Journal of Hydrology 540 (2016) 469–487

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Probabilistic modeling of flood characterizations with parametric and minimum information pair-copula model

Alireza Daneshkhah [a], Renji Remesan [b,d], Omid Chatrabgoun [c], Ian P. Holman [b,*]

[a] Warwick Centre for Predictive Modelling, School of Engineering, The University of Warwick, CV4 7AL, UK
[b] Cranfield Water Science Institute, Cranfield University, Cranfield MK43 0AL, UK
[c] Department of Statistics, Faculty of Mathematical Sciences and Statistics, Malayer University, Malayer, Iran
[d] Centre for Ecology and Hydrology, UK

ARTICLE INFO

Article history:
Received 19 December 2015
Received in revised form 26 April 2016
Accepted 20 June 2016
Available online 21 June 2016
This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Fateh Chebana, Associate Editor

Keywords:
Flood frequency analysis
Flood hazard characterization
Return period
D-vine model
Minimum information pair-copula model
Himalaya (India)

ABSTRACT

This paper highlights the usefulness of the minimum information and parametric pair-copula construction (PCC) to model the joint distribution of flood event properties. Both of these models outperform other standard multivariate copula in modeling multivariate flood data that exhibiting complex patterns of dependence, particularly in the tails. In particular, the minimum information pair-copula model shows greater flexibility and produces better approximation of the joint probability density and corresponding measures have capability for effective hazard assessments. The study demonstrates that any multivariate density can be approximated to any degree of desired precision using minimum information pair-copula model and can be practically used for probabilistic flood hazard assessment.

1. Introduction

Operational planning and design of flood defence systems, irrigation water management systems and hydroelectric schemes requires accurate estimation of flood hazard and/or specified exceedance probabilities of river flow. Flood frequency analysis (FFA) is traditionally used to assess flood hazard with an assumption that annual maximum floods are a stationary, independent, identically distributed random process (Kidson and Richards, 2005). Conventionally, FFA is performed using either 'Block (annual) maxima' or 'peaks over threshold (POT)' methods on partial series of data (Hosking et al., 1985). Although, the univariate FFA is widely used in hydrology, many studies have highlighted its unreliability and suggested that univariate frequency analysis methods cannot sufficiently characterize inflow hydrographs or reduce uncertainty in flood analysis (Cunnane, 1988; Bobee and Rasmussen, 1994). Indeed, most hydrologic events are multivariate in nature and defined by a group of correlated random variables (e.g. flood peak, volume, and duration). Therefore, multivariate FFA would be more suitable to describe the uncertainties associated with these events.

By recognizing the limitations of univariate FFA, multivariate flood frequency analysis methods were developed. Many early multivariate studies focused on bivariate normal distribution to perform flood analysis with later researchers considering multivariate Gaussian (Krstanovic and Singh, 1987), gamma (Yue et al., 2001; Nadarajah and Gupta, 2006), exponential (Choulakian et al., 1990), Gumbel (Bacchi et al., 1994) and other distributions. Durrans et al. (2003) applied Pearson Type III distribution to perform joint frequency analysis. Yue and Wang (2004) developed Gumbel mixed and Gumbel logistic models; and compared their performances in flood analysis. However, distribution-based traditional univariate and multivariate analysis methods have mathematical weaknesses that limit their potential for practical applications. These flaws include that (a) the mathematical formulation is complicated when the number of variables are high (b) it is not possible to distinguish marginal and joint behavior of studied variables, (c) marginal distributions are of same type, or normal, or independent and (d) joint distributions hold validity in limited space (Song and Singh, 2010).

* Corresponding author.
E-mail address: I.holman@cranfield.ac.uk (I.P. Holman).

http://dx.doi.org/10.1016/j.jhydrol.2016.06.044
0022-1694/

Recently, the application of copulas in hydrology, as well as in other earth and environmental sciences, has received increasing attention. Copulas are efficient mathematical tools which are capable of combining several univariate marginal cumulative distribution functions into their joint cumulative distribution function (Sklar, 1959). The copula application in hydrology largely began after De Michele and Salvadori (2003) highlighted the suitability of the Frank copula for the joint distribution of negatively associated storm intensity and storm duration data, whilst Grimaldi and Serinaldi (2006a,b) applied several trivariate copulas for determining joint and conditional distributions among design hyetograph variables. Recent works on analysis of multivariate hydrological extreme events (Salvadori and De Michele, 2006, 2007, 2010) have popularized copulas as a tool for extreme value applications in rainfall (Évin and Favre, 2008; Wang et al., 2010; Zhang et al., 2012), floods (Zhang and Singh, 2007; Chowdhary et al., 2011), and droughts (Shiau, 2006; Song and Singh, 2010; Zhang et al., 2012; Ma et al., 2013). A brief review of the application of copula in various engineering and science fields can found in Genest and Favre (2007). They have also identified plausible Copula candidates for flood peak flow and volume data in FFA. Dupuis (2007) used 5 copulas (Normal, Student-*t*, Frank, Clayton, Gumbel, and associated Clayton) and warned about ignoring the tail dependence characteristics of flood data. Their analysis showed that the Frank copula performed relatively well in comparison to other approaches. Karmakar and Simonovic (2009) identified that the generalized hyperbolic copula is better at obtaining pair-wise joint distributions among flood peak flow, volume and duration. Leonard et al. (2008) used copula for bivariate analysis of rainfall and stream flow extremes accounting for seasonal and climatic partitions. Huard et al. (2006) and Silva and Lopes (2008) used Bayesian based copula selection method for estimating marginal and dependence parameters.

The set of higher dimensional copulas proposed in the literature is limited and is not rich enough to model all possible mutual dependencies among all variables (see Kurowicka and Cooke, 2006 for details). In addition, Aas et al. (2009) show that the multivariate copulas (in particular, multivariate *t*-copula) cannot efficiently be used to model multivariate data exhibiting complex patterns of dependence in the tails (which are common in analyzing the extreme events). These limits of the multivariate copula motivate Joe (1997) and Bedford and Cooke (2001, 2002), to propose a far efficient new way of constructing complex multivariate highly dependent models called vine or pair-copula (Aas et al., 2009). The principle behind this method is to model dependency using simple local building blocks based on conditional independence, known as the pair-copulae. The modeling scheme is then based on a decomposition of a multivariate density into a cascade of pair copulae, applied on the original variables and on their conditional and unconditional distribution functions. There is a growing literature of using the pair-copula models in the different real world applications including finance, economic and insurance studies (Aas et al., 2009; Czado and Min, 2010; Min and Czado, 2010; Bauer et al., 2012; Dissmann et al., 2013; Brechmann et al., 2014), risk management (Brechmann and Czado, 2013; Brechmann et al., 2014), energy (Czado et al., 2011), hydrological drought frequency analysis (Song and Singh, 2010). In addition to the above references which give an idea of recent advancements happening on pair-copula applications in the different fields. Recently, Gyasi-Agyei and Melching (2012) have used PCC to model the dependence structure of storm event properties using hourly rainfall data from Cook County, Illinois, USA. Song and Kang (2011) demonstrated pair-copula based trivariate discharge modeling considering variables like flood duration, severity, and severity peak. Vernieuwe et al. (2015) constructed a continuous rainfall model based on vine copulas and they compared the vine model with ensemble synthetic rainfall series. In a similar study, Xiong et al. (2014) have developed an annual rainfall-runoff model using the canonical vine copula derivation approach and employed in 40 watersheds in two large basins in China.

The multivariate copula models have also been used in different applications in the domain of spatial statistics. Bárdossy (2006) was one of the first who applied copulas in a geostatistical context. Gräler and Pebesma (2011) propose a more efficient approach for modeling spatial data (including extremes) using the vine copula model. One of the advantages of their approach was its flexibility in choosing appropriate parametric copula families through bivariate spatial copulas. Gräler (2014) extends this methodology further by adding several spatial trees at the foundation of the selected vine. These additional spatial trees add valuable information on the dependence of the higher order neighbors leading to an improved model of the spatial data. The predictive accuracy of the spatial vine copula outperforms other spatial multivariate copulas, including spatial Gaussian copula which used to be a very common method (as suggested by Bárdossy, 2006).

In a more relevant study, Gräler et al. (2013) use the vine copula model to construct a joint probability distribution for the flood variables, including peak discharge, duration, and volume. However, their main purpose of modeling the dependencies between the flood variables using the vine copula model and other multivariate copula models was to estimate design events for a given return period and to discuss their differences in a practical application. They concluded that the vine copula approach is the way to go for constructing flexible multivariate distribution functions for the same reasons mentioned above and discussed in further details in the next section.

It should be noticed that the use of a copula to model dependency is simply a translation of one difficult problem into another. By using (parametric) copula, the difficulty of specifying the full joint distribution will be reduced to the difficulty of specifying the copula. The advantage is the technical one that copulas are normalized to have support on the unit square and uniform marginals. As many authors restrict the copulas to a particular parametric class (Gaussian, multivariate t, etc.) the potential flexibility of the copula approach is not realized in practice. Bedford et al. (2015) proposed a so-called minimum informative pair-copula using the vine structure to approximate any given multivariate copula to any required degree of approximation, and to show how this can be operationalized for use in practice. The only technical assumptions required are that the multivariate copula density under study is continuous and is non-zero. This approach, by contrast to the parametric methods mentioned above, allows a lot of flexibility in copula specification. This new approach involves the use of minimum information copulas that can be specified to any required degree of precision based on the data available and are then stacked together to produce the multivariate copula and density function.

Based on the above discussion, we extend the parametric vine copula model (Gräler et al., 2013) in modeling flood characterizations with the minimum information pair-copula model. This model shows greater flexibility and produces better approximation of the joint probability density and corresponding measures have better capability for effective hazard assessments. We also present an approximation method at which any multivariate density can be approximated to any degree of desired precision using minimum information pair-copula model and practically be applied for assessing probabilistic flood hazard. We finally illustrate the methods described above by modeling the flood event properties of the Himalayan River Beas. Himalayan rivers in north India are highly influenced by both the monsoon and intra-annual release of stored water in the snow cover and glacier ice of the Himalayas and its nearby foothills. The response of Himalayan rivers to precipitation and temperature is highly variable as it depends on the extent of snow cover and volume of snowpack in their catchment,

snow melt behavior, and unpredictability in monsoon patterns in the region. Most Himalayan river basins have witnessed serious economic, agricultural and social impacts due to extreme hydrological events, such as floods, storms and droughts. To the best of our knowledge we could not find a proper study focusing on flood frequency analysis and flood hazard assessment. One of the main challenges of modeling flood characterizations of Himalayan rivers is the data scarcity. This problem is another reason of choosing the minimum information PCC for modeling the flood variables in the presence of limited data. More precisely, in this study we apply three methods to analyze the flood event data: common multivariate copulas; parametric PCC; and the minimum information pair-copula model and compare their performances in modeling flood characterizations.

The remainder of the paper is organized as follows. A brief description of copulas and their mathematical formulation is given in Section 2. We then introduce the parametric pair-copula model and how it can be fitted to the data. In Section 3, we introduce the non-informative pair-copula model and show that how the minimum information copulas can be used in approximating a multivariate density distribution to any required degree of precision based on the observed data. In Section 4, we present and analysis the results associated of fitting the copula models discussed in this paper to the stream flow data of the Himalayan River Beas. In this section, we also compare these models using various statistics and graphical tools to show the benefit of the pair-copula models (particularly, minimum information copulas) in uncertainty modeling in Risk analysis. Section 5 is dedicated in using the selected models in flood risk management by computing various measures which are widely used in risk analysis of the flood data. We finally conclude our study in Section 6.

2. Multivariate dependence modeling using vine constructions

In many areas of applied science and particularly in modeling flood data and other hydrological data, it is necessary to model multiple uncertain quantities using an appropriate multivariate distribution. The Bayesian networks and multivariate copulas are widely used for this purpose. However, the Bayesian networks are more popular for general decision support settings, but their usage is limited to the multivariate normal and multinomial distributions for the continuous and discrete variables, respectively. In the recent years, the multivariate copulas have been attracted by the users in other disciplines, particularly for modeling financial data, risk and uncertainty analysis associated with the extreme events (including flood risk assessment), due to their flexibilities in dependency modeling of multiple data consists of both discrete and continuous data.

There is a growing literature on the use of the copulas to model dependencies of multiple uncertain quantities (Joe, 1997; Nelsen, 2006). In particular, these models have been widely used in multivariate analysis of hydrological data (Genest et al., 2007, 2009; Grimaldi and Serinaldi, 2006; Salvadori et al., 2007; Serinaldi and Grimaldi, 2007; Yan et al., 2007; Zhang and Singh, 2007; Song and Singh, 2010). A copula is a joint distribution on the unit square which enables us to uniquely determine a joint distribution of n random variables by specifying their marginal distributions and an appropriate copula function. As a more formal definition, a copula is any multivariate distribution, C, with uniformly distributed marginals $U(0,1)$ on the unit square $[0,1]$, where the corresponding joint distribution function F of $(x_1,\ldots,x_n)$ can be written as

$$F(x_1,\ldots,x_n) = C(u_1,\ldots,u_n;\boldsymbol{\theta}), \tag{1}$$

where $C:[0,1]^n \to [0,1]$ is an appropriate copula distribution function, $\boldsymbol{\theta}$ denotes to the association parameters, $u_i = F_i(x_i),\ i = 1,\ldots,n$ and $F_1,\ldots,F_n$ are marginal distribution functions of $X_1,\ldots,X_n$, respectively. This formula can be constructively used to define F in terms of given a copula function C and marginals $F_1,\ldots,F_n$ under reasonable conditions. For example, the 'Gaussian copula' as a widely used copula in many applications, can be obtained from the joint normal distribution and parameterized by the correlation matrix. The details of constructing the multivariate Archimedean copulas (e.g., Clayton, Gumbel and Frank) and multivariate canonical copula (t-student and Gaussian) can be found in Nelsen (2006) and Joe (1997).

In addition, the joint density function $f(x_1,\ldots,x_n)$, given that F_i and C are differentiable, can be also presented as

$$f(x_1,\ldots,x_n) = f_1(x_1)\times\cdots\times f_n(x_n)\times c(F_1(x_1),\ldots,F_n(x_n)) \tag{2}$$

where $f_i(x_i)$ is the density function corresponding to $F_i(x_i)$, and $c = \partial^n C/(\partial F_1\ldots\partial F_n)$ is called the copula density function f (Nelsen, 2006).

Building/approximating a high-dimensional copula is generally considered as a difficult task. For instance, Venter et al. (2007) reported that most multivariate copula densities get increasingly difficult to approximate as the dimension increases. Joe (1997) (and later Kurowicka and Cooke, 2006) highlighted this point that however certain copula families, including the multivariate Gaussian, t-student, the exchangeable multivariate Archimedean copula or the nested Archimedean constructions, exhibit a huge improvement in modeling multivariate data, but they are still rather limited, computationally not tractable, and not rich enough to model all possible mutual dependencies among the variables.

