Hydrol. Earth Syst. Sci. Discuss., 7, 7563–7590, 2010
www.hydrol-earth-syst-sci-discuss.net/7/7563/2010/
doi:10.5194/hessd-7-7563-2010

Hydrology and Earth System Sciences Discussions

Estimating strategies for Multiparameter Multivariate Extreme value copulas

G. Salvadori[1] **and C. De Michele**[2]

[1]Dipartimento di Matematica, Università del Salento, Provinciale Lecce-Arnesano, P.O. Box 193, 73100 Lecce, Italy
[2]DIIAR (Sezione CIMI), Politecnico di Milano, Piazza Leonardo da Vinci 32, 20132 Milano, Italy

Received: 19 July 2010 – Accepted: 20 July 2010 – Published: 4 October 2010

Correspondence to: G. Salvadori (gianfausto.salvadori@unisalento.it)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Multivariate Extreme Value models are a fundamental tool in order to assess potentially dangerous events. Exploiting recent theoretical developments in the theory of Copulas, new multiparameter models can be easily constructed. In this paper we suggest several strategies in order to estimate the parameters of the selected copula, according to different criteria: these may use either a nearest neighbor or a nearest cluster approach, or exploit all the pair-wise relationships between the available gauge stations. An application to flood data is also illustrated and discussed.

1 Introduction

Multivariate extremes occur in several hydrologic problems (like, e. g., space-time precipitation and floods; Singh, 1986; Pons, 1992; Wilks, 1998; Kim et al., 2003; Herr and Krzysztofowicz, 2005; Keef et al., 2009, or hydraulic conductivity in porous media; Journel and Alabert, 1988; Russo, 2009), as well as in many environmental problems (like, e. g., water quality and pollution; Grenney and Heyse, 1985, or sea levels; Butler et al., 2007).

The investigation of multivariate phenomena is best carried out via copulas. The use of copulas in hydrology, as well as in other geophysical and environmental sciences, is recent and rapidly growing. Incidentally, we observe that all the multivariate distributions present in literature can be described in a straightforward way in terms of suitable copulas. For a thorough bibliography see Nelsen (2006); Salvadori et al. (2007).

In modeling multivariate extremes, central is the issue of how to measure the dependence between the variables involved. In literature, the pair-wise dependence is generally measured by the canonical Pearson's correlation coefficient. However, it may not be the best measure of dependence when dealing with extremes (Joe, 1997), since it does not exist for heavy-tailed variables with infinite variance, and only involves a linear kind of dependence. Other pair-wise measures were recently considered

Nelsen (2006) to model the association between pairs of random variables (hereafter, r.v.s): among others, Kendall's τ and Spearman's ρ rank correlation coefficients, or the Blomqvist's β medial correlation coefficient. These measures always exist (being based on the ranks), and model several types of association (for a practical discussion see, e. g., the case studies illustrated in Salvadori et al., 2007).

Instead, the notion of cluster-type dependence (when the size of the cluster size is larger than two), has only been partially explored. Generalizations of Kendall's τ (Nelsen, 1996), Spearman's ρ (Schmid and Schmidt, 2007a,b), and Blomqvist's β (Durante et al., 2007; Schmid and Schmidt, 2007c) to the d-variate case ($d>2$) were recently introduced – see below. These extensions may be of practical importance: on the one hand, they provide useful tools to quantify the dependence within clusters; on the other hand, they can be used to estimate the parameters of the multivariate model at play (see later). However, at present the application of these measures in actual case studies is still quite limited.

Another important issue is represented by the construction of Multivariate Extreme Value (hereafter, MEV) models involving a significant number of parameters. Using the results of Liebscher (2008), recent works (Durante and Salvadori, 2010; Salvadori and De Michele, 2010) have shown how multiparameter MEV models can be easily constructed via copulas and suitable techniques of extra-parameterization, leading to the formulation of new models, or the generalization of existing ones.

A further fundamental question is represented by the estimate of the parameters of the multivariate copulas considered (see Genest et al., 1995; Shih and Louis, 1995; Joe, 1997; Genest and Favre, 2007, and references therein). Maximum Likelihood (hereafter, ML) or Pseudo-likelihood procedures involving the ranks of the data are generally used to fit these parameters. Alternatively, the parameters may be sometimes estimated via the method of moments and some pair-wise measures of association (usually, the Kendall's τ, the Spearman's ρ, or the Blomqvist's β). Apparently, no application of the d-variate generalizations of these measures to the parameters' estimation is available in literature.

HESSD
7, 7563–7590, 2010

G. Salvadori and C. De Michele

In this paper we focus the attention on the estimation of the parameters in copula-based MEV models, presenting some new fitting strategies. Each procedure exploits a different source of information: (i) the nearest station, (ii) the closest cluster of stations, (iii) all the pair-wise dependencies of the available stations.

Below, in Sect. 2 we introduce the concept of Multivariate Extreme Value copulas, describing some of the mathematical features of interest here. In Sect. 3 we show several strategies for estimating the relevant parameters. In Sect. 4 an application to maximum annual flood data is presented and discussed.

2 MEV copulas: an overview

In this section we briefly outline the mathematics of copulas needed in the sequel; for a thorough theoretical and practical introduction see, respectively (Joe, 1997; Nelsen, 2006, and Salvadori et al., 2007). Hereafter, for any integer $d>1$, we use the vector notation in $\boldsymbol{R}^d$, i. e. $\boldsymbol{x}=(x_1,\ldots,x_d)$; operations and inequalities are to be intended componentwise. Also, $\boldsymbol{I}=[0,1]$ will denote the unit interval, and $\boldsymbol{I}^d$ the d-dimensional unit cube.

The main target pursued here is to provide a general multivariate framework for modeling non-independent extreme observations sampled via a network of gauge stations; the particular situation of independent ones will be included as a special case. As shown below, this can easily be achieved by using copulas. The r.v.s used in the sequel may represent, for instance, rainfall or flood measurements collected in a given basin, or pollution samples in a region, or wave measurements collected by marine buoys. Below, $\mathcal{S}=\{S_1,\ldots,S_d\}$ will denote a set of d gauge stations.

The problem of specifying a probability model for dependent multivariate observations can be simplified by expressing the corresponding d-dimensional joint distribution F in terms of its margins $F_1,\ldots,F_d$, and the associated copula $\boldsymbol{C}$, implicitly defined

HESSD
7, 7563–7590, 2010

through the following functional identity stated by Sklar's Theorem (Sklar, 1959):

$$F(x_1,\ldots,x_d) = \boldsymbol{C}(F_1(x_1),\ldots,F_d(x_d)). \tag{1}$$

A multivariate copula $\boldsymbol{C}(u_1,\ldots,u_d)$ is simply a joint distribution over $\boldsymbol{I}^d$ with uniform margins. The link between d-copulas and multivariate distributions is provided by Eq. (1). If $F_1,\ldots,F_d$ are all continuous, then $\boldsymbol{C}$ is unique.

