The goodness-of-fit (GOF) of each distribution was computed by using RMSE and AIC values to select the most appropriate distribution for fitting each individual variable. The results of GOF are presented in Table 3. The results demonstrate that all of the proposed four distribution models can be applied for processing the distributions of the variables (i.e., St-1, Pt-1, St-2, St-12, Tt, Pt, and St), except that the P-III distribution is not suitable for the average temperature (Tt). Specially, the P-III distribution are most suitable for the streamflow data series (i.e., St-1, St-2, St-12, and St), the Gamma distribution would perform best when fitting the distributions of precipitation data (Pt-1 and Pt), and the GEV method has advantages in quantifying the distributions of the average temperature (Tt).

Table 3. Comparison of RMSE and AIC values for marginal distribution estimation.

Name	RMSE							AIC						
	S_{t-1}	P_{t-1}	S_{t-2}	S_{t-12}	T_t	P_t	S_t	S_{t-1}	P_{t-1}	S_{t-2}	S_{t-12}	T_t	P_t	S_t
P-III	**0.0340**	0.0280	**0.0340**	**0.0315**	NaN	0.028	**0.0343**	**−3076.67**	−3249.25	**−3076.47**	**−3146.55**	NaN	−3259.76	**−3068.64**
Gamma	0.0486	**0.0214**	0.0485	0.0466	0.060	**0.021**	0.0488	−2754.30	**−3494.98**	−2756.02	−2792.48	−2555.94	**−3498.64**	−2751.05
Lognormal	0.0382	0.0550	0.0385	0.0368	0.069	0.055	0.0386	−2972.55	−2632.66	−2966.17	−3007.10	−2434.36	−2636.50	−2963.20
GEV	0.0409	0.0359	0.0414	0.0415	**0.050**	0.036	0.0415	−2908.32	−3016.77	−2898.84	−2896.32	**−2719.98**	−3029.44	−2897.00

Note: The RMSE and AIC values of the optimal fitting distribution are shown in bold.

4.2. Selection and Estimation of C-Vine Copula

In this section, we introduce how to define the C-vine structures according to the learning data obtained from Section 4.1. Figure 5 shows the pair plots of the learning data set. The histograms along on the diagonal represent the marginal distributions discussed in Section 4.1. Additionally, Figure 5 (above the diagonal) indicates the values of Kendall's τ between two pairs of the variables, and the results show that the correlation between the variable St-1 and other variables is approximately stronger than that other pair variables (i.e., Kendall's τ = 0.65, 0.46, 0.33, 0.40, 0.32, and 0.46). Therefore, we define the variable St-1 as the central variate 1 (e.g., in Figure 1) in the first tree. In detail, considering that the monthly streamflow (S) is affected by various climatic and hydrological factors, such as temperature and precipitation, the monthly streamflow at last month (St-1) is selected as the first root in the first tree. Moreover, the predicted variable (St) is placed last because it is the more convenient option to evaluate the probability of St and to predict the St. The rest of the tree structures follow this principle and so forth (e.g., as shown in Figure 1). In general, the order of these variables is 1-St-1, 2-Pt-1, 3-St-2, 4-St-12, 5-Tt, 6-Pt, and 7-St. Figure 5 (below the diagonal) shows scatter plots for each pair of learning data and provides a basis for revealing the dependence structures between the variables. For example, we may find that there exists a lower tail correlation between St-1 and St-2. Obviously, the Clayton copula can be used to fit the relationship between variables St-1 and St-2.

According to the process of construction of the bivariate copula, the vine copula is constructed by a series of pair-copulas iterated tree by tree. Table 4 presents the C-vine structures consisting of 6 trees, 21 nodes, and the corresponding bivariate copulas with the parameters for every edge and KS test statistics. As mentioned above, the variables from 1 to 7 correspond to St-1, Pt-1, St-2, St-12, Tt, Pt, and St, respectively. In fact, due to the flexibility of the vines' structure, this order of the variables above is only such structure. It is the best arrangement made by considering the dependence of the variables in practical applications in this study. Meanwhile, in the process of constructing the paired copula, the vine copulas are simplified by ignoring the conditional variables.

λ-function is used to test the goodness of fit for the estimation of bivariate copula in each C-vine structure. Figure 6 illustrates the dependence of St-1 and other variables with the main node in tree 1 using λ-function. The results indicate that the selected and empirical copula are consistent with each other in all edges of tree 1. As shown in Figure 6a, the empirical λ-function (black) of the observations and the theoretical λ-function (grey) of the fitted copula coincide with each other, which means that the fitted copula is consistent with the empirical values. Combined with the KS test results in Table 4, all other selected pair-copulas obtained the optimal fitting results with $p > 0.05$ for the KS test.

