Copula-based composite likelihood approach for frequency analysis of short annual precipitation records

Ting Wei and Songbai Song

ABSTRACT

Hydrological series lengths are decreasing due to decreasing investments and increasing human activities. For short sequences, a copula-based composite likelihood approach (CBCLA) has been employed to enhance the quality of hydrological design values. However, the Pearson type III (P-III) distribution for short annual precipitation records has not yet been thoroughly investigated using the CBCLA. This study used the CBCLA to incorporate the concurrent and non-concurrent periods contained in data of various lengths into an integrated framework to estimate the parameters of precipitation frequency distributions. The marginal distributions were fitted using the P-III distribution, and the joint probability was constructed using a copula which offers flexibility in choosing arbitrary marginals and dependence structure. Furthermore, the uncertainties in the estimated precipitation design values for the short series obtained from this approach were compared with those obtained from univariate analysis. Then, Monte-Carlo simulations were performed to examine the feasibility of this approach. The annual precipitation series at four stations in Weihe River basin, China, were used as a case study. Results showed that CBCLA with P-III marginals reduced the uncertainty in the precipitation design values for the short series and the reduction in the uncertainty became more significant with longer adjacent series.

Key words | copula, precipitation design values, precipitation frequency analysis, uncertainty

Ting Wei
Songbai Song (corresponding author)
College of Water Resources and Architectural Engineering,
Northwest A&F University,
Yangling 712100,
China
E-mail: *ssb6533@nwsuaf.edu.cn*

INTRODUCTION

Hydrologic frequency analysis is used to estimate the magnitude of a hydrologic event with a given frequency of occurrence or to estimate the frequency of occurrence of a hydrologic event with a given magnitude (Haan 1977; Rao & Hamed 1999). Sufficiently long data are essential for accurate parameter estimation. However, it is often difficult to obtain such long-term data, and hydrological design estimates based on data of insufficient length involve large uncertainties. Therefore, analyzing these uncertainties is of great significance for improving the calculation accuracy and reducing the uncertainties of hydrological design values.

In the past, most hydrologic frequency analyses were analyzed through the use of univariate distributions. Using an assumed probability distribution to estimate the frequency analysis of a hydrologic event, one can compute the statistics for various magnitudes, which can then be used for the planning, design, and management of water resource systems.

Research on the univariate flood frequency analysis has been reviewed by Bobée & Rasmussen (1995), Rao & Hamed (1999), Singh & Strupczewski (2002), and Vittal *et al.* (2015). Singh & Strupczewski (2002) classified frequency analysis methods into four groups: (1) empirical, (2) phenomenological, (3) dynamic, and (4) stochastic watershed in conjunction with Monte-Carlo simulation. In addition, the suggested distributions for fitting hydrological data and a number of parameter estimation methods were summarized by Rao & Hamed (1999). Two popular methods for regional frequency analysis are the index FFA (Flood Frequency Analysis) and regression analysis (Singh & Strupczewski 2002).

doi: 10.2166/nh.2017.033

Some studies on annual maximum flow frequency analysis show that the empirical probability plot often displays two or more distinct segments. Using a single probability distribution smooth function to fit all the segments may produce a compromised curve and may lead to serious flood estimation errors for large return periods. In most cases, the observed data of the smaller censored flows have no influence on the fitting of the flood frequency curve. If the smaller flows produce some (very marginal) influence, the final fitting still depends very heavily on the larger flows. Wang (1990a, 1990b, 1996a, 1996b, 1997a, 1997b) proposed using partial probability weighted moments and higher probability weighted moments to estimate flood distribution and presents the potential merits of this method.

Recently, a vast amount of research has been dedicated to investigating frequency analysis under conditions of climate change, land use change and river regulation, acting individually or together (Strupczewski & Kaczmarek 2001; Strupczewski *et al.* 2001; Cunderlik & Burn 2003; Khaliq *et al.* 2006; Leclerc & Ouarda 2007; Villarini *et al.* 2009, 2010; Gilroy & McCuen 2012; Salas & Obeysekera 2014; Zeng *et al.* 2014; Li & Tan 2015; Vasiliades *et al.* 2015; Xiong *et al.* 2015).

The above-mentioned studies still require data over a sufficiently long period to obtain sound hydrological designs based on statistical analysis. However, measuring and monitoring hydrologic events involves high costs, and the global hydrological gauge network is still unevenly distributed, with most measuring stations located in a few developed countries and regions and particularly poor gauge coverage in developing countries. The station density is relatively low in developing countries, and the length of the hydrological data is limited, resulting in critical data gaps or representativeness problems. As a result of various natural conditions and human activities, such as destruction and migration of stations, construction of water resource projects, and river regulation, these deficiencies have been exacerbated over recent decades. Inadequate data lengths can produce large uncertainties in hydrological design estimates (Chowdhary & Singh 2010). Even when employing new parameter estimation techniques, such as partial probability weighted moments and higher probability weighted moments, it is difficult to improve the estimation accuracy for shorter data sequences.

Therefore, to overcome the problem of insufficient data for frequency analysis, additional information has been used to reduce the uncertainty in hydrologic estimates. For example, Vogel & Stedinger (1985) lengthened short hydrologic records by employing the cross-correlation among nearby longer records. They also developed improved unbiased estimators of the mean and variance of the streamflow at the sites with short records, and these estimators could obtain equal or lower variance than the simple at-site sample maximum likelihood estimates. In addition, to increase the precision of flood quantile estimators for a relatively short data series, historical or paleofloods data have been incorporated into flood frequency analyses (Stedinger & Cohn 1986; Reis & Stedinger 2005; Halbert *et al.* 2016), and joint cumulative distribution functions (bivariate or trivariate normal distributions, extreme value distributions, Gumbel mixed model and bivariate Gamma distribution) have been adopted to simultaneously consider data from adjacent stations (Escalante-Sanboval & Raynal-Villasenor 1998; Yue 2000, 2001, 2002; Yue *et al.* 2001; Escalante-Sandoval 2007).

The conventional multivariate approaches are limited due to the marginal distributions and the joint distribution belonging to the same family and the mathematical formulation becoming increasingly complicated with the increasing number of parameters (Chen *et al.* 2010; Li *et al.* 2013). To overcome these shortcomings, copula functions have been proposed and widely used in multivariate hydrological frequency analyses (Zhang & Singh 2006, 2007a, 2007b; Song & Singh 2010; Salvadori *et al.* 2011; Ma *et al.* 2013; Salvadori *et al.* 2013; Zhang *et al.* 2013, 2015; Fu & Butler 2014; Huang *et al.* 2015; Chen *et al.* 2016).

Chowdhary & Singh (2009, 2010) noted that only the concurrent periods of data sets have been used in either conventional or copula-based multivariate hydrological frequency analysis, whereas a majority of non-concurrent data in the extensive amount of staggered hydrological data remains unutilized for hydrological analysis. These non-concurrent data offer untapped potential for reducing parameter uncertainty by considering a multivariate framework which can simultaneously consider partially concurrent information in an integrated manner. However, such studies are currently still limited. Sandoval & Raynal-Villasenor (2008) constructed a trivariate logistic model with generalized extreme value distribution marginals to analyze samples with equal or unequal record lengths. Their aim was to reduce the bias or uncertainty in flood

estimates by combining observations from several sites with an insufficiently long record to increase the available information and to provide a regional at-site design event. Chowdhary & Singh (2009, 2010) proposed a copula-based composite likelihood approach (CBCLA) to analyze various lengths of multivariate data and to reduce uncertainty in hydrologic design estimates. The results show that the CBCLA can provide necessary flexibility for admitting arbitrary marginals and can significantly reduce uncertainty in design flood quantiles for a relatively short flood series by utilizing associated downstream flood data. This approach can offset the impact of dwindling hydrological observation networks around the world by adding information that can be derived from existing networks.

