On the exact distribution of correlated extremes in hydrology

F. Lombardo[1,2], F. Napolitano[1], F. Russo[1], and D. Koutsoyiannis[1,3]

[1] Dipartimento di Ingegneria Civile, Edile e Ambientale, Sapienza Università di Roma, Via Eudossiana, 18 – 00184 Rome, Italy.

[2] Corpo Nazionale dei Vigili del Fuoco, Ministero dell'Interno, Piazza del Viminale, 1 – 00184 Rome, Italy.

[3] Department of Water Resources and Environmental Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece.

Corresponding author: Federico Lombardo (federico.lombardo@uniroma1.it)

Key Points:

- We propose non-asymptotic closed-form distribution for dependent maxima.
- We introduce a new efficient generator of Markov chains with arbitrary marginals.
- We contribute to develop more reliable data-rich-based analyses of extreme values.

Abstract

The analysis of hydrological hazards usually relies on asymptotic results of extreme value theory (EVT), which commonly deals with block maxima (BM) or peaks over threshold (POT) data series. However, data quality and quantity of BM and POT hydrological records do not usually fulfill the basic requirements of EVT, thus making its application questionable and results prone to high uncertainty and low reliability. An alternative approach to better exploit the available information of continuous time series and non-extreme records is to build the exact distribution of maxima (i.e., non-asymptotic extreme value distributions) from a sequence of low-threshold POT. Practical closed-form results for this approach do exist only for independent high-threshold POT series with Poisson occurrences. This study introduces new closed-form equations of the exact distribution of maxima taken from low-threshold POT with magnitudes characterized by an arbitrary marginal distribution and first-order Markovian dependence, and negative binomial occurrences. The proposed model encompasses and generalizes the independent-Poisson model and allows for analyses relying on significantly larger samples of low-threshold POT values exhibiting dependence, temporal clustering and overdispersion. To check the analytical results, we also introduce a new generator (called Gen2Mp) of proper first-order Markov chains with arbitrary marginal distributions. An illustrative application to long-term rainfall and streamflow data series shows that our model for the distribution of extreme maxima under dependence takes a step forward in developing more reliable data-rich-based analyses of extreme values.

1 Introduction

The study of hydrological extremes is one of long history in research applied to design and management of water supply (e.g. Hazen, 1914) and flood protection works (e.g. Fuller, 1914). Almost half a century after the first pioneering empirical studies, Gumbel (1958) provided a general framework linking the theoretical properties of probabilities of extreme values (e.g. Fisher and Tippet, 1928) to the empirical basis of hydrological frequency curves. Since then, extreme value theory (EVT) applied to hydrological analyses has been a matter of primary concern in the literature (see e.g. Papalexiou and Koutsoyiannis, 2013; Serinaldi and Kilsby, 2014 for detailed overview). EVT aims at modeling the extremal behavior of observed phenomena by asymptotic probability distributions, and observations to which such distributions are allegedly related should meet the following important conditions:

1. They should resemble the samples of independent and identically distributed (i.i.d.) random variables. Then, extreme events arise from a stationary distribution and are independent of one another.
2. Their number should be large. Defining how large their size should be depends on the characteristics of the parent distribution from which the extreme values are taken (e.g. the tail behavior) and the degree of precision we seek.

Most of these assumptions, commonly made in classical statistical analyses, are hardly ever realized in hydrological applications, especially when studying extremes. Specifically, the traditional analysis of hydrological extremes is based on statistical samples that are formed by selecting from the entire data series (e.g. at the daily scale) those values that can reasonably be considered as realizations of independent extremes, e.g. annual maxima or peaks over a certain high threshold. Thus, many observations are discarded and the reduction of the already small size of common hydrological records significantly affects the reliability of the estimates (Koutsoyiannis, 2004a,b; Volpi et al., 2019). In addition, Koutsoyiannis (2004a) showed that the convergence to the asymptotic distributions can be extremely slow and may require a huge number of events. Thus, a typical number of extreme hydrological events does not guarantee convergence in applications.

Furthermore, the long-term behavior of the hydrological cycle and its driving forces provide the context to understand that correlations between hydrological samples not only occur, but they also can persist for a long time (see O'Connell et al., 2016 for a recent review). While Leadbetter (1974, 1983) demonstrated that distributions based on dependent events (with limited long-term persistence at extreme levels) share the same asymptotic properties of distributions based on independent trials, there is evidence that correlation has strong influence on the exact statistical properties of extreme values and it slows down the already slow rate of convergence (e.g. Eichner et al., 2011; Bogachev and Bunde, 2012; Volpi et al., 2015; Serinaldi and Kilsby, 2016). In essence, correlation inflates the variability of the expected values and the width of confidence intervals (CIs) due to information redundancy, and a typical effect is reflected in the tendency of hydrological extremes to cluster in space and time (e.g. Serinaldi and Kilsby, 2018 and references therein). Moreover, focusing on extreme data values, such as annual maxima, hinders reliable retrieval of the dependence structure characterizing the underlying process because of sampling effects of data selection (Serinaldi et al., 2018; Iliopoulou and Koutsoyiannis, 2019). Then, correlation structures and variability of hydrological processes might easily be underestimated, further compromising the attempt to draw conclusions about trends spanning the period of records (see Serinaldi et al., 2018, for detailed discussion). In other words, the lately growing body of publications examining "nonstationarity" in hydrological extremes (see Salas et al., 2018 and references therein) may likely reflect time dependence of such extremes within a stationary setting, as observed patterns are usually compatible with stationary correlated random processes (Koutsoyiannis and Montanari, 2015; Luke et al., 2017; Serinaldi and Kilsby, 2018).

In classical statistical analyses of hydrological extremes, to form data samples we commonly use two alternative strategies referred to as “block maxima” (BM) and “peaks over threshold” (POT) methods. The former is to choose the highest of all recorded values at each year (for a given time scale, e.g. daily rainfall) and form a sample with size equal to the number of years of the record. The POT method is to form a sample with all recorded values exceeding a certain threshold irrespective of the year they occurred, allowing to increase the available information by using more than one extreme value per year (Coles, 2001; Claps and Laio, 2003).

The fact that observed hydrological extremes tend to cluster in time increases the arguments towards the use of the POT sampling method, instead of block maxima approaches which tend to hide dependence (Iliopoulou and Koutsoyiannis, 2019). Such clustering reflects dependence (at least) in the neighboring excesses of a threshold, invalidating the basic assumption of independence made in classical POT analyses. Therefore, the standard approach in case studies is to fix a (somewhat subjective) high threshold, and then filter the clusters of exceedances so as to obtain a set of observations that can be considered mutually independent. Such a declustering procedure involves using empirical rules to define clusters (e.g. setting a run length that represents a minimum timespan between consecutive clusters, meaning that a cluster ends when the separation between two consecutive threshold exceedances is greater than the fixed run length) and then selecting only the maximum excess within each cluster (Coles, 2001; Ferro and Segers, 2003; Bernardara et al., 2014; Bommier, 2014). Declustering results in significant loss of data that can potentially provide additional information about extreme values.

In this paper, we aim to overcome these problems by investigating the exact distribution of correlated extremes. Hence, we can set considerably lower thresholds with respect to the standard POT analyses and avoid declustering procedures whose effectiveness is called into question if we do not account for the process characteristics. The proposed approach provides new insight into probabilistic methods devised for extreme value analysis taking into account the clustering dynamics of extremes, and it is consistent with the general principle of allowing maximal use of information (Volpi et al., 2019).

In summary, hydrological applications have made wide recourse to asymptotes or limiting extreme value distributions, while exact distributions for real-world finite-size samples are barely used in stochastic hydrology because their evaluation requires the parent distribution to be known. However, the small size of common hydrological records (e.g. a few tens of years) and the impact of correlations on the information content of observed extremes cannot provide sufficient empirical evidence to estimate limiting extreme value distributions with precision. Therefore, we believe that non-asymptotic analytical models for extremes arising from correlated processes should receive renewed research interest (Iliopoulou and Koutsoyiannis, 2019).

This paper is concerned with a theoretical approach to the exact distribution of high extremes based on the pioneering work by Todorovic and Zelenhasic (1970), who proposed a general stationary stochastic model to describe and predict behavior of the maximum term among a random number of random variables in an interval of time $[0, t]$ assuming independence. As verified in several studies mentioned above, to make a realistic stochastic model of hydrological processes, we are forced to confront the fact that dependence should necessarily be taken into consideration. The dilemma is that dependence structures make for realistic models, but also reduce the possibility for explicit probability calculations (i.e., analytical derivations of joint probability distributions are more complicated than under independence). The challenge of this paper is to propose a stochastic model of extremes with dependencies allowing for acceptable realism, but also permitting sufficient mathematical tractability. In this context, short-range dependence structures, such as Pólya's and Markov's schemes, nicely make a trade-off between these two demands, when hydrological maxima satisfy Leadbetter's condition of the absence of long-range dependence (Koutsoyiannis, 2004a).

In the remainder of this paper, we first introduce a novel theoretical framework to model the exact distribution of correlated extremes in Section 2. In Section 3, we present a new generator, called Gen2Mp, of correlated processes with arbitrary marginal distributions and Markovian dependence, and use it to validate the theoretical reasoning described in Section 2. Then, Section 4 deals with case studies in order to test the capability of our model to reproduce the statistical behavior of extremes of long-term rainfall and streamflow time series from the real world. Concluding remarks are reported in Section 5.

2 Theoretical framework

We use herein the POT approach to analyze the extreme maxima, and assume the number of peaks (e.g., flood peak discharges or maximum rainfall depths) exceeding a certain threshold ξ and their magnitudes to be random variables. The threshold simplifies the study and helps focus the attention on the distribution tails, as they are important to know in engineering design (Papalexiou et al., 2013). In the following, we use upper case letters for random variables or distribution functions, and lower case letters for values, parameters or constants.

If we consider only those peaks Y_i in $[0, t]$ exceeding ξ, then we can define the strictly positive random variable

$$Z_i = Y_i - \xi > 0 \tag{1}$$

for all $i = 1, 2, \ldots, n$, where n is the number of exceedances in $[0, t]$. Clearly, n is a non-increasing function of ξ for a given t, but we assume herein that ξ is a fixed constant.

It is recalled from probability theory that, given a fixed number n of i.i.d. random variables $\{Z_i\}$, the largest order statistic $X = \max\{Z_1, Z_2, \ldots, Z_n\}$ has a probability distribution $H_n(x)$ fully dependent on the joint distribution function of $\{Z_i\}$ that is

$$H_n(x) = \Pr\{Z_1 \leq x, Z_2 \leq x, \ldots, Z_n \leq x\} = \left(F(x)\right)^n \tag{2}$$

In hydrological applications, it may be assumed that the number n of values of $\{Z_i\}$ in $[0, t]$ (e.g. the number of storms or floods per year), whose maximum is the variable of interest X (e.g. the maximum rainfall depth or flood discharge), is not constant but it is a realization of a random variable N $(= 0, 1, 2, \ldots)$. Therefore, we are interested in the maximum term X among a random number N of a sequence of random variables $\{Z_i\}$ in an interval of time $[0, t]$.

In the following, we attempt to determine the one-dimensional distribution function of X that is defined as $H(x) = \Pr\{X \leq x\}$. Since the magnitude of exceedances Z_i and their number N are supposed to be random variables, Todorovic (1970) derived the distribution of the extreme maximum of such a particular class of stochastic processes as

$$H(x) = \Pr\{N = 0\} + \sum_{k=1}^{\infty} \Pr\left\{\bigcap_{i=1}^{k} \{Z_i \leq x\} \cap \{N = k\}\right\} \tag{3}$$

which represents the probability that all exceedances $Z_i > 0$ in $[0, t]$ are less than or equal to x. If $x = 0$, then $H(0) = \Pr\{N = 0\}$ is the probability that there are no exceedances in $[0, t]$.

