Hydrological Sciences Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/thsj20

The tales that the distribution tails of non-Gaussian autocorrelated processes tell: Efficient methods for the estimation of the *k*-length block-maxima distribution

Ioannis Tsoukalas

To cite this article: Ioannis Tsoukalas (2021): The tales that the distribution tails of non-Gaussian autocorrelated processes tell: Efficient methods for the estimation of the *k*-length block-maxima distribution, Hydrological Sciences Journal, DOI: 10.1080/02626667.2021.2014056

To link to this article: https://doi.org/10.1080/02626667.2021.2014056

Publisher: Taylor & Francis & IAHS

Journal: *Hydrological Sciences Journal*

DOI: 10.1080/02626667.2021.2014056

The tales that the distribution tails of non-Gaussian autocorrelated processes tell: Efficient methods for the estimation of the *k*-length block-maxima distribution

Ioannis Tsoukalas

Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, 15780 Zographou, Greece (itsoukal@mail.ntua.gr)

Abstract

Focal point of this work is the estimation of the distribution of maxima without the use of classic extreme value theory and asymptotic properties, which may not be ideal for hydrological processes. The problem is revisited from the perspective of non-asymptotic conditions, and regards the so-called exact distribution of block-maxima of finite-sized *k*-length blocks. First, we review existing non-asymptotic approaches/models, and also introduce an alternative and fast model. Next, through simulations and comparisons (using asymptotic and non-asymptotic models), involving intermittent processes (e.g., rainfall), we highlight the capability of non-asymptotic approaches to model the distribution of maxima with reduced uncertainty and variability. Finally, we discuss an alternative use of such models that concerns the theoretical estimation of the multi-scale probability of obtaining a zero value. A useful finding when the scope is the multi-scale modeling of intermittent hydrological processes (e.g., intensity-duration-frequency models). The work also entails step-by-step recipes and an R-package.

Keywords: non-asymptotic distribution of *k*-length block maxima; non-Gaussian marginal distribution; autocorrelated processes; intermittent processes; Gaussian copula.

1 Introduction

"We have to remember that what we observe is not nature herself, but nature exposed to our method of questioning."

W. K. Heisenberg

The probabilistic laws that rule the magnitude and frequency of occurrence of extreme events, particularly maxima, have been for years an active topic of research in a variety of scientific fields, including finance, actuarial science, as well as engineering, due to the severe and disruptive consequences that such events may cause (Embrechts et al. 1997; Reiss et al. 1997; Smith 2003; Beirlant et al. 2006). Of course, hydrological sciences are not an exception, rather a pioneer in the topic, as evidenced by the fact that the first relevant studies (Fuller 1914; Hazen 1914) were published more than 100 years ago. Also, hydrological applications were some of the first ones where the statistical theory of extreme value theory (EVT) was applied (Gumbel 1941b, a, 1958 pp. 236-245). As remarked by Katz et al. (2002), this is highlighted by E. Gumbel (1958) where in his iconic book, *Statistics of Extremes,* it is stated that *"the oldest problems connected with extreme values arise from floods"*. Eventually, the works of Gumbel together with other pioneering works in the domain (Fréchet 1927; Fisher and Tippett 1928; von Mises 1936; Gnedenko 1943; Jenkinson 1955; de Haan 1971; Galambos 1972; Leadbetter 1974, 1983; Pickands III 1975) are considered nowadays cornerstones for the development EVT. Studies that remain influential and arguably standard references in multiple scientific domains. For a thorough historical survey see Kotz and Nadarajah, (2000 ch. 1.1), while complete treatments on the subject can be found in Resnick (1987), Reiss et al. (1997), Coles (2001), Smith (2003), Salvadori et al. (2007), and Koutsoyiannis (2020).

The importance of extremes in hydrological engineering is manifested by the critical nature of hydraulic infrastructures (e.g., reservoirs, spillways, dikes, etc.), which among the many purposes they serve (e.g., water supply, energy production), they are constructed to offer protection and security against rare, and extreme phenomena (e.g., floods). Thus, by definition their design and management have to take into consideration the probabilistic behaviour of extremes, i.e., account for the distribution's tails (in particular the right one for maxima), where the extremes *live.* This criticality has motivated a significant amount of research in the domain hydrological extremes, offering a variety of approaches (Buishand 1989, 1991; Pilon et al. 1991; Wilks 1993; Koutsoyiannis et al. 1998; Koutsoyiannis 1999, 2004, 2020; Katz et al. 2002; Park and Jung 2002; Coles et al. 2003; Favre et al. 2004; Wilson and Toumi 2005; Deidda and Puliga 2006; Calenda et al. 2009; Svensson and Jones 2010; Volpi and Fiori 2012, 2014; Cavanaugh et al. 2015; Marani and Ignaccolo 2015; Volpi et al. 2015, 2019; Zorzetto et al. 2016; Blum et al. 2017; Salas et al. 2018; Ye et al. 2018; De Michele and Avanzi 2018; Salas and Obeysekera 2019; Benestad et al. 2019; Courty et al. 2019; De Michele 2019; Lombardo et al. 2019; Iliopoulou and Koutsoyiannis 2020; Serinaldi et al. 2020), just to name a few. For a thorough discussion on hydroclimatic extremes, and associated methodological approaches, the interested reader is referred to the recent book of Koutsoyiannis (2020). Nonetheless, many of these approaches are typically build upon the assumption of stationarity, while more recently special attention is given to methods accounting for non-stationarity (for a detailed review on such approaches see

Salas et al. (2018)). It is remarked that this modelling choice is for years a matter of significant debate (Lins and Cohn 2011; Matalas 2012; Montanari and Koutsoyiannis 2014; Koutsoyiannis and Montanari 2015), with several works highlighting the importance of the hypothesis of stationarity.

Notwithstanding the latter assumption, the majority of research works involve the use of EVT and the associated asymptotic laws and distributions. That is, focusing on the asymptotic distribution of block maxima (BM) or peaks over threshold (POT) exceedances, for a large number (tending to infinity) of events (block size or data respectively), under the hypothesis of temporal independence (i.e., zero serial correlation). Another important concept that is worth mentioning, despite the fact it is not in the core of this work, is the so-called extremal index (Leadbetter 1983; Leadbetter et al. 1983; Hsing et al. 1988; Leadbetter and Rootzen 1988; Coles 2001 Ch. 5), which extends the classical asymptotic EVT for processes with zero serial correlation to serially dependent (stationary) ones. It is argued (Hsing et al. 1988) that the extremal index is an important quantity for dependent processes since it can be interpreted as the reciprocal of the expected size of an extremal cluster above high thresholds (i.e., quantify the extend of clustering of extremes). More details on the topic, as well as on estimation methods can be found in the above seminal works as well as in literature (e.g., Smith and Weissman 1994; Embrechts et al. 1997).

On the other hand, the use of non-asymptotic laws and distributions (also called *exact* in Gumbel's (1958) terminology) has received little attention in hydrology (Koutsoyiannis 2004). Probably due to, 1) the distribution of the base/parent processes has to be *a priori* known (or inferred from data), and 2) the convenience offered (in terms of data storage/management and computation) by utilizing limiting laws that imply the use of subsets of data (i.e., inference on the distribution of maxima using only BM or POT observations). Of course, this convenience comes at the *cost* of neglecting the effect of temporal dependence, as well as neglecting observations *per se* (e.g., the second and third larger maxima within a block), facts that arguably affect the inference about the extremes behaviour (Volpi et al. 2019; Lombardo et al. 2019; Koutsoyiannis 2020; Serinaldi et al. 2020).

In this line of thought, recent research efforts favor the use of the complete record of observations and also involve the derivation of the non-asymptotic distribution of maxima over a finite-size block. Representative works are those of Marani and Ignaccolo (2015), Zorzetto et al. (2016), De Michele and Avanzi (2018), Volpi et al. (2019) and Koutsoyiannis (2020), which are all however linked with the assumption of temporal independence. See also the recent review on the topic by De Michele (2019), focusing on the exact distribution of maximum annual daily precipitation. Notable exceptions are the works of Lombardo et al. (2019) and Serinaldi et al. (2020), that regard the *exact* distribution of k-length block maxima under the assumptions of Markov and general autocorrelation structures respectively. In particular, the latter work via thorough analyses and insightful discussions clarifies, and brings *order*, in many delicate matters that concern the non-asymptotic distribution of maxima of autocorrelated processes (highlighting also important links with the seminal works of Todorovic and Zelenhasic (1970), and Todorovic (1970)).

Having the above in mind, and motivated by the potential, as well as the growing interest in non-asymptotic distributions for the modelling of k-length block maxima herein we aim to:

1. Review (section 3) the existing models/approaches for the derivation of the exact distribution of k-length block maxima, under the assumption that they arise from a stationary, discrete-time, non-Gaussian autocorrelated process with underlying dependence given by the Nataf's joint distribution (i.e., the Gaussian copula).
2. Implement the reviewed models in an R package, called `bBextremes` (freely available), favoring this way further research on the topic, as well as easy application of these approaches in real-world studies.
3. Introduce (section 3.6) a fast model for the exact distribution of k-length block maxima by combining the exact distribution of an autoregressive process of order n (AR(n)) and a recent finding of Serinaldi et al. (2020) that regards the ability of Beta-Binomial ($\beta\mathcal{B}$) distribution to model the exact distribution of k-length block maxima of autocorrelated processes.
4. Conduct a series of simulation experiments (section 4), involving intermittent non-Gaussian autocorrelated processes, to demonstrate and also compare the performance of the reviewed models for the estimation of the distribution of k-length block maxima with that of more classical approaches that involve: a) the asymptotic distribution of block maxima (i.e., the Generalized Extreme Value distribution) and, b) alternative, non-extreme-value distribution models for the k-length block maxima series. A comparison that aims to provide answers to questions such as: *Which model/approach should we employ to model the distribution of k-length block maxima? Are the non-asymptotic models more robust than the asymptotic ones? Using which approach (asymptotic or non-asymptotic) we have better chances to identify the true distribution of k-length block maxima?*
5. Highlight an additional use of the $\beta\mathcal{B}$ distribution, and the related models, that regards intermittent processes (e.g., rainfall at fine time scales), and the estimation of the probability of obtaining a zero value (also called probability dry, expressing the probability of a dry/zero interval) at multiple temporal scales of aggregation (section 5). A finding that can be of particular use when the scope is the multi-scale modeling of intermittent hydrological processes, such as the development of intensity-duration-frequency models (see section 5.4).

The remaining of this work is structured as follows: section 2 provides a brief introduction to the necessary concepts and notions used in this work. Section 3 discusses and reviews the existing approaches for the derivation of the exact distribution of k-length block maxima, as well as introduces a new model that combines the exact distribution of maxima of an AR(n) process and the $\beta\mathcal{B}$ distribution. Section 4 concerns the simulation experiments and comparisons mentioned above (point #4), while section 5 focuses on point #5. Finally, section 6 summarizes the key points and findings of this work through a brief discussion, noting also directions for future research.

2 Notation and introduction of key concepts

It is remarked that unless stated otherwise, this work concerns univariate stationary discrete-time processes with continuous or zero-inflated marginal distributions with finite variance as well as valid (i.e., positive definite) autocorrelation structures. Also,

we focus only on non-negative autocorrelation structures, since they are abundant in hydrometeorological processes. In addition, an assumption made throughout this work is that the latter processes are characterized by the Nataf's joint distribution model (Nataf 1962; Liu and Der Kiureghian 1986), hence called Nataf-based processes (Tsoukalas et al. 2018a). Such processes are essentially characterized by Gaussian copula dependence structure (often termed meta-Gaussian, see Kelly and Krzysztofowicz (1997)); since they rely on the (typically non-linear) mapping (i.e., transformation, translation) of an appropriately parameterized Gaussian process.

Further to the brief introduction provided in this section, the interested reader is referred to Tsoukalas et al. (2020) for a general overview and historical roadmap of such developments, as well as computer software (R package `anySim`). Similar characteristic works of this kind can be found in hydrological domain, starting with the works of Matalas (1967), Klemeš and Borůvka (1974), and Bell (1987), as well as more recent ones (Tsoukalas et al. 2017, 2018a, b, 2019; Serinaldi and Lombardo 2017; Papalexiou 2018; Kossieris et al. 2019). Interestingly, similar developments can also be found in broader engineering literature (Gujar and Kavanagh 1968; Li and Hammond 1975; Grigoriu 1984, 1998; Liu and Der Kiureghian 1986; Yamazaki and Shinozuka 1988; Emrich and Piedmonte 1991; Cario and Nelson 1996; Popescu et al. 1998; Biller and Nelson 2003; Christakos 2012).

To provide context, let $\{X_t\}_{t\in Z^{\geq}}$, where t denotes the time index, be a discrete-time stationary process with arbitrary, continuous or zero-inflated, marginal distribution, $F_X(x;\boldsymbol{\theta}) \coloneqq \mathrm{P}\{X \leq x\}$, where $\boldsymbol{\theta}$ is a vector that denotes the distribution's parameters. Let also the autocorrelation structure (ACS) of the process be denoted by, $\rho_\tau \coloneqq \mathrm{Corr}[X_t, X_{t+\tau}]$, where $\tau \in \{0, \pm1, \pm2, \dots\}$ stands for the time lag. Hereafter, and without loss of generality the parameter vector $\boldsymbol{\theta}$ will be omitted for the sake of simplicity.

A realization (i.e., a time series) x_t of the process X_t can be generated by the non-linear mapping of an auxiliary standard Gaussian process $\{Z_t\}_{t\in Z^{\geq}}$, hereafter abbreviated as Gp, with autocorrelation structure, $\tilde{\rho}_\tau \coloneqq \mathrm{Corr}[Z_t, Z_{t+\tau}]$, hereafter termed *equivalent* ACS, through the mapping operation, $x = F_X^{-1}\big(\Phi(z)\big)$, where Φ denotes the cumulative distribution function (CDF) of the standard Gaussian distribution, and F_X^{-1} denotes the inverse of the desired CDF (ICDF), also known as quantile function. Note that the relationship $z = \Phi^{-1}\big(F_X(x)\big)$ also holds true.

A delicate and important detail to recall is that the ACSs of the Gaussian and the target process are related via a double infinite integral (see the references mentioned above; e.g., Tsoukalas et al. (2019)), defining a relationship between ρ_τ and $\tilde{\rho}_\tau$, abbreviated as Target-Equivalent correlation (TEC) relationship. The *shape* and non-linearity of the TEC relationship depends on the marginal distribution of X_t as well as its parameters $\boldsymbol{\theta}$, i.e., $F_X(x;\boldsymbol{\theta})$. Hence, it can be shorthanded as,

$$\rho_\tau = \mathcal{F}(\tilde{\rho}_\tau | F_X) \tag{1}$$

where $\mathcal{F}$ denotes the TEC relationship. Note that the dependence of TEC on the distribution's parameter vector $\boldsymbol{\theta}$ has been omitted for brevity. To establish a target process X_t with a desired ACS, ρ_τ, the auxiliary Gp has to be parameterized with an appropriate (called equivalent) ACS $\tilde{\rho}_\tau$, which can be found by inverting the above TEC relationship, i.e.,

$$\tilde{\rho}_\tau = \mathcal{F}^{-1}(\rho_\tau | F_X) \tag{2}$$

It is noted that the above procedure can also be applied for the simulation of processes with discrete marginal distributions, such as, the Bernoulli, the Poisson or the Beta-Binomial (e.g., Serinaldi and Lombardo 2017; Papalexiou 2018; Tsoukalas et al. 2019, 2020). Herein, let us briefly discuss the case of processes with Bernoulli marginal distribution, which will be used next for the estimation of the distribution of maxima of non-Gaussian autocorrelated processes (see section 3). A discussion also useful for the theoretical estimation of the probability of obtaining a zero value at multiple scales (see section 5) for intermittent processes.

Following the same notation, a process X_t can be dichotomized at any level $x_u = F_X^{-1}(u)$, and converted into a binary process, denoted by Y_t, having Bernoulli marginal distribution with parameter p (i.e., $\mathcal{B}e(p)$), with state space $\{0, 1\}$ and probabilities $\mathrm{P}\{Y = 0\} = \mathrm{P}\{X_t \leq x_u\} = u = 1 - p$ and $\mathrm{P}\{Y = 1\} = \mathrm{P}\{X_t > x_u\} = 1 - u = p$, by applying the following mapping procedure,

$$Y_t = \begin{cases} 1, & \text{if } X_t > x_u = F_X^{-1}(u) = F_X^{-1}(1-p) \\ 0, & \text{if } X_t \leq x_u = F_X^{-1}(u) = F_X^{-1}(1-p) \end{cases} \tag{3}$$

The above procedure will yield a process with $Y_t \sim \mathcal{B}e\big(p = 1 - F_X(x_u)\big)$ and ACS, $\ddot{\rho}_\tau = \mathrm{Corr}[Y_t, Y_{t+\tau}]$, which will be different from that of X_t, i.e., ρ_τ. This is due to the use of the latter non-linear mapping operation. In order to find $\ddot{\rho}_\tau$, which depends both on $\rho_\tau = \mathcal{F}(\tilde{\rho}_\tau | F_X)$ and the dichotomization level x_u, $\tilde{\rho}_\tau$ should be first estimated by inverting TEC (Eq. (2)), and then by setting as "*target*" the Bernoulli marginal distribution with parameter $p = 1 - F_X(x_u)$, convert the $\tilde{\rho}_\tau$ to $\ddot{\rho}_\tau$. The latter procedure is expressed as,

$$\ddot{\rho}_\tau(x_u, \rho_\tau) = \mathcal{F}_{\mathcal{B}e}\left(\mathcal{F}^{-1}(\rho_\tau | F_X) | F_Y \equiv \mathcal{B}e\big(p = 1 - F_X(x_u)\big)\right) \tag{4}$$

or more conveniently as a function of $\tilde{\rho}_\tau$ as,

$$\ddot{\rho}_\tau(x_u, \tilde{\rho}_\tau) = \mathcal{F}_{\mathcal{B}e}\left(\tilde{\rho}_\tau | F_Y \equiv \mathcal{B}e\big(p = 1 - F_X(x_u)\big)\right) \tag{5}$$

where $\mathcal{F}_{\mathcal{B}e}$ is an abbreviation for the TEC relationship specifically for the case of $\mathcal{B}e(p)$ marginal distribution. It is interesting to note that the ACS $\ddot{\rho}_\tau$ of a binary process is closely related with the notions of extremogram (Davis and Mikosch 2009), and tail dependence coefficients of a univariate process (Beirlant et al. 2006), which in turn can be viewed as analogues of the autocorrelation function of extreme values (i.e., values of a series exceeding a specified threshold x_u).

Nevertheless, to estimate the ACS of Y_t, i.e., $\ddot{\rho}_\tau$, it is required to resolve a double infinite integral (e.g., Emrich and Piedmonte 1991; Serinaldi and Lombardo 2017; Serinaldi et al. 2020), as in the general case of Nataf (or Gaussian copula) -based processes, or resort to alternative approximate relationships (Liu and Der Kiureghian 1986; Serinaldi and Lombardo 2017). For reasons related with computational speed, herein we follow the

latter rationale and thus introduce an approximate closed-form expression for the TEC relationship for processes with $\mathcal{B}e(p)$ marginal distribution. In particular we used the following function,

$$\ddot{\rho}_\tau = \left(1 - (1-\tilde{\rho}_\tau)^{\frac{1}{\gamma_2(p_*)}}\right)^{\frac{1}{\gamma_1(p_*)}} \tag{6}$$

where, $p_* = \min(1-p, p)$, $p \in (0,1)$, while $\gamma_1(p_*)$ and $\gamma_2(p_*)$ are parameters that depend on p_* (and hence to $\mathcal{B}e$ distribution parameter p). The parameters are given by,

$$\gamma_1(p_*) = 0.257 + 1.382{p_*}^{0.304+0.129p_*}\exp(-0.484p_*) \tag{7}$$

$$\gamma_2(p_*) = 1.972 + 84.246{p_*}^{1.588+8.698p_*}\exp(-21.101p_*) \tag{8}$$

The functional forms of Eq. (7) and Eq. (8), depicted in **Figure 1** for $0.001 \leq p \leq 0.5$, resemble those used by Serinaldi and Lombardo (2017). However, here they are combined with Eq. (6), which in turn resembles the form of the Kumaraswamy ICDF (Kumaraswamy 1980), which has been suggested earlier as a good candidate model for approximating TEC relationships (Papalexiou 2018).