Combining the CBCLA with the P-III distribution for short annual precipitation records has not yet been thoroughly investigated. The primary aim of this paper is to use the CBCLA to investigate the uncertainty of annual precipitation design values for a relatively short series fitted with the P-III distribution. In this study, the concurrent and exclusive parts of a multivariate data set with various lengths were combined in an integrated manner using copulas and the maximum likelihood function, and the uncertainty was quantified in terms of confidence intervals. Examples using simulated and measured data are presented in this paper. The precipitation sequences of four stations, namely Xi'an, Dali, Lintong and Huayin, in Weihe River basin (WRB) in China were selected as a case study. The data length at Huayin station was 28 years and is thus short series. The data lengths at the other three stations were 77, 53 and 50 respectively, and they are long series that can provide additional information. The confidence intervals and standard errors of the design values at Huayin station were derived and compared with the results from univariate estimations. The results show that the CBCLA can reduce the uncertainty in the precipitation design values for a relatively short data series.

METHODS

Bivariate copulas

A copula is a function that connects univariate marginal probability distributions to construct a multivariate distribution function (Sklar 1959; Salvadori & De Michele 2007). For a bivariate case, let X and Y be two random variables with cumulative distribution functions (cdfs) of $F_X(x)$ and $F_Y(y)$; the bivariate joint distribution function of (X, Y) is defined as:

$$F(x, y) = C_\theta[F_X(x), F_Y(y)] = C_\theta(u, v) \tag{1}$$

where $F(x,y)$ is the bivariate distribution of X and Y, C_θ denotes the copula function with the dependence parameter θ, $F_X(x)$ and $F_Y(y)$ are the marginal distributions of X and Y, respectively, and $F_X(x) = u$ and $F_Y(y) = v$ with u and $v \tilde{(}0, 1)$. From Equation (1), the copula-based joint probability density function (pdf) can be obtained for the bivariate variables (Shiau 2006):

$$f(x, y) = f_X(x) f_Y(y) c_\theta(u, v) \tag{2}$$

where $f_X(x)$ and $f_Y(y)$ are marginal pdfs and $c_\theta(u, v)$ is the copula density function of concurrent random samples.

In this paper, the P-III distribution was used to establish the marginal pdfs in Equation (2). For a random variable Y, the pdf of the P-III distribution can be written as:

$$f(y|\gamma, \alpha, \beta) = \frac{1}{\alpha\Gamma(\beta)}\left(\frac{y-\gamma}{\alpha}\right)^{\beta-1} e^{-((y-\gamma)/\alpha)}, \quad \gamma < y < \infty \tag{3}$$

where α, β and γ are the scale, shape and location parameters, respectively. The MLM (Maximum Likelihood Method) was used to estimate the P-III distribution parameters. Two general goodness of fit test statistics, Kolmogorov–Smirnov's (K-S) Dn and Anderson–Darling's (A-D) A_n^2, were used to determine whether the data were consistent with the P-III distribution (Song & Singh 2010).

Parameter estimation and goodness-of-fit test for a bivariate distribution

The Archimedean copula family is a popular model for hydrologic analyses (Nelsen 2007) because it is easy to construct and it can be applied to both positively and negatively correlated hydrologic variables. Therefore, three one-parameter Archimedean copulas, Gumbel–Hougaard (GH), Clayton and Frank, were selected to determine the joint

probability distribution of the precipitation variables in this study. The analytical expressions of these copulas have been well documented in previous pioneering studies (Shiau 2006; Zhang & Singh 2007a, 2007b).

The dependence parameters for these copulas were estimated using Kendall's tau τ (KTE) and maximum likelihood estimation (MLE) (Genest & Favre 2007). For the KTE, the dependence parameter θ can be estimated from the relationship between τ and θ. The MLE is obtained by maximizing the likelihood function, which involves the parameters of both the marginal distributions and the copula function.

The root mean square error (RMSE), Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to represent the bias of the probability distributions (Posada & Buckley 2004; Zhang & Singh 2007a, 2007b; Ma *et al.* 2013). Additionally, a bootstrap version of the K-S test and A-D test based on Rosenblatt's transformation was employed to test the goodness-of-fit for the copulas (Dobrić & Schmid 2007). The expressions and calculation steps of these statistics have been documented in previous studies (Song & Singh 2010).

Copula-based composite likelihood parameter estimation

The MLM is applicable to the parameter estimation of complex density functions. The arrangement of a typical composite event, comprising some overlapping and some exclusive periods of X and Y, is shown in Figure 1. N_X and N_Y are the full lengths of the two data series, respectively. n_{XY} is the length of the concurrent period, and n_X and n_Y are the lengths of the individual nonconcurrent periods. Correspondingly, e_{XY} is the concurrent event, e_Y and e_Y are the nonconcurrent events. These three events constitute a composite event e_C, and the likelihood function of e_C is called the composite likelihood function (Chowdhary & Singh 2010). Let $f_X(x; \boldsymbol{\delta})$ and $f_Y(y; \boldsymbol{\eta})$ represent the pdfs of marginal distributions, and $f_{XY}(x, y; \boldsymbol{\psi})$ be the bivariate pdf of (X, Y), which takes the same form as $f(x,y)$ in Equation (2). Here, $\boldsymbol{\delta} = \{\delta_1, \delta_2, \ldots, \delta_{r_\delta}\}$, $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_{r_\eta}\}$ and $\boldsymbol{\psi} = \{\boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\theta}\} = \{\psi_1, \psi_2, \ldots, \psi_r\}$ are the parameter vectors of these distributions, respectively; r_δ, r_η and r represent the numbers of parameters; $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_{r_\theta}\}$ is the dependence parameter vector of the bivariate pdf; and r_θ is a parameter number. For the sake of brevity, the pdfs are denoted as $f(x)$, $f(y)$, and $f(x, y)$, respectively.

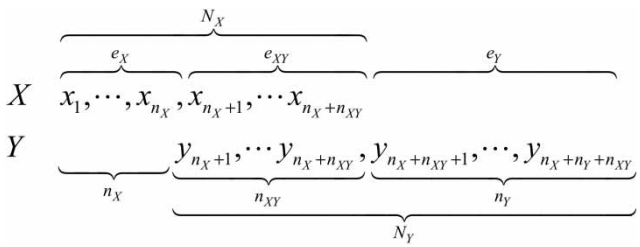

Figure 1 | Arrangement of a typical composite event of bivariate random variables with different lengths.

Considering the univariate case Y shown in Figure 1, the likelihood function L_Y and the log-likelihood function LL_Y with respect to N_Y independent and identically distributed (iid) observations are given as:

$$L_Y = \prod_{i=1}^{N_Y} f(y_i); \quad LL_Y = \sum_{i=1}^{N_Y} \log f(y_i) \tag{4}$$

The maximum likelihood estimators $\hat{\boldsymbol{\eta}}$ are obtained by maximizing LL_Y and solving the system of equations given by (Bobée 1979; Rao & Hamed 1999):

$$\frac{\partial LL_Y}{\partial \eta_p} = \sum_{i=1}^{N_Y} \frac{\partial \log f(y_i)}{\partial \eta_p} = 0; \quad p = 1, 2, \ldots, r_\eta \tag{5}$$

For the case in which the data lengths are not identical, Raynal-Villasenor (1985) proposed a general form of the likelihood function that is applicable to all possible composite arrangements. Following his work, the likelihood function L_C and log-likelihood function LL_C for a composite event can be expressed as:

$$L_C = \left[\prod_{i=1}^{n_X} f(x_i)\right]^{I_1} \left[\prod_{i=1}^{n_{XY}} f(x_i, y_i)\right]^{I_2} \left[\prod_{i=1}^{n_Y} f(y_i)\right]^{I_3} \tag{6}$$

$$LL_C = I_1 \sum_{i=1}^{n_X} \log f(x_i) + I_2 \sum_{i=1}^{n_{XY}} \log f(x_i, y_i) + I_3 \sum_{i=1}^{n_Y} \log f(y_i) = I_1 LL_X + I_2 LL_{XY} + I_3 LL_Y \tag{7}$$

In the above equations, $f(x)$ and $f(y)$ are the univariate densities for the exclusive periods that are given by their marginal density functions and LL_X and LL_Y are the corresponding log-likelihood functions (Escalante-Sandoval 2007; Raynal-Villasenor & Salas 2008). Unlike the composite likelihood function in Raynal-Villasenor's work, the bivariate pdf $f(x, y)$ of the concurrent period in Equation (7) is modeled using the copula function shown in Equation (2), and LL_{XY} is its log-likelihood function. I_i, $(i = 1, 2, 3)$ represents an indicator variable, such as $I_i = 1$ if $n_j > 0$ $(j = X, XY, Y)$ or $I_i = 0$ if $n_j = 0$.

