water

MDPI

Article

A Stochastic Simulation Model for Monthly River Flow in Dry Season

Wenzhuo Wang [1], Zengchuan Dong [1,*], Feilin Zhu [1], Qing Cao [1], Juan Chen [1,2] and Xiao Yu [3]

[1] College of Water Resources and Hydrology, Hohai University, Nanjing 210098, China; wzwang25@hhu.edu.cn (W.W.); zhufeilin@hhu.edu.cn (F.Z.); qingcaohhu@163.com (Q.C.); chenjuanhhu@163.com (J.C.)
[2] School of Earth Sciences and Engineering, Hohai University, No.1 Xikang Road, Nanjing 210098, China
[3] Shanghai Hydraulic Engineering Group Co., Ltd., Shanghai 201600, China; yuxiao@shslgc.com
* Correspondence: zcdong@hhu.edu.cn; Tel.: +86-137-0518-5693

Received: 11 October 2018; Accepted: 12 November 2018; Published: 14 November 2018

Abstract: Streamflow simulation gives the major information on water systems to water resources planning and management. The monthly river flows in dry season often exhibit high autocorrelation. The headwater catchment of the Yellow River basin monthly flow series in dry season exhibits this clearly. However, existing models usually fail to capture the high-dimensional, nonlinear dependence. To address this issue, a stochastic model is developed using canonical vine copulas in combination with nonlinear correlation coefficients. Kendall's tau values of different pairs of river flows are calculated to measure the mutual correlations so as to select correlated streamflows for every month. Canonical vine copula is used to capture the temporal dependence of every month with its correlated streamflows. Finally, monthly river flow by the conditional joint distribution functions conditioned upon the corresponding river flow records was generated. The model was applied to the simulation of monthly river flows in dry season at Tangnaihai station, which controls the streamflow of headwater catchment of Yellow River basin in the north of China. The results of the proposed method possess a smaller mean absolute error (MAE) than the widely-used seasonal autoregressive integrated moving average model. The performance test on seasonal distribution further verifies the great capacity of the stochastic-statistical method.

Keywords: monthly river flow simulation; canonical vine copula; Kendall's tau value; Akaike information criteria

1. Introduction

With the global population continuing to increase, water resources are becoming ever more vital by more demand for urbanization and agricultural intensification [1,2]. In water resources planning, streamflow simulation in dry season is a paramount process in water and drought management, determination of river water flow potentials, environmental flow analysis, agricultural practices, and hydro-power generation [3,4].

Compared with models which consider relatively steady physiographic, geological, soil, land use, and plant cover attributes in a site or watershed, the statistical models are simpler and more reliable for their principle of identifying relations between output variables with their predictors without any explicit knowledge of the physical processes [5,6]. The traditional statistical model consists of the parametric and nonparametric models. The most famous parametric models are autoregressive moving average models and autoregressive integrated moving average models proposed by Box and Jenkins [7]. They are established based on the linear regression method with auto-correlation function and partial autocorrelation function. The models and their variants are widely used for their practical nature, but also show limits such as normal assumption and linearity or inaccuracy coursed by transformation [8]. Lall and Sharma [9] proposed the nonparametric model of nearest neighbor resampling method to fit the skewed marginal distribution and nonlinear structure of river flow. This kind of nonparametric model performs well in inheriting statistic features of historical record in case of the high-class dataset. However, the temporal evolution of streamflow is highly non-linear and involves uncertainty, which could be the major hindrances to synthesize accurate and reliable river flow time series for those traditional models [10,11].

Copula functions have a flexible structure with different families for different dependence structures and do not restrict the shape of posterior distributions permitting separate analysis of marginal distributions and dependence structure [12–14]. They exhibit a powerful capacity of detecting correlations as they allow for nonlinear and asymmetric cross-sectional and serial dependence [15–17]. In order to escape the gap of inaccuracy on skewed distribution, nonlinearity, and tail dependence for river flow distribution, copulas have become a popular approach to detect temporal dependence of streamflow. Madadgar and Moradkhani [18] developed an approach to integrate copula functions into a Bayesian model averaging (BMA). The model overcomes the limits of certain distribution and biased forecasts in BMA. The results of streamflow simulations for 10 river basins demonstrate that the predictive distributions are more accurate and reliable, less biased, and more confident with smaller uncertainty after Cop-BMA application. Kong, et al. [19] proposed a maximum entropy-Gumbel-Hougaard copula (MEGHC) method for monthly streamflow simulation. The marginal distributions of monthly streamflows are estimated through the maximum entropy (ME) method, and the joint distributions of two adjacent monthly streamflows are constructed using the Gumbel-Hougaard copula (GHC) method. The goodness-of-fit statistical tests of a case study of monthly streamflow simulation in Xiangxi river show that the MEGHC method can reflect dependence structure in adjacent monthly streamflows of Xiangxi river, China. Singh and Zhang [20] combined the entropy theory and the copula theory in river flow simulation. The entropy theory was extensively applied to derive the most probable univariate distribution, and bivariate copulas were applied to multivariate modeling in water engineering. This study evaluated the copula–entropy theory using a flood dataset from the experimental watershed at Walnut Gulch, Arizona. The most entropic canonical copula (MECC) successfully modeled the joint distribution of bivariate random variables.

Research reviewed above proves the powerful function of copulas for river flow simulation. However, traditional, vine-based simulation models are limited to bivariate or D-vine copulas. More effort is supposed to be paid to multivariate copulas so as to excavate the capacity of different copulas with stronger functions. Canonical vine copula detects high-dimensional structures with a more flexible structure composed of a hierarchy of conditional bivariate copulas [21]. They differ from traditional Markov trees and Bayesian belief nets in that the concept of conditional independence is weakened to allow for various forms of conditional dependence [22]. Canonical vine copula is supposed to be a more useful and skillful tool for simulation of monthly river flow in dry season which is closely related to previous months [23].

The goal in this study is to apply canonical vine copulas to the simulation of monthly river flows in dry season which highly depends on previous months. The proposed model was applied to monthly streamflow data in dry season (Nov–May) at Tangnaihai station in Yellow River basin to test the performance of the model.

2. Materials and Methods

2.1. Study Area and Data

The Yellow River is located in northern China with a length of 5464 km. It is the second longest river in China and the fifth longest river in the world. With the rapid development of agriculture and industry and population growth, water resources are under increasing pressure, particularly, in arid and semi-arid areas, including the Yellow River basin [24]. A series of cascade reservoirs has been built along the Yellow River which plays an important role in the basin water resources comprehensive utilization. Longyangxia Reservoir is the leading reservoir with carryover storage capacity located in the Upper Yellow River. The control site of input flow for the reservoir is the Tangnaihai Site which has been selected as a case study. The catchment upstream of the Tangnaihai hydrological station covers an area of 121,972 km^2 that accounts for 16% of the total area of the Yellow River Basin and yields 35% of the total runoff of the Yellow River [25]. Runoff in this region also undergoes large seasonal fluctuations consisting of a peak in July and a trough in February [26]. The rational use of the storage is vital to the operation of the whole cascade reservoirs. The locations of the site and the reservoir at the Yellow River basin are shown as Figure 1. Fifty-five years of monthly river flow data in dry season (Jan, Feb, Mar, Apr, May, Nov, and Dec) spanning over a period of 1956–2010, without missing values used, which were obtained from Yellow River Conservancy Commission of the Ministry of Water Resources.