In 2002, Bedford and Cooke introduce a probabilistic construction of multivariate distributions based on a flexible graphical model called *vine* which was also later called *pair-copula* construction (*PCC*) in Aas et al. (2009). This flexible structure allows for a free specification of (at least) $n(n-1)/2$ bivariate copulas between n given variables. In other words, a vine on n variables is a nested set of trees, where the edges of the tree j are the nodes of the tree $j+1$ (for $j = 1,\ldots,n-2$), and each tree has the maximum number of edges. A vine in which two edges in tree j are joined by an edge in tree $j+1$ only if these edges share a common node, $j = 1,\ldots,n-2$, is called *regular vine*. The formal definition of vine and regular vine can also be found in Kurowicka and Cooke (2006).

The class of regular vines is generally quite broad and consists of many possible pair-copula decompositions. Among them, the *canonical* vine and the *D-vine* are two special regular vines where each one gives a specific way of decomposing a multivariate density function. The D-vine is more widely used in practice. In a D-vine with n variables, no node in any tree is connected to more than two edges. The canonical vine is more useful when a particular variable is known to be a key variable that controls interactions in the data. It is then recommended to place this variable at the root of this vine. As a result, each tree T_j in canonical vine has a unique node that is connected to $n-j$ edges. We briefly introduce these two vines and how they can be used to model multivariate data by a simple example (further details can be found in Aas et al., 2009; Kurowicka and Cooke, 2006).

As an example, a D-vine structure first will be used to model multivariate density function associated with the following random variables (X_1,X_2,X_3) with the given marginal densities f_1,f_2,f_3, respectively. A D-vine structure is normally selected based on the association measures between variables (see also Section 2.1). The one that is chosen for these variables is shown in Fig. 1 with the following joint density decomposition

$$f(x_1,x_2,x_3) = \left(\prod_{i=1}^{3} f_i(x_i)\right)\times c_{12}(F(x_1),F(x_2))c_{23}(F(x_2),F(x_3)) \times c_{13|2}(F(x_1|x_2),F(x_3|x_2)) \tag{3}$$

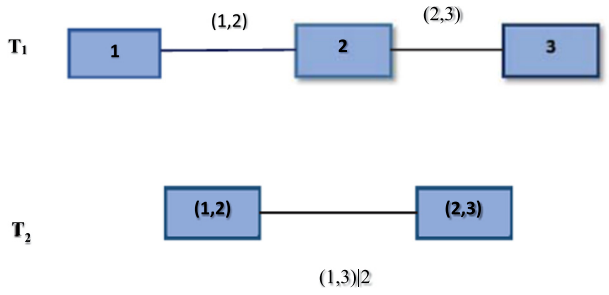

Fig. 1. A D-vine structure with 3 variables, 2 trees and 4 edges, where each edge may be associated with a pair-copula.

where $c_{ij}(F(x_i), F(x_j))$ denote the bivariate copula between x_i and x_j, and $c_{ik|j}(F(x_i|x_j), F(x_k|x_j))$ denote the bivariate copula fitted to the conditional distributions $F(x_i|x_j)$ and $F(x_k|x_j)$ (see Appendix A, for the details of this decomposition).

Similarly, a multivariate density decomposition can be derived based on a canonical vine structure. The canonical vine structure is very dependent on the root node. In other words, in a canonical vine, each tree T_j has a unique node that is connected to $n-j$ edges. The D-vine structure shown in Fig. 1 can be converted to a canonical vine with the same density factorisation given in (3) if x_2 is considered as the root node (see also Aas et al., 2009).

The above decomposition of the joint density gives us a constructive approach to build a multivariate distribution given a vine structure: If we make choices of marginal densities and copulae then the above formula will give us a multivariate density. In other words, we associate a vine distribution to a vine by specifying a copula to each edge of the first tree T_1 (as shown in Fig. 1) and a family of conditional copulas for the conditional variables given the conditioning variables in the second tree T_2.

On of the main objectives of this paper is to address the advantages of the vine or PCC models over the standard multivariate copula in modeling hydrology data, and particularly flood risk management. One of the advantages of the vine models is that various bivariate copulas can be used in fitting a copula to any pair of variables instead of fitting a fixed multivariate copula to all variables. As a specific example, the multivariate t-copula is widely being used in finance and hydrology, where multivariate data exhibit complex patterns of dependencies in the tails. The issue with the multivariate t-copula is that only a single degree of freedom parameter which drives the tail dependence of all pairs of variables is used. This problem can be dealt with using the vine model. Aas et al. (2009) demonstrate the superiority of a D-vine copula with bivariate t-copulas over a single multivariate t-copula approach. We adopt, extend and address superiority and flexibility of this approach to model the flood event data over the current alternative. In the next section, we address the methods that can be used to fit a parametric pair-copula model to the data.

2.1. Estimation methods for pair-copula models

Fitting a vine or PCC model to the data involves a number of steps. The first step is to identify an appropriate vine tree structure. Such a structure may either be given by the data itself or has to be selected manually or through expert knowledge. This step is quite similar to structure learning in graphical models[1] with this difference that a vine structure can be easily determined in terms of the association measures or appropriate graphical tools. For a given vine structure, adequate copulas have to be selected, and in the next step estimated. This step is shared with fitting a single multivariate parametric copula to the data.

There are different tools that can be used to determine an appropriate vine tree structure: scatter, chi (χ), Kendall and λ plots, the correlation coefficients, and the independent tests. It should be noticed that some of these tools can also be used to select a suitable parametric bivariate copula family. The definitions and further details of these tools can be found in Schirmacher and Schirmacher (2008).

In the next step of analyzing the data, we need to select suitable bivariate copula models describing the dependencies between the variables illustrated in terms of the selected vine structure. There are various graphical and analytical methods available to select an appropriate copula for the underlying data. Among the graphical methods, Kendall's plot (K-plot) and the chi-plot are more appropriate to select the best bivariate copula models directly (see Genest and Favre (2007) and references therein for more details, and Section 4 for an illustration). The λ-function (proposed by Genest and Rivest, 1993) is another useful analytical measure to select a suitable bivariate copula. This function provides a characteristic for each copula family and is defined as follows

$$\lambda(\upsilon, \boldsymbol{\theta}) := \upsilon - K(\upsilon, \boldsymbol{\theta}), \tag{4}$$

where $K(\upsilon, \boldsymbol{\theta}) = Pr(C(U, V; \boldsymbol{\theta}) \leqslant \upsilon)$ is Kendall's distribution function for a copula C with parameters $\boldsymbol{\theta}, \upsilon \in [0, 1]$, and U and V are Uniform distributions over the interval $[0, 1]$ and jointly distributed according to C.

There are also a range of test-based analytical tools that are able to evaluate the dependency strength between the variables of interests and select the most appropriate copula. These are: independence test; and goodness-of-fit (GOF) tests. The Cramer–von Misses and Kolmogorov–Smirnov test, and the Vuong and Clarke tests are among the most well-known GOF tests (see also Genest and Favre, 2007; Vuong, 1989; Clarke, 2007).

Once the appropriate pair-copula families were selected, the estimation of the parameters via maximum likelihood can be derived. A brief explanation of the parameters estimation procedure for the vine structure shown in Fig. 1 is presented below (further details can be found in Aas et al., 2009). Suppose, there is a 3-dimensional distribution function (as presented above) along with N observations. We denote x_j as the vector of observations for the j-th point, with $j = 1, 2, \ldots, N$. The parameterised likelihood function for the D-vine decomposition given in (3) is as follows

$$L(D; \boldsymbol{\theta}) = \prod_{j=1}^{N} f(x_{1j}) f(x_{2j}) f(x_{3j}) \times c_{12}(F_1(x_{1j}), F_2(x_{2j}); \boldsymbol{\theta}_{12}) c_{23}(F_2(x_{2j}), F_3(x_{3j}); \boldsymbol{\theta}_{23}) c_{13|2}(F_1(x_{1j}|x_{2j}), F_3(x_{3j}|x_{2j}); \boldsymbol{\theta}_{13|2})$$

where $D = \{(x_{1j}, x_{2j}, x_{3j});\ j = 1, \ldots, N\}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{23}, \boldsymbol{\theta}_{13|2})$.

By taking logarithms and removing each of the marginal distribution term from the log-likelihood term, the log-likelihood function is given by

$$l(D; \boldsymbol{\theta}) = \log(c_{12}(F_1(x_{1j}), F_2(x_{2j}); \boldsymbol{\theta}_{12})) + \log(c_{23}(F_2(x_{2j}), F_3(x_{3j}); \boldsymbol{\theta}_{23})) + \log(c_{13|2}(F_1(x_{1j}|x_{2j}), F_3(x_{3j}|x_{2j}); \boldsymbol{\theta}_{13|2}))$$

We can then use numerical optimization techniques to maximize the log-likelihood over all parameters simultaneously. Aas et al. (2009) presented a detailed algorithm for likelihood evaluation and estimating the parameters of a D-vine construction model. We briefly explain this algorithm adopted to the D-vine model shown in Fig. 1.

The parameters of the copulas in the first tree (i.e., $(\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{23})$) can be estimated from the original data, simply by fitting the bivariate copulas to the observations. For the copula parameters identified in the second tree, one first has to transform the data using the

[1] However, unlike the graphical models, the vine models can benefit from using different models of conditional dependence as building blocks in building the multivariate distribution.

conditional distribution function also known as h-function which can be derived from the corresponding bivariate copula using the following formula

$$h(x,y,\theta) = F(x|y) = \frac{\partial C_{x,y}(x,y;\theta)}{\partial y}$$

The h-function is needed to derive the appropriate conditional distribution function using estimated parameters to determine realizations needed in the second tree. For instance, in order to estimate the parameters of $c_{13|2}$, we first need to transform the observations $\{u_{1j} = F_1(x_{1j}), u_{2j} = F_2(x_{2j}),\ u_{3j} = F_3(x_{3j}),\ j = 1,\dots,N\}$ to $u_{1|2,j} := h(u_{1j}|u_{2j},\hat{\theta}_{12})$ and $u_{3|2,j} := h(u_{3j}|u_{2j},\hat{\theta}_{23})$, where $\hat{\theta}_{12}$ and $\hat{\theta}_{23}$ are the estimated parameters in the first tree. We can now estimate $\theta_{13|2}$ based on $\{u_{1|2,j}, u_{2|3,j};\ j = 1,\dots,N\}$ (see Aas et al. (2009) for further details).

In this section, we present a constructive approach to build a multivariate distribution given a vine structure and selected appropriate marginal densities and bivariate copulas. That means, the vine models can be used to model general multivariate densities. In practice, the copulas must be chosen from a convenient class, and this class should ideally be one that allows us to estimate any copula to an arbitrary degree. By having this class of copulas, we can approximate any multivariate distribution using any vine structure. This issue will be investigated in the following.

3. Building minimum information pair-copula model

As demonstrated in the previous section, the vine models are flexible enough for modeling high-dimensional multivariate data by cascading different fitted parametric bivariate copulas together to construct the corresponding joint density function. Bedford et al. (2015) show that building higher-dimensional copulas by fitting a parametric copula family to the data is generally a difficult problem, and choosing the parametric family for this purpose is even more difficult. They argue that if the copulae are restricted to be chosen from a particular parametric class (Archimedean, t-student, Gaussian, etc.), their potential flexibility will not be acknowledged. To overcome this difficulty, a new non-parametric vine model is introduced that can be easily implemented in practice and is able to approximate the underlying multivariate copula density to any arbitrary degree of precision.

The presented method in this section is similarly constructive and involves the use of *minimum information* technique to approximate the copula density as precisely as possible. This approximation method is very flexible and allows the use of a fixed finite dimensional family of copulas to be used in a vine construction, with the promise of a uniform level of approximation (Bedford et al., 2015).

We first need to introduce the minimum information copula, and then briefly explain how this copula can be approximated based on the observed data (or experts' stated information). Assuming that f_1 and f_2 are bivariate densities, the relative information of f_1 with respect to f_2 is then defined (Bedford and Meeuwissen, 1997) as

$$I(f_1|f_2) = \iint \ln\left(\frac{f_1(x_1,x_2)}{f_2(x_1,x_2)}\right) f_1(x_1,x_2)dx_1\,dx_2$$

This information is a measure of the degree of deviation of f_1 from f_2 and is minimized 0 when $f_1 = f_2$. It is trivial to show that relative information of f_1 with respect to f_2 is the same as that between the copula for f_1 with respect to f_2. Therefore, it can be used to scale the strength of dependency in a copula in the sense that if the marginal distributions associated with f_1 and f_2 are similar, then $I(f_1|f_2)$ will be equal to the information measure derived in terms of the copula of f_1 relative to the independent copula.

A natural way to build a minimum information copula or specifying dependency constraints is through the use of moments. Follow Bedford et al. (2015), we consider moment constraints in which real-valued functions $\phi_1,\dots,\phi_k$ are required to take expected values $\alpha_1,\dots,\alpha_k$, respectively. A minimum information copula can be then fitted to satisfy these constraints. The fitted copula has minimum information, with respect to the uniform copula $c(u,v) = uv$, among the class of all copulas satisfying those constraints. Before presenting a general computational framework for constructing a minimum information copula satisfying the constraints, we explain the idea behind this methodology used in this paper. Suppose we have uniform variables u, v and the copula density we wish to find is $c(u,v)$. Further suppose that we wish to find a copula which, for some functions of uniform variables $\phi_1(u,v),\dots,\phi_k(u,v)$ which are assumed to be continuous on $[0,1]^2$, satisfies the constraints $E[\phi_i(u,v)] = \alpha_i$, for some values α_i. If we make the assumption that a copula satisfying the constraints exists then this problem is, in general, underdetermined. To select a unique copula distribution we wish to find the copula with minimum information with respect to the uniform copula satisfying these expectations. The relative information of $c(u,v)$ with respect to the uniform copula is given by

$$\int_0^1\int_0^1 c(u,v)\log(c(u,v))du\,dv$$

It is trivial to show that if $c(u,v)$ needs to be a copula density, the marginal distributions for u and v must be uniforms which results in additional constraints:

$$\int_0^1 c(u,v)du = 1, \quad \forall v \in [0,1]$$

$$\int_0^1 c(u,v)dv = 1, \quad \forall u \in [0,1]$$

In order to find a copula density function satisfying the constraints introduced above, we need to solve the continuous optimization problem. However, to do so, we shall first consider the associated measurable optimization problem. We can then use this to give a solution in the continuous case. Thus, the measurable optimization problem we wish to solve is

$$\begin{aligned}
\text{minimize}\quad & \int_0^1\int_0^1 c(u,v)\log(c(u,v))du\,dv \\
\text{subject to}\quad & \int_0^1 c(u,v)du = 1, \quad \forall v \in [0,1] \\
& \int_0^1 c(u,v)dv = 1, \quad \forall u \in [0,1] \\
& \int_0^1\int_0^1 \phi_i(u,v)c(u,v)du\,dv = \alpha_i, \quad = 1,\dots,k \\
& c(u,v) \geqslant 0, \quad \forall u,v \in [0,1]
\end{aligned} \tag{5}$$

We shall determine the unique solution to this measurable optimization problem. The solution of this optimization problem is called minimum information bivariate copula (Bedford et al., 2015) which can lead us to the minimum information pair-copula construction model. It is trivial to show that if a minimal information copula satisfied each of the constraints (based on moments, rank correlation, etc.), then the approximated multivariate density will also be minimally informative given those constraints (see also Bedford et al., 2015).

