International Conference "Water, Environment, Energy and Society" (WEES–2009)
New Delhi, 12–16 January 2009

Copulas for Multivariate Flood Frequency Analysis

Hemant Chowdhary
Department of Civil and Environmental Engineering
Louisiana State University, Baton Rouge, LA 70803-6405, USA
E-mail: hchowd1@lsu.edu

Luis A. Escobar
Department of Experimental Statistics
Louisiana State University, Baton Rouge, LA 70803-6405, USA
E-mail: luis@lsu.edu

Vijay P. Singh
Department of Biological and Agricultural Engineering, Texas A&M University
College Station, Texas 77843-2117, USA
E-mail: vsingh@tamu.edu

ABSTRACT: Flood frequency analysis is commonly accomplished by fitting univariate distributions to annual peak flows. Lognormal, gamma, log-Pearson, extreme value, logistic and Wakeby are the commonly employed flood frequency distributions. Hydrological processes, however, exhibit multivariate characteristics and simultaneous consideration two or more component processes may be required and advantageous in a variety of applications. Multivariate flood frequency analysis, involving flood peaks, volumes and durations has been done in the past on a limited basis and has been traditionally accomplished by employing readily available bivariate and multivariate frequency distributions that have marginals from the same family of distributions. Such an approach is highly restrictive in situations where underlying processes are characterized by marginals from different distribution families. This difficulty is usually overcome, in a limited way, by first normalizing or transforming the variables into similar marginals and then employing the available functional distribution forms. The concept of copula overcomes this limitation by allowing combination of arbitrarily chosen marginal types for obtaining joint and conditional multivariate distributions. It also provides a wider choice of admissible dependence structure and easier procedure for generating multivariate random samples, as compared to the conventional approach. A variety of copula families have been evolved and thus the selection of appropriate copula family for different applications is an important first step. The use of copula-based multivariate distributions in the field of hydrology has started only recently and optimal copula structures for hydrological applications are yet to be identified. This paper highlights the merits of copula concept and illustrates its application to multivariate flood frequency analysis by way of investigating relative applicability of six copulas families.

INTRODUCTION

Flood frequency analysis typically involves fitting univariate distributions to annual maximum flows or peak flows or to peak over threshold flows, observed at a location of interest along a river. Commonly employed distributions include 2 and 3-parameter lognormal (LN and LN3), 2-paramter gamma (G2), Pearson type III (P3), log-Pearson type III (LP3), the extreme value type I (EV1) or Largest Extreme Value (LEV) or Gumbel, the Generalized Extreme Value (GEV), and Wakeby distributions, among others. The main objective of various hydrological designs, e.g., for dam spillways, bridges, etc., has been to estimate a flow that shall have an average inter-arrival period more than a specified design period. Hydrological processes, however, exhibit multivariate characteristics and simultaneous consideration of various component processes may be required and crucial in certain situations. Flood phenomenon involves important multivariate hydrologic features such as peak flood discharge, corresponding volume and duration, time to flood peak, rate of rise and recession of flood. The decision to employ univariate, bivariate or multivariate distribution is, however, made primarily on the basis of the objectives of the 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 significant effect on the flood attenuation or where failure mechanism is affected by flood duration and/or volume then these variables are also required to be

considered along with the peak flood discharges. Flood peaks, associated flood volumes and durations are also considered simultaneously in operational flood management measures, such as design of flood retention basins (Sackl and Bergmann, 1987) and in analyzing the risk of damage due to floods. Conditional probability of failure functions, based on both flood peak and duration, are studied for risk assessment of levees and embankments (U.S.A.C.E., 1999). Multivariate analysis, considering flood peak and volume and/or duration is thus essential and would result in improving management strategies and better assessment of potential risk (Michele *et al.*, 2005, and Salvadori and Michele, 2004). Application of multivariate flood frequency analysis involving flood peak, volume and duration has been made in past on a limited basis. This has mostly been bivariate frequency analysis of flood peak and volume. Conventional bivariate frequency distribution models have certain limitations. The copula concept which is emerging as a new way of multivariate frequency distribution analysis overcomes some of the restrictions posed by the conventional multivariate distributions. This study highlights the merits of copula concept and presents its application in the field of multivariate flood frequency analysis by way of investigating relative suitability of six copulas families.

CONVENTIONAL MULTIVARIATE FLOOD FREQUENCY APPROACH

Traditionally, bivariate normal, lognormal, exponential, or Gumbel (called mixed Gumbel) distributions have been applied for hydrological variables such as flood peaks, and associated flood volume and duration. Gupta *et al.* (1976), Todorovic and Woolhiser (1972), and Todorovic (1978) have discussed distributions for time of occurrence of peak flow in relation to the flood event. Ashkar and Rousselle (1982) discussed the multivariate nature of flood peak and corresponding volume and duration involving exponential conditional distributions for flood duration and volume for given peak flow levels. Bivariate stochastic model for flood peak and volume based on the principle of maximum entropy has been suggested by Krstanovic and Singh (1987). Sackl and Bergmann (1987) employed bivariate normal distribution on transformed flood peaks and volumes 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 marginal distribution for duration to be exponentially distributed.

An application of the general form of the logistic model for a bivariate extreme value distribution was demonstrated for obtaining flood frequency distribution at a downstream station on the basis of information from two stations upstream of the junction by Raynal and Salas (1987). In an indirect approach, Rosbjerg (1987) obtained frequency distribution of annual maximum flood from successive peak floods, employing Marshall-Olkin bivariate exponential distribution. A bivariate meta-Gaussian distribution proposed by Kelly and Krzysztofowicz (1997) allows specification of arbitrary marginals and covers full dependence range and is based on the assumption that normal quantile transformed (NQT) variates of original hydrologic variables follow bivariate normal distribution. Extending the work of Raynal and Salas (1987) and Escalante and Raynal (1994), Escalante-Sandoval (1998) employed a multivariate extreme value distribution with mixed Gumbel marginals and applied to data from 42 gaging stations in northern Mexico. Another instance of bivariate consideration of flood peak and volume is available in Goel *et al.* (1998) in which frequency analysis of transformed peak flows and volumes, from an Indian river, are modeled using bivariate normal distribution. Yue *et al.* (1999) employed bivariate Gumbel mixed model, originally proposed by Gumbel (1960), for obtaining joint and conditional probabilities for flood peak and volume and for volume and duration. Yue (2001) applied the bivariate extreme value distribution and bivariate lognormal distribution for multivariate flood frequency analysis. Most of these applications involve multivariate distributions that restrict marginals from the same distribution families. This and some other limitations posed by the conventional multivariate approach are described in the following subsection.

