Clustering of concurrent flood risks via Hazard Scenarios

R. Pappadà [a,*], F. Durante [b], G. Salvadori [c], C. De Michele [d]

[a] *Department of Economics, Business, Mathematics and Statistics "Bruno De Finetti", University of Trieste, Trieste, Italy*
[b] *Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy*
[c] *Dipartimento di Matematica e Fisica "Ennio De Giorgi", Università del Salento, Lecce, Italy*
[d] *Dipartimento di Ingegneria Civile e Ambientale (DICA), Politecnico di Milano, Milano, Italy*

Keywords:
Clustering
Copula
Risk assessment
Statistical hydrology
Flood risk

A B S T R A C T

The study of multiple effects of a number of variables, and the assessment of the corresponding environmental risks, may require the adoption of suitable multivariate models when the variables at play are dependent, as it often happens in environmental studies. In this work, the flood risks in a given region are investigated, in order to identify specific spatial sub-regions (clusters) where the floods show a similar behavior with respect to suitable (multivariate) criteria. The reason of the work is three-fold, and the outcomes have deep implications in the hydrological practice: (i) such a regionalization (as it is called in hydrology) may provide useful indications for deciding which gauge stations have a similar (stochastic) behavior; (ii) the spatial clustering may represent a valuable tool for investigating ungauged basins present in a given "homogeneous" Region; (iii) the estimate of extreme design values may be improved by using all the observations collected in a cluster (instead of only single-station data). For this purpose, a Copula-based Agglomerative Hierarchical Clustering algorithm – a key tool in geosciences for the analysis of the dependence information – is proposed. The procedure is illustrated via a case study involving the Po river basin, the largest Italian one. A comparison with a previous attempt to cluster the gauge stations present in the same spatial region is also carried out. The sub-regions picked out by the clustering procedure outlined here agree with previous

* Corresponding author.
E-mail addresses: rpappada@units.it (R. Pappadà), fabrizio.durante@unisalento.it (F. Durante), gianfausto.salvadori@unisalento.it (G. Salvadori), carlo.demichele@polimi.it (C. De Michele).

results obtained via heuristic hydrological and meteorological reasonings, and identify spatial areas characterized by similar flood regimes.

1. Introduction

The impact of extreme events such as floods, droughts and tropical cyclones, can be due to a single climate or weather variable being in an extreme state, but more often it is the result of a combination of variables not all of which are necessarily extreme (Leonard et al., 2014). The combination of variables or events that may yield extreme impacts is usually referred to as a *compound event*: in particular, in IPCC (2012, page 118), it is stressed that "impacts on the physical environment are often the result of compound events". As such, the study of correlation and dependence may be helpful to improve the knowledge of the occurrence of extremes and their impact on our societies, although "much of the analysis of changes of extremes has, up to now, focused on individual extremes of a single variable" (ibidem, page 118). Thus, in order to be prepared to face up to possible future climatic challenges, we need to understand the interactions between extreme events (and even not necessarily extreme ones) and other natural hazards (Vahedifard et al., 2016; Zscheischler and Seneviratne, 2017). As discussed in Leonard et al. (2014), any given compound event depends upon the nature and the number of the physical variables at play, as well as on the range of spatial and temporal scales, and the strength of dependence between processes. Thus, by its very nature, the study of compound (extreme) events is a tricky and complex task, that requires specific methodologies to be dealt with.

Recently, copula models are gaining ground in geo- and environmental-sciences, since engineers and practitioners need flexible theoretical frameworks to handle the diversity of the interplaying variables (see, for instance, Salvadori et al., 2007 and references therein, for an extensive discussion on both theoretical and practical aspects of the mathematical theory of extremes and the opportunities offered by copulas in many applied contexts). Shortly, a multivariate copula is the restriction of a joint cumulative distribution function over $[0, 1]^d$ with Uniform margins. Thanks to copulas, the behavior of a compound event can be conveniently decomposed into the marginal effects, given by the individual variables, and the linkage effects, as described by the copula uniquely associated with the involved (continuous) variables. As underlined in Guthke and Bárdossy (2017), besides taking into account non-Gaussianity, the main advantage of copula-based methods to treat geostatistical problems is the "descriptive power and standardized interpretability".

Here, we are mainly interested in hydrological issues, and more specifically in flood risks: these may potentially cause fatalities, displacement of people, damage the environment, severely compromise the economic development, and undermine the economic activities (as underlined in the Directive 2007/60/EC of The European Parliament and The Council, 2007). According to the framework proposed in Salvadori et al. (2016) (inspired by the previously cited Directive), flood risks should require: (a) the development of suitable hazard scenarios that takes into account the multivariate nature of a flood event; (b) the estimation of the probabilistic occurrence of the above scenarios. Both aspects have been considered in Salvadori et al. (2016), where various methodologies have been analyzed and discussed. Furthermore, a related aspect of interest is to determine the interdependencies among different flood risks across regions. In fact, as reported in Jongman et al. (2014), recent major flood disasters have shown that extreme events can affect multiple regions/countries simultaneously, which puts high pressure on trans-regional risk-reduction and risk-transfer mechanisms. The determination of suitable regions that are affected simultaneously by extreme events has a great impact in hydrological practice. In fact, the design of water engineering works like urban sewers may require the evaluation of the flood quantile for a Return Period T equal to 5 or 10 years. The design of levees, or water detention basins, usually requires 100- or 200-year flood quantiles, and for dams 1000-year or larger flood quantiles. These values refer to Italian engineering practice: see Chow et al. (1988, page 419) for general indications about the design Return Periods of water-control structures. As indicated by Benson (1962) and also De Michele and Rosso (2001), reliable quantile estimates

can be obtained for Return Periods $T = 2N$, where N denotes the length of the annual flood series. Thus, when the time series are too short, it is not always possible to use at-site data to evaluate the quantiles of interest for assigned frequency or Return Period. Empirically, Regionalization Techniques, including Clustering procedures, are used in hydrology, to substitute "space for time", in order to pool together a group of watersheds with sufficient "homogeneity" in the flood generating mechanisms, which may then represent a homogeneous region or pooling-group. Several approaches have been developed, based on the analysis of the "similarity" between sites involving catchment attributes like physiographic characteristics, seasonality, and at-site flood statistics (see Ramachandra Rao and Srinivas, 2008). These techniques allow to cope with two main problems: (1) the paucity, or, more frequently, the total absence, of data at the site of interest (the ungauged site problem); (2) the reliability of design estimates involving the calculation of the design quantiles for large Return Periods (say, 100–1000 years)—see De Michele and Rosso (2001).

It is worth stressing that these techniques have almost exclusively focused on the Flood Peak as the variable of interest for design purposes (i.e., a univariate approach), even if other variables – like, e.g., the Flood Volume – may play a significant role in the determination of flood hazards.

Here, we show how a clustering procedure could be implemented in order to model how flood risks may be interconnected in a given region. The proposed methodology is grounded on the concept of Hazard Scenarios, as developed in Salvadori et al. (2016), and some novel insights in copula-based clustering algorithms (for a description of the general framework, see Di Lascio et al., 2017). The final output of the cluster algorithm shows how (concurrent) flood events, characterized by a number of physical variables (like Flood Peak and Flood Volume), may be interrelated within a given geographical area. This might identify possible regions where risk managers and policy managers should adopt a cross-border strategy to mitigate and prevent the hydrological risks.

The paper is organized as follows. Section 2 outlines the general methodology. Section 3 illustrates the case study and the clustering procedures adopted. In Section 4 the results are discussed, and a comparison with previous (univariate) outcomes is presented. Finally, some conclusions are given in Section 5.

2. The methodology

The concept of compound event emphasizes the multivariate aspect of environmental risks, arising from the joint occurrence of multiple hazards or the interaction of several variables ruling a single phenomenon. As such, the identification of "similar" contributing events requires the use of non-standard techniques, in order to reflect the degree of dependence between variables or events. In this section, a novel clustering approach is proposed, based on a suitable multivariate similarity measure.

Usually, cluster analysis plays an important role in extracting information from a set of different observations that can be interpreted, in a model-based setting, as realizations of a given stochastic model/process related to different random objects. Clustering methods may also be used to perform an analysis of the dependence information, which is a key tool in geosciences (and, in particular, in hydrology) in order to understand the relationships between different variables.

In general, most algorithms require the choice of a *similarity (proximity) measure* between the objects. Within the time series framework, for instance, clustering procedures may use the information about the trends and/or similar sub-patterns (Caiado et al., 2015; Maharaj, 2000), or may consider a Pearson-correlation based distance metric. In this latter case, recent studies have underlined that classical correlation measures are often inadequate to capture the actual dependence structure between individual risk factors, especially in a financial and environmental context (see, among others, Embrechts et al., 2002; Poulin et al., 2007; Salvadori and De Michele, 2010; Salvadori and De Michele, 2011). As such, several investigations have been carried out exploiting tools from extreme-value analysis (see, for instance, De Luca and Zuccolotto, 2011; Durante et al., 2014, 2015; Mornet et al., 2016). Within the class of model-based clustering methods, copula-based algorithms (see, e.g., Di Lascio et al., 2017, and references therein) use the copula information to derive the specific criterion that determines the clustering composition. For an alternative copula-based approach, useful for identifying "similar/compatible" hydrological basins (in a broad sense), see Grimaldi et al. (2016).