A copula $\boldsymbol{C}$ is MEV if it is max-stable, i. e. if it satisfies the equation

$$\boldsymbol{C}\left(u_1^t,\ldots,u_d^t\right) = \left[\boldsymbol{C}(u_1,\ldots,u_d)\right]^t \tag{2}$$

for all $\boldsymbol{u} \in \boldsymbol{I}^d$ and all $t>0$. As a simple example consider the following two copulas:

$$\boldsymbol{\Pi}_d(\boldsymbol{u}) = u_1 \cdots u_d, \tag{3}$$

$$\boldsymbol{M}_d(\boldsymbol{u}) = \min\{u_1,\ldots,u_d\}. \tag{4}$$

The former one models independent variates, while the latter one models comonotone dependent ones, where each variable is a monotone increasing function of the others. Evidently, both $\boldsymbol{\Pi}_d$ and $\boldsymbol{M}_d$ are max-stable, and hence MEV. A distribution F is MEV if, and only if, all its margins F_i's are Generalized Extreme Value laws (hereafter, GEV), and the corresponding copula $\boldsymbol{C}$ is MEV. Note that not all copulas are MEV (i. e., satisfy the max-stability property Eq. 2), and consequently should not be used to construct consistent MEV models. In addition, since the GEV law is continuous, the representation $F=\boldsymbol{C}(F_1,\ldots,F_d)$ of a MEV distribution F is unique. Most importantly, by exploiting the invariance property of copulas (Nelsen, 2006), we may restrict our attention to copulas only, and do not worry about the GEV margins, as we shall do hereinafter.

The construction of multivariate measures of association and/or dependence is an involved mathematical problem, and is still an open question in statistics. Several ideas were developed in the last few years, and various measures were introduced in order to describe concepts like, e. g., concordance for random vectors (Joe, 1990; Nelsen, 1996, 2002; Úbeda-Flores, 2005; Schmid and Schmidt, 2007a; Taylor, 2007).

HESSD
7, 7563–7590, 2010

Estimating strategies for M-MEV copulas

G. Salvadori and C. De Michele

For bivariate problems, several measures of association are available (Joe, 1997; Nelsen, 2006). Among others, Kendall's τ and Spearman's ρ are frequently used in applications. The former one is the difference between the probability of concordance and discordance of the variables, the latter one measures the average distance between $\mathbf{\Pi}_2$ (i. e., independence) and the bivariate copula of interest. As is well known, these measures only depend upon the copula joining the variables under investigation, and not upon the margins (i. e., they are scale invariant). As already mentioned above, a further advantage is that, if the variables involved are characterized by heavy-tailed distributions, then the second order moment (and, in turn, Pearson's coefficient) may not exist, whereas these latter measures always exist, being based on the ranks.

Interesting extensions of Kendall's τ (Nelsen, 1996) and Spearman's ρ (Schmid and Schmidt, 2007a,b) to a general d-variate framework ($d>2$) were recently proposed, and several new measures involving the generic d-copula $\boldsymbol{C}$ were introduced:

$$\tau_d = \frac{1}{2^{d-1}-1}\left(2^d \int_{\boldsymbol{I}^d} \boldsymbol{C}(\boldsymbol{u})\,d\boldsymbol{C}(\boldsymbol{u}) - 1\right), \tag{5}$$

$$\rho_{d,1} = h(d)\left(2^d \int_{\boldsymbol{I}^d} \boldsymbol{C}(\boldsymbol{u})\,d\boldsymbol{u} - 1\right), \tag{6}$$

$$\rho_{d,2} = h(d)\left(2^d \int_{\boldsymbol{I}^d} \mathbf{\Pi}_d(\boldsymbol{u})\,d\boldsymbol{C}(\boldsymbol{u}) - 1\right), \tag{7}$$

$$\rho_{d,3} = h(2)\left(2^2 \sum_{i<j} \binom{d}{2}^{-1} \int_{\boldsymbol{I}^2} \boldsymbol{C}_{ij}(u,v)\,du\,dv - 1\right), \tag{8}$$

where $h(d)=(d+1)/(2^d-(d+1))$, and $\boldsymbol{C}_{ij}$ is the bivariate (i,j)-margin of $\boldsymbol{C}$. Note that $\rho_{d,3}$ is essentially the average Spearman's ρ for all the pairs in a set of d variables.

HESSD
7, 7563–7590, 2010

Estimating strategies for M-MEV copulas

G. Salvadori and C. De Michele

Another useful multivariate measure of association is the medial correlation coefficient β_d (see Durante et al., 2007; Schmid and Schmidt, 2007c, and references therein), which generalizes the well known Blomqvist's β coefficient (Nelsen, 2006):

$$\beta_d = \frac{2^{d-1}(\boldsymbol{C}(\mathbf{1/2}) + \overline{\boldsymbol{C}}(\mathbf{1/2})) - 1}{2^{d-1} - 1}, \tag{9}$$

where $\overline{\boldsymbol{C}}$ is the survival function associated with $\boldsymbol{C}$, given by $\overline{\boldsymbol{C}}(\boldsymbol{u}) = \boldsymbol{P}\{\boldsymbol{C} > \boldsymbol{u}\}$, and $\mathbf{1/2} = (1/2, \ldots, 1/2)$. Clearly, also β_d is invariant with respect to the distributions of the margins. As pointed out in Schmid and Schmidt (2007c), β_d has some advantages over competing measures such as τ_d or $\rho_{d,i}$'s. In fact, it can explicitly be derived whenever the copula is of explicit form, which is often not possible for other measures, and its estimation requires a low computational complexity. Thus, β_d may represent a fast alternative for estimating the copula parameters (see below).

A further notion of interest is represented by Pickands' dependence function A (Pickands, 1981). Recall that a bivariate copula $\boldsymbol{C}$ is MEV if there exists a convex function $A : \boldsymbol{I} \to [1/2, 1]$, satisfying the constraint $\max\{t, 1-t\} \le A(t) \le 1$ for all $t \in \boldsymbol{I}$, such that

$$\boldsymbol{C}(u, v) = \exp\left[\ln(uv) A\left(\frac{\ln v}{\ln(uv)}\right)\right] \tag{10}$$

for all $(u, v) \in \boldsymbol{I}^2$. In particular, if $A(t) \equiv 1$ then $\boldsymbol{C} = \boldsymbol{\Pi}_2$, and if $A(t) = \max\{t, 1-t\}$ then $\boldsymbol{C} = \boldsymbol{M}_2$. Conversely, given a bivariate MEV copula $\boldsymbol{C}$, the corresponding dependence function A is given by

$$A(t) = -\ln \boldsymbol{C}(e^{-(1-t)}, e^{-t}), \tag{11}$$

where $t \in \boldsymbol{I}$. It is worth noting that the value τ_C of the Kendall's τ associated with $\boldsymbol{C}$, as well as the one of the Spearman's ρ, can be expressed in terms of A via (Nelsen, 2006; Salvadori et al., 2007)

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD

7, 7563–7590, 2010

Estimating strategies for M-MEV copulas

G. Salvadori and C. De Michele

$$\tau_C = \int_0^1 \frac{t(1-t)}{A(t)} dA'(t) \tag{12}$$

and

$$\rho_C = 12 \int_0^1 \frac{1}{(1+A(t))^2} dt - 3\,. \tag{13}$$

A generalization of Pickands' dependence function to the multivariate case is shown in Falk and Reiss (2005). Since A can be estimated via empirical data (Genest and Segers, 2009), then it may be used to check the statistical adequacy of different models. We shall see later how to use Pickands' dependence function.

