Hydrologic Risk Analysis in the Yangtze River basin through Coupling Gaussian Mixtures into Copulas

Y.R. Fan[a], W.W. Huang[b], G.H. Huang[c,d*], K. Huang[a], Y.P. Li[d]

[a] Faculty of Engineering and Applied Science, University of Regina, Regina, Saskatchewan, Canada S4S 0A2b

[b] Department of Civil Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada

[c] Institute for Energy, Environment and Sustainability Research, UR-NCEPU, University of Regina, Regina, Saskatchewan, Canada S4S 0A2; Tel: +13065854095; Fax: +13065854855; E-mail: huang@iseis.org

[d] MOE Key Laboratory of Regional Energy and Environmental Systems Optimization, North China Electric Power University, Beijing 102206, China

*Correspondence: Dr. G. H. Huang
Institute for Energy, Environment and Sustainability Research, UR-NCEPU,
University of Regina, Regina, Saskatchewan, Canada S4S 0A2,
Tel: +1061773889;
Fax: +1061773885;
E-mail: huang@iseis.org

Abstract:

In this study, a bivariate hydrologic risk framework is proposed through coupling Gaussian mixtures into copulas, leading to a coupled GMM-copula method. In the coupled GMM-Copula method, the marginal distributions of flood peak, volume and duration are quantified through Gaussian mixture models and the joint probability distributions of flood peak-volume, peak-duration and volume duration are established through copulas. The bivariate hydrologic risk is then derived based on the joint return period of flood variable pairs. The proposed method is applied to the risk analysis for the Yichang station on the main stream of the Yangtze River, China. The results indicate that (i) the bivariate risk for flood peak-volume would keep constant for the flood volume less than 1.0×10^5 m^3/s day, but present a significant decreasing trend for the flood volume larger than 1.7×10^5 m^3/s day; (ii) the bivariate risk for flood peak-duration would not change significantly for the flood duration less than 8 days, and then decrease significantly as duration value become larger. The probability density functions (pdfs) of the flood volume and duration conditional on flood peak can also be generated through the fitted copulas. The results indicate that the conditional pdfs of flood volume and duration follow bimodal distributions, with the occurrence frequency of the first vertex decreasing and the latter one increasing as the increase of flood peak. The obtained conclusions from the bivariate hydrologic analysis can provide decision support for flood control and mitigation.

Keywords: Flood risk; Copula; Flood frequency analysis; Distribution; Conditional distribution; Gaussian mixture model.

1. Introduction

Extreme hydrologic events, such as floods, droughts and storms, have been leading to extensive property losses in recent decades. Specifically, floods have become one of the most common natural disasters, posing significant risks to human beings and environment [22, 31, 41, 67, 71-72, 81]. Hydrological frequency analysis procedures are widely adopted to estimate the occurrence probabilities of floods, providing decision support for many water resources management practices, such as reservoir management, dam design and flood insurance studies [6, 16-20, 43, 45, 82-85]. Moreover, a flood is associated with multidimensional characteristics. Consequently, flood frequency analysis under consideration of multiple flood variables would be desired to provide a full screen for a flood.

Copula functions, in recent years, have been widely used for multivariate hydrologic modeling, such as multivariate flood frequency analysis [7, 18, 22, 26-27, 65, 78-79], drought assessments [14, 35, 42, 59, 60, 64], storm or rainfall dependence analysis [2-4, 66], streamflow simulation [32, 39-40, 58]. De Michele and Salvador [13] initially introduced the concept of copulas into hydrological simulation, which described the dependence between storm duration and average rainfall intensity by means of a suitable 2-Copula. Salvadori and De Michele [51] characterized the dependence between storm duration and intensity via a suitable 2-Copula with the marginal distributions endowed with Generalized Pareto laws. Recently, Salvador and De Michele [52] conducted multivariate real-time assessment of droughts via copula-based multi-site Hazard Trajectories and Fans. The main advantage of copula functions over classical bivariate frequency analyses is that the selection of marginal distributions and multivariate dependence modelling are two separate processes, giving additional flexibility to the practitioner in choosing different marginal and joint probability functions [24, 36, 65, 78]. Consequently, the selection of marginal distributions would definitely impact the performance of the copula in modelling multivariate hydrologic simulation.

In multivariate hydrologic frequency analysis through copula functions, the flood variables under consideration include: the annual maximum peak discharges, and the associated hydrograph volumes and durations. Consequently, the distributions for modelling these flood variables would be various. For example, for modelling the annual maximum flood series, the used distributions over the world include extreme value type 1 (EV1), general extreme value (GEV), extreme value type 2 (EV2), two component extreme value, normal, lognormal (LN), Pearson type 3 (P3), Log Pearson type 3 (LP3), Gamma, exponential, Weibull, generalised Pareto and Wakeby distributions [5, 12]. Previous studies have shown that, in modelling multivariate flood frequency through copula functions, the marginal distributions of peak, volume and duration were different at different sites. For example, Sraj et al. [65] took bivariate flood frequency analysis using copula function for the Litija station on the Sava River, in which log-Pearson 3 distribution was chosen for modelling discharge peaks and hydrograph durations, and the Pearson 3 distribution was selected for hydrograph volumes. Reddy and Ganguli [48] applied Archimedean copulas for bivariate flood frequency analysis, where the normal kernel density function was used for quantifying the distributions of peak flow and duration, and quadratic kernel density function was applied for volume.

A Gaussian mixture model (GMM) is a mixture statistical model of a finite number of Gaussian distributions with unknown parameters. It is a semiparametric probability density function expressed as a weighted sum of Gaussian component densities, and all samples are assumed to be generated from this mixture model. GMMs are commonly used to model the probability distributions of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system [50]. The finite Gaussian mixture model can theoretically approximate any continuous distribution very closely if properly given a sufficient number of components [34]. Several research works have been reported to apply mixed distribution models to analyze hydrological and environmental data [15, 21, 34, 61-62, 68]. For example, Yue et al. [76] proposed a Gumbel mixed model for flood frequency analysis. Singh et al. [62] proposed a mixed distribution method for nonidentically distributed hydrologic flood data. He [34] applied the GMM for analyzing the multiply censored environmental data. Although GMM and other mixed model methods have been widely proposed to model the water and environmental samples, these proposed methods have some limitations in practical multivariate flood risk analysis. For instance, the Gumbel mixed distribution proposed by Yue et al. [76] can only applied to positively correlated random variables with the correlation coefficient less or equal to 2/3 [77].

As an extension of previous research, this paper aims to couple the GMM into copulas, leading to a coupled GMM-copula method for multivariate hydrologic risk analysis. The advantages of the proposed method are that i) the GMM can provide good estimations for the marginal distributions and ii) the copula method can relax the assumptions in previous mixed models such as same type distribution, correlation restriction [78]. Moreover, an integrated multivariate risk indicator is proposed to reveal significance of effects from persisting high risk levels due to impacts from multiple interactive flood variables. Such an analysis will be based on provision of the coupled GMM-copula method. Finally, the conditional probability density distributions (pdfs) of flood volume and duration under peak flows with different return periods will be characterized, intending to explore potential control and management practices once a flood has occurred. The proposed method will be applied to the Yangtze River (Chang Jiang), China

2. Methodology

2.1 Gaussian Mixture Model

The mixture model is a useful tool for density estimation, and can be viewed as a kind of kernel method [33]. Mixture models can use any component densities but the Gaussian mixture model (GMM) is the most popular [33]. The probability density function of a Gaussian mixture model is expressed by a weighted sum of M-component Gaussian probability densities as given below:

$$p(x) = \sum_{j=1}^{M} \alpha_j N_j(x; \mu_j, \sigma_j) \tag{1}$$

where x are one-dimensional measurement samples; α_j (j = 1, 2, …, M) denote the mixture weights; $N_j(x; \mu_j, \sigma_j)$ (i = 1, 2, …, M) are the component Gaussian densities, which can be expressed as:

$$N_j(x; \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(\frac{-(x-\mu_j)^2}{2\sigma_j^2}\right) \tag{2}$$

where μ_j and σ_j respectively denote the mean and standard deviation for the j^{th} Gaussian distribution model. The weights α_j are nonnegative and must satisfy $\sum_{j=1}^{M} \alpha_j = 1$. The GMM has two main advantages in practical applications in many engineering fields: (i) it can sufficiently approximate a broad class of distribution functions encountered in practice, if an appropriate size of components are given in the mixture; (ii) the form of the GMM simplifies the derivation of the subsequent estimation method and avoids the identifiability problem [34].

Let $\theta_j = (\alpha_j, \mu_j, \sigma_j)$, then $p(x_i)$ has M Gaussian models, and M sets of parameters are needed to be estimated. If $\Theta = (\theta_1, \theta_2, ..., \theta_M)$, The likelihood function of the GMM model can be expressed as:

$$l(x|\Theta) = \log \prod_{i=1}^{N} \sum_{j=1}^{M} \alpha_j N_j(x;\mu_j,\sigma_j) = \sum_{i=1}^{N} \log \sum_{j=1}^{M} \alpha_j N_j(x;\mu_j,\sigma_j) \tag{3}$$

The analytical solution to maximize Equation (3) is generally impractical due to the composite operation of component wise product. The Expectation-Maximization (EM) algorithm is usually applied to generate the unknown parameters (i.e. α_j,μ_j, σ_j) in a Gaussian mixture model. The EM algorithm is an iterative procedure for estimating the parameter θ_i of a target distribution that maximize the probability under consideration of a given set of realizations, $\{x_1, x_2, \ldots, x_N\}$ [63]. The EM algorithm is an iterative succession of expectation and maximization steps for obtaining the maximum likelihood (ML) estimate, which involves two steps: E-step and M-step. A brief description of the EM algorithm can expressed as follows:

E-step: Calculate the posterior probability of mixture component j having generated realization $\boldsymbol{x}_i$ based on the present estimates:

$$\beta_{ij} = E(\alpha_j | x_i;\Theta) = \frac{\alpha_j N_j(x_i;\Theta)}{\sum_{j=1}^{M} \alpha_j N_j(x_i;\Theta)}, \ 1 \le i \le \mathrm{N},\ 1 \le j \le \mathrm{M}. \tag{4}$$

M-step: Update the model parameters in accordance with their weighted averages across all realizations:

$$\alpha_j^{'} = \frac{\sum_{i=1}^{N} \beta_{ij}}{N} \tag{5}$$

$$\mu_j' = \frac{\sum_{i=1}^{N} \beta_{ij} x_i}{\sum_{i=1}^{N} \beta_{ij}} \tag{6}$$

$$\sigma_j' = \frac{\sum_{i=1}^{N} \beta_{ij} (x_i - \mu_j')^2}{\sum_{i=1}^{N} \beta_{ij}} \tag{7}$$

2.2. Copula Method for Bivariate Flood Frequency Analysis

2.2.1. Concept of Copula

A copula function is a multivariate probability distribution with its marginal distribution being uniform. Sklar's Theorem states that any n-dimensional distribution function F can be formulated through a copula and its marginal distributions, which is expressed as follows:

$$F(x_1, x_2, ..., x_n) = C(F_{X_1}(x_1), F_{X_2}(x_2), ..., F_{X_n}(x_n)) \tag{8}$$

where $F_{X_1}(x_1), F_{X_2}(x_2), ..., F_{X_n}(x_n)$ are marginal distributions of random vector (X_1, X_2, …, X_n). If these marginal distributions are continuous, then a single copula function C exists, which can be written as [46, 56]:

$$C(u_1, u_2, ..., u_n) = F(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2), ..., F_{X_n}^{-1}(u_n)) \tag{9}$$

More details on theoretical background and properties of various copula families can be found in [46] and [56].

A number of copula functions have been developed, mainly including the Archimedean, elliptical, extreme value copulas. Among them, the Archimedean copulas are quite attractive in hydrologic frequency analysis, because they can be easily generated, and are capable of capturing a wide range of dependence structure with several desirable properties, such as, symmetry and associativity [22]. The Ali-

Mikhail-Haq, Cook-Johnson and Gumbel-Hougaard and Frank copulas are most widely used Archimedean copulas for probabilistic assessment of flood risk. Table 1 presents some basic characteristics of the applied single-parameter bivariate Archimedean copulas.

Place Table 1 here

2.2.2. Conditional Distribution

If an appropriate copula function is selected, the conditional joint distribution can then be obtained. Following [46] and [56], the conditional distribution function of U_1 given $U_2 = u_2$ can be expressed as:

$$C_{U_1|U_2=u_2}(u_1) = P(U_1 \le u_1 \,|\, U_2 = u_2) = \frac{\partial}{\partial u_2} C(u_1, u_2) \tag{10}$$

Similar conditional cumulative distribution for U_2 given $U_1 = u_1$ can be obtained. Moreover, the conditional cumulative distribution function of U_1 given $U_2 \le u_2$ can be expressed as:

$$C_{U_1|U_2\le u_2}(u_1) = P(U_1 \le u_1 \,|\, U_2 \le u_2) = \frac{C(u_1, u_2)}{u_2} \tag{11}$$

Likewise, an equivalent formula for the conditional distribution function for U_2 given $U_1 \le u_1$ can be obtained.