Adopting a general setting, in the following we consider a set $\mathcal{P} = \{P_1, \ldots, P_d\}$ of occurrences of a random Phenomenon (e.g., floods) observed at $d > 1$ different sites (e.g., gauge stations). Each

occurrence is characterized by a set of p different variables (typically dependent on one another): for instance, a flood can be described by the triple Peak–Volume–Duration. For each ith site, the occurrences are observed at times $1, \ldots, t_i$. Formally, we denote by $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$ the data observed at the ith site, with $i = 1, \ldots, d$, where

$$\mathbf{x}_{ij} = (x_{ij}^1, \ldots, x_{ij}^{t_i}) \tag{1}$$

is the observed time series related to the jth variable ($j = 1, \ldots, p$) at times $1, \ldots, t_i$.

Following a model-based approach, we assume that, for $i = 1, \ldots, d$, the behavior of P_i can be represented by a random vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$, whose joint probability distribution function $\mathbf{F}_i$ can be expressed in the form (via Sklar's Theorem Sklar, 1959)

$$\mathbf{F}_i = \mathbf{C}_i(F_{i1}, \ldots, F_{ip}), \tag{2}$$

where F_{ij} is the univariate probability distribution function of the jth variable ($j = 1, \ldots, p$) observed at the ith site, and $\mathbf{C}_i$ is the copula associated with the phenomenon at the same site.

Remark 2.1. In this setting, the observations $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$ in (1) are considered as independent realizations from the model (2), and the copula $\mathbf{C}_i$ can be directly derived via a rank transformation (Genest and Favre, 2007). Otherwise, one may first filter the original time series (e.g., for removing trends and seasonality effects), and then investigate the dependence among the resulting residuals. This procedure is described in its full generality in Patton (2012), and is applied to cluster analysis in Durante et al. (2014, 2015), among others.

In the sequel, a clustering procedure, based on a suitably defined similarity criterion, is used in order to identify a partition of $\mathcal{P}$ into non-empty and non-overlapping subsets, exploiting the fact that the dependence between the random phenomena $P_1, \ldots, P_d$ can be expressed in terms of the $(d \times p)$-dimensional copula of the vector $(\mathbf{X}_1, \ldots, \mathbf{X}_d)$: viz., such a copula models the dependence between the p variables associated with the d gauge stations.

Such a procedure can be implemented via two main steps.

1. First, define a *dissimilarity measure* between the ith and the kth site as a measure of the deviation of their $2p$-dimensional copula from the co-monotonicity copula

$$\mathbf{M}_{2p}(\mathbf{u}) = \min\{u_1, \ldots, u_{2p}\}, \tag{3}$$

   i.e., the Fréchet–Hoeffding upper-bound (Nelsen, 2006), which models the comonotone dependence between the variables at play. Here, the copula $\mathbf{C}_{ik}$, associated with the pair of sites (P_i, P_k), can be estimated from the available data either parametrically (by properly choosing a suitable family) or non-parametrically (via the empirical copula, or related smoothed versions).
2. Secondly, classical clustering techniques can be applied to the dissimilarity matrix defined above. In this work, we adopt an Agglomerative Hierarchical approach, where each observation initially forms a cluster, and then pairs of clusters are merged at each step of the algorithm (see Hastie et al., 2009; Ward Jr, 1963).

Concerning the first step, two copula-based dissimilarity measures are outlined in the sequel, grounded on the notion of *Hazard Scenario* (shortly, HS) recently formalized in Salvadori et al. (2016), and briefly recalled below.

Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a m-dimensional random vector describing the phenomenon of interest, with joint distribution function $\mathbf{F}$ (respectively, joint survival function $\mathbf{F}^*$), and let $\mathbf{x} \in \mathbb{R}^m$. We denote by F_i (respectively, F_i^*) the distribution (respectively, survival) function associated with X_i ($i = 1, \ldots,$ m), and by $\mathbf{C}$ (respectively, $\mathbf{C}^*$) the copula (respectively, survival copula) associated with $\mathbf{X}$. In turn, by virtue of Sklar's Theorem, $\mathbf{F} = \mathbf{C}(F_1, \ldots, F_m)$ and $\mathbf{F}^* = \mathbf{C}^*(F_1^*, \ldots, F_m^*)$.

The Kendall distribution function $\mathbf{K}_\mathbf{C}$ of $\mathbf{X}$ (see, e.g., Barbe et al., 1996; Genest and Rivest, 1993) is the (univariate) distribution function of the random variable $\mathbf{C}(\mathbf{U})$, where $\mathbf{U} = (U_1, \ldots, U_m)$ with $U_i = F_i(X_i)$, viz.

$$\mathbf{K}_\mathbf{C}(t) = \mathbb{P}(\mathbf{C}(U_1, \ldots, U_m) \leq t), \tag{4}$$

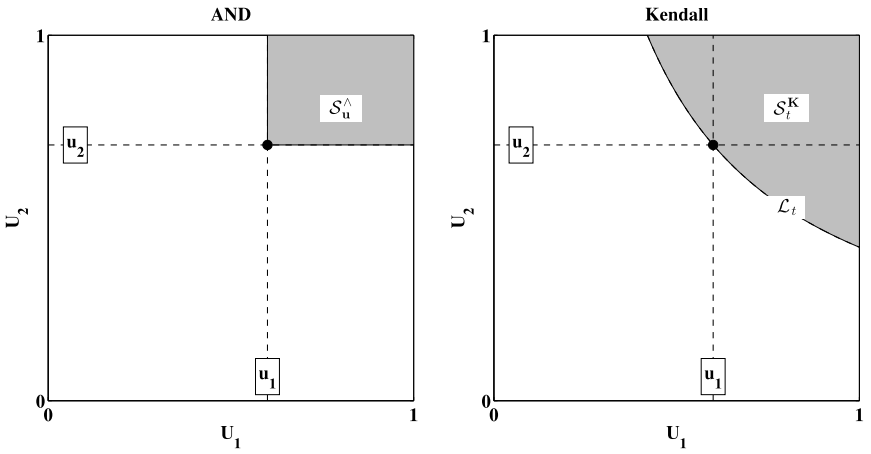

Fig. 1. The shape (shaded region) of a bivariate AND HS $\mathcal{S}^{\wedge}_{u_1=F_1(x_1),u_2=F_2(x_2)}$ (*left*), and a bivariate Kendall HS $\mathcal{S}^{\mathbf{K}}_{t=\mathbf{C}(u_1=F_1(x_1),u_2=F_2(x_2))}$ (*right*) in the copula domain—see text.

with $t \in [0, 1]$. This function also appears in Genest and Rivest (2001) and Nelsen et al. (2001) as a Multivariate Probability Integral Transform. Eq. (4) yields the probability that the random vector $\mathbf{U}$ belongs to the region of $[0, 1]^m$ identified by the inequality $C(\mathbf{u}) \leq t$.

Below, the two Hazard Scenarios used in the sequel are defined: here, $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{u} = (F_1(x_1), \ldots, F_m(x_m))$, i.e. the image of $\mathbf{x}$ in the copula domain $[0, 1]^m$ via the Probability Integral Transform.

"**AND**" **scenario** $\mathcal{S}^{\wedge}$. A m-dimensional AND HS is given by the region

$$\mathcal{S}^{\wedge}_{\mathbf{x}} = \bigcap_{i=1}^{m} (\mathbb{R} \times \cdots \times (x_i, +\infty) \times \cdots \times \mathbb{R}), \tag{5}$$

and the associated probability is

$$\mathbb{P}\left(\mathbf{X} \in \mathcal{S}^{\wedge}_{\mathbf{x}}\right) = \mathbf{C}^*(F_1^*(x_1), \ldots, F_m^*(x_m)). \tag{6}$$

For the realization of the event $\{\mathbf{X} \in \mathcal{S}^{\wedge}_{\mathbf{x}}\}$ it is necessary that all the variables X_i's, with $i = 1, \ldots,$ m, exceed the corresponding thresholds x_i's. The shape of a bivariate AND HS is illustrated in Fig. 1-left in the copula domain.

"**Kendall**" **scenario** $\mathcal{S}^{\mathbf{K}}$. Let $t = \mathbf{F}(\mathbf{x}) = \mathbf{C}(\mathbf{u})$, and let $\mathcal{L}_t$ be the related level set (also called *critical layer*). A m-dimensional Kendall HS is given by the region

$$\mathcal{S}^{\mathbf{K}}_t = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{F}(\mathbf{y}) > t\} = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{C}(F_1(y_1), \ldots, F_m(y_m)) > t\}, \tag{7}$$

and the associated probability is

$$\mathbb{P}\left(\mathbf{X} \in \mathcal{S}^{\mathbf{K}}_t\right) = 1 - \mathbf{K}_{\mathbf{C}}(t). \tag{8}$$

A bivariate Kendall HS is illustrated in Fig. 1-right in the copula domain. Roughly speaking, $\mathcal{S}^{\mathbf{K}}_t$ is the region "exceeding" $\mathcal{L}_t$: it is in this very sense that $\mathcal{L}_t$ may represent a critical multivariate threshold.

