© 3rd IAHR Europe Congress, Book of Proceedings, 2014, Porto - Portugal. ISBN 978-989-96479-2-3

MULTIVARIATE STATISTICAL ANALYSIS OF FLOOD VARIABLES BY COPULAS: TWO ITALIAN CASE STUDIES

MATTEO BALISTROCCHI, ROBERTO RANZI & BALDASSARE BACCHI

University of Brescia, DICATAM, Brescia, Italy
matteo.balistrocchi@unibs.it, roberto.ranzi@unibs.it, baldassare.bacchi@unibs.it

Abstract

Multivariate statistics are important to determine the flood hydrograph for the design of hydraulic structures and for the hydraulic risk assessment. In the last decade, the copula approach has been investigated in hydrological practice to assess the design flood hydrograph in terms of flood peak, volume and duration. In this paper, the copula approach is exploited to perform pair analyses of these three random variables for two Italian watersheds, in the Apennine and the Alps respectively. The criterion to separate continuous flow series into independent events is discussed along with its implications on the dependence structure. The goodness-of-fits of the proposed copulas are then assessed by non-parametric tests. Marginal distributions to derive joint distributions are briefly suggested. The possibility of generating flood events according to the proposed model and potential applications to hydraulic structure design and flood management are finally examined.

Keywords: Flood frequency analysis; Copula; Joint distributions; Goodness-of-fit tests.

1. Introduction

In many practical engineering applications regarding the wet weather discharge management, a certain number of hydrograph features play a significant role. On one hand, the peak flow discharge represents the key variable in any issue involving the conveyance capacity of the drainage network. On the other hand, when dealing with storage facilities, the most important aspect is undoubtedly the flood volume (Bergman and Sackl, 1989; Bacchi *et al.*, 1992; Fiorentino and Margiotta, 1999; Franchini and Galeati, 2000; De Michele *et al.*, 2005; Balistrocchi *et al.*, 2013). Additionally, time pattern characteristics, such as time to peak or flood duration, are relevant information, for instance, in warning system implementation, in real time management, in hydraulic vulnerability assessment or in ecological and sediment transport issues.

Nonetheless, conventional approaches to flood frequency analysis, essentially based on the seminal works by Gumbel (1958) and Todorovic (1978), mainly accounted for the peak flow discharge statistics. Unfortunately, all above mentioned hydrologic variables, though slightly or strongly associated, generally feature diverse return periods even within the same event, owing to the multivariate non stationary nature of the discharge process.

Despite many valuable research efforts accomplished to employ the multivariate approach (Goel *et al.*, 1998; Yue, 2000; Yue, 2001), inferring popular multivariate distributions by conventional methods showed to be complex and affected by several drawbacks.

The most important ones consist in the necessity to utilize marginal distributions belonging to the same family and in the impossibility of separating the assessment of the dependence structure from those of the marginal distributions. As a consequence, their effectiveness has remained questionable and such models found scarce applications in the real world practice.

Since then, a significant advance has been attained by introducing the copula approach (Joe, 1997; Nelsen, 2006) in the hydrological statistical analysis (Salvadori *et al.*, 2007), through which most of previous concerns are effectively overcome. By using copula functions, in fact, various dependence structures can be easily detected, while different and complex marginal distributions can be implemented into a unique probabilistic model. More recently, non-parametric statistical tests have been made available to verify in a quantitative statistically rigorous manner the goodness-of-fit of the selected copulas (Genest *et al.*, 2009).

In the present work copula functions have been exploited to perform pair analyses of flood variables with regard to two Italian river stations belonging to Panàro River and Tagliamento River, for which long flow rate records exist. Bearing in mind potential hydraulic engineering applications, such as routing reservoir or safety spillway design, three variables have been analyzed: the peak flow discharge, the flood volume and the flood duration.

2. Case studies

Although analyzed watersheds both belong to northern Italy, they offer quite diverse case studies, thanks to their varying climatic and physiographic characteristics. Figure 1 shows their location while table 1 reports some key characteristics: the catchment area A, the main river length L, the difference between the average watershed elevation and the outlet elevation ΔH and the time of concentration t_c estimated according to the Giandotti formula (Giandotti, 1934), a commonly applied tool in the Italian hydrologic practice, the average annual cumulative rainfall depth h and the average annual runoff coefficient Φ.