Compared to the more exact, but time consuming, numerical integration approach (see **Figure 2**), the proposed approximate solution provides a mean absolute difference error equal to ~0.0065, while the maximum error, in terms of simple difference is equal to ~ 0.04 for $p = 10^{-4}$. This error can be considered small for practical applications, given the computational gains that this approximation offers. Furthermore, in comparison with the solution provided by Serinaldi and Lombardo (2017), the proposed approximation offers two main advantages: 1) it consists a much simpler closed-form formula, since it avoids the use of Beta's distribution CDF that does not has an explicit expression (as in the aforementioned paper), and 2) it can be easily inverted to find the equivalent (i.e., Gaussian) correlation coefficients. The latter advantage is of particular use when the objective is the simulation of processes with $\mathcal{B}e(p)$ marginal distribution and any (valid) ACS, which in such cases, $\rho_\tau \equiv \ddot{\rho}_\tau$. In particular, the inverse of Eq. (6), is given by,

$$\tilde{\rho}_\tau = 1 - \left(1 - \ddot{\rho}_\tau^{\gamma_1(p_*)}\right)^{\gamma_2(p_*)} \tag{9}$$

Further to the above approach an alternative solution would be the use of an approximation formula for the bivariate Gaussian CDF (such as the one proposed by Koutsoyiannis (2020)), which is *hidden* in the double integral that links the Gaussian and target correlations in the case of $\mathcal{B}e$ processes (see Eq. (2.1) in Emrich and Piedmonte 1991; Serinaldi and Lombardo 2017). In particular, the ACS of a $\mathcal{B}e(p)$ and a standard Gaussian process are linked by,

$$\ddot{\rho}_\tau = \frac{\Phi_2(z_p;\tilde{\rho}_\tau) - p^2}{p(1-p)} \quad (10)$$

where $z_p = \Phi^{-1}(p)$, and $\Phi_2(z_p;\tilde{\rho}_\tau)$ stands for the bivariate standard Gaussian CDF with identical inputs z_p, i.e., $\Phi_2(z_p, z_p;\tilde{\rho}_\tau) \equiv \Phi_2(z_p;\tilde{\rho}_\tau)$. The function Φ_2 can be solved through numerical integration, or more conveniently approximated (for purposes of computational speed-up). Herein we suggest the use of a novel and accurate approximation (see Eq. (5.48) in Koutsoyiannis (2020), as well as the relevant comparisons highlighting the approximation's accuracy), which reads as follows (for the sake of simplicity the indices p and τ have been omitted),

$$\Phi_2(z) \approx \Phi(z) - \Phi(-|z|) + \exp\left(\frac{m^2}{2}\right)\frac{\Phi(-s|z| - m)}{s} \quad (11)$$

where, $m \coloneqq 2\sqrt{\frac{1-\tilde{\rho}}{17+\tilde{\rho}}}$ and $s \coloneqq \frac{1}{3}\sqrt{\frac{17+\tilde{\rho}}{1+\tilde{\rho}}}$. To illustrate the accuracy of this approximation for the purposes of this work we formulated an additional comparison with the numerical integration approach (similar to that of **Figure 2**), yet this time focusing on very low values of p spanning from 10^{-2} to 10^{-9}. In this case the mean absolute difference error is equal to ~0.0023, while the maximum error, in terms of simple difference, is equal to -0.023, for $p = 10^{-9}$ (see **Figure 3**). When cross-comparing the two approximation approaches it is found that, the former (i.e., Eq. (6)) performs better for values of p in the range of 0.5 to 10^{-3} (which was used to calibrate the parameters of Eq. (7) and (8)) while the latter approach (i.e., Eq. (11)) provides smaller approximation errors for values of p smaller than 10^{-3}. Having the above empirical results in mind, and aiming to minimize the approximation error, for all subsequent analyses we employ the first approach for $p \in [10^{-3}, 0.5]$, while employ the second one for values of $p < 10^{-3}$.

Finally, it is noted that for convenience and subsequent use, hereafter we denote the autocorrelation matrices (with dimensions $k \times k$) of the target, Gaussian and binary processes, by $\boldsymbol{R}_k$, $\tilde{\boldsymbol{R}}_k$ and $\ddot{\boldsymbol{R}}_k$ respectively. By definition, the above matrices are Toeplitz ones, and can be established using the corresponding ACSs, i.e., ρ_t, $\tilde{\rho}_\tau$ and $\ddot{\rho}_\tau$ respectively. For instance, the elements of $\boldsymbol{R}_k$ are $[\boldsymbol{R}_k]_{i,j} = \rho_{|i-j|}$, where i and j are indices denoting rows and columns respectively. The element of $\boldsymbol{R}_k$ in most right (left) column and bottom (top) row is equal to $\rho_{|1-k|}$ ($\rho_{|k-1|}$). Therefore, for convenience we may write, $\boldsymbol{R}_k = toeplitz(\rho_{1:(k-1)})$, where $toeplitz(x)$ is a function that transforms a vector x into a Toeplitz matrix.

3 The distribution of k-length block maxima of non-Gaussian autocorrelated processes

3.1 The general case: in view of the Gaussian copula

The distribution of k-length block maxima $M_k = \max_{1\le i\le k}\{X_i\}$ over a block of k time steps (i.e., of k random variables) is denoted by $F_{M_k}(x) \coloneqq \mathrm{P}\{X_1 \le x, \dots, X_k \le x\}$, and expresses the probability of not exceeding the value x in a block of size k, (i.e., in k time steps). Its reciprocal, or complementary function, $\bar{F}_{M_k}(x) = 1 - F_{M_k}(x)$, expresses the probability of exceeding the value x (i.e., observing at least one exceedance of x in a k-

sized block). By construction the distribution F_{M_k} is identical to the multivariate distribution of order k of the so-called data-generating, base or parent process X_t, i.e.,

$$F_{M_k}(x) = F_{\boldsymbol{X}_k}(x; \boldsymbol{R}_k) = F_{\boldsymbol{X}_k}(x, .., x; \boldsymbol{R}_k) = \mathrm{P}\{X_1 \leq x, \dots, X_k \leq x\} \tag{12}$$

which in turn can be expressed using the notion of copulas, and in particular using a k-dimensional copula function denoted by $C_k(\cdot)$, i.e.,

$$F_{M_k}(x) = F_{\boldsymbol{X}_k}(x; \boldsymbol{R}_k) = \mathrm{P}\{X_1 \leq x, \dots, X_k \leq x\} = C_k(F_X(x), \dots, F_X(x); \boldsymbol{\theta}) \tag{13}$$

where $\boldsymbol{\theta}$ is a vector that contains the parameters of the copula. Also note that in the case of stationary processes, the marginal distribution of X_t is identical for all t and denoted by F_X (hence the index t can be safely omitted), while, since we are concerned with the distribution of k-length block maxima (hereafter denoted simply as the distribution of maxima) the input in the multivariate distributions is just x.

It is recalled that a copula (Sklar 1959, 1973) is a multivariate distribution with uniform ($\mathcal{U}$) marginals, which allows the establishment of a plethora of multivariate distributions, by modelling separately the dependence structure and the marginal distributions of the random variables (provided that both the copula and the marginals are differentiable). During the recent years copulas have gained significant popularity in a variety of scientific domains (e.g., Schweizer et al. 1991; Nelsen 2007), including that of hydrology (e.g., De Michele and Salvadori 2003; Favre et al. 2004; Salvadori and De Michele 2004, 2007; Salvadori et al. 2007; Chen and Guo 2019; Zhang and Singh 2019), since they have been proven particularly useful modelling *tools* to characterize and model a variety of dependence structures.

In this work, the focus is given to the Gaussian copula for three main reasons. 1) It enables the relatively easy and straighthood modelling of multivariate distributions with more than two dimensions, compared to alternatives requiring pair-copula constructions (Aas et al. 2009; Joe 2014). 2) It consists the dependence structure of Nataf-based processes, which are used herein to generate synthetic time series, as well as validate the reviewed non-asymptotic models of F_{M_k}. 3) It has been widely used in hydrological domain for a variety of modelling purposes; see for instance the reviewed works of Section 2, the work of Renard and Lang (2007) who demonstrate the use of this copula in four different applications (some of them related with extremes), as well as works related with the meta-Gaussian distribution model (Kelly and Krzysztofowicz 1997; Fang et al. 2002; Herr and Krzysztofowicz 2005; Genest et al. 2007), which can be viewed as a Gaussian copula (for more applications of this kind see the brief review on meta-elliptical copulas presented in section 2.3.3 of Chen and Guo (2019)). However, and despite the above reasoning, it is remarked that the Gaussian copula is related with the property of asymptotic independence, and thus it is argued that the plausibility of this assumption is a topic requiring further research. With this in mind, all formulas presented in this section are first introduced using general copula functions, and then specialized for the specific case of Gaussian copula.

In particular, using the Gaussian copula, i.e., $C_k(\cdot) = \Phi_k\left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_k); \widetilde{\boldsymbol{R}}_k\right)$, where $u_i \sim \mathcal{U}(0,1)$, and $\Phi_k\left(\dots; \widetilde{\boldsymbol{R}}_k\right)$ denotes the k^{th} order multivariate standard Gaussian CDF with zero mean and correlation matrix $\widetilde{\boldsymbol{R}}_k$, we obtain the following relationships for the CDF of M_k,

$$\mathrm{P}\{X_1 \le x, \dots, X_k \le x\} = F_{M_k}(x) = F_{\boldsymbol{X}_k}(x; \boldsymbol{R}_k) = \Phi_k\left(\Phi^{-1}\left(F_X(x)\right), \dots, \Phi^{-1}\left(F_X(x)\right); \widetilde{\boldsymbol{R}}_k\right) \quad (14)$$

It is also remarked that Eq. (14) does not have a closed-form solution, and requires the use of numerical integration methods, something that it is not a trivial task. For demonstration, as well as verification purposes, herein we obtained estimates of the latter formula using the `copula` R-package (Yan 2007; Kojadinovic and Yan 2010; Hofert et al. 2014). Also, it is noted that the latter package provides functions applicable up to dimension $k = 1000$, hence it cannot be used when the block size is larger than 1000 (e.g., for the estimation of annual maxima distribution of an hourly process; where $k = 8760$).

3.2 The case of independence

It is well known that the distribution of maxima M_k over a block of k time steps of a stationary stochastic process with zero ACS, i.e., consisted by k independent identically distributed (i.i.d.) random variables, has an explicit and tractable solution given by (Gumbel 1958 ch. 3.1, Eq. (1)),

$$F_{M_k}(x) = \prod_{i=1}^{k} F_{X_i}(x) = \left(F_X(x)\right)^k \quad (15)$$

where $F_X(x)$ is the marginal distribution of the base process X_t. The latter formula consists also the basis for the derivation of classical asymptotic results (i.e., $k \to \infty$) that regard the probability distribution of extremes (e.g., maxima). In particular, it has been shown (Fisher and Tippett 1928; Gnedenko 1943) that when $k \to \infty$ there are just three types of extreme value asymptotes, in Gumbel's terms, (or more simply, extreme value distributions), that is, the Fréchet (1927), the Gumbel (1958), and the reversed Weibull. All of them contained into the iconic Generalized Extreme Value distribution ($\mathcal{GEV}$; see Eq. (B8) in Appendix B) of von Mises (1936). It is recalled that $\mathcal{GEV}$ is the asymptotic distribution of M_k under the assumptions of independence and stationarity.

3.3 The case of AR(1) process

As elaborated in, the arguably not widely popularized, work of Hirtzel et al. (1982) (focusing though only in the Gaussian case), and recently in hydrological domain by Lombardo et al. (2019), the distribution of k-length block maxima M_k of a first order autoregressive process (i.e., AR(1) or Markov) can be obtained by a simple and tractable formula, i.e.,

$$F_{M_k}(x) = \left(\frac{F_{\boldsymbol{X}_2}(x)}{F_X(x)}\right)^{k-1} F_X(x) = \left(\frac{C_2\left(F_X(x), F_X(x)\right)}{F_X(x)}\right)^{k-1} F_X(x) \quad (16)$$

where $C_2\left(F_X(x), F_X(x)\right)$ is a bivariate copula. In the case of the Gaussian copula the above relationship is re-written as,

$$F_{M_k}(x) = \left(\frac{\Phi_2\left(\Phi^{-1}\left(F_X(x)\right), \Phi^{-1}\left(F_X(x)\right); \widetilde{\boldsymbol{R}}_2\right)}{F_X(x)}\right)^{k-1} F_X(x) \tag{17}$$

As in the more general case of Eq. (14), Φ_2 also requires the use of numerical integration techniques (herein implemented using the `copula` R package), which in this case impose a minimal computational overhead. An alternative, more computationally efficient route would be the use an approximation formula for Φ_2, such as a recently proposed one (Koutsoyiannis 2020 pp. 166-167), which proved to be quite accurate (see Figure 5.5 therein) for a wide range of parameters/values.

3.4 The case of AR(n) process

The previous model of M_k is being further extended herein for autoregressive processes of order n (AR(n)). In such cases, the distribution of maxima is provided by,

$$\begin{aligned} F_{M_k}(x) &= \left(\frac{F_{\boldsymbol{X}_{n+1}}(x)}{F_{\boldsymbol{X}_n}(x)}\right)^{k-n} F_{\boldsymbol{X}_n}(x) \\ &= \left(\frac{C_{n+1}\left(F_X(x), \dots, F_X(x)\right)}{C_n\left(F_X(x), \dots, F_X(x)\right)}\right)^{k-n} C_n\left(F_X(x), \dots, F_X(x)\right) \end{aligned} \tag{18}$$

where $C_k\left(F_X(x), \dots, F_X(x)\right)$ denotes a k-dimensional copula. For the Gaussian copula the above expression reads as follows,

$$F_{M_k}(x) = \left(\frac{\Phi_{n+1}\left(\Phi^{-1}\left(F_X(x)\right), \dots, \Phi^{-1}\left(F_X(x)\right); \widetilde{\boldsymbol{R}}_{n+1}\right)}{\Phi_n\left(\Phi^{-1}\left(F_X(x)\right), \dots, \Phi^{-1}\left(F_X(x)\right); \widetilde{\boldsymbol{R}}_n\right)}\right)^{k-n} \Phi_n\left(\Phi^{-1}\left(F_X(x)\right), \dots, \Phi^{-1}\left(F_X(x)\right); \widetilde{\boldsymbol{R}}_n\right) \tag{19}$$

However, the direct use of this model is particularly complicated and quite unstable, since it involves the ratio of two high-order multivariate distributions (i.e., $F_{\boldsymbol{X}_{n+1}}$ and $F_{\boldsymbol{X}_n}$), which in the case of Gaussian copula require the use of numerical integration, which may lead to accuracy problems. A potential, and computationally efficient remedy to this issue is discussed next at section 3.6.

3.5 An approximation based on the Beta-Binomial distribution

As shown by Serinaldi et al. (2020), the distribution of maxima over a block of size k, i.e., F_{M_k}, as well as the k-dimensional standard Gaussian distribution with identical inputs, can be very well approximated using the Beta-Binomial distribution (βB), which is used herein for computational purposes. It is also interesting to note that the abovementioned work found that same methodological approach can be used to estimate with satisfactory accuracy the *exact* distribution of maxima of AR(1) processes with other types of dependence structures, such those provided by the Clayton and Gumbel copulas (e.g., see Figure 1 in Serinaldi et al. (2020), as well as the relevant discussion therein).

Nonetheless, it is recalled that the βB distribution is typically used to describe the number of successes over a certain number of trials (i.e., exceedances over a block of size k). A random variable $W \sim \beta B$ has probability mass function (PMF) and CDF given by,

$$f_{\beta\mathcal{B}}(w; k, \alpha, \beta) = \mathrm{P}\{W = w\} = \binom{k}{w} \frac{\mathrm{B}(w + \alpha, k - w + \beta)}{\mathrm{B}(\alpha, \beta)}, \qquad w \in \{0,1, \dots, k\} \tag{20}$$

$$F_{\beta\mathcal{B}}(w; k, \alpha, \beta) = \mathrm{P}\{W \leq w\} = \sum_{i=0}^{w} \binom{k}{i} \frac{\mathrm{B}(i + \alpha, k - i + \beta)}{\mathrm{B}(\alpha, \beta)}, \qquad w \in \{0,1, \dots, k\} \tag{21}$$

where $\mathrm{B}(\alpha, \beta)$ denotes the complete beta function, while $k \in \mathbb{N}_0$ is a parameter denoting the number of trials, as well as $\alpha > 0$ and $\beta > 0$ are shape parameters. For subsequent use, it is noted that for $w = 0$, the PMF and CDF of βB result into an identical expression (since $\binom{k}{0} = 1$), that is,

$$f_{\beta B}(0; k, \alpha, \beta) = F_{\beta B}(0; k, \alpha, \beta) = \frac{\mathrm{B}(\alpha, k + \beta)}{\mathrm{B}(\alpha, \beta)} \tag{22}$$

According to the aforementioned work, the distribution of maxima F_{M_k} over a block of size k of an autocorrelated stationary process X_t can be very well approximated using the βB distribution, in particular the distribution of maxima M_k is estimated by,

$$\begin{aligned} F_{M_k}(x) = F_{\boldsymbol{X}_k}(x; \boldsymbol{R}_k) &= \Phi_k\big(\Phi^{-1}\big(F_X(x)\big), \dots, \Phi^{-1}\big(F_X(x)\big); \widetilde{\boldsymbol{R}}_k\big) \\ &\approx F_{\beta B}\big(0; k, \alpha_{\beta B}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big), \beta_{\beta B}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big)\big) \end{aligned} \tag{23}$$

As hinted by notation, both parameters of βB distribution depend on $F_X(x)$ and $\widetilde{\boldsymbol{R}}_k$ which can be found by the following relationships,

$$\alpha_{\beta B}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big) = \frac{1 - \ddot{\rho}_{IC}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big)}{\ddot{\rho}_{IC}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big)}\big(1 - F_X(x)\big) \tag{24}$$

$$\beta_{\beta B}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big) = \frac{1 - \ddot{\rho}_{IC}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big)}{\ddot{\rho}_{IC}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big)} F_X(x) \tag{25}$$

where, $\ddot{\rho}_{IC}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big)$ denotes the so-called intra-cluster correlation coefficient $\left(\ddot{\rho}_{IC} \in \left[-\frac{1}{k-1}, 1\right]\right)$, of the associated binary process, dichotomized at value x (hence the use of "*double dot*" notation for correlation coefficient); see Serinaldi et al. (2020), and references therein, for further details. In particular, $\ddot{\rho}_{IC}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big)$ can be estimated by,

$$\ddot{\rho}_{IC}\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big) = \frac{\mathbf{1}^{\mathrm{T}} \ddot{\boldsymbol{R}}_k\big(F_X(x), \widetilde{\boldsymbol{R}}_k\big)\mathbf{1} - k}{k(k-1)} \tag{26}$$

where $\mathbf{1}$ is a k-dimensional vector of ones, and $\ddot{\boldsymbol{R}}_k\left(F_X(x), \widetilde{\boldsymbol{R}}_k\right) = \mathcal{F}_{\mathcal{B}e}\left(\widetilde{\boldsymbol{R}}_k | F_Y \equiv \mathcal{B}e\left(p = 1 - F_X(x)\right)\right)$, which can be easily and well approximated using the closed-form TEC relationship of Eq. (6).

It is remarked that in the above discussion we used a slightly different notation than the one used by Serinaldi et al. (2020), aiming to further highlight the links between different *types* of correlation coefficients (i.e., those in the target, the Gaussian and the binary domain) as well as clarify their appropriate use to estimate F_{M_k}.

3.6 An alternative and fast approximation based on the Beta-Binomial distribution and the distribution of maxima of AR(n) processes

A potential issue with the above βB model for F_{M_k} is related with the fact that when k is large enough, then the Toeplitz operator required to establish $\ddot{\boldsymbol{R}}_k = toeplitz\left(\ddot{\rho}_{1:(k-1)}\right)$, where $\ddot{\rho}_{1:(k-1)} = \mathcal{F}_{\mathcal{B}e}\left(\tilde{\rho}_{1:(k-1)} | F_Y \equiv \mathcal{B}e\left(p = 1 - F_X(x)\right)\right)$, can prove a significant computational barrier in several programming languages (e.g., R), given that this operation has to be performed for any given value of x, for which we need to estimate $F_{M_k}(x)$. For instance, in R, the function `stats::toeplitz(x)`, where `x` is vector of size 10^4, requires about 5 sec (in a 2.2GHz quad-core Intel Core i7 processor), hence a remedy may be needed for such cases. Of course, a potential way around is to use a faster function for such operation (e.g., written in a low-level programming language, such as C/C++). However, herein we follow a different route (which of course can be combined with alternative/faster `toeplitz` functions) and propose an approximation that combines Eq. (19) and Eq. (23) into a single relationship (abbreviated as, AR(n)+βB). This model reads as follows (assuming $k \geq n$),

$$F_{M_k}(x) = \left(\frac{F_{\beta B}\left(0; n+1, \alpha_{\beta B}\left(F_X(x), \widetilde{\boldsymbol{R}}_{n+1}\right), \beta_{\beta B}\left(F_X(x), \widetilde{\boldsymbol{R}}_{n+1}\right)\right)}{F_{\beta B}\left(0; n, \alpha_{\beta B}\left(F_X(x), \widetilde{\boldsymbol{R}}_n\right), \beta_{\beta B}\left(F_X(x), \widetilde{\boldsymbol{R}}_n\right)\right)}\right)^{k-n} F_{\beta B}\left(0; n, \alpha_{\beta B}\left(F_X(x), \widetilde{\boldsymbol{R}}_n\right), \beta_{\beta B}\left(F_X(x), \widetilde{\boldsymbol{R}}_n\right)\right) \quad (27)$$

The above model essentially assumes that the distribution of maxima $F_{M_k}(x)$ of a process $X_t \sim F_X$, with ACS ρ_τ, can be approximated by that of a high order AR process, where instead of numerically estimating (via integration) high-order copula functions (as in Eq. (14)), we approximate them using the βB distribution. Also, the above formula reduces the dimension of the vector to which the Toeplitz operator is applied, since the maximum Toeplitz matrix involved has dimension $n+1$, where n denotes the order of the AR process. As a rule of thumb (heuristic), and after several numerical investigations (not shown herein), we suggest the use of n in the range of, $\lfloor k/3 \rfloor \leq n \leq \lfloor k/2 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the *floor* operator. The rationale behind this heuristic rule is that a value of n lying within the suggested range can account for a significant *portion* of the process' second-order properties (through the autocorrelation coefficient matrices of $\widetilde{\boldsymbol{R}}_{n+1}$ and $\widetilde{\boldsymbol{R}}_n$), and thus can provide a good approximation of the $F_{M_k}(x)$ distribution. Of course, the use of $n < k$ results to information loss, and thus when maximum precision is required, the use of the approach discussed in section 3.5 should be considered preferable. However, the proof-of-concept examples of section 4 illustrate that the loss of precision is limited, and the proposed approximation can yield results

almost identical to that of the method of section 3.5, even for processes with strong dependence structure (see section 4 for more details and results).