Inspired by the previous definitions, we can now introduce the following *dissimilarity measures* between two phenomena/sites P_i and P_k, associated with the p-dimensional random vectors $\mathbf{X}_i$ and $\mathbf{X}_k$, whose $2p$-dimensional copula is $\mathbf{C}_{ik}$.

"**AND" dissimilarity measure** $\sigma^{\wedge}_{ik}$. It is given by

$$\sigma^{\wedge}_{ik}(\mathbf{C}_{ik}) = \int_a^b \left((1-t) - \mathbf{C}^*_{ik}(1-t,\ldots,1-t)\right)^2 dt \tag{9}$$

for suitable a and b, with $0 \le a < b \le 1$ (see below). If we denote by

$$\mathbf{x}(t) = (F^{-1}_{i1}(t),\ldots,F^{-1}_{ip}(t),F^{-1}_{k1}(t),\ldots,F^{-1}_{kp}(t)) \tag{10}$$

the $2p$-dimensional vector of univariate quantiles of order t associated with all the components of $\mathbf{X}_i$ and $\mathbf{X}_k$, then it follows that $\sigma^{\wedge}_{ik}(\mathbf{C}_{ik})$ is a L^2-type distance between the probability of occurrence of a HS of type $\mathcal{S}^{\wedge}_{\mathbf{x}(t)}$ under the co-monotonicity copula $\mathbf{M}_{2p}$ (equal to $1-t$) and under the copula $\mathbf{C}_{ik}$ (equal to $\mathbf{C}^*_{ik}(1-t,\ldots,1-t)$).
In the case $p=1$, $\mathbf{C}_{ik}$ is a bivariate copula. Let $(\mathbf{X}^+_i,\mathbf{X}^+_k)$ be a pair of one-dimensional random vectors whose copula is $\mathbf{M}_2$: the dissimilarity measure can then be expressed as

$$\begin{aligned}\sigma^{\wedge}_{ik}(\mathbf{C}_{ik}) &= \int_a^b \left((1-t) - \mathbf{C}^*_{ik}(1-t,1-t)\right)^2 dt \\ &= \int_a^b \left(\mathbb{P}\left((\mathbf{X}^+_i,\mathbf{X}^+_k)\in\mathcal{S}^{\wedge}_{(F_i^{-1}(t),F_k^{-1}(t))}\right) - \mathbb{P}\left((\mathbf{X}_i,\mathbf{X}_k)\in\mathcal{S}^{\wedge}_{(F_i^{-1}(t),F_k^{-1}(t))}\right)\right)^2 dt \\ &= \int_a^b (t-\mathbf{C}_{ik}(t,t))^2 dt,\end{aligned}$$

since (see also Fig. 1-left)

$$\mathbb{P}\left((\mathbf{X}_i,\mathbf{X}_k)\in\mathcal{S}^{\wedge}_{(F_i^{-1}(t),F_k^{-1}(t))}\right) = 1-2t+\mathbf{C}_{ik}(t,t).$$

In general, it follows that $\sigma^{\wedge}_{ik}(\mathbf{C}) = 0$ if $\mathbf{C}$ is the co-monotonicity copula $\mathbf{M}$, and for all pairs (i,k), $i,k=1,\ldots,d$, $i\neq k$, $0\le\sigma^{\wedge}_{ik}(\mathbf{C})\le\sigma^{\wedge}_{ik}(\mathbf{W}_{2p})$ for all copulas $\mathbf{C}$, where the upper bound corresponds to the limiting case when the integrable function

$$\mathbf{W}_{2p}(\mathbf{u}) = \max\left\{\sum_{i=1}^{2p} u_i - 2p + 1, 0\right\}, \tag{11}$$

i.e. the Fréchet–Hoeffding lower bound (Nelsen, 2006) (which is not a copula for $p>1$), is used in Eq. (9) in place of $\mathbf{C}^*$.
"**Kendall" dissimilarity measure** $\sigma^{\mathbf{K}}_{ik}$. It is given by

$$\sigma^{\mathbf{K}}_{ik}(\mathbf{C}_{ik}) = \int_a^b \left((1-t) - (1-\mathbf{K}_{\mathbf{C}_{ik}}(t))\right)^2 dt \tag{12}$$

for suitable a and b, with $0 \le a < b \le 1$ (see below), where $\mathbf{K}_{\mathbf{C}_{ik}}$ denotes the Kendall distribution function associated with the copula $\mathbf{C}_{ik}$. If $\mathbf{x}(t)$ denotes the vector in (10), which identifies a Kendall HS via the critical layer $\mathcal{L}_{t=\mathbf{F}(\mathbf{x}_t)}$, then it follows that $\sigma^{\mathbf{K}}_{ik}(\mathbf{C}_{ik})$ is a L^2-type distance between the probability of occurrence of a HS of type $\mathcal{S}^{\mathbf{K}}_t$ under the co-monotonicity copula $\mathbf{M}_{2p}$ (equal to $1-t$) and under the copula $\mathbf{C}_{ik}$ (equal to $1-\mathbf{K}_{\mathbf{C}_{ik}}(t)$).
In view of the properties of the Kendall function $\mathbf{K}$ (Nelsen et al., 2003), it holds that $\mathbf{K}_{\mathbf{C}}(t)\ge t$ for all $t\in[0,1]$ and all copulas $\mathbf{C}$, with $\mathbf{K}_{\mathbf{C}}(t)=t$ for all $t\in[0,1]$ only if $\mathbf{C}=\mathbf{M}$. Thus, $\sigma^{\mathbf{K}}_{ik}=0$ if $\mathbf{C}_{ik}$ is the co-monotonicity copula $\mathbf{M}_{2p}$, and $0\le\sigma^{\mathbf{K}}_{ik}(\mathbf{C}_{ik})\le L$, a fixed constant, that is attained when in Eq. (12) we set $\mathbf{K}_{\mathbf{C}}(t)=1$ for all $t\in[0,1]$, which corresponds to the Kendall function associated with the Fréchet–Hoeffding lower bound $\mathbf{W}_2$ in the case $p=1$.

In the following, we are mainly concerned with extreme risks such that the largest values of a time series correspond to the most risky situations. Therefore, in (9) and (12), we set $a=0.5$ and $b=1$, i.e. we focus on the upper-right tail/orthant of the joint distributions of interest.

Table 1
Geographical information concerning the gauge stations recorded in the Po river basin. The coordinates are referred to the UTM system zone 32N. Also shown is the temporal span of each time series, and the number of available flood episodes—see text.

#	Station	River	Easting (m)	Northing (m)	Elevation (m a.s.l.)	Period (years)	Size
1	Capriolo	Oglio	571648.60	5054257.50	136.5	1937–2002	262
2	Ponte Briolo	Brembo	545836.90	5061366.50	230.0	1940–2002	502
3	Fuentes	Adda	531934.80	5111550.90	198.0	1923–2000	331
4	Lavello	Adda	533891.00	5070252.50	194.9	1946–2002	221
5	Miorina	Ticino	473271.11	5061682.41	189.9	1923–1993	296
6	Tavagnasco	Dora Baltea	407931.10	5043839.00	265.0	1928–2008	337
7	Lanzo	SturaLanzo	381321.50	5013761.50	446.9	1930–2008	390
8	San Martino	Chisone	364479.80	4971478.00	400.0	1937–2008	163
9	Moncalieri	Po	395976.00	4984080.00	212.5	1928–1992	299
10	Farigliano	Tanaro	412684.10	4928921.50	235.0	1942–2008	328
11	Montecastello	Tanaro	475186.00	4977273.00	79.5	1943–2008	305
12	Serravalle	Scrivia	488819.80	4951838.00	195.9	1931–2008	218
13	Ponte Bacchello	Secchia	657402.00	4957027.00	21.5	1942–2007	362
14	Bomporto	Panaro	662026.00	4954877.00	18.4	1923–2007	397
15	Piacenza	Po	555446.00	4990114.00	42.2	1924–2007	558
16	Cremona	Po	578307.00	4997855.00	34.3	1972–2007	217
17	Boretto	Po	623258.00	4973892.00	20.0	1943–2007	400
18	Roncocorrente	Po	635742.66	4990362.55	15.3	1924–1988	431
19	Borgoforte	Po	638617.00	4989721.00	14.6	1924–2007	533
20	Pontelagoscuro	Po	705774.00	4973904.00	8.5	1923–2007	509

3. The case study

In this section, we analyze flood data recorded at $d = 20$ gauge stations spread across the Po river basin (see Table 1), located in Northern Italy: Fig. 2 shows the geographical area and the sites of interest. The Po river is the largest Italian basin, covering an area of 74,000 km^2 (70,000 km^2 in Italy, 4,000 km^2 in Switzerland and France), characterized by a main river length of 650 km. The Po river basin is a strategic area for the Italian economy, producing 40% of the national Gross Domestic Product, with a population of more than 16 Millions people. Water uses concern industrial activities (principally, the electricity sector, with 48% of the national hydroelectric production, and 31% of the thermo-electric production), agriculture, livestock, and inland navigation. As a consequence, floods may provoke serious harms and detriments.