As can be seen in the left-hand side of figure 1, the Panàro River is the last right-bank tributary to the Po River and springs from the northern edge of the Apennine chain, close to the watershed divide. Its course follows an almost straight north-east direction to the outlet located next to Ficarolo (where the first historical bank break of the Po River took place in the XII century). The catchment area is thus elongated and has a rather constant width because of two relevant left-bank tributaries.

The Tagliamento River drains a watershed belonging to the eastern range of Carnic Alps, as illustrated in the right-hand side of figure 1. It is characterized by a compact mountain catchment area, that represents most of the total, in which the flow is directed eastward. Downstream of this area, the river runs in the south direction to the estuary in the northern coast of the Adriatic sea. Here, the additional catchment area is only given by a narrow strip along the river course.

Despite the larger catchment area, the shorter main stream length and the greater elevation drop make the hydrologic response time of the Tagliamento watershed more rapid than that of the Panàro watershed, as evidenced by the times of concentration in table 1.

The Panaro hydrologic regime is essentially driven by the rainfall one, which features two maxima: the main one in autumn and the secondary one in spring (classified as sub-Apennine by Bandini, 1931). During summer, hydrologic losses are relevant so that null discharges can be often observed and, in autumn, the discharge increasing trend is delayed with respect to the rainfall one.

As a consequence, major floods typically occur in November-December, when suitable soil moisture conditions are reestablished. The monthly runoff coefficient variability is relatively high, since it includes values of 0.08 in August and 0.86 in February, while the average annual value amounts to 0.47.

The Tagliamento hydrologic regime is sensitive to both rainfall and snowmelt. Moreover, according to the classification by Bandini, the rainfall regime varies from the continental-Alpine one in the mountain area (a maximum in summer and a minimum in winter), to the sub-Alpine one (two maxima in spring and autumn) in the plain area. This results in a quite perennial flow regime which shows two relevant maxima in May, related to snowmelt, and in October, related to heavy precipitations. Monthly runoff coefficients are subject to a low variability, ranging from 0.44 in February, up to 1.26 in May; the average annual value is assessed in about 0.80.

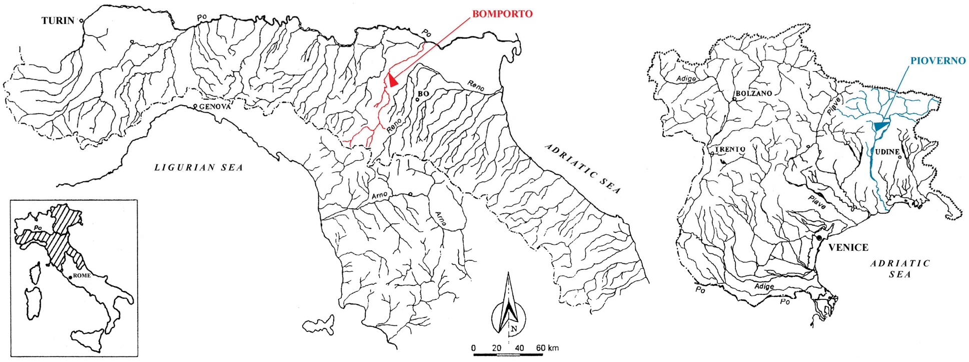

Figure 1. Locations of the Panàro River (red network) and the Tagliamento River (blue network), evidencing corresponding stream gauge stations (redrawn from Bacchi et al., 2000).

In addition to its higher impermeability, the Tagliamento watershed is affected by significantly greater precipitations. The annual cumulative rainfall depth in fact is less than 1200 mm for the Panàro watershed but almost 1800 mm for the Tagliamento watershed.

Table 1. Main hydrologic characteristics of studied watersheds (Servizio Idrografico, 1980).