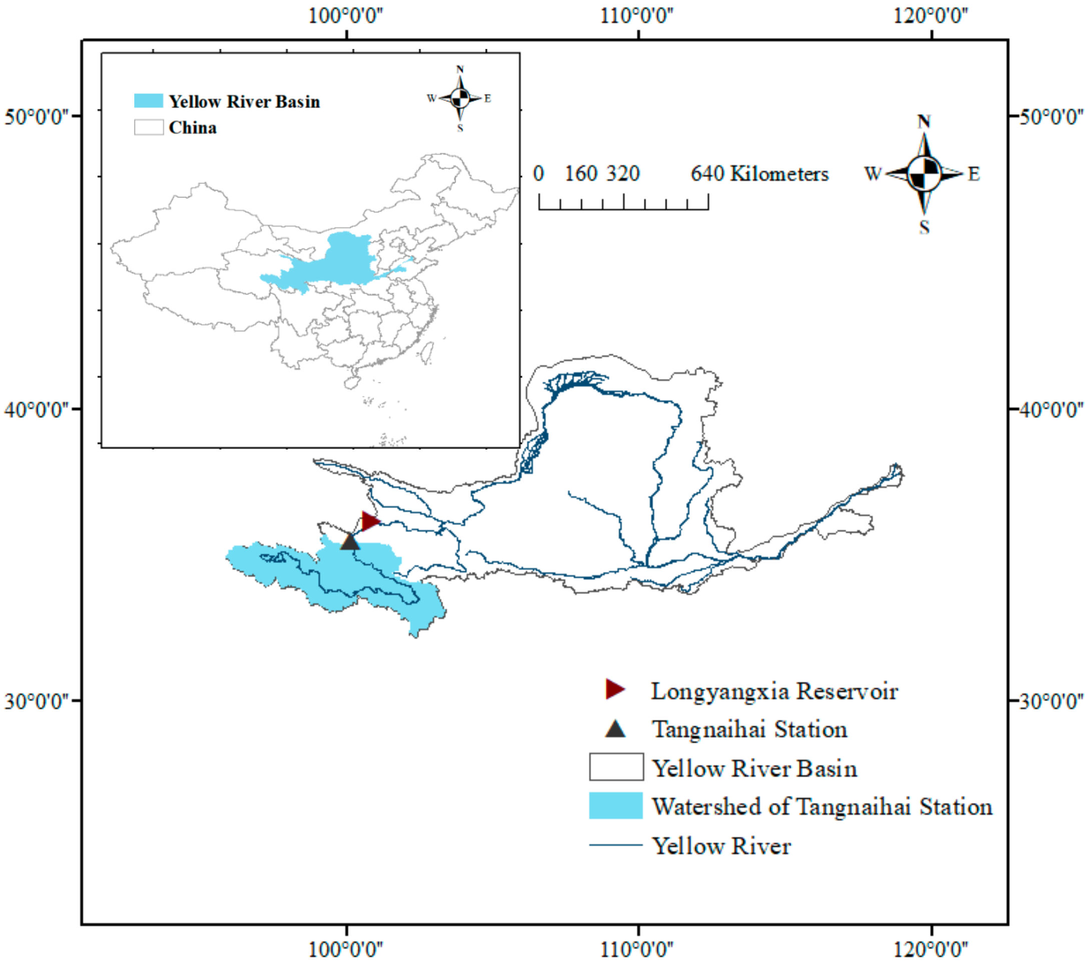

Figure 1. Map of the Yellow River and the location of Tangnaihai Site.

2.2. Methods

A general framework is developed to simulate monthly river flow in dry season. The approach establishes sub-models for different months. Every sub-model consists of selecting corelative months based on Kendall's tau values which are based upon a predetermined significance level (herein $\alpha = 1\%$). Then, the dependence structure of the current month with the selected related months in each sub-model is exploited using canonical vine copulas. Finally, the monthly river flow is generated by the conditional quantile functions of canonical vine copulas built for 7 months conditioned upon historic flows of correlated months.

2.2.1. Canonical Vine Copulas

Copulas are multivariate cumulative distributions on the unit hypercube $[0,1]^{\mathrm{d}}$ with arbitrary marginal distributions [27]. The famous theorem of Sklar [28] gives the strong function for copulas that a copula is capable of linking the joint cumulative distribution function to their marginal distribution functions. Therefore, they often used in modeling multivariate distributions.

Bivariate copulas are simple forms, including Gaussian copula, Student t copula, Clayton copula [29], Gumbel copula [30], Frank copula [31], Joe copula, BB1 copula, BB6 copula, BB7 copula, BB8 copula, and the rotated ones [27]. The general bivariate copulas can be expressed as:

$$F(x_1,x_2) = C(F_1(x_1),F_2(x_2)), \tag{1}$$

where $F(\cdot,\cdot)$ is the joint distribution and for the continuous marginal distribution F_1 for X_1 and F_2 for X_2, copula function $C(\cdot,\cdot)$ is unique. If $F(\cdot,\cdot)$ is absolutely continuous, the density of a bivariate copula, $c(\cdot,\cdot)$, is given by Joe [32] and Neslen [33]:

$$c(u_1,u_2) = \frac{\partial C(u_1,u_2)}{\partial u_1 \partial u_2}, \tag{2}$$

where $u_1 = F_1(x_1)$ and $u_2 = F_2(x_2)$. The detail inference of bivariate copulas is shown in Table 1 [27].

Table 1. Bivariate copula families.

Copula	$C(u,u^*)$	Generator $\varphi(t)$	Tail Dep. (Lower, Upper)	Parameter Range
Gaussian			0	$\theta > (-1,1)$
Student-t			$2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-\theta}{1+\theta}}\right)$	$\theta > (-1,1), \nu > 2$
Clayton Copula	$\max\left[\left(u^{-\theta}+u^{*-\theta}-1\right)^{-\frac{1}{\theta}},0\right]$	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$	$\left(2^{-\frac{1}{\theta}},0\right)$	$\theta > 0$
Gumbel Copula	$\exp\left\{-\left[(-\ln u)^{\theta}+(-\ln u^*)^{\theta}\right]^{\frac{1}{\theta}}\right\}$	$(-\ln t)^{\theta}$	$\left(0,2-2^{\frac{1}{\theta}}\right)$	$\theta \geq 1$
Frank Copula	$-\frac{1}{\theta}\ln\left[1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta u^*}-1\right)}{e^{-\theta}-1}\right]$	$-\ln\frac{e^{-\theta t}-1}{e^{-\theta}-1}$	$(0,0)$	$\theta \in R\{0\}$
Joe		$-\ln\left[1-(1-t)^{\theta}\right]$	$\left(0,2-2^{\frac{1}{\theta}}\right)$	$\theta > 1$
Clayton-Gumbel		$\left(t^{-\theta}-1\right)^{\delta}$	$\left(2^{-\frac{1}{\theta\delta}},2-2^{\frac{1}{\theta}}\right)$	$\theta > 0, \delta \geq 1$
Joe-Gumbel		$\left(-\ln\left[1-(1-t)^{\theta}\right]\right)^{\delta}$	$\left(0,2-2^{\frac{1}{\theta\delta}}\right)$	$\theta \geq 1, \delta \geq 1$
Joe-Clayton		$\left(1-(1-t)^{\theta}\right)^{-\delta}-1$	$\left(2^{-\frac{1}{\theta}},2-2^{\frac{1}{\theta}}\right)$	$\theta \geq 1, \delta > 0$
Joe-Frank		$-\ln\left[\frac{1-(1-\delta t)^{\theta}}{1-(1-\delta)^{\theta}}\right]$	$(0,0)$	$\theta \geq 1, \delta \in (0,1]$

A vine is one of the multivariate copulas with a graphical structure for dependent random variables which generalize the Markov trees [22]. Joe [32] first derived the class of m-variable distributions with given margins and $m \times (m-1)/2$ dependence parameters, one parameter corresponding to each bivariate margin. Bedford and Cooke [34] expressed these high-dimension joint distributions graphically with sequences of trees with undirected edges and nodes which are called vine trees. The nodes are actually the marginal distributions or conditional distributions of variables and the edges represent the correlations denoting the indices used for the conditional copula densities. Since trees in vine copulas can be decomposed into a series of pair copulas (building blocks), vine copulas are also called pair copulas. Canonical vine copula is a special case of the regular vine copulas which share special structures as Figure 2 shows.

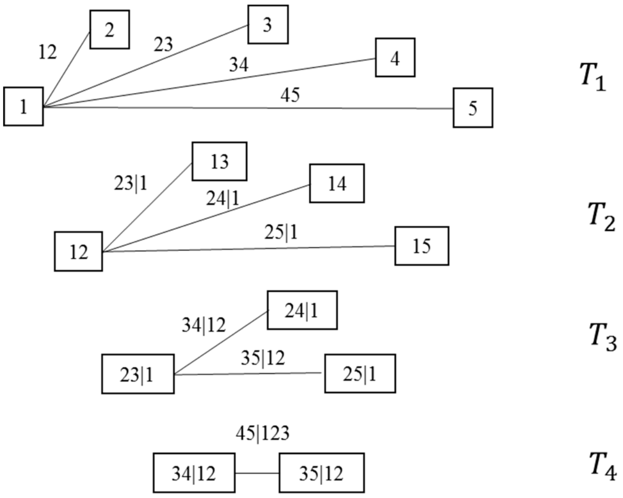

Figure 2. A canonical vine with 5 variables, 4 trees, and 10 edges. T_1 has nodes $N_1 = \{1,2,3,4,5\}$ and edges E_1. For $i = 2, \ldots, 4$ the tree T_i has nodes $N_i = E_{i-1}$ and edge set E_i. The edge represents the conditional distribution, e.g. 23 | 1 represent the bivariate conditional density copula $c(F_1, F_3|F_2)$.