LIMITATIONS OF CONVENTIONAL MULTIVARIATE APPROACH

All the traditional multivariate distribution approaches stated above, except for the meta-Gaussian method (Kelly and Krzysztofowicz, 1997), have limitations of allowing marginals from the same family. However, different hydrological applications may involve multiple variables, not all of which belong to the same distribution type. Transformation to normal distribution and consequent fitting of multivariate normal distribution has often been resorted in such situations. Extensive efforts, spanning decades of research work in the area of flood frequency analysis, has resulted in identification of some plausible candidate distribution functions. The lack of multivariate distributions featuring

marginals from different distributions restricts the ability to directly utilize such suitable distribution functions as marginals. This makes migration, from univariate to multivariate flood frequency analysis, sub-optimal, as noted for example by Choulakian *et al.* (1990).

Furthermore, several multivariate distributions also do not allow for a full coverage of possible dependence between different variables. Few examples that could be readily mentioned in this regard are the bivariate exponential and bivariate Gumbel distributions. The bivariate exponential distribution imposes a critical restriction of the variables to be negatively associated such that the Pearson's correlation coefficient ρ is between –0.404 and 0. On the other hand, the bivariate exponential distribution given by Moran and studied by Nagao and Kadoya (1971) provides an alternative formulation admitting full positive correlation coefficient. Although both these formulations complement each other in terms of covering wider range of correlation coefficient but are not versatile enough individually. Similarly, the bivariate extreme value distribution given by Gumbel (1960) admits only a partial positive range of ρ to the extent of 0 to 2/3. Although not applied for hydrological variables, Farlie-Gumbel-Morgenstern (F-G-M) family of distributions, as applied for rainfall variables by Singh and Singh (1991) and studied later by Long and 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 that of Pearson's linear correlation measure being linked to the dependence parameter, either directly or indirectly. The Pearson's linear correlation coefficient is not invariant to non-linear monotonic transformations and depicts linear correlation rather than the functional association. It may also not be even estimable in certain situations involving heavy-tailed distributions. These restrictions are overcome by copula-based procedure and the same is outlined in the following section.

COPULA CONCEPT

Copula is a simple concept wherein bivariate and multivaraiate probabilities are expressed in terms of marginal probabilities and more advantageously in terms of uniform marginals. This theory has been in vogue for some time now, especially with respect to actuarial science and finance applications, and in recent years has also made an impressive beginning in the field of hydrological engineering. Several illustrative and review studies, such as Favre *et al.* (2004), Salvadori and De Michele (2004), De Michele *et al.* (2005), Genest and Favre (2007), Poulin *et al.* (2007), Salvadori and De Michele (2007), Serinaldi and Grimaldi (2007), Zhang and Singh (2007) provide elaborate discussion on copula applications related to flow variables and explain its advantages and limitations. There are several other application studies covering rainfall variables that have been reported but not listed here. 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,

$$F(x,y)=C\left[F_X(x),F_Y(y)\right],\ x,y\in R \qquad \ldots (1)$$

where $F_X(x)$ and $F_Y(y)$ 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 model for (X,Y) is obtained whenever the three constituents (C, F_X, and F_Y) are chosen from given parametric families, viz,

$$F_X(x;\delta),\quad F_Y(y;\eta),\quad C(u,v;\theta) \qquad \ldots (2)$$

where δ and η are the parameter vectors of marginal distributions and θ is the dependence structure parameter vector. u and v are the quantiles of the uniformly distributed variables $U=F_X(x)$ and $V=F_Y(y)$ respectively. Several classes and families of copulas, such as the meta-elliptic copulas, extreme value copulas, and Archimedean copulas, exist. For an elaborate introduction about copulas reference may be made to Joe (1997), Cherubini *et al.* (2004), Nelsen (2006), and/or Genest and Favre (2007), among others. These references have been used substantially in collating the material for this article. Of the several families of copulas that exist, the one that has been frequently applied in the field of hydrology is the Archimedean family. This copula family has the form,

$$\begin{aligned}\phi\left[F(x,y)\right]&=\phi\left\{C\left[F_X(x),F_Y(y)\right]\right\}\\&=\phi\left[F_X(x)\right]+\phi\left[F_Y(y)\right]\end{aligned} \qquad \ldots (3)$$

where ϕ, a continuous, strictly decreasing function from $I[0,1]\rightarrow[0,\infty]$ and with $\phi(1)=0$, is called a generating function. The joint probability function can then be written as,

$$F(x,y) = C\left[F_X(x), F_Y(y)\right] = \phi^{[-1]}\left\{\phi\left[F_X(x)\right] + \phi\left[F_Y(y)\right]\right\} = C(u,v) = \phi^{[-1]}\left\{\phi(u) + \phi(v)\right\} \quad \ldots (4)$$

Parameter θ is hidden in the generating function, for example, the Clayton copula family, the one that has been employed for several hydrological applications, involves θ in the generating function in the form,

$$\phi(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right), \quad \theta \in [-1, \infty), \theta \neq 0 \quad \ldots (5)$$

Employing the above generating function and the form of Archimedean copulas given in Eqn. (4), the bivariate cumulative probability distribution is obtained as,

$$F(x,y) = \left\{\left[F_X(x)\right]^{-\theta} + \left[F_Y(y)\right]^{-\theta} - 1\right\}^{-\frac{1}{\theta}} = \left\{u^{-\theta} + v^{-\theta} - 1\right\}^{-\frac{1}{\theta}} = C(u,v) \quad \ldots (6)$$

$C(u,v)$ here is called the copula probability. Double differentiating the above bivariate probability function results in the joint density function as,

$$f(x,y) = f_X(x) f_Y(y) (1+\theta)\left[F_X(x) F_Y(y)\right]^{-\theta-1} \left\{\left[F_X(x)\right]^{-\theta} + \left[F_Y(y)\right]^{-\theta} - 1\right\}^{-\frac{1}{\theta}-2} = f_X(x) f_Y(y)\left[(1+\theta)(uv)^{-\theta-1}\left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}-2}\right] = f_X(x) f_Y(y)\, c_\theta(u,v) \quad \ldots (7)$$

where $f_x(x)$ and $f_y(y)$ are the marginal densities and $c_\theta(u, v)$ is the copula density. Expressions for probabilities of six copula families and corresponding parameter space and generating functions are given in Table 1. Generating function is not applicable for F-G-M and Galambos copulas as these belong to non-Archimedean families.

