Chapter 3
Copula-Based Flood Frequency Analysis

3.1 Introduction

Flood frequency analysis is a constant concern in the hydrological practice. The sizing of bridges, culverts and other facilities, the design capacities of levees, spillways and other control structures, and reservoir operation or management depend upon the estimated magnitude of various design flood values (ASCE 1996). Nowadays, the general methodology based on the univariate distribution is to derive the fitted distribution representing the probability of an annual maximum flood being exceeded (USWRC 1981; MWR 2006).

As the duration of gauged record rarely exceeds 50 years, estimates corresponding to high return period obtained from the systematic data alone are subject to large sampling errors. Furthermore, the existence of a cyclic variation over periods longer than the duration of the records might well introduce further bias (Leese 1973; Stedinger and Cohn 1986; Guo and Cunnane 1991). Therefore, to overcome the problem of relatively short data series for frequency analysis, the need to augment the flow record with historical is widely acknowledged in the hydrological community. Several methods for incorporating historical information into flood frequency studies have been suggested, including historically weighted moments, maximum likelihood, probability weighted moments and L-moments (USWRC 1982; Guo and Cunnane 1991; Hosking 1995).

The hydrologic extreme values and critical thresholds derived from complex hydrological events for engineering design are usually obtained from single site characteristics (e.g., annual maximum peak discharge). Therefore, conventional hydrological frequency analysis has also mainly focused on one characteristic value and univariate distributions that cannot provide a complete description of hydrologic events with multi-characteristics. Many hydrological frequency problems, such as design flood hydrograph that includes flood peak and flood volumes, should be solved by the multivariate distributions (Dupuis 2007; Xiao et al. 2008, 2009).

L. Chen and S. Guo, *Copulas and Its Application in Hydrology and Water Resources*, Springer Water, https://doi.org/10.1007/978-981-13-0574-0_3

In this chapter, the multivariate frequency analysis has been carried out. One of the main difficulties in the multivariate quantile estimation is how to choose the proper combinations of design values of the concerned random variables for a given multivariate return period of hydrologic structure design. Take the bivariate case (peak discharge Q and flood volume W) as an example. The combinations can differ greatly regarding their values: moving along the multivariate quantile curve to an asymptote, one of the two variables will approach its marginal value, while the other tends to increase indefinitely (for unbounded random variables). Chebana and Ouarda (2011) proposed the decomposition of the level curve into a naive part (tail) and the proper part (central); they assumed that the naive part was composed of two segments starting at the end of each extremity of the proper part. Salvadori et al. (2011) introduced two basic design realizations, i.e., component-wise excess design realization and most-likely design realization. Li et al. (2016) used the conditional expectation combination method to derive the quantiles of flood peak and 7-day volume under different JRPs, and they found that the bivariate design values have smaller flood volume and larger flood peak than bivariate equivalent frequency combination results.

3.2 Annual Maximum Flood Frequency Analysis Based on Copula

Annual maximum (AM) flood series can be characterized by flood occurrence dates and flood magnitudes. The marginal distribution of flood occurrence dates, peak discharges, and flood volumes are established.

3.2.1 Margin Distribution of AM Flood Occurrence Dates

The AM flood occurrence dates can be described by the directional statistics (DS) method. The date firstly should be converted to the angle of a circle by

$$\alpha_i = D_i \frac{2\pi}{L} \quad 0 \leq \alpha_i \leq 2\pi \tag{3.1}$$

where L is the length of flood season; D_i is the flood occurrence date.

The x and y coordinates of the flood dates described by the angles is determined by

$$(a_i, b_i) = (\cos\ \alpha_i, \sin\ \alpha_i) \tag{3.2}$$

$$\bar{a} = \sum_{i=1}^{n} \cos x_i / n \tag{3.3}$$

$$\bar{a} = \sum_{i=1}^{n} \sin x_i / n \tag{3.4}$$

where n is the sample size.

The mean direction of the circular data (denoted by $\bar{\theta}$) is estimated by

$$\bar{\theta} = \begin{cases} \arctan \bar{b}/\bar{a} & \bar{a} > 0, \quad \bar{\mathrm{b}} > 0 \\ 2\pi + \arctan \bar{b}/\bar{a} & \bar{a} > 0, \quad \bar{\mathrm{b}} < 0 \\ \pi + \arctan \bar{b}/\bar{a} & \bar{a} < 0 \\ \pi/2 & \bar{a} = 0, \quad \bar{\mathrm{b}} > 0 \\ 3\pi/2 & \bar{a} = 0, \quad \bar{\mathrm{b}} < 0 \\ unkown & \bar{a} = 0, \quad \bar{\mathrm{b}} = 0 \end{cases} \tag{3.5}$$

A measure of the variability of the flood occurrences about the mean date is determined by defining the mean resultant vector as:

$$\bar{r} = \sqrt{\bar{a}^2 + \bar{b}^2} \quad 0 \le r \le 1 \tag{3.6}$$

where $\bar{r}$ describes the dispersion measure (Black and Werritty 1997).

Since the distribution of dates is on a circle, rather than along a line, the use of the normal distribution is no longer appropriate. Therefore, the von Mises distribution is introduced and used to describe seasonal data with a single peak.

Fisher (1993) termed the von Mises distribution as the "natural" analog of the normal distribution for seasonal data with a single peak. It is the most commonly used and has some similar characteristics to the normal distribution (Mardia 1972). The probability density function of von Mises distribution is given by:

$$f(x) = \frac{1}{2\pi I_0(\kappa)} \exp[\kappa \cos(x - \mu)] \; 0 \le x \le 2\pi, \; 0 \le \mu \le 2\pi, \; \kappa \ge 0 \tag{3.7}$$

It is symmetric and unimodal, with a mean direction at μ and the dispersion given by a concentration parameter $\kappa = A^{-1}(r) \bullet A^{-1}(r)$ is the inverse function of $A \bullet I_0(\kappa)$ is the modified Bessel function of order zero. For large values of κ, the distribution is concentrated around the mean. When $\kappa = 0$, the density gives the uniform distribution on [0, 2].

3.2.2 Margin Distribution of AM Flood Peaks and Volumes

For the AM flood series, the Pearson type III (P-III) has been recommended by MWR (2006) as a uniform procedure for flood frequency analysis in China. The PDF of the P-III distribution is given in Table 1.1 of Chap. 1.

3.2.3 Bivariate Distribution of AM Flood Occurrence Dates and Magnitudes

For estimating the design flood, the bivariate joint distributions of AM flood occurrence dates and magnitudes (or flood peaks and volumes) need to be built. Every joint distribution can be written regarding a copula and its univariate marginal distributions. The copula is a function that links univariate marginal distribution functions to construct a multivariate distribution function. The definition and establishment of copulas can be seen in Chap. 2. The Gumbel copula is used to establish the joint distribution in this section.

3.2.4 Case Study

As an illustrative example, the Geheyan reservoir is selected as a case study. The Geheyan reservoir is a key control and multi-purpose water resources engineering project in the Qingjiang Basin, which is one of the main tributaries of the Yangtze River in China. The basin encompasses an area of 17,000 km^2 with the annual average rainfall 1500 mm. The annual average discharge and runoff at dam site are 393 m^3/s and 124 × 108 m^3 (from 1951 to 2005), respectively. The flood season lasts for five months from 1 May to 30 September (153 days).

3.2.4.1 Computation of Empirical Probability

The empirical probabilities can be computed by the Gringorten plotting–position formula

$$P(j) = \frac{j - 0.44}{n + 0.12} \tag{3.8}$$

where $P(j)$ is the cumulative frequency, indicating the probability that a given value is less than the jth smallest observation in the data set of n observations.

Observed joint probabilities are computed based on the same principle as in the case of a single variable. A two-dimensional table is constructed first in which the variables X and Y are arranged in ascending order. The joint cumulative frequency (non-exceedance joint probability) is then given by (Yue et al. 1999):

$$F(t_k, q_j) = P(X \leq t_k, Y \leq q_j) = \frac{\sum_{m=1}^{k} \sum_{l=1}^{j} n_{m,l} - 0.44}{n + 0.12} \tag{3.9}$$

3.2.4.2 Evaluation Criteria

A Chi-Square Goodness-of-fit test (χ^2), mean *Rbias* and *RRMSE* are selected to test the fitting descriptive ability of flood frequency curve, which can be calculated by

$$\chi^2 = \sum_{i=1}^{n} \left(P_{the}(i) - P_{emp}(i)\right)^2 / P_{emp}(i) \tag{3.10}$$

$$Rbias = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{Q}(i) - Q(i)\right)/Q(i) \tag{3.11}$$

$$RRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\hat{Q}(i) - Q(i)}{Q(i)}\right)^2} \tag{3.12}$$

where P_{the} and P_{emp} are the theoretical and empirical frequencies; and $\hat{Q}(i)$ and $Q(i)$ are the estimated and observed values, respectively.

3.2.4.3 Conditional Probability

The parameters of Von Mises and P-III distribution are estimated by L-moments method for given AM flood series of occurrence dates, peak discharges or volumes, respectively. A Chi-Square Goodness-of-fit test is performed to test the assumption, H_0, that the flood occurrence dates and magnitudes follow the Von Mises and P-III distributions. Table 3.1 shows that the assumption cannot be rejected at the 0.5% significance level. It is shown that the values of *Rbias* and *RRMSE* are very small, which mean that the marginal distribution can fit data set very well.

Table 3.2 lists the conditional probability of $P(X > x_p|Y > y_{1\%})$ given x_p. Under the condition of annual maximum flood magnitude $Y > y_{1\%}$, the probability corresponding occurrence date after May 27 is 98.45%, the probability of annual maximum flood occurred during May 27 to 29 is (98.45−29.86%) = 68.59%, and during July 18 to 29 is (81.16−75.29%) = 5.87%.