Finally, below we shall also use the Kendall's measure function K_C (Genest and Rivest, 1993, 2001) given by

$$K_C(t) = \boldsymbol{P}\{W \le t\} = \boldsymbol{P}\{\boldsymbol{C}(U_1,\ldots,U_d) \le t\}\,, \tag{14}$$

where $t \in \boldsymbol{I}$ is a probability level, $W = \boldsymbol{C}(U_1,\ldots,U_d)$ is a univariate r.v. taking value on $\boldsymbol{I}$, and the U_i's are Uniform r.v.s on $\boldsymbol{I}$ with copula $\boldsymbol{C}$. In the bivariate Extreme Value case, K_C is given by Ghoudi et al. (1998)

$$K_C(t) = t - (1-\tau_C)t \ln t\,, \tag{15}$$

where τ_C is the value of the Kendall's τ associated with the copula $\boldsymbol{C}$. Clearly, bivariate MEV copulas with the same value of τ share the same function K_C. Unfortunately, at present no useful expressions similar to Eq. (15) are known for the general multivariate case $d>2$.

The Kendall's measure K_C is a fundamental tool for introducing a mathematically consistent (copula-based) definition of the return period for multivariate events (see

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD

7, 7563–7590, 2010

also the discussion in Salvadori, 2004; Salvadori and De Michele, 2004, 2010; Salvadori et al., 2007; Durante and Salvadori, 2010). In fact, Eq. (14) represents a multivariate quantile relationship, since it corresponds to a multidimensional Probability Integral Transform (Genest et al., 2006).

Let μ be the average interarrival time of the events in the sequence observed (e. g., μ=1 year for annual maxima), and let $p \in \boldsymbol{I}$ be an arbitrary critical probability level (usually, p=90, 95, 99%, or any other threshold of interest). The multivariate return period T_{p} associated with p is defined as

$$T_{\mathrm{p}} = \frac{\mu}{1-p} = \frac{\mu}{1-K_{\boldsymbol{C}}(t)} = \frac{\mu}{1-\boldsymbol{P}\{\boldsymbol{u} \in \boldsymbol{I}^d : \boldsymbol{C}(\boldsymbol{u}) \leq t\}}, \tag{16}$$

where the critical threshold $t \in \boldsymbol{I}$ is given by

$$t = \inf\{s \in \boldsymbol{I} : K_{\boldsymbol{C}}(s) = p\} = K_{\boldsymbol{C}}^{[-1]}(p), \tag{17}$$

by analogy with the correct definition of quantile. Here $K_{\boldsymbol{C}}^{[-1]}$ indicates the generalized (or pseudo-) inverse (Nelsen, 2006) of the corresponding function. Since $K_{\boldsymbol{C}}$ is generally non-linear ($K_{\boldsymbol{C}}(t)=t$ only if $\boldsymbol{C}=\boldsymbol{M}_d$), then $t \neq p$. More particularly, the relation $K_{\boldsymbol{C}}(t) \geq t$ holds (Capéraà et al., 1997), and therefore

$$T_{\mathrm{p}} = T_{K_{\boldsymbol{C}}(t)} = \frac{\mu}{1-K_{\boldsymbol{C}}(t)} \geq \frac{\mu}{1-t} = \frac{\mu}{1-\boldsymbol{C}(\boldsymbol{u})}, \tag{18}$$

where $\boldsymbol{u} \in \boldsymbol{I}^d$ is such that $\boldsymbol{C}(\boldsymbol{u})=t$. The right-most term corresponds to the standard definition of multivariate return period (for a thorough review see Zhang, 2005; Singh et al., 2007, and references therein). Evidently, the traditional approach may yield an incorrect calculation of the return period, and, in turn, a wrong estimation of the risk. Since empirical estimators of the Kendall's measure function are available (Genest et al., 2009), we shall see later how to use them to perform a return period analyses of practical utility.

3 Parameters' estimation

As is well known, the estimate of the parameters of multivariate distributions is an involved problem. Usually, procedures like Maximum Likelihood are used to fit simultaneously all the parameters of interest. However, if, e. g., the copulas under investigation have singular components, then ML may be difficult to implement and use.

Below, we outline several approaches for estimating the parameters of interest: each procedure will exploits different sources of information, and estimations achieved via different techniques will generally differ from one another. For instance, the estimate may rely only upon the information drawn from the nearest station (Sect. 3.1), or the closest cluster of stations (Sect. 3.2), or all the pair-wise dependencies of the d stations (Sect. 3.3). The methods are quite general, and can be applied to any MEV copula, including those with singular components. Clearly, other approaches are possible, depending upon the specific needs. Note that the estimates calculated via the methods mentioned above could be used as starting guesses for running other procedures (e. g., ML).

It is worth pointing out that below we shall mainly use the Pickands' dependence function: this approach is not affected by the presence of singular components in the copula model. As an alternative, also the Kendall's measure function could be used. However, while the former is specific for any copula, the latter is not, for it only depends upon the corresponding value of the Kendall's τ – see Eq. (15), and the comment following it. Therefore, we prefer to use Pickands' representation.

Overall, the strategies presented below represent a physically based approach to the estimation problem. Generally, the fitting criterion is given by the best agreement, in the Least Squares sense (hereafter, LS), with the local dependence structures: clearly, this may yield estimates different from the ones achieved via other fitting procedures (e. g., the global ML). However, the overall fitting ability will always be certified via global Goodness-of-Fit tests (see Sect. 4), in order to verify whether the resulting model could be accepted or not.

3.1 The nearest neighbor approach

The first approach we propose for the estimate of the parameters of interest exploits a nearest neighbor principle, i. e. we only use the information drawn by the closest station. Let S_i be the i-th gauge station, and let $S_j = S_{j(i)}$ be the station closest to it. Note that, except for mathematically "pathological" cases of no interest here, usually S_j is unique: a counter-example is given by a (practically improbable) situation in which several stations are exactly positioned on a circle centered in S_i. Using the standard Euclidean metrics for the distance Δ_{ij} between S_i and S_j, only two things may happen:

1. either $\{S_i, S_j\}$ forms a pair-cluster (i. e., S_i is the station closest to S_j, and vice-versa),
2. or, there exists another station S_k closer to S_j than S_i; clearly, S_k may belong to a pair-cluster as defined in Eq. (1).