PARAMETER ESTIMATION

The parameters describing the copula dependence structure can be estimated by non-parametric, semi-parametric and parametric methods. The non-parametric method and semi-parametric method (called pseudo-maximum likelihood method) rely on the relative ranks of the joint variates. The parametric method utilizes the classical maximum likelihood procedure for estimating parameters of marginals and the dependence structure. These are outlined here below.

Estimate Based on Non-Parametric Measures of Association

This approach is based on the pretext that the dependence structure is fully defined by the relative ranks of individual variables and by a single parameter. Such basis also renders the dependence structure completely independent of the choice of the marginals. Non-parametric estimates of θ based on Kendall's Tau (τ) and Spearman's rho (ρ_s) are obtainable using the formulations given by Genest and Mackay (1986) as,

$$\tau = 4\int_{[0,1]^2} C(u,v)\, c_\theta(u,v)\, du\, dv - 1 \quad \ldots (8)$$

$$\rho_s = 12\int_{[0,1]^2} C(u,v)\, du\, dv - 3 \quad \ldots (9)$$

Table 1: Probability Function, Parameter Space, Generating Function and Relationship with Non-parametric Measure of Association for Six Copula Families under Consideration

Copula	$C_\theta(u,v)$	*Parameter Space*	*Generator* $\phi(t)$	*Kendall's tau* τ
A-M-H[1]	$\frac{uv}{1-\theta(1-u)(1-v)}$	[–1, 1)	$\ln\frac{1-\theta(1-t)}{t}$	A–$B\ln(1-\theta)$
Clayton	$\left[\max\left(u^{-\theta}+v^{-\theta}-1, 0\right)\right]^{-1/\theta}$	[–1, ∞\{0}	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$	θ/(θ + 2)
F-G-M	$uv\left[1+\theta(1-u)(1-v)\right]$	[–1, 1]	n.a.	2θ/9
Frank[2]	$-\frac{1}{\theta}\ln\left[1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{\left(e^{-\theta}-1\right)}\right]$	(–∞,∞)\{0}	$-\ln\frac{e^{-\theta t}-1}{e^{-\theta}-1}$	$1+\frac{4}{\theta}\left[D_1(\theta)-1\right]$
Galambos[3]	$uv\exp\left[\left(\tilde{u}^{-\theta}+\tilde{v}^{-\theta}\right)^{-1/\theta}\right]$	[1, ∞)	n.a.	n.a.
G-H[3]	$\exp\left[-\left(\tilde{u}^{-\theta}+\tilde{v}^{-\theta}\right)^{1/\theta}\right]$	[1, ∞)	$(-\ln t)^\theta$	1 – 1/θ

[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 $\tilde{u} = -\ln u$ and $\tilde{v} = -\ln v$

Estimates of dependence parameters for some copulas families such as those of Ali-Mikhail-Haq, Clayton, Frank, and Gumbel-Hougaard, among others are available in closed form while others can be obtained numerically. For example, for Clayton and Farlie-Gumbel-Morgenstern copula families the relationship of Kendall's tau and Spearman's rho with dependence parameter is given as,

$$\tau = \frac{\theta}{\theta+2} \text{ and } \rho_s = \frac{\theta}{3} \quad \ldots (10)$$

Based on the above, a sample-based estimate of θ, much like a moment based estimate, is obtained respectively as,

$$\hat{\theta} = \frac{2\hat{\tau}}{(1-\hat{\tau})} \text{ and } \hat{\theta} = 3\hat{\rho}_s \quad \ldots (11)$$

Such relationships of dependence parameter θ with Kendall tau for few copula families are given in Table 1.

Maximum Pseudo-Likelihood Estimator (PMLE)

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

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

where $\tilde{F}_X(x) = R_i/(n+1)$ and $\tilde{F}_Y(y) = S_i/(n+1)$ are the non-parametric marginal probabilities based on the bivariate ranks (R_i, S_i). In other words, Maximum Likelihood (ML) estimate of only θ is obtained.

Maximum Likelihood Estimator

The classical ML estimation of parameters of copulas involves maximization of log-likelihood function given by,

$$l(\theta,\delta,\eta) = \sum_{i=1}^{n} \log\left\{ c_\theta\left[F_X(x;\delta), F_Y(y;\eta)\right]\right\} \quad \ldots (13)$$

where δ and η are parameters of the marginals $F_X(x,\delta)$ and $F_Y(y,\eta)$, and θ is the parameter vector of the dependence structure. All these parameter are estimated simultaneously in this method. A variant of this ML approach is called "Inference From Margins" (IFM) method, wherein univariate ML estimates of δ and η are first obtained separately and then ML estimate of θ is obtained. The log-likelihood in this case is expressed as,

$$l(\theta) = \sum_{i=1}^{n} \log\left\{ c_\theta\left[\breve{F}_X(x;\delta), \breve{F}_Y(y;\eta)\right]\right\} \quad \ldots (14)$$

where $\breve{F}_X(x;\delta)$ and $\breve{F}_Y(y;\eta)$ indicate margins having parameters δ and η that are obtained on univariate basis using ML method. IFM approach is advocated for multivariate copulas of larger dimensions when estimation through classical approach becomes unwieldy. It is interesting to note that though the classical ML approach is more general but smallest mean squared errors are reported for the maximum pseudo-likelihood method (Tsukahara, 2005).

APPLICATION

The hydraulic infrastructure along a river, such as dams, bridges, levees etc. are designed in order to safely carry the maximum flows that may occur with the designed probability of non-exceedence. However, 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 etc. warrant simultaneous consideration of multiple flood related variables such as peak river flows, associated volumes and durations. Such considerations are very important for the insurance companies in order to be aware of the actual risk due to flooding and associated damages. An application in that direction is presented here with the objective of obtaining plausible joint densities and probabilities of river flow peaks and associated flood volumes. Six different copula types, namely Ali-Mikhail-Haq (A-M-H), Clayton, Farlie-Gumbel-Morgenstern (F-G-M), Frank, Galambos, and Gumble-Hougaard (G-H) have been considered in this study. Four of these, A-M-H, Clayton, Frank and G-H are Archimedean in nature. F-G-M copulas on the other hand are non-Archimedean and involve quadratic sections. Galambos and G-H (which is Archimedean also) belong to the extreme-value copula families.