Table 3.1 The goodness of fit and χ^2 test statistics

Index	Rbias	RRMSE	χ^2	c	$\chi^2_{0.995}(N-c-1)$
Von Mises	−4.378	0.982	0.253	2	82.001
P-III	0.254	0.327	0.903	3	80.747
Bivariate			4.400	6	76.969

Table 3.2 Conditional probability of X given $Y > y_{1\%}$

P (%)	0.01	0.1	1	10	20	30	40	50	70	90	99
x_p (Arc)	6.28	6.27	6.09	4.48	3.71	3.27	2.92	2.62	2.02	1.11	0.17
Dates	9/29	9/28	9/25	8/17	7/29	7/18	7/10	7/2	6/18	5/27	5/4
CP (%)	0.84	6.17	29.86	65.42	75.29	81.16	85.47	88.93	94.36	98.45	99.87

Note CP means the conditional probabilities $P(X > x_p|Y > y_{1\%})$

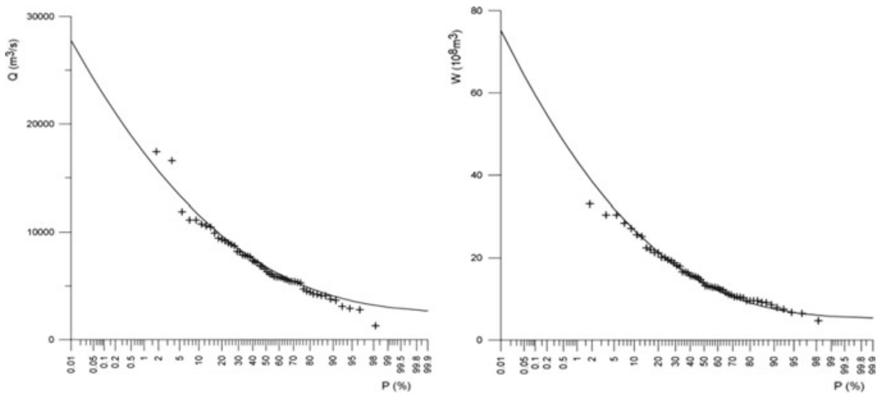
3.2.4.4 Fitting Marginal Distributions

The marginal distribution frequency curves of flood peaks and 7-day flood volumes are shown in Fig. 3.1, in which the line represents the theoretical distribution, and the crossing represents the empirical probabilities. Figure 3.1 indicates that these theoretical distributions can fit the observed data reasonably well.

The Gumbel copula is used to model the dependence between the extreme maximum annual flood peaks and 7-day flood volume in this study. The probability plot of joint distribution is shown in Fig. 3.2, in which the Gumbel copula can fit the empirical bivariate distribution very well.

3.3 Copula-Based Flood Frequency Considering Historical Information

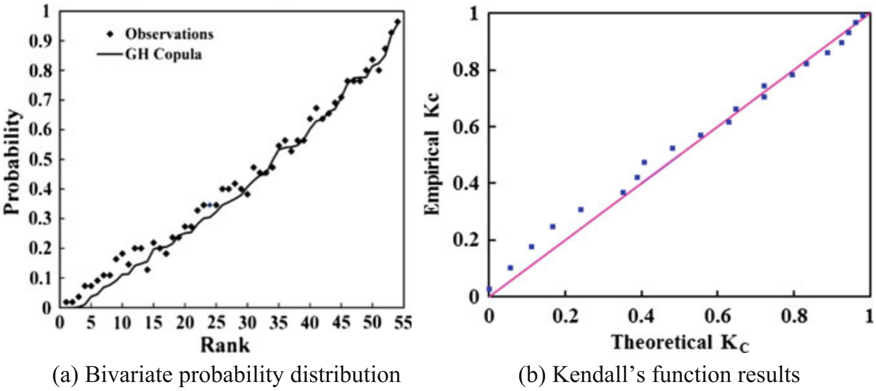
Flood events consist of flood peaks and flood volumes that are mutually correlated and need to be described by multivariate analysis methods, of which the copula functions are most desirable ones. Until now, the multivariate flood frequency



(a) Flood peak

(b) 7-day flood volume

Fig. 3.1 Probability curves of flood peak and 7-day flood volume



(a) Bivariate probability distribution (b) Kendall's function results

Fig. 3.2 Comparison of observed and theoretical bivariate probability distribution

analysis methods based on copulas doesn't consider the historical flood information. This may underestimate or overestimate the flood quantiles or conditional probabilities corresponding to high return periods, especially when the length of gauged record data series is relatively short.

3.3.1 Maximum Likelihood Estimation for Censored Samples

In certain sampling situations, the exact values of a proportion of the sample are unknown, although their range may be specified. Usually, the range consists of all points above or below a threshold level. Under these circumstances, the sample is said to be censored. Censored samples occur, for example, when instruments are not calibrated for measurements above or below a certain level. Both historical data and recent flood data (i.e., systematic record) may give rise to censored samples, but because the censoring is generally above a threshold in the former and below in the latter, they must be treated separately (Leese 1973).

Censored-sample maximum likelihood estimators were initially developed by Hald (1949) and Cohen (1976) for the normal and lognormal distributions. They were subsequently adapted by Leese (1973), Condie and Lee (1982), and Stedinger and Cohn (1986) for common case in hydrology where one have both a censored-sample historical flood record and also a systematic gaged record. The maximum likelihood estimation method for type-I censoring is described as follows.

In the annual maximum flood series of Fig. 3.3, there is a total of g known floods. Of these, k is known to be the k largest in the period of n years. The n year period contains within it a systematic record (recently gauged data) of s years ($s \leq n$) length. Of the k largest floods, c occurred during the systematic record

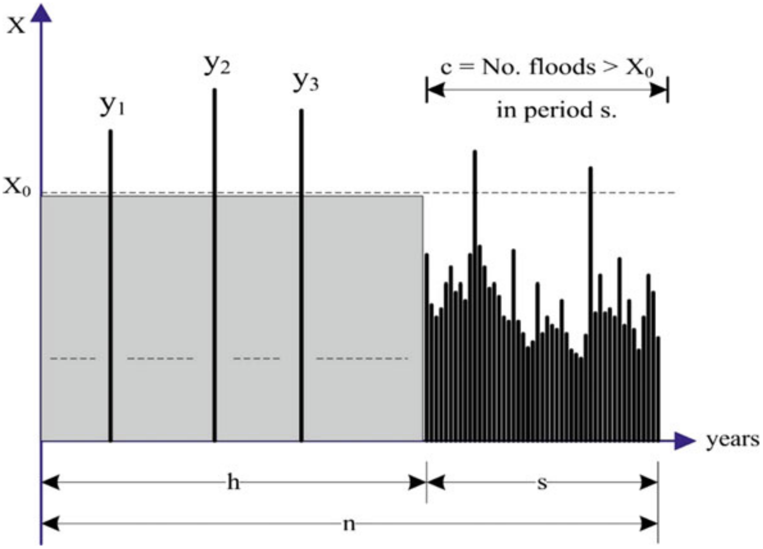


Fig. 3.3 Sketch of the annual maximum flood series when historical floods are available. Notations: s—the length of the systematic record; h—the length of the pre-gauging period; y_1, y_2, y_3—historical flood events; X_0—perception threshold

($c \leq k$ and $c < s$, and also $g = s+k-c$). Assume a fixed threshold X_0 exceeded by the k largest floods and not exceeded by any of the remaining $n–k$ floods, recorded or not (i.e., the k values which exceed X_0 form a type I censored sample). It is also noted that the m ($m = k-c$) floods in the pre-gauging period h ($h = n-s$) are known as they are included in the k values which exceed X_0, and it is assumed that no other floods exceeded the threshold during that period.

Let f_X and F_X denote the probability density function (PDF) and the cumulative distribution function (CDF) of variable X, respectively. The resulting likelihood function for the whole sample of s+m known and $h-m$ unknown values is given by (Leese 1973; Condie 1986; Stedinger and Cohn 1986; Guo and Cunnane 1991)

$$l(\boldsymbol{\alpha}) = \prod_{i=1}^{s+m} f_X(x_i)[\int_{-\infty}^{X_0} f_X(x)dx]^{h-m} \tag{3.13}$$

where α is the parameter vector of f_X and F_X.

Since c flood events exceeding the perception threshold X_0 occur among the systematic data (analogously to the sketch in Fig. 3.3), the c events are virtually removed from the period s and are treated as historical data (Bayliss and Reed 2001). Then, Eq. 3.13 can be expressed as

$$l(\boldsymbol{\alpha}) = \prod_{i=1}^{s-c} f_X(x_i) \prod_{j=1}^{k} f_X(y_j) [\int_{-\infty}^{X_0} f_X(x)dx]^{h-m} \tag{3.14}$$

where $x_i(i = 1, 2 \ldots s - c)$ denotes the systematic data less than the threshold X_0 and $y_i(j = 1, 2 \ldots k)$ denotes the k $(k = m+c)$ largest floods exceeding the threshold X_0; $\prod_{i=1}^{s-c} f_X(x_i)$ and $\prod_{j=1}^{k} f_X(y_j)$ are the likelihood functions of $s-c$ systematic records and the k largest floods, respectively; and $[\int_{-\infty}^{X_0} f_X(x)dx]^{h-m}$ represents the likelihood function for the $h-m$ unknown values, which has been defined and applied by Leese (1973), Condie (1986), Stedinger and Cohn (1986), and Guo and Cunnane (1991).

The log-likelihood function for the univariate distribution can be expressed as

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{s-c} \log f_X(x_i) + \sum_{j=1}^{k} \log f_X(y_j) + (h-m) \log F_X(X_0) \tag{3.15}$$

The maximum likelihood estimates are those values of α that maximize Eq. 3.15.

3.3.2 *Bivariate Flood Frequency Analysis with Historical Information*

The conventional flood frequency analysis incorporation with historical information is based on univariate distribution. To overcome the shortcomings of univariate frequency analysis, a multivariate copula-based flood frequency analysis model that considers historical information was proposed and discussed by Li et al. (2013). As the historic flood events occurred hundreds of years ago, the durations of them are hard to measure or investigate. There is no publication or any gauged record related to the duration samples of historical floods. Besides, the perception threshold of flood duration is also difficult to fix for maximum likelihood estimation. Thus, only the distribution of flood peak and volume with historical information is studied.

