

DOI:10.3876/j.issn.1000-1980.2016.05.011

Copula 熵方法及其在三变量洪水频率计算中的应用

李帆¹, 郑骞², 张磊³

(1. 扬州大学水利与能源动力工程学院, 江苏 扬州 225000;

2. 衢州市水利局, 浙江 衢州 324000; 3. 徐州市水利建筑设计研究院, 江苏 徐州 221116)

摘要: 为解决传统 Copula 方法在进行联合概率分布拟合过程中要先进行函数类型选择的问题, 将 Copula 函数和最大熵原理进行耦合, 通过求解具有最大熵的 Copula 方程, 求得二维联合分布函数, 即 Copula 熵方法。用求得的 Copula 函数对洪水事件的 3 个相关变量(洪峰流量、洪水总量和洪水历时)进行两两配对的二维联合分布拟合, 并利用 Gibbs 采样方法和 Copula 函数实现三变量洪水事件的随机模拟。以淮河鲁台子水文站的实测洪水资料为研究对象, 进行实例分析, 并通过拟合优度的计算, 证明 Copula 熵方法对多维相关变量概率拟合的有效性以及 Gibbs 采样方法在三变量洪水事件模拟过程中的有效性。

关键词: 洪水频率; 联合分布函数; Copula 函数; 最大熵原理; Copula 熵方法; 淮河; 鲁台子水文站

中图分类号: P333.2 **文献标志码:** A **文章编号:** 1000-1980(2016)05-0443-06

Copula entropy method and its application to trivariate flood frequency calculation

LI Fan¹, ZHENG Qian², ZHANG Lei³

(1. School of Hydraulic, Energy, and Power Engineering, Yangzhou University, Yangzhou 225000, China;

2. Quzhou Water Conservancy Bureau, Quzhou 324000, China;

3. Xuzhou Hydraulic Construction and Design Research Institute, Xuzhou 221116, China)

Abstract: In order to avoid selection of the type of functions when using traditional copula methods to fit the joint probability distribution, the copula function and the maximum entropy principle were jointly used (in a method also called the copula entropy method), and the two-dimensional joint distribution functions of flood variates were obtained by solving the copula function with the maximum entropy. With the solved copula function, the two-dimensional joint distribution functions of flood variates (peak discharge, flood volume, and flood duration) were mutually constructed. The Gibbs sampling method and copula functions were used to stochastically simulate trivariate flood events. A case study was conducted using the observed flood data from the Lutaizi Hydrological Station on the Huaihe River. A goodness-of-fit analysis shows that the copula entropy method is effective in fitting multivariate probability distributions and the Gibbs sampling method is effective in simulating trivariate flood events.

Key words: flood frequency; joint distribution function; copula function; maximum entropy principle; copula entropy method; Huaihe River; Lutaizi Hydrological Station

洪水事件的概率估计是防洪规划、水利工程建设和水资源管理的重要依据。一个洪水事件一般可以由 3 个具有相关关系的随机变量来描述, 即洪峰流量(P)、洪水总量(V)和洪水历时(D)。Copula 函数方法是一种实现多变量洪水事件联合分布拟合的有效途径, 并在水文领域中被广泛应用。郭生练等^[1]对 Copula 函数在水文中的应用进行了综述; 宋松柏等^[2]系统地探讨了各种 Copula 函数结构在西北典型流域中的应用,

收稿日期: 2015-12-07

基金项目: 江苏省高校自然科学研究项目(15KJD170003); 扬州大学科技创新培育基金(2015CXJ027)

