Impact of Data Length on the Uncertainty of Hydrological Copula Modeling

X. Tong[1]; D. Wang[2]; V. P. Singh, F.ASCE[3]; J. C. Wu[4]; X. Chen[5]; and Y. F. Chen[6]

Abstract: Three Archimedean copulas were employed to model annual maximum flood peak data of different lengths. Estimation methods based on ranks were employed for parameter estimation. Marginals were modeled with the generalized extreme value (GEV) distribution. Then, uncertainty in modeling results was investigated with the change in data length. The joint and conditional return periods were also analyzed with the selected copula model to see how it varied with data length. Results showed that the accuracy of modeling deteriorated with the decrease in data length and that the best-fitting copula model depended on the data length. The uncertainty of modeling results may be due to the uncertainty of the flow itself when the data length is shortened. The data length has a negative effect not only on copula modeling but may also have an adverse effect on the marginal, which is an important factor when using a copula model to do bivariate analysis. **DOI: 10.1061/(ASCE)HE.1943-5584.0001039.**

Author keywords: Archimedean copulas; Hydrological time series; Data length; Uncertainty.

Introduction

In recent years, copulas have been used for multivariate hydrologic modeling, such as frequency analysis of flood volume, duration, and magnitude (Zhang and Singh 2006; Parent et al. 2014); drought modeling (Song and Singh 2010; Kao and Govindaraju 2010; Ma et al. 2013); modeling the dependence between storm or rainfall variables (Zhang and Singh 2007; Vandenberghe et al. 2010); streamflow simulation (Lee and Salas 2011); and assessing the interaction between chlorophyll *a* and environmental variables (Wang et al. 2012). The main advantage of the copula method is that it can capture the dependence of variables irrespective of their marginal distributions. Copulas have also been discussed relating to uncertainty in many areas, including probabilistic uncertainty addressed by copulas (Kumar 2011), precipitation uncertainty modeling (Bárdossy and Pegram 2009; AghaKouchak et al. 2010), and expression of uncertainty measured with copulas (Possolo 2010; Serinaldi 2013).

Data of sufficient length are needed for dependence modeling in hydrology, whether a copula or any other method is used. Long-term data are essential for reliable modeling and accurate parameter estimation. However, in developing countries, it is difficult to obtain data of sufficient length. The question then arises: What is the influence of data length on the modeling results? Therefore, it is important to analyze the impact of data length on the uncertainty in the copula modeling results. Xia et al. (2004) investigated the impact of data length on the uncertainty of a land surface model—i.e., the chameleon surface model—and found that different data lengths were required for obtaining optimal parameters. Genest et al. (2009) investigated the effect of sample size on the results of goodness of fit for various combinations of the degrees of dependence and families of copulas. Su and Tung (2013) evaluated the uncertainty due to sampling errors in flood-damage mitigation and showed that long data length improved the reliability of estimators and reduced sampling error and uncertainty. Hao and AghaKouchak (2014) used a nonparametric copula as a replacement of the parametric copula (Hao and AghaKouchak 2013) to establish a multi-index drought monitoring model and found that the model needed at least 30 years of data to avoid bias. However, it is not clear how the results of the Archimedean copula modeling in hydrology are influenced by the data length. Although uncertainty can be attributed to many factors, including physical measurements, the uncertainty in this paper is considered from the perspective of statistical inference. Therefore, the objective of this paper is to determine how the hydrological data lengths impact the bivariate Archimedean copula modeling results. The Archimedean copulas with two-parameter estimation methods will be used to model the bivariate distributions, and the best-fitting copula model will be selected to analyze the joint and conditional return periods of the data in order to see how the data length would affect the results.

Methodology

Archimedean Copulas

A copula is a function linking a multivariate probability distribution to its marginal probability distributions (Nelsen 1999;

[1]Key Laboratory of Surficial Geochemistry, Ministry of Education, Dept. of Hydrosciences, School of Earth Sciences and Engineering, State Key Laboratory of Pollution Control and Resource Reuse, Nanjing Univ., Nanjing 210046, China.

[2]Professor, Key Laboratory of Surficial Geochemistry, Ministry of Education, Dept. of Hydrosciences, School of Earth Sciences and Engineering, State Key Laboratory of Pollution Control and Resource Reuse, Nanjing Univ., Nanjing 210046, China (corresponding author). E-mail: wangdong@nju.edu.cn

[3]Professor, Dept. of Biological and Agricultural Engineering and Zachry Dept. of Civil Engineering, Texas A&M Univ., College Station, TX 77843.

[4]Professor, Key Laboratory of Surficial Geochemistry, Ministry of Education, Dept. of Hydrosciences, School of Earth Sciences and Engineering, State Key Laboratory of Pollution Control and Resource Reuse, Nanjing Univ., Nanjing 210046, China.

[5]Professor, State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, School of Hydrology and Water Resources, Hohai Univ., Nanjing 210098, China.

[6]Professor, State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, School of Hydrology and Water Resources, Hohai Univ., Nanjing 210098, China.

Note. This manuscript was submitted on January 2, 2014; approved on June 10, 2014; published online on August 6, 2014. Discussion period open until January 6, 2015; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Hydrologic Engineering***, © ASCE, ISSN 1084-0699/05014019(10)/$25.00.

Salvadori and De Michele 2013). For a bivariate case, let X and Y be two random variables with cumulative distribution functions (CDFs) $F_X(x) = U$ and $F_Y(y) = V$; here, U and V are uniformly distributed random variables. Then, a two-dimensional (2D) copula can be constructed as

$$F_{X,Y}(x, y) = C_\theta[F_X(x), F_Y(y)] = C_\theta(U, V) \tag{1}$$

where $F_{X,Y}(x, y)$ = bivariate distribution of X and Y; and C_θ = copula function with parameter θ.

The Archimedean copulas have been widely used in hydrology, for they can be easily constructed; there is a great variety of copulas that belong to this family, and they possess several desirable properties (Nelsen 1999). Let a generating function $\varphi(t)$ be a continuous, strictly decreasing function that satisfies $\varphi(1) = 0$ and $\varphi^{-1}(t) = 0$ when $\varphi(0) \le t \le \infty$, $t = u$, v. Then, the Archimedean copula can be expressed as

$$C_\theta(u, v) = \varphi^{-1}[\varphi(u) + \varphi(v)], \quad 0 < u, \quad v < 1 \tag{2}$$

Commonly used are three copulas, including the Gumbel-Hougaard (G-H) copula, the Clayton copula, and the Frank copula, which are from the Archimedean copula family (Genest and Favre 2007). These three copulas were applied in this study. The corresponding $C_\theta(u, v)$ for the three copulas can be expressed as

$$\begin{aligned}
&\text{Clayton} \quad [\max(u^{-\theta} + v^{-\theta} - 1, 0)]^{-1/\theta}, \quad \theta \in [-1, \infty)/\{0\} \\
&\text{G-H} \quad \exp(-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta}), \quad \theta \in [1, \infty) \\
&\text{Frank} \quad -1/\theta \ln[1 + (e^{-\theta u - 1})(e^{-\theta v - 1})/(e^{-\theta - 1})], \quad \theta \in (-\infty, +\infty)
\end{aligned} \tag{3}$$

