Identification of suitable copulas for bivariate frequency analysis of flood peak and flood volume data

Hemant Chowdhary, Luis A. Escobar and Vijay P. Singh

ABSTRACT

Multivariate flood frequency analysis, involving flood peak flow, volume and duration, has been traditionally accomplished by employing available functional bivariate and multivariate frequency distributions that have a restriction on the marginals to be from the same family of distributions. The copula concept overcomes this restriction by allowing a combination of arbitrarily chosen marginal types. It also provides a wider choice of admissible dependence structure as compared to the conventional approach. The availability of a vast variety of copula types makes the selection of an appropriate copula family for different hydrological applications a non-trivial task. Graphical and analytic goodness-of-fit tests for testing the suitability of copulas are beginning to evolve and are being developed; there is limited experience of their usage at present, especially in the hydrological field. This paper provides a step-wise procedure for copula selection and illustrates its application to bivariate flood frequency analysis, involving flood peak flow and volume data. Several graphical procedures, tail dependence characteristics, and formal goodness-of-fit tests involving a parametric bootstrap-based technique are considered while investigating the relative applicability of six copula families. The Clayton copula has been identified as a valid model for the particular flood peak flow and volume data set considered in the study.

Key words | bivariate flood frequency, Chi-plot, Clayton copula, copula, generating functions, K-plot

Hemant Chowdhary (corresponding author)
Department of Civil and Environmental Engineering,
Louisiana State University,
Baton Rouge, Louisiana,
USA
E-mail: *hchowd1@lsu.edu*

Luis A. Escobar
Department of Experimental Statistics,
Louisiana State University,
Baton Rouge, Louisiana,
USA

Vijay P. Singh
Department of Biological and Agricultural Engineering & Department of Civil & Environmental Engineering,
Texas A & M University,
College Station, Texas,
USA

INTRODUCTION

Flood frequency analysis typically involves fitting univariate distributions to annual maximum (AM) flows or to peak over threshold (PoT) flows or to successive peak flows, observed at a location of interest along a river. The main objective of various hydrological designs, e.g., for dam spillways, levees, bridges, etc., has been to ensure safety with respect to flows that would have an average inter-arrival period, also called the return period, less than or equal to a specified design period. Hydrological processes, however, exhibit multivariate characteristics and a simultaneous consideration of various component processes may be crucial and required in certain situations. The flood phenomenon is also a multidimensional process, having peak flood flow, corresponding volume and duration, time to flood peak, rate-of-rise, and rate of recession as important hydrologic features. The decision to employ univariate, bivariate or multivariate distribution is, however, made primarily on the basis of the objectives of any particular application. For example, for risk assessment for a small- to moderate-sized flood protection structure, a univariate flood frequency analysis of annual flood peaks may suffice. On the other hand, in situations where storage has a significant effect on flood attenuation, flood duration and/or volume are also required to be considered along with the peak flood discharge. Such analysis is helpful in determining the spillway capacity and associated maximum water levels for proposed dams, as well as for assessing the adequacy of spillways of existing dams. Thus, wherever failure probability or variables of interest are a function of two or more hydrological

doi: 10.2166/nh.2011.065

variables, a multivariate consideration is desirable. Such a broader treatment of flow variables has been done in the past by employing available functional multivariate distributions. In recent years, copula-based distributions are being employed for the same purpose, as they overcome some of the limitations of traditional methods. Indicating the usefulness of multivariate hydrological frequency analysis, a brief review of studies done in the past using functional distributions and those done recently employing copula-based distributions is presented next.

Traditionally, bivariate normal, lognormal, exponential, or Gumbel (called mixed Gumbel) distributions have been applied for hydrological variables, such as flood peak, and associated volume and duration. Gupta *et al.* (1976), Todorovic & Woolhiser (1972), and Todorovic (1978) have discussed distributions for time of occurrence of peak flow in relation to flood events. Ashkar & Rousselle (1982) discussed the multivariate nature of flood peak and corresponding volume and duration. Bivariate stochastic model for flood peak and volume based on the principle of maximum entropy has been suggested by Krstanovic & Singh (1987). Sackl & Bergmann (1987) employed bivariate normal distribution on transformed flood peak and volume in order to estimate the design volume for retention basins. Correia (1987) obtained the bivariate density function for flood peak and duration by assuming their conditional distribution to be normally distributed and the marginal distribution for duration to be exponentially distributed. The application of the general form of logistic model for a bivariate extreme value distribution was demonstrated for obtaining a flood frequency distribution at a station on the basis of flood information from two stations immediately upstream of the junction by Raynal-Villasenor & Salas (1987). In an indirect approach, Rosbjerg (1987) obtained a frequency distribution of annual maximum flood from successive peak floods, employing Marshall–Olkin bivariate exponential distribution (Marshall & Olkin 1967). Escalante-Sandoval & Raynal-Villasenor (1994) showed the usefulness of a logistic model for trivariate general extreme value distribution for flood peak flows from adjoining stations. A bivariate meta-Gaussian distribution proposed by Kelly & Krzysztofowicz (1997) allowed the specification of arbitrary marginals and covered the full dependence range and is based on the assumption that normal quantile transformed (NQT) variates follow bivariate normal distribution. Extending the work of Raynal-Villasenor & Salas (1987) and Escalante-Sandoval & Raynal-Villasenor (1994), Escalante-Sandoval (1998) employed a multivariate extreme value distribution with mixed Gumbel marginals. Another instance of bivariate consideration of flood peak and volume, normalized using a two step Box–Cox transformation, is reported by Goel *et al.* (1998). Conditional probability of failure functions, based on both flood peak and duration, has been studied for risk assessment of levees and embankments by USACE (1999). Yue *et al.* (1999) employed bivariate Gumbel mixed model (Gumbel 1960) for obtaining pair-wise joint and conditional probabilities for flood peak, volume and duration. Yue (2000, 2001) applied bivariate lognormal and bivariate extreme value distributions for multivariate flood frequency analysis.

Almost all of the above multivariate distribution applications, except for the meta-Gaussian method of Kelly & Krzysztofowicz (1997), had limitations of admitting marginals from the same families. However, different hydrological applications may involve multiple variables, not all of which may belong to the same distribution type. Transformation to normal distribution and consequent fitting of multivariate normal distribution has often been resorted to in such situations. Extensive efforts, spanning decades of research work in the area of flood frequency analysis, has resulted in the identification of some plausible candidate distribution functions. The lack of multivariate distributions featuring marginals from different distributions restricts the ability to directly utilize such potential univariate distributions. This makes migration from univariate to multivariate flood frequency analysis sub-optimal (Choulakian *et al.* 1990). Furthermore, several multivariate distributions do not also allow for a full coverage of possible dependence between different variables. A few examples in this regard could be the bivariate exponential and bivariate Gumbel distributions. The bivariate exponential distribution imposes a critical restriction on the variables to be negatively associated such that Pearson's correlation coefficient is between -0.404 and 0. On the other hand, the bivariate exponential distribution given by Moran and studied by Nagao & Kadoya (1971) admits the full positive correlation coefficient. Although both these formulations complement each other, they are not comprehensive enough individually. Similarly, the bivariate extreme value distribution given by Gumbel (1960) admits only a partial positive range of Pearson's correlation coefficient to the extent of 0 to 2/3 only. Although not applied for hydrological variables, Farlie–Gumbel–Morgenstern (FGM) family of distributions, as applied for rainfall variables by Singh & Singh (1991) and studied later by Long & Krzysztofowicz (1992), are applicable for only weakly associated variates having Kendall's tau between −2/9 and 2/9 and may thus be of limited use in hydrological applications. Another concern, while using conventional multivariate formulations, is of Pearson's linear correlation coefficient being linked to the dependence parameter either directly or indirectly. Pearson's linear correlation coefficient is not invariant to non-linear monotonic transformations and depicts linear correlation rather than the functional association and may also not be even estimable in certain situations that involve heavy-tailed distributions (Genest & Favre 2007). The copula concept overcomes some of these restrictions posed by the conventional multivariate distributions and is emerging as a new way of multivariate frequency distribution analysis.

Copulas are essentially the mapping functions that combine uniformly distributed marginals in order to represent the joint distribution and dependence structure of arbitrarily distributed dependent variables. The copula theory has been in vogue for some time now, especially with respect to actuarial science and finance applications, and in recent years has also been applied in the field of hydrological engineering. Among the first copula-based hydrological studies, De Michele & Salvadori (2003) indicated the suitability of Frank copula for joint distribution of negatively associated storm intensity and storm duration data. Favre *et al.* (2004) employed FGM, Clayton, and Frank copulas for two applications; the first involving peak flow for a run-of-river power station that was modeled as a combination of peak flows from an upstream dam and the intermediate watershed, and the second for a better assessment of the joint and conditional processes of peak flow and volume. The Frank copula was adjudged marginally better than the Clayton copula. A comprehensive account of definitions and formulae for various types of return periods, namely conditional and secondary return periods in the context of different types of bivariate events, and copula-based frequency analysis was presented by Salvadori & De Michele (2004). De Michele *et al.* (2005) employed the Gumbel–Hoogaurd (GH) copula for generating a large number of pairs of generalized extreme value distributed annual peak flow and volume data in order to assess the adequacy of dam spillway capacity. The suitability of a fully nested asymmetric Frank copula over symmetric Frank copula and three-dimensional Gumbel logistic distribution for flood peak, volume and duration was established by Grimaldi & Serinaldi (2006*a*). Grimaldi & Serinaldi (2006*b*) employed several trivariate copulas for determining joint and conditional distributions among design hyetograph variables, such as critical depth, total depth and time to peak. Shiau *et al.* (2006) presented explicit relationships between joint return periods and univariate return periods for flood peak and volume data by considering the Ali–Mikhail–Haq (AMH), Clayton, Frank, Galambos, GH, and Plackette copulas. Zhang & Singh (2006) reported that of the four Archimedean copulas (AMH, Clayton, Frank and GH), the GH copula provided a better fit for the pair-wise distributions among flood peak, volume and duration and that the fit was also better than the conventional bivariate Gumbel mixed and bivariate normal distributions.

An elaborate review and illustration of copula estimation and inference procedures provided by Genest & Favre (2007) is helpful for the end users in applied fields, such as hydrology. Among a multitude of possible copula types they found five copulas – GH, Galambos, Husler–Reiss, BB1, and BB5 – to be plausible candidates for flood peak flow and volume data. The advantage of meta-elliptical copulas over Archimedean copulas with respect to admittance of different and/or negative dependence among flood variables has been demonstrated by Genest *et al.* (2007). For this trivariate case of the peak flow, volume and duration variables, important for inundation and flood management practices, various estimation and inference methods were illustrated by considering eight meta-elliptical copula types. Salvadori & De Michele (2007) discussed alternative trivariate copula models for characterizing temporal properties of storms, involving mean intensity, and wet and dry period durations. Along with a comprehensive review of storm models available in the literature they presented a storm volume distribution model that was based on a bivariate copula involving storm intensity and storm duration. Serinaldi & Grimaldi (2007) highlighted the difference of symmetric and asymmetric dependence structures in fitting Archimedean copulas to trivariate flood and trivariate sea wave data. Singh & Zhang (2007) employed the Frank copula for obtaining intensity–duration–frequency (IDF) curves on the basis of probabilities of rainfall depth, conditioned on duration. It was, however, mentioned by Chowdhary (2009) that since storm depth–duration data did not conform to the requirements of the IDF curves, the proposed procedure may not be applicable. Zhang & Singh (2007*a*, *b*) fitted the bivariate (AMH, Clayton, Frank and GH) and trivariate (GH) Archimedean copulas to rainfall intensity, depth, and duration data, respectively. Zhang & Singh (2007*c*) also employed the GH copula for obtaining a trivariate distribution of flood peak flow, volume and duration data. It may, however, be mentioned that the use of a single parameter GH copula for a trivariate rainfall variables case, having both positive and negative pair-wise associations, may pose limitations in light of the inadequacy of single parameter trivariate copulas to represent such differing pair-wise associations (Grimaldi & Serinaldi 2006*a*; Kao & Govindaraju 2008). Furthermore, Chowdhary (2008) indicated that for this particular case a bivariate distribution may suffice as the mean rainfall intensity was functionally related to the other two variables.