Data Set

The Annual peak flows and average daily flows of Greenbrier River at Alderson station (USGS station # 03183500) in United States' West Virginia state are obtained from the USGS website. The Greenbrier

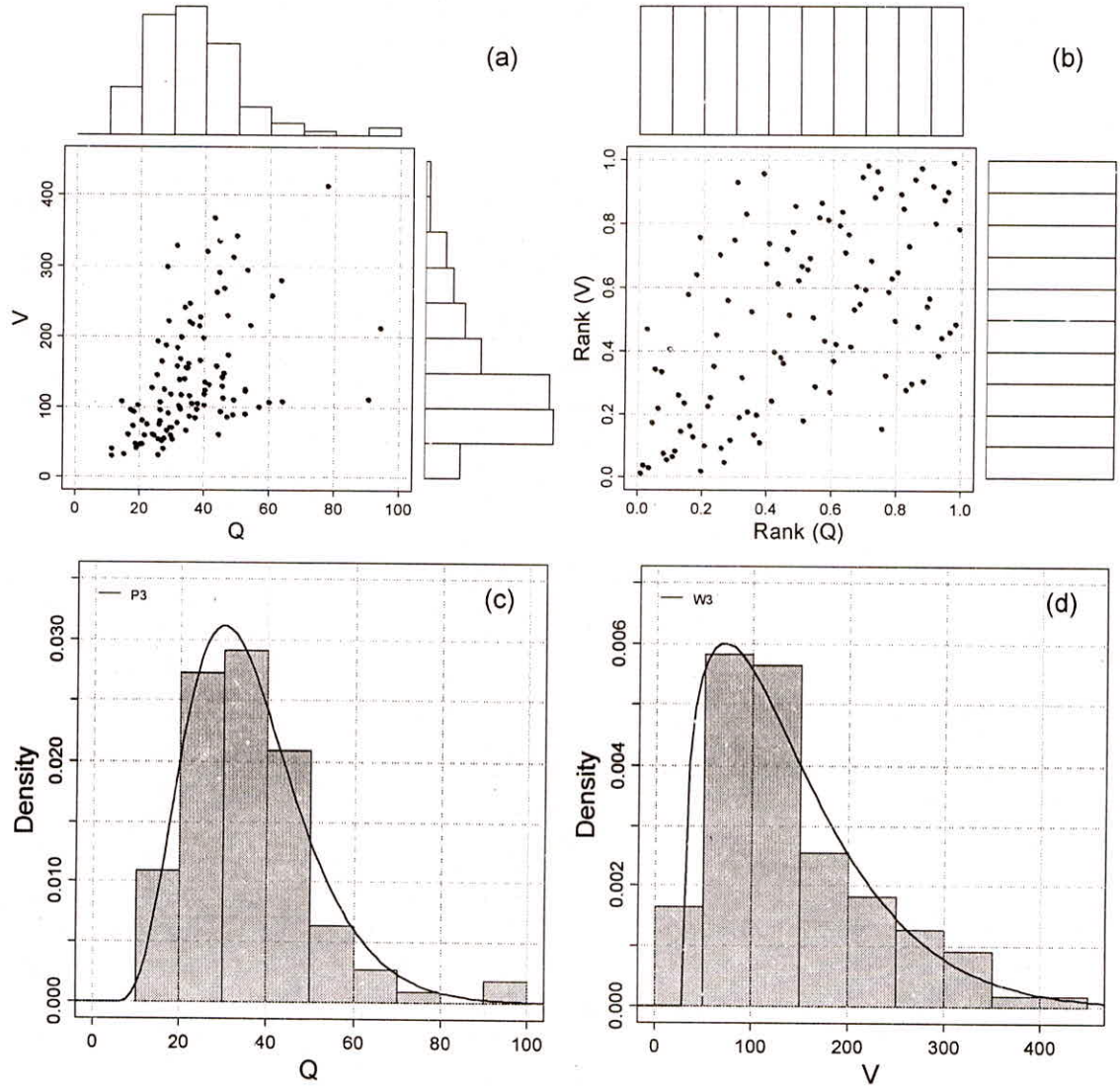

Fig. 1: Characteristics of observed bivariate annual peak flow (Q in 10^3 cusec) and volume (V in 10^3 cusec-day) data of Greenbrier River at Alderson gauging station—(a) scatter plot and histograms in original domain; (b) scatter plot of ranks and corresponding uniform histograms; (c and d) histograms along with P3 and W3 probability density curves respectively

River is a tributary of the New River in 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. The station commands a drainage area and contributing area of 1,364 square miles. A length of 110 years of data, from 1896 to 2005, is considered for this analysis. Volumes of flood events associated with the peak flows are obtained from the record of average daily flows. The scatter plot of this bivariate data and of their ranks, along with the respective histograms, is shown in Figure 1. As ranks are the scaled empirical probabilities, they are approximately uniformly distributed between 0 and 1, as seen in Figure 1.

Potential Marginal Distributions

Several candidate distributions, such as 2 and 3-parameter lognormal (LN2 and LN3), 2-paramter gamma (G2), Pearson type III (P3), log-Pearson type III (LP3), and largest extreme value (LEV) are considered for fitting the annual peak flows and volumes on univariate basis. On the basis of Kolmogorov-Smirnov, Anderson Darling, and Chi-Squared fit statistics and overall fit of the Q–Q plots, Pearson Type III and 3-parameter Weibull distributions were taken as the marginals for flood peak and volume respectively. The overlay of density curves of these distributions and the corresponding histograms is

shown in Figure 1. The density functions for P3 and W3, $f_X(x)$ and $f_Y(y)$, are

$$f_x(x)=\frac{1}{\alpha_X\,\Gamma(\beta_X)}\left(\frac{x-\gamma_X}{\alpha_X}\right)^{\beta_X-1}\exp\left[-\left(\frac{x-\gamma_X}{\alpha_X}\right)\right] \quad \ldots (15)$$

$$f_Y(y)=\frac{\alpha_Y}{\beta_Y}\left(\frac{y-\gamma_Y}{\beta_Y}\right)^{\alpha_Y-1}\exp\left[-\left(\frac{x-\gamma_Y}{\beta_Y}\right)^{\alpha_Y}\right] \quad \ldots (16)$$

where α_X, β_X and $\alpha_Y, \beta_Y > 0$ are scale and shape and $\gamma_X \le x < +\infty$ $\gamma_Y \le y < +\infty$ are the location parameters. The ML parameters estimates for these two marginals are obtained as $\hat{\gamma}_X = 4.601$, $\hat{\alpha}_X = 6.197$, $\hat{\beta}_X = 5.101$, and $\hat{\gamma}_Y = 28.361$, $\hat{\alpha}_Y = 122.185$, $\hat{\beta}_Y = 1.326$. And the corresponding standard errors are $Se_{\hat{\gamma}_X} = 4.332$, $Se_{\hat{\alpha}_X} = 1.365$, $Se_{\hat{\beta}_X} = 1.715$, and $Se_{\hat{\gamma}_Y} = 1.907$, $Se_{\hat{\alpha}_Y} = 6.093$, $Se_{\hat{\beta}_Y} = 0.107$.