Todorovic and Zelenhasic (1970) proposed the simplest form of the general model in eq. (3) for use in hydrological statistics, which is now the benchmark against which we measure frequency analysis of extreme events (e.g. Koutsoyiannis and Papalexiou, 2017). Its basic assumptions are that $\{Z_i\}$ is a sequence of N independent random variables with common parent distribution $F(x) = \Pr\{Z_i \leq x\}$, and N is a Poisson-distributed random variable independent of $\{Z_i\}$ with mean λ, i.e. $\Pr\{N = k\} = (\lambda^k/k!)\exp(-\lambda)$. Then, recalling that $\sum_{k=0}^{\infty} y^k/k! = \exp(y)$, eq. (3) becomes

$$H(x) = \sum_{k=0}^{\infty} \left(F(x)\right)^k \frac{\lambda^k}{k!} \exp(-\lambda) = \exp\left(-\lambda\left(1 - F(x)\right)\right) \tag{4}$$

It can be shown that $H(x) \approx H_n(x)$ with satisfactory approximation (Koutsoyiannis, 2004a).

As stated above, the derivation of eq. (4) includes strong assumptions, such as independence, and the purpose of this paper is to modify and test this equation under suitable dependence conditions.

Firstly, we suppose that $\{Z_i\}$ is a sequence of N random variables with common parent distribution $F(x) = \Pr\{Z_i \leq x\}$ and a particular Markovian dependence that give rise to the two-state Markov-dependent process (2Mp, see next Section for further details). Specifically, we let the occurrences of the event $\{Z_i \leq x\}$ evolve according to a Markov chain with two states, whose probabilities are:

$$\begin{cases} p_0 = \Pr\{Z_i \leq x\} \\ p_1 = \Pr\{Z_i > x\} = 1 - p_0 \end{cases} \tag{5}$$

and the transition probabilities (see also Lombardo et al., 2017, appendix C) are:

$$\begin{cases} \pi_{00} = \Pr\{Z_i \le x | Z_{i-1} \le x\} = p_0 + \rho_1(1 - p_0) \\ \pi_{01} = \Pr\{Z_i \le x | Z_{i-1} > x\} = p_0(1 - \rho_1) \\ \pi_{10} = \Pr\{Z_i > x | Z_{i-1} \le x\} = 1 - \pi_{00} \\ \pi_{11} = \Pr\{Z_i > x | Z_{i-1} > x\} = 1 - \pi_{01} \end{cases} \tag{6}$$

where ρ_1 is the lag-one autocorrelation coefficient of the Markov chain.

It follows that, for the process $\{Z_i\}$, the probability of the state $\{Z_n \le x\}$ at a given time n depends solely on the state $\{Z_{n-1} \le x\}$ at the previous time step $n - 1$. Then, for a fixed number of exceedances $N = n$, the Markov property yields:

$$\Pr\{Z_n \le x | Z_{n-1} \le x, \dots, Z_1 \le x\} = \Pr\{Z_n \le x | Z_{n-1} \le x\} \tag{7}$$

Applying the chain rule of probability theory to the distribution function of the maximum term X, $H_n(x) = \Pr\{Z_1 \le x, Z_2 \le x, \dots, Z_n \le x\}$, we obtain

$$H_n(x) = \Pr\{Z_n \le x | Z_{n-1} \le x\} \cdots \Pr\{Z_2 \le x | Z_1 \le x\} \Pr\{Z_1 \le x\} \tag{8}$$

From the above it follows that $H_n(x)$ can be determined in terms of the conditional probabilities $\Pr\{Z_i \le x | Z_{i-1} \le x\}$ and the parent univariate distribution function $F(x) = \Pr\{Z_i \le x\}$. As the random variables $\{Z_i\}$ are identically distributed, they correspond to a stationary stochastic process, and then the function $\Pr\{Z_i \le x | Z_{i-1} \le x\}$ is invariant to a shift of the origin. In this case, $H_n(x)$ is determined in terms of the second-order (bivariate) distribution $H_2(x) = \Pr\{Z_1 \le x, Z_2 \le x\} = \Pr\{Z_2 \le x | Z_1 \le x\} F(x)$ and the first-order (univariate) parent distribution $F(x)$. Indeed, from eq. (8) we obtain

$$H_n(x) = F(x) \left(\frac{H_2(x)}{F(x)} \right)^{n-1} = \frac{\left(F(x)\right)^2}{H_2(x)} \left(\frac{H_2(x)}{F(x)} \right)^{n} \tag{9}$$

It can be easily shown that eq. (9) reduces to eq. (2) in case of independence, i.e. $H_2(x) = \left(F(x)\right)^2$.

Secondly, we assume that exceedances $\{Z_i\}$ have positively correlated occurrences causing a larger variance than if they were independent, i.e. the occurrences are overdispersed with respect to a Poisson distribution, for which the mean is equal to the variance. Therefore, we assume that the random number of occurrences N in a specific interval of time $[0, t]$ follows the negative binomial distribution (e.g. Calenda et al., 1977; Eastoe and Tawn, 2010), which allows adjusting the variance independently of the mean. The negative binomial distribution (known as the limiting form of the Pólya distribution, cf. Feller, 1968, p. 143) is a compound probability distribution that results from assuming that the random variable N is distributed according to a Poisson distribution whose mean λ_j varies randomly following a gamma distribution with shape parameter $r > 0$ and scale parameter $\alpha > 0$, so that its density is

$$g(\lambda_j) = \frac{\lambda_j^{r-1}}{\Gamma(r)\alpha^r} \exp\left(-\frac{\lambda_j}{\alpha}\right) \tag{10}$$

Then, the probability distribution function of N conditional on $\Lambda = \lambda_j$ is

$$\Pr\{N = k | \Lambda = \lambda_j\} = \frac{\lambda_j^k}{k!} \exp(-\lambda_j) \tag{11}$$

We can derive the unconditional distribution of N by marginalizing over the distribution of Λ, i.e., by integrating out the unknown parameter λ_j as

$$\Pr\{N = k\} = \int_0^\infty \Pr\{N = k | \Lambda = \lambda_j\} \, g(\lambda_j) \, \mathrm{d}\lambda_j \tag{12}$$

Substituting eqs. (10) and (11) into eq. (12), we have

$$\Pr\{N = k\} = \frac{1}{k! \, \Gamma(r)\alpha^r} \int_0^\infty \lambda_j^{r+k-1} \exp\left(-\lambda_j \left(\frac{\alpha + 1}{\alpha}\right)\right) \mathrm{d}\lambda_j \tag{13}$$

Recalling that the gamma function is defined as $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x)\, \mathrm{d}x$, then multiplying and dividing eq. (13) by $(\alpha/(\alpha+1))^{r+k}$ and integrating by substitution, we obtain after algebraic manipulations

$$\Pr\{N = k\} = \left(\frac{\alpha}{\alpha+1}\right)^k \frac{\Gamma(r+k)}{k!\,\Gamma(r)} \left(\frac{1}{\alpha+1}\right)^r \tag{14}$$

To summarize, we specialize the general model in eq. (3) for the following conditions:

1. $\{Z_i\}$ is a sequence of N correlated random variables with 2Mp dependence and common parent distribution $F(x) = \Pr\{Z_i \leq x\}$.
2. N is a negative binomial random variable independent of $\{Z_i\}$ with mean $\mu = r\alpha$ and variance $\sigma^2 = r\alpha(\alpha+1)$

Under the above assumptions, from eq. (3) we can derive the conditional distribution function of the maximum X as

$$H(x|\lambda_j) = \Pr\{N = 0|\Lambda = \lambda_j\} + \sum_{k=1}^{\infty} \Pr\left\{\bigcap_{i=1}^{k} \{Z_i \leq x\}\right\} \Pr\{N = k|\Lambda = \lambda_j\} \tag{15}$$

where for $\{Z_i\}$ of 2Mp

$$\Pr\left\{\bigcap_{i=1}^{k} \{Z_i \leq x\}\right\} = \frac{\left(F(x)\right)^2}{H_2(x)} \left(\frac{H_2(x)}{F(x)}\right)^k \tag{16}$$

Substituting eqs. (11) and (16) in eq. (15), we obtain

$$H(x|\lambda_j) = \exp(-\lambda_j) + \frac{\left(F(x)\right)^2}{H_2(x)} \sum_{k=1}^{\infty} \left(\frac{H_2(x)}{F(x)}\right)^k \frac{\lambda_j^k}{k!} \exp(-\lambda_j) \tag{17}$$

Then, adding and subtracting the term $\left(\left(F(x)\right)^2/H_2(x)\right)\exp(-\lambda_j)$ yields

$$H(x|\lambda_j) = \exp(-\lambda_j) - \frac{\left(F(x)\right)^2}{H_2(x)} \exp(-\lambda_j) + \frac{\left(F(x)\right)^2}{H_2(x)} \sum_{k=0}^{\infty} \left(\frac{H_2(x)}{F(x)}\right)^k \frac{\lambda_j^k}{k!} \exp(-\lambda_j) \tag{18}$$

and thus

$$H(x|\lambda_j) = \exp(-\lambda_j) - \frac{\left(F(x)\right)^2}{H_2(x)} \exp(-\lambda_j) + \frac{\left(F(x)\right)^2}{H_2(x)} \exp\left(-\lambda_j \left(1 - \frac{H_2(x)}{F(x)}\right)\right) \tag{19}$$

which is the conditional distribution function of the maximum term X among a Poisson-distributed random number N with gamma-distributed mean $\Lambda = \lambda_j$ of 2Mp random variables $\{Z_i\}$ in an interval of time $[0, t]$. It can be shown that eq. (4) is easily recovered assuming independence, i.e. $H_2(x) = \Pr\{Z_1 \le x, Z_2 \le x\} = \left(F(x)\right)^2$ and $\Lambda = \lambda$ is a fixed constant.

The unconditional distribution of X is derived by substituting eqs. (14) and (16) into eq. (3) as follows

$$H(x) = \left(\frac{1}{\alpha+1}\right)^r + \frac{\left(F(x)\right)^2}{H_2(x)} \left(\frac{1}{\alpha+1}\right)^r \sum_{k=1}^{\infty} \left(\frac{H_2(x)}{F(x)}\right)^k \left(\frac{\alpha}{\alpha+1}\right)^k \frac{\Gamma(r+k)}{k!\,\Gamma(r)} \tag{20}$$

Then, adding and subtracting the term $\left(\left(F(x)\right)^2 / H_2(x)\right) / (\alpha+1)^r$ and denoting by $(r)_k = \Gamma(r+k)/\Gamma(r)$ the Pochhammer's symbol (Abramowitz and Stegun, 1972, p. 256) yields

$$H(x) = \left(\frac{1}{\alpha+1}\right)^r \left(1 - \frac{\left(F(x)\right)^2}{H_2(x)} + \frac{\left(F(x)\right)^2}{H_2(x)} \sum_{k=0}^{\infty} \frac{(r)_k}{k!} \left(\frac{\alpha H_2(x)}{(\alpha+1)F(x)}\right)^k\right) \tag{21}$$

Since $\alpha H_2(x) / \left((\alpha+1)F(x)\right) \in [0, 1)$ and $r > 0$ is a real number, then this series is known as a binomial series (Graham et al., 1994, p. 162), and, setting $y = \alpha H_2(x) / \left((\alpha+1)F(x)\right)$, it converges to $(1-y)^{-r} = \sum_{k=0}^{\infty} \frac{(r)_k}{k!} (y)^k$, thus

$$H(x) = (\alpha+1)^{-r}\left(1 - \frac{\left(F(x)\right)^2}{H_2(x)} + \frac{\left(F(x)\right)^2}{H_2(x)}\left(1 - \frac{\alpha H_2(x)}{(\alpha+1)F(x)}\right)^{-r}\right) \tag{22}$$

which is the unconditional distribution of the extreme maximum X. The parameters of the model in eq. (22) are α and r along with those of the models chosen for both the parent distribution, $F(x)$, and the bivariate distribution $H_2(x)$ (see Sect. 4 for further details).