Parameter Estimation

The copula parameter estimation methods can be divided into three types: fully parametric, semiparametric, and nonparametric. Fully parametric estimation contains the classical maximum likelihood estimator (MLE) and the inference function for marginal (IFM) (Joe 1997) method, among others. These methods assume parametric copula distribution and its marginal distributions. Semiparametric estimation (SPE) is derived from parametric estimation, which is expressed as

$$l(\theta) = \sum_{i=1}^{n} \log[c_\theta(F_X(x_i), F_Y(y_i)] \tag{4}$$

where $F(x_i)$ and $F_Y(y_i)$ = empirical CDFs of random variables X and Y. Because SPE substitutes the assumed marginal CDFs of the random variables with their empirical CDFs, it can be regarded as a special case of the IFM method. Kim et al. (2007) have shown that SPE is superior to the IFM method in most situations when marginal distributions are unknown, which is the case in this study. Nonparametric estimation methods include estimation from Kendall's tau (EKT). Relationships between Kendall's tau and parameter θ for the three copulas can be calculated as follows:

$$\begin{aligned}
&\text{Clayton} \quad \tau = \theta/(\theta + 2) \\
&\text{G-H} \quad \tau = 1 - 1/\theta \\
&\text{Frank} \quad \tau = 1 - 4/\theta + D_1(\theta)/\theta
\end{aligned} \tag{5}$$

where τ = Kendall's tau; and $D_1()$ = first Debye function.

Another nonparametric estimation method is based on the kernel estimator (Chen and Huang 2007).

For discussing the impact of hydrological data length on model results, rank-based parameter estimation was used in this paper. To that end, SPE and EKT were chosen because the bivariate distribution of two variables should not be affected by their marginal distributions, and models based on ranks are more reasonable than those inferred from some assumed distributions.

Goodness of Fit for Various Data Lengths

In order to evaluate how data length impacts the uncertainty of copula modeling, one must first determine the uncertainty of the time series itself due to the change in data length. The uncertainty of a system is related to the information it yields. Gaining information means reducing freedom of choices—i.e., reducing uncertainty. In another words, as information is added to a system, the uncertainty of the system decreases. Therefore, it can be inferred that if a system exhibits more uncertainty, more information should be collected in order to better describe the system. Entropy represents the uncertainty of a system before receiving any information. As a measurement of uncertainty, entropy has been used in hydrology (Wang et al. 2007, 2009; Wang 2010; Zeng et al. 2012; Singh 2013; Hao and Singh 2013). This study calculated marginal entropy to see how the uncertainty of the two variables changes. The entropy for a random variable X is expressed as

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i) \tag{6}$$

where $p(x_i)$ = probability of $X = x_i$. Joint entropy is also implemented to determine the variation of total uncertainty contained in the combination of the two variables with the variation in data length. The joint entropy of X and Y can be denoted as

$$H(X, Y) = -\sum_{i,j=1}^{N} P_{ij} \log P_{ij} \tag{7}$$

where $H(X, Y)$ = joint entropy of X and Y; and P_{ij} = joint probability of $X = x_i$, $Y = y_i$. If X and Y are independent, then the joint entropy of the two variables would be

$$H(X, Y) = H(X) + H(Y) \tag{8}$$

Because in many cases the two variables can be dependent to some extent, here, the mutual information is discussed, which is a measure of the dependence between the two variables shown as follows:

$$T(X, Y) = H(X) + H(Y) - H(X, Y) \tag{9}$$

where $T(X, Y)$ = mutual information between X and Y.

In order to investigate the effect of data length on modeling results, data of different lengths from the original data were selected. Then, using the aforementioned estimation methods, the simulation for each bivariate data length was carried out as follows: (1) calculate parameter θ of the three Archimedean copulas using the EKT estimation method, (2) calculate parameter θ using the SPE method, and (3) construct copula models with the calculated θ. Both Akaike information criterion (AIC) and formal goodness-of-fit methods were used.

The AIC was applied to determine the goodness of fit of copula modeling (Zhang and Singh 2006). It can be expressed in two ways

$$\text{AIC} = 2m - 2\log(L) \tag{10}$$

$$\text{AIC} = n\log(\text{MSE}) + 2m \tag{11}$$

where L = maximum value of the likelihood of the model; n = the length of data; m = number of parameters, which in this

paper equals 1; and MSE = mean square error, which can be written as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(p_{ei} - p_i)^2 \tag{12}$$

where p_i = theoretical joint probability; and p_{ei} = empirical joint probability (Gringorten 1963) given as

$$p_{ei} = p(X \le x_i, Y \le y_i) = \frac{\sum_{j=1}^{n} \text{Number of } (x_j \le x_i, y_j \le y_i) - 0.44}{n + 0.12} \tag{13}$$

Following Zhang and Singh (2006), Eq. (11) was chosen to evaluate the goodness of fit of the copula models. The modeling result is most accurate when the corresponding AIC value reaches the minimum.

In order to better justify how well the copula models perform, formal goodness-of-fit tests were also employed. The Cramér-von Mises functional S_n (Genest et al. 2009) was used for testing, and the corresponding approximate P-value was deduced through the Monte Carlo method (Genest et al. 2009). A small value of S_n represents good modeling performance. With the AIC value and S_n value derived from different data lengths and estimators, the impacts of data lengths on the Archimedean copula modeling results were compared and analyzed.

Marginal Distribution

The generalized extreme value (GEV) distribution was used to model the marginals. Its cumulative probability distribution and density function can be expressed, respectively, as

$$F(x;\mu,\sigma,k) = \exp\left\{-\left[1 + k\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/k}\right\} \tag{14}$$

$$f(x;\mu,\sigma,k) = \frac{1}{\sigma}\left[1 + k\left(\frac{x-\mu}{\sigma}\right)\right]^{(-1/k)-1} \times \exp\left\{-\left[1 + k\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/k}\right\} \tag{15}$$

where μ = location parameter; σ = scale parameter; and k = shape parameter.

Joint Return Period and Conditional Return Period

The joint return periods of two data sets are of practical use in hydrological bivariate modeling (Gräler et al. 2013; De Michele et al. 2013; Volpi and Fiori 2014). When the change in data length affects copula modeling, it consequently may have an impact on the bivariate distribution of the two data sets. Further, because the joint return periods are derived from the joint distribution, one would wonder how the joint return periods of two data sets will change when their bivariate distributions are affected by the change in data length. In this paper, the marginal return periods are considered to be $T_X(x)$ and $T_Y(y)$, respectively. Their joint return period would be

$$T_{X,Y}(x,y) = \frac{1}{F_{X,Y}(X > x \cup Y > y)} = \frac{1}{1 - F_{X,Y}(x,y)} = \frac{1}{1 - C_\theta\left\{\left[1 - \frac{1}{T_X(x)}\right], \left[1 - \frac{1}{T_Y(y)}\right]\right\}} \tag{16}$$

where $C_\theta(\{1 - [1/T_X(x)]\}, \{1 - [1/T_Y(y)]\})$ = cumulative distribution of X and Y obtained from copula modeling.