The solution, $c(u,v)$, to the measurable optimization problem presented above can be written in the form

$$c(u,v) = d^{(1)}(u)d^{(2)}(v)A(u,v) \tag{6}$$

where the kernel is given by

$$A(u,v) = \exp(\lambda_1\phi_1(u,v) + \cdots + \lambda_k\phi_k(u,v)). \quad (7)$$

for Lagrange multipliers $\lambda_1,\ldots,\lambda_k$ and measurable functions $d^{(1)}(u), d^{(2)}(v) : [0,1] \to \mathbb{R}$.

The representation given in (6) with the kernel given (7) forms a minimum information copula satisfying the constraints, $E[(\phi_i(u,v)))] = \alpha_i,\ i = 1,\ldots,k$. In other words, the copula given in (6) is a unique solution of the optimization problem introduced in (5).

There is a non-linear relationship between the set of $(\lambda_1,\ldots,\lambda_k)$ and $(\alpha_1,\ldots,\alpha_k)$. Bedford et al. (2015) give a detailed discussion about how this relationship can be determined. They also present a discrete version of the optimization problem given in (5) in terms of matrices that will be briefly explained below.

Suppose that both (u,v) are discretized into n points, as $\{(u_i,v_j),\ i,j = 1,\ldots,n\}$. We denote $A = (a_{ij})_{n\times n}, D_1 = diag(d_1^{(1)},\ldots,d_n^{(1)}), D_2 = diag(d_1^{(2)},\ldots,d_n^{(2)})$, where $a_{ij} = A(u_i,v_j), d_i^{(1)} = D_1(u_i)$, $d_j^{(2)} = D_2(v_j)$, and '*diag*' stands for a diagonal matrix. We define the matrix, D_1AD_2 with the uniform marginals as follows

$$\forall i = 1,\ldots n \quad \sum_j d_i^{(1)} d_j^{(2)} a_{ij} = 1/n, \quad and$$

$$\forall j = 1,\ldots n \quad \sum_i d_i^{(1)} d_j^{(2)} a_{ij} = 1/n,$$

The idea behind the D_1AD_2 algorithm is very simple, which starts with arbitrary positive initial matrices for D_1 and D_2, and the new vectors will then be successively defined by iterating the following maps

$$d_i^{(1)} \mapsto \frac{1}{n\sum_j d_j^{(2)} a_{ij}} \quad (i = 1,\ldots,n),$$

$$d_j^{(2)} \mapsto \frac{1}{n\sum_i d_i^{(1)} a_{ij}}, \quad (j = 1,\ldots,n)$$

It can be shown that this iteration scheme converges geometrically to the requested vectors (Bedford et al. (2015) and references therein).

Note that to compare different discretizations (for different n) we should multiply each cell weight $d_i(1)d_j(2)a_{ij}$ by n^2 as this quantity approximates the continuous copula density with respect to the uniform distributions.

The mapping from the set of vectors of λ's onto the set of vectors of resulting expectations of functions $(\phi_1,\ldots,\phi_k)$ has to be found numerically. Bedford et al. (2015) propose an optimization procedure to determine the λ_i's and corresponding copula for the given expectations α_i, where the expectations have been calculated using the discrete copula density D_1AD_2. Hence, to determine λ_i's whilst satisfying the constraints, the following set of equations has to be numerically solved

$$L_l(\lambda_1,\ldots,\lambda_k) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d^{(1)}(u_i)d^{(2)}(v_j)A(u_i,v_j)\phi_l(u_i,v_j) - \alpha_l,$$
$$l = 1,2,\ldots,k. \quad (8)$$

The left hand side of the above equations are just functions of λ's and, their roots can be found with optimization algorithms. Therefore, we must find the simultaneous roots of these functions and so minimize

$$L_{sum}(\lambda_1,\ldots,\lambda_k) = \sum_{l=1}^{k} L_l^2(\lambda_1,\ldots,\lambda_k).$$

One of the possible solvers for this task would be FSOLVE - MATLAB's optimization routine. An alternative method is to use another MATLAB's optimization procedure called FMINSEARCH, which implements the Nelder-Mead simplex method (see Lagarias et al., 1998).

Specifying the basis functions, $(\phi_1,\ldots,\phi_k)$, would greatly influence the copula density approximation describe above. A two-dimensional ordinary polynomial series is normally used to approximate the bivariate copula density. This approximation can be improved by using the orthonormal polynomial series or Legender multiwavelate which is studied in details in Daneshkhah et al. (2015) and will be also investigated in this paper to improve the minimum information pair-copula model fitted to the flood data.

4. Application

4.1. Study area and datasets

A study was performed with daily discharge data from the Beas River which originates in the Himalayas (Fig. 2) and flows for approximately 470 km before joining the Sutlej River. The Beas River, on which major two dams (Pong dam and Pandoh dam) are located, is one of the five major rivers of the Indus basin, India. The downstream Pong reservoir drains a catchment area of 12,561 km^2, of which the permanent snow cover is 780 km^2 (Jain et al., 2007). The active storage capacity of the Pong reservoir is 7051 Mm^3. Monsoon rainfall between July and September is a major source of water inflow into the reservoir in addition to snow and glacier melt. The dam acts as a store for flood flows, and reservoir regulation prevents the inundation of downstream areas from flooding during the monsoon season. Apart from its use for generating hydropower, the Pong reservoir meets irrigation water demands of 8896 Mm^3/year, which is spread relatively uniformly throughout the year. The Pandoh dam is a diversion dam which diverts nearly 4716 Mm^3 of Beas waters into the Sutlej River. Daily reservoir inflows to Pong reservoir for January 1998 to December 2010 (12 years) were used in this study. The Peak over threshold method is suitable for the Beas, as flash floods are common in the Himalayan region. Criterions for selection of independent POT data can be found in Bayliss (1999) and Bacova-Mitkova and Onderka (2010), but the threshold value is typically chosen so that the POT data series contains an average of around 4 values per year. For this study, the data series of peak discharges of 500 m^3/s and above with corresponding hydrograph volumes and durations were used for the analyses. The graphical method was used for independent event separation to obtain hydrograph volumes and durations (Fig. 3). Table 1 provides descriptive statistics of the flood event variables (flood peak discharge, P; hydrograph volume, V; and hydrograph duration, D). The kurtosis coefficients are quite high, and their skewness coefficients are positive indicating that these flood variables can be best modeled by non-symmetric heavy tailed distributions.

It should be noted that the number of available flood episodes extracted from the database are 109 data events due to data scarcity. Evidently, from a statistical point of view, the size of data could be small for investigating a multivariate problem. Unfortunately, this is a typical situation when multivariate copulas are used for modeling extreme data (e.g., Gaál et al., 2015; Genest et al., 2007; Favre et al., 2004; etc.). However, here the target is not to provide an ultimate extreme flood model, and no practical project of hydrological works (as for example considered in Gräler et al., 2013) is undertaken. Instead, one of our main motivations of this study is to demonstrate how the methodology proposed in this paper can be used in practice and exhibit its potential flexibility and efficiency over alternative multivariate copulas in modeling the flood data, particularly when the data is limited. In other words, this is rather a methodological paper.

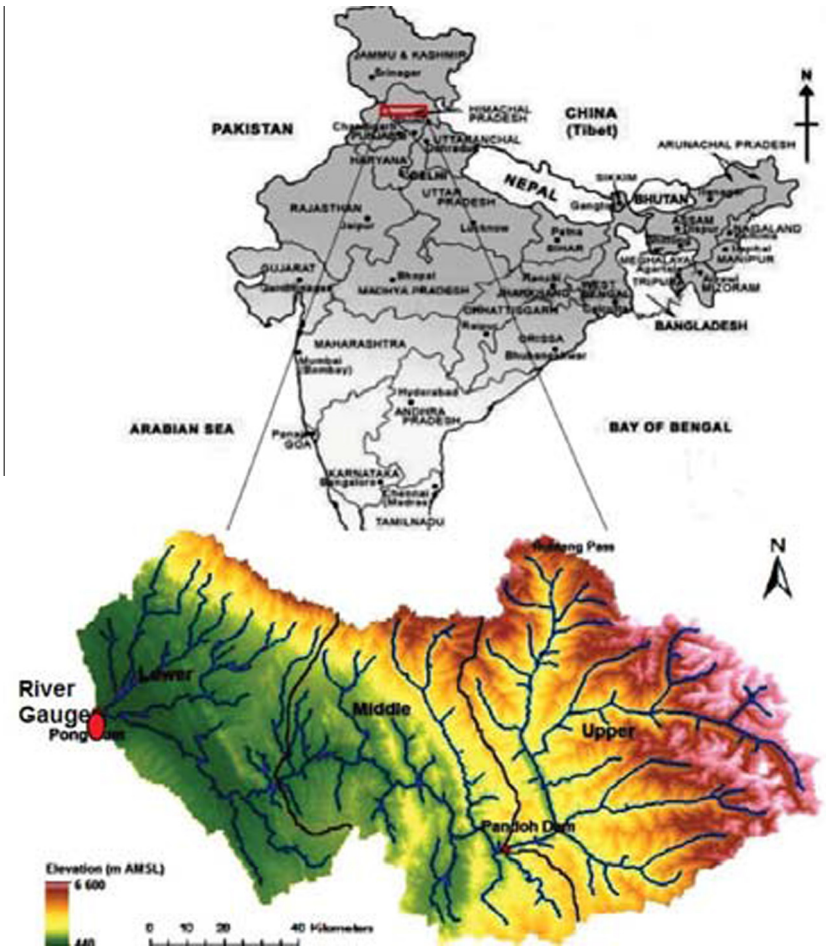

Fig. 2. Location of the Beas River.

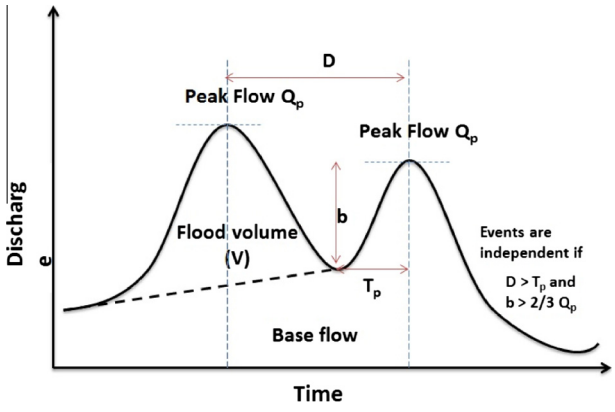

Fig. 3. Flood flow characteristics in hydrographs.

Indeed, another motivation of the methodology developed in our paper is modeling joint uncertainties in a probabilistic way and particularly when the data is limited or no data is available. For the latter case, the presented methodology can be viewed as an expert elicitation approach where the expert is asked to specify the expected values of some functions which is beyond the scope of this paper (see Bedford et al. (2015) for further details). However, this methodology can be more effective and efficient when it is used for approximating uncertainty modeling of the limited data which is very common in extreme value theory and risk analysis.

The size of observed data could be considered as a source of potential error when the minimum information copula is applied for modeling a high-dimensional problem. As the dimensionality (or number of uncertain variables) increases, the number of trees representing the structure of pair-copula model will also increase. The conditional distributions/expectations at lower levels of a deeper pair-copula model must then be estimated based on fewer data points which can be then less accurate and noisier (see also Gräler (2014) reported a sort of similar problem in modeling extreme data using spatial vine copula). This problem could be resolved by ignoring some unnecessary conditional dependencies (the so-called simplifying assumption) in the sense discussed in Acar et al. (2012) and Stöeber et al. (2013). An alternative method is

Table 1
Summary statistics of the flood event variables.

Variables	Max	Min	Mean	Std	Skewness cof.	Kurtosis cof.
Duration (days)	38	2	13.2	7.36	1.2	1.12
Peak (m^3/s)	10196	505	1956.7	1458.14	2.36	8.4
Volume (Mm^3/days)	2122	59.5	458.8	404	1.43	1.96

Table 2
The Q-statistics and their corresponding p-values.

Variables	Q-Statistics	p-values
Duration	26.0461	0.1643
Peak	16.7663	0.6681
Volume	12.4685	0.8990

to approximate fully conditional pair-copula models using Gaussian processes (Lopez-Paz et al., 2013). This simplifying pair-copula model is more appropriate for high-dimensional problems and is beyond the scope of this study. However, based on the demonstrated results, the approximations based on the minimum information pair-copula models for 3 variables are quite accurate (and can be made more accurate by adding more base functions and making grid discretization grid finer) and its performance in comparison with other methods is much better as discussed in Section 4.3.

Before modeling the dependencies between flood variables using the multivariate copula models, it is necessary to check whether the individual time series associated with each flood variable is stationary and exhibits no autocorrelation. Ljung and Box (1978) develop a statistical test, known as the Ljung-Box test, to check whether any of a group of autocorrelations of a time series are different from zero. In this test, instead of testing randomness at each distinct lag, the "overall" randomness based on a number of lags will be tested. The null and alternative hypothesis of this test are defined as:

H_0: The data are independently distributed (or there is no autocorrelation: $\rho_k = 0$);
H_1: The data are not independently distributed & they exhibit serial correlation.

The statistics to test these hypotheses which is known as Q-statistics, is defined as:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

where n is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k, and h is the number of lags being tested. Under the null hypothesis, this statistics follows a chi-square distribution with h degrees of freedom.

The Q-statistics and their corresponding p-values for each time series are shown in Table 2. Based on the computed p-values given in this table, the null hypothesis that there is no autocorrelation cannot be rejected at the 5% significance level. In other words, there is no serial correlation in the time series associated with the flood variables.

4.2. Trivariate copula models

In this section, we model the dependencies between the flood event variables by fitting a trivariate copula model. A wide range of multivariate copulas suitable to model the flood data including the well-known Archimedean and elliptical copulas introduced above have been evaluated. The marginal distribution of each variable is first selected based on the computed Akaike information criterion (AIC) given in Table 3 along with the estimated parameters (using maximum likelihood method). Using the results presented in this table, the Inverse Gaussian distribution is best fitted to the peak flow and flood volume, while the best fitted distribution to the flood duration is Log-Normal. Fig. 4 shows the cumulative distribution functions (cdfs), pdfs and q-q plots of the selected distributions to the data which supports our choices of distributions reported in Table 3.