The maximum likelihood estimators $\hat{\boldsymbol{\psi}}$ are derived by maximizing LL_C and solving the following equation:

$$\frac{\partial LL_C}{\partial \psi_p} = \frac{\partial (I_1 LL_X + I_2 LL_{XY} + I_3 LL_Y)}{\partial \psi_p} = 0; \quad p = 1, 2, \ldots, r \tag{8}$$

The estimators $\hat{\boldsymbol{\psi}}$ contain the new parameters of the marginal distributions for both the short and long series and the dependence parameter of the copula function. These parameters are used in the calculations in the following sections.

Variance-covariance matrix of parameter estimation

To calculate the confidence intervals of the design values, the standard errors of the distribution parameters should be estimated first. An asymptotic variance-covariance matrix usually provides a good description of the parameter error (Rueda 1981). In this section, following Chowdhary & Singh (2010), we summarize the procedures of the variance-covariance matrix for parameter estimation.

The asymptotic variance-covariance matrix of parameter estimation is the inverse of the Fisher information matrix, which has elements that are equal to the expectations of the second partial derivatives of the log-likelihood function distribution with respect to the parameters (Kamwi 2005; Nadarajah 2006). The main diagonal elements of the variance-covariance matrix correspond to the asymptotic variances of the parameters.

Taking the univariate variable as a special case, the elements of the information matrix using Equation (4) is given as:

$$\begin{aligned} i_Y^{p,q} &= E\left[\frac{\partial LL_Y}{\partial \eta_p}\frac{\partial LL_Y}{\partial \eta_q}\right] = E\left[\sum_{i=1}^{N_Y}\frac{\partial \log f(y_i)}{\partial \eta_p}\sum_{i=1}^{N_Y}\frac{\partial \log f(y_i)}{\partial \eta_q}\right] \\ &= \sum_{i=1}^{N_Y} E\left[\frac{\partial \log f(y_i)}{\partial \eta_p}\frac{\partial \log f(y_i)}{\partial \eta_q}\right] = N_Y E\left[\frac{\partial \log f(y)}{\partial \eta_p}\frac{\partial \log f(y)}{\partial \eta_q}\right] \\ &= N_Y E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_p \partial \eta_q}\right] \end{aligned} \tag{9}$$

where $p = 1, 2, \ldots, r_\eta$ and $q = 1, 2, \ldots, r_\eta$.

The resulting Fisher information matrix $\mathbf{I}_Y$ is:

$$\mathbf{I}_Y = \left\|i_Y^{p,q}\right\|_{r_\eta \times r_\eta} = N_Y \left\|a_Y^{p,q}\right\|_{r_\eta \times r_\eta} = N_Y \mathbf{A}_Y \tag{10}$$

where $\mathbf{A}_Y = \left\|a_Y^{p,q}\right\|_{r_\eta \times r_\eta}$ and $a_Y^{p,q}$ is the information content derived from a single observation of Y (Chowdhary & Singh 2010). Using Equation (9), $\mathbf{A}_Y$ is given by:

$$\begin{aligned} \mathbf{A}_Y &= \left\|a_Y^{p,q}\right\|_{r_\eta \times r_\eta} \\ &= \begin{bmatrix} E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_1^2}\right] & E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_1 \partial \eta_2}\right] & \cdots & E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_1 \partial \eta_{r_\eta}}\right] \\ E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_2 \partial \eta_1}\right] & E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_2^2}\right] & \cdots & E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_2 \partial \eta_{r_\eta}}\right] \\ \vdots & \vdots & \vdots & \vdots \\ E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_{r_\eta} \partial \eta_1}\right] & E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_{r_\mu} \partial \eta_2}\right] & \cdots & E\left[-\frac{\partial^2 \log f(y)}{\partial \eta_{r_\eta}^2}\right] \end{bmatrix} \end{aligned} \tag{11}$$

The details of the derivation of $\mathbf{I}_Y$ are given in Appendix A (available with the online version of this paper). Inverting $\mathbf{I}_Y$ leads to the corresponding variance-covariance matrix:

$$\begin{aligned} \mathbf{VC}_U = \mathbf{VC}_Y &= \begin{bmatrix} Var(\hat{\eta}_1) & Cov(\hat{\eta}_1, \hat{\eta}_2) & \cdots & Cov(\hat{\eta}_1, \hat{\eta}_{r_\eta}) \\ Cov(\hat{\eta}_2, \hat{\eta}_1) & Var(\hat{\eta}_2) & \cdots & Cov(\hat{\eta}_2, \hat{\eta}_{r_\eta}) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\hat{\eta}_{r_\eta}, \hat{\eta}_1) & Cov(\hat{\eta}_{r_\eta}, \hat{\eta}_2) & \cdots & Var(\hat{\eta}_{r_\eta}) \end{bmatrix} \\ &= (\mathbf{I}_Y)^{-1} = \frac{1}{N_Y}(\mathbf{A}_Y)^{-1} \end{aligned} \tag{12}$$

where $Var(\cdot)$ and $Cov(\cdot)$ are the asymptotic variances and covariances of the parameters, respectively.

As shown in Equation (9), the information matrix elements for the case of the bivariate random variable can be written as:

$$\begin{aligned} i_{XY}^{p,q} &= E\left[\frac{\partial LL_{XY}}{\partial \psi_p}\frac{\partial LL_{XY}}{\partial \psi_q}\right] \\ &= n_{XY}E\left[\frac{\partial \log f(x,y)}{\partial \psi_p}\frac{\partial \log f(x,y)}{\partial \psi_q}\right] \\ &= n_{XY}E\left[-\frac{\partial^2 \log f(x,y)}{\partial \psi_p \partial \psi_q}\right] \end{aligned} \tag{13}$$

where $p = 1, 2, \ldots, r$ and $q = 1, 2, \ldots, r$.

If the joint pdf is based on the copula, as shown in Equation (2), then the elements in Equation (13) are expressed as:

$$\begin{aligned} \frac{\partial \log f(x,y)}{\partial \psi_p} &= \frac{\partial}{\partial \psi_p}[\log f(x) + \log f(y) + \log c_\theta(u,v)] \\ &= \frac{\partial \log f(x)}{\partial \psi_p} + \frac{\partial \log f(y)}{\partial \psi_p} + \frac{\partial \log c_\theta(u,v)}{\partial \psi_p} \end{aligned} \tag{14}$$

The Fisher information matrix and variance-covariance matrix are given by:

$$\mathbf{I}_{XY} = \left\|i_{XY}^{p,q}\right\|_{r\times r} = n_{XY}\left\|a_{XY}^{p,q}\right\|_{r\times r} = n_{XY}\mathbf{A}_{XY} \tag{15}$$

$$\mathbf{VC}_{XY} = (\mathbf{I}_{XY})^{-1} = \frac{1}{n_{XY}}(\mathbf{A}_{XY})^{-1} \tag{16}$$

Similar to the above cases, the information matrix elements calculated from the composite likelihood function of Equation (7) are given as:

$$\begin{aligned} i_C^{p,q} &= E\left[\frac{\partial LL_C}{\partial \psi_p}\frac{\partial LL_C}{\partial \psi_q}\right] = I_1 E\left[\frac{\partial LL_X}{\partial \psi_p}\frac{\partial LL_X}{\partial \psi_q}\right] \\ &+ I_2 E\left[\frac{\partial LL_{XY}}{\partial \psi_p}\frac{\partial LL_{XY}}{\partial \psi_q}\right] + I_3 E\left[\frac{\partial LL_Y}{\partial \psi_p}\frac{\partial LL_Y}{\partial \psi_q}\right] \\ &= I_1 N_X E\left[-\frac{\partial^2 \log f(x)}{\partial \psi_p \partial \psi_q}\right] + I_2 n_{XY} E\left[-\frac{\partial^2 \log f(x,y)}{\partial \psi_p \partial \psi_q}\right] \\ &+ I_3 N_Y E\left[-\frac{\partial^2 \log f(y)}{\partial \psi_p \partial \psi_q}\right] \end{aligned} \tag{17}$$

