DOI: 10.24850/j-tyca-15-02-01

Articles

Bivariate frequencies analysis of annual floods through practical approach of the Copula functions

Análisis de frecuencias bivariado de crecientes anuales mediante enfoque práctico de las funciones Cópula

Daniel Francisco Campos-Aranda[1], ORCID: https://orcid.org/0000-0002-9876-3967

[1]Retired professor of the Autonomous University of San Luis Potosi, Mexico, campos_aranda@hotmail.com

Corresponding author: Daniel Francisco Campos-Aranda, campos_aranda@hotmail.com

Abstract

The study of hydrological safety in *reservoirs* is carried out with the so–called *Design Flood Hydrograph*. The simplest estimation of such a graph is based on the *bivariate frequency analysis* (BFA), by defining its maximum flow (*Q*) and volume (*V*), associated with a *joint* design return period. The *Copula functions* (*CF*) are based on the dependence between *Q* and *V*, and define the bivariate distribution by means of the previously adopted marginal univariate functions. The *practical approach* adopted uses *CF* with a single fit parameter and selects the most appropriate, based on the dependence shown by the joint record of *Q* and *V*, at the far-right end of its empirical distribution. Moreover, the practical approach compares and ratifies the adopted *CF*, against several of common use in the BFA. The above, using the fitting errors between the empirical and theoretical bivariate probabilities. Due to their importance, the search for marginal distributions was carried out based on the *L* quotient diagram, adopting the best three and comparing them to the highly versatile Kappa and Wakeby functions. The BFA of the 55 annual floods registered in the La Cuña hydrometric station of the Hydrological Region No. 12-3 (Santiago River), Mexico was carried out. Four *joint* design return periods were defined and the calculation of their AND curves is detailed. Finally, several Conclusions are cited which highlight the advantages of the use of *CF* in flood BFAs.

Keywords: Families of Copula functions, Kendall's tau quotient, Spearman's rho coefficient, joint empirical probabilities, right-tail dependency, joint return periods, critical events.

Resumen

El estudio de la seguridad hidrológica en los *embalses* se realiza con el llamado *hidrograma de la creciente de diseño*. El proceso más sencillo y aproximado para estimar tal gráfica se basa en el *análisis de frecuencias bivariado* (AFB) para definir su gasto máximo (*Q*) y volumen (*V*), asociados con el periodo de retorno *conjunto* de diseño. Las *funciones Cópula* (*FC*) se fundamentan en la dependencia entre *Q* y *V*, y definen la distribución bivariada por medio de las funciones univariadas marginales previamente adoptadas. El *enfoque práctico* adoptado utiliza *FC* de un solo parámetro de ajuste y selecciona la más adecuada a partir de la dependencia que muestra el registro conjunto de *Q* y *V* en el extremo derecho de su distribución empírica. Además, contrasta y ratifica la *FC* adoptada contra varias de uso común en los AFB. Lo anterior, por medio de los errores de ajuste entre las probabilidades bivariadas empíricas y teóricas. La búsqueda de las distribuciones marginales se realizó con base en el diagrama de cocientes *L* para adoptar las tres mejores y confrontarlas con las funciones Kappa y Wakeby de gran versatilidad. Se realizó el AFB de las 55 crecientes anuales registradas en la estación hidrométrica La Cuña de la Región Hidrológica No. 12-3 (río Santiago), México. Se definen cuatro periodos de retorno *conjuntos* de diseño y se detalla el cálculo de sus curvas de tipo AND. Por último, se citan varias conclusiones que destacan las ventajas del uso de las *FC* en los AFB de crecientes.

Palabras clave: familias de funciones Cópula, cociente tau de Kendall, coeficiente rho de Spearman, probabilidades empíricas conjuntas, dependencia en el extremo derecho, periodos de retorno conjuntos, eventos críticos.

Received: 16/11/2021

Accepted: 05/07/2022

Published online: 18/07/2022

Introduction

Generalities

In our country, the *Floods* caused by rising waters are one of the natural disasters that cause the highest economic, environmental and social damage to human settlements. Therefore, risk reduction through hydraulic works and the elaboration of damage mitigation plans, without such structures, have become vital for society. The planning and hydrological design of protection works such as dikes and retaining walls, rectifications, canalizations, bridges and some urban drainage works, are dimensioned based on the *Design Floods* (CD, by its acronym in Spanish). The CDs are maximum annual flows of the river, where the hydraulic works will be located, associated with low probabilities of being exceeded (Aldama, Ramírez, Aparicio, Mejía-Zermeño, & Ortega-Gil, 2006; Rao & Hamed, 2000; Meylan, Favre, & Musy, 2012).

The most reliable CD estimation is carried out through the univariate *Flood Frequency Analysis* (AFC, by its acronym in Spanish), which attempts to represent the maximum annual flow record of a river, with a probability distribution function (PDF) or probabilistic model, from where the sought *predictions* or CD are made. AFCs lead to reliable results when several conditions are met, among them, that the available record of flows is random, that the PDF is the most representative of such data and that its fit is made with an efficient and robust method (Hosking & Wallis, 1997; Meylan *et al*., 2012; Stedinger, 2017).

On the other hand, *reservoirs* or storages can be classified in a simple way, in large, medium and small. The large reservoirs have multiple purposes and multi–year operation, and for protection against floods among other uses. On the other hand, small reservoirs have a single purpose, generally irrigation, and therefore they store water from the rainy season for use in the following dry season. Whether large, medium or small, the design of reservoirs for hydrological safety requires the *hydrograph* of the CD estimation (Aldama, 2000; Aldama *et al.*, 2006; Gómez, Aparicio, & Patiño, 2010).

The hydrograph is the graphical representation of the flood evolution, with time on the abscissa and flow on the ordinate. Then, the area that it defines is the drained volume (*V*) and its amplitudes in time are the total duration (*D*) and in the ordinates the maximum flow (*Q*); whose time elapsed from the start is called peak time (*tp*). The hydrograph clearly establishes that floods are a *multivariate* hydrological phenomenon, whose elements are correlated (Favre, El Adlouni, Perreault, Thiémonge, & Bobée, 2004; Aldama *et al*., 2006; Gräler *et al*., 2013; Chowdhary & Singh, 2019).

Aldama (2000) proved that the reservoirs are not sensitive to the *tp value,* and since *Q* and *V* are correlated and *V* is correlated with *D*, the multivariate study of the floods could be reduced to a *bivariate analysis* of *Q* and *V*, to roughly define the CD hydrograph. The use of bivariate AFC initiated at the end of the last century with the works of Goel, Seth and Chandra (1998), and Yue, Ouarda, Bobée, Legendre and Bruneau (1999), who processed *Q* and *V* data of the annual floods, the first based on the bivariate Normal distribution, like Yue (1999), and the latter applying the so–called logistic model (Yue & Wang, 2004), with marginal Gumbel distributions.

The first bivariate AFCs were limited by the scarce availability of multivariate PDFs, whose first models used were extensions of univariate distributions, for example, the bivariate Normal, Log–normal, Exponential, Pareto and Gamma. Favre *et al*. (2004) indicate that this type of distributions have the following three disadvantages: (1) their marginal distributions belong to the same family; (2) the extension beyond the bivariate case, they are not simple or do not exist and (3) the fitting parameters of such marginals are also used to model the dependence between the random variables *Q* and *V*.

The limitations and disadvantages of the existing bivariate distributions are easily eliminated by applying the so–called *Copula Functions* (*CF*), which allow modeling the dependency between the variables *Q* and *V*, in addition to *joining* any type of univariate marginal distribution previously adopted, to build the new bivariate distribution (Shiau, Wang, & Tsai, 2006; Sraj, Bezak, & Brilly, 2015; Genest & Chebana, 2017; Zhang & Singh, 2019).

The first practical applications of *CF* were logically bivariate and began with the works of Favre *et al*. (2004), and Salvadori and De Michele (2004). Then there are the studies by Zhang and Singh (2006), and Genest and Favre (2007). Simultaneously, trivariate applications arrive with the analyzes of Grimaldi and Serinaldi (2006), and Zhang and Singh (2007). Currently, the applications of *CF* in the analysis of frequencies of the various hydrological data have been consolidated, as shown in the texts by Salvadori, De Michele, Kottegoda and Rosso (2007), Zhang and

Singh (2019), Chen and Guo (2019), and the chapters in Salvadori and De Michele (2013), and Chowdhary and Singh (2019).

In the present study, with the *practical CF approach,* an *operative similarity* with the univariate AFC is attempted, when performing the *bivariate* ones. To achieve the above, *CF*s with a single fit parameter are used, which have already been tested in bivariate AFCs and whose selection is made based on fit indicators and not through random simulation. In addition, at the beginning it is defined if the available sample of joint data of *Q* and *V*, show dependency in the right tail of their empirical distribution and if so, *CF*s that show such dependency are first searched and then tested.

Objectives

The present study comprises seven *objectives*: (1) briefly enunciate the operative concept of *CF*s; (2) expose each family of *CF*s that have been used in the AFCs; (3) describe the fit of the *CF*, based on the association measures; (4) explain the concept of right tail dependency, both empirical and theoretical, of each *CF* exposed; (5) select and test the *CF*s that are applicable in the described numerical example; (6) estimate and contrast the univariate and bivariate return periods and (7) conduct a numerical application with the 55 maximum flows and volumes of the annual floods registered at the hydrometric station La Cuña, of Hydrological Region No. 12–3 (Santiago River), Mexico.