作者简介: 李帆(1982—), 男, 辽宁营口人, 讲师, 博士, 主要从事水文统计研究。E-mail: lifan@yzu.edu.cn

分析了 Copula 函数在多变量水文计算领域的适用性及优越性。Copula 函数存在多种形式,其中二维 Archimedean Copula 函数应用最广泛。常用的 Archimedean Copula 函数包括 Frank Copula、Gumbel Copula 和 Clayton Copula 等类型,此外还有属于椭圆 Copula 函数的 Gaussian Copula 等。针对具体问题选取合适的 Copula 函数进行多维联合概率分布拟合。Salvadori 等^[3]用 Archimedean Copula 函数建立了最高洪水位与洪水总量、洪峰流量与洪水总量、降水强度与降水历时的 3 组二维联合分布,分析了 Gumbel Copula 和 Frank Copula 函数在水文二维极值联合分布构建上的有效性。Zhang 等^[4]用 4 种 Archimedean Copula 函数构造了洪峰与洪水总量、洪水总量与洪水历时的 2 组二维联合分布,并比较了他们的拟合优度,认为 Gumbel Copula 对所选河流洪水的拟合最好。肖义等^[5]、李天元等^[6]用 Gumbel Copula 函数构造边缘分布为 P-Ⅲ型分布的两变量联合分布,用以描述洪峰和洪水总量。Renard 等^[7]探讨了 Gaussian Copula 在洪水频率分析上的应用。李计等^[8]比较了 4 种对称的 Archimedean Copula 和 5 种非对称的 Archimedean Copula 对于干旱历时、干旱烈度和烈度峰值 3 个干旱特征变量联合分布的拟合效果。虽然 Copula 函数的应用已十分广泛,但在进行 Copula 函数参数的估算之前,必须先要对 Copula 函数的类型进行选择,然后再进行拟合分析。这种选型过程往往并不完全客观,从而造成计算结果不准确。为了避免这种由于选型而造成的偏差,本文将最大熵原理和 Copula 函数理论进行耦合,即 Copula 熵方法,以求得联合概率分布密度函数。另外,传统的 Copula 方程都是两变量的联合分布方程,在向三维变量拓展时往往比较困难,计算也比较复杂。笔者基于蒙特卡洛算法,利用 Gibbs 采样方法和 Copula 函数,生成大量洪水事件,从而实现对三变量的洪水事件的随机模拟,进而分析洪水事件的发生概率,并以淮河鲁台子水文站实测资料为对象,进行了示例研究。

1 理论与方法

1.1 Copula 函数

由 Sklar 定理可得二维 Copula 函数:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)) = C(u, v) \quad (1)$$

式中: $F(x_1, x_2)$ ——联合概率分布函数; $F_1(x_1) = u$ 、 $F_2(x_2) = v$ ——边缘分布函数。

根据 Sklar 定理,若边缘分布函数是连续的,则 Copula 函数 $C(u, v)$ 唯一。

1.2 最大熵原理与 Copula 熵

根据熵理论,对于一个连续的随机变量 $X \in [a_1, a_2]$,若概率密度函数为 $f(x)$,则其熵可定义为

$$H = - \int_{a_1}^{a_2} f(x) \ln f(x) dx \quad (2)$$

根据最大熵原理(POME),当 H 取得最大值时,人为假定最小,从而可以得到最合乎自然、最为超然、偏差最小的 $f(x)$ ^[9]。对于 Copula 函数,其对应的 Copula 熵为

$$W = - \int_0^1 \int_0^1 c(u, v) \ln c(u, v) du dv \quad (3)$$

其中

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}$$

式中: $c(u, v)$ ——二维 Copula 概率密度函数。

根据 POME,式(3)中使得 W 最大的 $c(u, v)$ 即为最适合的联合概率密度函数。

1.3 求解 Copula 概率密度函数

为计算最优的 $c(u, v)$,可以利用拉格朗日乘子法对式(3)进行求解。根据 Copula 概率密度函数的性质,可列出以下约束条件:

$$\int_0^1 \int_0^1 c(u, v) du dv = 1 \quad (4)$$

$$\int_0^1 \int_0^1 u^r c(u, v) du dv = \frac{1}{r+1} \quad \int_0^1 \int_0^1 v^r c(u, v) du dv = \frac{1}{r+1} \quad (5)$$

$$\int_0^1 \int_0^1 uvc(u,v) du dv = \frac{\rho + 3}{12} \quad (6)$$

式中: r —— r 阶矩; ρ ——Spearman 秩相关系数。

Chu^[10]证明,使得式(3)取得最大值的 $c(u,v)$ 的表达式为

$$c(u,v) = \lim_{n \rightarrow \infty} \varepsilon(u,v) \quad (7)$$

$$\int_0^1 \int_0^1 \varepsilon(u,v) du dv$$

其中

$$\varepsilon(u,v) = \exp \left[- \sum_r^n (\alpha_r u^r + \beta_r v^r) - \lambda uv \right]$$