The probability density function (pdf) of a copula function can be expressed as:

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2} \tag{12}$$

and the joint pdf of the two random variables can be obtained as:

$$f(x_1, x_2) = \frac{\partial^2 C(u_1, u_2)}{\partial x_1 \partial x_2} = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2} \frac{\partial u_1}{\partial x_1} \frac{\partial u_2}{\partial x_2} = f_{X_1}(x_1) f_{X_2}(x_2) c(u_1, u_2) \tag{13}$$

Consequently, the conditional pdf of X_1, given the value of X_2, can be formulated as:

$$f(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)} = f_{X_1}(x_1) c(u_1, u_2) \tag{14}$$

And the conditional pdf of X_2, given the value of X_1, can be expressed as:

$$f(x_2 \mid x_1) = \frac{f(x_1, x_2)}{f_{X1}(x_1)} = f_{X_2}(x_2) c(u_1, u_2) \tag{15}$$

2.2.3. Primary and Secondary Return Period

If appropriate copula functions are specified to reflect the joint probabilistic characteristics among peak, duration and volume of the flood, some conditional, primary and secondary return periods can be obtained. Specifically, Joint (primary) return periods called OR and AND can be formulated as [28, 56, 65]:

$$T_{u_1,u_2}^{OR} = \frac{\mu}{1 - C_{U_1U_2}(u_1, u_2)} \tag{16}$$

$$T_{u_1,u_2}^{AND} = \frac{\mu}{1 - u_1 - u_2 + C_{U_1U_2}(u_1, u_2)} \tag{17}$$

where μ is the mean inter arrival time of the two consecutive flood events.

The secondary return period, called Kendall's return period, is firstly introduced by Salvadori and De Michele [53] to characterize probability of occurrence of an event in the area over the copula level curve of value t. This concept has been successively elaborated and extended by many research works [14, 54-55, 57]. The secondary return period can be expressed as follows:

$$\overline{T}_{u_1,u_2} = \frac{\mu}{1 - K_C(t)} \tag{18}$$

where K_C is the Kendall's distribution, associated with theoretical copula function C_θ. For Archimedean copulas, K_C can be expressed as [46]:

$$K_C(t) = t - \frac{\phi(t)}{\phi'(t^+)} \tag{19}$$

where $\phi'(t^+)$ is the right derivative of the copula generator function $\phi(t)$, as presented in Table 1.

2.2.4. Bivariate Hydrologic Risk Analysis

Risk is the probability of occurrence of an extreme, dangerous, hazardous, or (more generally) undesirable event [38]. In engineering design of hydrologic infrastructures, risk can be explained as the chance of downstream flood attributable to uncontrolled water release from upstream flood facilities (e.g. a reservoir), leading to life and property losses [23]. Yen [73] proposed a formulation for the risk of failure associated with the return period of a flood event, which can be expressed as:

$$R = 1 - (1 - p)^n = 1 - q^n = 1 - (1 - 1/T)^n \tag{20}$$

where R is the risk of failure; p and q is the exceedance and nonexceedance probability, respectively; T is the return period of a flood event; n is the design life of the hydraulic structure.

In practical flood control practice, it is necessary to characterize the flood event through multiple aspects (e.g. peak and duration) rather than only one flood variable (e.g. peak). For example, a flood event with high peak flow and long duration may result in serious losses in properties, while a short-duration event with high peak may only cause a flash flood. Consequently, bivariate hydrologic risk would be much helpful in taking nonstructural safety measures, and developing flood mitigation strategies. In this study, the joint return period in "AND" case is applied to define the bivariate risk analysis as follows:

$$R_{u_1,u_2} = 1-(1-\frac{1}{T_{u_1,u_2}^{AND}})^n \tag{21}$$

2.3. Goodness-of-fit Statistical Tests

After parameter estimation for both the marginal and joint distributions, the goodness-of-fit statistic tests would be performed to determine whether those estimated distributions are satisfied. The root mean square error (RMSE), Akaikes Information Criterion (AIC) and the Kolmogorov-Smirnov (K-S) goodness-of-fit tests would be employed to evaluate the performance of the marginal distributions obtained through the parametric distributions and the Gaussian mixture model (GMM). And the Rosenblatt transformation [49] would be applied to investigate the performance of joint distributions in describing the dependency between flood variable pairs.

In the process of evaluating the performance of marginal distributions obtained through the parametric methods and GMM, the empirical nonexceedance probabilities would be obtained through the Gringorten plotting position formula [30], which is expressed as:

$$P\left(K \le k\right) = \frac{k-0.44}{N+0.12} \tag{22}$$

where N stands for the sample size; k stands for the k^{th} smallest observation in the data set; and the data set is arranged in an increasing order.

The RMSE, Akaikes Information Criterion (AIC) and the K-S test are used to evaluate fitting effect of different probability distributions to the flood variables. The RMSE can be expressed as [69]:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}\left(x_k^{est} - x_k^{obs}\right)^2}{N}} \tag{23}$$

where x_k^{est} denote theoretical values from the fitted probability distribution; x_k^{obs} denote the empirical probabilities obtained through Equation (22); N is the sample size. Based on RMSE, the AIC value can be obtained as follows:

$$AIC = N * \ln((RMSE)^2) + 2k \tag{24}$$

where k is the number of unknown parameters in the probability distribution.

The K-S test is a nonparametric probability distribution free test [80]. The statistic of K-S test quantifies the largest vertical difference between the estimated and empirical distributions [44, 47]. Given n increasing ordered data points, $x_{(\cdot)}$, the K-S test statistic is defined as [11]:

$$T = \sup_x \left|F^*(x) - F_n(x)\right| \tag{25}$$

where $F^*(x)$ means the estimated distribution, $F_n(x)$ denotes the empirical distribution, and 'sup' stands for supermum. The P-value for K-S test was approximated using Miller's approximation [80].

For evaluating the performance of copulas, the goodness-of-fit test based on Rosenblatt transformation would be employed based on the recommendation of Genest et al. [25]. They argued that test statistics based on the Cramér von Mises functional of a process tend to be more powerful than those based on the Kolmogorov–Smirnov distance taken on the same process [25]. Consequently, Cramér von Mises statistic will be adopted to test the performance of the copulas with the corresponding p-values being approximated through Monte Carlo simulation. The detailed procedures for performing goodness-of-fit test for copulas based on Rosenblatt transformation are provided by Genest et al. [25].

3. Study Area and Data

3.1. Overview of the Studied Watershed

The proposed GMM-copula method would be applied to the Yangtze River to demonstrate the applicability of the proposed method in analyzing multivariate flood risk. Yangtze River is the longest river in Asia, and the third longest river in the world, with a length of 6,300 km for the main stream, flowing from Qinghai Province eastward to the East China Sea at Shanghai. Floods of the Yangtze River in central and eastern China have occurred periodically and often caused considerable destruction of property and loss of life [10]. For example, in 1998, the entire Yangtze River basin suffered from tremendous flood—the largest flood since 1954, which led to the economic loss of 166 billion Chinese Yuan [74]. Hence, multivariate flood risk analysis for Yangtze River is very important for flood prevention and disaster relief.

For the Yangtze River, floods are caused by temporal-spatial variation in precipitation. A large part of the Yangtze River Basin has subtropical monsoon climate, with the precipitation being concentrated during summer reason. Consequently, summer is the main flood season due to the heavy monsoon rainfall [10]. The floods in the middle and lower reaches of the Yangtze River mainly stem from the upper region of the Yichang Station. The Yichang station plays a vital role for flood control in the middle and lower reaches of Yangtze River. It is also the control station for the Three Georges Reservoir. The flood from Yichang station contributes about 50% of the total flow volume of the middle and lower reaches of Yangtze River. Moreover, The Jingjiang reach (Figure 1), located in the middle reach of the Yangtze River from Zhicheng to Chenglingji with a length of 340 km, is the most prone area to suffer floods in the Yangtze River basin. Approximate 90% flood in Jingjiang reach comes from the flood in Yichang station [9].

Due to the key role of the Yichang station in controlling the flood in the middle and lower reaches of Yangtze River, the daily streamflow data from Yichang station would be applied to analyze the bivariate flood risk in Yangtze River. Figure 1 shows the location of Yichang station, which is also the control site of the Three Gorges Gam (TGD). The Three Gorges Dam (TGM) is the largest hydraulic project in terms of design capacity over the world. It has produced dramatic benefits in flood control, power generation and navigation. Recently, the impacts of the TGM project on hydrology and environment have been attracting the world's attention. The Yichang Station is the control site of TGD, which also divided the Yangtze River into the upper and middle reaches. This study mainly focused on the flood from the upper Yangtze River, which is 4,529 km long, up to 3/4 of the whole length of the Yangtze River, with a drainage area of 1,006,000 km^2 [8].

3.2. Historical Flood Characteristics at Yichang station

Based on the daily flow data, the annual maximum peak discharges and the corresponding hydrograph volumes and durations values can be obtained. Hence, although the peak discharges are definitely annual maximums, the hydrograph volumes and durations are not necessarily also annual maximums [65]. The single-peaked flood hydrograph is shown in Figure 2. Flood duration (D) can be determined by identifying the time of rise (point "s" in Figure 2) and fall (point "e" in Figure 2) of the flood hydrograph. The start of the surface runoff is marked by the sharp rise of the hydrograph and end of the flood runoff is identified by the inflection point on the receding limb of the hydrograph. Between these two points, the total flood volume is estimated. If time of rise of the flood hydrograph is denoted by SD (day) and fall by ED (day), the flood volume (V) of each flood event is determined using following expression (Yue 2001):

$$V_i = (V_i^{total} - V_i^{baseflow}) = \sum_{j=SD_i}^{ED_i} Q_{ij} - \frac{1}{2}(Q_{is} + Q_{ie})(1 + D_i) \qquad (29)$$

For a flood with multiple peaks, the peak flow would be the maximum peak value in the flood. The corresponding duration is identified based on Figure 2 and the associated volume is calculated through Equation (29). Moreover, when multiple floods happen in one year, the flood with maximum peak is only considered since risk analysis pays attention to flood extremes. Once the flood characteristics are obtained from daily streamflow data, then flood frequency analysis can be analyzed. Figure 3 shows the variations in flood peak discharge (i.e., Q (m^3/s)), hydrograph volume, (i.e., V (m^3/s day)) and hydrograph duration, (i.e., D (day)) from 1882-2007.

Place Figures 2 and 3here

----

4. Result Analysis

4.1. Marginal Probability Distribution Functions of Flood Variables

One of the main advantages for the copula method is that the marginal distributions and multivariate dependence modelling are two separate processes. Consequently, to analyze the multivariate flood frequency in the Yangtze River, the marginal distributions of flood variables can be quantified firstly. In this study, the Gaussian mixture model would be applied to quantify the marginal distributions of flood peak, volume and duration. Besides, many parametric distributions have been used to estimate flood frequencies from observed annual flood series, such as the general extreme value distribution in the United Kingdom, Log-Pearson Type-III in the U.S. and Pearson Type III in China [1, 37, 70, 75]. To demonstrate the performance of GMM in modeling the marginal distributions of flood variables, the GMM would be compared with four parametric methods, including Gamma, GEV, Lognormal distributions and Pearson Type III. The expressions for probability functions (pdfs) for Gamma, GEV, Lognormal, Pearson Type III and the values of their associated unknown parameter are presented in Table 2. These parameters are obtained through maximum likelihood estimation method. Table 3 shows the marginal distributions of flood variables obtained through GMM, in which the unknown parameters are obtained through the EM algorithm.

Place Tables 2 and 3 here

Figure 4 illustrates the fitted marginal distributions for the three flood variables through Gamma, GEV, Lognormal, Pearson Type III (i.e. P3), and GMM-based distribution functions. The cdfs and pdfs for the marginal distributions of flood variables (in Figure 3) show good agreement between the theoretical and the empirical distributions. Generally, the flood peak and volume can be well quantified through the proposed four parametric distributions and GMM-based distributions. For the flood duration, there are some deviations between the theoretical and observed values, especially for the four parametric distributions. To further evaluate the GMM and four parametric distributions in quantifying the probability distributions of flood variables, the Kolmogorov-Smirnov (K-S) test would be conducted. Table 4 presents the results of K-S tests. The results indicate that all the proposed five methods can be employed to model the distributions of flood peak, volume and duration, with the P-values larger than 0.05. However, the performance of the four parametric distributions in modelling the flood duration is not as well as those in quantifying the flood peak and volume, since the P-values are less than 0.1. The root mean square error (RMSE) and AIC values, which are respectively expressed as Equations (23) and (24), would then adopted to compare the performance of those four distributions. As shown in Table 4, the GMM-based distributions perform best in quantifying the three flood variables, with lowest RMSE and AIC values. Especially for flood duration, the GMM-based distribution performs much better than the other four parametric distributions.