The return period of variate $X > x$ under the condition of variate $Y > y$ can be deduced by the conditional return period expression as

$$T(X > x/Y > y) = 1/F(X > x/Y > y) = \frac{F_Y(Y > y)}{F(X > x, Y > y)} = \frac{1 - F_Y(y)}{1 - F_X(x) - F_Y(y) + C_\theta[F_X(x), F_Y(y)]} \tag{17}$$

where $C_\theta[F_X(x), F_Y(y)]$ = cumulative probability of X and Y calculated by the fitted copula model; and $F_X(x)$ and $F_Y(y)$ = cumulative probabilities of X and Y obtained from their marginal distributions.

Application of Copula Models

Data Selection and Description

Annual maximum flood magnitude (AMFM) data from Cuntan Station and Yichang Station were collected. Both stations are on the Yangtze River, which is the largest river in China. The data sets have a length of 109 years from 1893 to 2004 (1942 to 1944 not included) and were plotted against years, as shown in Fig. 1. The annual maximum floods at Yichang Station and Cuntan Station exhibited similar changing variations throughout the 109-year record.

When selecting the data, the possibility that data from different parts of the original series will bias the modeling results should be ruled out. For this purpose, 80-year-long, 60-year-long, 50-year-long, 40-year-long, and 30-year-long data were selected from different parts of the original 109-year record from Yangtze River. Then, the three Archimedean copulas were employed with empirical-based and Kendall's tau–based parameter estimators. Table 1 gives the AIC values of the three copula models. The minimum AIC values are marked in boldface. For both estimators, for randomly extracted data lengths of 80, 60, and 50 years, the best-fitting copula models almost remained the same. However, for data lengths of 40 and 30 years, the best-fitting copula models varied from the G-H copula to the Frank copula for the empirical-based estimator and from the G-H copula to the Frank copula and the Clayton copula for the Kendall's tau–based estimator. However, the best-fitting copula models remained the same for over 60% of all

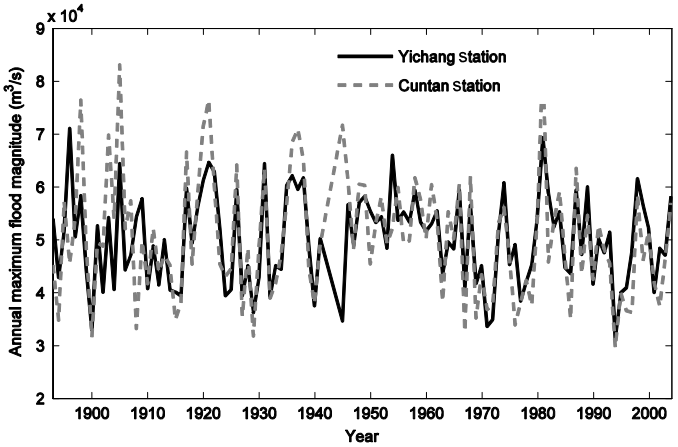

Fig. 1. Plot of the AMFM data sets from the Yangtze River

Table 1. Akaike Information Criterion Values of Copula Models for Different Parts of Original Data Length of the Yangtze River

Data length	Period (years)	Empirical-based			Kendall-based		
		Clayton	Frank	G-H	Clayton	Frank	G-H
80	1893–1975	−555.82	**−702.42**	−643.44	−605.19	**−697.24**	−675.16
	1895–1977	−548.88	**−702.39**	−635.99	−601.92	**−696.03**	−670.22
	1899–1981	−526.35	**−694.96**	−702.26	−583.33	−691.03	**−704.27**
	1902–1984	−509.02	**−699.74**	−692.95	−585.88	−695.17	**−700.29**
	1904–1986	−497.85	−686.1	**−694.49**	−577.23	−682.32	**−700.14**
	1907–1989	−497.09	**−700.78**	−651.6	−591.9	**−698.23**	−677.37
	1910–1992	−489.92	**−708.71**	−639.22	−596.48	**−705.79**	−671.57
	1912–1994	−517.86	**−713.31**	−656.55	−600.83	**−710.79**	−682.82
	1917–1999	−519.73	**−706.71**	−640.59	−602.42	**−707.45**	−671.72
	1922–2004	−516.06	**−698.24**	−630.14	−593.19	**−697.24**	−663.36
60	1893–1955	−421.66	**−492.03**	−469.18	−433.38	**−487.99**	−482.48
	1897–1959	−416.91	**−496.59**	−477.58	−432.7	**−493.05**	−491.81
	1902–1964	−406.71	**−511.52**	−503.7	−434.15	−507.5	**−519.8**
	1904–1966	−401.75	−493.59	**−506.44**	−426.67	−490.72	**−516.91**
	1909–1971	−388.35	**−539.64**	−465.18	−461.51	**−534.09**	−496.03
	1915–1977	−369.72	**−517.76**	−442.36	−451.32	**−510.83**	−472.54
	1923–1985	−381.59	**−530.5**	−483.74	−454.46	**−524.16**	−494.42
	1931–1993	−387.11	**−521.03**	−460.54	−465.98	**−515.86**	−477.8
	1936–1998	−393.61	**−510.52**	−458.39	−453.05	**−506.24**	−470.81
	1945–2004	−403.53	**−509.06**	−446.81	−452.69	**−504.33**	−462.28
50	1893–1943	−352.95	**−406.14**	−403.26	−359.9	−403.55	**−404.14**
	1902–1954	−335.67	**−401.26**	−397.85	−352.02	−398.18	**−403.28**
	1907–1959	−333.58	**−402**	−387.16	−362.16	**−414.47**	−405.18
	1912–1964	−342	**−439.13**	−382.53	−380.31	**−432.09**	−406.02
	1915–1967	−333.55	**−438.02**	−378.95	−377.33	**−429.74**	−400.95
	1927–1979	−313.3	**−433.35**	−366.89	−372.64	**−422.16**	−385.66
	1930–1982	−326.82	**−409.89**	−372.85	−385.57	**−405.52**	−380.03
	1937–1989	−326.82	**−409.89**	−372.86	−385.57	**−405.52**	−380.03
	1945–1994	−366.14	**−422.69**	−374.63	−398.17	**−417.04**	−381.39
	1955–2004	−314.09	**−392.79**	−371.05	−355.99	**−395.88**	−384.91
40	1893–1932	−276.02	−304.9	**−318.76**	−279.1	−304.49	**−319.36**
	1897–1936	−273.53	−307.17	**−323.72**	−278.3	−307.02	**−324.64**
	1904–1946	−259.42	−311.82	**−322.5**	−273.84	−309.69	**−321.38**
	1911–1953	−257.46	**−316.46**	−302.76	−281.48	**−314.26**	−308.19
	1918–1960	−270.44	**−331.1**	−290.22	−297.51	**−326.24**	−306.62
	1927–1969	−281.46	**−341.66**	−298.25	−306.57	**−332.8**	−307.22
	1937–1979	−259.34	**−331.69**	−286.11	−303.44	**−323.18**	−297.51
	1953–1992	−251.15	**−310.48**	−278.97	−298.79	**−312.44**	−293.34
	1960–1999	−237.35	**−294.7**	−285.81	−267.93	**−299.88**	−298.43
	1965–2004	−227.95	−277.91	**−281.49**	−255.49	−284.13	**−289.82**
30	1893–1922	−193.79	−220.81	**−238.94**	−197.31	−220.23	**−235.27**
	1897–1926	−192.49	−222.69	**−246.54**	−198.23	−222.47	**−243.5**
	1902–1931	−192.42	−215.2	**−233.09**	−196.01	−215.15	**−232.73**
	1911–1940	−201.99	**−234.55**	−224	−221.2	**−237.4**	−232.08
	1919–1951	−197.2	**−236.56**	−219.96	−215.4	**−234.02**	−223.25
	1926–1958	−201.2	**−242.45**	−211.26	−214.02	**−234.81**	−221.11
	1935–1967	−208.59	**−229.92**	−225.44	−212.72	**−225.17**	−224.56
	1945–1976	−231.64	**−228.32**	−205.89	**−236.07**	−223.24	−208.79
	1953–1982	−198.59	**−223.22**	−209.99	**−224.19**	−223	−212.65
	1960–1989	−169.78	**−222.91**	−211.13	−204.63	**−226.89**	−221.85