Estimation of Dependence Structure

The dependence parameters for the six copula families under consideration are estimated by (a) empirical and (b) pseudo-ML methods and the results are summarized in the following sections.

Empirically-based Parameter Estimation

The sample estimates of ρ, τ, and ρ_s are 0.466, 0.389, and 0.555 respectively. And with corresponding p-values of 1.78e–09, 3.0e–10, and 2.9e–07 these indicate significant positive dependence. A qualitative assessment of the dependence structure between the two variables can be made by way of scatter plot of the raw data and the ranks of the data. The plot of ranks, in Figure 1, is a better assessment in view of its invariant properties to non-linear monotonic transformation. The nature of positive orientation of scatter point in these plots corroborates the inference of positive dependence between Q and V. Additionally, Chi-plots and K-plots, proposed by Fisher and Switzer (1985) and Genest and Boies (2003) respectively, can be constructed for an objective graphical assessment. Chi-plots are based on the chi-squared statistics for independence in a two-way table. The plot also includes the control limits corresponding to a chosen significance level. Scatter of the chi-plot predominantly within these control limits indicates independence and vice-versa. When the scatter is largely on the upper (lower) side of the control limits, it indicates positive (negative) dependence. K-plot is another graphical tool, much like a Q–Q plot, that involves plotting of various order statistics of bivariate probabilities against the expected values of the same from a random sample of $W = C(U,V) = F(X,Y)$, of the same size n as the observed data, under the null hypothesis of independence between U and V or X and Y. The diagonal line indicates independence, whereas the curve given by $K_0(w) = w - w\log(w)$ corresponds to perfect positive dependence. In case of perfect negative dependence all the points lie on the x-axis. The Chi and K-plots for the data under consideration are given in Figure 2. It may be seen that similar to the assessment from the quantitative estimates and plot between ranks, both Chi and K-plots indicates significant positive dependence.

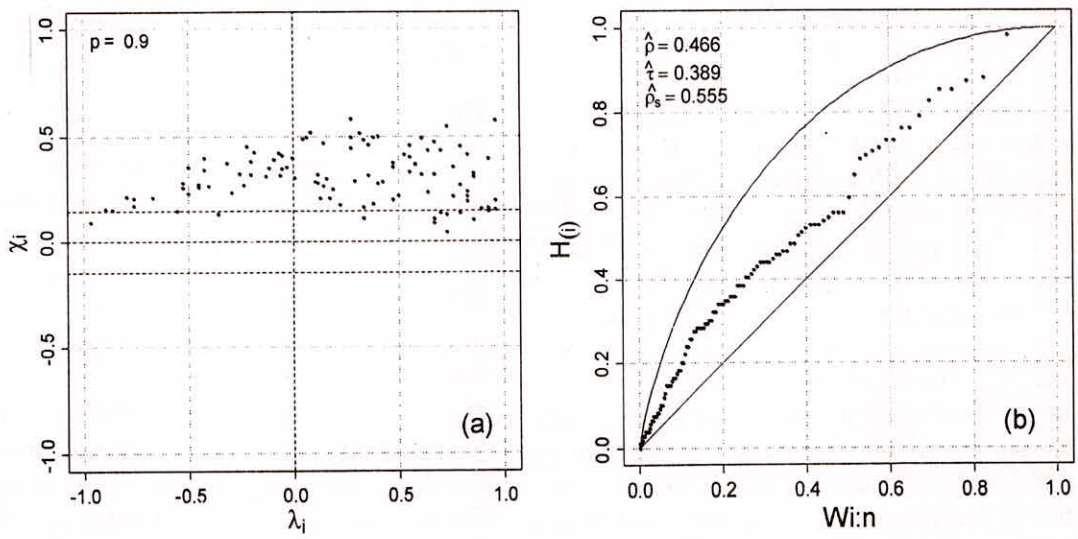

Fig. 2: Characterization of dependence structure using (a) Chi and (b) K plots

Table 2: Point and Interval dependence Parameter Estimates Based on Empirical and Pseudo-Maximum Likelihood Methods (p-value = 0.95)

Copula Family	Empirical Interval Estimate				PMLE based Interval Estimate				
	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
A-M-H	–	–	–	–	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
F-G-M	–	–	–	–	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
G-H	1.642	1.361	1.922	0.143	16.118	1.529	1.388	1.670	0.072

Based on the relationship between τ and dependence parameter θ, as given in Eqn. (9) and utilizing the available closed forms given in Table 1, the dependence parameter is estimated. These point estimates along with the corresponding standard errors and the interval estimates at a significance level of 0.95 are given in Table 2. The estimates for A-M-H and F-G-M copulas are not obtainable for this data set as the value of τ is beyond the admissible limit. The A-M-H copula requires τ to be between –0.1817 and 0.3333 whereas F-G-M copula admits it in the range of –2/9 and 2/9 (–0.222 to 0.222) only. This illustrates limitations of these copula structures, similar to that faced by some of the conventional distributions.

Pseudo-Maximum Likelihood Based Estimation

The dependence parameter based on Pseudo-Maximum Likelihood Estimation (PMLE) is computed by employing maximum likelihood function given in Eqn. (13). The maximized likelihood function value, and point and interval estimate of dependence parameter for the six copula families is given in Table 2. It may be seen from these results that the standard errors of the dependence parameter estimates from this method is an order of magnitude lower than those obtained using the empirically-based method and thus are preferable. It is noted that non-optimal estimates are obtained for the A-M-H and F-G-M copula types and considered here primarily to see how adversely the final outputs are affected.