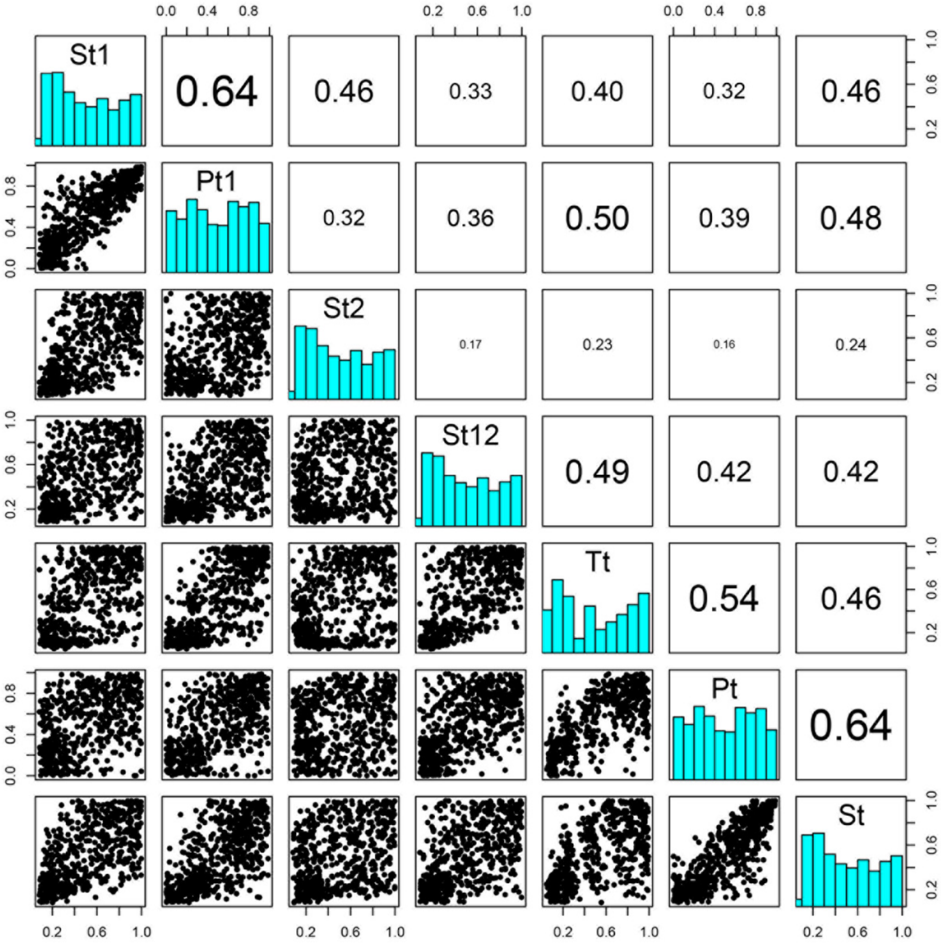

Figure 5. Pair plots of the learning data set with scatter plots below and Kendall's τ above the diagonal and histograms on the diagonal.

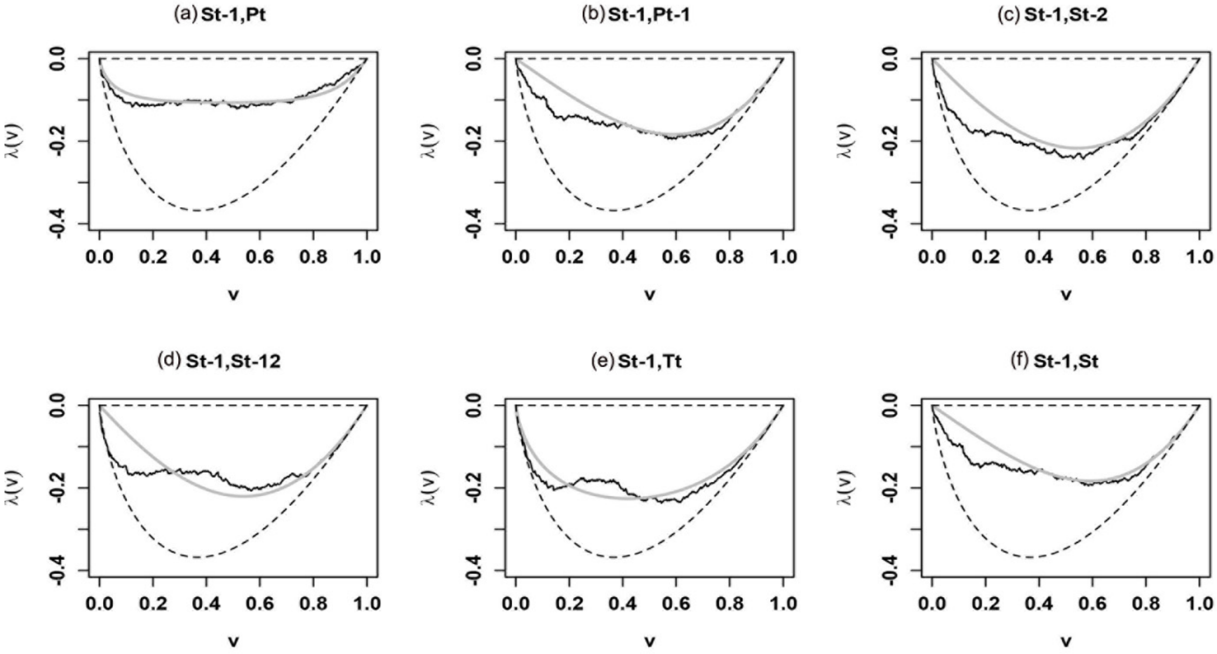

Figure 6. Correlation diagram of S_{t-1} with other variables of the λ-function with the main node in tree 1 (empirical function (black line), theoretical function of a fitted copula with parameters (grey line), as well as independence and comonotonicity limits (dashed lines)).

Table 4. Estimation of the 7-d C-vine model with bivariate copula-corresponding parameters of every node and the KS test.

Trees	C-Vine			KS Test	
	Nodes	Copulas	Parameters	p	Sn
Tree 1	12	F	9.50	0.94	0.01
	13	C	2.22	0.63	0.27
	14	C	1.52	0.68	0.19
	15	C	1.45	0.55	0.39
	16	F	3.12	0.74	0.12
	17	C	2.22	0.65	0.17
Tree 2	23 ∣ 1	N	−0.21	0.53	0.05
	24 ∣ 1	N	0.25	0.59	0.04
	25 ∣ 1	N	0.39	0.68	0.05
	26 ∣ 1	F	2.11	0.98	0.00
	27 ∣ 1	F	1.95	0.58	0.07
Tree 3	34 ∣ 12	F	−0.74	0.68	0.03
	35 ∣ 12	F	−0.51	0.54	0.00
	36 ∣ 12	F	−0.62	0.53	0.01
	37 ∣ 12	F	−0.69	0.55	0.13
Tree 4	45 ∣ 123	T	0.46, 13.95	0.98	0.07
	46 ∣ 123	T	0.41, 8.72	0.61	0.28
	47 ∣ 123	T	0.39, 5.40	0.65	0.17
Tree 5	56 ∣ 1234	F	2.81	0.75	0.11
	57 ∣ 1234	F	1.26	0.73	0.12
Tree 6	67 ∣ 12345	G	1.94	0.68	0.28

Notes: 1–7 represent St-1, Pt-1, St-2, St-12, Tt, Pt, and St, respectively; F—Frank, C—Clayton, G—Gumbel, N—Normal, and T—t-copula.