From a geometrical point of view, at least a couple of stations must form a pair-cluster. In fact, the set of $N_d = d(d-1)/2$ pair distances Δ_{ij}'s is finite, and hence it has (at least) a minimum: this corresponds to a pair-cluster.

Now, we may estimate the parameters via a LS fit, involving the empirical estimates of the Pickands' dependence functions A_{ij}'s of the (nearest neighbour) pairs. The procedure is as follows. Let n be the sample size, i. e. the number of available d-dimensional observations; we assume that $n \geq d$. For each station S_i, $i = 1, \ldots, d$, the closest station $S_j = S_{j(i)}$ is identified, and an estimate $\hat{A}_{ij}$ of the dependence function A of the model under investigation is calculated (Genest and Segers, 2009). In particular, in order to use all the information, since only n bivariate pairs are available, and given the constraints $A(0) = A(1) = 1$, the unit interval $\boldsymbol{I}$ is partitioned into n uniformly spaced intervals via the set of abscissas $x_k = k/n$, $k = 0, \ldots, n$ (clearly, other choices are possible).

Then, the $\hat{A}_{ij}(x_k)$'s are estimated over the given grid, and the LS objective function

$$Z^{(1)} = \sum_{i=1}^{d} Z_i^{(1)} = \sum_{i=1}^{d} \sum_{k=1}^{n-1} \left| A_{i,j(i)}(x_k) - \hat{A}_{i,j(i)}(x_k) \right|^2 \tag{19}$$

is minimized, yielding the LS estimates of the parameters of interest. Note that, if $\{S_i, S_{j(i)}\}$ is a pair-cluster, then there is no need to compute also the (symmetric) contribution of the pair $\{S_{i'=j(i)}, S_{j(i')=i}\}$: this may reduce the computational burden. Essentially, the nearest neighbor approach (hereafter, 1-MEV) exploits the relationships of the S_i's with the closest station, i. e. it uses the local (station based) bivariate dependence structures.

It is worth pointing out that, if the model involves global parameters (i. e., common to all stations), and these can be estimated a priori via other techniques, then the local parameters (if any) can be calculated as follows. For each station S_i, the closest station S_j is identified, and the local parameters are estimated via a LS fit of the dependence function A_{ij}, using the values of the global parameters already estimated (i. e., only $Z_i^{(1)}$ is minimized). If $\{S_i, S_{j(i)}\}$ is a pair-cluster, then all the estimates of the local parameters associated with S_i and S_j are kept; otherwise, only those associated with S_i are stored. This latter strategy can easily deal with sets of stations of any size: in fact, only two stations at a time are considered for estimating the local parameters. In other words, a global estimate is necessary only if the global parameters cannot be estimated otherwise.

3.2 The cluster approach

The nearest neighbor approach adopted in the previous section only exploited the information drawn by the closest station. This strategy can be extended to a larger set of stations closest to the one of interest. Let S_i be the i-th gauge station, and let $C_i^{(m)}$ be the cluster of the m stations closest to S_i, with $1<m<d$ (the case $m=1$ was already dealt with via the 1-MEV approach). Clearly, the choice of m can be made dependent

upon, e. g., an arbitrary distance of influence, or specific basin characteristics, or may be changed when considering different stations. Again, except for "pathological" cases, usually $C_i^{(m)}$ is uniquely defined.

Here the idea is to estimate the parameters by exploiting a multivariate measure of association ϕ_C calculated over the family of stations $\mathcal{F}_i = \{S_i \cup C_i^{(m)}\}$, having size $m+1$. For instance, any of the five measures outlined in Eqs. (5)–(9) could be used. First, for each station S_i, an estimate $\hat{\phi}_i$ of ϕ_C is calculated over the cluster $\mathcal{F}_i$. Then, the LS objective function

$$Z^{(m)} = \sum_{i=1}^{d} Z_i^{(m)} = \sum_{i=1}^{d} \left| \phi_C^{(\mathcal{F}_i)} - \hat{\phi}_i \right|^2 \tag{20}$$

is minimized, yielding the LS estimates of the parameters of interest. Note that, if $\mathcal{F}_i$ is a closed $(m+1)$-cluster (i. e., the closest m stations to any station in $\mathcal{F}_i$ again belong to $\mathcal{F}_i$), then the contribution of the cluster can be calculated only once: this may reduce the computational burden. Essentially, the nearest cluster approach (hereafter, c-MEV) exploits the relationships of the S_i's with the closest m-cluster of stations, i. e. it is based on the local m-variate dependence structures.

Again, it is worth pointing out that, if the model involves global parameters, and these can be estimated a priori via other techniques, then the local parameters (if any) can be calculated as follows. For each station S_i, the closest m-cluster $C_i^{(m)}$ is identified, and the local parameters are estimated via a LS fit of $\phi_C^{(\mathcal{F}_i)}$, using the values of the global parameters already estimated. If $\mathcal{F}_i$ is a closed $(m+1)$-cluster, then all the estimates of the local parameters associated with the stations in $\mathcal{F}_i$ are kept; otherwise, only those associated with S_i are stored. Thus, a global estimate is necessary only if the global parameters cannot be estimated otherwise.

3.3 The p-MEV approach

A further approach for the estimate of the parameters may rely upon the use of all the $d(d-1)/2$ bivariate margins, by simultaneously considering the dependence structures of all the pairs of stations. The simplest strategy is to fix all the parameters in such a way that the Pickands' dependence functions A_{ij}'s best fit (in the LS sense) the corresponding empirical ones. The LS objective function to be minimized is given by

$$Z^{(p)} = \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \sum_{k=1}^{n-1} \left| A_{i,j}(x_k) - \hat{A}_{i,j}(x_k) \right|^2 , \tag{21}$$

yielding the LS estimates of the parameters of interest. We call this method p-MEV approach.

By exploiting the same strategy, a faster alternative would be to calculate the parameters by simultaneously fitting all the bivariate Kendall's τ, or Spearman's ρ, or Blomqvist's β coefficients (or any other measure of association): essentially, this corresponds to a method of moments procedure. However, while the use of Pickands' function involves the full functional form of the dependence structure (which is specific for every copula), the coefficients mentioned above may not distinguish between different copulas. For this reason, we shall not investigate this point here.

4 Case study

For the sake of illustration, here we consider the same data and copulas used in Salvadori and De Michele (2010), to which we make reference for further details: a short summary is reported below.

The data are maximum annual flood measurements collected in the Spey catchment (Northern Highlands of Scotland). The basin is equipped with a network of 17 flow gauge stations, and is managed by the Scottish Environment Protection Agency (2009). Further details can be found in Gilvear (2004) and Black and Fadipe (2009). In

HESSD
7, 7563–7590, 2010

this study we consider four gauge stations located in the middle and lower part of the Spey catchment (see Fig. 1): three on the main stream (i. e., S_2, S_{10}, and S_6), and one on Dulnain tributary (i. e., S_9).