Note: Bold values indicate the minimum AIC values.

the random samples. Even for the most unstable case—i.e., the data length of 30 years—the differences in the AIC values between each copula model were small. Thus, on the whole, it can be inferred that the modeling results for a given data length do not depend on the period that is selected from the original record.

The authors deducted 5 years each time from 109 to 20 years, and 18 different lengths of data sets were obtained. Table 2 provides the statistical characteristics, including mean, standard deviation, coefficient of skewness, and lag-1 autocorrelation relating to different selected lengths of data for both stations. The statistics of the data sets varied when the data length was reduced. For each data length, the extracted data had a small autocorrelation, which proved that the data could be assumed as independent and could be modeled using a copula model. Then, the correlation coefficients of the two sets of data were also calculated for different lengths, as given in Table 2. The AMFMs of the two stations always exhibited strong positive correlations, but the correlation coefficient decreased as the data length was reduced.

The marginal entropies of the AMFM data series of both Yichang Station and Cuntan Station were plotted against data lengths, as were the joint entropy and mutual information between the two series, as shown in Fig. 2. For each data length, the entropy

Table 2. Statistics of the Data Sets from the Yangtze River

Data length (years)	Yichang Station				Cuntan Station				Correlation coefficient
	Mean (cm)	Standard deviation	Coefficient of skewness	Autocorrelation	Mean (cm)	Standard deviation	Coefficient of skewness	Autocorrelation	
109	50,228	8,613.6	0.041	0.147	50,385	11,440.1	0.587	0.163	0.717
100	50,203	8,742.6	0.043	0.136	50,799	11,610.0	0.553	0.136	0.720
95	50,515	8,691.2	0.060	0.121	51,199	11,633.8	0.548	0.121	0.707
90	50,512	8,736.2	0.052	0.151	51,386	11,688.5	0.547	0.151	0.700
85	50,172	8,654.0	0.040	0.123	51,002	11,417.2	0.436	0.123	0.688
80	50,429	8,748.2	−0.010	0.108	51,616	11,363.2	0.393	0.108	0.686
75	50,765	8,536.4	0.024	0.067	51,985	11,464.8	0.365	0.067	0.670
70	50,896	8,578.1	−0.002	0.103	52,360	11,303.7	0.418	0.103	0.648
65	50,955	8,826.3	−0.017	0.107	52,418	11,546.7	0.430	0.107	0.644
60	50,640	9,084.2	0.073	0.094	52,308	11,923.4	0.444	0.094	0.645
55	50,220	9,210.2	0.127	0.090	52,105	12,377.9	0.479	0.090	0.660
50	49,702	9,447.5	0.267	0.096	51,804	12,760.6	0.540	0.096	0.659
45	50,036	9,358.7	0.286	0.082	51,591	12,747.7	0.575	0.082	0.734
40	49,480	9,385.8	0.392	0.014	50,903	12,732.2	0.679	0.014	0.712
35	50,009	9,190.8	0.303	0.036	51,566	12,864.2	0.688	0.036	0.678
30	50,610	9,210.4	0.243	0.076	52,400	13,159.6	0.607	0.076	0.650
25	48,932	8,876.2	0.564	−0.140	50,408	12,666.4	0.890	−0.140	0.568
20	49,650	8,952.8	0.495	−0.200	51,415	12,845.7	0.869	−0.200	0.484

of Yichang Station was larger than that of Cuntan Station, which means that the AMFM data series of Yichang Station contained more uncertainty than did Cuntan Station. As the data length became shorter, the marginal entropy decreased for both stations. This means that the AMFM data series of Yichang and Cuntan Stations become less uncertain and yielded less information as the data length was shortened. It can be deduced from Fig. 2 that the total uncertainty of the two stations, which is represented by the joint entropy, decreased as the data length was reduced. As mutual information decreased, the AMFMs of the two stations became less dependent when the data length shortened.

Copula Modeling Results

Intuitively, longer data lead to better modeling results—i.e., smaller AIC (S_n) values. However, it may be worth discussing how AIC (S_n) values of different Archimedean copulas change with the change in data length for different parameter estimators.

As given in Table 3, for the data of various lengths extracted from the original record, the three Archimedean copulas performed differently for the two estimation methods. The smallest values—i.e., the best-fitting copulas and estimation methods—are shown in boldface. The AIC value, in general, becomes larger as the data length becomes shorter, which is expected. For each data length, the best-fitting copula model and estimation method were not invariable. In general, SPE performed better than EKT, except for the data length of 70, 55, 45, and 40 years. When using SPE, the best-fitting copula model changed from the Frank copula to the G-H copula, as the data set length reduced. In the case of EKT, the best-fitting copula model exhibited the same variation, as did SPE. From another perspective, when using the Clayton copula, EKT yielded better results than SPE, whereas for the Frank copula, SPE performed better than EKT. When using the G-H copula, EKT performed better for a data length longer than 35 years, and SPE performed better for a data length of less than 35 years. Therefore, it

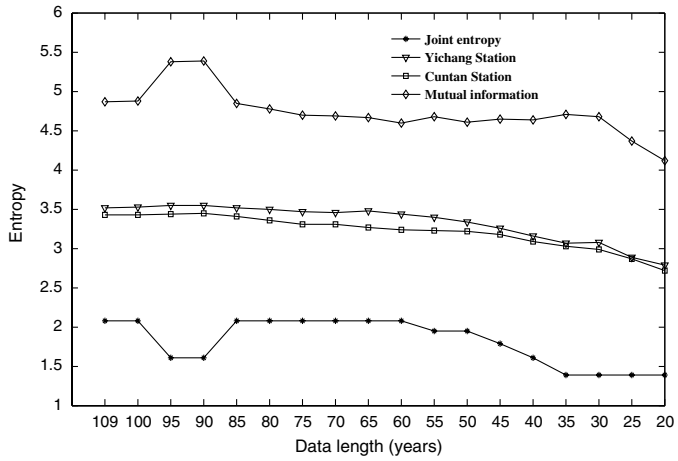

Fig. 2. Marginal entropy, joint entropy, and mutual information for various data lengths of data sets from the Yangtze River region

Table 3. Akaike Information Criterion Values of Copula Models for Data Sets from the Yangtze River Based on SPE and EKT Estimators