3.7 An algorithmic step-by-step recipe

The use of Eq. (15) (the case of independent processes) and Eq. (16) (the case of AR(1) processes), is rather straightforward, hence it is not further discussed herein. On the other hand, Eq. (18) is also omitted since its *raw* use is numerically unstable (for reasons explained above). Therefore, herein, we focus on the use of the procedures described in sections 3.5 and 3.6, both involving the use of the βB distribution, which is a relatively new development in the domain of hydrology, thus, deserving some extra attention to ensure the clear understanding of the underlying *mechanics*. With this in mind, and assuming that the marginal distribution F_X and ACS ρ_τ of the process are known, a step-by-step recipe is provided below for the estimation of the distribution of maxima M_k over a block of k time steps.

Step 1. Given the the marginal distribution F_X and the ACS ρ_τ, estimate the equivalent correlation coefficients $\tilde{\rho}_\tau$ up to $\tau = k-1$, by first establishing the corresponding TEC relationship, and then by inverting it, i.e., obtain $\tilde{\rho}_{1:k-1} = \mathcal{F}^{-1}(\rho_{1:k-1}|\, F_X)$.

Step 2. Create a vector $\boldsymbol{x} = [x_1, \dots, x_\nu]$, of length ν, that contains the values for which we need to estimate $F_{M_k}(x)$.

Step 3. For each $x_{i\,\in\,1,\dots,\nu}$, first estimate $\ddot{\boldsymbol{R}}_k\big(F_X(x_i), \widetilde{\boldsymbol{R}}_k\big) = \mathcal{F}_{\mathcal{B}e}\Big(\widetilde{\boldsymbol{R}}_k | F_Y \equiv \mathcal{B}e\big(p = 1 - F_X(x_i)\big)\Big)$, next the quantity $\ddot{\rho}_{IC}\big(F_X(x_i), \widetilde{\boldsymbol{R}}_k\big)$, and finally the parameters $\alpha_{\beta B}\big(F_X(x_i), \widetilde{\boldsymbol{R}}_k\big)$ and $\beta_{\beta B}\big(F_X(x_i), \widetilde{\boldsymbol{R}}_k\big)$ of the βB distribution.

Step 4. Finally, for each $x_{i\,\in\,1,\dots,\nu}$, estimate the distribution $F_{M_k}(x_i)$ using Eq. (23), i.e., $F_{M_k}(x_i) = F_{\boldsymbol{X}_k}(x_i; \boldsymbol{R}_k) \cong F_{\beta B}\big(0; k, \alpha_{\beta B}\big(F_X(x_i), \widetilde{\boldsymbol{R}}_k\big), \beta_{\beta B}\big(F_X(x_i), \widetilde{\boldsymbol{R}}_k\big)\big)$.

The above step-by-step recipe concerns the model of section 3.5, while the one of section 3.6 (i.e., AR(n)+βB model for F_{M_k}) is almost identical, with the main differences being on the use of Eq. (27) in step 4, as well as the estimation of $\ddot{\boldsymbol{R}}_k\big(F_X(x_i), \widetilde{\boldsymbol{R}}_k\big)$ and the related quantities at step 3, up to $k = n+1$, where n is the order of the selected approximating AR(n) process. Also, it is noted that the inverse cumulative distribution function (ICDF) of F_{M_k} does not attain an analytical form for the models of section 3 (apart from the i.i.d. case which is given by, $x_u = F_{M_X} = F_X^{-1}\big(u^{1/k}\big)$). Therefore, in order to estimate x for a given probability $u \in (0,1)$, one has to resort in solving the equation $F_{M_k}(x) = u$ for a fixed value of u, that is, $x_u = F_{M_k}^{-1}(u) = \inf\{x \in \mathbb{R}: u \le F_{M_k}(x)\}$. If the notion of return period, $T = 1/(1-u)$ is employed (typically expressed in years), which is the norm in hydrology, the above expression is re-written as, $x_T = F_{M_k}^{-1}\left(1 - \frac{1}{T}\right) = \inf\left\{x \in \mathbb{R}: 1 - \frac{1}{T} \le F_{M_k}(x)\right\}$ which denotes the return level x for a given T (hence the abbreviation x_T). Finally, and for further convenience, it is noted that the above *recipe*, as well as the algorithmic implementation of $F_{M_k}^{-1}$ are both implemented in the `bBextremes` R package. It is remarked that an algorithmic implementation of $F_{M_k}^{-1}$ would not be possible, in terms of reasonable computational time, without the use of the approximations introduced in section 2 and 3.6 (i.e., the closed-form formula of the TEC relationship for the case of $\mathcal{B}e(p)$ processes and the AR(n)+βB model, respectively). As

an example, it is mentioned that the required computational time (in R programming language, and in the same computer mentioned above) to estimate $F_{M_k}^{-1}$ for a single value of u, for the case of the βB model and for $k = 365$, is 2 sec, without the use of the above approximations. On the contrary, the computational time is reduced to 0.06 sec (~33 times faster) when employing the approximations introduced herein. It is noted that the computational effort required by the non βB-based approaches discussed in this section is not mentioned since it is negligible.

4 Simulation studies

4.1 Validation of the distribution of *k*-length block maxima models

Aiming to validate, as well as demonstrate the performance of the F_{M_k} models presented in section 3, we employ four (4) simulation studies, all of them concerning processes with zero-inflated ($\mathcal{ZI}$) marginal distributions. The simulated studies are detailed in **Table 1**. The selection of parameters of the ACS models and the marginal distributions was made on the basis of stress-testing the models of section 3 (i.e., under persistent ACSs – see appendix C for details about the employed ACS models). The selection of a $\mathcal{ZI}$ marginal distribution stems from the intermittent behaviour that characterizes hydrometeorological processes, such as rainfall, at fine time scales (e.g., daily or finer); see also Appendix A for a brief overview of a $\mathcal{ZI}$ distributions. In particular, we employed a $\mathcal{ZI}$ variant of the Burr type-XII distribution (Burr 1942; Singh and Maddala 1976; Tadikamalla 1980). The parameters of this distribution, abbreviated as $\mathcal{ZIB}rXII$ (see Appendix B for more details), are set equal to those found by Koutsoyiannis (2020), for the daily rainfall record at the station of Bologna, Italy (using a time series with more than 200 years of daily data). For each of the four simulation studies, we generated[1] synthetic time series with the target distribution and ACS, each of length $L = k \times 10\,000$, where $k = 365$. Mimicking this way, a realization of a daily process of 10 000 years (e.g., see panel (a) in **Figure 4** and **Figure 5**, where a randomly selected window of 365×30 time steps is depicted), and a distribution F_{M_k} of the annual block maxima of a daily rainfall process.

The analysis showed that similar results were obtained for all four cases, and due to this in the main text we focus on the first two cases, which concern persistent processes with fractional Gaussian noise (fGn) (e.g., Mandelbrot and Van Ness 1968; Beran 1992; Koutsoyiannis 2010), with H parameter, controlling the *degree of persistence* of the process, equal to $H = 0.65$ and $H = 0.8$, respectively (see Appendix C for details). The results of analysis for the other two simulation studies are presented in Appendix D (specifically in **Figure D1**, and **Figure D2**).

The results of case studies #1 and #2 are visualized in **Figure 4** and **Figure 5**, respectively. In these plots, panels (b) and (c) validate the ability of the `anySim` R package (Tsoukalas et al. 2020) to generate realizations with the target marginal and ACS, while panel (d) depicts the series of k-length block maxima obtained using $k = 365$ as a block size. Panel (e) provides a comparison between the empirical distribution of k-

[1] The simulations were performed using the `anySim` R package (Tsoukalas et al. 2020), which is capable of simulating Nataf-based (i.e., Gaussian copula) processes with any marginal distribution (with finite variance) and (valid; i.e., positive definite) correlation structure (temporal, spatial or combination of them).

length block maxima (orange dots) and the theoretical models of F_{M_k} of section 3. The comparison is conducted on the basis of return period, $T = 1/(1 - F_{M_k})$ (note that since we set $k = 365$, T is interpreted as referring to years; assuming that X_t refers to a daily process). It is interesting to observe that, despite the use of particularly *strong* ACSs (especially simulation studies #2 and #3; **Figure 5** and **Figure D1**), all F_{M_k} models converge to about the same return levels, x, for $T > 10$. This observation is also in accordance with tail-independence property that characterizes the Gaussian copula (Embrechts et al. 1999; Beirlant et al. 2006; Davis and Mikosch 2009). Also, it is observed that apart from the i.i.d. and AR(1) models of F_{M_k}, that do not reproduce the distribution of F_{M_k} with high accuracy for $T < 10$, all other models, that is, the Gaussian copula (Eq. (14)), the βB (Eq. (23)), and the AR(n)+βB (Eq. (27)), are almost indistinguishable for all T. This observation also validates the performance of the proposed AR(n)+βB model, which is also the faster among the three. In particular, for the estimation of the return period of fifty (50) return levels the AR(n)+βB, the βB and the Gaussian copula models require about 0.22 sec, 5.5 sec, and 145 sec respectively.

4.2 The GEV or not the GEV?

Naturally, one may wonder what kind of approach should we eventually use to model the distribution of block maxima. A thought that leads to the following questions:

- How does the reviewed non-asymptotic models compare to the iconic asymptotic approach that implies the use of the Generalized Extreme Value ($\mathcal{GEV}$) over the subset of block maxima values?
- How does the reviewed non-asymptotic models compare to the use of a non-extreme value distribution (any other but GEV or GPD) directly to the subset of block maxima?
- Which approach is more robust (i.e., has lower uncertainty or variability), and thus allows for better identifiability of the probabilistic behaviour of the maxima?
- Using which modelling approach (asymptotic or non-asymptotic), we have better chances to identify the *true* distribution of k-length block maxima?

Aiming to provide answers to the above questions, we setup a computational experiment to compare different approaches for the estimation of the distribution of block maxima. In particular, we formulate a Monte-Carlo experiment composed by the following steps:

1. We define a process with a given marginal distribution and ACS. In this case, we employ the intermittent process with moderate fGn ACS (e.g., rainfall) of simulation study #1, i.e., $\mathcal{ZIB}r XII(p_o = 0.75, \beta = 7.07, \zeta = 0.928, \xi = 0.098)$ and $\rho_\tau = \rho_\tau^{fGn}(H = 0.65)$. It is remarked that the distribution parameters resemble those identified by Koutsoyiannis (2020)) for the daily rainfall recorded at the station of Bologna, Italy, while the parameter of the ACS is set a bit higher than the average value (H~0.58) identified by the large-scale analyses of Kantelhardt et al. (2006) and Iliopoulou et al. (2016) for rainfall records.
2. We generate (using `anySim` R package) a large number of realizations from the above process, with length $L = k \times m$. In this case, we produced 1000 realizations, and by setting $k = 365$ and $m = \{50, 100\}$, we mimic daily records of 50 and 100 years (records with typical and rich length, which are nowadays become more and more available).

3. We perform a set of diagnostics and analysis on the basis of the produced time series ensemble.

The approaches that are examined and compared herein are:

I. The non-asymptotic distribution of block maxima (in particular AR(n) + βB for fixed k), identified using the complete record of observations. The employed approach entails the fitting of the $\mathcal{ZIB}rXII$ to each one of the realizations, as well as the fitting of the Cauchy-type correlation structure (CAS). The fitting of the distribution was done using the method of L-moments[2] (Hosking 1990), while the empirical probability of zero is estimated as the count of zero values to the total length of the series. On the other hand, CAS was fitted by minimizing the sum of least square error between the CAS model (see Appendix C) and the empirical autocorrelation coefficients. It is remarked that the CAS is not the *true* ACS of the process (i.e., the fGn with $H = 0.65$ which was used for the generation of the timeseries), and has been selected herein due to its increased flexibility to describe a wide range of ACSs.

II. The asymptotic distribution of block maxima, i.e., the $\mathcal{GEV}$ model. This entails the fitting of the $\mathcal{GEV}$ only on the record of block maxima (for each one of the realizations). The fitting was also performed using the L-moments method[2].

III. The use of a non-extreme value distribution model fitted directly to the record of block maxima (see, Chen and Singh 2018). In particular, we fitted (through L-moments) the $\mathcal{B}rXII$ distribution on the block maxima record of each one of the realizations; a distribution often employed in hydrology for such purposes (e.g., Shao et al. 2004; Hao and Singh 2009; Usta 2013).

Hereafter, the three above approaches are abbreviated as, ZIBrXII+AR(n)+βB, GEV+BM, and BrXII+BM, respectively.

Since the characteristics of the process are *a priori* known we can estimate (using the AR(n) + βB model described in section 3.6) the *true* distribution of k-length block maxima (since $k = 365$, we resemble annual daily maxima) of the process. We then use this distribution, as well as the generated realizations (see point #2 above), to compare the performance of the three alternative approaches on the basis of the following diagnostics:

1. Based on the time series ensemble compare the estimates of the expected (mean) return level x, as well as estimates for selected confidence levels (here, 10% and 90%).
2. For each alternative approach and each realization estimate the absolute relative difference between the obtained empirical level $\hat{x}_T$ and the theoretical one, x_T (which is known *a priori*), i.e., $\epsilon_T = |(\hat{x}_T - x_T)/x_T|$. Furthermore, on the basis of these estimates construct, and compare the empirical CDF of obtaining an error lower than ϵ_T for each approach.

[2] The distribution fitting was performed using the `lmom` R package (Hosking 2019). Particularly, for the $\mathcal{B}rXII$ distribution we employed the `pelp` function which, as described, provides "*Parameter estimation for a general distribution by the method of Lmoments*". It is noted that for computational efficiency, and optimization-related reasons, in the case of $\mathcal{B}rXII$ distribution (as well as for other models comprised by a scale and multiple shape parameters) the argument type should be set to "`s`" (for more details see p. 37 of the package manual).

3. Compare the mean and standard deviation values of the absolute relative difference of ϵ_T (estimated using all 1000 realizations) between the theoretical and alternative approaches for a range of return periods T.
4. Compare the asymptotic behaviour of the distribution of maxima as identified by the three approaches. It is noted that since we employ the $\mathcal{B}rXII$ distribution for the continuous part of the base process, the distribution of the maxima should exhibit the same asymptotic behaviour dictated by the tail index ξ (=0.098 in our case), see also Koutsoyiannis (2020). Also, since in our experiment we use the $\mathcal{GEV}$ and the $\mathcal{B}rXII$ distributions (fitted either to block maxima or the complete record), that share the same asymptotic behaviour ($\sim x^{-1/\xi}$), they should in theory converge to the same value of ξ.

Figure 6 to **Figure 9** synopsize the diagnostics #1 to #4, respectively. It is remarked that when performing the diagnostics #2 and #3, we included in our analysis the distribution of k-length block maxima of the i.i.d. model of Eq. (15), aiming to assess the effect of neglecting the autocorrelation for the estimation of the distribution of maxima. This modelling approach is abbreviated as ZIBrXII+iid.

Starting by **Figure 6**, that presents the comparison between the distribution of maxima obtained by the three approaches, it is observed that the mean (coloured solid lines) of the return levels obtained by the ZIBrXII+AR(n)+βB model and the GEV+BM model are both very close to the true values, while the mean of the BrXII+BM approach implies significantly larger return levels. On the other hand, the 80% confidence intervals (similarly coloured shaded areas) of the ZIBrXII+AR(n)+βB model are much narrower than those the other two models; with those of BrXII+BM approach being the largest of all. The above observations hold for both $m = 50$ and $m = 100$, and are being further confirmed by **Figure 7**, where we compare the absolute relative mean square error ϵ_T for the return periods $T \in \{10, 50, 100, 1000\}$. The CDFs of ϵ_T highlight that for all examined T, the probability of not exceeding a specified value of error ϵ_T is higher when using the ZIBrXII+AR(n)+βB model, implying that it has better chances to obtain the minimum ϵ_T among all methods under comparison (the same applies also for the ZIBrXII+iid model). On the other hand, the BrXII+BM gives much lower estimates of probability of not exceeding ϵ_T, especially for $T \geq 50$, implying that there is a significant chance to exceed that threshold value.

The previous observation is also validated in **Figure 8** where we plot the mean and standard deviation of ϵ_T (estimated on the basis of all 1000 realizations). In this figure, it is clear that the non-asymptotic models of BM, i.e., the ZIBrXII+AR(n)+βB and ZIBrXII+iid, are those with the lower values of mean and standard deviation of ϵ_T for $T \geq 10$. Their estimates differ for lower values of T, since the ZIBrXII+iid is linked with the assumption of temporal independence (a characteristic that influences particularly small T). This observation highlights the importance of identifying a good distribution model for the base process in order to devise a good model for the maxima corresponding to $T \geq 10$. Furthermore, when comparing the two other models, i.e., the GEV+BM and the BrXII+BM, it can be argued that the former shows better, and more stable performance since it exhibits lower values for both the mean and standard deviation of ϵ_T.

Moving to the final diagnostic, i.e., comparison of the obtained asymptotic behaviour of the three approaches, we depict in **Figure 9** the empirical CDF of the estimates of the tail indices (parameter ξ, in both $\mathcal{B}rXII$ and $\mathcal{GEV}$ distributions), obtained by the three methods (the true value, $\xi = 0.098$, is depicted with an orange vertical line). This plot indicates clearly that for both $m = 50$ and $m = 100$, the ZIBrXII+AR(n)+βB model is the one with smaller variability of ξ estimates, ranging in a much narrower spectrum compared to the other two approaches.

On the other hand, the GEV+BM approach exhibits a tendency to negative values of ξ ($\mathrm{P}\{\hat{\xi} \leq 0\} \cong 0.25$ for $m = 50$, and $\mathrm{P}\{\hat{\xi} \leq 0\} \cong 0.16$ for $m = 100$), which is not in harmony with the expected tail behaviour. This tendency may also provide a possible explanation for the increased variability exhibited by the GEV+BM approach in **Figure 9** (and similar plots in SM) which could be attributed to the fact that the $\mathcal{GEV}$'s shape parameter ξ (i.e., tail index) can admit values in $(-\infty, \infty)$, thus includes the case of reversed Weibull distribution for $\xi < 0$. It is recalled that the reversed Weibull is an upper-bounded distribution with finite upper support (see Eq. (B8) in Appendix B), an assumption that is considered unrealistic for most hydroclimatic processes (e.g., precipitation and runoff processes are considered physically bounded from below, i.e., at zero, and unbounded from above); see for instance Koutsoyiannis (2004a).

Moving to the case of BrXII+BM approach, it is observed that the obtained estimates of ξ exhibit a significant shift from the true value, since the median estimates are $\hat{\xi}_{0.5} \cong 0.17$ and $\hat{\xi}_{0.5} \cong 0.22$ for $m = 50$ and $m = 100$ respectively (all other percentiles are shifted accordingly). This shift towards larger values of ξ implies a distribution of maxima with much heavier right tail than the *a priori* defined distribution, and thus leading to misspecification of the process's asymptotic behaviour. To the author's view, this may be due to the fact that the $\mathcal{B}rXII$ distribution is a particularly flexible model (since it has two shape parameters) that can take a lot of forms. This flexibility may be also the root of this problem since the fitting of the $\mathcal{B}rXII$ distribution to the BM subset (i.e., though matching the empirical L-moments with their theoretical counterparts), may provide spurious parameter estimates, just to fit the data. After all, it is important to recall that distribution functions are models, and thus can be often *wrong* or over-parameterized. Also, fitting a distribution model to a set of data is nothing more than solving an optimization problem (e.g., minimizing the difference between empirical and theoretical quantities, such as moments and quantiles, or alternatively maximizing the likelihood function), which may result into suboptimal or non-informative solutions (i.e., local optima). A potential remedy to this issue, which may be worth exploring (left out for future research), would be the identification of the tail index of the $\mathcal{B}rXII$ distribution on the basis of the complete record (e.g., using L-moments), and then estimating the other two parameters using the BM subset (of course a similar approach could be followed using other distribution models parameterized using a tail index parameter, e.g., the Dagum distribution, which is the Burr III model after the introduction of an additional scale parameter; see Kleiber (2008)). Formulating this way, a multi-level distribution fitting method for extremes.

The above analysis and diagnostics were performed for all four simulation studies of **Table 1**, yielding similar results, hence not included herein. As a proof of concept, we included in Appendix D, only plots similar to the **Figure 6** (**Figure D3** to **Figure D5**). Furthermore, we a provide a Supplementary Material (SM) document which contains four (4) additional simulation studies complementing those presented in this section (see **Figure S1** to **Figure S15** of SM). All of them concern a base/parent process with fractional Gaussian noise (fGn) autocorrelation structure (ACS) with $H = 0.65$ and differ at the marginal distribution of the process, which is continuous or zero-inflated (see **Table S1** of SM for details). The results presented in the SM are similar and alignment with those presented in the main manuscript.