In the case of independence, where $H_2(x) = \left(F(x)\right)^2$, eq. (22) reduces to

$$H(x) = \left(1 + \alpha\left(1 - F(x)\right)\right)^{-r} \tag{23}$$

As shown in later examples and case studies, eq. (22) yields probabilities of non-exceedance that are systematically larger than those under independence, i.e. $H_{\text{dep}}(x) > H_{\text{indep}}(x)$.

3 Gen2Mp: An Algorithm to Simulate the Two-State Markov-Dependent Process (2Mp) with Arbitrary Marginal Distribution

To check the performance of our stochastic model for correlated extremes, we need to simulate a random process $\{Z_i\}$ with any marginal distribution and Markovian dependence. Nevertheless, we must better clarify what the "Markovian dependence" refers to here. As stated in the previous Section, we assume that a Markov chain with two states (which may represent for example flood or no flood, dry or wet year, etc.) governs the excursions above/below any level (threshold) x of the process $\{Z_i\}$ (see e.g. Fernández and Salas, 1999). We refer to this process as 2Mp (Volpi et al. 2015). For such a process, the Markov property is valid because the probability of the state $\{Z_n \leq x\}$ at a given time n depends solely on the state $\{Z_{n-1} \leq x\}$ at the previous time step $n-1$, i.e., $\Pr\{Z_n \leq x | Z_{n-1} \leq x, \dots, Z_1 \leq x\} = \Pr\{Z_n \leq x | Z_{n-1} \leq x\}$.

One can be tempted to use the classical AR(1) (first-order autoregressive) model to simulate the 2Mp. However, this is not appropriate in general, as we show in the following by a numerical experiment that provides insights into an effective simulation strategy. Let us define the random variable S_j in such a way that for $j = 1, 2, \ldots$, it is

$$\Pr\{S_j = j\} = \Pr\{Z_j \leq x, Z_{j-1} \leq x, \ldots, Z_1 \leq x\} \tag{24}$$

Then, by definition of conditional probability, we may write e.g. for $j = 3$

$$\Pr\{Z_3 \leq x | Z_2 \leq x, Z_1 \leq x\} = \frac{\Pr\{Z_3 \leq x, Z_2 \leq x, Z_1 \leq x\}}{\Pr\{ Z_2 \leq x, Z_1 \leq x\}} = \frac{\Pr\{S_3 = 3\}}{\Pr\{S_2 = 2\}} \tag{25}$$

In our case the Markov property yields

$$\Pr\{Z_3 \leq x | Z_2 \leq x, Z_1 \leq x\} = \Pr\{Z_3 \leq x | Z_2 \leq x\} = \frac{\Pr\{S_2 = 2\}}{\Pr\{S_1 = 1\}} \tag{26}$$

where $\Pr\{S_2 = 2\} = \Pr\{Z_2 \leq x, Z_1 \leq x\} = \Pr\{Z_3 \leq x, Z_2 \leq x\}$ because $\{Z_i\}$ is stationary. From eqs. (25) and (26), it is easily understood that we seek a modelling framework for which the ratio $\mathrm{rt}_j(x) = \Pr\{S_{j+1} = j + 1\}/\Pr\{S_j = j\}$ should be constant for every j, depending solely on the value of the threshold x. In order to show that this is generally not valid for AR(1) processes, we compute such a ratio from a sequence of 100000 random numbers generated by a standard Gaussian AR(1) model with lag-one correlation equal to 0.85. In particular, we calculate four ratios ($j = 1, \ldots, 4$) for various threshold values x_k ($k = 1, \ldots, 100$) selected randomly over the entire range of the standard Gaussian distribution. Then, as the ratio values depend on the threshold, for each x_k we "standardize" the results by taking the absolute difference between each ratio $\mathrm{rt}_j(x_k)$ and its mean $\mu_{\mathrm{rt}}(x_k)$ computed over $j = 1, \ldots, 4$, i.e.

$\mu_{\mathrm{rt}}(x_k) = (1/4)\sum_{j=1}^{4} \mathrm{rt}_j(x_k)$, then dividing all by $\mu_{\mathrm{rt}}(x_k)$; hence, we obtain the relative difference $e_j(x_k) = \left|\left(\mathrm{rt}_j(x_k) - \mu_{\mathrm{rt}}(x_k)\right)/\mu_{\mathrm{rt}}(x_k)\right|$.

We seek a model with a particular Markovian dependence so that $e_j(x) = 0$ for all j and x. In Fig. 1, we show the boxplots depicting the variability of (percent) $e_j(x_k)$ over all threshold values x_k with $j = 1, \dots, 4$. In the left panel, we display the results for the AR(1) model described above. In contrast it can be noted that $e_j(x_k)$ values are not only significantly different from zero (especially if compared with results shown in the right panel of Fig. 1, based on simulation algorithm described below), but their variability also changes strongly with the index j. Then, we conclude that AR(1) models are not appropriate for our purposes. As shown later, despite sharing similar dependence structures (see Fig. 2), Gen2Mp outperforms AR(1) in terms of $e_j(x) = 0$.

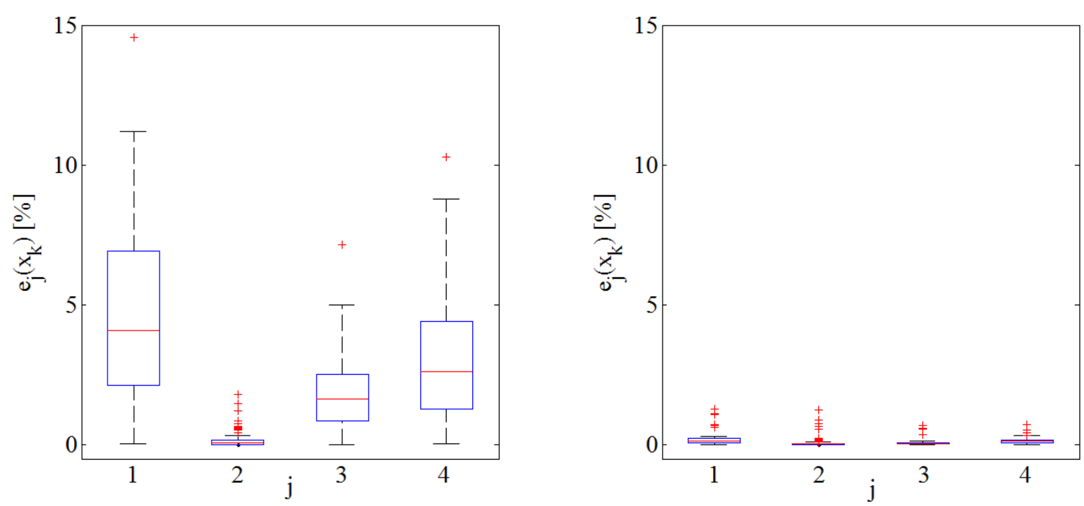

Figure 1. Box plots of four ($j = 1, \dots, 4$) relative differences $e_j(x_k) = \left|\left(\mathrm{rt}_j(x_k) - \mu_{\mathrm{rt}}(x_k)\right)/\mu_{\mathrm{rt}}(x_k)\right|$ for various threshold values x_k ($k = 1, \dots, 100$) selected at random from the parent (standard Gaussian) distribution, where $\mathrm{rt}_j(x) = \Pr\{S_{j+1} = j + 1\}/\Pr\{S_j = j\}$ and $\mu_{\mathrm{rt}}(x_k) = (1/4)\sum_{j=1}^{4} \mathrm{rt}_j(x_k)$. The red line inside each box is the

median and the box edges are the 25th and 75th percentiles of the samples. The left panel depicts results for AR(1) model, while right panel shows boxplots of synthetic data from Gen2Mp algorithm.

3.1 Description of the Gen2Mp simulation algorithm

We introduce herein a new generator, which enables the Monte Carlo materialization of a 2Mp with any arbitrary marginal distribution. It is worth stressing that the theoretical considerations discussed above result in a conceptually simple simulation algorithm, whose scheme consists of an iteration procedure with the following steps:

a) We start by generating two sequences $\{a_i\}_{i=1}^{n}$ and $\{b_i\}_{i=1}^{n}$ of n independent random numbers with the same arbitrary distribution but conditional on being higher ($\{a_i\}_{i=1}^{n}$) or lower ($\{b_i\}_{i=1}^{n}$) than the median.

b) Then, we generate the series $\{c_i\}_{i=1}^{n}$ sampled from i.i.d. Bernoulli random variables taking values 1 and 0 with probability p and $(1-p)$, respectively.

c) The events $\{c_i = 1\}$ in the Bernoulli series determine the alternation between the two states of our target process, i.e. higher (state 1) and lower (state 2) than the median. In other words, the series $\{c_i\}_{i=1}^{n}$ determines the "holding times" before our process switches (jumps) from a state to the other one, because we assume that the state remains the same up to the "time" when there comes a state change $\{c_i = 1\}$. We can now simulate the state-of-generation sequence $\{d_i\}_{i=1}^{n}$ taking values 1 when the state of our process is higher than the median (i.e., $\{a_i\}_{i=1}^{n}$) and 2 otherwise (i.e., $\{b_i\}_{i=1}^{n}$).

d) Consequently, the sequence $\{d_i\}_{i=1}^{n}$ is a sample of a Markov chain $\{D_i\}$ with state space $\{1, 2\}$. Since the holding times of each state are completely random, the state probabilities are $\Pr\{D_i = 1\} = \Pr\{D_i = 2\} = 0.5$. On the other hand, as the jumps arrive randomly according to the Bernoulli process, the transition probabilities are $\Pr\{D_i = 1|D_{i-1} = 2\} = \Pr\{D_i = 2|D_{i-1} = 1\} = p$ and $\Pr\{D_i = 1|D_{i-1} = 1\} = \Pr\{D_i = 2|D_{i-1} = 2\} = 1 - p$. Therefore, the dependence structure of $\{d_i\}_{i=1}^{n}$ is completely specified in terms of the lag-one autocorrelation coefficient $\rho_1 = 1 - 2p$ (see e.g. Lombardo et al., 2017).

e) We can now obtain the target correlated sequence $\{z_i\}_{i=1}^{n}$ as follows:

$$z_i = \begin{cases} a_i & \text{if } d_i = 1 \\ b_i & \text{otherwise} \end{cases} \tag{27}$$

f) As the resulting sequence $\{z_i\}_{i=1}^{n}$ generally does not satisfy the properties of the process we are interested in, we must subdivide each of the cases "> median" and "< median" into two subcases. Specifically, we generate the i.i.d. sequences $\{a_i'\}_{i=1}^{n}$, $\{b_i'\}_{i=1}^{n}$ and $\{a_i''\}_{i=1}^{n}$, $\{b_i''\}_{i=1}^{n}$ conditional on being, respectively, " > 75th percentile", "(median, 75th percentile)", (25th percentile, median) and " < 25th percentile". Then we generate other two Bernoulli series $\{c_i'\}_{i=1}^{n}$ and $\{c_i''\}_{i=1}^{n}$ with same parameter as above, and consequently derive the corresponding state-of-generation sequences $\{d_i'\}_{i=1}^{n}$ (taking values 1 when the state of our process is higher than the 75th percentile, and 2 if it belongs to the interval (median, 75th percentile)) and $\{d_i''\}_{i=1}^{n}$ (taking values 1 when the state belongs to the interval (25th percentile, median), and 2 if it is lower than the 25th percentile). We can now obtain the target correlated sequence $\{z_i\}_{i=1}^{n}$ as follows:

$$z_i = \begin{cases} a_i' & \text{if } d_i = 1 \text{ and } d_i' = 1 \\ b_i' & \text{if } d_i = 1 \text{ and } d_i' = 2 \\ a_i'' & \text{if } d_i = 2 \text{ and } d_i'' = 1 \\ b_i'' & \text{if } d_i = 2 \text{ and } d_i'' = 2 \end{cases} \tag{27'}$$

g) We continue to subdivide until the relative difference $e_j(x_k)$ converges to zero for any j. In any subdivision step, we follow the same procedure as that described above with a fixed parameter p, until a convergence threshold is achieved (here, a mean absolute error equal to 0.002 for $e_j(x_k)$ is used in the numerical examples below, which is obtained after 9 subdivision steps for $p = 0.06$).