Place Figure 4 and Table 4 here

4.2. Joint Distributions Based on Copula Method

The dependence of flood variables was evaluated through the Pearson's linear correlation (r), and one non-parametric dependence measure, Kendall's tau. Table 5 presents the values of Pearson's linear correlation coefficient and Kendall's tau among flood peak, volume and duration. The values of Pearson's r and Kendall's tau between duration and volume are highest, followed by the flood pairs of peak-volume, and peak-duration. In detail, the Pearson, Kendall correlation coefficient values are

0.55 and 0.66 for peak-volume, 0.68 and 0.75 for volume-duration, and 0.27 and 0.35 for peak-duration. These results indicate that the correlation between the flood duration and volume would be higher than the other two flood variable pairs. In our case, the correlation coefficient for peak and duration is much smaller than for the other two pairs (i.e. peak-volume and volume-duration), which is consistent with conclusions from previous studies [29, 36, 48, 65].

The Archimedean copulas are the most attractive copulas for multivariate hydrologic risk analysis due to their ease for construction and capability of capturing dependence structure with several desirable properties. The Cook-Johnson (Clayton), Gumbel-Hougaard, Frank and Ali-Mikhail-Haq copulas are the four widely used Archimedean copulas. However, the Ali-Mikhail-Haq copula is only applicable with the Kendall's tau value varied within [-0.18, 0.33] [46]. In this study, the flood pair of peak-duration exhibits the lowest Kendall's tau values with a value being 0.35. For the flood pairs of peak-volume and volume-duration, the corresponding Kendall's tau values are 0.66 and 0.78, respectively. Consequently, the Ali-Mikhail-Haq copula is excluded and the Cook-Johnson (Clayton), Gumbel-Hougaard and Frank copulas would be selected to model the dependence among flood variables. The unknown parameters in these four copulas are estimated by method-of-moments-like (MOM) estimator based on inversion of Kendall's tau.

The joint distribution functions for flood peak and volume, obtained through the three above-mentioned copulas, are shown in Figure 5; the joint distributions for peak-duration, and volume-duration are shown in Figures 6 and 7, respectively. Also, comparison between empirical and theoretical copula functions for the flood pairs of peak-volume, peak-duration and volume-duration can be found in Figure 5, 6 and 7, respectively. In Figures 5 to 7, the red dashed contour lines represent the empirical copula obtained through $C_n(u,v) = 1/n\sum_{i=1}^{n} 1(R_i/(n+1) \leq u, S_i/(n+1) \leq v)$, where $u, v \in [0,1]$, R_i and S_i denote the ranks of the ordered sample, and the solid contour lines represent the theoretical copula. The results indicate that the empirical and the three theoretical copulas can match well for flood peak-volume. For flood peak-duration, there are some deviations between theoretical copulas and empirical copula at low probability levels. This may due to the discrete characteristic of the duration sample and the relative low accuracy of the obtained marginal distribution. However, at high probability levels, the theoretical values can fit well with the empirical copula values. Also, similar characteristic can be found for the flood pair of volume-duration.

Since there are three candidate copulas, investigating the differences among the three chosen copulas and identifying the most appropriate copulas for further analysis are necessary. In this study, the Rosenblatt transformation with Cramér von Mises statistic is employed to evaluate performance of the proposed three copulas in modelling joint distributions of flood variable pairs. Table 6 presents the results of statistic test results for the three flood pairs. It can be seen that the proposed Cook-Johnson (Clayton), Gumbel-Hougaard and Frank copulas can be applicable for modelling the dependence of flood peak-volume, peak-duration and volume-duration, with the p-values larger than 0.05. To further identify the most appropriate one, the root mean square error (RMSE) (expressed by Equation (23)) is used to test the goodness of fit of sample data for the theoretical joint distribution obtained using copula functions. Table 6 shows the RMSE values for joint distributions obtained through different copula functions for flood peak-volume, peak-duration and volume-duration. The differences among these three copulas in quantifying the joint probabilities of the three flood pairs are rarely small. Take the flood pair of peak-volume as an example, the RMSE value for the Gumbel-Hougaard and Cook-Johnson copula is 0.0168 and 0.0199 respectively, while the RMSE value of Frank copula is 0.0149. Based on the values of RMSE, it can be concluded that the Frank copula would be best for quantifying the joint distribution of flood peak-volume. Similarly, the Frank copula would be the most appropriate copula for modelling the joint distribution of flood peak-duration and volume-duration.

Place Figures 5 - 7 here and Table 6

4.3. Bivariate and Conditional Risk Analysis

4.3.1. Conditional Cumulative Distribution Functions and Return Periods of Flood Characteristics

Based on the results presented in Table 6, the Frank copula would be chosen to model the dependence between the three flood pairs. Consequently, the conditional cumulative distribution functions (cdfs) of one flood variable, given the value of the other flood variable value, can be derived based on the fitted copula function. Figure 8 shows the conditional cdfs of flood variables, which are obtained through Equations (10) and (11). It can be seen that, among the flood pairs of peak-volume, peak-duration and volume-duration, the values of conditional cdf for one flood variable would decrease as the value of other flood variable increase. This indicates positive correlation structures between peak-volume, peak-duration, and volume-duration. Besides, the decreasing trend of conditional cdfs for peak-duration is less than the other two pairs, indicating less correlation structures between peak and duration. This is consistent with the results presented in Table 5.

Place Figure 8 Here

The concurrence probabilities of various combination of flood variable would be more helpful for actual flood control and management than the univariate flood frequency analysis. As expressed as Equations (16) – (19), the joint return period and second return period can be derived based on the selected copula functions. Table 7 presents the joint return periods of "AND" and "OR" cases for different flood pairs. In general, the joint return period in "AND" case is much longer than the joint return period in "OR" case. For example, if both the flood peak and duration are in 100-year return period, the "OR" joint return period of flood peak-duration would be 50.9 years, while, in contrast, the "AND" joint return period is 2809.4. Furthermore, the "AND" return period for flood peak-duration is longest among the three flood variable pairs due to the low correlation between flood peak and duration, followed by the "AND" return periods of peak-volume and volume-duration. Correspondingly, the "OR" joint return period of peak-duration is shorter than the "OR" return periods of the other two flood variable pairs. Figure 9 shows the contour plot of the joint return periods in "OR" and "AND" cases for different flood pairs. Also, the secondary return periods are presented in Table 7, which can be useful for analyzing risk of supercritical flood events. The secondary return period is defined as the average time between the concurrence of two supercritical flood events, which would appear more rarely than the given design return period. As the primary return period increases, the probability of supercritical flood events decreases, leading to increase of the secondary return period. Furthermore, the secondary return period is higher than the joint return period in T^{OR} case but less than the joint return period in T^{AND} case.

Place Table 7 and Figure 9 here

4.3.2. Bivariate Hydrologic Risk Analysis

The damages caused by a flood, such as the failure of hydraulic structures, mainly due to the high peak flow of the flood. The annual maximum peak discharge would be the central issue to be considered for hydrologic risk analysis. Moreover, the flood discharge volume and duration would be also under consideration in practical flood control and mitigation, in which the flood duration is the vital factor for decision maker in characterizing the flood control pressure, and the flood volume is related to flood diversion practices. Consequently, multivariate flood risk analysis, which involves more flood variables than just considering flood peak, would be more helpful for actual flood control. Therefore, in this study, a bivariate hydrologic risk analysis method would be proposed to identify the inherent flood characteristics in Yangtze River. In particular, three flow amounts, with a return period of 50, 70, and 100-year, respectively are considered as designed standard for the river levee around the Yichang Station. Four service time scenarios are also assumed for the river levee, namely 30, 50, 70 and 100 years.

(1) Bivariate flood risk under different flood peak-volume scenarios

The bivariate hydrologic risk for flood peak flow and volume indicates the concurrence probabilities of flood peak flow and volume values. Figure 10 shows the bivariate flood risk under different flood peak-volume scenarios. For the univariate hydrologic risk expressed as Equation (20), its value would decrease as the increase of designed peak flow or the service time of the river levee. As can be seen from Figure 10, if the flood volume is less than 1×10^5 (m^3/s day), the bivariate risk values for flood peak-volume would not decrease significantly for all designed flows and service time periods. This suggests that the occurrence of one flood peak flow would usually be accompanied with a flood volume up to 1×10^5 (m^3/s day). However, for one designed flow and service time period, the values of the bivariate risk for flood peak-volume would decrease when the associated flood volume is larger than 1×10^5 (m^3/s day). This indicates that the probabilities of concurrence of large flood volumes and high peak flows would be generally less than the occurrence probabilities of high peak flows.

Place Figure 10 here

The implication for the bivariate risk of flood peak flow and volume is to provide decision support for hydrologic facility design and establishment of flood diversion areas. In actual flood control practices, the excess water of floods can be redirected temporary holding ponds or other bodies of water with a lower risk or impact to flood. For example, in China, the flood diversion areas are rural areas that are deliberately flooded in emergencies in order to protect cities. In flood diversion practice, the bivariate risk for flood peak flow and volume would be an important reference for the design of flood diversion areas. For example, as shown in Figure 9, for the river levee with a designed flow of 50-year return period and 30-year service period, the flood risk value would be about 45, 43, 35, 22% with a flood volume being 0.5, 1, 1.5, and 2×10^5 m^3/s, respectively. Based on these bivariate risk values, the flood manager can design corresponding scales of the flood diversion areas.

(2) Bivariate flood risk under different flood peak-duration scenarios

Figure 11 shows the variations in the failure risk of river levee around Yichang Station under different flood peak-duration scenarios. The bivariate hydrologic risk can reflect the failure risk of river levee with respect to the variation of flood durations. In Figure 11, the initial risk values (points on the y-coordinate) are obtained through Equation (19) without considering impacts of the flood duration, while the points on the solid, dashed and asterisk lines are derived based on Equation (21). The results in Figure 10 indicate that the bivariate risk of flood peak-duration would not decrease at the flood duration less than 8 days, and then decrease as the increase of flood duration. Such results suggest that the once a flood occurs at Yichang Station, this flood would last up to 8 days without significant decrease in the occurrence probability. However, the concurrence of a flood with high peak flow and long duration would not appear frequently.

Place Figure 11

The bivariate risk of the flood peak flow and duration can provide useful information for actual hydrologic facility design and potential flood control. In practical engineering hydrologic facility construction, the return period of peak flow would be the key factor to be considered. Moreover, the flood duration would be related to flood defense preparation, in which longer flood duration would generally require more flood defense materials such as sand, wood, bags. Consequently, the bivariate flood risk values under different flood peak-duration scenarios would be considered as references for decision makers to determine how much materials would be prepared for flood defense.

4.3.3. Conditional Probability Density Functions of Flood Characteristics

In addition to derive the conditional cdfs and joint return periods based on the best-fitted copula for the historical flood data, the conditional probability density functions (pdfs) of the flood variable can also be generated based on Equations (12) - (15). In flood risk analysis, the peak flow would be the critical factor to judge whether a flood appears. However, once the flood occurred, the severity of the flood would also influenced by flood duration and volumes. In detail, the flood duration would be related to the flood control pressure in which flood defense materials should be prepared for strengthening the river levee and inspection should be conducted for the safety of the river levee. The flood volume would generally influence the flood diversion practices, in which excess water would be diverted to temporary holding ponds with lower risk in order to protect cities.

Figure 12 shows the distributions of flood volume conditional on the flood peak flows with different return periods. In this study, the flood peak flows with return periods of 10, 20, 50, and 100-year are under consideration. Each curve represents the probability distribution function (pdf) of flood volume associated with the flood peak flow with a particular return period. It can be seen that, once a flood appears, the conditional pdf of flood volume would approximately follow a bimodal distribution, with the two vertexes appearing around 1.2 and 2.0×10^5 m^3/s day, respectively. More specifically, the former vertex would appear more frequently for small floods while the latter one is more frequent for large floods. Moreover, as the peak flows increases, the two vertexes of the flood volume would also increase correspondingly, but the latter vertex seems to increase more than the former vertex, as shown in Figure 12. Finally, the conditional pdf of flood volume also shows that the occurrence probability of the first vertex would decrease while the occurrence probability of the latter vertex would increase when the return period of the flood peak increases. Such pdfs of flood volume conditional on different flood peak flows can provide support information for flood diversion practices and be involved in the flood optimization models to determine the capacities of flood diversion. For instance, once a flood occurs and excessive flood is required to be diverted to some flood discharge area, the associated flood volume should be estimated before conducting flood diversion. From Figure 12, it can be concluded that two flood volumes would be primarily under consideration, around 1.2 and 2.0×10^5 m^3/s day, respectively. Particularly, the flood volume of 1.2×10^5 m^3/s day would be paid more attention for small floods while the volume of 2.0×10^5 m^3/s day would be paid more attention for large floods. These results can provide useful information for flood managers to prepare appropriate flood diversion schemes. Moreover, Table 8 shows the statistical characteristics for the PDFs of flood volume conditional on different floods. The results indicate that, as the increase of the flood peak return period, the mean value of the conditional pdf of flood volume would generally increase, while the standard deviation of the conditional pdfs would not change significantly.