The available observations amount to 37 quadruples of maximum annual floods. Evidently, from a statistical point of view, the sample size is very small for investigating a multivariate problem: unfortunately, this is a typical situation when extreme data bases are considered. However, here the target is not to provide an ultimate extreme flood model, and no design of actual structures is involved. Instead, our point is only to show, in a relatively simple case, how the techniques outlined above can be used in practice: in other words, this is a methodological paper.

As a dependence model, here we use the multiparameter MEV copula $\boldsymbol{H}$ introduced in Salvadori and De Michele (2010):

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{u}) &= \boldsymbol{G}_{\xi}(\boldsymbol{u}^{\boldsymbol{a}}) \times \boldsymbol{G}_{\chi}(\boldsymbol{u}^{\boldsymbol{1}-\boldsymbol{a}}) \\ &= \boldsymbol{G}_{\xi}\left(u_1^{a_1}, u_2^{a_2}, u_3^{a_3}, u_4^{a_4}\right) \times \boldsymbol{G}_{\chi}\left(u_1^{1-a_1}, u_2^{1-a_2}, u_3^{1-a_3}, u_4^{1-a_4}\right), \end{aligned} \tag{22}$$

with Gumbel parameters $\xi, \chi \geq 1$, and "extra-parameters" $a_1, a_2, a_3, a_4 \in \boldsymbol{I}$, which represents a 4-variate generalization of the well known Gumbel copula $\boldsymbol{G}_{\theta}$ (Nelsen, 2006; Salvadori et al., 2007)

$$\boldsymbol{G}_{\theta}(\boldsymbol{u}) = \exp\left\{-\left[\sum_{i=1}^{4}(-\ln u_i)^{\theta}\right]^{1/\theta}\right\}, \tag{23}$$

with parameter $\theta \geq 1$. Note that the Gumbel copula $\boldsymbol{G}_{\theta}$ represents a sort of "standard" MEV model in hydrology (see, e. g., Yue, 2000a,b; Zhang and Singh, 2007, and references therein). A straightforward interpretation of the parameters a_i's is as follows. Suppose that $\boldsymbol{a}=\boldsymbol{1}$: then, $\boldsymbol{H}=\boldsymbol{G}_{\xi}$. Conversely, should it be $\boldsymbol{a}=\boldsymbol{0}$, then $\boldsymbol{H}=\boldsymbol{G}_{\chi}$. For other values of $\boldsymbol{a}$, $\boldsymbol{H}$ is a sort of "mixture" between $\boldsymbol{G}_{\xi}$ and $\boldsymbol{G}_{\chi}$: in particular, the a_i's play the role of "local" mixing parameters.

HESSD
7, 7563–7590, 2010

G. Salvadori and C. De Michele

The generic bivariate dependence function A_{ij} of $\boldsymbol{H}$ is

$$A_{ij}(t) = \left\{[(1-a_i)(1-t)]^{\chi} + [(1-a_j)t]^{\chi}\right\}^{1/\chi} + \left\{[a_i(1-t)]^{\xi} + [a_j t]^{\xi}\right\}^{1/\xi}, \tag{24}$$

i.e. a non-linear, possibly asymmetric, function, able to model non-exchangeable variables (an important feature in applications, not shared by $\boldsymbol{G}_\theta$ – see, e.g., the discussion in Grimaldi and Serinaldi, 2006).

In Table 1 (upper triangular) we show the inter-station distances Δ's. It is then immediate to identify, for each site, the nearest neighbor station (or cluster), as outlined in Sect. 3.1 (or Sect. 3.2): namely, $S_2 \leftarrow S_9$, $S_6 \leftarrow S_{10}$, $S_9 \leftrightarrow S_{10}$, i.e. the latter two stations form a pair-cluster (see Table 1, diagonal). In Table 1 (lower triangular) we show the empirical estimates of the bivariate Kendall's τ, for all the pairs of the four stations of interest here. It is interesting to note that the coefficient is very small for the two farthest stations $\{S_2, S_6\}$: this means that the association between the two is negligible, as confirmed by the corresponding p-values, though this does not imply that the stations are statistically independent (as, instead, is commonly misinterpreted). On the contrary, the analysis of the p-values shows that the estimates of the coefficients for all the other pairs are statistically significantly different from zero: this means that the corresponding stations are definitely dependent.

Below we shall statistically compare the performances of the copula model provided by Eq. (22), using sets of parameters fitted via different methods. The estimates are reported in Table 2. For the sake of shortness, here we do not use the c-MEV approach: as a matter of facts, the necessarily small size of the clusters in a set of only four stations would be of little practical interest.

The six parameters of $\boldsymbol{H}$ have been estimated in Salvadori and De Michele (2010) via ML (see the first row of Table 2): this will give us the possibility to compare and discuss the results of fitting techniques different from the standard one. The estimates of the same parameters according to, respectively, the 1-MEV and the p-MEV strategies are also reported in Table 2. It is interesting to note that, independently of the fitting procedure, $\boldsymbol{G}_{\xi} \approx \boldsymbol{\Pi}_4$ (the copula of independence – see Eq. 3), whereas $\boldsymbol{G}_{\chi} \approx \boldsymbol{M}_4$ (the copula of full dependence – see Eq. 4). Thus, as already mentioned, the extra-parametrized copula $\boldsymbol{H}$ is a sort of "mixture" between $\boldsymbol{\Pi}_4$ and $\boldsymbol{M}_4$, ruled by the "local" mixing parameters a_i's.

In Fig. 2 we plot the empirical and fitted Pickands' functions A's for all the pairs of stations and the models of interest. It must be stressed that the empirical estimates of the true (but unknown) dependence functions do not generally respect the convexity constraint (see the discussions in Genest and Segers, 2009, and references therein). Apparently, the 1-MEV strategy (the one using the least information) and the ML technique show the worst performances, whereas the p-MEV method overall provides the best fits. However, the lacks of fit are more apparent than real: in fact, due to the small sample size, the confidence bands are expected to be quite large. Most interestingly, the copula $\boldsymbol{H}$ fitted via the "local" strategies is well able to match the asymmetries shown by the empirical functions, and adapts itself to the "local" behaviors of the data: in particular, the "degree of dependence", as measured via the Kendall's τ, ranges from ≈ 0.1 to ≈ 0.6 (see Table 3), whereas the corresponding values fitted via ML only range from ≈ 0.2 to ≈ 0.4 (see Salvadori and De Michele, 2010).