Based on bivariate copulas and their conditional forms, the density distribution of an n-dimension canonical vine copula can be expressed as [22]:

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} f_k(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j,j+i|1,\ldots,j-1}(F(x_j|x_1, \ldots, x_{j-1}), F(x_{j+i}|x_1, \ldots, x_{j-1})), \tag{3}$$

where index j identifies the trees, while i runs over the edges of each tree X_t; $c_{j,j+i|1,\ldots,j-1}$ varies according to subscript; $F(\cdot|\cdot)$ represents the conditional distributions of density distribution for canonical vine given by Aas, et al. [35]:

$$F\left(x_j|x_1, \ldots, x_{j-1}\right) = \frac{\partial C_{j,j-1|1,\ldots,j-2}\left\{F\left(x_j|x_1,\ldots,x_{j-2}\right), F\left(x_{j-1}|x_1,\ldots,x_{j-2}\right)\right\}}{\partial F\left(x_{j-1}|x_1,\ldots,x_{j-2}\right)} \tag{4}$$

where $j = 2, \ldots, n$.

2.2.2. Monthly River Flow Simulation Using Canonical Vine Copulas

Streamflows in dry season exhibit high autocorrelation, which could be explained using the linear reservoir model for base flow generation. Base flow at the catchment scale could be represented as [36,37]:

$$S_t = aQ_t{}^{\mathrm{b}}, \tag{5}$$

where S_t represents the aggregate storage of the basin in month t; Q_t is the amount of outflow from groundwater in month t, which is equal to base flow of month t in dry season; a, b are watershed parameters. In dry season, mass balance can be written as:

$$S_t = S_{t-1} - Q_{t-1}, \tag{6}$$

for the linear reservoir, $b = 1$, and hence $Q_t = \frac{S_t}{a}$. This allows (6) to be rewritten as:

$$Q_t = Q_{t-1}(a-1), \tag{7}$$

more generally, the autocorrelation function would be

$$\rho_k = (a-1)^k, \tag{8}$$

since a is typically small, i.e., $S \gg Q$, the autocorrelation in the dry season is high.

Suppose X_t, denoting the variable representing the streamflow of month t, is correlated to river flows of $(d-1)$ previous successive months. The dependence structure could be built using the canonical vine copula with the nodes in an order of $X_{t-1}, \ldots, X_{t-d+1}, X_t$, which is shown in Figure 3, and the multivariate density distribution can be expressed by Equation (3).

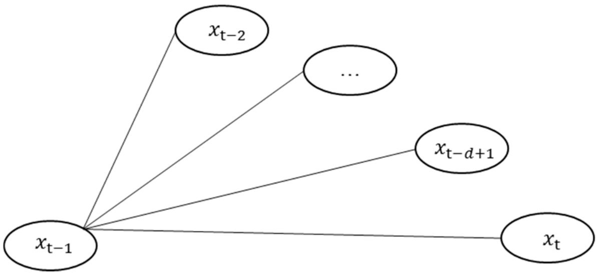

Figure 3. Tree 1 structure of month t with its correlated months.

In a case of $d = 4$, the multivariate density distribution for the month t can be expressed using canonical vine copulas as:

$$\begin{aligned} &f(x_{t-1}, x_{t-2}, x_{t-3}, x_t) = f_{t-1}(x_{t-1}) \cdot f_{t-2}(x_{t-2}) \cdot f_{t-3}(x_{t-3}) \cdot f_t(x_t) \cdot \\ &c(F(x_{t-1}), F(x_{t-2})) \cdot c(F(x_{t-2}), F(x_{t-3})) \cdot c(F(x_{t-3}), F(x_t)) \cdot \\ &c_{t-2,t-3|t-1}(F(x_{t-2}|x_{t-1}), F(x_{t-3}|x_{t-1})) \cdot c_{t-2,t|t-1}(F(x_{t-3}|x_{t-1}), F(x_t|x_{t-1})) \cdot \\ &c_{t-3,t|t-1,t-2}(F(x_{t-3}|x_{t-1}, x_{t-2}), F(x_t|x_{t-1}, x_{t-2})) \end{aligned} \tag{9}$$

where for $t = 1, 2, \ldots, 12$, the lower case letter $x_t$refers to the value taken by the corresponding variable X_t.

The general algorithm to generate river flow of month t, which is conditioned on the $(d-1)$observations, can be expressed as:

$$x_t = F^{-1}(w_t | x_{t-d+1}, \ldots, x_{t-1}), \tag{10}$$

where w_t is a uniform random number; $F^{-1}(\cdot|\cdot)$ is the inverse function of Equation (4) for canonical vine copulas.

The procedure to determine marginal distribution, choose corelative variables, specify tree structure, and estimate parameters is shown by the following steps:

Step 1—Determination of marginal distributions and corresponding parameters.

The common families of marginal distributions for river flows of different months are determined by cumulative distribution plots. The corresponding parameters are supposed to be estimated by the maximum likelihood method expressed as:

$$\hat{\Theta} = \mathrm{argmax} L(\Theta), \tag{11}$$

where Θ is the vector of parameters of marginal distributions.

Step 2—Selection of historical monthly river flows correlated to month t.

The correlated flow variables are selected from previous months which are successive using Kendall's tau method. The empirical version of Kendall's tau measuring dependence based on ranks for n observations is given by Genest and Favre (2007):

$$\tau_n = \frac{P_n - Q_n}{\begin{pmatrix} n \\ 2 \end{pmatrix}} = \frac{4}{n(n-1)} P_n - 1, \tag{12}$$

where P_n and Q_n represent the number of concordant and discordant pairs, respectively.

The null hypothesis $H_0 : C = \Pi$ is independent between X_1 and X_2, the distribution of τ_n is close to normal with zero mean and variance $\frac{2(2n+5)}{9n(n-1)}$. Therefore, H_0 would be rejected at the approximate level $\alpha = 5\%$ if

$$|Z| = \sqrt{\frac{9n(n-1)}{2(2n+5)}} |\tau_n| > 1.96, \tag{13}$$

Step 3—Specification of vine structures and estimation of parameters.

The families and parameters of copulas are estimated using the Akaike information criteria (AIC) and the maximum log-likelihood method, respectively. For streamflow simulation, the bivariate building block copulas are chosen from different bivariate copulas, such as Gaussian copula, Student t copula, Clayton copula, Gumbel copula, Frank copula, Joe copula and the rotated forms of Clayton copula, and Gumbel copula [21,38]. The commonly used fitting error functions consist of the root mean square error (RMSE), the Akaike information criteria (AIC), and Bayesian information criteria (BIC). AIC considers both the likelihood function and the number of free parameters in such a way as to maximize the probability that the candidate model has generated the observed data [39,40]. Therefore, the bivariate copula family with the lowest AIC value will be chosen as building blocks in canonical vine structures. AIC for the vine copula with $(d+1)$ variables can be calculated as [41,42]:

$$AIC = -2\log(L(\theta|data)) + 2V, \tag{14}$$

where L is the likelihood function and V is the number of free parameters. For canonical vine of the month t, the log-likelihood of the chosen bivariate copula with parameters θ given the data vectors x_1 and x_2 is:

$$\log(L(\theta|x_1, x_2)) = \sum_{k=1}^{n} \log(c(u_{1,t}, u_{2,t}, \theta)), \tag{15}$$

Step 4—Generation of monthly river flow in dry season.

3. Results and Discussion

Monthly river flow in dry season of Tangnaihai station is simulated using the proposed model as a case study. Dry season of river flow of Tangnaihai station are from November to June in which runoff is highly dependent on previous months. For the first step, the river flows in dry season were fitted into the appropriate marginal distribution. Lognormal distribution, gamma distribution, and Weibull distribution were selected as candidates since they were widely used for river flow simulation [43]. According to the empirical probability and theoretical cumulative distributions for different months in Figure 4 and results of Kolmogorov-Smirnov test shown in Table 2, lognormal distribution is better fitted for Jan, Feb, Mar, Apr, May, and Nov, and gamma distribution is better fitted for Dec of Tangnaihai site. The parameters of different months were estimated by Equation (11).