All in all, the above diagnostics (including those of SM) seems to agree, and highlight that estimating the distribution of k-length block maxima using non-asymptotic distribution models results into quite stable and less variable estimates of return levels (compared to the other two approaches examined). Highlighting this way the suitability of the models of section 3, such as the AR(n)+βB, and the original βB model of Serinaldi et al. (2020) for use in real-world applications. On the other hand, the competence and utility of the $\mathcal{GEV}$ model is further remarked under the premise of limited data availability (i.e., assuming that only the subset of BM is available), since it exhibits less variability and more consistent behaviour compared to the approach that fits a non-extreme value distribution (i.e., the Burr type-XII) directly to the record of BM.

Further to the above, it is remarked that the performance of non-asymptotic models of maxima could be further improved by using alternative and more robust distribution fitting methods (beyond the L-moments method used herein) and more advanced methods (than the simple least square error minimization used herein) for ACS identification (e.g., see, Koutsoyiannis 2020). Finally, an incidental advantage of using a non-asymptotic distribution model is that the supports of the resulting BM distribution are in agreement with those of the distribution employed for the base process (e.g., if the distribution of the base process is defined in $[0, \infty)$, then the resulting non-asymptotic BM distribution will have the same supports). A property that is also in-line with the nature of several hydrological variables defined in $[0, \infty)$, e.g., rainfall and streamflow. Recall that the supports of the $\mathcal{GEV}$ distribution depend on the value of tail index parameter (see Appendix B), and can result into bounded (from below or above) or completely unbounded distributions.

5 The probability of a zero value over multiple temporal scales

5.1 Theoretical background and key concepts

Another particularly interesting use of the F_{M_k} models presented in section 3, especially of those based on the βB distribution, concerns processes characterized by intermittent behaviour (i.e., modelled using $\mathcal{ZI}$ marginal distributions). In such case, these models can be used to estimate the probability of observing a zero value over multiple scales of (e.g., temporal) aggregation d. For simplicity, hereafter called probability of zero at scale d, and denoted by $p_0^{(d)}$. While not so apparent at first sight, and as detailed later on (see section 5.4), this quantity can be of particular use when modelling the behaviour of

extremes (i.e., maxima). It is noted that in the hydrological domain, $p_0^{(d)}$ is also called probability dry at scale d, denoting the probability of a dry interval at scale d.

However, before getting into the implementation details, let us first provide some necessary context. Let $\left\{X_t^{(1)}\right\}_{t\in\mathbb{Z}^>}$ or simply $\{X_t\}$ be an intermittent discrete-time stationary stochastic process at some time scale $d = 1$.

Let us also define the averaged aggregated process $X_l^{(d)}$ at a higher time scale $d = 2,3, \ldots$, obtained by:

$$X_l^{(d)} = \frac{1}{d}\sum_{t=(l-1)d+1}^{dl} X_t \tag{28}$$

where l is the *new* time index of the averaged aggregated process. Note that while herein we focus on the averaged aggregated process, the same rationale can be applied for a non-averaged aggregated process, obtained as $\check{X}_l^{(d)} = \sum_{t=(l-1)d+1}^{dl} X_t$.

Further, let also the process be characterized at all scales d by:

- a $\mathcal{ZI}$ marginal distribution $F_{X^{(d)}}(x) = \mathrm{P}\{X^{(d)} \le x\} = p_0^{(d)} + \left(1 - p_0^{(d)}\right)G_{X^{(d)}}(x)$, where $G_{X^{(d)}} = F_{X^{(d)}|X^{(d)}>0} = \mathrm{P}\{X^{(d)} \le x|X^{(d)} > 0\}$. For simplicity $F_X(x) = F_{X^{(1)}}(x) = \mathrm{P}\{X \le x\} = p_0 + (1 - p_0)G_X(x)$, where, $p_0 = p_0^{(1)} = \mathrm{P}\{X = 0\}$ is a parameter controlling the inflation of zeros (i.e., the discrete part of the $\mathcal{ZI}$ distribution – the probability of observing a zero value), and $G_X = G_{X^{(1)}} = F_{X^{(1)}|X^{(1)}>0} = \mathrm{P}\{X^{(1)} \le x|X^{(1)} > 0\}$ denotes the distribution to be inflated (i.e., the continuous part of the $\mathcal{ZI}$ distribution, herein assumed to be defined in the positive half line $(0, \infty)$). For completeness and subsequent use, it is noted that $p_1^{(d)} \coloneqq \mathrm{P}\{X^{(d)} > 0\} = 1 - p_0^{(d)}$, while for simplicity let $p_1 = p_1^{(1)}$. See Appendix A for more details.
- an autocorrelation structure $\rho_\tau^{(d)} = \mathrm{Corr}\left[X_t^{(d)}, X_{t+\tau}^{(d)}\right]$, $\tau > 0$, for simplicity $\rho_\tau = \rho_\tau^{(1)}$.

The mean and variance of the process at scale $d = 1$, are $\mu = \mu^{(1)} = \mathrm{E}\left[X_t^{(1)}\right]$, and $\gamma = \gamma^{(1)} = \sigma^2 = \sigma^{2(1)} = \mathrm{Var}\left[X_t^{(1)}\right]$. The mean of positive values, i.e., $\mu_p = \mathrm{E}[X|X > 0\,]$ can be obtained by, $\mu_p = p_1\mu$, and the variance of positive values, i.e., $\gamma_p = \mathrm{Var}[X|X > 0\,] = (\gamma p_1 + \mu^2 p_1 - \mu^2)/p_1^2$. Similarly, the autocovariance structure is, $c_\tau = c_\tau^{(1)} = \mathrm{Cov}\left[X_t^{(1)}, X_{t+\tau}^{(1)}\right] = \sigma^2\rho_\tau$, while $c_0 = \gamma = \sigma^2$.

The properties of process at scale $d = 1$ are related with those of the averaged aggregated process at higher time scale $d = 2,3, \ldots$ In particular, it is straightforward to obtain the mean and autocovariance by (Koutsoyiannis 2005),

$$\mu^{(d)} = \mu = \mathrm{E}\left[X_t^{(d)}\right] \tag{29}$$

$$c_\tau^{(d)} = \mathrm{Cov}\left[X_l^{(d)}, X_{l+\tau}^{(d)}\right] = \frac{1}{d}\sum_{i=1-d}^{d-1} c_{\tau d+i}\left(1-\frac{|i|}{d}\right) = \frac{1}{d^2}\sum_{t=1}^{d}\sum_{r=\tau d+1}^{(1+\tau)d} c_{|t-r|} \tag{30}$$

The variance of the averaged aggregated process at scale d, i.e., $\gamma^{(d)}$, is $\gamma^{(d)} = c_o^{(d)}$. Also, note that $\sigma^2 = c_0 = \gamma = \gamma^{(1)}$. The autocorrelation of the averaged aggregated process at scale d is, $\rho_\tau^{(d)} = c_\tau^{(d)}/c_o^{(d)}$.

Arguably, the above quantities are not enough to characterize the behaviour of an intermittent processes, since no information is provided about the degree of intermittency at scales of aggregation $d > 1$. This is also apparent by the need of $p_0^{(d)}$ to estimate the mean and the variance of positive values at scale d. The formulas for the estimation of these quantities at scales of aggregation $d > 1$ are given by,

$$\mu_p^{(d)} = \frac{\mu}{p_1^{(d)}} \tag{31}$$

$$\gamma_p^{(d)} = \frac{\gamma^{(d)} p_1^{(d)} + \mu^2 p_1^{(d)} - \mu^2}{\left(p_1^{(d)}\right)^2} \tag{32}$$

Solving the above equation for $\gamma^{(d)}$, yields

$$\gamma^{(d)} = \frac{\mu^2 - \mu^2 p_1^{(d)} + \gamma_p^{(d)}\left(p_1^{(d)}\right)^2}{p_1^{(d)}} \tag{33}$$

which highlights the link between the variance of the whole process (including zeros) with the mean and the variance of positive values, as well as with probability of zero at scale d. Of course, when $p_1^{(d)} = 1$ (i.e., $p_0^{(d)} = 0$), $\gamma_p^{(d)}$ and $\gamma^{(d)}$ are identical quantities.

To estimate $p_0^{(d)}$ let us recall that by definition, $p_0^{(d)} = \mathrm{P}\{X_1 = 0, \dots, X_d = 0\}$, which expresses the probability of having d consecutive zeros in an equally d-sized block at the base time scale $d = 1$. Thus, this probability is identical to Eq. (14), for $x = 0$, hence,

$$\begin{aligned} p_0^{(d)} = \mathrm{P}\left\{X^{(d)} = 0\right\} &= \mathrm{P}\{X_1 \le 0, \dots, X_d \le 0\} = F_{\boldsymbol{X}_d}(0; \boldsymbol{R}_d) \\ &= \Phi_k\left(\Phi^{-1}\left(F_X(0)\right), \dots, \Phi^{-1}\left(F_X(0)\right); \widetilde{\boldsymbol{R}}_d\right) \end{aligned} \tag{34}$$

which in turn allows us to use the models of section 3 to estimate $p_0^{(d)}$. Yet, this time with fixed x (equal to 0) and varying $d(= k)$, denoting the scale at which we wish to estimate $p_0^{(d)}$. Noting also that for the $\mathcal{ZI}$ distribution, $F_X(0) = p_0$. In particular, the βB and AR(n)+βB of section 3.5 and 3.6 respectively, read as follows,

$$p_0^{(d)} = F_{M_d}(0) = F_{\boldsymbol{X}_d}(0;\boldsymbol{R}_d) = \Phi_d\left(\Phi^{-1}\left(F_X(0)\right),\dots,\Phi^{-1}\left(F_X(0)\right);\widetilde{\boldsymbol{R}}_d\right) \cong F_{\beta B}\left(0;k,\alpha_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_d\right),\beta_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_d\right)\right) \quad (35)$$

$$p_0^{(d)} = F_{M_d}(0) = \left(\frac{F_{\beta B}\left(0;n+1,\alpha_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_{n+1}\right),\beta_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_{n+1}\right)\right)}{F_{\beta B}\left(0;n,\alpha_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_n\right),\beta_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_n\right)\right)}\right)^{\min(d,d-n)} F_{\beta B}\left(0;n,\alpha_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_n\right),\beta_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_n\right)\right) \quad (36)$$

Notice that apart from fixing $x = 0$, and varying d, depending on which scale one wishes to estimate $p_0^{(d)}$, the only modification concerns Eq. (36), where the exponent of the first part of the formula now becomes, $\min(d, d-n)$, instead of the original value of $d-n$ (recall that in this case, $d = k$). This is due to the fact that there is no point of using an F_{M_d} model of an AR(n) process when the target scale d is smaller than the order of the AR model, i.e., n.

5.2 An algorithmic step-by-step recipe

Step 1. Given the the marginal distribution F_X and the ACS ρ_τ, estimate the equivalent correlation coefficients $\tilde{\rho}_\tau$ up to $\tau = d_\nu - 1$, where d_ν is the maximum scale we are interested in (see step #2), by first establishing the corresponding TEC relationship, and then by inverting it, i.e., obtain $\tilde{\rho}_{1:d_\nu-1} = \mathcal{F}^{-1}\left(\rho_{1:d_\nu-1}|\, F_X\right)$.

Step 2. Create a vector $\boldsymbol{d} = [d_1,\dots,d_\nu]$, of length ν, that contains the scales d for which we wish to estimate p_0^d (in accenting order).

Step 3. For each $d_{i\in 1,\dots,\nu}$, first estimate $\ddot{\boldsymbol{R}}_{d_i}\left(F_X(0),\widetilde{\boldsymbol{R}}_{d_i}\right) = \mathcal{F}_{\mathcal{B}e}\left(\widetilde{\boldsymbol{R}}_d|F_Y \equiv \mathcal{B}e\left(p = 1 - F_X(0)\right)\right)$, next the quantity $\ddot{\rho}_{IC}\left(F_X(0),\widetilde{\boldsymbol{R}}_{d_i}\right)$, and finally the parameters $\alpha_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_{d_i}\right)$ and $\beta_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_{d_i}\right)$ of the βB distribution.

Step 4. Finally, for each $d_{i\in 1,\dots,\nu}$, estimate the quantity $p_0^{d_i} = F_{M_{d_i}}(0)$ using Eq. (35), i.e., $p_0^{d_i} = F_{M_{d_i}}(0) = F_{\boldsymbol{X}_{d_i}}(0;\boldsymbol{R}_d) \cong F_{\beta B}\left(0;d,\alpha_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_{d_i}\right),\beta_{\beta B}\left(F_X(0),\widetilde{\boldsymbol{R}}_{d_i}\right)\right)$.

As in the case of the recipe of the distribution of block maxima, F_{M_k}, given in section 3.7, the above recipe concerns the use of the βB model, while the recipe for implementing AR(n)+βB model for the estimation of $p_0^{(d)}$ is similar.

Finally, it is noted that instead of $x = 0$, alternative, low, threshold values can be used, e.g., for the case of rainfall to account for measurement accuracy of rain gauges (e.g., not logging rainfall below 0.1 mm). This recipe is also implemented in the `bExtremes` R package.

5.3 Proof of concept by simulation

To illustrate the proposed methods for $p_0^{(d)}$ estimation, we employ the first two simulation studies of **Table 1**. It is reminded that both processes concern a $\mathcal{ZJBr}XII$ distribution with $p_o = 0.75, \beta = 7.07, \zeta = 0.928, \xi = 0.098$ and differ in the ACS. The 1st

process is modelled using $\rho_\tau^{fGn}(H = 0.65)$, while the second $\rho_\tau^{fGn}(H = 0.80)$. The results are visually synopsized in **Figure 10**, where panels (a) and (b) provide respectively for simulation studies #1 and #2, a comparison between the empirical (as obtained by simulation), and theoretical (as obtained by the formulas of section 5.1) probability of a zero value over multiple temporal scales (i.e., $p_0^{(d)} \coloneqq \mathrm{P}\{X^{(d)} = 0\}$) $d \in \{2, \dots, 365\}$. Mimicking this way, an intermittent process with daily time step. It is noted that the results are similar for the other simulation studies (including #3 and #4 of **Table 1,** hence not presented herein).

As in the case of the distribution of k-length block maxima, the βB and AR(n)+βB models prove to be capable of establishing the multiscale behaviour of $p_0^{(d)}$, since they are in absolute agreement with the values obtained by both the simulation, as well as the Gaussian copula (obtained by numerical integration).

5.4 Further considerations and potential applications

As hinted earlier, the multiscale description of the probability of observing a zero value, can be of particular use within the context of modelling extreme values, such as k-length block maxima.

With rainfall intensity processes in mind, $p_0^{(d)}$, can be used in conjunction with $\mu^{(d)} = \mu$, and ρ_τ (hence $\gamma^{(d)}$; since they are interrelated quantities; see Eq. (30)) in order to devise parsimonious multiscale probabilistic models for rainfall. For instance, this information solely suffices to fit a multi-scale ZI distribution; where $p_0^{(d)}$ determines the degree of zero-inflation at scale d, while the continuous part (i.e., the positive values) can be modelled using a two-parameter marginal distribution whose parameters can be directly obtained by the classical methods of moments (since $\mu_p^{(d)}$ and $\gamma_p^{(d)}$ are known for all scales d; see Eq. (31) and Eq. (32)). Further to this, it is noted that it is possible to use three-parameter marginals, using for instance a property that characterizes distributions with Pareto-type tails (such as the Burr type XII (Burr 1942; Singh and Maddala 1976; Tadikamalla 1980) or the Dagum (Dagum 1977; Kleiber 2008)). This property implies that the tail index remains constant across multiple scales of aggregation d (Koutsoyiannis 2020). The total number of parameters to construct such a multiscale distribution model, i.e., $F_{X^{(d)}}(x) = \mathrm{P}\{X^{(d)} \leq x\} = p_0^{(d)} + \left(1 - p_0^{(d)}\right) G_{X^{(d)}}(x)$, where $G_{X^{(d)}} = \mathrm{P}\{X^{(d)} \leq x | X^{(d)} > 0\}$ is the distribution of positive quantities, is equal to the number of parameters of the base process's marginal distribution and ACS. For instance, if a three-parameter marginal is selected and a two-parameter ACS, the total number of parameters is five (5). Regardless the number of parameters, once $F_{X^{(d)}}$ has been determined, it is also straightforward to turn this information into an intensity-duration-frequency (IDF) model; which are models widely used in hydrology for the design of hydraulic engineering works that need to be resilient under extremes. For a thorough overview on the topic of IDF models see the works of Koutsoyiannis et al. (1998) and Koutsoyiannis (2020), as well as the recent work of Courty et al. (2019), who aim at providing a global IDF model.

For instance, assuming that a procedure similar to the one described above has been followed, thus $F_{X^{(d)}}$ and $\rho_\tau^{(d)}$ could be determined, a rainfall IDF model can be constructed with the help of the formulas of section 3. To provide an example, let us

assume that we work with data expressed in terms of hourly intensities (which is the norm for IDF models), and we used as a basic time scale (i.e., $d = 1$), the one corresponding to duration $D^* = 0.25$ h (i.e., 15 min). Let us assume also that it was decided to use the i.i.d. model of Eq. (15) for the distribution of k-length block maxima (i.e., F_{M_k}). Under these assumptions, the return period, T [years] of rainfall intensity x [mm/h] and duration $D = dD^*$ [h], where $d = D/D^* \in \mathbb{N}$, will read as follows,

$$T(x,d) = \frac{1}{1 - F_{M_k}(x; p_0^{(d)}, \boldsymbol{\theta}^{(d)})} = \frac{1}{1 - \left(F_{X^{(d)}}\left(x; p_0^{(d)}, \boldsymbol{\theta}^{(d)}\right)\right)^{k(d)}} = \frac{1}{1 - \left(p_0^{(d)} + \left(1 - p_0^{(d)}\right) G_{X^{(d)}}(x; \boldsymbol{\theta}^{(d)})\right)^{k(d)}} \quad (37)$$

where, $k(d) = k^*/d$ denotes the block size (number of time steps) that corresponds to 1 year at time scale d, and k^* stands for the block size of 1 year at the basic time scale (i.e., for $d = 1$ and D^*), also identical to the number of hours of 1 year divided by D^*. In our working example (since $D^* = 0.25$ h) $k^* = 4 \times 24 \times 365$, hence $k(d) = 4 \times 24 \times 365/d$ (e.g., when $d = 4$, i.e., $D = 1$ [h], $k(d) = 8760$, and when $d = 96$, i.e., $D = 24$ [h], $k(d) = 365$). Also note that $\boldsymbol{\theta}^{(d)}$ is a vector containing the distribution's parameters at scale d.

In addition, the inverse relationship that expresses rainfall intensity x [mm/h] as a function of return period T [years] and duration $D = dD^*$ [h], where $d = D/D^* \in \mathbb{N}$, reads as follows,

$$x(T,d) = F_{X^{(d)}}^{-1}\left(\left(1 - \frac{1}{T}\right)^{1/k(d)}; p_0^{(d)}, \boldsymbol{\theta}^{(d)}\right) = G_{X^{(d)}}^{-1}\left(\frac{\left(1 - \frac{1}{T}\right)^{1/k(d)} - p_0^{(d)}}{1 - p_0^{(d)}}; \boldsymbol{\theta}^{(d)}\right) \quad (38)$$

Similar, yet somewhat more difficult to handle, IDF models can be constructed using the other models for the distribution of k-length block maxima presented in section 3 (accounting also for the effect of the process's autocorrelation in F_{M_k},).

However, and without having performed a thorough analysis, it may be argued that the above simple model can prove to be sufficient for most hydrologic engineering studies, since,

1) The ACS of daily rainfall, when modelled using the fGn ACS, has been identified to exhibit an average value of $H \sim 0.58$ (e.g., see the large-scale analyses of Kantelhardt et al. (2006) and Iliopoulou et al. (2016)); which implies a rather weak ACS compared to those used herein ($H = 0.65$ and $H = 0.80$) to demonstrate the performance of F_{M_k} models.
2) The investigated models for the distribution of k-length block maxima converge to the same return level x for return periods $T >$ 10-20 years (at least for processes ACS similar or weaker than those explored herein, e.g., fGn with $H = 0.8$).

3) Typical hydraulic works are designed for $T \gg$ 10-20 years (e.g., the standard for reservoir spillways is $T = 10\,000$ years).
4) It is a relatively simple and fast model, an important aspect for real-world engineering purposes.
5) The effect of autocorrelation is not completely neglected, since it is used for the estimation of $p_0^{(d)}$, and thus $F_{X^{(d)}}$.
6) The i.i.d. model of F_{M_k} provides larger return levels x for the same T, compared to those accounting for the autocorrelation of the base process. Thus, lies on the safe side from the perspective of hydrological engineering and hydraulic infrastructures design.

We remark that the above procedure is not illustrated in this work, hence the above arguments should not be taken for granted, rather verified or falsified, using extensive analysis and simulations (left out for a sequel work). However, it is noted, that a similar, yet not identical, approach to the one sketched above, has been recently presented by Koutsoyiannis (2020), for the construction of a seven-parameter IDF model (also called ombrian model), achieving a very good accuracy between observations. With the main differences being that this work, 1) uses an alternative definition for return period, linked with the so-called complete time series analysis (CTA) of extremes (see also section 3 in Serinaldi et al. (2020) where the delicate difference between the one used herein is highlighted), as well as 2) uses parametric functions (Koutsoyiannis 2020) to establish the scaling laws implied by $p_0^{(d)}$ and $\gamma^{(d)}$, using notions similar to the one of lower scale extrapolation (Kossieris 2020). Of course, various elements of these approaches may be combined into different ways, formulating alternative IDF models, whose performance needs to be further verified under different real-world situations (e.g., using long, fine time scale, quality-controlled historical rainfall datasets comprised by records from all around the world).