Place Figure 12 and Table 8 here

Figure 13 shows the distributions of flood duration conditional on the flood peak flows with different return periods. It is indicated that, the conditional pdfs of flood duration would also obey bimodal distributions, with two vertexes appearing around 11 and 15 days. Specifically, as the increase of flood return period, the former vertex around 11 days would not change significantly, but the latter vertex (around 15 days) would show a remarkable increase. For instance, the latter vertex of the duration pdf conditional on a flood with 10 years would be about 14.8 days, while such a vertex would increase to around 15.7 days when the return period of the flood peak increase to 100 years. Moreover, the latter vertex show a more frequent occurrence probability than the former vertex except for a small flood with a 10-year return period. The engineering implications of the pdfs of flood duration conditional on flood peak is to provide an insightful screening for the duration time once a flood occurs, which will further be considered as a reference for flood defense materials preparation and river levee safety inspection. The statistical characteristics of the conditional pdfs are presented in Table 8. The results indicate that the mean values of the conditional pdfs would increase while the standard deviations are nearly constants. Furthermore, for a flood with a return period larger than 50 years, associated mean value of the flood duration would not change significantly, even though the latter vertex shows apparent increase with the increase in flood peak return period.

Place Figure 13 here

----

5. Conclusions

In this study, a bivariate hydrologic risk analysis method is proposed through coupling Gaussian mixtures into copulas. In the bivariate hydrologic risk analysis framework, the bivariate frequency analysis, which considered the flood variables pairs of flood peak, duration and volume, was firstly conducted through coupling Gaussian mixture models into copulas, leading to a coupled GMM-copula method. This method improved upon previous methods through providing better estimation for marginal distributions through Gaussian mixture models. The primary, conditional and secondary return periods were then derived based on the selected copula. The bivariate hydrologic risk was defined based on the joint return period of flood variables to reflect the hydrologic risks of flood peak-duration and flood peak-volume pairs. Besides, the conditional probability distribution functions (pdfs) of flood volume and duration under different flood peak scenarios were also derived to explore the variation in pdfs of flood volume and duration corresponding to different flood peak flows.

The proposed method was applied for quantifying the bivariate hydrologic risk in the Yangtze River based on the daily discharge measurements at Yichang Station. The results indicated that, compared with the parametric distributions such as Gamma, GEV and Lognormal and Pearson Type III functions, the Gaussian mixture model could perform much better for quantifying the marginal distributions of flood peak, volume and duration. Such conclusions has been demonstrated through the K-S test, the RMSE and AIC values.

For the dependence among flood variables, the Frank copula would be best for quantifying the joint distributions of the three flood variable pairs. The bivariate risks of flood peak-volume and flood peak-duration were evaluated based on the joint return period in “AND”, revealing significance of effects from persisting high risk levels due to impacts from multiple interactive flood variables. The results show that the bivariate risk of flood peak-volume would keep constant for the corresponding volume less than 1.0×10^5 m^3/s day, show apparent decrease for the flood volume varying between 1.0 and 1.7×10^5 m^3/s day, and present most significant decreasing rates for the volume lager than about 1.7×10^5 m^3/s day. For the bivariate risk of flood peak-duration, it would not change significantly for the flood duration less than about 8 days and then show significant decreasing rate. Moreover, the pdfs of flood volume and duration conditional on flood peak appeared to be bimodal. The two vertexes for the conditional pdfs of flood volume were located at around 1.2 and 2.0×10^5 m^3/s day; the occurrence probability for the former vertex would decrease and that for the latter one would increase with the return period of the flood peak increases. The two vertexes for the conditional pdfs of flood duration appeared at around 11 and 15 days, respectively, with the associated occurrence probabilities respectively decreasing and increasing with the increase of the flood peak return period.

In engineering applications, the bivariate risk can be applied for actual flood management. Specifically, the bivariate risk of flood peak-volume can provide support for design of flood diversion area, and the bivariate risk of flood peak-duration can be considered as a reference for preparation of flood defense materials. Moreover, the pdfs of flood volume and duration conditional on different flood flows can help flood mitigation and control once a flood has occurred, in which the conditional pdfs of flood volume can provide useful information for flood diversion, and the conditional pdfs of flood duration can help decision maker arrange related people for river levee inspection.

Acknowledgement

This research was supported by the Natural Sciences Foundation (51190095, 51225904), the 111 Project (B14008) and the Natural Science and Engineering Research Council of Canada.

References

[1] Adamowski K. (1989). A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies. *Journal of Hydrology*, 108, 295-308

[2] AghaKouchak, A., A. Bárdossy, and E. Habib (2010a), Conditional simulation of remotely sensed rainfall data using a non-gaussian v- transformed copula, Advances in Water Resources, 33 (6), 624 – 634, doi:DOI:10.1016/j.advwatres.2010.02.010.

[3] AghaKouchak, A., A. Bárdossy, and E. Habib (2010b), Copula-based uncertainty modelling: application to multisensor precipitation estimates, Hydrological Processes, 24, 2111–2124, doi:10.1002/hyp.7632.

[4] Aghakouchak, A., G. Ciach, and E. Habib (2010c), Estimation of tail dependence coefficient in rainfall accumulation fields, Advances in Water Resources, 33(9), 1142-1149, doi:10.1016/j.advwatres.2010.07.003.

[5] Bobee B, Cavidas G, Ashkar F, Bernier J, Rasmussen P (1993). Towards a systematic approach to comparing distributions used in flood frequency analysis. *Journal of Hydrology* 142:121–136

[6] Chebana, F., S. Dabo-Niang, and T. B. M. J. Ouarda (2012), Exploratory functional flood frequency analysis and outlier detection, *Water Resources Research*, 48, W04514.

[7] Chebana, F., and T. B. M. J. Ouarda (2009), Index flood-based multivariate regional frequency analysis, Water Resources Research, 45, doi:10.1029/2008WR007490.

[8] Chen L., Singh V.P., Guo S., (2013). Measure of correlation between river flows using the copula-entropy method. *Journal of Hydrologic Engineering*, 18, 1591-1606.

[9] Chen, L., Ye, L.,Singh, V., Zhou, J., and Guo, S. (2012). Flood coincidence risk analysis using multivariate copula functions. Journal of Hydrologic Engineering, 17(6), 742-755

[10] Chen, L., Ye, L.,Singh, V., Zhou, J., and Guo, S. (2014). Determination of Input for Artificial Neural Networks for Flood Forecasting Using the Copula Entropy Method. Journal of

Hydrologic Engineering, doi,10.1061/(ASCE)HE.1943-5584.0000932,

[11] Conover W.J. (1999). *Practical nonparametric statistics*. Third Edition, John Wiley & Sons, Inc. New York, pp.428-433

[12] Cunnane C (1989). Statistical distributions for flood frequency analysis. WMO No. 718, WMP, Geneva

[13] De Michele C, Salvadori G (2003) A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-copulas. Journal of Geophysical Research, 108(D2), 4067, doi:10.1029/2002JD002534.

[14] De Michele C., G. Salvadori, R. Vezzoli, and S. Pecora. Multivariate assessment of droughts: frequency analysis and Dynamic Return Period. Water Resources Research, 49(10):6985–6994, 2013. doi: 10.1002/wrcr.20551.

[15] Diehl T., Potter K. W. Mixed flood distribution in Wisconsin. Hydrologic Frequency Modelling. Netherlands: D. Reidel Publishing Company, 1987, 213-226.

[16] Fan Y.R., Huang G.H., Huang K., Baetz B.W., (2015a). Planning Water Resources Allocation under Multiple Uncertainties through A Generalized Fuzzy Two-Stage Stochastic Programming Method. IEEE Transactions on Fuzzy Systems, doi 10.1109/TFUZZ.2014.2362550

[17] Fan Y.R., Huang G.H., Li Y.P., (2012). Robust interval linear programming for environmental decision making under uncertainty. *Engineering Optimization* 44 (11), 1321-1336

[18] Fan Y.R., Huang W.W., Huang G.H., Huang K., Li Y.P., Kong X.M., (2015b). Bivariate hydrologic risk analysis based on a coupled entropy-copula method for the Xiangxi River in the Three Gorges Reservoir area, China. *Theoretical and Applied Climatology*, doi: 10.1007/s00704-015-1505-z

[19] Fan Y.R., Huang W.W., Huang G.H., Huang K., Zhou X., (2015c). A PCM-based stochastic hydrological model for uncertainty quantification in watershed systems. *Stochastic Environmental Research and Risk Assessment*, 29, 915-927

[20] Fan Y.R., Huang W.W., Li. Y.P., Huang G.H., Huang K., (2015d). A coupled ensemble filtering and probabilistic collocation approach for uncertainty quantification of hydrological models. Journal of Hydrology, doi: 10.1016/j.jhydrol.2015.09.035

[21] Feng P., Li X. Bivariate frequency analysis of non-stationary flood time series based on Copula methods. Journal of Hydraulic Engineering, 2013, 44(10), 1137-1147, doi:10.13243/j.cnki.slxb.2013.10.005 (in Chinese with English abstract).

[22] Ganguli P., Reddy M.J., (2013). Probabilistic assessment of flood risks using trivariate copulas. *Theoretical and Applied Climatology*, 111(1-2), 341-360.

[23] Gebregiorgis A.S., Hossain F., (2012). Hydrological risk assessment of old dams: case study on Wilson dam of Tennessee River basin, *Journal of Hydrologic Engineering*, 17, 201-212.

[24] Genest C, Favre AC. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4): 347–368.

[25] Genest C., Rémillard B., Beaudoin D., (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics* 44:199-213.

[26] Ghizzoni, T., G. Roth, and R. Rudari (2010), Multivariate skew-t approach to the design of accumulation risk scenarios for the flooding hazard, Advances in Water Resources, 33 (10), 1187-1290, doi:10.1016/j.advwatres.2010.08.003.

[27] Ghizzoni T., G. Roth and R. Rudari (2012), Multisite flooding hazard assessment in the Upper Mississippi River, *Journal of Hydrology*, 412-413, 101-113

[28] Graler B., van den Berg M.J., Vandenberghe S., Petroselli A., Grimaldi S., De Baets B., Verhoest N.E.C., (2013). Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. *Hydrology and Earth System Sciences* 17: 1281–1296.

[29] Grimaldi S., Serinaldi F., (2006). Asymmetric copula in multivariate flood frequency analysis. *Advances in Water Resources*, 29, 1281-1296.

[30] Gringorten I.I. (1963). A plotting rule for extreme probability paper. *Journal Geophysical Research* 68:813-814.

[31] Han J.C., Huang G.H., Zhang H., Li Z., Li Y.P., (2014). Bayesian uncertainty analysis in hydrological modeling associated with watershed subdivision level: a case study of SLURP model applied to the Xiangxi River watershed, China. *Stochastic Environmental Research and Risk Assessment*, 28(4), 973-989.

[32] Hao Z. and V.-P. Singh (2012), Entropy-copula method for single-site monthly streamflow simulation, Water Resources Research, 48, W06604, DOI:10.1029/2011WR011419.

[33] Hastie T., Tibshirani R., Friedman J., (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2009

[34] He J., Mixture model based multivariate statistical analysis of multiply censored environmental data. Advances in Water Resources, 59, 15-24.

[35] Kao S.C., Govindaraju R.S., (2010). A copula-based joint deficit index for droughts. *Journal of Hydrology*, 380, 121-134.

[36] Karmakar S., Simonovic S.P., (2009). Bivariate flood frequency analysis. Part 2: a copula-based approach with mixed marginal distributions. *Journal of Flood Risk Management*, 2, 32-44.

[37] Kidson R., Richards K.S., (2005). Flood frequency analysis: assumption and alternatives. *Progress in Physical Geography*, 29(3), 392-410.