Assessment of Copula Fitting

More than one copula structures may adequately fit the data at hand. It is imperative to ascertain relative suitability of such plausible copula families. Comparisons are normally made using (a) graphical methods, (b) error statistics, and (c) formal goodness of fit statistics. Although the typical probability plots used for univariate frequency analysis are not applicable, owing to the bivariate or multivariate nature of data, plots with similar basis are however proposed for such cases. Plot involving comparison of empirical and computed probability distribution of random variable $W = C(U, V)$, as discussed with respect to k-plots, is one option for the graphical evaluation. Another option is to draw a Q–Q type of plot between the order statistics $W_{(1)} \le W_{(2)} \le \ldots \le W_{(n)}$ of W, as has been illustrated by Genest and Favre (2007). Comparison of observed data with the large number of generated random samples is yet another option. For this study, six sets of random samples of size 500 are generated employing the approach outlined in Nelsen (2006) for the six copula families under consideration, respectively. The dependence parameters obtained above by PMLE method are used while utilizing the copula-based conditional bivariate distributions for generating bivariate random samples. The comparison of observed data and 500 generated random samples for the six copula families is shown in Figure 3. It may be seen from these plots that the general nature of spread of observed data matches with that of 500 random samples. However, a closer look reveals that for Galambos and G-H copulas there are certain random samples in the upper tail that do not have similar representation by the observed data. In that sense, Clayton and Frank copulas may be adjudged having better representation of the observed data.

A simple quantitative performance assessment of various copula families is made by comparing error statistic such as Root Mean Square Error (RMSE) from empirical and computed bivariate probabilities. Similarly, other error statistics like mean absolute error (ME-A-ERR), mean error (MN-ERR) and maximum absolute error (MX-A-ERR) reflect other important characteristics of this comparison. An account of this comparison is given in Table 3. It may be seen that

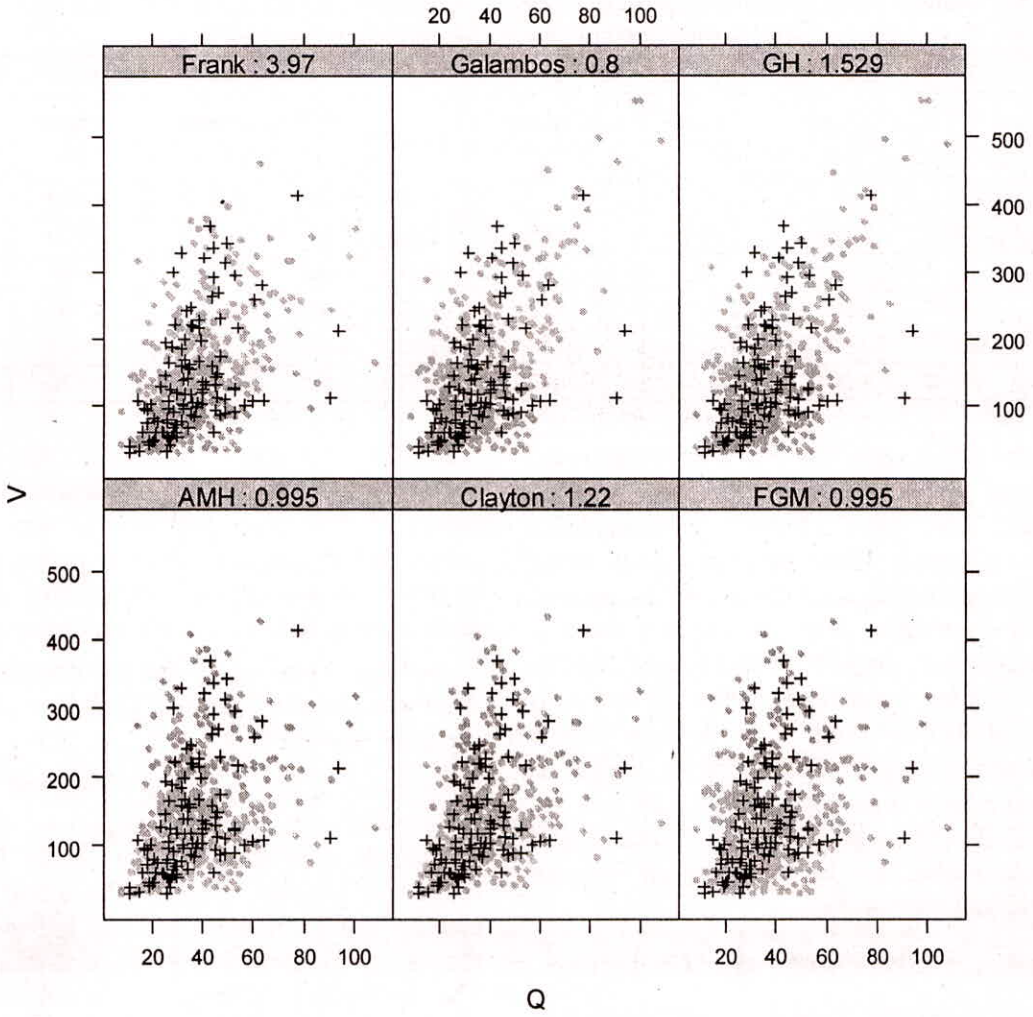

Fig. 3: Sets of 500 random samples based on pseudo-maximum likelihood estimates of dependence parameter ($\hat{\theta}$) for the six copula families under consideration. The solid circles in grey color are the random samples whereas "+" symbols represent observed data

Table 3: Error Estimates for Empirical and Pseudo-Maximum Likelihood-Based Copula Models

Copula Family	*Empirically-based Copula Model**				*PMLE-based Copula Model*			
	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	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	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

* - A-M-H and F-G-M copulas consider same dependence parameter as obtained from the PMLE method.