WATERSHED	STATION	A (km^2)	L (km)	ΔH (m)	t_c (h)	h (mm)	Φ
Panàro	Bomporto	1036	106	643.6	14.2	1150	0.47
Tagliamento	Pioverno	1880	65	936.7	11.1	1780	0.80

3. Data analysis

As evidenced in table 2, several decades of discharge observations (Bacchi et al., 2000) can be exploited for the above described stations. Since only the first and last years of the record are listed and some years are missing, the number of actually available years is stated, as well.

To make such data suitable for the copula analysis, the continuous record had to be separated into independent flood events. Herein, a two-step criterion was utilized.

Firstly, according to the partial duration series approach delineated by Todorovic (1978), a discharge threshold q_s was set. Individual flood discharges and base flow discharges were accordingly discriminated.

Secondly, the independence of such floods was verified. To do so, time periods included between the peak of the antecedent event and the onset of the following one were compared to a minimum interevent period t_i: if greater, the hydrographs were assumed to be independent. Conversely, the complete hydrograph between the beginning of the first one and the end of the latter one was identified as an individual independent flood.

Thus, the previously mentioned random variables were computed. In this formulation, the total discharge, not only the portion exceeding the threshold discharge q_s, was accounted for. That is, the peak flow discharge q_p and the flood volume v were calculated, respectively, as the maximum and the integral of the observed discharge, over the partial duration t of the identified flood event. Although only the flow rate above the threshold is commonly considered, this choice is intended to analyze the behavior of the actual discharge conveyed by a canal and entering, for instance, a hypothetical on-line storage facility.

Table 2. Consistencies of the analyzed data.

WATERSHED	STATION	OBSERVED YEARS	OBSERVATION PERIOD	SAMPLING INTERVAL
Panàro	Bomporto	52	1923:1983	1 h
Tagliamento	Pioverno	49	1886:1975	1 h

4. Dependence structure assessment

Thanks to the Sklar theorem (Sklar, 1959), copula functions operate in the probability space, so that pseudo-observations must be derived from the original ones by means of the probability integral transformations, as shown in the equation set written below.

In these equations F_Q, F_V and F_T are the Weibull standard plotting positions of the variables q_p, v and t, while x, y and z are the corresponding pseudo-observations, uniformly and identically distributed.

$$x = F_Q(q_p); \quad y = F_V(v); \quad z = F_T(t) \qquad [1]$$

The assessment of the overall dependence structure was dealt with by pair analyses, since variable sensitivities with respect to the parameter set employed to separate the independent events can be evaluated by association measures such as the Kendall coefficient τ_K (Kendall, 1937).

A visual comparison between bivariate theoretical copulas and empirical copulas can also guide the prescreening of best fitting functions. Most important, asymmetric association degrees can be accounted for, before the complete probabilistic model is constructed.

4.1 Sensitivity to discretization parameter

The sensitivity analysis was chiefly aimed at identifying the effect of the threshold discharge on the association degree and the mean annual event number. In fact, the minimum interevent time can be reasonably set on a hydrologic or operational basis, with regard to the watershed time of concentration or residence time in a floodplain: in this study, the hypothesis that the flood recession phase must last more than the double of t_c to ensure adjacent event independence was assumed (only major floods were herein taken into consideration). In consideration of diverse discharge regimes q_s was varied between 10:100 m^3/s in the Bomporto station and between 50:1200 m^3/s in the Pioverno station.

The τ_K coefficient trends are illustrated in figure 2a. Significant concordant associations are always detected for the Panàro case, as *p-values*, obtained from testing the no-association hypothesis, are definitely close to zero for every value of the threshold discharge. Increasing trends are related to low values of q_s, whereas an almost constant association degree is reached for unreasonably high threshold values.

In both cases, the flood volume - flood duration pair and the peak flow discharge-flood duration pair systematically feature the strongest (τ_K greater than 0.80) and the weakest (τ_K less than about 0.60) concordances, respectively. For the peak rate-duration pair, τ_K takes values in an intermediate range: closer to 0.80 in the Panàro watershed. Unlike studies by other authors (Grimaldi and Serinaldi, 2006), intersections with the flood volume – flood duration curve cannot be observed when the q_s threshold is increased.