3.2 Numerical simulations

We show some Monte Carlo experiments assuming the standard Gaussian probability model as parent distribution, but it can be changed to any distribution function. We generate a correlated series of 100000 standard Gaussian random numbers using Gen2Mp with parameter $p = 0.06$. Such a parameter completely determines the dependence structure of the 2Mp process. For $0 < p < 0.5$ the process is positively correlated, while it reduces to white noise for $p = 0.5$. For $0.5 < p < 1$ we get an anticorrelated series. The particular value of $p = 0.06$ is chosen in order to have the dependence structure of the generated series similar to that of the AR(1) model with lag-one correlation equal to 0.85 (see Fig. 2). Such a value of p has been determined numerically exploiting the fact that the dependence structure of the generated series is closely related (showing slight downward bias) to that of the Markov chain $\{D_i\}$ defined above, whose lag-one autocorrelation is $\rho_1 = 1 - 2p$ (see Fig. 2). Then, to a first approximation, we start assuming $\rho_1 = 0.85$, and progressively increase it until the dependence structures of the 2Mp and AR(1) match.

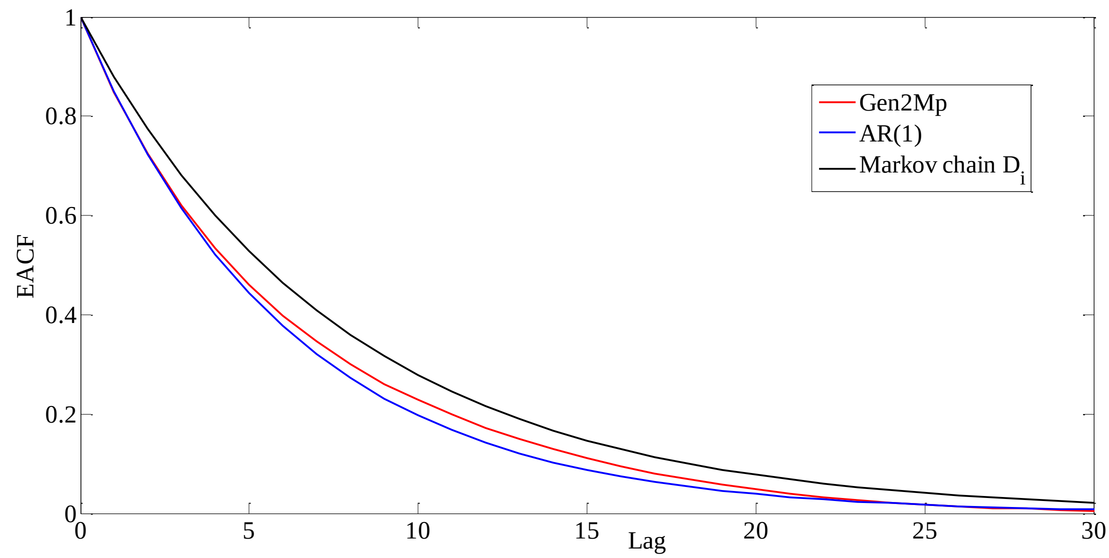

Figure 2. Comparison of the empirical autocorrelation functions (EACFs) resulting from time series generated by Gen2Mp $\{z_i\}_{i=1}^{n}$ and the Markov chain $\{d_i\}_{i=1}^{n}$ with parameter $p = 0.06$, and by AR(1) model with lag-one correlation equal to 0.85.

Then, even though Gen2Mp and the classical AR(1) algorithms generate time series exhibiting analogous dependence structures, the former significantly outperforms the latter in terms of $e_j(x) = 0$, as shown in Fig. 1 (right panel). Furthermore, we generate an independent series of 100000 standard Gaussian random numbers as a benchmark using classical generators (e.g. Press et al., 2007). As it can be noticed from the probability-probability (PP) and quantile-quantile (QQ) plots in Fig. 3, the marginal distribution of the final dependent series (corresponding to a 2Mp) is the same as that of the benchmark series. In summary, the important achievement is that Gen2Mp does not alter the parent distribution, but it only induces time dependence in a Markov chain sense.

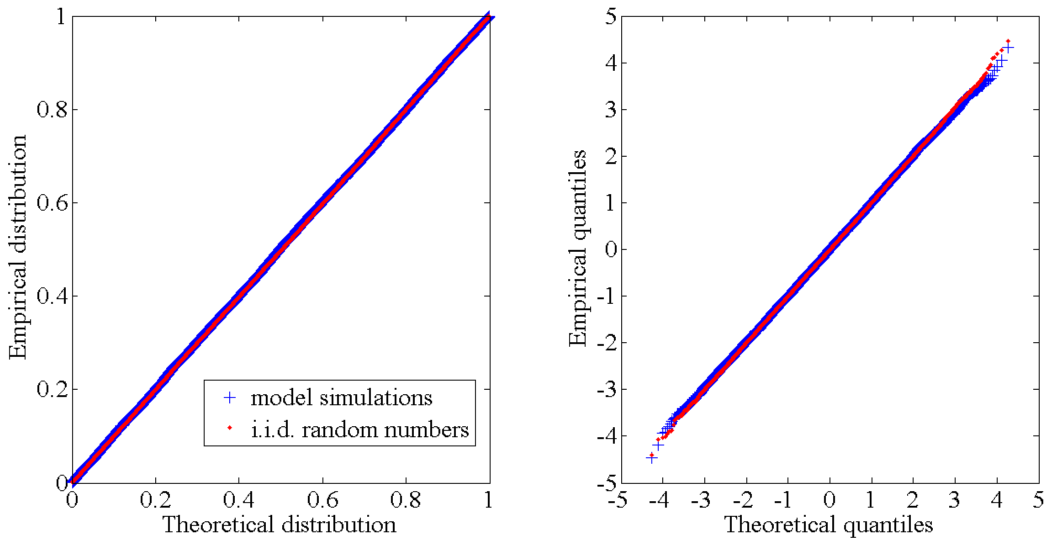

Figure 3. Probability–Probability plot (left) and Quantile–Quantile plot (right) comparing the marginal distribution of a benchmark series (i.i.d. standard Gaussian random numbers) to that of the correlated series generated using Gen2Mp.

Focusing on the frequency analysis of maxima, we investigate the distribution of the maximum term X among a random number N of a sequence of standard Gaussian random variables $\{Z_i\}$. Specifically, we assume that N follows a negative binomial distribution in eq. (14), while the variables $\{Z_i\}$ form a 2Mp stochastic process. Based on such hypotheses, in the previous Section we derived the corresponding theoretical probability distribution function $H(x) = \Pr\{X \leq x\}$ given by eq. (22). To check this numerically, we generate the random numbers $\{n_k\}_{k=1}^{m}$ (where $m = 450$) from the negative binomial distribution with parameters $r = 4$ and $\alpha = 25$, then we form the target sample $\{x_k\}_{k=1}^{m}$ by taking the maximum of m non-overlapping sequences of n_k consecutive random numbers $\{z_i\}_{i=1}^{n_k}$. We allow two different dependence structures for $\{z_i\}_{i=1}^{n_k}$. In the first case we assume that $\{z_i\}_{i=1}^{n_k}$ are sampled from i.i.d. random variables; while in the second case $\{z_i\}_{i=1}^{n_k}$ are sampled from a 2Mp stochastic process with parameter $p = 0.06$, which is simulated by Gen2Mp.

Results in the form of PP plots are depicted in Fig. 4. In the left panel, we show the independent case, and it can be noticed how the empirical distribution of $\{x_k\}_{k=1}^{m}$ is closely matched by eq. (23), i.e. the PP plot (blue line) follows a straight line configuration oriented from $(0,0)$ to $(1,1)$. In other words, when $\{Z_i\}$ are i.i.d. eq. (23) proves to be a good model for the theoretical distribution of X.

In the right panel of Fig. 4, we show the dependent case where the joint probability $H_2(x) = \Pr\{Z_n \leq x, Z_{n-1} \leq x\}$ in eq. (22) is determined numerically. Clearly, if we apply eq. (23) to the correlated sample $\{x_k\}_{k=1}^{m}$, then the corresponding plot (blue line) shows a marked departure from the 45° line (i.e., the line of equality). By contrast, the theoretical distribution that we propose in eq. (22) reasonably models the empirical distribution of correlated maxima $\{x_k\}_{k=1}^{m}$ in all respects (see black line). Therefore, when the $\{Z_i\}$ belong to 2Mp eq. (22) (black line) largely outperforms eq. (23) (blue line) in modelling the extreme maxima

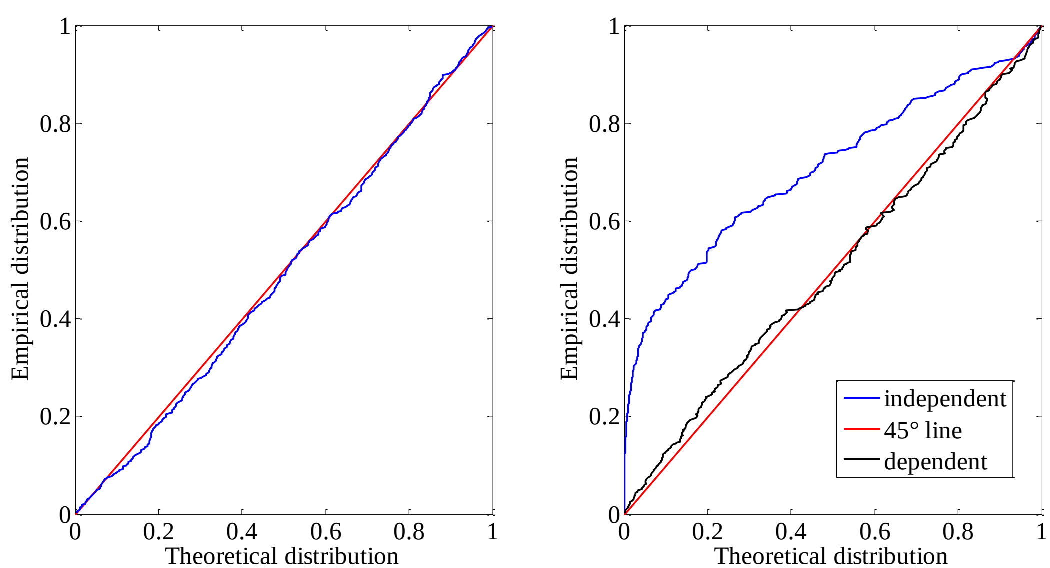

Figure 4. Probability–Probability plots of the maximum term X among a (negative binomial) random number N of a sequence of i.i.d. (left panel) and 2Mp (right panel) standard Gaussian random variables $\{Z_i\}$.