Data length (years)	SPE			EKT		
	Clayton	Frank	G-H	Clayton	Frank	G-H
109	−739.96	**−960.57**	−925.9	−809.28	−957.75	−959.02
100	−691.15	**−884.9**	−849.73	−750.68	−881.63	−879.86
95	−634.93	**−843.73**	−795.50	−705.48	−839.01	−826.12
90	−610.94	**−792.74**	−763.17	−671.8	−788.29	−784.11
85	−573.37	**−751.94**	−691.56	−632.14	−745.5	−725.63
80	−555.82	**−702.42**	−643.44	−605.19	−697.24	−675.16
75	−519.99	**−649.67**	−611.83	−552.7	−644.41	−636.56
70	−499.22	−590.13	−572.76	−515.83	−586.62	**−592.17**
65	−455.67	**−534.83**	−514.19	−470.23	−531.02	−530.99
60	−421.66	**−492.03**	−469.18	−433.38	−487.99	−482.48
55	−382.28	−438.1	−442.12	−390.52	−436.21	**−444.26**
50	−352.95	**−406.14**	−403.26	−359.90	−403.55	−404.14
45	−314.18	−354.8	−363.04	−319.1	−354.27	**−365.67**
40	−276.02	−304.9	−318.76	−279.1	−304.49	**−319.36**
35	−229.12	−262.84	**−281.43**	−235.1	−262.13	−278.75
30	−193.79	−220.81	**−238.94**	−197.31	−220.23	−235.27
25	−165.92	−178.24	**−188.65**	−166.06	−177.8	−185.62
20	−128.56	−133.2	**−141.15**	−127.54	−133.06	−137.67

Note: Bold values indicate the smallest values—i.e., the best-fitting copulas and estimation methods.

Table 4. Cramér-von Mises Statistic S_n and the Corresponding P-Value ($\alpha = 5\%$) for Data Sets from the Yangtze River Based on SPE and EKT

Data length (years)	SPE			EKT		
	Clayton copula	Frank copula	G-H copula	Clayton copula	Frank copula	G-H copula
109	0.1345 (0.0005)	**0.0180** (0.4940)	0.0258 (0.1683)	0.0677 (0.0015)	0.0191 (0.4191)	0.0181 (0.4161)
100	0.1133 (0.0005)	0.0175 (0.5559)	0.0245 (0.1683)	0.0601 (0.0005)	0.0187 (0.4371)	**0.0173** (0.4900)
95	0.1341 (0.0005)	**0.0161** (0.6908)	0.0257 (0.1713)	0.0619 (0.0005)	0.0176 (0.5440)	0.0177 (0.5110)
90	0.1165 (0.0005)	**0.0165** (0.6788)	0.0218 (0.3442)	0.0575 (0.0015)	0.0181 (0.5519)	0.0165 (0.6249)
85	0.1144 (0.0005)	**0.015** (0.8197)	0.0289 (0.1054)	0.0558 (0.0025)	0.0171 (0.6469)	0.0181 (0.5360)
80	0.0910 (0.0035)	**0.0158** (0.7917)	0.0311 (0.1114)	0.0473 (0.0065)	0.0177 (0.6229)	0.0197 (0.4261)
75	0.0860 (0.0045)	**0.0169** (0.7368)	0.0266 (0.1563)	0.0534 (0.0025)	0.0189 (0.5569)	0.0181 (0.5819)
70	0.0670 (0.0095)	0.0197 (0.5350)	0.0255 (0.2433)	0.0504 (0.0025)	0.0217 (0.4271)	**0.0186** (0.6119)
65	0.0698 (0.0085)	**0.0212** (0.4950)	0.0301 (0.1294)	0.0531 (0.0065)	0.0236 (0.3501)	0.0220 (0.3771)
60	0.0627 (0.0175)	**0.0196** (0.6399)	0.0294 (0.1653)	0.0493 (0.0065)	0.0226 (0.4491)	0.0226 (0.4031)
55	0.0629 (0.0215)	0.0223 (0.5190)	0.0223 (0.4331)	0.0512 (0.0075)	0.0251 (0.3771)	**0.0212** (0.5360)
50	0.0532 (0.0465)	**0.0184** (0.8097)	0.0205 (0.6059)	0.0443 (0.0245)	0.022 (0.6119)	0.0203 (0.6918)
45	0.0516 (0.0604)	0.0214 (0.6838)	0.019 (0.7168)	0.0439 (0.0135)	0.0231 (0.5390)	**0.0176** (0.8826)
40	0.0505 (0.0594)	0.0247 (0.5549)	0.0196 (0.7328)	0.0443 (0.0205)	0.0264 (0.4141)	**0.0195** (0.8506)
35	0.065 (0.0385)	0.0263 (0.5519)	**0.0169** (0.9476)	0.0514 (0.0195)	0.0282 (0.4281)	0.0198 (0.9236)
30	0.0626 (0.0584)	0.0301 (0.4431)	**0.0188** (0.9406)	0.0533 (0.0245)	0.032 (0.3861)	0.0232 (0.8706)
25	0.0438 (0.2522)	0.0321 (0.5699)	**0.0233** (0.8337)	0.0448 (0.1484)	0.0346 (0.4830)	0.0290 (0.7687)
20	0.0474 (0.2443)	—	—	0.0530 (0.1523)	0.0465 (0.2802)	0.0414 (0.4231)

Note: The P-value is given in parentheses. Bold values indicate the smallest S_n values.

can be concluded that the best-fitting copula model varies not only with the change in data length but also as different estimators are used.

The S_n values and the corresponding P-values in Table 4 exhibit the same changing pattern as the AIC values. However, the Clayton copula model was rejected for most of the data lengths, as shown by the approximate P-value, which indicates that the Clayton copula is not adequate for modeling the bivariate feature of the AMFMs of the Yangtze River.

In order to further evaluate the effect on modeling, the AIC values were plotted against data length for the three copulas and the two estimation methods, as shown in Fig. 3. For all the copulas, the AIC values increased with the decrease in data length, no matter which estimator—SPE or EKT—was used. In other words, the Clayton, Frank, and G-H copulas performed better for longer data lengths. The Frank copula performed better than the other two copulas for both SPE and EKT, followed by the G-H copula. The Clayton copula was found to yield the worst results. In addition, as the data length became shorter, the differences between the results of different copulas became blurred.

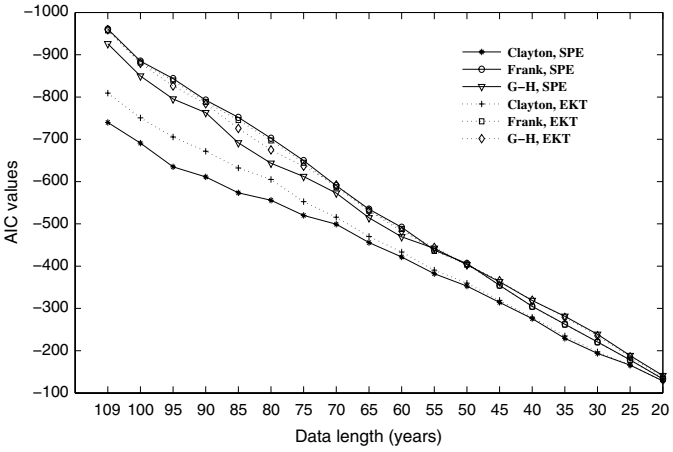

Fig. 3. Akaike information criterion values of copula models for data sets from the Yangtze River based on SPE and EKT

To illustrate the appropriateness of various copulas, the authors selected 109-year and 40-year data lengths from the original data and compared the joint cumulative probabilities of the empirical copula and the three copulas, as shown in Fig. 4. The figure demonstrates the inadequacy of the Clayton copula and the good performance of the Frank and G-H copulas.