6 Summary and concluding remarks

The distribution of *k*-length block maxima is arguably of particular interest for the design and management of hydraulic infrastructures, thus consists a valuable *tool* for hydrological design (i.e., estimate the design rainfall of a hydraulic work for a given return period, *T*) and flood risk assessment and security. Until very recently the focus was almost exclusively given on approaches that utilize the notions of temporal independence, limiting laws (e.g., for the block size), subsets of data (e.g., BM or POT) and asymptotic properties (convergence to the $\mathcal{GEV}$ or generalized Pareto distribution; $\mathcal{GPD}$). While these assumptions are convenient in terms of data storage, data management and computations they may *hide* important information about the probabilistic behaviour of extremes.

This work provides an alternative view, by focusing on the non-asymptotic distribution of the *k*-length block maxima that arise from a base/parent, discrete-time, stationary, non-Gaussian autocorrelated processes under the assumption of Nataf's joint distribution (i.e., Gaussian copula). Assuming that k is constant, and in view of the Gaussian copula, we reviewed five non-asymptotic distribution models for the distribution of *k*-length block maxima. In particular, we discussed models based on the assumptions that the base process is characterized by, independence (i.i.d.), AR(1) and AR(n) autocorrelation structure (ACS), as well as detailed two alternative models that

rely on the Beta-Binomial (βB) distribution, i.e., the βB-based model of Serinaldi et al. (2020) and the newly proposed AR(n)+βB model, which can be used as an alternative and fast approximation (~33 times faster than the original βB-based model) for processes with general ACSs (as the original βB-based model). In addition, throughout the manuscript special effort was given to highlight the links between the various *types* of correlation coefficients (i.e., target, equivalent, and binary) that are used to characterize the base process (after transformations/mappings) across different domains (i.e., target, Gaussian, and Bernoulli, respectively), thus clarify delicate steps, avoid potential pitfalls, and *algorithmize* the underlying computational procedure (by providing step-by-step recipes). It is remarked that the assumption of Gaussian copula can be relaxed and other types of copulas can be employed with satisfactory performance (see Serinaldi et al. (2020)), an aspect which can be a potential topic of future research.

The performance of the models has been validated through four simulation studies, concerning intermittent processes modelled with zero-inflated ($\mathcal{ZI}$) distributions, and particularly *strong* theoretical ACS. The simulations highlight the performance of the two βB-based models, being able to describe the distribution of k-length block maxima with high accuracy. Interestingly, and as confirmed by the examples of Serinaldi et al. (2020) (solely though for the βB and i.i.d. models), the distribution of k-length block maxima obtained by the five models converge as the probability of exceedance decreases (i.e., large values of return periods), as a result of the tail independence property of the Gaussian copula (Embrechts et al. 1999; Beirlant et al. 2006; Davis and Mikosch 2009). A result that if investigated further can be used to relax, in some extent, the need for precise identification of the ACS structure of the base process (i.e., when using data-based inference from historical records); when the target is the identification of the design rainfall for a given return period. Highlighting on the other hand the importance of selecting and fitting an appropriate marginal distribution to the observed records of observations. In our view, this is actually a major point of interest for future studies, since throughout this work it was assumed that the marginal distribution F_X and the ACS ρ_τ of the base process X_t are *a priori* known.

Next, using as examples intermittent processes (e.g., daily rainfall), we performed a series of simulation experiments (see also the simulations provided in the Supplementary Material) and comparisons (using asymptotic or non-asymptotic models for the block maxima distribution), aiming to provide answers to questions such as, w*hich model/approach should we employ to model the distribution of k-length block maxima? Are the non-asymptotic models more robust than the asymptotic ones? Using which approach (asymptotic or non-asymptotic), we have better chances to identify the true distribution of k-length block maxima?* The results of the devised Monte-Carlo-type experiments remark the ability of the non-asymptotic approaches to model the distribution of maxima with reduced uncertainty and variability, since in all *diagnostics* the latter approaches demonstrated better and more stable performance in terms of identifying the *true* probability distribution of maxima (which was *a priori* know). A finding that highlights that if the complete record of observations is available (i.e., not just a subset of BM or POT data), then our best bet would be the use of non-asymptotic models for the distribution of maxima. However, it is important to remark that when only a subset of BM is available, the simulation experiments and diagnostics performed herein verify the competence of $\mathcal{GEV}$ distribution to describe the probabilistic

behaviour of maxima (compared to fitting a non-extreme value distribution model, such as the Burr type-XII, directly to the record of BM).

An additional topic of interest of this work, also regarding intermittent processes (i.e., modelled by a $\mathcal{ZI}$ distribution), concerns the use of the two aforementioned βB-based models for the estimation of the probability of observing a zero value across multiple levels of aggregation *d.* As demonstrated via two simulation studies the two models proved to be capable of determining the probability of observing a zero value across multiple scales d, solely based on the marginal distribution and correlation structure of the base process. It is also highlighted that this topic, which at first sight might be considered irrelevant with the behaviour of extremes, can be (see the discussion of section 5.4) a crucial step for the multi-scale modeling of the marginal distribution of hydrometeorological processes, as well as used for the development of intensity-duration-frequency (IDF) models (e.g., rainfall), such as the one *sketched* in section 5.4.

With all the above in mind, and by acknowledging that this work did not covered real-world applications, (since the focus was on methods and algorithmic developments) it is argued that it would be of particular interest to further assess and cross-compare the performance of such modelling approaches (e.g., asymptotic Vs non-asymptotic models for the distribution of maxima, multi-scale modeling of the process' distribution construction of IDF models) by employing real-world datasets (e.g., rainfall or runoff) comprised by multiple stations and long records of observations. Furthermore, such large-scale datasets could also provide a unique testbed to assess the plausibility of the Gaussian copula assumption for the dependence structure of the process among consecutive time steps, an assumption made throughout the analyses presented herein. It is recalled that the latter copula is linked with the property of asymptotic tail independence and thus its plausibility for hydrometeorological processes should be further investigated.

Incidental developments of this work are: 1) the closed-form approximation for establishment of the TEC relationship (i.e., target vs equivalent correlations) for (binary) processes with Bernoulli marginal distribution, which can be used either for computational speed-up of βB-based block maxima models, or for the generation of correlated binary processes. 2) The brief note on the moments, probability weighted moments, and L-moments of zero-inflated distributions (presented in Appendix A). This distribution is widely used in the hydrological domain and is particularly suitable for processes characterized by intermittency (e.g., fine time scale ones). Appendix A is filling this way the gap in hydrological literature which, to the best of the author's knowledge, was lacking from a similar effort. 3) Finally, and in the spirit of reproducible research, all the procedures and recipes described herein have been implemented in R programming language (R Core Team 2017), in the form of an R package, named `bBextremes`, which to the best of the author's knowledge is currently the only open-source implementation of these models.

Acknowledgements

I thank the Associate Editor and the three reviewers for their constructive comments and suggestions that significantly improved the content and quality of the manuscript. Also, I acknowledge the motivational and always fruitful discussions on the broader topic of stochastics with the NTUA colleagues, Panagiotis Kossieris, Andreas

Efstratiadis, Demetris Koutsoyiannis and Christos Makropoulos. This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY).

Data and software availability: The analysis, simulations, as well as visualization were performed in R language (R Core Team 2017). The source code of `anySim` R package (Tsoukalas et al. 2020) is available at the GitHub repository: https://github.com/itsoukal/anySim. The source code of `bBextremes` R package (introduced in this work) is available at the GitHub repository: https://github.com/itsoukal/bBextremes. Also, all the figures/plots presented herein were produced via `ggplot2` R package (Wickham 2016).

References

Aas K, Czado C, Frigessi A, Bakken H (2009) Pair-copula constructions of multiple dependence. Insur Math Econ. doi: 10.1016/j.insmatheco.2007.02.001

Bárdossy A, Pegram G (2009) Copula based multisite model for daily precipitation simulation. Hydrol Earth Syst Sci Discuss 6:4485–4534. doi: 10.5194/hessd-6-4485-2009

Bárdossy A, Pegram GGS (2016) Space-time conditional disaggregation of precipitation at high resolution via simulation. Water Resour Res 52:920–937. doi: 10.1002/2015WR018037

Beirlant J, Goegebeur Y, Segers J, Teugels JL (2006) Statistics of extremes: theory and applications. John Wiley & Sons

Bell TL (1987) A space-time stochastic model of rainfall for satellite remote-sensing studies. J Geophys Res 92:9631. doi: 10.1029/JD092iD08p09631

Benestad RE, Parding KM, Erlandsen HB, Mezghani A (2019) A simple equation to study changes in rainfall statistics. Environ Res Lett 14:084017. doi: 10.1088/1748-9326/ab2bb2

Beran J (1992) Statistical Methods for Data with Long-Range Dependence. Stat Sci 7:404–416. doi: 10.1214/ss/1177011122

Biller B, Nelson BL (2003) Modeling and generating multivariate time-series input processes using a vector autoregressive technique. ACM Trans Model Comput Simul 13:211–237. doi: 10.1145/937332.937333

Blum AG, Archfield SA, Vogel RM (2017) On the probability distribution of daily streamflow in the United States. Hydrol Earth Syst Sci 21:3093–3103. doi: 10.5194/hess-21-3093-2017

Buishand TA (1991) Extreme rainfall estimation by combining data from several sites. Hydrol Sci J 36:345–365. doi: 10.1080/02626669109492519

Buishand TA (1989) Statistics of extremes in climatology. Stat Neerl 43:1–30. doi: 10.1111/j.1467-9574.1989.tb01244.x

Burr IW (1942) Cumulative Frequency Functions. Ann Math Stat 13:215–232. doi: 10.1214/aoms/1177731607

Calenda G, Mancini CP, Volpi E (2009) Selection of the probabilistic model of extreme floods: The case of the River Tiber in Rome. J Hydrol 371:1–11. doi: 10.1016/j.jhydrol.2009.03.010

Cario MC, Nelson BL (1996) Autoregressive to anything: Time-series input processes for simulation. Oper Res Lett 19:51–58. doi: 10.1016/0167-6377(96)00017-X

Cavanaugh NR, Gershunov A, Panorska AK, Kozubowski TJ (2015) The probability distribution of intense daily precipitation. Geophys Res Lett 42:1560–1567

Chen L, Guo S (2019) Copulas and Its Application in Hydrology and Water Resources. Springer

Chen L, Singh VP (2018) Entropy-based derivation of generalized distributions for hydrometeorological frequency analysis. J Hydrol 557:699–712. doi: 10.1016/j.jhydrol.2017.12.066

Christakos G (2012) Random field models in earth sciences. Courier Corporation

Coles S (2001) An introduction to statistical modeling of extreme values. Springer

Coles S, Pericchi LR, Sisson S (2003) A fully probabilistic approach to extreme rainfall modeling. J Hydrol 273:35–50. doi: 10.1016/S0022-1694(02)00353-0

Courty LG, Wilby RL, Hillier JK, Slater LJ (2019) Intensity-duration-frequency curves at the global scale. Environ Res Lett 14:084045. doi: 10.1088/1748-9326/ab370a

Dagum C (1977) New model of personal income-distribution-specification and estimation. Econ appliquée 30:413–437

Davis RA, Mikosch T (2009) The extremogram: A correlogram for extreme events. Bernoulli 15:977–1009. doi: 10.3150/09-BEJ213

de Haan L (1971) A form of regular variation and its application to the domain of attraction of the double exponential distribution. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 17:241–258

De Michele (2019) Advances in Deriving the Exact Distribution of Maximum Annual Daily Precipitation. Water 11:2322. doi: 10.3390/w11112322

De Michele C, Avanzi F (2018) Superstatistical distribution of daily precipitation extremes: A worldwide assessment. Sci Rep 8:14204. doi: 10.1038/s41598-018-31838-z

De Michele C, Salvadori G (2003) A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas. J Geophys Res 108:4067. doi: 10.1029/2002JD002534

Deidda R, Puliga M (2006) Sensitivity of goodness-of-fit statistics to rainfall data rounding off. Phys Chem Earth, Parts A/B/C 31:1240–1251. doi: 10.1016/j.pce.2006.04.041

Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling Extremal Events. Springer Berlin Heidelberg, Berlin, Heidelberg

Embrechts P, McNeil AJ, Straumann D (1999) Correlation and Dependence in Risk Management: Properties and Pitfalls. In: Dempster MAH (ed) Risk Management. Cambridge University Press, Cambridge, pp 176–223

Emrich LJ, Piedmonte MR (1991) A Method for Generating High-Dimensional Multivariate Binary Variates. Am Stat 45:302–304. doi: 10.1080/00031305.1991.10475828

Fang H-B, Fang K-T, Kotz S (2002) The Meta-elliptical Distributions with Given Marginals. J Multivar Anal 82:1–16. doi: 10.1006/jmva.2001.2017

Favre A-CA, El Adlouni S, Perreault L, et al (2004) Multivariate hydrological frequency analysis using copulas. Water Resour Res 40:. doi: 10.1029/2003WR002456

Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. Math Proc Cambridge Philos Soc. doi: 10.1017/S0305004100015681

Fréchet M (1927) Sur la loi de probabilité de l'écart maximum. Ann Soc Math Pol 6:93–116

Fuller WE (1914) Flood flows. In: Proceedings of the American Society of Civil Engineers. ASCE, pp 1011–1064

Galambos J (1972) On the distribution of the maximum of random variables. Ann Math Stat 516–521

Genest C, Favre A, Béliveau J, Jacques C (2007) Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. Water Resour Res 43:. doi: 10.1029/2006WR005275

Gnedenko B (1943) Sur la distribution limite du terme maximum d'une serie aleatoire. Ann Math 423–453

Greenwood JA, Landwehr JM, Matalas NC, Wallis JR (1979) Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. Water Resour Res 15:1049–1054. doi: 10.1029/WR015i005p01049

Grigoriu M (1984) Crossings of Non-Gaussian Translation Processes. J Eng Mech 110:610–620. doi: 10.1061/(ASCE)0733-9399(1984)110:4(610)

Grigoriu M (1998) Simulation of stationary non-Gaussian translation processes. J Eng Mech 124:121–126

Gujar U, Kavanagh R (1968) Generation of random signals with specified probability

density functions and power density spectra. IEEE Trans Automat Contr 13:716–719

Gumbel EJ (1941a) The Return Period of Flood Flows. Ann Math Stat 12:163–190. doi: 10.1214/aoms/1177731747

Gumbel EJ (1941b) Probability-interpretation of the observed return-periods of floods. Eos, Trans Am Geophys Union 22:836–850

Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York, USA

Hao Z, Singh VP (2009) Entropy-based parameter estimation for extended Burr XII distribution. Stoch Environ Res Risk Assess 23:1113–1122. doi: 10.1007/s00477-008-0286-7

Hazen A (1914) Storage to be provided in impounded reservoirs for municipal water supply. Trans ASCE 77: 1539:

Herr HD, Krzysztofowicz R (2005) Generic probability distribution of rainfall in space: The bivariate model. J Hydrol 306:234–263. doi: 10.1016/j.jhydrol.2004.09.011

Hirtzel CS, Corotis RB, Quon JE (1982) Estimating the maximum value of autocorrelated air quality measurements. Atmos Environ 16:2603–2608. doi: 10.1016/0004-6981(82)90341-9

Hofert M, Kojadinovic I, Maechler M, Yan J (2014) copula: Multivariate dependence with copulas. R Packag version 0999-9, URL http//CRAN R-project org/package= copula C225

Hosking JRM (1990) L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. J R Stat Soc Ser B 52:105–124

Hosking JRM (2019) Package 'lmom'

Hsing T, Husler J, Leadbetter MR (1988) On the exceedance point process for a stationary sequence. Probab Theory Relat Fields 78:97–112. doi: 10.1007/BF00718038

Iliopoulou T, Koutsoyiannis D (2020) Projecting the future of rainfall extremes: Better classic than trendy. J Hydrol 588:125005. doi: 10.1016/j.jhydrol.2020.125005

Iliopoulou T, Papalexiou SM, Markonis Y, Koutsoyiannis D (2016) Revisiting long-range dependence in annual precipitation. J Hydrol 6:399–401. doi: 10.1016/j.jhydrol.2016.04.015

Jenkinson AF (1955) The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Q J R Meteorol Soc 81:158–171. doi: 10.1002/qj.49708134804

Joe H (2014) Dependence modeling with copulas. CRC Press

Kantelhardt JW, Koscielny-Bunde E, Rybski D, et al (2006) Long-term persistence and multifractality of precipitation and river runoff records. J Geophys Res 111:D01106. doi: 10.1029/2005JD005881

Katz RW, Parlange MB, Naveau P (2002) Statistics of extremes in hydrology. Adv Water Resour. doi: 10.1016/S0309-1708(02)00056-8

Kelly KS, Krzysztofowicz R (1997) A bivariate meta-Gaussian density for use in hydrology. Stoch Hydrol Hydraul 11:17–31. doi: 10.1007/BF02428423

Kleiber C (2008) A Guide to the Dagum Distributions. In: Chotikapanich D (ed) Modeling Income Distributions and Lorenz Curves. Springer New York, New York, NY, pp 97–117

Klemeš V, Borůvka L (1974) Simulation of Gamma-Distributed First-Order Markov Chain. Water Resour Res 10:87–91. doi: 10.1029/WR010i001p00087

Kojadinovic I, Yan J (2010) Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. J Stat Softw 34:. doi: 10.18637/jss.v034.i09

Kossieris P (2020) Multi-scale stochastic analysis and modelling of residential water demand processes. PhD Thesis, Department of Water Resources and Environmental Engineering, National Technical University of Athens

Kossieris P, Tsoukalas I, Makropoulos C, Savic D (2019) Simulating Marginal and Dependence Behaviour of Water Demand Processes at Any Fine Time Scale. Water 11:885. doi: 10.3390/w11050885

Kotz S, Nadarajah S (2000) Extreme Value Distributions. Imperial College Press

Koutsoyiannis D (2010) A random walk on water. Hydrol Earth Syst Sci 14:585–601. doi: 10.5194/hess-14-585-2010

Koutsoyiannis D (2020) Stochastics of Hydroclimatic Extremes – A Cool Look at Risk, Edition 0. National Technical University of Athens, Athens, Greece

Koutsoyiannis D (2004) Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation / Statistiques de valeurs extrêmes et estimation de précipitations extrêmes: I. Recherche théorique. Hydrol Sci J 49:. doi: 10.1623/hysj.49.4.575.54430

Koutsoyiannis D (1999) A probabilistic view of hershfield's method for estimating probable maximum precipitation. Water Resour Res 35:1313–1322. doi: 10.1029/1999WR900002

Koutsoyiannis D (2005) Uncertainty, entropy, scaling and hydrological stochastics. 2. Time dependence of hydrological processes and time scaling / Incertitude, entropie, effet d'échelle et propriétés stochastiques hydrologiques. 2. Dépendance temporelle des processus hydrologiq. Hydrol Sci J 50:381–404. doi: 10.1623/hysj.50.3.405.65028

Koutsoyiannis D (2000) A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. Water Resour Res 36:1519–1533. doi: 10.1029/2000WR900044

Koutsoyiannis D, Kozonis D, Manetas A (1998) A mathematical framework for studying rainfall intensity-duration-frequency relationships. J Hydrol 206:118–135. doi:

10.1016/S0022-1694(98)00097-3

Koutsoyiannis D, Montanari A (2015) Negligent killing of scientific concepts: the stationarity case. Hydrol Sci J 60:1174–1183. doi: 10.1080/02626667.2014.959959

Kumar P, Guttarp P, Foufoula-Georgiou E (1994) A probability-weighted moment test to assess simple scaling. Stoch Hydrol Hydraul 8:173–183. doi: 10.1007/BF01587233

Kumaraswamy P (1980) A generalized probability density function for double-bounded random processes. J Hydrol 46:79–88. doi: 10.1016/0022-1694(80)90036-0

Lanza LG (2000) A conditional simulation model of intermittent rain fields. Hydrol Earth Syst Sci 4:173–183. doi: 10.5194/hess-4-173-2000

Leadbetter MR (1983) Extremes and local dependence in stationary sequences. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65:291–306. doi: 10.1007/BF00532484

Leadbetter MR (1974) On extreme values in stationary sequences. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete. doi: 10.1007/BF00532947

Leadbetter MR, Lindgren G, Rootzén H (1983) Extremes and Related Properties of Random Sequences and Processes. Springer New York, New York, NY

Leadbetter MR, Rootzen H (1988) Extremal Theory for Stochastic Processes. Ann Probab 16:431–478

Li ST, Hammond JL (1975) Generation of Pseudorandom Numbers with Specified Univariate Distributions and Correlation Coefficients. IEEE Trans Syst Man Cybern SMC-5:557–561. doi: 10.1109/TSMC.1975.5408380

Lins HF, Cohn TA (2011) Stationarity: Wanted dead or alive? J Am Water Resour Assoc. doi: 10.1111/j.1752-1688.2011.00542.x

Liu P-L, Der Kiureghian A (1986) Multivariate distribution models with prescribed marginals and covariances. Probabilistic Eng Mech 1:105–112. doi: 10.1016/0266-8920(86)90033-0

Lombardo F, Napolitano F, Russo F, Koutsoyiannis D (2019) On the Exact Distribution of Correlated Extremes in Hydrology. Water Resour Res 55:10405–10423. doi: 10.1029/2019WR025547

Mandelbrot B, Van Ness J (1968) Fractional Brownian Motions, Fractional Noises and Applications. SIAM Rev 10:422–437. doi: 10.1137/1010093

Marani M, Ignaccolo M (2015) A metastatistical approach to rainfall extremes. Adv Water Resour 79:121–126. doi: 10.1016/j.advwatres.2015.03.001