Kao & Govindaraju (2007*a*) employed survival copula to derive zero-runoff probabilities involving exponentially distributed rainfall intensity and duration data in order to obtain a probabilistic structure of rainfall excess using the SCS method. It was opined that the probabilistic structure of storm surface runoff was highly sensitive to the dependence between intensity and duration and the copula method provided easier estimators for the same. Kao & Govindaraju (2007*b*) presented a copula-based description of the dependence structure for annual extreme rainfall events selected on the basis of three criteria: annual maximum volume (AMV), annual maximum peak intensity (AMI), and annual maximum cumulative probability (AMP). Four Archimedean copulas – AMH, Clayton, Frank, and Genest–Ghoudi (GG) – were explored for plausible alternatives out of which the Frank copula performed better than others. Renard & Lang (2007) presented case studies that utilized the Gaussian copula-based multivariate hydrological distributions for field significance determination, regional risk estimation, discharge–duration–frequency (QdF) curves, and for regional frequency analysis. Nikoloulopoulos & Karlis (2007) applied a model averaging approach in order to reduce effects of copula misspecification for testing the seismic gap hypothesis involving dependence among two geophysical variables: earthquake intensity and elapsed time.

Tail dependence characteristics constitute important features that differentiate extreme value copulas from other copula structures. Employing extreme value copulas when tail dependence properties indicate otherwise, or vice versa, can lead to serious overestimation or underestimation, respectively. Poulin *et al.* (2007) emphasized this need by considering seven copulas from three copula families – Archimedean, extreme value and meta-elliptical – while selecting copulas for joint distribution of flood peak flow and volume. Along with citing benefits of copulas, Dupuis (2007) also cautioned against ignoring the tail dependence characteristics by illustrating the consequences of misspecification of copulas by employing a simulation technique involving six copula families – Normal, Student-*t*, Frank, Clayton, Gumbel, and associated Clayton. Simulation results showed that the Frank copula performed relatively well when the true copula was any of the other five copulas. Another important conclusion reported in the study was that bivariate modeling led to a small gain in parameter estimation efficiency as compared with the univariate approach. Renard & Lang (2007) also cautioned against using the Gaussian copula when extrapolation was involved and data exhibited an asymptotic dependence.

Kao & Govindaraju (2008) showed the applicability of the Plackett copula family for a trivariate hydrological application involving AMP-based extreme rainfall variables. They stated that for three- or more-dimensional distributions, the Plackett copula family has an advantage over the Archimedean copulas in that they allow for the retention of pair-wise dependencies. Karmakar & Simonovic (2009) found representation by the GH copula better while obtaining pair-wise joint distributions among flood peak flow, volume and duration variables using the AMH, Clayton and GH copulas. Wang *et al.* (2009) reported a satisfactory application of the copula approach for obtaining flood quantiles downstream of a confluence, assuming concurrence among annual flood peaks in two upstream tributaries. They selected the Frank copula on the basis of better AIC values from three Archimedean copulas: Clayton, Frank, and GH. Exploring possibilities of its applicability, Huard *et al.* (2006) presented a Bayesian-based copula selection method. The Bayesian method has the advantage that it is not conditioned on parameter estimates. Silva & Lopes (2008) also employed the Bayesian method for estimating marginal and dependence parameters utilizing a deviance information criterion among a few others.

This paper illustrates a copula application for a bivariate data set involving flood peak flow and volume by considering six copula families and employing a number of graphical and formal goodness-of-fit tests. After a brief review of multivariate hydrological studies using conventional functional forms and copula-based methods in this "Introduction" section, the "Methods" section provides a short background of the copula concept, parameter estimation methods and a few graphical and analytical goodness-of-fit procedures that have been recently proposed. The "Application" section presents an elaborate illustration of the copula selection process by considering an observed flood peak flow and volume data set. The "Conclusions" section summarizes the main results and provides necessary explanations.

METHODS

The copula is a function that provides bivariate or multivariate probability functions in terms of constituent marginal probabilities. Although the development and application potential of copulas is a topic of current research, it is rooted in the theorem due to Sklar (1959), stating that the joint distribution function of any randomly distributed pair (X, Y) may be written as

$$H(x,y) = C[F(x), G(y)], \qquad x, y \,\epsilon\, R \tag{1}$$

where $F(x)$ and $G(y)$ are marginal probability distributions and $C = [0,\ 1] \times [0,\ 1] \rightarrow [0,\ 1]$, a mapping function, is the "copula". In turn it means that a valid probabilistic model for (X, Y) is obtained whenever the three constituents (C, F, and G) are chosen from given parametric families:

$$F(x; \boldsymbol{\delta}), G(y; \boldsymbol{\eta}), C(u, v; \boldsymbol{\theta}) \tag{2}$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are parameter vectors of marginal distributions, and $\boldsymbol{\theta}$ is the parameter vector for the dependence structure. u and v are the quantiles of the uniformly distributed variables $U = F(X)$ and $V = G(Y)$, respectively.

There is a multitude of copula types, broadly categorized in four classes – Archimedean, extreme value, elliptical, and other miscellaneous class. Copulas may also be categorized as single parameter or vector parameter copulas, depending upon the comprehensiveness with which the dependence structure can be defined by it. Joe (1997) and Nelsen (2006) provide a theoretical background and properties of a large number of copula types. Salvadori *et al.* (2007) makes a useful reference for the end users working on copula applications in the field of geosciences, as it includes necessary theoretical details besides providing pertinent examples from the fields of hydrology and geophysics. The overview of copula estimation and inference procedures given by Genest & Favre (2007) provides details of several important aspects of copula modeling, including recently proposed goodness-of-fit tests. It is acknowledged that this article makes substantial use of these sources. For the sake of simplicity, this study, however, is restricted to single-parameter bivariate copulas only.

Of the several copula families, the Archimedean family has been frequently applied in various fields including hydrology, owing to the ease of their construction, wider range of choice for the strength of dependence, and for several other nicer properties. This copula family has the following form:

$$\phi[H(x,y)] = \phi\{C[F(x), G(y)]\} = \phi[F(x)] + \phi[G(y)] \tag{3}$$

where ϕ, called a generator of copula, is a continuous strictly decreasing mapping function from [0, 1] to [0, ∞] such that $\phi(1) = 0$. The joint probability function for a bivariate random variable (X, Y) can then be written as

$$\begin{aligned} H(x,y) &= C[F(x), G(y)] = \phi^{[-1]}\{\phi[F(x)] + \phi[G(y)]\} \\ &= C(u,v) = \phi^{[-1]}\{\phi(u) + \phi(v)\} \end{aligned} \tag{4}$$

Here, $U = F(X)$ and $V = G(Y)$ are uniformly distributed probability integral transform variates. The function $\phi^{[-1]}(t)$: $[0,\ \infty] \rightarrow [0,\ 1]$ is the pseudo-inverse of the generating function. It is continuous and non-increasing on [0, ∞] and strictly decreasing on [0, $\phi(0)$] and is given by

$$\phi^{[-1]}(t) = \begin{cases} \phi^{-1}(t) & \forall \quad 0 \le t \le \phi(0) \\ 0 & \forall \quad \phi(0) \le t < \infty \end{cases}$$

The generator is termed "strict" and the resulting copula a strict copula when $\phi(0) = \infty$. The dependence parameter θ is hidden in the generating function $\phi(t)$, e.g., for the Frank copula, the one that has been employed for several hydrological applications, the generating function involves θ in the form

$$\phi(t) = -\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right), \qquad \theta \,\epsilon\, (-\infty, \infty)$$

{\tf="Pi3"\ \char"5C}{0}

Employing the above generating function and the form of the Archimedean copulas given in Equation (4), the bivariate cumulative probability distribution function for the Frank copula is obtained as

$$\begin{aligned} C_\theta(u,v) &= -\frac{1}{\theta}\ln\left[1-\frac{(1-e^{-\theta u})(1-e^{-\theta v})}{1-e^{-\theta}}\right] \\ &= -\frac{1}{\theta}\ln\left[1-\frac{(1-e^{-\theta F_X(x)})(1-e^{-\theta F_Y(y)})}{1-e^{-\theta}}\right] \\ &= H(x,y) \end{aligned} \tag{6}$$

$C_\theta(u, v)$ here is called the copula probability function. Double differentiating this probability function we get the copula density as

$$c_\theta(u,v) = \frac{\theta e^{-\theta(u+v)}}{(1-e^{-\theta})[\exp(-\theta C_\theta)]^2} \tag{7}$$

The joint density function in the original domain, taking $f(x)$ and $g(y)$ as marginal densities, can then be obtained as

$$\begin{aligned} h(x,y) &= \frac{\partial^2 C_\theta(u,v)}{\partial u\,\partial v}\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} = \frac{\partial^2 C_\theta(u,v)}{\partial u\,\partial v}\frac{\partial F(x)}{\partial x}\frac{\partial G(y)}{\partial y} \\ &= f(x)g(y)\, c_\theta(u,v) \end{aligned} \tag{8}$$

The Archimedean copula is a fairly large class, owing to easier evolution of newer copulas by coining valid generating functions as defined above. Nelsen (2006) enumerated 22 single-parameter bivariate Archimedean copulas along with their generating functions and probability distribution functions. AMH, Clayton, Frank, GH, Joe, Genest–Ghoudi, and Gumbel–Barnett are some of the commonly used Archimedean copulas.

Extreme value copulas are suitable when associated random variables formed by component-wise maxima are of interest. A copula C_* is considered an extreme value copula if there exists a copula C such that

$$C_*(u,v) = \lim_{n\to\infty} C^n(u^{1/n}, v^{1/n}) \tag{9}$$

In the above, C implies the copula representing a set of independent and identically distributed random variables $(X_i,\ Y_i)$; $i = 1{:}n$, and C_* is the joint distribution of their component-wise maxima $X_{(n)}$ and $Y_{(n)}$. Commonly used extreme value copulas are GH, Galambos and Husler–Reiss. It may be noted that GH is also an Archimedean copula along with being an extreme value copula. Durante & Salvadori (2009) provide an elaborate discussion on the construction of extreme value copulas for non-independent annual maximum peak flows from adjoining river gauging stations.

Meta-elliptical copulas consist of families that have elliptically contoured distributions, such as normal, Student-*t* and Cauchy copulas. The FGM, Plackett and Raftery copulas fall under the miscellaneous class. Reference may be made to Salvadori *et al.* (2007) for definitions and construction of copulas in extreme value, meta-elliptical and miscellaneous copula classes. Expressions for probability functions of a few copula families, their parameter space and generating functions are given in Table 1. The generating function is not applicable for the FGM and Galambos copulas, as these do not belong to the Archimedean families.

Parameter estimation methods

The copula dependence structure can be estimated using methods such as (a) moment-like method (MOM) based on the inversion of non-parametric dependence measures, (b) canonical or maximum pseudo-likelihood (MPL) method and (c) exact maximum likelihood (EML) method. The first two methods completely rely on the relative ranks of joint variates and thus render the determination of dependence structure completely independent of the choice of marginals. These methods are now outlined.