We then select the best fitted trivariate copula model using the common goodness-of-fit measures including log-likelihood and AIC values which are given in Table 4. Based on the results given in this table, it can be concluded that the trivariate t-student outperforms the other proposed copula models (including, Frank, Gumbel, Clayton, etc.). The parameters' estimations of the selected copula (the pairwise correlation measures and the degree of freedom) are given as follows

$$\hat{\rho}_{VD} = 0.6534, \quad \hat{\rho}_{DP} = 0.2668, \quad \hat{\rho}_{VP} = 0.7839, \quad \text{and} \quad \hat{\nu} = 10.5168.$$

These results suggest that an elliptical copula is more suitable to model dependencies of the flood variables. The n-dimensional t-Student copula has been widely used for modeling of the hydrological (Ganguli and Reddy, 2013; Sraj et al., 2014). As mentioned above (and demonstrated in Aas et al., 2009), the main issue with

Table 3
Performance of various probability models for fitting marginal distributions for flood variables, where the best fitted distribution to each flood variable is highlighted with boldface.

Flood variables	Distributions	Estimated parameter	AIC
Peak flow	Gamma	$\hat{\alpha} = 2.54258,\ \hat{\beta} = 769.562$	1832.3
	Generalized extreme value	$\hat{k} = 0.397395,\ \hat{\sigma} = 681.169,\ \hat{\mu} = 1222.06$	1820.6
	Log-Normal	$\hat{\mu} = 7.36965,\ \hat{\sigma} = 0.633164$	1819.3
	Inverse Gaussian	$\hat{\mu} = 1956.7,\ \hat{\lambda} = 4066.3$	**1817.4**
	Normal	$\hat{\mu} = 1956.67,\ \hat{\sigma} = 1457.14$	1900.4
Volume	Gamma	$\hat{\alpha} = 1.63144,\ \hat{\beta} = 297.778$	1556.8
	Generalized extreme value	$\hat{k} = 0.4689,\ \hat{\sigma} = 200.464,\ \hat{\mu} = 257.538$	1561.8
	Log-Normal	$\hat{\mu} = 5.84904,\ \hat{\sigma} = 0.853755$	1553
	Inverse Gaussian	$\hat{\mu} = 485.8,\ \hat{\lambda} = 489.95$	**1550.8**
	Normal	$\hat{\mu} = 485.806,\ \hat{\sigma} = 404.006$	1620.6
Duration	Gamma	$\hat{\alpha} = 3.64367, \hat{\beta} = 3.6207$	713.4080
	Generalized extreme value	$\hat{k} = 0.157246,\ \hat{\sigma} = 4.80359,\ \hat{\mu} = 9.5802$	712.4440
	Log-Normal	$\hat{\mu} = 2.44,\ \hat{\sigma} = 0.54$	**710.36**
	Inverse Gaussian	$\hat{\mu} = 13.1927,\ \hat{\lambda} = 38.7913$	711.2240
	Normal	$\hat{\mu} = 1301927, \hat{\sigma} = 7.35663$	747.3700

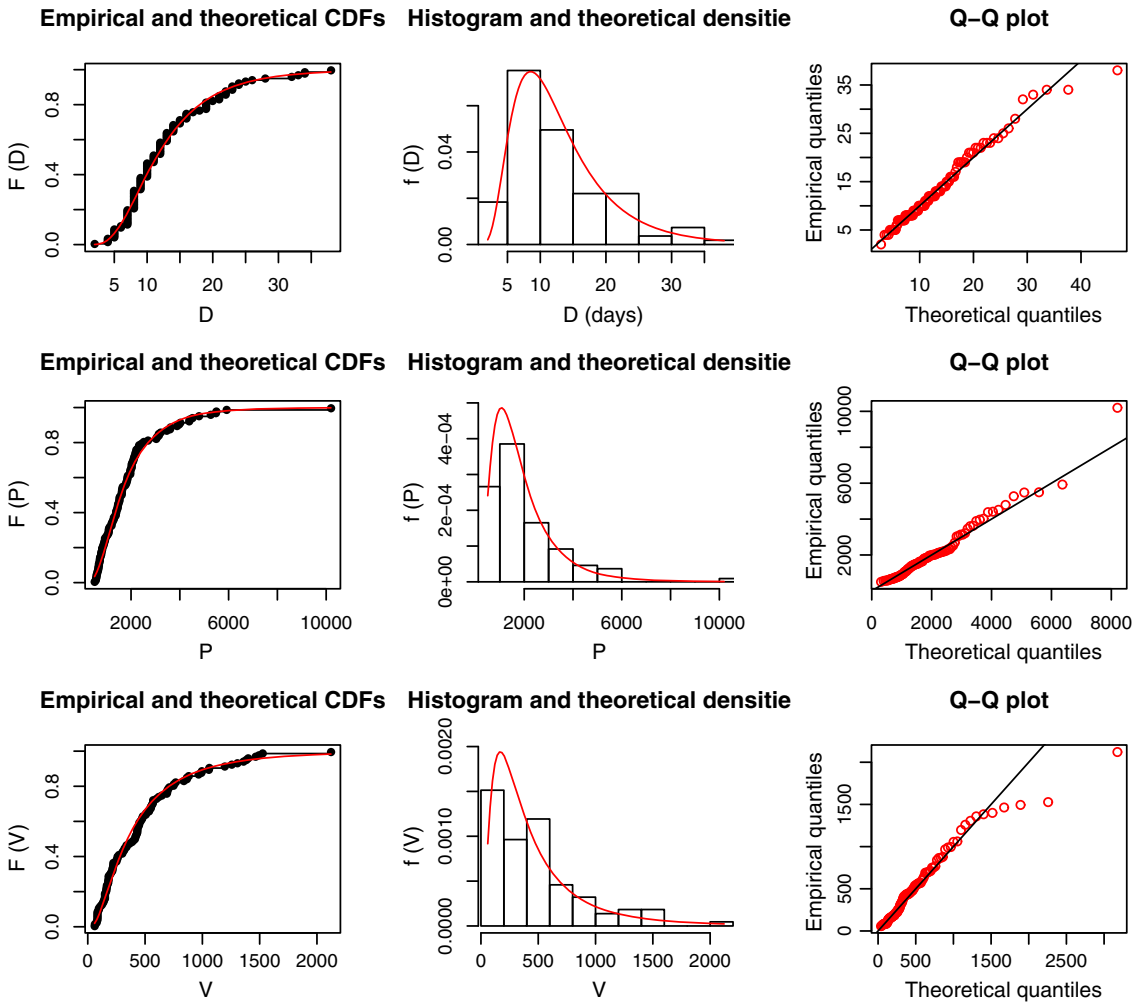

Fig. 4. Fitted distributions to the flood variables, D, P, V, are respectively shown in the 1st to 3rd rows. The 1st to 3rd columns illustrate cdfa, pdfs and q-q plots of the flood variables, respectively.

Table 4
The results of fitting different trivariate copula functions to the flood data. The best fitted copula, with the lowest AIC, is highlighted with boldface.

Multivariate copula function	Log-likelihood	AIC
Gaussian	92.2	−179.4
t-student	**96.64**	**−185.27**
Frank	16.4	−28.9
Gumbel	11.6	−19.2
Clayton	15.19	−28.4

the multivariate t-copula is that only a single degree of freedom parameter which drives the tail dependence of all pairs of variables is used. Therefore, if the tail dependencies of different pairs of the flood event variables are different, the dependence structure can be better described by the pair-copula models which will be discussed in the next section.

4.3. Modeling flood data using PCC models

In this section, we study the flood data using the PCC models and compare the fitted pair-copula model with the trivariate copula model selected in the previous section to verify the claim reported in the literature that the PCC model is generally superior to that of other multivariate copula models (Bedford et al., 2015; Aas et al., 2009; Joe et al., 2010).

In order to fit a PCC model to the flood data, we use the methods described in Section 2.1 to first identify an appropriate vine tree structure, then select the most appropriate copula families for the pair-copulas and estimate their parameters. Finally, the derived model will be evaluated and compared to the alternatives.

The first impression of the dependency structure of the flood event data is given in Fig. 5. The upper diagonal part of this figure show scatter plots, and the lower diagonal part shows the contour plots. There is evidently stronger dependence between (V, P) than between other pairs of variables (D, P) and (D, V). The correlation coefficients and p-values reported in Table 5 support the similar conclusions taken from the pairs plot. The strongest dependencies are between (P, V) and (V, D). That means, V should be placed between the other two variables as illustrated in Fig. 6 to model the flood event data. That means a D-vine copula model will be used for modeling the flood data. Aas et al. (2009) also reported that D-vines are indeed more flexible than canonical vines. This is mainly because for the canonical vines we should specify the relationships between one specific pilot variable and the others, while in the D-vine structure we can select more freely which pairs to model as demonstrated above (see also Czado et al. (2013) for a detailed discussion of regular vine model class selection).

Fig. 7 shows the chi-plots (first row) and Kendall's plots (second row) of the variables (D, V) (first column), (V, P) (second column) and (D, P) (third column) which indicate strong positive dependencies between these pairs of variables. Evidence of symmetric tail

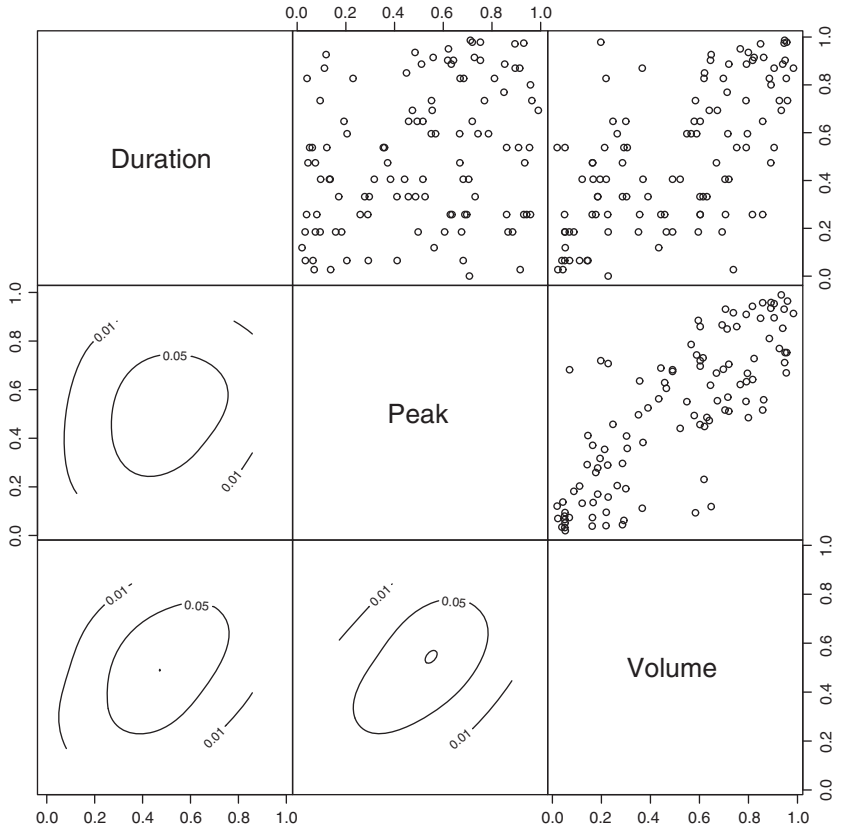

Fig. 5. The scatter plots (plots in the first row) and contour plots (presented in the second row) of the flood data.

Table 5
The correlation coefficients between the flood variables.

Dependence measure	$(P-V)$	$(V-D)$	$(D-P)$
Pearson r	0.66	0.60	0.19
Spearman ρ	0.80	0.63	0.29
Kendall's τ	0.60	0.47	0.20
p-value (2-tailed)	0.00	0.00	0.04

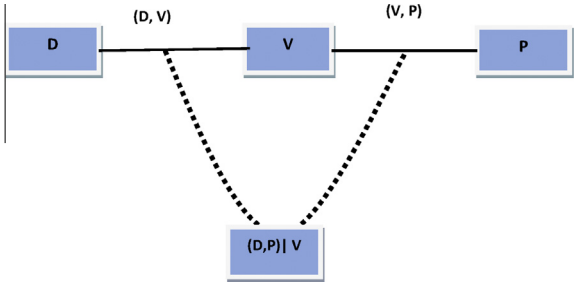

Fig. 6. Selected vine structure for the flood data set with 3 variables: Duration (D), flood peak (P) and flood volume (V).

dependence between the flood variables is also visible in these plots. Based on the properties of the different plausible copula candidates and their chi and Kendall's plots, we can conclude that t-Student, Gaussian or Frank copulas are most appropriate for these pairs of variables. In addition to these plots, by comparing empirical and theoretical λ-functions (given in Eq. (4)), an indication can be given as to which copula family is more suitable to describe the observed dependencies. On the left panel of Fig. 8, we present the empirical λ-function (black line) and theoretical λ-function of a Gaussian copula fitted to the pair of variables (P,V) with the estimated parameters (gray line) as well as independence and comonotonicity limits (dashed lines). The right panel of this figure shows the theoretical λ-function of a t-student copula fitted to (D,V). The closeness of the theoretical λ-functions of the suggested copulas with the empirical λ-functions support our choices yielded by using the chi and Kendall's plots. An R package called *CDvine* has been developed which provides the functions and tools used above for statistical inference of canonical vine and D-vine copulas (see Brechmann and Schepsmeier, 2013).

The scoring test based on the Vuong and Clarke tests described above strongly tends to select a Gaussian copula for the pair variables, (V,P) with the estimated parameters, $\hat{\rho}_{VP} = 0.7971014$. The same method selects a bivariate t-student copula between (V,D) with the following estimated parameters

$$\hat{\rho}_{(D,V)} = 0.6386490 \quad \text{and} \quad \hat{\nu}_{(D,V)} = 7.572639.$$

The similar copula models will be chosen if the AIC, log-likelihood, Cramer–von Misses or Kolmogorov–Smirnov test statistics are applied as the goodness-of-fit measures. It should be noted that the selected copulas are chosen from a wide range of alternative copulas including Frank, Gumbel, Frank, Joe, etc., and the reported copula represents the best fit among others.