Essentially akin behaviors for τ_K coefficients are revealed in the Pioverno station in terms of both trend and significance, figure 2b, even if some irregularities are shown. This could be partly due to a worse quality of the original data series. In this case, water stage data were employed to reconstruct flood hydrographs by a discharge rating curve. Observations collected in a downstream station, for which a broader data set was available, were used as well. To do so, water stages in the station of interest were obtained by a regression curve from downstream ones. A certain influence of an additional catchment area is expected as stations are not very close. In addition, difficulties arose for low flows both in rising limbs and in recession limbs because of the braided nature of the riverbed.

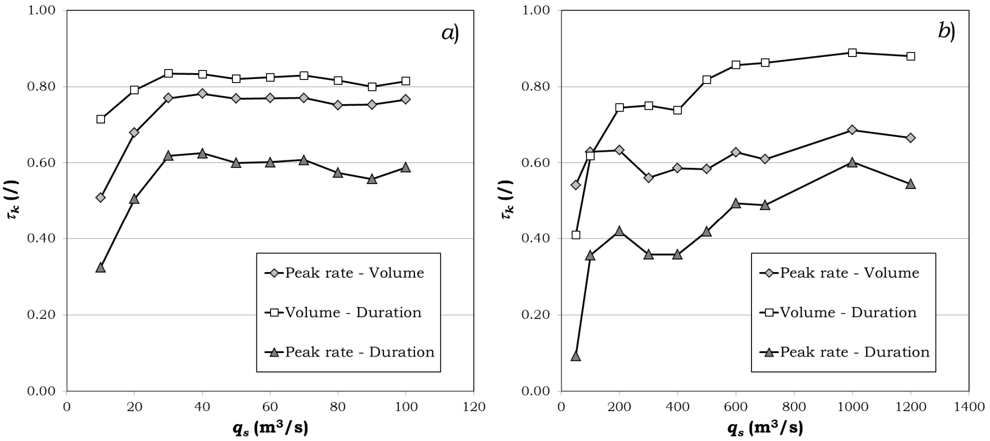

Figure 2. Trend of Kendall coefficients with respect to increasing discharge thresholds for the Bomporto station *a*) (t_i = 1.2 d) and the Pioverno station *b*) (t_i = 0.9 d).

The behavior of the mean annual event number □, which plays a relevant role in the return period estimation, is depicted in figure 3. Although fewer flood events are expected in the Tagliamento watershed, trends are qualitatively analogous since in both cases a maximum is shown. This can be explained by taking into account the combination of two opposite occurrences as q_s increases: the event number augments as multi peak floods are splitted into distinct floods, whilst it diminishes as minor floods are suppressed when the peak flow discharge is less than the threshold. For low q_s values the first one prevails leading to the □ rising trend, conversely for large q_s values the decreasing trend is due to the predominance of the second one.

In view of the variability evidenced by □ in figure 3, an accurate parameter selection during the identification of independent floods seems to be crucial in developing reliable probabilistic models mainly in consideration of this second aspect. In fact, small variations of the mean annual event number results in huge variations of the return period estimates. The use of the minimum inter event time t_i however contributes to the reduction of □ variability.

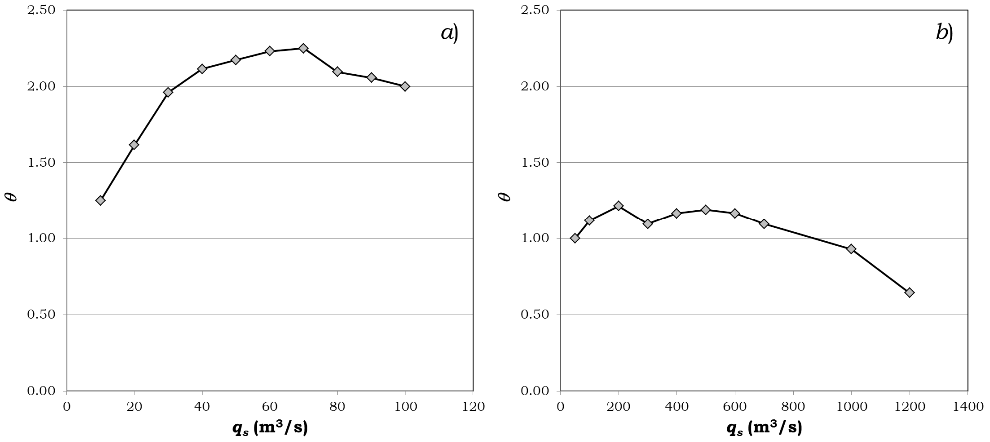

Figure 3. Variability of the mean annual event number θ with respect to increasing discharge thresholds for the Bomporto station *a*) (t_i = 1.2 d) and the Pioverno station *b*) (t_i = 0.9 d).