Moment-like method (MOM) based on inversion of dependence measures

This approach is based on the pretext that bivariate dependence structure is fully defined by the relative ranks of the

Table 1 | Probability function, parameter space, generating function and relationship of non-parametric dependence measure with association parameter for the six copula families under consideration

Copula	$C_\theta(u, v)$	Parameter Space	Generator $\phi(t)$	Kendall's tau τ
AMH[1]	$\dfrac{uv}{1-\theta(1-u)(1-v)}$	$[-1, 1)$	$\ln\dfrac{1-\theta(1-t)}{t}$	$A{-}B \ln(1{-}\theta)$
Clayton	$[\max(u^{-\theta}+\mathrm{v}^{-\theta}-1, 0)]^{-1/\theta}$	$[-1, \infty)\backslash\{0\}$	$\dfrac{1}{\theta}(t^{-\theta}-1)$	$\theta/(\theta+2)$
FGM	$uv[1+\theta(1-u)(1-v)]$	$[-1, 1]$	n.a.	$2\theta/9$
Frank[2]	$-\dfrac{1}{\theta}\ln\left[1+\dfrac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right]$	$(-\infty, \infty)\backslash\{0\}$	$-\ln\dfrac{e^{-\theta t}-1}{e^{-\theta}-1}$	$1+\dfrac{4}{\theta}[D_1(\theta)-1]$
Galambos[3]	$uv\exp[(\tilde{\mathrm{u}}^{-\theta}+\tilde{\mathrm{v}}^{-\theta})^{-1/\theta}]$	$[1, \infty)$	n.a.	n.a.
GH[3]	$\exp[-(\tilde{\mathrm{u}}^{\theta}+\tilde{\mathrm{v}}^{\theta})^{1/\theta}]$	$[1, \infty)$	$(-\ln \mathrm{t})^{\theta}$	$1{-}1/\theta$

[1] $A = \frac{3\theta-2}{3\theta}$ and $B = \frac{2(1-\theta)^2}{3\theta^2}$;
[2] $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t^k}{\exp(t-1)}dt$ is a Debye function;
[3] Expression involves $\tilde{\mathrm{u}} = -\ln u$ and $\tilde{\mathrm{v}} = -\ln v$

constituent variables. The non-parametric estimates of θ based on Kendall's tau τ and Spearman's rho ρ_s are obtainable from the relationships reported by Nelsen (2006) or Genest & Favre (2007) as

$$\tau = 4\int_{[0,1]^2} C(u,v)\ c_\theta(u,v)du\, dv - 1 \tag{10}$$

$$\rho_s = 12\int_{[0,1]^2} C(u,v)du\, dv - 3 \tag{11}$$

The above relationships for some copula families are available in closed form. For example, for the FGM copula, these relationships among dependence parameter and Kendall's tau and Spearman's rho are given as

$$\tau = \frac{2\theta}{9} \text{ and } \rho_s = \frac{\theta}{3} \text{ for } -1 \le \theta \le 1$$

The above results in a restricted admissible dependence space of $-0.2222 \le \tau \le 0.2222$ or $-0.3333 \le \rho_s \le 0.3333$. Based on this, a sample-based estimate of θ, much like a moment-based estimate, is obtained respectively as

$$\hat{\theta} = \frac{9\hat{\tau}}{2} \text{ and } \hat{\theta} = 3\hat{\rho}$$

Such relationships between association parameter θ with Kendall's tau for a few copula families are given in Table 1. For the cases for which closed forms are not forthcoming, numerical integration is done for relating τ and/or ρ_s with the association parameter θ. These relationships also define the dependence space for each copula, corresponding to the domain of association parameter θ. It must, however, be realized that this method is applicable for the single-parameter copula families only.

Maximum pseudo-likelihood (MPL) method

In this method, the dependence structure is again kept completely independent of the margins that are represented non-parametrically by the respective scaled ranks. Only the dependence parameter is obtained by maximizing the likelihood function. The log-likelihood function, assuming that C_θ is absolutely continuous with density c_θ, is of the form

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \log[c_\theta(\tilde{F}(x_i), \tilde{G}(y_i))] \\ &= \sum_{i=1}^{n} \log\left[c_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right)\right] \end{aligned} \tag{12}$$

where $\tilde{F}(x) = R_i/(n+1)$ and $\tilde{G}(y) = S_i/(n+1)$ are non-parametric marginal probabilities solely based on ranks. In other words, the maximum likelihood estimate of only θ is obtained in this method.

Exact maximum likelihood (EML) method

In this classical or exact maximum likelihood method, all parameters appearing in the log-likelihood function

$$l(\theta, \boldsymbol{\delta}, \boldsymbol{\eta}) = \sum_{i=1}^{n} \log\{c_\theta[F(x; \boldsymbol{\delta}), G(y; \boldsymbol{\eta})]\} \tag{13}$$

are simultaneously estimated. Here, $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are parameter vectors of the marginals $F(x;\ \boldsymbol{\delta})$ and $G(y;\ \boldsymbol{\eta})$, and θ is the association parameter. Another variant of this approach is referred to as the "Inference From Margins" (IFM) method, wherein univariate maximum likelihood estimates of $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are first obtained separately and then the estimate of θ is obtained by maximizing the likelihood function. The log-likelihood function for this can be expressed as

$$l(\theta) = \sum_{i=1}^{n} \log\{c_\theta[\breve{F}(x; \boldsymbol{\delta}), \breve{G}(y; \boldsymbol{\eta})]\} \tag{14}$$

where $\breve{F}(x; \boldsymbol{\delta})$ and $\breve{G}(y; \boldsymbol{\eta})$ indicate margins having parameters $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ that are obtained on a univariate basis. The IFM approach is advocated for multivariate copulas of larger dimensions when estimation through classical approach becomes computationally unwieldy. Furthermore, Joe (2005) found the IFM method to be nearly as efficient as EML method. However, caution has to be exercised while employing the IFM method, as misspecification of marginals may affect the dependence estimation. It is interesting to note that although classical maximum likelihood approach is more general, smaller mean squared errors were reported for the MPL method in a simulation study reported by Tsukahara (2005).

Dependence structure and plausible copulas

The fundamental objective in the copula selection process is to adequately represent the dependence structure of the data under consideration. There is a popular notion that the copula method overcomes various limitations faced by functional distributional forms, including those of restricted dependence space and of difficulty in having their multivariate extensions. However, this is not entirely true, as most copulas also are not comprehensive and cover a limited dependence space individually. Their multivariate extensions also invariably come with a variety of additional restrictions, e.g., extension to multivariate single-parameter copulas entails all pair-wise dependence levels to be equal; the fully nested Archimedean copulas require certain dependence compatibility conditions to be met. Another important aspect in the copula selection process is that of ensuring suitability in terms of tail dependence characteristics. Certain copulas may exhibit similar overall dependence features while possessing different lower and/or upper tail dependence characteristics. The compatibility of copula tail dependence characteristics with that exhibited by the process under consideration thus becomes an important goal. Furthermore, while functional multivariate distributions lack in terms of variety in their forms, there is a problem of a different nature with copulas and that is of a vast solution space (Michiels & Schepper 2008). There are numerous classes and types of copulas, making identification of suitable ones a non-trivial task.

Intuitively, the copula selection process can be split into two parts. In the first stage, the plausible copula types can be screened from the pool of all available copulas on the basis of admissible dependence ranges and tail dependence characteristics of individual copula types vis-à-vis the dependence characteristics of the data under consideration. Parameter estimation and goodness-of-fit tests for only those copulas that are screened in the first stage can then follow into the second stage, in order to select a final set of suitable copulas.

The inventory of admissible dependence space for 29 copulas provided by Michiels & Schepper (2008) becomes handy, while screening for the plausible copulas in the first stage. Each of these copula types may further have three more associated copulas, making the inventory even richer. An adapted and abridged version of this inventory, for purposes of illustration, is given in Table 2 for ten commonly used copula types. The shaded cells in this table imply admissibility of corresponding dependence ranges for different copula types. The table also lists lower and upper tail dependence coefficients λ_L and λ_U as defined and given in Nelsen (2006). On the basis of the knowledge of the strength of dependence and tail dependence characteristics of the data under consideration, a short-listing of plausible copula types can be done. An assessment of the dependence, along with their p-values, can be made by computing the values of Spearman's rho ρ_s and/or Kendall's tau τ based on observed data. The tail dependence characteristics of the process under consideration can be known from the available knowledge base about that process.

Additionally, a qualitative graphical assessment and re-affirmation of the strength of dependence can be done by plotting Chi-Plots and K-Plots as proposed by Fisher & Switzer (2001) and Genest & Boies (2003), respectively. Whereas Chi-Plots are akin to chi-square statistics for independence in a two-way table, K-Plots are similar to QQ-Plots. Conceptually, a Chi-Plot is a scatter plot of the measure of distance between an observation and the centre of all observations, and the chi-square test statistic for independence in a two-way frequency table generated by four regions delineated by that observation. Formally, this is a plot of (λ_i, χ_i), where

$$\lambda_i = 4\ \mathrm{sign}(\tilde{F}_i\ \tilde{G}_i)\ \max\ (\tilde{F}_i^2, \tilde{G}_i^2)\ \text{and}$$

$$\chi_i = \frac{H_i - F_i\ G_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}}$$

where $\tilde{F}_i = F_i - 1/2, \tilde{G}_i = G_i - 1/2\ \forall\ i \epsilon 1 : n$ and

$$H_i = \frac{1}{n-1}\#\{j \neq i : X_j \leq X_i, Y_j \leq Y_i\} = \frac{n\,W_i - 1}{n-1},$$

$$F_i = \frac{1}{n-1}\#\{j \neq i : X_j \leq X_i\}\ \text{and}\ G_i = \frac{1}{n-1}\#\{j \neq i : Y_j \leq Y_i\}$$

In the above, W_i is the bivariate probability integral transform (BIPIT) variate given by

$$W_i = \frac{1}{n}\#\{j : X_j \leq X_i, Y_j \leq Y_i\} = C_n\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right)$$

The plot may also include control limits at ordinates $\pm c_p/\sqrt{n}$. A scatter of the Chi-Plot predominantly within these control limits indicates independence among variables and vice versa. Based on a simulation study, Fisher & Switzer (2001) provided values of $c_p = 1.54$, 1.78, and 2.18 corresponding to p-values of 0.90, 0.95, and 0.99, respectively. When the scatter is largely on the upper side of control limits then it indicates a positive dependence, whereas when it is in the lower side of the limits it indicates a negative dependence. K-Plots, on the other hand, are scatter plots of observed order statistics and the expected value of the corresponding order statistics of the bivariate probability integral transformed (BIPIT) variable $W = H(X, Y) = C(U, V)$ of the same size, under the null hypothesis of independence among its components X and Y or U and V. Formally, Genest & Boies (2003) suggested plotting the pairs $(W_{1:n}, H_{(i)})$ for $i \epsilon 1{:}n$, where

$$H_{(1)} < H_{(2)} < \ldots < H_{(n)}$$

are the order statistics associated with H_i as defined for the Chi-plot above. $W_{1:n}$ is the expected value of the ith statistic of the variable W from a random sample of size n under the assumption of independence between X and Y or U and V. The diagonal line indicates independence, whereas the curve given by $K_0(w) = w - 1\ \log(w)$ corresponds to a perfect positive dependence. In the case of perfect negative dependence all the points would lie on the x-axis. Reference may be made to Genest & Favre (2007) for further details and properties of these plots. As illustrated by Abberger (2005), Chi-Plots can also help establish the significance of tail dependence qualitatively. Although, quantitative estimates of tail dependence are not obtained in this study, such estimators are available in literature and have been employed by Schmidt & Stadtmüller (2006), Poulin *et al.* (2007), Serinaldi (2008, 2009), and Villarini *et al.* (2008). On the basis of such an assessment of the dependence level and the tail dependence features, copulas offering such a range of

Table 2 | Copula test space based on admissible dependence range, and tail dependence characteristics. The shaded cells imply admissibility of corresponding dependence range

	Admissible Dependence Range in terms of Kendall's Tau													**Tail Dependence Coefficients**	
Copula Family/ Copula	**[−1, −0.6109)**	**[−0.6109, −0.5649)**	**[−0.5649, −0.4674)**	**[−0.4674, −0.3613)**	**[−0.3613, −0.3333)**	**[−0.3333, −0.2222)**	**[−0.2222, −0.1817)**	**[−0.1817,0)**	**0**	**[0,0.2222]**	**(0.2222, 0.3333)**	**0.3333**	**(0.3333,1]**	λ_L	λ_U
Archimedean															
AMH								▨	▨	▨	▨	▨		0	0
Clayton	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	$2^{-1/\theta}$	0
Frank	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	0	0
GG	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	0	$2-2^{1/\theta}$
Joe									▨	▨	▨	▨	▨	0	$2-2^{1/\theta}$
Extreme Value															
GH*									▨	▨	▨	▨	▨	0	$2-2^{1/\theta}$
Galambos									▨	▨	▨	▨	▨	0	$2^{-1/\theta}$
Meta-elliptical															
Normal	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	0	0
Other															
FGM							▨	▨	▨	▨				0	0
Plackett	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	0	0

*GH copula is, both, an Archimedean and an extreme value copula.

dependence and tail dependence characteristics can be short-listed. The second stage can then proceed in order to complete the copula selection process.