The time series investigated in the following consist of discharge measurements observed, at a daily scale, at the d gauge stations of interest. Overall, the observations cover the period 1923–2008, although some series may start after 1923, and/or end before 2008: viz., only a partial temporal overlapping is present (see Table 1).

As is typical in hydrology, here we consider a flood as an episode during which the discharge is larger than a given threshold, as recommended by the guidelines of the Watershed Council (AA. VV., 2011). Specifically, each episode is characterized by two variables: the Flood Peak Q (in m^3/s), and the Flood Volume V (in 10^6 m^3). According to the Run Method (Yevjevich, 1967; Zelenhasić and Salvai, 1987), a traditional procedure used in hydrology to extract flood episodes from a time series, a flood can be identified as follows: (i) it starts when the discharge exceeds a given threshold, and (ii) it ends when the discharge falls short of the threshold for at least Δ time units. As a discharge threshold, in the following the so-called "Q300" is used: this denotes an Alarm-threshold according to the guidelines of the Watershed Council, and corresponds to the discharge value that is not exceeded for 300 days in a year (practically, the empirical quantile of the annual discharge series of order $300/365 \approx 82\%$), assuming suitable stationarity conditions of the considered flood processes. The reason for choosing such a threshold is twofold, and represents a valuable compromise between hydrological and statistical needs: on the one hand, it gives the possibility to extract extreme floods (the most interesting ones); on the other hand, the size of the flood samples is sufficiently large for carrying out a proper statistical analysis (see Table 1).

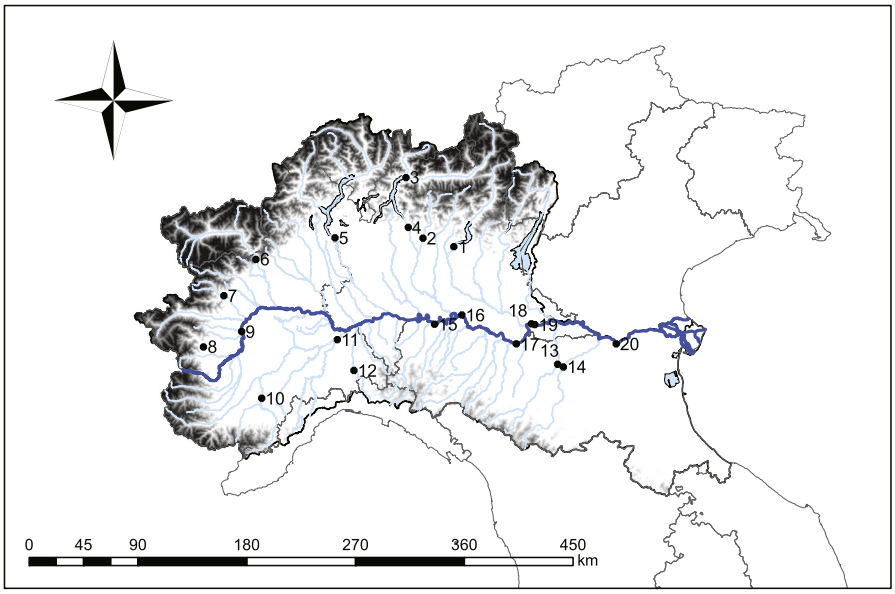

Fig. 2. Map of the Po river basin, and locations of the gauge stations of interest—see text and Table 1.

Finally, in order to guarantee the physical independence of successive floods episodes (as generated by different meteorological events), here Δ is set equal to 3 days, according to the information provided by the Meteo Office. In order to investigate the possible presence of serial dependence in the time series, the autocorrelation function is used, as well as the Box–Pierce and the Ljung–Box tests, in order to check the null hypothesis of independence in the given time series. Here, 20 stations are considered, each associated with two data bases (flood Peak and flood Volume), for a total amount of 40 time series. The results indicate that the null hypothesis of (serial) independence cannot be rejected in 36 out of 40 series separately for each variable (the p-values are much larger than 10%), while in one case (involving the station of Tavagnasco) both the p-values are smaller than 5%, but larger than 1%. In the remaining case (involving the station of Bomporto), the p-value associated with the Volume data is larger than 1%, while the one associated with the Peak data is smaller than 1%. Note that, upstream the Bomporto gauge station, a water detention basin was built in 1999. This may represent an anthropogenic disruption that may have affected the discharges collected at Bomporto, and could explain the serial dependence observed for the Peak time series. However, for the sake of completeness, in order not to discard information concerning the important river basin of Panaro, the series related to the station of Bomporto are kept anyway in our study.

Using the same notation as in Section 2, hereinafter we consider the time series of the flood events P_i, with $i = 1, \ldots, 20$: these consist of the pairs $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$ representing the Flood Peak Q and the Flood Volume V observed at the ith site. Each series has length t_i.

In order to perform the cluster analysis, a suitable dissimilarity matrix is computed by using all the pairwise information from the flood events P_i's. In particular, we focus on two cases:

1. both the Peak and the Volume are considered for each flood episode;
2. only the Peak is considered for each flood episode.

In the following, we provide detailed explanations concerning the former case, since the latter can be treated analogously.

As a preliminary step for the clustering investigation, for each pair (P_i, P_k), with $i, k \in \{1, \ldots, 20\}$, $i \neq k$, the observations of the pairs (Q, V)'s are identified and selected in such a way that all of the

occurrence dates of the floods events at the ith and the kth site fall within the same time period: shortly, we only consider *temporally concurrent episodes*. Here, driven by specific meteo-hydrological indications provided by the Meteo Office, we empirically set the temporal interval for potential overlap equal to 5 days: this choice identifies flood episodes generated by the same meteorological event over the Po river basin, acting at (short) different times at (close) spatially separated sites.

The procedure for calculating the dissimilarity matrix between $P_1, \ldots, P_{20}$, as based on the AND HS approach, is outlined below.

Algorithm 3.1 (*Dissimilarity Matrix Based on the AND HS*)**.** For each $i, k \in \{1, \ldots, 20\}$, $i \neq k$, proceed as follows.

1. Extract the observations of concurrent episodes, as given by the time series
$$(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{k1}, \mathbf{x}_{k2})_s,$$
with $s = 1, \ldots, s_{ik}$: here, s_{ik} is the size of the sample of temporally matching pairs.
2. Test the independence between $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$ and $(\mathbf{x}_{k1}, \mathbf{x}_{k2})$ according to the procedure described in Genest and Rémillard (2004).
3. If the independence assumption is rejected (at a given significance level α), then set
$$\widehat{\sigma_{ik}^{\wedge}} = \int_{a=0.5}^{b=1} \left((1-t) - \widehat{\mathbf{C}}^*_{ik}(1-t, 1-t, 1-t, 1-t)\right)^2 dt,$$
where $\widehat{\mathbf{C}}^*_{ik}$ is an estimate of the 4-dimensional empirical survival copula associated with the observations.
4. Otherwise, set
$$\begin{aligned}\widehat{\sigma_{ik}^{\wedge}} &= \int_{a=0.5}^{b=1} \left((1-t) - \Pi_4^*(1-t, 1-t, 1-t, 1-t)\right)^2 dt \\ &= \int_{a=0.5}^{b=1} \left((1-t) - (1-t)^4\right)^2 dt \approx 0.0366753,\end{aligned}$$
where Π_4^* is the 4-dimensional survival version of the Independence Copula, modeling the joint behavior of independent variables.

Remark 3.1. In the present case study, the level of the independence test is fixed to $\alpha = 10\%$. In the case $p = 1$, when only one variable is associated with each flood event, Step 2 is replaced by a test of independence based on the estimated Kendall τ (Genest and Favre, 2007). Moreover, if the number of concurrent events extracted from the original time series is less than 2, then the assumption of independent events is (reasonably) used by default.

Similarly, the computation of the dissimilarity matrix based on the Kendall HS approach can be carried out as outlined below.

Algorithm 3.2 (*Dissimilarity Matrix Based on the Kendall HS*)**.** For each $i, k \in \{1, \ldots, 20\}$, $i \neq k$, proceed as follows.

1. Extract the observations of concurrent episodes as in Algorithm 3.1.
2. Test the independence as in Algorithm 3.1.
3. If the independence assumption is rejected (at a given significance level α), then set
$$\widehat{\sigma_{ik}^{\mathbf{K}}} = \int_{a=0.5}^{b=1} \left((1-t) - (1 - \widehat{\mathbf{K}}_{\mathbf{C}_{ik}}(t))\right)^2 dt,$$
where $\widehat{\mathbf{K}}_{\mathbf{C}_{ik}}$ is the estimate of the Kendall function of $\mathbf{C}_{ik}$ calculated according to the procedure outlined in Barbe et al. (1996).

4. Otherwise, set

$$\widehat{\sigma_{ik}^{\mathbf{K}}} = \int_{a=0.5}^{b=1} \left((1-t) - (1 - \mathbf{K}_{\Pi_4}(t))\right)^2 dt$$
$$= \int_{a=0.5}^{b=1} \left((1-t) - \left(1 - t\sum_{i=0}^{3} \frac{\log^i(1/t)}{i!}\right)\right)^2 dt \approx 0.0412593,$$

where $\mathbf{K}_{\Pi_4}$ is the Kendall function of the 4-dimensional Independence Copula calculated using (McNeil and Nešlehová, 2009, Proposition 4.5).