However, when the problem is multivariate, what should always be analyzed is the full dependence structure, and its global ability to fit the actual data. For this purpose, we exploit some robust Goodness-of-Fit tests for multivariate copulas (Genest et al., 2009). These tests use Cramér-von-Mises statistics, and acceptance or rejection of a model is based on the p-values calculated via bootstrap techniques: small ones suggest to discard the corresponding copula, whereas large ones support its suitability. In our case, the p-values are as reported in Table 2. In turn, all the models investigated here should be accepted, since the p-values are much larger than 5%. It is worth mentioning that the p-values should only be used to reject a copula, according to some standard criterion (like, e.g., a value smaller than 1%). It is a common error to consider as "better" those models yielding the highest p-values: mathematically speaking, this is

HESSD
7, 7563–7590, 2010

G. Salvadori and C. De Michele

generally false.

A further issue of interest concerns the investigation of the multivariate return period: this is a fundamental point in applications, since it provides crucial information of practical utility. In particular, it may be possible to use the estimates of the multivariate return period function in order to choose between models fitted via different strategies. In Fig. 3 we show the empirical and the fitted return periods for all the four stations and the models of interest: the plot shows the return periods associated with all critical probability levels $t \in \boldsymbol{I}$. Note that, due to the limited sample size, the estimates of the largest empirical return periods are spoiled (as is well evident in Fig. 3). Visually, both the ML and the p-MEV fits are valuable, whereas the 1-MEV one apparently fails to provide a consistent approximation: this may not be surprising, since this latter strategy uses the least amount of information.

As illustrated and discussed in Salvadori and De Michele (2010), these multivariate return periods are generally much larger than the ones calculated via the formulas usually found in literature (see Eq. 18, and the following discussion). Clearly, the underestimates provided by the standard approach, i. e. a return period much smaller than the correct one, may have sizable consequences. Instead, following the Kendall's measure approach illustrated here, a correct risk analysis can be performed, and a practical criterion for the choice of a suitable multivariate model is provided.

5 Conclusions

In order to properly assess the risk, MEV models are fundamental in all areas of geophysics. This paper is of methodological nature, and introduces new estimation techniques for dealing with extremes. In particular, we outline several strategies in order to estimate the parameters of the selected copula, according to different criteria: we use either a nearest neighbor or a nearest cluster approach, or exploit all the pairwise relationships between the available gauge stations. The techniques suggested are physically based, and may offer interesting alternatives to standard fitting methods

HESSD

7, 7563–7590, 2010

(e. g., ML). An application to flood data is also shown, and a comparison of different estimating strategies is illustrated: this shows how the techniques outlined in the paper can be used in practice.

Acknowledgements. The authors thank C. Sempi (Università del Salento, Lecce, Italy) and F. Durante (Free University of Bozen-Bolzano, Bolzano, Italy) for invaluable helpful discussions and suggestions. The research was partially supported by the Italian M.I.U.R. via the project "Metodi stocastici in finanza matematica". The support of "Centro Mediterraneo Cambiamenti Climatici" (CMCC – Lecce, Italy) is also acknowledged.

References

Black, A. R. and Fadipe, D.: Use of historic water level records for re-assessing flood frequency: case study of the Spey catchment, Water Environ. J., 23, 23–31, 2009. 7576

Butler, A., Heffernan, J. E., Tawn, J. A., and Flather, R. A.: Trend estimation in extremes of synthetic North Sea surges, J. Roy. Stat. Soc.-App., 56, 395–414, 2007. 7564

Capéraà, P., Fougères, A.-L., and Genest, C.: A stochastic ordering based on a decomposition of Kendall's tau, in: Distributions with Given Marginals and Moment Problems, edited by: Beneš, V. and Štěpán, J., Kluwer Academic, Dordrecht, 81–86, 1997. 7571

Durante, F. and Salvadori, G.: On the construction of multivariate extreme value models via copulas, Environmetrics, 21, 143–161, 2010. 7565, 7571

Durante, F., Quesada-Molina, J., and Úbeda-Flores, M.: On a family of multivariate copulas for aggregation processes, Inform. Sciences, 177, 5715–5724, 2007. 7565, 7569

Falk, M. and Reiss, R. D.: On Pickands coordinates in arbitrary dimensions, J. Multivariate Anal., 92, 426–453, 2005. 7570

Genest, C. and Favre, A.: Everything you always wanted to know about copula modeling but were afraid to ask, J. Hydrol. Eng., 12, 347–368, 2007. 7565

Genest, C. and Rivest, L.-P.: Statistical inference procedures for bivariate Archimedean copulas, J. Am. Stat. Assoc., 88, 1034–1043, 1993. 7570

Genest, C. and Rivest, L.-P.: On the multivariate probability integral transformation, Stat. Probabil. Lett., 53, 391–399, 2001. 7570

Genest, C. and Segers, J.: Rank-based inference for bivariate extreme value copulas, Ann. Stat., 37, 2990–3022, 2009. 7570, 7573, 7579

Genest, C., Ghoudi, K., and Rivest, L.-P.: A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, Biometrika, 82, 543–552, 1995. 7565

Genest, C., Quessy, J.-F., and Rémillard, B.: Goodness-of-fit procedures for copula models based on the probability integral transformation, Scand. J. Stat., 33, 337–366, 2006. 7571

Genest, C., Rémillard, B., and Beaudoin, D.: Goodness-of-fit tests for copulas: a review and a power study, Insur. Math. Econ., 44(2), 199–213, 2009. 7571, 7579

Ghoudi, K., Khoudraji, A., and Rivest, L.: Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles, Can. J. Stat., 26, 187–197, 1998. 7570

Gilvear, D.: Patterns of channel adjustment to impoundment of the upper River Spey, Scotland (1942–2000), River Res. Appl., 20, 151–165, 2004. 7576

Grenney, W. and Heyse, E.: Suspended sediment – river flow analysis, J. Environ. Eng., 111, 790–803, 1985. 7564

Grimaldi, S. and Serinaldi, F.: Asymmetric copula in multivariate flood frequency analysis, Adv. Water Resour., 29, 1155–1167, 2006. 7578

Herr, H. and Krzysztofowicz, R.: Generic probability distribution of rainfall in space: the bivariate model, J. Hydrol., 306, 234–263, 2005. 7564

Joe, H.: Multivariate concordance, J. Multivariate Anal., 35, 12–30, 1990. 7567

Joe, H.: Multivariate Models and Dependence Concepts, Chapman and Hall, London, 1997. 7564, 7565, 7566, 7568

Journel, A. and Alabert, F.: Non-gaussian data expansion in the earth sciences, Terra Nova, 1, 123–134, 1988. 7564

Keef, C., Svensson, C., and Tawn, J.: Spatial dependence in extreme river flows and precipitation for Great Britain, J. Hydrol., 378, 240–252, 2009. 7564

Kim, T.-W., Valdes, J., and Yoo, C.: Nonparametric approach for estimating return periods of droughts in arid regions, ASCE – J. Hydrol. Eng., 8, 237–246, 2003. 7564

Liebscher, E.: Construction of asymmetric multivariate copulas, J. Multivariate Anal., 99, 2234–2250, 2008. 7565