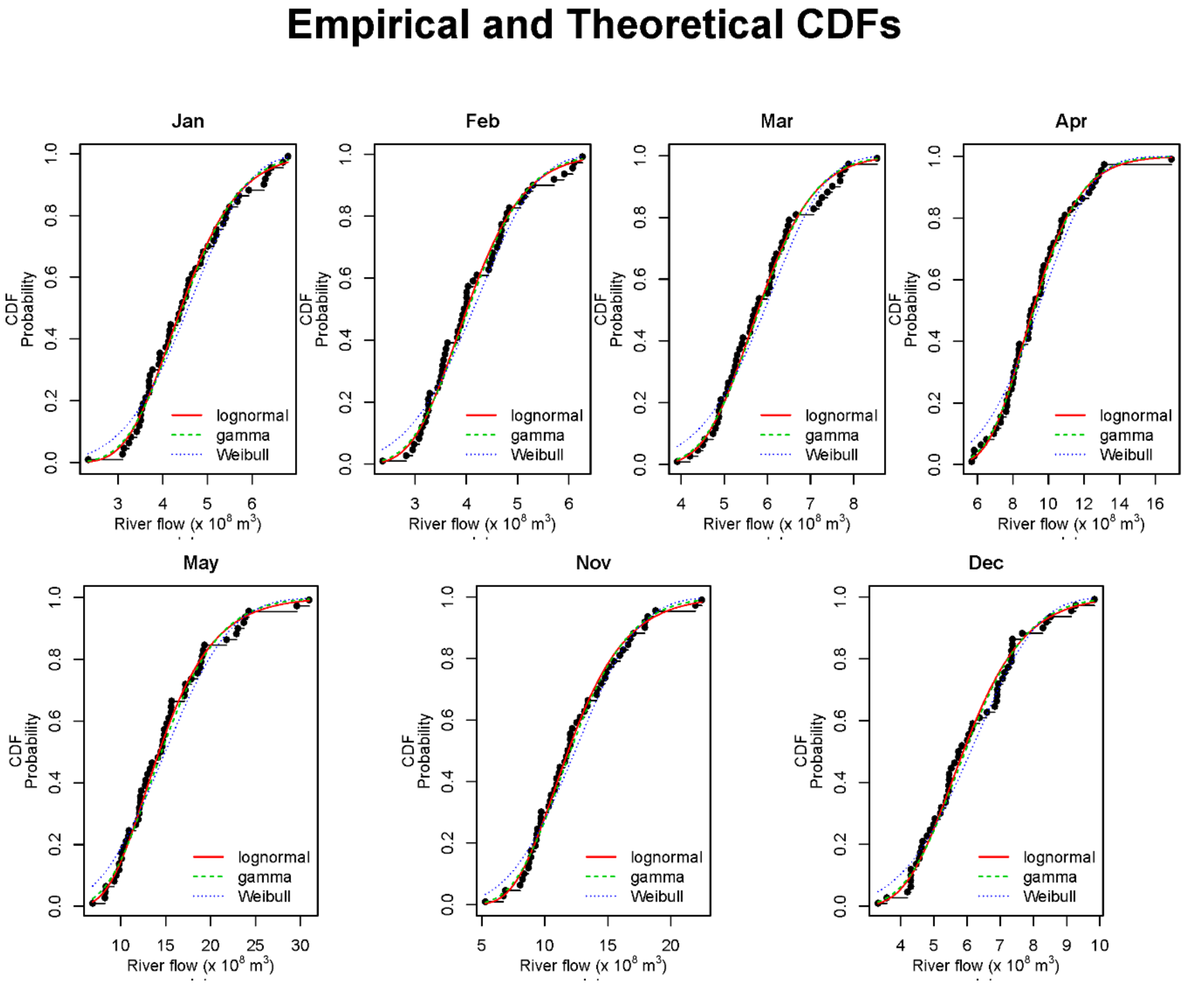

Figure 4. The plot of CDFs for gamma and lognormal distributions fitted to continuous observed data of river flow of months in dry season.

Table 2. Kolmogorov-Smirnov test for different theoretical distributions.

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Nov**	**Dec**
Gamma	0.93	0.75	0.97	0.98	0.80	0.93	0.79
Lognormal	0.95	0.84	0.99	0.99	0.98	0.99	0.68
Weibull	0.66	0.34	0.59	0.62	0.35	0.51	0.61

For the second step, the dependence between different pairs of streamflows in adjacent months was measured by Kendall’s tau value. According to Equation (13), Kendall’s tau values, which are higher than 0.24, represent correlation relationship. The lag months with bold Kendall’s tau values in Table 3 were selected as corelative months for the head month. The Kendall’s tau values for different pairs indicate that river flows in dry season (Nov–June) is highly dependent on a series of previous months. In spite of high lag 1 correlation between streamflows in wet season, we only focus on the application of high-dimensional canonical vine copulas to multivariate dependence among streamflows in dry season in this paper.

Table 3. Correlation matrix measured by Kendall's tau values.

	Lag1	Lag2	Lag3	Lag4	Lag5	Lag6	Lag7	Lag8	Lag9	Lag10	Lag11	Lag12
Jan	**0.747**	**0.722**	**0.702**	**0.586**	**0.459**	**0.449**	0.147	0.146	0.117	0.213	0.092	0.135
Feb	**0.803**	**0.746**	**0.681**	**0.665**	**0.574**	**0.447**	**0.423**	0.126	0.223	0.174	0.233	0.135
Mar	**0.681**	**0.693**	**0.656**	**0.617**	**0.594**	**0.501**	**0.393**	**0.310**	0.113	0.116	0.075	0.167
Apr	**0.340**	**0.336**	**0.371**	**0.280**	**0.333**	**0.308**	0.217	0.255	0.198	0.114	0.096	0.142
May	**0.454**	**0.284**	**0.312**	**0.311**	**0.262**	**0.287**	**0.256**	**0.241**	0.061	0.099	0.040	0.111
Jun	**0.356**	0.187	0.308	0.345	0.298	0.342	0.350	0.350	0.268	0.152	0.159	0.022
Jul	**0.372**	0.203	0.154	0.361	0.374	0.350	0.368	0.354	0.363	0.286	0.139	0.126
Aug	**0.384**	0.055	0.115	0.142	0.108	0.073	0.092	0.076	0.126	0.099	0.022	−0.052
Sep	**0.382**	0.222	0.005	0.096	0.100	0.091	0.045	0.086	0.031	0.082	0.087	0.066
Oct	**0.602**	**0.356**	**0.282**	0.023	0.024	0.045	0.103	0.044	0.082	0.008	0.066	0.050
Nov	**0.781**	**0.530**	**0.416**	**0.286**	0.097	0.073	0.094	0.150	0.059	0.092	0.017	0.096
Dec	**0.795**	**0.695**	**0.549**	**0.417**	**0.296**	0.126	0.153	0.126	0.235	0.111	0.160	0.094

For the third step, the temporal dependence of the successive months was exploited using canonical vine copulas, and the corresponding parameters were estimated. The canonical vine structure of streamflow in Jan is shown as an example in Figure 5.

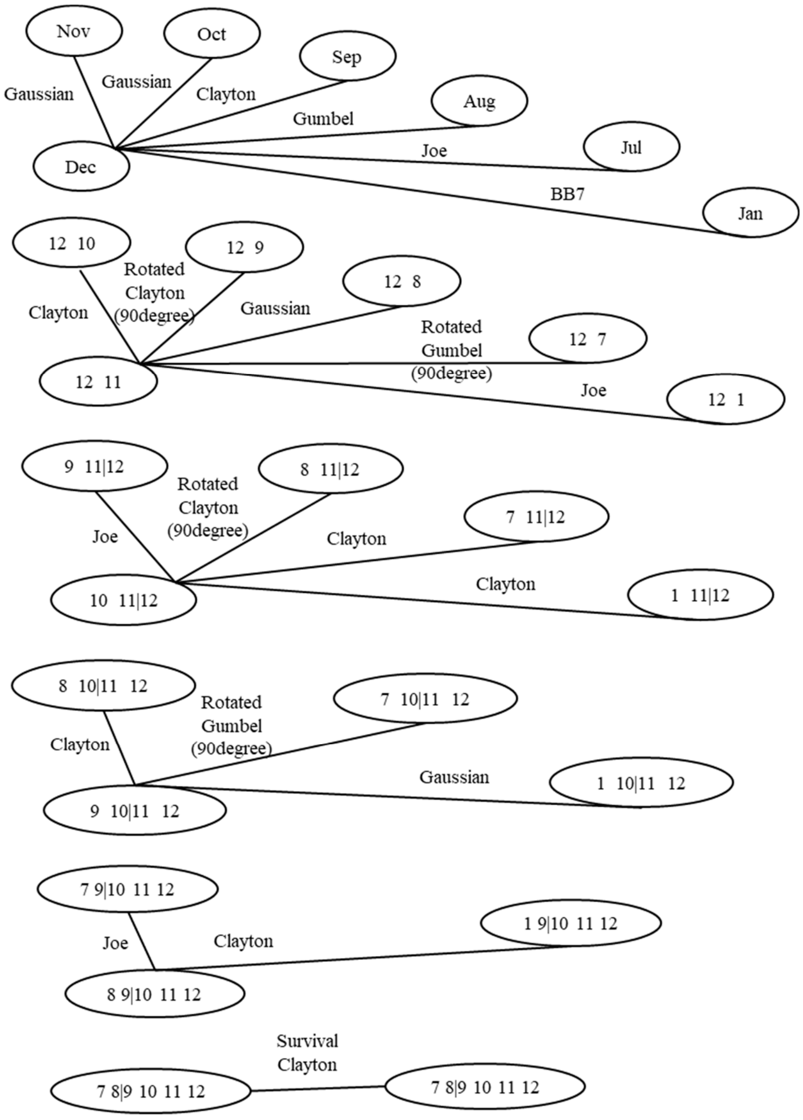

Figure 5. Canonical vine structure of streamflow in Jan.