4 Applications to Rainfall and Streamflow Data

In order to provide some insights into the capability of the proposed methodology to reproduce the statistical pattern of observed hydrological extremes, the datasets used in the applications comprise long-term daily rainfall and streamflow time series with no missing values or as few as possible, to fulfil the requirements of POT analyses. In more detail, we use three daily precipitation time series recorded by rain gages located at Groningen (north-eastern Netherlands), Middelburg (south-western Netherlands) and Bologna (northern Italy) respectively ranging from 1847 to 2017 (171 years, no missing values), from 1855 to 2017 (163 years, no missing values) and from 1813 to 2018 (206 years, only three missing values). Raw data, retrieved through the Royal Netherlands Meteorological Institute (KNMI) Climate Explorer web site, are available at https://climexp.knmi.nl/data/bpeca147.dat (accessed on 26 October 2019) for Groningen station, at https://climexp.knmi.nl/data/bpeca2474.dat (accessed on 26 October 2019) for Middelburg station and at https://climexp.knmi.nl/data/pgdcnITE00100550.dat (accessed on 26 October 2019) for Bologna station in the period 1813-2007 (see Klein Tank et al., 2002; Menne et al., 2012). For the most recent period, 2008-2018, daily data for Bologna station are provided by the Dext3r public repository (http://www.smr.arpa.emr.it/dext3r/) (accessed on 26 October 2019) of the Regional Agency for Environmental Protection and Energy (Arpae) of Emilia Romagna, Italy (retrieved and processed by Koutsoyiannis for the book: Stochastics of Hydroclimatic Extremes, in preparation for 2020).

Furthermore, we analyze one daily streamflow time series of the Po River recorded at Pontelagoscuro, northern Italy (see Montanari, 2012 for further details). The data series, spanning from 1920 to 2017 (98 years, no missing values), is made publicly available by Prof. Alberto Montanari at https://distart119.ing.unibo.it/albertonew/sites/default/files/uploadedfiles/po-pontelagoscuro.txt (accessed on 26 October 2019) for the period 1920-2009, while the remainder (2010-2017) has been retrieved through the Dext3r repository.

Since it has been shown that seasonality affects the distribution of hydrological extremes (Allamano et al., 2011), our analyses are performed on a seasonal basis; we distinguish four seasons, each consisting of three months such that the autumn comprises September, October, and November. Winter, spring, and summer are defined similarly. We prefer not to use deseasonalization procedures to avoid possible artifacts that may affect the results. Furthermore, as daily rainfall and streamflow processes exhibit very different marginal distributional properties, all recorded values exceeding a certain threshold are transformed to normality by normal quantile transformation (NQT) for the sake of comparison (Krzysztofowicz, 1997). In practice, observed exceedances $\{z_i\}_{i=1}^{n}$ are transformed to $\psi_i = \Phi^{-1}\big(F_n(z_i)\big)$, where Φ^{-1} is the quantile function of the standard Gaussian distribution and F_n is the Weibull plotting position of the ordered sample. In addition, all datasets used in this study have been preprocessed by removing leap days, because the February 29th was already removed from all leap years of the 1920-2009 Po river discharge dataset.

We now investigate the frequency analysis of observed hydrological maxima. For each season of any dataset, we use for example the value of the threshold corresponding to the 5th percentile (excluding zeros for rainfall datasets for simplicity, but we checked that results do not vary considerably if we include zeros), whose exceedances $\{z_i\}$ are normalized to $\{\psi_i\}$ for each sample. As stated in Sect. 1, we are interested in the statistical behavior of the maximum term X among a random number of equally distributed random variables (i.e., belonging to a certain season) in an interval of time (we assume one year). Then, first we form the POT samples for each year of the record, consisting of m (i.e., number of years) sequences of threshold excesses $\{\psi_i\}_{i=1}^{n_k}$ each of size n_k (for $k = 1, \dots, m$); second we form the sample of annual extremes $\{x_k\}_{k=1}^{m}$ by taking the maximum of each POT series. In other words, $\{x_k\}_{k=1}^{m}$ is a sample of annual maxima of size m (i.e., the number of years of the given dataset) taken from annual POT series of size n_k (i.e., the number of exceedances in the k-th year for the considered season). It follows that the sample size used in classical BM analysis is m, while that used in our approach is $\sum_{k=1}^{m} n_k$. As detailed below, all parameter values (see, e.g., Tables 1 and 2) are estimated from the POT series by maximum likelihood method.

We compare the empirical distribution of X to the theoretical probability distribution function $H(x) = \Pr\{X \leq x\}$ given by eq. (4) (i.e., the classical method) assuming Poisson occurrences of independent exceedances, and by eq. (22) (i.e., the proposed method) assuming negative binomial occurrences of 2Mp exceedances. Parameters of Poisson and negative binomial distributions are derived through a process of maximum likelihood estimation from the annual counts $\{n_k\}_{k=1}^{m}$ for each season of each dataset. To a first approximation, we assume statistical independence of $\{n_k\}_{k=1}^{m}$ by checking that, for each dataset, the empirical autocorrelations between the numbers of exceedances of subsequent years are negligible (not shown). Furthermore, we assume that the joint probability of exceedances $H_2(x) = \Pr\{Z_1 \leq x, Z_2 \leq x\}$ in eq. (22) can be written in terms of the univariate marginal distribution $F(x)$ (which is the standard normal in case of normal quantile transformation) and a bivariate copula that describes the dependence structure between the variables (Salvadori et al., 2007). Several bivariate families of copulas have been presented in the literature, allowing the selection of different dependence frameworks (Favre et al., 2004). For the sake of simplicity, we choose the following three types of copulas that have been in common use:

1. The Gaussian copula (Salvadori et al., 2007 pp. 254-256), which implies the elliptical shape of isolines of the pairwise joint distribution $H_2(x)$ that in our case is given by a bivariate normal distribution $\mathcal{N}_2(\mathbf{0},\mathbf{\Sigma})$ with zero mean and covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where the parameter ρ is the average (over m years) lag-one autocorrelation coefficient of the annual POT series $\{\psi_i\}$.
2. The Clayton copula (Salvadori et al., 2007 pp. 237-240), which exhibits upper tail independence and lower tail dependence (Salvadori et al., 2007 pp. 170-175), and in our case yields

$$H_2(x) = \max\left(\left(2F(x)^{-\beta} - 1\right)^{-\frac{1}{\beta}}, 0\right) \quad (28)$$

where the parameter β can be written in terms of the Kendall's tau correlation coefficient as $\beta = 2\tau/(1-\tau)$, which is the average (over m years) of lag-one Kendall's tau autocorrelation coefficient of the annual POT series $\{\psi_i\}$.

3. The Gumbel-Hougaard copula (Salvadori et al., 2007 pp. 236-237), which exhibits upper tail dependence and lower tail independence, and in our case yields

$$H_2(x) = \exp\left(-\left(2\left(-\ln\left(F(x)\right)\right)^{\beta}\right)^{\frac{1}{\beta}}\right) \quad (29)$$

where the parameter β is again written in terms of the Kendall's tau correlation coefficient as $\beta = 1/(1-\tau)$.

All parameter values for all seasons and datasets are reported in Table 1.

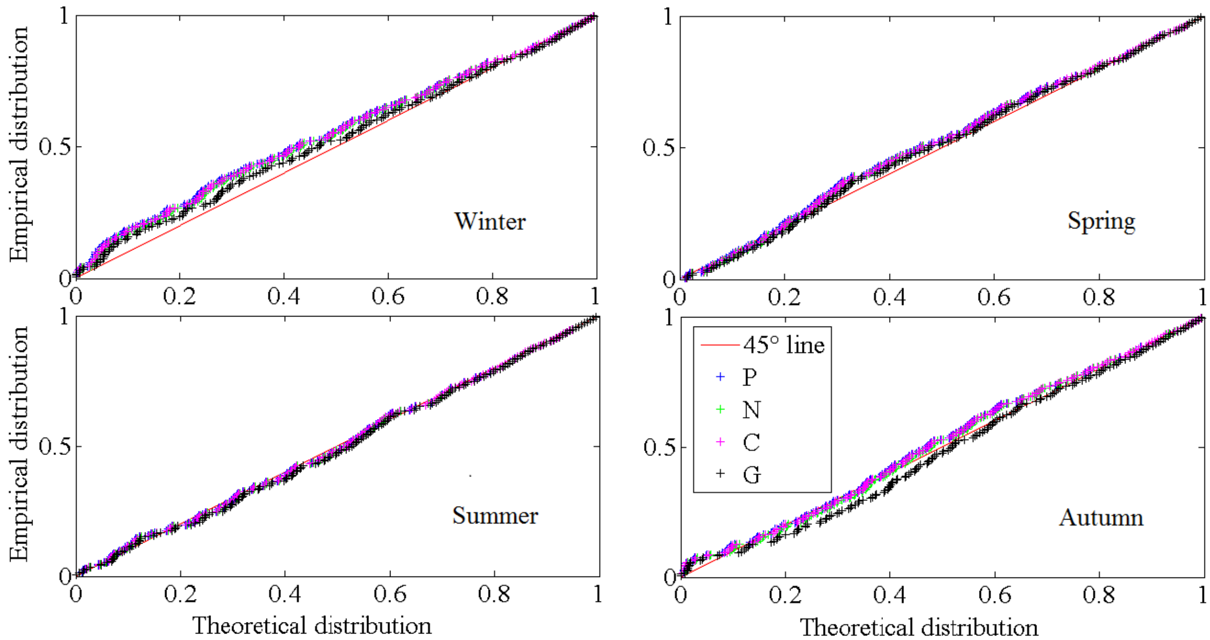

Figure 5. Probability–Probability plots of Groningen dataset of daily rainfall. The empirical distributions of maximum terms $\{x_k\}_{k=1}^{m}$ among annual exceedances of the 5th percentile threshold for winter (top left), spring (top right), summer (bottom left) and autumn (bottom right) seasons are compared to the corresponding theoretical distributions assuming both Poisson (P) occurrences (with parameter λ) of independent exceedances (eq. 4), and negative binomial occurrences (with parameters r and α) of correlated exceedances (eq. 22) with pairwise joint distribution described by the Gaussian (N), Clayton (C, eq. 28) and Gumbel (G, eq. 29) copulas, with parameters ρ and τ as detailed in the text. All parameter values are reported in Table 1.