Nelsen, R.: Nonparametric measures of multivariate association, in: Distributions with fixed marginals and related topics (Seattle, WA, 1993), IMS Lecture Notes Monogr. Ser., Inst. Math. Statist., Hayward, CA, 223–232, 1996. 7565, 7567, 7568

Nelsen, R.: Concordance and copulas: a survey, in: Distributions with Given Marginals and Statistical Modelling, Kluwer Acad. Publ., Dordrecht, 169–177, 2002. 7567

Nelsen, R.: An Introduction to Copulas, 2nd edition, Springer-Verlag, New York, 2006. 7564, 7565, 7566, 7567, 7568, 7569, 7571, 7577

Pickands, J.: Multivariate extreme value distributions, B. Int. Statist. Inst., 49, 859–878, 1981. 7569

Pons, F.: Regional flood frequency analysis based on multivariate lognormal models, Ph.D. thesis, Colorado State University, Fort Collins, 1992. 7564

Russo, D.: On probability distribution of hydraulic conductivity in variably saturated bimodal heterogeneous formations, J. Vadose Zone, 8, 611–622, 2009. 7564

Salvadori, G.: Bivariate return periods via 2-copulas, Stat. Method., 1, 129–144, 2004. 7571

Salvadori, G. and De Michele, C.: Frequency analysis via Copulas: theoretical aspects and applications to hydrological events, Water Resour. Res., 40, W12511, doi:10.1029/2004WR003133, 2004. 7571

Salvadori, G. and De Michele, C.: Multivariate multiparameter extreme value models and return periods: a copula approach, Water Resour. Res., in press, doi:10.1029/2009WR009040, 2010. 7565, 7571, 7576, 7577, 7578, 7579, 7580

Salvadori, G., De Michele, C., Kottegoda, N., and Rosso, R.: Extremes in nature, An approach using copulas, Water Science and Technology Library, vol. 56, Springer, 2007. 7564, 7565, 7566, 7569, 7571, 7577

Schmid, F. and Schmidt, R.: Multivariate extensions of Spearman's rho and related statistics, Stat. Prob. Lett., 77, 407–416, 2007a. 7565, 7567, 7568

Schmid, F. and Schmidt, R.: Multivariate conditional versions of Spearman's rho and related measures of tail dependence, J. Multivariate Anal., 98, 1123–1140, 2007b. 7565, 7568

Schmid, F. and Schmidt, R.: Nonparametric inference on multivariate versions of Blomqvist's beta and related measures of tail dependence, Metrika, 66, 323–354, 2007c. 7565, 7569

Scottish Environment Protection Agency: http://www.nwl.ac.uk/ih/nrfa/station_summaries/crg.html, last access: October, 2010. 7576

Shih, J. and Louis, T.: Inferences on the association parameter in copula models for bivariate survival data, Biometrics, 51, 1384–1399, 1995. 7565

Singh, V.: Hydrologic Frequency Modeling, Reidel Publishing Company, 1986. 7564

Singh, V., Jain, S., and Tyagi, A.: Risk and Reliability Analysis, ASCE Press, Reston, Virginia, 2007. 7571

Sklar, A.: Fonctions de répartition à *n* dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris, 8, 229–231, 1959. 7567

HESSD

7, 7563–7590, 2010

Estimating strategies for M-MEV copulas

G. Salvadori and C. De Michele

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◄| | ►|
◄ | ►
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

Taylor, M.: Multivariate measures of concordance, Ann. Inst. Stat. Math., 59, 789–806, 2007. 7567
Úbeda-Flores, M.: Multivariate versions of Blomqvist's beta and Spearman's footrule, Ann. Inst. Stat. Math., 57, 781–788, 2005. 7567
Wilks, D.: Multisite generalization of a daily stochastic precipitation generation model, J. Hydrol., 210, 178–191, 1998. 7564
Yue, S.: The Gumbel mixed model applied to storm frequency analysis, Water Resour. Manage., 14, 377–389, 2000a. 7577
Yue, S.: The Gumbel logistic model for representing a multivariate storm event, Adv. Water Resour., 24, 179–185, 2000b. 7577
Zhang, L.: Multivariate hydrological frequency analysis and risk mapping, Ph.D. thesis, Louisiana State University, Baton Rouge, Louisiana, USA, 2005. 7571
Zhang, L. and Singh, V.: Gumbel-Hougaard copula for trivariate rainfall frequency analysis, ASCE – J. Hydrol. Eng., 12, 409–419, 2007. 7577

HESSD
7, 7563–7590, 2010
Estimating strategies for M-MEV copulas
G. Salvadori and C. De Michele

Table 1. (*Upper triangular*) Inter-station distances (in km). (*Diagonal*) Labels of the nearest neighbor station. (*Lower triangular*) Empirical estimates of the Kendall's τ for all the pairs of the four stations, with the p-values in parentheses – see text.

Station	S_2	S_6	S_9	S_{10}
S_2	S_9	61.7	19.1	24.0
S_6	0.06 (0.62)	S_{10}	43.6	37.9
S_9	0.25 (0.03)	0.34 (4e−3)	S_{10}	6.0
S_{10}	0.43 (2e−4)	0.29 (0.01)	0.54 (3e−6)	S_9

HESSD
7, 7563–7590, 2010

Table 2. Estimates of the parameters of the 4-variate copula $\boldsymbol{H}$ using different fitting techniques – see text. Also shown are the p-values of the corresponding models.

Method	$\hat{\xi}$	$\hat{\chi}$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	p-v.
ML	1.55	11.04	0.97	0.36	0.78	0.89	0.40
1-MEV	2.73	11.03	0.99	0.12	0.48	0.79	0.77
p-MEV	1.99	11.03	1.00	0.15	0.71	0.82	0.77

HESSD
7, 7563–7590, 2010

Table 3. Values of the Kendall's τ for all the pairs of the four stations – see text: (*upper triangular*) estimates using the 1-MEV strategy; (*lower triangular*) estimates using the p-MEV strategy.