式中: α, β, λ ——待求系数; n ——矩的阶数。

式(7)没有解析解,可以使用牛顿迭代法计算待求系数^[11],进而求得 Copula 概率密度函数的表达式。

1.4 Gibbs 采样方法生成三变量洪水事件

Copula 熵方法在理论上可以直接从二维扩展到三维,得到三变量 Copula 概率密度函数,但计算量较大,比较耗时。笔者采用 Gibbs 采样方法^[12]和二维 Copula 函数进行三维随机变量模拟,这种方法原理简单,计算速度快,并已在海浪波况的随机模拟中成功运用^[13-14]。

令 X, Y, Z 为洪峰流量、洪水总量和洪水历时的随机变量, x, y, z 分别表示洪峰流量、洪水总量、洪水历时的累计频率, c_{XY} 和 c_{XZ} 分别表示洪峰流量与洪水总量的 Copula 密度函数和洪峰流量与洪水历时的 Copula 密度函数。随机生成 4 个 0~1 之间的随机数 (a, b, c, d) , 令 $y_0 = a$, 并使得

$$P_r(X \leq x_1 | Y = y_0) = \int_0^{x_1} c_{XY}(x, y_0) dx = b \quad (8)$$

由式(8)可求得 x_1 , 把 x_1 代入

$$P_r(Y \leq y_1 | X = x_1) = \int_0^{y_1} c_{XY}(x_1, y) dy = c \quad (9)$$

$$P_r(Z \leq z_1 | X = x_1) = \int_0^{z_1} c_{XZ}(x_1, z) dz = d \quad (10)$$

可分别求得 y_1 和 z_1 。重复以上步骤可以得到任意多组三维模拟值 (x, y, z) , 然后利用边缘概率分布函数的反函数将 (x, y, z) 转换成其所对应的物理值, 得到任意多个三变量洪水事件, 从而利用模拟的洪水事件进行概率的估计和分析。

2 结果与分析

以淮河干流鲁台子水文站多年(1954—1998 年)洪水观测资料为研究对象, 验证 Copula 熵方法的有效性。鲁台子水文站位于淮河中游, 控制面积 88 630 km², 多年平均流量 693 m³/s, 资料长度 45 a。为了充分利用观测数据, 采用超阈值法选取洪水事件。流量超过 2 000 m³/s 的时刻, 定义为洪水事件的起始时间, 流量低于 2 000 m³/s 的时刻, 定义为洪水事件的结束时间。为了保证洪水事件的独立性, 规定 2 个洪水事件之间的间隔不小于 3 d, 否则, 两场洪水将合并为一场洪水。根据该定义方法, 共从原始观测资料中提取出 56 场洪水事件。

2.1 边缘分布拟合

分别用 P-Ⅲ 分布、广义帕累托分布 (Generalized Pareto Distribution, GPD)、广义极值分布 (Generalized Extreme Value distribution, GEV)、Gamma 分布和指数分布对观测值进行拟合, 并通过计算归一化均方差 (Normalized Mean Square Error, NMSE) 与经验分布进行比较。结果表明 P-Ⅲ 分布的拟合效果最好 (表 1)。图 1

表 1 各分布函数的拟合优度

Table 1 Goodness-of-fit for different distribution functions

变量	NMSE				
	P-Ⅲ	GEV	Gamma	GPD	指数分布
P	0.981	0.479	0.877	0.977	0.795
V	0.986	0.377	0.951	0.968	0.850
D	0.967	0.577	0.953	0.955	0.951