4.3. Predicted Monthly Streamflow of MLR, ANN, and C-Vine Models

Figure 7 shows a comparison of the predicted and observed streamflow acquired by the MLR, ANN, and CVQR models. For the MLR model, the results indicate that the values of R^2, NSE, and RMSE are 0.73, 0.72, and 16.16 in the calibration period and 0.73, 0.66, and 16.72 in the validation period. For the MLR model (Figure 7a), the predicted value is slightly underestimated in the case of high flow observation values (1980–1986), and vice versa, the predicted value is slightly overestimated during 2004–2009. Due to the inherent characteristics of the algorithm, the predicted values even become negative at some low-flow records (e.g., 1999 and 2000).

The ANN model performs better than the MLR model in the calibration period (Figure 7b). The ANN model obtains an R^2 of 0.75, an NSE of 0.73, and an RMSE of 15.57 in the calibration period. Similar to the results of the MLR model, the ANN model, with values of R^2 at 0.72, NSE at 0.69, and RMSE at 16.53, performs worse in the validation period than that in the calibration period. Moreover, as presented in Figure 7b, the ANN model also underestimates some streamflow during the high flow periods (e.g., 1963–1964) but overestimates more records during 2004–2009.

As presented in Figure 7c, the predicted monthly streamflow using the CVQR model could satisfy the observed values well. In the calibration period, the values of R^2, NSE, and RMSE obtained by the CVQR model are 0.73, 0.70, and 16.75, respectively. In the validation period, the values are 0.74, 0.71, and 16.13, which shows that the performance of CVQR model in the validation period is similar to that in the calibration period. The CVQR model underestimates some high flow values (e.g., during 1980–1986). Generally, compared with MLR and ANN models, the CVQR model performs best in the calibration period for monthly streamflow prediction. The CVQR model can effectively capture both linear and nonlinear dependence of these input variables (e.g., temperature, precipitation, and streamflow). Additionally, the CVQR model based on the multivariate copula functions could effectively reveal the correlation structures between predictor–response variables, which provides a potent and adaptable tool to model the dependence of such complex and jointly correlated variables.

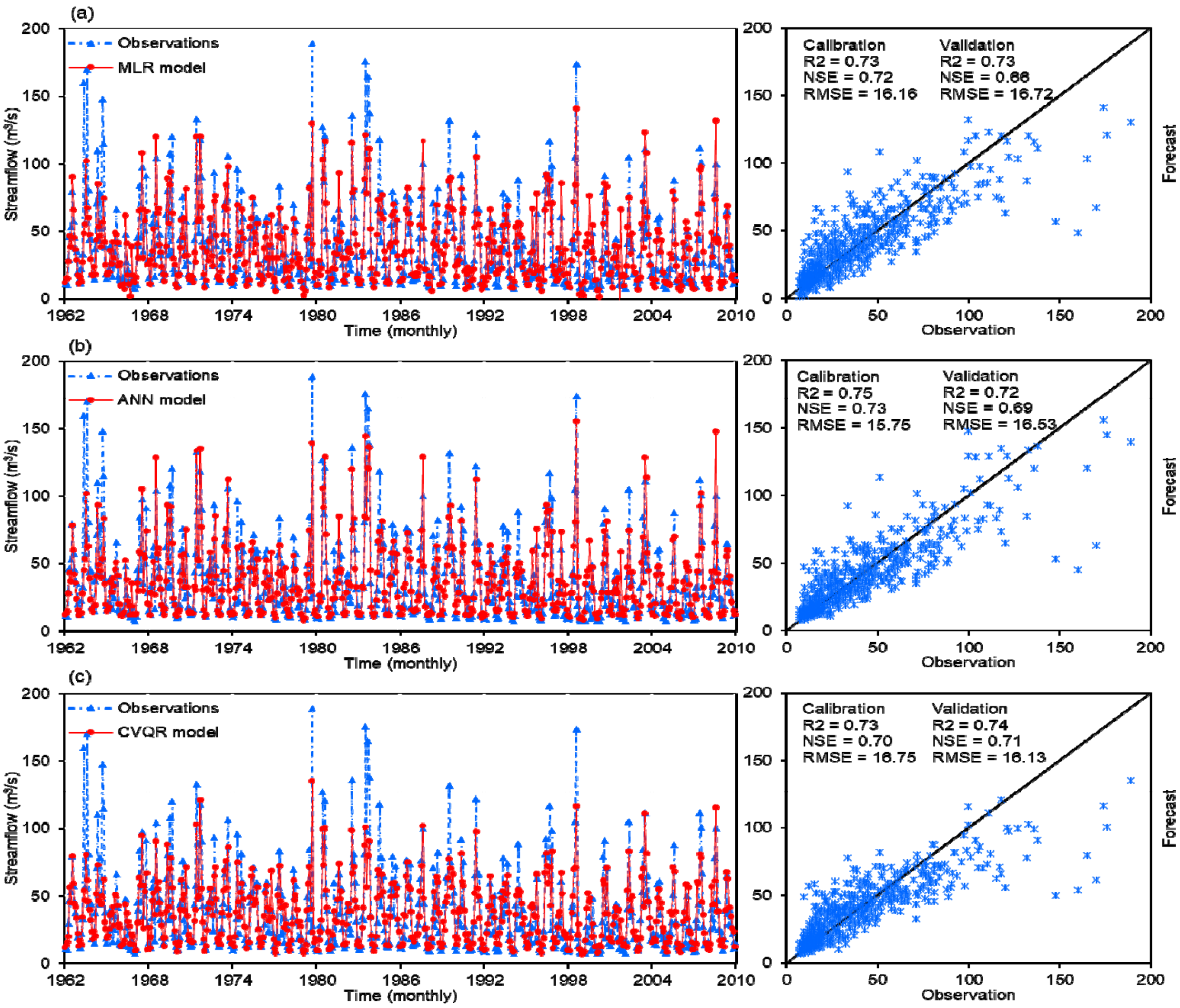

Figure 7. Comparison of predicted and observed monthly streamflow using the MLR (**a**), ANN (**b**), and CVQR (**c**) models.