In the next step, an appropriate copula between $(P|V, D|V)$ will be selected. We select this copula using the goodness-of-fit methods. Based on the computed goodness-of-fit measures, we select a t-copula with the following estimated parameters as the best fitted copula to (P,D) conditional on V:

$$\hat{\rho}_{(P|V,D|V)} = -0.5185675 \quad \text{and} \quad \hat{\nu}_{(P|V,D|V)} = 5.830839.$$

The AIC for this PCC based on the fitted bivariate copulas presented above is -192 which is less than the best fitted trivariate

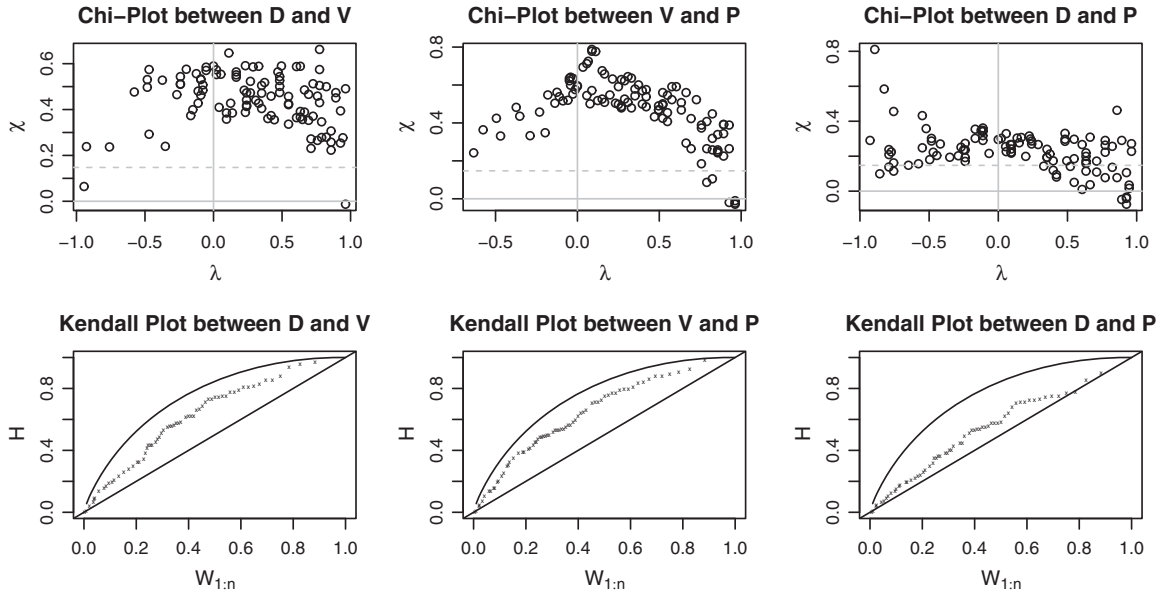

Fig. 7. Chi- and K-plots of the pairwise variables: 1st column these plots Chi- and K-plots for (D,V) variables, respectively which shows positive dependent between these variables. The 2nd and 3rd columns show these graphs for (V,P) and (D,P), respectively.

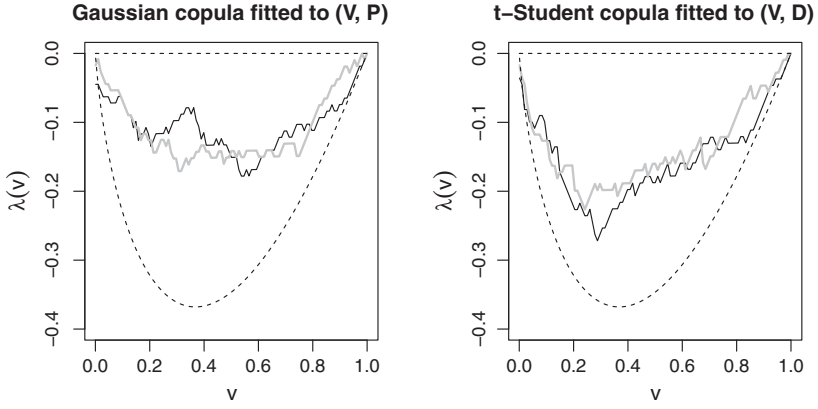

Fig. 8. Left panel: empirical λ-function (black line), theoretical λ-function of a Gaussian copula with the estimated parameters fitted to (P,V) (gray line) as well as independence and comonotonicity limits (dashed lines). Right panel: empirical λ-function (black line), theoretical λ-function of a t-student copula with the estimated parameter fitted to (D,V) (gray line) as well as independence and comonotonicity limits (dashed lines).

copula (i.e., t-student copula with AIC = -185.27). That means the PCC model is a more appropriate to model the flood data. Unlike the trivariate t-copula for all flood variables, this model enable us to use different copula models for each pair of the flood variables. Furthermore, the PCC model represents generally a more flexible and intuitive way of extending bivariate copulae to higher dimensions. Several studies have reported considerable improvement in modeling multivariate data which exhibit complex patterns of dependence in the tails using PCC model than the standard multivariate copula (particularly, multivariate t-copula), including Aas et al. (2009), Bauer et al. (2012), and Kurowicka and Joe (2011). Both of the models compared against each other suffer from this drawback that the chosen copulas are restricted to a particular parametric class (Gaussian, multivariate t, etc.) so that the potential flexibility of the copula approach is not realized in practice. The minimum information pair-copula model applied to analysis the flood even data, by contrast, allows a lot of flexibility in copula specification and results in a better fit.

4.4. Modeling flood data using minimum information pair-copula

In this section, we fit a joint probability distribution to the flood data using minimum information pair-copula described in Section 3. The same pair-copula structure illustrated in Fig. 6 will be used here. It is more convenient to present the minimum information copula in terms of functions of the so-called copula variables, denoted by $X = F_1(D)$, $Y = F_2(V), Z = F_3(P)$, where $F_i(.)$ denote to the marginal CDF of the flood variables derived above. These functions to construct a minimum information between (D,V) are given by

$$\phi_i(X,Y) = \phi_i'(F_1^{-1}(D), F_2^{-1}(V)), \quad i = 1,\ldots,k,$$

and these should clearly have the same specified expectation, that is, $E[\phi_i'(D,V)] = E[\phi_i(X,Y)]$.

We begin constructing the minimum information copulas between each set of two adjacent variables in the first tree, that

is, $C(D,V)$ and $C(V,P)$. In order to implement this, one needs to decide which basis functions should be chosen for each pair of these variables. We show only the detailed procedure of estimating the copula between (D,V).

As mentioned above, a two-dimensional ordinary polynomial series can be used to approximate log-density of a bivariate copula function by truncating the series at an appropriate point until they were satisfied with the approximation. As Daneshkhah et al. (2015) show that this approximation can be improved by using the orthonormal polynomial series, we also use the orthonormal polynomial basis functions to approximate the bivariate copula function of interest. We therefore briefly define the orthonormal polynomial functions in $[0,1]$ and then give a procedure to select an appropriate series of these basis functions.

Two polynomial functions h_1 and h_2 are called orthonormal in $[0,1]$, if

$$\int_0^1 h_1(x)h_2(x)dx = \begin{cases} 1 & \text{for } h_1(x) = h_2(x); \\ 0 & \text{for } h_1(x) \neq h_2(x). \end{cases} \quad (9)$$

We follow Gram-Schmidt procedure to construct the orthonormal polynomial (OP) basis functions. Using this method, OP series can be defined as

$$\phi_0(x) = 1, \quad \phi_1(x) = \sqrt{3}(-1+x), \quad \phi_2(x) = \sqrt{5}(1-6x+6x^2), \quad \phi_3(x) = \sqrt{7}(-1+12x-30x^2+20x^3)$$

$$\phi_4(x) = \sqrt{9}(1-20x+90x^2-140x^3+70x^4), \ldots$$

The two-dimensional OP basis functions are then given by

$$\{\phi_i(x)\phi_j(y)\}_{i,j\geqslant 1}$$

In order to choose the most suitable basis functions to approximate the density of interest, we use an optimal method which is similar to the stepwise regression procedure (Bedford et al., 2015). In this method, at each stage, we evaluate the log-likelihood changes after adding each additional basis function. We then choose the basis with the largest increase in the log-likelihood. By applying this method on the proposed OP basis functions, we select the following four bases

$$\phi_1(D)\phi_1(V), \quad \phi_2(D)\phi_2(V), \quad \phi_4(D)\phi_5(V), \quad \phi_5(D)\phi_1(V)$$

It should be noted however there is no longer a jump in the log-likelihood when adding the fifth basis function, but the approximation can be slightly improved by adding more basis which will not be considered here. We use this step-wise technique to choose all of the remaining basis functions for other pairs in this case study.

The calculated expected values of these basis functions based on the observed data are given by

$$\alpha_1 = \frac{1}{109}\sum_{i=1}^{109}\phi_1(D_i)\phi_1(V_i) = 0.6315, \quad \alpha_2 = \frac{1}{109}\sum_{i=1}^{109}\phi_2(D_i)\phi_2(V_i) = 0.3694,$$

$$\alpha_3 = \frac{1}{109}\sum_{i=1}^{109}\phi_4(D_i)\phi_5(V_i) = -0.0973, \quad \alpha_4 = \frac{1}{109}\sum_{i=1}^{109}\phi_5(D_i)\phi_1(V_i) = -0.1254,$$

The minimum information copula $C(D,V)$ with respect to the uniform distribution given the constraints above can be now constructed. In order to do this, we also need to decide on the number of discretization points or grid size. It is shown that a larger grid size will provide a better approximation to the log-density of copula but would increase the computational time (Bedford et al., 2015). They also illustrate that the more iterations of the D_1AD_2 would result in a more accurate density approximation. In order to make a balance between the level of accuracy and the computational time, we choose a grid size of 200×200 and fixed the approximation error at 1×10^{-12}.

The Lagrange coefficients of this density approximation, satisfied in Eq. (8), are given by

$$\lambda_1 = 0.8074, \quad \lambda_2 = 0.2229, \quad \lambda_3 = -0.1266, \quad \lambda_4 = -0.1889$$

The approximated minimum information copula, $C(D,V)$ is shown in the left plot of Fig. 9.

The copula density between (V,P) can be similarly approximated. First, we select the most suitable OP basis functions using the stepwise like method, as described above. These functions are as follows

$$\phi_1(P)\phi_1(V), \quad \phi_2(P)\phi_2(V), \quad \phi_1(P)\phi_4(V), \quad \phi_4(P)\phi_3(V)$$

The corresponding constraints as the mean of the above functions are calculated using the observed data as

$$\alpha_1 = 0.7750, \quad \alpha_2 = 0.4796, \quad \alpha_3 = 0.1141, \quad \alpha_4 = -0.1501$$

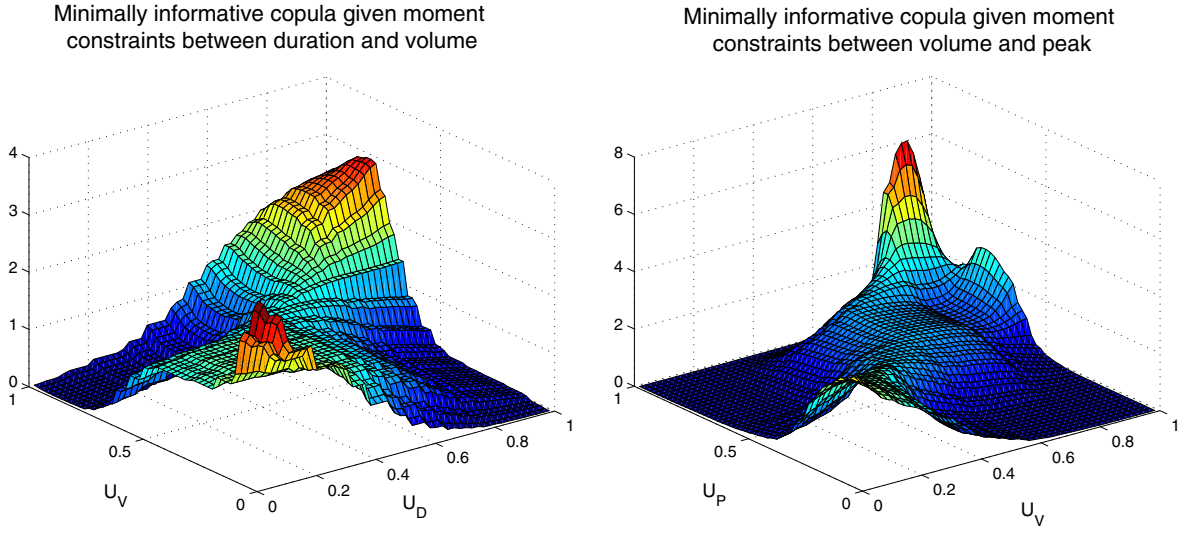

Fig. 9. The minimally informative copula given moment constraints between the variables: Left plot, minimum information copula between (D,V); Right plot, minimum information copula between (V,P).

By fitting the minimum information copula to these data and constraints, the following Lagrange multipliers are obtained

$$\lambda_1 = 1.7189, \quad \lambda_2 = 0.26523, \quad \lambda_3 = 0.45487, \quad \lambda_4 = -0.1945$$

The corresponding approximated minimum information copula, $C(V,P)$ is shown in the right plot of Fig. 9.

The conditional copula, $C(D|V,P|V)$, located in the second tree of the PCC illustrated in Fig. 6 can be similarly approximated. In order to calculate the minimum information copula between $D|V$ and $P|V$, we first split the support of V into some arbitrary sub-intervals or bins (4 bins in this example) and then approximate the corresponding copula on each bin using the minimum information copula. The basis functions will be selected in the same way discussed above. Table 6 shows the selected basis functions, their corresponding expected values, Lagrange coefficients and log-likelihoods for each bin.

We now compare the methods used in this paper to model the dependencies between the flood variables based on the computed AIC of the fitted copula illustrated in Table 7. The AIC of the overall minimum information pair-copula model is considerably less than the AICs of the trivariate copula and less than the parametric D-vine copula model. That means the minimum information PCC model fits the observed data better than other models, and all dependencies are better captured using this method (see Table 8).

In addition to the correlation measures reported in Table 8, we can validate the proposed approximation method based on the minimum information copula based on the simulations drawn from the fitted models. In the next section, we first introduce a simulation method which will be used to validate our approximation and then to compare our approximation versus other alternative methods.

It should be noted that the best model should be selected by trading-off between the goodness of fit of the candidate model and the complexity of the model (e.g., AIC). The proposed approximation method is a general and can be applied to approximate any multivariate distribution with any degree of complexity to any required degree of approximation. Indeed the flexibility of vines gives us the potential to capture any fine-grain structure within a multivariate distribution, and unlike the Bayesian networks, the PCC can be modeled in terms of the conditional dependence aspects which could result in much simpler model structure search. In addition, unlike Multivariate Gaussian copulas, the proposed method in this paper allows the explicit modeling of non-constant conditional dependence. However, Serinaldi (2013) extends this widespread belief that the increasingly refined mathematical structures of probability functions increase the accuracy and credibility of the fitted models (particularly, in extrapolating upper tails of the fitted models), but we have found some mixed conclusions of simplifying vine models and surrounding assumptions. It is evident that the deeper a bivariate copula is in the vine hierarchy, more variables will be conditioned on. Thus, if the aforementioned conditional dependencies are neglected, the pair-copula constructions models are a direct method to build a flexible multivariate models using standard parametric bivariate copulas as building blocks. Acar et al. (2012) argue that however the ignoring conditional dependencies (so-called simplifying assumption) can lead to reasonably precise approximations of the underlying copula (as claimed by Haff et al., 2013), but this can generally be misleading, and develop an approach to condition parametric bivariate copulas on a single scalar variable. Stöeber et al. (2013) repeated this concern after studying several examples that the simplifying assumption for the pair-copula construction models is often too restrictive, and also the assumption of dealing with absolutely continuous pair-copula construction model is sometimes too strong. The latter assumption is used to make the pair-copula models tractable for inference and model selection (a pair-copula construction model is called an absolutely continuous if all bivariate copula families occurring in the construction have densities with a parameter vector). Lopez-Paz et al. (2013) also reported that the simplifying assumption can lead to a totally oversimplified estimates in practice. They then extended the work of Acar et al. by developing a method for estimation of fully conditional vines using Gaussian Process. This model shows promising results with better predictive performance than the method that ignores conditional dependencies.