OPEN ACCESS

Check for updates

Operative theory

The Copula Functions

Concept and definition

The essential characteristic of the *Copula Functions* (*CF*) consists in allowing to express a joint distribution of correlated random variables, as a function of their marginal distributions. So, a *CF* links or relates the univariate marginal distributions to form a multivariate distribution. A basic advantage of *CF*s for forming multivariate distributions is the fact that they separate the dependence effect between the random variables from the effects of the marginal distributions.

Due to the above, the construction of the multivariate distribution is reduced to studying the relationship between the dependent variables, when the univariate marginal distributions are known. In short, the use of *CF* offers complete freedom to adopt or select the best univariate marginal models that represent the data (Shiau *et al*., 2006; Kottegoda & Rosso, 2008; Meylan *et al*., 2012; Zhang & Singh, 2019; Chen & Guo, 2019).

As in this study *CF* will be applied in the bivariate frequency analysis of annual floods, the following definition refers to two correlated *X* and *Y* *random variables*, whose joint cumulative probability distribution function is $F_{X,Y}(x,y)$ with univariate marginal distributions $F_X(x)$ and $F_Y(\mathrm{y})$; then the *Copula Function C* exists and is such that:

$$F_{X,Y}(x,y) = C[F_X(x), F_Y(y)] \tag{1}$$

The Copula function C is unique if the random variables X and Y are continuous, and otherwise it is uniquely determined on the product of the rank of X by the rank of Y. Furthermore, a Copula is a bivariate distribution function with uniform marginals at the interval (0, 1).

Equation (1) defines the basic concept for the development of Copula functions and is known *Sklar's Theorem* exposed in 1959 (Sklar, 1959; Shiau *et al.*, 2006; Salvadori *et al*., 2007; Kottegoda & Rosso, 2008; Meylan *et al.*, 2012; Chen & Guo, 2019; Zhang & Singh, 2019).

Copula Families

The *Copula Functions* (*CF*) that have been developed have been classified into four classes (Meylan *et al*., 2012; Chowdhary & Singh, 2019): Archimedean, extreme value, elliptical and miscellaneous. They are also classified as one–parameter or multi–parameter copulas, depending on the extent or thoroughness to which the dependency structure between the variables Q and V is defined.

Salvadori *et al*. (2007) provide a comprehensive and useful summary of *CF*s that have been applied in the field of hydrology. In this regard, the following relationship has not included the families of Ali–Mikhail–Haq Copulas because it limits the dependence to a range of the Kendall tau ratio from −0.1817 to 0.3333 and the Farlie–Gumbel–Morgenstern family (FGM) which further limits it, to the range of ±0.2222.

Archimedean copulas have had broad application due to their simple construction, single parameter, wide range, and acceptance of both types of dependencies. Making $F_X(x) = u$, $F_Y(y) = v$ and θ the parameter that measures the dependence or association between u and v, we have the following two families of Archimedean Copulas, which accept negative and positive dependence (Shiau *et al*., 2006; Genest & Favre, 2007; Salvadori *et al.*, 2007; Zhang & Singh, 2019; Chen & Guo, 2019; Chowdhary & Singh, 2019).

1. Clayton. This Copula Function is called Cook–Johnson by Zhang and Singh (2006). Sraj *et al*. (2015) cite it with both names. Its equation and variation space of θ are (Nelsen, 2006):

$$C(u,v) = \max\left\{\left[u^{-\theta} + v^{-\theta} - 1\right]^{-1/\theta}, 0\right\} \; [-1,\infty) \setminus \{0\} \qquad (2)$$

for $\theta = 0$, $C(u,v) = \Pi(u,v) = uv$, is the product copula that uniquely defines independence. The relationship of θ to Kendall's tau ratio goes as follows:

$$\tau_n = \frac{\theta}{\theta+2} \qquad (3)$$

2. Frank. Its equation and variation space of θ are:

$$C(u,v) = -\frac{1}{\theta}\ln\left[1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right] \; (-\infty,\infty) \setminus \{0\} \qquad (4)$$

for $\theta = 1$, $C(u,v) = \Pi(u,v) = uv$, independence is defined. The relationship of θ with τ_n is as follows:

$$\tau_n = 1 + \frac{4}{\theta}[\mathrm{D}_1(\theta) - 1] \tag{5}$$

where $\mathrm{D}_1(\theta)$ is the Debye function of order 1, whose expression is:

$$D_1(\theta) = \frac{1}{\theta}\int_0^{\theta} \frac{t}{e^t - 1} dt \tag{6}$$

The previous equation was estimated with numerical integration, ratifying its results with the values tabulated by Stegun (1972).

Next, a family of Extreme Value Copulas is cited, which accepts only positive dependence.

3. Gumbel-Hougaard. Its equation and variation space of θ are:

$$C(u,v) = \exp\left\{-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{1/\theta}\right\} \; [1, \infty) \tag{7}$$

for $\theta = 1$, $C(u,v) = \Pi(u,v) = uv$, independence is defined. The relationship of θ to Kendall's tau ratio is as follows:

$$\tau_n = \frac{\theta - 1}{\theta} \tag{8}$$

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Lastly, two families of Miscellaneous class Copulas are cited.

4. Plackett accepts negative ($\theta < 1$) and positive ($\theta > 1$) dependence:

$$C(u,v) = \frac{1+(\theta-1)(u+v)}{2(\theta-1)} - \frac{\sqrt{[1+(\theta-1)(u+v)]^2 - 4uv\theta(\theta-1)}}{2(\theta-1)} \text{ con } 0 < \theta < \infty, \theta \neq 1 \quad (9)$$

For $\theta = 1$, $C(u,v) = \Pi(u,v) = uv$ independence is defined. The relationship of θ to Spearman's rho is as follows:

$$\rho_n = \frac{(\theta+1)}{(\theta-1)} - \frac{2\theta \ln(\theta)}{(\theta-1)^2} \quad (10)$$

5. Raftery accepts only positive dependence and θ varies between 0 and 1:

$$C(u,v) = min(u,v) + \frac{1-\theta}{1+\theta}(uv)^{1/(1-\theta)}\left\{1 - [max(u,v)]^{-(1+\theta)/(1-\theta)}\right\} \quad (11)$$

for $\theta = 0$, $C(u,v) = \Pi(u,v) = uv$, independence is defined. The relationship of θ to Kendall's tau ratio and to Spearman's rho are as follows:

$$\tau_n = \frac{2\theta}{3-\theta} \quad (12)$$

$$\rho_n = \frac{\theta(4-3\theta)}{(2-\theta)^2} \tag{13}$$

Families of Associated Copulas

The four classes of Copula families that exist and for the bivariate case, have three *associated families*, defined by the following strictly increasing or decreasing transformations, according to Theorem 2.4.4 in Nelsen (2006). Such families are (Salvadori & De Michele, 2004; Michiels & De Schepper, 2008; Chowdhary & Singh, 2019):

$$C'(u,v) = u - C(u,1-v) \tag{14}$$

$$C''(u,v) = v - C(1-u,v) \tag{15}$$

$$\hat{C}(u,v) = u + v - 1 + C(1-u,1-v) \tag{16}$$

The first two transformations change the nature of the agreement measures; that is, Kendall's tau and Spearman's rho, from negative to positive and vice versa. The third transformation is called the *Survival Copula*, because it is related to the joint distribution of the probabilities of exceedance, of the pair of random variables *u* and *v* whose Copula is *C*.

Dupuis (2007) highlighted the importance of the associated Clayton Copula, as it has a significant dependence on its right tail defined by the following Equation (17) and whose formula, according to equations (2) and (16), is Equation (18):

$$\lambda_U = 2^{-1/\theta} \tag{17}$$

$$\hat{C}(u,v) = u + v - 1 + \left[(1-u)^{-\theta} + (1-v)^{-\theta} - 1\right]^{-1/\theta} \tag{18}$$

Association measures

Concordance

Since the Copula Function (*CF*) characterizes the *dependence* between random variables, it is necessary to study the association measures, to count with a method that allows estimating its parameter θ. In general terms, a random variable is *concordant* with another, when its large values are associated with the large values of the other and the small values of one with the small values of the other. Some variables with direct linear correlation will be *concordant*, since when one increases the other also does. Variables with an inverse linear correlation will be *discordant*, since large values of one correspond to small values of the other.

Then, (x_1, y_1) and (x_2, y_2) are said to be concordant if $x_1 < x_2$ and $y_1 < y_2$, or $x_1 > x_2$ and $y_1 > y_2$. They are discordant if $x_1 < x_2$ and $y_1 > y_2$, or $x_1 > x_2$ and $y_1 < y_2$. This implies that the pairs $(x_1 - x_2)(y_1 - y_2) > 0$ are concordant (c) and discordant (d) when $(x_1 - x_2)(y_1 - y_2) < 0$ (Salvadori *et al*., 2007; Chowdhary & Singh, 2019).

A numerical association measure is a statistic that indicates the degree of dependence or association of the variables. For comparison purposes, such measures range from zero to +1 or −1, indicating perfect association positive at +1 or negative at −1. Kendall's tau quotient and Spearman's rho coefficient are two nonparametric measures that provide information on a special form of association or dependence, known as concordance (Salvadori *et al*., 2007).