Goodness-of-fit tests

Generally, there are more than one feasible copula structures that may constitute the copula test space. In the second stage, parameters of all the short-listed copula families are estimated using one or more parameter estimation methods mentioned above. It is imperative to now ascertain the adequacy of the hypothesized copulas. This can be normally accomplished in three ways: (a) graphical methods, (b) error statistics, and (c) formal goodness of fit statistics.

A number of graphical approaches can be employed that utilize different features for making comparisons. Graphical comparison of the superimposed scatter plots of observed and simulated data is an intuitive way of qualitatively assessing the suitability of the hypothesized copulas. The data generation method, employing conditional distribution, as outlined by Nelsen (2006) can be employed for simulating a fairly large sample and its scatter plot together with that of the observed data provides a valuable comparative picture. It is important to note here that both very small and very large generated sample sizes can lead to misleading comparisons. Whereas too small sample sizes may not be adequate in reflecting the true nature of the distribution, too large sample sizes can result in very dense plotting in some portions, thereby obscuring the actual relative frequency of occurrence in those areas. This method, however, is better suited for bivariate cases only as similar comparisons in higher dimensions become difficult. Secondly, comparison of the ordered empirical probabilities with corresponding computed probabilities can be made, revealing the extent to which copula surface fits the scaled ranks of observed data. The other two graphical options are related to the K-plots mentioned above. In one option, the empirical and theoretical probability distributions, $K_n(w)$ and $K_{\theta_n}(w)$, of the BIPIT variate $W = C(U, V)$ can be compared, their closeness supporting non-rejection of the hypothesized copula. The second option is much like a Q-Q-plot: a scatter plot between the observed order statistics $W_{(1)} \leq W_{(2)} \leq \ldots \leq W_{(n)}$ of W and the corresponding expected order statistics $W_{1:n}$ based on the hypothesized copula. Again, conformation to the line from origin and having unit slope would suggest non-rejection of the hypothesized copula. This plot is also referred to as the "generalized K-plot".

A quantitative assessment of the performance of various copula families can be made by comparing maximized log-likelihood or Akaike information criterion (AIC) values. Other error statistics, such as the root mean square error (RMSE), mean absolute error (ME-A-ERR), mean error (MN-ERR) and maximum absolute error (MX-A-ERR), reflect other important characteristics of the comparison between empirical and computed probabilities.

Genest *et al.* (2009) presented a review of the available analytical goodness-of-fit tests for copulas, including those proposed by Wang & Wells (2000), Fermanian (2005), and Genest *et al.* (2006), among others, and recommended a few Cramer–von Mises type of test statistics based on Rosenblatt's transformation. The validity of parametric bootstrap procedure proposed by Genest *et al.* (2006) and for the empirical copula-based test statistics proposed by Fermanian (2005) has since been formally established by Genest & Remillard (2008). In this study, three goodness-of-fit test statistics, proposed by Fermanian (2005) and Genest *et al.* (2006), have been employed to formally test the adequacy of the hypothesized copulas. The first one is the Cramer–von Mises type of statistic proposed by Fermanian (2005). This is based on the comparison of empirical and parametric copula probabilities given by the process $\sqrt{n}(C_n - C_{\theta_n})$ and can be obtained as

$$\begin{aligned}\mathcal{CM}_n &= n\sum_{i=1}^{n}\left[C_n\left(\frac{R_i}{n+1},\frac{S_i}{n+1}\right) - C_{\theta_n}\left(\frac{R_i}{n+1},\frac{S_i}{n+1}\right)\right]^2 \\ &= n\sum_{i=1}^{n}\left[W_i - C_{\theta_n}\left(\frac{R_i}{n+1},\frac{S_i}{n+1}\right)\right]^2\end{aligned} \quad (15)$$

The other two Cramer–von Mises and Kolmogorov-type statistics are given by Genest *et al.* (2006) as variants of those proposed by Wang & Wells (2000). Providing an objective comparison of the empirical and theoretical probabilities of the BIPIT variate W, these are based on the process $\mathrm{K}_n(w) = \sqrt{n}\{K_n(w) - K_{\theta_n}(w)\}$ and can be obtained as

$$\begin{aligned}\mathcal{S}_n &= \int_0^1 |\mathrm{K}_n(w)|^2 k_{\theta_n}(w)\, dw \\ &= \frac{n}{3} + n\sum_{j=1}^{n-1} K_n^2\left(\frac{j}{n}\right)\left[K_{\theta_n}\left(\frac{j+1}{n}\right) - K_{\theta_n}\left(\frac{j}{n}\right)\right] \\ &\quad - n\sum_{j=1}^{n-1} K_n\left(\frac{j}{n}\right)\left[K_{\theta_n}^2\left(\frac{j+1}{n}\right) - K_{\theta_n}^2\left(\frac{j}{n}\right)\right] \qquad (16)\end{aligned}$$

and

$$\begin{aligned}\mathcal{T}_n &= \sup_{0\le w\le 1} |\mathrm{K}_n(w)| \\ &= \sqrt{n} \max_{i=0,1;0\le j\le n-1}\left\{\left|K_n\left(\frac{j}{n}\right) - K_{\theta_n}\left(\frac{j+1}{n}\right)\right|\right\} \qquad (17)\end{aligned}$$

The copula selection process can now be summarized by the following steps:

- Get an initial idea of the dependence level from the scatter plot of scaled ranks.
- Quantify the strength of dependence by computing non-parametric dependence measures, such as Spearman's rho and Kendall's tau.
- Reaffirm the significance of dependence using Chi-Plot and/or K-Plot.
- Observe the significance of tail dependence by using Chi-Plot and/or by obtaining quantitative estimates and compare this with the knowledge-base about the process under consideration.
- Pre-select one or more copula types admitting the dependence level and the tail dependence under consideration.
- Estimate copula parameters by one or more methods.
- Assess the adequacy of hypothesized copulas on the basis of graphical diagnostics plots.
- Assess the adequacy of hypothesized copulas on the basis of one or more analytic goodness-of-fit test statistics and error statistics.
- Identify suitable copula models on the basis of the above assessment.

APPLICATION

Hydraulic infrastructure along a river, such as dams, levees and bridges, is designed in order to safely carry high flows. For this, simultaneous consideration of the occurrence of flood peak flow and the corresponding volume and/or duration is important. This is typically done empirically by routing critical observed or synthetic hydrographs and determining spillway capacity and/or safer crest level on the basis of the expected maximum water levels. The multivariate statistical frequency analysis of such processes can provide a probabilistic assessment of the occurrence of critical events, enabling multivariate risk-based designs. Also, many situations, such as design of retention basins, extent of flooding due to levee breach and consequent property damage, serviceability of a highway bridge across a river, warrant a simultaneous consideration of multiple flood variables. Such considerations are important for agencies dealing with disaster management and for insurance companies, in order to be aware of the actual risk due to flooding and associated damage at a regional scale. An application in that direction is presented here, with an objective of finding suitable copula types that can adequately represent a data set of flood peak flow and volume.

Dataset

The annual peak and average daily flows of Greenbrier River at Alderson station (USGS Station # 03183500) in West Virginia, the United States, are obtained from the USGS website and considered for this application. The Greenbrier River is a tributary of the New River in the southeastern part of the state and is approximately 165 mi (265 km) long. Through the New, Kanawha and Ohio rivers, it is part of the Mississippi River watershed. A river gauging station is located at Alderson at Latitude 37°43'27" and Longitude 80°38'30", commanding a drainage area and contributing area of 1,364 square miles. The datum of the gauge is at 1,529.42 feet above sea level. A length of 110 years of data, from 1896 to 2005, has been considered for this analysis. Preparation of annual maximum flood data is an important first step as the selection of maximal events for bivariate data becomes slightly ambiguous. In the study of extreme rainfall events, Kao & Govindaraju (2007*b*) recommended selection of AMP events, arguing that such events would represent a wider range of durations as compared to AMI or AMV events. It may, however, be noted that as most hydrologic designs have to be safer against critical events rather than most probable events, such AMP-based events may result in the underestimation of risk. Extreme flood events with respect to the safety of a drainage system are invariably primarily associated with peak flows that cause overtopping of crests of dams or levees, or inundation of floodplains. Any high volume or long duration of flow by themselves may not be any cause of concern when flows are less than the design capacity of the system. Detrimental effects of high volume and/or duration of flow are also important but they typically come into play only when there is a primary failure or likelihood of such a failure due to higher peak flows. Annual maximum flood events have therefore been considered in this study on the basis of annual peak flows and associated volumes that have been obtained from the record of average daily flows. Time series of these two data sets, Q in 10^3 cusec and V in 10^3 cusec-days, are given in Figure 1. The scatter plots of this bivariate data and of their scaled ranks, along with the respective histograms, are shown in Figure 2. As scaled ranks are empirical probabilities, they are uniformly distributed between 0 and 1, as seen in Figure 2.

Potential marginal distributions

Several candidate distributions, such as two- and three-parameter lognormal (LN2 and LN3), two-parameter gamma (G2), Pearson type III (P3), log-Pearson type III (LP3), largest extreme value (LEV), two- and three-parameter Weibull (W2 and W3), are considered for fitting annual peak flow and volume on a univariate basis. On the basis of the Kolmogorov–Smirnov, Anderson Darling, and Chi-Squared fit statistics and the overall fit of the Q–Q plots, Pearson Type III and three-parameter Weibull distributions were selected as marginal distributions for annual flood peak and associated volume, respectively. The overlay of probability density curves of these distributions and the corresponding histograms is shown in Figure 2 and the corresponding Q–Q plots, along with 95% confidence intervals, are given in Figure 3. The density functions for P3 and W3, $f_X(x)$ and $g_Y(y)$ for flood peak flow $X = Q$ and volume $Y = V$, respectively, are given as

$$f_X(x) = \frac{1}{|\alpha_X|}\frac{1}{\Gamma(\beta_X)}\left(\frac{x-\gamma_X}{\alpha_X}\right)^{\beta_X-1}\exp\left(-\frac{x-\gamma_X}{\alpha_X}\right) \tag{18}$$

where $-\infty < \gamma_X < \infty$, $-\infty < \alpha_X < \infty$, and $\beta_X > 0$ are location, scale and shape parameters, respectively, and $x \geq \gamma_X \ \forall \ \alpha_X > 0$ *and* $x \leq \gamma_x \ \forall \ \alpha_x < 0$;

$$g_Y(y) = \frac{\beta_Y}{\alpha_Y}\left(\frac{y-\gamma_Y}{\alpha_Y}\right)^{\beta_Y-1}\exp\left[-\left(\frac{y-\gamma_Y}{\alpha_Y}\right)^{\beta_Y}\right] \tag{19}$$

where $-\infty < \gamma_Y < \infty$, and α_Y, $\beta_Y > 0$ *are location, scale and shape parameters, respectively, and* $y \geq \gamma_Y$.

The maximum likelihood parameter estimates for these two marginals are obtained as $\hat{\gamma}_X = 4.601$, $\hat{\alpha}_X = 6.197$, $\hat{\beta}_X = 5.101$, and $\hat{\gamma}_Y = 28.361$, $\hat{\alpha}_Y = 122.185$, $\hat{\beta}_Y = 1.326$. The corresponding standard errors are $Se_{\hat{\gamma}_X} = 4.332$, $Se_{\hat{\alpha}_X} = 1.365$ $Se_{\hat{\beta}_X} = 1.715$, and $Se_{\hat{\gamma}_Y} = 1.907$, $Se_{\hat{\alpha}_Y} = 6.093$, $Se_{\hat{\beta}_Y} = 0.107$, respectively.