The right side of the above equation includes the elements of the Fisher information matrix for two univariate distributions and one bivariate distribution, as given in Equations (9) and (13), respectively. Considering that the length of the concurrent period is $n_{XY} > 0$, the above result can be expressed in terms of $a^{p,q}$ using Equations (10) and (15):

$$\begin{aligned} i_C^{p,q} &= I_1 i_X^{p,q} + I_2 i_{XY}^{p,q} + I_3 i_Y^{p,q} = I_1 n_X \mathbf{A}_X + I_2 n_{XY}\mathbf{A}_{XY} + I_3 n_Y \mathbf{A}_Y \\ &= n_{XY}\left[I_1(m_X - 1)\left\|a_X^{p,q}\right\| + I_2\left\|a_{XY}^{p,q}\right\| + I_3(m_Y - 1)\left\|a_Y^{p,q}\right\|\right] \end{aligned} \tag{18}$$

where $m_X = (N_X/n_{XY})$ and $m_Y = (N_Y/n_{XY})$ are the ratios of the total lengths of X and Y to the concurrent period, respectively, and the Fisher information matrix for composite events is given by:

$$\mathbf{I}_C = \left\|i_C^{p,q}\right\|_{r\times r} = n_{XY}\left\|a_C^{p,q}\right\|_{r\times r} = n_{XY}\mathbf{A}_C \tag{19}$$

where $\left\|a_C^{p,q}\right\| = I_1(m_X - 1)\left\|a_X^{p,q}\right\| + I_2\left\|a_{XY}^{p,q}\right\| + I_3(m_Y - 1)\left\|a_Y^{p,q}\right\|$.

The variance-covariance matrix of the composite likelihood function parameter estimates is expressed as:

$$\mathbf{VC}_C = (\mathbf{I}_C)^{-1} = \frac{1}{n_{XY}}(\mathbf{A}_C)^{-1} \tag{20}$$

Information matrix for the P-III distribution and bivariate distribution

According to the previous section, the information matrix for the scale, shape and location parameters of the P-III distribution for the univariate case of series Y in Figure 1 is given as:

$$I_Y(\alpha,\beta,\gamma) = \begin{bmatrix} \frac{N_Y}{\alpha^2(\beta-2)} & \frac{N_Y}{\alpha^2} & \frac{N_Y}{\alpha(\beta-1)} \\ \frac{N_Y}{\alpha^2} & \frac{N_Y\beta}{\alpha^2} & \frac{N_Y}{\alpha} \\ \frac{N_Y}{\alpha(\beta-1)} & \frac{N_Y}{\alpha} & N_Y\psi'(\beta) \end{bmatrix} \tag{21}$$

where $\psi'(\beta) = \dfrac{d^2 \log \Gamma(\beta)}{d\beta^2}$.

For the copula-based bivariate joint probability distribution with P-III marginals, the number of parameters in Equation (13) is $r = 7$, including the parameters of marginal distributions $\alpha_x, \beta_x, \gamma_x, \alpha_y, \beta_y, \gamma_y$ and the dependence parameter of the copula θ. The Fisher information matrix of the parameters is given by:

$$\mathbf{I}_{XY} = n_{XY}\begin{bmatrix} \frac{\beta}{\alpha_x^2} & \frac{1}{\alpha_x} & \frac{1}{\alpha_x^2} & 0 & 0 & 0 & 0 \\ \frac{1}{\alpha_x} & \psi'(\beta_x) & \frac{1}{\alpha_x(\beta_x-1)} & 0 & 0 & 0 & 0 \\ \frac{1}{\alpha_x^2} & \frac{1}{\alpha_x(\beta_x-1)} & \frac{1}{\alpha_x^2(\beta_x-2)} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\beta}{\alpha_y^2} & \frac{1}{\alpha_y} & \frac{1}{\alpha_y^2} & 0 \\ 0 & 0 & 0 & \frac{1}{\alpha_y} & \psi'(\beta_y) & \frac{1}{\alpha_y(\beta_y-1)} & 0 \\ 0 & 0 & 0 & \frac{1}{\alpha_y^2} & \frac{1}{\alpha_y(\beta_y-1)} & \frac{1}{\alpha_y^2(\beta_y-2)} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & D \end{bmatrix} \tag{22}$$

where

$$D = E\left[-\frac{\partial^2 \log f(x,y)}{\partial\theta^2}\right] = E\left[-\frac{\partial^2 \log\left(f(x)f(y)c_\theta(u,v)\right)}{\partial\theta^2}\right]$$
$$= E\left[-\frac{\partial^2\left(\log f(x) + \log f(y) + \log c_\theta(u,v)\right)}{\partial\theta^2}\right]$$
$$= E\left[-\frac{\partial^2 \log f(x)}{\partial\theta^2} - \frac{\partial^2 \log f(y)}{\partial\theta^2} - \frac{\partial^2 \log c_\theta(u,v)}{\partial\theta^2}\right]$$
$$= E\left[-\frac{\partial^2 \log c_\theta(u,v)}{\partial\theta^2}\right]$$

Uncertainty estimation of the design values

The confidence interval is a convenient approach for quantifying the uncertainty and indicating the accuracy of design values (Rao & Hamed 1999; Silva *et al.* 2012). The distribution parameters of the short series Y were obtained from: (a) the univariate log-likelihood function LL_Y given in Equation (3) and (b) the composite log-likelihood function LL_C given in Equation (7), and the standard errors of the design values were compared for the two cases. The variances of the design values for case (a) were calculated using the delta method (Coles *et al.* 2001):

$$Var(q_{Uy}|T) = \nabla_{q_{Uy}}^T \mathbf{VC}_U \nabla_{q_{Uy}} \tag{23}$$

where $(q_{Uy}|T)$ is the quantile function of series Y for a given return period, $\nabla_{q_{Uy}}$ is the derivative of q_{Uy} with respect to parameter $\boldsymbol{\eta}$, $\nabla_{q_{Uy}}^T$ is the corresponding transposed matrix, and $\mathbf{VC}_U$ is the variance-covariance matrix of the parameter vector $\hat{\boldsymbol{\eta}}$ estimated using the univariate likelihood function given in Equation (5). The standard error is written as:

$$std(q_{Uy}|T) = \sqrt{Var(\mathrm{q}_{Uy}|T)} \tag{24}$$

Using a similar procedure, the uncertainty estimates of the design values were obtained when the distribution parameters were computed using a composite log-likelihood function:

$$std(q_{Cy}|T) = \sqrt{Var(q_{Cy}|T)} = \sqrt{\nabla_{q_{Cy}}^T \mathbf{VC}_C \nabla_{q_{Cy}}} \tag{25}$$

where $(q_{Cy}|T)$ is the quantile function of series Y for a given return period, which takes the same form as $(q_{Uy}|T)$. The derivatives of q_{Cy} and $\mathbf{VC}_C$ were determined according to the estimated values of the parameters $\hat{\boldsymbol{\psi}}$. The detailed derivations are given in Appendix B (available with the online version of this paper).

After deriving the standard errors of the design values estimated using Equations (24) and (25), the confidence intervals under confidence level $1-\alpha$ were calculated using $\hat{x}_T \pm u_{1-(\alpha/2)}\hat{s}_T$, where $u_{1-(\alpha/2)}$ is the quantile of the standard normal distribution for confidence levels equal to $1-(\alpha/2)$, $\hat{x}_T$ is the design value for the return period T which is calculated using q_{Uy} and q_{Cy}, and $\hat{s}_T$ is the standard error ($std(q_{Uy}|T)$ and $std(q_{Cy}|T)$) of $\hat{x}_T$.