Clayton copula yields lowest errors and F-G-M copula yields largest errors in all these error categories. The reasoning for the poor performance of F-G-M copula is obvious as this copula admits τ up to 0.222 only whereas the sample estimate was much higher at 0.389. Similarly, poor but slightly better performance of A-M-H copula than F-G-M may be attributed to the fact that this copula also allows τ to be 0.333 at the most. Another important observation in regard to Frank, Galambos and G-H copulas is that although

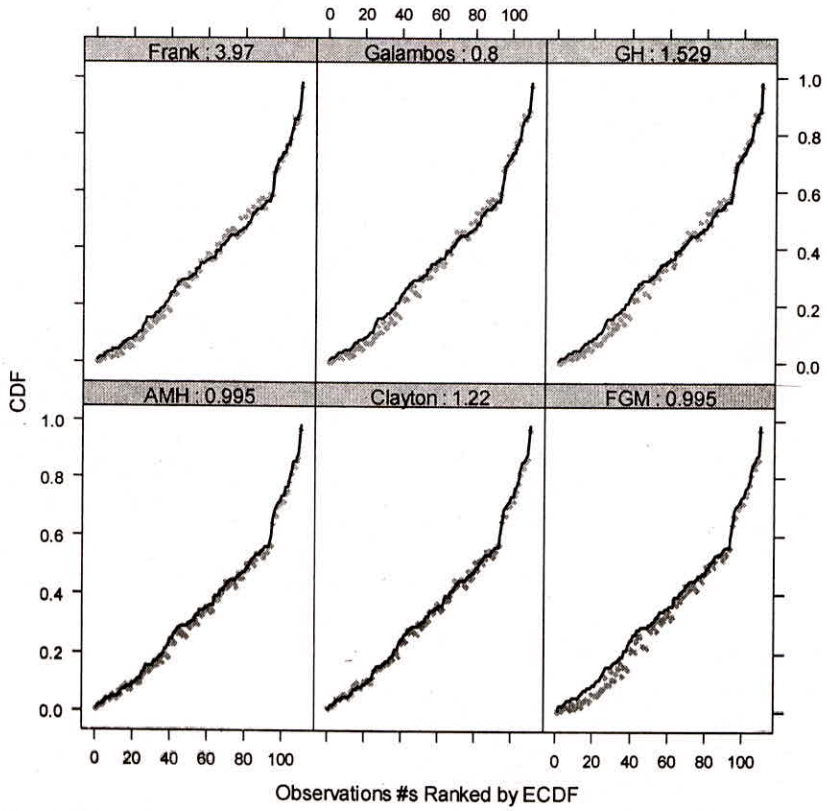

Fig. 4: Comparison of empirical and PMLE-based computed probabilities

they performed slightly inferiorly than A-M-H 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 Frank, Galambos, and G-H copulas may be considered to have given a comparatively better fit. It may also be seen from the table that the errors from both the methods are of similar order of magnitude. This comparison is graphically depicted in Figure 4, wherein ordered empirical probabilities are shown as solid black line and corresponding computed probabilities are plotted as grey points. The inference from this graphical comparison is similar to that derived from the tabular results. The Clayton copula-based joint probabilities and densities, both in 3-D and contour forms, are given in Figure 5. Figure 5(a) illustrates the close match between the computed and empirical probabilities.

CONCLUSIONS

At present, the use of copula-based multivariate distributions in the field of hydrological engineering is in the initial stages of development and application. There have been few studies made in this regard and this number is growing at a faster pace. The above study presents a successful application in this direction and reinforces that copula-based bivariate frequency distribution can be effectively applied in this domain. The major advantage of copula usage is the possibility of able to combine arbitrary margins as per the wishes of the analyst which is normally dictated by the specific nature of the margins. This overcomes the main limitation of the conventional approach wherein limited number of functional distributional forms involving similar marginals is available. This study also highlights the fact that certain copula structures may not be comprehensive enough in terms of covering the entire dependence space. The inferior performance of A-M-H and F-G-M copulas is attributable to this limitation and it parallels similar restrictions posed by some of the conventional distributions. However, since there is availability of several copula types, the ones that cover the required dependence can be employed. Considering the graphical and quantitative methods of performance evaluation presented in this study it may be concluded that although Clayton, Frank, Galambos, and G-H copulas all performed almost equally well, performance of Clayton copula was best among all in various ways, followed closely by Frank copula. Clayton copula provided the least errors among all the

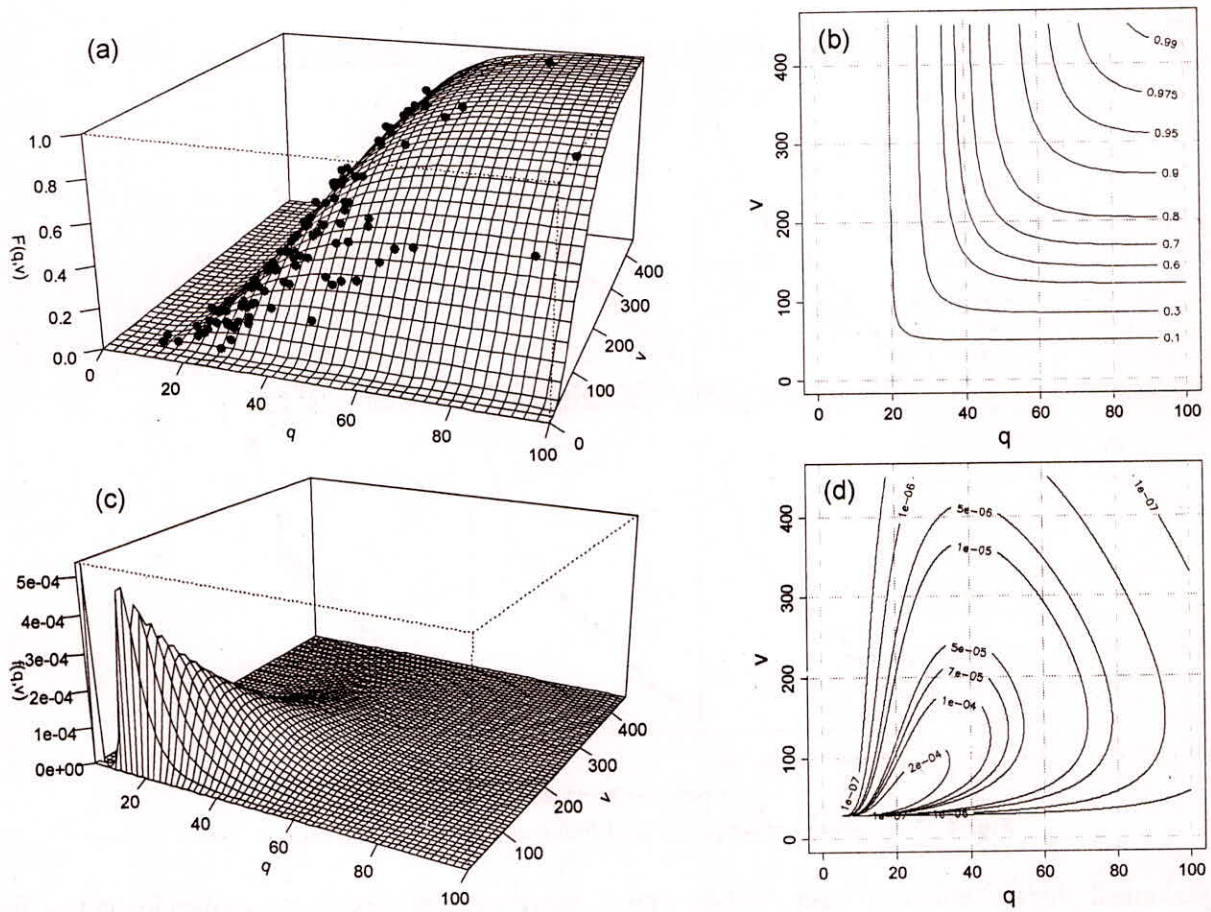

Fig. 5: Plots of joint probability (a and b) and joint density (c and d)