Table 5 illustrates the general resulting statistics from the ANN, MLR, and CVQR models for forecasting during the calibration and validation periods. For the results of R^2, NSE, and RMSE, these results indicate that the ANN model performs best in the calibration period compared to the MLR and CVQR models while the proposed CVQR achieves the best results among the validation period compared to other models. However, the results show that ANN and CVQR performed best in terms of 90% confidence interval prediction (CR90 and DI) while MLR performed worst. The result, on the other hand, shows that MLR is not effective in quantifying nonlinear relationships among hydrological variables. In general, the results show that CVQR performs best in the calibration period for monthly streamflow prediction compared to ANN and MLR models. Moreover, the CVQR and ANN models can reflect the complex nonlinear relationships between the hydrological and meteorological factors. Therefore, in order to understand the prediction performance of CVQR in the tail correlations, the comparison of regression predictions between the CVQR and ANN models at different quantiles are explored in the next section.

Table 5. Summary statistics of streamflow forecasting during the validation period through different models.

Models	Calibration				Validation			
	R^2	NSE	RMSE	CR90/DI	R^2	NSE	RMSE	CR90/DI
MLR	0.73	0.72	16.16	0.43/0.46	0.73	0.66	16.72	0.47/0.48
ANN	0.75	0.73	15.57	0.89/1.14	0.72	0.69	16.53	0.81/1.32
CVQR	0.73	0.70	16.75	0.88/1.18	0.74	0.71	16.13	0.83/1.27

4.4. Probabilistic and Interval Predictions Obtained by the CVQR Model

As mentioned in Section 2.3, according to the C-vine copula-based quantile regression (CVQR) model, for any quantile $\tau \in (0,1)$, the τth conditional quantile function of the predicted variable can be obtained. In this section, the relationships between the streamflow (St) abnormalities and other hydrometeorological indices at different levels of quantiles τ (i.e., τ = 0.05, 0.25, 0.50, 0.75, and 0.95) are explored.

The median prediction (i.e., α = 0.5) provides a general level about the monthly streamflow, while extreme values (e.g., flood, drought) in the upper tail ($\tau \geq 0.75$) or lower tail ($\tau \leq 0.25$) indicate the worst forecast scenarios. Table 6 describes the relative performance of the ANN model with respect to the CVQR model at different quantiles. It can be seen that the proposed CVQR model outperforms the ANN model at quantiles τ = 0.75 and 0.95 and that the ANN model performs better than the CVQR model at quantiles τ = 0.25 and 0.50, which indicate that the proposed CVQR model could perform better at upper extreme events (i.e., τ = 0.75 and 0.95 quantile levels) and that the ANN model provides good results in some cases of the mean and lower quantile values.

Table 6. The performance RRMSE and RMAE of the ANN model with respect to the CVQR model at different quantiles.

τ	All		Calibration		Validation	
	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE
0.05	0.93	0.97	0.92	0.95	0.97	0.95
0.25	0.95	0.92	0.93	0.92	1.00	0.96
0.50	0.96	0.95	0.93	0.93	1.06	1.02
0.75	1.01	1.05	1.01	1.03	1.05	1.11
0.95	1.03	1.02	1.02	1.00	0.99	1.07

A scatter diagram of the simulated streamflow at different quantiles (τ = 0.05, 0.25, 0.5, 0.75, and 0.95) by the ANN and CVQR models with five-fold K cross-validations is depicted in Figure 8. The results also show that the proposed CVQR model performs a better fit in most cases, especially in the process of upper tail predictions, which are consistent with the earlier study of Kong in Xiangxi River basin [51]. While the ANN tends to overfit overestimated in the aspect of upper tail prediction. In general, the CVQR model shows a higher accuracy at upper tail levels while the ANN model provides overestimation predictions. The results indicate that the CVQR model can effectively capture upper tail dependences and has a relatively accurate assessment of the impact of upper extreme conditions (i.e., flood) in Xiangxi watershed.