4.2 Copula selection and testing

The pair analysis strategy was exploited for selecting the most suitable copula functions, as well. Despite climatic and hydrologic diversities, very similar dependence structures were detected in both case studies. Herein, the ability of the Gumbel-Hougaard copula to suit pair data samples was tested.

Copulas belonging to this family model dependence structures characterized by concordant association and by upper tail dependence. As already noticed by other authors (De Michele *et al.*, 2005), who exploited the same copula in multivariate flood analysis, the stronger association of extreme events accords well with the empirical evidence. Indeed, occurrences of random variables related to the most severe floods are likely to have the same frequency.

Considering a generic couple of random variables ξ and η uniformly distributed in I = [0,1], the expression of the bivariate Gumbel-Hougaard theoretical copula C_α can be written in the terms of equation [2], where $\alpha \geq 1$ is the dependence parameter.

$$C_{\alpha}(\xi,\eta)=\exp\left\{-\left[(-\ln\xi)^{\alpha}+(-\ln\eta)^{\alpha}\right]^{1/\alpha}\right\} \qquad [2]$$

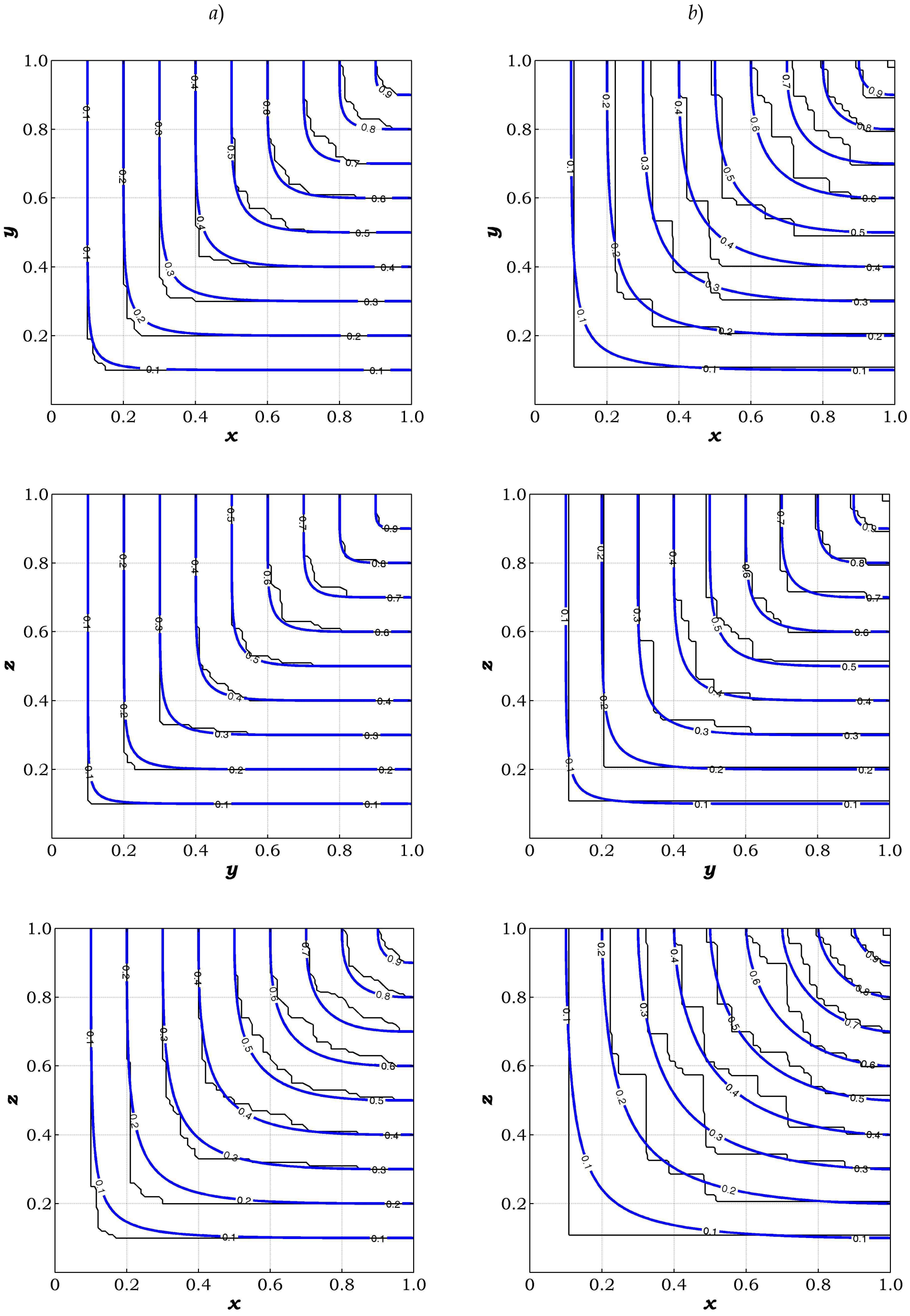

Figure 4. Contours of fitted Gumbel-Hougaard copulas (blue thick line) compared to those of empirical copulas (black thin line) for the Bomporto station *a*), and for the Pioverno station *b*).

The sample counterpart is instead represent by the empirical copula C_n, whose expression is provided by the sum in [3], in which N is the sample size, $\mathbf{1}(.)$ is the indicator function and $\hat{\xi}_i, \hat{\eta}_i$ are pseudo-observations (i = 1 … N).

$$C_n(\xi,\eta)=\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(\hat{\xi}_i \le \xi, \hat{\eta}_i \le \eta\right) \qquad [3]$$