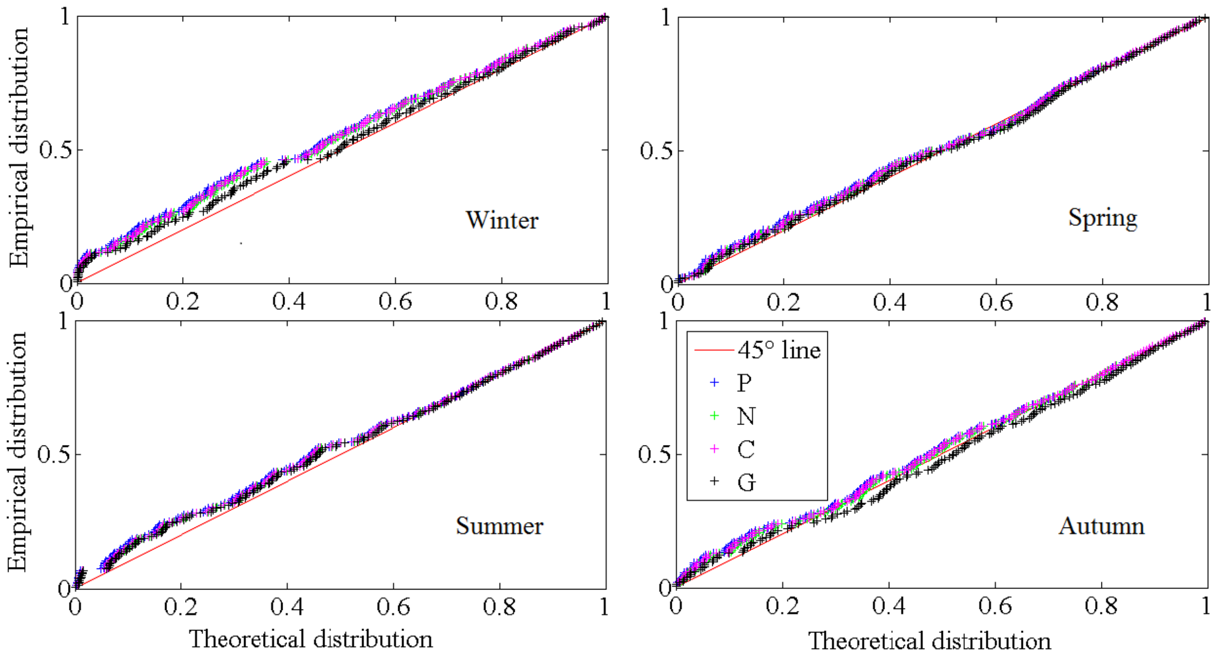

Figure 6. Same as Fig. 5 for Middelburg dataset of daily rainfall.

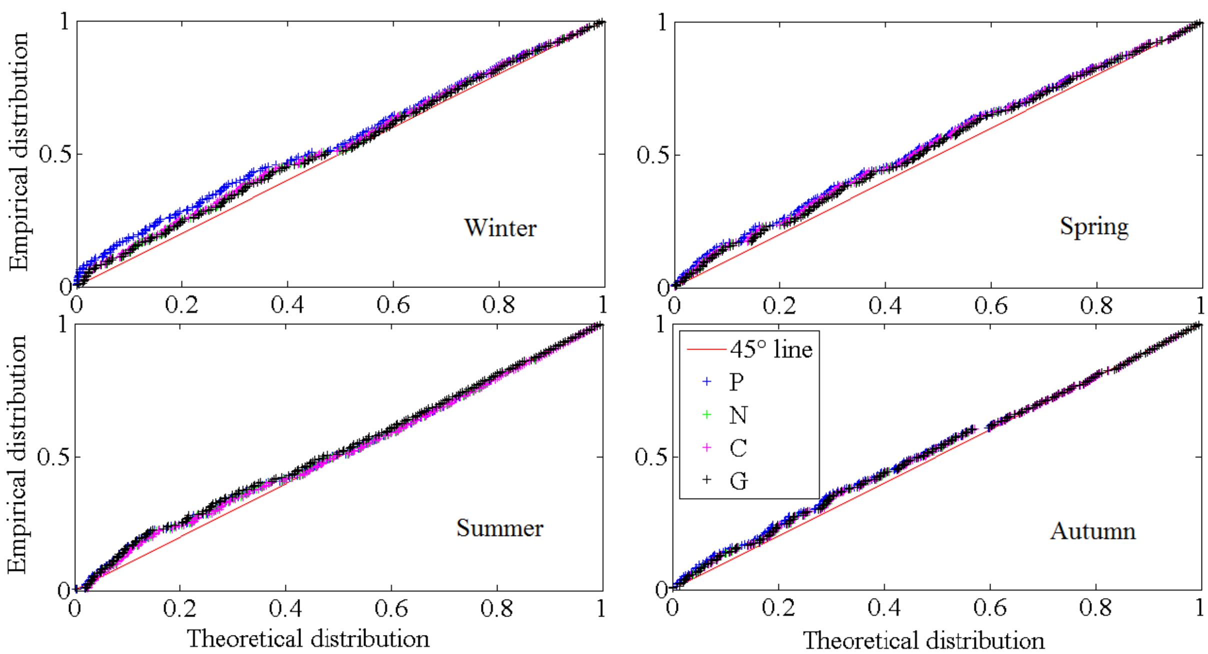

Figure 7. Same as Fig. 5 for Bologna dataset of daily rainfall.

In Figs. 5-7 we may observe that for all daily rainfall datasets the magnitudes of extreme events taken from excesses of a low threshold (the 5th percentile of the nonzero sample) can be considered independent and identically distributed, and this is consistent with the results shown in the literature using different approaches (see e.g. Marani and Ignaccolo, 2015; Zorzetto et al., 2016; De Michele and Avanzi, 2018). In addition, we may notice that the classical model of POT analyses assuming Poisson occurrences (see eq. (4)) seems to be appropriate to study rainfall extremes. Analogous considerations obviously apply to higher thresholds (not shown). Our model of correlated extremes in eq. (22) is capable of capturing such a behavior with precision.

After showing the results with daily rainfall, we also analyze rainfall records at finer time resolution (hourly scale) whose correlation can be stronger than that pertaining to daily data. To this end, we use hourly rainfall data of "Bologna idrografico" station for the period 1990-2013 provided by the Dext3r repository (23 years full coverage, while the entire 2008 is missing). We checked that such hourly rainfall data aggregated at the daily scale are consistent with the daily data recorded in the same period by Bologna station above (not shown).

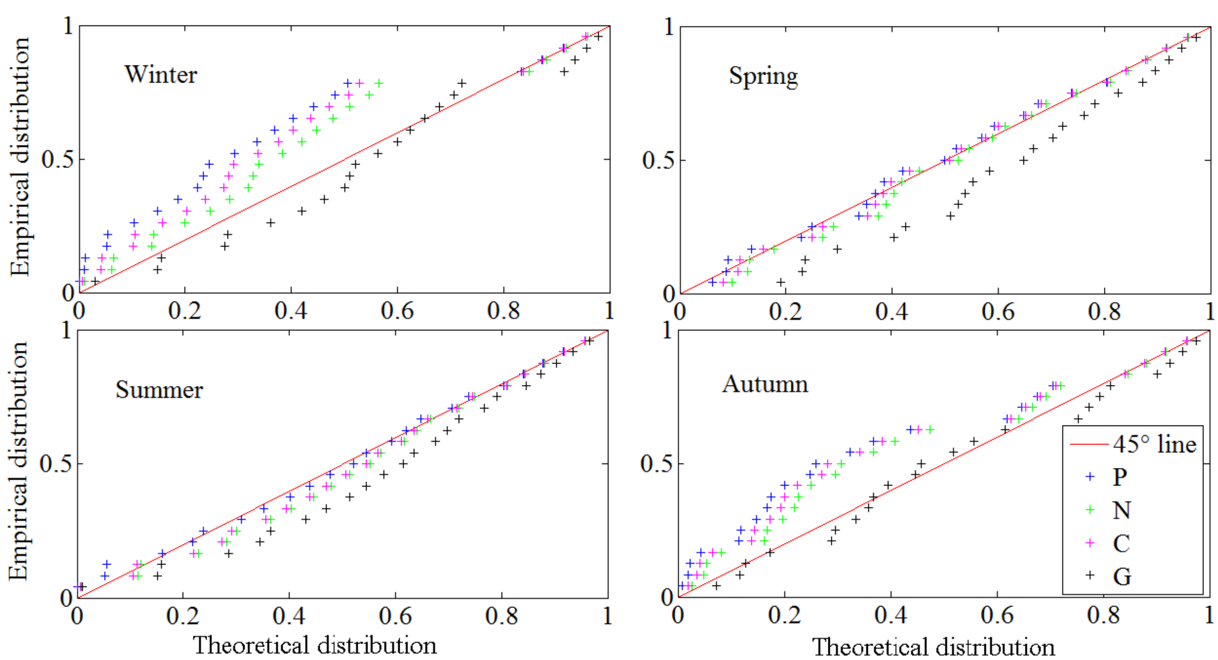

Figure 8. Same as Fig. 5 for Bologna dataset of hourly rainfall.

Comparing Figs. 7 and 8, it is noted that extremes of hourly rainfall data are more affected by correlation than daily data (see e.g. winter and autumn seasons, respectively top left and bottom right panels). This is also the case if we consider the same period of record (1990-2013) for both datasets (not shown). Then, we may conclude that low thresholds can be used for classical POT analyses (assuming independence) of rainfall time series at the daily scale (or above), while further investigations of different datasets are required to describe the impact of dependence on the extremal behavior of the rainfall process at finer time scales. Besides, other interesting future analyses could investigate the extremes of areal rainfall, as for example weather radar data will become more reliable and will accumulate in time providing samples with lengths adequate enough to enable reliable investigation of the probability distribution of areal rainfall (Lombardo et al., 2006a,b; Lombardo et al., 2009).

By contrast, results change significantly when analyzing extremes of streamflow time series. In fact, we present a case study that shows how models assuming independence among magnitudes of extreme events prove to be inadequate to study the probability distribution of discharge maxima.

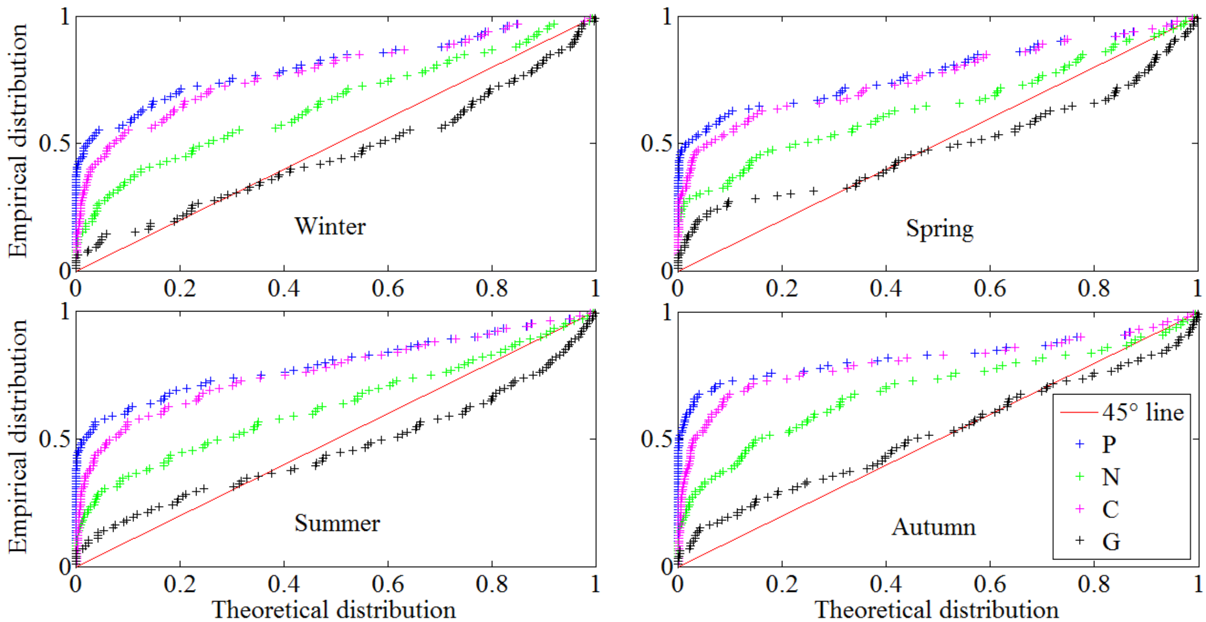

Figure 9. Same as Fig. 5 for the Po River dataset of daily discharge.