为 P-III 理论分布曲线与经验点据的拟合情况。

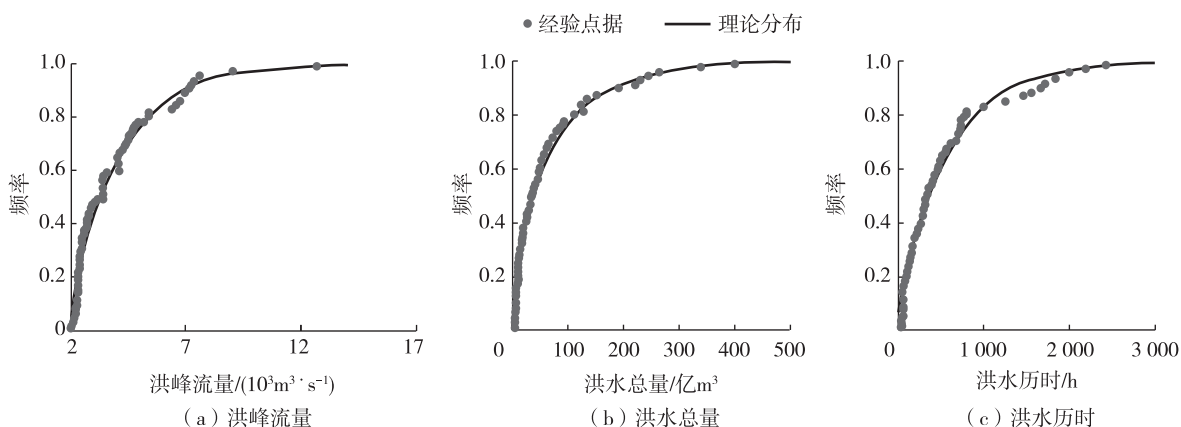


图1 洪水事件三要素的理论分布曲线与经验点据

Fig. 1 Theoretical distribution curves and empirical cumulative distribution functions of flood variables

2.2 Copula 密度函数计算

按照 1.3 中方法进行系数估计,当式(7)中的 n 分别取 2 和 3 时,计算得到的各系数见表 2。当 $n=3$ 时,系数 α_3 和 β_3 非常小,表明 $\varepsilon(u, v)$ 只需展开到 $n=2$ 即可。于是可得到 3 个二元 Copula 密度函数:

$$\left\{ \begin{aligned} c_{PV}(u, v) &= \frac{1.26u + 34.3u^2 + 1.26v + 34.3v^2}{\int_0^1 \int_0^1 (1.26u + 34.3u^2 + 1.26v + 34.3v^2) dudv} \\ c_{PD}(u, w) &= \frac{1.53u + 19.8u^2 + 1.53w + 19.8w^2}{\int_0^1 \int_0^1 (1.53u + 19.8u^2 + 1.53w + 19.8w^2) dudw} \\ c_{VD}(v, w) &= \frac{0.463v + 320v^2 + 0.463w + 320w^2}{\int_0^1 \int_0^1 (0.463v + 320v^2 + 0.463w + 320w^2) dv dw} \end{aligned} \right. \quad (11)$$

则 Copula 累积概率分布函数可表示为

$$\left\{ \begin{aligned} C_{PV}[F_1(x), F_2(y)] &= \int_0^{F_1(x)} \int_0^{F_2(y)} c_{PV}(u, v) dudv \\ C_{PD}[F_1(x), F_3(z)] &= \int_0^{F_1(x)} \int_0^{F_3(z)} c_{PD}(u, w) dudw \\ C_{VD}[F_2(y), F_3(z)] &= \int_0^{F_2(y)} \int_0^{F_3(z)} c_{VD}(v, w) dv dw \end{aligned} \right. \quad (12)$$

利用得到的累积概率分布函数及联合重现期公式得到图 2 中的联合重现期等值线(重现期单位为 a)。联合重现期表示其中一个变量超过某一定值的概率,其公式见文献[3],本文平均每年发生的洪水次数取为 56/45。

2.3 三变量洪水事件模拟

由于使用了超阈值方法进行洪水事件的选择,所以每年发生洪水事件的次数也需要进行估计。笔者使用经验频率来估计每一年洪水的发生次数。在 45 a(1954—1998)间,没有发生超阈

表 2 两变量 Copula 函数系数

Table 2 Parameters of bivariate copula functions

系数	(P, V)		(P, D)		(V, D)	
	$n=2$	$n=3$	$n=2$	$n=3$	$n=2$	$n=3$
α_1	1.26	1.26	1.53	1.53	0.463	0.463
α_2	34.3	34.3	19.8	19.8	320	320
α_3		1.24×10^{-5}		-2.52×10^{-5}		-1.79×10^{-5}
β_1	1.26	1.26	1.53	1.53	0.463	0.463
β_2	34.3	34.3	19.8	19.8	320	320
β_3		-6.71×10^{-6}		3.66×10^{-5}		1.08×10^{-5}
λ	-71.2	-71.2	-42.7	-42.7	-641	-641

值洪水的年数是 10,发生 1 次的有 14 a,发生 2 次的有 17 a,发生 3 次的有 6 a,即 $P(n=0)=10/45$, $P(n=1)=17/45$, $P(n=2)=12/45$, $P(n=3)=6/45$ 。结合 1.3 节中的 Gibbs 采样方法,得到 1 000 a 的模拟洪水事件(图 2)。