[38] Kite G.W., (1988). *Frequency and risk analysis in water resources*, Water Resources Publications, Littleton, CO

[39] Kong X.M., Huang G.H., Fan Y.R., Li Y.P., (2015). Maximum entropy-Gumbel-Hougaard copula method for simulation of monthly streamflow in Xiangxi river, China. *Stochastic Environmental Research and Risk Assessment*, 29 (3), 915-927

[40] Lee T, Salas JD. 2011. Copula-based stochastic simulation of hydrological data applied to Nile River flows. *Hydrology Research* 42(4): 318–330.
[41] Liu M., Xu X., Sun A.Y., Wang K., Liu W., Zhang X., (2014). Is southwestern China experiencing more frequent precipitation extremes? *Environmental Research Letter*, 9, 064002.
[42] Ma M., Song S., Ren L., Jiang S., Song J., (2013). Multivariate drought characteristics using trivariate Gaussian and Student copula. *Hydrological Processes*, 27, 1175-1190.
[43] Ma Z.Z., Wang Z.J., Xia T., Gippel C.J., Speed R., (2014). Hydrograph-based hydrologic alteration assessment and its application to the Yellow River. *Journal of Environmental Informatics*, 23(1), 1-13.
[44] Massey J.F., (1951) The Kolmogorov-Smirnov test for goodness of fit. Journal of American Statistical Association 46 (253):68-78.
[45] Miao D.Y., Huang W.W., Li Y.P., Yang Z.F., (2014). Planning Water Resources Systems under Uncertainty Using an Interval-Fuzzy De Novo Programming Method, Journal of Environmental Informatics, 24(1), 11-23.
[46] Nelsen R.B., (2006). *An Introduction to Copulas*. Springer: New York.
[47] Razali N.M., Wah Y.B. (2011), Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics* 2(1): 21-33
[48] Reddy J.M., Ganguli P., (2012). Bivariate flood frequency analysis of upper Godavari river flows using Archimedean copulas. Water Resources Management, 26(14), 3995-4018.
[49] Rosenblatt M (1952) Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* 23(3):470-472.
[50] Reyonlds D. (2007). Gaussian mixture models, Encyclopedia of Biometric Recognition
[51] Salvadori G, De Michele C (2004) Analytical calculation of storm volume statistics involving Pareto-like intensity-duration marginals. Geophysical Research Letters, 31, L04502, doi:10.1029/2003GL018767.
[52] Salvadori G. and C. De Michele. Multivariate real-time assessment of droughts via Copula-based multi-site Hazard Trajectories and Fans. Journal of Hydrology, 2015. (in press).
[53] Salvadori, G., De Michele, C. 2004. Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. Water Resources Research, 40, 12, W12511.
[54] Salvadori, G., De Michele, C. 2010.Multivariate multiparameter extreme value models and return periods: a copula approach. Water Resources Research, 46, 10, W10501.
[55] Salvadori G., De Michele C., Durante F., (2011). On the return period and design in a multivariate framework. *Hydrology and Earth System Sciences* 15: 3293–3305
[56] Salvadori G., De Michele C., Kottegoda N.T., Rosso R., (2007). *Extremes in Nature: an Approach using Copula*. Springer: Dordrencht; 292

[57] Salvadori, G., Durante F., De Michele, C. 2013.Multivariate return period calculation via survival functions. Water Resources Research, 49, 2308–2311

[58] Samaniego, L., A. Bárdossy, and R. Kumar (2010), Streamflow prediction in ungauged catchments using copula-based dissimilarity measures, Water Resources Research, 46, W02506, doi:10.1029/2008WR007695.

[59] Serinaldi, F. (2009), A multisite daily rainfall generator driven by bivariate copula-based mixed distributions, Journal of Geophysical Research - Atmospheres, 114, doi:10.1029/2008JD011258.

[60] Shiau JT (2006) Fitting drought duration and severity with two-dimensional copulas. Water Resources Management, 20, 795–815, 2006.

[61] Singh V. P., Sinclair R. A. Two-distribution method for flood frequency analysis. Journal of Hydraulics Division, 1972, 98(1), 29-44.

[62] Singh V. P., Wang S. X., Zhang L. Frequency analysis of nonidentically distributed hydrologic flood data. Journal of Hydrology, 2005, 307(1-4), 175-195.

[63] Sondergaard T. and Lermusiaux, P.F.J. (2013). Data assimilation with Gaussian mixture models using the dynamically orthogonal field equations. Part I: Theory and scheme. Monthly Weather Review, 141, 1761–1785

[64] Song S., Singh V.P., (2010). Meta-elliptical copulas for drought frequency analysis of periodic hydrologic data. *Stochastic Environmental Research and Risk Assessment*, 24(3), 425-444.

[65] Sraj M., Bezak N., Brilly M. (2014). Bivariate flood frequency analysis using the copula function: a case study of the Litija station on the Sava River. *Hydrological Processes*, DOI: 10.1002/hyp.10145

[66] Vandenberghe S, Verhoest NEC, De Baets B. (2010). Fitting bivariate copulas to the dependence structure between storm characteristics: a detailed analysis based on 105 year 10 min rainfall. *Water Resources Research*, 46. DOI: 10.1029/2009wr007857.

[67] Wang L.Z., Huang Y.F., Wang L., Wang G.Q., (2014). Pollutant flushing characterization of stormwater runoff and their correlation with land use in a rapidly urbanizing watershed. *Journal of Environmental Informatics*, 23(1), 37-43.

[68] Waylen P., Woo M. K. Prediction of annual floods generated by mixed processes. Water Resources Research, 1982, 18(4), 1283-1286.

[69] Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30(1):79.

[70] Wu Z.Y., Lu G.H., Liu Z.Y., Wang J.X., Xiao H., (2013). Trends of extreme flood events in the Pearl river basin during 1951-2010. *Advances in Climate Change Research*, 4(2), 110-116

[71] Xu H., Taylor R.G., Kinston D.G., Jiang T., Thompson J.R., Todd M.C., (2010). Hydrological modeling of River Xiangxi using SWAT2005: A comparison of model parameterizations using

station and gridded meteorological observations. *Quaternary International*, 226, 54-59.
[72] Yang W., Yang Z.F., (2014). Evaluation of Sustainable Environmental Flows Based on the Valuation of Ecosystem Services: a Case Study for the Baiyangdian Wetland, China. *Journal of Environmental Informatics*, 24(2), 90-100
[73] Yen B.C., (1970). Risk Analysis in design of engineering projects. *Journal of Hydrologic Engineering*, 96(4), 959-966.
[74] Yin H.F., Li C.A., (2001). Human impact on floods and flood disasters on the Yangtze River. *Geomorphology*, 41(2–3), 105–109
[75] Yue S., (2001). A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrological Processes*, 15(6), 1033-1045
[76] Yue S., T.B.M.J., Ouarda, B. Bobee, P. Legendre, P. Bruneau, (1999). The Gumbel mixed model for flood frequency analysis. Journal of Hydrology, 226, 88-100.
[77] Zhang L., (2005). Multivariate hydrological frequency analysis and risk mapping. Ph.D dissertation, Louisiana State University. Baton Rouge, Louisiana (USA).
[78] Zhang L., Singh V.P., (2006). Bivariate flood frequency analysis using the copula method. *Journal of Hydrologic Engineering*, 11, 150-164.
[79] Zhang L, Singh VP (2007) Bivariate rainfall frequency distributions using Archimedean copulas. *Journal of Hydrology* 332(1–2):93–109
[80] Zhang L, Singh VP (2012). Bivariate rainfall and runoff analysis using entropy and copula theories. *Entropy* 14:1784-1812.
[81] Zhang N., Li Y.P., Huang W.W., Liu J., An Inexact Two-Stage Water Quality Management Model for Supporting Sustainable Development in a Rural System. Journal of Environmental Informatics, 24(1), 52-64.
[82] Zhang Y., Li H., Yang D., (2012a). Simultaneous estimation of relative permeability and capillary pressure using ensemble-based history matching techniques. *Transport in porous media* 94 (1), 259-276.
[83] Zhang Y, Song C, Zheng S, Yang D. (2012b). Simultaneous estimation of relative permeability and capillary pressure for tight formations from displacement experiments. SPE paper 162663, *the SPE Canadian Unconventional Resources Conference*; 2012 October 30–November 1; Calgary, AB
[84] Zhang Y., Yang D., (2013). Simultaneous estimation of relative permeability and capillary pressure for tight formations using ensemble-based history matching method. *Computers & Fluids* 71, 446-460
[85] Zhang Y., Yang D., (2014). Estimation of relative permeability and capillary pressure for tight formations by assimilating field production data. *Inverse Problems in Science and Engineering* 22 (7), 1150-1175

Captions of Figures

Figure 1: the location of the studied watershed

Figure 2. Typical flood hydrograph showing flood flow characteristics

Figure 3. Variations of flood variables during the studied period

Figure 4. Comparison of different probability density estimates with observed frequency.

Figure 5. The copula estimation between flood peak and volume

Figure 6. The copula estimation between flood peak and duration

Figure 7. The copula estimation between flood volume and duration

Figure 8. The conditional cumulative distribution functions.

Figure 9. Comparison of the joint return periods.

Figure 10. Bivariate flood risk under different flood peak-volume scenarios

Figure 11. Bivariate flood risk under different flood peak-duration scenarios

Figure 12. Probability density functions of volume under different peak flow return periods.

Figure 13. Probability density functions of duration under different peak flow return periods.

Captions of Tables

Table 1. Basic properties of applied copulas

Table 2. Parameters of marginal distribution functions of flood variables

Table 3. Marginal distributions for flood variables through GMM

Table 4. Statistical test results for marginal distribution estimation

Table 5. Dependence evaluations among flood variables

Table 6. Statistical test results for the flood pairs of peak-volume, peak-duration and volume-duration

Table 7. Comparison of univariate, bivariate return periods for flood characteristics

Table 8. Statistical characteristics of the conditional PDFs of flood duration and volume under different peak flow return periods.

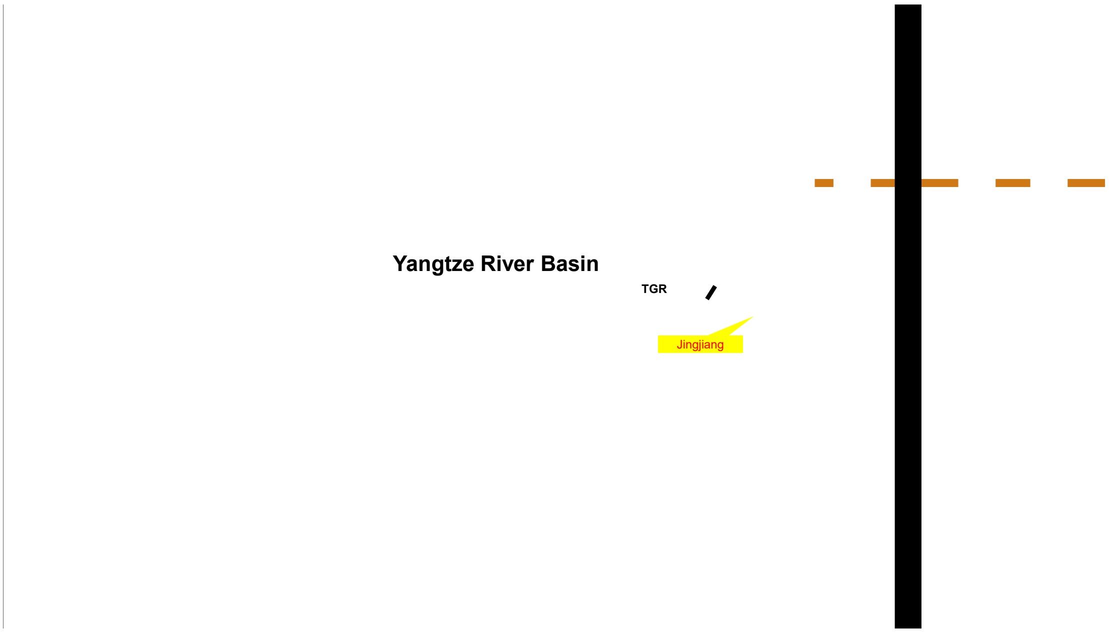

Figure 1: the location of the studied watershed

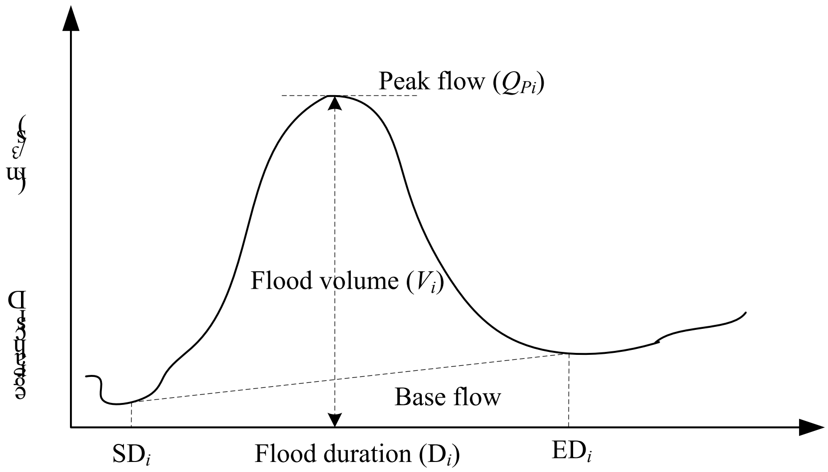

Figure 2. Typical flood hydrograph showing flood flow characteristics (adapted from Ganguli and Reddy, [22])