Finally, as anticipated above, Agglomerative Hierarchical Clustering with *average*-linkage method for computing distances between each cluster (which usually represents a good compromise between *single*- and *complete*-linkage) can be applied directly to the dissimilarity matrix. This procedure may be represented by a two-dimensional diagram known as a *dendrogram*, which illustrates the aggregation made at each stage of the process. In standard agglomerative clustering, partitions are achieved by selecting one of the solutions in the nested sequence of clusterings forming the hierarchy, i.e. by cutting the dendrogram at the specific level of aggregation (viz., the Height on the vertical axis), so that clusters below that height are distant from each other by at least that amount. The dendrograms obtained from hierarchical clustering applied to Flood data are shown in Figs. 3 and 6.

The proposed algorithms have been implemented by means of the software R (R Core Team, 2017); in particular, the functions available in the R "copula"-package (Yan, 2007; Kojadinovic and Yan, 2010) were used for the estimation of the copula (and related quantities) associated with the phenomena of interest, and to carry out multivariate independence tests based on the empirical copula process.

4. Results and discussion

In the following, we first discuss the results related to a bivariate approach involving both the variables Flood Peak and Flood Volume (see Section 4.1). Here, aggregations into 3- and 4-clusters are presented, since these represent "natural" choices. In fact, on the one hand, such partitions are the ones suggested by the structure of the dendrograms shown in Fig. 3, which statistically indicate (even visually) the plausibility of such aggregations. On the other hand, these clusters also make sense from a hydrological point of view, e.g. distinguishing between Alpine and Apennine fluvial regimes. In particular, eventually this way of data pooling makes a distinction between different precipitation forcings (rainfall, snow) and hydrological flow contributions (rainfall-runoff, snow melting, groundwater flow) occurring in the considered river sites, which affect and control the dynamics of floods.

Then, in Section 4.2, a historical (univariate) contrast will be presented by considering the variable Flood Peak only, and comparing the clustering of several Po river basin stations previously outlined in De Michele and Rosso (2002) with the one resulting from the AND and Kendall approaches proposed in this work.

The outcomes of the cluster analyses discussed in this section represent an innovative contribution concerning the regionalization issues mentioned in the introduction. For instance, the new information regarding the joint dynamics of Flood Peak and Flood Volume (gained via the multivariate clustering approach outlined here) may provide new, useful guidance for evaluating several (multivariate) design quantiles of interest in ungauged sites, a problem of utmost relevance in practical applications. However, this latter issue is outside the scope of the present paper, and is left for future works.

4.1. Flood peak and volume

The results are commented, and are shown in Figs. 4–5.

- [**3-cluster, AND & Kendall:** Fig. 4] In this case, the AND and Kendall criteria yield the same cluster solution: viz., a macro-cluster grouping the central Alpine stations 1–5 (viz., Capriolo, Ponte Briolo, Fuentes, Lavello, and Miorina), located in Lombardia region, the western Alpine

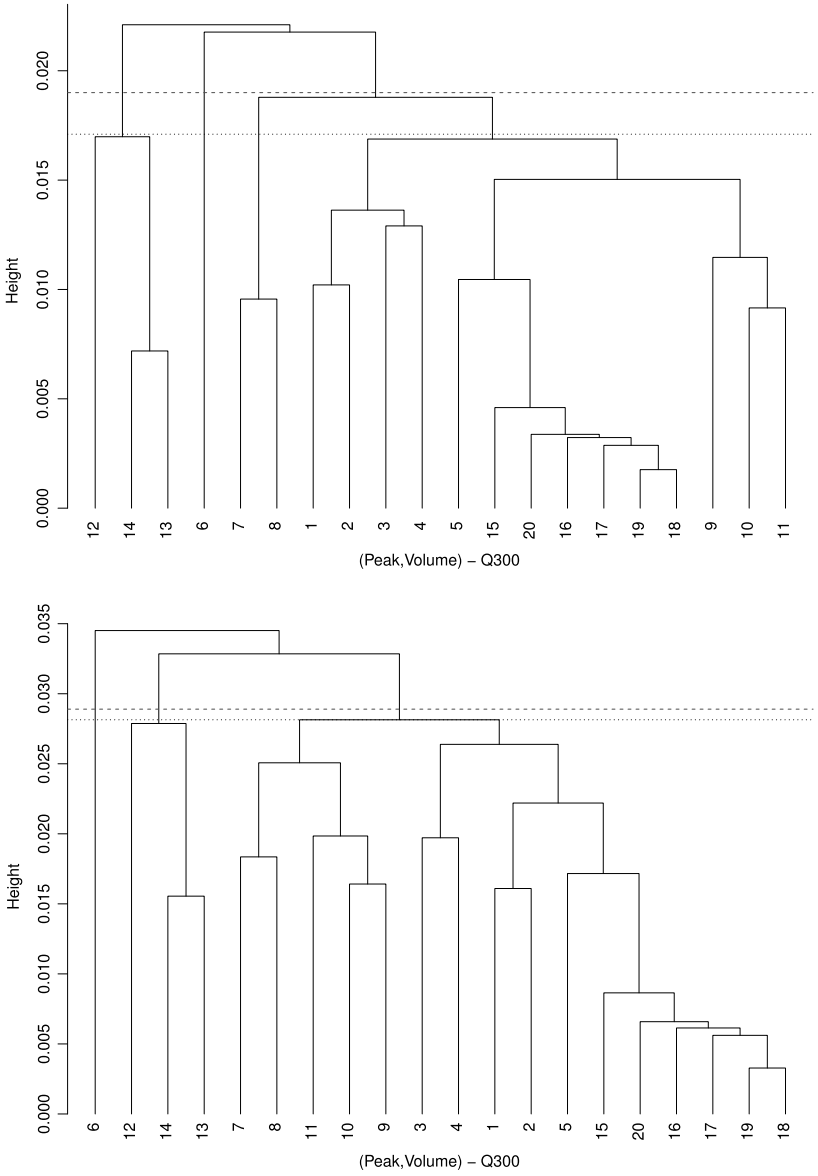

Fig. 3. Flood Peak and Volume. Dendrogram showing average-linkage clustering of flood data based on the AND (*top*) and the Kendall (*bottom*) dissimilarity measure—see text. The labels on the horizontal axis correspond to the stations' reference numbers (see Table 1) and are placed according to the aggregation process. In both cases, the cut of the dendrogram corresponding to the 3- and 4-group solution is indicated by the dashed and dotted lines, respectively.

stations 7–11 (with the exception of Tavagnasco (6) alone), located in Piedmont region, and the main stream stations 15–20 (viz., Piacenza, Cremona, Boretto, Borgoforte, Roncorrente, and Pontelagoscuro), whereas stations 12–14 (viz., Serravalle, Ponte Bacchello and Bomporto) make a cluster, which we denominate "Apennine" cluster, since these sites belong to the Apennine part of the Po river basin.

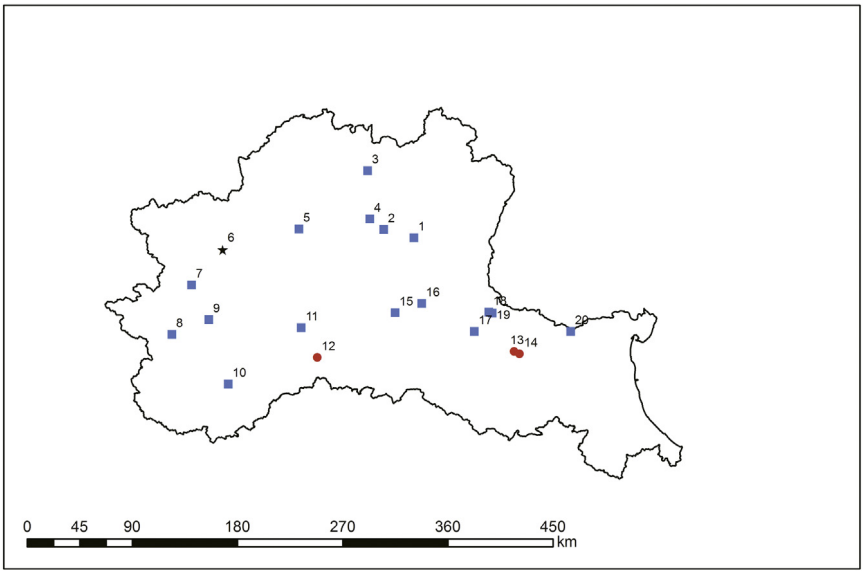

Fig. 4. Flood Peak & Volume. Map of 3-cluster aggregation as extracted from the dendrograms shown in Fig. 3 concerning the AND dissimilarity measure approach and the Kendall one—see text. The clusters are indicated by using different markers and colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- [**4-cluster, AND:** Fig. 5 **(top)**] This case is similar to the 3-cluster solution presented above, but a Piedmont sub-cluster is now identified (stations 7–8): the different behavior of the Central and the Western Alpine stations could be explained by the (occasional) regional dissimilarity of the meteorological events generating the corresponding floods.
- [**4-cluster, Kendall:** Fig. 5 **(bottom)**] In this case, a cluster is identified including the sites located in central Alpine region (1–5) and those located along the main stream (15–20), station 6 is a cluster alone, stations 7–11 make kind of a western Alpine cluster, and lastly the Apennine cluster is identified (stations 12–14).