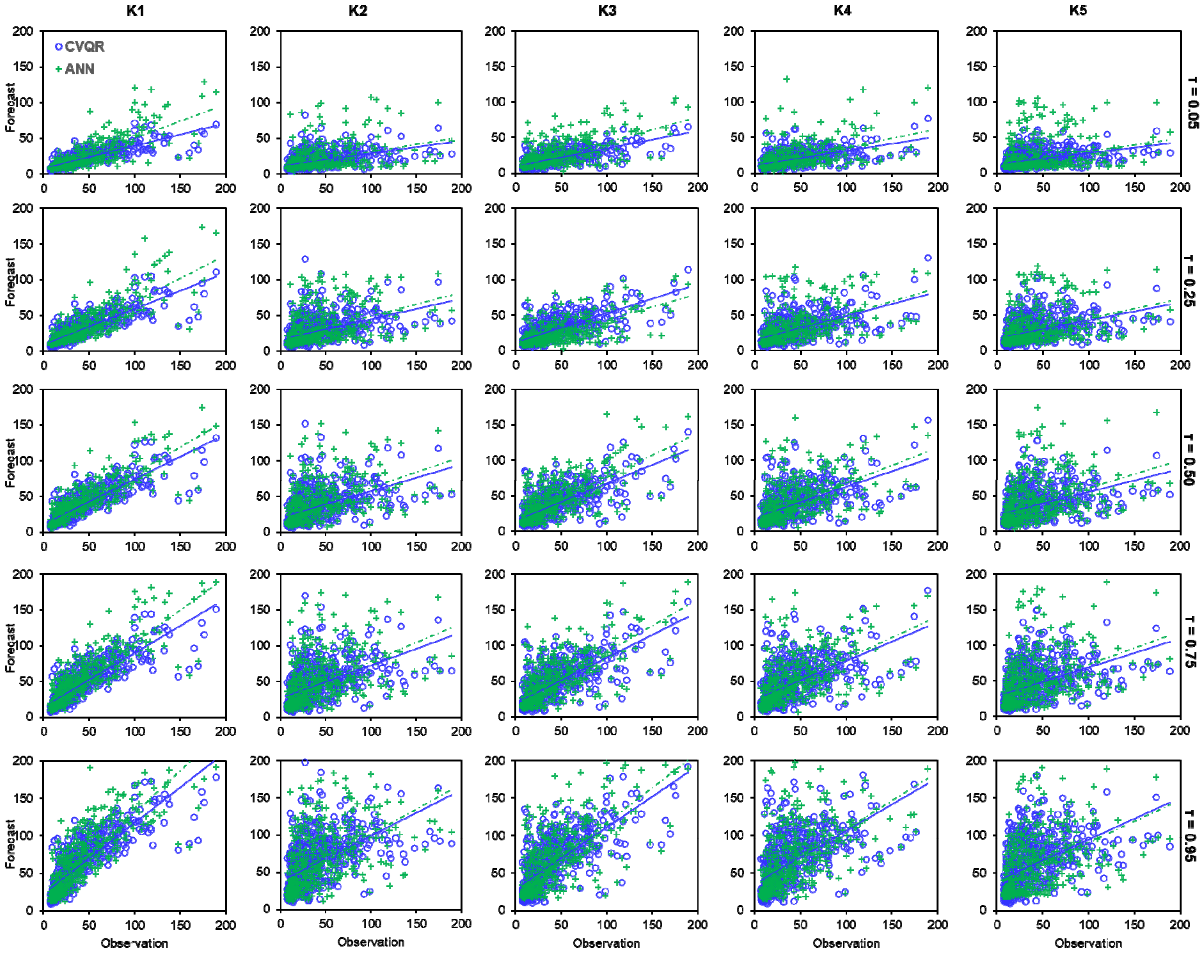

Figure 8. Scatter diagram of predicted and observed monthly streamflow using the CVQR and ANN models and their corresponding fitting lines at different quantiles (τ = 0.05, 0.25, 0.5, 0.75, and 0.95) with the five-fold K cross-validation models.

Figure 9 depicts the simulated streamflow with quantiles of 5% and 95% (90% uncertainty prediction intervals) using the ANN and CVQR models. The results indicate that the quantiles τ = 5% and 95% values of the predicted variable cover most of the observations and effectively reflect the fluctuation of the actual streamflow for the two models. Usually, hydrological forecasting in extreme cases can help policy makers make timely policy responses within the maximum risk range. The predicted 90% CI can reflect the fluctuation trend and abnormal value of the records well, whereas compared with the CVQR model, the ANN model often overestimates peaks in the prediction of flood events. Therefore, the CVQR model can effectively capture the complex nonlinear dependences among hydrological meteorological factors. This is of great significance to the practice of water resource management, for example, in rainy and dry seasons, managers can well prevent and control the occurrence of flood and make timely corresponding countermeasures.

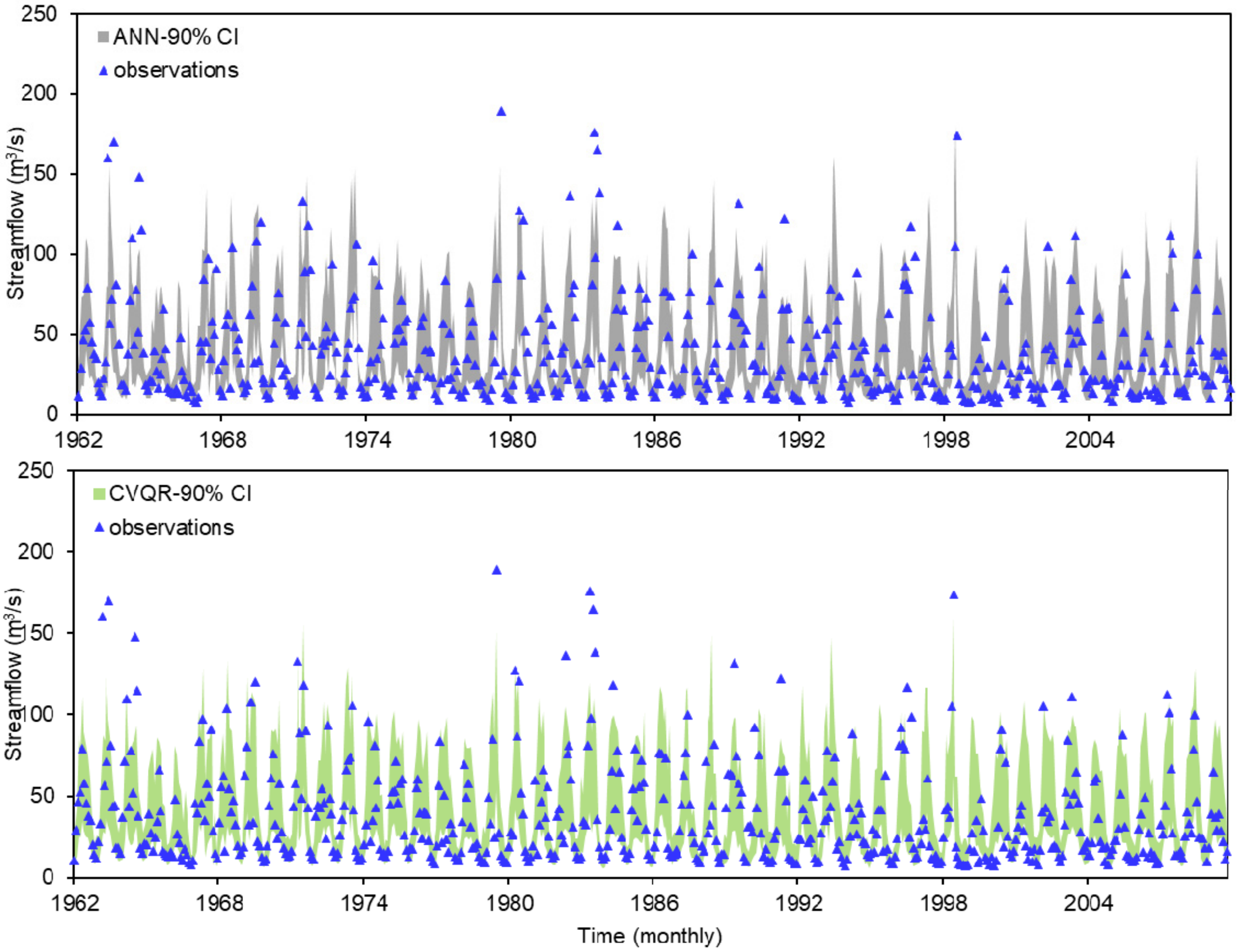

Figure 9. Comparison of the predicted and observed monthly streamflow using the CVQR and ANN models with τ = 5% and 95% (90% uncertainty prediction intervals).