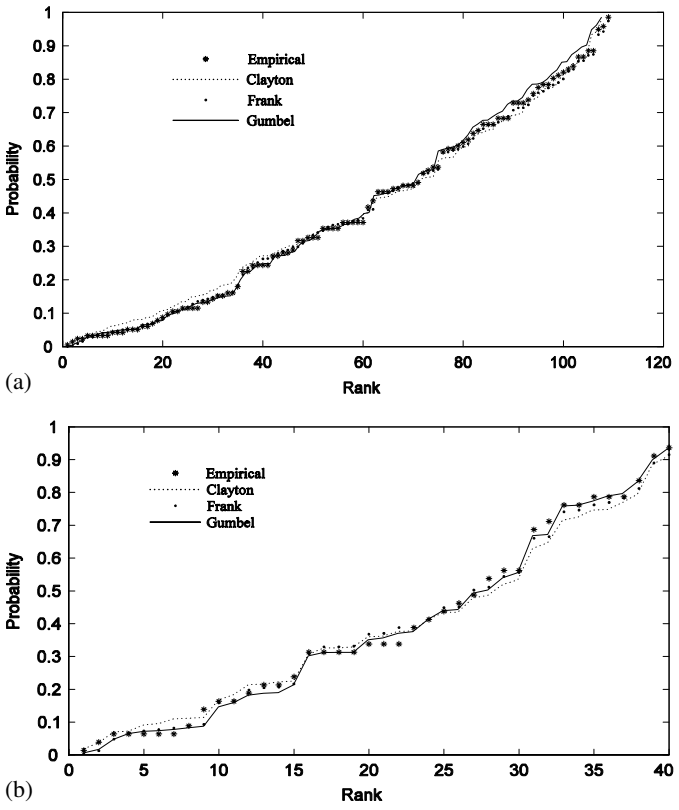

Fig. 4. Joint cumulative probability obtained from different copula models of the: (a) 109-year; (b) 40-year data sets from the Yangtze River region

Table 5. Parameters of GEV Distribution of the Data Sets from the Yangtze River

Data length (years)	Yichang Station				Cuntan Station			
	k	σ	μ	P-value	k	σ	μ	P-value
109	−0.280	8,521.680	47,171.042	0.525	−0.097	9,933.661	45,483.641	0.838
100	−0.281	8,649.342	47,105.307	0.509	−0.110	10,195.858	45,880.218	0.765
95	−0.278	8,562.299	47,421.843	0.706	−0.105	10,169.224	46,242.223	0.814
90	−0.280	8,623.798	47,410.238	0.869	−0.108	10,247.990	46,424.105	0.782
85	−0.283	8,553.879	47,111.207	0.524	−0.137	10,243.575	46,266.476	0.821
80	−0.301	8,740.865	47,407.615	0.908	−0.158	10,375.144	46,999.392	0.908
75	−0.288	8,465.885	47,766.557	0.778	−0.169	10,554.174	47,377.993	0.962
70	−0.296	8,557.001	47,912.861	0.772	−0.157	10,308.627	47,772.983	0.697
65	−0.309	8,844.068	47,941.456	0.719	−0.150	10,455.568	47,702.946	0.843
60	−0.288	8,942.983	47,450.117	0.533	−0.138	10,662.493	47,371.915	0.833
55	−0.265	8,921.665	46,888.927	0.664	−0.109	10,796.197	46,832.029	0.644
50	−0.186	8,669.886	45,949.891	0.682	−0.070	10,779.344	46,176.390	0.909
45	−0.170	8,497.109	46,253.144	0.662	−0.069	10,754.114	45,976.764	0.992
40	−0.104	8,149.282	45,427.966	0.447	−0.042	10,488.131	45,196.361	0.921
35	−0.173	8,395.651	46,310.340	0.497	−0.034	10,514.661	45,762.540	0.970
30	−0.213	8,665.602	47,069.602	0.921	−0.074	11,141.561	46,645.421	0.973
25	−0.097	7,684.769	45,119.350	0.919	−0.017	10,137.530	44,689.351	0.897
20	−0.145	8,091.505	45,982.356	0.965	−0.056	10,658.171	45,804.794	0.827

Note: Denoted P-value is calculated from the Kolmogorov-Smirnov test under a 95% confidence level.

Uncertainty Analysis and Return Periods

The uncertainty of the copula models may be due to the uncertainty of the data series itself. Fig. 2 shows that the marginal entropy decreases as data length shortens, as does the dependence. This suggests that longer data length contains more uncertainty and, as a result, is more reliable and can exhibit more natural flow characteristics. Because the three copula models represent the joint behavior of two random variables (i.e., AMFMs of Yichang and Cuntan Stations), when the total uncertainty of the two variables changes, it surely has some effect on the modeling results. As the data length is reduced, the total uncertainty—i.e., the joint entropy—of the AMFM series obviously decreases for the case of the Yangtze River. The effect on the modeling results can be seen from the variation in the choice of the best-fitting copula model and parameter estimator in the case of the Yangtze River.

As the joint behavior is affected by data length, it can also be of interest to see how the joint and conditional return periods change with data length. According to the previous discussion, the Frank copula and SPE were chosen for modeling the data sets from the Yangtze River. The marginals were modeled using GEV distributions. The bivariate distributions were calculated for various data lengths.

Table 5 provides the parameters of the marginals for each data length. The table shows that under a 95% confidence level, the GEV distributions can be accepted for each data length. The location parameter μ became smaller as the data length was shortened. This trend can also be detected in the mean value of the data sets

Table 6. Joint Return Period of the Data Sets from the Yangtze River

	Marginal return period (years)								
	10			100			1,000		
Data length (years)	Joint return period (years)	Flood peak of Yichang Station (cm)	Flood peak of Cuntan Station (cm)	Joint return period (years)	Flood peak of Yichang Station (cm)	Flood peak of Cuntan Station (cm)	Joint return period (years)	Flood peak of Yichang Station (cm)	Flood peak of Cuntan Station (cm)
109	6.39	61,397	82,373	51.81	65,574	73,199	501.87	69,207	95,541
100	6.44	61,530	82,719	51.89	66,214	73,463	501.96	69,433	95,273
95	6.40	61,747	83,360	51.82	66,629	73,713	501.89	69,652	96,229
90	6.39	61,814	83,553	51.80	66,891	73,786	501.86	69,733	96,265
85	6.39	61,345	81,217	51.80	66,102	73,041	501.86	69,105	92,000
80	6.40	61,692	80,945	51.82	66,655	72,798	501.89	69,164	90,663
75	6.37	61,780	81,158	51.77	67,144	73,113	501.83	69,329	90,449
70	6.32	61,971	81,570	51.69	67,324	73,081	501.74	69,415	91,281
65	6.33	62,281	82,423	51.71	67,665	73,163	501.76	69,646	92,632
60	6.31	62,264	83,716	51.67	68,008	74,267	501.72	70,255	94,910
55	6.32	62,006	85,902	51.68	68,381	75,131	501.74	70,590	99,252
50	6.29	61,889	88,577	51.65	68,621	79,646	501.70	72,741	105,225
45	6.39	62,134	88,395	51.80	68,400	80,744	501.86	73,345	105,126
40	6.37	61,782	89,052	51.78	67,713	85,603	501.83	75,233	108,043
35	6.26	61,960	90,550	51.59	68,545	80,151	501.64	72,944	110,516
30	6.15	62,560	90,092	51.43	69,743	78,403	501.46	72,477	106,910
25	5.92	60,656	89,535	51.09	67,069	83,806	501.11	73,637	110,731
20	5.72	61,514	89,050	50.81	68,346	81,262	500.82	73,131	106,903