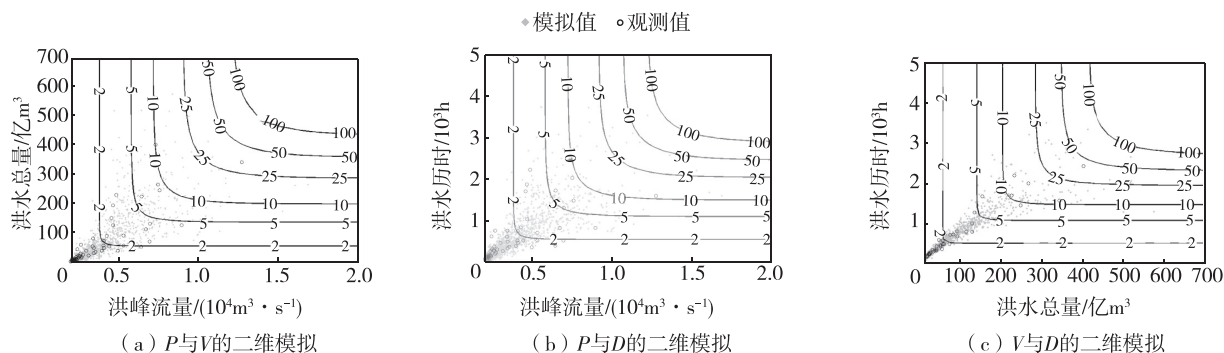


图 2 1 000 a 的洪水模拟事件与观测值以及两变量联合重现期等值线

Fig. 2 Observations and simulated flood events with a 1000-year return period and bivariate joint return period contours

2.4 洪水事件的拟合优度评估

为定量估计 Copula 熵方法与 Gibbs 采样方法的有效性,分别计算观测值与模拟值的经验频率,并进行比较。其中,三变量经验频率定义为^[15-16]

$$F(x, y, z) = P_r(X \leq x, Y \leq y, Z \leq z) = \frac{K - 0.44}{m + 0.12} \quad (13)$$

式中: K ——满足的样本个数, m ——总的样本个数。

两变量经验频率的定义与式(13)相似。利用卡方检验与 K-S(Kolmogorov-Sminov)检验来进行模拟值与实测值的比较。表 3 给出了 2 种检验方法的计算结果,如果卡方统计值小于临界值, K-S 检验的 P 值大于 0.05,则表明在 5% 的显著水平下通过假设检验,说明拟合情况较好,反之则不通过假设检验,说明拟合情况无法达到要求。

计算结果表明,模拟值在 5% 的显著性水平下与样本的拟合较好,能够反映实测洪水的统计特征。

3 结论与展望

a. Copula 熵方法与 Copula 方法类似,可以通过任意边缘分布和变量间的相关性构建二维的联合概率分布函数,与单纯的 Copula 方法比较, Copula 熵方法省去了 Copula 方程类型选择这一过程,计算效率更高。

b. 结合 Gibbs 采样方法和二维 Copula 函数,可以利用二维联合概率分布函数和条件概率函数进行三维随机变量的模拟,从而实现对三变量洪水事件的概率模拟。以淮河干流鲁台子水文站为研究对象,应用上述方法结合 1954—1998 年间的洪水实测资料,对洪水事件进行了概率模拟,模拟结果和检验结果显示,利用 Copula 熵方法对二维联合分布进行拟合适用且有效。此外, Gibbs 采样方法可以实现对三变量洪水事件的随机模拟,模拟结果通过了拟合优度检验。该方法可以为该地区的水利工程规划、洪水风险评估以及水资源合理利用提供参考依据。

c. Copula 熵方法在水文领域中的应用还处于起步阶段,相关的研究还比较少,有必要对该方法进行更为深入的研究。本文仅对 Copula 熵方法进行了简单的介绍,并用一个实例进行了验证。 Copula 熵方法在其他地区及领域的有效性以及与传统 Copula 函数拟合效果的比较还需要进一步研究与分析。此外, Copula 在求解联合分布方程的过程中,不同约束条件对拟合结果的影响等问题也需要进一步讨论。

参考文献:

[1] 郭生练, 闫宝伟, 肖义, 等. Copula 函数在多变量水文分析计算中的应用及研究进展[J]. 水文, 2008, 28(3): 1-7. (GUO

表 3 卡方检验与 K-S 检验结果

Table 3 Results of Chi squared test and K-S test

变量	卡方统计值	临界值	K-S 统计值	P 值
P, V	0.0605	34.76	0.0714	0.9982
P, D	0.0969	34.76	0.0893	0.9727
V, D	0.1178	34.76	0.0714	0.9982
P, V, D	0.0788	33.10	0.0714	0.9982

- Shenglian, YAN Baowei, XIAO Yi, et al. Multivariate hydrological analysis and estimation[J]. Journal of China Hydrology, 2008, 28(3): 1-7. (in Chinese))
- [2] 宋松柏, 蔡焕杰, 金菊良, 等. Copulas 函数及其在水文中的应用[M]. 北京: 科学出版社, 2012.
- [3] SALVADORI G, DE MICHELE C. Frequency analysis via Copulas: theoretical aspects and applications to hydrological events [J]. Water Resources Research, 2004, 40(12): 1-17.
- [4] ZHANG L, SINGH V P. Bivariate flood frequency analysis using the Copula method [J]. Journal of Hydrologic Engineering, 2006, 11(2): 150-164.
- [5] 肖义, 郭生练, 熊立华, 等. 一种新的洪水过程随机模拟方法研究[J]. 四川大学学报(工程科学版), 2007, 39(2): 55-60. (XIAO Yi, GUO Shenglian, XIONG Lihua, et al. A new random simulation method for constructing synthetic flood hydrographs[J]. Journal of Sichuan University(Engineering Science Edition), 2007, 39(2): 55-60. (in Chinese))
- [6] 李天元, 郭生练, 刘章君, 等. 基于峰量联合分布推求设计洪水[J]. 水利学报, 2014, 45(3): 269-276. (LI Tianyuan, GUO Shenglian, LIU Zhangjun et al. Design flood estimation based on bivariate joint distribution of flood peak and volume [J]. Journal of Hydraulic Engineering, 2014, 45(3): 269-276. (in Chinese))
- [7] RENARD B, LANG M. Use of a Gaussian Copula for multivariate extreme value analysis: some case studies in hydrology [J]. Advances in Water Resources, 2007, 30: 897-912.
- [8] 李计, 李毅, 宋松柏, 等. 基于 Copulas 函数的多维干旱变量联合分布[J]. 自然资源学报, 2013, 28(2): 312-320. (LI Ji, LI Yi, SONG Songbai, et al. Multivariate joint distributions of drought variables based on Copulas function[J]. Journal of Natural Resources, 2013, 28(2): 312-320. (in Chinese))
- [9] 王栋, 朱元胜. 最大熵原理在水文水资源科学中的应用[J]. 水科学进展, 2001, 12(3): 424-430. (WANG Dong, ZHU Yuansheng. Principle of maximum entropy and its application in hydrology and water resources[J]. Advances in Water Science, 2001, 12(3): 424-430. (in Chinese))
- [10] CHU B. Recovering Copulas from limited information and an application to asset allocation[J]. Journal of Banking and Finance, 2011, 35: 1824-42.
- [11] AGHAKOUCHAK A. Entropy-Copula in hydrology and climatology [J]. Journal of Hydrometeorology, 2014, 15: 1-13.
- [12] GEMAN S, GEMAN D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 20: 25-62.
- [13] LI F, VAN GELDER P H A J M, RANASINGHE R, CALLAGHAN D P, JONGEJAN R B. Probabilistic modelling of extreme storms along the Dutch coast[J]. Coastal Engineering, 2014, 86(4): 1-13.
- [14] CALLAGHAN D P, NIELSEN P, SHORT A, et al. Statistical simulation of wave climate and extreme beach erosion[J]. Coastal Engineering, 2008, 55(5): 375-390.
- [15] SALVADORI G, DE MICHELE C. Frequency analysis via Copulas: theoretical aspects and applications to hydrological events [J]. Water Resources Research, 2004, 40: 1-17.
- [16] ZHANG L, SINGH V P. Trivariate flood frequency analysis using the Gumbel-Hougaard Copula[J]. Journal of Hydrological Engineering, 2007, 12: 431-439.