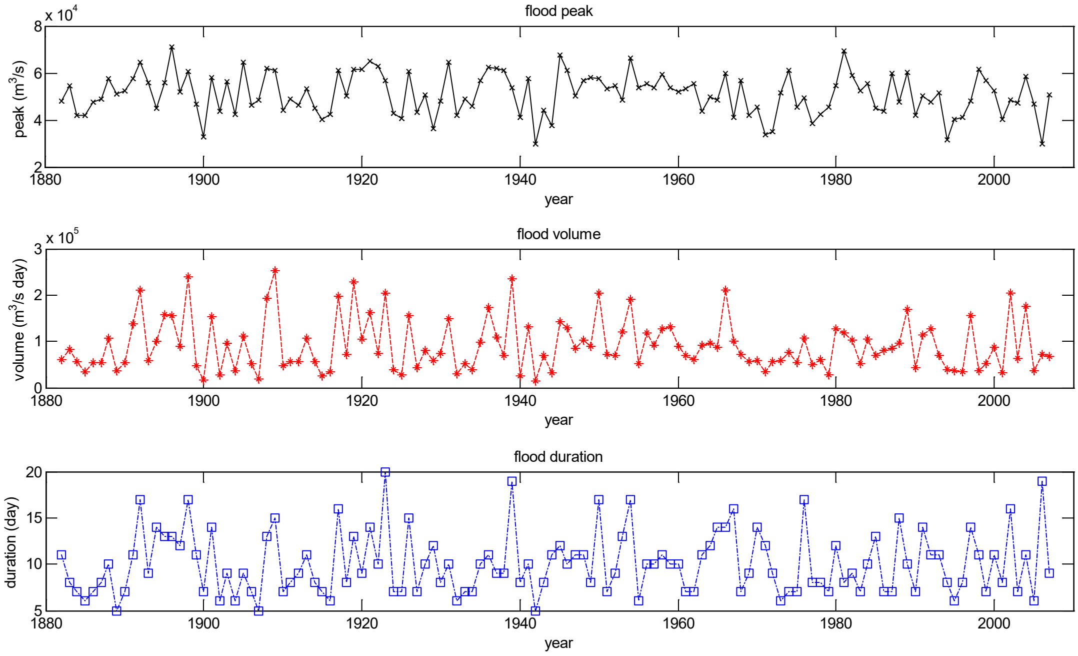

Figure 3. Variations of flood variables during the studied period

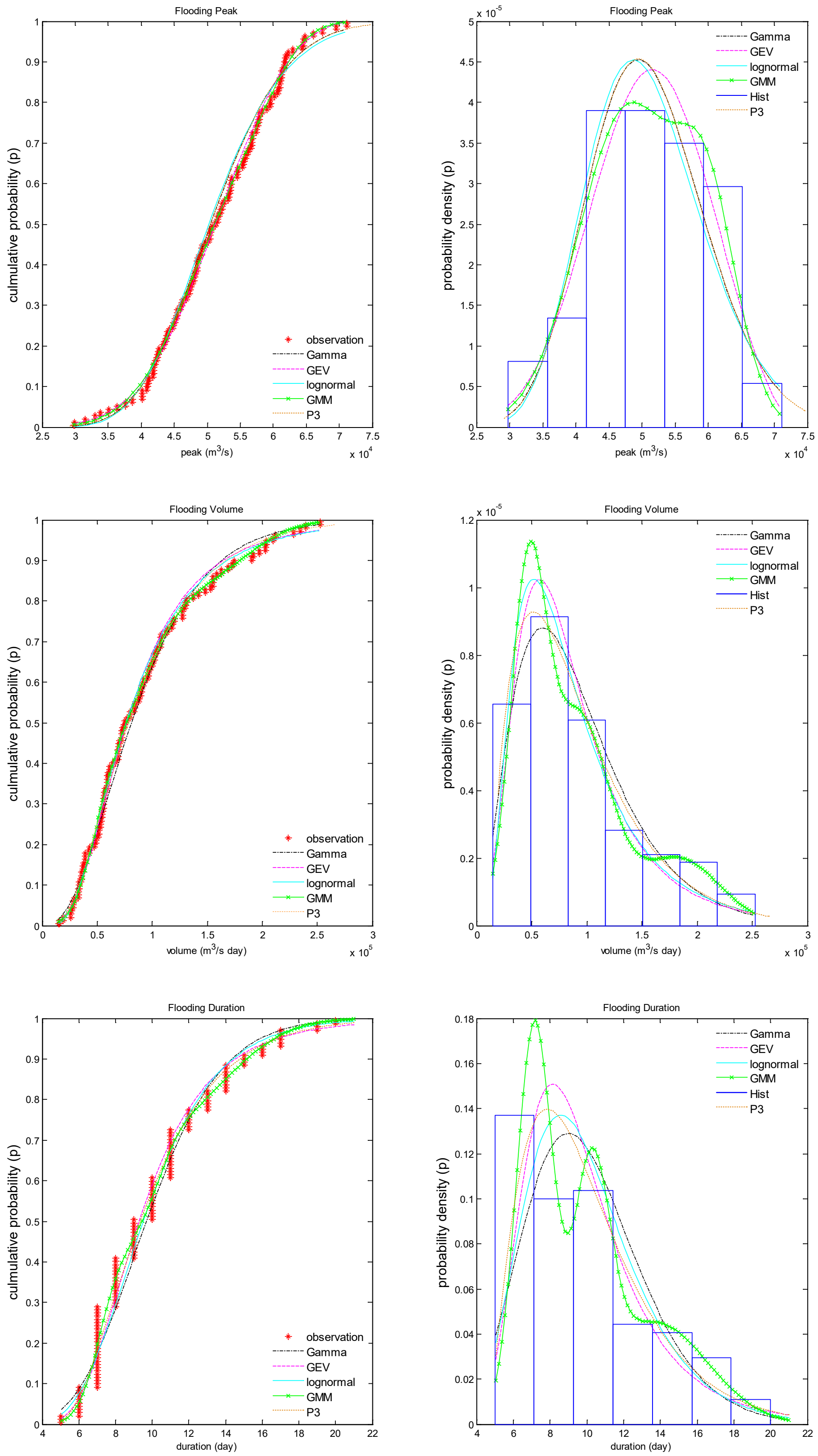

Figure 4. Comparison of different probability density estimates with observed frequency.