In Fig. 9, we show the PP plots of the distribution of extreme maxima taken from annual exceedances of the 5th percentile thresholds for the four seasons of the Po River discharge dataset, recorded at Pontelagoscuro station. Contrary to the rainfall case studies, the classical model assuming independent magnitudes with Poisson (P) occurrences shows marked departures from the 45° line. The theoretical distribution is usually much lower than its empirical counterpart, meaning that, under the popular assumption of independent extremes, the theoretical probability of an extreme event of given magnitude being exceeded is significantly higher than the corresponding observed frequency of exceedance. Fig. 9 shows that our 2Mp model of correlated extremes outperforms the widely used independent model. In particular, the distribution of maxima that has a Gumbel copula seems to be more consistent with observed extreme values, denoting dependence in the upper tail of the bivariate distribution $H_2(x) = \Pr\{Z_1 \leq x, Z_2 \leq x\}$ (Schmidt, 2005). In summary, daily streamflow extremes may exhibit noteworthy departures from independence which are consistent with a stochastic process characterized by a 2Mp behavior and upper tail dependence.

Table 1. Parameters values for all normalized case studies detailed in the text: λ for Poisson (P) occurrences (eq. 4); r and α for negative binomial occurrences (eq. 22); τ for Clayton (C) and Gumbel (G) copulas (eqs. 28-29); ρ for Gaussian copula.

Station	**Parameter / Season**	**Winter**	**Spring**	**Summer**	**Autumn**
Groningen	λ	50.04	41.56	45.04	50.05
	r	76.24	73.15	150.54	164.94
	α	0.66	0.57	0.30	0.30
	τ	0.08	0.04	0.02	0.1
	ρ	0.10	0.05	0.04	0.13
Middelburg	λ	48.41	40.00	38.42	47.16
	r	35.71	40.22	35.47	61.68
	α	1.36	0.99	1.08	0.76
	τ	0.09	0.04	0.02	0.09
	ρ	0.12	0.06	0.02	0.14
Bologna daily	λ	20.92	25.39	16.59	24.67
	r	7.20	22.57	20.98	21.14
	α	2.91	1.13	0.79	1.17
	τ	0.03	0.02	-0.05	-0.01
	ρ	0.05	0.02	-0.06	0.01
Bologna hourly	λ	127.59	128.87	54.09	129.74
	r	5.27	14.14	4.55	12.32
	α	24.22	9.12	11.90	10.53
	τ	0.43	0.30	0.17	0.33
	ρ	0.54	0.38	0.20	0.41
Pontelagoscuro	λ	85.48	87.40	87.39	86.41
	r	67.02	136.59	81.95	245.15
	α	1.28	0.64	1.07	0.35
	τ	0.82	0.81	0.84	0.84
	ρ	0.92	0.92	0.94	0.93

The above results are also evident if we compare theoretical and empirical distributions of streamflow maxima by plotting their quantiles against each other. We use real values for this example (i.e., we do not apply the normal quantile transformation to the data series); therefore, empirical quantiles equal the observed annual maxima. Theoretical quantiles referring to eqs. (4) and (22) (the latter specializes for Gaussian, Clayton and Gumbel copulas) are computed by numerically solving for the root of the equation $H(x) - p = 0$ for a given probability value, p

(i.e., the Weibull plotting position of observed annual maxima), assuming the classical generalized Pareto (GPD) with zero lower bound as parent distribution of threshold excesses:

$$F(x) = \begin{cases} 1 - \left(1 + \gamma \dfrac{x}{\sigma}\right)^{-\frac{1}{\gamma}} & \text{for } \gamma \neq 0 \\ 1 - \exp\left(-\dfrac{x}{\sigma}\right) & \text{otherwise} \end{cases} \tag{30}$$

where γ is the shape parameter and σ is the scale parameter, which we estimate through the maximum likelihood method applied to the entire POT series of each season.

In Fig. 10, QQ plots of Po river discharge for the spring season are shown when varying the threshold ξ (from the 5th, Q_5, to the 75th, Q_{75}, percentiles) to form POT series. It can be noticed that for low thresholds there is a shift in variance between theoretical (i.e., derived from eq. (22) with Gumbel copula) and empirical quantiles, namely the variance of theoretical annual maxima underestimates its empirical counterpart. This can be due to the fitting performance of the marginal generalized Pareto, which does not reproduce well the tail behavior of observed data (not shown). Fig. 10 shows that increasing the threshold value helps focus the attention on the distribution tail to better capture the behavior of maxima. This is also the case if we compare streamflow quantiles resulting from our model with those estimated through "classical" Generalized Extreme Value (GEV) distribution fitted to the observed annual maxima. All parameter values are reported in Table 2. We note that the three GEV parameters are estimated on $m = 98$ data points, while the five parameters of our model in eq. (22) (α, r, τ or ρ, and the two parameters of the GPD with zero lower bound) are estimated on $\sum_{k=1}^{m} n_k$ data, which are 8565, 6759, 4497, and 2253 for Q_5, Q_{25}, Q_{50}, Q_{75}, respectively.

As threshold increases evidence of persistence is progressively reduced as expected, but, we also note in Fig. 10 that the theoretical quantiles derived from the classical independent

Poisson method always show a shift in mean with respect to observed maxima (i.e., under independence, theoretical streamflow quantiles systematically and significantly overestimate observed streamflow maxima).

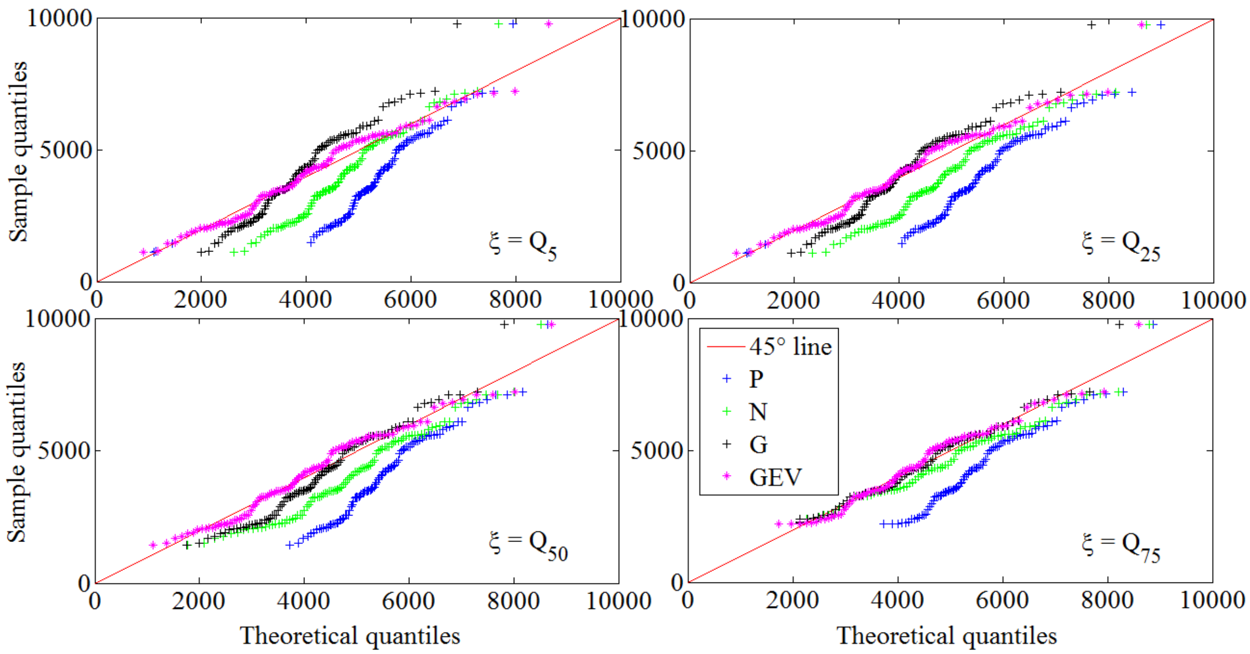

Figure 10. Quantile–Quantile plots of Po river discharge (m^3/s) for spring season. The observed maximum terms among annual peaks over the 5th percentile (top left), 25th percentile (top right), 50th percentile (bottom left) and 75th percentile (bottom right) thresholds are compared to the corresponding theoretical quantiles. In all cases, we assume the Generalized Pareto as parent distribution of daily streamflow (with shape $\gamma \in \mathbb{R}$, scale $\sigma > 0$ and threshold $\xi > 0$ parameters), and compute quantiles specializing eq. (22) for Poisson (P) occurrences (with parameter λ, eq. 4) of independent exceedances, and for negative binomial occurrences (with parameters r and α) of correlated exceedances with pairwise joint distribution described by the Gaussian (N), Clayton (C) and Gumbel (G) copulas, with parameters ρ and τ as detailed in the text. We also plot theoretical quantiles from GEV distribution (with shape $\chi \in \mathbb{R}$, scale $\theta > 0$ and location $\mu > 0$ parameters) fitted to the observed annual maxima. All parameter values are reported in Table 2.

To summarize, our model provides a closed-form expression of the exact distribution for dependent hydrological maxima, which is capable of capturing the behavior of observed extremes of long-term hydrological records. In particular, while rainfall extremes do not seem to be significantly affected by correlation at the daily scale so that the classical Poisson model can be appropriate for use in POT analyses of daily rainfall time series, the influence of correlation is prominent in the streamflow process at the daily scale and it is important to preserve in simulation and analysis of extremes.

Table 2. Parameters values for all models used in the QQ plots of Fig. 10.

Model	**Parameter / Threshold**	Q_5	Q_{25}	Q_{50}	Q_{75}
Generalized Pareto	γ	-0.10	-0.03	-0.05	-0.03
	σ	1220.16	1044.03	1065.80	998.06
	ξ	653.00	998.00	1410.00	2133.00
Poisson	λ	87.40	68.97	45.89	22.99
Negative Binomial	r	136.59	5.89	1.74	0.71
	α	0.64	11.71	26.45	32.22
Clayton & Gumbel copulas	τ	0.82	0.76	0.63	0.48
Gaussian copula	ρ	0.91	0.86	0.75	0.61
GEV	χ	-0.11	-0.11	-0.08	-0.07
	θ	1463.94	1463.94	1399.01	1273.31
	μ	3309.91	3309.91	3369.76	3739.46

5 Conclusions

The study of hydrological extremes faces the chronic lack of sufficient data to perform reliable analyses. This is partly related to the inherent nature of extreme values, which are rare by definition, and partly related to the relative shortness of systematic records from hydro-meteorological gauge networks. The limited availability of data poses serious problems for an effective and reliable use of asymptotic results provided by EVT.