3.3.3 *Inference Function for Margins Method*

In classical statistics, the inference function for margins (IFM) method was first defined as a terminology by McLeish and Small (1988). Compared with other estimation methods, the IFM method is the preferred fully parametric method for multidimensional parameter estimation because it is close to maximum likelihood (ML) in approach and is easier to implement (Joe and Xu 1996; Joe 1997). Comparisons of various types have been made in Xu (1996) for some multivariate models which suggest that the IFM method is highly efficient compared to maximum likelihood. Similar comparisons have also been made by Joe (1997), (2005) and the derived conclusions are: (1) the ML estimation is much more time-consuming than IFM method, (2) the IFM method allows one to do inference and modelling starting with univariate and lower-dimensional margins, (3) there is some robustness against misspecification of the dependence structure and also there should be more robustness against outliers or perturbations of the data, compared with the ML method; and (4) the IFM rather than the ML method avoids the sparseness problem to a certain degree, especially if parameters can all be estimated from univariate and bivariate likelihoods. Therefore, the IFM method is selected and described briefly as follows:

Under the assumption that the marginal distributions are continuous with probability density functions $f_X(x;\boldsymbol{\alpha_1})$ and $f_Y(y;\boldsymbol{\alpha_2})$, the joint PDF then becomes

$$f_{X,Y}(x,y;\boldsymbol{\alpha_1},\boldsymbol{\alpha_2},\theta) = c_\theta[F_X(x;\boldsymbol{\alpha_1}),F_Y(y;\boldsymbol{\alpha_2})]f_X(x;\boldsymbol{\alpha_1})f_Y(y;\boldsymbol{\alpha_2}) \tag{3.16}$$

where F_X and F_Y are univariate CDFs with respective parameter vectors $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$, and c_θ is the density of C_θ parametrized by a parameter θ, defined as

$$c_\theta(u,v) = \frac{\partial^2 C_\theta(u,v)}{\partial u \partial v} \tag{3.17}$$

For the observed bivariate series (x_1, y_1),…, (x_s, y_s) with a sample size s, we can consider the two log-likelihood functions for the univariate marginal distribution, i.e.

$$L_1(\boldsymbol{\alpha_1}) = \sum_{i=1}^{s} \log f_X(x_i;\boldsymbol{\alpha_1}) \tag{3.18a}$$

$$L_2(\boldsymbol{\alpha}_2) = \sum_{i=1}^{s} \log f_Y(y_i;\boldsymbol{\alpha}_2) \tag{318b}$$

and the log-likelihood function for the joint distribution,

$$L(\theta,\boldsymbol{\alpha_1},\boldsymbol{\alpha_2}) = \sum_{i=1}^{s} \log f_{X,Y}(x_i,y_i;\boldsymbol{\alpha_1},\boldsymbol{\alpha_2},\theta) \tag{3.19}$$

The IFM method consists of two separate optimizations of univariate likelihoods, followed by an optimization of multivariate likelihood as a function of the dependence parameter vector. More specifically,

(a) The log-likelihoods $L_1(\boldsymbol{\alpha}_1)$ and $L_2(\boldsymbol{\alpha}_2)$ of the two univariate marginal distributions are separately maximized by Eq. 3.18a, 318b to get estimates $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$;

(b) The function $L(\theta, \hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2)$ is maximized over θ to get $\hat{\theta}$ in Eq. 3.19.

That is, under regularity conditions, $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, \hat{\theta})$ is the solution of

$$(\partial L_1/\partial \boldsymbol{\alpha}_1, \partial L_2/\partial \boldsymbol{\alpha}_2, \partial L/\partial \theta) = 0 \tag{3.20}$$

This procedure is computationally simpler than that of estimating all parameters $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \theta$ simultaneously in Eq. 3.19.

3.3.4 Modified IFM Method with Incorporation of Historical Information

Since the current IFM method can only be used for systematic data series, a modified IFM (MIFM) method with an incorporation of historical and paleological information is proposed and described as follows.

Let x_i and y_i (i = 1,..., s-c) respectively denote the systematic data of marginal distributions (flood peak and volume); g_j and p_j (j = 1,..., k) respectively denote the k largest floods of marginal distributions (flood peak and volume) with the same years of occurrence. Of the k largest floods, c occurred during the systematic record and m occurred during the pre-gauging period h (k = m+c and $h = n - s$); X_0 (or Y_0) is the fixed threshold of margin exceeded by the k largest flood peaks (or volumes) and not exceeded by any of the remaining $n - k$ flood peaks (or volumes). Furthermore, let f_x, and f_y denote the univariate marginal PDFs, and F_x, and F_y denote the univariate marginal CDFs of variables X and Y, respectively. f_{XY} denotes the joint PDF.

Referring to Eq. 3.14, the likelihood function with historical floods for joint distributions can be described as

$$\begin{aligned} l(\theta, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) &= \prod_{i=1}^{s-c} f_{XY}(x_i, y_i) \prod_{j=1}^{k} f_{XY}(g_j, p_j) [\int_{-\infty}^{X_0} \int_{-\infty}^{Y_0} f_{XY}(x, y) dxdy]^{h-m} \\ &= \prod_{i=1}^{s-c} f_{XY}(x_i, y_i) \prod_{j=1}^{k} f_{XY}(g_j, p_j) \{C_\theta[F_X(X_0), F_Y(Y_0)]\}^{h-m} \end{aligned} \tag{3.21}$$

Then, the log-likelihood function for joint distribution can be expressed as:

$$\begin{aligned} L(\theta, \boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}) = & \sum_{i=1}^{s-c} \log c_\theta[F_X(x_i), F_Y(y_i)] + \sum_{j=1}^{k} \log c_\theta[F_X(g_j), F_Y(p_j)] \\ & + (h-m) \log C_\theta[F_X(X_0), F_Y(Y_0)] + \sum_{i=1}^{s-c} \log f_X(x_i) + \sum_{j=1}^{k} \log f_X(g_j) \\ & + \sum_{i=1}^{s-c} \log f_Y(y_i) + \sum_{j=1}^{k} \log f_Y(p_j) \end{aligned} \tag{3.22}$$

In which, the two log-likelihood functions for the univariate marginal distribution are

$$L_1(\boldsymbol{\alpha_1}) = \sum_{i=1}^{s-c} \log f_X(x_i) + \sum_{j=1}^{k} \log f_X(g_j) \tag{3.23}$$

$$L_2(\boldsymbol{\alpha_2}) = \sum_{i=1}^{s-c} \log f_Y(y_i) + \sum_{j=1}^{k} \log f_Y(p_j) \tag{3.24}$$

Similar to the IFM method, the MIFM method also consists of two separate procedures:

(a) The log-likelihoods $L_1(\boldsymbol{\alpha_1})$ and $L_2(\boldsymbol{\alpha_2})$ are separately maximized by Eqs. 3.23 and 3.24 to get estimates $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$;

(b) The function $L(\theta, \hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2)$ is maximized by Eq. 3.22 over θ to get $\hat{\theta}$.

As a consequence, the precious historical information is used to estimate not only the parameters of marginal distributions but also the dependence parameters of joint distribution that is based on the correlation of the marginal distributions. The more additional information of marginal distribution provides, the more precise dependence structure will be obtained.

3.3.5 *Case Study*

The Three Gorges reservoir (TGR) in China is selected as an illustrative example. The basin area of TGR is one million km^2, and the annual average discharge and runoff volume at the dam site are 14,300 m^3/s and 4510 × 10^8 m^3, respectively. The TGR located on middle reaches of the Yangtze River is the largest water conservancy project in the world, with a normal pool level at an elevation of 175 m. The total storage capacity of the TGR is 393 × 10^8 m^3, of which 221.5 × 10^8 m^3

is flood control storage, and 165 × 10^8 m^3 is the conservation regulating storage volume. With 26 hydro-generators installed, the mean annual electricity output of the TGR reaches up to 847 × 10^8 kW•h. The TGR also plays a key role in the flood prevention of Yangtze River basin which is the richest area in China (Li et al. 2010).

3.3.5.1 Systematic Record and Historical Floods

The annual maximum peak discharge (Q), 3-day flood volume (W_3), and 15-day flood volume (W_{15}) are available with a systematic record of 128 years (1882–2009, i.e., no systematic data are formally gauged before 1882). Besides the systematic observations, a lot of historical flood events had been investigated by CWRC (Changjiang Water Resources Commission) in the last century for the design of the Three Gorges Project. The gathered information from gauging authority records, historical documents, archives, flood marks and stone inscriptions showed the concrete positions of high water stages recorded. As a result, the eight largest historical floods since 1153 were quantificationally evaluated by CWRC and other relevant units (CWRC 1996).

As the same notations defined previously, the length of the systematic observations is unequivocally given: $s = 128$ years; since no extraordinary flood occurred during the systematic record, $c = 0$ and $k = m$; for the joint distribution of flood peak (Q) and 3-day flood volume (W_3), $k = m = 8$; for the joint distribution of flood peak and 15-day flood volume (W_{15}), $k = m = 3$; the perception thresholds of peak discharge, 3-day flood volume and 15-day flood volume are $X_{0Q} = 80{,}000$ m^3/s, $X_{0w3} = 200 \times 10^8$ m^3 and $X_{0w15} = 780 \times 10^8$ m^3, respectively; and the pre-gauging period, $h = 730$ (i.e. from 1153 to 1882). These data settings are also listed in Table 3.3.

3.3.5.2 Parameter Estimation for Marginal Distributions

The empirical probabilities of univariate discontinuous series can be computed by Weibull formula recommended by MWR (2006)

$$P_i = P(x \geq x_i) = \begin{cases} P_h(i) = \frac{i}{n+1} & i = 1, \cdots, k \\ P_s(i) = P_h(k) + (1 - P_h(k)) \times \frac{i}{s-c+1} & i = 1, \cdots, s-c \end{cases} \tag{3.25}$$

Table 3.3 Data settings for the modified IFM method

Variables	Threshold X_0/Y_0	h	s	k	m
Q (m^3/s)	80,000	730	128	8	8
W_3 (10^8 m^3)	200				
Q (m^3/s)	80,000	730	128	3	3
W_{15} (10^8 m^3)	780				

where P_i represents the exceedance probability; $P_h(i)$ is the empirical probabilities of historical floods for $i = 1,\ldots, k$; $P_s(i)$ is the empirical probabilities of systematic data for $i = 1,\ldots, s{-}c$; and the meanings of n, k, s, c are the same as those defined in Fig. 3.3.

The parameters of the P-III marginal distributions estimated by the first stage of the MIFM method in Eqs. 3.23 and 3.24 are listed in Table 3.4. A Chi-Square Goodness-of-fit test is performed to test the assumption, H_0, that the flood magnitudes follow the P-III distribution. Table 3.5 shows that the assumption cannot be rejected at the 5% significance level. The marginal distribution frequency curves of flood peak and flood volumes are drawn in Fig. 3.4, in which the line represents the theoretical distribution, the crossings and circles represent systematic record and historical flood data, respectively. Figure 3.4 indicates that all the theoretical distributions can fit the observed data reasonably well.