Kendall's tau quotient

It measures the probability of having matching pairs, so it is the ratio from $c-d$ to $c+d$. Its expression to estimate it with bivariate data is (Zhang & Singh, 2006; Zhang & Singh, 2019):

$$\tau_n = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{signo}\left[\left(x_i - x_j\right)\left(y_i - y_j\right)\right] \quad (19)$$

in the above equation, n is the number of observations and the sign [·] is +1 if such pairs are concordant and −1 if they are discordant.

Genest and Favre (2007) present a test for the tau ratio, accepting that the null hypothesis H_0 indicates that X and Y are independent and then the statistic has an approximately Normal distribution with zero mean and variance $2(2n+5)/[9n(n-1)]$. Then, H_0 will be rejected with a confidence level $\alpha = 5$ % if:

$$\sqrt{\frac{9n(n-1)}{2(2n+5)}}|\tau_n| > Z_{\alpha/2} = 1.96 \quad (20)$$

Spearman's rho coefficient

It measures the correlation between rank pairs (R_i, S_i) of the continuous random variables X_i and Y_i. Therefore, it is equivalent to the Pearson correlation coefficient (r_{xy}). Its expression to estimate it in a bivariate record of size n is as follows (Chowdhary & Singh, 2019; Zhang & Singh, 2019):

$$\rho_n = \frac{12}{n(n+1)(n-1)} \sum_{i=1}^{n} R_i \cdot S_i - 3\frac{n+1}{n-1} \quad (21)$$

Genest and Favre (2007) also present a similar test for the rho coefficient, whose distribution is close to Normal with zero mean and variance $1/(n-1)$; therefore, H_0 is rejected with a level α=5%, if:

$$\sqrt{n-1} \cdot |\rho_n| > 1.96 \quad (22)$$

Estimation of the dependency parameter

The simplest method to estimate the parameter θ of the Copula Functions is similar to the method of moments and is based on the inversion of the equation that relates θ to Kendall's tau quotient or Spearman's rho coefficient. The other two available techniques are called: (1) maximum pseudo–likelihood method and (2) exact maximum likelihood method (Meylan *et al*., 2012; Chowdhary & Singh, 2019; Zhang & Singh, 2019; Chen & Guo, 2019). In equations (5), (10), (12) and (13) we proceed by trial and error to obtain θ; instead, in formulas (3) and (8) its value is cleared.

Estimation of empirical probabilities

The bivariate empirical probabilities were estimated based on the Gringorten formula (Equation (23)), applied by Yue *et al*. (1999), Chen and Guo (2019), and Zhang and Singh (2019). Such a formula is:

$$p = \frac{i - 0.44}{n + 0.12} \tag{23}$$

in which, *i* is the number of each data, when they are ordered progressively and *n* is the total number of them, or amplitude of the processed register. The previous expression was applied in the two–dimensional plane, with the data ordered progressively; the maximum flows (*Q*) in the rows and the volumes (*V*) in the columns. The plane formed is a square of *n* by *n* cells, with *n* cells on its main diagonal, when the order number of the row is equal to that of the column. Then each pair of annual data (*Q* and *V*) is located in the two-dimensional plane and the box defined by the intersection of the row and column is identified with the number *i* that corresponds to the historical year drawn.

When the *n* pairs of data are drawn, look for year 1 and define a rectangular or square area of minor values of *Q* and *V*, whose *count* of numbered cells within, is NM_1 or combinations of minor *Q* and *V*. Once the *n* values of NM_i have been calculated, the Gringorten graphical position formula is applied to estimate the joint or bivariate empirical probability:

$$F(x,y) = P(Q \leq q, V \leq v) = \frac{NM_i - 0.44}{n + 0.12} \tag{24}$$

Selection of the Copula function

A simple approach to selecting the Copula Function is based on the fit error statistics, by comparing the observed empirical probabilities (w_i^o) with the theoretical ones calculated (w_i^c) with the Copula Function being tested. The indicators applied are the standard mean error (*EME*,), the absolute mean error (*EMA*) and the maximum absolute error (*EAM*); their expressions are (Chowdhary & Singh, 2019; Chen & Guo, 2019):

$$EME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(w_i^o - w_i^c\right)^2} \tag{25}$$

$$EMA = \frac{1}{n}\sum_{i=1}^{n}|(w_i^o - w_i^c)| \tag{26}$$

$$EAM = \max_{i=1:n} |(w_i^o - w_i^c)| \tag{27}$$

Tail dependence of the *CF*

Generalities

Another criterion that is applied to select a *CF* is based on the magnitude of the dependency in joint distribution's upper tail, which affects the accuracy of the extreme predictions. The upper right tail dependency (λ_U) is the conditional probability that *Y* is greater than a certain percentile (*s*) of $F_Y(y)$, given that *X* is greater than that percentile in $F_X(x)$, as *s* approaches unity. The lower left tail dependency(λ_L) compares that *Y* is less than *X*, when *s* approaches zero (Chowdhary & Singh, 2019).

In relation to the exposed *CF*s, those of Frank and Plackett have insignificant dependencies in their extreme zones: therefore, λ_l = 0 and λ_U = 0. In contrast, the Gumbel-Hougaard Copula has significant dependence on the upper tail, equal to:

$$\lambda_U = 2 - 2^{1/\theta} \tag{28}$$

Clayton and Raftery Copulas have it in their lower tail and equal to:

$$\lambda_L = 2^{-1/\theta} \tag{29}$$

$$\lambda_L = \frac{2\theta}{\theta+1} \tag{30}$$

As already indicated (Equation (17)), the associated Clayton Copula has $\lambda_U = 2^{-1/\theta}$. Families of unexposed Copulas, with $\lambda_U > 0$, correspond to Galambos, Genest-Ghoudi, Hüsler-Reiss, Joe and *t*–Student, the latter belonging to the elliptic family (Salvadori *et al*., 2007; Chowdhary & Singh, 2019; Chen & Guo, 2019).

Dupuis (2007) tested six Copula families and found that their ability to estimate extreme events ranges from poor to good, in the following order: Clayton, Frank, Normal, *t*–Student, Gumbel-Hougaard, and Associated Clayton (*Survival Clayton*). Similar conclusions are reached by Poulin, Huard, Favre and Pugin (2007), when comparing the same six families of Copulas and the one called A12, which has significant right tail dependency.

Estimation of the observed dependency

To address the estimation of the upper tail dependency (λ_U) shown by the available data, the *Empirical Copula* must first be defined. As the *CF* that characterizes the dependence between the random variables *X* and *Y*; then the pair of ranges R_i and S_i coming from such variables, are the statistic that retains the greatest amount of information and its scaling with the factor $1/(n+1)$ generates a series of points in the unit square $[0,1]^2$, forming the domain of the Empirical Copula (Chowdhary & Singh, 2019), defined as follows:

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v\right) \tag{31}$$

In the above equation, 1(·) denotes a function of the random variables *u* and *v*, which are a continuously increasing transformation of *X* and *Y*, relative to the empirical probability integrals $F_n(X)$ and $F_n(Y)$, whose equations are:

$$U_i = \frac{\text{Rango}(X_i)}{n+1} = F_n(x_i)\ V_i = \frac{\text{Rango}(Y_i)}{n+1} = F_n(Y_i) \tag{32}$$

Equation (31) is a function from $(0, 1)^2$ to (0, 1) that has leaps of magnitude 1/*n* in each pair of the form of Equation (32), for $i \in \{1,2,\cdots, n\}$. Actually, Equation (31) is not a Copula, but a *subcopula* that serves to approximate Copulas (Nelsen, 2006).

Poulin *et al*. (2007) use the estimator proposed by Frahm, Junker and Schmidt (2005), which is based on a random sample obtained from the Empirical Copula, its expression is:

$$\lambda_U^{CFG} = 2 - 2\exp\left\{\frac{1}{n}\sum_{i=1}^{n} \ln\left[\sqrt{\ln\frac{1}{U_i}\cdot\ln\frac{1}{V_i}}/\ln\left(\frac{1}{\max(U_i,V_i)^2}\right)\right]\right\} \tag{33}$$

This estimator accepts that the *CF* can be approximated by one of the extreme values class and has the advantage of not requiring a threshold value for its estimation.

Ratification of the selected Copula Function

This is the most important stage in the process of practical application of the *CF*, since it is verified that such a model faithfully reproduces the observed joint probabilities (Equation (24)). Yue (2000) indicates that a simple and practical way of representing the empirical and theoretical joint probabilities consists of taking the first one to the abscissa axis and the second to the ordinate axis; logically, in such a graph, each pair of data defines a point that coincides with or moves away from the line at 45°.

Yue and Rasmussen (2002) apply the Kolmogorov–Smirnov test with a significance level (α) of 5 %, to accept or reject the maximum absolute difference (*dma*) between the joint empirical and theoretical probabilities. To evaluate the statistic (D_n) of the test, the expression proposed by Meylan *et al.* (2012), for α = 5 % this is:

$$D_n = \frac{1.358}{\sqrt{n}} \tag{34}$$

n is the number of data. If the *dma* is less than D_n, the adopted *CF* is ratified.

Bivariate return periods

The first *bivariate return period* of the event (X, Y) is defined under the OR condition, which indicates that the limits x or y, or both can be exceeded and then, the classical return period equation or inverse probability of exceedance will be (Shiau *et al*., 2006; Vogel & Castellarin, 2017):

$$T_{XY} = \frac{1}{P(X>x \vee Y>y)} = \frac{1}{1-F_{X,Y}(x,y)} = \frac{1}{1-C[F_X(x),F_Y(y)]} \tag{35}$$

where, $C[F_X(x), F_Y(y)]$ is the selected *CF*.