In figure 4 contour lines of the empirical copulas and fitted Gumbel-Hougaard copulas are drawn for the three pairs giving a visual evaluation of the goodness-of-fit. In this example the identification of independent floods was performed by assuming q_s = 40 m^3/s t_i = 1.2 d in the Panàro case and q_s = 250 m^3/s t_i = 0.9 d in the Tagliamento case. Bearing in mind the discussion in previous subsection, such values were considered to be suitable for a meaningful evaluation in the selected cases studies. The calibration of the Gumbel-Hougaard copulas was carried out by exploiting the relationship between the dependence parameter α and the Kendall coefficient τ_k [4]. Calibration parameters are listed in table 3.

$$\alpha = 1/\left(1-\tau_k\right) \qquad [4]$$

Despite the good agreement evidenced by the contour overlaps shown in figure 4, test statistics are however mandatory to quantitatively assess the goodness-of-fit of the selected functions. Particular attention has been recently paid to blanket tests (Genest *et al.*, 2006), because they can be applied to every copula family and do not require any subjective choice for their implementation, such as those regarding smoothing parameters, weight functions or arbitrary data categorizations. An effective blanket test proposed by Genest *et al.* (2009) is based on the Cramer-Von Mises statistics S_n, whose expression in a bivariate case can be written as in the sum [5].

$$S_n = \sum_{j=1}^{N}\left[C_n(\hat{\xi}_j,\hat{\eta}_j)-C_\alpha(\hat{\xi}_j,\hat{\eta}_j)\right]^2 \qquad [5]$$

If a large number m of data samples is generated by copula simulations (Salvadori *et al.*, 2007) under the hypothesis that the selected copula is the underlying copula (null hypothesis H_0), an appropriate estimate of the *p-value* is given by expression [6]. In this sum $S_{n,k}$ are Cramer-Von Mises statistics evaluated for simulated samples.

$$p-value = \frac{1}{m}\sum_{k=1}^{m}\mathbf{1}\left(S_{n,k} > S_n\right) \qquad [6]$$

Table 3. Copula calibration parameters and *p-values* testing the null hypothesis H_0.