3.3.5.3 Empirical Joint Probabilities of Dependence Flood Variables

Empirical (observed) joint probabilities of flood peak (Q) and volume (W) are computed in a manner analogous to that for a univariate variable. A two-dimensional table is constructed in which the variable X and Y are arranged in descending order. The joint probabilities (exceedance) of k historical floods and s-c systematic data are empirically computed separately, which are expressed as

$$\begin{array}{l} F(x_i, y_i) = \\ P(X \geq x_i, Y \geq y_i) = \end{array} \begin{cases} P_h(i) = \dfrac{\sum\limits_{l=1}^{i}\sum\limits_{p=1}^{i} N_{lp}}{n+1} & i = 1,\ldots,k \\[2ex] P_s(i) = P_h(k) + (1 - P_h(k)) \times \dfrac{\sum\limits_{l=1}^{i}\sum\limits_{p=1}^{i} M_{lp}}{s-c+1} & i = 1,\ldots,s-c \end{cases} \tag{3.26}$$

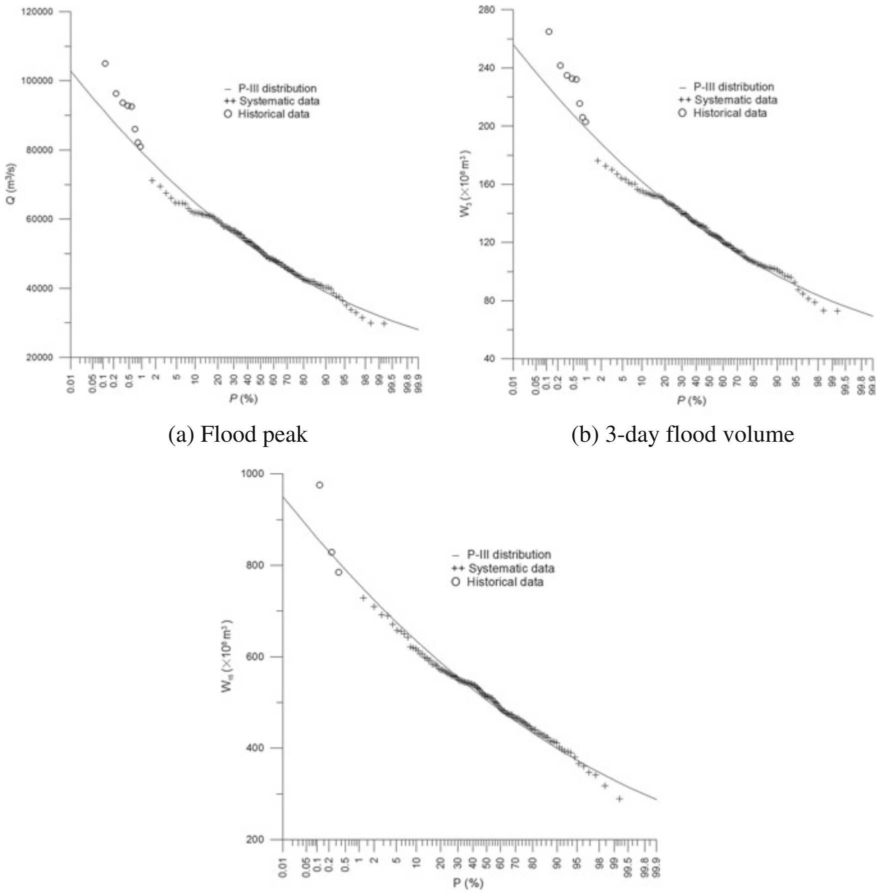
where $F(x_i,\ y_i)$ is obtained by arranging the number of $(x_i,\ y_i)$ by either x_i or $y_{i;}$ $P_h(i)$ is the empirical joint probabilities of historical floods and N_{lp} is the number of $(x_i,\ y_i)$ counted as $x_j\ \geq\ x_i$ and $y_j\ \geq\ y_i$, $i = 1,\ldots, k$, $1\ \leq\ j \leq i$; $P_s(i)$ is the empirical

Table 3.4 Estimated parameters of P-III marginal distributions for flood peak and volumes by MIFM

Variables	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$
Q (m^3/s)	11.11	0.0003	17066.7
W_3 (10^8 m^3)	11.89	0.1348	39.7
W_{15} (10^8 m^3)	18.26	0.0463	118.22

Table 3.5 Hypothesis test results of P-III marginal distributions for flood peak and volumes

Variables	$\chi_{0.05}$	Chi-Square statistics, χ^2
Q (m^3/s)	7.815	4.924
W_3 (10^8 m^3)	9.488	5.048
W_{15} (10^8 m^3)	7.815	4.110



(a) Flood peak

(b) 3-day flood volume

(c) 15-day flood volume

Fig. 3.4 P-III distributions fitted to flood peak and volumes with historical information

joint probabilities of systematic data and M_{lp} is the number of (x_i, y_i) counted as $x_j \geq x_i$ and $y_j \geq y_i$, $i = 1,\ldots, s-c$, $1 \leq j \leq i$; and n is the total length of the analyzed time period ($n = s+h$).

3.3.5.4 Identification of Copula

The parameters of marginal distributions are estimated in the first stage of MIFM method. The dependence parameter θ is obtained by maximizing the log-likelihood function of the joint distribution. For Gumbel copula, the estimation results are $\theta = 16.2524$ for the joint distribution of flood peak and 3-day flood volume, and $\theta = 3.2977$ for that of flood peak and 15-day flood volume. For Student copula, the

Table 3.6 RMSE of Gumbel and student's copulas and upper TDC estimated by parametric and nonparametric methods

Variables	RMSE		λ_U of copula		$\hat{\lambda}_U^{LOG}$	$\hat{\lambda}_U^{SEC}$	$\hat{\lambda}_U^{CFG}$
	Gumbel	Student	Gumbel	Student			
Q (m^3/s)	0.0262	0.0874	0.9564	0.8954	0.9442	0.9511	0.9482
W_3 (10^8 m^3)							
Q (m^3/s)	0.0413	0.2149	0.7661	0.5262	0.7218	0.7618	0.7109
W_{15} (10^8 m^3)							

estimation results are ($\theta = 0.9947$, $v = 6$) for the joint distribution of flood peak and 3-day flood volume, and ($\theta = 0.8598$, $v = 5$) for that of flood peak and 15-day flood volume. The root mean square errors (*RMSE*) of Gumbel and Student copulas are listed in Table 3.6. The comparison results show that the Gumbel copula represents the bivariate distribution of correlated flood peak and volumes better than that of Student copula.

The upper tail dependence coefficients (TDC) of Gumbel copula ($\lambda_U = 2 - 2^{1/\theta}$) and student's t copula ($\lambda_U = 2t_{v+1}\left(-\sqrt{(v+1)(1-\theta)/(1+\theta)}\right)$) are computed by the estimated parameters and listed in Table 3.6. The upper TDC can also be estimated by the nonparametric estimation, which is a much more general as no assumption is made about copula and marginal distributions (Poulin et al. 2007). The Log, Sec and CFG estimators of upper TDC (Coles et al. 1999; Joe et al. 1997; Poulin et al. 2007; Frahm et al. 2005) are respectively determined as follows.

$$\overset{\frown}{\lambda}_U^{LOG} = 2 - \frac{\log C_n((n-k)/n, (n-k)/n)}{\log((n-k)/n)}, \quad 0<k<n \tag{3.27}$$

$$\overset{\frown}{\lambda}_U^{SEC} = 2 - \frac{1 - C_n((n-k)/n, (n-k)/n)}{1-(n-k)/n}, \quad 0<k<n \tag{3.28}$$

$$\overset{\frown}{\lambda}_U^{CFG} = 2 - 2\exp\left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\sqrt{\log\frac{1}{U_i}\log\frac{1}{V_i}}\Big/\log\frac{1}{\max(U_i, V_i)^2}\right)\right] \tag{3.29}$$

in which

$$C_n(u, v) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}\left(\frac{R_i}{n+1} \le u, \frac{S_i}{n+1} \le v\right) \tag{3.30}$$

where $C_n(u, v)$ is the empirical copula, I denote the indicator function, R_i and S_i are the ranks of block maxima x_i and y_i, respectively. $\{(U_1, V_1), \ldots, (U_n, V_n)\}$ denote random sample obtained from the copula C.

The nonparametric estimation results of upper TDC are calculated and also listed in Table 3.6. The comparison results of Table 3.7 show that the upper TDC of Gumbel copula is much closer to the nonparametric estimation results than that of

Table 3.7 Parameters of marginal distributions and copula estimated by different data and methods

Variables	IFM				MIFM			
	P-III			Copula	P-III			Copula
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\theta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\theta}$
Q (m^3/s)	13.72	0.0004	16933.3	15.1545	11.11	0.0003	17066.7	16.2524
W_3 (10^8 m^3)	15.75	0.1736	36.28		11.89	0.1348	39.7	
Q (m^3/s)	13.72	0.0004	16933.3	3.0962	11.11	0.0003	17066.7	3.2977
W_{15} (10^8 m^3)	22.15	0.0541	102.38		18.26	0.0463	118.22	

student copula. This indicates that Gumbel copula reproduces better the observed tail dependence coefficient, and the extreme behavior of Gumbel copula is more similar to that of the sample. Therefore, the Gumbel copula is used to model the dependence between the extreme maximum annual flood peak and volumes in this study.

3.3.5.5 Copula-Based Conditional Distributions

The conditional flood distributions with historical flood data can be easily derived if the copula-based bivariate flood distribution is constructed. For instance, the conditional distributions for flood volume given that the peak discharge exceeding a certain threshold q_{x0} can be expressed as

$$\begin{aligned} P(W \leq w|Q > q_{X0}) &= \frac{P(W \leq w, Q > q_{X0})}{P(Q > q_{X0})} \\ &= \frac{F_Y(w) - C_\theta[F_X(q_{X0}), F_Y(w)]}{1 - F_X(q_{X0})} \end{aligned} \tag{3.31a}$$

$$\begin{aligned} P(W > w|Q > q_{X0}) &= \frac{P(W > w, Q > q_{X0})}{P(Q > q_{X0})} \\ &= \frac{1 - F_X(q_{X0}) - F_Y(w) + C_\theta[F_X(q_{X0}), F_Y(w)]}{1 - F_X(q_{X0})} \end{aligned} \tag{3.31b}$$

where F_x and F_Y represent the marginal distributions, and θ represents the dependence parameter of the bivariate distribution.