Finally, the 7 quantile conditional distribution functions were developed for 7 monthly river flows in dry season. 500 monthly river flow simulation ensembles in the dry season were generated.

The seasonal autoregressive integrated moving average model (SARIMA) was also developed in order to compare the performance of the proposed model which can be expressed as:

$$x_t - x_{t-12} = e_t - 0.8565e_{t-12}, \tag{16}$$

where x_t represents streamflow in month t; e_t represents random errors.

Figure 6 shows the average of 500 monthly river flow simulation ensembles in dry season by canonical vine copulas compared with observed streamflows and simulated streamflows by SARIMA at Longyangxia station. The results prove the much better performance of the proposed model than SARIMA since the time series synthesized based on canonical vines is closer to observation, especially for the Jan, Feb, Mar, Apr, and Dec. The accurate descriptions of nonlinear and multivariate correlations of streamflows in different months are the major cause of this much better performance of the proposed model. Moreover, these better simulations, which occurred in months with lower streamflows, should be contributed by the greater capacity of capturing seasonality characteristics since the model develops different canonical vine structures for different months.

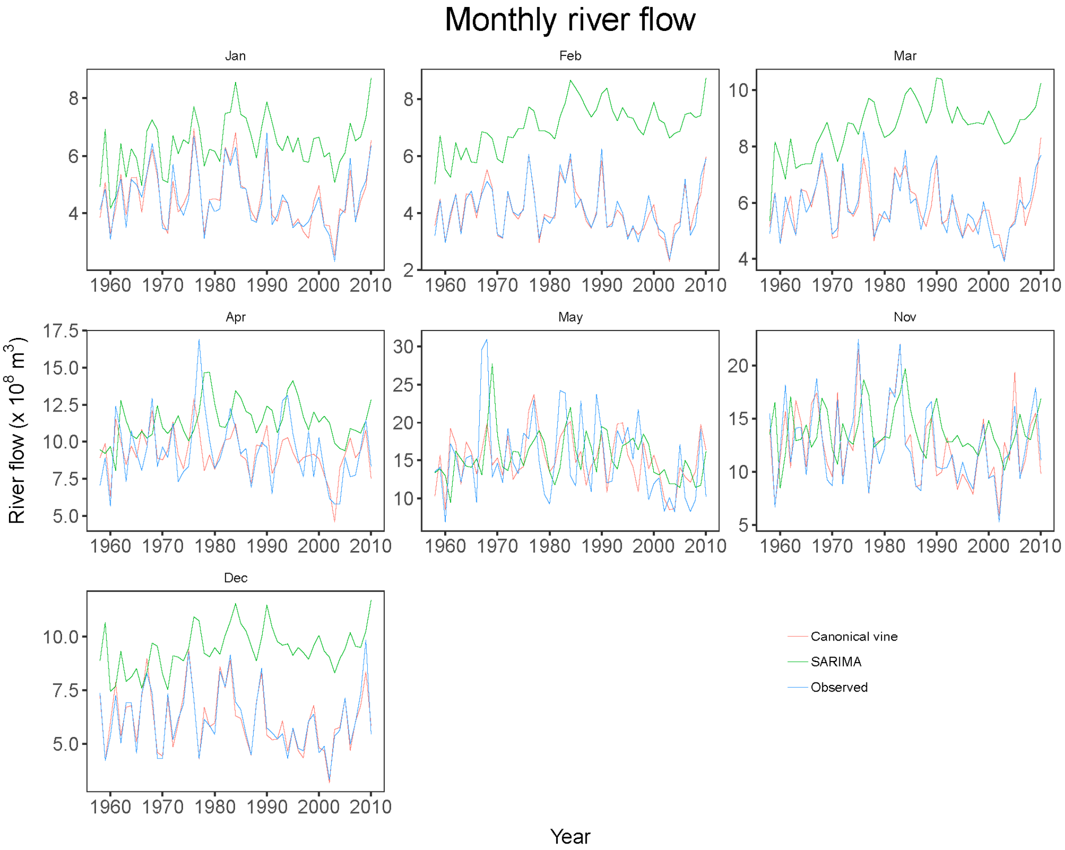

Figure 6. Average of monthly river flow simulation ensembles in dry season using the vine-based model and observed runoff, compared with SARIMA.

Furthermore, three fitting error functions were calculated to test performance. The mean squared error (MSE) and the related normalization, the Nash–Sutcliffe efficiency (NSE), are the two criteria most widely used for calibration and evaluation of hydrological models with observed data [44]. The MAE values and NSE values are reported in Table 4, where the RMSE values are also shown to give more reference. The stochastic model based on canonical vine copulas exhibits lower values of both MAE and RMSE, and higher NSE than SARIMA. The general low NSE values could be because of the misrepresenting of extreme values as the fitting errors caused by extreme values are the majority of the sum error. These errors are much smaller when calculated between mean observed streamflow and extreme values since the simulated streamflow is closer to the mean of normal observed values than mean of whole observation. Although the NSE value of the proposed model is low, the simulated streamflow is capable of keeping more consistent with normal values, which serve as the main content of the streamflow time series.

Table 4. Mean MAE, RMSE, and NSE of time series ensembles generated by different models.

	SARIMA	Canonical Vine
MAE	3.14	1.21
RMSE	3.79	2.48
NSE	−0.76	0.11

A set of distributional statistics were calculated to evaluate the capacity of capturing seasonality of the proposed model based on the canonical vine copulas. Five distributional statistics, including monthly (1) mean, (2) standard deviation, (3) skewness, (4) maximum, (5) minimum, and (6) lag 1 correlation, were chosen which were interpreted using boxplots in Figure 7. The observed runoff falls within boxplots in most dry season. There is one limitation on lag 1 correlation.

Figures 7 and 8 show that the mean and the standard deviation of the simulation are highly consistent with observation in Jan, Feb, Mar, Apr, May, and Nov. Dec exhibits higher mean and lower standard deviation; the skewness, as Figure 9 shows, in Jan, Mar, May, Nov, and Dec falls within the boxplot, while Feb and Apr are lower caused by the mispresenting of extreme values; the max and the min, shown in Figures 10 and 11, are accurate in most months; however, the performance on lag 1 correlation is poor except Mar and Apr as Figure 12 shows.