Table 6
Bases, parameter values and log-likelihoods for $C(D|V,P|V)$.

Interval	Bases	α_i	λ_i	Log-likelihood
$0 < V < 0.25$	$\varphi'_1(D)\varphi'_1(P)$	0.9388	1.3525	
	$\varphi'_2(D)\varphi_3(P)$	−0.4108	−0.65797	10.39
	$\varphi'_4(D)\varphi'_2(P)$	0.2457	−0.33647	
	$\varphi'_2(D)\varphi'_5(P)$	−0.0397	−0.53411	
$0.25 < V < 0.5$	$\varphi'_1(D)\varphi'_2(P)$	0.4333	2.1203	
	$\varphi'_2(D)\varphi'_1(P)$	0.2736	0.54487	5.27
	$\varphi'_4(D)\varphi'_3(P)$	−0.0149	−0.49954	
	$\varphi'_1(D)\varphi'_5(P)$	−0.1409	−0.7591	
$0.5 < V < 0.75$	$\varphi'_1(D)\varphi'_2(P)$	−0.3588	−0.75247	
	$\varphi'_2(D)\varphi'_3(P)$	0.2044	0.58484	7.17
	$\varphi'_1(D)\varphi'_5(P)$	0.3391	1.2127	
	$\varphi'_4(D)\varphi'_5(P)$	0.0936	0.44209	
$0.75 < V < 1$	$\varphi'_3(D)\varphi'_5(P)$	−0.2474	−0.18412	
	$\varphi'_1(D)\varphi'_4(P)$	−0.1147	0.24618	3.29
	$\varphi'_2(D)\varphi'_5(P)$	−0.1565	−0.10098	
	$\varphi'_4(D)\varphi'_5(P)$	−0.1547	−0.06666	

Table 7
The results of fitting different copula functions to the flood data.

Type of copula	AIC
t-student	−185.27
Gaussian	−179.4
Parametric pair-copula	−192
Minimum information Copula	−204.8

Table 8
Simulated correlations using different methods.

Method	ρ_{DV}	ρ_{VP}	ρ_{DP}
Observed data	0.628	0.7913	0.2853
Trivariate t-student copula	0.625	0.745	0.221
Pair-copula model	0.6263	0.781	0.266
Minimum information copula	0.6279	0.7950	0.2882

4.5. Validation by simulation

We now discuss the simulation of data taken from the PCC model. We follow the simulation method proposed by Kurowicka and Cooke (2006) based on sampling from the cumulative distributions. Their sampling strategy is as follows: sample three independent variables distributed uniformly on intervals $[0,1]$, denoted by U_1, U_2, U_3, and calculate values of the original variables using the following equations:

$$x_1 = u_1, \quad x_2 = F^{-1}_{2|1}(u_2|x_1), \quad x_3 = F^{-1}_{3|1,2}(u_3|x_1,x_2)$$

where x_i is realization values of X_i, and u_i is realization value of U_i. More details and pseudo code can be found in Daneshkhah et al. (2015) (see also Cooke et al., 2015).

Table 8 shows the rank correlations between the pairs of the flood variables calculated from the original observed data, and

Table 9
Comparison of return periods for flood characteristics calculated based on trivariate t-copula (denoted by T_{TT}), pair-copula model (T_{PC}) and minimum information pair-copula model (T_{MI}).

T	Peak	Duration	Volume	T_{TT}^{OR}	T_{PC}^{OR}	T_{MI}^{OR}	T_{TT}^{AND}	T_{PC}^{AND}	T_{MI}^{AND}
5	154.7	13.3	500.7	4.6	5.1	6.2	15.6	18.1	19.3
10	200.6	17.1	673.7	12.3	11.7	12.6	66.5	72.5	80.1
20	224.9	22.5	789.4	24.8	26.7	27.9	110.78	121.78	130.14
50	244.8	26.4	905.3	44.11	42.19	45.04	300.11	293.11	307.4

based on the simulated data of size 1000 taken from the fitted trivariate t-student copula; the parametric PCC model; and the minimum information pair-copula. Both methods (PCC model and minimum information copula) reproduce the overall correlation structure fairly well. We further investigate and compare the tail dependence of the minimum information copula with the other copulas proposed above based on simulation study in the following section.

5. Probabilistic analysis of flood variables

The frequency analysis of multivariate extreme events is very useful for understanding critical hydrologic behavior of flood events at a river basin scale through consideration of multiple interacting flood characteristics. The understanding gained from such analyses would be very helpful in measuring nonstructural safety, and in developing flood hazard mitigation strategies, as the impacts of extreme flood events with similar peak flows can differ greatly depending on event duration and hence volume (i.e. long duration-high volume floods compared to short duration-moderate volume flash floods).

The objective of frequency analysis of hydrologic data is then to relate the magnitude of extreme events to their frequency of occurrence through the use of probability distributions (Chow et al., 1988; Ganguli and Reddy, 2013). For multivariate case, in which the flood variables, D, V, P exceeds their respective thresholds ($D > d^*, V > v^*, P > p^*$), the joint return period is computed using inclusive probability ("OR" and "AND" cases) of all three events, known as primary return periods (Salvadori, 2004). The joint primary return period in "OR" case denoted by $T_{(D,V,P)}^{OR}$ (for annual flood analysis) is defined as

$$T_{(D,V,P)}^{OR}(d^*, v^*, p^*) = \frac{1}{P(D \geqslant d^*, \vee\, V \geqslant v^*, \vee P \geqslant p^*)} = \frac{1}{1 - P(D \leqslant d^*, V \leqslant v^*, P \leqslant p^*)} = \frac{1}{1 - F_{D,V,P}(d^*, v^*, p^*)} = \frac{1}{1 - C(u_1, u_2, u_3)}$$

where $u_1 = F_D(d^*)$, $u_2 = F_V(v^*)$, $u_3 = F_P(p^*)$ and $C(u_1, u_2, u_3)$ is a trivariate copula.

The joint primary return period in "AND" case denoted by $T_{X_1X_2X_3}^{AND}$ (for annual flood analysis) is defined as,

$$T_{(D,V,P)}^{AND}(d^*, v^*, p^*) = \frac{1}{P(D \geqslant d^*, V \geqslant v^*, P \geqslant p^*)} = \frac{1}{1 - F_D(d^*) - F_V(v^*) - F_P(p^*) + F_{D,V}(d^*, v^*) + F_{D,P}(d^*, p^*) + F_{V,P}(v^*, p^*) - F_{D,V,P}(d^*, v^*, p^*)} = \frac{1}{1 - F_D(d^*) - F_V(v^*) - F_P(p^*) + C(u_1, u_2) + C(u_1, u_3) + C(u_2, u_3) - C(u_1, u_2, u_3)}$$

where $C(u_1, u_2)$, $C(u_2, u_3)$, and $C(u_1, u_3)$ are bivariate copulas between the cdfs of the flood variables.

Table 9 exhibits return period obtained using univariate marginal distributions of peak flow, volume, and duration; and joint return periods for "AND" and "OR" cases for the different trivariate distributions presented in this paper. In this table, T_{TT}^{AND}, T_{PC}^{AND} and T_{MI}^{AND} present the joint return periods for "AND" case approximated by trivariate t-copula, PCC model and the minimum information pair-copula model, respectively. The differences between the joint return periods for "AND" (and "OR") case are due to the approximation methods of trivariate and bivariate copulas required in the definitions of $T_{D,V,P}^{AND}$ and $T_{D,V,P}^{OR}$. The joint return period in "AND" case, using any approximation method, is greater than the joint return period in "OR" case. Hence, it also infers that the occurrence of trivariate flood characteristics simultaneously is less frequent in "AND" case and more frequent in "OR" case.

Fig. 10 shows the joint bivariate return periods for the OR and AND cases for the pairs of flood variables. Ganguli and Reddy (2013) reported that the joint bivariate return period in "AND" case is greater than the joint bivariate return period in "OR" case. A similar finding is concluded here.

The study shows the joint return period, $T^{AND}(D, V, P)$, in the case of minimum information copula is larger than other copulas and the values are followed by parametric pair copula and trivariate t-copula indicating that other two methods are underestimating the flood hazard under high value combinations. In Table 9, the maximum return periods for $T^{AND}(D, V, P)$ and $T^{OR}(D, V, P)$, based on individual flood event characteristics with return periods of 100 years ranged from 1101 years and 94 years represent the range of possible Beas river flood hazards in the case of minimum information copula.

5.1. Analyzing tail dependence: a simulation study

Based on the results presented above and demonstrated in Bedford et al. (2015) and Daneshkhah et al. (2015), the pair-copula model constructed based on the minimum information copulas can model any dependence structure. In many fields, including hydrology, extreme weather forecast, financial risk prediction that the fitted copula would lie within non-Gaussian multivariate families (Joe, 1997), tail dependence properties and behavior are more important. We therefore investigate the tail behavior of the minimum information copula for the data simulated from the fitted copulas introduced above.

Tail dependence in a bivariate distribution can be represented by the probability that the first variable exceeds its q-quantile,

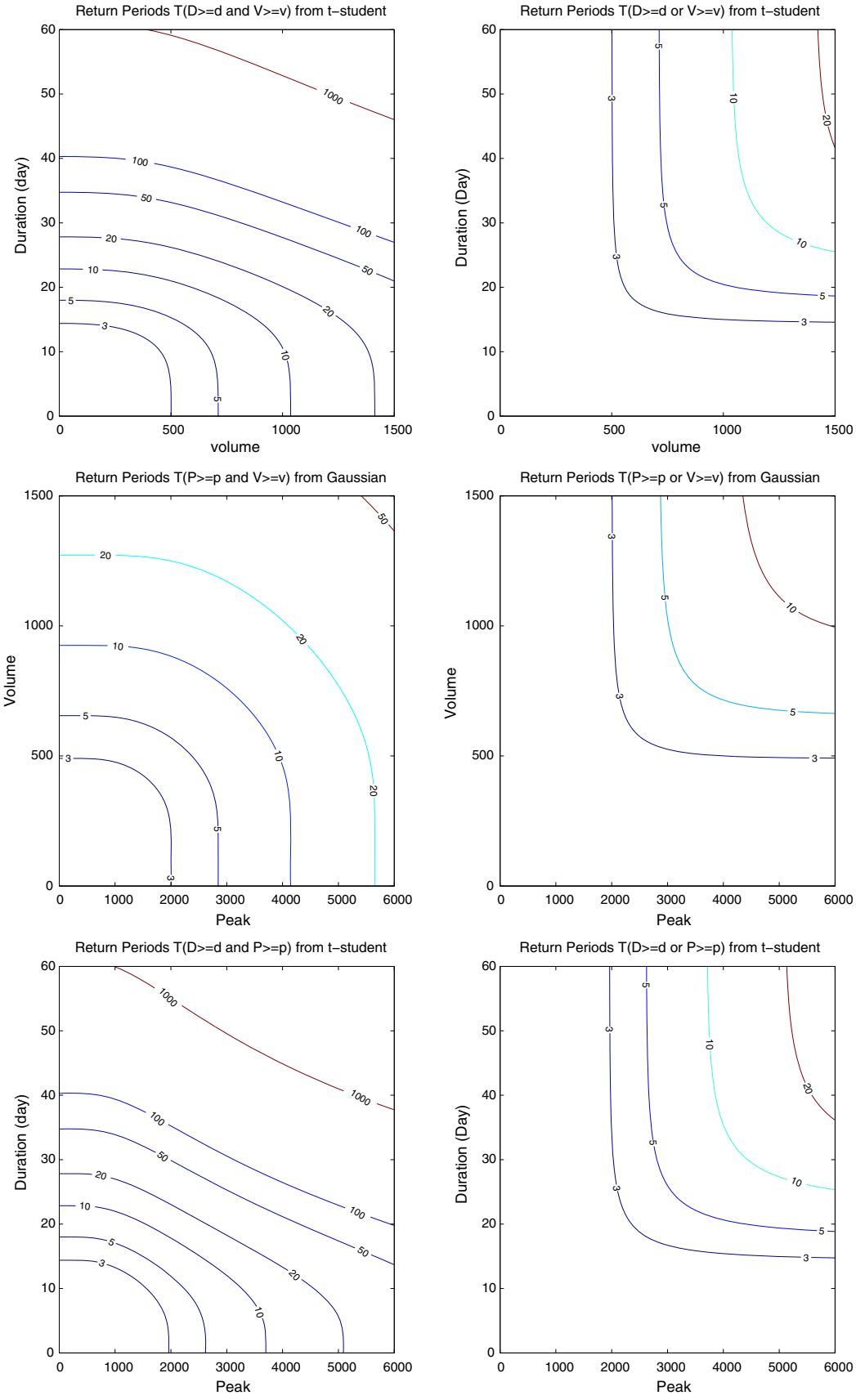

Fig. 10. Contour plots for joint return period values (in years) for OR and AND cases for the pairs of flood event variables.

given that the other exceeds its own q-quantile. In order to study the tail behavior of the fitted minimum information copulas, we first utilize scatter-plot, Chi-plot and K-plot which can detect bivariate dependence using the ranks of the data as explained in Section 2. The first column of Fig. 11 illustrates a scatter-plot of a random sample (of size 1000) taken from the fitted Normal copula (as fitted to (V,P) variables) with correlation coefficient of $\hat{\rho} = 0.7971014$, and the corresponding Chi and K-plots. The second column demonstrates the same plots of a random sample with the same size taken from the minimum information copula fitted to (V,P). By comparing the scatter-plots, it can be concluded that the minimum information copula is well capturing the general behavior of the Normal copula. The upper and lower tail dependency can be derived from the Chi and K-plots. For example, If there is no upper or lower tail dependence, the χ values rightmost of the Chi-plot have to return to the zero line. This can be clearly observed in the Chi-plots of the Normal and corresponding minimum information copulas. The same tail dependencies behaviors can be observed from the K-plots of these copulas.