Likewise, the conditional distribution functions for peak discharge given that the flood volumes exceeding a certain threshold W_{Y0} can be expressed as

$$\begin{aligned} P(Q \leq q|W > w_{Y0}) &= \frac{P(Q \leq q, W > w_{Y0})}{P(W > w_{Y0})} \\ &= \frac{F_X(q) - C_\theta[F_X(q), F_Y(w_{Y0})]}{1 - F_Y(w_{Y0})} \end{aligned} \tag{3.32a}$$

$$\begin{aligned}P(Q > q|W > w_{Y0}) &= \frac{P(Q > q, W > w_{Y0})}{P(W > w_{Y0})} \\ &= \frac{1 - F_X(q) - F_Y(w_{Y0}) + C_\theta[F_X(q), F_Y(w_{Y0})]}{1 - F_Y(w_{Y0})}\end{aligned} \tag{3.32b}$$

The historical floods, which usually occurred as extraordinary events, may help exposit the correlation of variables with high return period. As a consequence, the incorporation of historical information into bivariate frequency analysis can provide better insight into the dependence structure of variables. The conditional probabilities accounting for historical floods can provide more comprehensive and adequate information, which is useful in evaluating the flood prevention capability.

3.3.5.6 Comparative Study and Discussions

The comparative study and discussions of MIFM and IFM methods are conducted in this section. First, the parameters of marginal distributions (Q, W_3, and W_{15}) and copulas are estimated by IFM and MIFM methods, respectively. Table 3.7 shows that the different data and methods lead to different parameter estimation results of both marginal distributions and copula. Second, the quantiles of flood peak (Q), 3-day flood volume (W_3) and 15-day flood volume (W_{15}) are estimated by univariate distribution (Chinese design flood guidelines), MIFM and IFM methods, respectively.

The Relative Errors (RE) of T-year quantile estimator are calculated by

$$\mathrm{RE} = \frac{\hat{X}_T - X_T}{X_T} \times 100\% \tag{3.33}$$

where X_T is the univariate quantile estimated by univariate distribution (Chinese design flood guidelines) with an incorporation of historical information; $\hat{X}_T$ represents the bivariate quantiles estimated by MIFM method with an incorporation of historical information or by IFM method using systematic records alone.

The relative errors (RE) of flood peak, 3-day flood volume, and 15-day flood volume are calculated and listed in Tables 3.8, 3.9, and 3.10, respectively. The results of these tables indicate that the bivariate quantiles estimated by MIFM

Table 3.8 Comparison of quantile Q estimated by univariate and bivariate distributions

T (years)	Univariate quantile Q_T (m^3/s)	MIFM		IFM	
		$\hat{Q}_T$ (m^3/s)	RE (%)	$\hat{Q}_T$ (m^3/s)	RE (%)
10,000	102,900	103,100	0.19	95,900	−6.80
1000	91,700	91,900	0.22	86,400	−5.78
100	79,400	79,700	0.38	75,800	−4.53
Mean relative error			0.26		−5.70

Table 3.9 Comparison of quantile W_3 estimated by univariate and bivariate distributions

T (years)	Univariate quantile W_{3T} (10^8 m^3)	MIFM		IFM	
		$\hat{W}_{3T}$ (10^8m^3)	RE (%)	$\hat{W}_{3T}$ (10^8m^3)	RE (%)
10,000	255.9	256.3	0.16	246.0	−3.87
1000	228.4	228.9	0.22	220.8	−0.33
100	198.0	198.6	0.30	193.0	−2.53
Mean relative error			0.23		−3.24

Table 3.10 Comparison of quantile W_{15} estimated by univariate and bivariate distributions

T (years)	Univariate quantile W_{15T} (10^8 m^3)	MIFM		IFM	
		$\hat{W}_{15T}$ (10^8 m^3)	RE (%)	$\hat{W}_{15T}$ (10^8 m^3)	RE (%)
10,000	950.3	958.2	0.83	924.4	−2.73
1000	859.5	868.1	1.00	842.5	−1.97
100	757.9	767.8	1.31	750.7	−0.95
Mean relative error			1.05		−1.88

approach is much closer to the univariate quantiles than that estimated by IFM method. The quantiles estimated by IFM method are much smaller than that of Chinese design flood guidelines. The mean relative errors are equal to −5.70, −3.24, and −1.88% for flood peak, 3-day flood volume, and 15-day flood volume, respectively.

3.4 Bivariate Design Flood Quantile Selection Using Copulas

To derive the feasible range, a boundary identification method is suggested, which is inspired by the ideas of Chebana and Ouarda (2011) and Volpi and Fiori (2012). Li et al. (2016) estimated the bivariate feasible ranges of flood peak and flood volume suitable for combination in the critical level curve. Two combination methods for estimating unique bivariate flood quantiles, i.e., the EFC method and the CEC method, are proposed based on the assumption of the relationship between u and v (or q and w).

3.4.1 Bivariate Return Period

In the conventional univariate analysis, flood events of interest are often defined by return periods. In the bivariate domain, however, it is still discussed by the community as to which method is most suitable to transform the joint exceedance probability to a bivariate joint return period (JRP). Different JPRs estimated by copula function have been developed for the case of a bivariate flood frequency analysis. Eight types of possible joint events were presented by Salvadori and De Michele (2004) using "OR" and "AND" operators, of which, two cases are of the greatest interest in hydrological applications (Shiau et al. 2006; Salvadori and De Michele 2004):

(1) (OR case) either $Q > q$ or $W > w$, i.e.,

$$E_{or} = \{Q > q \,\mathrm{or}\, W > w\} \tag{3.34}$$

(2) (AND case) both $Q > q$ and $W > w$, i.e.,

$$E_{and} = \{Q > q \,\mathrm{and}\, W > w\} \tag{3.35}$$

In simple words: for E_{or} to happen it is sufficient that either peak discharge Q or flood volume W (or both) exceed given thresholds; instead; for E_{and} to happen it is necessary that both Q and W are larger than prescribed values. Thus, two different JRPs can be defined accordingly (De Michele et al. 2005):

$$T_{or} = \frac{\mu}{P[Q > q\, or\, W > w]} = \frac{\mu}{1 - F(q,w)} \tag{3.36}$$

$$T_{and} = \frac{\mu}{P[Q > q\, and\ W > w]} = \frac{\mu}{1 - F_Q(q) - F_W(w) + F(q,w)} \tag{3.37}$$

where μ is the mean inter-arrival time between two consecutive events (in the case of annual maxima μ = 1 year), and $F(q,\ w) = P(Q\ \leq\ q,\ W\ \leq\ w)$.

The Kendall JRP was introduced by Salvadori and De Michele (2004) to identify the univariate critical threshold in a multivariate context, which is given by:

$$\theta_t = \frac{\mu_T}{1 - K_C(t)} \tag{3.38}$$

where K_C is the Kendall's distribution function associated with the joint cumulative distribution function of the copula's level curves: $K_C(\mathrm{t}) = P[C(\mathrm{u},\ \mathrm{v})\ \leq\ t]$. It allows for the calculation of the probability that a random point (u, v) in the unit square has a smaller (or larger) copula value than a given critical probability level t. In other words, it is related to the probability of occurrence of an event in the area over the copula level curve of value t.

Different definitions of the multivariate return period are available in the literature, based on regression analysis, bivariate conditional distributions, survival Kendall distribution function, and structure performance function. For instance, some studies have focused on a structure-based return period for the design and or risk assessment of hydrological structures in a bivariate environment (Volpi and Fiori 2014).

A comprehensive review of the JRP estimation methods was given by Volpi and Fiori (2014).

The OR return period given in Eq. 3.36 has been extensively applied in multivariate hydrological frequency analysis (e.g., Shiau et al. 2006; Salvadori and De Michele 2004; Chebana and Ouarda 2011; Volpi and Fiori 2012; Li et al. 2013). In this study, we focus on the OR case for quantile estimation in a bivariate context.

3.4.2 Feasible Range Identification for Bivariate Quantile Curve

The critical level curve, as shown in Fig. 3.5, was defined as a bivariate quantile curve by Chebana and Ouarda (2011). As previously stated, for the case of OR return period, the function that describes the level curve for any given return period T or critical probability level p has two asymptotes, $q = q_p$ and $w = w_p$, where $q_p = F_Q^{-1}(p)$ and $w_p = F_W^{-1}(p)$ are the quantiles of the marginal distribution for the given probability level p. According to Eq. 3.36 in the bivariate case, the choice of an appropriate return period T or a critical probability level p for hydraulic structure design will lead to the infinite combinations of flood peak and volume. However, all the bivariate flood events with the same value of T or p along the level curve differ greatly not only in terms of their quantile values, but also in terms of their probability of occurrence, which is measured by the joint probability density function (PDF), i.e., $f(q, w)$, evaluated along the critical level curve (Volpi and Fiori 2012). Meanwhile, different combinations of Q and W are generally not equivalent from a practical point of view, although they all satisfy the flood prevention standards. The boundaries (see points B and C in Fig. 3.5) for selection of design flood peak and volume are necessary in the case that the flood combinations are outside the boundaries with unrealistically low occurrence probabilities.