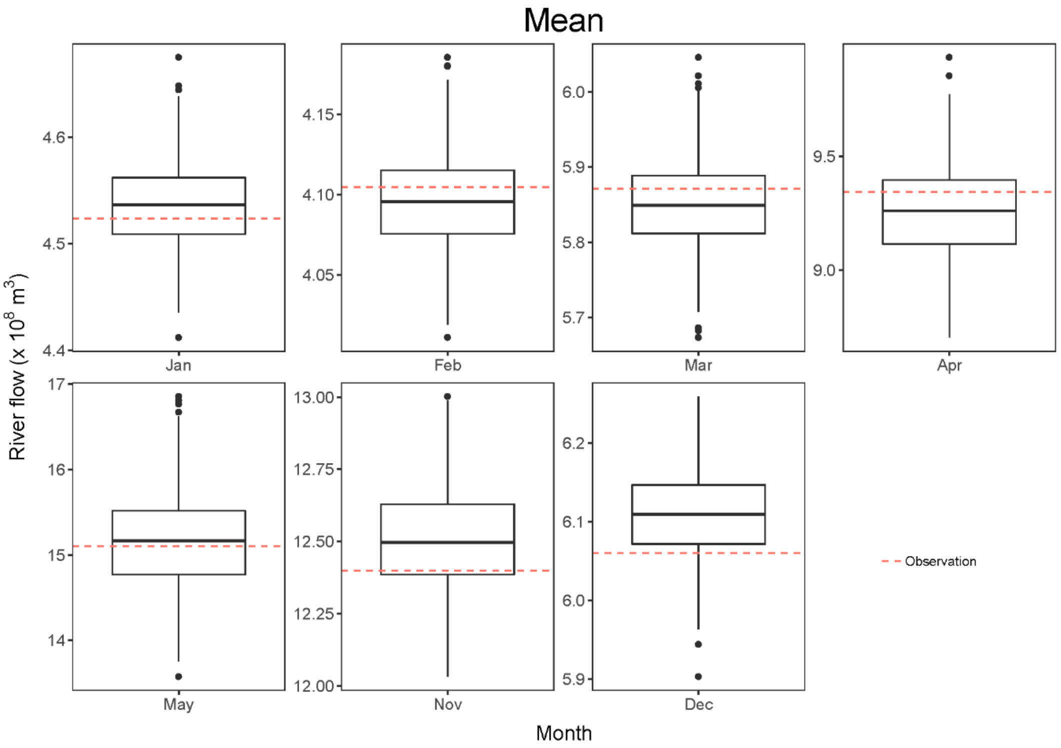

Figure 7. Box plots of mean of observed and synthesized sequences by the model based on canonical vine copulas for streamflow at Tangnaihai Site.

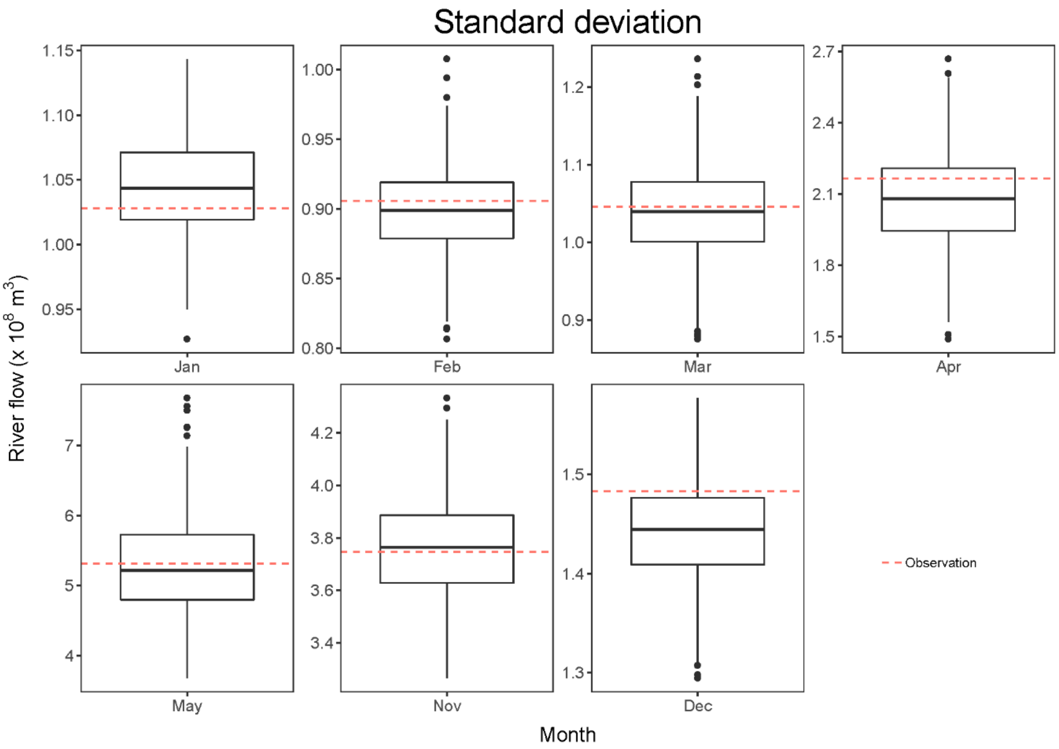

Figure 8. Box plots of standard deviation of observed and synthesized sequences by the model based on canonical vine copulas for streamflow at Tangnaihai Site.

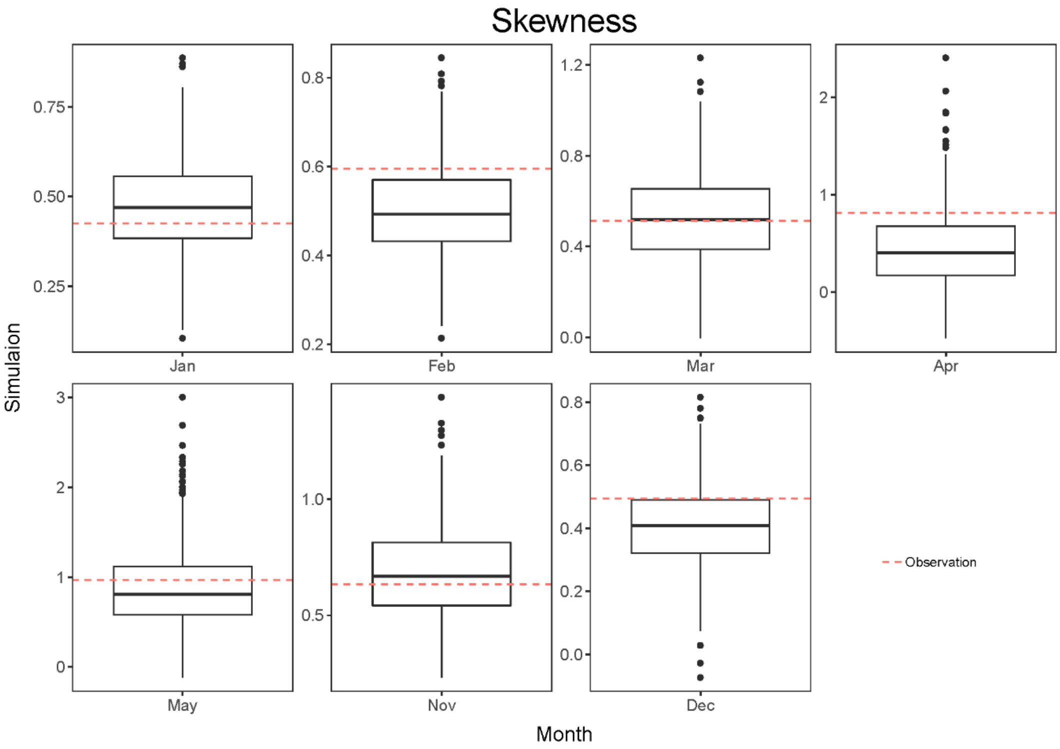

Figure 9. Box plots of skewness of observed and synthesized sequences by the model based on canonical vine copulas for streamflow at Tangnaihai Site.

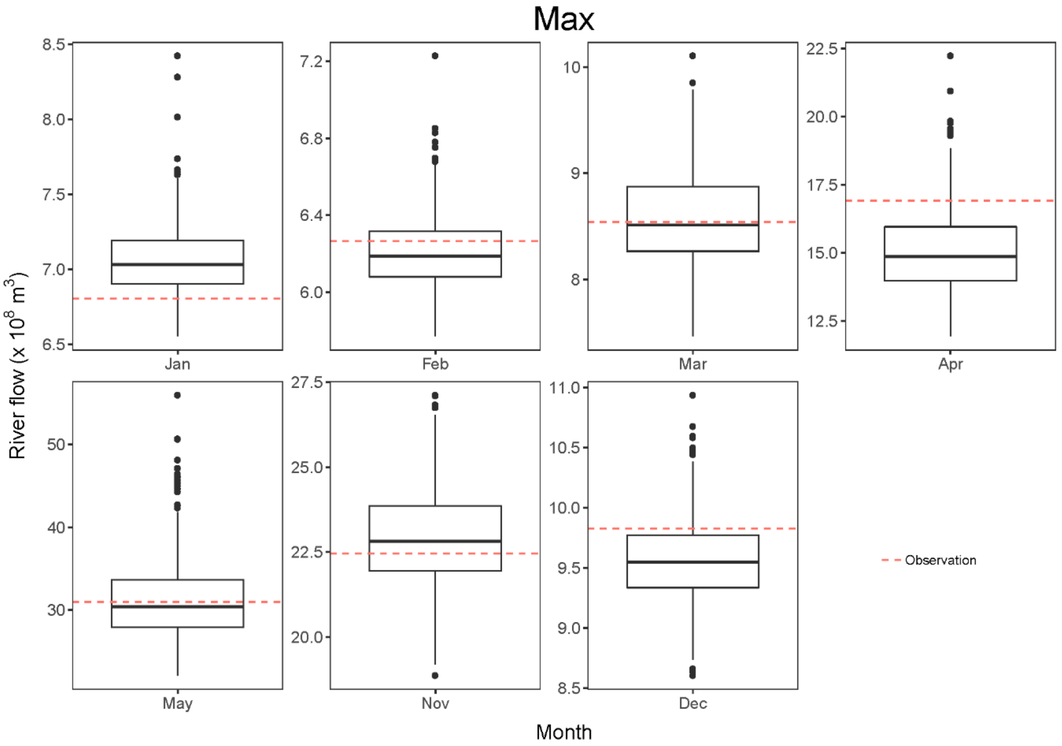

Figure 10. Box plots of max of observed and synthesized sequences by the model based on canonical vine copulas for streamflow at Tangnaihai Site.