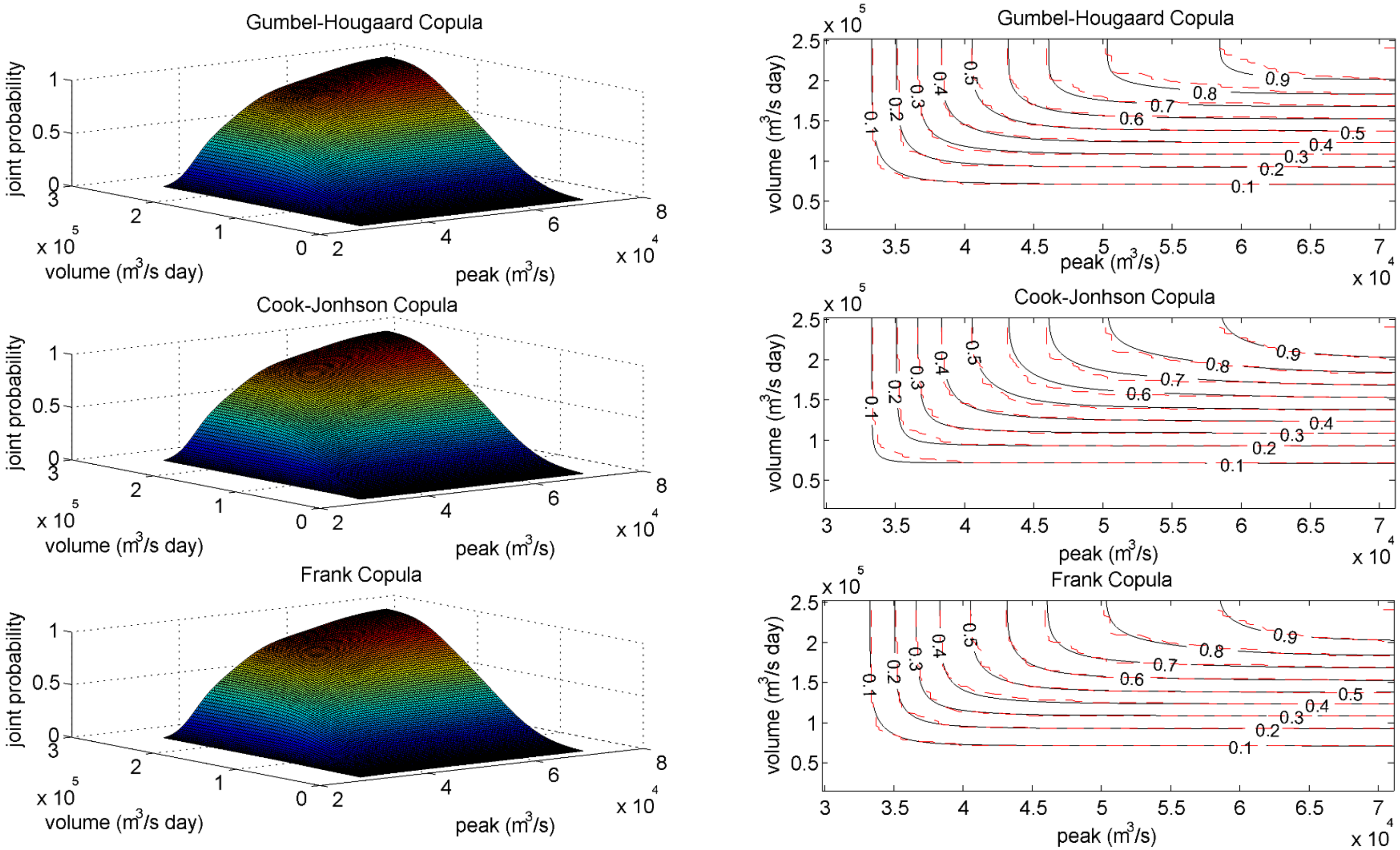

Figure 5. The copula estimation between flood peak and volume

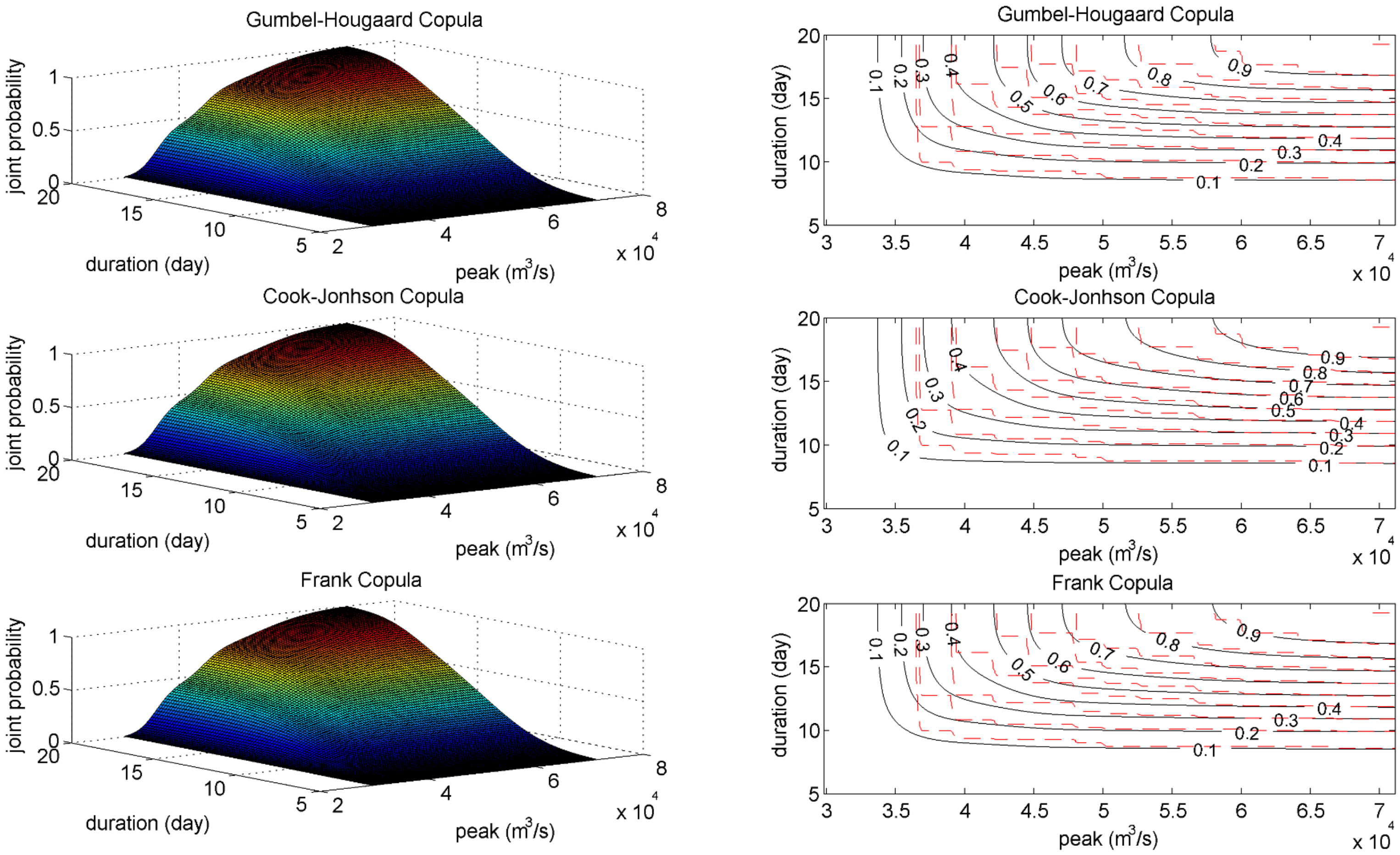

Figure 6. The copula estimation between flood peak and duration

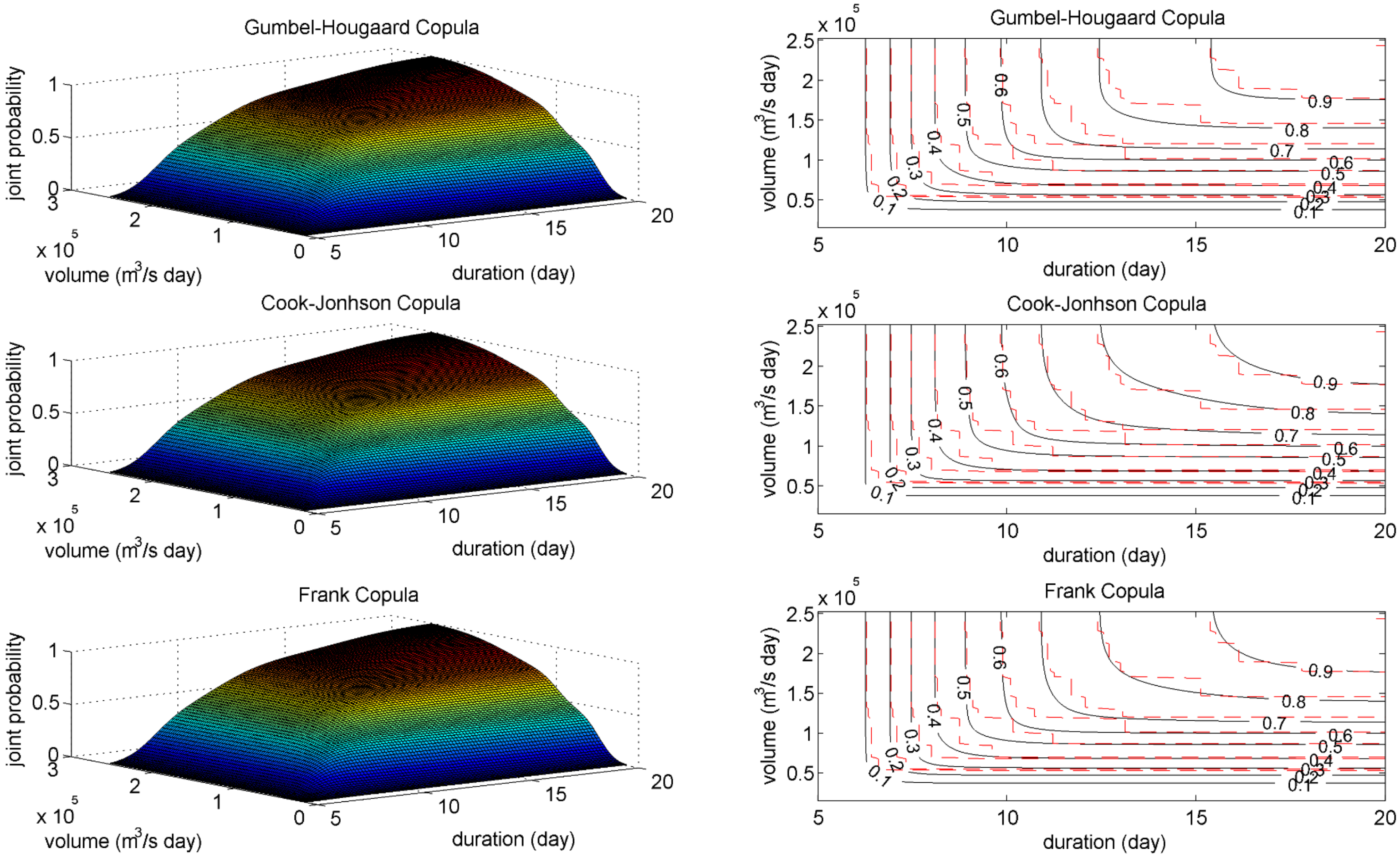

Figure 7. The copula estimation between flood volume and duration

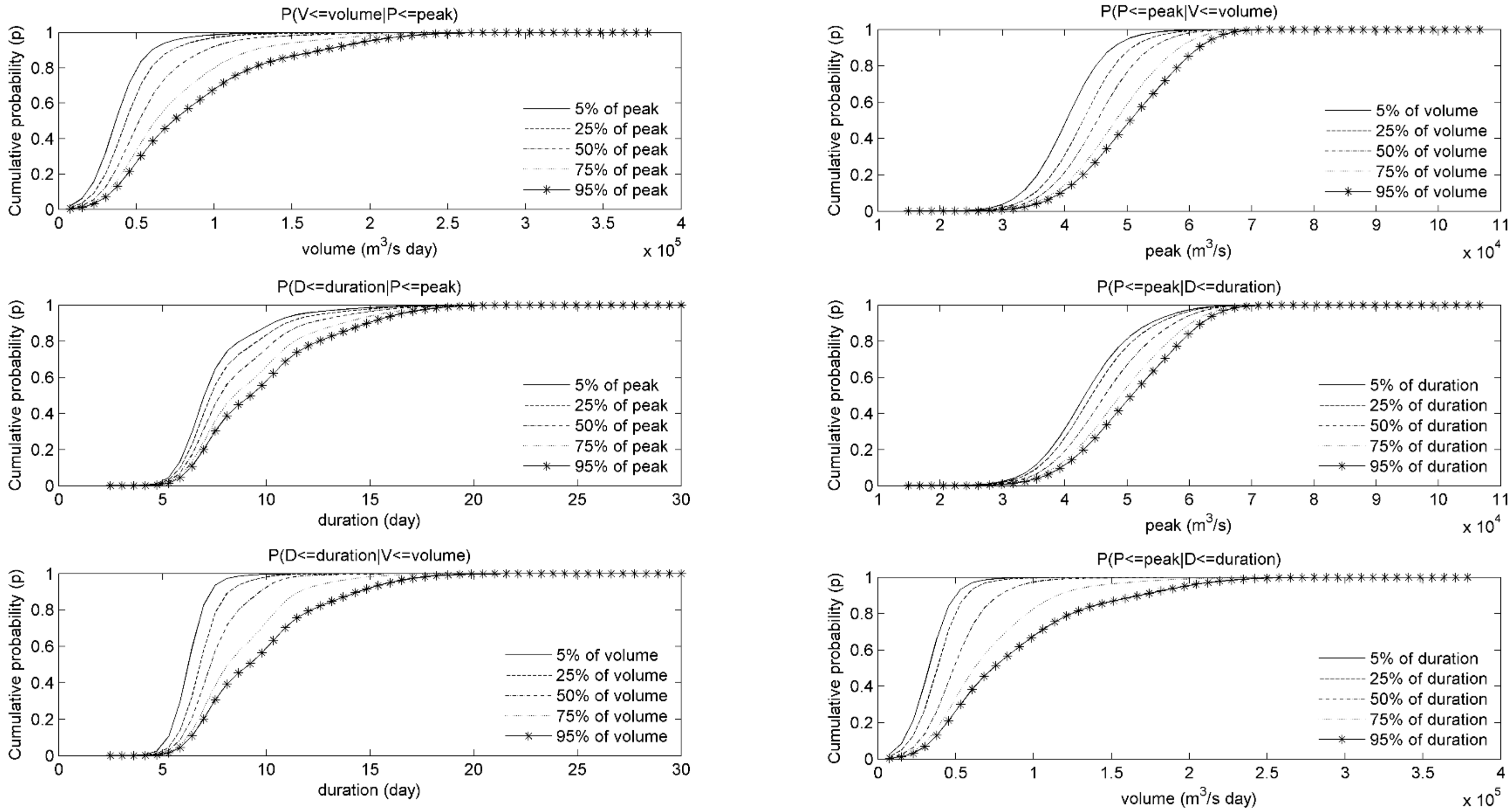

Figure 8. The conditional cumulative distribution functions.

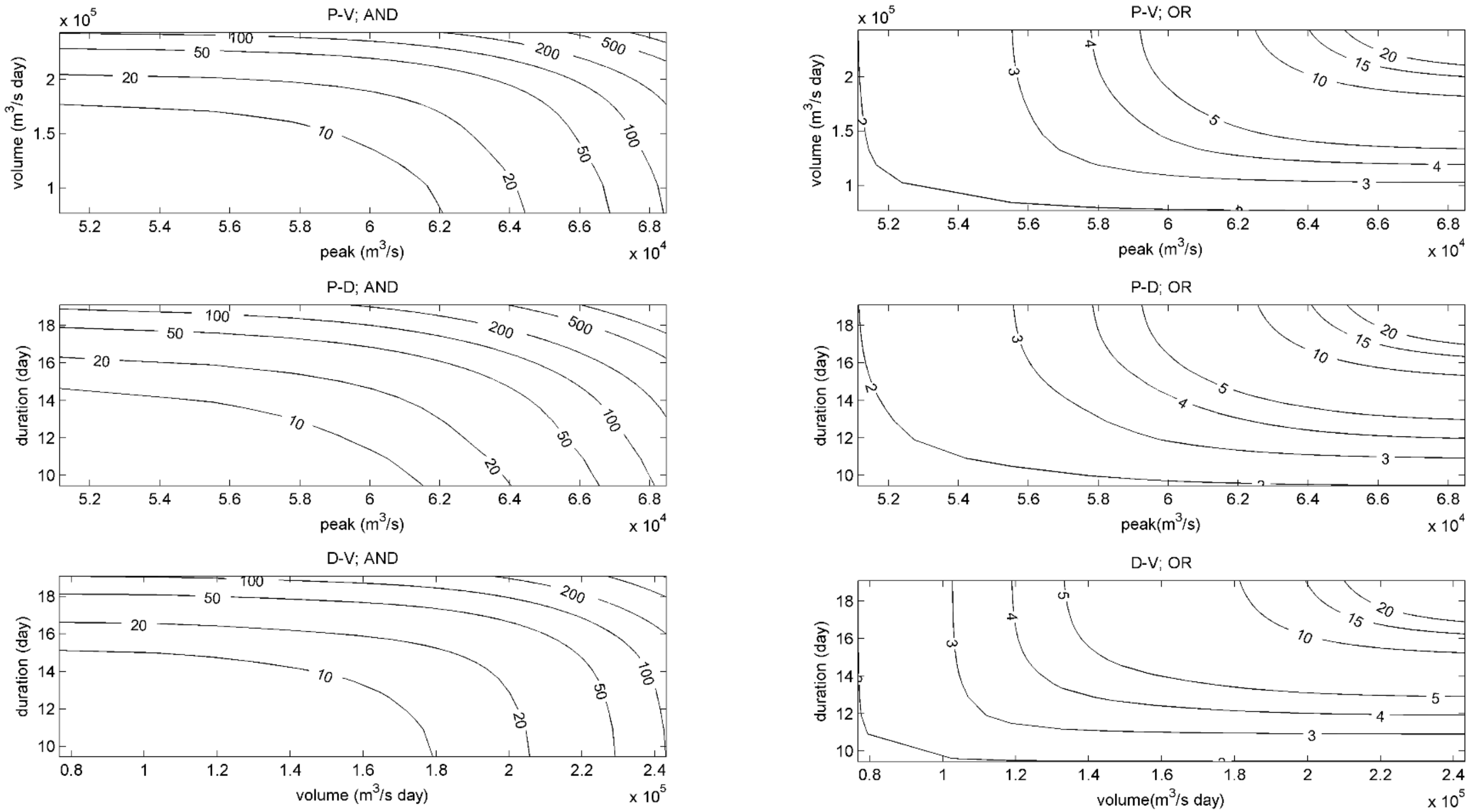

Figure 9. Comparison of the joint return periods.