Matalas NC (1967) Mathematical assessment of synthetic hydrology. Water Resour Res 3:937–945. doi: 10.1029/WR003i004p00937

Matalas NC (2012) Comment on the Announced Death of Stationarity. J Water Resour Plan Manag 138:311–312. doi: 10.1061/(ASCE)WR.1943-5452.0000215

Montanari A, Koutsoyiannis D (2014) Modeling and mitigating natural hazards: Stationarity is immortal! Water Resour Res 50:9748–9756. doi: 10.1002/2014WR016092

Mooley DA (1973) Gamma Distribution Probability Model for Asian Summer Monsoon Monthly Rainfall. Mon Weather Rev 101:160–176. doi: 10.1175/1520-0493(1973)101<0160:GDPMFA>2.3.CO;2

Nataf A (1962) Statistique mathematique-determination des distributions de probabilites dont les marges sont donnees. C R Acad Sci Paris 255:42–43

Nelsen RB (2007) An introduction to copulas. Springer Science & Business Media

Papalexiou SM (2018) Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. Adv Water Resour 115:234–252. doi: 10.1016/j.advwatres.2018.02.013

Park J-S, Jung H-S (2002) Modelling Korean extreme rainfall using a Kappa distribution and maximum likelihood estimate. Theor Appl Climatol 72:55–64. doi: 10.1007/s007040200012

Pickands III J (1975) Statistical inference using extreme order statistics. Ann Stat 3:119–131

Pilon PJ, Adamowski K, Alila Y (1991) Regional analysis of annual maxima precipitation using L-moments. Atmos Res 27:81–92. doi: 10.1016/0169-8095(91)90009-L

Popescu R, Deodatis G, Prevost JH (1998) Simulation of homogeneous nonGaussian stochastic vector fields. Probabilistic Eng Mech 13:1–13. doi: 10.1016/S0266-8920(97)00001-5

R Core Team (2017) R Development Core Team. R A Lang. Environ. Stat. Comput. 55:275–286

Reiss R-D, Thomas M, Reiss RD (1997) Statistical analysis of extreme values. Springer

Renard B, Lang M (2007) Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. Adv Water Resour 30:897–912. doi: 10.1016/j.advwatres.2006.08.001

Resnick SI (1987) Extreme Values, Regular Variation and Point Processes. Springer New York, New York, NY

Roche M (1963) Hydrologie de surface

Salas JD, Obeysekera J (2019) Probability Distribution and Risk of the First Occurrence of k Extreme Hydrologic Events. J Hydrol Eng. doi: 10.1061/(asce)he.1943-5584.0001809

Salas JD, Obeysekera J, Vogel RM (2018) Techniques for assessing water infrastructure for nonstationary extreme events: a review. Hydrol Sci J. doi: 10.1080/02626667.2018.1426858

Salvadori G, De Michele C (2007) On the Use of Copulas in Hydrology: Theory and Practice. J Hydrol Eng 12:369–380. doi: 10.1061/(ASCE)1084-0699(2007)12:4(369)

Salvadori G, De Michele C (2004) Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. Water Resour Res 40:. doi: 10.1029/2004WR003133

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

Schweizer B, Dall'Aglio G, Kotz S, Salinetti G (1991) Thirty years of copulas. In advances in probability distributions with given marginals: beyond the copulas. Kluwer Acad Publ

Serinaldi F, Lombardo F (2017) BetaBit: A fast generator of autocorrelated binary processes for geophysical research. EPL (Europhysics Lett 118:30007. doi: 10.1209/0295-5075/118/30007

Serinaldi F, Lombardo F, Kilsby CG (2020) All in order: Distribution of serially correlated order statistics with applications to hydrological extremes. Adv Water Resour 144:103686. doi: 10.1016/j.advwatres.2020.103686

Shao Q, Wong H, Xia J, Ip W-C (2004) Models for extremes using the extended three-parameter Burr XII system with application to flood frequency analysis / Modèles d'extrêmes utilisant le système Burr XII étendu à trois paramètres et application à l'analyse fréquentielle des crues. Hydrol Sci J 49:. doi: 10.1623/hysj.49.4.685.54425

Singh S, Maddala G (1976) A function for size distribution of incomes. Econometrica 44:963–970. doi: 10.2307/1910422

Sklar A (1973) Random variables, joint distribution functions, and copulas. Kybernetika 9:449–460

Sklar M (1959) Fonctions de repartition an dimensions et leurs marges. Publ Inst Stat Univ Paris 8:229–231

Smith RL (2003) Statistics of extremes, with applications in environment, insurance, and finance. In: Extreme Values in Finance, Telecommunications, and the Environment

Smith RL, Weissman I (1994) Estimating the Extremal Index. J R Stat Soc Ser B 56:515–528

Svensson C, Jones DA (2010) Review of rainfall frequency estimation methods. J Flood Risk Manag 3:296–313. doi: 10.1111/j.1753-318X.2010.01079.x

Tadikamalla PR (1980) A look at the Burr and related distributions. Int Stat Rev Int Stat 337–344

Thom HCS (1968) Approximate convolution of the gamma and mixed gamma distributions. Mon Weather Rev 96:883–886. doi: 10.1175/1520-0493(1968)096<0883:ACOTGA>2.0.CO;2

Todorovic P (1970) On Some Problems Involving Random Number of Random Variables. Ann Math Stat 41:1059–1063. doi: 10.1214/aoms/1177696981

Todorovic P, Zelenhasic E (1970) A Stochastic Model for Flood Analysis. Water Resour Res 6:1641–1648. doi: 10.1029/WR006i006p01641

Tsoukalas I (2018) Modelling and simulation of non-Gaussian stochastic processes for optimization of water-systems under uncertainty. PhD Thesis, Department of Water Resources and Environmental Engineering, National Technical University of Athens (Defence date: 20 December 2018)

Tsoukalas I, Efstratiadis A, Makropoulos C (2018a) Stochastic Periodic Autoregressive to Anything (SPARTA): Modeling and Simulation of Cyclostationary Processes With Arbitrary Marginal Distributions. Water Resour Res 54:161–185. doi: 10.1002/2017WR021394

Tsoukalas I, Efstratiadis A, Makropoulos C (2017) Stochastic simulation of periodic processes with arbitrary marginal distributions. In: 15th International Conference on Environmental Science and Technology. CEST 2017. Rhodes, Greece

Tsoukalas I, Efstratiadis A, Makropoulos C (2019) Building a puzzle to solve a riddle: A multi-scale disaggregation approach for multivariate stochastic processes with any marginal distribution and correlation structure. J Hydrol 575:354–380. doi: 10.1016/j.jhydrol.2019.05.017

Tsoukalas I, Kossieris P, Makropoulos C (2020) Simulation of Non-Gaussian Correlated Random Variables, Stochastic Processes and Random Fields: Introducing the anySim R-Package for Environmental Applications and Beyond. Water 12:1645. doi: 10.3390/w12061645

Tsoukalas I, Makropoulos C, Koutsoyiannis D (2018b) Simulation of Stochastic Processes Exhibiting Any-Range Dependence and Arbitrary Marginal Distributions. Water Resour Res 54:9484–9513. doi: 10.1029/2017WR022462

Usta I (2013) Different estimation methods for the parameters of the extended Burr XII distribution. J Appl Stat 40:397–414. doi: 10.1080/02664763.2012.743974

Volpi E, Fiori A (2012) Design event selection in bivariate hydrological frequency analysis. Hydrol Sci J 57:1506–1515. doi: 10.1080/02626667.2012.726357

Volpi E, Fiori A (2014) Hydraulic structures subject to bivariate hydrological loads: Return period, design, and risk assessment. Water Resour Res 50:885–897. doi: 10.1002/2013WR014214

Volpi E, Fiori A, Grimaldi S, et al (2015) One hundred years of return period: Strengths and limitations. Water Resour Res 51:8570–8585. doi: 10.1002/2015WR017820

Volpi E, Fiori A, Grimaldi S, et al (2019) Save hydrological observations! Return period estimation without data decimation. J Hydrol 571:782–792. doi: 10.1016/j.jhydrol.2019.02.017

von Mises R (1936) La distribution de la plus grande de n valeurs. Am Math Soc

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, New York, NY

Wilks DS (1993) Comparison of three-parameter probability distributions for representing annual extreme and partial duration precipitation series. Water Resour Res 29:3543–3549. doi: 10.1029/93WR01710

Wilson PS, Toumi R (2005) A fundamental probability distribution for heavy rainfall. Geophys Res Lett 32:n/a-n/a. doi: 10.1029/2005GL022465

Yamazaki F, Shinozuka M (1988) Digital generation of non-Gaussian stochastic fields. J Eng Mech 114:1183–1197. doi: 10.1061/(ASCE)0733-9399(1988)114:7(1183)

Yan J (2007) Enjoy the Joy of Copulas: With a Package copula. J Stat Softw 21:. doi: 10.18637/jss.v021.i04

Ye L, Hanson LS, Ding P, et al (2018) The probability distribution of daily precipitation at the point and catchment scales in the United States. Hydrol Earth Syst Sci 22:6519–6531. doi: 10.5194/hess-22-6519-2018

Zhang L, Singh VP (2019) Copulas and their Applications in Water Resources Engineering. Cambridge University Press

Zorzetto E, Botter G, Marani M (2016) On the emergence of rainfall extremes from ordinary events. Geophys Res Lett 43:8076–8082. doi: 10.1002/2016GL069445

Appendix A: Moments, probability weighted moments, and L-moments of zero-inflated distributions

The CDF, denoted as F_X, and ICDF, denoted as F_X^{-1} of a zero-inflated ($\mathcal{ZI}$) distribution are given respectively by,

$$u = F_X(x) = \begin{cases} p_0, & x = 0 \\ p_0 + (1 - p_0)G_X(x), & x > 0 \end{cases} \tag{A1}$$

$$x_u = F_X^{-1}(u) = \begin{cases} 0, & 0 \le u \le p_0 \\ G_X^{-1}\left(\dfrac{(u - p_0)}{(1 - p_0)}\right), & p_0 < u \le 1 \end{cases} \tag{A2}$$

where, $p_0 \coloneqq \mathrm{P}\{X = 0\}$ is a parameter controlling the inflation of zeros (i.e., the discrete part of the $\mathcal{ZI}$ distribution – the probability of observing a zero value), and $G_X \coloneqq F_{X|X>0} = \mathrm{P}\{X \le x | X > 0\}$ denotes the distribution to be inflated (i.e., the continuous part of the $\mathcal{ZI}$ distribution, herein assumed to be defined in the positive half line $(0, \infty)$). For completeness and subsequent use, it is noted that, $p_1 \coloneqq \mathrm{P}\{X > 0\} = 1 - p_0$.

For applications of this distribution in hydrological domain see for instance the early works of Roche (1963 p. 69), Thom (1968), Mooley (1973), Bell (1987), Lanza (2000), and the more recent ones of Bárdossy and Pegram (2009, 2016). See also, Papalexiou (2018), Tsoukalas (2018), Tsoukalas et al. (2018b, 2019, 2020) who used it within the context of Nataf-based (i.e., Gaussian copula) stochastic models, aiming at the simulation of intermittent hydrometeorological processes, such as rainfall at fine time scales, as well as Kossieris et al. (2019) for modelling non-physical processes, such as water demand.

Moments of a zero-inflated ($\mathcal{ZI}$) distribution model

Denoting by $f_X(x)$ the probability density function of the RV X, the theoretical raw moments (or moments about the origin), $\mu'_X(r) \coloneqq \mathrm{E}[X^r] = \int_{-\infty}^{\infty} x^r f_X(x)\mathrm{d}x = r\left(-\int_{-\infty}^{0} x^{r-1}F_X(x)\mathrm{d}x + \int_0^{\infty} x^{r-1}\big(1 - F_X(x)\big)\mathrm{d}x\right) = \int_0^1 \big(F_X^{-1}(u)\big)^r \,\mathrm{d}u$ of order $r = 0,1,\ldots$ of the zero-inflated distribution are given by,

$$\mu'_X(r) = p_1 \int_0^{\infty} x^r g_X(x)\mathrm{d}x = \int_{p_0}^{1} \left(G_X^{-1}\left(\frac{u - p_0}{1 - p_0}\right)\right)^r \mathrm{d}u = p_1 \mu'_{X|X>0}(r) \tag{A3}$$

where $\mu'_{X|X>0}(r)$ is the r^{th} raw moment of $G_X = F_{X|X>0} = \mathrm{P}\{X \le x | X > 0\}$. Noting that, $\mu'_X(0) = 1$ and $\mu'_X(1) = \mu_X = \mathrm{E}[X]$.

The corresponding r^{th} order central moments, $\mu_X(r) \coloneqq \mathrm{E}[(X - \mu_X)^r]$, can be obtained by the following general relationship that the links raw and central moments.

$$\mu_X(r) = \mathrm{E}[(X-\mu_X)^r] = \sum_{j=0}^{r}\binom{r}{j}(-1)^{r-j}\mu'_X(j)\mu_X^{r-j} \tag{A4}$$

The inverse relationship is given,

$$\mu'_X(r) = \mathrm{E}[(X)^r] = \sum_{j=0}^{r}\binom{r}{j}\mu_X(j)\,\mu_X^{r-j} \tag{A5}$$

where $\binom{r}{j}$ denotes the Binomial coefficient. Note that $\mu_X(0)=1$ and $\mu_X(1)=0$. For instance, for $r=2,3,4$, the central moments are given by,

$$\mu_X(2) = \mu'_X(2) - \left(\mu'_X(1)\right)^2 \tag{A6}$$

$$\mu_X(3) = \mu'_X(3) - 3\mu'_X(1)\mu'_X(2) + 2\left(\mu'_X(1)\right)^3 \tag{A7}$$

$$\mu_X(4) = \mu'_X(4) - 4\mu'_X(1)\mu'_X(3) + 6\left(\mu'_X(1)\right)^2\mu'_X(2) - 3\left(\mu'_X(1)\right)^4 \tag{A8}$$

Particularly, the mean (μ_X), variance (σ_X^2) and skewness $\left(C_{s_X} = \mu_X(3)/\left(\mu_X(2)\right)^{3/2}\right)$ of a zero-inflated distribution are given by,

$$\mu_X = p_1\mu_p \tag{A9}$$

$$\sigma_X^2 = p_1(\sigma_p^2 + (1-p_1)\mu_p^2) \tag{A10}$$

$$C_{s_X} = \frac{\left(p_1 C_{s_p}(\sigma_p^2)^{\frac{3}{2}} + 3(p_1-p_1^2)\mu_p\sigma_p^2 + \frac{1-2p_1}{\sqrt{(p_1-p_1^2)}}\left((p_1-p_1^2)^{\frac{3}{2}}\right)\mu_p^3\right)}{(\sigma_X^2)^{3/2}} \tag{A11}$$

where, $\mu_p \coloneqq \mu_{X|X>0}$, $= \sigma_p^2 \coloneqq \sigma_{X|X>0}^2$ and $C_{s_p} \coloneqq C_{s_{X|X>0}}$ denote the mean, variance and skewness of the continuous distribution $G_X = F_{X|X>0}$. It is noted that an analytical relationship can be established for the kurtosis coefficient $\left(C_{k_X} \coloneqq \mu_X(4)/\left(\mu_X(2)\right)^2\right)$, but it is too complex, and probably of limited use, hence not included herein.

Probability weighted moments (PWMs) of a zero-inflated ($\mathcal{ZI}$) distribution model

Assuming that the random variable X has finite mean, the PWMs are defined as (Greenwood et al. 1979), $M_X(p,r,s) = \int_0^1\left(F_X^{-1}(u)\right)^p u^r(1-u)^s\mathrm{d}u$, where, in the most general case, p, r and s, are real numbers (typically though denoting non-negative integers). When $r=0$, $s=0$, and $p \in Z^+$, PWMs reduce to classical moments, i.e., $M_X(p,0,0) = \mu'_X(r)$. Following the above definition, the PWMs of a $\mathcal{ZI}$ distribution are given by,

$$M_X(p,r,s) = \int_{p_0}^{1} \left(F_X^{-1}\left(\frac{u-p_0}{1-p_0}\right)\right)^p u^r (1-u)^s \mathrm{d}u \tag{A12}$$

The two most common types of PWMs are the so-called, α- and β-type moments (Greenwood et al. 1979), which for the $\mathcal{ZI}$ distribution are given by (note that hereafter when convenient, the subscript X is omitted for simplicity),

$$\alpha(s) = M_X(1,0,s) = \int_{p_0}^{1} \left(F_X^{-1}\left(\frac{u-p_0}{1-p_0}\right)\right)(1-u)^s \mathrm{d}u \tag{A13}$$

$$\beta(r) = M_X(1,r,0) = \int_{p_0}^{1} \left(F_X^{-1}\left(\frac{u-p_0}{1-p_0}\right)\right) u^r \mathrm{d}u \tag{A14}$$

It is also noted that the quantities $\alpha(s)$ and $\beta(r)$ are related by,

$$\alpha(s) = \sum_{i=0}^{s} \binom{s}{i} \beta(i) \tag{A15}$$

$$\beta(r) = \sum_{i=0}^{r} \binom{r}{i} \alpha(i) \tag{A16}$$

Hence providing interchangeably the same information about a distribution (Kumar et al. 1994). Furthermore, there is also a link between α- and β-type PWMs and L-moments (Hosking 1990) of order r, which allows us to express the latter as function of the former, i.e.,

$$\lambda_{r+1} = \sum_{m=0}^{r} p_{r,m}^{*}\, \beta(m) = (-1)^r \sum_{m=0}^{r} p_{r,m}^{*}\, \alpha(m) \tag{A17}$$

where,

$$p_{r,m}^{*} = (-1)^{r-m} \binom{r}{m} \binom{r+m}{m} \tag{A18}$$

The first four L-moments of a distribution can be obtained by,

$$\lambda_1 = \alpha(0) = \beta(0) \tag{A19}$$

$$\lambda_2 = \alpha(0) - \alpha(1) = 2\beta_X(1) - \beta_X(0) \tag{A20}$$

$$\lambda_3 = \alpha(0) - 6\alpha(1) + 6\alpha(2) = 6\beta(2) - 6\beta(1) + \beta(0) \tag{A21}$$

$$\lambda_4 = \alpha(0) - 12\alpha(1) + 30\alpha(2) - 20\alpha(2) = 20\beta(3) - 30\beta(2) + 12\beta(1) - \beta(0) \tag{A22}$$

while, the L-coefficient of variation (τ_2), L-skewness (τ_3) and L-kurtosis (τ_4) can be obtained respectively by (note that, $\lambda_1 \equiv \mu_X$),

$$\tau_2 = \lambda_2/\lambda_1\,, \qquad \tau_3 = \lambda_3/\lambda_2\,, \qquad \tau_4 = \lambda_4/\lambda_2 \tag{A23}$$

Following a recent result related with so-called knowable, or for short, k-moments (Koutsoyiannis 2020) as well as simple reasoning, it is possible to derive formulas that link the α- and β-type moments of the continuous distribution (i.e., $\alpha_{X|X>0}(s)$ and $\beta_{X|X>0}(r)$ respectively) with those of the $\mathcal{ZI}$ distribution (i.e., $\alpha_X(s)$ and $\beta_X(r)$). In particular, the quantities $\alpha_{X|X>0}(s)$ and $\alpha_X(s)$ are related by,

$$\alpha_X(s) = \alpha_{X|X>0}(s)(1-p_0)^{s+1} \tag{A24}$$

On the other hand, the the quantities $\beta_{X|X>0}(r)$ and $\beta_X(r)$ are related by,

$$\beta_X(r) = \frac{{p_0}^{r+1}}{r+1}\sum_{i=1}^{r+1}\binom{r+1}{i} i\beta_{X|X>0}(i-1)\left(\frac{1-p_0}{p_0}\right)^i \tag{A25}$$

The above equation along with Eq. (A17)-(A22) provide the means to calculate the theoretical L-moments of a $\mathcal{ZI}$ distribution (Eq. A1), on the basis of $\alpha_{X|X>0}(s)$ or $\beta_{X|X>0}(r)$ and p_0, even analytically, if $\alpha_{X|X>0}(s)$ or $\beta_{X|X>0}(r)$ have an explicit form (see Appendix C, for the case of a zero-inflated Burr type XII distribution).

Appendix B: Distribution functions

The CDF and ICDF of a zero-inflated Burr type-XII ($\mathcal{ZIB}rXII$) distribution are given by,

$$F(x) = 1 - p_1\left(1 + \xi\zeta\left(\frac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\xi\zeta}} \tag{B1}$$

$$x_u = F^{-1}(u) = \lambda\left(\frac{\left(\frac{p_1}{1-u}\right)^{\xi\zeta} - 1}{\xi\zeta}\right)^{1/\zeta} \tag{B2}$$

where $p_1 = 1 - p_0$. p_0 is a parameter that controls the inflation of zeros, while $\lambda > 0$, $\zeta > 0$ and $\xi > 0$ denote the scale, shape, and tail index parameters respectively. Of course, by setting $p_1 = 1$ (i.e., $p_0 = 0$) we obtain the classical $\mathcal{B}rXII$ distribution (i.e., not inflated with zeros) (Burr 1942; Singh and Maddala 1976; Tadikamalla 1980).