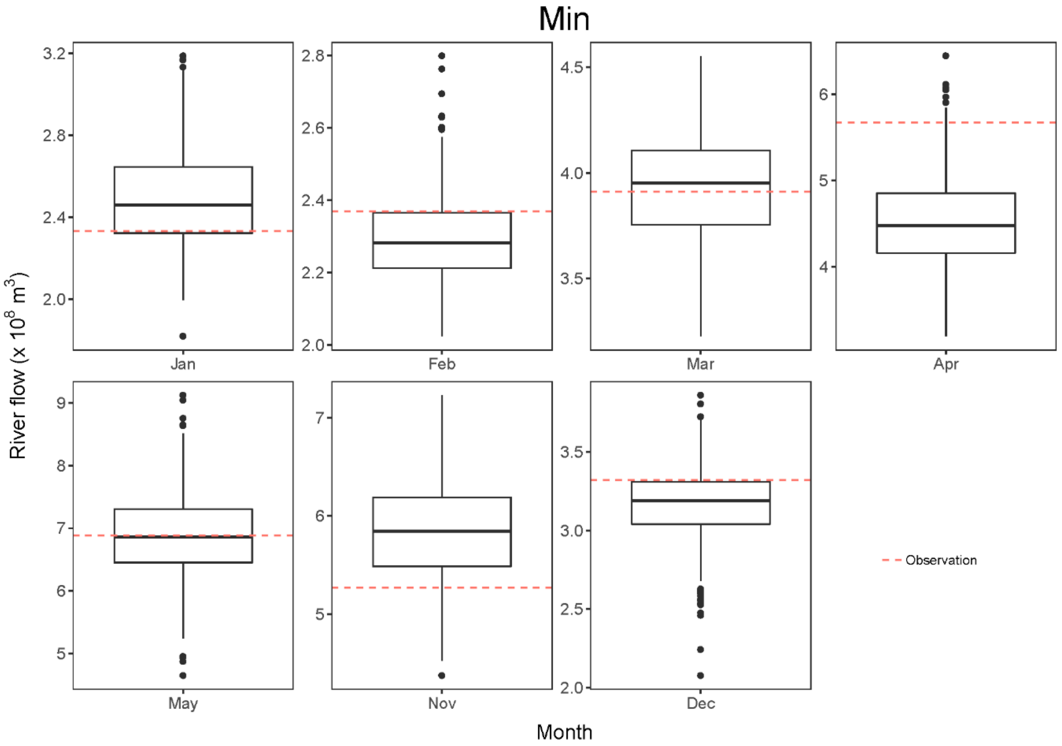

Figure 11. Box plots of min of observed and synthesized sequences by the model based on canonical vine copulas for streamflow at Tangnaihai Site.

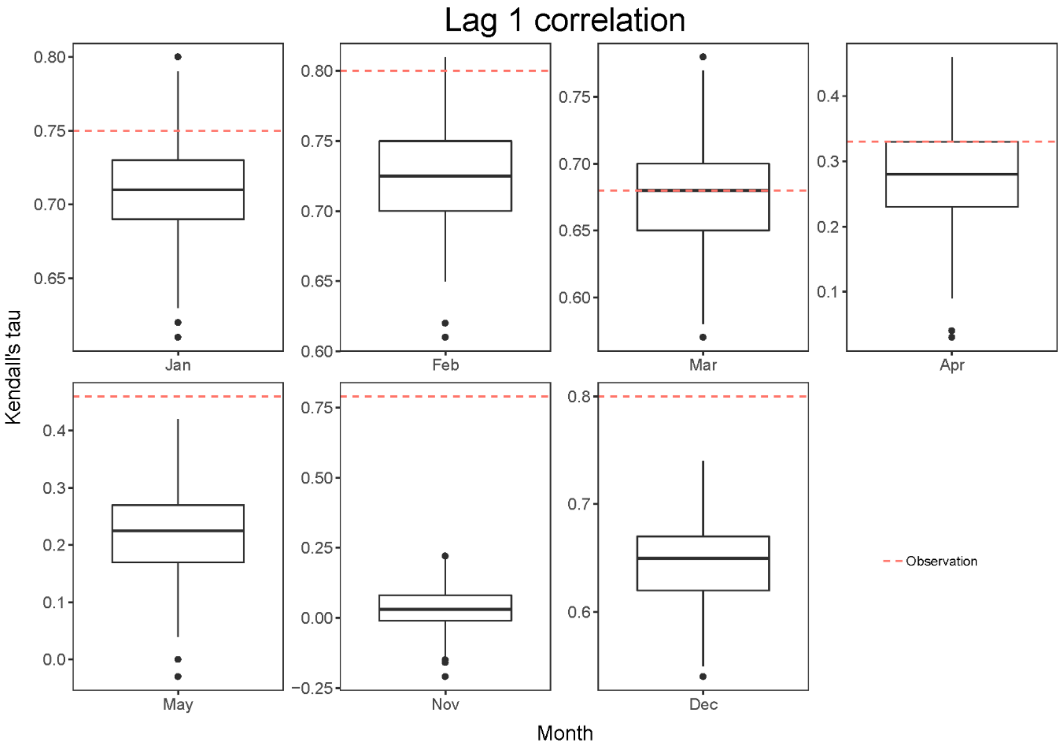

Figure 12. Box plots of lag 1 correlation of observed and synthesized sequences by the model based on canonical vine copulas for streamflow at Tangnaihai Site.

4. Conclusions

This study develops a stochastic-statistical model based on canonical vine copulas, which is capable of simulating monthly river flows in dry season. The main advantages are that the model allows for arbitrary marginal distributions for river flow and is capable of exploiting the temporal correlation in a nonlinear, high-dimensional scheme.

This paper explains the main mechanism and shows the step-by-step framework of the proposed model. The correlations of different pairs of monthly river flow are first measured by Kendall's tau values based on ranks. Then, the joint distributions are detected by canonical vine trees comprised by different bivariate copulas. Finally, the time series of monthly river flow is generated using the quantile conditional functions of canonical vine copulas conditioned upon corelative flow records. Once the model is trained, we can use this model to simulate different scenarios of river flow time series, which are capable of inheriting the temporal dependence of historical data including consistent distributional statistics and lag correlations. These streamflow time series scenarios can serve as different inputs for reservoir operation and water allocation strategies in water resources planning and management.

A case study is carried out on the simulation of monthly river flows in dry season at Tangnaihai station in the headwater catchment of the Yellow River basin. The MAE, RMSE, and NSE values were calculated to compare the performance of the proposed model with the seasonal autoregressive integrated moving average model (SARIMA). Moreover, a set of distributional statistics were calculated to evaluate the capacity of capturing seasonality. The results verified the accuracy and powerful function of the proposed model. The results also show the limitation to lag 1 correlation of the proposed model. Further studies may focus on this issue.

Author Contributions: Conceptualization, W.W. and Z.D.; methodology, W.W. and Z.D.; software, W.W.; formal analysis, F.Z., Q.C., and J.C.; investigation, F.Z., Q.C., and J.C.; resources, Z.D.; data curation, X.Y.; writing—original draft preparation, W.W.; writing—review and editing, Z.D.; visualization, W.W.; supervision, Z.D.; project administration, Z.D.; funding acquisition, W.W., Z.D., and J.C.