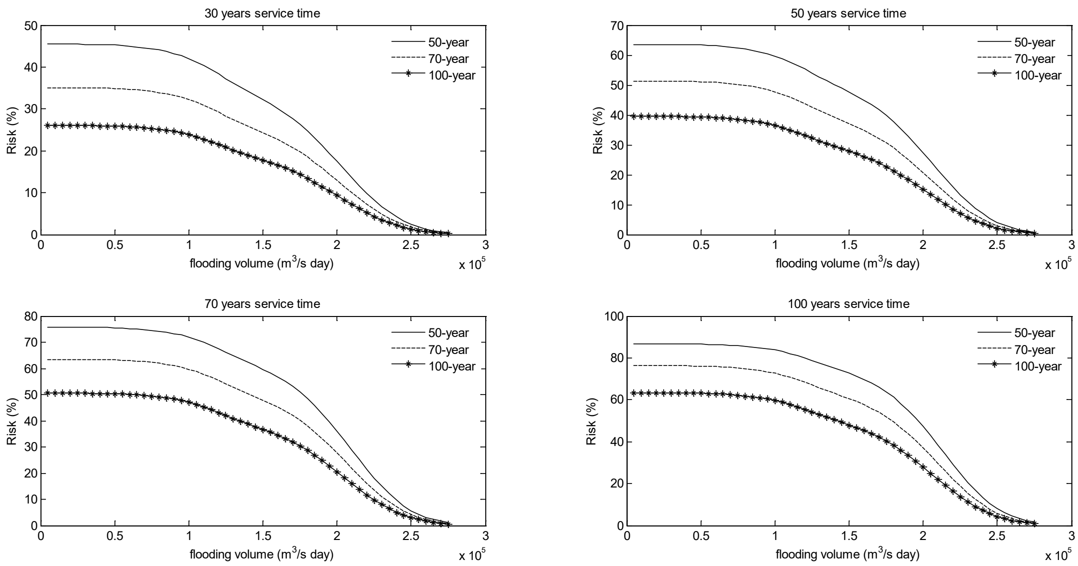

Figure 10. Bivariate flood risk under different flood peak-volume scenarios

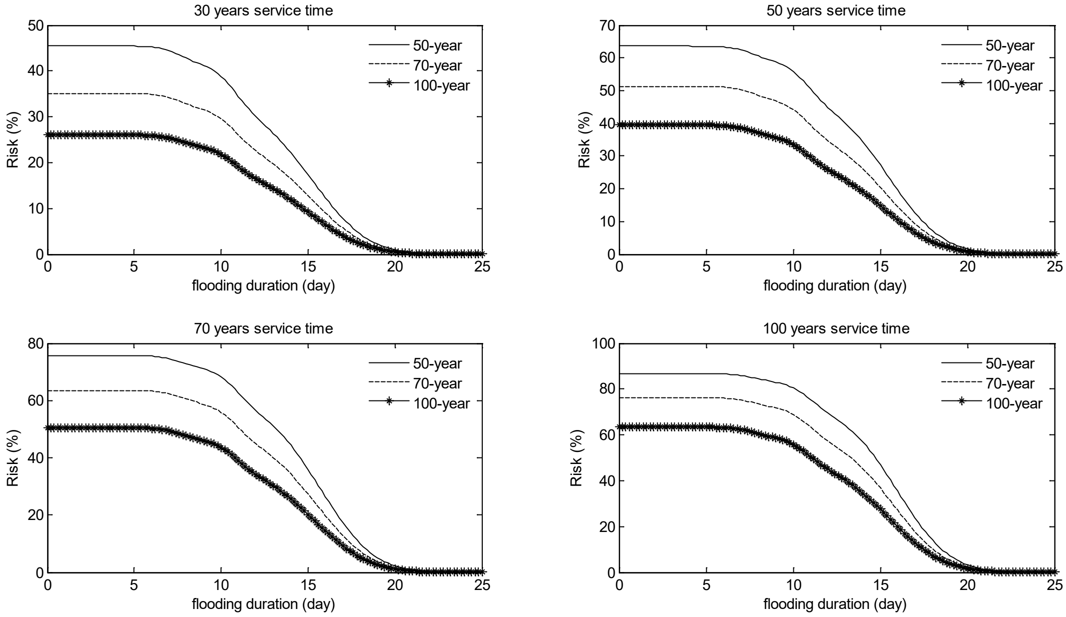

Figure 11. Bivariate flood risk under different flood peak-duration scenarios

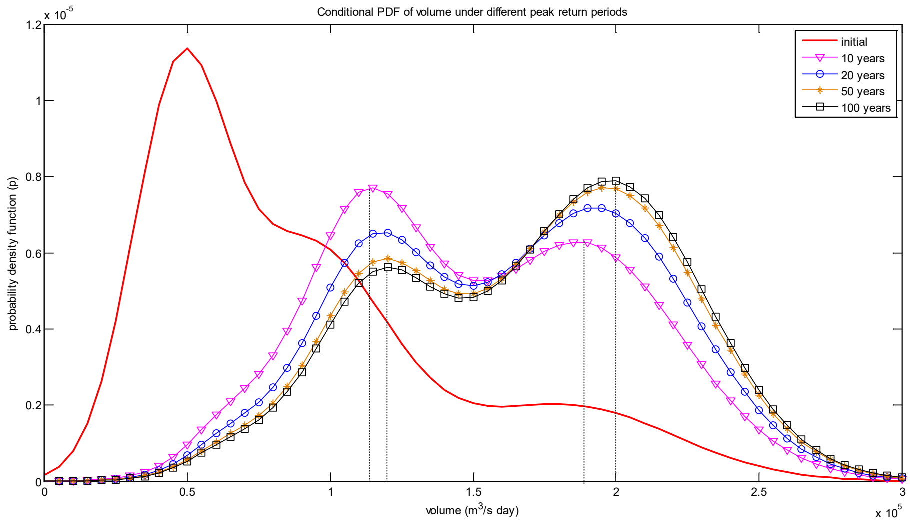

Figure 12. Probability density functions of volume under different peak flow return periods.

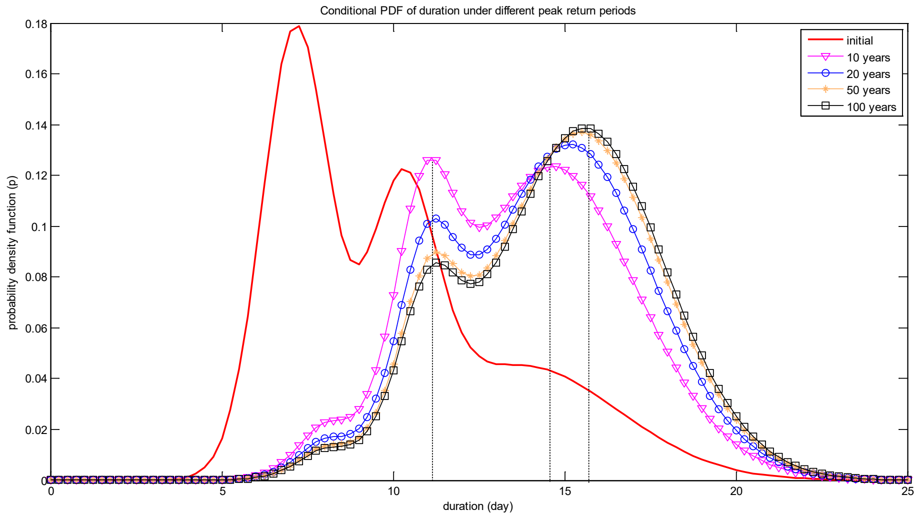

Figure 13. Probability density functions of duration under different peak flow return periods.

Table 1. Basic properties of applied copulas

Copula Name	Function[$C_\theta(u_1, u_2)$]	$\theta \in$	Generating function [$\phi(t)$]	$\tau = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)} dt$
Cook- Johnson	$[u_1^{-\theta} + u_2^{-\theta} - 1]^{-1/\theta}$	[-1, ∞)\{0}	$t^{-\theta} - 1$	$\frac{\theta}{\theta + 2}$
Gumbel-Hougaard	$\exp\{-[(-\ln u_1)^\theta + (-\ln u_2)^\theta]^{1/\theta}\}$	[1, ∞)	$(-\ln t)^\theta$	$1 - \theta^{-1}$
Frank	$-\frac{1}{\theta}\ln\{1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\}$	[-∞, ∞)\{0}	$\ln[\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}]$	$1 - \frac{4}{\theta}[D_1(-\theta) - 1]^*$

Note: * D_1 is the first order Debye function, and for any positive integer *k*, $D_k(x) = \frac{k}{x^k}\int_0^k \frac{t^k}{e^t - 1} dt$

Table 2. Parameters of marginal distribution functions of flood variables

Name	Probability density function		Parameters		
			Peak	Volume	Duration
Gamma	$\frac{1}{b^a\Gamma(a)}x^{a-1}e^{-\frac{x}{b}}$, $\Gamma(a)=\int_0^\infty u^{a-1}e^{-u}du$	a	32.76	2.9	7.90
		b	1557.5	31363.5	1.24
GEV	$(\frac{1}{\sigma})\exp(-(1+k\frac{(x-\mu)}{\sigma})^{-\frac{1}{k}})(1+k\frac{(x-\mu)}{\sigma})^{-1-\frac{1}{k}}$	k	-0.336	0.18	0.04
		μ	8899.6	36511.23	2.74
		σ	48177	63161.81	8.07
Lognormal	$\frac{1}{x\sigma_y\sqrt{2\pi}}\exp(-\frac{(y-\mu_y)}{2\sigma_y^2})$	μ_y	10.82	11.24	2.22
	y = log(x), x>0, $-\infty < \mu_y < \infty$, $\sigma_y > 0$	σ_y	0.18	0.62	0.36
Pearson Type III	$\frac{1}{b^a\Gamma(a)}(x-\alpha)^{a-1}e^{-\frac{x-\alpha}{b}}$, $\Gamma(a)=\int_0^\infty u^{a-1}e^{-u}du$	a	32.85	1.98	2.48
		b	1554.1	40002.4	2.22
		α	-21.91	12249.8	4.57

Table 3. Marginal distributions for flood variables through GMM

Flood Variables	Weights	Mean	Standard Deviation
Volume	0.4232	91987.0	27586.0
	0.1882	182387.1	37581.9
	0.3886	46691.3	16094.4
Peak	0.7436	47928	7551.2
	0.2564	60020	4480.5
Duration	0.2681	5.98	0.7
	0.4533	9.43	1.7
	0.2785	14.03	2.8

Table 4. Statistical test results for marginal distribution estimation

Flood variables	Marginal distribution	K-S test		RMSE	AIC
		T	P-value		
Peak	Gamma	0.0570	0.4253	0.0247	-401.0
	GEV	0.0362	0.7017	0.0176	-436.1
	Lognormal	0.0612	0.3740	0.0287	-384.6
	Pearson Type III	0.0610	0.7369	0.0246	-399.5
	GMM	**0.0380**	**0.6776**	**0.0119**	**-473.0**
Volume	Gamma	0.0611	0.3753	0.0266	-392.9
	GEV	0.0459	0.5705	0.0213	-415.2
	Lognormal	0.0390	0.6648	0.0174	-439.4
	Pearson Type III	0.0417	0.9808	0.0166	-442.5
	GMM	**0.0434**	**0.6049**	**0.0148**	**-443.1**
Duration	Gamma	0.1009	0.0716	0.0375	-355.3
	GEV	0.1023	0.0666	0.0403	-345.5
	Lognormal	0.0996	0.0769	0.0376	-355.1
	Pearson Type III	0.1113	0.0881	0.0378	-352.5
	GMM	**0.0703**	**0.2754**	**0.0297**	**-366.9**

Table 5. Dependence evaluations among flood variables

No.	Flood characteristics	Kendall's tau	Pearson's r
1	Peak – Volume	0.5509	0.6598
2	Volume – Duration	0.6756	0.7529
3	Peak - Duration	0.3561	0.2902

Table 6. Statistical test results for the flood pairs of peak-volume, peak-duration and volume-duration

Site	Copulas	Cramér von Mises statistic		RMSE
		S_n	P-value	
Peak - Volume	G-H	70.8224	0.3365	0.0168
	C-J	69.3597	0.3085	0.0199
	Frank	70.3817	0.3495	0.0149
Peak - Duration	G-H	66.2142	0.1215	0.0349
	C-J	64.9940	0.1165	0.0342
	Frank	65.7948	0.1325	0.0334
Volume-Duration	G-H	77.2530	0.1156	0.0302
	C-J	75.8958	0.1096	0.0305
	Frank	76.8450	0.1216	0.0291

Table 7. Comparison of univariate, bivariate return periods for flood characteristics (year)

T	Peak (m^3/s)	Volume (m^3/s day)	Duration (day)	T_{PV}^{AND}	T_{PD}^{AND}	T_{DV}^{AND}	T_{PV}^{OR}	T_{PD}^{OR}	T_{DV}^{OR}	$\bar{T}_{PV}$	$\bar{T}_{PD}$	$\bar{T}_{DV}$
5	59120.7	132815.5	12.9	8.5	11.4	7.2	3.5	3.2	3.8	6.4	7.7	5.9
10	62281.3	179597.7	15.2	24.5	36.6	19.2	6.3	5.8	6.8	16.2	22.0	13.8
20	64567.8	205920.7	16.6	78.8	127.6	58.0	11.5	10.8	12.1	46.8	70.8	36.7
50	66950.9	229240.1	18.1	419.6	726.4	290.3	26.6	25.9	27.4	227.3	380.1	163.1
100	68475.7	243091.0	19.1	1579.9	2809.4	1063.1	51.6	50.9	52.5	824.1	1438.0	566.2

Table 8. Statistical characteristics of the conditional PDFs of flood duration and volume under different peak flow return periods.

Flood variables	Index	initial	Return periods of peak flow (year)			
			10	20	50	100
Volume	Mean	91356.3	151750.3	161656.9	167571.8	169531.0
	Std	54840.2	51379.1	51441.9	51214.0	51095.1
	Kurtosis	0.4	-0.7	-0.8	-0.7	-0.7
	Skewness	1.0	0.1	-0.1	-0.2	-0.2
Duration	Mean	10.0	13.7	14.3	14.6	14.7
	Std	3.4	3.0	3.0	3.0	2.9
	Kurtosis	0.1	-0.4	-0.4	-0.3	-0.3
	Skewness	0.8	0.1	0.0	-0.1	-0.1