The second *bivariate return period* of the event (X, Y) is associated with the case in which both limits will be exceeded ($X > x$, $Y > y$) or AND condition, its equation is (Shiau *et al*., 2006; Vogel & Castellarin, 2017):

$$T'_{XY} = \frac{1}{P(X>x \wedge Y>y)} = \frac{1}{F'_{X,Y}(x,y)} = \frac{1}{1-F_X(x)-F_Y(y)+C[F_X(x),F_Y(y)]} \tag{36}$$

Aldama (2000) obtains the expression $F'_{X,Y}(x,y)$ of the bivariate probability of exceedance through a simple and logical reasoning of probabilities applied in the Cartesian plane. Instead, Yue and Rasmussen (2002) use the Cartesian plane to define the bivariate event (X, Y) conceptually, which can occur in any of the four quadrants.

The relationship between bivariate and univariate return periods is as follows (Yue & Rasmussen, 2002; Shiau *et al*., 2006; Vogel & Castellarin, 2017):

$$T_{XY} \leq \min[T_X, T_Y] \leq \max[T_X, T_Y] \leq T'_{XY} \tag{37}$$

being:

$$T_X = \frac{1}{F'_X(x)} = \frac{1}{1-F_X(x)} \quad (38)$$

$$T_Y = \frac{1}{F'_Y(y)} = \frac{1}{1-F_Y(y)} \quad (39)$$

Critical or design events

Volpi and Fiori (2012) highlight that the plot of the AND-type joint return period —shown as Figure 1— presents a severe inconsistency as it contains, in a bivariate context, univariate critical thresholds. Due to the above, such a graph is considered to be made up of two portions, the two designated *simple* (*naive part*) and the *correct one* (*proper part*). The straight parts are the tails or straight lines asymptotes to the curved part. The probability of an event occurrence or pair of *Q* and *V*, is variable in the curved part and decreases along the straight part, although all the values define the same joint return period. In short, the pairs of values of the asymptote lines have low probabilities of occurrence and therefore should not be included in the analysis of the search for critical or severe peaks (*Q* and *V*). For practical purposes, the extreme points of the curved part can be defined according to their empirical distribution or close to the beginning of the asymptotic lines.

Tecnología y Ciencias del Agua
OPEN ACCESS
Check for updates

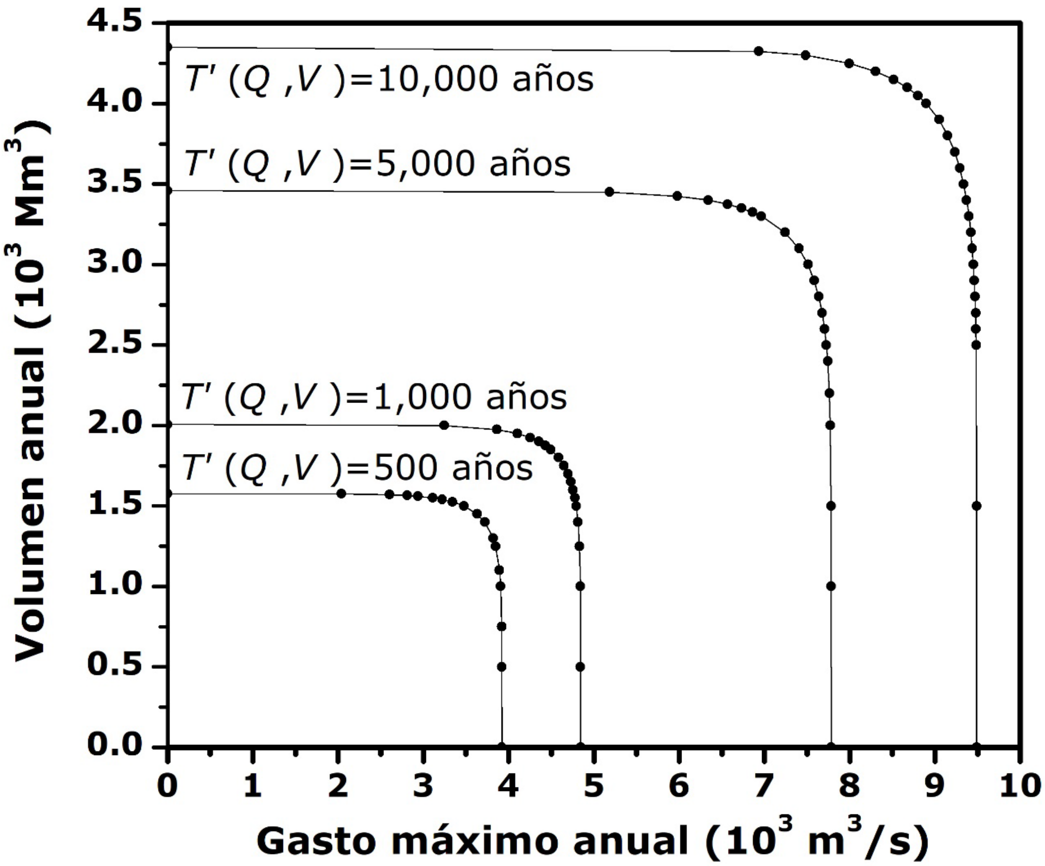

Figure 1. Graphs or the four design joint return periods T'_{XY} ontained with the Gumbel–Hougaard *CF*, from the floods of the hydrological station La Cuña, Mexico.

Wald-Wolfowitz test

This non–parametric test has been exposed and applied by Bobée and Ashkar (1991), Rao and Hamed (2000) and Meylan *et al*. (2012) to verify *independence* and *stationarity* in records of maximum annual flows (X_i).

Due to the above, it was proposed to apply the test to the records of maximum flows and annual volumes, which must be random samples.

Selection of marginal distributions

Eight applied probability distributions

The search approach for the ideal marginal distributions took into account the statistical characteristics of the hydrological data to be processed, maximum flow and volumes of the annual floods. The later, through the quotients L of asymmetry (t_3) and kurtosis (t_4), which allow defining in the L–Moment Ratio Diagram of Hosking and Wallis (1997), the best distributions, due to their less proximity to the five curves that are displayed on such a graph. These distributions have three fit parameters and they are: Generalized Logistic (LOG), General Extreme Values (GVE), Log–Normal (LGN), Pearson type III (PE3) and Generalized Pareto (PAG). The PE3 curve allows testing the Log–Pearson type III (LP3) distribution.

In addition, two probabilistic models were applied that have shown great versatility and universality to represent series of extreme hydrological data, due to the fact that they have four and five fit parameters; such distributions are the Kappa (Hosking, 1994; Kjeldsen, Ahn, & Prosdocimi, 2017; Campos-Aranda, 2023) and the Wakeby (Hosking & Wallis, 1997).

The LP3 distribution was the only one that was applied with the method of moments in the logarithmic (WRC, 1977) and real (Bobée, 1975) domains, selecting the one with the best fit. The remaining seven were adjusted to the records of maximum flows and annual volume of floods, through the method of moments L, according to procedures exposed by Hosking and Wallis (1997).

L moments of the sample

The L–moments are linear combinations of the probability–weighted (b_r) moments, defined by Greenwood, Landwehr, Matalas, and Wallis (1979); therefore, they are robust to scattered sample values. Its calculation begins by ordering the available series of maximum flows (x_i) and annual volumes (y_i) of the floods from smallest to largest ($x_1 \leq x_2 \leq \cdots \leq x_n$) and estimating the b_r with the following equation (Hosking & Wallis, 1997; Rao & Hamed, 2000; Asquith, 2011; Stedinger, 2017):

$$b_r = \frac{1}{n} \sum_{i=r+1}^{n} \frac{(i-1)(i-2)\cdots(i-r)}{(n-1)(n-2)\cdots(n-r)} x_i \qquad (40)$$

In the previous expression, the order number r varies from 0 to 3 and n is the number of data of the annual series. It follows that b_0 is equal to the arithmetic mean. The moments L of the sample (l) and their respective *quotients* (t) of similarity with the coefficients of variation, asymmetry and kurtosis are:

$$l_1 = b_0 \qquad (41)$$

$$l_2 = 2 \cdot b_1 - b_0 \qquad (42)$$

$$l_3 = 6 \cdot b_2 - 6 \cdot b_1 + b_0 \quad (43)$$

$$l_4 = 20 \cdot b_3 - 30 \cdot b_2 + 12 \cdot b_1 - b_0 \quad (44)$$

$$t_2 = l_2/l_1 \quad (45)$$

$$t_3 = l_3/l_2 \quad (46)$$

$$t_4 = l_4/l_2 \quad (47)$$

L-Moments Ratio Diagram

It has t_3 on the abscissa axis and t_4 on the ordinate axis. The three-parameter fit PDFs are curved lines with the following polynomial-type equations (Hosking & Wallis, 1997):

Generalized logistic (LOG):

$$t_4^{\mathrm{LOG}} = 0.16667 + 0.83333 \cdot t_3^2 \quad (48)$$

Generalized Pareto (PAG):

$$t_4^{\mathrm{PAG}} = 0.20196 \cdot t_3 + 0.95924 \cdot t_3^2 - 0.20096 \cdot t_3^3 + 0.04061 \cdot t_3^4 \quad (49)$$

Log-normal (LGN):

$$t_4^{\mathrm{LGN}} = 0.12282 + 0.77518 \cdot t_3^2 + 0.12279 \cdot t_3^4 - 0.13638 \cdot t_3^6 + 0.11368 \cdot t_3^8 \quad (50)$$

Pearson type III (PT3):

$$t_4^{\mathrm{PT3}} = 0.12240 + 0.30115 \cdot t_3^2 + 0.95812 \cdot t_3^4 - 0.57488 \cdot t_3^6 + 0.19383 \cdot t_3^8 \quad (51)$$

General extreme value (GVE):

$$t_4^{\mathrm{GVE}} = 0.10701 + 0.11090 \cdot t_3 + 0.84838 \cdot t_3^2 - 0.06669 \cdot t_3^3 + SF \quad (52)$$

being $SF = 0.00567 \cdot t_3^4 - 0.04208 \cdot t_3^5 + 0.03763 \cdot t_3^6$

Using the natural logarithms of the data in equations (40) to (47), the logarithmic L ratios are obtained, and Equation (51) can then be used to estimate the Log-Pearson type III distribution.