Funding: This research was funded by the National Key Research Program of China, grant number 2016YFC0401306; the Postgraduate Research & Practice Innovation Program of Jiangsu Province, grant number 2017B615X14 and grant number KYCX17_0420; and the Fundamental Research Funds for the Central Universities, grant number 2018B18214; and the China Postdoctoral Science Foundation, grant number 2017M621612.

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

References

1. Hutchins, M.; Abesser, C.; Prudhomme, C.; Elliott, J.; Bloomfield, J.; Mansour, M.; Hitt, O. Combined impacts of future land-use and climate stressors on water resources and quality in groundwater and surface waterbodies of the upper Thames river basin, UK. *Sci. Total Environ.* **2018**, *631*, 962–986. [CrossRef] [PubMed]
2. Zhu, F.; Zhong, P.A.; Sun, Y.; Yeh, W.W.G. Real-Time Optimal Flood Control Decision Making and Risk Propagation Under Multiple Uncertainties. *Water Resour. Res.* **2017**, *53*, 10635–10654. [CrossRef]
3. Yaseen, Z.M.; Fu, M.; Wang, C.; Mohtar, W.H.M.W.; Deo, R.C.; El-Shafie, A. Application of the Hybrid Artificial Neural Network Coupled with Rolling Mechanism and Grey Model Algorithms for Streamflow Forecasting Over Multiple Time Horizons. *Water Resour. Manag.* **2018**, *32*, 1883–1899. [CrossRef]
4. Yin, J.; Guo, S.; Liu, Z.; Yang, G.; Zhong, Y.; Liu, D. Uncertainty analysis of bivariate design flood estimation and its impacts on reservoir routing. *Water Resour. Manag.* **2018**, *32*, 1795–1809. [CrossRef]
5. Ahani, A.; Shourian, M.; Rad, P.R. Performance Assessment of the Linear, Nonlinear and Nonparametric Data Driven Models in River Flow Forecasting. *Water Resour. Manag.* **2018**, *32*, 383–399. [CrossRef]
6. Solomatine, D.P.; Maskey, M.; Shrestha, D.L. Instance-based learning compared to other data-driven methods in hydrological forecasting. *Hydrol. Process.* **2008**, *22*, 275–287. [CrossRef]
7. Box, G.E.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: San Francisco, CA, USA, 1970.
8. Song, X.; Zhang, J.; Zhan, C.; Xuan, Y.; Ye, M.; Xu, C. Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. *J. Hydrol.* **2015**, *523*, 739–757. [CrossRef]
9. Lall, U.; Sharma, A. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. *Water Resour. Res.* **1996**, *32*, 679–693. [CrossRef]
10. Mukherjee, S.; Mishra, A.; Trenberth, K.E. Climate Change and Drought: A Perspective on Drought Indices. *Curr. Clim. Chang. Rep.* **2018**, *4*, 145–163. [CrossRef]
11. Zhu, F.; Zhong, P.-A.; Sun, Y. Multi-criteria group decision making under uncertainty: Application in reservoir flood control operation. *Environ. Model. Softw.* **2018**, *100*, 236–251. [CrossRef]
12. Balistrocchi, M.; Bacchi, B. Derivation of flood frequency curves through a bivariate rainfall distribution based on copula functions: Application to an urban catchment in northern Italy's climate. *Hydrol. Res.* **2017**, *48*, 749–762. [CrossRef]
13. Yin, J.; Guo, S.; Liu, Z.; Chen, K.; Chang, F.-J.; Xiong, F. Bivariate seasonal design flood estimation based on copulas. *J. Hydrol. Eng.* **2017**, *22*, 05017028. [CrossRef]
14. Yin, J.; Guo, S.; He, S.; Guo, J.; Hong, X.; Liu, Z. A copula-based analysis of projected climate changes to bivariate flood quantiles. *J. Hydrol.* **2018**, *566*, 23–42. [CrossRef]
15. Wang, W.; Dong, Z.; Si, W.; Zhang, Y.; Xu, W. Two-Dimension Monthly River Flow Simulation Using Hierarchical Network-Copula Conditional Models. *Water Resour. Manag.* **2018**, *32*, 3801–3820. [CrossRef]
16. 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]
17. Smith, M.S. Copula modelling of dependence in multivariate time series. *Int. J. Forecast.* **2015**, *31*, 815–833. [CrossRef]
18. Madadgar, S.; Moradkhani, H. Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resour. Res.* **2014**, *50*, 9586–9603. [CrossRef]
19. Kong, X.; Huang, G.; Fan, Y.; Li, Y. 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]
20. Singh, V.P.; Zhang, L. Copula–entropy theory for multivariate stochastic modeling in water engineering. *Geosci. Lett.* **2018**, *5*, 6. [CrossRef]

21. 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*. [CrossRef]
22. 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]
23. Pereira, G.A.A.; Veiga, Á.; Erhardt, T.; Czado, C. A periodic spatial vine copula model for multi-site streamflow simulation. *Electr. Power Syst. Res.* **2017**, *152*, 9–17. [CrossRef]
24. Wang, G.; Zhang, J.; Jin, J.; Weinberg, J.; Bao, Z.; Liu, C.; Liu, Y.; Yan, X.; Song, X.; Zhai, R. Impacts of climate change on water resources in the Yellow River basin and identification of global adaptation strategies. *Mitig. Adapt. Strateg. Glob. Chang.* **2017**, *22*, 67–83. [CrossRef]
25. Wang, T.; Yang, D.; Qin, Y.; Wang, Y.; Chen, Y.; Gao, B.; Yang, H. Historical and future changes of frozen ground in the upper Yellow River Basin. *Glob. Planet. Chang.* **2018**, *162*, 199–211. [CrossRef]
26. Hu, Y.; Maskey, S.; Uhlenbrook, S.; Zhao, H. Streamflow trends and climate linkages in the source region of the Yellow River, China. *Hydrol. Process.* **2011**, *25*, 3399–3411. [CrossRef]
27. Joe, H. *Multivariate Models and Multivariate Dependence Concepts*; CRC Press: New York, NY, USA, 1997.
28. Sklar, M. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **1959**, *8*, 229–231.
29. Clayton, D.G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **1978**, *65*, 141–151. [CrossRef]
30. Gumbel, E.J. Distributions des valeurs extrêmes en plusieurs dimensions. *Publ. Inst. Stat. Univ. Paris* **1960**, *9*, 171–173.
31. Frank, M.J. On the simultaneous associativity of F (x, y) andx+y − F (x, y). *Aequ. Math.* **1979**, *19*, 194–226. [CrossRef]
32. Joe, H. Families of m-Variate Distributions with Given Margins and m(m−1)/2 Bivariate Dependence Parameters. *Lect. Notes-Monogr. Ser.* **1996**, *28*, 120–141.
33. Neslen, R. *An Introduction to Copulas*; Springer: New York, NY, USA, 2006.
34. Bedford, T.; Cooke, R.M. Vines: A New Graphical Model for Dependent Random Variables. *Ann. Stat.* **2002**, *30*, 1031–1068. [CrossRef]
35. Aas, K.; Czado, C.; Frigessi, A.; Bakken, H. Pair-copula constructions of multiple dependence. *Insur. Math. Econ.* **2009**, *44*, 182–198. [CrossRef]
36. Gao, S.; Liu, P.; Pan, Z.; Ming, B.; Guo, S.; Xiong, L. Derivation of low flow frequency distributions under human activities and its implications. *J. Hydrol.* **2017**, *549*, 294–300. [CrossRef]
37. Wang, D.; Cai, X. Detecting human interferences to low flows through base flow recession analysis. *Water Resour. Res.* **2009**, *45*. [CrossRef]
38. Pereira, G.; Veiga, Á. PAR(p)-vine copula based model for stochastic streamflow scenario generation. *Stoch. Environ. Res. Risk Assess.* **2017**, *32*, 833–842. [CrossRef]
39. Wagenmakers, E.J.; Farrell, S. AIC model selection using Akaike weights. *Psychon. Bull. Rev.* **2004**, *11*, 192–196. [CrossRef] [PubMed]
40. Posada, D.; Buckley, T.R. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* **2004**, *53*, 793–808. [CrossRef] [PubMed]
41. Akaike, H. Information theory and an extension of the maximum likelihood principle. *Int. Symp. Inf. Theory* **1973**, 267–281.
42. Akaike, H. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* **1979**, *66*, 237–242. [CrossRef]
43. Li, C.; Singh, V.P.; Mishra, A.K. Monthly river flow simulation with a joint conditional density estimation network. *Water Resour. Res.* **2013**, *49*, 3229–3242. [CrossRef]
44. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [CrossRef]