5. Conclusions

In this study, a C-vine copula-based quantile regression (CVQR) model was developed to model the relationship between streamflow and other hydrometeorological variables, such as temperature and precipitation. The proposed CVQR model couples vine copulas (known as pair copula constructions) with a quantile regression method, which was applied to monthly streamflow forecasting in the Xiangxi River basin.

Specifically, the CVQR model could process multidimensional data problems while satisfying the wide range of dependence. Meanwhile, the CVQR model can effectively capture the upper correlations between independent and dependent variables (i.e., flood events). In this paper, comparisons between the proposed CVQR model and the MLR and ANN models for monthly streamflow prediction are explored. The results indicate that the performance of the CVQR model is most effective for monthly streamflow forecasting in the calibration period. The performance of the MLR model in extreme quantile (flood events) and confidence intervals is the worst and is mainly determined by the inherent characteristics of the algorithm. Compared with the MLR model, the ANN model has good advantages in this aspect of flood events and confidence intervals, but it tends to be over-fit in the process of peaks prediction. Undeniably, the CVQR model can effectively capture both the linear and nonlinear dependence of these input variables and to perform best when dealing with upper tail correlation issues (i.e., flood events) in this study.

In summary, this proposed method can effectually depict the complicated dependencies between the hydrometeorological variables. However, there still remain some flaws in the process of model building. Pair-copula is joined by marginal distributions irrespective of the conditional variables, which simplifies the construction of vine copulas [65]. The structure of PCCs is often not unique due to the flexibility of vine copulas [66]. Moreover, the proposed model can be used to explore temporal and spatial dependencies among hydrological series while spatial dependence is not considered in this study [67]. Consequently, the model will be explored further in the application process of future extensions.

Author Contributions: Conceptualization, H.L.; methodology, H.L.; software, H.L.; validation, P.G. and J.S.; formal analysis, H.L.; investigation, P.G.; resources, G.H. and Y.L.; data curation, J.S.; writing—original draft preparation, H.L.; writing—review and editing, J.S. and Y.L.; visualization, H.L.; supervision, G.H. and Y.L.; project administration, G.H. and Y.L.; funding acquisition, G.H. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant Number: XDA20060302).

Data Availability Statement: Publicly available datasets were analyzed in this study. The data presented in this study are available at http://data.cma.cn/, accessed on 20 January 2021.

Acknowledgments: The authors gratefully acknowledge all the reviewers and editors for their insightful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, Y.P.; Huang, G.H.; Nie, S.L.; Liu, L. Inexact multistage stochastic integer programming for water resources management under uncertainty. *J. Environ. Manag.* **2008**, *88*, 93–107. [CrossRef]
2. Gu, H.; Yu, Z.; Wang, G.; Wang, J.; Ju, Q.; Yang, C.; Fan, C. Impact of climate change on hydrological extremes in the Yangtze River Basin, China. *Stoch. Environ. Res. Risk Assess.* **2015**, *29*, 693–707. [CrossRef]
3. Zhu, F.L.; Zhong, P.A.; Sun, Y.; Yeh, W.-G. Real-Time Optimal Flood Control Decision Making and Risk Propagation Under Multiple Uncertainties. *Water Resour. Res.* **2017**, *53*, 10635–10654. [CrossRef]
4. Brooks, K.N.; Ffolliott, P.F.; Magner, J.A. *Hydrology and The Management of Watersheds*, 4th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2012.
5. Fu, Z.H.; Zhao, H.J.; Wang, H.; Lu, W.T.; Wang, J.; Guo, H.C. Integrated planning for regional development planning and water resources management under uncertainty: A case study of Xining, China. *J. Hydrol.* **2017**, *554*, 623–634. [CrossRef]
6. Chen, J.; Zhong, P.-A.; An, R.; Zhu, F.; Xu, B. Risk analysis for real-time flood control operation of a multi-reservoir system using a dynamic Bayesian network. *Environ. Model. Softw.* **2019**, *111*, 409–420. [CrossRef]
7. Craig, J.R.; Brown, G.; Chlumsky, R.; Jenkinson, R.W.; Jost, G.; Lee, K.; Mai, J.; Serrer, M.; Sgro, N.; Shafii, M.; et al. Flexible watershed simulation with the Raven hydrological modelling framework. *Environ. Model. Softw.* **2020**, *129*, 104728. [CrossRef]
8. Ghobadi, Y.; Pradhan, B.; Sayyad, G.A.; Kabiri, K.; Falamarzi, Y. Simulation of hydrological processes and effects of engineering projects on the Karkheh River Basin and its wetland using SWAT2009. *Quat. Int.* **2015**, *374*, 144–153. [CrossRef]
9. Zhang, D.; Lin, J.; Peng, Q.; Wang, D.; Yang, T.; Sorooshian, S.; Liu, X.; Zhuang, J. Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. *J. Hydrol.* **2018**, *565*, 720–736. [CrossRef]
10. Hrachowitz, M.; Clark, M.P. HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 3953–3973. [CrossRef]
11. Baroni, G.; Schalge, B.; Rakovec, O.; Kumar, R.; Schüler, L.; Samaniego, L.; Simmer, C.; Attinger, S. A Comprehensive Distributed Hydrological Modeling Intercomparison to Support Process Representation and Data Collection Strategies. *Water Resour. Res.* **2019**, *55*, 990–1010. [CrossRef]
12. Yifru, B.A.; Chung, I.-M.; Kim, M.-G.; Chang, S.W. Assessment of Groundwater Recharge in Agro-Urban Watersheds Using Integrated SWAT-MODFLOW Model. *Sustainability* **2020**, *12*, 6593. [CrossRef]
13. Yang, S.; Yang, D.; Chen, J.; Santisirisomboon, J.; Zhao, B.A. Physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. *J. Hydrol.* **2020**, *590*, 125206. [CrossRef]
14. Sharma, S.; Siddique, R.; Reed, S.; Ahnert, P.; Mejia, A. Hydrological model diversity enhances streamflow forecast skill at short-to medium-range timescales. *Water Resour. Res.* **2019**, *55*, 1510–1530. [CrossRef]
15. Zounemat-Kermani, M.; Mahdavi-Meymand, A.; Alizamir, M.; Adarsh, S.; Yaseen, Z.M. On the complexities of sediment load modeling using integrative machine learning: Application of the great river of Loíza in Puerto Rico. *J. Hydrol.* **2020**, *585*, 124759. [CrossRef]
16. Amaranto, A.; Munoz-Arriola, F.; Solomatine, D.P.; Corzo, G. A Spatially Enhanced Data-Driven Multimodel to Improve Semiseasonal Groundwater Forecasts in the High Plains Aquifer, USA. *Water Resour. Res.* **2019**, *55*, 5941–5961. [CrossRef]
17. Luo, X.G.; Yuan, X.H.; Zhu, S.; Xu, Z.Y.; Meng, L.S.; Peng, J. A hybrid support vector regression framework for streamflow forecast. *J. Hydrol.* **2019**, *568*, 184–193. [CrossRef]