Similarly, in the first column of Fig. 12, a scatter-plot of a random sample (of size 1000) taken from the fitted t-copula (as fitted to (D,V) variables) with parameters of $\hat{\rho} = 0.6386490$ and $\hat{\nu} = 7.572639$, and the corresponding Chi and K-plots are shown, the corresponding plots associated with the fitted minimum information copula are illustrated in the second column. By comparing the scatter-plots, it can be concluded that the minimum information copula is well capturing the general behavior of the t copula. A similar upper tail dependency can be observed for these two copula by comparing their Chi and K-plots. The minimum information copula is able to capture the upper tail behavior which can be found in other copulas including Gumbel and Tawn Copulas, and lower tail dependency such as Frank copula (see also Bedford et al. (2015) for similar findings). In addition to these graphical tools to detect and study the tail behavior of the fitted copulas above, we also present some analytical tools to measure tail dependency.

In order to study occurrence of extreme events like flood, the pair-wise analysis of upper tail dependence of flood variables can be implemented using the fitted copula models. The coefficient of upper tail dependence of two variables of interests X and Y is denoted by $\lambda_U(X,Y)$ and defined as follows

$$\lambda_U(X,Y) = \lim_{\alpha \to 1^-} P(Y > F_Y^{-1}(\alpha) \Big| X > F_X^{-1}(\alpha)) \tag{10}$$

where α is considered as a threshold value associated with the upper tail dependence between these variables.

This coefficient can be also presented in terms of copula as given in (Joe, 1997). It can be shown that if $0 < \lambda_U \leqslant 1$, the corresponding variables are said to be asymptotically dependent in the upper tail or the corresponding copula, C coupling these variables has upper tail dependence; if $\lambda_U = 0$, the variables are said to be independent in the upper tail.

In flood hazard management, it is very crucial to take into account the tail-dependence coefficient in the modeling of joint flood characterizations. Otherwise, it can lead to a serious underestimation of the hazard and under design of flood protection works, with well-known consequences. Therefore, computing the tail-dependence coefficients as precise as possible would reduce the associated hazard. The method of approximating a bivariate copula using the minimum information technique can be used to estimate the tail-dependence coefficient by any level of approximation as desired. In this section, we analysis the tail dependencies between the flood variables using the different methods of modeling copulas demonstrated above.

The tail dependence may be studied either graphically using the chi-plot or numerically from an empirical copula, a given group of multivariate distributions, and a given group of copula functions. There are closed formulas for tail dependence of the bivariate t-student and Gaussian copulas given in Table 10.

In order to calculate the tail dependence associated with the fitted minimum information pair-copula, we can use the

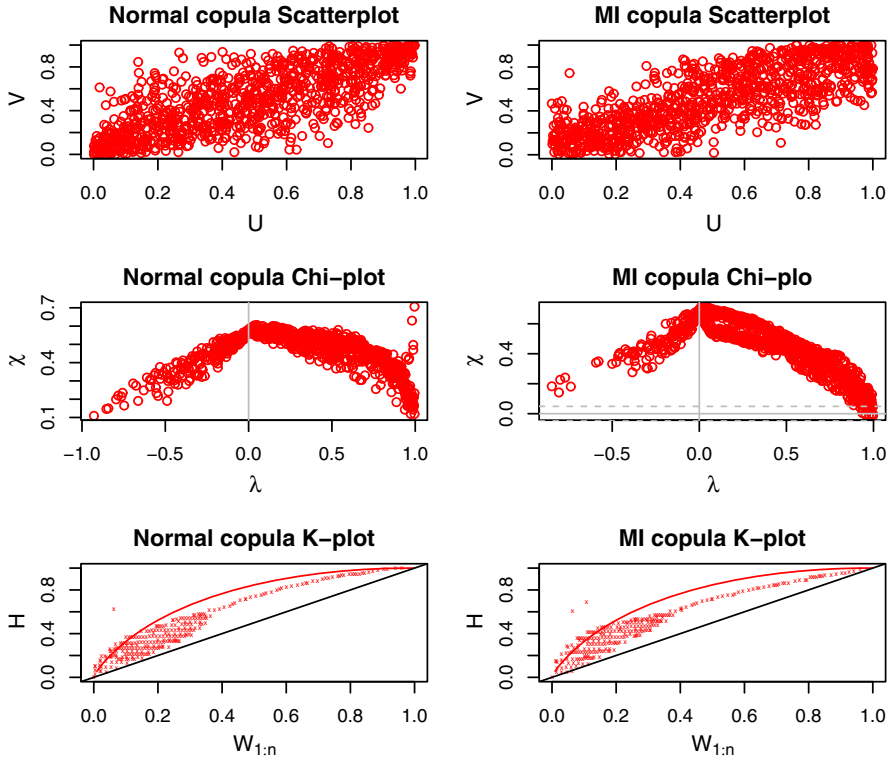

Fig. 11. Scatter plots, Chi-plots and K-plots of the normal copula and the minimum information copula fitted to (V,P) variables.

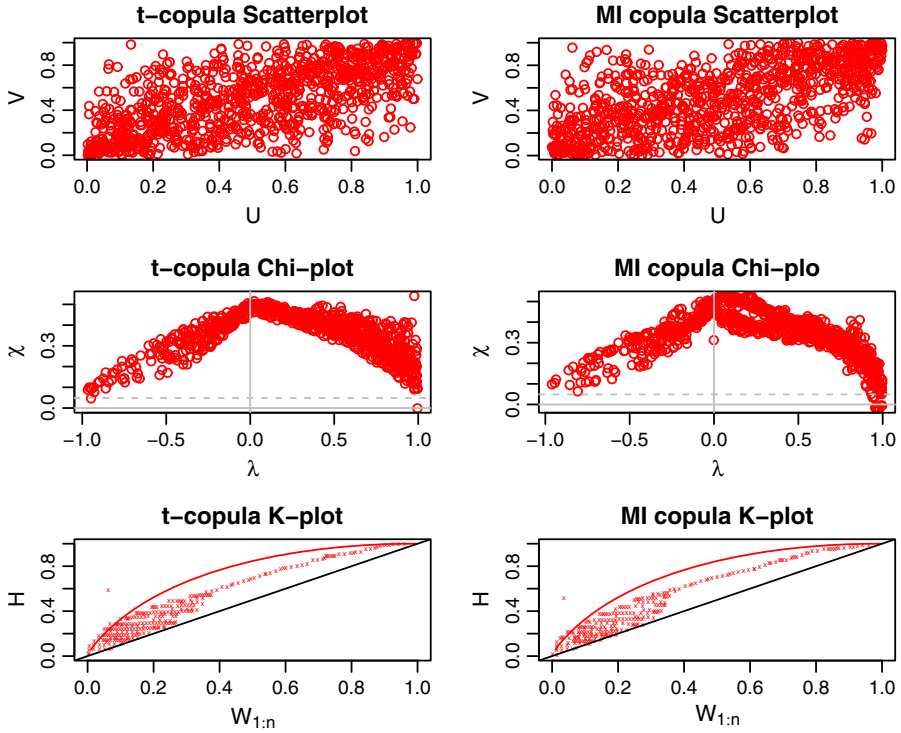

Fig. 12. Scatter plots, Chi-plots and K-plots of the t-student copula and the minimum information copula fitted to (D, V) variables.

Table 10
Properties and denotation of bivariate elliptical copula families.

Elliptical distribution	Parameter range	Kendall's τ	Tail dependence
Gaussian	$\rho \in (-1,1)$	$\frac{2}{\pi}\arcsin(\rho)$	0
Student-t	$\rho \in (-1,1), \nu > 2$	$\frac{2}{\pi}\arcsin(\rho)$	$2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\right)$

Table 11
Tail dependence coefficients for the flood variable pairs computed based on the different methods.

Variables	$\hat{\lambda}_U^{TT}$	$\hat{\lambda}_U^{PC}$	$\hat{\lambda}_U^{MI}$
(D,V)	0.202	0.208	0.2106
(V,P)	0	0	0.0000012
(D,P)	0.0001	0.0024	0.00291

non-parametric estimations of the tail dependence. We use the C apéraá–Fougéres–Genest estimator denoted by $\hat{\lambda}_U^{CFG}$ and suggested by Capéraá et al. (1997) to compute the tail dependencies between the pairs of flood variables fitted by the minimum information PCC. In order to calculate $\hat{\lambda}_U^{CFG}$, a random sample $\{(U_1,V_1),\ldots,(U_n,V_n)\}$ taken from the underlying copula $C(\cdot,\cdot)$ is required. The bivariate upper tail dependence, $\hat{\lambda}_U^{CFG}$ is then given by

$$\hat{\lambda}_U^{CFG} = 2 - 2\exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{\sqrt{\log\left(\frac{1}{U_i}\right)\log\left(\frac{1}{V_i}\right)}}{\log\left(\frac{1}{\max(U_i,V_i)^2}\right)}\right]\right) \tag{11}$$

Table 11 shows the tail dependence coefficients for the different pairs of the flood variables and different types of copula models to capture the dependency structure. These coefficients are calculated based on the samples taken from the multivariate copulas fitted to the flood data in this paper. For instance, the tail dependence coefficients λ_U^{TT} for each pair of the flood variables are calculated using (11) and based on a sample taken from the trivariate t-copula fitted to the flood data. In this table, we denote *TT* as the trivariate t-copula, *PC* stands for the pair-copula model, and *MI* denotes the minimum information copula.

Based on the results shown in Table 11, for the bivariate copula between (D,P), the value for the pair-copula distribution is 24 times and for the minimum information copula 29 times higher than the corresponding one for the trivariate t-copula. The practical implication of this difference in tail dependence is that the probability of observing a long duration flood is much higher for the PCC model and the minimum information pair-copula model than it is for the trivariate t-copula.

6. Conclusions

The aim of this paper was to present the use and usefulness of pair-copulas and minimum information pair-copula in flood hazard management. We developed a flexible D-vine and minimum information PCC with the same structure to model multivariate data exhibiting complex patterns of dependence in the tails. The developed methodology was used to analyze the dependency structure among flood data collected from Beas basin. In these analyses the developed models in this paper were carefully compared to relevant benchmark models such as multivariate copula model, and particularly multivariate t-copula. However, standard multivariate copulas have added some flexibility, this flexibility is insufficient in higher dimensional applications or the extreme events applications. The pair-copula models can fill this gap by benefiting from the rich class of existing bivariate parametric copula families or more flexible class of non-informative pair-copulas.

In order to compare the proposed models to the standard multivariate copulas, we first select the best trivariate copula to model the joint density of the flood variables. Using the different graphical and analytical goodness-of-fit criterions, the t-copula was chosen as the best trivariate copula. This copula has been chosen as

the most appropriate model in analyzing multivariate flood data in several other studies (see Ganguli and Reddy, 2013 and reference therein). We show that the drawbacks of this copula explained above can be resolved by using the D-vine copula model and minimum information D-vine copula. In addition to the general statistical comparisons between these models, we also computed the primary return periods of the flood data using these copulas and analyzed them in details concluding that the minimum informative pair-copula prediction of the primary return periods was the best and the trivariate t-copula was the worst among these three models. We also calculated the tail dependence coefficients between any pair of the flood variables using these three models and the same results as above were concluded.

We show that the vine model constructed from minimum information copulas cab represent any dependence structure. The minimum information copula can be used to model the multivariate data with various tail dependency, including heavy, symmetric, and nonsymmetric tails. can model from weak to strong upper tail dependence in all of the parametric copulas chosen. The minimum information copula can model from weak to strong upper tail dependence in all of other suitable parametric copulas, including t, Gumbel, and Tawn copulas (see also Bedford, et al., 2015). In this study, we show that the minimum information copula is very useful to precisely estimate the tail dependence coefficients and primary return periods which are very vital in flood hazard management, and would allow improved representation of the interdependencies between flood event peak, event duration and volume to be taken into account in efficient flood analysis.

The minimum information copula we propose here to approximate uncertainty modeling in flood hazard management allows for the common correlation-based approaches to determining dependence, as well as providing a precise probabilistic approximation given a wide range of constraints and uncertainty available in the data. Our approach can be considered as subjectivist approach which follows a tradition in which expectation values are used to specify uncertain quantities. For instance, within a Bayesian approach, the proposed method in this paper may be thought of as a way to generate an informative prior distribution. In the Bayesian framework of risk assessment, the elicitation of a joint probability distribution from experts is among the key research areas, and the minimum information pair-copulas can be considered as a promising way to approximate a multivariate prior distribution based on the experts probabilistic statements. In addition, the pair-copula models can be used in conjunction of MCMC methods to update the models in a probabilistic way which is useful for detailed uncertainty analysis (see Min and Czado (2010) for further details).

Acknowledgements

Special thanks are extended to Dr. Sanjay Jain (National Institute of Hydrology, Roorkee) and Bhakra Beas Management Board for supplying stream flow data of Beas river. In addition, the authors would like to thank the Natural Environment Research Council (NERC) and Indian Ministry of Earth Sciences-funded project on "Mitigating Climate Change Impacts on India Agriculture through Improved Irrigation Water Management" (NE/I022329/1) for funding this work.

Enquiries for access to the data referred to in this article should be directed to researchdata@cranfield.ac.uk.

Appendix A

Assume that we decompose a given three-dimensional $f(x_1,x_2,x_3)$ as follows:

$$f(x_1,x_2,x_3)=f(x_1)f(x_2|x_1)f(x_3|x_1,x_2) \tag{12}$$

Using (2), the following expression can be easily derived

$$f(x_2|x_1)=f(x_2)c_{12}(F(x_1),F(x_2)). \tag{13}$$

where c_{12} is the copula density and $F_1(x_1),F_2(x_2)$ are the marginal distributions.

In addition, we have

$$\begin{aligned} f(x_3|x_1,x_2) &= \frac{f(x_1,x_3|x_2)}{f(x_1|x_2)} \\ &= \frac{c_{13|2}(F(x_1|x_2),F(x_3|x_2))f(x_3|x_2)f(x_1|x_2)}{f(x_1|x_2)} \\ &= c_{13|2}(F(x_1|x_2),F(x_3|x_2))f(x_3|x_2) \end{aligned} \tag{14}$$

Similar to the expression given in (13), $f(x_3|x_2)$ can be written as

$$f(x_3|x_2)=f(x_3)c_{23}(F(x_2),F(x_3)). \tag{15}$$

By substituting (15) into (14), we have

$$f(x_3|x_1,x_2)=c_{13|2}(F(x_1|x_2),F(x_3|x_2))c_{23}(F(x_2),F(x_3))f(x_3) \tag{16}$$

Now, by substituting (13) and (16) into (12), the expression given in (3) will be derived.