error statistics considered in the study. These two copula types also provided a closer match between the large sized random sample and observed data. The Galambos and G-H copulas, although from the extreme value copulas families, did not perform satisfactorily in terms of the match between the random samples and the observed data. Some of the extreme values generated by these copula types are significantly greater than those in the 110 years of observed data. As Clayton and Frank copulas are also comprehensive in nature, these copula types appear to be better suited for applications involving annual flood peaks and volumes.

REFERENCES

Ashkar, F. and Rousselle, J. (1982). "A multivariate statistical analysis of flood magnitude, duration and volume." Statistical analysis of rainfall and runoff. V.P. Singh ed., Fort Collins, Colorado, *Water Resources Publication*, 651–669.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula methods in finance*, Wiley, NY.

Choulakian, V., EI-Jabi, N. and Moussi, J. (1990). "On the distribution of flood volume in partial duration series analysis of flood phenomena." *Stochastic Hydrology and Hydraulics*, 4: 217–226.

Correia, F.N. (1987). "Multivariate partial duration series in flood risk analysis." *Proc. Hydrologic frequency modeling*, Vijay P. Singh ed., Reidel, Dordrecht, The Netherlands.

De Michele, C. and Salvadori, G. (2003). "A generalized Pareto intensity-duration model of storm rainfall exploiting 2-copulas." *Journal of Geophysical Research*, 108(D2): 4067.

De Michele, C., Salvadori, G., Canossi, M., Petaccia, A. and Rosso, R. (2005). "Bivariate statistical approach to check adequacy of dam spillway." *J. Hyd. Engg., ASCE,* Vol. 10(1): 50–57.

Escalante-Sandoval, C. (1998). "Multivariate extreme value distribution with mixed Gumbel marginals." Journal of the American Water Resources Association, 34, 321–33.

Escalante-Sandoval, C.A. and Raynal-Villasenor, J.A. (1994). "A trivariate extreme value distribution applied to flood frequency analysis." Journal of Research, 99, 369–75.

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

Genest, C. and Favre, A.-C. (2007). "Everything you always wanted to know about copula modeling but were afraid to ask." *J. Hyd. Engg., ASCE,* 12(4), 347–368.

Goel, N.K., Seth, S.M. and Chandra, S. (1998). "Multivariate modeling of flood flows." *J. Hyd. Engrg., ASCE,* 124(2), 146–155.

Gumbel, E.J. (1960). "Multivariate extreme distributions." Bulletin of the International Statistical Institute, 39(2), 471–475.

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

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

Kelly, K.S. and Krzysztofowicz, R. (1997). "A bivariate meta-Gaussian density for use in hydrology." Stochastic Hydrology and Hydraulics, 11, 17–31.

Krstanovic, P.F. and Singh, V.P. (1987). "A multivariate stochastic flood analysis using entropy." *Proc. Hydrologic frequency modeling*, Vijay P. Singh ed., Reidel, Dordrecht, The Netherlands.

Long, D. and Krzysztofowicz, R. (1992). "Farlie-Gumbel-Morgenstern bivariate densities: Are they applicable in hydrology?" Stochastic Hydrology and Hydraulics, 6, 47–54.

Nagao, M. and Kadoya, M. (1971). Two-variate exponential distribution and its numerical table for engineering application. Bulletin of the Disaster Prevention Research Institute. Kyoto University, 20(3), 183–215.

Nelsen, R.B. (2006). *An introduction to copulas*, Springer Verlag.

Poulin, A., Huard, D., Favre, A.-C. and Pugin, S. (2007). "Importance of tail dependence." *J. Hyd. Engg., ASCE,* 12(4), 394–403.

Raynal, J.A. and Salas, J.D. (1987). A multivariate stochastic flood analysis using entropy. *Proc. Hydrologic frequency modeling*, Vijay P. Singh ed., Reidel, Dordrecht, The Netherlands.

Rosbjerg, D. (1987). "On the annual maximum distribution in dependent partial duration series." *Stochastic Hydrology and Hydraulics*, 1, 3–16.

Sackl, B. and Bergmann, H. (1987). A bivariate flood model and its application. *Proc. Hydrologic frequency modeling*, Vijay P. Singh ed., Reidel, Dordrecht, The Netherlands.

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

Salvadori, G. and De Michele, C. (2006). "Statistical characterization of temporal structure of storms." *Advances in Water Resour.*, 29(6), 827–842.

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

Serinaldi, F. and Salvatore, G. (2007). "Fully nested 3-copula: Procedure and application on hydrologic data." *J. Hyd. Engg., ASCE,* 12(4), 420–430.

Sklar, A. (159). "Fonctions de repartition a n dimensions et leurs marges." *Publ. Inst. Stat. Univ. Paris*, 8, 229–231.

Singh, K. and Singh, V.P. (1991). "Derivation of bivariate probability density functions with exponential marginals." *Stochastic Hydrology and Hydraulics*, 5, 55–68.

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

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

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

U.S.A.C.E. (1999). Report no. ETL 1110–2–556, "*Risk-based analysis in geotechnical engineering for support of planning studies*".

Yue, S. (1999a). "Applying Bivariate Normal Distribution to Flood Frequency Analysis." *Water International*, 24(3), 248–254.

Yue, S. (2001a). "The Gumbel logistic model for representing a multivariate storm event" *Advances in Water Resources*, 24(2), 179–185.

Zhang, L. and Singh, V.P. (2007). "Trivariate flood frequency analysis using the Gumbel-Hougaard copula." *J. Hyd. Engg., ASCE,* 12(4), 431–439.