Absolute distance

One of the recent approaches for selecting the best PDF of a data series consists of taking the sample values (t_3 and t_4) to the ratio diagram L and defining their proximity to any of the curves, to obtain the best probabilistic model. To avoid subjectivity in the selection of the PDF, it has been proposed to evaluate the Absolute Distance (*DA*) with the following expression (Yue & Hashino, 2007):

$$DA = t_4\left(t_3^{obs}\right) - t_4^{obs} \qquad (53)$$

where, t_3^{obs} and t_4^{obs} are the asymmetry and kurtosis ratios L of the analyzed series and $t_4\left(t_3^{obs}\right)$ is the theoretical value of the kurtosis quotient L calculated with each PDF (equations (48) to (52)), for the observed value of the asymmetry ratio L. A PDF with the smallest *DA* value is the best for the processed data.

Fit errors

The first criteria applied for the selection of the best PDF to some available data or series were the so–called *fit errors* (Kite, 1977; Willmott & Matsuura, 2005; Chai & Draxler, 2014). This criterion will be applied after selecting the three best PDFs in the L–moment ratio diagram, according to their minimum absolute distance and after applying the Kappa and Wakeby distributions.

Changing in equations (25) and (26), the probabilities observed by the ordered data of the analyzed series (x_i or y_i) and the probabilities calculated by the estimated values with the PDF that is tested or contrasted, the standard error of fit (*EEA*) and the mean absolute error

(*EAM*). The values that are estimated ($\hat{x}_i$ or $\hat{y}_i$) are searched for the same probability of non–exceedance, assigned to the data by Gringorten's empirical formula (Equation (23)).

Data to be processed

The hydrometric station La Cuña with code 12054 and a basin area of 19097 km^2, is located in the Verde River of Hydrological Region No. 12–3 (Santiago River), Mexico. Gomez *et al*. (2010) expose the record of maximum flow in m^3/s and annual volumes in millions of m^3 (Mm^3), from the period from 1947 to 2004, with 55 pairs of data (n), since the years 1983, 1984 and 1985 do not have volume information. Such data is shown in Table 1.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Table 1. Maximum flows, volumes and their bivariate order numbers of the annual floods of the hydrometric station La Cuña, Mexico (Gómez *et al*., 2010).

Year	*Q* (m^3/s)	*V* (Mm^3)	NM_i	Año	*Q* (m^3/s)	*V* (Mm^3)	NM_i
1947	784.0	146.80	36	1975	622.1	249.07	40
1948	736.8	155.12	37	1976	1374.0	527.96	52
1949	510.0	111.40	28	1977	439.7	111.77	28
1950	461.0	94.06	24	1978	280.2	66.23	15
1951	411.0	111.55	27	1979	267.2	45.80	9
1952	326.0	70.82	19	1980	287.3	99.60	20
1953	349.8	144.75	28	1981	280.7	28.70	5
1954	130.4	23.22	3	1982	156.5	35.37	4
1955	690.0	203.31	39	1986	698.2	193.51	38
1956	266.0	106.76	17	1987	184.7	55.39	7
1957	199.0	45.92	9	1988	595.2	242.21	38
1958	690.0	188.71	37	1989	110.2	42.49	3
1959	340.6	47.91	12	1990	523.9	248.07	37
1960	249.6	91.58	14	1991	1636.3	443.30	52
1961	350.0	130.68	28	1992	1168.0	172.49	40
1962	317.0	51.27	12	1993	295.0	96.50	20
1963	732.6	127.90	33	1994	212.8	53.55	10
1964	265.1	82.75	15	1995	367.4	114.61	27
1965	743.6	295.34	46	1996	144.6	57.43	5
1966	463.9	202.90	35	1997	78.4	16.55	1
1967	1474.9	598.38	53	1998	261.9	66.17	13
1968	323.0	118.25	25	1999	196.3	41.15	7
1969	160.4	32.22	4	2000	46.8	18.62	1
1970	763.8	187.75	39	2001	313.8	75.78	18
1971	578.0	166.61	36	2002	319.6	153.79	25
1972	191.8	26.39	4	2003	621.1	326.28	40
1973	2440.0	920.30	55	2004	824.5	384.45	50
1974	238.4	66.66	13	*Med*	340.6	111.40	–

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Results and their discussion

Search of the marginal distributions

Verification of the randomness

First, the randomness of the records to be processed was verified, based on the Wald–Wolfowitz Test; the statistic led to values of 0.284 and 0.213, for the flows and volumes in Table 1.

Calculated absolutes distances

Table 2 has concentrated the results of equations (48) to (52), highlighting the three best PDFs of each data record in Table 1.

Table 2. Order numbers of the best PDF for the indicated records of the hydrometric station La Cuña, Mexico, according to absolute distance in the diagram of L ratios (Hosking & Wallis, 1997).

Annual maximum flows (m^3/s)		**Annual volumes (Mm^3)**	
t_3 = 0.38713	ln t_3 = 0.01249	t_3 = 0.43326	ln t_3 = 0.01382
t_4 = 0.25523	ln t_4 = 0.15953	t_4 = 0.28466	ln t_4 = 0.12466
LOG distribution	0.0363 (3)	LOG distribution	0.0384 (5)
GVE distribution	0.0179 (2)	GVE distribution	0.0240 (3)
LGN distribution	0.0139 (1)	LGN distribution	0.0128 (2)
PT3 distribution	0.0680 (6)	PT3 distribution	0.0755 (6)
PAG distribution	0.0440 (5)	PAG distribution	0.0320 (4)
LP3 distribution	0.0371 (4)	LP3 distribution	0.0022 (1)

Distribution of maximum annual flows

Table 3 shows the adjustment errors and predictions (m^3/s) obtained with the three best distributions and the Kappa and Wakeby, applied to the record of maximum flow in Table 1. It is observed that the two best distributions, the LGN and the GVE, report the lowest and similar fitting errors; but their predictions are quite different in the four high return periods. On the other hand, the fit errors of the Kappa distribution are almost identical to those of the LGN, but with intermediate predictions and for this reason, it was adopted.

Table 3. Fit errors and predictions (m^3/s) of the three best PDF, the Wakeby and the Kappa in the record of maximum annual flows of the hydrometric station La Cuña, Mexico.

PDF	*EEA* (m^3/s)	*EAM* (m^3/s)	Return periods in years					
			50	100	500	1 000	5 000	10 000
LOG (3)	61.3	39.4	1 816	2 393	4 504	5 902	11 036	14 441
GVE (2)	*55.7*	*35.8*	1 838	2 372	4 160	5 254	8 926	11 172
LGN (1)	55.8	36.5	1 839	2 293	3 592	4 270	6 167	7 139
KAPPA	55.7	36.7	1 835	2 335	3 920	4 844	7 784	9 489
WAKEBY	35.2	55.7	1 844	2 304	3 646	4 367	6 469	7 599

The location (u_1), scale (a_1) and shape (k_1 and h_1) parameters of the selected Kappa distribution are: 258.7462, 228.381, −0.2685394 and 0.2888472, whose expression is:

$$F(x) = \left\{1 - h_1 \left[1 - \frac{k_1(x-u_1)}{a_1}\right]^{1/k_1}\right\}^{1/h_1} \quad (54)$$

Distribution of annual volumes

In Table 4, similar to the previous one for the annual volumes of Table 1. It is observed that the best distribution, the LP3, leads to the lowest fit errors; but its predictions are quite lower than those of the GVE. In contrast, the Kappa distribution leads to low fit errors and its predictions are intermediate, therefore, it was adopted. In this case, the Wakeby distribution ratifies the selection.

Table 4. Fit errors and predictions (Mm^3) of the three best PDF, Wakeby and Kappa in the record of annual volumes of the station hydrometric La Cuña, Mexico.

PDF	*EEA* (Mm^3)	*EAM* (Mm^3)	Return periods in years					
			50	**100**	**500**	**1 000**	**5 000**	**10 000**
GVE (3)	16.6	9.0	662	888	1 702	2 234	4 156	5 411
LGN (2)	14.5	7.6	670	863	1 439	1 752	2 661	3 142
LP3 (1)	*10.5*	*6.3*	694	907	1 574	1 952	3 101	3 736
KAPPA	15.5	8.1	663	873	1 576	2 007	3 459	4 350
WAKEBY	15.7	8.3	664	872	1 561	1 978	3 358	4 193

The location (u_2), scale (a_2) and shape (k_2 and h_2) parameters of the adopted Kappa distribution are: 60.39858, 78.31082, −0.3155518 and 0.4287021, whose expression is:

$$F(y) = \left\{1 - h_2\left[1 - \frac{k_2(y-u_2)}{a_2}\right]^{1/k_2}\right\}^{1/h_2} \quad (55)$$

Selection and ratification of the Copula Function

The bivariate data processing in Table 1 led to the following three indicators of association: r_{xy} = 0.9302, τ_n = 0.7199 and ρ_n = 0.9088. Equation (20), with n = 55 and the quoted tau, gives a value of 7.76; therefore, Kendall's quotient is significant. Equation (22) generates a value of 6.68 and so Spearman's coefficient is also significant. Equation (33) provides a value for the observed right tail dependence of $\lambda_U^{CFG} =$ 0.782.