Chebana and Ouarda (2011) proposed a method to decompose the quantile curve in Fig. 3.5 into a naive part (i.e., the subset BC) and a proper part (outside subset

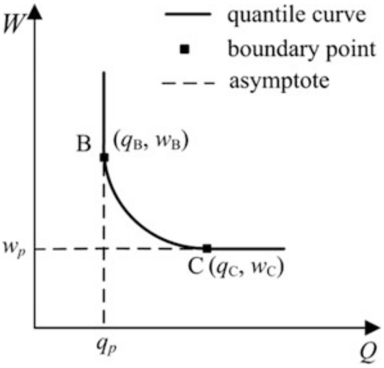


Fig. 3.5 Bivariate quantile curve with a critical probability level p

BC). They assumed that the naive part is composed of two segments starting at the end of each extremity of the proper part. They also suggested selecting these boundary points according to the empirical version or as close as to the asymptotes (the naive part). Volpi and Fiori (2012) defined the distance of each point along the quantile curve in Fig. 3.5 from its vertex as a random variable (s) and derived its PDF. The boundary points of the quantile curve are identified with a chosen percentage in the probability of the events. They also proposed a way of decomposition of the quantile curve into the naive part and proper part. However, the procedure presented by Volpi and Fiori (2012) is difficult to apply in the curvilinear coordinate system [$s(x, y)$, $n(x, y)$] or to derive the expression of a random variable (s). To overcome these limitations, an approach to identify the boundary points (i.e., B and C) of the quantile curve is developed. A new density function $\varphi(q)$ is used to measure the relative likelihood of flood events, which is a non-curvilinear variable in the procedure.

To derive the new density function with a chosen probability level to decompose the quantile curve, a joint distribution of annual maximum flood peak (Q) and flood volume (W) should be built by copula functions. The joint distribution function $F(q, w)$ can be expressed in terms of its marginal functions and $F_W(w)$ by using an associated dependence function C, $F(q, w) = C[F_Q(q), F_W(w)]$.

It is found that flood peak and volumes are usually upper-tailed dependent variables and the Gumbel copula can reproduce best the observed tail dependence coefficient (e.g., Poulin et al. 2007) Therefore, the Gumbel copula is taken as an example to illustrate the developed boundary identification method because of its easy expression and wide applications (Li et al. 2013).

For the Gumbel copula function, the relationship of joint distribution $C_\theta(u, v)$ and bivariate return period T can be expressed as ($\mu = 1$ for annual maxima flood series):

$$C_\theta(u, v) = \exp\{-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta}\} = 1 - \frac{1}{T} \tag{3.39}$$

where θ is the dependence parameter of the Gumbel copula, $u = F_Q(q)$, $v = F_W(w)$.

Thus, the relationship between u and v with the given bivariate return period T can be derived as:

$$v = \exp\left\{-[(-\ln u)^\theta - (-\ln(1 - \frac{1}{T}))^\theta]^{1/\theta}\right\} \tag{3.40}$$

Replacing $u = F_Q(q)$, and $v = F_W(w)$into the above equation yields:

$$F_W(w) = \exp\left\{-[(-\ln F_Q(q))^\theta - (-\ln(1 - \frac{1}{T}))^\theta]^{1/\theta}\right\} = \eta(F_Q(q)) \tag{3.41}$$

in which, $\eta(x) = \exp\left\{-[(-\ln x)^\theta - (-\ln(1 - \frac{1}{T}))^\theta]^{1/\theta}\right\}$

Thus, the relationship between Q and W with the fixed bivariate return period T can be derived as:

$$w = F_W^{-1}(v) = F_W^{-1}(\eta(F_Q(q)) = \varsigma(q) \tag{3.42}$$

where $F_W^{-1}(v)$ is the inverse CDF of flood volume W. The above equation reveals that W can be derived by Q if the bivariate return period T is fixed.

It should be noted that other copulas with more complicated formulas sometimes may be needed. For the Frank copula, Clayton copula and several two-parameter copulas, the implicit expression for describing the relationship between Q and W in Eqs. 3.39 to 3.42 can be derived. For copulas with more complicated expressions, the numerical method should be applied. For example, the unique value of w could be obtained with given q by a trial and error method.

After obtaining the corresponding relationship of the values of w and q for the flood events along the critical level curve, the bivariate joint PDF of w and q can be expressed according to Sklar's theory as (Nelsen 2006):

$$f(q, w) = c_\theta(F_Q(q), F_W(w)) \cdot f_Q(q) \cdot f_W(w) \tag{3.43}$$

where $f_Q(q)$ and $f_W(w)$ are univariate PDFs of flood peak and volume, respectively, and $c_\theta(u, v)$ is the density of $C_\theta(u, v)$ and defined as:

$$c_\theta = \frac{\partial^2 C_\theta(u, v)}{\partial u \partial v} \tag{3.44}$$

Referring to Eqs. 3.41 and 3.42, the bivariate joint PDF of flood peak and volume can be finally described as the function of the single random variable of flood peak Q for the fixed bivariate return period T, i.e.,

$$f(q, w) = c_\theta(F_Q(q), \eta(F_Q(q))) \cdot f_Q(q) \cdot f_W(\varsigma(q)) \tag{3.45}$$

According to Eq. 3.45, there is a curve that can describe the relationship between joint PDF $f(q, w)$and flood peak Q for a given bivariate return period T or a critical probability level p. Assume that the area between the curve of $f(q, w)$ and the horizontal axis of flood peak Q is A, i.e.,

$$A = \int_{q_p}^{+\infty} f(q, w) dq = \int_{q_p}^{+\infty} c(F_Q(q), \eta(F_Q(q))) \cdot f_Q(q) \cdot f_W(\varsigma(q)) dp \tag{3.46}$$

where q_p represents univariate design value of flood peak, i.e., $q_p = F_Q^{-1}(p)$, which is chosen as the lower bound of flood peak in the estimation of the bivariate design flood values.

As $f(q, w)$ is a joint density function of q and w, area A does not equal to 1 if only q is taken as an integral variable (i.e., $A \neq 1$). A new density function $\varphi(q)$ over the area A which has proper density characters is constructed and expressed as follows:

$$\varphi(q) = \frac{f(q,w)}{A} = \frac{f(q,w)}{\int_{q_p}^{+\infty} f(q,w)dq} \tag{3.47}$$

Obviously, there is a one-to-one correspondence between the density function $\varphi(q)$ and bivariate PDF $f(q, w)$. The density function $\varphi(q)$ varies with the horizontal axis and $\int_{q_T}^{+\infty} \varphi(q)dq = 1$.

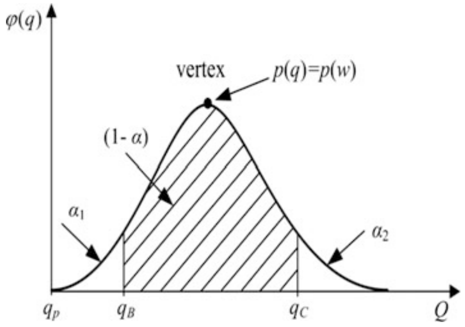
As previously stated, the bivariate design flood combinations near the upper and lower bounds of the quantile curve have lower occurrence probability than that near the middle of the quantile curve. As a consequence, the bivariate PDF $f(q, w)$ of bivariate design flood combination near the upper and lower bounds of quantile curve is smaller than that near the middle of the quantile curve. The density function $\varphi(q)$ has the same property as the bivariate PDF $f(q, w)$. As the design flood peak (or flood volume) varies from the lower bound, i.e., (q_p) to infinitely great, the density function $\varphi(q)$ increases to the maximum value and then decreases gradually, as shown in Fig. 3.6. The vertex of the density function $\varphi(q)$ describing the full dependence (Chebana and Ouarda 2011; Volpi and Fiori 2012) between peak and volume has the highest density. In other words, this is the most likely bivariate design flood event.

Once the density function $\varphi(q)$ along Q is defined by Eq. 3.43, we can evaluate the lower and upper bounds that contain $\varphi(q)$ with probability of $1-\varepsilon$, for a given probability level ε. The quantiles of lower and upper bounds (q_{B} and q_{C}) are specified respectively by (Volpi and Fiori 2012):

$$\int_{q_p}^{q_B} \varphi(q)dq = \alpha_1 \tag{3.48}$$

$$\int_{q_p}^{q_C} \varphi(q)dq = 1 - \alpha_2 \tag{3.49}$$

Fig. 3.6 Relationship between density function $\varphi(q)$ and flood peak Q



where $\alpha_1 + \alpha_2 = \varepsilon$. The lower and upper bounds q_B and q_C identify a feasible range on the quantile curve, bounded by the points of coordinates (q_B, $\zeta(q_B)$) and (q_C, $\zeta(q_C)$), that excludes the ε percentage in the probability of the critical events. The probability levels α_1 and α_2 can be arbitrarily chosen, taking account of the specific problem under investigation (Volpi and Fiori 2012).

3.4.3 Bivariate Flood Quantile Selection

For a given bivariate return period T, there are countless combinations of u and v that satisfy Eq. 3.39. To derive the design values of flood peak q and flood volume w, the unique combination of u and v (or q and w) should be determined. Hence besides Eq. 3.39, one more equation that can establish the relationship between u and v (or q and w) is necessary. Two combination methods were proposed to derive the quantiles of flood peak and flood volume for given multivariate return periods, and they are now outlined.

3.4.3.1 Equivalent Frequency Combination Method

With a given bivariate return period T, we assume that the flood peak and flood volume have the same probability of occurrence, i.e., $u = v$ (or $F_Q(q) = F_W(w)$). This assumption is usually taken as a uniform procedure for the derivations of design flood values and design flood hydrograph in China (MWR 2006; Xiao et al. 2008, 2009; Chen et al. 2010). Then, the design frequency of bivariate equivalent frequency combination can be obtained by jointly solve the equation $u = v$ and Eq. (3.39).

Taking the Gumbel copula for example, the relationship between u and v with the given bivariate return period T is described in Eq. 3.39. Based on the assumption that $u = v$, the probabilities of occurrence of flood peak and volume (i.e., u and v) can be estimated by the solution of the following equation.

$$u = v = (1 - \frac{1}{T})^{\varsigma} \tag{3.50}$$

where $\varsigma = 2^{-\frac{1}{\theta}}$, and θ is the dependence parameter of the Gumbel copula.