18. Besaw, L.E.; Rizzo, D.M.; Bierman, P.R.; Hackett, W.R. Advances in ungauged streamflow prediction using artificial neural networks. *J. Hydrol.* **2010**, *386*, 27–37. [CrossRef]
19. Guo, J.; Zhou, J.; Qin, H.; Zou, Q.; Li, Q. Monthly streamflow forecasting based on improved support vector machine model. *Expert Syst. Appl.* **2011**, *38*, 13073–13081. [CrossRef]
20. Terzi, Ö.; Ergin, G. Forecasting of monthly river flow with autoregressive modeling and data-driven techniques. *Neural Comput. Appl.* **2014**, *25*, 179–188. [CrossRef]
21. Fan, Y.R.; Huang, G.H.; Li, Y.P.; Wang, X.Q.; Li, Z. Probabilistic prediction for monthly streamflow through coupling stepwise cluster analysis and quantile regression methods. *Water Resour. Manag.* **2016**, *30*, 5313–5331. [CrossRef]
22. Hassani, B.K. *Dependencies and Relationships between Variables. Scenario Analysis in Risk Management*; Springer: Berlin/Heidelberg, Germany, 2016.
23. Ayantobo, O.O.; Li, Y.; Song, S.; Javed, T.; Yao, N. Probabilistic modelling of drought events in china via 2-dimensional joint copula. *J. Hydrol.* **2018**, *559*, 373–391. [CrossRef]
24. Chen, L.; Singh, V.P.; Guo, S.; Zhou, J.; Zhang, J. Copula-based method for multisite monthly and daily streamflow simulation. *J. Hydrol.* **2015**, *528*, 369–384. [CrossRef]
25. Grimaldi, S.; Serinaldi, F. Asymmetric copula in multivariate flood frequency analysis. *Adv. Water Resour.* **2006**, *29*, 1155–1167. [CrossRef]
26. Bessa, R.J.; Miranda, V.; Botterud, A.; Zhou, Z.; Wang, J. Time-adaptive quantile-copula for wind power probabilistic forecasting. *Renew. Energy* **2012**, *40*, 29–39. [CrossRef]
27. Schepsmeier, U. Efficient information based goodness-of-fit tests for vine copula models with fixed margins: A comprehensive review. *J. Multivar. Anal.* **2015**, *138*, 34–52. [CrossRef]
28. Koenker, R.; Bassett, G. Regression quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
29. Volpi, E.; Fiori, A. Design event selection in bivariate hydrological frequency analysis. *Hydrol. Sci. J.* **2012**, *57*, 1506–1515. [CrossRef]
30. Ye, W.; Luo, K.; Liu, X. Time-varying quantile association regression model with applications to financial contagion and var. *Eur. J. Oper. Res.* **2017**, *256*, 1015–1028. [CrossRef]
31. Machado, J.A.F.; Mata, J. Counterfactual decomposition of changes in wage distributions using quantile regression. *J. Appl. Econom.* **2005**, *20*, 445–465. [CrossRef]
32. Baur, D.; Schulze, N. Coexceedances in financial markets—A quantile regression analysis of contagion. *Emerg. Mark. Rev.* **2005**, *6*, 21–43. [CrossRef]
33. Boucai, L.; Hollowell, J.G.; Surks, M.I. An approach for development of age-, gender-, and ethnicity-specific thyrotropin reference limits. *Thyroid* **2011**, *21*, 5–11. [CrossRef]
34. Yan, X.; Su, X. *Linear Regression Analysis: Theory and Computing*; World Scientific: Singapore, 2009.
35. Adamowski, J.; Chan, H.F.; Prasher, S.O.; Bogdan, O.Z.; Sliusarieva, A. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour. Res.* **2012**, *48*, 273–279. [CrossRef]
36. Sharifi, E.; Saghafian, B.; Steinacker, R. Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques. *J. Geophys. Res. Atmos.* **2019**, *124*, 789–805. [CrossRef]
37. Mouatadid, S.; Raj, N.; Deo, R.C.; Adamowski, J.F. Input selection and data-driven model performance optimization to predict the Standardized Precipitation and Evaporation Index in a drought-prone region. *Atmos. Res.* **2018**, *212*, 130–149. [CrossRef]
38. Tan, Q.-F.; Lei, X.-H.; Wang, X.; Wang, H.; Wen, X.; Ji, Y.; Kang, A.-Q. An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. *J. Hydrol.* **2018**, *567*, 767–780. [CrossRef]
39. Bedford, T.; Cooke, R.M. Vines—A new graphical model for dependent random variables. *Ann. Stat.* **2002**, *30*, 1031–1068. [CrossRef]
40. Kurowicka, D.; Cooke, R.M. Distribution-free continuous bayesian belief nets. In *Modern Statistical and Mathematical Methods in Reliability*; World Scientific: London, UK, 2005; pp. 309–322.
41. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. [CrossRef]
42. Sklar, A. Fonctions de Repartition a n Dimensions et Leurs Marges. *Publ. Inst. Stat. Univ. Paris* **1959**, *8*, 229–231.
43. Aas, K.; Czado, C.; Frigessi, A.; Bakken, H. Pair-copula constructions of multiple dependence. *Insur. Math. Econ.* **2009**, *44*, 182–198. [CrossRef]
44. Joe, H. Distributions with fixed marginals and related topics | | families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. *Lect. Notes Monogr. Ser.* **1996**, *28*, 120–141.
45. Bedford, T.; Cooke, R.M. Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines. *Ann. Math. Artif. Intell.* **2001**, *32*, 245–268. [CrossRef]
46. Serinaldi, F.; Grimaldi, S. Fully nested 3-copula: Procedure and application on hydrological data. *J. Hydrol. Eng.* **2007**, *12*, 420–430. [CrossRef]
47. Trivedi, P.K.; Zimmer, D.M. Copula Modeling: An Introduction for Practitioners. *Found. Trends Econom.* **2006**, *1*, 1–111. [CrossRef]
48. Brechmann, E.; Schepsmeier, U. Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine. *J. Stat. Softw.* **2013**, *52*, 1–27. [CrossRef]