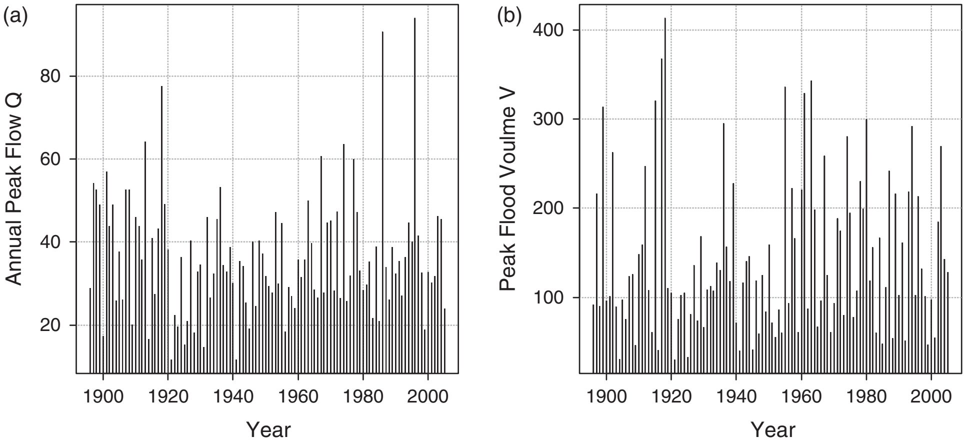

Figure 1 | Time series of annual peak flows (Q, in 10^3 cusec) and corresponding flood volumes (V, in 10^3 cusec-days) at Alderson gauging station on Greenbrier River.

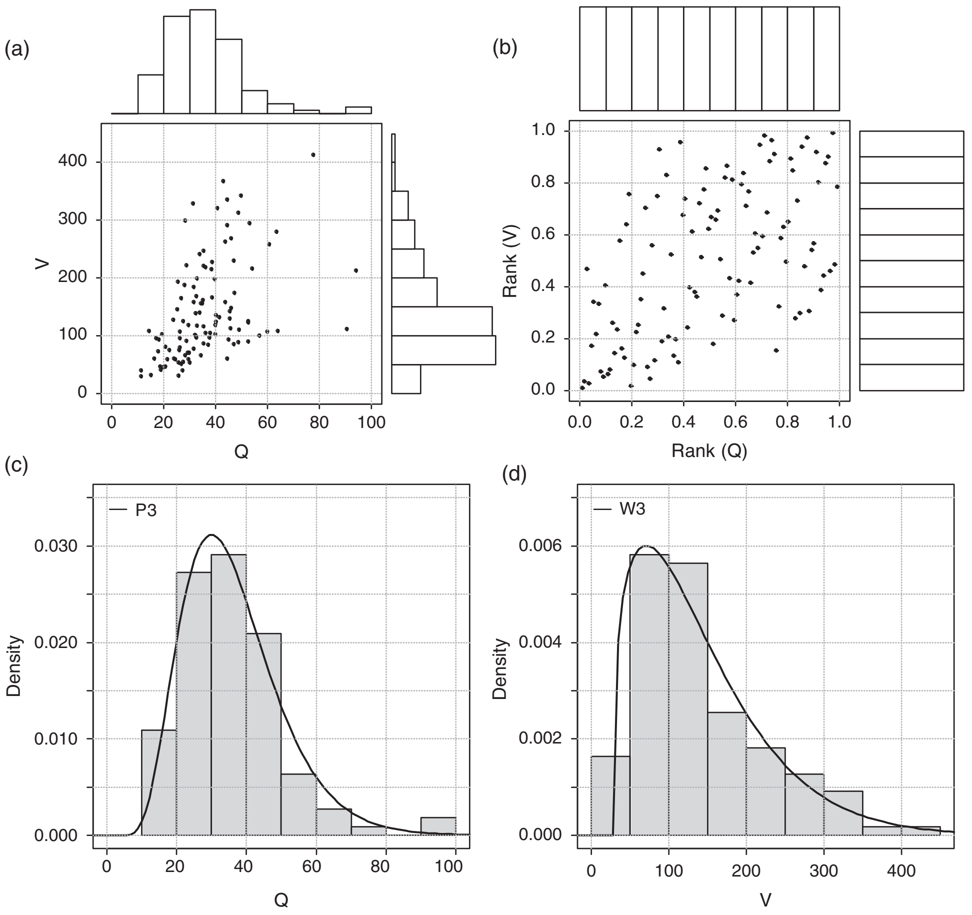

Figure 2 | Characteristics of observed bivariate annual peak flow (Q in 10^3 cusec) and volume (V in 10^3 cusec-day) data of Greenbrier River at Alderson gauging station: (a) scatter plot and histograms in original domain; (b) scatter plot of scaled ranks and corresponding uniform histograms; (c), (d) histograms along with P3 and W3 probability density curves, respectively.

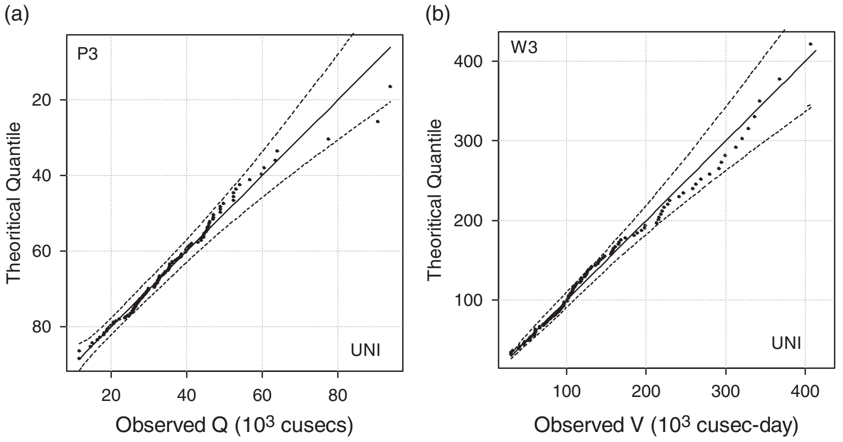

Figure 3 | Q-Q plots for peak flow (Q) and volume (V) data fitted with Pearson type III (P3) and three-parameter Weibull (W3) distributions, respectively.

Dependence structure and copula test space

The scatter plots in Figure 2(a) and (b) indicate a positive association between annual peak flow and volume data. The sample estimates of the Pearson correlation coefficient, Kendall's tau and Spearman's rho, 0.466, 0.391, and 0.557 with corresponding *p*-values 2.9e-07, 1.78e-09, and 3.0e-10, respectively, corroborate this assertion. A significant positive dependence is also indicated by both Chi- and K-plots given in Figure 4. Considering data exclusively from the lower-left and upper-right quadrants, as suggested by Abberger (2005), the Chi-plots in Figure 5 exhibit lower and upper tail dependence features. It is apparent from Figure 5(a) that there is evidence of lower tail dependence as a few points close to $\lambda_i = 1$ show significance. More importantly, Figure 5(b) indicates upper tail independence, as points in the end zone are within control limits with a *p*-value of 0.95. Based on sample Kendall's tau value of 0.391 and the features of lower and upper tail dependence, two Archimedean copulas, Clayton and Frank, and two extreme value copulas, the Galambos and GH, are selected. In fact, GH is an Archimedean copula also. In order to appreciate the problems that arise due to misspecification, two more copulas, AMH and FGM, are also short-listed, noting that the sample dependence is beyond the admissible ranges of these copulas. Although more copulas could have been considered in the initial screening stage, only these six are included in the copula test space, in order to keep the selection process shorter.

Estimation of dependence parameter

The dependence parameters for the six copula families under consideration are estimated by (a) moment-like method of inversion of dependence measures and (b) maximum pseudo-likelihood method. Based on the relationship between Kendall's tau τ and dependence parameter θ as given in Equation (10) and utilizing their closed forms given in Table 1, wherever available, the dependence parameters are estimated for each of the six copulas. These point estimates along with standard errors and the interval estimates, corresponding to a coverage probability of 0.95, are given in Table 3. The estimates for AMH and FGM copulas are not obtainable for this data set, as the values of τ are beyond the admissible limits. The AMH copula requires τ to be between -0.1817 and 0.3333, and FGM copula to be in the range of $-2/9$ to $2/9$ (-0.222 to 0.222). This illustrates limitations of these copula structures, similar to that faced by some of the conventional functional distributions.

The dependence parameter, based on the maximum pseudo-likelihood (MPL) estimation method, is computed by employing the log-likelihood function given in Equation (12). The maximized log-likelihood value ($\mathrm{LL}_{\max}$), and the point and interval estimates of the dependence parameter for the six copula families are given in Table 3. It may be seen from these results that the standard error of the dependence parameter estimates from this method is much lower and thus may be preferable. It is noted that for the AMH and FGM copula types the sub-optimal values of dependence

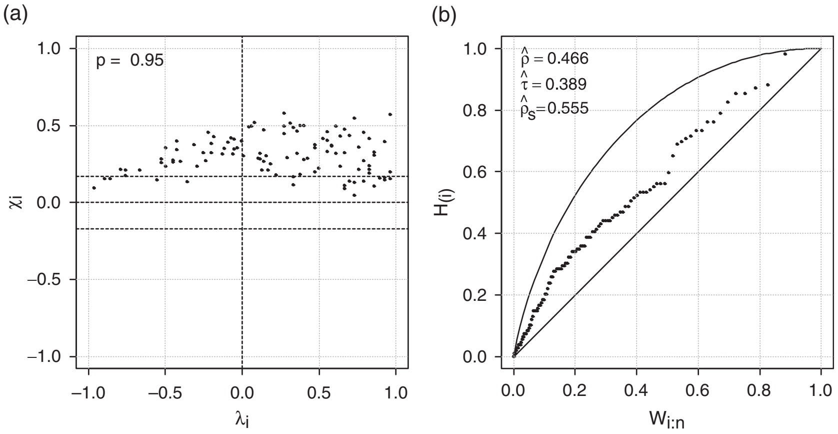

Figure 4 | Characterization of dependence between annual flood peak flow and volume, using (a) Chi-plot and (b) K-plot.

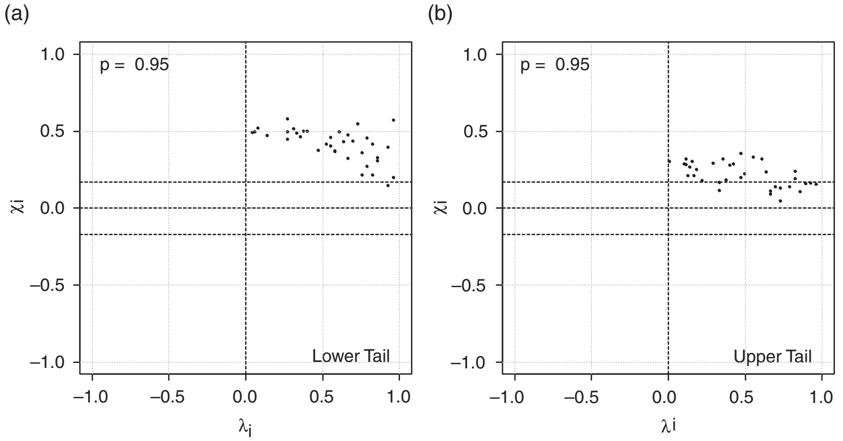

Figure 5 | Characterization of (a) lower and (b) upper tail dependence between annual flood peak flow and volume, using Chi-plots.

parameters lie at the end of the parameter space and correspond to much lower values of τ than that obtained from the sample data set.

Assessment of copula fitting

The relative suitability of plausible copula families is ascertained in multiple ways, as outlined in "Methods" section, by employing (a) graphical methods, (b) error statistics, and (c) formal goodness of fit statistics.

Graphical goodness-of-fit tests

First, observed data are compared with a set of large number of generated random samples. For this study, a set of random samples of size 500 is generated for each of the six copula families under consideration, utilizing MPL method-based parameters and employing the approach outlined in Nelsen (2006). This comparison of observed data and generated random samples is shown in Figure 6. It may be seen from these plots that the general nature of spread of observed data matches with that of random samples. However, a closer look reveals that the Galambos and GH copulas exhibit upper tail dependence that does not have a similar representation in the observed data. Also, very high flows with moderate volumes in the observed data are not represented by the simulated set. The simulated sets of the other four copulas provide adequate representation, except for the differences around the lower tail where the AMH and Clayton copulas seem to be performing better.