As a general comment, the following important outcomes are evident.

- An Apennine cluster is always well identified (stations 12–14). From a hydrological point of view (i.e., in terms of the corresponding fluvial regimes), this may represent an interesting upshot of the clustering strategies outlined in this work, identifying an Apennine dynamics different from the Alpine one.
- The stations located in the Central Alps (1–5) are always clustered together with the sites 15–20, on the middle-final part of the Po river main stream: again, this has a clear hydrological meaning, since the Alpine basins mainly influence the behavior of the sites located along the main stream.
- Station 6 (Tavagnasco), in the Piedmont Alpine region, always forms a cluster alone: apparently, this station has a peculiar hydrological regime (see also the discussion in Section 4.2).
- While the 3-cluster solutions under the AND and the Kendall criteria are identical, in the 4-cluster case the Kendall approach identifies a Piedmont Alpine cluster "larger" than in the AND case: viz., all the stations 7–11 instead of only the stations 7–8. As above, the identification of such a cluster, showing a behavior different from the one of the Central Alpine stations, could be supported by meteorological considerations concerning the different etiology of floods. From a hydrological point of view, this may represent an interesting result, since the AND

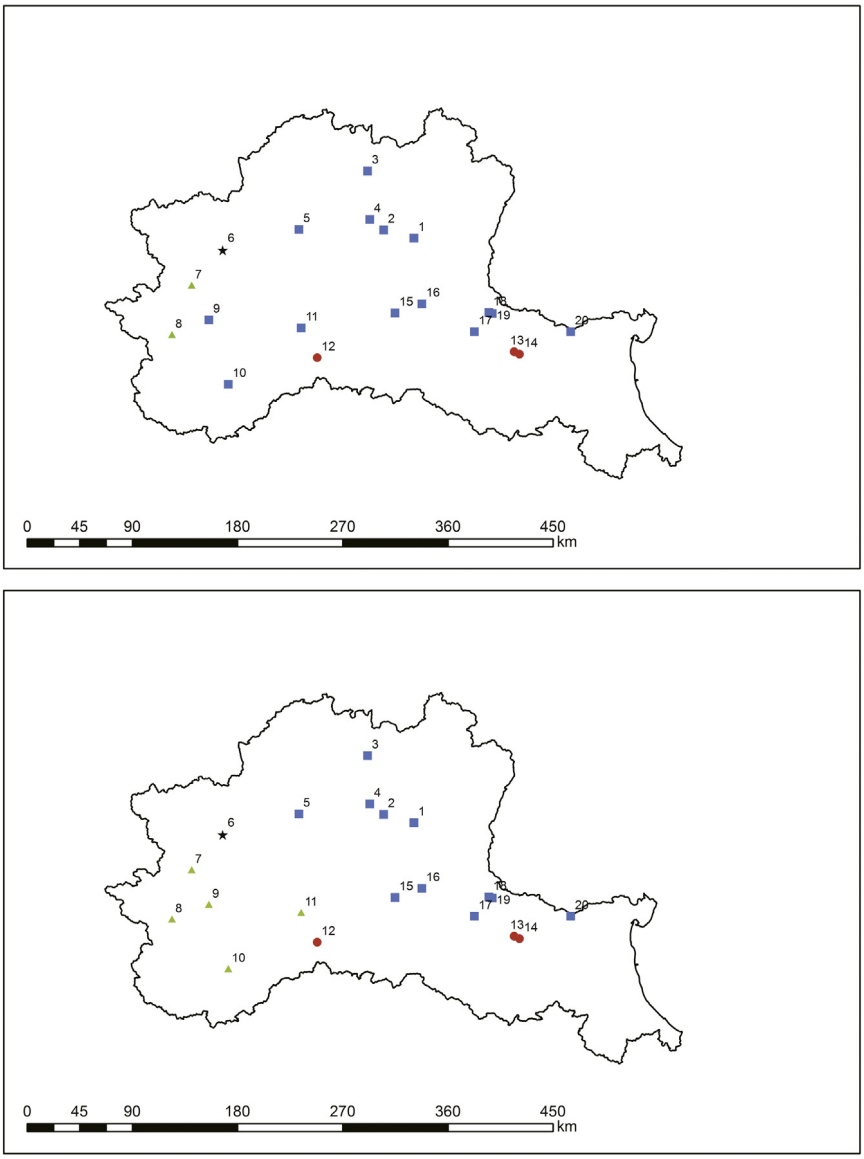

Fig. 5. Flood Peak & Volume. Maps of 4-cluster aggregation as extracted from the dendrograms shown in Fig. 3 concerning the AND dissimilarity measure approach (*top*) and the Kendall one (*bottom*)—see text. The clusters are indicated by using different markers and colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and the Kendall approaches are based on different Hazard Scenarios (see Fig. 1), i.e. different mechanisms/dynamics of aggregation, a feature that may not be (or become) evident using a too coarse and rough 3-cluster partition.

As an external criterion for the evaluation of the clustering results we may consider a measure of internal cohesion and between-cluster separation, using only the geospatial features of the data.

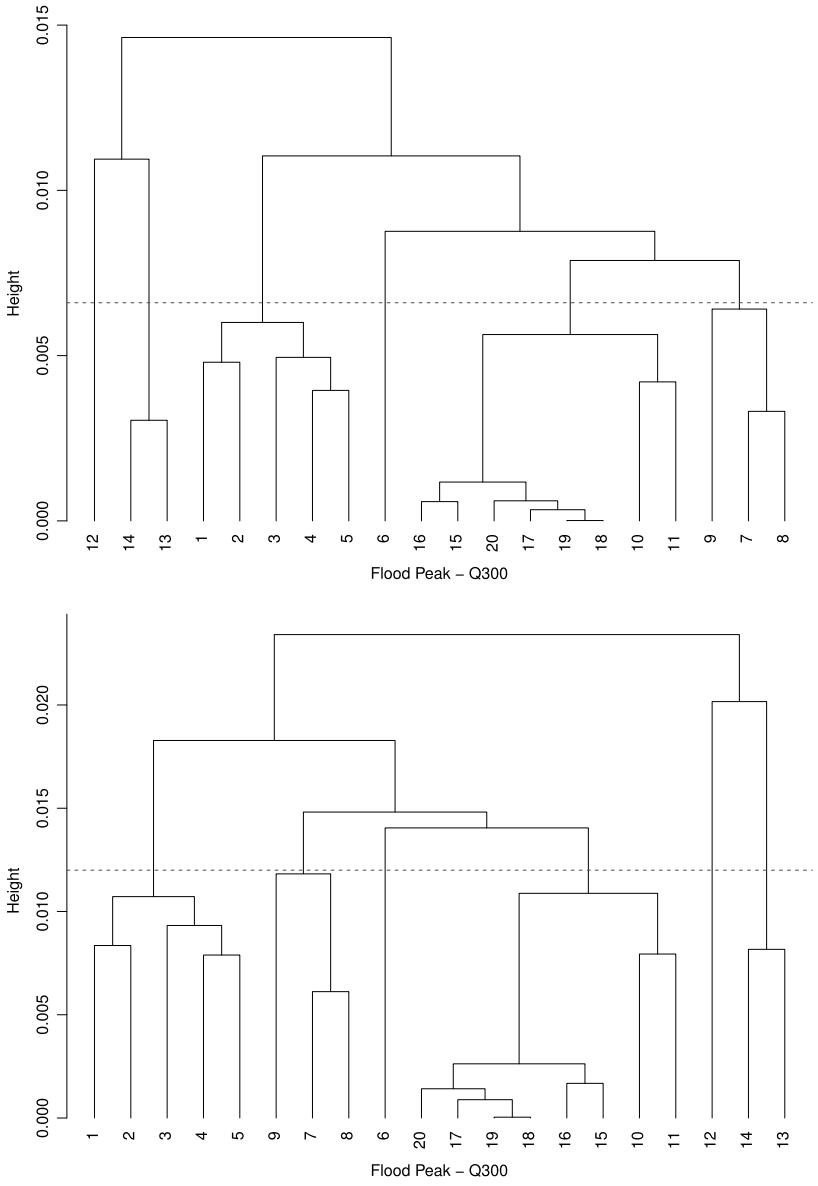

Fig. 6. Flood Peak. Dendrogram showing average-linkage clustering of flood data based on the AND (*top*) and the Kendall (*bottom*) dissimilarity measure—see text. The labels on the horizontal axis correspond to the stations' reference numbers (see Table 1) and are placed according to the aggregation process. The cut of the dendrogram corresponding to the 6-cluster solution is indicated by the dashed line.