On the other hand, Table 5 shows the statistical indicators of fit that were obtained by applying the six exposed *CF* of Clayton, Frank, Gumbel–Hougaard (GH), Plackett, Raftery and associated Clayton. In equations (25) to (27), the empirical bivariate probabilities were estimated with Equation (24) and the theoretical ones with equations (2), (4), (7), (9), (11) and (18).

Table 5. Statistical indicators of the Copula Functions fit indicated in the annual floods of the station hydrometric La Cuña, Mexico.

Copula	**θ**	***EME***	***EAM***	**DP**	**DN**	**MDP**	**MDN**	λ_U
Clayton (C.)	5.1394	0.0245	0.0189	28	27	0.0714	*−0.0424*	0.000
Frank	12.386	*0.0227*	*0.0169*	29	26	0.0581	−0.0561	0.000
GH	3.5697	0.0255	0.0192	28	27	0.0541	−0.0676	0.786
Plackett	75.040	0.0247	0.0187	23	32	0.0495	−0.0647	0.000
Raftery	0.7940	0.0250	0.0193	27	28	0.0746	−0.0457	0.000
C. Asociada	5.1394	0.0288	0.0225	29	26	*0.0477*	−0.0788	0.874

Meaning of the new acronyms:

DP, DN: Number of positive and negative differences.

MDP, MDN: Maximum positive and negative differences.

As the *Q* and *V* values of the annual floods recorded at the hydrometric station La Cuña, show a magnitude of λ_U^{CFG} of 0.782 for the dependence observed in its right tail, there is no difficulty in selecting the best *CF* that of Gumbel–Hougaard, for the data in Table 1, because it reports a value for the significant dependence on its right end, practically the same as that observed or shown by the data.

When the λ_U^{CFG} value is much lower than the Gumbel–Hougaard *CF* one, it is possible to resort to applying the elliptic class *CF* called *t*–Student; which, according to the results of Dupuis (2007) and Poulin *et al*. (2007) has an equal dependency on both queues, but of lesser value.

In Table 5 the lowest statistical indicators of the *EME* and *EAM*, as well as the maximum differences between the bivariate probabilities, have been shown in italics. It is observed that the adopted Gumbel–Hougaard *CF* has fit indicators located between the smallest obtained with the Frank *CF* and the largest generated by the Associated Clayton *CF*.

Table 6, Table 7 and Table 8 show a part of the bivariate probabilities of non-exceedance, empirical observed (w_i^o) and theoretical calculated (w_i^c) with the Frank, Gumbel-Hougaard and Associated Clayton *CF*s. The maximum positive and negative differences are also shaded.

Table 6. Joint non–exceedance probabilities and their differences for a part of the floods of La Cuña station, Mexico; with Frank's *CF*.

No.	w_i^o	w_i^c	Differences	No.	w_i^o	w_i^c	Differences
1	0.6451	0.6477	−0.0026	30	0.9354	0.9344	0.0011
5	0.4819	0.4872	−0.0053	35	0.0827	0.0783	0.0044
10	0.3004	0.3139	−0.0134	40	0.0464	0.0375	0.0089
12	0.6633	0.7194	−0.0561	45	0.1734	0.1704	0.0226
20	0.6270	0.6096	0.0174	50	0.1190	0.1169	0.0021
25	0.6451	0.6609	−0.0157	55	0.8991	0.8410	0.0581

Table 7. Joint non–exceedance probabilities and their differences for a part of the floods of the station La Cuña, Mexico; with the Gumbel–Hougaard *CF*.

No.	w_i^o	w_i^c	Differences	No.	w_i^o	w_i^c	Differences
1	0.6451	0.6512	−0.0061	30	0.9354	0.9520	−0.0166
5	0.4819	0.4765	0.0053	35	0.0827	0.0777	0.0050
12	0.6633	0.7309	−0.0676	40	0.0464	0.0401	0.0063
15	0.5000	0.4459	0.0541	45	0.1734	0.1624	0.0111
20	0.6270	0.6116	0.0154	50	0.1190	0.1143	0.0048
25	0.6451	0.6661	−0.0210	55	0.8991	0.8557	0.0435

Table 8. Joint non-exceedance probabilities and their differences for a part of the floods of La Cuña station, Mexico; with Associated Clayton *CF*.

No.	w_i^o	w_i^c	Differences	No.	w_i^o	w_i^c	Differences
1	0.6451	0.6540	−0.0088	30	0.9354	0.9548	−0.0194
5	0.4819	0.4863	−0.0044	35	0.0827	0.0727	0.0100
12	0.6633	0.7421	−0.0788	40	0.0464	0.0286	0.0178
15	0.5000	0.4523	0.0477	45	0.1734	0.1454	0.0280
20	0.6270	0.6170	0.0100	50	0.1190	0.0944	0.0246
25	0.6451	0.6798	−0.0346	55	0.8991	0.8578	0.0414

On the other hand, Equation (34) defines D_n = 0.1831 and since the absolute maximum differences of Table 6, Table 7 and Table 8 are 0.0581, 0.0676 and 0.0788, the Kolmogorov-Smirnov test confirms the three contrasted *CF*. The correlation coefficients (r_{xy}) between the 55 empirical probabilities and the theoretical ones, estimated with the Frank, Gumbel-Hougaard and Associated Clayton *CF*s, are the following: 0.9968, 0.9964 and 0.9962; therefore, they correspond, in order of magnitude, with the maximum differences observed.

The graphic contrast between both probabilities, to ratify the adoption of the Gumbel-Hougaard *CF* is shown in Figure 2 for the complete data in Table 7.

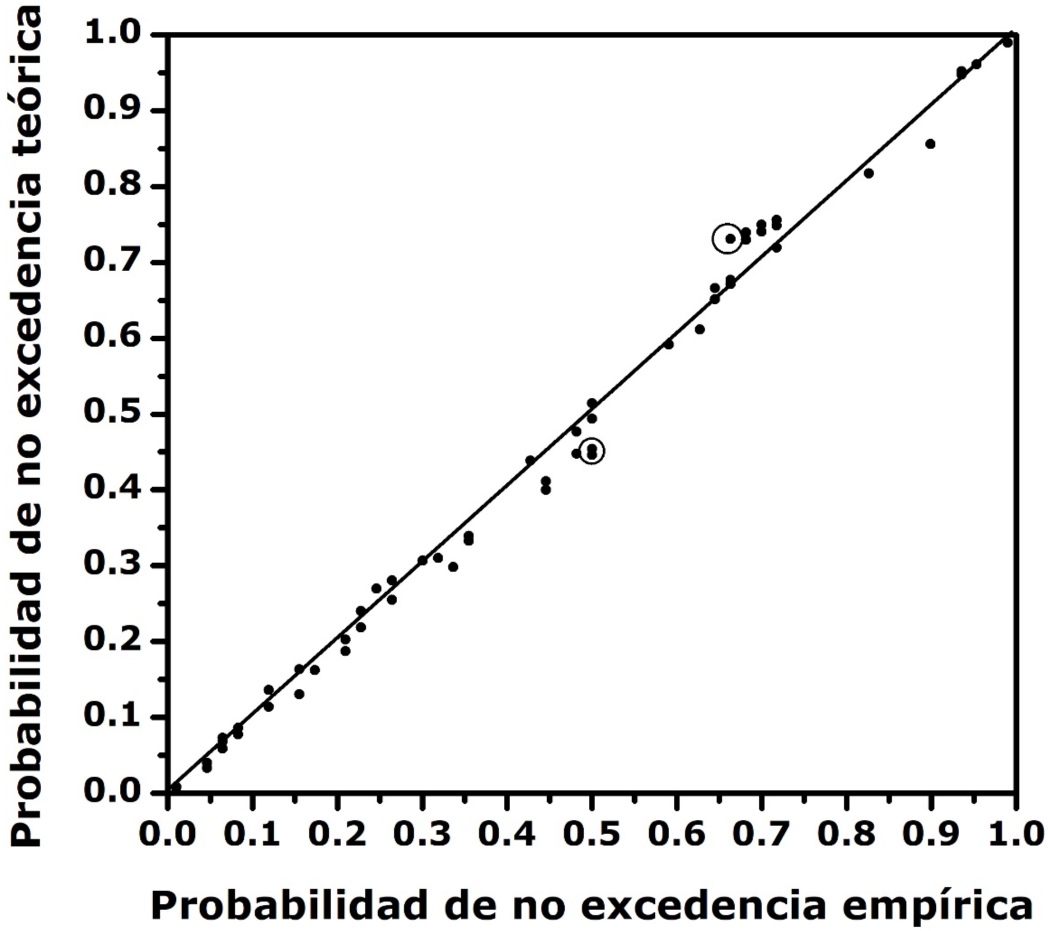

Figure 2. Graphical contrast of joint probabilities of the peak flow and volumes of the floods of the La Cuña station, Mexico; with the Gumbel–Hougaard *CF*.