Secondly, comparison of ordered empirical probabilities with corresponding computed probabilities, as depicted in Figure 7, reveals the extent to which the copula surface fits

Table 3 | Point and interval dependence parameter estimates based on the inversion of dependence measure (MOM) and maximum pseudo-likelihood (MPL) methods. Interval estimates correspond to a coverage probability of 0.95

	MOM-based Interval Estimate				MPL-based Interval Estimate				
Copula Family	**Theta ($\hat{\theta}_n$)**	**Lower C.L.**	**Upper C.L.**	**Standard Error**	**LL_{max}**	**Theta ($\hat{\theta}$)**	**Lower C.L.**	**Upper C.L.**	**Standard Error**
AMH	-	-	-	-	24.490	0.995	0.900	1.090	0.049
Clayton	1.283	0.722	1.844	0.286	**25.451**	1.220	1.031	1.409	0.097
FGM	-	-	-	-	14.811	0.995	0.823	1.167	0.088
Frank	4.036	2.631	5.441	0.717	19.646	3.970	3.807	4.133	0.083
Galambos	0.917	0.631	1.202	0.146	16.367	0.800	0.665	0.935	0.069
GH	1.642	1.361	1.922	0.143	16.118	1.529	1.388	1.670	0.072

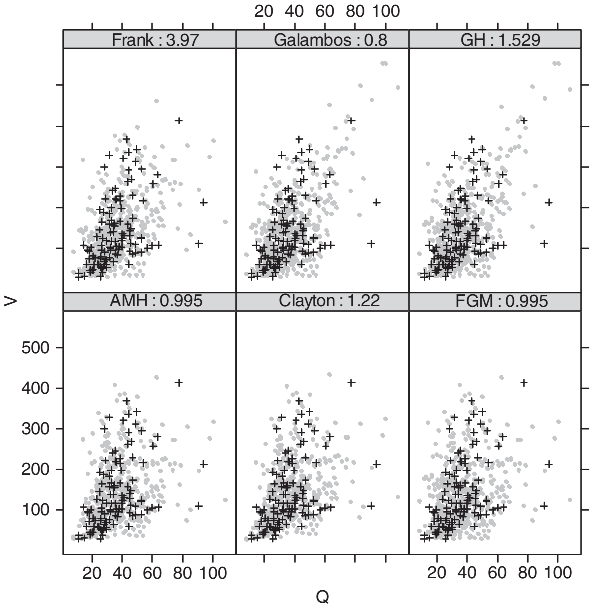

Figure 6 | Comparison of observed data with sets of 500 generated random samples based on dependence parameters obtained by the MPL method. Solid circles in gray are random samples, whereas "+" symbols represent observed data. The numbers after the copula names are the MPL method based dependence parameter estimates.

the scaled ranks of observed data. The matching by the AMH, Clayton, and Frank copulas is better than the other three copulas, with differences being minimal for the Clayton copula.

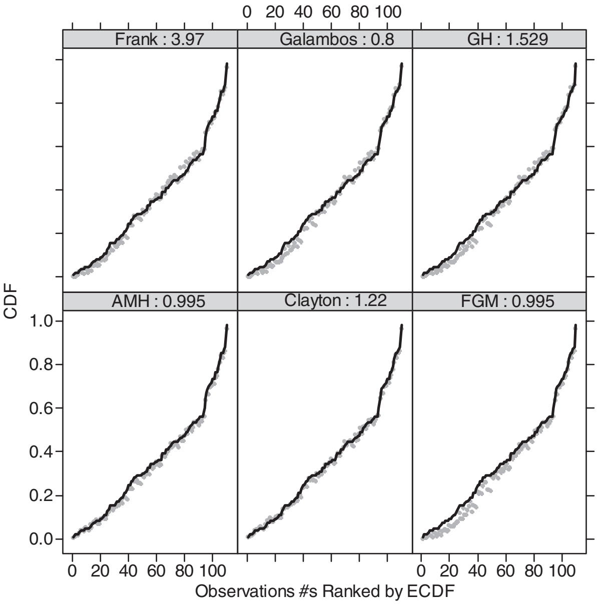

Figure 7 | Comparison of empirical and MPL method based computed probabilities. Solid circles in gray are empirical probabilities and solid line is the computed probability. The numbers after the copula names are the MPL method based dependence parameter estimates.

Comparison of empirical and computed probability distributions of the BIPIT variate, $K_n(w)$ and $K_{\theta_n}(w)$, in Figure 8 shows that matching of the two distributions is best for the Clayton copula, followed by the AMH and Frank copulas. Lastly, the generalized K-plot in Figure 9 provides a comparison of observed and expected order statistics of the BIPIT variate and it may be seen that the matching is again best for the Clayton copula, followed again by the AMH and Frank copulas. The graphical fit for the Clayton copula is clearly the best in all the four graphical methods, among the six copulas considered. The relative superiority can however be further established by looking at the formal test results.

Various error statistics of fit

A quantitative assessment of the performance of various copula families is made by comparing maximized log-likelihood values. As all the copula models considered in this study involve fitting a single parameter, comparing the maximized log-likelihood value or AIC value would be equivalent. The maximized log-likelihood value for the Clayton copula, as given in Table 3, is the maximum among the six copula models considered and supports the conclusions

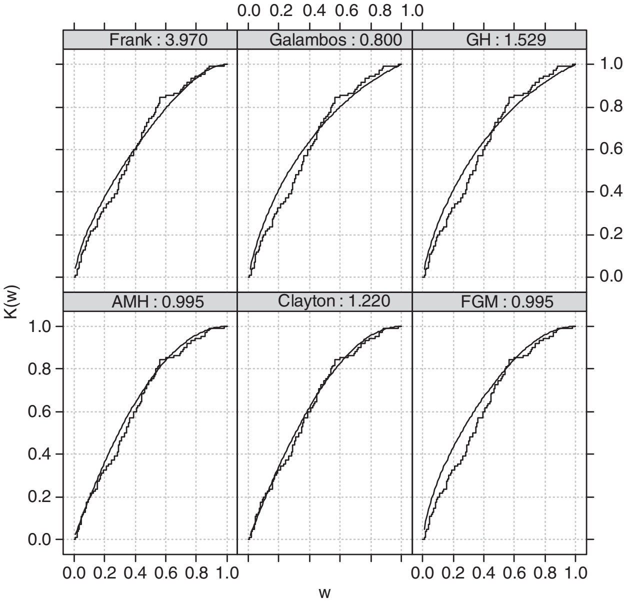

Figure 8 | Comparison of empirical, $K_n(W)$, and theoretical, $K_{\theta_n}(W)$, probability distribution functions of the bivariate integral transform variable $W = (C, V)$ for various copulas. The step functions are empirical distributions and the curves are the MPL method based theoretical distributions. The numbers after the copula names are the MPL method based dependence parameter estimates.

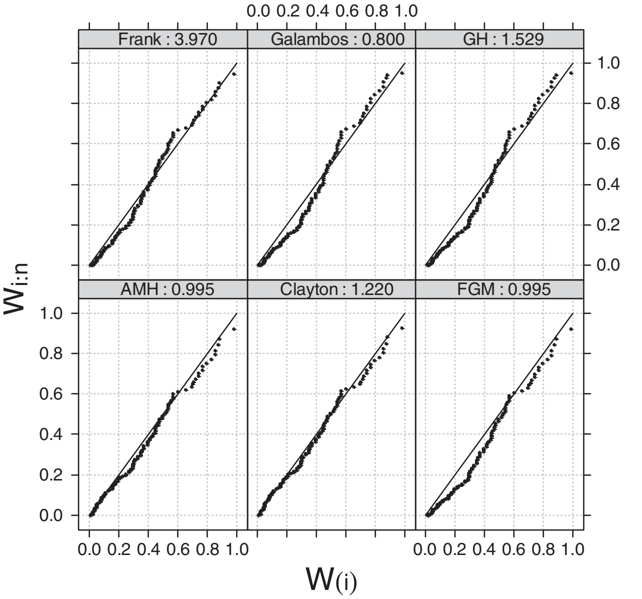

Figure 9 | Comparison, in terms of generalized K-plots, between observed $W_{(i)}$ and corresponding expected W_{in} order statistics of the MPL method based bivariate probability integral transform variable $W = C(U, V)$ for various copulas. The line through origin with unit slope indicates equivalence between entities of the two axes. The numbers after the copula names are the MPL method based dependence parameter estimates.

based on the above graphical goodness-of-fit tests. An account of the comparison of other error statistics, RMSE, ME-A-ERR, MN-ERR, and MX-A-ERR, is given in Table 4 and it may be seen that the Clayton copula yields lowest errors and FGM copula yields largest errors in all these error categories. The reasoning for the poor performance of the FGM copula is obvious as this copula admits τ up to 0.222 only, whereas the sample estimate is much higher at 0.391. Similarly, a slightly better performance of the AMH copula than FGM may be attributed to the fact that although this copula also does not cover the desired range of τ, the shortfall from the largest permissible value of 0.333 is not severe. Another important observation with regard to the Frank, Galambos and GH copulas is that although they performed slightly inferiorly than AMH with respect to most error statistics, they show a far less mean error statistic. Since a lower mean error indicates a better balance between positive and negative deviations, this may be a desirable feature and that way the Frank, Galambos, and GH copulas may be considered to have given a better fit than AMH. It may also be seen from the table that the errors from both the methods, MOM-based and MPL, are of similar order of magnitude.

Analytical goodness-of-fit tests

The formal goodness-of-fit tests are carried out by evaluating the Cramer-von Mises type statistics, $\mathcal{CM}_n$, $\mathcal{S}_n$, and $\mathcal{T}_n$ as given in Equations (15)–(17), for four copulas: Clayton, Frank, Galambos, and GH. For this, a parametric bootstrap procedure is employed for simulating random samples of sizes 100, 1000, 10,000 and 100,000. The values of these three statistics, their p-values and the critical values at the 5% significance level are computed. Tables 5 and 6 list these results for both methods of parameter estimation considered in this study. Three simulations are run for each combination of sample size, copula model and method of estimation, except for the Galambos copula for which only one run was made for the sample sizes of up to 1000 due to large computational requirements. An important observation in all cases is that the values of these statistics stabilize sufficiently, even when the sample size is 10,000 and this is also in agreement with the observations made by Genest & Favre

Table 4 | Various error statistics for the six copulas under consideration with respect to inversion of dependence measures (MOM) method and maximum pseudo-likelihood (MPL) method

	MOM-based Copula Model				MPL-based Copula Model			
Copula Family	RMSE	ME-A-ERR	MN-ERR	MX-A-ERR	RMSE	ME-A-ERR	MN-ERR	MX-A-ERR
AMH	-	-	-	-	0.0184	0.0153	0.0143	0.0470
Clayton	**0.0132**	**0.0109**	**0.0066**	**0.0338**	**0.0139**	**0.0115**	**0.0081**	**0.0364**
FGM	-	-	-	-	0.0340	0.0289	0.0287	0.0820
Frank	0.0187	0.0159	**0.0066**	0.0469	0.0188	0.0160	**0.0072**	0.0479
Galambos	0.0224	0.0183	**0.0067**	0.0593	0.0246	0.0196	0.0123	0.0678
GH	0.0224	0.0183	**0.0065**	0.0587	0.0244	0.0196	0.0121	0.0670

Table 5 | Goodness-of-fit statistics for the inversion of dependence measure (MOM) method. S^* implies the critical value of test statistic at a significance level of 5% and P-val indicates the *p*-values of the observed test statistic