References

Aas, K., Czado, K.C., Frigessi, A., Bakken, H., 2009. Pair-copula constructions of multiple dependence. Insur. Math. Econ. 44, 182–198.
Acar, E.F., Genest, C., Neslehova, J., 2012. Beyond simplified pair-copula constructions. J. Multivariate Anal. 110, 74–90.
Bacova-Mitkova, V., Onderka, M., 2010. Analysis of extreme hydrological events on the danube using the peak over threshold method. J. Hydrol. Hydromech. 58, 88–101.
Bayliss, A., 1999. Validation and update of flood peak data. In: Robson, A., Reed, D. (Eds.), Flood Estimation Handbook. Institute of Hydrology, Wallingford, UK.
Bacchi, B., Becciu, G., Kottegoda, N.T., 1994. Bivariate exponential model applied to intensities and durations of extreme rainfall. J. Hydrol. 155 (1/2), 225–236.
Bárdossy, A., 2006. Copula-based geostatistical models for groundwater quality parameters. Water Resour. Res. 42 (11), W11416. http://dx.doi.org/10.1029/2005WR004754.
Bauer, A., Czado, C., Klein, T., 2012. Pair-copula constructions for non-Gaussian DAG models. Can. J. Stat. 40 (1), 86–109.
Bedford, T., Cooke, R.M., 2001. Probability density decomposition for conditionally dependent random variables modeled by vines. Ann. Math. Artif. Intell. 32, 245–268.
Bedford, T., Cooke, R.M., 2002. Vines – a new graphical model for dependent random variables. Ann. Stat. 30 (4), 1031–1068 (random variables modeled by vines. Annals of Mathematics and Artificial Intelligence 32, 245–268).
Bedford, T., Daneshkhah, A., Wilson, K.J., 2015. Approximate uncertainty modeling in risk analysis with vine copulas. Risk Anal. http://dx.doi.org/10.1111/risa.12471.
Bedford, T., Meeuwissen, A., 1997. Minimally informative distributions with given rank correlation for use in uncertainty analysis. J. Stat. Comput. Simulat. 57, 143–174.
Bobee, B., Rasmussen, P.F., 1994. Statistical analysis of annual flood series. In: Trends in Hydrology (1). Council of Scientific Research Integration, India, pp. 117–135.
Brechmann, E.C., Czado, C., 2013. Risk management with high-dimensional vine copulas: an analysis of the Euro Stoxx 50. Stat. Risk Model. 30 (4), 307–342.
Brechmann, E.C., Czado, C., Paterlini, S., 2014. Flexible dependence modeling of operational risk losses and its impact on total capital requirements. J. Bank. Financ. 40, 271–285.
Brechmann, E.C., Schepsmeier, U., 2013. Modeling dependence with C- and D-vine copulas: the R package CDVine. J. Stat. Softw. 52 (3), 1–27.
Capéraá, P., Fougéres, A.L., Genest, C., 1997. A non-parametric estimation procedure for bivariate extreme value copulas. Biometrika 84 (3), 567–577.
Cooke, R.M., Kurowickac, D., Wilson, K., 2015. Sampling, conditionalizing, counting, merging, searching regular vines. J. Multivariate Anal. 138, 4–18.
Czado, C., Min, A., 2010. Bayesian inference for D-vines: estimation and model selection. In: Dependence Modelling. World Scientific.
Czado, C., Gärtner, F., Min, A., 2011. Analysis of Australian electricity loads using joint Bayesian inference of D-Vines with autoregressive margins. In: Kurowicka, Dorota, Joe, Harry (Eds.), Handbook on Vines. World Scientific.
Czado, C., Jeske, S., Hofmann, M., 2013. Selection strategies for regular vine copulae. J. Soc. Fr. Stat. 154 (1), 174–190.
Chowdhary, H., Escobar, L.A., Singh, V.P., 2011. Identification of suitable copulas for bivariate frequency analysis of flood peak and flood volume data. Hydrol. Res. 42 (2–3), 193–216.

Choulakian, V., El-Jabi, N., Moussi, J., 1990. On the distribution of flood volume in partial duration series analysis of flood phenomena. Stoch. Hydrol. Hydraul. 4, 217–226.
Chow, V.T., Maidment, D.R., Mays, L.W., 1988. Applied Hydrology. McGraw Hill, New York.
Clarke, K.A., 2007. A simple distribution-free test for non-nested model selection. Polit. Anal. 15 (3), 347–363.
Cunnane, C., 1988. Methods and merits of regional flood frequency analysis. J. Hydrol. 100, 269–290.
Daneshkhah, A., Parham, G.A., Chatrabgoun, O., Jokar, M., 2015. Approximation multivariate distribution with pair copula using the orthonormal polynomial and Legendre multiwavelets basis functions. Commun. Stat. – Simulat. Comput. http://dx.doi.org/10.1080/03610918.2013.804557.
De Michele, C., Salvadori, G., 2003. A generalized Pareto intensity-duration model of storm rainfall exploiting 2-copulas. J. Geophys. Res.: Atmos. 108 (D2), 40–67.
Dissmann, J., Brechmann, E.C., Czado, C., Kurowicka, D., 2013. Selecting and estimating regular vine copulae and application to financial return. Comput. Stat. Data Anal. 59, 52–69.
Dupuis, D.J., 2007. Using copulas in hydrology: benefits, cautions, and issues. J. Hydrol. Eng. 12 (4), 381–393.
Durrans, S., Eiffe, M., Thomas, W., Goranfio, H., 2003. Joint seasonal/annual flood frequency analysis. J. Hydol. Eng. 8, 181–189.
Évin, G., Favre, A.C., 2008. A new rainfall model based on the Neyman Scott process using cubic copulas. Water Resour. Res. 44, W03433. http://dx.doi.org/10.1029/2007WR006054.
Favre, A.C., El Adlouni, S., Perreault, L., Thiémonge, N., Bobée, B., 2004. Multivariate hydrological frequency analysis using copulas. Water Resour. Res. 40, W01101. http://dx.doi.org/10.1029/2003WR002456.
Gaál, L., Szolgay, J., Kohnová, S., Hlavcová, K., Parajka, J., Viglione, A., Merz, R., Blöschl, G., 2015. Dependence between flood peaks and volumes – a case study on climate and hydrological controls. Hydrol. Sci. J. http://dx.doi.org/10.1080/02626667.2014.951361.
Ganguli, G., Reddy, M.J., 2013. Probabilistic assessment of flood hazards using trivariate copulas. Theor. Appl. Climatol. 111, 341–360.
Genest, C., Favre, A.C., 2007. Everything you always wanted to know about copula modeling but were afraid to ask. J. Hydrol. Eng. 12 (4), 347–368.
Genest, C., Favre, A.C., Béliveau, J., Jacques, C., 2007. Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. Water Resour. Res. 43. http://dx.doi.org/10.1029/2006WR005275.
Genest, C., Rivest, L.P., 1993. Statistical inference procedures for bivariate Archimedean copulas. J. Am. Stat. Assoc. 88, 1034–1043.
Genest, C., Remillard, B., Beausoin, D., 2009. Goodness-of-fit tests for copulas: a review and a power study. Insur. Math. Econ. 44, 199–213.
Gräler, B., 2014. Modelling skewed spatial random fields through the spatial vine copula. Spatial Stat. 10, 87–102.
Gräler, B., Pebesma, E.J., 2011. The pair-copula construction for spatial data: a new approach to model spatial dependency. Proc. Environ. Sci. 7, 206–211. http://dx.doi.org/10.1016/j.proenv.2011.07.036.
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. Hydrol. Earth Syst. Sci. 17, 1281–1296.
Grimaldi, S., Serinaldi, F., 2006a. Asymmetric copula in multivariate flood frequency analysis. Adv. Water Resour. 29 (8), 1155–1167.
Grimaldi, S., Serinaldi, F., 2006b. Design hyetograph analysis with 3-copula function. Hydrol. Sci. J. 51 (2), 223–238.
Gyasi-Agyei, Y., Melching, C.S., 2012. Modelling the dependence and internal structure of storm events for continuous rainfall simulation. J. Hydrol., 249–261
Haff, I., Aas, K., Frigessi, A., 2013. On the simplified pair-copula construction – simply useful or too simplistic? J. Multivariate Anal. 101, 1296–1310.
Hosking, J.R.M., Wallis, J.R., Wood, E.F., 1985. Estimation of the general extreme value distribution be the method of probability weighted moments. Technometrics 27 (3), 251–261.
Huard, D., Évin, G., Favre, A.C., 2006. Bayesian copula selection. Comput. Stat. Data Anal. 51 (2), 809–822.
Jain, S.K., Agarwal, P.K., Singh, V.P., 2007. Indus basin, hydrology and water resources of India. Water Sci. Technol. Libr. 57, 473–511.
Karmakar, S., Simonovic, S.P., 2009. Bivariate flood frequency analysis. Part 2: A copula-based approach with mixed marginal distributions. J. Flood Hazard Manage. 2 (1), 32–44.
Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman & Hall, London.
Joe, H., Li, H., Nikoloulopoulos, A.K., 2010. Tail dependence functions and vine copulas. J. Multivariate Anal. 101, 252–270.
Kidson, R., Richards, K.S., 2005. Flood frequency analysis: assumptions and alternatives. Prog. Phys. Geogr. 29, 392–410.
Krstanovic, P.F., Singh, V.P., 1987. A multivariate stochastic flood analysis using entropy. In: Singh, V.P. (Ed.), Hydrologic Frequency Modeling. Reidel, Dordrecht, The Netherlands, pp. 515–539.
Kurowicka, D., Cooke, R., 2006. Uncertainty Analysis with High Dimensional Dependence Modeling. John Wiley.
Kurowicka, D., Joe, H., 2011. Dependence Modeling: Vine Copula Handbook. World Scientific, Singapore.
Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E., 1998. Convergence properties of the Nelder–Mead simplex method in low dimensions. SIAM J. Optimiz. 9 (1), 112–147.
Leonard, M., Metcalfe, A., Lambert, M., 2008. Frequency analysis of rainfall and streamflow extremes accounting for seasonal and climatic partitions. J. Hydrol. 348, 135–147.
Ljung, G.M., Box, G.E.P., 1978. On a measure of a lack of fit in time series models. Biometrika 65 (2), 297–303.
Lopez-Paz, D., Hernandez-Lobato, J.M., Ghahramani, Z., 2013. Gaussian process vine copulas for multivariate dependence. Proceedings of the 30th International Conference on Machine Learning, vol. 28, pp. 10–18.
Ma, M.W., Song, S.B., Ren, L.L., Jiang, S.H., Song, J.L., 2013. Multivariate drought characteristics using trivariate Gaussian and Student t copulas. Hydrol. Process. 27, 1175–1190.
Min, A., Czado, C., 2010. Bayesian inference for multivariate copulas using pair-copula constructions. J. Financ. Econom. 8, 511–546.
Nadarajah, S., Gupta, A.K., 2006. Some bivariate gamma distributions. Appl. Math. Lett. 19, 767–774.
Nelsen, R.B., 2006. An Introduction to Copulas. Springer, New York.
Salvadori, G., 2004. Bivariate return periods via 2-copulas. Stat. Methodol. 1, 129–144.
Salvadori, G., De Michele, C., 2006. Statistical characterization of temporal structure of storms. Adv. Water Resour. 29 (6), 827–842.
Salvadori, G., De Michele, C., 2007. On the use of copulas in hydrology: theory and practice. J. Hydrol. Eng. 12, 369–380.
Salvadori, G., De Michele, C., 2010. Multivariate multiparameter extreme value models and return periods: a copula approach. Water Resour. Res. 46, W10501.
Salvadori, G., De Michele, C., Kottegodam, N.T., Rosso, R., 2007. Extremes in Nature: An Approach Using Copulas. Springer Velag.
Schirmacher, D., Schirmacher, E., 2008. Multivariate Dependence Modeling using Paircopulas. Tech. Rep., Society of Acturaries, Enterprise Risk Management Symposium, April 14-16, Chicago, 2008. <http://www.soa.org/library/monographs/othermonographs/2008/april/2008-erm-toc.aspx>.
Serinaldi, F., 2013. An uncertain journey around the tails of multivariate hydrological distributions. Water Resour. Res. 49 (10), 6527–6547.
Serinaldi, F., Grimaldi, S., 2007. Fully nested 3-copula: procedure and application on hydrological data. J. Hydrol. Eng. 12 (4), 420–430.
Shiau, J.T., 2006. Fitting drought duration and severity with two-dimensional copulas. Water Resour. Manage. 20, 795–815.
Silva, R.S., Lopes, H.F., 2008. Copula, marginal distributions and model selection: a Bayesian note. Stat. Comput. 18, 313–320.
Sklar, A., 1959. Fonctions de repartition a n dimensions e leurs marges. Publ. l'Inst. Stat. l'Univiversit Paris 8, 229–231.
Song, S.B., Kang, Y., 2011. Pair-copula decomposition constructions for multivariate hydrological drought frequency analysis. Proc. 2011 International Symposium on Water Resource and Environmental Protection (ISWREP), vol. 4, pp. 2635–2638.
Song, S., Singh, V.P., 2010. Meta-elliptical copulas for drought frequency analysis of periodic hydrologic data. Environ. Res. Hazard Assess. 24 (3), 425–444.
Sraj, M., Bezak, N., Brilly, M., 2014. Bivariate flood frequency analysis using the copula function: a case study of the Litija station on the Sava River (in Press). Hydrol. Process. http://dx.doi.org/10.1002/hyp.10145.
Stöeber, J., Joe, H., Czado, C., 2013. Simplified pair copula constructions – limitations and extensions. J. Multivariate Anal. 119, 101–118.
Vernieuwe, H., Vandenberghe, S., De Baets, B., Verhoest, N.E.C., 2015. A continuous rainfall model based on vine copulas. Hydrol. Earth Syst. Sci. Discuss. 12, 489–524. http://dx.doi.org/10.5194/hessd-12-489-2015.
Venter, G., Barnett, J., Kreps, R., Major, J., 2007. Multivariate copulas for financial modeling. Variance 1 (1), 103–119.
Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 (2), 307–333.
Wang, X., Gebremichael, M., Yan, J., 2010. Weighted likelihood copula modeling of extreme rainfall events in connecticut. J. Hydrol. 390 (1–2), 108–115.
Xiong, L., Yu, K.X., Gottschalk, L., 2014. Estimation of the distribution of annual runoff from climatic variables using copulas. Water Resour. Res. 50, 7134–7152.
Yan, B.W., Guo, S.G., Xiao, Y., 2007. Analysis on drought characteristics based on bivariate joint distribution. Arid. Zone. Res. 24, 537–542.
Yue, S., Ouarda, T.B.M.J., Bobèe, B., 2001. A review of bivariate gamma distributions for hydrological application. J. Hydrol. 246, 1–18.
Yue, S., Wang, C.Y., 2004. A comparison of two bivariate extreme value distributions. Stoch. Environ. Res. 18, 61–66.
Zhang, L., Singh, V.P., 2007. Bivariate rainfall frequency distributions using Archimedean copulas. J. Hydrol. 332, 93–109.
Zhang, Q., Li, J., Singh, V.P., 2012. Application of Archimedean Copulas in the analysis of the precipitation extremes: effects of precipitation changes. Theoret. Appl. Climatol. 107 (1), 255–264.