The classical, raw moments of order r of the $\mathcal{ZIB}rXII$ distribution, are given by,

$$\mu'(r) = \frac{p_1 r\lambda^r}{\zeta(\xi\zeta)^{r/\zeta}}\,\mathrm{B}\left(\frac{r}{\zeta}, \frac{1}{\xi\zeta} - \frac{r}{\zeta}\right) \tag{B3}$$

where B(·,·) denotes the Beta function. It is noted that, the r^{th} moment of the $\mathcal{ZIB}rXII$ (and $\mathcal{B}r$XII) distribution is finite, only if, $r > 1/\xi$. Thus, to have finite variance $\xi < 0.5$.

The a-type PWMs of order s of the $\mathcal{ZIB}rXII$ distribution can be obtained by (see also Koutsoyiannis (2020) for a similar formula concerning K-moments),

$$\alpha(s) = \frac{p_1^{s+1}\lambda}{s+1}\frac{1}{\zeta(\zeta\xi)^{1/\zeta}}\,\mathrm{B}\left(\frac{s+1}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right) \tag{B4}$$

The above equation provides also the basis to analytically calculate the L-moments of a $\mathcal{ZIB}rXII$ distribution. In particular, the L-mean (λ_1), the L-coefficient of variation (τ_2) and L-skewness (τ_3) can be obtained by (L-kurtosis, τ_4, is omitted due to its complexity),

$$\lambda_1 = \alpha(0) = \frac{\lambda p_1}{\zeta}\frac{1}{(\zeta\xi)^{\frac{1}{\zeta}}}\,\mathrm{B}\left(\frac{1}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right) \tag{B5}$$

$$\tau_2 \coloneqq \frac{\lambda_2}{\lambda_1} = 1 - \frac{2\alpha(1)}{\alpha(0)} = 1 - \frac{p_1\mathrm{B}\left(\frac{2}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right)}{\mathrm{B}\left(\frac{1}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right)} \tag{B6}$$

$$\tau_3 \coloneqq \frac{\lambda_3}{\lambda_2} = \frac{\mathrm{B}\left(\frac{1}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right) - 3p_1\mathrm{B}\left(\frac{2}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right) + 2{p_1}^2\mathrm{B}\left(\frac{3}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right)}{\mathrm{B}\left(\frac{1}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right) - p_1\mathrm{B}\left(\frac{2}{\zeta\xi} - \frac{1}{\zeta}, \frac{1}{\zeta}\right)} \tag{B7}$$

The a-type PWMs and L-moments of the classical $\mathcal{B}rXII$ distribution (i.e., not inflated with zeros) can be estimated by simply setting $p_1 = 1$ (i.e., $p_0 = 0$).

The CDF of the Generalized Extreme Value ($\mathcal{GEV}$) distribution is given by,

$$F(x) = \begin{cases} \exp\left(-\left(1 + \xi\frac{x - c}{\lambda}\right)^{-\frac{1}{\xi}}\right), & \xi \neq 0 \\ \exp\left(-\exp\left(-\frac{x - c}{\lambda}\right)\right), & \xi = 0 \end{cases} \tag{B8}$$

where $\xi, c \in \mathbb{R}$ and $\lambda > 0$ are the tail index (shape), location and scale parameters respectively. $\mathcal{GEV}$ encompasses three distributions, the Fréchet ($\xi > 0$ with $X \in [c - \lambda/\xi\,, +\infty)$), the Gumbel ($\xi = 0$ with $X \in (-\infty, +\infty)$) and the reversed Weibull ($\xi < 0$ with $X \in (-\infty, c - \lambda/\xi]$); the latter case is not considered herein, since it regards upper bounded RVs. The moments of the $\mathcal{GEV}$ are finite up to orders smaller than $1/\xi$ (i.e., $\xi = 0.25$ implies that the distribution's moments that are greater or equal to 4 ($1/\xi$) are infinite).

Appendix C: Autocorrelation structures

The autocorrelation structure (ACS) of a fractional Gaussian noise (fGn) model (e.g., Mandelbrot and Van Ness 1968; Beran 1992), also referred to as Hurst-Kolmogorov (HK; Koutsoyiannis 2010), to credit seminal works, is given by,

$$\rho_\tau^{\mathrm{fGn}}(H) = \frac{1}{2}\left(|\tau-1|^{2H} - 2|\tau|^{2H} + |\tau+1|^{2H}\right) \tag{C1}$$

where, τ denotes the time lag, and $H \in (0,1)$ is the so-called Hurst parameter that controls the degree of persistence of the ACS. It is noted that a process with fGn ACS with $H = 0.5$, is identical to white noise one (i.e., i.i.d.), while for $H \in (0.5, 1)$, the ACS is positive for any τ, and exhibits long range dependence.

The Cauchy-type correlation structure (CAS), has been originally proposed by Koutsoyiannis (2000) in an effort to provide a parsimonious yet flexible ACS model capable of modelling a wide range of processes, entailing both short- and, long-range dependence. The ACS of CAS is given by:

$$\rho_\tau^{\mathrm{CAS}}(\beta,\kappa) = (1+\kappa\beta\tau)^{-1/\beta}, \qquad \tau \geq 0 \tag{C2}$$

where $\beta \geq 0$ and $\kappa > 0$ are model parameters. It is noted that $\beta = 0$ implies an ACS identical to that of an AR(1) processes (i.e., short-range dependence), and $\beta > 0$ implies an ACS with long-range dependence. The increased flexibility provided by CAS is demonstrated in several other studies, involving the simulation of physical (Tsoukalas et al. 2018b, 2019) and non-physical processes (Kossieris et al. 2019; Kossieris 2020), as well as spatiotemporal random fields (Tsoukalas et al. 2020).

Appendix D: Additional figures for the simulation studies of section 4

Tables

Table 1. Selected marginal distributions and autocorrelation structures (ACSs) for the simulation studies of this section.

Simulation study ID #	1	2	3	4
Marginal distribution	Zero-inflated Burr type-XII*: $\mathcal{ZIBr}XII(p_0 = 0.75, \beta = 7.07, \zeta = 0.928, \xi = 0.098)$**			
ACS*, ρ_τ	$\rho_\tau^{fGn}(H = 0.65)$	$\rho_\tau^{fGn}(H = 0.8)$	$\rho_\tau^{CAS}(\beta = 1.5, k = 0.5)$	$\rho_\tau^{\mathrm{AR(3)}}$***
Relevant Figure	**Figure 4**	**Figure 5**	**Figure D1**	**Figure D2**

* For further details about the employed marginal distributions and ACSs, i.e., the fGn, and Cauchy-type correlation structure (CAS), see Appendix B and C respectively. Also, Appendix A provides a brief overview on zero-inflated marginal distributions, as well as provides formulas for the calculation of their moments, probability weighted moments (Greenwood et al. 1979) and L-moments (Hosking 1990).

** The selected parameters resemble the distributional properties (as identified in Koutsoyiannis (2020)) of daily rainfall recorded at the station of Bologna, Italy.

*** The parameters of the AR(3) ACS are, $a_1 = 0.4, a_2 = 0.2$ and $a_3 = 0.1$, where a_i's denotes the autoregressive parameters of an AR process (in this case of order 3). Its ACS is given by, $\rho_\tau^{AR(n)} = \sum_{l=1}^{n} a_l \rho_{\tau-l}$. See also **Figure D2c** for a graphical illustration of this ACS.

List of captions

Figure 1. Graphical illustration of (a) Eq. (7), and (b) Eq. (8), for various values of $p^* = \min(1-p, p)$, where p is the parameter of the Bernoulli distribution (i.e., $\mathcal{B}e(p)$).

Figure 2. (a) Graphical comparison between the approximate closed-form expression of Eq. (6) and numerical integration for establishment of the TEC relationship for processes with $\mathcal{B}e(p)$ marginal distribution. (b) Graphical depiction of the error (simple difference) between the approximate closed-form expression of Eq. (6) and numerical integration. This plot focuses on *typical* values of p spanning from 0.5 to 10^{-4}.

Figure 3. (a) Graphical comparison between the approximate closed-form expression of Eq. (11) and numerical integration for establishment of the TEC relationship for processes with $\mathcal{B}e(p)$ marginal distribution. (b) Graphical depiction of the error (simple difference) between the approximate closed-form expression of Eq. (11) and numerical integration. This plot focuses on very low values of p spanning from 10^{-2} to 10^{-9}.

Figure 4. (a) Simulated realization (randomly selected window of 365×30 time steps); (b) comparison between empirical (i.e., simulated) and target theoretical distribution (in terms of probability of exceedance, i.e., $P\{X > x\} = 1 - F_X(x)$); (c) comparison between empirical (i.e., simulated) and target theoretical ACS (in this case $\rho_\tau = \rho_\tau^{fGn}(H = 0.65)$); (d) time series of maximum values of a block with size $k = 365$; (e) comparison of the distribution of maxima over a k-length block (in terms of return period, T), obtained by the models of section 3. Note that the axes of panel (e) are logarithmic.

Figure 5. As in **Figure 4**, but for simulation study #2 (see **Table 1**), which differs at the type of the ACS structure, in this case $\rho_\tau = \rho_\tau^{fGn}(H = 0.8)$. Note that the axes of panel (e) are logarithmic.

Figure 6. Comparison of the distribution of maxima over a 365-length block (in terms of return period, $T = 1/1 - F_{M_k}$), as obtained by different methods, assuming an intermittent base process with $\mathcal{ZIB}rXII(p_0 = 0.75, \beta = 7.07, \zeta = 0.928, \xi = 0.098)$ marginal distribution and ACS structure $\rho_\tau = \rho_\tau^{fGn}(H = 0.65)$. Panel (a) regards estimates (the median, and the 80% confidence intervals) based on 1000 realizations with length $L = 365 \times 50$, while panel (b) with length $L = 365 \times 100$. The axes of both panels are logarithmic.

Figure 7. Comparison (in terms of empirical CDF) of the absolute relative difference [%] between the theoretical the return level x_T (estimated using the AR$(n) + \beta B$ model; see section 3.6) and the estimated return level x_T as obtained by different methods for the estimation of the distribution of maxima over a 365-length block (estimated using each of the 1000 realizations). The results concern an intermittent base process with $\mathcal{ZIB}rXII(p_0 = 0.75, \beta = 7.07, \zeta = 0.928, \xi = 0.098)$ marginal distribution and ACS structure $\rho_\tau = \rho_\tau^{fGn}(H = 0.65)$. From left to right, each column corresponds to $T = 10, 50, 100$ and 1000. Also, the first row corresponds to estimates based on 1000 realizations with length $L = 365 \times 50$, while the second one to $L = 365 \times 100$. For the sake of comparison, this plot includes an additional model for the distribution of maxima, i.e., the i.i.d. model of Eq. (15) combined with a $\mathcal{ZIB}rXII$ marginal (abbreviated in the plot as, ZIBrXII+iid). Note, that the black and magenta lines overlap in several cases.

Figure 8. Panels (a) and (b): mean absolute relative difference [%] between the theoretical distribution of maxima (estimated using the AR$(n) + \beta B$ model; see section 3) and the investigated approaches. Panels (c) and (d): standard deviation of the absolute relative difference [%] between the theoretical distribution of maxima and the investigated approaches. The estimates were derived using 1000 realizations of an intermittent base process with $\mathcal{ZIB}rXII(p_0 = 0.75, \beta = 7.07, \zeta = 0.928, \xi = 0.098)$ marginal distribution and ACS structure

$\rho_\tau = \rho_\tau^{fGn}(H = 0.65)$. Panels (a) and (d) regards realizations with length $L = 365 \times 50$, while panels (b) and (d) with $L = 365 \times 100$. Note, that for the sake of comparison, this plot includes and an additional model for the distribution of maxima, i.e., the i.i.d. model of Eq. (15) combined with a $\mathcal{ZIBr}XII$ (abbreviated in the plot as, ZIBrXII+iid).

Figure 9. Comparison (in terms of empirical CDF) between the estimates of the tail indices (parameter ξ, in both $\mathcal{B}rXII$ and $\mathcal{GEV}$ distributions), obtained by the different methods. The *true* value ($\xi = 0.098$) is illustrated with the vertical orange line. The estimates were derived using 1000 realizations of an intermittent base process with $\mathcal{ZIBr}XII(p_o = 0.75, \beta = 7.07, \zeta = 0.928, \xi = 0.098)$ marginal distribution and ACS structure $\rho_\tau = \rho_\tau^{fGn}(H = 0.65)$. Panel (a) regards realizations with length $L = 365 \times 50$, while panel (b) with $L = 365 \times 100$.

Figure 10. Comparison between the empirical (as obtained by simulation), and theoretical (as obtained by the formulas of section 5.1) probability of a zero value over multiple temporal scales $d \in \{2, \dots, 365\}$, (i.e., $p_0^{(d)} \coloneqq \mathrm{P}\{X^{(d)} = 0\}$) for simulation study #1 (panel a) and #2 (panel b).

Figure D1. As in **Figure 4**, but for simulation study #3 (see **Table 1**), which differs at the type of the ACS structure, in this case $\rho_\tau = \rho_\tau^{CAS}(\beta = 1.5, \kappa = 0.5)$. Note that the axes of panel (e) are logarithmic.

Figure D2. As in **Figure 4**, but for simulation study #4 (see **Table 1**). which differs at the type of the ACS structure, in this case $\rho_\tau = \rho_\tau^{AR(3)} = \sum_{l=1}^{3} a_l \rho_{\tau-l}$, where $a_1 = 0.4, a_2 = 0.2$ and $a_3 = 0.1$. Note that the axes of panel (e) are logarithmic.

Figure D3. As in **Figure 6**, but for simulation study #2 (see **Table 1**). which differs at the type of the ACS structure, in this case $\rho_\tau = \rho_\tau^{fGn}(H = 0.8)$. The axes of both panels are logarithmic.

Figure D4. As in **Figure 6**, but for simulation study #3 (see **Table 1**). which differs at the type of the ACS structure, in this case $\rho_\tau = \rho_\tau^{CAS}(\beta = 1.5, \kappa = 0.5)$. The axes of both panels are logarithmic.

Figure D5. As in **Figure 6**, but for simulation study #4 (see **Table 1**). which differs at the type of the ACS structure, in this case $\rho_\tau = \rho_\tau^{AR(3)} = \sum_{l=1}^{3} a_l \rho_{\tau-l}$, where $a_1 = 0.4, a_2 = 0.2$ and $a_3 = 0.1$. The axes of both panels are logarithmic.

Figures

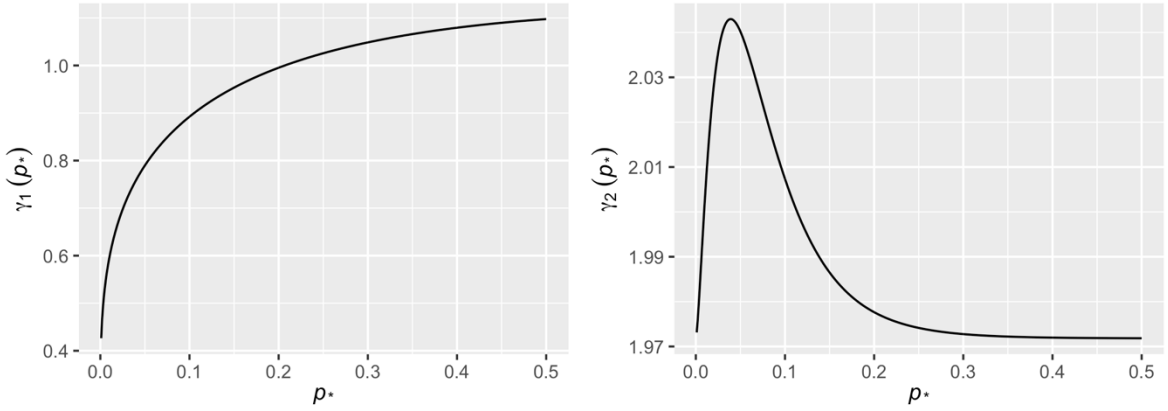

Figure 1.

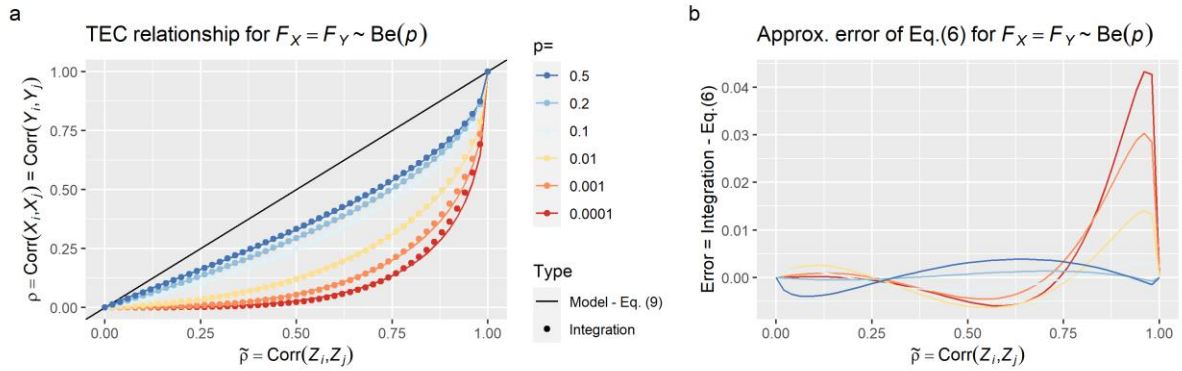

Figure 2.

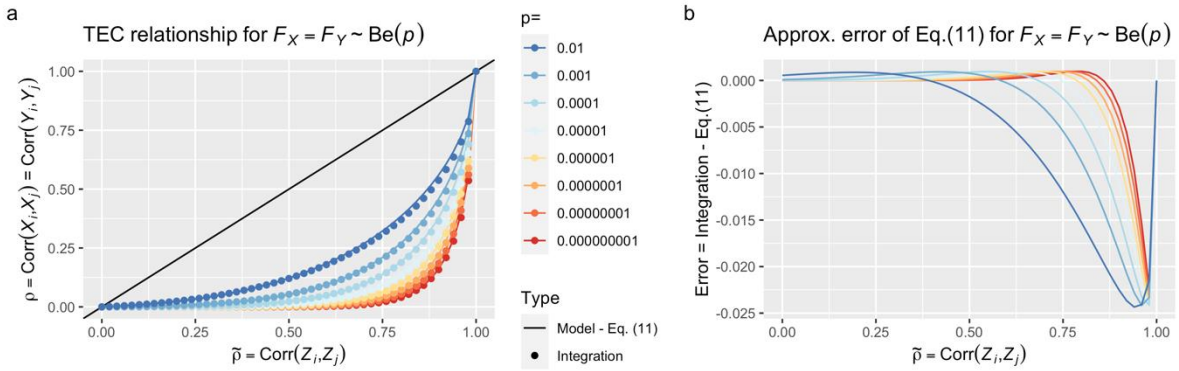

Figure 3.

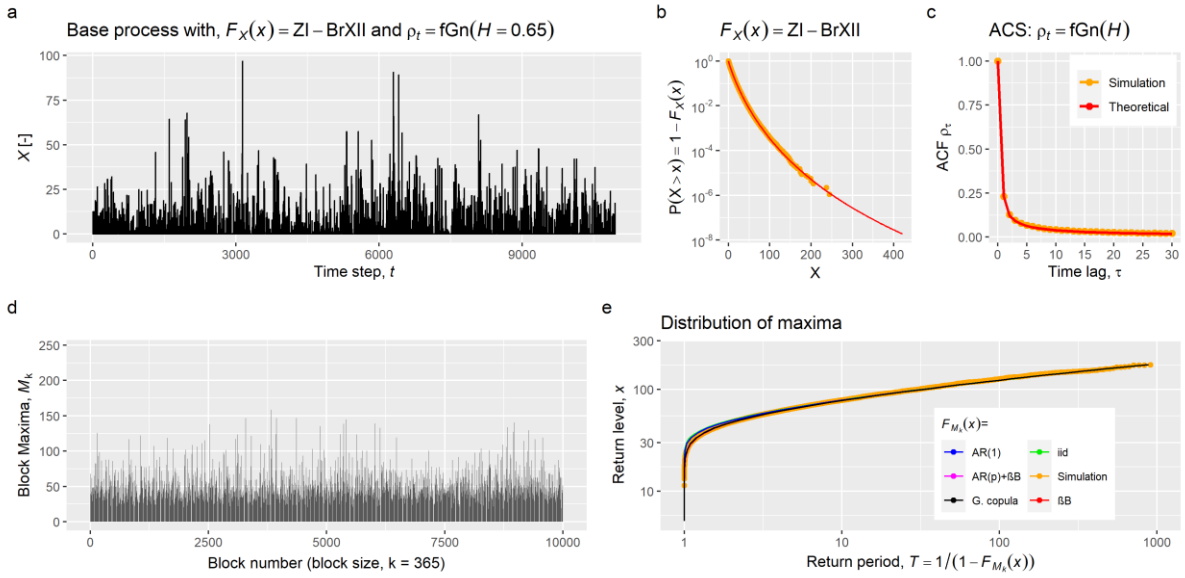

Figure 4.