CBCLA simulation

In this section, CBCLA was used in a simulated series to validate its applicability for improving the accuracy of parameter estimation for shorter series. Assuming that two data series, X and Y, were P-III distributed and that the

joint probability of their concurrent part was constructed following a Frank copula, the Monte-Carlo method was used to generate a data set with length $N = 80$ and the following distribution parameters: $\alpha_1 = 50, \beta_1 = 10, \gamma_1 = 200$ ($EX = 700$, $Cv = 0.2259$, $Cs = 0.6325$) for X; $\alpha_2 = 25$, $\beta_2 = 30$, $\gamma_2 = -150$ ($EX = 600$, $Cv = 0.2282$, $Cs = 0.3651$) for Y; and the Frank copula dependence parameter is $\theta = 3$, which corresponds to a Kendall correlation coefficient of $\tau = 0.6342$. Based on the arrangement of the composite event shown in Figure 1, we constructed eleven different composite event cases S1, S2, S3, S4, S5, S6, S7, S8, S9, S10 and S11, as shown in Table 1, using the simulated data. Here, variables X and Y are called the 'longer series' and the 'shorter series,' respectively. The objective here was to see whether using CBCLA decreases the uncertainty in the design values for Y.

Table 1 | Length of the data series for different composite cases

Case	Y N_Y	X N_X	n_X	e_C n_Y	n_{XY}
S1	30	50	20	30	0
S2		60	30	30	0
S3		60	20	40	0
S4	40	70	30	40	0
S5		80	40	40	0
S6		30	15	15	5
S7	20	40	25	15	5
S8		50	35	15	5
S9		40	5	35	5
S10	40	45	10	35	5
S11		50	15	35	5

The distribution parameters of Y in different composite cases were estimated using CBCLA, and the design values and 95% confidence widths corresponding to return periods of 50, 100 and 200 years were calculated based on these parameters. Table 2 compares the calculation results with the estimations from the univariate case. For both the univariate and composite case, the standard errors and confidence widths increased with increasing return periods. For the same size of Y, the standard errors and confidence widths obtained based on CBCLA were less than that of the

Table 2 | Design values, standard errors and confidence widths for different composite cases using Monte-Carlo simulation compared with the estimation results using the univariate method

N_Y	Case	Design value	T-50[b] Standard error	Confidence width	Design value	T-100[b] Standard error	Confidence width	Design value	T-200[b] Standard error	Confidence width
	UL[a]	824.9	55.2	216.5	854.2	65.3	256.0	880.3	75.7	296.8
30	S1	816.6	52.1	204.2	845.7	61.8	242.2	871.7	71.9	281.7
	S2	818.5	51.9	203.6	847.4	61.6	241.4	873.2	71.6	280.8
	UL[a]	815.2	47.4	185.9	840.6	55.6	218.1	862.9	64.0	251.1
40	S3	810.1	44.6	174.6	835.4	52.4	205.6	857.5	60.6	237.5
	S4	807.4	44.5	174.5	832.7	52.4	205.5	854.8	60.5	237.3
	S5	809.9	44.5	174.3	835.1	52.4	205.2	857.2	60.5	237.1
	UL[a]	836.1	87.9	344.5	879.5	99.9	391.7	920.5	111.7	438.0
20	S6	818.9	74.4	291.5	857.2	87.3	342.3	893.4	100.5	394.1
	S7	819.4	74.0	290.1	857.6	87.0	341.0	893.8	100.2	393.0
	S8	818.3	74.0	290.1	856.5	87.0	340.9	892.6	100.2	392.7
	UL[a]	842.5	48.4	189.7	879.9	58.0	227.2	914.9	67.9	266.3
40	S9	850.0	47.9	187.9	887.6	57.5	225.3	922.9	67.5	264.6
	S10	849.6	47.9	187.8	887.2	57.5	225.3	922.4	67.5	264.6
	S11	849.1	47.9	187.7	886.7	57.4	225.2	921.9	67.5	264.4

[a]UL implies that the results were estimated from the shorter series using the univariate maximum likelihood method.
[b]T-50, T-100 and T-200 indicate the given return period.

estimations using the univariate likelihood method. For example, in the case of $N_Y = 30$, the standard errors and confidence widths in cases S1 and S2 are smaller than those in the univariate case. Therefore, integrating additional observations with a series of insufficient length can increase the available information, can obtain more accurate parameter estimations from CBCLA, and can improve the accuracy of the design values.

Table 2 shows that for a constant length of Y, the standard errors and confidence widths decrease as the length of the additional series X increases. For example, in the case of $N_Y = 40$, the standard errors and confidence widths in cases S3, S4 and S5 gradually decrease, which indicates that the reduction in uncertainty increases as the length of the additional series increases. Therefore, CBCLA can reduce the uncertainty in the design values for inadequate series and can provide a more precise design basis for engineering projects.

STUDY AREA AND DATA

To illustrate the application of the CBCLA as outlined above, we analyzed the annual precipitation data collected in the WRB, which is the longest tributary of the Yellow River in China. The river originates from Gansu province and flows 818 km eastward into the Yellow River, and the catchment covers an area of approximately 134,800 km^2. The WRB is located in the transitional zone between arid and humid areas with an average annual precipitation of 572 mm. Due to the typical temperate continental monsoon climate, the total annual precipitation exhibits significant spatial and seasonal variations.

In this study, the observed annual precipitation series at four rainfall gauges (Figure 2), namely, Xi'an, Dali, Lintong and Huayin, were used as a case study. All the data were obtained from the National Climate Center of China Meteorological Administration. The observation periods at these four stations were 1932–2008, 1956–2008, 1959–2008 and 1981–2008, respectively. All these data have low first-order series correlation coefficients (−0.1140, −0.0785, 0.0480 and 0.0503 respectively) and no missing records. Anderson's test of independence showed that these data have an independent structure at a 90% confidence level. The characteristics of the annual precipitation data are listed in Table 3.

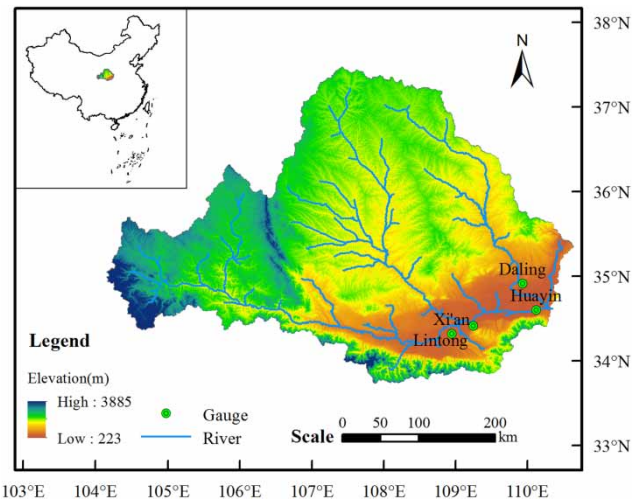

Figure 2 | Study area and locations of the stations in China.

Among the four stations, Huayin (the shortest data series) served as the base station, and the other three stations were regarded as reference stations to provide additional information. Three composite events were constructed based on the four series and were labelled case I, case II, and case III. The relevant information on the annual precipitation data series and the arrangement of composite events are summarized in Table 4.

RESULTS AND DISCUSSION

Marginal distribution and goodness-of-fit

The P-III distribution was applied to fit each data series on a univariate basis. The MLM was used to estimate

Table 3 | Statistics characteristics of the precipitation data at the selected gauging stations

	Xi'an	Dali	Lintong	Huayin
Min precipitation (mm)	285.2	240.8	302.3	302.5
Max precipitation (mm)	903.2	843.5	954.9	898.8
Average precipitation (mm)	571.9	500.8	579.5	567.8
Standard deviation	126.957	124.926	129.201	159.269

Table 4 | Information about the composite events

Case	X Station	N_X	n_X	Y Station	N_Y	e_{XY} n_{XY}
I	Xi'an	77	49		28	28
II	Dali	53	25	Huayin	28	28
III	Lintong	50	22		28	28

the distribution parameters. Using the estimated parameters, we compared the theoretical and empirical probabilities of the observed data (Figure 3). The empirical nonexceedance probabilities were calculated using the Weibull formula recommended by the Ministry of Water Resources of China (MWR 2006). The P-III distribution in Figure 3 shows a satisfactory fitting of the observed data series.

The K-S test and A-D test were used to evaluate the goodness-of-fit of the precipitation observations. The computed sample values, Dn and A_n^2, are listed in Table 5. The critical values of Dn and A_n^2 are also shown in Table 5 for simulation times $N = 5{,}000$ and significance levels of $\alpha =$ 0.20, 0.15, 0.10, 0.05, 0.02, 0.01. Because all the sample statistics are smaller than the corresponding critical values at various significance levels, the P-III distribution is acceptable as a marginal distribution.