Specifically, for each clustering solution derived from the hierarchical structures in Figs. 3 and 6, we look at the ratio, W_b, of the within-cluster to the between-cluster dispersion, defined in terms of sums of squares, computed using the geographical information (Easting, Northing and Elevation) reported in Table 1. Such index decreases monotonically as the number of clusters k increases, but a drastic

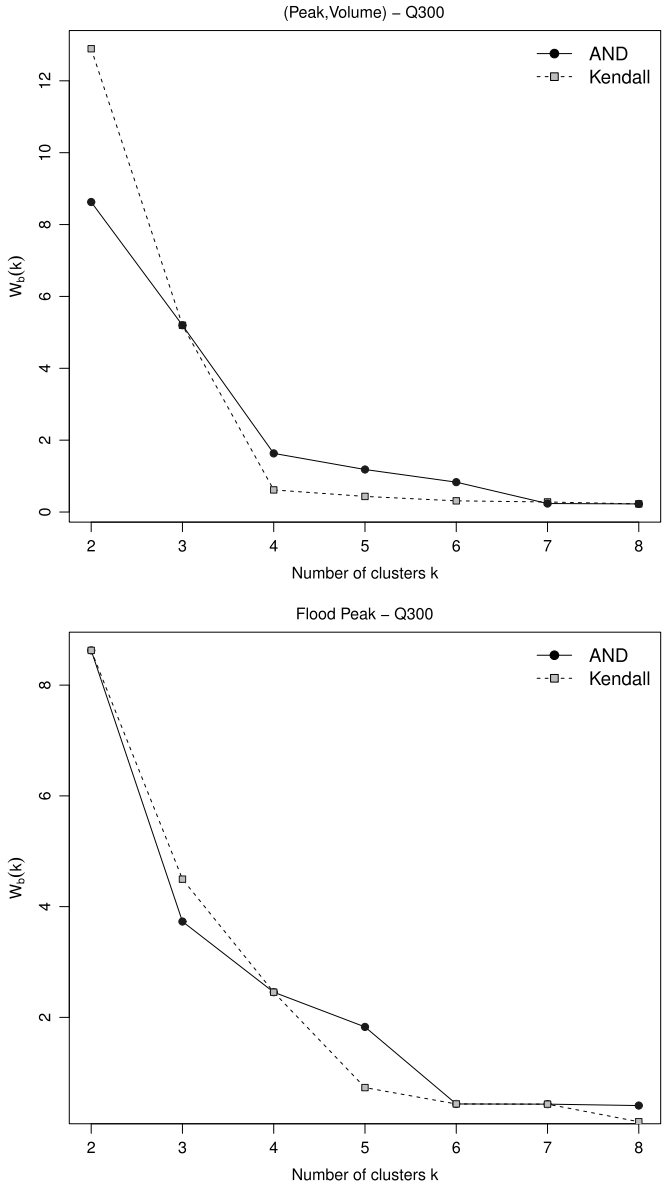

Fig. 7. The index $W_b(k) = W(k)/B(k)$ vs. the number of clusters k, where $W(k)$ is the total within-cluster sum of squares around the cluster means and $B(k)$ is the between-cluster sum of squares, computed by using geographical coordinates of the corresponding cluster members. The graphs of $W_b(k)$ are obtained from hierarchical clustering of Flood Peak and Volume data (see Fig. 3) with AND (solid line) and Kendall (dashed line) dissimilarity measure (*top*), and hierarchical clustering of Flood Peak data (see Fig. 6) with AND (solid line) and Kendall (dashed line) approach (*bottom*).

change in the graph structure (e.g., an elbow shape) may indicate an appropriate number of groups. As can be seen from Fig. 7 (top), the graph of W_b related to the dendrograms in Fig. 3 identifies the four-cluster solution as a valuable choice, since $k = 4$ corresponds to the value beyond which the decrease of W_b flattens markedly for both dissimilarity approaches.

4.2. Comparison with previous literature

In order to compare our results with those present in literature, here we compare the outcomes obtained considering the variable Flood Peak with those outlined in De Michele and Rosso (2002), where a clustering of discharge stations located in the Po river basin was attempted considering (i) the maximum annual Flood Peak as the variable of interest, and (ii) a combination of seasonality indices of flood and precipitation occurrences, and the statistical scale invariance of Flood Peak with respect to the basin area, as aggregation criteria. However, note that the database considered by De Michele and Rosso (2002) is different from the present one, in terms of (a) the variable of interest (in De Michele and Rosso (2002), it is the annual maximum Flood Peak), and (b) the number of sites considered (57 sites in De Michele and Rosso (2002), with drainage area ranging from 10 to 2500 km^2, not accounting for those located along the main stream, instead of the 20 sites used here, with at least 20 years of observations).

In De Michele and Rosso (2002), four distinct "homogenous" Regions (an analogous of the notion of cluster) were identified – two in the Alps, and two in the Apennines, intended as geographical regions – according to seasonality indices and the scale invariant behavior of floods with respect to the basin area. In particular (see also Table 1 and Fig. 2): (A) Region A, or Central Alps and Prealps, which includes the Po sub-basins from Chiese to Sesia river basins; (B) Region B, or Western Alps and Prealps, from Dora Baltea river to Rio Grana; (C) Region C, or North-Western Apennines and Tyrrhenian basins, which include the Po sub-basins from Scrivia river basin to Taro river basins, and in addition Ligurian basins with outlet to the Tyrrhenian sea; (D) Region D, or North-Eastern Apennines, from Parma to Panaro river basin. Furthermore, a "Transition Zone" was identified, represented by the Tanaro river basin, a catchment where the sites have characteristics similar to (two, or more) neighboring clusters.

According to the map drawn by De Michele and Rosso (2002), the 20 sites considered in this study are spread as follows:

- Region A includes the sites 1–5,
- Region B includes the sites 6–9,
- Region C includes the site 12,
- Region D includes the sites 13–14, and
- the Transition Zone includes sites 10–11.

The dendrograms presented in Fig. 6, concerning the Flood Peak variable only, and the AND and Kendall approaches, show that two well-separated groups can be distinguished in both cases, while several smaller sub-clusters are also evident. At a finer level of aggregation, it can be seen that a 6-cluster solution is a feasible one; this is also confirmed by the graph of W_b against the number of clusters (see Fig. 7 (bottom)), which is practically constant from $k = 6$ onwards. The resulting partitions are presented in Fig. 8, yielding the same AND and Kendall clusters.

Interestingly enough, the eventual configuration is quite close to the one outlined in De Michele and Rosso (2002): as above, the behavior of site 6 (Tavagnasco) makes an exception, showing a dynamics different from the one of the other closest sites located in the Piedmont region, viz. sites 7–9. Moreover, sites 10–11 (Farigliano and Montecastello, also belonging to the Piedmont region) are assigned a behavior different from the one of sites 7–9 located in the same region, which agrees with the distinction between Region B and the Transition Zone made in De Michele and Rosso (2002). It should be point out that such agreement is not trivial and is particularly remarkable, since the variable considered (i.e., discharge) is not the same in the two studies, being sampled at different time resolutions (annual (maximum) and daily).

5. Conclusions

In this work, inspired by the EU Directive 2007/60/EC on the assessment and the management of flood risks, we investigated the concurrence of flood episodes in a given region, in order to identify common patterns of flood dynamics: this represents an essential information for the evaluation of the potential threatening of flood occurrences. In particular, each flood event has been considered

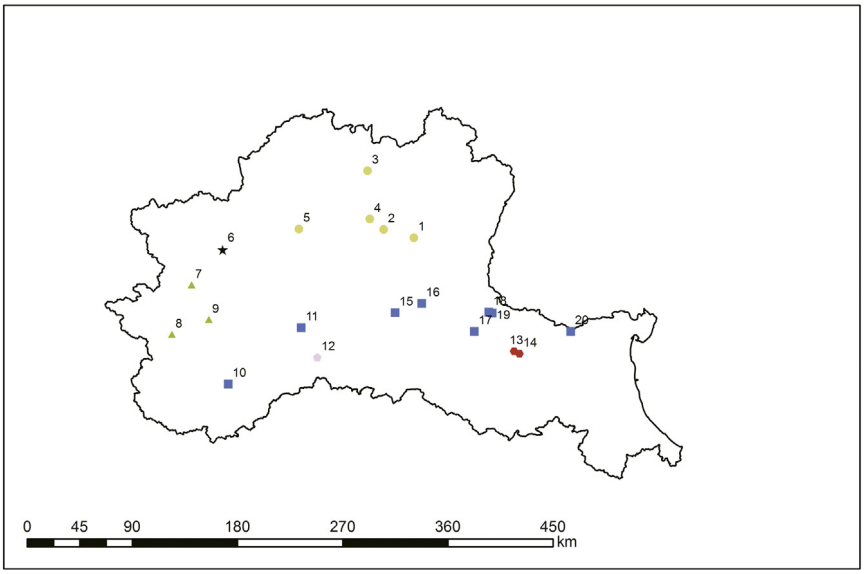

Fig. 8. Flood Peak. Map of 6-cluster solution as extracted from the dendrograms shown in Fig. 6, for the AND and the Kendall dissimilarity measure approach—see text. The clusters are indicated by using different markers and colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

as a multivariate object characterized by a number of physical variables, like Flood Peak and Flood Volume.

The methodology proposed aims at identifying spatial sub-regions characterized by similar flood regimes, by taking into account the interaction between the variables at play. It is grounded on the concept of Hazard Scenarios (namely, the AND and the Kendall ones), and a Copula-based Agglomerative Hierarchical Clustering approach, thus representing an innovative contribution concerning the regionalization problem, of utmost importance in hydrological applications. The procedure has been applied to 20 gauge stations spread across the Po river basin (the largest Italian one), a strategic area for Italian economy.