Consequently, the design value of bivariate equivalent frequency combination can be derived by the inverse function of marginal distributions:

$$q = F_Q^{(-1)}(u) \tag{3.51a}$$

$$w = F_W^{(-1)}(v) \tag{3.51b}$$

3.4.3.2 Conditional Expectation Combination Method

Since the flood peak Q and flood volume W are dependent variables, one may wish to predict the value of W based on an observed value of Q. Let $g(Q)$ be a predictor, i.e., $g \in N = \{$all Borel functions g with $E[g(Q)]^2 < \infty$ Each predictor is assessed by the "mean squared prediction error" $E[W - g(Q)]^2$. The conditional expectation $E(W|Q)$ is the best predictor of W in the sense that

$$E[W - E(W|Q)]^2 = \min_{g \in N} E[W - g(Q)]^2 \tag{3.52}$$

Herein, during a flood event, when the flood peak $Q = q$ takes place; the conditional expectation $E(w|q)$ is used to estimate the value of flood volume, which can be derived by

$$E(w|q) = \int_{-\infty}^{+\infty} w f_{W|Q}(w) dw \tag{3.53}$$

where $f_{W|Q}(w)$ is the density function of the conditional CDF $F_{W|Q}(w)$ and defined as (Zhang and Singh 2006).

$$f_{W|Q}(w) = \frac{f(q,w)}{f_Q(q)} = \frac{c_\theta(u,v) f_Q(q) f_W(w)}{f_Q(q)} = c_\theta(u,v) f_W(w) \tag{3.54}$$

Hence, Eq. 3.53 can be expressed by

$$E(w|q) = \int_{-\infty}^{+\infty} w f_{W|Q}(w) dw = \int_{-\infty}^{+\infty} w c_\theta(u,v) f_W(w) dw = \int_0^1 F_W^{-1}(v) c_\theta(u,v) dv \tag{3.55}$$

where $F_W^{-1}(\cdot)$ is the inverse CDF of W.

Then, the flood peak q and $E(w|q)$ will be the conditional expectation combination if the following equations are satisfied

$$\begin{cases} u = F_Q(q) \\ v = F_W[E(w|q)] \\ \frac{1}{1 - C_\theta(u,v)} = T \end{cases} \tag{3.56}$$

The above equation can be solved by trial and error method with different values of q.

3.4.4 Case Study

3.4.4.1 Bivariate Quantile Curve and Feasible Range Identification

The return period of design flood of Geheyan reservoir, i.e., $T = 1000$-year, is selected as the bivariate return period and $T = 200$-year is also chosen for comparison. The bivariate quantile curves of the two return periods are shown in Fig. 3.7. Even if the Gumbel copula model is symmetric, the probability density function $\varphi(q)$ is not symmetrical due to the difference in the marginal distributions.

The upper and lower bounds on the level curve are estimated numerically by solving Eqs. 3.48 and 3.49, and assuming for simplicity (although other assumptions are possible) $\alpha_1 = \alpha_2 = \varepsilon/2$, with $\varepsilon = 0.05$. The upper and lower bounds are denoted as B_1 and C_1, respectively, in Fig. 3.7. It is found that the bounds are close to the horizontal asymptote (i.e., $w_7 = 61.49 \times 10^8$ m^3 for $T = 1000$ and $w_7 = 50.23 \times 10^8$ m^3 for $T = 200$) and vertical asymptote (i.e., $q_p = 22{,}800$ m^3/s for $T = 1000$ and $q_p = 19{,}300$ m^3/s for $T = 200$) due to the small value assumed for the probability level ε. The upper and lower bounds are also calculated by the boundary identification method proposed by Volpi and Fiori (2012). The results are also presented in Table 3.11, and the derived bounds are denoted as B_2 and C_2, as shown in Fig. 3.7. It is shown that the bounds estimated by the proposed method and that proposed by Volpi and Fiori (2012) are very similar.

3.4.4.2 Estimation of Bivariate Flood Quantiles

The bivariate EFC and CEC methods are used to estimate flood peak and 7-day flood volume quantiles with return periods of $T = 1000$ and $T = 200$ years,

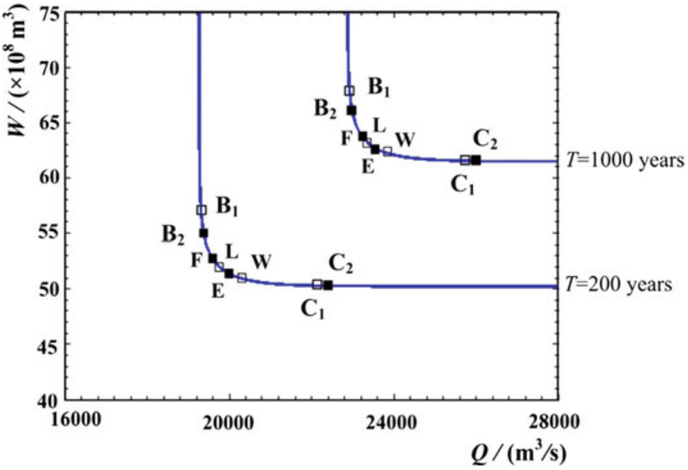


Fig. 3.7 Bivariate quantile curve of joint distribution of flood peak and 7-day flood volume

Table 3.11 Comparison of the lower and upper bounds of the quantile curve

Boundary identification method	Return period	Lower bound		Upper bound	
		Q_p(m^3/s)	W_7 (10^8 m^3)	Q_p(m^3/s)	W_7 (10^8 m^3)
Volpi and Fiori (2012)	1000	22,930	65.84	26,080	61.54
	200	19,350	50.27	22,460	55.86
Li et al. (2016)	1000	23,000	65.76	26,100	61.52
	200	19,400	54.49	22,500	50.26

respectively. For comparison, the univariate flood quantiles (called marginal quantiles by Chebana and Ouarda 2011) are estimated by marginal distributions, assuming that the univariate return periods (T_Q and T_W) are equal to the bivariate return period (i.e., $T_Q = T_W = T$). The univariate flood quantiles can be obtained from the equations $q = F_Q^{-1}(p) = F_Q^{-1}(1 - \frac{1}{T})$ and $w = F_W^{-1}(p) = F_W^{-1}(1 - \frac{1}{T})$. The results of the component-wise excess realization and the most likely realization proposed by Salvadori et al. (2011) are also estimated. The estimation results of bivariate and univariate quantiles are listed in Table 3.12. It is shown that the design values of bivariate quantiles are larger than those of univariate quantiles. The quantiles estimated by the four bivariate event selection methods are also shown in Fig. 3.7, and the estimation points of the EFC method are denoted as point E, while the quantiles estimated by the CEC method are denoted as point F. For the results of selection approaches proposed by Salvadori et al. (2011), the events of component-wise excess realization are denoted as point W, and the events of most likely realization are denoted as point L. From Fig. 3.7, we find that the joint design values estimated by the four event-selection methods are within the feasible regions. Consequently, the two proposed methods and selection approaches proposed by Salvadori et al. (2011) can be selected as an option of deriving unique flood quantiles, and they can satisfy the inherent law of hydrologic events and have a statistical basis to some degree. It can be seen from Table 3.12 and Fig. 3.7 that

Table 3.12 Design flood values and corresponding highest water levels estimated by bivariate quantile combinations and univariate distribution

T	Method	Q_p(m^3/s)	W_7 ($\times 10^8$ m^3)	Z_{max} (m)
1000	EFC	23,390	63.09	202.97
	CEC	23,420	62.98	202.92
	Component-wise excess realization	23,510	62.78	202.90
	Most-likely realization	23,400	63.05	202.95
	Univariate distribution	22,800	61.49	202.58
200	EFC	19,800	51.87	198.10
	CEC	20,130	51.11	197.79
	Component-wise excess realization	20,200	51.03	197.59
	Most-likely realization	19,940	51.50	197.82
	Univariate distribution	19,300	50.23	197.30

the estimated events of the EFC method and that of the most likely realization are similar. The bivariate EFC results have larger flood volume and smaller flood peak than bivariate CEC results. As well, the results estimated by the component-wise excess realization have larger flood peak and smaller flood volume than the other three methods.

3.4.4.3 Design Flood Hydrograph Based on Joint Distribution

The two combination methods are applied to derive the design flood hydrograph (DFH), and the resulting highest reservoir water level is selected as an index to evaluate the effects of different hydrological loads on the structure. The DFH for a dam is the flood of suitable probability and magnitudes adopted to ensure safety of the dam in accordance with appropriate design standards. The annual maximum flood hydrograph of 1997, which has a high peak and large volume with a posterior-peak shape, is selected as a typical flood hydrograph (TFH). The DFH with bivariate combinations is amplified from a TFH by the following method (Xiao et al. 2008):

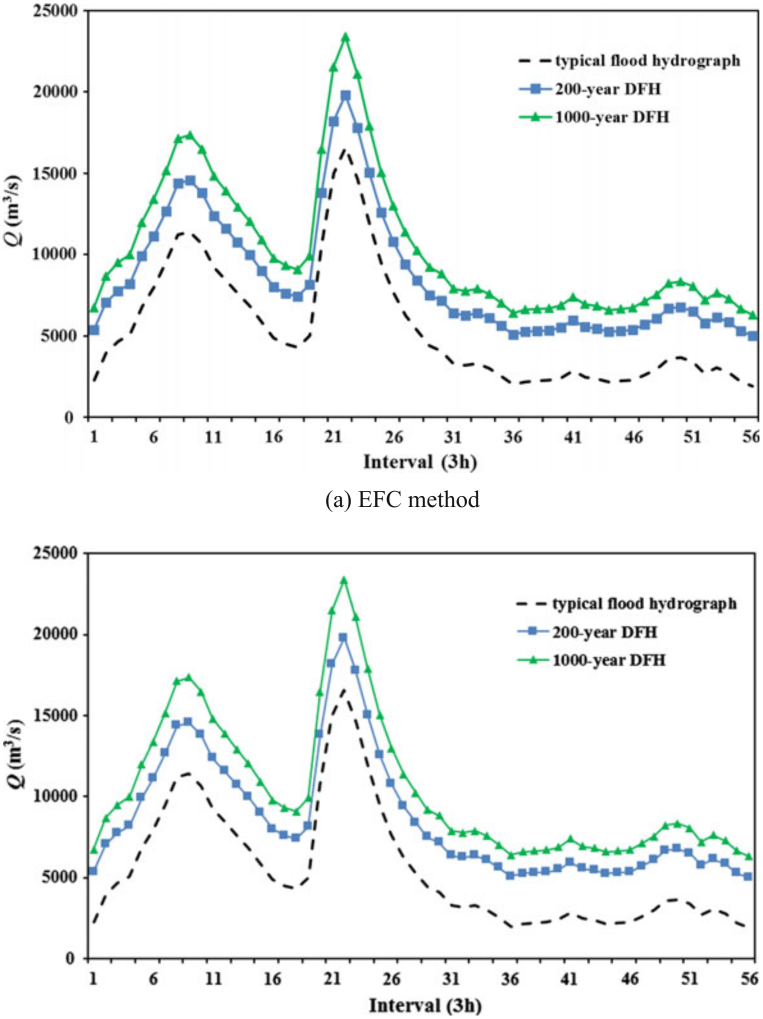
$$DFH(t) = (TFH(t) - Q_{TFH}) \times (w/DT - q)/(W_{TFH}/DT - Q_{TFH}) + q \qquad (3.57)$$

where $DFH(t)$ and $TFH(t)$ are the flood discharges of the DFH and TFH for time t respectively; Q_{TFH} is flood peak discharge of TFH; W_{TFH} is 7-day flood volume of TFH for flood duration DT; q and w are flood peaks and 7-day flood volumes of bivariate design flood combination, respectively. Nevertheless, other DFH generation methods based on flood peak and volume are also available and can be applied with the bivariate design value combinations.