			Critical Test Statistic S^* and P-value for a run of N =							
			100		1000		10,000		100,000	
Statistic	Copula	Observed Statistic	S^*	P-val	S^*	P-val	S^*	P-val	S^*	P-val
$\mathcal{CM}_n$	Clayton	2.115	3.970	0.490	3.621	0.526	3.787	0.517	3.799	**0.523**
			3.713	0.520	3.861	0.526	3.830	0.527	3.798	**0.521**
			3.987	0.610	3.809	0.502	3.772	0.528	3.816	**0.522**
	Frank	4.228	4.333	0.060	3.911	0.028	3.835	0.029	3.832	0.029
			3.704	0.040	3.801	0.024	3.858	0.029	3.838	0.029
			3.686	0.040	3.712	0.022	3.816	0.029	3.833	0.029
	Galambos	6.095	3.521	0.000	3.710	0.002	-	-	-	-
	GH	6.051	3.734	0.010	3.578	0.003	3.755	0.002	3.737	0.002
			3.571	0.000	3.559	0.005	3.739	0.002	3.738	0.002
			3.512	0.010	3.689	0.002	3.698	0.003	3.723	0.002
$\mathcal{S}_n$	Clayton	0.062	0.130	0.460	0.146	0.474	0.145	0.518	0.146	**0.516**
			0.147	0.500	0.138	0.522	0.144	0.515	0.146	**0.513**
			0.148	0.570	0.143	0.511	0.145	0.515	0.147	**0.517**
	Frank	0.211	0.135	0.000	0.129	0.005	0.132	0.004	0.132	0.004
			0.123	0.010	0.130	0.004	0.130	0.004	0.131	0.004
			0.128	0.010	0.127	0.002	0.132	0.004	0.132	0.004
	Galambos	0.331	0.166	0.000	0.157	0.000	-	-	-	-
	GH	0.315	0.139	0.000	0.128	0.001	0.134	0.000	0.134	0.000
			0.133	0.000	0.131	0.001	0.134	0.000	0.133	0.000
			0.114	0.000	0.132	0.000	0.135	0.000	0.133	0.000
$\mathcal{T}_n$	Clayton	0.706	0.963	0.360	0.970	0.394	0.977	0.412	0.967	**0.414**
			0.932	0.440	0.971	0.425	0.966	0.419	0.968	**0.414**
			0.957	0.440	0.964	0.403	0.970	0.411	0.971	**0.415**
	Frank	0.900	0.864	0.050	0.905	0.053	0.921	0.062	0.916	0.060
			0.920	0.070	0.915	0.058	0.915	0.057	0.917	0.060
			0.917	0.060	0.919	0.063	0.907	0.055	0.917	0.060
	Galambos	1.141	1.062	0.030	1.021	0.005	-	-	-	-
	GH	1.136	0.918	0.020	0.916	0.006	0.943	0.006	0.942	0.006
			0.964	0.010	0.951	0.007	0.944	0.006	0.941	0.006
			0.899	0.000	0.961	0.005	0.936	0.006	0.939	0.005

(2007). Simulation runs with sample sizes of 100,000 were made only as a confirmatory step and the results indicate that it can confidently be avoided for actual applications. The results for all the three statistics from both the methods do not provide any evidence for rejecting the hypothesis of the Clayton copula as a valid model for the peak flow and volume data under consideration. Significantly higher *p*-values leading to this inference are highlighted in Tables 5 and 6. At the same time, there is an overwhelming basis for the rejection of hypothesis of the Frank, Galambos, and GH copulas being viable models at the 5% significance level. Barring a feeble support for the Frank copula in terms of $\mathcal{T}_n$ statistic, this inadequacy of support for these three copulas is based on the results of all the three statistics and for both the methods.

Table 6 | Goodness-of-fit statistics for maximum pseudo-likelihood (MPL) method. S* implies the critical value of the test statistic at a significance level of 5% and P-val indicates the *p*-values of the observed test statistic

			Critical Test Statistic S* and P-value for a run of N =							
			100		**1000**		**10,000**		**100,000**	
Statistic	**Copula**	**Observed Statistic**	**S***	**P-val**	**S***	**P-val**	**S***	**P-val**	**S***	**P-val**
$\mathcal{CM}_n$	Clayton	2.356	4.109	0.470	4.807	0.445	4.843	0.432	4.867	**0.433**
			4.311	0.500	4.900	0.430	4.731	0.425	4.866	**0.436**
			4.316	0.510	5.090	0.461	4.908	0.438	4.843	**0.433**
	Frank	4.277	3.725	0.010	3.829	0.023	3.808	0.027	3.825	0.027
			3.745	0.040	3.870	0.031	3.751	0.024	3.845	0.028
			3.540	0.030	3.787	0.027	3.833	0.026	3.810	0.027
	Galambos	7.309	3.550	0.000	4.180	0.004	-	-	-	-
	GH	7.206	4.053	0.000	4.125	0.000	4.023	0.001	3.953	0.001
			4.354	0.000	4.072	0.000	3.962	0.001	3.967	0.001
			4.044	0.000	3.970	0.002	3.966	0.001	3.945	0.001
$\mathcal{S}_n$	Clayton	0.072	0.136	0.440	0.153	0.414	0.151	0.402	0.152	**0.402**
			0.146	0.550	0.152	0.399	0.148	0.398	0.151	**0.402**
			0.131	0.340	0.161	0.414	0.152	0.403	0.151	**0.400**
	Frank	0.216	0.104	0.000	0.133	0.006	0.134	0.003	0.131	0.003
			0.106	0.010	0.133	0.007	0.129	0.002	0.132	0.003
			0.126	0.000	0.133	0.001	0.129	0.003	0.131	0.003
	Galambos	0.425	0.182	0.000	0.177	0.002	-	-	-	-
	GH	0.406	0.161	0.000	0.157	0.000	0.155	0.000	0.154	0.000
			0.170	0.000	0.155	0.000	0.153	0.000	0.155	0.000
			0.158	0.000	0.153	0.000	0.155	0.000	0.153	0.000
$\mathcal{T}_n$	Clayton	0.762	0.945	0.270	0.982	0.272	0.973	0.283	0.969	**0.282**
			0.991	0.340	0.960	0.291	0.958	0.271	0.966	**0.281**
			0.919	0.210	0.987	0.273	0.969	0.284	0.966	**0.278**
	Frank	0.906	0.870	0.040	0.923	0.059	0.916	0.055	0.912	0.053
			0.858	0.030	0.897	0.045	0.907	0.051	0.913	0.054
			0.889	0.020	0.898	0.048	0.911	0.053	0.911	0.053
	Galambos	1.310	1.066	0.000	1.063	0.002	-	-	-	-
	GH	1.304	0.961	0.000	0.983	0.000	0.982	0.002	0.981	0.001
			0.951	0.000	0.981	0.000	0.975	0.001	0.983	0.001
			0.998	0.000	1.000	0.001	0.980	0.001	0.981	0.001

Thus all the graphical and analytical goodness-of-fit test results indicate non-rejection of the Clayton copula as a suitable copula for the flood peak flow and volume data under consideration. At the same time, these results also provide sufficient basis for rejecting the Frank, Galambos and GH copulas as viable options.

Joint distribution

The MPL method-based Clayton copula is taken as the finalized copula model for the joint distribution of peak flow and volume data under consideration. The joint probabilities and densities in the domains of the original variables,

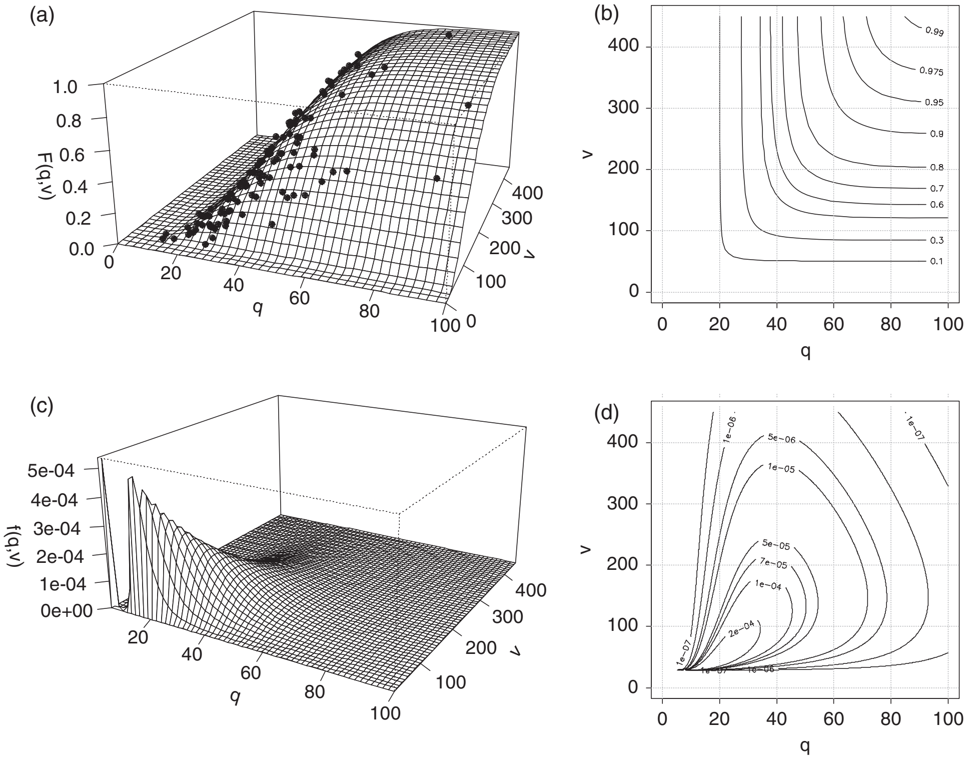

Figure 10 | (a), (b) Plots of joint probability functions, along with superimposed empirical probabilities in 3D plot; (c), (d) joint probability density functions.

both in 3D and contour forms, are given in Figure 10. Figure 10(a) illustrates the close match between computed and empirical probabilities.

CONCLUSIONS

The copula approach allows all the valuable experience gained thus far in univariate hydrological frequency analysis to be carried forward to multivariate analysis, without resorting to data transformation techniques or working with restrictive functional distributional forms. The use of copula-based multivariate distributions in the field of hydrological engineering is presently in the initial stages. This study adds a valuable experience and provides insight into the copula selection process for hydrological data involving peak flows and associated volumes. Several graphical and formal goodness-of-fit tests are employed in order to identify suitable copula structures for the data under consideration and the Clayton copula emerges as a clear choice. This assertion cannot however be generalized for all cases of peak flow and volume data on the basis of this one study. In view of such data not conforming to the definition of bivariate extreme value processes, there is, in general, a strong likelihood of the Clayton copula to be a strong contender. It is opined that more studies in this direction would be useful in ascertaining the suitability of the Clayton copula in general. The following specific conclusions can be drawn on the basis of this study:

- In the absence of any evidence from any of the graphical and formal goodness-of-fit test statistics against the hypothesis of the Clayton copula being a valid model, it is considered a suitable model for this particular dataset of peak flow and volume data. The Clayton copula also had the least values of all the error statistics considered for comparison.
- Various analytic tests provide an overwhelming evidence for rejection of the hypothesis of an extreme value copula, such as Galambos and GH, being a valid model. This is also explained on the basis that the 110 years of a fairly long record of peak flow and volume data under consideration does not show any evidence of a significant upper

tail dependence. Moreover, such peak flow and corresponding volume data do not conform to the definition of component-wise maximal order statistic that is required for a joint extreme value process. A misspecification of an extreme value copula as a valid copula would result in substantial overestimation in hydrologic designs pertaining to this particular dataset.

- The Frank copula could have been considered nearly as good as the Clayton copula on the basis of all the graphical tests, but such a hypothesis is rejected on the basis of formal goodness-of-fit tests, except for a feeble support it received with respect to the $\mathcal{T}_n$ statistic at 5% significance level. Furthermore, a significant lower tail dependence, evidenced by the data, makes yet another valid basis for such rejection.
- Acceptance of the Clayton copula as a suitable model for this flood peak flow and volume data is at odds with the results reported by De Michele *et al.* (2005), Zhang & Singh (2006), and Genest & Favre (2007), Poulin *et al.* (2007) and Karmakar & Simonovic (2009), wherein GH copula had been adjudged better than other copulas, including the Clayton. This may be due to the nature of this particular data set under consideration. However, it is recommended that this variation in inference be researched further by studying more cases.
- Formal test statistics, such as Cramer–von Mises statistic, obtained by the parametric bootstrap method prove effective in objectively discriminating between various copulas while the graphical tests do not make such a distinction obvious.
- There are many copula families, e.g., AMH and FGM, that are not comprehensive enough to cater to the range of dependence required for hydrological applications such as the one in this study. This is similar to the restrictions posed by some of the functional distributional forms. The graphical goodness-of-fit tests may sometimes not able to effectively discern the inadequacies of such models and hence should not constitute the sole basis for copula selection.