The analyses outlined in this work comprise both the bivariate approach, where the joint behavior of the variables Flood Peak and Flood Volume is considered, and the univariate case (where only the Flood Peak is used), the latter approach being useful for comparing the corresponding outcomes with those present in literature. The clusters detected by the proposed algorithms seem to adequately capture the distinction between different meteorological forcings and hydrological flow contributions, although the dissimilarity measure adopted is obtained via two different approaches. As outlined in Section 4.1, the integration of Flood Peak and Flood Volume information may provide a more comprehensive picture of flood dynamics and threatening with respect to univariate approaches/scenarios focusing the attention on one variable only (as in Section 4.2), with obvious advantages in terms of flood risk assessment—e.g. for the estimate of design quantiles in ungauged basins. This may represent a progress over, e.g., the past work by De Michele and Rosso (2002), and provide a promising and valuable investigation tool to identify possible regions where risk and water managers should adopt a cross-border strategy to mitigate and prevent the hydrological risks.

Notably, the methodology proposed is a general one, and could also be applied to other kind of environmental events of interest, like droughts or heat waves.

Acknowledgments

The first Author thanks F. Pauli (University of Trieste, Trieste, Italy) for helpful discussions and software recommendations. Third author's helpful discussions with C. Sempi (Università del Salento,

Italy) are acknowledged. The support of the CMCC (Centro Euro-Mediterraneo sui Cambiamenti Climatici, Lecce (Italy)) is acknowledged.

References

AA. VV., 2011. Le magre del Po. Conoscerle per prevederle, cooperare per prevenirle. Fondazione CIMA, Parma (Italy), (in Italian: www.cimafoundation.org).

Barbe, P., Genest, C., Ghoudi, K., Rémillard, B., 1996. On Kendall's process. J. Multivariate Anal. 58 (2), 197–229.

Benson, M.A., 1962. Evolution of methods for evaluating the occurrence of floods. U.S. Geological Survey Water Supply Paper (1580A).

Caiado, J., Maharaj, E.A., D'Urso, P., 2015. Time series clustering. In: Hennig, C., Meila, M., Murtagh, F., Rocci, R. (Eds.), Handbook of Cluster Analysis. Chapman and Hall/CRC, pp. 241–264.

Chow, V.T., Maidment, D.R., Mays, L.W., 1988. Applied Hydrology. In: McGraw-Hill Series in Water Resources and Environmental Engineering, Tata McGraw-Hill Education.

De Luca, G., Zuccolotto, P., 2011. A tail dependence-based dissimilarity measure for financial time series clustering. Adv. Data Anal. Classif. 5 (4), 323–340.

De Michele, C., Rosso, R., 2001. Uncertainty assessment of regionalized flood frequency estimates. J. Hydrol. Eng. 6 (6), 453–459.

De Michele, C., Rosso, R., 2002. A multi-level approach to flood frequency regionalization. Hydrol. Earth Syst. Sci. 62 (2), 185–194.

Di Lascio, F.M.L., Durante, F., Pappadà, R., 2017. Copula–based clustering methods. In: Úbeda Flores, M., de Amo, E., Durante, F., Fernández-Sánchez, J. (Eds.), Copulas and Dependence Models with Applications. Springer, Berlin, pp. 49–67.

Durante, F., Pappadà, R., Torelli, N., 2014. Clustering of financial time series in risky scenarios. Adv. Data Anal. Classif. 8, 359–376.

Durante, F., Pappadà, R., Torelli, N., 2015. Clustering of time series via non–parametric tail dependence estimation. Statist. Papers 56 (3), 701–721.

Embrechts, P., McNeil, A.J., Straumann, D., 2002. Correlation and dependence in risk management: properties and pitfalls. In: Dempster, M. (Ed.), Risk Management: Value at Risk and Beyond. Cambridge University Press, Cambridge, pp. 176–223.

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., Rémillard, B., 2004. Tests of independence and randomness based on the empirical copula process. TEST 13 (2), 335–370.

Genest, C., Rivest, L.P., 1993. Statistical inference procedures for bivariate Archimedean copulas. J. Amer. Statist. Assoc. 88 (423), 1034–1043.

Genest, C., Rivest, L.P., 2001. On the multivariate probability integral transformation. Statist. Probab. Lett. 53 (4), 391–399.

Grimaldi, S., Petroselli, A., Salvadori, G., De Michele, C., 2016. Catchment compatibility via copulas: A non-parametric study of the dependence structures of hydrological responses. Adv. Water Resour. 90, 116–133.

Guthke, P., Bárdossy, A., 2017. On the link between natural emergence and manifestation of a fundamental non-Gaussian geostatistical property: Asymmetry. Spat. Stat. 20, 1–29.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed. In: Springer Series in Statistics, Springer, New York.

IPCC, 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Cambridge, UK, a Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change.

Jongman, B., Hochrainer-Stigler, S., Feyen, L., Aerts, J.C.J.H., Mechler, R., Botzen, W.J.W., Bouwer, L.M., Pflug, G., Rojas, R., Ward, P.J., 2014. Increasing stress on disaster-risk finance due to large floods. Nature Clim. Change 4, 264–268.

Kojadinovic, I., Yan, J., 2010. Modeling multivariate distributions with continuous margins using the copula R package. J. Stat. Softw. 34 (9), 1–20. URL http://www.jstatsoft.org/v34/i09/.

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., Stafford-Smith, M., 2014. A compound event framework for understanding extreme impacts. Wiley Interdiscip. Rev. Clim. Change 5 (1), 113–128.

Maharaj, E., 2000. Clusters of time series. J. Classification 17 (2), 97–314.

McNeil, A.J., Nešlehová, J., 2009. Multivariate archimedean copulas, d-monotone functions and -norm symmetric distributions. Ann. Statist. 3059–3097.

Mornet, A., Opitz, T., Luzi, M., Loisel, S., Bailleul, B., 2016. Wind storm risk management: sensitivity of return period calculations and spread on the territory. Stoch. Environ. Res. Risk Assess. 1–19.

Nelsen, R.B., 2006. An Introduction to Copulas, second ed. In: Springer Series in Statistics, Springer, New York.

Nelsen, R.B., Quesada-Molina, J.J., Rodrıguez-Lallena, J.A.,Úbeda-Flores, M., 2001. Distribution functions of copulas: a class of bivariate probability integral transforms. Statist. Probab. Lett. 54 (3), 277–282.

Nelsen, R.B., Quesada-Molina, J.J., Rodrıguez-Lallena, J.A.,Úbeda-Flores, M., 2003. Kendall distribution functions. Statist. Probab. Lett. 65 (3), 263–268.

Patton, A.J., 2012. A review of copula models for economic time series. J. Multivariate Anal. 110, 4–18.

Poulin, A., Huard, D., Favre, A.-C., Pugin, S., 2007. Importance of tail dependence in bivariate frequency analysis. J. Hydrol. Eng. 12, 394–403.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramachandra Rao, A., Srinivas, V.V., 2008. Regionalization of Watersheds: An Approach Based on Cluster Analysis. Springer, Netherlands.

Salvadori, G., De Michele, C., 2010. Multivariate multiparameter extreme value models and return periods: A copula approach. Water Resour. Res. 46, W10501. http://dx.doi.org/10.1029/2009WR009040.
Salvadori, G., De Michele, C., 2011. Estimating strategies for multiparameter Multivariate Extreme Value copulas. Hydrol. Earth Syst. Sci. 15, 141–150. http://dx.doi.org/10.5194/hess-15-141-2011.
Salvadori, G., De Michele, C., Kottegoda, N.T., Rosso, R., 2007. Extremes in Nature. An Approach using Copulas. In: Water Science and Technology Library Series, vol. 56, Springer, Dordrecht, ISBN: 978-1-4020-4415-1.
Salvadori, G., Durante, F., De Michele, C., Bernardi, M., Petrella, L., 2016. A multivariate copula-based framework for dealing with hazard scenarios and failure probabilities. Water Resour. Res. 52 (5), 3701–3721.
Sklar, A., 1959. Fonctions de répartition à *n* dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, 229–231.
The European Parliament and The Council, 6.11.2007. Directive 2007/60/EC: on the assessment and management of flood risks. Official Journal of the European Union. URL eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007L0060&from=EN.
Vahedifard, F., AghaKouchak, A., Jafari, N.H., 2016. Compound hazards yield Louisiana flood. Science 353 (6306), 1374.
Ward Jr., J.H., 1963. Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc. 58 (301), 236–244.
Yan, J., 2007. Enjoy the joy of copulas: with a package copula. J. Stat. Softw. 21 (4), 1–21. URL http://www.jstatsoft.org/v21/i04/.
Yevjevich, V., 1967. An objective approach to definitions and investigations of continental hydrologic droughts. Hydrologic Paper 23, Colorado State University, Fort Collins.
Zelenhasić, E., Salvai, A., 1987. A method of streamflow drought analysis. Water Resour. Res. 23 (1), 156–168.
Zscheischler, J., Seneviratne, S.I., 2017. Dependence of drivers affects risks associated with compound events. Sci. Adv. 3 (6),