The DFHs of 1000-year and 200-year return periods are constructed, respectively, with the bivariate EFC method and bivariate CEC method as shown in Fig. 3.8. It is found in Fig. 3.8 that only a few differences exist between the DFHs estimated by the EFC and CEC methods. This is because that the differences between the bivariate design values vary within a small range. Volpi and Fiori (2012) found that the feasible range on a p-level curve strongly depends on the correlation coefficient of Q and W. In the limiting case of full dependence, the level curve reduces to its vertex and the width of the feasible range tends to 0 (Volpi and Fiori 2012). Since the Kendall correlation coefficient between flood peak and 7-day volume in Geheyan reservoir equals to 0.66, the differences of quantiles estimated by EFC and CEC methods are relatively small in this case study.

The DFH rescaled by univariate distribution design values and two realizations proposed by Salvadori et al. (2011) is also derived from TFH by Eq. 3.59. These DFHs are routed through the Geheyan reservoir with initial water level (flood control limiting water level, 192.2 m). The corresponding highest reservoir water levels (Z_{max}) are calculated and are listed in Table 3.12.

It is shown in Table 3.12 that the design values of flood peak and 7-day flood volume obtained by univariate distribution method are both smaller than those



(a) EFC method

(b) CEC method

Fig. 3.8 DFHs derived by EFC method and CEC method

obtained by four bivariate methods. The resulting Z_{max} of the univariate method is relatively lower than those of bivariate approaches. Since flood events are naturally multivariate phenomena and flood peak and flood volume are mutually correlated, the quantiles estimated by bivariate distribution are more rational than these by univariate distribution (Chebana and Ouarda 2011).

The comparison results listed in Table 3.12 also show that Z_{max} obtained by bivariate EFC method is larger than that obtained by the other three bivariate methods, while the component-wise excess method reaches the lowest Z_{max}. The results of Z_{max} calculated by most-likely realization are a little lower than those of the EFC method, and the CEC method obtains a slightly higher Z_{max} than the component-wise excess method. Comparing the results of 200-year and 1000-year return period, it is found that the differences among the four bivariate methods decrease as the return period increases. The water level reaches 202.97 m by the EFC method and is slightly higher than other methods for the 1000-year return period. Since the Geheyan reservoir has a large amount of flood control storage with annual regulation ability, the design flood volume is relatively more important than peak discharge for flood prevention safety. As a consequence, the bivariate EFC method with slightly larger 7-day flood volume is safer for reservoir design than other methods.

3.5 Conclusion

According to the bivariate joint distribution of annual maximum flood occurrence dates and magnitudes, flood peaks and volumes, a flood frequency analysis model with an incorporation of historical floods are established based on GH copula. Modified inference functions for the margins (MIFM) method and the quantile curve boundary identification method are developed. The following conclusions are drawn from this Chapter:

(1) The Von Mises and Pearson Type III distributions can fit observed data series very well. The goodness-of-fit tests indicate a good agreement between observed and theoretical probabilities for both marginal and joint distributions.
(2) The proposed MIFM method may reduce the uncertainties of parameter estimation in flood frequency analysis, since the historical floods have been taken into account.
(3) The quantile combination methods provide a simple but effective way for bivariate quantile estimation with given bivariate return period. The results illustrate that the joint design values estimated by the two proposed combination methods are within the feasible regions, and the equivalent frequency combination method perform satisfactorily.

References

ASCE (American Society of Civil Engineers) (1996) Hydrology handbook. In: ASCE manuals and reports on engineering practices no. 28. American Society of Civil Engineers, New York, USA

Bayliss AC, Reed DW (2001) The use of historical data in flood frequency estimation. Report to MAFF.CEH Walingford

Black AR, Werritty A (1997) Seasonality of flooding: a case study of North Britain. J Hydrol 195:1–25

Chebana F, Quarda TBMJ (2011) Multivariate quantiles in hydrological frequency analysis. Environmetrics 22:441–455

Chen L, Guo SL, Yan BW, Liu P, Fang B (2010) A new seasonal design flood method based on bivariate joint distribution of flood magnitude and date of occurrence. Hydrol Sci J 55(8):1264–1280

Cohen AC (1976) Progressively censored sampling in the three parameters log-normal distribution. Technometrics 18(1):99–103

Coles S, Heffernan J, Tawn J (1999) Dependence measures for extreme value analysis. Extremes 2(4):339–365

Condie R (1986) Flood samples from a three-parameter lognormal population with historical information: the asymptotic standard error of estimate of the T-year flood. J Hydrol 85: 139–150

Condie R, Lee K (1982) Flood frequency analysis with historical information. J Hydrol 58:47–62

CWRC (Changjiang Water Resources Commission) (1996) Hydrologic inscription cultural relics in Three Gorges Reservoir. Science Press, Beijing (in Chinese)

De Michele C, Salvadori G, Canossi M, Petaccia A, Rosso R (2005) Bivariate statistical approach to check adequacy of dam spillway. J Hydrol Eng 1:50–57

Dupuis DJ (2007) Using copulas in hydrology: benefits, cautions, and issues. J Hydrol Eng 12 (4):381–393

Fisher NI (1993) Statistical analysis of circular data. Cambridge University Press, Cambridge

Frahm G, Junker M, Schmidt R (2005) Estimating the tail dependence coefficient: properties and pitfalls. Insur Math Econ 37(1):80–100

Guo SL, Cunnane C (1991) Evaluation of the usefulness of historical and paleological floods in quantile estimation. J Hydrol 129:245–262

Hald A (1949) Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. Skand Aktuarietidskrift 32(1/2):119–134

Hosking JRM (1995) The use of L-moments in the analysis of censored data". In: Balakrishnan N (ed) Recent advances in life-testing and reliability. CRC Press, Boca Raton, Fla, pp 545–564

Joe H (1997) Multivariate models and dependence concepts. Chapman & Hall, London

Joe H (2005) Asymptotic efficiency of the two-stage estimation method for copula-based models. J Multivariate Anal 94:401–419

Joe H, Xu JJ (1996) The estimation method of inference functions for margins for multivariate models. Technical Report no. 166, Department of Statistics, University of British Columbia

Leese MN (1973) Use of censored data in the estimation of Gumbel distribution parameters for annual maximum flood series. Water Resour Res 9(6):1534–1542

Li T, Guo S, Chen L, Guo J (2013) Bivariate flood frequency analysis with historical information based on Copula. J Hydrol Eng 18(8):1018–1030

Li T, Guo S, Liu Z, Xiong L, Yin J (2016) Bivariate design flood quantile selection using copulas. Hydrol Res. https://doi.org/10.2166/nh.2016.049

Li X, Guo SL, Liu P, Chen G (2010) Dynamic control of flood limited water level for reservoir operation by considering inflow uncertainty. J Hydrol 391:124–132

Mardia KV (1972) Statistics of directional data. Academic Press, London

McLeish DL, Small CG (1988) The theory and applications of statistical inference functions. Lecture Notes in Statistics, 44. Springer-verlag, New York

MWR (Ministry of Water Resources) (2006) Regulation for calculating design flood of water resources and hydropower projects. Chinese Water Resources And Hydropower Press, Beijing (in Chinese)

Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York

Poulin A, Huard D, Favre AC, Pugin S (2007) Importance of tail dependence in bivariate frequency analysis. J Hydrol Eng 12(4):L394–L403

Salvadori G, De Michele C, Durante F (2011) Multivariate design via Copulas. Hydrol Earth Syst Sci Discuss. 8:5523–5558

Salvadori G, De Michele C (2004) Frequency analysis via copulas: theoretical aspects and applications to hydrological events. Water Resour Res 40:W12511. https://doi.org/10.1029/2004WR003133

Shiau JT, Wang HY, Tsai CT (2006) Bivariate frequency analysis of floods using copulas. J Am Water Resour Assoc 42(6):1549–1564

Stedinger JR, Cohn TA (1986) The value of historical and paleoflood information in flood frequency analysis. Water Resour Res 22(5):785–793

USWRC (US Water Resources Council) (1981) Guidelines for determining flow frequency, Bulletin 17B. D. C, Washington

USWRC (US Water Resources Council) (1982) Guidelines for determining flood flow frequency, Bull. 17B (revised), U.S. Gov. Print. Off., Washington, D. C

Volpi E, Fiori A (2012) Design event selection in bivariate hydrological frequency analysis. Int Assoc Sci Hydrol 57(8):1506–1515

Volpi E, Fiori A (2014) Hydraulic structures subject to bivariate hydrological loads: return period, design, and risk assessment. Water Resour Res 50(2):885–897

Xiao Y, Guo SL, Liu P, Yan B, Chen L (2009) Design flood hydrograph based on multi characteristic synthesis index method. J Hydrol Eng 14(12):1359–1364

Xiao Y, Guo SL, Liu P, Fang B (2008) A new design flood hydrograph method based on bivariate joint distribution. In: Chen XH, Chen YD, Xia J, Zhang H (eds) Hydrological sciences for managing water resources in the asian developing world, IAHS Press, IAHS Publications 319, Wallingford, pp 75–82

Xu JJ (1996) Statistical Modelling and inference for multivariate and longitudinal discrete response data. Ph.D. thesis, Department of Statistics, University of British Columbia

Yue S, Quarda TBMJ, Bobée B, Legendre P, Bruneau P (1999) The Gumbel mixed model for flood frequency analysis. J Hydrol 226:88–100

Zhang L, Singh VP (2006) Bivariate flood frequency analysis using the copula method. J Hydrol Eng 11(2):150–164