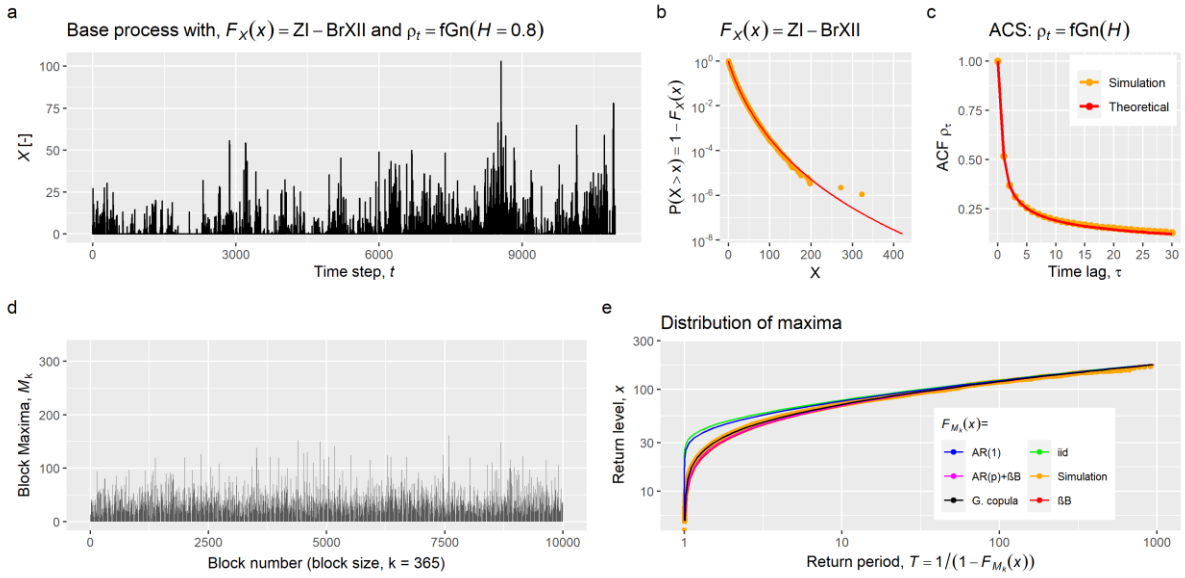

Figure 5.

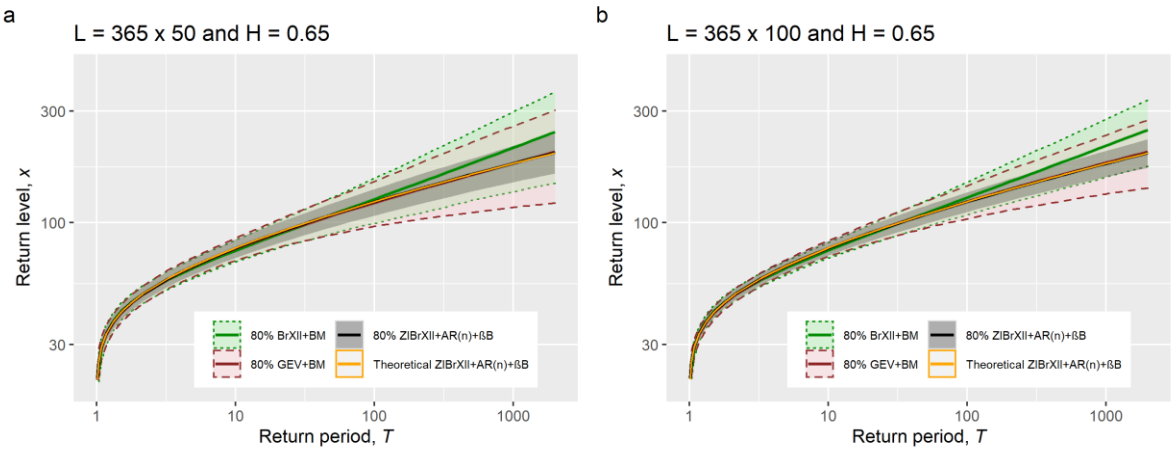

Figure 6.

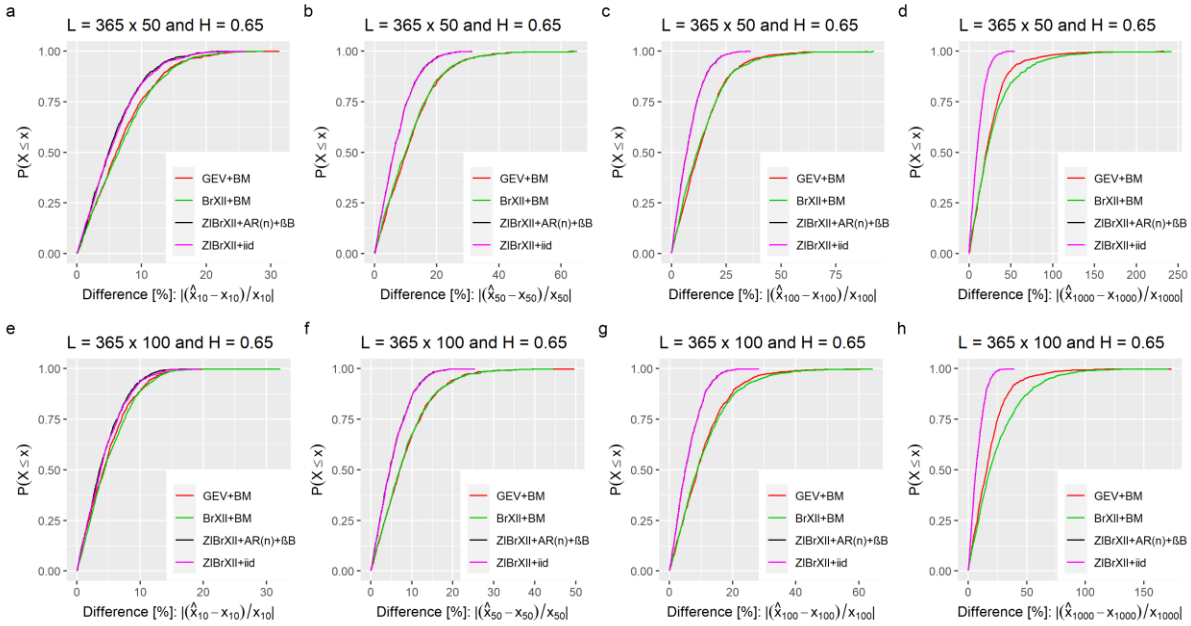

Figure 7.

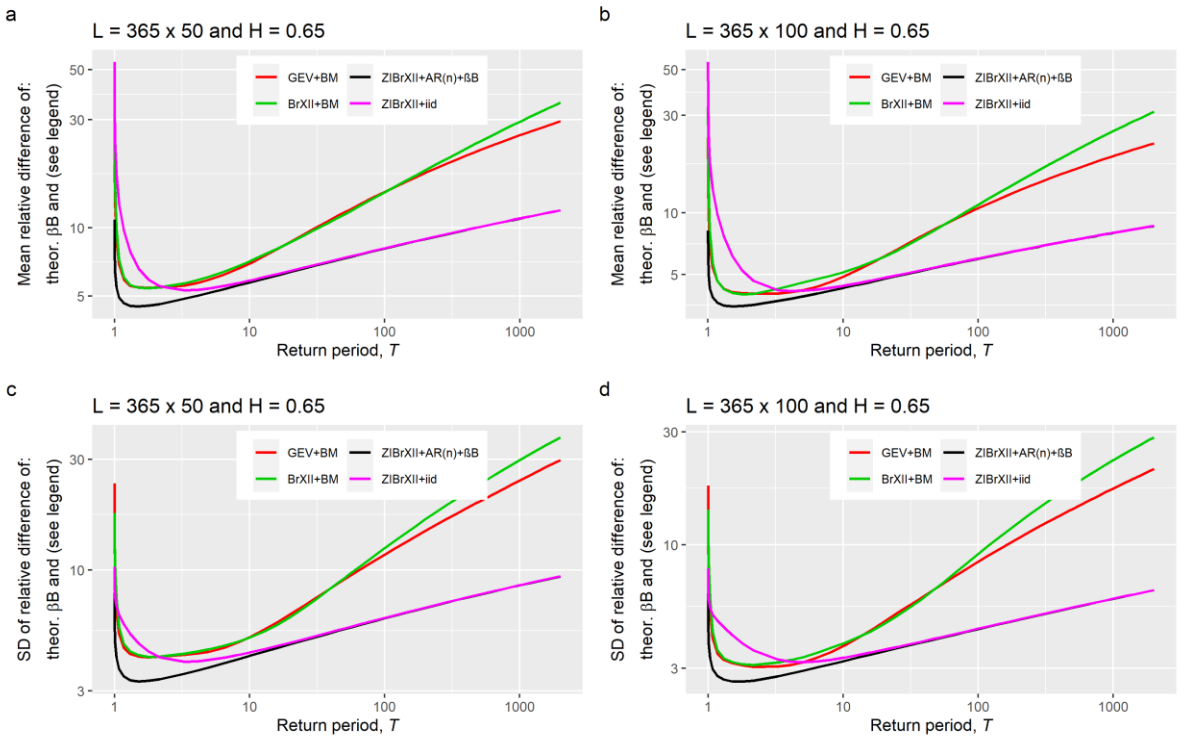

Figure 8.

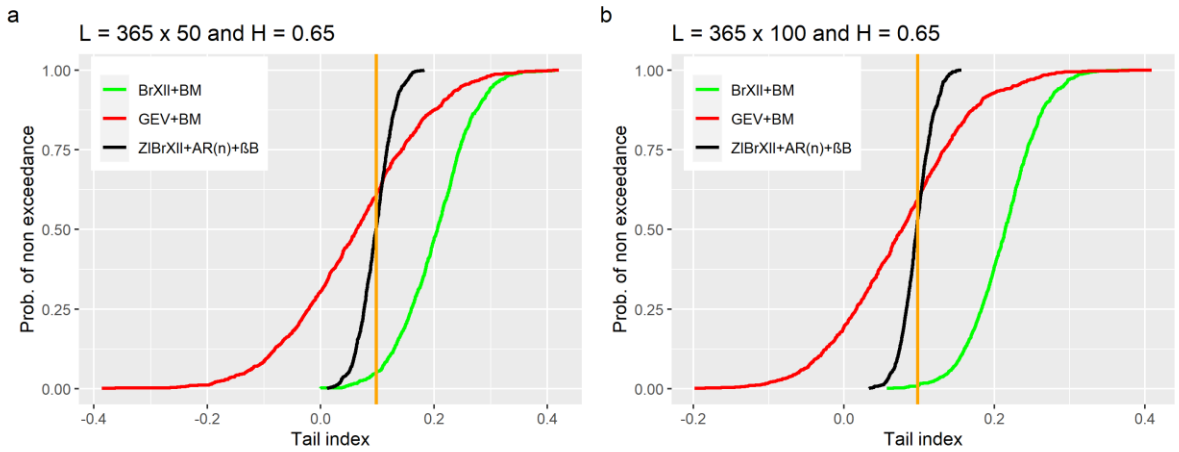

Figure 9.

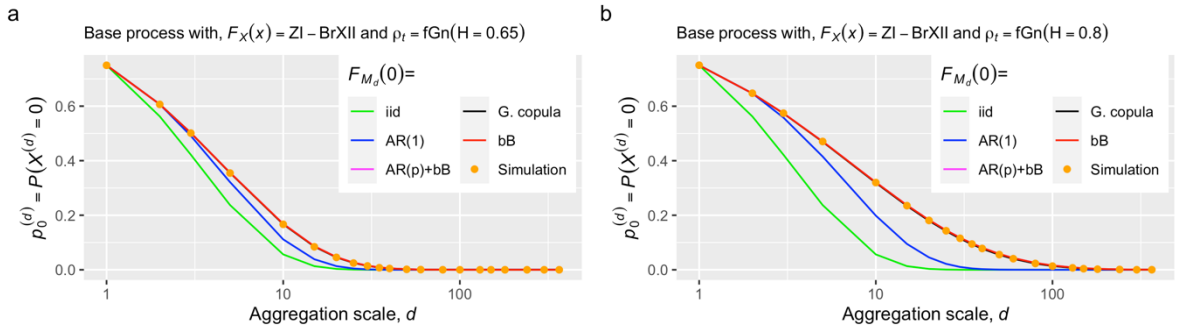

Figure 10.

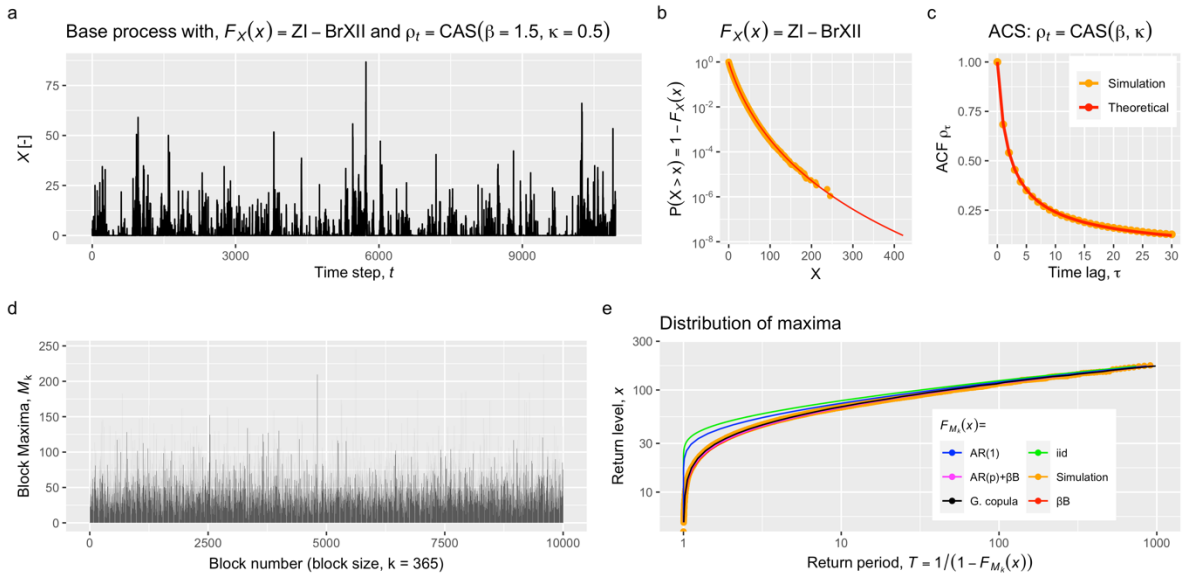

Figure D6.

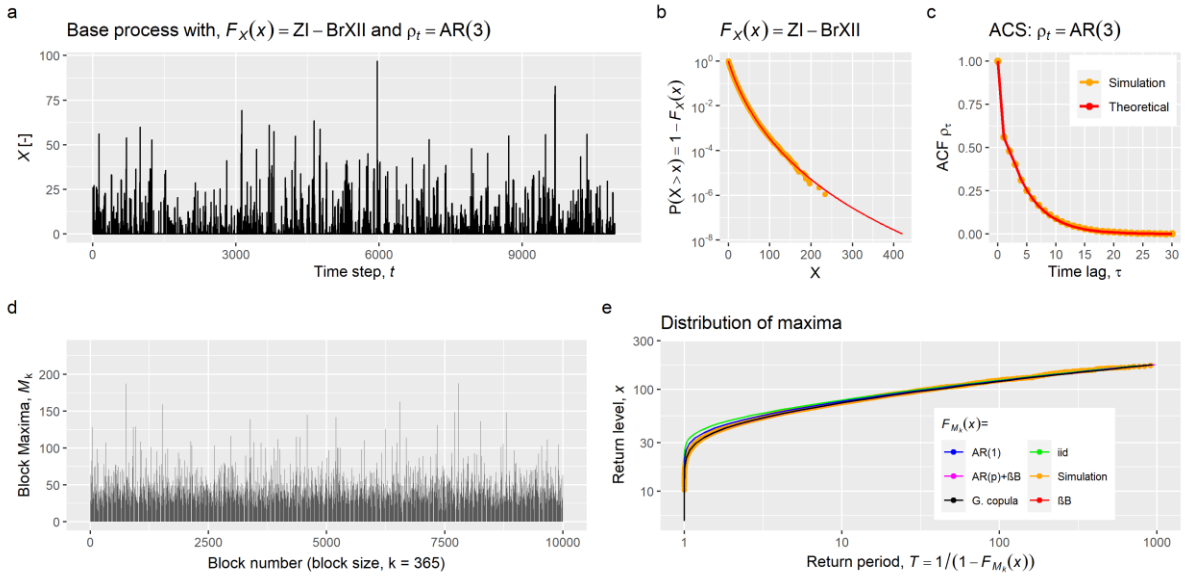

Figure D7.

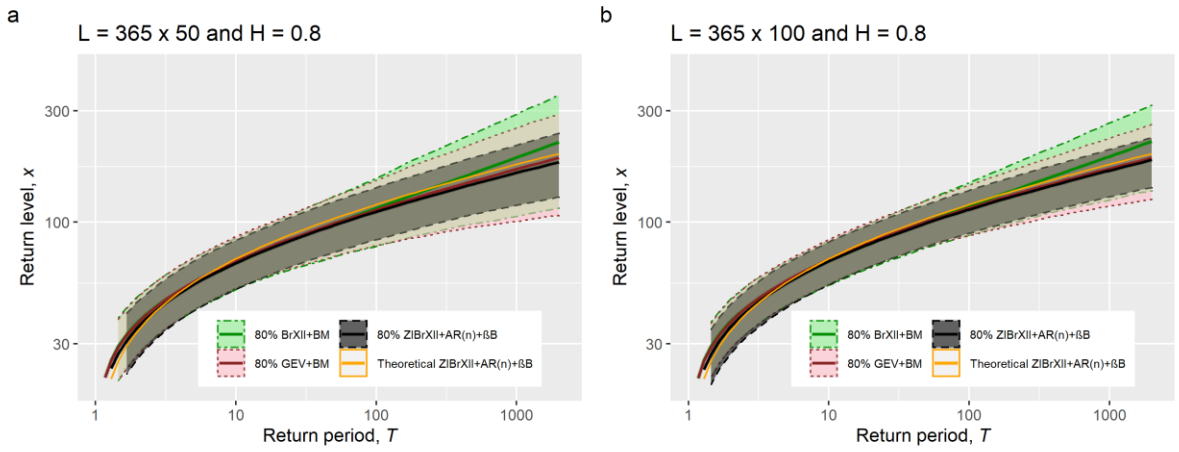

Figure D8.

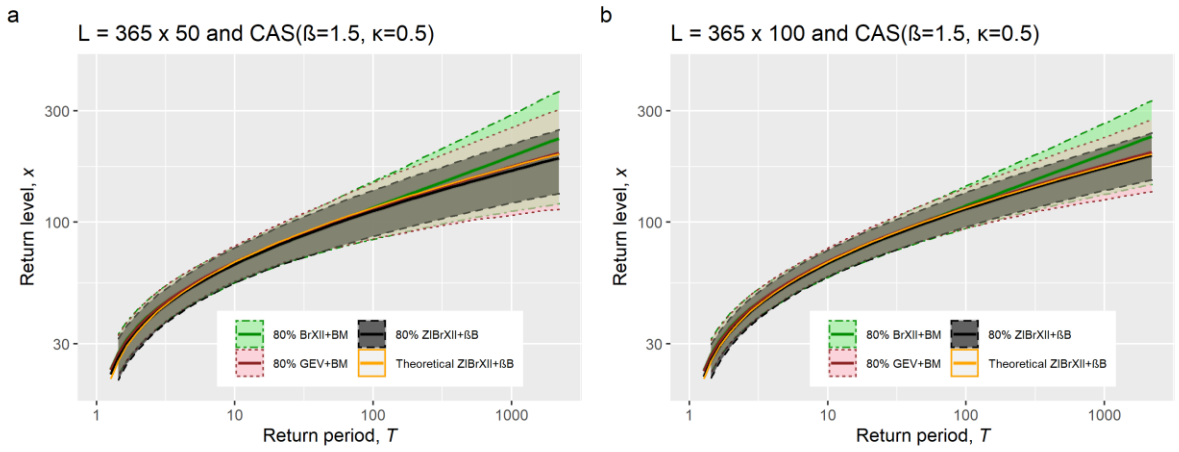

Figure D9.

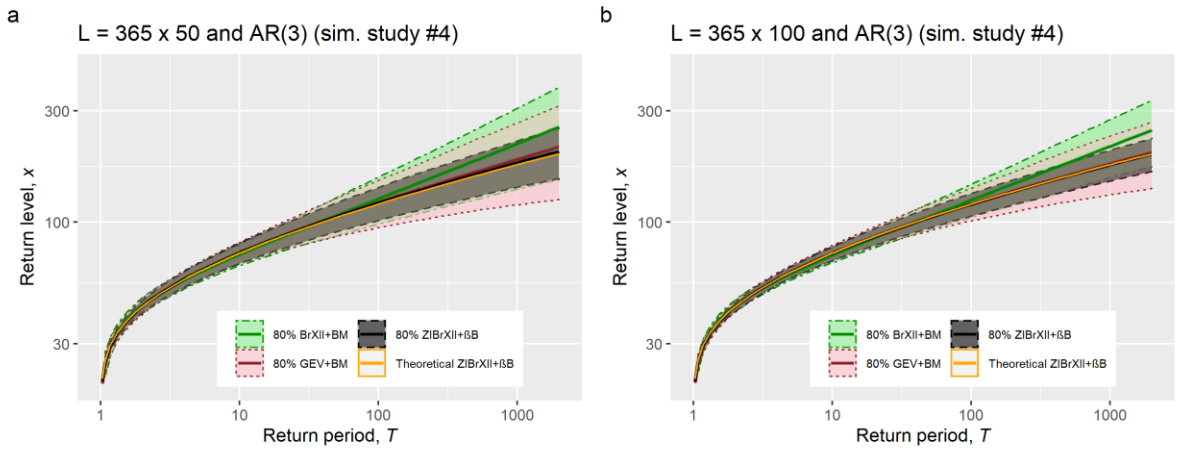

Figure D10.