49. Genest, C.; Favre, A.-C. Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *J. Hydrol. Eng.* **2007**, *12*, 347–368. [CrossRef]
50. Genest, C.; Rivest, L.-P. Statistical Inference Procedures for Bivariate Archimedean Copulas. *J. Am. Stat. Assoc.* **1993**, *88*, 1034–1043. [CrossRef]
51. Kong, X.M.; Huang, G.H.; Fan, Y.R.; Li, Y.P. Maximum Entropy-Gumbel-Hougaard copula method for simulation of monthly streamflow in Xiangxi river, China. *Stoch. Environ. Res. Risk Assess.* **2015**, *29*, 833–846. [CrossRef]
52. Zhang, J.L.; Li, Y.P.; Huang, G.H.; Baetz, B.W.; Liu, J. Uncertainty analysis for effluent trading planning using a bayesian estimation-based simulation-optimization modeling approach. *Water Res.* **2017**, *116*, 159–181. [CrossRef] [PubMed]
53. Xu, H.; Taylor, R.G.; Kingston, D.G.; Jiang, T.; Thompson, J.R.; Todd, M.C. Hydrological modeling of river Xiangxi using SWAT2005: A comparison of model parameterizations using station and gridded meteorological observations. *Quat. Int.* **2010**, *226*, 54–59. [CrossRef]
54. Rosenberg, E.A.; Wood, A.W.; Steinemann, A.C. Statistical applications of physically based hydrologic models to seasonal streamflow forecasts. *Water Resour. Res.* **2011**, *47*, 1995–2021. [CrossRef]
55. Robertson, D.E.; Pokhrel, P.; Wang, Q.J. Improving statistical forecasts of seasonal streamflows using hydrological model output. *Hydrol. Earth Syst. Sci.* **2013**, *17*, 579–593. [CrossRef]
56. Gómez, M.; Concepción Ausín, M.; Carmen Domínguez, M. Seasonal copula models for the analysis of glacier discharge at King George Island, Antarctica. *Stoch. Environ. Res. Risk Assess.* **2017**, *31*, 1107–1121. [CrossRef]
57. Shao, Q.; Wong, H.; Li, M.; Ip, W.C. Streamflow forecasting using functional-coefficient time series model with periodic variation. *J. Hydrol* **2009**, *368*, 88–95. [CrossRef]
58. Fan, Y.R.; Huang, G.H.; Li, Y.P.; Wang, X.Q.; Li, Z.; Jin, L. Development of PCA-based cluster quantile regression (PCA-CQR) framework for streamflow prediction: Application to the Xiangxi river watershed, China. *Appl. Soft Comput.* **2016**, *51*, 280–293. [CrossRef]
59. Liu, Z.; Zhou, P.; Chen, X.; Guan, Y. A multivariate conditional model for streamflow prediction and spatial precipitation refinement. *J. Geophys. Res. Atmos.* **2015**, *120*, 10116–10129. [CrossRef]
60. Darbandsari, P.; Coulibaly, P. Introducing entropy-based Bayesian model averaging for streamflow forecast. *J. Hydrol.* **2020**, *591*, 125577. [CrossRef]
61. Kraus, D.; Czado, C. D-vine copula based quantile regression. *Comput. Stat. Data Anal.* **2017**, *110*, 1–18. [CrossRef]
62. Adamowski, K. A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies. *J. Hydrol.* **1989**, *108*, 295–308. [CrossRef]
63. Shiau, J.T. Fitting Drought Duration and Severity with Two-Dimensional Copulas. *Water Resour. Manag.* **2006**, *20*, 795–815. [CrossRef]
64. Šraj, M.; Bezak, N.; Brilly, M. Bivariate flood frequency analysis using the copula function: A case study of the Litija station on the Sava River. *Hydrol. Process.* **2015**, *29*, 225–238. [CrossRef]
65. Acar, E.F.; Genest, C.; Neslehova, J. Beyond simplified pair-copula constructions. *J. Multivar. Anal.* **2012**, *110*, 74–90. [CrossRef]
66. Geidosch, M.; Fischer, M. Application of vine copulas to credit portfolio risk modeling. *J. Risk Financ. Manag.* **2016**, *9*, 4. [CrossRef]
67. Armando, D.; Veiga, A. Periodic copula autoregressive model designed to multivariate streamflow time series modelling. *Water Resour. Manag.* **2019**, *33*, 3417–3431.