Alternative methods focusing on the exact distribution of extreme maxima extracted from POT sequences of random size over fixed time windows have been proposed in the past. However, closed-form analytical results were developed only for independent data with Poisson occurrences. Even though these assumptions may be sufficiently reliable for high-threshold POT values, this type of data still generates relatively small sample size. In order to better exploit the available information, it can be convenient to consider lower thresholds. However, the effect of lower thresholds is twofold: on the one side the sample size increases, but on the other side the hypotheses of independent magnitudes and Poisson occurrences of POT values are no longer reliable.

In this study, we have introduced closed-form analytical formulae for the exact distribution of maxima from POT sequences that generalize the classical independent model, overcoming its limits and enabling the study of maxima taken from dependent low-threshold POT values with arbitrary marginal distribution, first-order Markov dependence structure, and negative binomial occurrences, and tested real data against this hypothesis. Even though the framework can be further generalized by introducing arbitrary dependence structures and models for POT occurrences, first-order Markov chains and negative binomial distributions provide a good compromise between flexibility and the possibility to obtain simple ready-to-use formulae. In this respect, it should be noted that our model of correlated extremes can cover a sufficient range of cases. We have shown that the modulation of the lag-one autocorrelation coefficient of the annual sequences of POT values (i.e. the Markov chain parameter) gives a set of extremal distributions that include the empirical distribution of maxima for rainfall data series, and for highly correlated low-threshold discharge POT series. On the other hand, the negative binomial model is a widely used and theoretically well-established model for occurrences exhibiting clustering and overdispersion, which are common characteristics of POT events resulting from persistent processes, such as river discharge.

The relationship between our model and its classical independent version (i.e. eqs. (22) and (4)) along with results of the case studies show that distribution of extreme maxima under dependence yields probabilities of exceedance that are systematically lower than those under independence, and are also consistent with traditional approaches (GEV), based on extreme value theory, applied to long annual maxima series.

Finally, we stress that our model of the exact distribution of correlated extremes requires knowledge or fitting of a bivariate distribution (and therefore its univariate marginal distribution). In particular, while the extremal behavior of the rainfall process does not seem to be significantly affected by dependence at the daily scale so that the classical Poisson model can be appropriate for use in POT analyses of daily rainfall time series, the influence of correlation is prominent in the streamflow process at the daily scale and it appears also in the rainfall process at the hourly scale. Then, it is important to account for such dependence in the extreme value analyses, which are crucial to hydrological design and risk management because critical values can be less extreme and more frequent than expected under the classical independent models. Comparing the Gaussian, Clayton and Gumbel bivariate copulas, describing different dependence structures, and the standard Gaussian and Generalized Pareto marginal distributions, we found that the distribution of maxima that has a Gumbel copula seems to be more consistent with streamflow extreme values, denoting dependence in the upper tail of the bivariate distribution. However, these aspects require further investigation form both theoretical and empirical standpoints, and will be the subject of future research. In the spirit of the recent literature on the topic, we believe that the present study will contribute to develop more reliable data-rich-based analyses of extreme values.

Acknowledgments

All data used in this study are freely available online, as described in Section 4 above. The associate editor, an eponymous reviewer, Geoff Pegram, and two anonymous reviewers are gratefully acknowledged for their constructive comments that helped to substantially improve the paper. We also thank Alessio Domeneghetti for providing the first author with detailed information on the Dext3r public repository.

References

Abramowitz, M., and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas,* Graphs, and Mathematical Tables, 9th printing. New York: Dover.

Allamano, P., Laio, F., and Claps, P. (2011). Effects of disregarding seasonality on the distribution of hydrological extremes. *Hydrology and Earth System Sciences*, 15, 3207-3215.

Bernardara, P., Mazas, F., Kergadallan, X., & Hamm, L. (2014). A two-step framework for over-threshold modelling of environmental extremes. *Natural Hazards and Earth System Sciences*, 14(3), 635-647.

Bogachev, M. I., and Bunde, A. (2012). Universality in the precipitation and river runoff. *EPL (Europhysics Letters)*, 97(4), 48011.

Bommier, E. (2014). *Peaks-Over-Threshold Modelling of Environmental Data*. U.U.D.M. Project Report 2014:33, Department of Mathematics, Uppsala University.

Calenda, G., Petaccia, A., and Togna, A. (1977). Theoretical probability distribution of critical hydrologic events by the partial-duration series method. *Journal of Hydrology*, 33(3-4), 233-245.

Claps, P., and Laio, F. (2003). Can continuous streamflow data support flood frequency analysis? An alternative to the partial duration series approach. *Water Resources Research*, 39(8), 1216.

Coles S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics, Springer, London.

De Michele, C., and Avanzi, F. (2018). Superstatistical distribution of daily precipitation extremes: A worldwide assessment. *Scientific reports*, 8, 14204.

Eastoe, E. F., and Tawn, J. A. (2010). Statistical models for overdispersion in the frequency of peaks over threshold data for a flow series. Water Resources Research, 46(2).

Eichner, J. F., Kantelhardt, J. W., Bunde, A., and Havlin, S. (2011). The statistics of return intervals, maxima, and centennial events under the influence of long-term correlations. In J. Kropp & H.-J. Schellnhuber (Eds.), *Extremis* (pp. 2–43). Berlin, Heidelberg: Springer.

Favre, A. C., El Adlouni, S., Perreault, L., Thiémonge, N., and Bobée, B. (2004). Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, 40(1).

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, vol. I, 3rd edition, London-New York-Sydney-Toronto, John Wiley & Sons.

Fernández, B., and Salas, J. D. (1999). Return period and risk of hydrologic events. I: mathematical formulation. *Journal of Hydrologic Engineering*, 4(4), 297-307.

Ferro, C. A., and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2), 545-556.

Fisher, R., & Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2), 180-190.

Fuller, W. E. (1914). Flood flows. *Transactions of the American Society of Civil Engineers*, 77, 564-617.

Graham, R. L., Knuth, D. E., and Patashnik, O. (1994). Concrete Mathematics: A Foundation for Computer Science, 2nd ed. Reading, MA: Addison-Wesley.

Hazen, A. (1914). The storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers*, 77, 1539-1669.

Klein Tank, A.M.G. et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22(12), 1441-1453.

Koutsoyiannis, D. (2004a). Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. *Hydrological Sciences Journal*, 49(4), 575–590.

Koutsoyiannis, D. (2004b). Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records. *Hydrological Sciences Journal*, 49(4), 591–610.

Koutsoyiannis, D., and Montanari, A. (2015). Negligent killing of scientific concepts: the stationarity case. *Hydrological Sciences Journal*, 60(7-8), 1174-1183.

Koutsoyiannis, D., and Papalexiou, S.M. (2017). Extreme rainfall: Global perspective, *Handbook of Applied Hydrology*, Second Edition, edited by V.P. Singh, 74.1–74.16, McGraw-Hill, New York.

Krzysztofowicz, R. (1997). Transformation and normalization of variates with specified distributions. *Journal of Hydrology*, 197(1-4), 286-292.

Iliopoulou, T., and Koutsoyiannis, D. (2019). Revealing hidden persistence in maximum rainfall records. *Hydrological Sciences Journal*, doi: 10.1080/02626667.2019.1657578.

Leadbetter M. R. (1974). On extreme values in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 28, 289–303.

Leadbetter M. R. (1983). Extremes and local dependence in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65, 291–306.

Lombardo, F., Volpi, E., Koutsoyiannis, D., and Serinaldi, F. (2017). A theoretically consistent stochastic cascade for temporal disaggregation of intermittent rainfall. *Water Resources Research*, 53(6), 4586-4605.

Lombardo, F., Montesarchio, V., Napolitano, F., Russo, F., and Volpi, E. (2009). Operational applications of radar rainfall data in urban hydrology. In *Proceedings of a symposium on the role of hydrology in water resources management, Capri, Italy, October 2008.* (pp. 258-265). IAHS Press.

Lombardo, F., Napolitano, F., & Russo, F. (2006a). On the use of radar reflectivity for estimation of the areal reduction factor. *Natural Hazards and Earth System Sciences*, *6*(3), 377-386.

Lombardo, F., Napolitano, F., Russo, F., Scialanga, G., Baldini, L., and Gorgucci, E. (2006b). Rainfall estimation and ground clutter rejection with dual polarization weather radar. *Advances in Geosciences*, 7, 127-130.

Luke, A., Vrugt, J. A., AghaKouchak, A., Matthew, R., and Sanders, B. F. (2017). Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the United States. *Water Resources Research*, 53(7), 5469-5494.

Marani, M., and Ignaccolo, M. (2015). A metastatistical approach to rainfall extremes. *Advances in Water Resources*, 79, 121-126.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7), 897-910.

Montanari, A. (2012). Hydrology of the Po River: looking for changing patterns in river discharge. *Hydrology and Earth System Sciences*, 16, 3739-3747.

O'Connell, P. E., Koutsoyiannis, D., Lins, H. F., Markonis, Y., Montanari, A., and Cohn, T. (2016). The scientific legacy of Harold Edwin Hurst (1880–1978). *Hydrological Sciences Journal*, 61(9), 1571-1590.

Papalexiou, S. M., and Koutsoyiannis, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, 49, 187-201.

Papalexiou, S. M., Koutsoyiannis, D., and Makropoulos, C. (2013). How extreme is extreme? An assessment of daily rainfall distribution tails. *Hydrology and Earth System Sciences*, 17, 851-862.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge University Press.

Salas, J. D., Obeysekera, J., & Vogel, R. M. (2018). Techniques for assessing water infrastructure for nonstationary extreme events: a review. *Hydrological Sciences Journal*, 63(3), 325-352.

Salvadori, G., De Michele, C., Kottegoda, N. T., & Rosso, R. (2007). *Extremes in nature: an approach using copulas*. Vol. 56. Springer Science & Business Media.

Schmidt, R. (2005). Tail dependence. In *Statistical Tools for Finance and Insurance* (pp. 65-91). Springer, Berlin, Heidelberg.

Serinaldi, F., and Kilsby, C. G. (2014). Rainfall extremes: Toward reconciliation after the battle of distributions. *Water Resources Research*, 50(1), 336-352.

Serinaldi, F., and Kilsby, C. G. (2016). Understanding persistence to avoid underestimation of collective flood risk. *Water*, 8(4), 152.

Serinaldi, F., and Kilsby, C. G. (2018). Unsurprising Surprises: The Frequency of Record-breaking and Overthreshold Hydrological Extremes Under Spatial and Temporal Dependence. *Water Resources Research*, 54(9), 6460-6487.

Serinaldi, F., Kilsby, C. G., and Lombardo, F. (2018). Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology. *Advances in Water Resources*, 111, 132-155.

Todorovic, P. (1970). On some problems involving random number of random variables. *The Annals of Mathematical Statistics*, 41(3), 1059–1063.

Todorovic, P., and Zelenhasic, E. (1970). A stochastic model for flood analysis. *Water Resources Research*, 6(6), 1641–1648.

Volpi, E., Fiori, A., Grimaldi, S., Lombardo, F., and Koutsoyiannis, D. (2015). One hundred years of return period: Strengths and limitations. *Water Resources Research*, 51(10), 8570-8585.

Volpi, E., Fiori, A., Grimaldi, S., Lombardo, F., and Koutsoyiannis, D. (2019). Save hydrological observations! Return period estimation without data decimation. *Journal of Hydrology*, 571, 782-792.

Zorzetto, E., Botter, G., and Marani, M. (2016). On the emergence of rainfall extremes from ordinary events. *Geophysical Research Letters*, 43(15), 8076-8082.