ACKNOWLEDGEMENTS

The authors are thankful for the constructive comments made and corrections suggested by two anonymous reviewers that substantially improved the quality of the paper. This work is based on first author's Master's project and Ph.D. dissertation research and the partial financial support received by him from the Graduate School of Louisiana State University to pursue the same is most sincerely acknowledged.

REFERENCES

Abberger, K. 2005 A simple graphical method to explore tail-dependence in stock-return pairs. *Appl. Fin. Econ.* **15**(1), 43–51.

Ashkar, F. & Rousselle, J. 1982 A multivariate statistical analysis of flood magnitude, duration and volume. In: *Statistical Analysis of Rainfall and Runoff*. (Singh, V. P. ed.), Water Resources Publications, Fort Collins, CO, pp. 651–669.

Choulakian, V., El-Jabi, N. & Moussi, J. 1990 On the distribution of flood volume in partial duration series analysis of flood phenomena. *Stoch. Hydrol. Hydraul.* **4**, 217–226.

Chowdhary, H. 2008 Discussion of "Gumbel–Hougaard copula for trivariate rainfall frequency analysis". *J. Hydrol. Engng.* **13**(10), 992–994.

Chowdhary, H. 2009 Discussion of "IDF curves using the Frank Archimedean copula". *J. Hydrol. Engng* **14**(1), 107–108.

Correia, F. N. 1987 Multivariate partial duration series in flood risk analysis. In: *Hydrologic Frequency Modelling*, Singh, V. P. (Ed.), Reidel, Dordrecht, The Netherlands, pp. 541–554.

De Michele, C. & Salvadori, G. 2003 A generalized Pareto intensity-duration model of storm rainfall exploiting 2-copulas. *J. Geophys. Res.* **108**(D2), 4067.

De Michele, C., Salvadori, G., Canossi, M., Petaccia, A. & Rosso, R. 2005 Bivariate statistical approach to check adequacy of dam spillway. *J. Hydrol. Engng* **10**(1), 50–57.

Dupuis, D. J. 2007 Using copulas in hydrology: Benefits, cautions, and issues. *J. Hydrol. Engng* **12**(4), 381–393.

Durante, F. & Salvadori, G. 2009 On the construction of multivariate extreme value models via copulas. *Environmetrics* **21**, 143–161.

Escalante-Sandoval, C. A. 1998 Multivariate extreme value distribution with mixed Gumbel marginals. *J. Am. Water Resour. Assoc.* **34**, 321–33.

Escalante-Sandoval, C. A. & Raynal-Villasenor, J. A. 1994 A trivariate extreme value distribution applied to flood frequency analysis. *J. Res. Natl Inst. Stand. Technol.* **99**, 369–375.

Favre, A.-C., El Adlouni, S., Perreault, L., Thiemonge, N. & Bobee, B. 2004 Multivariate hydrological frequency analysis using copulas. *Water Resour. Res.* **40**, 1–12.

Fermanian, J.-D. 2005 Goodness-of-fit tests for copulas. *J. Multivariate Anal.* **95**, 119–152.

Fisher, N. I. & Switzer, P. 2001 Statistical computing and graphics. *Am. Stat.* **55**(3), 233–239.

Genest, C. & Boies, J.-C. 2003 Detecting dependence with Kendall plots. *Am. Stat.* **57**(4), 275–284.

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

Genest, C., Favre, A.-C., Béliveau, J. & Jacques, C. 2007 Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resour. Res.* **43**, 1–12.

Genest, C., Quessy, J.-F. & Remillard, B. 2006 Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Stat.* **33**(2), 337–366.

Genest, C. & Remillard, B. 2008 Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Probab. Stat.* **44**(6), 1096–1127.

Genest, C., Remillard, B. & Beaudoin, D. 2009 Goodness-of-fit tests for copulas: A review and a power study. *Insur. Math. Econ.* **44**, 199–213.

Goel, N. K., Seth, S. M. & Chandra, S. 1998 Multivariate modeling of flood flows. *J. Hydrol. Engng* **124**(2), 146–155.

Grimaldi, S. & Serinaldi, F. 2006*a* Asymmetric copula in multivariate flood frequency analysis. *Adv. Water Resour.* **29**, 1155–1167.

Grimaldi, S. & Serinaldi, F. 2006*b* Design hyetograph analysis with 3-copula function. *Hydrol. Sci.* **51**(2), 223–238.

Gumbel, E. J. 1960 Multivariate extreme distributions. *Bull. Int. Stat. Inst.* **39**(2), 471–475.

Gupta, V. K., Duckstein, L. & Peebles, R. W. 1976 On the joint distribution of the largest flood and its time of occurrence. *Water Resour. Res.* **12**(2), 295–304.

Huard, D., Évin, G. & Favre, A.-C. 2006 Bayesian copula selection. *Comput. Stat. Data Anal.* **51**, 809–822.

Joe, H. 1997 *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

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

Kao, S. & Govindaraju, R. S. 2007*a* Probabilistic structure of storm surface runoff considering the dependence between average intensity and storm duration of rainfall events. *Water Resour. Res.* **43**, 1–15.

Kao, S. & Govindaraju, R. S. 2007*b* A bivariate frequency analysis of extreme rainfall with implications for design. *J. Geophys. Res.* **112**, 1–15.

Kao, S.-C. & Govindaraju, R. S. 2008 Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas. *Water Resour. Res.* **44**, 1–19.

Karmakar, S. & Simonovic, S. P. 2009 Bivariate flood frequency analysis. Part 2: a copula-based approach with mixed marginal distributions. *J. Flood Risk Mgmnt* **2**(1), 32–44.

Kelly, K. S. & Krzysztofowicz, R. 1997 A bivariate meta-Gaussian density for use in hydrology. *Stoch. Hydrol. Hydraul.* **11**: 17–31.

Krstanovic, P. F. & Singh, V. P. 1987 A multivariate stochastic flood analysis using entropy. In: *Hydrologic Frequency Modelling* Singh, V. P., (Ed.), Reidel, Dordrecht, The Netherlands, pp. 515–539.

Long, D. & Krzysztofowicz, R. 1992 Farlie–Gumbel–Morgenstern bivariate densities: Are they applicable in hydrology? *Stoch. Hydrol. Hydraul.* **6**, 47–54.

Marshall, W. M. & Olkin, I. 1967 A multivariate exponential distribution. *Am. Stat. Assoc.* **62**, 30–44.

Michiels, F. & Schepper, A. D. 2008 A copula test space model: how to avoid the wrong copula choice. *Kybernetika* **44**, 864–878.

Nagao, M. & Kadoya, M. 1971 Two-variate exponential distribution and its numerical table for engineering application. *Bull. Disaster Prevention Res. Inst., Kyoto University* **20**(3), 183–215.

Nelsen, R. B. 2006 *An Introduction to Copulas*. Springer, Berlin.

Nikoloulopoulos, A. K. & Karlis, D. 2007 Fitting copulas to bivariate earthquake data: the seismic gap hypothesis revisited. *Environmetrics* **19**(3), 251–269.

Poulin, A., Huard, D., Favre, A. -C. & Pugin, S. 2007 Importance of tail dependence. *J. Hydrol. Engng* **12**(4), 394–403.

Raynal-Villasenor, J. A. & Salas, J. D. 1987 A multivariate stochastic flood analysis using entropy. In: *Hydrologic frequency modeling*, Singh, V. P. (Ed.), Springer, Dordrecht, The Netherlands, pp. 595–602.

Renard, B. & Lang, M. 2007 Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Adv. Water Resour.* **30**, 897–912.

Rosbjerg, D. 1987 On the annual maximum distribution in dependent partial duration series. *Stoch. Hydrol. Hydraul.* **1**, 3–16.

Sackl, B. & Bergmann, H. 1987 A bivariate flood model and its application. In: *Hydrologic Frequency Modelling*, Singh, V. P., (Ed.), Reidel, Dordrecht, The Netherlands, pp. 571–582.

Salvadori, G. & De Michele, C. 2004 Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resour. Res.* **40**, 1–17.

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

Salvadori, G., De Michele, C., Kottegoda, N. T. & Rosso, R. 2007 *Extremes in Nature: An Approach using Copulas*. Springer, Dordrecht, The Netherlands.

Schmidt, R. & Stadtmüller, U. 2006 Non-parametric estimation of tail dependence. *Scand. J. Stat.* **33**, 307–335.

Serinaldi, F. 2008 Analysis of inter-gauge dependence by Kendall's τ_K, upper tail dependence coefficient, and 2-copulas with application to rainfall fields. *Stoch. Environ. Res. Risk Assess.* **22**(6), 671–688.

Serinaldi, F. 2009 A multisite daily rainfall generator driven by bivariate copula-based mixed distributions. *J. Geophys. Res. Atmos.* **114**, D10103.

Serinaldi, F. & Grimaldi, S. 2007 Fully nested 3-copula: Procedure and application on hydrologic data. *J. Hydrol. Engng* **12**(4), 420–430.

Shiau, J. T., Wang, H. Y. & Tsai, C. T. 2006 Bivariate frequency analysis of floods using copulas. *J. Am. Water Resour. Assoc.* **42**(6), 1549–1564.

Silva, R. S. & Lopes, H. F. 2008 Copula, marginal distributions and model selection: a Bayesian note. *Stat. Comput.* **18**, 313–320.

Singh, K. & Singh, V. P. 1991 Derivation of bivariate probability density functions with exponential marginals. *Stoch. Hydrol. Hydraul.* **5**, 55–68.

Singh, V. P. & Zhang, L. 2007 IDF curves using the Frank Archimedean copulas. *J. Hydrol. Engng* **12**(6), 651–662.

Sklar, A. 1959 Fonctions de repartition a *n* dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8**, 229–231.

Todorovic, P. 1978 Stochastic models of floods. *Water Resour. Res.* **14**(2), 345–356.

Todorovic, P. & Woolhiser, D. A. 1972 On the time when the extreme flood occurs. *Water Resour. Res.* **8**(6), 1433–1438.

Tsukahara, H. 2005 Semiparametric estimation in copula models. *Can. J. Stat.* **33**(3), 357–375.

USACE 1999 Risk-based analysis in geotechnical engineering for support of planning studies. *Report #ETL 1110-2-556*.

Villarini, G., Serinaldi, F. & Krajewski, W. F. 2008 Modeling radar rainfall estimation uncertainties using parametric and non-parametric approaches. *Adv. Water Resour.* **31**(12), 1674–1686.

Wang, C., Chang, N.-B. & Yeh, G.-T. 2009 Copula-based flood frequency (COFF) analysis at the confluences of river systems. *Hydrol. Process.* **23**, 1471–1486.

Wang, W. & Wells, M. T. 2000 Model selection and semiparametric inference for bivariate failure-time data. *J. Am. Stat. Assoc.* **95**(449), 62–72.

Yue, S. 2000 The bivariate lognormal distribution to model a multivariate flood episode. *Hydrol. Process.* **14**, 2575–2588.

Yue, S. 2001 A bivariate extreme value distribution applied to flood frequency analysis. *Nordic Hydrol.* **32**(1), 49–64.

Yue, S., Bobee, B., Legendre, P. & Bruneau, P. 1999 The Gumbel mixed model for flood frequency analysis. *J. Hydrol.* **226**(1–2), 88–100.

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

Zhang, L. & Singh, V. P. 2007*a* Bivariate rainfall frequency distributions using Archimedean copulas. *J. Hydrol.* **332**, 93–109.

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

Zhang, L. & Singh, V. P. 2007*c* Trivariate flood frequency analysis using the Gumbel–Hougaard copula. *J. Hydrol. Engng* **12**(4), 431–439.

First received 31 May 2009; accepted in revised form 5 January 2010. Available online February 2011