Station	S_2	S_6	S_9	S_{10}
S_2	1	0.12	0.37	0.55
S_6	0.12	1	0.60	0.32
S_9	0.39	0.39	1	0.57
S_{10}	0.43	0.29	0.54	1

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD

7, 7563–7590, 2010

Estimating strategies for M-MEV copulas

G. Salvadori and C. De Michele

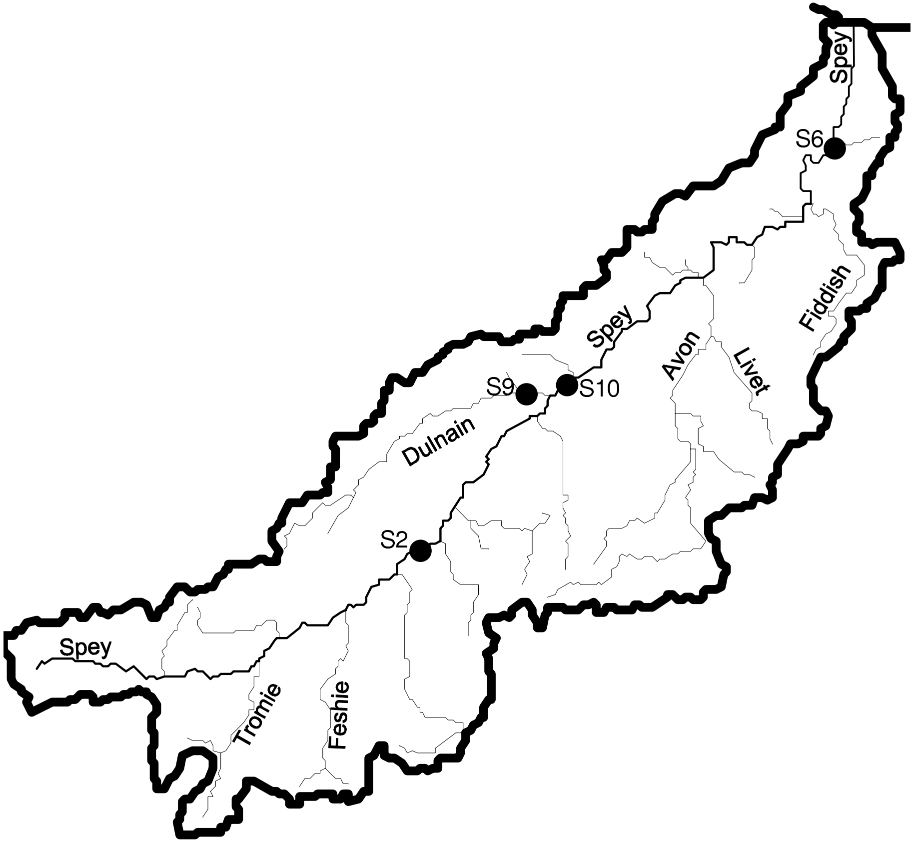

Fig. 1. Map of the Spey catchment. The black circles indicate the four gauge stations of interest – see text.

HESSD

7, 7563–7590, 2010

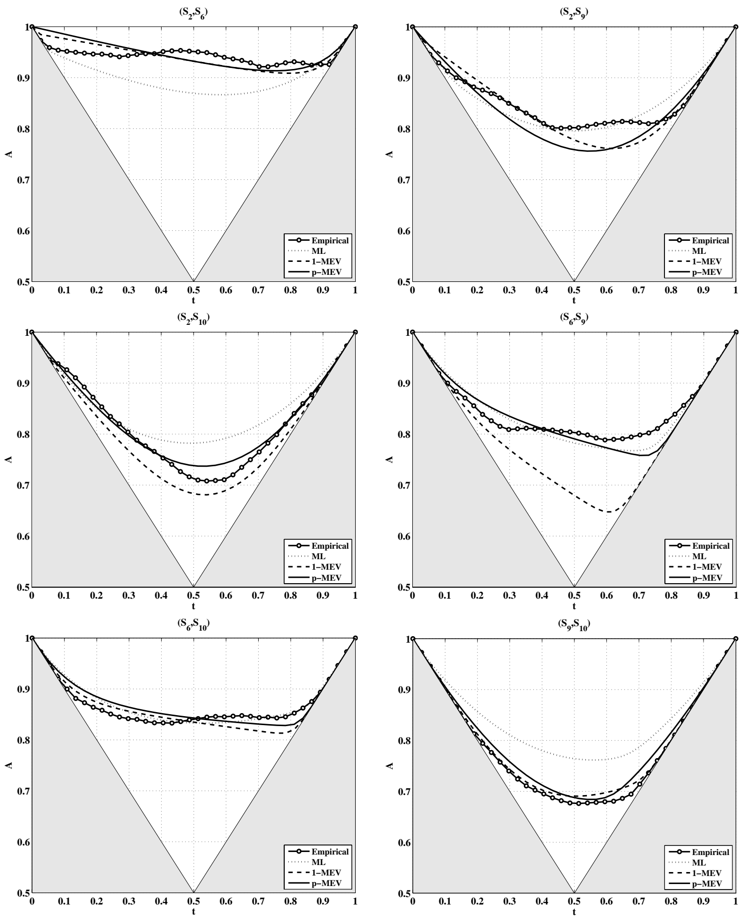

Fig. 2. Plots of empirical and fitted Pickands' dependence functions for all the pairs of stations and the models of interest – see text.

HESSD

7, 7563–7590, 2010

G. Salvadori and C. De Michele

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

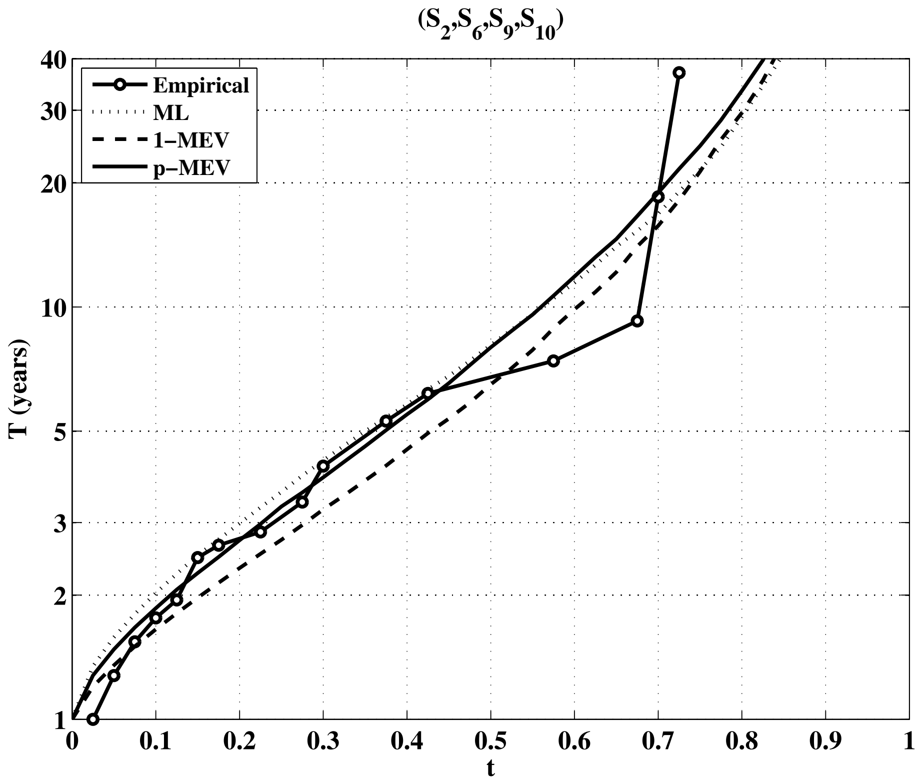

Fig. 3. Plot of empirical and fitted return periods for all the four stations and the models of interest – see text.

HESSD
7, 7563–7590, 2010