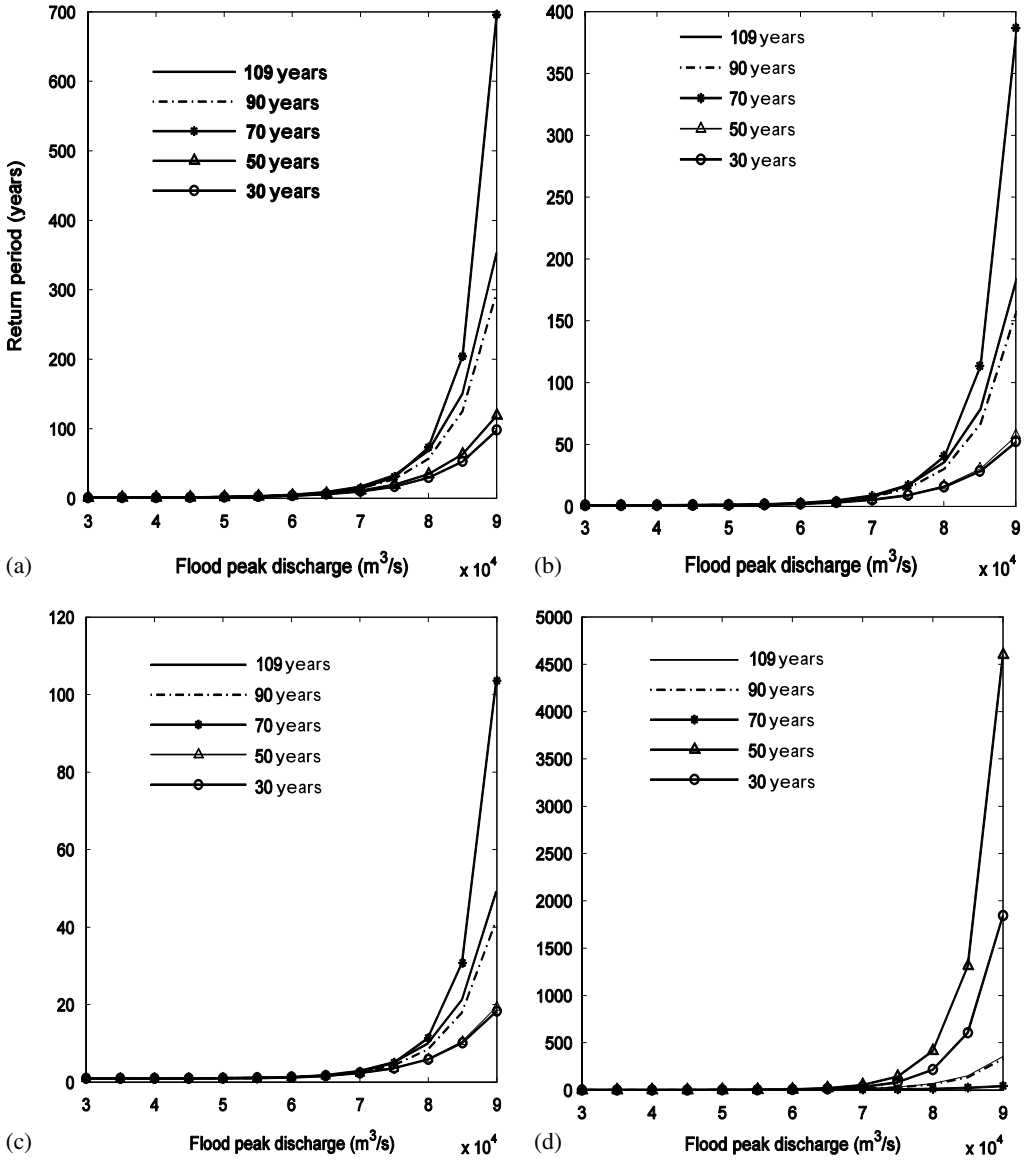

Fig. 5. Return periods of Cuntan Station under different conditions and different data lengths: (a) conditional return period of Cuntan Station given the AMFM of Yichang Station is 30,000 cm; (b) conditional return period of Cuntan Station given the AMFM of Yichang Station is 50,000 cm; (c) conditional return period of Cuntan Station given the AMFM of Yichang Station is 70,000 cm; (d) marginal return periods of Cuntan Station of different data lengths

provided in Table 2. The shape parameterincreased when the data length was reduced, which suggests a more right-skewed distribution of the data. The same quality can be found in the coefficient of the skewness in Table 2. For the data from Yichang Station, the scale parameter σ became larger when the data length was reduced, whereas for Cuntan Station, σ became smaller. It can be inferred that as the data length decreased, the AMFM values of Yichang Station became more scattered, whereas those of Cuntan Station became concentrated.

Marginal return periods of both stations were set as 10, 100, and 1,000 years. Table 6 provides the joint return periods of the two stations and their corresponding flood peak discharge values. For marginal return periods of 10, 100 and 1,000 years, the joint return period for the data sets changed slightly as the data length was reduced; the value was approximately 6 years, 51–52 years, and 500–501 years, respectively. The marginal AMFM values corresponding to each return period tended to increase when the data length was reduced. The values exhibited fluctuating characteristics, especially at shorter lengths.

For each data length, the conditional return periods of data from Cuntan Station, given the AMFM values of Yichang Station, were obtained using the chosen copula model. The marginal return periods were also calculated from the GEV distribution. As shown in Fig. 5, the data lengths of 109, 90, 70, 50, and 30 years were selected to illustrate how the change in data length would affect the conditional return period. Given specific flood peak values, the conditional return periods generally became larger with a longer data length. However, there were some variations, and the data length of 70 years had the largest return period compared to other data lengths. Fig. 5(d) shows the marginal return periods of Cuntan Station under each data length. The figure shows that the marginal return periods were affected by the change in data length. Given a fixed flood peak value, the marginal return periods obtained under each data length tended to vary significantly, especially for larger

flood peaks. Therefore, while the fluctuation of conditional return periods may be partly caused by copula modeling, it can also be caused by the variation of marginal distributions.

Conclusions

This paper evaluates the effect of data length on the uncertainty in the modeling results obtained by the use of the Clayton, Frank, and G-H copulas with SPE and EKT for data sets collected from the Yangtze River region. The uncertainty is considered from a statistical point of view via entropy. Two criteria—AIC and S_n values—are used.

Copula modeling results deteriorate as the data set length becomes shorter. While the best-fitting copula and estimator change as the data length is reduced, the change may be due to the uncertainty of the marginal data sets and the joint uncertainty—i.e., the dependence between the data of the two stations for both cases.