OPEN ACCESS

Bivariate return period T'_{XY} plots

Considering that a reservoir is going to be built in the riverbed and near the site of the La Cuña gauging station, which can be small, medium or large, depending on the capacity assigned to it, then bivariate return periods are required of type AND of 500, 1 000, 5 000 and 10 000 years (Campos-Aranda, 2008). Such magnitudes are estimated with Equation (36). For which, volumes and peak flows are arbitrarily selected, to obtain their marginal non–exceedance probabilities (equations (54) and (55)) and joint (Equation (7)) with the Gumbel–Hougaard *CF*. Table 9 shows results to define the four graphs of Figure 1.

Table 9. Pairs of maximum flow and annual volume used to define the graphs of the joint return period type AND with the *CF* of Gumbel-Hougaard, in the floods of the hydrometric station La Cuña, Mexico.

T'_{XY} **500 years**		T'_{XY} **1 000 years**		T'_{XY} **5 000 years**		T'_{XY} **10 000 years**	
V **Mm³**	***Q*** **m³/s**	***V*** **Mm³**	***Q*** **m³/s**	***V*** **Mm³**	***Q*** **m³/s**	***V*** **Mm³**	***Q*** **m³/s**
0	3920	0	4844	0	7784	0	9489
500	3919	500	4843	1000	7782	1500	9489
750	3918	1000	4841	1500	7782	2500	9485
1000	3906	1250	4830	2000	7773	2600	9480
1100	3892	1400	4813	2200	7763	2700	9478
1250	3847	1500	4791	2400	7742	2800	9469
1300	3818	1550	4776	2500	7727	2900	9459
1400	3720	1600	4756	2600	7705	3000	9449
1450	3630	1650	4729	2700	7678	3100	9436
1500	3475	1700	4695	2800	7638	3200	9421
1525	3342	1750	4649	2900	7585	3300	9400
1540	3222	1800	4586	3000	7511	3400	9370
1550	3108	1850	4495	3100	7405	3500	9335
1560	2938	1875	4433	3200	7241	3600	9290
1565	2809	1900	4355	3300	6963	3700	9232
1570	2604	1925	4252	3325	6860	3800	9148
1575	2039	1950	4104	3350	6731	3900	9049
1576	0	1975	3862	3375	6566	4000	8898
		2000	3244	3400	6340	4050	8800
		2007	0	3425	5981	4100	8674
				3450	5183	4150	8515
				3459	0	4200	8306
						4250	7996
						4300	7482
						4325	6935
						4350	0

Contrast of bivariate return periods

Table 10 shows the calculations made to verify Equation (37). It is observed and verified that in all cases the T_{XY} is less than the univariate return periods (T) and on the contrary, the T'_{XY} is always greater.

Table 10. Univariate and bivariate return periods estimated with the Gumbel–Hougaard *CF* in the floods of the hydrometric station La Cuña, Mexico.

$T = T_X = T_Y$	Q_T	V_T	$C[F_X(x),F_Y(y)]$	T_{XY}	T'_{XY}
50	1835	663	0.9757031	41.2	63.3
100	2335	873	0.9878718	82.5	127.1
500	3920	1576	0.9975724	411.9	636.3
1000	4844	2007	0.9987859	823.7	1272.5
5000	7783	3459	0.9997571	4116.9	6364.6
10000	9489	4350	0.9998783	8216.9	12700.4

Selection of design events

In Figure 1 or in Table 9, infinite pairs of Q and V can be selected, which satisfy the joint design return period and which are defined as a *subgroup of critical pairs*, as they are within the curved portion of each graph of the T'_{XY}, outside the asymptote lines (Volpi & Fiori, 2012).

The combinations of peak flow and volume that have the same *design bivariate return period*, establish floods or *hydrographs* that will produce different effects in the reservoir that is designed or revised; adopting for security, the one that generates the most critical or severe conditions. To form each design hydrograph, there are theoretical and empirical methods (Aldama, 2000; Aldama *et al*., 2006; Campos-Aranda, 2008; Xiao, Guo, Liu, & Fang, 2008; Gómez *et al*., 2010; Gräler *et al*., 2013).

Contrast with results of the bivariate GVE

Campos-Aranda (2022) processed the records in Table 1, based on the bivariate GVE distribution, adjusted by means of maximum likelihood; therefore, its marginal functions are of the GVE type and its fit parameters are involved in the best fit achieved. Table 11, in rows 1 and 2, shows the joint univariate predictions that were obtained.

Table 11. Predictions obtained with the bivariate GVE and Kappa distributions as marginal of the floods reports from the hydrometric station La Cuña, Mexico.

Row	Rec.	PDF	Univariate return periods, in years					
			50	100	500	1 000	5 000	10 000
1	*Q*	GVEb	1550	1958	3254	4009	6418	7823
2	*V*	GVEb	568	739	1317	1673	2877	3619
3	*Q*	Kappa	1835	2335	3920	4844	7784	9489
4	*V*	Kappa	663	873	1576	2007	3459	4350

On the other hand, rows 3 and 4 of Table 11 indicate the predictions adopted in Tables 3 and 4 above. It can be observed, for the indicated return periods, that all the new pairs of predictions of Q and V are larger, this affects the definition of the critical pairs in Figure 1, which are now larger, compared to those obtained when applying the bivariate GVE distribution.

Conclusions

The fundamental advantage of engaging the *Copula Functions* (*CF*) in the analysis of bivariate increasing frequencies consists in the ability to easily construct the *joint* probability distribution, based on the same or different marginal univariate distributions, even from mixed populations; prior estimation of the dependence between the random variables: maximum flow (Q) and volume (V) of the annual floods.

The *practical approach to applying CF* requires a careful selection of marginal distributions, and for this reason, the L–ratio diagram was used to search for the three best probability distributions and then contrast them, based on their fit errors and magnitude of its predictions and thus adopt the most representative ones for the available Q and V records. In this contrast, two highly versatile distributions were included, as they have four and five fit parameters, the Kappa and Wakeby models.

This *practical approach to CF application*, uses single fit parameter Copulas (θ), which is estimated based on Kendall's tau quotient or

Spearman's rho coefficient, which are calculated with the joint registration of *Q* and *V*. Such an approach first estimates λ_U^{CFG} or right tail dependency of the available record. Then a *CF* is searched that reproduces such value of λ_U^{CFG}. *CF*s that do not have significant right tail dependency, such as those of Clayton, Frank, Plackett and Raftery, are also applied in order to compare and judge the quality of the fit of the previously adopted *CF*.

The numerical application described, in the 55 annual data of *Q* and *V* of the floods registered in the hydrometric station La Cuña of the Hydrological Region No. 12-3, of Mexico; showed a reliable reproduction of the empirical and theoretical bivariate probabilities, through the Gumbel-Hougaard *CF*, with a linear correlation coefficient of 0.9964. On the other hand, in Figure 1, relative to the *design joint return periods* of the AND type, infinite pairs of critical *Q* and *V* can be defined, since they are in the curved region of each graph.

Acknowledgment

The authors thank anonymous referees F, G and H for their observations and suggested corrections, which allowed correcting important theoretical and practical omissions. Their comments helped to improve the general writing and to highlight the contributions of the proposed approach for the application of the Copula functions, in the bivariate frequency analyzes of annual floods.

References

Aldama, A. A. (2000). Hidrología de avenidas. Conferencia Enzo Levi 1998. *Ingeniería Hidráulica en México,* 15(3), 5-46.

Aldama, A. A., Ramírez, A. I., Aparicio, J., Mejía-Zermeño, R., & Ortega-Gil, G. E. (2006). *Seguridad hidrológica de las presas en México.* Jiutepec, México: Instituto Mexicano de Tecnología del Agua.

Asquith, W. H. (2011). Chapter 6. L–moments. In: *Distributional analysis with L–moments statistics using the R environment for statistical computing* (pp. 87-122). Lubbock, USA: Copyright by William H. Asquith.

Bobée, B. (1975). The Log-Pearson type 3 distribution and its application to Hydrology. *Water Resources Research*, 11(5), 681-689.

Bobée, B., & Ashkar, F. (1991). Chapter 1. Data requirements for hydrologic frequency analysis. In. *The gamma family and derived distributions applied in hydrology* (pp. 1-12). Littleton, USA: Water Resources Publications.

Campos-Aranda, D. F. (2008). Procedimiento para revisión (sin hidrometría) de la seguridad hidrológica de presas pequeñas para riego. *Agrociencia*, 42(5), 551-563.

Campos-Aranda, D. F. (noviembre-diciembre, 2022). Aplicación de la distribución GVE bivariada en el análisis de frecuencias conjunto de crecientes. *Tecnología y ciencias del agua*, 13(6), 534-602. DOI: 10.24850/j–tyca-13-6-11

Campos-Aranda, D. F. (enero-febrero, 2023). Análisis de frecuencias comparativo con momentos L entre la distribución Kappa y seis de aplicación generalizada. *Tecnología y ciencias del agua*, 14(1), 432-469. DOI: 10.24850/j-tyca-14-01-05

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247-1250. DOI: 10.5194/gmd–7–1247–2014

Chen, L., & Guo, S. (2019). Chapter 2. Copula theory. Chapter 3. Copula–based flood frequency analysis. In: *Copulas and its application in hydrology and water resources* (pp. 13-38, 39-71). Gateway East, Singapore: Springer.