STATION	θ	PAIR	τ_k	α	*p-value* (%)
Bomporto	2.12	x-y	0.78	4.55	96.74
		y-z	0.83	5.88	99.50
		x-z	0.63	2.70	76.99
Pioverno	1.14	x-y	0.60	2.50	98.88
		y-z	0.76	4.17	99.95
		x-z	0.40	1.67	98.45

The H_0 hypothesis was tested by using m = 10000 and *p-values* reported in table 3 were obtained. It is evident that H_0 cannot be rejected for very large levels of significance. Despite this suitability, a comparative analysis should be conducted to evaluate whether other copula functions featuring different dependence structure, namely symmetric tail dependence (*t*-Student), no tail dependence (Frank or Gaussian copulas) or lower tail dependence (Clayton copula), are more representative.

5. Marginal distributions

To derive bivariate joint distributions by means of the Sklar theorem, marginal distributions must be defined. Indeed, a number of studies has already been conducted with regard to the peak flow discharge statistics by this time, see for example Cunanne (1987) and references therein. Popular solutions usually rely on the log-normal distribution (LN), the Gumbel distribution (EV1) and the Generalized Extreme Value distribution (GEV).

Nevertheless, none of these functions usually suits upper outlier occurrences. In such a case, two component models, namely Two Component Extreme Value distribution (TCEV), should be adopted. Owing to the greater number of calibration parameters, samples of adequate size are however needed to achieve a reasonable model reliability.

Referring to selected case studies, LN distribution and EV1 distributions were found to be the best models in the Bomporto station and in the Pioverno station, respectively. In addition to such distributions, the Weibull function can be suggested as an alternative solution in order to represent flood volume and, especially, flood duration distributions.

6. Conclusions

Pair statistical analyses of three flood variables, peak flow, flood volume and flood duration, were herein performed by means of the copula approach with reference to two Italian watersheds. Despite the relevant climatic and geo-morphological diversities, the Gumbel-Hougaard copula demonstrated to yield very similar fits for each of the analyzed pairs. Good agreements have been shown as *p-values*, that is "empirical significances", higher than 95% for 5 out of 6 investigated pairs and equal to 77% for the peak and duration pair in the Panàro River case were estimated.

At the same time, if proper ranges are chosen, comparable sensitivities with respect to thresholds employed to extract the independent events from the continuous discharge series are evidenced. Nonetheless, the possibility of expressing such dependence structures by copulas other than the Gumbel-Hougaard remains to be investigated in more detail.

Practical applications of multivariate distributions may be addressed towards their implementation into stochastic processes devoted to the generation of extended discharge time series. Thus, continuous simulations of storage capacities, spillway devices or, in general, flood control structures could be utilized for planning, design or management purposes.

Firstly, a hydrograph shape depending on three variables must be chosen; secondly the trivariate copula can be constructed by means of the bivariate ones. In fact, as made clear by association trends in figure 2, such a trivariate copula needs to take into account different association degrees among the three pairs.

However, popular shapes of flood hydrographs are expressed by two parameter functions: the Gamma distribution function by Ranzi (2005), the triangular shape introduced by Wycoff and Singh (1976). On the other hand, given the very strong association demonstrated by the flood volume - flood duration pair, a first attempt could be undertaken by using only the peak rate and the flood volume and expressing the duration as a function of the volume by an analytical relationship.

Finally, since the initial condition always play a key role in any device performance assessment, the interevent period variability must be accounted for. To do so, this additional random variable can be correctly assumed as independent of the others and distributed according to a lower bounded Weibull function (Balistrocchi and Bacchi, 2011).

References

Bacchi B, Brath A, Kottegoda N T, 1992. ‘Analysis of the relationship between flood peaks and flood volumes based on crossing properties of river flow processes’, *Water Resour Res*, 28, 10, 2773-2782.

Bacchi B, Franchini M, Galeati G, Ranzi R, 2000. ‘Parametrizzazione e regionalizzazione della curva di riduzione dei massimi annuali delle portate medie su assegnata durata’, Proc. XXVII Convegno di Idraulica e Costruzioni Idrauliche, 12-15 September, Genova.