On the whole, the data sets from the Yangtze River can be modeled with the Frank copula and SPE estimator. Joint return periods and conditional return periods are studied. While the joint return periods tend to be slightly affected by the reduced data length, the conditional return periods fluctuate when the data length is shortened. Marginal flood peak discharge exhibits higher variations when the data length is shorter, as is the marginal return period. Thus, it can be concluded that data length affects not only copula modeling merely through the bivariate behavior, but it may also have an adverse effect on the marginal, which is an important factor when using a copula model to do bivariate analysis.

In practice, when using short lengths of records, practitioners should pay attention to the uncertainty contained in the data set itself before constructing their joint distribution.

Acknowledgments

The authors gratefully acknowledge the helpful comments and suggestions from the editor and anonymous reviewers. This study was supported by the National Natural Science Fund of China (No. 51190091 and 41071018), the Program for New Century Excellent Talents in University (NCET-12-0262), the China Doctoral Program of Higher Education (20120091110026), the Qing Lan Project, the Skeleton Young Teachers Program, and the Excellent Disciplines Leaders in Midlife-Youth Program of Nanjing University.

References

AghaKouchak, A., Bárdossy, A., and Habib, E. (2010). "Copula-based uncertainty modeling: Application to multisensor precipitation estimates." *Hydrol. Process.*, 24, 2111–2124.

Bárdossy, A., and Pegram, G. G. S. (2009). "Copula based multisite model for daily precipitation simulation." *Hydrol Earth Syst. Sci.*, 13, 2299–2314.

Chen, S. X., and Huang, T.-M. (2007). "Nonparametric estimation of copula functions for dependence modeling." *Canadian J. Stat.*, 35(2), 265–282.

De Michele, C., Salvadori, G., Vezzoli, R., and Pecora, S. (2013). "Multivariate assessment of droughts: Frequency analysis and dynamic return period." *Water Resour. Res.*, 49(10), 6985–6994.

Genest, C., and Favre, A.-C. (2007). "Everything you always wanted to know about copula modeling but were afraid to ask." *J. Hydrol. Eng.*, 10.1061/(ASCE)1084-0699(2007)12:4(347), 347–368.

Genest, C., Rémillard, B., and Beaudoin, D. (2009). "Goodness-of-fit test for copulas: A review and a power study." *Insur. Math. Econ.*, 44(2), 199–213.

Gräler, B., et al. (2013). "Multivariate return periods in hydrology: A critical and practical review focusing on synthetic design hydrograph estimation." *Hydrol. Earth Syst. Sci.*, 17(4), 1281–1296.

Gringorten, I. I. (1963). "A plotting rule for extreme probability paper." *J. Geophs. Res.*, 68(3), 813–814.

Hao, Z., and AghaKouchak, A. (2013). "Multivariate standardized drought index: A parametric multi-index model." *Adv. in Water Resour.*, 57, 12–18.

Hao, Z., and AghaKouchak, A. (2014). "A nonparametric multivariate multi-index drought monitoring framework." *J. Hydrometeorol.*, 15(1), 89–101.

Hao, Z., and Singh, V. P. (2013). "Modeling multisite streamflow dependence with maximum entropy copula." *Water Resour. Res.*, 49(10), 7139–7143.

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

Kao, S.-C., and Govindaraju, R. S. (2010). "A copula-based joint deficit index for droughts." *J. Hydrol.*, 380(1–2), 121–134.

Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). "Comparison of semiparametric and parametric methods for estimating copulas." *Comput. Stat. Data Annu.*, 51(6), 2836–2850.

Kumar, P. (2011). "Copula functions: Characterizing uncertainty in probabilistic systems." *Appl. Math. Sci.*, 5(30), 1459–1472.

Lee, T., and Salas, J. D. (2011). "Copula-based stochastic simulation of hydrological data applied to Nile River flows." *Hydrol. Res.*, 42(4), 318–330.

Ma, M. W., Song, S. B., Ren, L. L., Jiang, S. H., and Song, J. L. (2013). "Multivariate drought characteristics using trivariate Gaussian and Student t copulas." *Hydrol. Process.*, 27(8), 1175–1190.

Nelsen, R. B. (1999). *An introduction to copulas*, Springer, New York.

Parent, E., Favre, A.-C., Bernier, J., and Perreault, L. (2014). "Copula models for frequency analysis what can be learned from a Bayesian perspective?" *Adv. in Water Resour.*, 63, 91–103.

Possolo, A. (2010). "Copulas for uncertainty analysis." *Metrol.*, 47(3), 262–271.

Salvadori, G., and De Michele, C. (2013). "Multivariate extreme value methods." *Extremes Changing Clim.*, 65, 115–162.

Serinaldi, F. (2013). "An uncertain journey around the tails of multivariate hydrological distributions." *Water Resour. Res.*, 49(10), 6527–6547.

Singh, V. P. (2013). *Entropy theory and its application to environmental and water engineering*, John Wiley, Chichester, U.K.

Song, S. B., and Singh, V. P. (2010). "Frequency analysis of droughts using the Plackett copula and parameter estimation by genetic algorithm." *Stoch. Environ. Res. Risk Assess.*, 24(5), 783–805.

Su, H.-T., and Tung, Y.-K. (2013). "Incorporating uncertainty of distribution parameters due to sampling errors in flood-damage-reduction project evaluation." *Water Resour. Res.*, 49(3), 1680–1692.

Vandenberghe, S., Verhoest, N. E. C., and 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 Resour. Res.*, 46, W01512.

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

Wang, D. (2010). "Accelerating entropy theory: New approach to the risks of risk analysis in water issues." *Hum. Ecol. Risk Assess.*, 16(1), 4–9.

Wang, D., Singh, V. P., and Zhu, Y. (2007). "Hybrid fuzzy and optimal modeling for water quality evaluation." *Water Resour. Res.*, 43, W05415.

Wang, D., Singh, V. P., Zhu, Y., and Wu, J. (2009). "Stochastic observation error and uncertainty in water quality evaluation." *Adv. Water Resour.*, 32(10), 1526–1534.

Wang, Y. K., Ma, H., Sheng, D., and Wang, D. (2012). "Assessing the interactions between chlorophyll a and environmental variables using copula method." *J. Hydrol. Eng.*, 10.1061/(ASCE)HE.1943-5584.0000387, 495–506.

Xia, Y., Yang, Z.-L., Jackson, C., Stoffa, P. L., and Sen, M. K. (2004). "Impacts of data length on optimal parameter and uncertainty estimation of a land surface model." *J. Geophys. Res.*, 109, D07101.

Zeng, X. K., Wang, D., and Wu, J. C. (2012). "Sensitivity analysis on the probability distribution parameters of water level series based on information entropy." *Stoch. Environ. Res. Risk Assess.*, 26(3), 345–356.

Zhang, L., and Singh, V. P. (2006). "Bivariate flood frequency analysis using the copula method." *J. Hydrol. Eng.*, 10.1061/(ASCE)1084-0699(2006)11:2(150), 150–164.

Zhang, L., and Singh, V. P. (2007). "Bivariate rainfall frequency distributions using Archimedean copulas." *J. Hydrol.*, 332(1–2), 93–109.