Chowdhary, H., & Singh, V. P. (2019). Chapter 11. multivariate frequency distributions in hydrology. In: Teegavarapu, R. S. V., Salas, J. D., & Stedinger, J. R. (eds.). *Statistical analysis of hydrologic variables* (pp. 407-489). Reston, USA: American Society of Civil Engineers.

Dupuis, D. J. (2007). Using Copulas in hydrology: Benefits, cautions, and issues. *Journal of Hydrologic Engineering*, 12(4), 381-393. DOI: 10.1061/(ASCE)1084–0699(2007)12:4(381)

Favre, A. C., El Adlouni, S., Perreault, L., Thiémonge, N., & Bobée, B. (2004). Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, 40(1), 1-12. DOI: 10.1029/2003WR002456

Frahm, G., Junker, M., & Schmidt, R. (2005). Estimating the tail–dependence coefficient: Properties and pitfalls. *Insurance: Mathematics and Economics*, 37(1), 80-100. DOI: 10.1016/j–insmatheco.2005.05.008

Genest, C., & Favre, A. C. (2007). Everything you always wanted to know about Copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4), 347-368. DOI: 10.1061/(ASCE)1084–0699(2007)12:4(347)

Genest, C., & Chebana, F. (2017). Copula modeling in hydrologic frequency analysis. In: Singh, V. P. (ed.). *Handbook of applied hydrology* (2nd ed.) (pp. 30.1-30.10). New York, USA: McGraw-Hill Education.

Goel, N. K., Seth, S. M., & Chandra, S. (1998). Multivariate modeling of flood flows. *Journal of Hydraulic Engineering*, 124(2), 146-155.

Gómez, J. F., Aparicio, M., & Patiño, C. (2010). Capítulo 6. Análisis de frecuencias bivariado para la estimación de avenidas de diseño. En: *Manual de análisis de frecuencias en hidrología* (pp. 106-127). Jiutepec, México: Instituto Mexicano de Tecnología del Agua.

Gräler, B., van den Berg, M. J., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., & Verhoest, N. E. C. (2013). Multivariate return periods in hydrology: A critical and practical review focusing on synthetic design hydrograph estimation. *Hydrology and Earth System Sciences*, 17(4), 1281-1296. DOI: 10.5194/hess–17–1281–2013

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., & Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research*, 15(5), 1049-1054.

Grimaldi, S., & Serinaldi, F. (2006). Design hyetograph analysis with 3–copula function. *Hydrological Sciences Journal*, 51(2), 223-238. DOI: 10.1623/hysj.51.2.223

Hosking, J. R. M. (1994). The four–parameter Kappa distribution. *IBM Journal of Research and Development*, 38(3), 251-258.

Hosking, J. R., & Wallis, J. R. (1997). Appendix: *L*–moments for some specific distributions. In: *Regional frequency analysis. An approach based on L–moments*. (pp. 191-209). Cambridge, UK: Cambridge University Press.

Kite, G. W. (1977). Chapter 12. Comparison of frequency distributions. In: *Frequency and risk analyses in hydrology* (pp. 156-168). Fort Collins, USA: Water Resources Publications.

Kjeldsen, T. R., Ahn, H., & Prosdocimi, L. (2017). On the use de a four-parameter kappa distribution in regional frequency analysis. *Hydrological Sciences Journal*, 62(9), 1354-1363. DOI: 10.1080/02626667.2017.1335400

Kottegoda, N. T., & Rosso, R. (2008). Theme 3.5. Copulas. In: *Applied statistics for civil and environmental engineers* (2nd ed.) (pp. 154-157). Oxford, UK: Blackwell Publishing.

Meylan, P., Favre, A. C., & Musy, A. (2012). Chapter 1. Introduction. Chapter 3. Selecting and checking data series. Theme 9.2. Multivariate frequency analysis using Copulas. In: *Predictive hydrology. A frequency analysis approach* (pp. 1-13, 29-70, 164-176). Boca Raton, USA: CRC Press.

Michiels, F., & De Schepper, A. (2008). A Copula test space model. How to avoid the wrong copula choice. *Kybernetika*, 44(6), 864-878.

Nelsen, R. B. (2006). *An introduction to Copulas*. *Lecture notes in statistics 139* (2nd ed.). New York, USA: Springer.

Poulin, A., Huard, D., Favre, A. C., & Pugin, S. (2007). Importance of tail dependence in bivariate frequency analysis. *Journal of Hydrologic Engineering,* 12(4), 394-403. DOI: 10.1061/(ASCE)1084–0699(2007)12:4(394)

Rao, A. R., & Hamed, K. H. (2000). Chapter 1. Introduction. Chapter 3. Probability weighted moments and L–moments. In: *Flood frequency analysis* (pp. 1-21, 53-72). Boca Raton, USA: CRC Press.

Salvadori, G., & De Michele, C. (2004). Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resources Research*, 40(W12511), 1-17. DOI: 10.1029/2004WR003133

Salvadori, G., & De Michele, C. (2013). Multivariate extreme value methods. In: AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., & Sorooshian, S. (eds.). *Extremes in a changing climate* (pp. 115-162). London, UK: Springer.

Salvadori, G., De Michele, C., Kottegoda, N. T., & Rosso, R. (2007). Appendix B. Dependence. Appendix C: Families fo Copulas (pp. 219-232, 233-269). In: *Extremes in nature. An approach using Copulas*. Dordrecht, The Netherlands: Springer.

Shiau, J. T., Wang, H. Y., & Tsai, C. T. (2006). Bivariate frequency analysis of floods using Copulas. *Journal of the American Water Resources Association*, 42(6), 1549-1564.

Sklar, A. (1959). Functions de repartition à *n* dimensions et leur marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8, 229-231.

Sraj, M., Bezak, N., & Brilly, M. (2015). Bivariate flood frequency analysis using the copula function. A case study of the Litija station on the Sava River. *Hydrological Processes*, 29(2), 225-238. DOI: 10.1002/hyp.10145

Stedinger, J. R. (2017). Flood frequency analysis. In: Singh, V. P. (ed.). *Handbook of applied hydrology* (2nd ed.) (pp. 76.1-76.8). New York, USA: McGraw-Hill Education.

Stegun, I. A. (1972). Chapter 27. Miscellaneous functions. In: Abramowitz, M., & Stegun, I. A. (eds.). *Handbook of Mathematical Functions* (pp. 997-1010). New York, USA: Dover Publications.

Vogel, R. M., & Castellarin, A. (2017). Risk, reliability, and return periods and hydrologic design. In: Singh, V. P. (ed.). *Handbook of applied hydrology* (2nd ed.). (pp. 78.1-78.10). New York, USA: McGraw-Hill Education.

Volpi, E., & Fiori, A. (2012). Design event selection in bivariate hydrological frequency analysis. *Hydrological Sciences Journal*, 57(8), 1506-1515. DOI: 10.1080/02626667.2012.726357

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79-82. DOI: 10.3354/cr030079

WRC, Water Resources Council. (1977). *Guidelines for determining flood flow frequency* (revised edition) (Bulletin #17A) Hydrology Committee. Washington, DC, USA: Water Resources Council.

Xiao, Y., Guo, S., Liu, P., & Fang, B. (2008). A new design flood hydrograph method based on bivariate joint distribution. In: *Hydrological sciences for managing water resources in the Asian developing world* (pp. 75-82). UK: IAHS Publication.

Yue, S. (1999). Applying bivariate Normal distribution to flood frequency analysis. *Water International*, 24(3), 248-254.

Yue, S., Ouarda, T. B. M. J., Bobée, B., Legendre, P., & Bruneau, P. (1999). The Gumbel mixed model for flood frequency analysis. *Journal of Hydrology*, 226(1-2), 88-100.

Yue, S. (2000). Joint probability distribution of annual maximum storm peaks and amounts as represented by daily rainfalls. *Hydrological Sciences Journal*, 45(2), 315-326. DOI: 10.1080/02626660009492327

Yue, S., & Rasmussen, P. (2002). Bivariate frequency analysis: Discussion of some useful concepts in hydrological application. *Hydrological Processes*, 16(14), 2881-2898. DOI:10.1002/hyp.1185

Yue, S., & Hashino, M. (2007). Probability distribution of annual, seasonal and monthly precipitation in Japan. *Hydrological Sciences Journal*, 52(5), 863-877. DOI: 10.1623/hysj.52.5.863

Yue, S., & Wang, C. Y. (2004). A comparison of two bivariate extreme value distributions. *Stochastic Environmental Research and Risk Assessment*, 18(2), 61-66. DOI. 10.1007/s00477–003–0124–x

Zhang, L., & Singh, V. P. (2006). Bivariate flood frequency analysis using the Copula method. *Journal of Hydrologic Engineering*, 11(2), 150-164. DOI: 10.1061/(ASCE)1084–0699(2006)11:2(150)

Zhang, L., & Singh, V. P. (2007). Trivariate flood frequency analysis using the Gumbel-Hougaard Copula. *Journal of Hydrologic Engineering*, 12(4), 431-439. DOI: 10.1061/(ASCE)1084–0699(2007)12:4(431)

Zhang, L., & Singh, V. P. (2019). Chapter 3. Copulas and their properties. Chapter 11. Flood frequency analysis. In: *Copulas and their applications in water resources engineering* (pp. 62-122, 396-440). Cambridge, UK: Cambridge University Press.