Balistrocchi M, Bacchi B, 2011. ‘Modelling the statistical dependence of rainfall event variables through copula functions’, *Hydrol Earth Syst Sci*, 15, 1959–1977.

Balistrocchi M, Grossi G, Bacchi B, 2013. ‘Deriving a practical analytic-probabilistic method to size routing reservoirs’, *Adv Water Resour*, 62, 37–46.

Bandini A, 1931. ‘Tipi pluviometrici dominanti sulle regioni italiane’, Servizio Idrografico Italiano, Roma, IT.

Bergman H, Sackl B, 1989. ‘Determination of design flood hydrographs based on regional hydrological data’, Proc. Symp. on New Directions for Surface Water Modelling, IAHS pub. 181, Baltimora, May, 261-269.

Cunanne C, 1987. ‘Review of statistical model for flood frequency estimation’, Hydrologic frequency modelling, Singh VP ed., Reidel, Dordrecht, NR, 49-95.

De Michele C, Salvadori G, Canossi M, Petaccia A, Rosso R, 2005. ‘Bivariate statistical approach to check adequacy of dam spillway’, *J Hydrol Eng*, 10, 1, 50–57.

Fiorentino M, Margiotta MR, 1999. ‘La valutazione dei volumi di piena e il calcolo semplificato dell'effetto di laminazione dei grandi invasi’, In Tecniche per la difesa dall’inquinamento, Frega G (ed.), Bios, Cosenza, IT, 203-222.

Franchini M, Galeati G, 2000. ‘Comparative analysis of some methods for deriving the expected flood reduction curve in the frequency domain’, *Hydrol Earth Syst Sc*, 4, 1, 155-172.

Genest C, Quessy JF, Rémillard B, 2006. ‘Goodness-of-fit procedures for copula models based on the probability integral transformation’, *Scand J Stat*, 32, 2, 337–366.

Genest C, Rémilland B, Beaudoin D, 2009. ‘Goodness-of-fit tests for copulas: a review and a power study’, *Insur Math Econ*, 44, 2, 199–213.

Giandotti M, 1934. ‘Previsione delle piene e delle magre dei corsi d’acqua’, Servizio Idrografico Italiano, Rome, IT.

Goel NK, Seth SM, Chandra S, 1998. ‘Multivariate modeling of flood flows’, *J Hydraul Eng-ASCE*, 124, 2, 146–155.

Grimaldi S, Serinaldi F, 2006. ‘Asymmetric copula in multivariate flood frequency analysis’, *Adv Water Resour*, 29, 8, 1115–1167.

Gumbel EJ, 1958. 'Statistics of extremes', Columbia University Press, New York.

Kendall MG, 1938. 'A new measure of the rank correlation', *Biometrika*, 30, 81–93.

Joe H, 1997. 'Multivariate models and dependence concepts', Chapman & Hall, London, UK.

Nelsen RB, 2006. 'An introduction to copulas. Second edition', Springer, New York, NY.

Ranzi R, 2005. 'Structural and non-structural methods for flood hazard mitigation', Proc. Workshop on Natural Environment, Sustainable Protection and Conservation: Italy-Vietnam Cooperation perspectives, 15-17 November 2004, Hanoi, Viet-Nam, 108-118.

Salvadori G, De Michele C, Kottegoda NT, Rosso R, 2007. 'Extremes in nature: an approach using copulas', Springer, Dordrecht, NL.

Servizio Idrografico, 1980. 'Dati caratteristici dei corsi d'acqua italiani. Pubblicazione 17. V edizione aggiornata al 1970', Istituto Poligrafico dello Stato, Roma, IT.

Sklar A, 1959. 'Fonctions de répartition à n dimensions et leures marges', Publ. Inst. Statist. Univ. Paris 8, 229–231.

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

Wycoff RL, Singh UP, 1976. 'Preliminary hydrologic design of small flood detention reservoirs', *Water Resour Bull*, 12, 2, 337-349.

Yue S, 2000. 'The bivariate lognormal distribution to model a multivariate flood episode', *Hydrol Process*, 14, 14, 575–588.

Yue S, 2001. 'A bivariate lognormal distribution to model a multivariate flood episode', *Hydrol Process*, 15, 6, 1033–1045.