Bivariate distribution of concurrent period based on the copula

As shown in Table 4, the length of the concurrent periods in all composite events is 28 years. Common measures, such as the Pearson correlation coefficient γ, Kendall's τ, and Spearman's ρ, were used to investigate the dependence of the concurrent variables. The results of the correlation coefficients and the corresponding p-values shown in Table 6 reveal the presence of significant dependent relationships between the concurrent periods in cases I, II, and III.

The joint distribution was constructed using the GH, Clayton and Frank copulas. The dependence parameters estimated based on the MLE and KTE method are listed in Table 7, and the RMSE, AIC and BIC for three copulas are also compared in this table. According to the smallest values of these three statistics, the best-fit copula and parameter estimation methods for the three composite cases are GH (MLE), Clayton (KTE) and GH (KTE). Table 8 lists the goodness-of-fit results of the best-fit copulas for the three cases. Given a significance level of $\alpha = 0.05$, all the test statistics based on the observed data are smaller than the corresponding critical values, indicating that GH, Clayton and GH

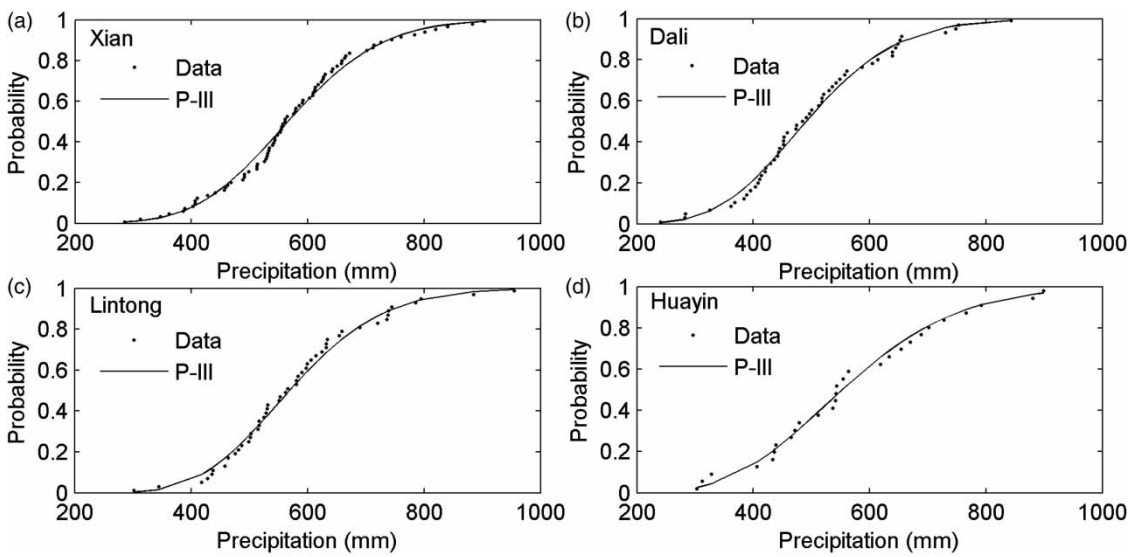

Figure 3 | Fitting distributions to the full length annual precipitation series: (a) Xi'an; (b) Dali; (c) Lintong; (d) Huayin.

Table 5 | Goodness-of-fit test for univariate probability distribution using the K-S and A-D tests

			Critical values of various significance levels *a*					
Station	**Type**	**Sample statistic**	**0.20**	**0.15**	**0.10**	**0.05**	**0.02**	**0.01**
Xi'an	Dn	0.0733	0.1196	0.1268	0.1366	0.1519	0.1692	0.1797
	A_n^2	0.3346	1.3756	1.5790	1.8912	2.4793	3.2856	3.9696
Dali	Dn	0.0607	0.1441	0.1525	0.1653	0.1834	0.2013	0.2188
	A_n^2	0.2794	1.4356	1.6459	1.9710	2.5110	3.2464	3.8063
Lintong	Dn	0.0701	0.1494	0.1585	0.1701	0.1888	0.2096	0.2249
	A_n^2	0.2561	1.3832	1.5679	1.8546	2.4415	3.3945	4.0102
Huayin	Dn	0.0803	0.1977	0.2097	0.2260	0.2522	0.2797	0.2975
	A_n^2	0.1934	1.4283	1.6449	1.9447	2.4772	3.2023	3.7704

Table 6 | Correlation coefficients of concurrent data

	Pearson		**Spearman**		**Kendall**	
Case	γ	***p*1**	ρ	***p*2**	τ	***p*3**
I	0.7825	8.66E-07	0.6579	0.0002	0.5079	8.34E-05
II	0.8590	4.93E-09	0.7942	2.10E-06	0.6032	1.60E-06
III	0.8396	2.33E-08	0.7690	4.18E-06	0.6085	4.18E-06

copulas are acceptable for modeling the bivariate joint distributions of the concurrent data in cases I, II, III, respectively.

Composite likelihood-based design values and uncertainty estimation

Considering P-III as a marginal distribution and the best-fit copulas selected above, the estimates of the distribution parameters and the variance-covariance matrix based on the CBCLA were obtained from Equations (8) and (25), respectively. The precipitation design values along with the associated standard errors and widths of the 95% confidence intervals for 10, 20, 50, 100, 200 and 500 year return periods based on these parameters and the dispersion

Table 7 | Copula dependence parameter estimates and fitting evaluation

Case	**Method**	**Copula**	$\hat{\theta}$	**AIC**	**BIC**	**RMSE**
I	KTE	Frank	5.8824	−180.4813	−179.1491	0.0377
		GH	2.0323	−182.5350	−181.2028	0.0364
		Clayton	2.0645	−182.2199	−180.8877	0.0366
		Frank	5.9547	−180.5580	−179.2258	0.0377
	MLE	GH	2.0430	−182.6685	−181.3363	0.0363
		Clayton	1.6118	−178.7845	−177.4523	0.0389
	KTE	Frank	8.0150	−182.2286	−180.8964	0.0366
II		GH	2.5200	−177.2635	−175.9313	0.0400
		Clayton	3.0400	−183.6340	−182.3018	0.0357
		Frank	7.6888	−181.5102	−180.1780	0.0371
	MLE	GH	2.4162	−176.0000	−174.6678	0.0409
		Clayton	1.5949	−166.9416	−165.6094	0.0481
	KTE	Frank	8.1601	−200.2240	−200.2240	0.0265
III		GH	2.5541	−204.5003	−203.1681	0.0246
		Clayton	3.1081	−193.9300	−192.5978	0.0297
		Frank	7.8457	−200.3987	−199.0665	0.0265
	MLE	GH	2.5534	−204.4974	−203.1652	0.0246
		Clayton	2.1037	−189.1661	−187.8339	0.0323

Table 8 | Goodness-of-fit test for copulas using the K-S and A-D tests

Case	Type	Sample statistic	Critical values of various significance level *a*					
			0.20	0.15	0.10	0.05	0.02	0.01
I	Dn	0.1139	0.9351	0.9467	0.9550	0.9614	0.9745	0.9883
	A_n^2	1.0468	1.2392	1.3789	1.5587	2.0436	2.2835	2.5604
II	Dn	0.1423	0.9283	0.9408	0.9598	0.9896	0.9907	0.9961
	A_n^2	0.4408	1.5004	2.1661	2.6674	3.3539	4.8191	5.0050
III	Dn	0.0895	0.9489	0.9581	0.9706	0.9882	0.9969	0.9985
	A_n^2	0.3370	1.7051	2.0485	2.3323	3.0700	3.9724	4.7634

characteristics are shown in Table 9. The precipitation design values, standard errors and 95% confidence interval widths derived for only a univariate basis are also given in Table 9.

In addition, the standard errors and confidence widths obtained from the composite cases were smaller than those derived from the univariate case (see Table 9), indicating that the CBCLA yields a more precise parameter estimation. Table 10 compares the differences in the uncertainty reductions for the standard errors and confidence widths under different situations. For example, for composite cases I, II, and III, the standard error and confidence interval for a return period of $T = 50$ years decreased 10.71, 2.48, and 2.73%, respectively, compared with the univariate case when using CBCLA.

Furthermore, the reductions in the standard errors and confidence widths in case I (10–20%) are greater than those in cases II and III (1.5–3%) due to the length of the exclusive period in the additional data. For the same concurrent period length, the exclusive period in case I is longer than those in cases II and III (Table 4). The increased information for a shorter series is more significant when the length of the associated data is longer. As a result, integrating the multivariate data set in the CBCLA is more advantageous than the univariate analysis.

Overall, the uncertainty in the design values decreases when using the CBCLA, which validates the merits of using this approach to improve the accuracy and precision of design values with insufficient data.

CONCLUSIONS

To summarize, this study presented a CBCLA for the parameter estimation of shorter series with a P-III distribution. Using this method improved the accuracy of

Table 9 | Results of the annual precipitation design values, standard errors and confidence widths based on the univariate method and CBCLA (cases I, II, and III) for Huayin station (mm)

Case	Return period (years)	Design value	Standard error	Confidence width	Case	Return period (years)	Design value	Standard error	Confidence width
Univariate	10	775.2	51.2	200.7	I	10	757.8	45.6	178.8
	20	841.8	64.8	253.9		20	818.3	57.7	226.1
	50	919.8	85.0	333.2		50	889.1	75.9	297.5
	100	973.6	101.4	397.5		100	937.8	90.8	355.7
	200	1024.0	118.5	464.3		200	983.6	106.2	416.3
	500	1086.8	141.7	555.5		500	1040.6	127.4	499.2
II	10	788.2	49.8	195.1	III	10	775.0	49.7	194.9
	20	854.8	62.9	246.7		20	840.7	62.9	246.5
	50	932.8	82.9	325.0		50	917.5	82.7	324.1
	100	986.5	99.2	389.0		100	970.4	98.8	387.4
	200	1037.0	116.3	455.7		200	1020.1	115.6	453.2
	500	1099.8	139.6	547.1		500	1081.9	138.6	543.1

Table 10 | Reduction in the uncertainty in the annual precipitation design values based on CBCLA (cases I, II, and III) compared with the univariate case for Huayin station (%)

	Case I to Univariate			Case II to Univariate			Case III to Univariate		
Return period (years)	Design value	Standard error	Confidence width	Design value	Standard error	Confidence width	Design value	Standard error	Confidence width
10	−2.24	−10.93	−10.93	1.68	−2.81	−2.81	−0.02	−2.92	−2.92
20	−2.79	−10.93	−10.93	1.55	−2.83	−2.83	−0.14	−2.92	−2.92
50	−3.34	−10.71	−10.71	1.41	−2.48	−2.48	−0.26	−2.73	−2.73
100	−3.67	−10.52	−10.52	1.33	−2.15	−2.15	−0.33	−2.56	−2.56
200	−3.95	−10.33	−10.33	1.27	−1.85	−1.85	−0.38	−2.41	−2.41
500	−4.25	−10.12	−10.12	1.19	−1.50	−1.50	−0.45	−2.22	−2.22

the parameter estimation, as reflected in the confidence intervals of the design precipitation values. Four annual precipitation series in the WRB of China were selected as a case study. The following conclusions were drawn:

1. Three bivariate composite events were constructed based on four precipitation series of different lengths. The bivariate data set in each composite event consisted of two parts, i.e., the concurrent period and the exclusive part in the longer series. The marginal distributions of the concurrent and exclusive part were fitted using a P-III distribution, and the corresponding parameters were estimated using the MLM. The dependence structure of the overlapping records was modeled using Archimedean copulas, and a copula-based bivariate composite likelihood function for parameter estimation was established based on the likelihood function. Finally, the precision of the estimates calculated with this approach was compared with the precision of the univariate maximum likelihood estimates. The results show that the uncertainty in the precipitation design values due to inadequate data decreased and that this reduction in uncertainty became more significant as the length of the exclusive data series increased.
2. Monte-Carlo simulation results showed that using CBCLA with P-III marginals achieved improved design values for short sequences in terms of both standard error and confidence intervals.
3. The main advantage of this approach is that it considers the unused data in a longer data series to build a multivariate data set with different lengths, which cannot be implemented using conventional methods. This integration enhances the information for short series and yields improved parameter estimation and estimation precision. Furthermore, copulas were used to build joint distribution functions in this approach to avoid the assumption that the marginal distribution must belong to the same type, as in the traditional multivariate distribution. The marginal and bivariate distributions used the P-III distribution and Archimedean copulas, respectively. Further studies on this approach with other marginals and copula models related to other hydrologic designs should be conducted.

ACKNOWLEDGEMENTS

The present study is financially supported by the National Natural Science Foundation of China (Grant Nos 51479171, 51179160 and 51579059). The authors also wish to express their cordial gratitude to the editor and anonymous reviewers for their helpful comments which have greatly helped to improve the quality of this paper.

REFERENCES

Bobée, B. 1979 Comment on 'The log Pearson type 3 distribution: the T-year event and its asymptotic standard error by maximum likelihood theory', by R. Condie. *Water Resour. Res.* **15** (1), 189–190.

Bobée, B. & Rasmussen, P. F. 1995 Recent advances in flood frequency analysis. *Rev. Geophys.* **33** (S2), 1111–1116.

Chen, L., Guo, S. L., Yan, B. W., 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.

Chen, L., Singh, V. P., Lu, W., Zhang, J., Zhou, J. & Guo, S. 2016 Streamflow forecast uncertainty evolution and its effect on real-time reservoir operation. *J. Hydrol.* **540**, 712–726.

Chowdhary, H. & Singh, V. 2009 *Copula Approach for Reducing Uncertainty in Design Flood Estimates in Insufficient Data Situations.* Report, World Environmental and Water Resources Congress 2009, Great Rivers. ASCE, Reston, VA.

Chowdhary, H. & Singh, V. P. 2010 Reducing uncertainty in estimates of frequency distribution parameters using composite likelihood approach and copula-based bivariate distributions. *Water Resour. Res.* **46** (11), W11516.

Coles, S., Bawa, J., Trenner, L. & Dorazio, P. 2001 *An Introduction to Statistical Modeling of Extreme Values.* Springer, London.

Cunderlik, J. M. & Burn, D. H. 2003 Non-stationary pooled flood frequency analysis. *J. Hydrol.* **276** (1–4), 210–223.

Dobrić, J. & Schmid, F. 2007 A goodness of fit test for copulas based on rosenblatt's transformation. *Comput. Stat. Data Anal.* **51** (9), 4633–4642.

Escalante-Sandoval, C. 2007 Application of bivariate extreme value distribution to flood frequency analysis: a case study of northwestern Mexico. *Nat. Hazards* **42** (1), 37–46.

Escalante-Sanboval, C. A. & Raynal-Villasenor, J. A. 1998 Multivariate estimation of floods: the trivariate gumbel distribution. *J. Stat. Comput. Sim.* **61** (4), 313–340.

Fu, G. & Butler, D. 2014 Copula-based frequency analysis of overflow and flooding in urban drainage systems. *J. Hydrol.* **510**, 49–58.

Genest, C. & Favre, A. 2007 Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.* **12** (4), 347–368.

Gilroy, K. L. & McCuen, R. H. 2012 A nonstationary flood frequency analysis method to adjust for future climate change and urbanization. *J. Hydrol.* **414**, 40–48.

Haan, C. T. 1977 *Statistical Methods in Hydrology.* Iowa State University Press, Iowa, USA.

Halbert, K., Nguyen, C. C., Payrastre, O. & Gaume, E. 2016 Reducing uncertainty in flood frequency analyses: a comparison of local and regional approaches involving information on extreme historical floods. *J. Hydrol.* **541**, 90–98.

Huang, S., Huang, Q., Chang, J., Chen, Y., Xing, L. & Xie, Y. 2015 Copulas-based drought evolution characteristics and risk evaluation in a typical arid and semi-arid region. *Water Resour. Manag.* **29** (5), 1489–1503.

Kamwi, I. S. 2005 *Fitting Extreme Value Distributions to the Zambezi River Flood Water Levels Recorded at Katima Mulilo in Namibia*. PhD thesis, University of the Western Cape, South Africa.

Khaliq, M. N., Ouarda, T. B. M. J., Ondo, J. C., Gachon, P. & Bobée, B. 2006 Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: a review. *J. Hydrol.* **32** (3–4), 534–552.

Leclerc, M. & Ouarda, T. B. M. J. 2007 Non-stationary regional flood frequency analysis at ungauged sites. *J. Hydrol.* **343** (3–4), 254–265.

Li, J. & Tan, S. 2015 Nonstationary flood frequency analysis for annual flood peak series, adopting climate indices and check dam index as covariates. *Water Resour. Manag.* **29** (15), 5533–5550.

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.

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

MWR (Ministry of Water Resources) 2006 *Regulation for Calculating Design Flood of Water Resources and Hydropower Projects.* Chinese Water Resources and Hydropower Press, Beijing, China (in Chinese).

Nadarajah, S. 2006 Information matrix for the bivariate Gumbel distribution. *Appl. Math. Comput.* **172** (1), 394–405.

Nelsen, R. B. 2007 *An Introduction to Copulas.* Springer Science & Business Media, New York, USA.

Posada, D. & Buckley, T. R. 2004 Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Bio.* **53** (5), 793–808.

Rao, A. R. & Hamed, K. 1999 *Flood Frequency Analysis.* CRC Press, Washington, DC, USA.

Raynal Villasenor, J. A. 1985 *Bivariate Extreme Value Distributions Applied to Flood Frequency Analysis*. PhD thesis, Colorado State University, Colorado, USA.

Raynal-Villasenor, J. A. & Salas, J. D. 2008 *Using Bivariate Distributions for Flood Frequency Analysis Based on Incomplete Data. Report, World Environmental and Water Resources Congress*. ASCE, Honolulu, Hawii.

Reis Jr, D. S. & Stedinger, J. R. 2005 Bayesian MCMC flood frequency analysis with historical information. *J. Hydrol.* **313** (1–2), 97–116.

Rueda, E. 1981 *Transfer of Information for Flood Related Variables*. MS thesis, Colorado State Univ., Fort Collins, Colorado, USA.

Salas, J. D. & Obeysekera, J. 2014 Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events. *J. Hydrol. Eng.* **19** (3), 554–568.

Salvadori, G. & De Michele, C. 2007 On the use of copulas in hydrology: theory and practice. *J. Hydrol. Eng.* **12** (4), 369–380.

Salvadori, G., De Michele, C. & Durante, F. 2011 On the return period and design in a multivariate framework. *Hydrol. Earth Syst. Sci.* **15** (11), 3293–3305.

Salvadori, G., Durante, F. & De Michele, C. 2013 Multivariate return period calculation via survival functions. *Water Resour. Res.* **49** (4), 2308–2311.

Sandoval, C. E. & Raynal-Villasenor, J. 2008 Trivariate generalized extreme value distribution in flood frequency analysis. *Hydrolog. Sci. J.* **53** (3), 550–567.

Shiau, J. T. 2006 Fitting drought duration and severity with two-dimensional copulas. *Water Resour. Manag.* **20** (5), 795–815.

Silva, A. T., Portela, M. M., Baez, J. & Naghettini, M. 2012 Construction of confidence intervals for extreme rainfall

quantiles. *WIT Trans. Inform. Commun. Technol.* **44**, 293–304.

Singh, V. P. & Strupczewski, W. G. 2002 On the status of flood frequency analysis. *Hydrol. Process.* **16** (18), 3737–3740.

Sklar, M. 1959 *Fonctions de Répartition À N Dimensions Et Leurs Marges. Université Paris 8*. Publ. Inst. Stat. Univ., Paris, France, pp. 229–231.

Song, S. & Singh, V. P. 2010 Meta-elliptical copulas for drought frequency analysis of periodic hydrologic data. *Stoch. Env. Res. Risk A.* **24** (3), 425–444.

Stedinger, J. R. & Cohn, T. A. 1986 Flood frequency analysis with historical and paleoflood information. *Water Resour. Res.* **22** (5), 785–793.

Strupczewski, W. G. & Kaczmarek, Z. 2001 Non-stationary approach to at-site flood frequency modelling II. Weighted least squares estimation. *J. Hydrol.* **248** (1–4), 143–151.

Strupczewski, W. G., Singh, V. P. & Feluch, W. 2001 Non-stationary approach to at-site flood frequency modelling I. Maximum likelihood estimation. *J. Hydrol.* **248** (1–4), 123–142.

Vasiliades, L., Galiatsatou, P. & Loukas, A. 2015 Nonstationary frequency analysis of annual maximum rainfall using climate covariates. *Water Resour. Manag.* **29** (2), 339–358.

Villarini, G., Smith, J. A., Serinaldi, F., Bales, J., Bates, P. D. & Krajewski, W. F. 2009 Flood frequency analysis for nonstationary annual peak records in an urban drainage basin. *Adv. Water Resour.* **32** (8), 1255–1266.

Villarini, G., Smith, J. A. & Napolitano, F. 2010 Nonstationary modeling of a long record of rainfall and temperature over Rome. *Adv. Water Resour.* **33** (10), 1256–1267.

Vittal, H., Singh, J., Kumar, P. & Karmakar, S. 2015 A framework for multivariate data-based at-site flood frequency analysis: essentiality of the conjugal application of parametric and nonparametric approaches. *J. Hydrol.* **525**, 658–675.

Vogel, R. M. & Stedinger, J. R. 1985 Minimum variance streamflow record augmentation procedures. *Water Resour. Res.* **21** (5), 715–723.

Wang, Q. J. 1990a Estimation of the GEV distribution from censored samples by method of partial probability weighted moments. *J. Hydrol.* **120** (1–4), 103–114.

Wang, Q. J. 1990b Unbiased estimation of probability weighted moments and partial probability weighted moments from systematic and historical flood information and their application to estimating the GEV distribution. *J. Hydrol.* **120** (1), 115–124.

Wang, Q. J. 1996a Using partial probability weighted moments to fit the extreme value distributions to censored samples. *Water Resour. Res.* **32** (6), 1767–1771.

Wang, Q. J. 1996b Direct sample estimators of L moments. *Water Resour. Res.* **32** (12), 3617–3619.

Wang, Q. J. 1997a LH moments for statistical analysis of extreme events. *Water Resour. Res.* **33** (12), 2841–2848.

Wang, Q. J. 1997b Using higher probability weighted moments for flood frequency analysis. *J. Hydrol.* **194** (1), 95–106.

Xiong, L., Du, T., Xu, C. Y., Guo, S., Jiang, C. & Gippel, C. J. 2015 Non-stationary annual maximum flood frequency analysis using the norming constants method to consider non-stationarity in the annual daily flow series. *Water Resour. Manag.* **29** (10), 3615–3633.

Yue, S. 2000 The Gumbel logistic model for representing a multivariate storm event. *Adv. Water Resour.* **24** (2), 179–185.

Yue, S. 2001 A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrol. Process.* **15** (6), 1033–1045.

Yue, S. 2002 The bivariate lognormal distribution for describing joint statistical properties of a multivariate storm event. *Environmetrics* **13** (8), 811–819.

Yue, S., Ouarda, T. B. M. J. & Bobée, B. 2001 A review of bivariate gamma distributions for hydrological application. *J. Hydrol.* **246** (1–4), 1–18.

Zeng, H., Feng, P. & Li, X. 2014 Reservoir flood routing considering the non-stationarity of flood series in north China. *Water Resour. Manag.* **28** (12), 4273–4287.

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

Zhang, L. & Singh, V. P. 2007a Gumbel–Hougaard copula for trivariate rainfall frequency analysis. *J. Hydrol. Eng.* **12** (4), 409–419.

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

Zhang, Q., Xiao, M., Singh, V. P. & Chen, X. 2013 Copula-based risk evaluation of hydrological droughts in the East River basin, China. *Stoch. Env. Res. Risk A.* **27** (6), 1397–1406.

Zhang, Q., Xiao, M. & Singh, V. P. 2015 Uncertainty evaluation of copula analysis of hydrological droughts in the East River basin, China. *Glob. Planet Change* **129**, 1–9.

First received 22 February 2017; accepted in revised form 11 September 2017. Available online 28 November 2017